Management for Professionals

Hao Zhang *Editor*

Models and Methods for Management Science

清華大学出版社
TSINGHUA UNIVERSITY PRESS

Springer

Management for Professionals

The Springer series *Management for Professionals* comprises high-level business and management books for executives. The authors are experienced business professionals and renowned professors who combine scientific background, best practice, and entrepreneurial vision to provide powerful insights into how to achieve business excellence.

Hao Zhang
Editor

Models and Methods for Management Science

清華大学出版社
TSINGHUA UNIVERSITY PRESS

Springer

Editor
Hao Zhang
School of E-commerce and Logistics
Beijing Technology and Business University
Beijing, China

ISSN 2192-8096 ISSN 2192-810X (electronic)
Management for Professionals
ISBN 978-981-19-1613-7 ISBN 978-981-19-1614-4 (eBook)
https://doi.org/10.1007/978-981-19-1614-4

Jointly published with Tsinghua University Press.
The print edition is not for sale in China Mainland. Customers from China Mainland please order the print book from Tsinghua University Press.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Welcome to this book! We are familiar with the term management science, and its role in management is becoming more and more important. With management science, our decision-making basis will be more detailed, clear, and accurate. With the popularization of management science knowledge, especially the rise of computer technology, big data analysis, and simulation technology in recent years, the application of management science in reality has become more and more convenient, the application ideas have become clearer, and the application effects have become more and more obvious.

What is the essence of management science, then? In the book *Data, Models and Decisions* by Prof. Frederick S. Hillier, its definition is given: Management science is a discipline that attempts to aid managerial decision-making by applying a scientific approach to managerial problems that involve quantitative factors.

Thus, management science is first a management discipline, in Hillier's definition, and a collection of knowledge and technology based on science to solve management decision-making problems. Here, it also mentions "a scientific approach," so what exactly is a scientific approach? What approaches are defined as scientific?

In Chinese *Ci Hai* (Word-Ocean), science is a knowledge system of nature, society, and concepts. In French *La Grande Encyclopédie*, science seeks the order in things through classification, and seeks to explain things by revealing the laws governing things. In Japan's *Encyclopedia Nipponica*, science is a systematic academic understanding with objectivity and truthfulness, specific and universal, that is, science is a department with the highest level of knowledge. Albert Einstein believed that, science is "methodical thinking directed toward finding regulative connections between our sensual experiences." From these different definitions, we find the same which is, science is used to reveal regularities. Science means knowledge, concreteness, and reliability. The scientific approach is the more specific and reliable approach to knowledge. The approaches involved in management science are often based on mathematics, computer, social science, system science, etc., and are recognized as reliable after long-term verification.

Management is also supposed as an art. Indeed it is. The science of management and the art of management are two sides of a coin. Only these two sides together make a complete and smart work.

There is another keyword, "quantitative," mentioned in the definition of management science. The research object of management science is "management problems related to quantitative factors," and quantitative analysis methods are the cure to such problems. Quantitative analysis methods in management are mainly used to study the quantitative characteristics, quantitative relationships, and quantitative changes of management phenomena, and to discover laws therefrom. Quantitative analysis methods are often used in conjunction with models, where a model refers to a simplified description of a management phenomenon or process. Models and methods provide tools to more deeply describe various laws in management.

Management science is traditionally called operations research, which is familiar to many. The bud of management science can be traced back to the middle of the eighteenth century, when in the process of the industrial revolution, the pursuit of efficiency and effectiveness prompted managers to pay more and more attention to the application of mathematics and natural sciences. Until the beginning of the twentieth century, Frederick Winslow Taylor et al. founded scientific management theory and published *Principles of Scientific Management*, which is usually regarded as the starting point for the formation of a theoretical system of management science. With the rapid development of the market economy, many managers put forward their views on management from different angles, forming many schools. Among them, some scholars use mathematics, statistics, and other scientific quantitative methods to study and solve management problems from a systematic point of view, so the research on management problems has added a quantitative analysis management school on the basis of the previous qualitative analysis. In 1939, P. M. S. Blackett, a professor at the University of Manchester, established an operations research group, represented by Elwood Spencer Buffa, Horace C. Levencon, et al. which is considered to be the official establishment of the management science school. Since then, management science has gradually become an independent discipline, and research in this field has become more and more active.

The development of management science in China can be traced back to the research on operations research and systems engineering advocated by Hua Luogeng and Qian Xuesen. They put forward a series of theories and methods, such as optimization method, overall planning method, and open complex giant system, to solve many practical management problems. After the 1970s, management scientific research focused on regional planning, project management, forecasting and evaluation management, and other fields, with a combination of various management decision-making theories and methods. After entering the twenty-first century, the application of new technologies and methods such as management information systems, data mining, computer simulation, and artificial intelligence has accelerated the development of management scientific research, and has further intersected and integrated with many social science disciplines and natural science disciplines.

In teaching and research work, I found that many students are willing to learn some quantitative models and methods, and their lack of a systematic understanding made me come up with the idea of writing a book. The models and methods in the field of management science are extensive and profound. Here, I have selected some quantitative models and methods commonly used in economic and management research. This book is rich in content, easy to understand, with representative examples and detailed calculation process, all of which makes a textbook for postgraduates and undergraduates in economics and management in colleges and universities, as a learning material for management science enthusiasts, and as a reference book for management workers. Systematic data analysis methods are not included in this book, as I plan to have a separate textbook on business data analysis in the future. This book does not repeat the knowledge content in the operations research textbook, and focuses on the quantitative models and methods that are not covered in the operations research textbook. Combining this book with operations research, business data analysis, etc. is very beneficial for cultivating high-end management personnel.

This book has 12 chapters. Chapter 1 is an introduction to systems science, including the formation and development of systems science, and the definition, characteristics, and classification of systems. Chapter 2 is decision-making methods, such as assured decision, risk decision, unsure decision, and multi-objective decision. Chapter 3 is forecasting methods, mainly including moving average method, exponential smoothing method, regression analysis forecasting method, and trend extrapolation forecasting method. Chapter 4 is evaluation methods, which introduces DEA, AHP, fuzzy comprehensive evaluation, entropy evaluation method, and set-pair analysis method. Chapter 5 is optimization algorithm, which introduces particle swarm optimization and genetic algorithm. Chapter 6 is system reliability, which introduces the calculation method of system reliability and fault analysis. Chapter 7, game theory, focuses on cooperative and evolutionary games. Chapter 8, management simulation, introduces system dynamics and focusing models. Chapter 9 is complexity science, introducing dissipative structure, synergetics theory, chaos theory, catastrophe theory, hypercycle theory, fractal theory, and self-organized criticality. Chapter 10 is about structural equations. Chapter 11 is Markov chains. Chapter 12 is grey systems theory.

In the process of writing and publishing this book, I have received guidance and help from many scholars, and I would like to express my gratitude. Thanks to my master's supervisor, Prof. Zhang Tienan, who brought me into the world of complexity science. Thanks to my doctoral supervisor, Prof. Shen Jihong, who gave me a comprehensive understanding of systems science and management science. Salute to both mentors! I would like to thank Prof. He Mingke, a well-known logistics scientist in China. His book *Logistics System Theory* has benefited me a lot. Thanks to Prof. Yang Haoxiong, who often discussed quantitative analysis methods with me and gave me a lot of inspiration. Thanks to Prof. Mao Xinshu, who contributed to the publication of this book. I would like to thank Profs. Zuo Min, Wang Guoshun, Ouyang Aiping, Sun Yongbo, Pang Yi, Zhang Yong, Wang Zhen, Li Yekun, Liu Wengang, and Zhang Lulu, the editor of the FLTRP, and

other scholars for giving me a lot of guidance and help. Thanks to my colleagues, Cui Li, Wang Jing, Chen Kai, Hou Hanpo, Lv Junjie, Zhang Jingmin, Du Xinjian, Zhou Yongsheng, Sun Hongxia, Wang Linsheng, Guo Daxin, Li Wendong, Zhang Linlin, Wang Yong, Zhang Yunlai, Guo Chongyi, Deng Chunping et al., it is you who gave me great support and help, and made many valuable and useful suggestions. Thanks to the graduate students who helped me, Xu Shensi, Liu Kuo, Zhu Liyu, Wang Mingkun, Zhang Yiheng, Wang Liling et al., for their hard work in document sorting, numerical simulation, and proofreading. Thanks to the Springer Nature Press for supporting the publication of this book. At the same time, I would like to thank Zuo Yubing, the editor of Tsinghua University Press, and related staff, who have devoted a lot of labor to the publication of this book.

In the process of writing this book, a large number of relevant literature were referenced, including the works of domestic and foreign experts and scholars, teaching materials and papers, and these literature were listed in the references attached, and some are inevitably omitted. Here are my special apologies to the authors that were omitted, and my sincerest thanks to all the authors.

There may be some mistakes in this book, please correct them from scholars and experts. In the future, I will continue to enrich and improve the content of this book, and make teaching materials with PowerPoint. My email id is zhbtbu@126.com, and I sincerely look forward to discussing with friends who are interested in management science.

Beijing, China Hao Zhang

Contents

1 System Science 1
Hao Zhang
1.1 Formation and Development of System Science 3
1.2 Definition and Characteristics of System 18
1.3 Classification of System 20
Further Readings 23

2 Decision-Making Methods 25
Hao Zhang
2.1 Overview of Decision-Making 27
2.2 Certainty Decision-Making 31
2.3 Risk Decision-Making 32
2.4 Uncertainty Decision-Making 38
2.5 Multi-Attribute Decision-Making 42
Further Readings 52

3 Prediction Methods 53
Hao Zhang
3.1 Overview of Prediction 54
3.2 Qualitative Prediction Methods 56
3.3 Quantitative Prediction Methods 59
Further Readings 84

4 Evaluation Methods 87
Hao Zhang
4.1 Overview of Evaluation Methods 89
4.2 DEA (Data Envelopment Analysis) 93
4.3 AHP (Analytic Hierarchy Process) 101
4.4 Fuzzy Comprehensive Evaluation 120
4.5 Entropy Evaluation 127
4.6 Set Pair Analysis 138
Further Readings 160

5 Optimization Algorithm 161
Hao Zhang
5.1 Overview of Optimization Algorithm 162

5.2 PSO (Particle Swarm Optimization) 165
5.3 GA (Genetic Algorithm) 176
5.4 FOA (Fruit Fly Optimization Algorithm) 180
5.5 WOA (Whale Optimization Algorithm) 184
5.6 GWO (Grey Wolf Optimization) 190
Further Readings 206

6 System Reliability 207
Hao Zhang
6.1 Overview of Reliability 208
6.2 Reliability Eigenvector 212
6.3 System Reliability and Calculation 213
6.4 System Reliability Fault Analysis 218
6.5 Application Case 225
Further Readings 233

7 Game Theory 235
Hao Zhang
7.1 Overview of Game Theory 237
7.2 Non-cooperative Game 244
7.3 Cooperative Game 256
7.4 Evolutionary Game 265
Further Readings 272

8 Management Simulation 273
Hao Zhang
8.1 Overview of Simulation 275
8.2 System Dynamics 278
8.3 Multi-agent System 285
Further Readings 298

9 Complexity Science 301
Hao Zhang
9.1 Overview of Complexity Science 303
9.2 Chaos Theory 315
9.3 Catastrophe Theory 334
9.4 Hypercycle Theory 345
9.5 Self-organized Criticality 355
Further Readings 361

10 Structural Equation Modeling 363
Hao Zhang
10.1 Overview of Structural Equation Modeling 364
10.2 Composition of Structural Equation Modeling 366
10.3 Application Case of Structural Equation Modeling 370
Further Readings 381

11 Markov Chain 383
Hao Zhang
11.1 Markov Process 385
11.2 Markov Chain 386
11.3 Classification of Markov Chain Models 392
11.4 Application of Markov Chain Models 394
Further Readings 403

12 Grey Systems Theory 405
Hao Zhang
12.1 Basic Concepts of Grey Systems 406
12.2 Grey Correlation Analysis 407
12.3 Gm (1,1) 410
12.4 Grey Prediction Model 414
Further Readings 419

About the Editor

Hao Zhang, Ph.D. in engineering, is deputy director of science and Technology Department, doctor, Professor of the School of E-commerce and Logistics, Beijing Technology and Business University, doctoral supervisor. Dr. Zhang is the member and special research fellow of China Logistics Association, the senior member of Chinese Society of Optimization and Overall Planning Economical Mathematics, the winner of "youth top individual project" in Beijing and also enjoys membership both in Systems Engineering Society of China and Operations Research Society of China. He participates in International E-Commerce Association (Chinese branch) as a councilor and takes part in China Food Safety Research Center of BTBU, Capital Flow Circulation Research Center and Enterprise Growth and Innovation Research Center as a committee member. His research interests include supply chain and logistics management, digital management, and modeling and simulation of complex systems. Dr. Zhang teaches courses of modeling of supply chain systems, simulation of management system, models and methods of management decision, project management and management economics for undergraduate, postgraduate, MBA, MPACC and international MBA level.

Dr. Zhang has participated in more than ten national and provincial research projects such as National Social Science Fund Project, and Beijing Philosophy and Social Science Project. He also took part in more than 50 projects commissioned by government departments or large enterprises. More than 50 papers were published in journals at home and abroad, and 30 of them were indexed by SSCI, SCI, EI, CSSCI, and CSCD. He has published the following monographs and teaching materials: "city logistics system reliability optimization research", "enterprise strategic synergy mechanism optimization" and "models and methods of management science", "City Commercial Block Logistics Solutions", and "Blue book of China supply chain management". He was awarded the outstanding paper award by China Society of Logistics, Beijing Teaching Achievement award by BTBU, "BaoGong" logistics award, and Humanities and Social Sciences award by Hei Long Jiang Province.

1 System Science

Hao Zhang

Learning Objectives

1. Understanding how system science is formed and developed.
2. Grasping the definition and characteristics of the system.
3. Understanding the classifications of the system.

Literature Review

Chinese Aerospace and System Science

At the beginning of the establishment of China's missile and rocket research system, Qian Xuesen realized that the development of modern complex engineering systems is very different from traditional engineering development. It is necessary to establish a general research organization which combines macro planning guidance, system design control, and management functions. Accordingly, he proposed to establish a missile research institute. Based on national realities, Qian Xuesen adopted engineering cybernetics into his aerospace scientific research practice to realize the standardization and scientification of China's research and management of aerospace. Under the guidance of Qian Xuesen's thought, the theory and methods of system engineering were creatively adopted to build up a scientific organization and management system of Chinese aerospace, which was realized through creating a new role chief designer, and administrative directions, as well as implementing matrix management, project management, and "zero defect" management comprehensively. This creative move also helps to enrich and improve the core concepts of system engineering such as overall optimization, system coordination, environmental adaptation, innovative development, risk management, and optimization assurance.

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_1

Under the guidance of these systematic engineering methods, an efficient and systematic missile and rocket project development organization and management system was established. Thousands of researchers, a large number of departments and units, countless equipment, huge amounts of funds, strict requirements and a wide variety of materials and equipment, were coordinated and organized to be invested in the research, design, trial production, test and production process of this project in an orderly manner.

Although China's systems engineering still falls into the category of engineering systems engineering, it also includes factors such as people, teams, management, and culture. A typical example of systems engineering thinking with Chinese characteristics is the "Two General System" of China's Aerospace. The "Two General System" was established in accordance with China's system at that time. The chief division guarantees technology and business, and the chief commander is responsible for administrative directions and logistics support. Factors including people, team, and culture are well planned and managed in this system.

In 2011, China Aerospace Engineering Consulting Center officially changed its name to China Aerospace Academy of Systems Science and Engineering (CAASSE). The Academy is the main consulting research and information technology research and development institution for China's aerospace development. It also plays important roles as the group's headquarters think tank and development strategy staff, the information construction department, the information system technology and operation guarantee department, the common software development department and the software industry promotion department. In addition, the Academy has the technical reserves and engineering experience of developing large-scale application software systems, and has strong capabilities to develop various information management systems. The Academy also accumulates rich experience in in enterprise portal systems, integrated office management, information management application systems, and enterprise ERP.

In light of increasingly complex challenges, systems and scientific engineering methods are an important guarantee for the success of China's aerospace engineering projects. System complexity theory provides new insights into the solution of complex aerospace project organization and management problems. At the same time, through the understanding of organization complexity management in the practice of aerospace engineering, we can gradually summarize the social organization system engineering method with the goal of forming overall creativity, and ultimately promote the development of system science and management science.

References

[1] Hu Shixiang, Zhang Qingwei. China's manned spaceflight project: a model of successful system engineering [J]. *Aerospace China*, 2004, (10): 3–6.

[2] Guo Baozhu. The application of the theory and method of system science in the management of aerospace projects [J]. *Journal of Astronautics*, 2008, 29 (1): 29–33.

1.1 Formation and Development of System Science

Systematic thoughts have emerged long ago in ancient times. For example, Aristotle's idea that "the whole is greater than the sum of parts" represented an embryonic systematic view of holism. After that, Democritus and Plato's ideas, as well as some eastern classics such as the *Huangdi Neijing* (Yellow Emperor's Inner Canon) and the *Tao Te Ching* also reflected the concept of system. It was not until the 1920s that systems science gradually formed a theoretical system. From 1924 to 1928, the American Austrian theoretical biologist Ludwig von Bertranffy (1901–1972) published several articles to talk about the ideas of system theory. He put forward the concept of organism in the field of biology, and emphasized that the organism must be studied as a whole or system in order to discover the principles of organization at different levels. In the *Theoretical Biology* published in 1932 and the *Modern Development Theory* published in 1934, Bertranffy proposed using mathematical models to investigate biological methods and the concepts of organism system theory. He introduced the mathematical concepts of coordination, order, and objective into the study of organisms, which further develops into, three basic points of view in the study of living organisms, namely, system viewpoint, dynamic viewpoint, and hierarchical viewpoint. His arguments thus became the foundation of general system theory. Afterwards, with the continuous enrichment of system science theories, an increasing number of experts such as Kenneth Boulding, Paul A. Samuelson, A. D. Hall, etc. devoted themselves to the study of system science. In addition, more and more subject knowledge such as mathematics, physics, logic, cybernetics, information theory, sociology, etc. has been gradually integrated into system sciences. At the same time, system science has been widely used in a variety of fields such as biology, management, economics, medicine, engineering, etc (Fig. 1.1).

Fig. 1.1 Ludwig von Bertalanffy

1.1.1 The Development Process of System Science

The development of system science can be divided into the following three stages:

(1) 1940s–1950s, the formation and proposition of system science
Early systems science includes general system theory, information theory, and cybernetics. These three theories are commonly known as the "old three theories", and they are the core knowledge system in the early formation stage of system science.

① General system theory. The general system theory was founded by the American Austrian theoretical biologist Ludwig von Bertalanffy. In 1937, Bertalanffy proposed the concept of general system theory at a philosophical seminar at the University of Chicago. By pointing out the limitations of its analytical methods, he challenged the theories of mechanism and reductionism formed since the Renaissance, and put forward the systematic methods. General system theory involves many basic concepts, such as system integrity, relevance, order, direction, and purpose. Bertalanffy also pointed out that the general system theory took an approach of organism, and thus we cannot adopt the analytical methods of mechanism in the field of the inorganic theory to solve this kind of problems. In addition, he suggested that we cannot simply combine the individual parts into a whole. Instead, we need to take the interrelationship between individual systems and the whole system into consideration so that we can better understand the behavior of each part and the whole. In the 1950s, Bertalanffy, together with economist Kenneth Ewart Boulding, biologist Ralph Gerard, and biomathematician Anatol Rapoport, established the International Society for the Systems Sciences (ISSS), which attracted a large number of scientists and had a profound influence in the Western world. These four persons are also known as the "founding fathers of general system theory movement".

② Information theory. Information theory came into being with the development of general system theory. Since 1948, the American mathematician Claude Elwood Shannon (April 30, 1916–February 26, 2001), successively published *The Mathematical Theory of Communication* and *Communication in Noise* with a focus on the communication process. He investigated the definition, measurement, conversion, transmission, reception, processing, storage and control of information in the communication process, using mathematical statistical methods to describe the transmission and reading of information in terms of quantity. Furthermore, Shannon put forward the theory of information amount and coding theory, and established information entropy formula. These two articles are the foundation of modern information theory. Meanwhile Shannon further expanded the scope of information theory, enabling it to deal with generalized information such as pragmatic information, semantic information and non-probabilistic information (Fig. 1.2).

③ Cybernetics. In 1948, the American mathematician Norbert Wiener (1894–1964) combined the research results of related control problems in physiology,

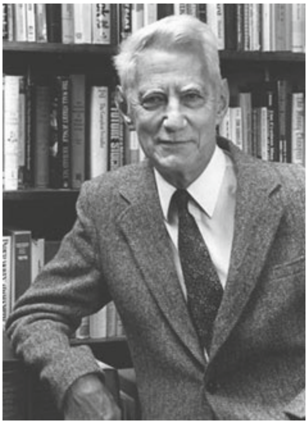


Fig. 1.2 Claude Elwood Shannon

biology, behavioral sciences and other disciplines with the servo system theory of machine control principles, based on which he published *Cybernetics* and *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, and finally established the science of cybernetics. From the perspective of system science, he summarized concepts, principles, methods and models with general significance. Cybernetics was considered as a comprehensive and basic subject, which was different from general system theory. It only focuses on information and only studies the behavior of the system. The development of cybernetics was closely integrated with complexity research, and had gradually been applied to many fields such as engineering control, neural control, economic control, military control, biological control, population control, ecological control, and social control to deal with complex system control problems (Fig. 1.3).

(2) 1960s–1970s, the great development of system science

Dissipative structure theory, synergetics, and catastrophe theory are three subdisciplines of system science that have been established and developed rapidly since the 1960s and 1970s, collectively called the "new three theories". These three disciplines and chaos theory constitute the core of self-organization theory, which greatly enriches the theoretical system of system science.

① Dissipative structure theory. In 1969, Professor Ilya Prigogine, a Belgian physical chemist and representative of the Brussels School, established the total entropy change formula by absorbing the basic ideas of general system theory and investigating the self-organization phenomenon of complex systems, which further developed into a set of unique self-organization theories—dissipative structure theory. Dissipative structure theory was the study of the

Fig. 1.3 Norbert Wiener

characteristics, stability, and evolution of dissipative structures. It extended statistics, thermodynamics, and physics from equilibrium to non-equilibrium, and establishes an orderly evolutionary state in time, space, and function. He was well-known for his research including *Structural Dissipation and Life*, *Time, Irreversibility and Structure*, and he won the Nobel Prize in Chemistry in 1977 (Fig. 1.4).

② Synergetics. In 1969, German physicist Hermann Haken (1927–) proposed the concept of "synergy". In 1988, Harken published the book *Advanced Synergetics* based on the laser theory and founded the science of Synergetics. Synergetics explored how systems work simultaneously. Haken summarized a set of self-organization theories and methods of processing laser systems

Fig. 1.4 Ilya Prigogine

through drawing upon the results of equilibrium and non-equilibrium phase transition theories to describe self-organization phenomena and combining the findings of laser theory. He discovered the laser was in an orderly state when it is far from the equilibrium state and within the equilibrium state. In addition, he found out that the interaction between the various subsystems within the system can spontaneously produce a stable and ordered structure. Synergetics proposed three principles: dominance principle, instability principle, and order parameter principle. It provided a profound insight into the stability analysis, the establishment of equations, the optimization of models, and the description of self-organization mechanisms are more profound. Synergetics also took an approach of classification and analogy to describe the common law of the transition from disorder to order in various systems and motion phenomena, and prepared us with the theories and methods to deal with self-organization problems in the system (Fig. 1.5).

③ Catastrophe theory. In 1972, the French mathematician René Thom (1923–2002) published the book *Structural Stability and Morphogenesis*, in which he put forward the theory of catastrophic phenomena in morphogenesis, namely catastrophe theory, by describing sudden shifts in behavior arising from small changes in circumstances in the real world. Thom maintained that in addition to smooth, gradual, and continuous movement, there were also large-scale transition phenomena in nature and human social activities. Catastrophe theory mainly concentrates on discontinuous results caused by the continuous action, and provides a broader mathematical framework for studying discontinuities, transitions and sudden deterioration. One characteristic of the catastrophe phenomenon is that slight changes in external conditions lead to drastic changes in the macroscopic state of the system, which only exist in nonlinear systems. Built upon Berta Langfei's arguments, René Thom employed the theory of

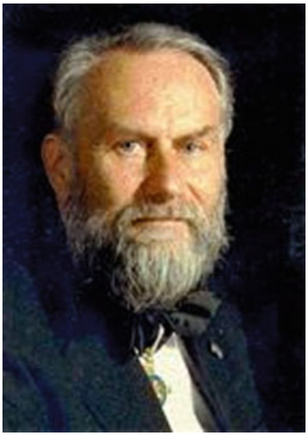


Fig. 1.5 Hermann Haken

Fig. 1.6 Rene Thom

structural stability and abandoned the characteristics of discontinuities to ultimately develop a unified way to solve the problem of system evolution in form formation in. Catastrophe theory is the mathematical foundation and tool of dissipative structure theory and system science. It takes a dialectical approach to deal with contradictions such as instability and stability, discontinuity and stability, sudden change and gradual change, and expresses it in a concise model (Fig. 1.6).

④ Chaos theory. In the 1960s, mathematicians Arnold, Kolmogorov and Mosher examined the stability of the system and brought forward the KAM theorem (KAM theorem). Meteorologist Lorenz analyzed the dissipative dynamic system and obtained a system of differential equations, the well-known Lorenz equation. He also discovered the first strange attractor in the study of chaos—the Lorenz attractor, which provided an important model for the study of chaos In 1975, Chinese American scholar Tien-Yien Li and his mentor Yorke J., a mathematician, published a famous article *Period 3 Means Chaos* in the US *Math Monthly* to reveal the evolution process from order to chaos. They established the famous Li-Yorke theorem and described the mathematical characteristics of chaos. In 1981, F. Takens et al. proposed an experimental method to determine singular attractors based on Whitney's topological embedding theorem. Since the 1990s, more and more attention has been paid to the application of chaos science. Chaos science provides a method to unify two mutually incompatible description systems, namely deterministic description and probability description), which thus enables an objective and in-depth study on system evolution. The study of chaos theory has been fruitful by taking an interdisciplinary approach to incorporate other disciplines, such as mathematics, physics, astronomy, chemistry, information science, geology, biological sciences, economics.

⑤ Systems management theory. With the development of general system theory, some scholars have applied related research findings to the study of management, which further developed into the systems management theory. American management scientists Richard A. Johnson, Fremont E. Kast and James E. Rosenzweig were the representative systems management theorists. In 1963, they published *The Theory and Management of Systems*, the foundational work of systems management theory together. In 1970, Kast and Rosenzweig jointly published *Organization and Management: A System and Contingency Approach*, providing a comprehensive discussion on system management. Systems management theory believes that an enterprise is an integrated system is composed of personnel, materials, machinery and other resources under specific goals. Personnel are dominant element while others are relatively passive. Businesses enterprise are open systems with dynamic interactions with the elements of their surrounding environment such as natural environment, customers, competitors, suppliers, governments. The internal and external information feedback mechanisms continuously and automatically allow businesses to adapt to environmental changes and the needs of their own growth. System management has four characteristics: First, it is goal-centered, always emphasizing the objective achievement and objective effect of the system; second, it revolves around the entire system, stressing the optimization of the entire system rather than the optimization of the subsystems; Additionally, system management is responsibility-oriented, which means each manager must be assigned certain tasks, and the corresponding input and output will also be measured; Ultimately, it is people-centered, and each employee is assigned to do challenging tasks and paid according to their performance. The systems management theory advocates that the goal of the system should be established first when conducting a system analysis. Moreover, the overall interests of the system should be prioritized rather than and the partial interests, and the partial interests should also be subordinated to the overall interests. Also, both current interests and long-term interests should be taken into consideration. The theory adopts, a mixed method of quantitative and qualitative analysis to investigate key issues. The system management school also uses systematic methods to study organization theory and believes that there are three subsystems that perform different tasks in a complex management system, namely the system-strategy subsystem, coordination subsystem and operation subsystem.

(3) The prosperity of systems science after the 1980s

The research of nonlinearity and complexity has emerged since the 1980s, promoting the prosperity of system science. Nonlinear science is the study of non-linear common problems in various scientific fields, which suggests that everything is a system, and the internal structure, system function as well as evolution process are the result of interaction. It not only discusses the balance of the system, but also discusses the problems of oscillation, chaos and bifurcation, which lays a scientific foundation for studying the complexity of the system.

Complexity research started in the mid-1980s. In 1984, scientists from different fields, led by Gell-man, Arrow, and Anderson, gathered together to establish the Santa Fe Institute in New Mexico, USA, to study the complexity of cross-discipline. It refers to a series of issues including "the chemical reactions that lead to life before the emergence of life on earth, the evolution of life itself, the functions of individual living organisms and ecological communities, the operation of life subsystems such as the mammalian immune system and the human brain, the evolution of human culture, and computer hardware, the functions of software, the evolution of various economic systems on earth, the evolution of organizations and communities, etc. Scientists argues that ecology, genetics, embryos, nerves, economy, and computer networks are complex adaptive systems. They believe that there are universal laws that control the behavior of complex adaptive systems.

In 1987, Chinese Professor Qian Xuesen proposed an open complex giant system. He found that there were both open and closed complex giant systems. For instance, the reform and opening-up in China is a typical example of openness while the Chinese closed-door policy in Qing dynasty as well as small peasant economy fall into the category of closedness. He integrated Marxist philosophy into systematic research, and employed materialism and dialectics to develop a new approach of systematic and sustainable research. Qian Xuesen argued that "We must try our best to integrate the knowledge of many people to form a whole thing." Combining system science, statistical data and expert experience, and taking electronic information technology as the core, Qian adopted a comprehensively integrated method from qualitative methods to quantitative approaches, to put forward scientific methods with advanced technology. The integrated method was a method that includes scientific, perceptual, empirical, rational, qualitative and quantitative knowledge. Although there were still many difficulties in theoretical analysis and application, the research was essential in terms of its implications on future research regarding thinking science, artificial intelligence, and brain science. He also pointed out that the study of open and complex giant systems not only needed logical thinking and scientific methods, but also required the combination of logical thinking and image thinking as well as the combination of theory and practice to give full play to the subjective initiative of people. Moreover, the study should be substantially tested in the understanding and practice to be able to achieve the leap from quantitative change to qualitative change. To sum up, the complex system theory emphasizes the dynamics, openness, nonlinearity, internal randomness, and autonomy of the system, and dialectically unifies determinism and non-determinism, reductionism and holism to form a kind of Comprehensive research program.

1.1.2 The Main Schools of System Science

Many schools of system science came into being during the formation period of systems science since scholars came from various fields, and their research objects, perspectives, methods, and tools were different, as shown in Table 1.1.

Table 1.1 Representative schools of system science theory

School	Representative	Masterpiece	Main research content
School of mathematical systems	M. Mesorovic, and Y. Takahara	*General System Theory: Mathematical Fundamentals*, *Abstract System Theory*	The mathematical system theory was applied to management science. It proposed the mathematical organization theory (MOT), and the target search system model as well as general organization model
School of system analysis	Arthur D. Hall	*System Engineering Methodology*	It constructed the "three-dimensional space structure" of time dimension, logic dimension and knowledge dimension, and summarized the general process of system engineering with an emphasis on clear goals. Its core content is optimization
School of operations research	Russell Ackoff	*Operations Research*, *Social Redesign*, *Beat the System: Rely on Creativity to Win Bureaucracy*	It mainly refined some general operational research problems that appeared in events such as production and management, and then adopted mathematical methods to solve them. Ekov created the management theory of the "social system era"
School of system dynamics	Jay W. Forrester	*Industrial Dynamics*	It was a quantitative method based on feedback control theory and computer simulation technology, aiming to examine complex social economic systems
Evolutionary school of complex systems	Peter Allen	*Evolution and Creativity Modeling of Complex Systems*	It created a new model to show how the dialogue between the creation mechanism of diversity and the competitive constraints of the material world creates an ecological structure
School of dissipative structure	I. Pringogine	*Unbalanced Statistical Mechanics*, *Self-Organization in Non-equilibrium Systems*, *The End of Certainty*	It proposed that openness was a prerequisite, "non-equilibrium is the source of order", as well as the view that fluctuations lead to order, which was of great significance to correcting the concept of "balanced and orderly"

(continued)

Table 1.1 (continued)

School	Representative	Masterpiece	Main research content
School of synergetic theory	Hermann Hakenn	*Synergy: A Collaborative Science*, *Synergy: An Introduction to Non-equilibrium Phase Transitions and Self-organization in Physics*, *Chemistry and Biology*, *Advanced Synergetics*	It adopted evolution equations to investigate various non-equilibrium steady states and instabilities of cooperative systems
School of catastrophe theory	R. Thom	*Dynamics of Morphogenesis*, *Topological Models in Biology*, *Structural Stability and Morphogenesis*	It concentrated on the phenomena and laws of sudden changes in system discontinuity, and it could effectively describe the phase transition process of system state changes
School of systems management	R. A. Johnson, F. E. Kast and J. E. Rosenzweig	*System Theory and Management*, *Organization and Management: System and Contingency Method*	The theory examined problems from a system point of view. In business management, it was necessary to adapt to the internal and external conditions of the enterprise as there is no fixed, universally applicable, and the so called best management theories or methods

These different schools of system science were all based on systematic thinking, and their research contents were interconnected and supplemented, forming a diversified systematic scientific system.

1.1.3 Development of System Science in China

The rise of China's system science research can be traced back to the 1950s. Qian Xuesen and Xu Guozhi introduced operations research into China and formed an operations research group at the Institute of Mechanics, Chinese Academy of Sciences. Since then, under the vigorous promotion of Qian Xuesen, Hua Luogeng, Xu Guozhi, Song Jian and others, China has gradually developed a wave of systems science research, and a large number of outstanding systems science researchers have emerged.

Qian Xuesen (1911–2009) (Fig. 1.7), was widely recognized as a world-renowned scientist, an aerodynamicist, the founder of China's manned spaceflight, an academician of the Chinese Academy of Sciences and Chinese Academy of Engineering. He was awarded the "Two Bombs and One Satellite Merit Award" by the Chinese government. He was also known as the "Father of China's Aerospace", "Father of Chinese Missiles", as well as the "Father of China's Automatic Control" and "King of Rockets." Qian's representative works regarding system science included *On System Engineering* and *Creating System Science.*

Qian Xuesen graduated from the School of Mechanical and Power Engineering at Jiaotong University in 1934, and he was once a professor at the Massachusetts Institute of Technology and California Institute of Technology. In 1955, he returned to China to initiate and lead the analysis and research of China's defense system.

Qian Xuesen was a great scientist in our country. During more than 70 years of scientific research, his research touched upon the fields of science, engineering,

Fig. 1.7 Qian Xuesen

technology and philosophy. He has made pioneering and outstanding contributions especially in regards to the systems science that intersects, utilizes and integrates different disciplines and fields.

Qian Xuesens research on systems science can be roughly divided into three stages. During the first stage, he mainly investigated natural science and technology in the United States for 20 years, and his research revolved around engineering cybernetics. However, the publication of *Engineering Cybernetics* in 1954 indicated that his research focus shifted from natural science to systems science. The second stage referred to the next thirty years of Qian's research concerning system engineering. After he returned to China, he had been concentrating on the development of Chinas rockets, missiles and aerospace. His thoughts and methods had been well practiced in the successful research and development of China's "Two Bombs and One Satellite". The third stage was systematic research in his later years. Qian was fully committed to academic research at that time, and ultimately created a systems science system and a modern science and technology system. Qian Xuesen has made great contributions to many levels of the system science structure, the most important of which was the creation of system science and the proposal of open complex giant systems and their related methodology. He introduced complex giant system science into the system science system, enriching the system of scientific research to a large extent.

① The establishment of systematics and comprehensive integration methods. Soon after Qian Xuesen proposed to establish systematics, his related academic research was carried out in the form of "systems seminars". Through these academic discussions, he first put forward new standards of systems classification, and introduced the thought and concept of control into system science based on the different levels of complexity. Furthermore, Qian Xuesen concretized the system theory method and proposed a comprehensive integration method, which organically combined the expert system, the information and knowledge system, and the computer system to form a highly intelligent human–computer combination system. The system could therefore deal with the problems of complex systems, complex giant systems and social system effectively.

② Creating a new field of science and technology for complex giant systems. Qian Xuesen pointed out that biological systems, human body systems, geographic systems, social systems, and galaxy systems were all representatives of complex giant systems while social systems, as the most complex system, was defined as a special complex giant system. These systems were open, and they constantly exchanged material, energy and information with the external environment, so they were also called open complex giant systems. Qian Xuesen not only came up with the related concepts of open complex giant systems, but also proposed methods to deal with the system, thus creating the field of science and technology of complex giant systems.

As a systems scientist, Qian Xuesen's research on systems science involved different levels and branches, and achieved some major research results. His achievements and contributions to the field of systems science not only fully embodied his scientific innovation spirit, but also deeply demonstrated his scientific ideas and methods. He is a veritable scientific master.

Hua Luogeng (1910–1985) was known as a world-renowned mathematician, a member of the Chinese Academy of Sciences, a foreign member of the National Academy of Sciences, a member of the Third World Academy of Sciences, and a member of the Bavarian Academy of Sciences of Germany. He was one of China's most influential mathematicians in the world and was listed as one of the 88 mathematics greats in current world in the Chicago Museum of Science and Technology. Hua Luogeng established the optimization method and the overall planning method, which were both new disciplines of system science. The optimization method explored how to quickly find the optimal solution with a small number of trials. The overall planning method was a mathematical method of arranging the work process, mainly used for plan management and schedule management. Reflected in the form of network diagrams, this method designed and optimized the work plan to select the best work plan. As a result the method helped to achieve the predetermined goal and obtain the best economic benefits. Hua Luogeng has been committed to promoting the "Double Method" for 20 years, and personally led students to some enterprise factories to promote and apply the "Double Method". In March 1981, the "China Optimization Method, Overall Planning Method and Economic Mathematics Research Association" (a national society under the Chinese Association for Science and Technology) was officially established, and Hua Luogeng was the first chairman of the board. The promotion of the optimization method and the overall planning method supported the system and science to serve the production and life in reality more effectively, and made outstanding contributions to the construction of the national economy.

Xu Guozhi (1919–2001) was a well-known system engineering and operations research expert in China, and an academician of the Chinese Academy of Engineering. He graduated from Shanghai Jiaotong University in 1943, after which received a Master of Science degree from the University of Kansas in 1949, and a Ph.D. degree in philosophy from the University of Kansas in 1953. Xu was a researcher of the Institute of Systems Science of, Chinese Academy of Sciences, the former chairman of the Chinese Society of Systems Engineering, as well as one of the main founders of China System Engineering. He started his research on systems engineering in the 1980s and participated in the preparations for the establishment of the Institute of Systems Science of the Chinese Academy of Sciences and the Department of Systems Engineering and Mathematics of National University of Defense Technology. Additionally, he initiated the Chinese Society of Systems Engineering and the first system engineering publication, *System Engineering Theory and Practice*. He also trained a large number of specialized professionals in the field of operations research and systems science, which played an important role in promoting the application of systems engineering in China's economic development, national defense construction, and decision analysis.

Dai Ruwei (1932–) was elected as an academician of the Chinese Academy of Sciences in 1991. He is currently the chairman of the Chinese Society of Automation and the convener of the Automation Discipline Review Group of the Academic Degrees Committee of the State Council. He has been employed as an adjunct professor or honorary professor in nearly 30 universities such as Tsinghua University, Shantou University, University of Science and Technology of China., and the chief scientist of the "System Medicine and Chinese Medicine Scientific Research Center" jointly established by the Institute of Automation of the Chinese Academy of Sciences and Qingdao University. In the 1960s, he solved the numerical calculation problem by using maximum value to understand the fastest control and final value control. In the 1970s, he was the first to carry out pattern recognition research in China. In the early 1980s, he combined statistical pattern recognition with syntactic pattern recognition, and proposed semantic and syntactic methods. In the late 1980s, he took the lead in conducting research on artificial neural networks in China. In the 1990s, he was engaged in the research of open and complex giant systems and the methods of dealing with such systems. He adopted the concept of integrated integration for pattern recognition, and put forward the academic viewpoint of integrated pattern recognition, which further advocated and promoted the research of complexity science in China. In the field of system science, his representative works include: *From Qualitative to Quantitative Metasynthesis: Open Complex Giant System Methodology*, *Complex Giant System Science: A 21st Century Science,* and *Comprehensive Integration of Intelligent Systems*, etc.

Wang Yingluo (1930–) is a well-known Chinese management science and engineering expert, an educator, as well as an academician of the Chinese Academy of Engineering. He was the former vice chairman of the second to fourth session of the Chinese Society of System Engineering. He takes the lead in integrating the three fields of management engineering, systems engineering and industrial engineering to form a system management school with Chinese characteristics. His theories and methods have been applied to engineering management and socio-economic issues, accomplishing outstanding achievements. Wang has conducted in-depth and systematic research on China's management science discipline system, and has published a series of academic works such as *Management Science Discipline Development Strategy, System Engineering System, Management, Strategy, Enterprise Management System Engineering, Strategic Research Theory and Corporate Strategy,* and *Strategic Decisions*, making him one of the important founders of the integrated development of China's system science and management science. Moreover, he initiates the establishment of a master's degree in business administration (MBA) in China, and he is the first to advocate the application of systems engineering methods to carry out educational system engineering research in China. He is also the first Chinese scholar to propose the harmony theory, to establish the harmony diagnosis model of the system state, and to optimize the system evolution process models. Both the theory and the models are widely used in the research of regional development and enterprise development.

Cheng Siwei (1935–2015) was a professor-level senior engineer, a well-known economist and social activist, the former vice chairman of the Standing Committee of the Ninth and Tenth National People's Congress, the former dean of the School of Management of the Graduate School of the Chinese Academy of Sciences, the former director of the Virtual Economy and Data Science Research Center of the Chinese Academy of Sciences, as well as the former chairman of the China Soft Science Research Association. He was known as the "Father of Chinese Venture Capital." Cheng's main research areas were chemical system engineering, soft science and management science. Complexity science, which emerged in the 1980s, was one of the frontiers in the development of system science. The combination of complexity science and management science has become a hot issue in this field. Cheng Siwei led the research on the complexity of management science in China. Many of his research findings focused on the nonlinearity and chaos of economic systems, and these findings had promoted further research on related areas at the same time. It is believed that Cheng has played a critical role in the development of Chinese management science.

Wang Shouyang (1958–) is well-acknowledged as a distinguished researcher of the Chinese Academy of Sciences, the vice dean of the Academy of Mathematics and Systems Science of the Chinese Academy of Sciences, as well as the director of the Predictive Science Research Center of the Chinese Academy of Sciences. At the same time, he also serves as the vice chairman of the Chinese Society of Operations Research and the vice chairman of the Chinese Society of Systems Engineering. He is committed to research regarding decision analysis, economic forecasting, conflict analysis and systems engineering. His findings, particularly those related to financial management, logistics and supply chain management, conflict analysis and countermeasures, as well as decision analysis and economic forecasting, have been highly praised by international counterparts and highly valued by relevant government decision-making departments. Furthermore, he has proposed an analytical method to resolve multi-sector conflicts based on the trade-off analysis of multi-objective decision-making, and this method is currently used by foreign environmental experts as a major tool to solve international environmental negotiations.

More than those introduces above, there are many outstanding Chinese scholars in the field of systems science, who have established many new disciplines or important branches of systems science, such as Deng Julong's grey system theory, Wu Xuemou's pansystems theory, Cai Wen's matter-element analysis method, and Wei Hongsen's generalized system theory, Wu Tong's self-organization theory. These theories have played a far-reaching role in the development of China's system science, and have had an important international impact at the same time.

In addition to these exceptional scholars, China also has many influential research organizations in the field of systems science, including the Chinese Society of Systems Engineering, the Professional Committee of Philosophy of Complexity and Systems Science, as well as the Chinese Society of Systems Science. The Chinese Society of Systems Engineering is an academic social organization for Chinese experts with an academic background of systems science and system engineering science and technology. It is also a component of the China Association for Science and Technology. The organization's main publications are

System Engineering Theory and Practice, *Journal of Systems Engineering*, and *Journal of Systems Science and Systems Engineering.* The Professional Committee of Philosophy of Complexity and System Science is an integral part of the Chinese Society of Dialectics of Nature. The Chinese Society for System Science is a national academic group that organizes and conducts research on system science theory and application, and its main publication is *Journal of System Science*. These organizations have facilitated scholars to exchange views on system science and promotes the development of the discipline.

China's system science research is still at the developmental stage. In the future, while diversifying the development of system science, the concepts, viewpoints, characteristics, and research methods of system science will be further improved towards a stable system. At the same time, more attention will be paid to the application and promotion of system science in practice.

1.2 Definition and Characteristics of System

1.2.1 Definition of System

The word "system" comes from the ancient Greek word *systema*, which refers to a collection or a whole with specific functions formed by interconnected, interacting, but different elements.

The word system first appeared in the book *The Great System of the World* (by Democritus). The research on systems began in the study of engineering and military, and then extended to many other fields such as economics, biology, and social management. The definitions of the system provided by scholar's center around its characteristic as an organic whole with specific functions. The definition of the system has the following three meanings:

(1) Elements are the basic units of the system. A single element cannot be called a system. There are at least two different elements in the system, and there is an interaction among elements (the so-called interaction means that the elements change each other's state and behavior).
(2) The system has a certain structure.
(3) The system has particular characteristics or functions, which are produced by the interaction or structure of various elements within the system.

The system environment suggests a collection of things related to a particular system outside the system. The system environment is divided into surrounding environment and general environment, while the system environment signifies the surrounding environment in general.

The system boundary refers to the boundary which is used to draw a distinction between the system and the environment.

System input denotes the impact of the environment on the system. On the contrary, system output refers to the effect of the system on its environment.

The interaction between the system and the environment is achieved through the transfer of matter, energy or information.

System structure refers to how elements are related in the system.

System function is the characteristics and capacity of the system revealed in its interaction with the environment. System structure and function are both independent and interactive.

System status refers to the status, situation, characteristics, etc. of the system in every moment. The change of the system status over time is called system behavior.

The life cycle of a system refers to the process through which the system undergoes generation, development, and extinction.

System evolution refers to the changes in the structure, state, characteristics, functions, and behaviors of the system over time. The system always moves and changes in a certain time and space.

1.2.2 Characteristics of System

Systems have the following characteristics:

(1) Organization. A system is composed of two or more distinct elements, and a single element cannot be called a system. For example, a single experimental platform is not a system. Only when this experimental platform is assembled with other different elements, such as experimenters, microscopes, measuring instruments, and biological agents, will it form an experimental system.
(2) Interaction. The internal elements of the system are interrelated with each other to form an organic whole with specific functions. At the same time, the functions and properties of the elements are not a simply addition but a synergistic effect of $1 + 1 > 2$. The change of a certain element will influence overall function.
(3) Objective. The system has a clear goal. For example, the monitoring system can monitor the situation of a particular thing or area, the education system can cultivate outstanding students, and the management system can enable the organization to achieve the desired goals.
(4) Hierarchy. The system contains subsystems with specific functions. The subsystems can be further divided into a certain number of secondary and tertiary subsystems, forming a hierarchical sequence from whole to part. In the social system, individual residents, groups, units, communities, provinces or municipalities, and countries constitute a hierarchical sequence from micro to macro level.

1.3 Classification of System

There are various standards for system classification, and the common classifications can be summarized as follows:

(1) Natural system and artificial system. Natural system refers to a system developed in nature without human intervention, such as virgin forest. Artificial system is a system designed and manufactured by humans for a specific purpose. For example, mechanical devices and enterprise systems are both artificial systems.
(2) Entity system and conceptual system. Entity system signifies system organized by entities such as minerals, biology, energy, and machinery while conceptual system refers to a system composed of non-physical substances such as concepts, principles, principles, methods, systems, and procedures. The entity system serves the conceptual system, and the conceptual system guides the entity system and provides solutions for the entity system. The entity system and the conceptual system are inseparable. To be more specific, when designing a car, the engine, tires, steel plate, etc. belong to the physical system while the design drawings falls into the category of conceptual system.
(3) Physical systems and non-physical systems. Physical systems suggests systems composed of physical objects and their processes, such as supply chain systems and production lines. Non-physical systems refer to systems composed of non-physical objects and their processes including economic systems and cultural systems. Physical systems and non-physical systems are intertwined under certain conditions to constitute a large system.
(4) Open system and closed system. An open system refers to a system that exchanges material, energy or information between the system and its external environment. The system obtains the necessary material, energy or information from the environment, and converts it into new material, energy or information output after processing. The environment influences system from two aspects. On one hand, the environment promotes the exchange of material, energy or information in the system. On the other hand, it plays a role in disturbing and restricting the system. Relatively speaking, a closed system means that there is no exchange of matter, energy or information between the system and the environment. In real life, all systems have some kind of connection with the external environment, and there is no absolute closed system. However, for the convenience of research, some systems that are weakly connected to the external environment can be approximately regarded as closed systems, such as a closed feudal country.
(5) Dynamic system and static system. Dynamic system suggests that the status and function of the system as the time goes by. It has the functions of input, conversion and output such as social system and economic system. A static system means that the state and function of the system remain unchanged for a certain period of time such as the layout of urban underground pipelines as well as the station setting of buses. The concept of static systems is also

relative, and there is no static system. For the purpose of research, when a system changes very little over a period of time, it can be regarded as a static system.

(6) Deterministic system and uncertain system. A deterministic system refers to a system that does not contain uncertain factors. A deterministic system allows real-time input to determine the status of the system in a clear and unique manner and generate real-time output. On the contrary, in an uncertain system, such as the climate system, real-time input and real-time status cannot clarify the status of the system at the next moment, nor can the real-time output be produced. The uncertain system is further divided into fuzzy systems and stochastic systems. In a fuzzy system, the result of system input, transformation as well as the output is a fuzzy subset. Under the influence of fuzzy input, the system changes from one fuzzy status to another fuzzy status, and produces fuzzy output. On the other hand, the stochastic system is the probability system. In this kind of systems, the probability distribution of the real-time output of the system at the next moment can be inferred according to the real-time input and real-time status.

(7) Simple system, simple giant system, and complex giant system. There are three different kinds of systems based on the system scale: little system, large system, and giant system. However, if divided by the degree of simplicity of the system structure, systems can be classified into two categories: simple system and complex system. Generally, more research have been conducted on simple systems, simple giant systems and complex giant systems.

A simple system refers to a system in which the subsystems or element relationships that make up the system are relatively simple, such as a bicycle or a small team.

A simple giant system is a system which has a large number of subsystems or elements, however, there is a limited variety of the corresponding subsystems or elements, and their relationship is relatively simple in a simple giant system. Statistical mechanics, dissipative structure theory, synergy, etc. can be used to in the study of simple giant systems.

A complex giant system is characterized by a variety of enormous subsystems or elements, as, which has a complex and multi-layered relationship. In some complex giant systems, such as ecosystems, the relationship among elements has a certain law, which is called general complex giant systems. Nevertheless, relationships among elements of other complex giant systems, such as social systems, are complex and changeable with uncertain and unobvious regularities, which are called special complex giant systems. In special complex giant systems, people are usually parts of the constituent elements.

Chapter Summary

This chapter serve as a foundation for this book. A general understanding of the system is conducive to the analysis and management of the system, and enables students to adopt different models and methods more effectively. The main content of this chapter includes the history of the development of system science, a variety of system science schools and their corresponding research findings, the representatives of system science in China and their contributions, the definition, characteristics and classification of systems.

Important Concepts and Terms

system
system science
general system theory
simple system
simple giant system
complex giant system
system hierarchy
concept system
fuzzy system.

Questions and Exercises

1. What theories do the "old three theories" and "new three theories" refer to in the development of system science? What is the core notion of these theories?
2. What are the main schools of system science? What are their corresponding research content?
3. Briefly describe the definition of the system and its three-level meaning.
4. How can we classify systems according to different standards? What are the concepts of different types of systems?

Case Study

The combined population of Beijing, Tianjin, and Hebei is more than 100 million, and the land area is 216,000 square kilometers. Beijing, Tianjin, and Hebei are geographically linked, their people are connected and their cultures enjoy a natural affinity. They also share a profound historical origin and enjoys suitable radius of communication. All these factors led to the integration and coordinated development of Beijing-Tianjin-Hebei. The coordinated development of the three places as a whole is the core of the integration plan. The basic starting point of this project is to relieve the core functions of the capital and solve Beijing's "urban diseases". It aims to adjust and optimize the urban layout and spatial structure, build a modern transportation network system, and expand the environmental capacity and ecological space. In addition, the project promotes industrial upgrading and transfer and the co-construction and sharing of public services, and accelerates the process of market integration to build a modern new capital area. It is hoped that Beijing-Tianjin-Hebei will achieve a new pattern of coordinated development with the same goals, integrated measures, complementary advantages, and mutual benefit.

Please apply the system theory to analyze the coordinated development of Beijing-Tianjin-Hebei in terms of system boundaries, elements, structure, and functions, combined with the relevant information you have found.

Further Readings

1. Xu Guozhi, Gu Jifa, etc. *System Science* [M]. Shanghai: Shanghai Science and Technology Education Press, 2005.
2. Qian Xuesen Research Center of Shanghai Jiaotong University. *Key to Wisdom-Qian Xuesen on System Science (2nd Edition)* [M]. Shanghai: Shanghai Jiao Tong University Press, 2015.
3. Yan Zexian, Fan Dongping, Zhang Huaxia. *Introduction to System Science-Exploration of Complexity* [M]. Beijing: People's Publishing House, 2006.

2 Decision-Making Methods

Hao Zhang

Learning Objectives

1. Students will be able to understand the definition, classification and process of decision-making.
2. Students will be able to master the basic elements of the decision-making system.
3. Students will be able to master the principle and calculation method of certainty decision-making, risk decision-making and uncertainty decision, and can use it flexibly to solve practical problems.
4. Students will be able to apply TOPSIS and VIKOR methods to deal with multi-attribute decision-making problems.

Introduction

Geely Buys Volvo
Li Shufu, known as the "car maniac", is subverting the traditional order of the global auto industry: Geely bought Volvo for $1.8 billion.

1999, Volvo was bought by Ford for $6.45 billion. January 2008, Geely offered to buy Volvo from Ford, and until the end of 2008, Ford just announced its intention to sell Volvo. February 2009, Geely–Volvo MBO was approved by the National Development and Reform Commission; in the following October, Geely was identified as a preferred bidder; in December, Geely and Ford agreed on all

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_2

the important business terms of the acquisition. On 28 March 2010, the two sides signed a final agreement. Let's review the acquisition process.

Making purchase determination and appointing financial consultant. The acquisition of Volvo was a strategic decision of Geely Group. In September 2007, Li Shufu explained to Ford the idea of acquiring Volvo through a public relations company, which did not attract Ford's attention; in 2008, Li Shufu first met Ford's staff with the help of a public relations company, but the conversation was unsuccessful; Li Shufu was determined to appointed NM Rothschild, the UK investment bank, as financial adviser to participate in the project.

Getting official support and reducing competitors. Around 2009, a number of domestic companies have revealed that they want to buy Volvo, Geely Group decided to seek official support, and in March 2009, Geely received a letter of support from the NDRC. After holding the support letter, Geely actually has no competitors at home.

Recognizing the opponent's strategic intention and dealing calmly. In July 2009, when the final bid came, Crown Holdings and a Swedish consortium suddenly pulled out, which prompted the offer climbing to $2.8 billion. Geely's M&A team made an analysis and decided not to compete with the opponent. The opponent's higher quoted price is not to Ford's taste. As a "responsible" seller, Ford would not smash the brand Volvo for money. Even if the opponent is strong, it's not easy to raise more than $2 billion in the financial crisis; and without the support of the Chinese market, Volvo cannot rely the quoted money to turn its losses into profits. In July 2009, Geely submitted a legally binding bid to Ford that was approved, and both consortia were rushed out of the bid.

Finalizing the financing plan. In September 2009, Geely registered Beijing Jili Kaisheng International Investment Co., Ltd. with a registered capital of 4.1 billion yuan. At that time Geely had two paths, one on the government line and the other on cooperation with the fund. Rothschild's M&A team is divided into two groups, one to the government and the other to the fund. But almost all of the funds use "exit" as a prerequisite for negotiations, which are not smooth. The key condition to negotiate with the regional government is to settle down the company, negotiate with the Beijing and Daqing municipal governments successively, considering the company location problem, Geely Group makes the decision of giving up Beijing and Daqing, and finally reaching an agreement with the Shanghai Jiading District government. After the acquisition, Geely would build a Volvo factory in Jiading District, Shanghai.

Making investigation. Before the formal delivery, the merger faced antitrust investigations in many countries, and Geely decided to actively cooperate with the investigation, and finally passed antitrust investigations in more than 40 countries, including the European Union and the United States. Since then until delivery, Geely did not encounter resistance.

Management master Harbert A. Simen said: management is decision-making. The decision-making is carried out in the management process, from the strategic problem of M&A to the piecemeal trifles in daily operation. The application of quantitative analysis tools to decision-making is helpful to improve the quality of decision-making, make the goal of decision-making more clear, make the information of options more sufficient, and make the criteria and process of scheme selection more scientific.

2.1 Overview of Decision-Making

In daily life people face a variety of decision-making issues ranging from daily trivialities to state affairs. For enterprises, the correctness of decision-making and otherwise directly determine the success or failure of business. Scientific decision-making methods provide essential tools for enterprise managers to make correct decisions. This chapter first introduces the definition of decision, the classification of decision problems and the process of decision making. After that, the decision-making methods of certainty, risk and uncertainty are introduced, and then TOPSIS and VIKOR methods are introduced.

2.1.1 Definition of Decision-Making

Harbert A. Simen (1916–2001), one of the founders of western management decision-making school, advocated limited rationality and put forward the principle of "satisfaction" in decision-making. In his opinion, management is equal to decision-making. Peter Drucker, the father of management, pointed out that decision-making is a judgment and a choice among several options. Professor Zhou Sanduo of Nanjing University, a well-known expert in Chinese management circles, defined decision-making as the process of choosing or adjusting the direction, content and mode of activities in the future for the purpose of achieving a certain goal. This book holds that decision-making refers to the application of decision-making theory to identify, analyze and solve decision-making problems in order to achieve a certain goal. Finally, an ideal decision-making scheme is selected.

Traditional decision-making often depends on the accumulation of personal or group knowledge and ability of decision makers, which belongs to empirical decision-making. It is obviously inappropriate to rely solely on past experience for decision-making in a large and complex economic and social system. The emergence of operational research, probability theory and other disciplines provides conditions for scientific decision-making. In general, the basic elements of scientific decision-making mainly include decision makers, decision objectives, natural states, options, decision results, decision criteria.

2.1.2 Classification of Decisions

According to different criteria, the classification of decisions is different. Several common decision classifications are briefly described below.

(1) **Classification according to the degree of influence of decision-making objectives**
According to the influence degree of the decision goal, the decision can be divided into strategic decision, tactics decision and executive decision, with the influence degree of the three kinds of decision on the goal from big to small. Strategic decision is a kind of decision with global, directional and principled characteristics, which involves the overall and long-term problems related to survival and development. Tactics decision is a kind of decision with local and periodic characteristics, and it is made to achieve the goal stipulated in strategic decision-making. Executive decision is the choice of execution scheme according to the requirement of policy decision.

(2) **Classification according to the degree of repeatability of decision-making issues**
According to the degree of repeatability of decision-making problems, decisions can be divided into procedural decisions and non-procedural decisions. Procedural decisions refer to solving routine problems that often recur and are very similar in nature. For this kind of problem, it can be dealt with according to the procedure step and the routine method, the decision of this kind of problem has the rule to follow, therefore is more standard. Non-procedural decisions usually deal with occasional, unconventional problems that decision makers have no precedent to follow.

(3) **According to the number of decision-making objectives**
According to the number of decision goals, the decision can be divided into single-objective decision and multi-objective decision. Single-objective decision refers to the decision with only one goal. For example, choosing the lowest cost scheme is a single goal decision. Multi-objective decision is the existence of multiple goals, these goals interact and restrict each other. For example, to meet the goal of cost and quality at the same time, the higher the quality, the higher the cost, and there is a conflict between the two goals. Compared with single-objective decision, multi-objective decision has the following characteristics (Table 2.1):

(4) **According to the number of decision-making attributes**
Attributes describe the characteristics, qualities, or performance parameters of the alternatives. Based on the number of attributes, the types of decision can be divided into two categories: single-attribute decisions and multi-attribute

Table 2.1 Differences between single-objective and multi-objective decisions

Project	Single-objective decision	Multi-objective decision
Number of targets	One	Two or more
Concept of solution	Optimal solution	Satisfactory solution, non-inferior solution, etc
Number of solutions	Maybe the only	Generally not the only
Target value	Optimal	Possible conflict between objectives

decisions. Single-attribute decision-making method is a decision method to select the optimal scheme in a limited option with only one attribute. Multi-attribute decision is the decision process of selecting an optimal scheme in two or more finite alternatives with multiple attributes. Multi-attribute decision is different from multi-objective decision. After that, we will focus on TOPSIS, VIKOR multi-attribute decision-making method.

(5) **According to decision quantification**
According to the degree of quantization, decisions can be divided into quantitative decisions and qualitative decisions. If the index describing the decision object can be quantified and the quantitative method is used to make the decision, it is the quantitative decision; otherwise, it belongs to the qualitative decision. Qualitative decisions mainly include Delphi method, nominal group technique, brainstorming method and so on. This chapter mainly introduces several representative quantitative decision-making methods.

(6) **According to the degree of control of natural state**
According to the degree of control of natural state, decisions can be divided into certainty decisions, uncertainty decisions and risk decisions. Certainty decisions refer to the complete determination of the natural state and the determination of the result of the choice. Risk decisions refer to which natural state cannot be completely determined in the future, but the probability of its occurrence can be predicted. Uncertainty decisions are not only unable to determine which natural state appears in the future, but also uncertain about the probability of its occurrence.

(7) **According to the continuity of the decision-making process**
According to the continuity of the decision-making process, the decision can be divided into single decisions and sequential decisions. Single decisions refer to the result of making only one decision in the whole decision-making process. Sequential decisions mean that the whole decision-making process consists of a series of decisions.

2.1.3 Decision-Making Process

In general, the decision-making process consists of the following four steps:

(1) **Setting decision-making goals**
The decision maker should make clear the expected result of the decision, that is, the goal of the decision, according to the problem to be solved. The determination of goals provides a direction for decision-making, and goals should be as clear, measurable and assessable as possible. If you need to meet multiple goals at the same time, prioritize them.

(2) **Preparation of Policy-making schemes**

After setting the target, two or more options need to be presented for comparison and selection. Options should be as detailed as possible, avoid duplication and omissions, and focus on the quality of the schemes.

(3) **Choice of decision-making schemes**
Analyze, evaluate and choose the most satisfactory scheme. The specific methods of decision-making scheme selection are empirical judgment method, mathematical analysis method and experimental method. This chapter mainly introduces the mathematical analysis method, that is, using the quantitative method of decision theory to select the scheme, including expectation value method, decision tree method, Bayesian decision and so on.

(4) **Implementation of decision-making schemes**
According to the selected program implementation, make specific plans to ensure the quality of the implementation process. Sometimes the programme is readjusted and refined during implementation.

2.1.4 Basic Elements of the Decision-Making System

The decision system includes three basic elements: natural state, decision scheme and profit and loss value.

(1) **Natural state**

Natural state, also called uncontrollable factor, refers to the objective factor that does not take the will of man as the transfer. Its set is called state space and can be recorded as

$$\theta = \{\theta_i\} \quad i = 1, 2, ..., m \tag{2.1}$$

θ_i is a state variable.

(2) **Decision-making scheme**

Decision-making scheme, also called controllable factor, refers to the subjective factor that needs people to choose. The set is called the decision-making (or policy, action, behavior, program, activity, etc.) space, recorded as

$$A = \{a_j\} \quad j = 1, 2, ..., n \tag{2.2}$$

a_j is a decision variable.

(3) **Profit and loss value**

When a certain state occurs in the external environment, the profit and loss value after the implementation of the decision plan a_j, if the profit type problem is considered as the income value obtained, if the cost type problem is considered as the cost value consumed, recorded as v_{ij}

$$v_{ij} = v(\theta_i, a_j) \quad i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n \tag{2.3}$$

When the state variable is discrete, the profit and loss values form the following matrix:

$$v = (v_{ij})_{m \times n} = \begin{bmatrix} v(\theta_1, a_1) & v(\theta_1, a_2) & \ldots & v(\theta_1, a_n) \\ v(\theta_2, a_1) & v(\theta_2, a_2) & \ldots & v(\theta_2, a_n) \\ \ldots & \ldots & \ldots & \ldots \\ v(\theta_m, a_1) & v(\theta_m, a_2) & \ldots & v(\theta_m, a_n) \end{bmatrix}$$

State space, policy space and profit and loss function constitute the decision system:

$$D = D(\theta, A, V) \tag{2.4}$$

The decision-making system exists in various aspects, for example, according to the market situation of a new product to decide whether to develop or not. In this decision-making system, the state space includes being salable and being unsalable, and the policy space includes to develop and not to -develop. Policy space can be determined, while state space cannot be determined by decision makers.

The value matrix assumes the following:

$$\begin{matrix} \text{Develop} & \text{Not develop} & \\ \end{matrix}$$
$$\begin{pmatrix} 15 & 0 \\ -2 & 0 \end{pmatrix} \begin{matrix} \text{Salable} \\ \text{UnSalable} \end{matrix}$$

For the development of a new product, the market selling or not are uncontrollable. When the new product is developed and best-selling, the income is 15. If the new product is developed and unsalable, the income is -2. If you do not develop new products, whether salable or unsalable, the income is 0. How to make decisions on developing new products depends on the managers' attitude to market risk and the judgment of the probability of being salable and unsalable.

2.2 Certainty Decision-Making

Certainty decision-making, also called standard decisions or structured decisions, refer to a kind of problem that the result of decision-making process is completely determined by the action taken by the decision-maker, and its result is also clear. It has the following four conditions:

(1) There is a clear objective.

(2) There is only one definite natural state.
(3) There are two or more action programs.
(4) The profit and loss values of different action programs can be calculated in a defined state.

Example 2.1 Datang Company needs 10 million yuan bank loan and can borrow from three banks with different annual interest rates, 8%, 7.5% and 8.5% respectively. Which bank does Datang should borrow from?

Solution: The interest to be repaid to the first bank is:

$10 \times 8\% = 0.8$(million yuan).

The interest to be repaid to the second bank is:

$10 \times 7.5\% = 0.75$(million yuan).

The interest to be repaid to the third bank is:

$10 \times 8.5\% = 0.85$(million yuan).

Obviously, borrowing from the second bank with the lowest interest rate is the best option. In this decision-making process, the state space is clear, that is, the loan interest rate information of the three banks is known, and the clear decision result can be calculated, which is an certainty decision-making.

2.3 Risk Decision-Making

In the risk decision-making, it is possible to determine which natural state will appear in the future, but the probability of natural state can be predicted according to experience or information. The commonly used risk decision-making methods are: expected value analysis, decision tree analysis, Bayesian decision.

2.3.1 Expected Value Analysis

The expected value analysis is to calculate the expected profit and loss value of each scheme, and then select it according to the calculated expected profit and loss value. The scheme with the largest income or the smallest cost is the optimal scheme.

Example 2.2 The mobile phone factory tries to produce three kinds of mobile phones M_i $(i = 1, 2, 3)$, there are two states in the market: being salable and being unsalable S_j $(j = 1, 2)$, the probability of being salable is 0.6, and the probability of being unsalable is 0.4. What kind of mobile phone should be produced?

	S_1	S_2
	0.6	0.4
M_1	50	– 30
M_2	25	5
M_3	30	10

Solution: first of all, calculate the profit and loss value of each mobile phone model:

$$E(M_1) = 50 \times 0.6 + (-30) \times 0.4 = 18$$
$$E(M_2) = 25 \times 0.6 + 5 \times 0.4 = 17$$
$$E(M_3) = 30 \times 0.6 + 10 \times 0.4 = 22$$

Compare the income of each model of mobile phone $E(M_3) > E(M_1) > E(M_2)$, so the expected value method chooses to produce M_3 mobile phone.

2.3.2 Decision Tree Analysis

Decision tree analysis is a kind of risk decision-making method which uses probability and graph theory to compare different schemes in decision making, so as to obtain the optimal scheme.

The tree in graph theory is a connected and unlooped graph. In the decision tree, □ represents decision-making points, the branch derived from it is called the scheme branch. ○ represents opportunity points, the branch derived from it is called the event (state) branch. △ represents the outcome point, it is the leaf node of the decision tree, and next to it is the profit and loss value in the corresponding state. The root node is the decision point and the decision of what scheme is adopted; the second layer is the scheme layer, which is the opportunity node; the last layer is the result layer and the leaf node.

Steps to apply the decision tree method:

(1) Drawing decision trees based on decision-making issues;
(2) Calculating the probability value of the probability branch and the return value of the corresponding result node;
(3) Calculating the expected return value of each probability point;
(4) Determining the optimal scheme.

The process of using decision tree to make decision needs to be analyzed step by step from right to left. According to the profit and loss value of the right end and the probability of the probability branch, the size of the expected value is calculated and the expected result of the scheme is determined. Then the choice is made according to the expected value of different schemes.

Example 2.3 To meet the needs of the market, ready to expand production capacity, there are two options available: the first option is to build a large plant, the second plan is to build a small factory first, then consider the expansion. If we build a big factory, need to invest 5 million yuan, when the market sells well, Annual income of 2 million yuan; when the market is poor, a loss of 500,000 yuan per year. In the second scenario, build a small factory, if it sells well, two years later to decide whether to expand. We need to invest 3 million yuan, when the market sells well, earning 900,000 yuan a year, when the market is poor, the annual income is 600,000 yuan. If the plant expands in two years, need to invest 2 million yuan, the income is consistent with the construction of a large plant. The odds of good sales in the future are 0.7, the probability of market difference is 0.3; if the first two years sell well, then the probability of good sales in the next three years is 0.9, the probability of market difference is 0.1. Either way, the service period is 5 years. In this case, ask the enterprise decision maker which option will choose.

Solution:

Specific decision-making steps are as follows:

(1) Drawing the decision tree as shown below
(2) Calculating the expected returns of each node (Fig. 2.1).

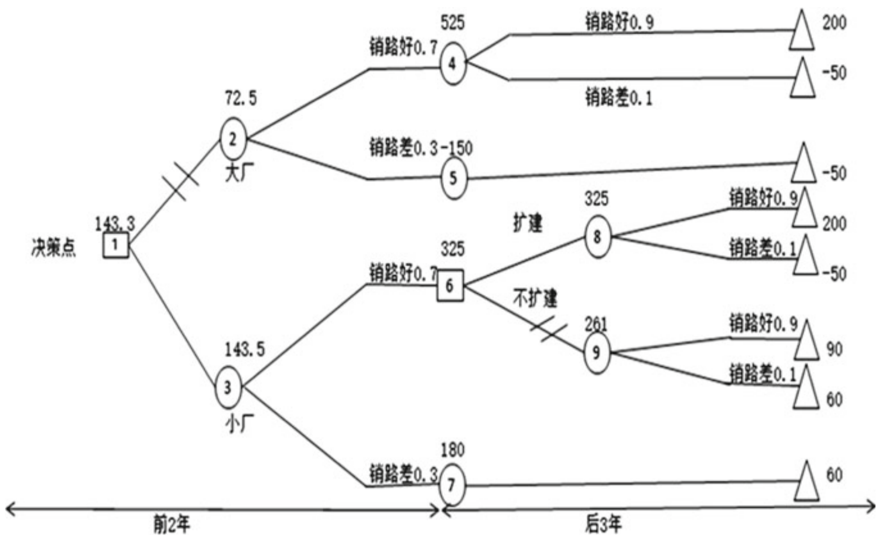


Fig. 2.1 Decision tree

Replace the Chinese in the Fig 2.1 with the corresponding English:

决策点 **decision node.**
销路好 **salable.**
销路差 **unsalable.**
扩建 **to expand.**
不扩建 **not to expand.**
大厂 **big factory.**
小厂 **small factory.**
前**3**年 **the former 3 years.**
后**2**年 **the latter 2 years.**

First, the expected income of Nodes 8 and 9 is calculated, and the decision is made at Node 6 according to the expected value of income.

Node 8: $[200 \times 0.9 + (-50) \times 0.1] \times 3 – 200 = 325$(ten thousand yuan).

Node 9: $(90 \times 0.9 + 60 \times 0.1) \times 3 = 261$(ten thousand yuan).

Node 6: $\max\{325, 261\} = 325$(ten thousand yuan).

The expansion should be selected by calculation, so if a small factory is built, it should be expanded two years later. Then continue to calculate the return value of each node from right to left.

Node 4: $[200 \times 0.9 + (-50) \times 0.1] \times 3 = 525$(ten thousand yuan).

Node 5: $-50 \times 3 = -150$(ten thousand yuan).

Node 7: $60 \times 3 = 180$(ten thousand yuan).

After calculating the expected return value of each node in the next three years, the expected income of Nodes 2 and 3 is calculated to the left.

Node 2: $[200 \times 0.7 + (-50) \times 0.3] \times 2 + 525 \times 0.7 + (-150) \times 0.3 – 500 = 72.5$(ten thousand yuan).

Node 3: $(90 \times 0.7 + 60 \times 0.3) \times 2 + 325 \times 0.7 + 180 \times 0.3 – 300 = 143.5$(ten thousand yuan).

Compared with the expected income of Nodes 2 and 3, the expected income of Node 3 is larger, so the branch of building a large factory is removed.

To sum up, the best plan is to build a small factory, if the market is good 2 years before expansion.

2.3.3 Bayesian Decision-Making

The basis of solving the risk decision-making problem is to set the probability distribution of natural state and the expected value function of consequence. The accuracy of the estimation of the probability distribution of the natural state will directly affect the expected profit and loss value of the decision. Obviously, the accuracy of the probability distribution (i.e., prior probability) of the set natural state cannot be greatly improved by only relying on the experience of the decision maker to make subjective judgment and estimation. Therefore, in order to improve the quality of decision analysis, decision makers can collect new information by sampling survey, scientific experiment and other methods according to the needs of decision-making, and then use more and more accurate new information to modify and improve the original estimation of the probability distribution of natural state, and make decisions using the modified probability distribution (i.e. posterior probability). Such a process is called Bayesian decision.

Its main steps are:

(1) Having the known conditional probability density parameter expressions and prior probabilities;
(2) Converting Bayesian formula into posterior probability;
(3) Making decisions based on posterior probability.

Using the conditional probability, multiplication formula and total probability formula, the Bayesian formula of posterior probability is as follows:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_i)p(B_i)} \quad (i = 1, 2, 3, ..., n) \tag{2.5}$$

Bayesian decision is one of the most important methods of decision analysis, which needs to solve two problems.

(1) How to use the new information to correct the prior probability.
(2) Since access to new information usually entails a certain cost, the value of new information needs to be estimated and the need for access to new information determined.

Whether the information is reliable or not will directly affect the quality of decision-making, but in the environment of asymmetric information, it is cost, time cost or economic cost to obtain reliable information. Only when the value of information is greater than the cost of obtaining information, it is necessary to obtain new information. Therefore, the new benefits that information itself can bring are usually called the value of information.

Example 2.4 Datang Company operates intelligent monitoring products, assuming that the market may have two states of being salable and being unsalable. If the market sells well, you can profit 300 million yuan; if the market is unsalable, the loss is 70 million yuan. According to the sales data of the past years, the probability of being salable is 0.8, and being unsalable is 0.2. In order to grasp the marketing situation of the product more accurately, Datang intends to hire Borui Consulting Company to conduct market research and analysis. According to the record of prediction accuracy in the past, the consulting company predicted that the best-selling accuracy of this kind of products was 0.95, and the accuracy of forecasting unsalable was 0.90. According to the results of market consultation and analysis, ask Datang whether it should hire Borui for consultation.

Solution

Datang has two options: A1- to operate the product, A2- not to operate the product. There are two kinds of marketing status of this product, that is, Q1- being salable and Q2- being unsalable. Prior probability $p(Q_1) = 0.8$, $p(Q_2) = 0.2$.

A prior probability is used to calculate the expected return value of each scheme:

$$\begin{aligned} &E(A_1) = 300 \times 0.8 + (-70) \times 0.2 = 226(\text{million yuan}) \\ &E(A_2) = 0 \end{aligned}$$

It is profitable to run the product, and the next step should be to decide whether or not to hire Borui. On the basis of the accuracy of the consulting company's market forecast, H1 = forecast the market salable, H2 = forecast the market unsalable, according to the meaning:

$$\begin{aligned} P(H_1|Q_1) = 0.95 \quad P(H_2|Q_1) = 0.05 \\ P(H_1|Q_2) = 0.10 \quad P(H_2|Q_2) = 0.90 \end{aligned}$$

From the full probability formula, the consulting company predicts that being salable and unsalable probability of the product are:

$$P(H_1) = \sum_{j=1}^{2} P(H_1|Q_j)P(Q_j) = 0.95 \times 0.8 + 0.10 \times 0.2 = 0.78$$

$$p(H_2) = \sum_{j=1}^{2} P(H_2|Q_j)P(Q_j) = 0.05 \times 0.8 + 0.90 \times 0.2 = 0.22$$

From the Bayesian formula, the conditional probability is calculated

$$\begin{aligned} P(Q_1|H_1) &= \frac{P(H_1|Q_1)P(Q_1)}{P(H_1)} = \frac{0.95 \times 0.8}{0.78} \approx 0.9744 \\ P(Q_2|H_1) &= \frac{P(H_1|Q_2)P(Q_2)}{P(H_1)} = \frac{0.10 \times 0.2}{0.78} \approx 0.0256 \\ P(Q_1|H_2) &= \frac{P(H_2|Q_1)P(Q_1)}{P(H_2)} = \frac{0.05 \times 0.8}{0.22} \approx 0.1818 \\ P(Q_2|H_2) &= \frac{P(H_2|Q_2)P(Q_2)}{P(H_2)} = \frac{0.05 \times 0.8}{0.22} \approx 0.8182 \end{aligned}$$

The probability distribution of state variables obtained by modifying state variables with supplementary information:

$$P(Q_j|H_i) \quad (i = 1, 2;\ j = 1, 2) \tag{2.6}$$

This is the posterior distribution of state variables.

If the forecast result of market consultation is best-selling, the expected return of A1 is obtained by replacing prior probability with posterior probability:

$$E(A_1|\,H_1) = 0.9744 \times 300 - 0.0256 \times 70 = 290.528(\text{million yuan})$$
$$E(A_2|\,H_1) = 0(\text{million yuan})$$

So when the consultant's forecast is best-selling, the best decision is to A1, the product should be operating.

Similarly, when the forecast result is unsalable, the posterior probability is used to replace the prior probability, and the expected income is:

$$E(A_1|H_2) = 0.1818 \times 300 - 0.8182 \times 70 = -2.734(\text{million yuan})$$
$$E(A_2|\,H_2) = 0(\text{million yuan})$$

As a result, when the forecast result of the consulting company is unsalable, the optimal decision is A2, not to operate the product.

$$\begin{aligned}\text{Total expected income is } E &= E(A_1|\,H_1) \times P(H_1) + E(A_2|\,H_2) \times P(H_2)\\ &= 290.528 \times 0.78 + 0 \times 0.22\\ &= 226.612(\text{million yuan}),\\ E - E(A_1) &= 226.612 - 226 = 0.612(\text{million yuan})\end{aligned}$$

Therefore, as long as Datang Company pays no more than 612,000 yuan to the consulting company, it can carry out market research, otherwise it should not carry out market research. If the survey results are that the product sells well, you should choose to operate the product. If the market survey results are unsalable, the product should not be operated.

2.4 Uncertainty Decision-Making

In most cases, the decision maker is faced with uncertainty, and the decision maker can only make the decision according to a certain simple principle, which we call the decision criterion. The commonly used decision criteria are optimistic law, Wald law, Hurwicz law, Laplace law, Savage law.

2.4.1 Optimistic Law

Optimistic law, also known as the greatest criterion, makes policy makers optimistic about the future. The decision maker first determines the return value of each scheme in the most optimistic natural state, then compares and selects the scheme with the largest return value as the optimal scheme.

Table 2.2 Unit of profit and loss of three production schemes Unit: 10,000 yuan

Production scheme	Market performance		
	Salable	Mediocre	Unsalable
Mass volume	9	3	−1
Medium volume	7	6	2
Small volume	5	4	1

Mass volume: max(9,3,−1) = 9(ten thousand yuan). Medium volume: max(7,6,2) = 7(ten thousand yuan). Small volume: max(5,4,1) = 5(ten thousand yuan). Select the best option from three options: max(9,7,5) = 9(ten thousand yuan)

Example 2.5 Datang Factory decided to invest a certain cost to produce a new product, need to decide whether the scale of production is large, medium or small batch, and the market for this kind of products have three cases of selling, general and unsalable, the three production scale in the three natural state of the income value as shown in the following table: (Table 2.2)

According to the optimistic criterion, we should choose the mass production plan.

2.4.2 Wald Law

Wald law is also called pessimistic criterion, also known as maximum minimum criterion, which is a safe-haven decision criterion. Contrary to the optimistic criterion, the decision maker is pessimistic about the future. First, the minimum income of each scheme in various states is determined, and then a maximum value is selected from these minimum values. The corresponding decision is the optimal decision.

Example 2.6 The pessimistic criterion is used to solve the example 2.5.

Solution:

Mass volume: min(9,3,−1) = −1(ten thousand yuan).

Medium volume: min(7,6,2) = 2(ten thousand yuan).

Small volume: min(5,4,1) = 1(ten thousand yuan).

Select the best option from the above three options: max(−1,2,1) = 2(ten thousand yuan).

According to the pessimistic criterion, the medium batch production plan should be selected.

Table 2.3 Expected income for three production schemes

Production scheme	Market performance			Expected return (10,000 yuan)
	Salable	Mediocre	Unsalable	
Mass volume	9	3	−1	$9 \times 0.8 + (-1) \times 0.2 = 7.0$
Medium volume	7	6	2	$7 \times 0.8 + 2 \times 0.2 = 6.0$
Small volume	5	4	1	$5 \times 0.8 + 1 \times 0.2 = 4.2$

2.4.3 Hurwicz Law

Hurwicz method is also called rule of compromise, also known as the optimistic coefficient method. The best and worst natural states appear in general decisions. When the decision maker is unable to estimate the probability of each state appearing and is willing to take a compromise, the best natural states can be given an optimistic coefficient α, and the corresponding worst natural states have a pessimistic coefficient $(1 - \alpha)$, to calculate their respective expected values. Finally, the scheme with the largest expected value is the optimal scheme. An optimistic coefficient is between [0, 1].

Example 2.7 Assuming that the decision maker estimates the optimistic coefficient $\alpha = 0.8$, the problem of example 2.5 is solved by compromise criterion.

Solution:

The optimistic coefficient is 0.8 and the pessimistic coefficient is 0.2. The expected income of each scheme is shown in the right column of Table 2.3 above, max (7.0, 6.0, 4.2) = 7.0 (10,000 yuan), and the corresponding scheme is mass production. Therefore, according to the compromise criterion, mass production is chosen as the optimal scheme.

2.4.4 Laplace Law

Laplace law is also called equal possibility criterion, because it is impossible to know the probability of occurrence of various natural states, it can be considered that the probability of occurrence of each state is equal. The maximum expected income or minimum expected loss is chosen as the optimal decision.

Example 2.8 Use the possibility criterion to make decisions on example 2.5.

Solution:

The probability of occurrence of various states is 1/3, and the expected income of each scheme is shown in Table 2.4 below, max (11/3, 5, 10/3) = 5 (10,000 yuan), and the corresponding scheme is medium batch production. Therefore, according to the equal possibility criterion, the medium batch is selected as the optimal production scheme.

Table 2.4 Expected return values for three production options

Production programme	Market conditions			Expected income (10,000 yuan)
	Salable	Mediocre	Unsalable	
Mass volume	9	3	-1	$9 \times 1/3 + 3 \times 1/3 + (-1) \times 1/3 = 11/3$
Medium volume	7	6	2	$7 \times 1/3 + 6 \times 1/3 + 2 \times 1/3 = 5$
Small volume	5	4	1	$5 \times 1/3 + 4 \times 1/3 + 1 \times 1/3 = 10/3$

2.4.5 Savage Law

Savage method is also called regret value criterion. Regret value is a kind of opportunity loss, which refers to the decrease of income or the increase of cost caused by improper decision-making. According to the regret value criterion, the maximum return value in each natural state is the ideal value (that is, the least regret choice), and the difference between the return value and the ideal value of each scheme in this state is taken as the regret value. The decision maker pursues the minimum regret value. First, the maximum regret value is selected in various schemes, then the maximum regret value of each scheme is compared, and the corresponding scheme of the smallest is selected as the optimal decision scheme.

Example 2.9 Try the regret value criterion to make decisions on Example 2.5.

Solution:

Step 1 is to calculate the maximum return value in each state, as shown in Table 2.5.

Step 2 is to calculate the regret value in each state (minus the actual value in each state with the maximum return value in each state), as shown in Table 2.6.

Step 3 is to find out the minimum result of the maximum regret value of the three schemes from Table 2.6: min (3, 2, 4) = 2 (10,000 yuan), corresponding to

Table 2.5 Maximum return values for each state

Programme	Salable	Mediocre	Unsalable
Maximum return	9	6	2

Table 2.6 Regret values for each scheme in each state

Production programme	Market conditions			Expected income (10,000 yuan)
	Salable	Mediocre	Unsalable	
Mass volume	$9–9 = 0$	$6–3 = 3$	$2–(-1) = 3$	3
Medium volume	$9–7 = 2$	$6–6 = 0$	$2–2 = 0$	2
Small volume	$9–5 = 4$	$6–4 = 2$	$2–1 = 1$	4

the medium batch production scheme. According to the regret value criterion, the production scheme of medium batch is the optimal scheme.

2.5 Multi-Attribute Decision-Making

Decision-making problems in socio-economic systems often involve multiple different attributes (commonly referred to as indicators). Generally speaking, multi-attribute comprehensive evaluation has two remarkable characteristics: first, there is no uniform dimension between indicators, that is, there is no uniform dimension between attributes, so it is difficult to measure by unified standard; second, there is some contradiction between some indexes. A scheme improves an index value, but may reduce another index value. Therefore, it is a problem to solve the multi-attribute decision-making method to overcome the difficulty of non-metricity between indicators and coordinate the contradiction between indicators. This book introduces two typical multi-attribute decision-making methods: TOPSIS method and VIKOR method.

2.5.1 TOPSIS Method

TOPSIS (technique for order preference by similarity to ideal solution) is also called ideal solution. This method constructs the ideal solution and the negative ideal solution of the multi-attribute problem, and takes the two benchmarks of approaching the ideal solution and away from the negative ideal solution as the basis for evaluating the feasible schemes. The so-called ideal solution is to assume that each index attribute reaches the most satisfactory solution. The so-called negative theoretical solution is to assume that each index attribute reaches the most unsatisfactory solution.

Multi-index attributes are different in dimension and order of magnitude, which often brings a lot of inconvenience to decision analysis. Generally speaking, the index value should be standardized to make the decision with the ideal solution.

Decision-making processes using the TOPSIS method are:

Let the scheme set of the multi-attribute decision problem is $X = \{X_1, X_2, \ldots, X_m\}$, attribute set is $F = \{f_1, f_2, \ldots, f_n\}$, decision matrix is $Y = \left(y_{ij}\right)_{m\times n}$, y_{ij} is the property value of the i number scheme under the j property, $i \in M$, $j \in N$. $M = \{1, 2, \ldots, m\}$ is the subscript set of the scheme, $N = \{1, 2, \ldots, n\}$ is the subscript set of the attribute. Weight vector of attributes is $W = (w_1, w_2, \ldots, w_n)$, satisfied $\sum_{j=1}^{n} w_j = 1, w_j \geq 0, j \in N$.

(1) Using Vector Normative Method to Solve Normalized Matrix

$$Z = (z_{ij})_{m\times n}$$

$$Z_{ij} = y_{ij} / \sqrt{\sum_{i=1}^{m} y_{ij}^2} \tag{2.7}$$

(2) Constructing weighted canonical matrix $X = \left(x_{ij}\right)_{m \times n}$

$$x_{ij} = w_j \cdot z_{ij} \tag{2.8}$$

(3) Determining the ideal solution and negative ideal solution
Ideal Solution of Benefit-based Properties

$$x_j^* = \max_i x_{ij} \tag{2.9}$$

Ideal Solution of Cost Type Attributes

$$x_j^* = \min_i x_{ij} \tag{2.10}$$

Negative ideal solution of Benefit Type Attributes

$$x_j^0 = \min_i x_{ij} \tag{2.11}$$

Negative ideal Solution of Cost Type Attributes

$$x_j^0 = \max_i x_{ij} \tag{2.12}$$

Constructing ideal solution set $X^* = \left\{x_1^*, x_2^*, \ldots, x_n^*\right\}$, a negative ideal solution set $X^0 = \left\{x_1^0, x_2^0, \ldots, x_n^0\right\}$.

(4) Calculating the distance between ideal solution and negative ideal solution

Distance to ideal solution

$$d_i^* = \sqrt{\sum_{j=1}^{m} \left(x_{ij} - x_j^*\right)^2} \tag{2.13}$$

Distance to the negative solution

$$d_i^0 = \sqrt{\sum_{j=1}^{m} \left(x_{ij} - x_j^0\right)^2} \tag{2.14}$$

(5) Calculating the proximity of each scheme to the ideal solution

$$C_i^* = d_i^o / \left(d_i^o + d_i^*\right) \tag{2.15}$$

(6) The smaller the square sum of the difference between the attribute value of each weighted specification x_{ij} and the x_j^* of the ideal solution of the attribute, the smaller the distance d_i^* between the scheme and the ideal solution, the better the scheme is. The larger the square sum of the difference between the attribute value of each weighted specification x_{ij} and the negative ideal solution of the attribute, the larger the distance d_i^0 between the scheme and the negative ideal solution, the better the scheme is. Therefore, the maximum value C_i^* will be the optimal scheme.

Example 2.10 Ms. Zhang is going to buy a house this year. The main factors affecting the purchase of a house include the following three aspects: the price of per square meter, the commuting distance and the comfort level of the house. The comfort level takes into account: area, orientation, storey, the pattern, lift etc. There are four alternative houses available, in which the weight of price is 0.65, the weight of distance from the company is 0.30, and the weight of comfort is 0.05. The specific data of the four alternative rooms are shown in Table 2.7 below.

Table 2.7 Attributes of the three properties of the selected rooms

Alternative	Price (10,000 yuan per square metre)($W_1 = 0.65$)	Distance (km)($W_2 = 0.30$)	comfort level ($W_3 = 0.05$)
X_1	2.50	15	10
X_2	1.80	20	5
X_3	2.04	18	6
X_4	2.24	10	8

Table 2.8 Standardized data

Alternative	Price (10,000 yuan) ($W_1 = 0.65$)	Distance (km)($W_2 = 0.30$)	Comfort level ($W_3 = 0.05$)
X_1	0.5787	0.4631	0.6667
X_2	0.4167	0.6175	0.3333
X_3	0.4722	0.5557	0.4000
X_4	0.5185	0.3087	0.5333

Solution:

(1) According to the steps, the data in Table 2.7 are normalized according to the formula $Z_{ij} = y_{ij}/\sqrt{\sum_{i=1}^{m} y_{ij}^2}$, as shown in Table 2.8.

Take the price data y for room selection X_1 as an example:

$$Z_{11} = y_{11}/\sqrt{\sum_{i=1}^{4} y_{i1}^2} = 2.50/\sqrt{2.50^2 + 1.80^2 + 2.04^2 + 2.24^2} = 0.5787$$

Normalize the rest of the data in the same way:

(2) Using Formula $x_{ij} = w_j \cdot z_{ij}$ to calculate Weighted Normative Array.

Take the standardized price data Z_{11} of alternative housing X_1 as an example to carry on the weighted standardization:

$$x_{11} = w_1 \cdot z_{11} = 0.65 \times 0.5787 = 0.3762.$$

The rest of the data were weighted according to the same methodology and the results were as follows: 0.3762 0.1389 0.0333.

0.27090.18530.0167.

0.30690.16670.0200.

0.33700.09260.0267.

(3) According to the weighted norm matrix, the ideal solution and the negative ideal solution are obtained. Among them, the ideal solution of the cost attribute (price, distance) are $x_1^* = \min_i x_{i1}$, $x_2^* = \min_i x_{i2}$, the negative ideal solution are $x_1^0 = \max_i x_{i1}$,$x_2^0 = \max_i x_{i2}$, the ideal solution for Benefit Attributes is $x_3^* = \max_i x_{i3}$, the negative ideal solution is $x_3^0 = \min_i x_{i3}$, the results are as follows:

$$X* = [0.2709\,,\ 0.0926\,,\ 0.0333]$$

$$X^0 = [0.3762\,,\ 0.1853\,,\ 0.0167]$$

(4) Calculating the distance between ideal solution and negative ideal solution.

Table 2.9 Distance data

Alternative	$d*$	d^0	$C*$
X_1	0.1150	0.0493	0.3001
X_2	0.0942	0.1053	0.5278
X_3	0.0834	0.0718	0.4626
X_4	0.0664	0.1011	0.6036

Take an example of the distance from alternative room X_1 to ideal solution and negative ideal solution:

$$d_1^* = \sqrt{(x_{11} - x_1^*)^2 + (x_{12} - x_2^*)^2 + (x_{13} - x_3^*)^2}$$
$$= \sqrt{(0.3762 - 0.2709)^2 + (0.1389 - 0.0926)^2 + (0.0333 - 0.0333)^2} = 0.1150$$

$$d_1^0 = \sqrt{(x_{11} - x_1^0)^2 + (x_{12} - x_2^0)^2 + (x_{13} - x_3^0)^2}$$
$$= \sqrt{(0.3762 - 0.3762)^2 + (0.1389 - 0.1853)^2 + (0.0333 - 0.0167)^2} = 0.0493$$

The remaining data are calculated and processed using the same method, and the distance is shown in Table 2.9.

(5) Calculating the proximity of each scheme to the ideal solution.

Taking as an example the degree of proximity of alternatives X_1 to the ideal solution:

$$C_1^* = d_1^0/(d_1^0 + d_1^*) = 0.0493/(0.0493 + 0.1150) = 0.3001$$

In the same way, we can calculate the value of C_2^*、C_3^*、C_4^*, according to the degree of proximity $C*$, the results of the alternative housing ranking are $X_4 > X_2 > X_3 > X_1$, the fourth suite is the most suitable.

2.5.2 VIKOR Law

VIKOR method is a multi-attribute decision-making method based on ideal point solution proposed by Opricovic in 1998. Its basic idea is to determine ideal solution and negative ideal solution based on optimal solution. Then the optimal scheme is selected according to the evaluation value of each option and the proximity of the ideal scheme. VIKOR method is similar to the classical TOPSIS method and adopts the compromise method close to the ideal point, but the optimal solution obtained by the TOPSIS method is not necessarily the solution closest to the ideal point, but the solution obtained by the VIKOR method is based on the result of mutual concession between the two attributes.

VIKOR method is characterized by the proper solution closest to the ideal scheme, which maximizes the group benefit and minimizes the individual loss. The scheme obtained by this method is more acceptable to the decision makers.

(1) Standardized processing of decision matrices.

Efficiency indicators:

$$x_{ij} = \frac{y_{ij}}{\sqrt{\sum_{i=1}^{m} y_{ij}^2}} \tag{2.16}$$

Cost indicators:

$$x_{ij} = \frac{1/y_{ij}}{\sqrt{\sum_{i=1}^{m} (1/y_{ij})^2}} \tag{2.17}$$

(2) Determination of ideal solution and negative ideal solution

$$x_j^* = \max_i x_{ij}$$

$$x_j^0 = \min_i x_{ij}$$

Constructing ideal solution set $X^* = \left\{x_1^*, x_2^*, \ldots, x_n^*\right\}$, negative ideal solution set

$$X^0 = \left\{x_1^0, x_2^0, \ldots, x_n^0\right\}$$

(3) Calculation of the distance ratio between ideal solution and negative ideal solution

$$M_i = \sum_{j=1}^{m} w_j \left(\frac{x_j^* - x_{ij}}{x_j^* - x_j^0}\right) \tag{2.18}$$

$$N_i = \max_j \left\{ w_j \left(\frac{x_j^* - x_{ij}}{x_j^* - x_j^0}\right) \right\} \tag{2.19}$$

(4) Calculation of benefit ratio Q_i

let $M^+ = \max\{M_1, M_2, ..., M_m\}$,$M^- = \min\{M_1, M_2, ..., M_m\}$.
$N^+ = \max\{N_1, N_2, ..., N_m\}$,$N^- = \min\{N_1, N_2, ..., N_m\}$

$$Q_i = v\frac{M_i - M^-}{M^+ - M^-} + (1 - v)\frac{N_i - N^-}{N^+ - N^-} \tag{2.20}$$

For the decision mechanism coefficient of most criterion strategies, the value of V reflects the most important degree of criterion or the preference of decision makers. When $V > 0.5$, it represents that it is a risk preference type to make decisions according to the opinions of the majority; when $V = 0.5$, it represents that it is a risk neutral type to take into account the interests of the majority and the objections of the minority at the same time; when $V < 0.5$, it represents that it is a risk aversion type to make decisions according to the objections of the minority. In general, it will take $V = 0.5$.

M_i represents group benefits of options, the smaller the M_i value, the greater the group benefit; N_i Value represents individual regret, the smaller the N_i value, the smaller the individual regret. Finally, the region is sorted according to the size of the Q_i value, and the scheme with the smallest Q_i value is the optimal scheme.

Example 2.11 Continue with example 2.10.

(1) Normalization of Decision Matrix

Take the standardization price data y for alternative room X_1 as an example, because the price is a cost index, so the formula (2.17) is standardized:

$$x_{11} = \frac{1/y_{11}}{\sqrt{\sum_{i=1}^{4}(1/y_{i1})^2}} = \frac{1/2.50}{\sqrt{(1/2.5)^2 + (1/1.80)^2 + (1/2.04)^2 + (1/2.24)^2}} = 0.4197$$

According to formula (2.16), (2.17), the remaining data are normalized to obtain the following matrix.

0.41970.47110.6667.
0.58300.35340.3333.
0.51440.39260.4000.
0.46840.70670.5333.

(2) The ideal solution and negative ideal solution are obtained.

Take the price attribute as an example,
$x_1^* = \max\{x_{11}, x_{21}, x_{31}, x_{41}\} = 0.5830$,$x_1^0 = \min\{x_{11}, x_{21}, x_{31}, x_{41}\} = 0.4197$.
In the same way, the ideal solution of distance and comfort is obtained.
Constructing ideal solution sets $X^* = \{0.5830, 0.7067, 0.6667\}$.
Constructing the Negative ideal solution set $X^0 = \{0.4197, 0.3534, 0.3333\}$.

(3) Calculating group benefit M_i, individual regret N_i and benefit ratio Q_i,

Take the alternative room X_1 as an example, $M_1 = \sum_{j=1}^{m} w_j(\frac{x_j^*-x_{1j}}{x_j^*-x_j^0}) = w_1(\frac{x_1^*-x_{11}}{x_1^*-x_1^0}) + w_2(\frac{x_2^*-x_{12}}{x_2^*-x_2^0}) + w_3(\frac{x_3^*-x_{13}}{x_3^*-x_3^0}) = 0.85$

$$N_1 = \max_j\left\{w_j(\frac{x_j^* - x_{1j}}{x_j^* - x_j^0})\right\}$$
$$= \max\left\{w_1(\frac{x_1^* - x_{11}}{x_1^* - x_1^0}), w_2(\frac{x_2^* - x_{12}}{x_2^* - x_2^0}), w_3(\frac{x_3^* - x_{13}}{x_3^* - x_3^0})\right\} = 0.65$$

Similarly, $M_2 = 0.35$, $M_3 = 0.58$, $M_4 = 0.48$, $N_2 = 0.30$, $N_3 = 0.23$, $N_4 = 0.38$.
$M^+ = \max\{M_1, M_2, M_3, M_4\} = 0.85$,$M^- = \min\{M_1, M_2, M_3, M_4\} = 0.35$.
$N^+ = \max\{N_1, N_2, N_3, N_4\} = 0.65$,$N^- = \min\{N_1, N_2, N_3, N_4\} = 0.27$

$$Q_1 = V\frac{M_1 - M^-}{M^+ - M^-} + (1 - V)\frac{N_1 - N^-}{N^+ - N^-}$$
$$= 0.5\frac{0.85 - 0.35}{0.85 - 0.35} + (1 - 0.5)\frac{0.65 - 0.27}{0.65 - 0.27} = 1$$

As shown in the table below (Table 2.10):

(4) According to the size of the Q, preparation room sorting results are $X_2 > X_3 > X_4 > X_1$. The second housing should be chosen.

TOPSIS and VIKOR methods can achieve the purpose of scheme sorting. Each of the two methods has its own advantages and disadvantages. No one is good or bad.

Table 2.10 Final ranking tables for alternatives

Alternative	M	N	Q	Rank
X_1	0.85	0.65	1	4
X_2	0.35	0.30	0.04	1
X_3	0.58	0.27	0.23	2
X_4	0.48	0.46	0.38	3

They are different in the principle, process and complexity of scheme sorting. The specific decision of which decision method is used depends on the preference of the decision maker on the one hand and the probability of success of each method in the past on the other.

Summary
In the development of the organization it will face a lot of decisions, and it's important to make scientific and reasonable decisions according to the decision-making method. This chapter introduces decision theory, decision process and common decision methods, with emphasis on decision tree method, Bayesian decision, TOPSIS method and VIKOR method.

Important concepts and terminology
decision-making.

single-object decision.

multi-objective decision.

certainty decision-making.

risk decision-making.

unsure decision.

natural state.

decision-making scheme.

profit and loss values.

decision tree.

Bayesian decision.

TOPSIS law(technique for order performance by similarity to ideal solution).

VIKOR law(VIsekriterijumska Optimizacija IKompromisno Resenie).

Questions and Exercises

1. What should be included in the basic elements of the decision-making system?
2. Briefly analyzes the applicable conditions of certainty decision-making, risk decision-making and uncertainty decision-making.
3. Datang intends to put into production a new electronic product, there are two options: one is to build a larger plant, the other is to build a smaller plant. Assume that the construction of a large plant investment of 500,000 yuan, small plant investment of 250,000 yuan. At the same time, the future sales of this electronic product will be best-selling and unsalable, the probability of occurrence is 0.7 and 0.3 respectively. The income situation is as follows: under the condition that the big factory is under construction and the product sells well,

can obtain 800,000 yuan income, unsalable loss 100,000 yuan; build small factory and the product sells well can obtain 500,000 yuan income, unsalable loss 200,000 yuan, which kind of plan is desirable?

4. In order to meet the development of logistics industry, a logistics park needs to be established. There are three options, from which an optimal scheme needs to be selected. The following four factors affect the choice of the scheme: the natural environment (topography, geological and hydrological conditions, etc.), the degree of transportation convenience, land cost and labor supply. The weights of each factor are:0.3, 0.25, 0.3 and 0.15 respectively. The specific data are shown in Table 2.11. Please choose the scheme by VIKOR and TOPSIS methods respectively.

Case Analysis

Suppose we choose one of the three alternatives of Beijing to expand into a logistics park, which are: Oriental Chemical Plant in Tongzhou District; Beijing Yanshan Petrochemical Product Storage and Transportation Center; Beijing Hongxing Aoxin Storage and Transportation Co., Ltd. In Daxing District.

Oriental Chemical Plant covers an area of 1.28 km^2. It is located in Zhangjiawan Town, with an area of 105.8 km^2. There are 625 industrial enterprises, 68 construction enterprises and 107 other industries in the town. The total population is 50,782 and the density is 48 km^2. Its traffic is as follows:

- 1.2 km west of Railway and Sixth Ring Road;
- 1.6 km north of the Grand Canal;
- 1.5 km east of S229 Songliang Road;
- 1.4 km south of G103 National Road;
- 0.621 km apart from its nearest residential area.

Beijing Yanshan Petrochemical products Storage and Transportation Center covers an area of 0.02 km^2, and is located in in Fangshan District of Beijing with an area of 37.8 km^2. It is located at the southeast end of Xishan Mountain of Yanshan Mountains Ranges. Mountains and hills each account for about half, the existing population of about 100,000, density of 2646 people per square kilometer. Its traffic is as follows:

Table 2.11 Specific data for programmes

Factor Programme	Natural environmental factors (0.3)	Transportation accessibility (0.25)	Land costs (0.3)	Labor supply(0.15)
Programme 1	31	1	200	5
Programme 2	18	2.5	150	3
Programme 3	25	2	180	8

- 0.43 km northwest of railway;
- Northeast from Yanshan-Fangshan northeast ring line 0.931 km;
- Next to Dingdong Road in the west;
- 2.6 km from the nearest neighborhood;
- 1.6 km from the nearest school;
- There is a large open space in the northeast.

Beijing Hongxing Aoxin Storage and Transportation Co., Ltd. is located in Beizangcun Town of Daxing District, with the town area of 60 km^2. It belongs to Yongding River alluvial plain with flat terrain, including 23 villages. The town has 15,544 people, with a density of 284 people per square kilometer. Its traffic is as follows:

- 0.738 km east of X032 Luqiu Road;
- 0.312 km south of S316 Huangliang Road;
- 0.87 km from Six Ring Road;
- Close to residential areas in the west;
- The nearest school is 0.2936 km.

According to your decision-making knowledge and more relevant information in research, choose the appropriate method to help the enterprise solve this location problem, and explain the reasons.

Further Readings

1. Zhang Suodi, Ji Yingdong, Hu Linna, et al. *Theory, Technology and Methods of Management Decision-making* [M]. Beijing: Tsinghua University Press, 2013
2. Tao Changqi et al. *Theory and Methods of Decision Making* [M]. Beijing: Renmin University Press, 2010
3. Fang Zhigeng, Liu Sifeng, Zhu Jianjun, Hu Mingli. *Theory and Methods of Decision Making* [M]. Beijing: Science Press, 2008

Prediction Methods

3

Hao Zhang

Learning Target

1. Understanding the meaning, classification and steps of prediction.
2. Being familiar with several typical qualitative prediction methods.
3. Being familiar with moving average method, exponential smoothing method, regression analysis prediction method, etc., and making predictions based on actual data.
4. Mastering the trend extrapolation prediction method and applying several typical functions to construct prediction models.

Introduction

Who is the next president of the United States?
For every U.S. presidential election, there will be a frenzy. The two major parties are fighting openly and secretly, and the public is also looking forward to the final candidate. However, big data technology can foresee the result of Donkey versus Elephant, which may no longer be mysterious.

Researchers at the University of South Florida have proposed a predictive model that can determine the final outcome of the election by analyzing TV program ratings. Two researchers from the Department of Information Systems and Decision Sciences at the school's business school, Alash Bafa and Balaji Padmanaban, used some political television programs in the four weeks before the 2012 US election as data samples. The states where the situation is basically clear, collected the ratings and duration of these programs, and calculated a correlation model between

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_3

the ratings and the election results. After that, we collect ratings data from other states and use this model to predict wins and losses in elections in these regions. The results of the study showed that the accuracy of prediction for 99 TV programs exceeded 59%, and the accuracy of even 3 programs exceeded 79%.

In addition, the researchers pointed out that the model can not only be applied to each election prediction, but also can tell candidates where to place campaign ads. In the 2012 U.S. election, the two sides spent a total of nearly 2 billion US dollars on campaign advertising, and relevant sources estimated that the 2016 election cost up to 10 billion US dollars. Using this model, candidates can choose the best TV programs and time slots for targeted advertising to maximize the effect of publicity.

For this prediction model, New York University School of Business Professor Vasan Tedal believes that this interesting study proves that the election results in the region can be predicted based on the ratings of the country or a certain state. It opened our minds to some of the actual factors that influence the election situation, and also saw the magic of big data prediction.

Source: *Science and Technology Daily*, November 2015.

3.1 Overview of Prediction

3.1.1 The Meaning of Prediction

Prediction is a process of rationally inferring the development trend of things based on the information and data of the history and current situation of objective things, using scientific methods and means.

Prediction was founded by Swiss mathematician Jakob Bernoulli (1654–1705). Contemporary prediction technology is generally believed to have originated in the early twentieth century. Chinese scientist Professor Weng Wenbo created a prediction theory with information prediction as the core, making a great contribution to the research and development of prediction theory.

This chapter focuses on a summary of several major quantitative prediction methods.

3.1.2 Classification of Predictions

Due to different factors such as prediction nature and prediction range, predictions can generally be divided into the following categories.

1. According to the nature of prediction
   (1) Quantitative prediction. Quantitative prediction refers to the use of modern mathematical methods for data processing based on the data related to the predicted object, and quantitative inferences about the future development trend of the predicted object.

   (2) Qualitative prediction. Qualitative prediction refers to making qualitative inferences on the future development trend of things based on the personal experience and analytical judgment ability of relevant experts, based on existing data, combined with the characteristics of the prediction object, without considering quantitative changes.
2. According to the prediction range
   (1) Macro prediction. Macro prediction refers to a comprehensive prediction based on a global perspective, mainly for countries, regions or large organizations.
   (2) Micro prediction. Micro prediction refers to the prediction of the specific activities of the grassroots units.
3. According to the length of prediction time
   (1) Short-term prediction. Short-term prediction refers to the prediction of the development status of the prediction object within 1 year.
   (2) Mid-term prediction. The mid-term prediction refers to the prediction of the development status of the prediction target in 1–5 years.
   (3) Long-term prediction. Long-term prediction refers to the prediction of the development status of the prediction object over 5 years.

3.1.3 Prediction Steps

For different prediction objects, the corresponding prediction steps may be different. Under normal circumstances, prediction activities include the following basic steps:

(1) Define prediction goals. Determining the prediction target is the premise of the entire prediction activity. The prediction target must meet the actual demand. Different prediction targets require different materials and methods.
(2) Collect and sort out prediction data. The accuracy of the prediction largely depends on the quantity and quality of the prediction data, so complete and accurate data is the basis of the prediction. When collecting information, we should start from many aspects and try to be as comprehensive and accurate as possible.
(3) Select prediction methods and models. Different prediction methods have their own characteristics and applicability. The appropriate method should be selected and a prediction model should be established on the premise of comprehensively considering the prediction target and data collection.
(4) Analyze prediction errors. Errors in the prediction process are inevitable, but the reasons for the errors can be analyzed, and then the prediction methods and models can be adjusted to make the prediction results as realistic as possible or to control the errors within a reasonable range to improve the accuracy of the prediction.

(5) Generate prediction reports. After completing the above process, a prediction report needs to be generated. The prediction report is the final result of the entire prediction activity and can be used as a reference for decision-making.

3.2 Qualitative Prediction Methods

1. Market research prediction method

The market survey prediction method refers to the fact that the investigator obtains information and collects data from relevant personnel familiar with the market, and then makes a judgment on the future market conditions of the product based on his own professional knowledge. The market survey prediction method includes four forms: (1) The factory director (manager) opinion survey method: mainly to understand the market trend from the factory director, manager, business supervisor and other personnel. (2) Marketers' opinion survey method: mainly to learn about market trends from those engaged in marketing or market work. (3) Investigation method for trade fairs: to understand market information at trade fairs. (4) Consumer survey method: to judge market trends through consumer surveys.

2. Expert prediction method

It is a method that experts analyze the characteristics and the external environment to make judgments about the future state of things based on their own knowledge, experience, etc. The expert prediction method is simpler and more intuitive than the quantitative prediction method, and is suitable for predictions that lack sufficient historical data. There are three specific forms:

(1) Expert meeting method. It is to convene experts in related fields to participate in seminars, exchange opinions and information through on-site discussions, and make judgments on the development trend of the predicted objects.
(2) Brainstorming method. It is also to convene relevant experts to participate in meetings, through exchanges to stimulate innovative views or plans, so as to select the best plan. Unlike the expert meeting method, this method cannot easily negate other people's plans in use, but it can improve on others' plans and encourage innovative thinking.
(3) Delphi method. It is also known as the "back-to-back method", which uses letter inquiries to solicit opinions from various experts. In the process of making judgments, experts do not interact with each other and make independent predictions. Then, the investigators will summarize the prediction results, and then anonymously feedback the comprehensive opinions to the experts, and ask them to modify their own opinions based on the summary results of the previous round prediction result. Repeat this for several rounds to finally determine the prediction result. The advantage of the Delphi method is that each

expert can fully and independently express his own opinions, which eliminates the influence of a few authoritative experts on the prediction conclusions in the expert meeting method. But its disadvantage is that the prediction period is long and the efficiency is low.

3. Omen prediction

This is a prediction method to judge the future development trend of things through the analysis of the early development of things. The key to this method is that the predictioner needs to accurately understand the internal connection and causality between precursor phenomena and future trends, and grasp the law of change of things. "One leaf's falling tells the coming of autumn." is to judge the future trend of changes through the precursors of things.

4. Subjective probability

Subjective probability method is usually used in combination with expert prediction method, market survey method, etc., and the results predicted by other methods are expressed by subjective probability to make judgments. The so-called subjective probability refers to an expert's expectation of the degree of occurrence of an event in the future. There are two specific forms:

(1) Weighted average method of subjective probability. This method uses subjective probability as the weight, weights average of various prediction results, and calculates a comprehensive result.
(2) Cumulative probability median method. This method is a prediction method that determines the estimated number under different probabilities based on the cumulative probability, and then performs point estimation and interval estimation on the estimated number.

Example: Motor Sales Company T intends to prediction the company's car sales in 2017, through the subjective probability method for prediction.

(1) Prepare relevant information

The sales data of the car sales company in the past few years and the current market conditions and other relevant data are compiled for reference by relevant experts.

(2) Compile the subjective probability questionnaire

Send the subjective probability questionnaire to experts to fill in, as shown in Table 3.1.

In Table 3.1, the number 0.01 below Column (1) is the cumulative probability, and 2110 below is the prediction made by experts based on historical sales data

Table 3.1 Subjective probability survey table

Expert name(A)									
Cumulative probability	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	0.01	0.125	0.25	0.375	0.5	0.625	0.75	0.875	0.99
Sales volume	2110	2134	2165	2200	2223	2341	2365	2498	2512

analysis. The cumulative probability when the predicted sales volume is 2110 is 0.01, that is, the sales volume is less than 2110. The probability is only 1%.

The corresponding cumulative probability under Column (9) is 0.99, and the following 2512 is the auto sales volume predicted by experts with a cumulative probability of 0.99. The probability that the sales volume is greater than this value is only 1%. The corresponding cumulative probability under Column (5) is 0.5, and the following 2223 is the auto sales volume predicted by experts with a cumulative probability of 0.5. The probability that the sales volume is greater than and less than this value is 50%.

(3) Compile and sort out the questionnaire and make predictions

Summarize the questionnaire filled out by various experts and calculate the average of each column. In this case, a total of 5 experts (A, B, C, D, E) were surveyed, and the survey summary data is shown in Table 3.2.

From Table 3.2, it can be concluded that the minimum sales volume prediction of the car sales company for 2017 is 1985 (rounded up), and the probability of less than this value is 1%; the sales volume prediction of the car sales company for 2017, the maximum is 2507, and the probability of greater than this value is 1%; the predicted value of 2264 corresponding to the cumulative probability of 0.5 can be used as the predicted value of the sales volume of the company in 2017, which is the best estimate of the expected value of sales volume number.

5. Scenario prediction

Table 3.2 Subjective probability method for prediction automobile sales

Expert name	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	0.01	0.125	0.25	0.375	0.5	0.625	0.75	0.875	0.99
A	2110	2134	2165	2200	2223	2341	2365	2498	2512
B	1965	2030	2132	2210	2285	2356	2400	2476	2563
C	1856	2100	2143	2198	2298	2349	2456	2500	2514
D	2007	2200	2210	2298	2314	2347	2412	2450	2467
E	1986	2020	2133	2168	2198	2356	2389	2400	2478
Average	1984.8	2096.8	2156.6	2214.8	2263.6	2349.8	2404.4	2464.8	2506.8

Scenario prediction method is also called prospect analysis prediction method. It is a method that assumes that a certain trend should continue to exist for a period of time in the future, so as to make a comprehensive description of the future state of things. The method considers the problems comprehensively, and the analysis is flexible. It can combine qualitative and quantitative analysis to make more reliable judgments on the predicted objects, so as to discover and solve possible problems in time.

3.3 Quantitative Prediction Methods

3.3.1 Time Series Prediction

Time series refers to a series of numbers of statistical indicators arranged in chronological order, generally represented by $y_1, y_2, \ldots, y_t$, where t means time.

Time series consists of four types of factors: long-term trends, seasonal changes, cyclical changes and irregular changes.

The long-term trend means that in a relatively long period of time, due to the influence of certain fundamental factors, the time series shows a continuous upward or downward trend, and a tendency to stay at a certain level, reflecting the trend of things Main trends.

Seasonal changes refer to the cyclical changes in the data as the seasons change.

Cyclic changes generally refer to fluctuating changes with an unfixed cycle. Long cycle changes may be several years, short cycle changes may be several hours, and each cycle may not be exactly the same.

Irregular changes refer to sudden changes caused by accidental factors or exceptional events.

This section mainly introduces two time series prediction methods: moving average method and exponential smoothing (ES) method.

1. **Moving average method**

The moving average method is a simple smoothing prediction technology. The basic idea is to calculate the time series averages containing a certain number of items in order to reflect the long-term trend according to the time series data. The advantage of the moving average method is that it has a small amount of calculation and can better reflect the trend and change of the time series; the disadvantage is that the importance of recent periods of data is treated equally, sometimes not in line with the actual situation.

(1) Simple moving average method. Let the time series be $y_1, y_2, \ldots, y_t$, and the calculation formula of the simple moving average method is:

$$M_t = \frac{y_t + y_{t-1} + \cdots + y_{t-N+1}}{N}, t \geq N \tag{3.1}$$

In this formula, M_t is the moving average of t period; y_t is the observation value of the tth period; N is the number of terms in the moving average.

Here, N generally takes 3, 4, 5, 6 and other integers, for example, when N = 3, $M_t = \frac{y_t + y_{t-1} + y_{t-2}}{3}$, $t \geq 3$.

From formula (3.1), we know: $M_{t-1} = \frac{y_{t-1} + y_{t-2} + \cdots + y_{t-N}}{N}$.

$$M_t = M_{t-1} + \frac{y_t - y_{t-N}}{N} \tag{3.2}$$

The prediction formula is:

$$\hat{y}_{t+1} = M_t \tag{3.3}$$

That is, the moving average of the t period is used as the predicted value of the $t+1$ period.

Example 3.1 The sales volume of Company DT's products in recent years is as follows, please use the one-time moving average method to predict its sales volume in 2016. (N = 3 here, unit: thousand pieces) (Table 3.3).

Solution:

Calculated from the data in the above table:

$$\hat{y}_{t+1} = M_t^{(1)} = \frac{y_t + y_{t-1} + \cdots + y_{t-n}}{N}$$

$$\hat{y}_{2012} = M_{2011}^{(1)} = \frac{y_{2011} + y_{2010} + y_{2009}}{3} = \frac{2041 + 2028 + 1980}{3} = 2016;$$

$$\hat{y}_{2013} = M_{2012}^{(1)} = \frac{y_{2012} + y_{2011} + y_{2010}}{3} = \frac{2020 + 2041 + 2028}{3} = 2030;$$

…

Table 3.3 DT's product sales volume in recent years

Years	Sales volume y_t	One-time moving average M_t
2009	1980	
2010	2028	
2011	2041	
2012	2020	2016
2013	2035	2030
2014	2010	2032
2015	2015	2022
2016		2020

Similarly, there are

$$\hat{y}_{2016} = M_{2015}^{(1)} = \frac{y_{2015} + y_{2014} + y_{2013}}{3} = \frac{2015 + 2010 + 2035}{3} = 2020$$

Therefore, it can be concluded that the predicted value of DT's product sales in 2016 is 2020.

(2) Weighted moving average method. In the one-time moving average method, the influence of each period's data on the predicted value is regarded as equivalent, that is, their weights are the same, but the weights are often different in practical applications. It is generally believed that in predicting future conditions, recent data is more predictive than previous data. Therefore, when the recent data has a greater impact on the predicted value, the recent data should be given a greater weight, that is, the weighted moving average method.

Let the time series be $y_1, y_2, \ldots, y_t$, and the calculation formula of the weighted moving average method is:

$$M_{tw} = \frac{w_1 y_t + w_2 y_{t-1} + \cdots + w_N y_{t-N+1}}{w_1 + w_2 + \cdots + w_N}, t \geq N \tag{3.4}$$

In this formula, M_{tw} is the weighted moving average of t period, and w_i is the weight of y_{t-i+1}.

The prediction formula is:

$$\hat{y}_{t+1} = M_{tw} \tag{3.5}$$

That is, the weighted moving average of the t period is used as the predicted value of the $t + 1$ period.

Example 3.2 Using the data in Example 3.1, assign different weights to the data of Company DT for each period, set the w_1, w_2, w_3 as 0.5, 0.3, and 0.2 respectively, using the weighted moving average method to calculate (Table 3.4):

Table 3.4 DT's product sales in recent years

Years	Sales volume	Weighted moving average
2009	1980	
2010	2028	
2011	2041	
2012	2020	2025
2013	2035	2028
2014	2010	2032
2015	2015	2020
2016		2018

Solution:

$$M_{2015} = \frac{w_1 y_t + w_2 y_{t-1} + w_3 y_2}{w_1 + w_2 + w_3} = \frac{0.5 * 2015 + 0.3 * 2010 + 0.2 * 2035}{0.5 + 0.3 + 0.2} = 2018$$
$$\hat{y}_{2016} = M_{2015} = 2018$$

Therefore, we can conclude that the predicted value of DT's product sales in 2016 is 2018.

(3) Trend moving average method. When the time series has a linear increase or decrease trend, there will be a hysteresis deviation. The trend moving average method can correct it by a second moving average, and use the law of moving average hysteresis deviation to establish a linear trend prediction model.

Let the one-time moving average be $M_t^{(1)}$, and

$$M_t^{(1)} = \frac{y_t + y_{t-1} + \cdots + y_{t-N+1}}{N}$$

A moving average of $M_t^{(1)}$ is the second moving average $M_t^{(2)}$, the formula is

$$M_t^{(2)} = \frac{M_t^{(1)} + M_{t-1}^{(1)} + \cdots + M_{t-N+1}^{(1)}}{N} \tag{3.6}$$

Recursion gets

$$M_t^{(2)} = M_{t-2}^{(2)} + \frac{M_t^{(1)} - M_{t-N}^{(1)}}{N} \tag{3.7}$$

Assuming that the time series $y_1, y_2, \ldots, y_t$ has a linear trend starting from a certain period, and it is believed that the future period will also change according to this linear trend, then the linear trend prediction model can be set as

$$\hat{y}_{t+T} = a_t + b_t T \tag{3.8}$$

In this formula, is the number of current periods; T is the number of periods from t to the prediction period; a_t is the intercept and b_t is the slope, both are also called smoothing coefficients.

After derivation, the prediction formula can be obtained:

$$M_t^{(1)} - M_t^{(2)} = \frac{N-1}{2} b_t \tag{3.9}$$

$$\begin{cases} a_t = 2M_t^{(1)} - M_t^{(2)} \\ b_t = \dfrac{2}{N-1}\left(M_t^{(1)} - M_t^{(2)}\right) \end{cases} \tag{3.10}$$

2. **Exponential smoothing method**

Exponential smoothing (ES) method refers to a prediction method in which a simplified weighting factor, that is, a smoothing coefficient, is introduced on the basis of the actual number of a certain index and the predictioned number of the current period to obtain the average.

According to the different smoothing times, it can be divided into single exponential smoothing method, double exponential smoothing method and triple exponential smoothing method. The basic idea is: the predicted value is the weighted sum of the previous observations, and different weights are given to different data, the new data is given a larger weight, and the old data is given a smaller weight.

(1) Single exponential smoothing method. Suppose the time series is $y_1, y_2, \ldots, y_t$, then the exponential smoothing formula is

$$S_t^{(1)} = \alpha y_t + (1 - \alpha) S_{t-1}^{(1)} \tag{3.11}$$

In this formula, $S_t^{(1)}$ is an exponential smoothing value; α is a weighting coefficient, and $0 < \alpha < 1$.

The prediction model is:

$$\hat{y}_{t+1} = S_t^{(1)} \tag{3.12}$$

In this formula, $\hat{y}_{t+1}$ is the predicted value of period $t + 1$:

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha) \hat{y}_t \tag{3.13}$$

In other words, the predicted value of period $t+1$ is the exponentially smoothed value of period t.

Example 3.3 Knowing that DT's product sales volume in the last 8 months is shown in the following table, try exponential smoothing to predict the next month's sales volume y_9 (where $\alpha = 0.6$, unit: thousand pieces) (Table 3.5).

Table 3.5 DT's product sales volume in the last 8 months

Time serial number (t)	1	2	3	4	5	6	7	8
Sales volume (y_t)	10	15	8	20	10	16	18	20

Table 3.6 Calculation table of an exponential smoothing index

Time serial number (t)	1	2	3	4	5	6	7	8
Sales volume (y_t)	10	15	8	20	10	16	18	20
$S_t^{(1)}$	10.40	13.16	4.80	13.92	6.00	12.00	10.80	16.32

Solution:

Let the initial value be the average of the earliest three data, taking an exponential smoothing value calculation of $\alpha = 0.6$ as an example,

$$S_0^{(1)} = \frac{y_1 + y_2 + y_3}{3} = \frac{10 + 15 + 8}{3} = 11$$
$$S_1^{(1)} = \alpha y_1 + (1-\alpha)S_0^{(1)} = 0.6 * 10 + (1-0.6) * 11 = 10.4$$
$$S_2^{(1)} = \alpha y_2 + (1-\alpha)S_1^{(1)} = 0.6 * 15 + (1-0.6) * 10.4 = 13.16$$
$$\ldots$$
$$S_8^{(1)} = \alpha y_8 + (1-\alpha)S_7^{(1)} = 0.6 * 20 + (1-0.6) * 10.8 = 16.32$$

After calculation, Table 3.6 is obtained.

If the actual value of the 8th period is 20, the predicted value of the 9th period can be obtained

$$\hat{y}_9 = 0.6 * 20 + (1-0.6) * 16.32 = 18.53$$

(2) Double exponential smoothing method. Although the single exponential smoothing method overcomes the two shortcomings of the moving average method, it still needs to be corrected. The correction method is the same as the trend moving average method, that is, the double exponential smoothing is performed again, and then the straight-line trend prediction model is established. This is the double exponential smoothing method. The calculation formula is

$$\begin{cases} S_t^{(1)} = \alpha y_t + (1-\alpha)S_{t-1}^{(1)} \\ S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha)S_{t-1}^{(2)} \end{cases} \tag{3.14}$$

In this formula, $S_{t-1}^{(2)}$ is the double exponential smoothing value of the $t-1$ period; $S_t^{(2)}$ is the double exponential smoothing value of the t period; α is the weighting coefficient (also called the smoothing coefficient).

Double exponential smoothing method is a method of performing exponential smoothing again on single exponential smoothing. It cannot be predictioned alone, and must be used in conjunction with an exponential smoothing method. When the time series $y_1, y_2, \ldots, y_t$ has a linear trend from a certain period, similar to the trend moving average method, the following linear trend model can be used to predict:

$$\hat{y}_{11+T} = a_t + b_t T, T = 1, 2, 3, \ldots \tag{3.15}$$

Table 3.7 Non-operating expenses of DT Company from 2006 to 2013

Years	t	Operating expenses	$S_t^{(1)}$	$S_t^{(2)}$
	0		23.0	25.0
2006	1	29	27.8	27.2
2007	2	36	34.4	32.9
2008	3	40	38.9	37.7
2009	4	48	46.2	44.5
2010	5	54	52.4	50.8
2011	6	62	60.1	58.2
2012	7	70	68.0	66.1
2013	8	76	74.4	72.7

$$\begin{cases} a_t = 2S_t^{(1)} - S_t^{(2)} \\ b_t = \dfrac{\alpha}{1-\alpha}\left(S_t^{(1)} - S_t^{(2)}\right) \end{cases} \tag{3.16}$$

Example 3.4 Company DT's non-operating expenditure data from 2006 to 2013 is shown in Table 3.7. The exponential smoothing method is used to solve the trend line equation and predict the non-operating expenditure in 2016 (Set the initial values $S_0^{(1)}$ and $S_0^{(2)}$ to 23 and 25 respectively, $\alpha = 0.8$, T = 3, unit: ten thousand yuan).

Solution:
According to the title:
$S_0^{(1)} = 23$, $S_0^{(2)} = 25$, available:
$S_8^{(1)} = \alpha y_8 + (1-\alpha)S_7^{(1)} = 0.8 * 76 + (1-0.8) * 68 = 74.4$
$S_8^{(2)} = \alpha S_8^{(1)} + (1-\alpha)S_7^{(2)} = 0.8 * 74.4 + (1-0.8) * 66.1 = 72.7$
Then

$$\begin{cases} a_8 = 2 * S_8^{(1)} - S_8^{(2)} = 2 * 74.4 - 72.7 = 76.1 \\ b_8 = \dfrac{0.8}{1-0.8}(74.4 - 72.7) = 6.8 \end{cases}$$

The required model is

$$\hat{y}_{11+T} = 76.1 + 6.8T$$

The final conclusion: Company DT's non-operating expenditure prediction in 2016 is:

$$\hat{y}_{11+3} = 76.1 + 6.8 * 3 = 96.5$$

(3) Triple exponential smoothing method

When the change of the time series shows a quadratic curve trend, the triple exponential smoothing method is required. The triple exponential smoothing method is to perform another smoothing based on the double exponential smoothing. Its calculation formula is:

$$\begin{cases} S_t^{(1)} = \alpha y_t + (1-\alpha) S_{t-1}^{(1)} \\ S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \\ S_t^{(3)} = \alpha S_t^{(2)} + (1-\alpha) S_{t-1}^{(3)} \end{cases} \tag{3.17}$$

$S_t^{(3)}$ is the triple exponential smoothing value.

The prediction model of the cubic exponential smoothing method is:

$$\hat{y}_{t+T} = a_t + b_t T + c_t T^2 \tag{3.18}$$

In this formula:

$$\begin{cases} a_t = 3S_t^{(1)} - 3S_t^{(2)} + S_t^{(3)} \\ b_t = \dfrac{\alpha}{2(1-\alpha)^2}\left[(6-5\alpha)S_t^{(1)} - 2(5-4\alpha)S_t^{(2)} + (4-3\alpha)S_t^{(3)}\right] \\ c_t = \dfrac{\alpha^2}{2(1-\alpha)^2}\left[S_t^{(1)} - 2S_t^{(2)} + S_t^{(3)}\right] \end{cases} \tag{3.19}$$

3.3.2 Regression Analysis Prediction Method

Regression analysis originated from biological research and is a statistical analysis method to determine the quantitative relationship between two or among more variables. Its purpose is to estimate and predict the total average value of dependent variables based on known independent variables.

In the fields of economics and management, there are two main relationships among variables, namely functional relationship and correlation relationship. The former refers to the uniquely determined influence and dependence relationship between variables. Given the value of one or more independent variables, there is a uniquely determined value of the dependent variable corresponding to it. The latter refers to the relationship of mutual influence between variables. This relationship is not uniquely determined and is not strictly dependent. The two relationships are different, and also closely related. Regression analysis mainly focuses on correlation.

Regression models can have different classifications according to different methods. According to the number of independent variables, it can be divided into unary regression model and multiple regression model. According to whether the regression model is linear, it can be divided into linear regression model and non-linear regression model. In this section, the regression analysis prediction method is introduced according to the univariate linear regression prediction model and the multiple linear regression prediction model.

1. Unary linear regression prediction model

The unary linear regression model is the simplest regression analysis model. It is a predictive method that analyzes the linear relationship between a dependent variable and an independent variable. It is the basis for studying multiple linear regression models and nonlinear regression models.

(1) Unary linear regression model and basic assumptions. The univariate linear regression model has only one independent variable, and the function form is linear, which means that the dependent variable is a linear function of the independent variable. The function expression is as follows:

$$y_i = \beta_0 + \beta_1 x_i + \mu_i \tag{3.20}$$

Among them, x is called the independent variable or explanatory variable; y is called the dependent variable or the explained variable; β_0 and β_1 are two parameters, called regression parameters; μ is the total effect of all the factors that have a small influence on y except the independent variable x, called the random interference term. Due to the existence of the random interference term μ, for the value of a given independent variable, the value of the dependent variable is not uniquely determined, but random. Therefore, μ is a random variable. The main reasons for its occurrence are: ignoring some independent variables. Errors caused by variables, errors caused by inaccurate mathematical expressions of models, measurement errors of variable observations, random errors, etc.

The establishment of the unary linear regression model needs to meet the following five assumptions:

1. Normal distribution assumption: every random interference term $\mu_i (i = 1, 2, \cdots, n)$ is a random variable that obeys a normal distribution.
2. Zero-mean assumption: the expected value of each random interference term $\mu_i (i = 1, 2, \cdots, n)$ is zero.
3. Homovariance assumption: the variance of each random interference term $\mu_i (i = 1, 2, \cdots, n)$ is the same constant.
4. Non-autocorrelation assumption: the random interference items corresponding to different observations of independent variables are not correlated with each other.
5. The random interference term $\mu_i (i = 1, 2, \cdots, n)$ is not correlated with any observation value $x_i (i = 1, 2, \cdots, n)$ of the independent variable.

Among them, the first three assumptions can be summarized as each random interference term $\mu_i (i = 1, 2, \ldots, n)$ obeys a normal distribution with an expectation of 0 and a variance of σ_μ^2, which can be expressed as $\mu_i \sim N\left(0, \sigma_\mu^2\right)(i =$

$1, 2, \ldots, n$). The above five basic assumptions must be met when using the unary linear regression model to predict. When one or more assumptions are not met, the model may need to be processed and predicted or cannot be predicted. The processing method and process will be carried out in subsequent chapters.

(2) The least squares estimation method and the statistical properties of parameter estimators. The implicit distribution parameters of β_0, β_1, and μ_i in Formula 3.20, such as the variance σ_μ^2 of the random interference term, are all unknown, and the existing observations of x and y need to be used for estimation, in order to make predictions be based on the final regression equation.

The basic idea of parameter estimation is to use the method of sample trend fitting to find a sample regression line that is similar to the overall regression line, and get the approximate value of the parameter. Because in fact the overall regression line cannot be obtained, it only exists in theory. This needs to be discussed: how to make a straight line to make it the best estimate of the overall regression line, that is, find reasonable estimates of β_0 and β_1.

The most commonly used and more accurate method of parameter estimation is the least squares method, also known as ordinary least squares (OLS). Use $\hat{\beta}_0$ and $\hat{\beta}_1$ to represent the estimators of β_0 and β_1 respectively. At this time, the estimator $\hat{y}_i$ of the ith observation value of the dependent variable can be expressed as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3.21}$$

where $\hat{y}_i$ and the actual $\hat{y}_i$ are not necessarily equal, the difference between the two is called the residual, expressed by ε_i, so there is

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i \tag{3.22}$$

In formula (3.22), ε_i can be regarded as an estimate of the random interference term μ_i.

The basic idea of the least square method is: choose a straight line so that it is as close to each sample point as possible. Mathematically, it is to ensure that the sum of squares of the longitudinal distances from the n sample points to the straight line is minimized, and find the values $\hat{\beta}_0$, $\hat{\beta}_1$ of b and b when the sum of squares of ε_i to all i is the smallest, namely:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2 = \min \tag{3.23}$$

It can be seen from calculus that the necessary condition for $\sum_{i=1}^{n} \varepsilon_i^2$ to take a minimum is that its partial derivative with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ is equal to zero. After derivation

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \overline{y})}{\sum_{i=1}^{n} x_i (x_i - \overline{x})} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{3.24}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{3.25}$$

After processing the observations of x and y, they can be brought into Eqs. (3.24) and (3.25) to obtain the following estimation equations:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3.26}$$

Among them, $\hat{y}_i$ can be regarded as the estimator of the mean value of y_i, or as the estimator of y_i, that is, the predicted value.

The straight line fitted by OLS method has the following properties: the sum of residuals is equal to zero; the true value of the explained variable y and the fitted value have a common mean; the residuals are not related to the fitted value, that is, they are not mutually related influences.

In addition to OLS method, there are other methods to estimate the regression parameters β_0 and β_1, like finding the sum of the absolute values of the minimized residuals to estimate the regression parameters β_0 and β_1. The parameter estimators obtained by this method have some good statistical properties, namely linearity, unbiasedness and optimality, making the OLS method the most widely used one.

Linearity means that the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can be expressed as a linear combination of the explained variable y_i or the random interference term μ_i. Therefore, the parameter estimator and the explained variable and random interference items obey the same distribution; unbiasedness means that the mean value of the estimator is equal to the true value. Unbiasedness is a very important statistical property and the basis of statistical inference. Because the parameter estimator is a random variable centered on the true value of the parameter, the true value can be obtained by repeated sampling estimation. It can also be said that the estimator is equal to the true value from an average point of view; the optimality is also called the Gauss-Markov theorem, Means that among all linear unbiased estimators of the regression parameters, the least squares estimator has the smallest variance, also known as the best linear unbiased estimator (BLUE), which determines the general least squares method widely used.

(3) Interval estimation of regression parameters. From the above discussion, it can be seen that the variance of the least squares estimator of the regression parameter is related to the variance σ_μ^2 of the random interference term, but since σ_μ^2 is unknown, it is necessary to estimate σ_μ^2 before the interval estimation of the regression parameters. From the comparison of Eqs. 3.21 and 3.22, it can be seen that the residual ε_i can be regarded as the estimator of the random interference term μ_i, and the variance of ε_i can be used to estimate the variance σ_μ^2 of μ_i. After a simple derivation

$$E\left(\sum_{i=1}^{n} \varepsilon_i^2\right) = (n-2)\sigma_\mu^2 \tag{3.27}$$

Therefore, the unbiased estimator of σ_μ^2 is:

$$\hat{\sigma}_\mu^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2} \tag{3.28}$$

According to the nature of the least squares estimator, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of the random interference term μ_i. It can be known from the assumption that each μ_i obeys a normal distribution and is independent of each other. It can be deduced that the least squares estimator obeys, the true value of the parameter is centered, and a ratio of the variance of the random interference term is the normal distribution of the variance, which can be expressed as:

$$\begin{cases} \hat{\beta}_0 \sim N\left(\beta_0, D\left(\hat{\beta}_0\right)\right) \\ \hat{\beta}_1 \sim N\left(\beta_1, D\left(\hat{\beta}_1\right)\right) \end{cases} \tag{3.29}$$

among them,

$$\begin{cases} D\left(\hat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right)\sigma_\mu^2 \\ D\left(\hat{\beta}_1\right) = \frac{1}{\sum (x_i - \overline{x})^2}\sigma_\mu^2 \end{cases} \tag{3.30}$$

Standardize the normal distribution to get

$$\begin{cases} \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right)\sigma_\mu^2}} \sim N(0, 1) \\ \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{\sum (x_i - \overline{x})^2}\sigma_\mu^2}} \sim N(0, 1) \end{cases} \tag{3.31}$$

When performing interval estimation and hypothesis testing, you need to replace the unknown σ_μ^2 with $\hat{\sigma}_\mu^2$ in Eq. 3.31. At this time, the statistic obtained is no longer a normal distribution, but a t distribution. According to the nature of the t distribution, the confidence interval of the regression parameters β_0 and β_1 for $1 - \alpha$

$$\begin{cases} \hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n-2)\sqrt{\hat{D}(\hat{\beta}_0)} \\ \hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2)\sqrt{\hat{D}(\hat{\beta}_1)} \end{cases} \tag{3.32}$$

(4) Hypothesis testing of regression parameters. Hypothesis testing of regression parameters, also known as significance testing, evaluates the corresponding economic theory based on the evidence of quantitative estimation. If a random variable obeys a particular distribution, then the probability of its value range taking the midpoint of various ranges is certain. Therefore, it can be judged whether

the parameter level is reasonable according to whether its actual value appears in the range that usually appears (95% or 99% probability, etc.). In univariate linear regression analysis, hypothesis testing is mainly used to test whether there is a significant linear correlation between the independent variable and the dependent variable, that is, whether β_1 in Eq. 3.21 is equal to 0.

To conduct hypothesis testing, we must first establish basic hypotheses. Let the null hypothesis and alternative hypothesis be:

$$H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$$

Then construct the statistics and calculate the value of the statistics:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{D}(\hat{\beta}_1)}} \sim t(n-2) \tag{3.33}$$

Assuming H_0 holds, then:

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{D}(\hat{\beta}_1)}} \sim t(n-2) \tag{3.34}$$

Next, the significance level α is given, and $t_{\frac{\alpha}{2}}(n-2)$ is obtained by querying the percentile table of the t distribution, and finally, a decision is made by comparing the value of $t_{\frac{\alpha}{2}}(n-2)$ with the value of Eq. 3.34.

If $|t| > t_{\frac{\alpha}{2}}(n-2)$, then negate the original hypothesis and believe $\beta_1 \neq 0$, that there is a linear correlation between the dependent variable and the independent variable; if $|t| < t_{\frac{\alpha}{2}}(n-2)$, then accept the null hypothesis and believe that there is no linear correlation between the dependent variable and the independent variable.

(5) Correlation coefficient and goodness of fit. In the above discussion, according to the t test, $\beta_1 \neq 0$, it can only show that there is a linear correlation between the dependent variable y and the independent variable x. The model is reasonable, but the strength of the correlation and the fit of the model cannot be quantified. Research, which requires correlation coefficient and goodness of fit to be measured.

In correlation analysis, the correlation coefficient is a measure of the closeness of the linear relationship between two variables to reflect the closeness of the correlation between the variables. Remember the correlation coefficient

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \tag{3.35}$$

In the above formula, the value range of the correlation coefficient r is between -1 and 1, and the sign indicates the direction of change of the two variables. $r > 0$ means positive correlation, that is, the dependent variable increases with the increase of the independent variable; $r < 0$ means negative correlation, that is, the dependent variable decreases with the increase of the independent variable. The closer $|r|$ is to 1, the higher the degree of linear correlation between the two variables; the closer $|r|$ is to 0, the lower the degree of correlation between the two variables; when $|r| = 0$, it can only indicate that there is no linearity between the two variables. The relationship does not mean that there is no correlation between them. Whether there is a nonlinear relationship between them needs further discussion.

Goodness of fit refers to the degree of agreement between the sample regression line and the sample data trend. Although the ordinary least square method has good properties, it does not guarantee that the parameter estimation results of the specific model are ideal. Because the model assumptions may not be true, or the model settings are inappropriate. The goodness of fit is an important indicator to judge the authenticity and quality of the fitted line.

There will be a deviation between the measured value and the estimated value. Record $TSS = \sum (y_i - \overline{y})^2$ as the sum of squared deviations of y, which reflects the variation of the dependent variable y_i; record $ESS = \sum \left(\hat{y}_i - \overline{y}\right)^2$ as the regression sum of squares, TSS is explained by the regression equation of y to x explained part; remember $RSS = \sum \varepsilon_i^2$ as the residual sum of squares, that is, the part that is not explained by the regression equation. Proven

$$TSS = ESS + RSS \tag{3.36}$$

That is, the total square deviation is equal to the sum of the regression sum of squares and the residual sum of squares. Through analysis, we can see that the closer ESS is to TSS or the closer RSS is to 0, the regression equation shows that the total variation of y is more successful and the sample equation fits the observed data better. Therefore, the ratio of ESS to TSS can be used to measure the degree of fit between the sample regression equation and the observed data, that is, the goodness of fit or the coefficient of determination, denoted as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{3.37}$$

In regression analysis, the goodness of fit is more meaningful than the correlation coefficient, because it can indicate the degree to which the explanatory variable explains the change of the dependent variable, and it fully measures the degree to which one variable determines the other variable, but the correlation coefficient cannot, and is in the subsequent discussion of multiple regression analysis, the interpretation of the correlation coefficient is also uncertain. Therefore, the use of goodness of fit is more extensive.

(6) Point prediction and interval prediction. Two types of predictions can be made using a univariate linear regression model: point prediction and interval prediction.

Point prediction is to predict the value of a variable with a certain known quantity; interval prediction is the probability that the predictor variable falls within a certain interval. The larger the predicted interval, the greater the probability of falling within the interval. But relatively, the accuracy of prediction will also decrease.

Set up a univariate linear regression model as in Formula 3.21, assuming that the independent variable value x_f of the prediction period f is known, and Formula 3.21 is also applicable to the fth period, then

$$y_f = \beta_0 + \beta_1 x_f + \mu_f \tag{3.38}$$

The formula for point prediction is

$$\hat{y}_f = \hat{\beta}_0 + \hat{\beta}_1 x_f \tag{3.39}$$

Among them, $\hat{y}_f$ is not only the prediction of a single value of y_f, but also the prediction of the y_f mean value $E(y_f)$.

Interval prediction can be divided into single value interval prediction and mean value interval prediction.

To perform interval prediction on y_f, you need to first transform $\left(\hat{y}_f - y_f\right)$ into a statistic that obeys the standard normal distribution, and then replace the unknown σ_μ^2 with $\hat{\sigma}_\mu^2$ to obtain a statistic that obeys the t distribution:

$$t = \frac{\hat{y}_f - y_f}{\sqrt{\left[1 + \frac{1}{n} + \frac{(x_f - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right]\hat{\sigma}_\mu^2}} \sim t(n-2) \tag{3.40}$$

Finally, the confidence interval of $100(1-\alpha)\%$ for y_f is:

$$\hat{y}_f \pm t_{\frac{\alpha}{2}}(n-2)\sqrt{\left[1 + \frac{1}{n} + \frac{\left(x_f - \overline{x}\right)^2}{\sum (x_i - \overline{x})^2}\right]\hat{\sigma}_\mu^2} \tag{3.41}$$

To perform interval prediction on $E(y_f)$, you need to first transform $\left(\hat{y}_f - E(y_f)\right)$ into a statistic that obeys the standard normal distribution, and then replace the unknown σ_μ^2 with $\hat{\sigma}_\mu^2$ to obtain a statistic that obeys the t distribution

$$t = \frac{\hat{y}_f - E(y_f)}{\sqrt{\left[\frac{1}{n} + \frac{(x_f - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right]\hat{\sigma}_\mu^2}} \sim t(n-2) \tag{3.42}$$

Finally, the $100(1-\alpha)\%$ confidence interval of $E(y_f)$ is

$$\hat{y}_f \pm t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\left[\frac{1}{n} + \frac{\left(x_f - \overline{x}\right)^2}{\sum (x_i - \overline{x})^2}\right]\hat{\sigma}_\mu^2} \tag{3.43}$$

Table 3.8 Statistics of consumption level of urban residents and per capita disposable income of households from 2004 to 2013

Year	Index	
	Per capita disposable income of urban households (yuan)	Consumption level of urban residents (yuan)
2004	9421.6	8912
2005	10,493.0	9593
2006	11,759.5	10,618
2007	13,785.8	12,130
2008	15,780.8	13,653
2009	17,174.7	14,904
2010	19,109.4	16,546
2011	21,809.8	19,108
2012	24,564.7	21,035
2013	26,955.1	22,880

Source China Statistical Yearbook

Example 3.5 Table 3.8 shows the consumption level of urban residents and per capita disposable income of households in a certain region from 2004 to 2013. Try to cooperate with an appropriate regression model to conduct a significance test; if the per capita disposable income of urban households in 2014 is 30,000 yuan, try to predict the consumption level of urban residents in 2014.

Solution:

(1) Set the regression prediction model:

Let the per capita disposable income of urban households be x and the consumption level of urban households be y, then the unary linear regression model is

$$y = \beta_0 + \beta_1 x$$

(2) Collect and bring in sample data

10 sample data questions have been given.

(3) Estimate the parameters in the regression prediction model

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.812$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 1059.610$$

Then the model is

$$y = 1059.610 + 0.812x$$

(4) Test regression prediction model

First carry out the significance test and establish the basic hypothesis

$$H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$$

Find the value of the statistic

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{D}(\hat{\beta}_1)}} = 6.155$$

After checking the table, you can get: $|t| > t_{\frac{\alpha}{2}}(n-2)$, then negate the null hypothesis, thinking $\beta_1 \neq 0$, and there is a linear correlation between the dependent variable and the independent variable.

Then perform a goodness of fit test:

$$R^2 = \frac{ESS}{TSS} = 0.999$$

The fit of the model is very good.

Finally, the economic significance test is carried out. The meaning of the model is that when the per capita disposable income of urban households increases by one unit, the consumption level of urban residents increases by 0.812 units; when the per capita disposable income of urban households is 0, the consumption level of urban residents is 1059.61 yuan, the model fits the economic sense.

In summary, the model is reasonable and can be predicted.

(5) Economic prediction

$$y = 1059.610 + 0.812x$$

When the per capita disposable income of urban households $x = 30000$,

$$y = 1059.610 + 0.812x = 25419.61$$

Therefore, if the per capita disposable income of urban households is 30,000 yuan in 2014, it is predicted that the consumption level of urban residents in 2014 will be 2,5419.61 yuan.

2. Multiple linear regression prediction model

The multiple linear regression model is the addition of one or more independent variables on the basis of the unary linear regression model. It is a predictive method that analyzes the linear relationship between a dependent variable and two or more independent variables. This book uses the knowledge of vectors and matrices to describe them. In addition, in order to express more clearly, the vector symbols used in this chapter are handwritten, like vector $\vec{a}$.

(1) Multiple linear regression model and basic assumptions. The multiple linear regression model has multiple independent variables, and the function form is linear, which means that the dependent variable $f(x)$ is a linear function of the independent variable x. Assuming that there are k independent variables, the function expression is as follows:

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu_i \tag{3.44}$$

Among them, x is called the independent variable or explanatory variable, y is called the dependent variable or the explained variable, $\beta_1, \beta_2, \ldots, \beta_k$ is called the partial regression parameter, and μ is the random interference item. It is represented by matrix and vector symbols as follows:

$$\vec{Y} = \vec{\beta}\vec{X} + \vec{\mu} \tag{3.45}$$

among them:

$$\vec{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \vec{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

The establishment of the multiple linear regression model needs to satisfy the following six assumptions. Among them, the first five assumptions are the same as those of the one-variable linear regression model, but they can be expressed by vectors, so here they are not repeated. The sixth assumption is:

There is no strict linear relationship between the explanatory variables, namely $r(X) = k + 1 < n$.

The use of multiple linear regression model for prediction must meet the above six basic assumptions. When one or more assumptions are not met, the model may need to be processed and then predicted, or it cannot be predicted.

(2) The OLS method and the statistical properties of parameter estimators. The parameter estimation of the multiple linear regression model also uses the method. The OLS estimation formula is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} \tag{3.46}$$

where $\hat{y}_i$ and the actual y_i are not necessarily equal, the difference between the two has a residual ε_i, so there is

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} + \varepsilon_i \tag{3.47}$$

In formula (3.47), ε_i can be regarded as an estimate of the random interference term μ_i, which can be expressed as a matrix:

$$\vec{Y} = \vec{\hat{\beta}} \vec{X} + \vec{\varepsilon} \tag{3.48}$$

where $\vec{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$, represents the estimated amount of $\vec{\beta}$; $\vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ is the residual vector; the remaining vectors are the same as above.

From the basic idea of the least square method, the residual sum of squares is minimized, namely

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} \right) \right]^2 = \min \tag{3.49}$$

After a simple derivation, the least square estimation vector of the parameters can be obtained

$$\vec{\hat{\beta}} = \left(X^T X \right)^{-1} X^T Y \tag{3.50}$$

For multiple linear regression models, the parameter estimators obtained by the OLS method still have good statistical properties, namely linearity, unbiasedness, and optimality. The meaning of the properties is basically the same as that of the above-mentioned unary linear regression model.

(3) Hypothesis testing of regression parameters. Multiple linear regression analysis is slightly different from the hypothesis test in univariate linear regression analysis. It not only needs to test whether there is a significant linear correlation between each independent variable and the dependent variable when

other variables remain unchanged, that is, test arbitrary. Whether the parameter $\beta_i (i = 1, 2, \ldots, k)$ of is equal to 0, the application is still the t-test in the unary linear regression analysis, and the process is the same. And also another test is needed—F-test. The purpose of applying F-test is to judge whether the independent variable $x_1, x_2, \cdots, x_k$ has a linear effect on the dependent variable y at the same time, that is, whether the model is significant as a whole.

To perform F-test, you need to establish basic hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Then construct the statistics and calculate the value of the statistics:

$$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}} \sim F_\alpha(k, n - k - 1) \tag{3.51}$$

Then the significance level α is given, and $F_\alpha(k, n - k - 1)$ is obtained by querying the upper percentile table of the F distribution. Finally, a decision is made by comparing the value of $F_\alpha(k, n - k - 1)$ with the value of formula (3.51).

If $F > F_\alpha(k, n - k - 1)$, the original hypothesis is rejected, and the regression is considered to be significant as a whole, that is, there is a linear correlation between the dependent variable and all independent variables; if $F < F_\alpha(k, n - k - 1)$, the null hypothesis is accepted and the regression is considered to be insignificant as a whole, and the dependent variable and all independent variables are not significant. There is a linear correlation.

(4) Goodness of fit and modified goodness of fit. By extending the definition of the goodness of fit of the univariate linear regression model, the goodness of fit of the multiple linear regression model refers to the percentage of the sum of square deviations of the dependent variable that is explained by all the independent variables.

Following the expression method in the unary linear regression model, the sum of squared deviations of the multiple linear regression model

$$TSS = \sum (y_i - \overline{y})^2 = \vec{Y}^T \vec{Y} - n\overline{y}^2 \tag{3.52}$$

Regression sum of squares:

$$ESS = \sum \left(\hat{y}_i - \overline{y}\right)^2 = \vec{\hat{\beta}}^T \vec{X}^T \vec{Y} - n\overline{y}^2 \tag{3.53}$$

Residual sum of squares:

$$RSS = \sum \varepsilon_i^2 = \vec{\varepsilon}^T \vec{\varepsilon} \tag{3.54}$$

Proved that:

$$TSS = ESS + RSS \tag{3.55}$$

That is, the total square deviation is equal to the sum of the regression sum of squares and the residual sum of squares.

Define goodness of fit:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\vec{\hat{\beta}}^T \vec{X}^T \vec{Y} - n\overline{y}^2}{\vec{Y}^T \vec{Y} - n\overline{y}^2} \tag{3.56}$$

After analysis, it can be found that although this method imitates the definition method of goodness of fit in the unary linear regression model, it has problems when applied to the multiple linear regression model: the observation value y_i of a set of dependent variable y, regardless of the new added explanatory variable (independent variable) has a significant effect on y, the goodness of fit R^2 will increase with the increase in the number of explanatory variables (independent variables) in the model, and the goodness of fit is no longer a good description, and the degree to which the newly added independent variable explains the dependent variable cannot be used to determine the choice of independent variable. Therefore, it is necessary to modify the goodness of fit, so as to propose the concept of correcting the goodness of fit.

Define the modified goodness of fit:

$$\overline{R}^2 = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-1}} \tag{3.57}$$

The modified goodness of fit is done by dividing the numerator and denominator on the right end of the Eq. 3.56 by their respective degrees of freedom, thereby eliminating the influence of the number of variables on the goodness of fit. At this time, if $\overline{R}^2$ increases after adding an independent variable to the model, it can be considered that the independent variable has a significant impact on the dependent variable; conversely, if $\overline{R}^2$ does not increase after adding an independent variable to the model, it cannot be considered that the independent variable has a significant effect on the dependent variable.

(5) Point prediction and interval prediction. Like the univariate linear regression model, two types of predictions can be made using multiple linear regression models: point prediction and interval prediction.

Set up a multiple linear regression model such as Eq. 3.44, assuming that a certain set of independent variable values $x_{1f}, x_{2f}, \ldots, x_{kf}$ in the prediction period f are known, and Eq. 3.44 is also applicable to the fth period, then:

$$y_f = \beta_0 + \beta_1 x_{1f} + \cdots + \beta_k x_{kf} + \mu_f \tag{3.58}$$

The formula for point prediction is:

$$\hat{y}_f = \hat{\beta}_0 + \hat{\beta}_1 x_{1f} + \cdots + \hat{\beta}_k x_{kf} = \vec{X}_f^T \vec{\hat{\beta}} \tag{3.59}$$

Interval prediction can be divided into single value interval prediction and mean value interval prediction.

To perform interval prediction on y_f, it is necessary to construct statistics that obey the t distribution:

$$t = \frac{\hat{y}_f - y_f}{\sqrt{\hat{\sigma}_\mu^2\left(x_f^T (X^T X)^{-1} x_f + 1\right)}} \sim t(n-k-1) \tag{3.60}$$

Finally, the confidence interval of a single value of $(1-\alpha)$ is

$$\hat{y}_f \pm t_{\frac{\alpha}{2}}(n-k-1)\sqrt{\hat{\sigma}_\mu^2\left(x_f^T (X^T X)^{-1} x_f + 1\right)} \tag{3.61}$$

For interval prediction of $E(y_f)$, we still need to construct statistics that obey the t distribution:

$$t = \frac{\hat{y}_f - E(y_f)}{\sqrt{\hat{\sigma}_\mu^2 x_f^T (X^T X)^{-1} x_f}} \sim t(n-k-1) \tag{3.62}$$

Finally, the confidence interval for $(1-\alpha)$ single value of $E(y_f)$'s confidence level is:

$$\hat{y}_f \pm t_{\frac{\alpha}{2}}(n-k-1)\sqrt{\hat{\sigma}_\mu^2 x_f^T (X^T X)^{-1} x_f} \tag{3.63}$$

3.3.3 Trend Extrapolation

1. Definition of trend extrapolation method

The trend extrapolation method is a general term for inferring future development trends based on past and current conditions. The basis of the trend extrapolation method is the continuity and regularity of the prediction object over time.

When the development of the prediction object is mainly gradual with time, the prediction object will have a certain trend. If the predictioner can find a function curve reflecting this trend, he can establish time t as the independent variable. The trend extrapolation model where the sequence value y is the dependent variable:

$$y = f(t)$$

Based on this deduction, we can predict the future trend and state of things.

2. Types of trend extrapolation methods

(1) Linear trend method. When the long-term trend of the prediction object's time series basically shows a linear trend, the straight-line trend method can be used for prediction.
(2) Curve trend method. When the changing trend of the prediction object presents a curve of different shapes, the curve trend method can be used to fit, and the prediction can be made according to the time trend of the curve equation. Commonly used curve equations mainly include quadratic curve method, cubic curve method, Goppitz curve method and exponential curve method.
(3) Function model method. You can also use more mature functional models to fit the changing trends of the predicted objects, such as linear models, exponential curves, growth curves, and envelope curves.

3. Example: City X population prediction model

Select the data of the total permanent population of City X from 1978 to 2014, the data comes from the statistical yearbook of City X. Set the time variable t = year-1977, y = the permanent population of X city. Use MATLAB for data fitting.

Solution:

(1) Make a scatter plot based on the data in the table, as shown in Fig. 3.1.

(2) Based on historical data, perform linear fitting, polynomial fitting and exponential function fitting on the data respectively.

Linear model:

$$Y(t) = 34.11t + 684.1,\ R^2 = 0.9123,\ \text{RMSE} = 116.1;$$

Quadratic model:

$$Y(t) = 1.032t^2 - 5.111t + 939,\ R^2 = 0.9883,\ \text{RMSE} = 42.96;$$

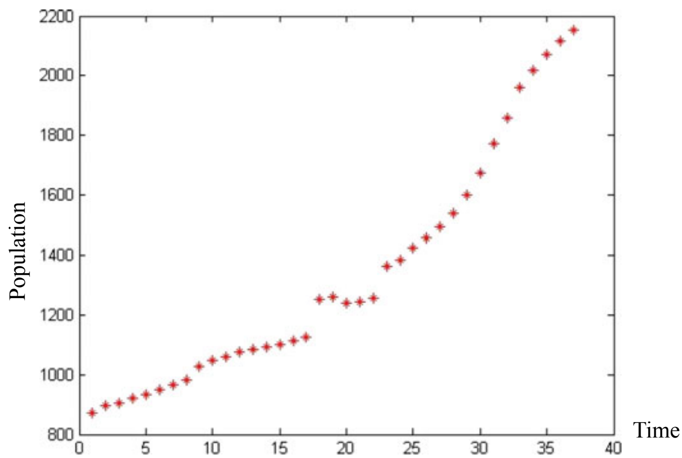


Fig. 3.1 Scatter plot of permanent residents in City X

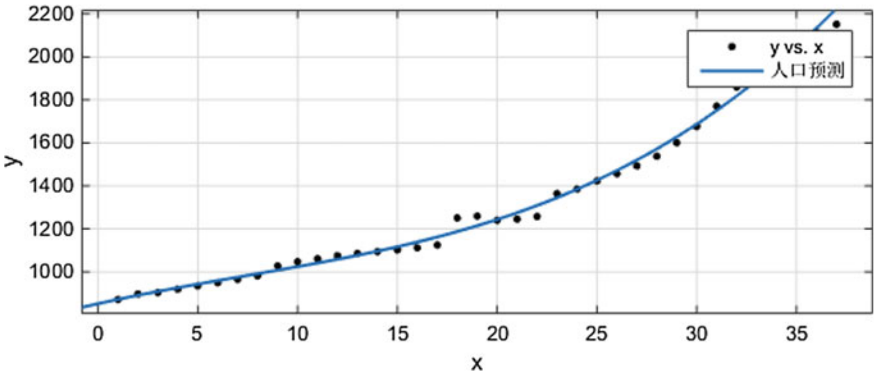


Fig. 3.2 Cubic model fitting curve

Cubic model:

$$Y(t) = 0.029t^3 - 0.645t^2 - 20.73t + 851.8, R^2 = 0.9937, \text{ RMSE} = 31.95;$$

Index model:

$$Y(t) = 790.1e^{0.004t} + 92.48e^{0.07t}, R^2 = 0.9929, \text{ RMSE} = 33.94.$$

Based on the above fitting process, the observation result shows that the fitting effect of the cubic model is the best. At this time, the fitting degree is 0.9937, and the RMSE is 31.95.

(3) According to the fitting effect of the above several prediction functions, the cubic model is selected as the final prediction function. The fitting effect is shown in Fig. 3.2.

According to the fitted model, we predict the situation of the permanent resident population in City X in the next 10 years, and the results are shown in Fig. 3.3. Of course, this kind of prediction should also take into account the possible impact of changes in other factors, such as changes in the permanent population policy, which may make the prediction results and the actual situation have a large deviation.

Chapter Summary

The primary function of management is planning, and prediction is the premise and basis of planning. Scientific prediction helps organizations improve the accuracy of planning and reduce the risk of plan implementation. The main points of this chapter include: prediction classification, moving average method, exponential smoothing method, regression analysis prediction method, trend extrapolation method.

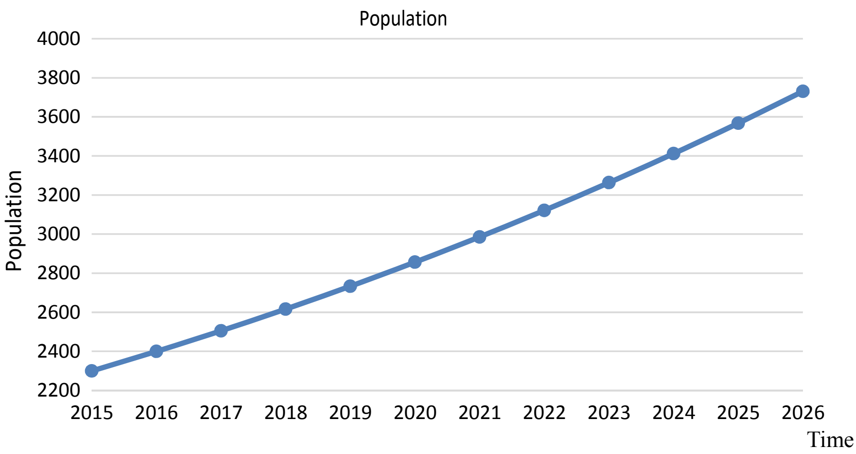


Fig. 3.3 Prediction of the permanent population of City X in the next 10 years

Important Concepts and Terms

- Prediction.
- Delphi Method.
- Subjective Probability Method.
- Moving Average Method.
- Exponential Smoothing, ES.
- Regression Analysis Prediction Method.
- Trend Extrapolation Method.
- Ordinary Least Square, OLS.
- Significance Testing.
- Goodness of Fit.

Questions and Exercises

1. What are the factors to consider when choosing a prediction method?
2. Briefly describe the meaning, principle, advantages and disadvantages of the moving average method.
3. Briefly describe the meaning, principle, advantages and disadvantages of exponential smoothing.
4. Regarding the application of the exponential smoothing method, the example question 3.4 in this chapter has been solved as $\alpha = 0.8$, the predicted value of Company DT's non-operating expenditure in 2016, then when $\alpha = 0.3$, $\alpha = 0.5$, how will the predicted value of DT's non-operating expenditure change? Explain the relationship between the exponential smoothing coefficient

Table 3.9 GDP of a certain city

Years	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
GDP	3557	4160	6765	7737	9207	10,325	11,972	13,904	16,014	17,616	19,213

Table 3.10 Per capita GDP and per capita disposable income of urban residents in a certain area

Years	GDP per capita	Disposable income per capita
2000	7902	6280
2001	8670	6859
2002	9450	7702
2003	10,600	8472
2004	12,400	9422
2005	14,259	10,493
2006	16,602	11,759
2007	20,337	13,786
2008	23,912	15,781
2009	25,963	17,175
2010	30,567	19,109
2011	36,018	21,809
2012	39,544	24,565
2013	43,320	26,955.0
2014	46,629	28,844

α and the predicted value, and consider what factors determine the exponential smoothing coefficient α ?

5. Given the GDP (unit: ten million yuan) of a city as shown in the table below, please select several different types of specific methods from the trend extrapolation method to predict the city's GDP development in the next five years, and compare the prediction results (Table 3.9).
6. The per capita GDP and per capita disposable income of urban households in a certain area are shown in the following table. Please use regression analysis and trend extrapolation to predict the data for the next five years. The specific methods can be selected by yourself and comparative analysis (Table 3.10).

Further Readings

1. Liu Sifeng. *Prediction Methods and Technology (2nd Edition)*. Beijing: Higher Education Press, 2015.

2. Wang Yu. *Classification prediction Method and Its Application in Economic Management Decision-making*. Beijing: Science Press, 2012.

Evaluation Methods

4

Hao Zhang

Learning Target

1. Understanding the definition, classification and process of evaluation, and understanding the connotation of typical evaluation methods.
2. Understanding the application conditions of the DEA, analytic hierarchy process, fuzzy comprehensive evaluation, entropy method and set pair analysis method introduced in the book of understanding, mastering the calculation process and being able to apply these methods to the practical evaluation problems.
3. Being familiar with the improved methods of evaluation or other typical evaluation methods introduced in the book.

Introduction

Controversial CPI

CPI is the abbreviation of consumer price index of residents. It is a relative number reflecting the trend and degree of price change of consumer goods and service items purchased by urban and rural residents in a certain period of time. It is the result of comprehensive summary and calculation of the consumer price index of urban residents and the consumer price index of rural residents. Through this index, we can observe and analyze the influence of retail price of consumer goods and price change of service items on the actual living expenses of urban and rural residents.

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_4

CPI has three main uses: First, as an economic index to measure inflation or contraction, it is often used to measure the general inflation of the whole economy, and provides the decision basis for the national macro-control. Second, it is used in national economic accounting. In the process of accounting, in order to eliminate the influence of price factors, the final current price consumption is reduced by CPI, and the final consumption of household constant price is obtained. Third, it is used for exponential adjustment. It is usually used to adjust monetary flows such as wages, interest, rent, taxes, reduce expenditure or monetary income at current prices, compensate in part or in whole for changes in the price or cost of living of consumer goods and services, and measure the true state of consumption and income.

When China's economy shows high growth and low inflation, no one doubts the rapid improvement of China's economic quality. But in the face of rising commodity prices every day, people have doubts about the stability of the CPI. The National Bureau of Statistics constantly adjusts the classification weight of CPI, and the focus of people's questioning is the weight setting of various commodity prices in CPI statistics. A white-collar worker at a foreign company complained to reporters: "oil prices are rising, house prices are rising, and the unit price of taxis has been raised successfully under heavy resistance. I don't understand why this CPI can continue to be stable. Although I don't understand CPI statistical methods, I think there are problems with statistical methods."

CPI is a very effective evaluation index, the compilation is also more complex. The formula of the CPI is:

$$\begin{aligned} \text{CPI} = {} & (\text{the value of a group of fixed commodities at the current price}/ \\ & \text{the value of a group of fixed commodities at the base price}) \times 100\%. \end{aligned}$$

The compilation includes: the determination of product basket, the collection of price data, the collection of weight data, the selection of compilation methods, the necessary adjustment of index compilation, etc. It covers the prices of food, housing, alcohol, tobacco and articles, clothing, household equipment and maintenance services, health care and personal effects, transportation and communications, recreational, educational, cultural and cultural goods and services consumed by urban and rural residents throughout the country. It is an international practice that the house price is not included in the statistics. In the compilation of our country, the cost method is used for the own housing of the residential category, that is to say, the consumption of the own housing in the residence is reflected by the estimated rent of the housing.

CPI reflects only changes in the prices of daily consumer goods and services purchased by residents, not all commodity prices. There is also a lot of controversy over whether housing prices should be included in the CPI. For ordinary people, buying housing is undoubtedly the largest consumption, but in CPI statistics, housing prices are not counted. This is incomprehensible to the common people.

Yi Xianrong, a researcher at the Institute of Finance of the Chinese Academy of Social Sciences, said that the existing housing consumption statistics do have the problem of low weight, and the method of price establishment is also open to question. Yi Xianrong told reporters, as to housing prices, the United States established a weight of 46%, Canada's weight of 36%, and our country now only 13%.

However, Professor Hua Ruxing of Tsinghua University expressed a different view: "Housing prices are transactions in the field of investment. We can't put the whole expenditure of residents in the CPI. Housing is the purchase of investment goods, all purchases of investment goods must be based on investment transactions, not in the CPI."

Yao Jingyuan, a special researcher at the State Council's Counsellor's Office and former chief economist of the National Bureau of Statistics, also said: "In economics and statistics, buying a house is called investment, so we don't call it consumption, so the price of a house is not in the CPI."

The CPI contains goods and services and weights will be controversial in the future, in order to enable CPI to evaluate the actual situation of consumption more truthfully.

In management, whether macro or micro, there are always a variety of evaluation problems. Scientific selection and design of evaluation methods and reasonable setting of evaluation indicators and weights are very important for managers to make correct decisions. However, if the evaluation method is not used properly, it will not only distort the actual situation, but also cause misunderstanding and contradiction.

Data: "Why house prices are not included in the CPI?", Yao Jingyuan, http://finance.cnr.cn/jjpl/201403/t20140307_515015087.shtml.

4.1 Overview of Evaluation Methods

4.1.1 Definition of Evaluation

Evaluation refers to the process of comprehensive investigation of various feasible schemes from the aspects of society, politics, economy, technology, etc., to determine the significance, value or state of the evaluation object, to give the evaluation results, and to provide scientific guidance for decision-making. It is essentially a judgment processing process and thus an important work in quantitative analysis. The purpose of evaluation is to realize the quantitative description of the whole level or function of the evaluation object through the quantitative measurement of the attributes of the evaluation object, so as to reveal the value or development law of things.

British scholar Francis Edgeworth published a paper "The Statistics of Examinations" in 1888, which was considered the earliest evaluation method. After the 1950s, with the development of multi-indicator comprehensive evaluation method, the dimensionless development of various indicators, formed the "weighted average of quantitative values" evaluation idea. After the '70s, a variety of evaluation methods are proposed and widely used, such as linear programming, analytic hierarchy process, data envelopment analysis, etc. After the '80s, With the introduction of fuzzy mathematics theory, grey system theory, information theory and so on, the convergence of evaluation methods began to emerge, various models and methods are proposed, such as fuzzy number based DEA model, stochastic DEA model, countermeasure DEA model, fuzzy matter element model and fuzzy analytic hierarchy process. After the '90s, a variety of deterministic and uncertain evaluation methods have been developed, such as fuzzy-AHP method, FHW decision system, grey-rough model combination application and so on.

4.1.2 Classification of Evaluation Methods

Evaluation methods can be divided into qualitative evaluation and quantitative evaluation.

Qualitative evaluation refers to the judgment of the nature, direction and degree of things made by the evaluators according to their own knowledge background and the actual situation and practical experience. The commonly used qualitative evaluation methods are as follows:

(1) Expert Meeting Law. Refers to the organization of a group of experts to participate in the forum, gradually formed a consensus evaluation method.
(2) Brainstorming. A method of stimulating creative thinking by organizing expert meetings. Through the direct exchange of information between experts and the full play of creative thinking, it is possible to obtain fruitful creative results in a relatively short time.
(3) Delphi method. Through several rounds of anonymous letters, the organizers collect and organize the opinions of each round, and then send them to each expert for their analysis and judgment, and put forward new arguments. Such a cycle, experts' opinions gradually converge, the reliability of the conclusion is also increasing.
(4) Comparison and evaluation of main indicators. It refers to a method to check the completion of the plan, analyze the reasons for the difference, and then tap the internal potential through the comparison of technical and economic indicators. When applying this method, attention must be paid to the comparability of technical and economic indicators.
(5) Logical framework method. This method provides an analytical framework for evaluators to analyze evaluation objects clearly and make them easier to understand. Starting with the determination of the core problems to be solved, we can expand up step by step, make clear its influence, deduce down layer by layer to find out the causes, and get the "problem tree". The causality described by the problem tree is transformed into the corresponding "means-goal" relationship to obtain the target tree.

Quantitative evaluation refers to the method of selecting and evaluating data resources from the angle of objective quantification according to the method of quantitative analysis and drawing the evaluation conclusion through mathematical calculation. The quantitative evaluation method makes the evaluation results more specific and reliable through objective quantitative analysis.

The commonly used quantitative evaluation methods are as follows:

(1) Analytic hierarchy process (AHP). According to the target-index system with hierarchical network structure, the weight of each evaluation object is obtained by comparing pairwise.
(2) Fuzzy comprehensive evaluation method. On the basis of fuzzy mathematics, the factors with unclear boundary are quantified and used for comprehensive evaluation.
(3) Multi-objective utility synthesis method. First, the goal-index system is established, starting from the bottom attribute, and the multilevel utility merging is carried out with certain rules, so as to better evaluate the research object.
(4) Simulation method. Through the simulation of the evaluation object, the impact of the implementation is evaluated from a dynamic and systematic point of view.
(5) Data envelopment analysis. According to the input data and output data of the evaluation object, the relative validity of the comparable units of the same type is evaluated.
(6) Computational intelligence methods. Through the study of a given sample, based on the biological system of biological evolution, cellular immunity, neural cell network and other mechanisms, using the abstract description of mathematical language, to obtain expert knowledge, experience, and subjective judgment, to establish a comprehensive evaluation method which is close to the qualitative and quantitative combination of human thinking mode.
(7) Combination evaluation method. A variety of evaluation methods are selected to improve the scientific and reliable evaluation process and results with the advantage of different evaluation methods.

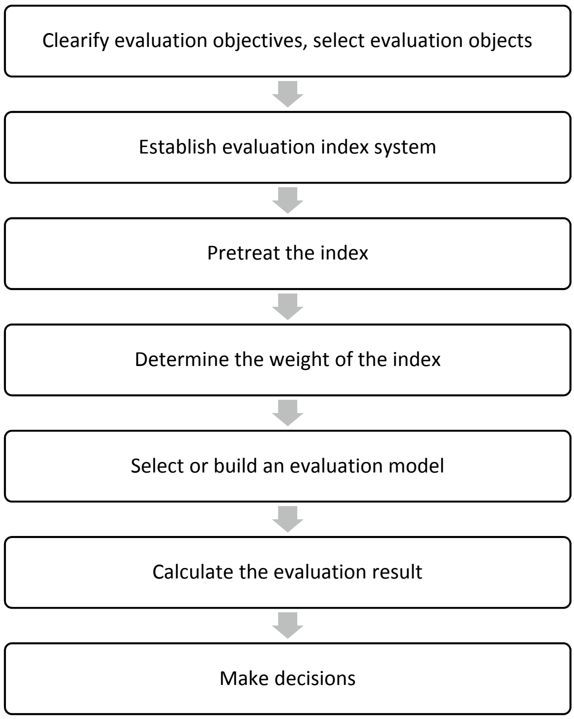
Qualitative evaluation methods require evaluators to have relevant knowledge and experience, while quantitative evaluation methods require a large number of reliable data. Therefore, scientific evaluation should combine qualitative and quantitative methods, complement each other, and carry out more systematic analysis and evaluation.

4.1.3 Evaluation Procedure

The evaluation process generally consists of the following basic steps (Fig. 4.1):

(1) Clarify evaluation objectives, select evaluation objects. According to the specific problems of decision makers, a clear evaluation goal is put forward, and the evaluation object is selected.
(2) Establish evaluation index system. Each index measures some characteristics or attributes of the evaluated object from different angles. The establishment of evaluation index system depends on specific evaluation problems, but generally follows the basic principles of integrity, science and maneuverability, including screening, optimization and testing of evaluation indicators.
(3) Pretreat the index. It includes the consistency of index types, nondimensionalization and non-differentiation of measure order of magnitude. In order to eliminate the comparability of each index and unify the trend of each index, the original data of each index should be dimensionless before evaluation, and the actual value of each evaluation index should be transformed into the evaluation value of the index.
(4) Determine the weight of the index. Weight is used to measure the relative importance of a single evaluation index in the whole evaluation system. The

Fig. 4.1 General process of evaluation



methods of determining weight mainly include subjective weighting method, objective weighting method and combination weighting method.

(5) Select or build an evaluation model. Through a certain model or algorithm, the evaluation a certain model or algorithm to obtain a holistic evaluation. The commonly used evaluation models mainly include linear evaluation model, nonlinear evaluation model and approximate rational point model.

(6) Calculate the evaluation results. By using the established model, the concrete evaluation results are obtained by calculation.

(7) Make decisions. According to the evaluation results, the selected evaluation object is made a decision.

4.2 DEA (Data Envelopment Analysis)

4.2.1 Definition of DEA

Data Envelopment Analysis (DEA) is a management method based on multi-input and multi-output to evaluate the relative effectiveness of work performance of the same type of organization (or project). It is based on multi-input and multi-output. DEA is widely used in the field of management.

This approach was proposed by leading American operations research scientists, University of Texas professors A. Charnes, W. W. Cooper and E. Rhodes in a joint article, “Measuring the efficiency of decision making units”. For short, the C^2R model, is used to evaluate the relative effectiveness of departments, that is, DEA effectiveness.

This method is often used to measure the relative efficiency of different organizations with the same goals, such as comparisons between the same type of company, business departments of a company, and cities of the same type. They have the same or similar inputs and outputs. Especially when there are many kinds of inputs (such as investment, personnel, land, equipment, etc.) and many kinds of outputs (such as profit, market share, innovation results, etc.), it is more suitable for applying DEA to carry out comparative analysis of inputs/outputs.

With the deepening of DEA research, C^2R, BC^2, FG and ST have become four classical DEA models. The C^2R model is specially used to judge whether the decision unit is technically effective and scale effective at the same time; the BC^2 model is used to judge whether the decision unit is technically effective; the FG model can not only evaluate the technical effectiveness of the decision unit, but also judge the non-decreasing problem.

Moreover, many scholars have carried out extended research on DEA models, such as GDEA model, comprehensive DEA model with “preference cone”, inverse DEA model, intersection DEA model, network DEA and so on, which enrich the types of DEA models.

Decision Making Units (DMU), an economic system or a production process can be regarded as a unit, which aims at maximizing benefits by inputting a certain number of factors of production and producing a certain number of results. Such units are called decision units.

DEA evaluation is based on input data and output data of the decision unit. According to these two kinds of data, the performance of decision-making units is evaluated, that is, the relative effectiveness of evaluation units. Through the comprehensive analysis of input and output, DEA can get the quantitative index of comprehensive efficiency of each DMU. According to this, each decision unit is graded and queued to determine the effective decision unit, and the causes and degrees of other decision units are given.

4.2.2 DEA Analysis Model

C^2R is an important model of DEA, which is mainly introduced in this chapter.

If the input vector of a DMU in a production activity is $x = (x_1, x_2, \ldots, x_m)^T$, the output vector is $y = (y_1, y_2, \ldots, y_s)$. Use (x, y) to represent the production activities of this DMU.

Assuming that there are n $DMU_j (1 \le j \le n)$, the corresponding DMU_j input and output vectors are:

$$\boldsymbol{x}_{\boldsymbol{j}} = \left(x_{1j}, x_{2j}, \ldots, x_{mj}\right)^T > 0, j = 1, 2, \ldots, n$$

$$\boldsymbol{y}_{\boldsymbol{j}} = \left(y_{1j}, y_{2j}, \ldots, y_{sj}\right)^T > 0, j = 1, 2, \ldots, n$$

and $x_{ij} > 0$, $y_{rj} > 0$, $i = 1, 2, \ldots, m$; $r = 1, 2, \ldots, s$

That is, each decision unit has m types of input and s types of output.

x_{ij} is the inputs into type i for the j decision unit;
y_{rj} is the output of the j decision unit to the r type.
x_{ij} and y_{rj} are known data can be obtained through historical data analysis or prediction.

Because of the different roles or effects between various inputs and outputs in the production process, to evaluate the DMU, it is necessary to make a comprehensive analysis of its input and output, so that each input and output need to be given the appropriate weight (see Fig. 4.2).

Because the information of input and output is not well understood, or we try to avoid the subjective influence of the analyst, we do not give the input and output weight vectors $\mathbf{v} = (v_1, v_2, \ldots, v_m)^T$, $\mathbf{u} = (u_1, u_2, \ldots, u_s)^T$ in advance, but first regard them as variable vectors, and then determine them according to some principle in the process of analysis. Here, v_i is a measure (weight) input for the i type; u_r is a measure (weight) output for the r type.

Because the weight coefficient of the importance of the item input is expressed $\mathbf{v} = (v_1, v_2, \ldots, v_m)^T$, the weight coefficient of the importance of the item output

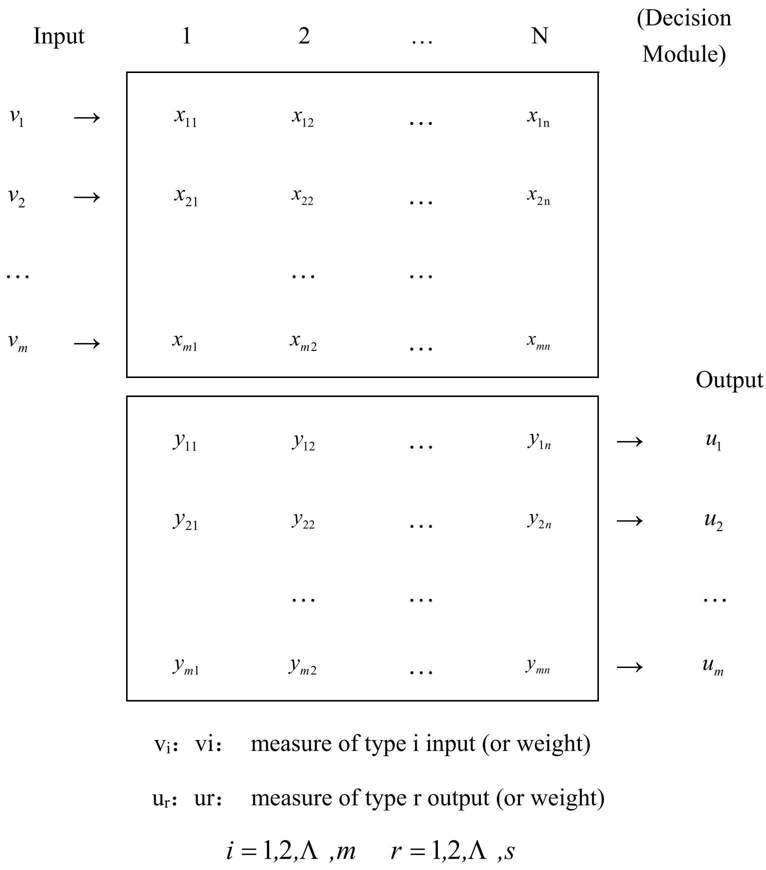


Fig. 4.2 DEA input–output block diagram

is expressed $\mathbf{u} = (u_1, u_2, \ldots, u_s)^T$, and at this time, under the weight **v** and **u**, it is transformed into a situation with an input and an output. Therefore, each decision unit DMUj has a corresponding relative efficiency value in terms of output per unit input (efficiency index):

$$h_j = \frac{u^T y_j}{v^T x_j} = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \quad j = 1, 2, \ldots, n$$

And we can take the weight coefficient v and u appropriately, so that $h_j \leq 1$.

C^2R is based on the invariance of scale income, that is, the output per unit of input is fixed and will not be changed by scale. In general, the larger the h_{j0}, it means the more DMU$_{j0}$ output can be obtained by relatively less input. To see DMU$_{j0}$ if it is optimal in these n DMU, you can calculate the maximum value h_{j0} when the weight changes.

If the efficiency index of the j_0 decision unit is taken as the target and the efficiency index of all the decision units (including the j_0 decision unit) is constrained, the relative efficiency optimization evaluation model is as follows:

$$\max h_{j_0} = \max \frac{\sum_{r=1}^{s} u_r y_{rj_0}}{\sum_{i=1}^{m} v_i x_{ij_0}}$$

$$st.\begin{cases} \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \le 1 \ j = 1, 2, \ldots, n \\ v = (v_1, v_2, \ldots, v_m)^T \ge 0 \\ u = (u_1, u_2, \ldots, u_m)^T \ge 0 \end{cases}$$

$v \ge 0$ means for $i = 1, 2 \ldots, m, v_i \ge 0$, and there's something $i_0(1 \le i_0 \le m), v_{i_0} > 0$. For the same meaning for $u \ge 0$.

The above formula is a fractional programming model, which must be transformed into a linear programming model to solve the problem, which can be:

$$t = \frac{1}{v^t x_0}, \quad \omega = tv, \quad \mu = tu$$

Then the linear programming model of C^2R can be obtained:

$$(P)\begin{cases} \max h_{j_0} = \mu^T y_0 \\ s.t. \omega^T x_0 - \mu^T y_j \ge 0 \quad j = 1, 2, \ldots, n \\ \omega^T x_0 = 1 \\ \omega \ge 0 \quad \mu \ge 0 \end{cases}$$

The above model can define the effectiveness of the j_0 decision unit by finding the optimal solution of linear programming.

When the model is used to evaluate the effectiveness of production efficiency of DMU_{j_0}, it is relative to all other decision units.

On the basis of the above derivation, the following mathematical model can be written, which is also the dual form D of planning P:

$$st.\begin{cases} \min \theta \\ \sum_{j=1}^{n} \lambda_j x_j \le \theta x_0 \\ \sum_{j=1}^{n} \lambda_j y_j \ge y_0 \\ \lambda_j \ge 0 \end{cases}, \quad \theta \text{ is no contraint}$$

By using the dual theory of linear programming, we can judge the validity of DMU_{j_0} by dual programming. In order to facilitate discussion and application, relaxation variables s^+ and residual variables s^- should be introduced, so the above inequality constraints can be transformed into equality constraints:

$$(D)\begin{cases} \min\theta \\ s.t.\ \sum_{j=1}^{n}\lambda_j x_j + s^+ = \theta x_0 \\ \sum_{j=1}^{n}\lambda_j y_j - s^- = y_0 \\ \lambda_j \geq 0 \quad j = 1,\ldots,n \\ s^+ \geq 0,\quad s^- \geq 0 \\ \theta \text{ is no constraint} \end{cases}$$

Linear programming (D) is directly called dual programming of planning (P).

Several theorems and definitions are given below in order to prepare for the application of later models.

Theorem 4.1 if linear programming (P) and dual programming (D) have feasible solutions, then their efficiency index must have optimal value. If the optimal values are respectively $h_{j_0}^*$ and θ^*, then $h_{j_0}^* = \theta^* \leq 1$.

Definition 4.1 if the optimal value of linear programming (P) $h_{j_0}^* = 1$, the decision unit DMU_{j_0} is called weak DEA effective.

Definition 4.2 if the solution of linear programming (P) exists, if and only if its optimal value $h_{j_0}^* = 1$, then the decision unit DMU_{j_0} is called DEA effective $\left(C^2R\right)$.

According to the definition of DEA effectiveness and weak DEA effectiveness, the relationship between them is as follows:

DEA effective ==> weak DEA effective

Theorem 4.2

(1) DMU_{j_0} is the sufficient necessary condition for weak DEA, which is the optimal value of linear programming (D) $\theta^* = 1$.
(2) DMU_{j_0} is the sufficient and necessary condition for DEA effectiveness, which is the optimal value of linear programming (D) $\theta^* = 1$, and for each optimal solution $\lambda^* {s^*}^- {s^*}^+ \theta^*$, there are ${s^*}^- = 0$, ${s^*}^+ = 0$.

Next, the economic meaning of DEA effectiveness is expounded. We can use the C^2R model to determine whether to achieve both technical effectiveness and scale effectiveness in production activities. The conclusions are as follows:

(1) If the optimal value of linear programming (D)$\theta^* = 1$, and the optimal solution satisfies $s^{*^+} = 0$ and $s^{*^-} = 0$. Then DMU_{j_0} called DEA effective. The DMU_{j_0} production activities have achieved both technical and scale effectiveness. s^+ represents the "deficit" of output and s^- represents the "excess" of input. At this time, there is no "excess" input and "deficit" output.
(2) When the optimal value of linear programming (D)$\theta^* = 1$, but $s^{*^-} > 0$ or $s^{*^+} > 0$, then DMU_{j_0} is called weak DEA effective. At this time, the DMU_{j_0} is not achieved technology and scale efficiency at the same time, technology and scale benefit of economic activities are not optimal at the same time, and some inputs may be redundant. In other words, there are still "excess" inputs in some aspects of production activities, and some outputs have "deficit".
(3) The DMU_{j_0} is not DEA valid if the optimal value of linear programming (D)$\theta^* < 1$. Its economic significance shows that the DMU_{j_0} technical efficiency of production activities is not optimal, and the scale efficiency is not the best, investment needs to be reduced to improve efficiency.

According to the measurement method of efficiency index, the DEA model can be divided into three types: input oriented, output oriented and non-conductive. The input-oriented type takes the input as the starting point to measure the degree of inefficiency of the DMU, and to the extent that technical inputs should be reduced under the condition of not reducing the output; the output-oriented type measures the degree of inefficiency of the DMU from the point of view of output, and achieves the degree of each output that should be increased without increasing the input; non-conductive model can measure from the point of output and input at the same time.

Check the DMU_{j_0} DEA validity, can use linear programming, but also dual linear programming. Neither method is convenient, and this test process can be simplified by constructing a slightly changed model. This is the C^2R model with non- Archimedes infinitesimal. This model can be used to determine whether the decision unit is DEA effective, weak DEA effective, or not DEA effective. Also, because of the diversity of activities in the actual production process, or the different role of decision makers in the evaluation activities, many scholars put forward some new DEA models on the basis of basic model C^2R.

DEA analysis steps:

(1) Clear evaluation objectives;
(2) Identify the target and select the DMU;
(3) Establish input index system and output index system respectively;
(4) Collection and collation of indicator data, qualitative indicators require quantitative assignment;
(5) Select the DEA model and calculate the evaluation results;
(6) Analysis of evaluation results and recommendations for decision-making.

4.2.3 Example

Datang Bank's four urban branches of the input and output as follows. Try to find out whether the operation of each branch DEA effective (Output unit: number of operational processing/months) (Table 4.1).

Solution: (1) In the constraints of linear programming DEA model, it is necessary to compare each element with the weighted average of all elements, so it is necessary to assume a synthetic branch as the weighted average of all elements. To make the Beijing branch relatively inefficient than the synthetic branch (the weighted average of the four branches), the output of the synthetic branch must be greater than or equal to that of the Beijing branch, while the input of the synthetic branch is less than or equal to that of the Beijing branch, that is, the synthetic branch has larger output and smaller input.

① In order to make the model logical, the output of the synthetic branch must be greater than or equal to that of the Beijing branch. That is:

$$\text{Output of synthetic branch} \geq \text{Output of Beijing branch}$$

② In order to make the model logical, the input of the synthetic branch must be less than or equal to that of the Beijing branch. That is:

$$\text{Input of synthetic branch} \leq \text{Input of Beijing branch}$$

③ Introduction of efficiency index E, if the number of employees in Beijing branch is 15, the number of employees in the synthetic branch is $15E$.

When $E = 1$, the synthetic branch needed the same investment resources as the Beijing branch;

When $E > 1$, the investment resources needed by the synthetic branch were larger than those of the Beijing branch;

When $E < 1$, the investment resources needed by the synthetic branch were smaller than those of the Beijing branch.

The logic of the DEA model is to seek whether a synthesis can achieve the same or more output with less input. If this synthesis can be obtained, a part of

Table 4.1 Input–output data for 4 branches

Branch	Input		Output		
	Number of employees	Operating area (m^2)	Deposit business	Loan transaction	Other business
Beijing	15	140	1800	200	1600
Zhengzhou	20	130	1000	350	1000
Shijiazhuang	21	120	800	450	1300
Wuhan	20	135	900	420	1500

the synthesis (such as the Beijing branch) will be judged to be less efficient than the synthesis branch.

(2) For the Beijing branch, the combined output value is $1800u_1 + 200u_2 + 1600u_3$, the total input value is $15v_1 + 140v_2$. The u_1, u_2, u_3, v_1, v_2 are weight coefficients of output and input respectively.

For the Bank of Beijing production efficiency of the highest optimization model as follows:

$$\max\ h_1 = \frac{1800u_1 + 200u_2 + 1600u_3}{15v_1 + 140v_2} \leq 1$$

$$st.\begin{cases} h_1 = \frac{1800u_1+200u_2+1600u_3}{15v_1+140v_2} \leq 1 \\ h_2 = \frac{1000u_1+350u_2+1000u_3}{20+130v_2} \leq 1 \\ h_3 = \frac{800u_1+450u_2+1300u_3}{21v_1+120v_2} \leq 1 \\ h_4 = \frac{900u_1+420u_2+1500u_3}{20v_1+135v_2} \leq 1 \end{cases} \tag{4.1}$$

Make $t = \frac{1}{15v_1+140v_2}$, $\mu_i = tu_i, \omega_i = tv_i$, the upper deformation is:

$$\max\ h_1 = 1800\mu_1 + 200\mu_2 + 1600\mu_3$$

$$st.\begin{cases} 1800\mu_1 + 200\mu_2 + 1600\mu_3 \leq 15\omega_1 + 140\omega_2 \\ 1000\mu_1 + 350\mu_2 + 1000\mu_3 \leq 20\omega_1 + 130\omega_2 \\ 800\mu_1 + 450\mu_2 + 1300\mu_3 \leq 21\omega_1 + 120\omega_2 \\ 900\mu_1 + 420\mu_2 + 1500\mu_3 \leq 20\omega_1 + 135\omega_2 \\ 15\omega_1 + 140\omega_2 = 1 \end{cases} \tag{4.2}$$

Its dual linear programming is:

$$\min V_D = E$$

$$st.\begin{cases} 15\lambda_1 + 20\lambda_2 + 21\lambda_3 + 20\lambda_4 \leq 15E \\ 40\lambda_1 + 130\lambda_2 + 120\lambda_3 + 135\lambda_4 \leq 140E \\ 1800\lambda_1 + 1000\lambda_2 + 800\lambda_3 + 900\lambda_4 \geq 1800 \\ 200\lambda_1 + 350\lambda_2 + 450\lambda_3 + 420\lambda_4 \geq 200 \\ 1600\lambda_1 + 1000\lambda_2 + 1300\lambda_3 + 1500\lambda_4 \geq 1600 \\ \sum_{i=1}^{4} \lambda_i = 1 \\ \lambda_i \geq 0 \end{cases} \tag{4.3}$$

(3) By the same way, the linear programming of Zhengzhou branch, Shijiazhuang branch and Wuhan branch can be established and solved by DEAP2.1 software. The results of each linear programming are shown in Table 4.2.

Table 4.2 Technical efficiency, pure technical efficiency and scale efficiency values of each branch

Branch	crste	vrste	scale	Return of scale
Beijing Branch	1.000	1.000	1.000	–
Zhengzhou Branch	0.892	0.966	0.924	Irs
Shijiazhuang Branch	1.000	1.000	1.000	–
Wuhan Branch	1.000	1.000	1.000	–

Note crste means technical efficiency, also known as comprehensive efficiency; vrste means pure technical efficiency; scale means scale efficiency (drs: scale returns decline; –: scale returns remain unchanged; irs: scale returns increase); crste = vrste × scale

For the Beijing branch, $E = 1$, it shows that the synthetic branch cannot obtain the same or more output of the Beijing branch without only less input, so the operation of the Beijing branch is DEA effective.

For the Zhengzhou branch, $E = 0.892$, this shows that the synthetic branch can obtain every output of the Zhengzhou branch and only use 89.2% of the input resources of the Zhengzhou branch. As a result, the Zhengzhou branch is relatively inefficient (or DEA invalid).

For Shijiazhuang branch, $E = 1$, this shows that the synthetic branch cannot obtain the same or more output of Shijiazhuang branch at the same time only need less input, so the operation of Shijiazhuang branch is DEA effective.

For Wuhan branch, $E = 1$, it shows that the synthetic branch cannot obtain the same or more output of Wuhan branch at the same time only need less input, so the operation of Wuhan branch is DEA effective.

4.3 AHP (Analytic Hierarchy Process)

4.3.1 Principle of AHP

Analytic hierarchy process (AHP), was proposed by Thomas L. Saaty, a famous American operations research scientist and professor at the University of Pittsburgh in the United States in the 1970s.

In real life, often encounter more complex decision-making problems. For example, there are three tourist destinations A, B and C, you will choose the tourist destination according to the cost, scenery, traffic condition, living condition, food condition and so on. AHP is a kind of multi-criteria decision-making method which combines qualitative and quantitative analysis to simulate the thinking process of a complex decision-making problem and to hierarchy and quantify the thinking process. The hierarchical analysis method has the characteristics of clear thinking, simple method, wide scope of application, strong systematicness, easy popularization, and provides a simple, rapid and practical solution to the general evaluation problems.

The basic idea of applying AHP to solve the problem is, first, to hierarchy the problem, that is, to decompose the problem into different factors, and to classify these factors into different levels of objectives, criteria, schemes, etc., to form a hierarchical, orderly, multi-level analytical structural model; secondly, to give a quantitative representation of the relative importance of each hierarchical factor in the structural model based on the judgment of objective reality, and then to use mathematical methods to determine the weights of the relative importance order of all factors in each layer; Finally, by synthetically calculating the weights of the relative importance of each layer of factors, the combined weights or ranking values of the relative importance order of the lowest layer (scheme layer) relative to the highest level (target layer) are obtained. As the basis of evaluation and selection, the final decision scheme is obtained.

When using analytic hierarchy process to analyze and solve problems, we usually go through the following steps: (1) to construct the hierarchical structure model; (2) to construct the judgment matrix; (3) to test the consistency of the judgment matrix; (4) hierarchical single sort; (5) hierarchical total sort; (6) to make decision. The implementation process of each step is specified below.

1. **Building Hierarchical Models**

First, the factors of the evaluation problem are decomposed into the objectives of the decision, the factors considered (decision criteria) and the decision options, and the hierarchical structure model is drawn according to the form of the highest, middle and bottom layers. As shown in Fig. 4.3.

(1) The highest level: the target layer, with only one element representing the purpose of the decision and the problem to be solved.
(2) The middle layer: the criterion layer, which represents the factors considered in decision-making and the criteria for decision-making.

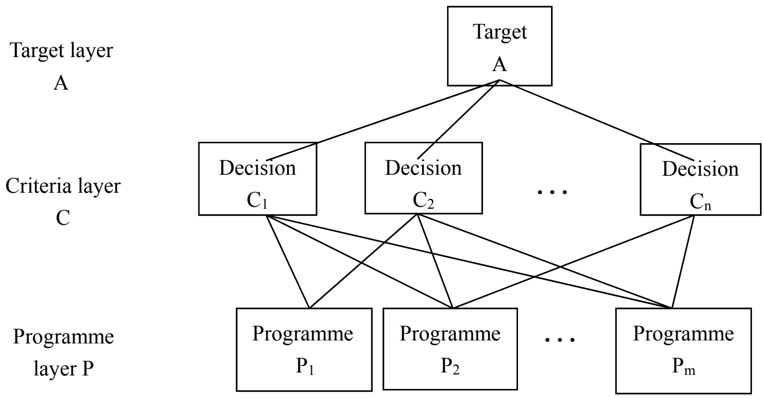


Fig. 4.3 Hierarchical model

(3) The bottom layer: the solution layer, which represents the solution to the problem.
(4) Connection: indicates the link between the upper and lower factors.

For adjacent two layers, the high layer is called the target layer, and the bottom layer is called the factor layer. The number of levels in the hierarchy model is related to the complexity of the problem and the degree of detail that needs to be analyzed, and the general number of levels is not limited. The number of elements in each level is usually not more than 9, because too many elements in the same level will bring difficulties to pairwise comparison judgment.

2. **Constructive judgment matrix**

Analytic hierarchy process (AHP) is mainly to judge the relative importance of each factor in each level. These judgments are expressed by introducing appropriate numerical values, and written into matrix form is the judgment matrix, which is established by the comparison of factors. Suppose that the factors at the previous level are A_k related to B_1, B_2, …, B_n next level, we should give B_1, B_2, …, B_n the corresponding weight according to the relative importance under the A_k. This process should consider which is more important B_i, B_j two factors for A_k and give a certain value to importance.

The pairwise comparison judgment matrix $\boldsymbol{B} = (B_{ij})_{n\times n}$ is obtained for n elements, where B_{ij} represent the important values of factor i and factor j relative to A_k. The judgment matrix is constructed as follows (Table 4.3).

A matrix $\boldsymbol{B}$ has the following properties:

(1) $b_{ij} > 0$;
(2) $b_{ij} = 1/b_{ji}$ $(i \neq j)$;
(3) $b_{ii} = 1$ $(i, j = 1, 2, \ldots n)$

In the AHP, in order to quantify the decision judgment and form the above numerical judgment matrix, the judgment is quantified according to a certain ratio scale. The following is the classical 1–9 scaling method, as shown in Table 4.4.

Table 4.3 A_k-B of the judgment matrix

A_k	B_1	B_2	…	B_n
B_1	b_{11}	b_{12}	…	b_{1n}
B_2	b_{21}	b_{22}	…	b_{2n}
⋮	⋮	⋮		⋮
B_n	b_{n1}	b_{n2}	…	b_{nn}

Table 4.4 Determination of matrix scale and its meaning

Order number	Level of importance	b_{ij} assignment
1	i, j two factors are equally important	1
2	i factors are more important than j factors	3
3	i factors are significantly important than j factors	5
4	i factors are extensively important than j factors	7
5	i factors are extremely important than j factors	9
6	i factors are less important than j factors	1/3
7	i factors are significantly less important than j factors	1/5
8	i factors are extensively less important than j factors	1/7
9	i factors are extremely less important than j factors	1/9
If the b_i assignment is 2, 4, 6, 8, or the reciprocal of each number, it has a similar meaning		

The basis of the assignment can be provided directly by the decision-maker, determined jointly by the decision-maker and the analyst, or obtained through expert advice.

3. **Judging Matrix Consistency Test**

Through the previous step, we have constructed the judgment matrix. For the judgment matrix B, if there is a $b_{ij} \cdot b_{jk} = b_{ik}$, for any i, j, k, then the matrix is called the complete consistency matrix. However, in dealing with practical problems, due to the complexity of judging objects and the limitations of human judgment ability, the judge may have the problem of inconsistent ranking of factors or inconsistent overall ranking when judging factors. For example, the emergence of A than B is extremely important, B than C is extremely important, and C is extremely important than A, which is obviously contrary to common sense. Therefore, in order to ensure that the conclusions obtained by AHP analysis conform to common sense, it is necessary to test the consistency of the constructed judgment matrix.

From the matrix theory, we can conclude that if the λ1, λ2, …, λn are satisfied the number of

$$Bx = \lambda x$$

that is, the characteristic root of the matrix B, and for all bii = 1,

$$\sum_{i=1}^{n} \lambda_i = n$$

When the matrix has complete consistency, $\lambda_1 = \lambda_{max} = n$, other characteristic roots are zero, and when the matrix B does not have complete consistency, $\lambda_1 =$

$\lambda_{max} > n$, other characteristic roots $\lambda_2, \lambda_3, \ldots, \lambda_n$ satisfy the relationship shown in the following formula:

$$\sum_{i=2}^{n} \lambda_i = n - \lambda_{max} \quad (4.4)$$

The above conclusion tells us that when the judgment matrix cannot guarantee complete consistency, the characteristic root of the judgment matrix will also change, so that the change of the characteristic root of the judgment matrix can be used to test the consistency of the judgment matrix. Therefore, the negative mean value of the other feature roots other than the maximum feature root of the judgment matrix is introduced in the hierarchical analysis method as an index to measure the deviation of the judgment matrix from the consistency.

$$CI = \frac{\lambda_{max} - n}{n - 1} \quad (4.5)$$

The larger the CI value, the greater the degree of deviation of the judgment matrix from the complete consistency; the smaller the CI value (close to 0), the better the consistency of the judgment matrix; CI = 0, when $\lambda_1 = \lambda_{max} = n$, the judgment matrix has complete consistency. However, it is obviously impossible to require each judgment matrix to be completely consistent, especially the problem of how large the factors are, so the judgment matrix is required to satisfy the consistency.

When the matrix ***B*** has satisfactory consistency, the λ_{max} is slightly larger than n, and the other characteristic roots are close to zero. In order to make the expression more accurate, a measure of satisfactory consistency should be given.

For different order judgment matrix, the judgment of consistency error is different, and the requirement of CI value is different. If the different order judgment matrix has satisfactory consistency, it is necessary to introduce the average random consistency index RI value of the judgment matrix. For the order 1–9 judgment matrix, the RI values are listed in Table 4.5 respectively.

For the first and second order judgment matrices, the RI is only formal. According to the definition of the judgment matrix, the first and second order judgment matrices always have complete consistency. When the order is greater than 2, the ratio of the consistency index of the judgment matrix CI, to the RI of the average random consistency index of the same order is called the random consistency ratio of the judgment matrix, which is recorded as the CR. When

$$CR = \frac{CI}{RI} < 0.10$$

Table 4.5 Average random consistency indicators

1	2	3	4	5	6	7	8	9
0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

it can be considered that the judgment matrix has satisfactory consistency, otherwise it is necessary to adjust the judgment matrix until it reaches satisfactory consistency.

4. **Hierarchical single sorting**

Hierarchical order refers to the weight of the order of importance of the factors associated with the previous layer according to the judgment matrix. Hierarchical order can be summed up as the maximum feature root of the judgment matrix and its feature vector problem. This chapter introduces a simple method to calculate the maximum characteristic root of matrix and its eigenvector: square root method. The calculation steps are as follows:

(1) Calculate the product M_i of each line of factors in the judgment matrix

$$M_i = \prod_{j=1}^{n} b_{ij} \quad i = 1, 2, \ldots, n$$

(2) Calculate the n root $\overline{W}_i$ of M_i

$$\overline{W}_i = \sqrt[n]{M_i}$$

(3) Normalize the vector $\overline{\boldsymbol{W}} = [\overline{W}_1, \overline{W}_2, \ldots, \overline{W}_n]^T$

$$W_i = \frac{\overline{W}_i}{\sum_{j=1}^{n} \overline{W}_j} \tag{4.6}$$

then $\boldsymbol{W} = [W_1, W_2, \ldots, W_n]^T$ is the desired eigenvector.

(4) Calculate the Maximum Characteristic Root λ_{max} of the Judgment Matrix

$$\lambda_{max} = \sum_{i=1}^{n} \frac{(AW)_i}{nW_i} \tag{4.7}$$

where A represents the judgment matrix, and $(AW)_i$ represents the i component of the vector $\boldsymbol{AW}$.

In addition to the root method, there are power method, sum product method and so on.

5. **Hierarchical Total Sort**

Table 4.6 Hierarchical total sort

Hierachy	A_1	A_2	…	A_m	Overall ranking of level B
	a_1	a_2	…	a_m	
B_1	b_1^1	b_1^2	…	b_1^m	$\sum_{i=1}^{m} a_i b_1^i$
B_2	b_2^1	b_2^2	…	b_2^m	$\sum_{i=1}^{m} a_i b_2^i$
⋮	⋮	⋮		⋮	⋮
B_n	b_n^1	b_n^2	…	b_n^m	$\sum_{i=1}^{m} a_i b_n^i$

By calculating the hierarchical structure from top to bottom in turn, we can calculate the relative advantages and disadvantages of the lowest factors relative to the highest evaluation target, that is, the total ranking of the hierarchy. Assume all factors A_1, A_2, …, A_m at the previous level total sorting is complete, a_1, a_2, …, a_m are the weights obtained, B_1, B_2, …, B_n of factors at this level corresponding to the a_i order results are:

$$b_1^i, b_2^i, \ldots, b_n^i$$

Here, if the B_j has nothing to do with A_i, then $b_j^i = 0$. The overall hierarchy is shown in Table 4.6.

Obviously,

$$\sum_{j=1}^{n} \sum_{i=1}^{m} a_i b_j^i = 1$$

So the hierarchical total sort is still a normalized normal vector.

6. **Make decision**

By calculating the ranking weights of the lowest schemes for the highest level targets, the options can be sorted by the weight size. It can also give each scheme a score and calculate the weighted average to get a comprehensive score, thus providing the basis for the evaluation scheme.

4.3.2 Example of AHP

4.3.2.1 Example

Example 4.1 The maximum characteristic root and its eigenvector of the following judgment matrix are calculated by square root method. The judgment matrix is shown in Table 4.7.

Solution: calculation steps by root method.

(1) Each line of elements multiplies to:

$$M_1 = 1 \times 2 \times 1/4 = 0.5 \quad M_2 = 1/2 \times 1 \times 1/6 = 0.083$$
$$M_3 = 4 \times 6 \times 1 = 24$$

(2) Obtain n second root to M_i:

$$\overline{W}_1 = \sqrt[3]{0.5} = 0.794 \quad \overline{W}_2 = \sqrt[3]{0.083} = 0.436 \quad \overline{W}_3 = \sqrt[3]{24} = 2.884$$

(3) Normalize the vector $\overline{\mathbf{W}} = (0.794, \quad 0.436, \quad 2.884)^T$ and the eigenvectors of the judgment matrix are:

$$\frac{0.794}{0.794 + 0.436 + 2.884} = 0.193$$
$$\frac{0.436}{0.794 + 0.436 + 2.884} = 0.106$$
$$\frac{2.884}{0.794 + 0.436 + 2.884} = 0.701$$

then $\mathbf{W} = (0.193, \quad 0.106, \quad 0.701)^T$.

(4) Calculating the maximum characteristic root λ_{max} of the judgment matrix:

$$\boldsymbol{AW} = \begin{bmatrix} 1 & 2 & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{6} \\ 4 & 6 & 1 \end{bmatrix} \begin{bmatrix} 0.193 \\ 0.106 \\ 0.701 \end{bmatrix}$$

Table 4.7 Judgment matrix

B	C_1	C_2	C_3
C_1	1	2	1/4
C_2	1/2	1	1/6
C_3	4	6	1

$$(AW)_1 = 1 \times 0.193 + 2 \times 0.106 + \frac{1}{4} \times 0.701 = 0.58$$

$$(AW)_2 = \frac{1}{2} \times 0.193 + 1 \times 0.106 + \frac{1}{6} \times 0.701 = 0.308$$

$$(AW)_3 = 4 \times 0.193 + 6 \times 0.106 + 1 \times 0.701 = 2.109$$

$$\lambda_{max} = \sum_{i=1}^{n} \frac{(AW)_i}{nW_i} = \frac{(AW)_1}{3W_1} + \frac{(AW)_2}{3W_2} + \frac{(AW)_3}{3W_3}$$
$$= \frac{0.58}{3 \times 0.193} + \frac{0.308}{3 \times 0.106} + \frac{2.109}{3 \times 0.701} = 3$$

The eigenvector of the matrix is $(0.193, 0.106, 0.701)^T$, and the maximum eigenroot λ_{max} is 3.

Example 4.2 Talent Selection Datang Company should select one of the three projects P_1, P_2, P_3 to invest. The criteria of the selection are technical ability, financial situation, environmental protection ability, organizational ability, economic strength and risk control ability. Comprehensive evaluation and quantitative ranking of these three projects are carried out by AHP method. A judgment matrix has been given, as shown in Tables 4.8, 4.9, 4.10, 4.11, 4.12, 4.13 and 4.14.

Solution:
According to the general steps of analytic hierarchy process.

Table 4.8 Judgment matrix A-C

A	C_1	C_2	C_3	C_4	C_5	C_6
C_1	1	1/2	1	3	2	1/2
C_2	2	1	2	4	4	1/6
C_3	1	1/2	1	5	3	1/3
C_4	1/3	1/4	1/5	1	1/3	1/4
C_5	1/2	1/4	1/3	3	1	1/5
C_6	2	6	3	4	5	1

Table 4.9 Judgement matrix C_1-P

C_1	P_1	P_2	P_3
P_1	1	1/5	1/3
P_2	5	1	3
P_3	3	1/3	1

Table 4.10 Judgment matrix C_2-P

C_2	P_1	P_2
P_1	1	1/5
P_2	5	1

Table 4.11 Judgment matrix C_3-P

C_3	P_1	P_2	P_3
P_1	1	1/6	3
P_2	3	1	7
P_3	1/3	1/7	1

Table 4.12 Judgment matrix C_4-P

C_4	P_1	P_2	P_3
P_1	1	1/3	5
P_2	3	1	7
P_3	1/5	1/7	1

Table 4.13 Judgment matrix C_5-P

C_5	P_1	P_2	P_3
P_1	1	1/2	2
P_2	2	1	7
P_3	1/2	1/7	1

Table 4.14 Judgment matrix C_6-P

C_6	P_1	P_2	P_3
P_1	1	6	9
P_2	1/6	1	3
P_3	1/9	1/3	1

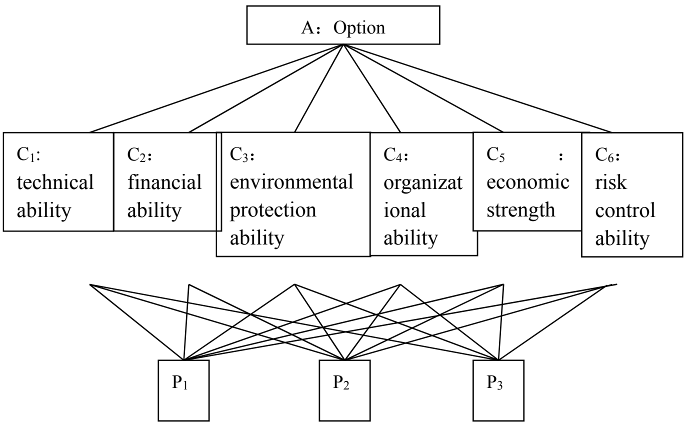


Fig. 4.4 Hierarchical model for selecting a project

Table 4.15 Maximum characteristic roots of each factor

Eigenvalue	Technical ability	Financial ability	Environmental protection ability	Organizing ability	Economic strength	Risk control ability
λ_{max}	3.04	2.00	3.10	3.06	3.03	3.05

(1) Build a hierarchical model, as shown in Fig. 4.4.
(2) Constructive judgment matrix.

A judgment matrix is shown in Tables 4.8, 4.9, 4.10, 4.11, 4.12, 4.13 and 4.14.

(3) Judging Matrix Consistency Test and Hierarchical Single Order

For example, Judgment matrix A-C, the maximum characteristic root $\lambda_{max} = 6.56$ is obtained by square root method, CI = 0.112, RI = 1.24, CR = 0.0903 < 0.10, Judgment matrix A-C pass the consistency test and get the eigenvector of the matrix $\boldsymbol{W} = [0.13, 0.19, 0.14, 0.04, 0.07, 0.42]^T$.

Similarly, the maximum characteristic roots of the judgment matrix C_i-P (i = 1, 2, …, 6) of the six criteria for the three items are shown in Table 4.15.

Through the calculation, we can pass the consistency test, and the corresponding eigenvectors are:

$$\boldsymbol{W_1} = [0.10, 0.64, 0.26]^T$$

$$\boldsymbol{W_2} = [0.17, 0.83]^T$$

$$\boldsymbol{W_3} = [0.17, 0.75, 0.07]^T$$

$$\boldsymbol{W_4} = [0.28, 0.65, 0.07]^T$$

$$\boldsymbol{W_5} = [0.36, 0.63, 0.11]^T$$

$$\boldsymbol{W_6} = [0.77, 0.16, 0.07]^T$$

Table 4.16 Hierarchical total ordering of a one-selected item

Hierachy	C_1	C_2	C_3	C4	C5	C_6	General ordering at the P level
	0.13	0.19	0.15	0.04	0.07	0.41	
P_1	0.1	0.17	0.17	0.28	0.36	0.77	0.43
P_2	0.64	0.83	0.75	0.65	0.63	0.16	0.49
P_3	0.26	–	0.07	0.07	0.11	0.07	0.08

(3) Hierarchical Total Sort

According to the above calculation, the total ranking is shown in Table 4.16.

$$0.13 \times 0.1 + 0.19 \times 0.17 + 0.15 \times 0.17 \\ + 0.04 \times 0.28 + 0.07 \times 0.36 + 0.41 \times 0.77 \approx 0.43$$

$$0.13 \times 0.64 + 0.19 \times 0.83 + 0.15 \times 0.75 \\ + 0.04 \times 0.65 + 0.07 \times 0.63 + 0.41 \times 0.16 \approx 0.49$$

$$0.13 \times 0.26 + 0.15 \times 0.07 + 0.04 \times 0.07 + 0.07 \times 0.11 + 0.41 \times 0.07 \approx 0.08$$

The weight of each factor C the target A is

$$\boldsymbol{W} = [0.43, 0.49, 0.08]^{\mathrm{T}}$$

At the time of selecting an item, the three items are sorted as follows: P_2, P_1, P_3. You can choose the first project to invest.

4.3.2.2 Yaahp Software Profile

Yaahp (Yet Another AHP) is an analytic hierarchy process auxiliary software, which has the functions of hierarchical model construction, judgment matrix data input, ranking weight calculation and analysis, data export and so on. Application of Yaahp software can easily complete the task of AHP, fuzzy comprehensive evaluation and multi-criteria decision analysis. The latest version of the software interface is shown in Fig. 4.5.

Applying Yaahp V10.3 solving AHP problem mainly includes three steps: AHP model, judgment matrix and calculation result. First, we should construct the hierarchical model, then input the judgment matrix involved in the hierarchical model, and finally find out the calculation results. It is worth noting that when using Yaahp V10.3 software to solve AHP, the next step can only be carried out after the completion of the previous step. The following software is used to solve the Example 4.2.

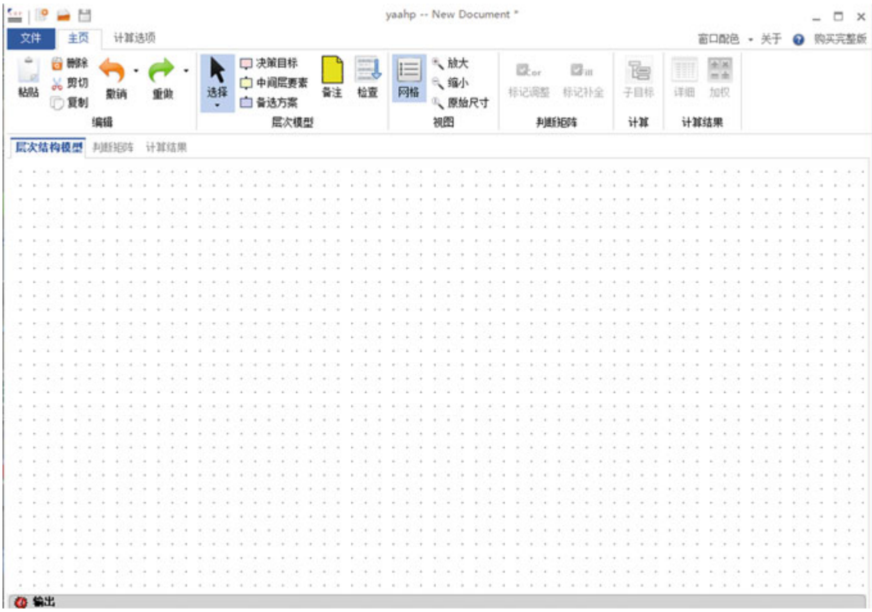


Fig. 4.5 Yaahp V10.3 interface (in Chinese)

1. **Building Hierarchical Models**

First, the hierarchical structure model is constructed. Enter the decision target, intermediate element layer, options and connect the adjacent two layers with the arrow pointing to the upper layer in the next layer. The hierarchical model is shown in Fig. 4.6.

2. **Input Judgment Matrix**

Enter the matrixes A-C, C_1-P, C_2-P, C_3-P, C_4-P, C_5-P, C_6-P separately, and each input matrix software automatically judges the consistency of the matrix (Fig. 4.7).

3. **Calculation result**

When all judgment matrices pass the consistency test, the software outputs the calculation results. As with Example 4.2, when selecting a project, the order of the three investment projects is: P_2, P_1, P_3. You can choose P_2 project to invest (Fig. 4.8).

The weights of the factor C to the target A are calculated by Yaahp software

$$\boldsymbol{W} = [0.4358, 0.4772, 0.0869]^{\mathrm{T}}$$

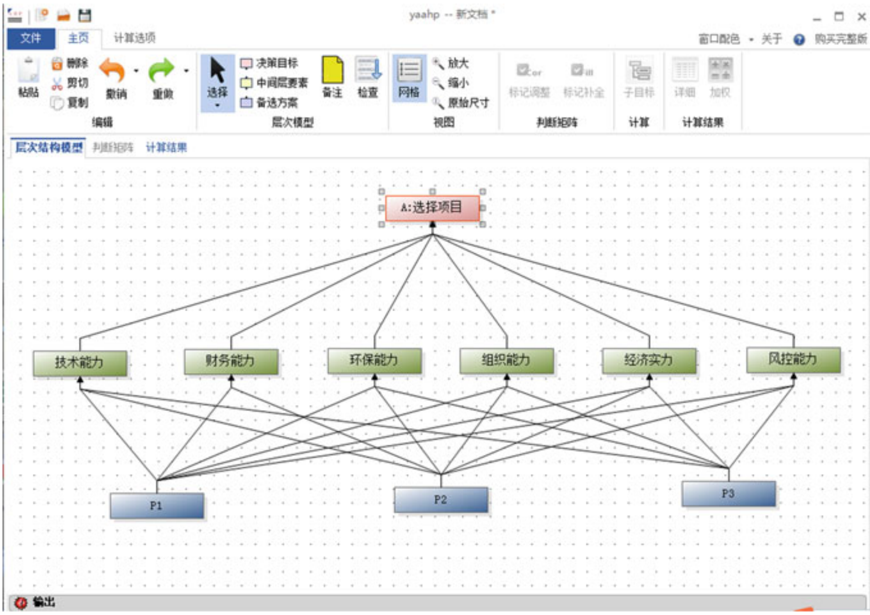


Fig. 4.6 Hierarchical model for selecting an investment project

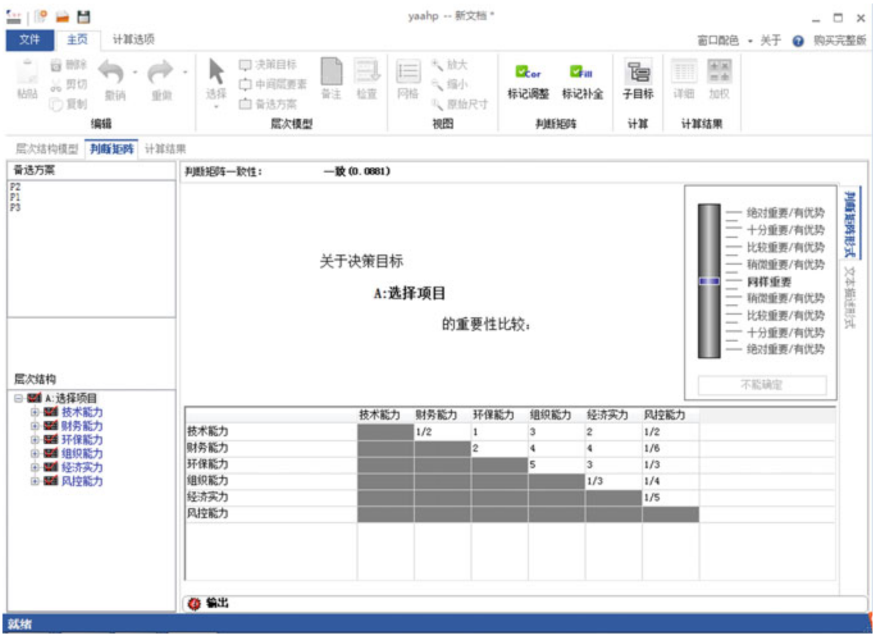


Fig. 4.7 Input judgment matrixes A–C

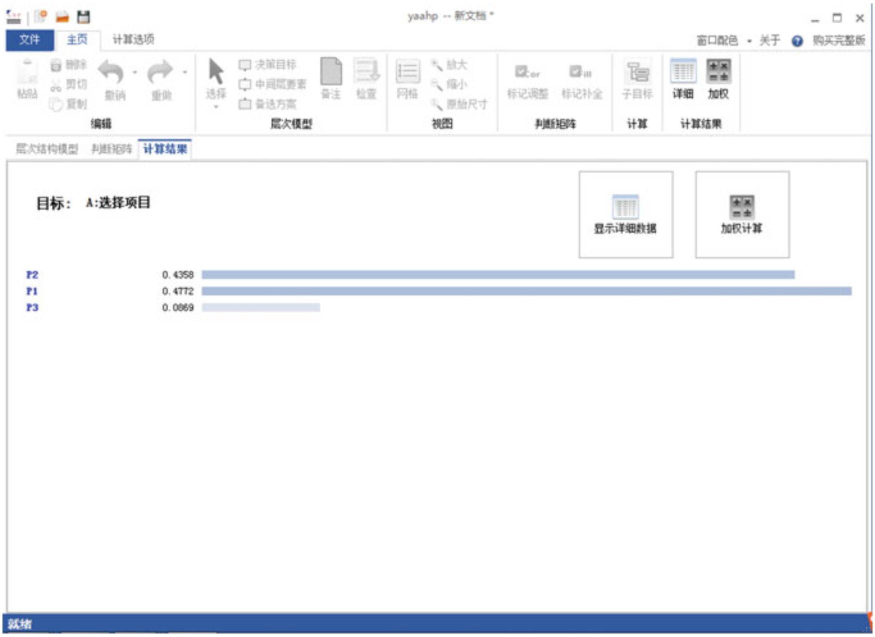


Fig. 4.8 Calculation results of the optional items

The reason why there is an error with the weights calculated above: the eigenvector calculated above keeps two decimal places all the time, and the weight of the calculation is also reserved for two decimal places, so there is an error.

4.3.3 Analytic Network Hierarchy Process

Professor Saaty proposed the Analytic Network Process (ANP) on the basis of AHP in the 1990s. ANP is often used to solve the system evaluation problem with network structure, which is a kind of evaluation method suitable for non-independent progressive hierarchical structure. The relationship between the indexes of the evaluated object is expressed as a similar network structure. Instead of a simple progressive hierarchy. ANP can more accurately describe this influence relationship when there is a possible interaction and mutual domination between indicators, it is a more effective evaluation method. During the analysis of application ANP, it is necessary to judge the relative importance of each index. In reality, it is often not to judge the relative importance of all evaluation indicators, but to judge the relative importance of some of them according to the information they have. At this time, there will be some vacancies in the pairwise judgment matrix. This situation is called incomplete information.

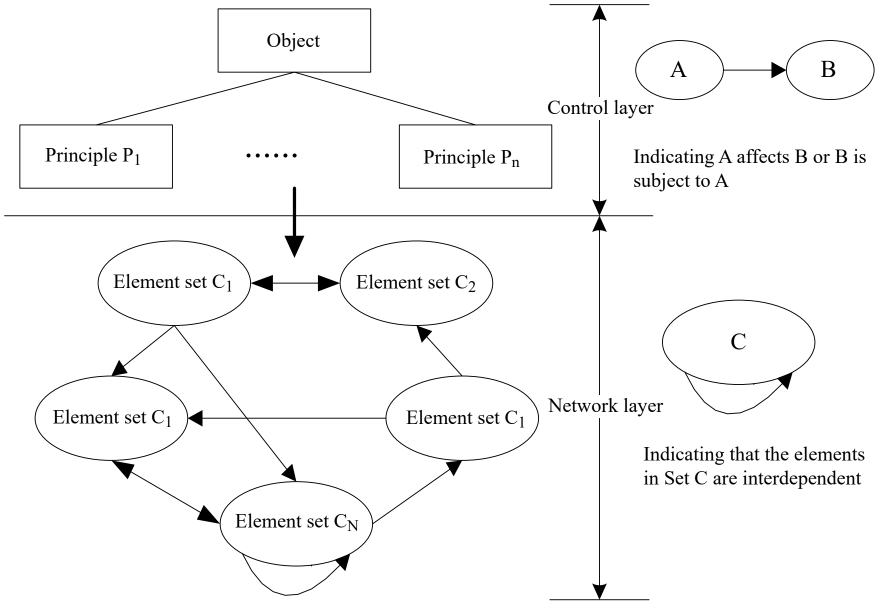


Fig. 4.9 ANP structure diagram

ANP considering the internal cycle of the progressive hierarchy and its dependence and feedback, the system elements (or evaluation index system) are divided into two parts. The first part is called the control layer, including the target layer and the criterion layer. The criteria are considered independent of each other and dominated by the target layer. There can be no guidelines in the control layer, but there should be at least one goal. If the control layer contains multiple criteria, the weight of each criterion can be obtained by the AHP method and the second part is the network layer, which is composed of all the elements dominated by the control layer and its interior is a network structure that affects each other and is interrelated. Figure 4.9 is a typical ANP structure. To make ANP method easy to calculate, Rozann W. Satty and William Adams introduced Super Decision software.

Example 4.3 Application of ANP to evaluate the logistics service level of Datang Company.

First, analyze the problem and form the evaluation index system.

When evaluating the logistics service level of Datang Company, the main indicators needed are: (Table 4.17).

1. Reserve capacity (B)

Table 4.17 Evaluation index system of logistics service level of Datang Company

Level 1 indicators	Level 2 indicators	Level 3 indicators
Logistics service level	Reserve capacity B	Inventory capacity B1
		Amount of reserve B2
		Variety assurance rate B3
	Traffic capacity T	Traffic volume T1
	Organizing ability C	Scale of organization C1
		Division of labor and coordination C2

Reserve capacity includes inventory capacity (B1), reserve amount (B2), variety assurance rate (B3).

2. Transport capacity (T)

Transport capacity (T1) is measured by volume of transport.

3. Organizational capacity (C)

Organizational capacity includes organizational size (C1), division of labor and coordination (C2).

Second, build dependency and feedback.

Based on the construction of the index system, in order to establish the ANP model, we must also study the interaction (feedback or dependence) between the evaluation indicators, that is, the correlation of the indicators. The correlation between evaluation indicators can usually be obtained through expert surveys or panel discussions. As shown below (Table 4.18) (Fig. 4.10).

Third, form a pairwise comparison matrix.

The pairwise comparison matrix is the judgment matrix, which is mainly used to describe the superiority between the indexes. The judgment matrix represents the comparison of the relative importance between the upper layer and the related indexes. The judgment matrix is the basic information of AHP and the basis of relative importance calculation.

Fourth, calculate weights.

After finishing the data obtained from the investigation, the judgment matrix between the indexes is obtained, and the weight of the index is calculated by Super Decision. as shown in Tables 4.19 and 4.20, Fig. 4.11.

Fifth, analysis of results.

Through the first column of Fig. 4.11, we can see the weight of each three-level index relative to its secondary index, for example, the weight of C1 and C2 are

0.44733, 0.55267 respectively. According to the third column of data, we can see the weight of each index relative to the first class index, for example B1 the weight of this index to the first class index is 0.228320. If the score value of the three-level index is obtained, the score value of the logistics service level of Datang Company can be calculated.

4.4 Fuzzy Comprehensive Evaluation

4.4.1 Definition of Fuzzy Comprehensive Evaluation

Fuzzy comprehensive evaluation is based on fuzzy mathematics, applying the principle of fuzzy transformation, quantifying some factors which are not clear and easy to quantify, and synthetically evaluating the subordinate grade of things from many factors.

Fuzzy mathematics came into being in the 1960s. Professor L. A. Zadeh of American cybernetics put forward the concept of fuzzy set for the first time on the basis of set theory. This paper studies how to use mathematical methods to deal with fuzzy things and phenomena of some “cognitive uncertainty” classes. “Fuzziness” refers to the indistinctness and uncertainty of the existence of things. For example, evaluating a person’s “beauty” or “ugliness” is a vague concept. Fuzzy comprehensive evaluation can take into account subjective cognition and objective measurement, which can better reflect the fuzziness of evaluation, and the mathematical model is simple, which is suitable for multi-factor and multi-level complex evaluation problems.

Table 4.18 Survey on the relation of evaluation index of logistics service level of Datang Company

Influence factors		Influencing factors					
		Reserve capacity B			Traffic capacity T	Organizing ability C	
		B1	B2	B3	T1	C1	C2
Reserve capacity B	B1		✓	✓	✓	✓	✓
	B2	✓		✓	✓		✓
	B3	✓	✓		✓	✓	
Traffic capacity T	T1			✓			
Organizing ability C	C1		✓	✓		✓	
	C2	✓		✓	✓		✓

Note The first line of indicators is the affected indicators, the left first column may cause impact indicators. Please type “✓” in the corresponding space. Such as B1 inventory capacity will affect the amount of B2 reserves

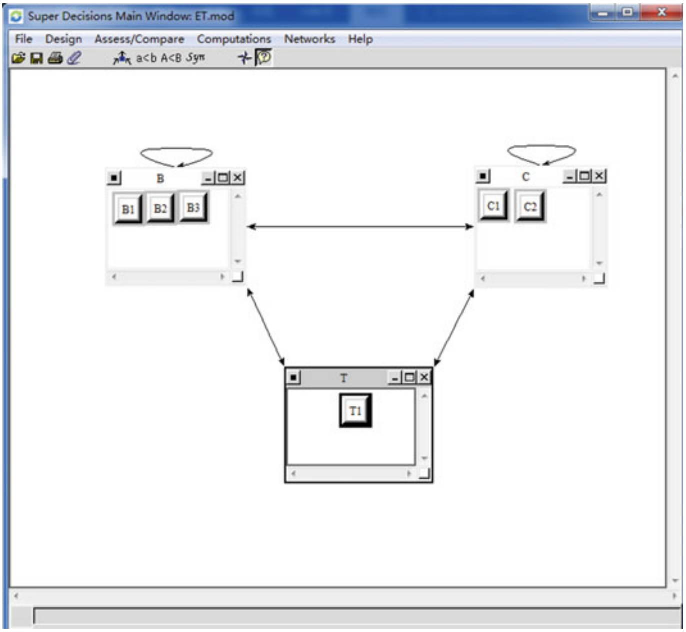


Fig. 4.10 Model of logistics service level evaluation ANP Datang Company

4.4.2 Fuzzy Mapping and Fuzzy Transformation

1. Fuzzy mapping

A set such as A = (1, 5, 6, 8, 9) is called a common set (the whole set is a domain). There is no clear boundary for subsets in fuzzy sets, as defined below.

Definition 4.3 Let A be a mapping of the domain X to the closed interval [0, 1], that is

$$A : X \rightarrow [0, 1]$$

Table 4.19 Comparison matrix of level 2 indicators

B	T	C	B
T	1	9	2
C	1/9	1	1/5
B	1/2	5	1

C	B	C
B	1	1/5
C	5	1

T	B	C
B	1	7
C	1/7	1

Table 4.20 Comparison matrix of level 3 indicators

B1	B2	B3
B2	1	4
B3	1/4	1

C1	**B1**	**B3**
B1	1	1/4
B3	4	1

T1	**B1**	**B2**	**B3**
B1	1	4	1/2
B2	1/4	1	1/9
B3	2	9	1

B2	**B1**	**B3**
B1	1	1/4
B3	4	1

C2	**B1**	**B2**
B1	1	3
B2	1/3	1

B3	**B1**	**B2**
B1	1	6
B2	1/6	1

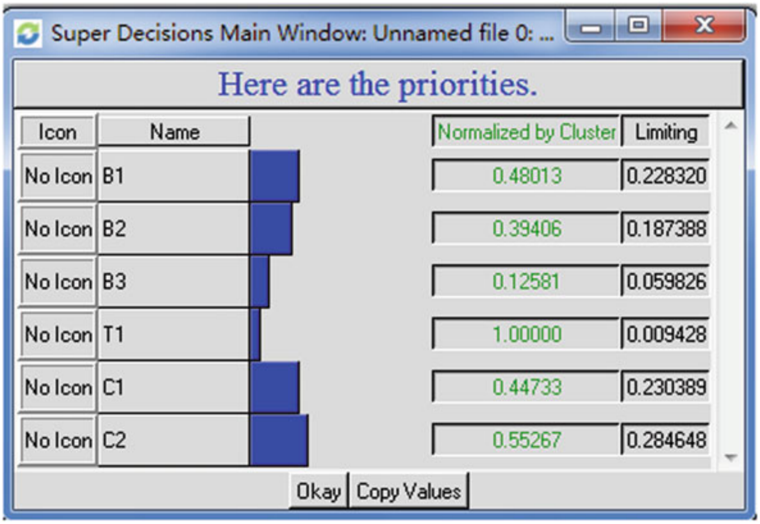


Fig. 4.11 Priorities

$$x \to A(x) \in [0, 1]$$

Then A is a fuzzy set of X (or A is a fuzzy subset of X), A (x) is called the membership function of fuzzy set A, and the value of A (x) is called the membership degree of x to the fuzzy set.

The fuzzy set on the domain X is recorded as F (x), which is called fuzzy power set. When A (x) values are only 0 and 1, A (x) degenerates into a feature function of a common set.

Definition 4.4 Name mapping

$$f : X \to F(Y)$$

$$x \to f(x) = B \in F(Y)$$

is a fuzzy mapping from X to Y.

Example 4.3 Assume $X = \{x_1, x_2, x_3\}$, $Y = \{y_1, y_2, y_3, y_4, y_5\}$, $R \in F(X \times Y)$, and.

$$R = \begin{bmatrix} 0.7 & 0.3 & 0.1 & 0.3 & 0 \\ 0.5 & 0.1 & 0.2 & 0.6 & 0.6 \\ 0.1 & 0 & 0.5 & 0.4 & 0.1 \end{bmatrix}$$

let

$$x_1 \to f(x)_1 = \frac{0.7}{y_1} + \frac{0.3}{y_2} + \frac{0.1}{y_3} + \frac{0.3}{y_4}$$
$$x_2 \to f(x)_2 = \frac{0.5}{y_1} + \frac{0.1}{y_2} + \frac{0.2}{y_3} + \frac{0.6}{y_4} + \frac{0.6}{y_5}$$
$$x_3 \to f(x)_3 = \frac{0.1}{y_1} + \frac{0.5}{y_3} + \frac{0.4}{y_4} + \frac{0.1}{y_5}$$

then f is called fuzzy mapping from X to Y.

Definition 4.5 Assume $R \in F(X \times Y)$, for arbitrary $x \in X$, there is a corresponding to a fuzzy set on the Y, recorded as $R|_x$, then the membership function is defined as follows:

$$R|_x(y) = R(x, y), \quad y \in Y \tag{4.8}$$

the $R|_x$ is called the R's x section.

Similarly, the $R|_y$ is called the R's y section, and

$$R|_y(x) = R(x, y), x \in X \tag{4.9}$$

There is a one-to-one correspondence between fuzzy relation and fuzzy mapping.

(1) For arbitrary $R \in F(X \times Y)$, R can determine fuzzy mapping f_R

$$f_R : X \to F(Y) \tag{4.10}$$

$$x \to f_R(x) = R|_x \in F(Y)$$

$$f_R(x)(y) = R|_x(y) = R(x, y) \tag{4.11}$$

(3) For arbitrary $f : X \to F(Y)$, can determine fuzzy relation

$$R_f \in F(X \times Y),$$

$$R_f(x, y) = f(x)(y) \tag{4.12}$$

There is a one-to-one correspondence between the fuzzy relation on the $X \times Y$ and the fuzzy mapping from X to Y, which can be written without confusion

$$R = R_f = f_R = f \tag{4.13}$$

Example 4.4 In Example 4.3

$$R|_x = (0.7, 0.3, 0.1, 0.3, 0)$$

$$R|_x = (0.5, 0.1, 0.2, 0.6, 0.6)$$

$$R|_x = (0.1, 0, 0.5, 0.4, 0.1)$$

And this is the fuzzy mapping f_R determined by fuzzy relations R, for example

$$f_R(x_1) = R|_{x_1} = (0.7, 0.3, 0.1, 0.3, 0) \in F(Y)$$

2. Fuzzy transformation

Definition 4.6 Name mapping

$$T : F(X) \rightarrow F(Y)$$

$$A \rightarrow T(A) = B \in F(Y)$$

is a fuzzy transformation from X to Y. The B is called the image of the A under the T of fuzzy transformation, A is the original image of the B.

Assume $X = (x_1, x_2, \ldots, x_n)$, $Y = \left(y_1, y_2, \ldots, y_m\right)$, then the fuzzy transformation T from X to Y is mapping. $T : M_{1\times n} \rightarrow M_{1\times m}$.

There are many applications of fuzzy transformation determined by fuzzy relation.

Definition 4.7 Assume $A, B \in F(X)$, if fuzzy transformation

$$T : F(X) \rightarrow F(Y)$$

satisfied:

(1) $T(A \cup B) = T(A) \cup T(B)$
(2) $T(\alpha A) = \alpha T(A), \quad \alpha \in [0, 1]$

then T is called fuzzy linear transformation.

The fuzzy linear transformation T is induced by fuzzy mapping f, which provides a theoretical basis for fuzzy comprehensive evaluation method.

4.4.3 Steps of the Fuzzy Comprehensive Evaluation Method

The modeling process of fuzzy comprehensive evaluation method can be carried out according to the following six steps:

(1) Determine the scope U of the influencing factors of the evaluation object

If there are n influencing factors, the scope $U = \{u_1, u_2, \ldots, u_n\}$.

(2) Determine the evaluation rating domain V of the evaluation object

Such as $V = \{v_1, v_2, \ldots, v_m\}$, is the evaluation level set. Each rating set is equivalent to a fuzzy subset. m is the number of comments, generally 3–5.

(3) Generate the fuzzy relation matrix R of single factor evaluation

Make a separate judgment $f(u_i)$ for each element, that is,

$$f: U \rightarrow F(V)$$

$$u_i \rightarrow f(u_i) = B \in F(V) \tag{4.14}$$

And $f(u_i) = (r_{i1}, r_{i2}, \ldots, r_{im})$. According to fuzzy mapping f can deduce fuzzy relation $R_f \in F(U \times V)$, that is $R_f(u_i, v_j) = f(u_i)(v_j) = r_{ij}$, hence, it can be represented by fuzzy matrix $R \in \mu_{n \times m}$:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{12} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{4.15}$$

(4) Determining the weight A of the influencing factors of the evaluation object

Relevant methods can be used to determine the weight of the influencing factors of the evaluation object:

$$A = (a_1, a_2, \ldots, a_n), \sum_{i=1}^{n} a_i = 1, a_i \geq 0. \tag{4.16}$$

(5) Calculation of Fuzzy Synthesis Value B

Using appropriate fuzzy operators to synthesize the factor weights A and the fuzzy matrix R of evaluation objects, the fuzzy synthetic values B of each evaluation object obtained. Namely:

$$A^*R = a_1, a_2, \ldots, a_n \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{12} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} = b_1, b_2, \ldots, b_m = B$$

b_i is the degree of membership of the fuzzy subset of v_i grade in general.

(6) Evaluation based on fuzzy synthetic values

B the fuzzy synthetic value obtained above is the degree description of the comprehensive status of each evaluated object, which cannot be directly ranked and evaluated, and can only be used after further analysis and processing. The principle of maximum membership is the most commonly used method.

4.4.4 Example of Fuzzy Comprehensive Evaluation

Example 4.5 Selection of R & D Scheme for New Energy Vehicles.

Datang plans to develop a new energy vehicle, and here are three R&D programs A, B and C. The influencing factors are shown in Table 4.21. One of the most suitable new energy vehicle R&D programs should be selected from the three schemes.

Solution:

(1) Determine the influencing factors scope U of the evaluation object

$$U = [\text{technical level, forecast market share, economic benefit}]$$

(2) Determine the evaluation rating domain V of the evaluation object

To simplify the calculation, the evaluation set is assumed to be V = [high, medium, low].

(3) Generate the fuzzy relation matrix R of single factor evaluation

Assuming that for each new energy vehicle R&D program, the comprehensive statistical results of each factor survey by experts are shown in Table 4.22.

The values in the table indicate the ratio of the number of experts in favor of such an evaluation to the total number of experts. For the economic benefits A the scheme, 40% of the experts think that the economic benefits are high, 40% think

Table 4.21 Relevant information of the three programs

Influence factor / Program	Technical level	Forecasting of market share (%)	Economic benefits
A	International advanced level	15	>15 million
B	Domestic advanced level	55	>30 million
C	Domestic general level	30	>5 million

Table 4.22 Expert evaluation findings

Evaluation	Technical level			Market share			Economic benefits		
	High	Medium	Low	High	Medium	Low	High	Medium	Low
A	0.5	0.4	0.1	0.15	0.15	0.7	0.4	0.4	0.2
B	0.1	0.6	0.3	0.1	0.1	0.8	0.5	0.5	0
C	0.3	0.3	0.4	0.9	0.1	0	0.15	0.15	0.7

that the economic benefits are medium and 20% think that the economic benefits are low. Therefore, the fuzzy relation matrix:

$$R_A = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.15 & 0.15 & 0.7 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}$$

$$R_B = \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.1 & 0.1 & 0.8 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

$$R_C = \begin{bmatrix} 0.3 & 0.3 & 0.4 \\ 0.9 & 0.1 & 0 \\ 0.15 & 0.15 & 0.7 \end{bmatrix}$$

(4) Determining the weight A of the influencing factors of the evaluation object

$$\text{Assume } A = \{0.1, 0.3, 0.6\}$$

(5) Calculation of Fuzzy Synthesis Value B

Using fuzzy operators $M = (\wedge, \vee)$ $(b_j = \max\{\min(a_1, r_{1j}), \min(a_1, r_{1j}), \ldots \min(a_1, r_{1j})\})$ to calculate, and $b_j = \vee_{i=1}^{n}\left(a_i \wedge r_{ij}\right)$ $(j = 1, 2, \ldots, m)$. The result is:

$$B_A = A \cdot R_A = (0.1, 0.3, 0.6)\begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.15 & 0.15 & 0.7 \\ 0.4 & 0.4 & 0.2 \end{bmatrix} = (0.4, 0.4, 0.3)$$

$$B_B = A \cdot R_B = (0.1, 0.3, 0.6)\begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.1 & 0.1 & 0.8 \\ 0.5 & 0.5 & 0 \end{bmatrix} = (0.5, 0.5, 0.3)$$

$$B_C = A \cdot R_C = (0.1, 0.3, 0.6)\begin{bmatrix} 0.3 & 0.3 & 0.4 \\ 0.9 & 0.1 & 0 \\ 0.15 & 0.15 & 0.7 \end{bmatrix} = (0.3, 0.15, 0.6)$$

(6) Evaluation based on fuzzy synthetic values

First, the normalization process

$$B_A = (0.36, 0.36, 0.28)$$

$$B_B = (0.38, 0.38, 0.24)$$

$$B_C = (0.29, 0.14, 0.57)$$

According to the principle of maximum membership, it can be obtained from the above results that the C scheme is the most suitable scheme in the new energy vehicle research scheme.

4.5 Entropy Evaluation

4.5.1 Fundamental Principle

"Entropy" is an important concept of thermodynamics, created by German physicist Rudolph Clausius in 1854 to indicate the uniformity of the distribution of any kind of energy in space. The more uniform the energy distribution, the greater the entropy. Entropy is also a measure of system chaos and disorder. By calculating entropy value to judge the randomness and disorder degree of an event, the greater the entropy value, the greater the degree of chaos and disorder. C. E. Shannon, the founder of information theory, put forward the concept of "information entropy" in 1948, and solved the problem of quantitative measurement of information. In information theory, entropy characterizes the uncertainty of information, the entropy of high information is very low, and the entropy of low information is high. Therefore, information entropy can also be used to judge the degree of uncertainty of response information of an index. Generally speaking, the higher the certainty of information reflected by the index, the smaller the information entropy, the greater the amount of information provided by the index, and the greater the role played in the comprehensive analysis. On the contrary, the larger the information entropy of the index, the smaller the amount of information provided by the index, and the smaller the role in the comprehensive analysis. The objective assignment of weight coefficient in comprehensive evaluation by entropy method can effectively reflect the change law of the system and reflect the evaluation effect objectively and truthfully.

The evaluation steps of the entropy method are as follows:

1. Standardization of indicators

If the statistical data of the evaluation object has n sample, the evaluation system index has m, x_{ij} $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m)$ is the index value of the number j index of the number i sample. Each evaluation index is standardized and processed into standard data of uniform order of magnitude.

Positive indicators: $x'_{ij} = x_{ij}/x^*_{j\,\max}$.

Negative indicators: $x'_{ij} = x^*_{j\,\min}/x_{ij}$.

2. Calculation of the proportion of indicators

$$p_{ij} = \frac{x_{ij}^{'}}{\sum_{i=1}^{n} x_{ij}^{'}} \tag{4.17}$$

3. Calculation of Entropy Value for Item j Index

$$e_j = -k\sum_{i=1}^{n} p_{ij}\ln p_{ij} \tag{4.18}$$

And $k > 0$, $e_j \geq 0$, $k = \frac{1}{\ln n}$.

4. Calculation of the coefficient of difference for item j

For the number j index, the greater the difference of index value, the greater the effect on sample evaluation, the smaller the entropy value. Definition of coefficient of difference:

$$g_j = 1 - e_j \tag{4.19}$$

5. Definition of weights

$$a_j = \frac{g_j}{\sum_{j=1}^{m} g_j} \tag{4.20}$$

6. Calculation of sample comprehensive evaluation values

$$v_i = \sum_{j=1}^{m} a_j p_{ij} \tag{4.21}$$

v_i is the comprehensive evaluation value of the number i sample.

Example 4.6 Taking Datang Company's strategic performance evaluation as an example, its purpose is to promote the effectiveness of strategic control, so it is necessary to fully consider the company's realistic viability and potential development ability. The evaluation index not only reflects the change of strategic performance, but also is the key point of strategic control. The selection of indicators should follow the key, growth, scientific and feasible principles.

Solution:

Dimension	Index	Unit	1	2	3	4
Finance	Return on net assets	%	3.8	6.5	9.3	12.7
	Ratio of selling profit	%	16	17	20	21
	Operation revenue	10,000	400	450	830	1150
Customer	Customer satisfaction	Value	80	85	90	90
	Market share	%	3	18	28	40
	Old customer retention rate	%	96	95	97	98
	New customer growth rate	%	80	85	87	95
Internal operations	Pass rate of products (services)	%	89	92	97	97
	Investment in brand construction	10,000	20	27	35	35
	Failure rate of production equipment	%	3	2	5	3
	Investment in public relations costs for major clients	10,000	10	15	15	20
Learning and growth	Employee satisfaction	Value	90	82	88	85
	Growth rate of new product sales	10,000	50	70	100	100
	Staff training expenditure	10,000	3	7	15	20
	Training pass	%	70	80	90	90
	Recommendation acceptance rate	%	30	25	20	20

The first step: Standardize the evaluation indicators due to the different indicators

$$\text{Positive indicators } x'_{ij} = x_{ij}/x_{j\,max}$$
$$\text{Negative indicators } x'_{ij} = x_{j\,min}/x_{ij}$$

Among them, the failure rate of production equipment is negative index, the rest are positive index.

Positive indicators (taking ROE as an example): sample 1 standardized values are:

$$3.8/12.7(\max) = 0.299$$

Negative indicators (for example, failure rate of production equipment): sample 1 standardized values are: (Table 4.23)

$$2(\min)/3 = 0.667$$

Table 4.23 Standardized data for indicators

Index	×1′	×2′	×3′	×4′
Return on net assets	0.299	0.512	0.732	1
Ratio of selling profit	0.762	0.81	0.952	1
Operation revenue	0.348	0.391	0.722	1
Customer satisfaction	0.889	0.944	1	1
Market share	0.075	0.45	0.7	1
Old customer retention rate	0.98	0.969	0.99	1
New customer growth rate	0.842	0.895	0.916	1
Pass rate of products (services)	0.918	0.948	1	1
Investment in brand construction	0.571	0.771	1	1
Failure rate of production equipment	0.667	1	0.4	0.667
Investment in public relations costs for major clients	0.5	0.75	0.75	1
Employee satisfaction	1	0.911	0.978	0.944
Growth rate of new product sales	0.5	0.7	1	1
Staff training expenditure	0.15	0.35	0.75	1
Training pass	0.778	0.889	1	1
Recommendation acceptance rate	1	0.833	0.667	0.667

Step 2: Calculate the weight of each index, that is, the proportion of a single sample to the total sample. The sum of all samples in an indicator divided by a sample, that is:

$$\mathrm{p_{ij}} = \mathrm{x'_{ij}} / \sum\nolimits_{i=1}^{n} \mathrm{x'_{ij}}$$

Take the Return on Net Assets as an example (Table 4.24):

$$\text{Sample 1} = 0.299/\,(0.299 + 0.512 + 0.732 + 1) = 0.118$$

$$\text{Sample 2} = 0.512/(0.299 + 0.512 + 0.732 + 1) = 0.201$$

$$\text{Sample 3} = 0.732/(0.299 + 0.512 + 0.732 + 1) = 0.288$$

$$\text{Sample 4} = 1/(0.299 + 0.512 + 0.732 + 1) = 0.393$$

Step 3: Calculate the entropy value of item j, according to the formula:

$$\mathrm{e_j} = -\mathrm{k} * \sum_{i=1}^{n} p_{\mathrm{ij}} * \ln\left(p_{\mathrm{ij}}\right)$$

Table 4.24 Percentage of indicators

Index	p1	p2	p3	p4
Return on net assets	0.118	0.201	0.288	0.393
Ratio of selling profit	0.216	0.230	0.270	0.284
Operation revenue	0.141	0.159	0.293	0.406
Customer satisfaction	0.232	0.246	0.261	0.261
Market share	0.034	0.202	0.315	0.449
Old customer retention rate	0.249	0.246	0.251	0.254
New customer growth rate	0.231	0.245	0.251	0.274
Pass rate of products (services)	0.237	0.245	0.259	0.259
Investment in brand construction	0.171	0.231	0.299	0.299
Failure rate of production equipment	0.244	0.366	0.146	0.244
Investment in public relations costs for major clients	0.167	0.250	0.250	0.333
Employee satisfaction	0.261	0.238	0.255	0.246
Growth rate of new product sales	0.156	0.219	0.313	0.313
Staff training expenditure	0.067	0.156	0.333	0.444
Training pass	0.212	0.242	0.273	0.273
Recommendation acceptance rate	0.316	0.263	0.211	0.211

$$\text{Among them, } k > 0,\ e_j > 0,\ k = 1/\ln(n)$$

Take the Return on Net Assets as an example (Table 4.25).

$$e_1 = -1/\ln 4^* \{0.118^* \ln(0.118) + 0.201^* \ln(0.201) + 0.288^* \ln(0.288) + 0.393^* \ln(0.393)\} = 0.938$$

Step 4: Calculate the difference coefficient of item j, the greater the difference coefficient, the greater the effect on sample evaluation, the smaller the entropy value, define the difference coefficient:

$$g_j = 1 - e_j$$

Table 4.25 Entropy of indicators

e1	e2	e3	e4	e5	e6	e7	e8
0.938	0.996	0.934	0.999	0.837	1	0.999	1
e9	e10	e11	e12	e13	e14	e15	e16
0.983	0.965	0.980	1	0.973	0.863	0.996	0.989

Table 4.26 Variance coefficients for indicators

g1	g2	g3	g4	g5	g6	g7	g8
0.062	0.004	0.066	0.001	0.163	0.000	0.001	0.000
g9	g10	g11	g12	g13	g14	g15	g16
0.017	0.035	0.020	0.000	0.027	0.137	0.004	0.011

Take the Return on Net Assets as an example: $g_1 = 1 - e_1 = 1 - 0.938 = 0.062$ (Table 4.26).

Step 5: Define the weight of the index, that is, the proportion of a single index to the total index. Divide the difference coefficient of an indicator by the sum of the difference coefficients of all sample indicators, that is:

$$a_j = g_j / \sum_{j=1}^{m} g_j$$

Take the Return on Net Assets as an example (Table 4.27).

$$a_1 = 0.062/(0.062 + 0.004 + 0.066 + 8E - 04 + 0.163 + 5E - 05 + 0.001 + 5E - 04 + 0.017 + 0.035 + 0.02 + 4E - 04 + 0.027 + 0.137 + 0.004 + 0.011) = 0.113$$

Step 6: Calculate the sample comprehensive evaluation value, according to the formula:

$$v_j = \sum_{j=1}^{m} a_j * p_{ij}$$

Take the sample1 as an example:

$$v_1 = 0.113^* 0.118 + 0.008^* 0.216 + \ldots + 0.02^* 0.316 = 0.102$$

Table 4.27 Weight of indicator

a1	a2	a3	a4	a5	a6	a7	a8
0.113	0.008	0.120	0.002	0.296	0.000	0.003	0.001
a9	a10	a11	a12	a13	a14	a15	a16
0.031	0.064	0.037	0.001	0.048	0.249	0.007	0.020

Table 4.28 Comprehensive evaluation of strategic performance

v1	v2	v3	v4
0.102	0.201	0.297	0.399

Table 4.28 reflects the company's strategic performance for the fourth consecutive year, from the development trend, the company's performance steadily improved. It can be seen from the weight of each index that the three indexes of market share, employee training expenditure and return on net assets have a large weight, which indicates that these three indexes have a great impact on the results of strategic performance evaluation in this group of data. The strategic adjustment makes the organization structure, assets, business, value chain and so on optimized, but the initial adjustment must be cost greater than income. But through continuous optimization, the strategic performance is improved.

4.5.2 Application Case

The Entropy Method for Measuring System Orderliness

Entropy, as a measure of information content, represents the uncertainty of the existence state and motion state of the system, and is widely used to determine the uncertainty and change measurement in phenomena. In the information age system, the internal and external environment of the system is changeable, and more information uncertainty problems appear than before. All these uncertainties can be described by the concept of information entropy. The degree of information communication and the loss rate of information flow can be evaluated by entropy transformation equivalent, which is called the information entropy H_1 of the system. The structure reflects the interrelation of information transmission between the units of the system, and the ability of the carrier to process information is called the structural entropy H_2 of the system. The function reflects the degree of interaction of each subsystem and can reflect some function of the system as a whole, which is called functional entropy H_3. Information entropy, structure entropy and function entropy together determine the order degree R of the system.

1. System Information Entropy

If there is a direct information communication relationship between elements in the system, it is called the inter-element connection. The number of connections between two elements is defined as the length of the connection between the two elements, which is expressed as L_{ij} (i, j representing the element number, $i, j = 1, 2, \ldots, N$).

Define the information entropy $H_1(ij)$ between any two elements in a system

$$H_1(ij) = -P_1(ij)\ln P_1(ij) \tag{4.22}$$

$H_1(ij)$ reflects the uncertainty of the flow of information between the two elements i, j in the system, and $P(ij)$ is the realization probability of the microscopic state of information between the elements i, j, which is defined as

$$P_1(ij) = L(ij)/A_1 \tag{4.23}$$

The connection length L_{ij} is the shortest path between the two elements in the system, the length of the direct connection is 1, the length of each transfer plus 1, from the definition, the greater the P_{ij} value, it represents that the larger the connection length between any two elements i, j in the system, the greater the probability of perfecting information and creating information. A_1 is the total number of microscopic states corresponding to the information in the macroscopic state of the system, which is related to the total amount of system elements and the length of the connection between elements

$$A_1 = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} L_{ij} \tag{4.24}$$

Information entropy of the system H_1

$$H_1 = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} H_{ij} \tag{4.25}$$

H_1 is the sum $H_1(ij)$ of all elements of the system, it reflects the uncertainty of the microscopic state corresponding to the macroscopic state information of the system.

$$H_{1max} = lnA_1 \tag{4.26}$$

$H_{1\max}$ is the maximum value of system information entropy. If the number of microscopic states A_1 corresponding to the macroscopic state of the system is more, the more elements of the system, the more diversified the contact methods, so it reflects the order degree of the system from the effectiveness of information circulation.

The information order of the system is R_1

$$R_1 = 1 - H_1/H_{1\max}, \quad R_1 \in [0, 1] \tag{4.27}$$

$H_1/H_{1\max}$ reflects the uncertainty probability of information in the system, so R_1 reflects the order degree of system information.

2. System structure entropy

Let the contact amplitude of each element in the system be the number of elements directly related to the element, expressed as K_i $(i = 1, 2, \ldots, N)$.

$$H_2(i) = -P_2(i)/\ln P_2(i) \tag{4.28}$$

The uncertainty of the number i microscopic state corresponding to the macroscopic state of the system is described as $H_2(i)$, which is related not only to the number of possible microscopic states in the macroscopic state, but also to the probability of the occurrence of different microscopic states. The probability of the appearance of the microscopic state of the i structure corresponding to the macroscopic state of the system is $P_2(i)$, defined as

$$P_2(i) = K_i/A_2 \tag{4.29}$$

Among them K_i is the connection range of system components, which reflects the real connection between system components. The larger K_i, the more complex the corresponding structure, the more extensive the connection and the stronger the nonlinear effect. A_2 is the total number of microscopic states corresponding to the structure in the macroscopic state of the system, which increases with the increase of the number of components and the increase of the contact amplitude.

$$A_2 = \sum_{i=1}^{N} K_i \tag{4.30}$$

$$H_2 = \sum_{i=1}^{N} H_2(i) \tag{4.31}$$

H_2 is the entropy of the system structure, which is the sum of all the elements $H_2(i)$ in the system, which reflects the disorder of the microstructure corresponding to the macroscopic state of the system.

$$H_{2\max} = \ln A_2 \tag{4.32}$$

$H_{2\max}$ is the maximum structural entropy of the system. If the number of microscopic states A_2 corresponding to a macroscopic state of the system is more, the larger the order degree of the system is. So it reflects the order of the system in terms of structure.

The structural order of the system is R_2

$$R_2 = 1 - H_2/H_{2\max}, \quad R_2 \in [0, 1] \tag{4.33}$$

$H_2/H_{2\max}$ reflects the probability of system structure disorder, so R_2 represents the degree of system structure order.

3. System capacity entropy

The function coupling of subsystems reflects the level of logical function between subsystems. Let the strength of a subsystem function in the system be the number of other subsystems directly related to the subsystem, and express it as C_i $i = 1, 2, \cdots, N)$

$$H_3(i) = -P_3(i)\ln P_3(i) \tag{4.34}$$

The uncertainty of the number i microscopic state corresponding to the macroscopic state of the system is described as $H_3(i)$, which is related not only to the number of possible microscopic states in the macroscopic state, but also to the probability of the occurrence of different microscopic states. The probability of the appearance of the microscopic state of the first structure corresponding to the macroscopic state of the system is $P_3(i)$, defined as

$$P_3(i) = C_i/A_3 \tag{4.35}$$

Among them, C_i is the system component function strength, which reflects the degree of promotion between the system components. The larger C_i, the more complex the function, the stronger the promotion and the stronger the nonlinear effect. A_3 is the total number of microscopic states corresponding to the function in the macroscopic state of the system, which increases with the increase of the number of components and the increase of the degree of promotion.

$$A_3 = \sum_{i=1}^{N} C_i \tag{4.36}$$

$$H_3 = \sum_{i=1}^{N} H_3(i) \tag{4.37}$$

H_3 is the function entropy of the system and the sum $H_3(i)$ of all the elements in the system, which reflects the disorder of the microscopic state corresponding to the function under the macroscopic state of the system.

$$H_{3\max} = \ln A_3 \tag{4.38}$$

$H_{3\max}$ is the maximum functional entropy of the system. If the number of microscopic states A_3 corresponding to a macroscopic state of the system is more, the more powerful the system is.

The functional order of the system is R_3

$$R_3 = 1 - H_3/H_{3max}, \quad R_2 \in [0, 1] \tag{4.39}$$

$H_3/H_{3\max}$ reflects the probability of system function disorder, so R_3 represents the degree of system function order.

Synthesize the information order degree, structure order degree and function order degree, the system whole order degree is

$$R = R_1 + R_2 + R_3 \tag{4.40}$$

4. Example

Suppose the structure of a system is shown in Fig. 4.12 and the functional connection is shown in Fig. 4.13 (Tables 4.29 and 4.30)

The H_{max} increase indicates that the increase of system order is promoted by increasing the total number of microscopic states of the system. From the above

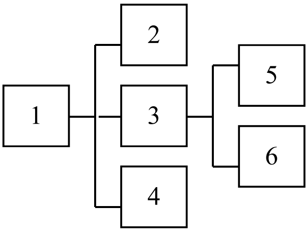


Fig. 4.12 System architecture

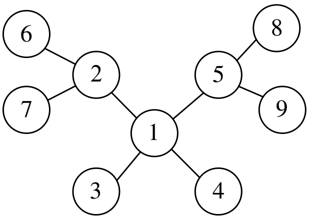


Fig. 4.13 Functional linkages

Table 4.29 Value of L_{ij}, K_i, C_i

$L_{12} = 1$	$L_{21} = 1$	$L_{31} = 1$	$L_{41} = 1$	$L_{51} = 2$	$L_{61} = 2$	$K_1 = 3$	$C_1 = 4$
$L_{13} = 1$	$L_{23} = 1$	$L_{32} = 1$	$L_{42} = 1$	$L_{52} = 2$	$L_{62} = 2$	$K_2 = 3$	$C_2 = 6$
$L_{14} = 1$	$L_{24} = 1$	$L_{34} = 1$	$L_{43} = 1$	$L_{53} = 1$	$L_{63} = 1$	$K_3 = 5$	$C_3 = 4$
$L_{15} = 2$	$L_{25} = 2$	$L_{35} = 1$	$L_{45} = 2$	$L_{54} = 2$	$L_{64} = 2$	$K_4 = 3$	$C_4 = 4$
$L_{16} = 2$	$L_{26} = 2$	$L_{36} = 1$	$L_{46} = 2$	$L_{56} = 1$	$L_{65} = 1$	$K_5 = 2$	$C_5 = 4$
						$K_6 = 2$	$C_6 = 2$
							$C_7 = 2$
							$C_8 = 2$
							$C_9 = 2$

Table 4.30 Orderliness

R_1	R_2	R_3	R
0.106	0.398	0.377	0.881

numerical results, it can be seen that $R_2 > R_3 > R_1$, the order degree of the structure contributes the highest to the macroscopic order degree of the system, followed by the higher contribution of the functional order degree to the macroscopic order degree of the system. By comparing the order of different periods of the same system, we can understand the changes of the order of the system in information, structure and function in time, which is helpful for the decision-making level to adjust the strategy in time and optimize the system.

4.6 Set Pair Analysis

4.6.1 Fundamental Principle

Set pair analysis was put forward by Chinese scholar Zhao Keqin in 1989. It is a systematic analysis method to deal with uncertainty. It studies the uncertainty of things from the same, different and inverse aspects, and comprehensively analyzes the relationship between the two things. The core idea of set pair analysis is to regard certainty uncertainty as a definite uncertainty system, and to analyze the relationship and transformation between things from positive and negative to definite uncertainty. Set pair, that is, the pair of different sets with certain interrelatedness, according to one of its characteristics according to the same, different, reverse three directions of analysis, the quantized expression of the relation degree of the set to a problem background is obtained,

$$\mu = a + bi + cj$$

μ represents the degree of connection of set pairs; a represents the same degree of two set pairs; b represents the degree of difference between two sets; c represents the degree of opposition of two sets; i represents the degree of difference uncertainty coefficient ($i \in [-1, 1]$), when $i = -1, 1$, it is deterministic, it is opposites and identities. When $i \in (-1, 1)$, it is uncertain, j represents the coefficient of opposites, its value is -1.

Among them, a, b and c satisfy the normalized condition,

$$a + b + c = 1$$

The set pair analysis method is widely used in the fields of mathematics, system science, information and so on. It has been in system risk, environmental bearing capacity and flood control system evaluation, and has also been used in multi-objective system evaluation.

1. **Construct evaluation matrix**

Construct evaluation matrix $D = (d_{rk})_{m \times n}$.

$$D = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{m1} & \cdots & d_{mn} \end{pmatrix} \tag{4.41}$$

Formula 4.41 represents systematic sustainability evaluation of m evaluation indicators for n evaluation subjects, d_{rk} of the evaluation matrix represents the r evaluation index value of the k evaluation object.

2. **Determination of the optimal and the worst scheme**

The optimal and inferior scheme U and V of the evaluation indicators are determined from the internal and external aspects of the evaluation index system according to the objectives of the system sustainability evaluation and the characteristics of the rating indicators

$$U = (u_1, u_2, \ldots, u_m)^T$$

$$V = (v_1, v_2, \ldots, v_m)^T$$

u_r represents the optimal value of the r evaluation index and v_r represents the worst value of the r evaluation index. The evaluation matrix D^* is constructed according to the optimal scheme and the worst scheme.

$$D^* = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} & u_1 & v_1 \\ d_{21} & d_{22} & \cdots & d_{2n} & u_2 & v_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} & u_m & v_m \end{pmatrix}$$

3. **Determination of linkages**

The d_{rk} value of the evaluation index r by any evaluation object k for the optimal scheme's linkages is:

$$\mu_{rk} = a_{rk} + b_{rk} i + c_{rk} j$$

Among them, a_{rk} represents consistency, b_{rk} represents diversity, c_{rk} represents opposites.

(1) When the evaluation indicator is of benefit type

$$a_{rk} = \frac{d_{rk}}{u_r + v_r}$$

$$b_{rk} = 1 - (a_{rk} + b_{rk}) = \frac{(u_r - d_{rk})(d_{rk} - v_r)}{(u_r + v_r)d_{rk}}$$

$$c_{rk} = \frac{u_r v_r}{(u_r + v_r)d_{rk}}$$

Among them, $\frac{d_{rk}}{u_r + v_r} \in [0, 1]$ represents the degree of proximity of d_{rk} and u_r, the greater the value, the greater the degree of proximity.

(2) When the evaluation indicator is cost-based

$$a_{rk} = \frac{u_r v_r}{(u_r + v_r)d_{rk}}$$

$$b_{rk} = 1 - (a_{rk} + b_{rk}) = \frac{(u_r - d_{rk})(d_{rk} - v_r)}{(u_r + v_r)d_{rk}}$$

$$c_{rk} = \frac{d_{rk}}{u_r + v_r}$$

Any evaluation object k the comprehensive relation degree for the optimal scheme is:

$$\mu_k = a_k + b_k i + c_k j$$

$$a_k = \sum_{r=1}^{m} \omega_r a_{rk}$$

$$b_k = \sum_{r=1}^{m} \omega_r b_{rk}$$

$$c_k = \sum_{r=1}^{m} \omega_r c_{rk}$$

ω_r represents the weights of evaluation indicators r.

Finally, according to the size of different comprehensive evaluation values R different objects to be selected, the order of advantages and disadvantages can be discharged. The larger the R value, the better the result is.

4.6.2 Application Case

Sustainability Assessment of Regional Innovation Capacity Based on Set-Pair Analysis

The sustainability of regional innovation ability is a complex organic system. First of all, the sustainability of regional innovation ability is a dynamic development process, not only to evaluate its development status, but also to ignore its development potential and development trend. Secondly, regional innovation ability is an organic whole in which the elements of the system interact with each other. In addition, innovation is a relatively abstract and potential concept, and some of the factors that affect the sustainability of regional innovation capabilities may be difficult to quantify, lack of data or cannot fully reflect regional innovation capabilities. Based on the evaluation index of regional innovation ability at home and abroad, this chapter explains the principle of establishing the index and constructs the sustainability index system of regional innovation ability.

1. Clear evaluation indicators

Based on the analysis of the connotation and structure of regional innovation ability sustainability, the evaluation of innovation ability sustainability in a region should decompose each basic element and construct different levels of index system. The evaluation index system is divided into three levels, the first is the main body of regional innovation, the regional innovation environment and the regional ecological development, which are the first level indicators; the second level is the index of regional innovation main body is innovation input and innovation output, and the regional innovation environment index is economic environment, infrastructure, science and technology development, government management ability and network maturity. The regional ecological development index is atmospheric environment and other. Continue to subdivide to form three-level indicators. As shown in Table 4.31, a regional innovation capacity sustainability evaluation index system is constructed.

2. Collection of origional data

This chapter selects several large cities in China as the evaluation objects, namely Beijing, Shanghai, Tianjin and Shenzhen. Based on the established regional innovation capacity sustainability evaluation indicators, based on 2012 data, the specific values of the evaluation indicators for each region are determined in Table 4.32.

3. Determination of indicator weights

Using analytic hierarchy process (AHP) to determine and evaluate the weight vector of indicators M_1, M_2, M_3 as

$$\omega = (\omega_1, \omega_2, \omega_3) = (0.4, 0.3, 0.3)$$

Table 4.31 Indicators for sustainability evaluation of regional innovation capacity

Level 1 indicators	Level 2 indicators	Level 3 indicators
Regional innovation main body	Innovative inputs	Percentage GDP R&D expenditure
		Estimated proportion of R&D personnel in industrial enterprises above
		R&D bodies
	Innovative outputs	Number of patents invented per 10,000 population
		Number of trademark applications
		Proportion of added value of high-tech industries to industrial added value
Regional innovation environment	Economic environment	Investment in fixed assets per capita
		Technology Enterprise Innovation Fund
		GDP per capita
		Total retail sales of consumer goods per capita
	Infrastructure	Number of mobile phones per 100 households
		Road growth rate
	Technological development	Proportion of expenditure on education
		High school penetration rate
		Per capita years of education
	Government management capacity	Proportion of expenditure on scientific research
	Network maturity	First class road mileage
		Employment mobility
		Number of high-tech enterprises
Regional ecological development	Atmospheric environment	PM2.5 pollution days
		Sulfur dioxide (mg/m^3)
		Nitrogen oxide (mg/m^3)
	Other environmental indicators	Product value of comprehensive utilization of three wastes
		Industrial wastewater treatment rate

(continued)

Table 4.31 (continued)

Level 1 indicators	Level 2 indicators	Level 3 indicators
		Industrial waste gas treatment facilities
		Percentage of forest cover

Table 4.32 Original tables for sustainability evaluation of regional innovation capacity

Level 1 indicators	Level 2 indicators	Level 3 indicators	Beijing	Tianjin	Shanghai	Shenzhen
M1	N_1	x_1	5.90%	2.70%	3.40%	3.90%
		x_2	23.40%	17.23%	54.70%	71.15%
		x_3	1114	1193	1389	1018
	N_2	x_4	33.6	2.38	21.6	4.9
		x_5	92,305	22,000	38,600	73,130
		x_6	34.60%	30.20%	75%	22.30%
M_2	N_3	x_7	233,482.7	62,776	22,073	20,805.4
		x_8	17,700	17,900	29,300	8415
		x_9	103,928	91,242	85,373	123,247
		x_{10}	37,224.2	27,749	31,138	38,007
	N_4	x_{11}	225.92	225.05	239.58	296
		x_{12}	0.30%	1.40%	3.80%	2.60%
	N_5	x_{13}	14.60%	17.67%	15.50%	13.10%
		x_{14}	54.30%	75.51%	65.51%	55.51%
		x_{15}	9.12	10.62	9.2	9.23
	N_6	x_{16}	5.95%	16.82%	16.20%	4.20%
	N_7	x_{17}	0.11	0.11	0.04	0.08
		x_{18}	5.86%	4.54%	7.85%	2.57%
		x_{19}	2711	1799	3589	2898
M_3	N_8	x_{20}	84	60	23	77
		x_{21}	0.028	0.048	0.023	0.01
		x_{22}	0.052	0.042	0.046	0.04
	N_9	x_{23}	34,366	192,650	170,379	624,265
		x_{24}	83.00%	87.50%	91.00%	96.00%
		x_{25}	2468	3126	4319	12,789
		x_{26}	38.60%	34.90%	38%	45%

Note Data derived from data published in the relevant statistical yearbooks of our country in 2012

The weight vector of secondary evaluation indicators $N_i (i = 1, 2, \ldots, 9)$ is,

$$\omega_{M1} = (0.5, 0.5), \quad \omega_{M2} = (0.25, 0.23, 0.24, 0.13, 0.15), \quad \omega_{NM} = (0.52, 0.48)$$

The weight vector of third level evaluation indicators $X_i (i = 1, 2, \ldots, 26)$ is,

$$\omega_{N1} = (0.4, 0.3, 0.3), \quad \omega_{N2} = (0.3, 0.4, 0.3), \quad \omega_{N3} = (0.22, 0.3, 0.28, 0.2),$$
$$\omega_{N4} = (0.48, 0.52), \quad \omega_{N5} = (0.34, 0.42, 0.24), \quad \omega_{N6} = (1), \omega_{N7} = (0.35, 0.3, 0.35),$$
$$\omega_{N8} = (0.34, 0.33, 0.33), \quad \omega_{N9} = (0.34, 0.22, 0.22, 0.22)$$

4. Calculation of evaluation results
   (1) Level 1 comprehensive evaluation

The evaluation matrix D_{N1w} on the N_1 of innovation ability M_1 of regional innovation subject is calculated from Table 4.1,

$$D_{N1w} = \begin{bmatrix} 5.9 & 2.7 & 3.4 & 3.9 \\ 23.4 & 17.23 & 54.7 & 71.15 \\ 1114 & 1193 & 1389 & 1018 \end{bmatrix}$$

Ideal program $U_0 = \left[5.9,\ 71.15,\ 1389\right]^T$

$$\mu_{N1} = \begin{bmatrix} 1 & 0.45763 & 0.57627 & 0.66102 \\ 0.32888 & 0.24216 & 0.76880 & 1 \\ 0.80202 & 0.85889 & 1 & 0.73290 \end{bmatrix}$$

Then we can obtain the comprehensive evaluation result R_{N1} on the N_1 of innovation ability M_1 of the regional innovation subject,

$$R_{N1} = \omega_{N1} \times \mu_{N1} = (0.73927, 0.51337, 0.76115, 0.78428)$$

In the same way, we can calculate the comprehensive evaluation results R_{N2} on the N_2 of the innovation ability M_1 of the regional innovation subject,

$$R_{N2} = \omega_{N2} \times \mu_{N2} = (0.78453, 0.25382, 0.71831, 0.40036)$$

The results of the comprehensive evaluation of the regional innovation environment M_2 on the N_3, N_4, N_5, N_6, N_7 are

$$R_{N3} = (0.83322, 0.59574, 0.67861, 0.58576)$$

$$R_{N4} = (0.40741, 0.55652, 0.90851, 0.83579)$$

Table 4.33 Results of the integrated evaluation at the level I

Level I evaluation results	Beijing	Tianjin	Shanghai	Shenzhen
Innovative inputs	0.73927	0.51337	0.76115	0.78428
Innovative outputs	0.78453	0.25382	0.71831	0.40036
Economic environment	0.83322	0.59574	0.67861	0.58576
Infrastructure	0.40741	0.55652	0.90851	0.83579
Technological development	0.78906	1.00000	0.87053	0.76941
Government management capacity	0.35375	1.00000	0.96314	0.24970
Network maturity	0.83833	0.69894	0.77727	0.63538
Atmospheric environment	0.46480	0.51337	0.77043	0.76156
Other environmental indicators	0.44009	0.52984	0.56288	1.00000

$$R_{N5} = (0.78906, 1, 0.87053, 0.76941)$$

$$R_{N6} = (0.35375, 1, 0.96314, 0.24970)$$

$$R_{N7} = (0.83833, 0.69894, 0.77727, 0.63538)$$

The results of the comprehensive evaluation of the regional ecological development capacity M_3 on the N_8, N_9 are:

$$R_{N8} = (0.46480, 0.51337, 0.77043, 0.76156)$$

$$R_{N9} = (0.44009, 0.52984, 0.56288, 1)$$

The results of the first-level integrated evaluation are presented in Table 4.33.

(2) Level 2 Comprehensive evaluation

From Table 4.33, the evaluation matrix D_{M1} on the M_1 of regional innovation capacity sustainability M can be obtained,

$$D_{M1} = \begin{bmatrix} 0.73927 & 0.51337 & 0.76115 & 0.78428 \\ 0.78453 & 0.25382 & 0.71831 & 0.40036 \end{bmatrix}$$

Ideal programme $U_0 = \begin{bmatrix} 0.78428 & 0.78453 \end{bmatrix}^T$

$$\mu_{M1} = \begin{bmatrix} 0.94261 & 0.65457 & 0.97051 & 1 \\ 1 & 0.32353 & 0.91559 & 0.51032 \end{bmatrix}$$

Table 4.34 Results of the secondary comprehensive evaluation

Level 2 evaluation results	Beijing	Tianjin	Shang hai	Shenzhen
Regional innovation main body	0.97131	0.48905	0.94305	0.75516
Regional innovation environment	0.73850	0.81470	0.90682	0.71815
Regional ecological development	0.52496	0.60082	0.79018	0.99401

Then the evaluation matrix R_{M1} on the M_1 of regional innovation capacity sustainability M can be obtained,

$$R_{M1} = \omega_{M1} \times \mu_{M1} = (0.971306, 0.489051, 0.943049, 0.75516)$$

In the same way, we can calculate the evaluation matrix R_{M2}, R_{M3} on the M_2, M_3 of the regional innovation capacity sustainability M,

$$R_{M2} = (0.738501, 0.814697, 0.906822, 0.71815)$$

$$R_{M3} = (0.524957, 0.600819, 0.790181, 0.994009)$$

The results of the secondary comprehensive evaluation are tabulated, as shown in Table 4.34.

(3) Level 3 Comprehensive evaluation

Table 4.34 shows that the evaluation matrix D of regional innovation capacity M is,

$$D = \begin{bmatrix} 0.97131 & 0.48905 & 0.94305 & 0.75516 \\ 0.73850 & 0.81470 & 0.90682 & 0.71815 \\ 0.52496 & 0.60082 & 0.79018 & 0.99401 \end{bmatrix}$$

Ideal program $U_0 = \begin{bmatrix} 0.97131 & 0.90682 & 0.99401 \end{bmatrix}^T$

$$\mu = \begin{bmatrix} 1 & 0.50350 & 0.97091 & 0.77747 \\ 0.81438 & 0.89841 & 1 & 0.79194 \\ 0.52812 & 0.60444 & 0.79494 & 1 \end{bmatrix}$$

Then the overall evaluation R of regional innovation capacity is,

$R = \omega \times \mu = (0.80275, 0.65225, 0.92685, 0.84857)$ Statistics the results into tables, as shown in Table 4.35:

It can be seen that the sustainability of regional innovation capacity from strong to weak ranking is: Shanghai, Shenzhen, Beijing, Tianjin.

5. Evaluation results analysis

Table 4.35 Results of the level 3 integrated evaluation

Level 3 evaluation results	Beijing	Tianjin	Shanghai	Shenzhen
Sustainability of regional innovation capacity	0.80275	0.65225	0.92685	0.84857

Analyzes the advantages and characteristics of four regional innovation from three aspects of the evaluation results of the secondary comprehensive indicators, and discusses the shortcomings and shortcomings of the region in improving the sustainability of innovation ability.

Level 2 evaluation results	Beijing	Tianjin	Shanghai	Shenzhen
Regional innovation main body	0.97131	0.48905	0.94305	0.75516
Regional innovation environment	0.73850	0.81470	0.90682	0.71815
Regional ecological development	0.52496	0.60082	0.79018	0.99401

(1) According to the evaluation results of regional innovation subject, the regional innovation input and output in Beijing and Shanghai are relatively high, the investment of R&D funds, the proportion of personnel is relatively high, and the ability of regional innovation subject is strong. In contrast, the innovation output in Tianjin area is less than the innovation input, especially in the number of trademarks and patents.

(2) From the results of regional innovation environment evaluation, the environment of the four regions is relatively average. With the development of economy and the attention of the government, the innovation environment of the four regions is conducive to the development of regional innovation, with good infrastructure and network. The management ability of government department is also relatively strong.

(3) According to the results of regional ecological development evaluation, the results of regional ecological environment evaluation in Beijing are low. Due to the serious haze weather in Beijing in recent years, the atmospheric environment needs to be improved, in addition, the industrial wastewater and waste treatment facilities in Beijing need to be built.

Appendix: Matlab program of DEA model

1. **Matlab program of PC^{2}R model**

```
clear
X=[];
Y=[];
n=size(X', 1); m=size(X,1); s=size(Y,1);
A=[-X'    Y'];
b=zeros(n, 1);
LB=zeros(m+s,1); UB=[];
for i=1:n;
    f= [zeros(1,m) -Y(:,i)'];
    Aeq=[X(:,i)' zeros(1,s)]; beq=1;
    w(:,i)=linprog(f,A,b,Aeq,beq,LB,UB);
        E(i, i)=Y(:,i)' *w(m+1:m+s,i);
end
w
E
Omega=w(1:m,:)
mu=w(m+1:m+s,:)
```

2. **Matlab program of model $D^{\varepsilon}{}_{C}{}^{2}{}_{R}$**

```
clear
X=[];
Y=[];
n=size(X',1); m=size(X,1); s=size(Y,1);
epsilon=10^-10;
f=[zeros(1,n) -epsilon*ones(1,m+s) 1];
A=zeros(1,n+m+s+1); b=0;
LB=zeros(n+m+s+1,1); UB=[];
LB(n+m+s+1)= -Inf;
for i=1:n;
        Aeq=[X    eye(m)      zeros(m,s)    -X(:,i)
                Y    zeros(s,m)      -eye(s)          zeros(s,1)];
           beq=[zeros(m, 1 )
                Y(:,i)];
           w(:,i)=linprog (f,A,b,Aeq,beq,LB,UB);
end
w
lambda=w(1:n,:)
s_minus=w(n+1:n+m,:)
s_plus=w(n+m+1:n+m+s,:)
theta=w(n+m+s+1,:)
```

Summary
The research on evaluation has always been one of the hot spots in management science, and it is also a basic management technology that students need to master. It is also essential in the practical activities of enterprise management, and it is related to the success or failure of many fields. Among the main points of this chapter analytic hierarchy process (AHP), fuzzy comprehensive evaluation, entropy method, set-to-analysis method, the principle and calculation process of these methods, and understand the improvement methods of these typical methods.

Important Concepts and Terminology
evaluation.

Logical Framework Approach (LFA).

data envelopment analysis (DEA).

C^2R model (CCR model).

Analytic Hierarchy Process (AHP).

judgment matrix.

consistency check.

level simple sequence.

total level sequence.

fuzzy mathematics.

fuzzy comprehensive evaluation.

entropy.

set-pair analysis.

Questions and Exercises

1. Compare the investment and output levels of four cities A, B, C and D, and the input-output indicators are shown in the table below. Try to determine whether each city is DEA effective (Table 4.36).
2. A customer to choose from 3 different mobile phone to buy, the selection criteria are price, photo function, brand, appearance, memory five indicators. Please use AHP to evaluate and quantify these three mobile phones. The structural model and judgment matrix are shown in figures and tables (Figs. 4.14, 4.15, 4.16, 4.17, 4.18, 4.19 and 4.20)
3. The school will have a comprehensive performance evaluation to 4 teachers A, B, C and D. Among them, the evaluation index has the teaching level, the scientific research ability, the student work, the weights are 0.5, 0.3 and 0.2. The evaluation grade is divided into five grades: excellent, good, good, general and poor. Please use M = ($\wedge$, $\vee$) fuzzy operator, according to the expert score

Table 4.36 Input–output indicators for cities

City	Input			Output	
	Fixed assets investment completion (Billions yuan)	Total electricity consumption (Billion KWH)	Employed population (Thousand)	Regional GDP (Billions)	Public revenue (Billions)
A	6.9	1.0	90	21.3	4.0
B	6.0	1.4	82	23.5	4.5
C	26.2	5.2	198	67.8	8.0
D	24.2	3.5	110	40.1	4.1

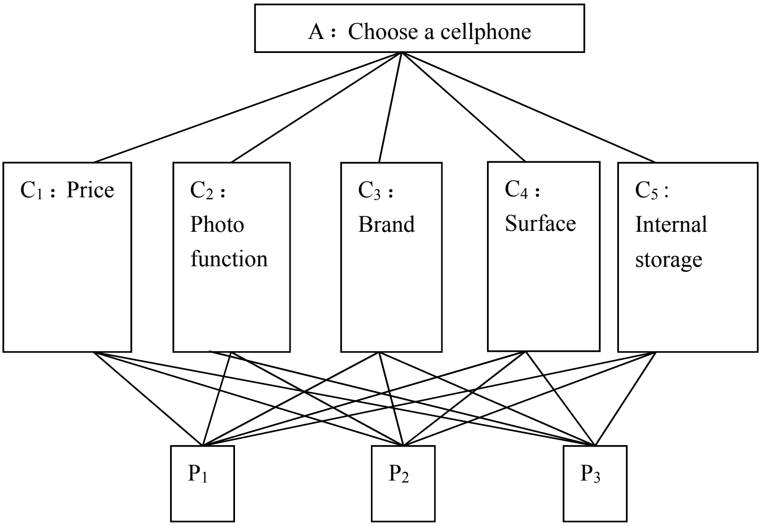


Fig. 4.14 Hierarchical model for selecting a mobile phone

A	C_1	C_2	C_3	C_4	C_5
C_1	1	3	5	3	1/2
C_2	1/3	1	5	4	1/5
C_3	1/5	1/5	1	1/2	1/4
C_4	1/3	1/4	2	1	1/5
C_5	2	5	4	5	1

Fig. 4.15 Judgment matrixes A-C

C_1	P_1	P_2	P_3
P_1	1	4	2
P_2	1/4	1	1/4
P_3	1/2	4	1

Fig. 4.16 Judgment matrixes C_1-P

C_2	P_1	P_2	P_3
P_1	1	1/6	1/5
P_2	6	1	3
P_3	5	1/3	1

Fig. 4.17 Judgment matrixes C_2-P

C_3	P_1	P_2	P_3
P_1	1	1/5	1/2
P_2	5	1	3
P_3	2	1/3	1

Fig. 4.18 Judgment matrixes C_3-P

C_4	P_1	P_2	P_3
P_1	1	1/4	3
P_2	4	1	7
P_3	1/3	1/7	1

Fig. 4.19 Judgment matrixes C_4-P

C_5	P_1	P_2	P_3
P_1	1	5	2
P_2	1/5	1	1/4
P_3	1/2	4	1

Fig. 4.20 Judgment matrixes C_5-P

(as shown in Table 4.37), give the performance of 4 teachers for comprehensive evaluation.

4. Please give brief description of application conditions of DEA, analytic hierarchy process, fuzzy comprehensive evaluation, entropy method and set-pair analysis method.
5. Empirical Analysis

(1) Enterprise profile

Huatang is a comprehensive company based on smart card technology, with smart card operating system as the core, engaged in the development, promotion and application of reading and writing machines, smart cards and related software products. Founded in 2004, around reading and writing machines and smart cards, the company has accumulated rich experience in product development, production, sales and application solutions, system integration, and formed a series of products with independent intellectual property rights.

(2) Data on indicators and standard values (Table 4.38)

(3) Determination of weights

The weight is determined by the weighted average method by Huatang senior management and consultants through the questionnaire. The relative weights of indicators in each dimension are shown in Table 7.4, and the relative weights between structure, ability and culture are $\alpha = 0.35$, $\beta = 0.35$, $\gamma = 0.3$.

(4) Evaluation results and analysis

If $k = 3$, the model is calculated as follows:

Step 1: Calculate standard values

Among them, the negative indicators are the number of customer complaints, the turnover rate of managers, the number of employee disputes, and the rest are positive indicators.

$$X_i = \begin{cases} \dfrac{x_i}{\lambda_{max\, x_i}} & x_i \text{ the bigger the better} \\ \dfrac{\lambda_{mix\, x_i}}{x_i} & x_i \textit{ the smaller the better} \end{cases}$$

Table 4.37 Expert evaluation findings

Evaluation	Teaching level					Scientific research ability					Student work				
	Excellent	Great	Good	General	Poor	Excellent	Great	Good	General	Poor	Excellent	Great	Good	General	Poor
A	0.2	0.5	0.2	0.1	0	0.15	0.15	0.6	0.1	0	0	0.5	0.1	0.2	0.2
B	0.1	0.6	0.2	0.1	0	0.1	0.1	0.7	0.1	0	0	0.1	0.3	0.3	0.3
C	0.1	0.3	0.4	0.1	0.1	0.9	0.1	0	0	0	0.1	0.1	0.1	0.4	0.3
D	0	0.2	0.2	0.4	0.2	0.5	0.2	0.3	0	0	0.5	0.3	0.2	0	0

Table 4.38 Evaluation data on strategic synergy of enterprises

Dimension	Index	Unit	Index value	Standard value	Normative value	Weight
Structure	Strategic objectives synergy	Value	70	85	0.82	0.16
	Support for strategic objectives	%	80	90	0.89	0.14
	Asset-liability ratio	%	30	30	1.00	0.11
	Return on net assets	%	25	30	0.83	0.14
	Customer loyalty	Value	80	85	0.94	0.14
	Number of customer complaints	Time	7	5	0.71	0.10
	Management turnover rate	%	10	5	0.50	0.11
	Educational background structure	%	96	96	1.00	0.10
Ability	Operating revenue	10,000	1150	1150	1.00	0.13
	Target market share	%	40	40	1.00	0.13
	Sales growth	%	20	20	1.00	0.13
	Growth rate of new product sales	%	25	25	1.00	0.13
	Turnover ratio of receivable	%	29	32	0.91	0.12
	Average wage level	Yuan	1500	1600	0.94	0.10
	Product percent of pass	%	98	98	1.00	0.13
	Supplier satisfaction	Value	90	95	0.95	0.13
Culture	Employee satisfaction	Value	85	90	0.94	0.20
	Number of staff disputes	Time	5	3	0.60	0.14
	Staff training expenditure	10,000	20	20	1.00	0.17
	Training pass	%	90	95	0.95	0.17
	Recommendation acceptance rate	%	20	30	0.67	0.14
	Information communication efficiency	Value	85	90	0.94	0.18

$$Y_i=\begin{cases}\dfrac{y_i}{\lambda_{max\ y_i}} & y_i \text{ the bigger the better}\\ \dfrac{\lambda_{mix\ y_i}}{y_i} & y_i \textit{ the smaller the better}\end{cases}$$

$$Z_i=\begin{cases}\dfrac{z_i}{\lambda_{max\ z_i}} & Z_i \text{ the bigger the better}\\ \dfrac{\lambda_{mix\ z_i}}{z_i} & Z_i \textit{ the smaller the better}\end{cases}$$

λ max, λmix is the standard value of the corresponding index.
Since the standard values are already given.
For positive indicators:

$$\bar{x} = \text{index value/standard value} = \text{normative value}$$

For negative indicators:

$$\bar{x} = \text{standard value/index value} = \text{normative value}$$

Table 4.39 is consistent with the given specification values in Table 4.38 (Tables 4.40 and 4.41)

Table 4.39 Normative values for indicators

1	2	3	4	5	6	7	8	9	10	11
0.82	0.89	1.00	0.83	0.94	0.71	0.50	1.00	1.00	1.00	1.00
12	13	14	15	16	17	18	19	20	21	22
1.00	0.91	0.94	1.00	0.95	0.94	0.60	1.00	0.95	0.67	0.94

Table 4.40 Effect function

f(x)	g(y)	h(z)
0.840	0.977	0.867

Table 4.41 Evaluation of strategic synergy

f(x)	0.840
g(y)	0.977
h(z)	0.867
C	0.981
T	0.896
D	0.937

Step 2: Calculate f (x), g (y), h (z) according to the weight given in Table 4.40, according to the formula:

$$f(x) = \sum_{i=1}^{m} ai\,\overline{xi}$$

$$g(y) = \sum_{i=1}^{m} bi\,\overline{yi}$$

$$h(z) = \sum_{i=1}^{m} ci\,\overline{zi}$$

The results are:

$$f(x) = 0.82 * 0.16 + 0.89 * 0.14 + \cdots + 1 * 0.1 = 0.840$$

$$g(y) = 1 * 0.13 + 1 * 0.13 + \cdots + 0.95 * 0.13 = 0.977$$

$$h(z) = 0.94 * 0.2 + 0.6^{*}0.14 + \cdots + 0.94 * 0.18 = 0.867$$

Step 3: Calculate the degree of coordination C, comprehensive evaluation index T and synergy D among the elements according to the given weight. According to the formula:

$$C = \left\{ f(x) * g(y) * h(z) / \left[\frac{f(x) + g(y) + h(z)}{3} \right]^{3} \right\}^{k}$$

$$T = \alpha * f(x) + \beta * g(y) + \gamma * h(z)$$

$$D = \sqrt{C * T}$$

The results are:

$$C = \{0.84 * 0.977 * 0.867/[(0.84 + 0.977 + 0.867)/3]^\wedge 3\}^\wedge 3 = 0.981$$

$$T = 0.35 * 0.840 + 0.35 * 0.977 + 0.3 * 0.867 = 0.896$$

$$D = \sqrt{0.981 * 0.896} = 0.937$$

The evaluation result of strategic synergy of Huatang Company is 0.937, which is excellent in evaluation grade. It shows that the company's strategic cooperative operation effect is good, each business unit can communicate well, the cooperation between business processes is smooth, and the operation efficiency is high. The degree of synergy is relative to other enterprises in the industry and changes with the overall trend of the industry. Compared with the same industry enterprises in the target market, most of the evaluation indexes of the empirical enterprises score higher, but the four indexes of customer complaint, manager turnover, employee dispute and suggestion acceptance are lower. Therefore, the empirical enterprise should start from the improvement management team, formulate the whole execution system composed of the strategy-oriented system, mechanism and procedure, build the excellent corporate culture, while improving the internal cohesion, realize the good synergy between the enterprise and the external elements such as customers and partners, and make the culture become the driving force to realize the strategic synergy of the enterprise.

Case Analysis
Site Evaluation of Agricultural Product Logistics Park.

With the continuous development of agricultural products market, agricultural products enterprises further improve the requirements of logistics specialization, distribution network is more complex. Compared with other industries logistics, agricultural products logistics has the following difficulties: First, agricultural products logistics quantity and variety are particularly large. In 2012, China's total agricultural product logistics reached a new high, exceeding 2.4 trillion yuan; in 2013, China's total agricultural product logistics increased by 4% on year. Second, agricultural product logistics is difficult. This difficulty mainly reflects the difficulty of packaging, transportation and storage. A large number of agricultural products in the transport process due to the lack of necessary protection and corruption, serious loss. Third, the short logistics cycle of agricultural products and the shorter shelf life of agricultural products and other products, the accuracy and punctuality of distribution services are high.

To solve the above problems, the establishment of specialized agricultural logistics park has become a key means. Logistics park is a place where logistics operations are concentrated. In several transportation modes, a variety of logistics facilities and different types of logistics enterprises are centralized in space. It is also a gathering point of logistics enterprises with certain scale and various service functions. The agricultural product logistics park is an important part of the whole agricultural product logistics system. Through the development of the agricultural product logistics park, many agricultural product logistics companies with different service functions are gathered together, and the cooperative function of the park is adopted. Realize logistics information and logistics infrastructure sharing, thus forming close cooperation. Therefore, the location of agricultural product logistics park has become the first problem to be solved in the establishment of specialized agricultural product logistics park. The location mainly considers natural factors, social factors, infrastructure and so on. According to the existing theory and the

Table 4.42 Factors influencing the location of agricultural product logistics park

Analytical factor		Evaluating indicators	Location requirements of agricultural product logistics park
Natural factor	Meteorologic condition	Factors such as annual rainfall, average humidity, general wind, temperature of each season	The agricultural products have biological characteristics, and the areas with moderate humidity should be selected when selecting the location, because the construction in the areas with too high humidity will accelerate the decay of some agricultural products, and the wind power in the tuyere area will accelerate the loss of green and make it lose freshness
	Geological conditions	Hard and persistent pressure soil texture	One of the important functions of the logistics park is to carry out loading and transshipment, so there will be large tonnage trucks in and out frequently, so the soil quality is required to have a certain bearing capacity
	Orographic condition	The external shape should be chosen in a regular and suitable area for planning	Easy access to large cars
Social factors	Business environment	Preferential industrial policies	Logistics is a labor-intensive industry, whether to attract high-quality and sufficient workers is also an important part of the location of agricultural logistics park
	Logistics costs	Transportation costs are a major component of logistics costs	Agricultural products have the characteristics of large quantity and low price base, so the logistics cost has a great influence on the price of agricultural products, so the agricultural product logistics park should be as close as possible to the demand place
	Service level	The degree to which the logistics park meets the logistics needs of users within the maximum service radius	Agricultural products themselves have the characteristics of perishable, if they cannot be sent to the demand point in time, it will directly affect the quality and price of agricultural products, resulting in heavy losses, so the service level should be paid special attention to by the logistics of agricultural products

(continued)

Table 4.42 (continued)

Analytical factor		Evaluating indicators	Location requirements of agricultural product logistics park
	Environmental requirements	Harmonious development of nature, economy and politics	Considering the requirements of preservation and cold storage, the agricultural product logistics park needs to build cold storage, and the interference to urban life should be reduced when selecting the location, and the construction of large transit hub should choose the place far from the center of the city to improve the traffic condition of the big city
Infrastructure	Transportation condition	Is it near the main road, port, railway, etc	The most important function of agricultural product logistics park is distribution. Whether the transportation is convenient directly determines the speed of agricultural product transportation and creates the time value of agricultural product
	Communal facilities	Supply electricity, heating, water, gas, communications	The construction of agricultural product logistics park often includes cold storage facilities, so it is necessary to ensure that there is sufficient electricity. The perishable characteristics determine that a large number of corrupt products will inevitably be produced, because the treatment of surrounding sewage and solid waste is required

particularity of agricultural product logistics, the factors affecting the location of agricultural product logistics park can be summarized as follows: (Table 4.42)

According to the information of the case, please select the appropriate evaluation method, establish the evaluation model, design the evaluation index system, find the relevant data, give the evaluation process and analyze the evaluation results.

Further Readings

1. Du Dong, Pang Qinghua, Wu Yan. *Modern Comprehensive Evaluation Methods and Case Selection*. Beijing: Tsinghua University Press, 2008.
2. Liu Sifeng. *Systematic Evaluation: Methods, Models & Applications*. Beijing: Science Press, 2015.
3. Zhang Bingjiang. *Analytic Hierarchy Process and Its Application*. Beijing: Electronic Industry Press, 2014.
4. Li Zuoyong, Wang Jiayang et al. *Evaluation Model and Application of Sustainable Development*. Beijing: Science Press, 2007.

5 Optimization Algorithm

Hao Zhang

Learning Target

1. Grasping the principles of particle swarm algorithm and genetic algorithm, and applying them to solve specific optimization problems.
2. Understanding the classification of optimization algorithms and understand the applicable conditions of different optimization algorithms.
3. Understanding the principles of the improved optimization algorithm.

Introduction

Big Problems in the Little Timetable

Everyone has used a timetable during his/her school years. The timetables vary with different schools, majors and studying phases. Have you ever thought the little timetable actually contains a big problem?

Unexpectedly, the University Timetable Problem (UTP) is actually a problem that has been bothering the multi-objective combination optimization of resources in various schools for a long time. This question is related to how many classes or hours we have to arrange in a week, how many weeks we have to arrange to ensure the quality of this course, whether the arrangement of all subjects in the current semester and whether the credit arrangement is reasonable, etc. The timetable problem may be something we never realized in the three or four years of university, but it is closely related to our study. A small class schedule may seem simple, but its truly reasonable planning has been proved by S. Even to be an NP-complete problem. The so-called NP-complete problem refers to the non-deterministic problem of polynomial complexity. In such a problem, many possible combinations will be produced. These

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_5

combinations may cause a combination “explosion”. Such an explosion cannot be solved by current computer technology. The problem. Although UTP is a complete NP problem, the reasonable arrangement of the timetable has great practical significance for solving the current situation of relatively scarce resources in education and relatively large numbers of students. Although the timetable problem is so difficult, genetic algorithms can solve it well! Create an initial candidate program randomly, and then screen it according to the relevant teaching goals set by the school and the situation of the students. After the screening, the corresponding operator operations are carried out, and the approximate optimal course arrangement results are found through iteration, and guidance is provided for the actual class schedule. Optimize the allocation of teaching resources. This is an example of the application of a typical optimization algorithm around us. Although we may feel relaxed because of fewer classes, pressured because of more classes, and bored because we can only follow the timetable step by step, every timetable we get with a simple appearance is optimized through intelligent algorithms. As a result, this result is a scientific and reasonable study plan since considering various important factors.

Source:

Teng Zi, Deng Huiwen, Yang Jiujun. Design and Implementation of a Course Arrangement System Based on Genetic Algorithm [J]. *Computer Applications.* 2007, 27(s2): 199–201.

5.1 Overview of Optimization Algorithm

5.1.1 Basic Concepts of Optimization Algorithms

1. **Optimization**

Optimization is an important research tool in the fields of engineering technology, scientific research, economic management, etc. It studies how to find the best solution among many solutions. For example, in engineering design, how to choose design parameters so that the design plan can not only meet the design requirements and reduce costs; in resource allocation, how to allocate limited resources so that the allocation plan can meet the basic requirements of all aspects and obtain a good economy benefit; in logistics transportation, how to choose the route can minimize the transportation cost. Optimizing this technology provides a theoretical basis and solution method for solving these problems. It is a widely used and practical science.

2. **Algorithm**

Algorithm refers to the accurate and complete description of the problem-solving scheme. It is a series of clear instructions to solve the problem. The algorithm represents a systematic method to describe the strategy of solving the problem. The algorithm can obtain the required output for a certain standard input within a limited time. Different algorithms can use different space, time or efficiency to accomplish the same task.

The instruction in the algorithm describes a calculation, which starts from an initial state and initial input, passes through a series of finite and clearly defined states, and finally produces output and stops in a final state. The transition from one state to another is not necessarily definite.

3. **Optimization Algorithm**

The algorithm that solves the optimization problem is called optimization algorithm, which can be divided into classic optimization algorithm and heuristic optimization algorithm. The establishment of the classic optimization algorithm can start from G. B. Dantzig, who proposed the simplex method for solving linear programming in 1947. Subsequently, Kamaka proposed the ellipsoid algorithm (polynomial algorithm) and the interior point method. For nonlinear problems, the scholars tried to use linear optimization theory to solve nonlinear problems at the beginning, but the effect is not ideal. Later nonlinear theories are mostly based on quadratic functions, using quadratic functions to solve other nonlinear functions. Based on this principle, many classic optimization algorithms have been proposed. Classical optimization algorithms can be divided into unconstrained optimization algorithms and constrained optimization algorithms. Among them, unconstrained optimization algorithms include: steepest, conjugate gradient method, Newton Algorithm, pseudo Newton Algorithm, trust region method. Constrained optimization algorithms include: Augmented Lagrangian Algorithms, Sequential Quadratic Programming (SQP), etc.

With the development of society, actual problems have become more and more complex. Scholars have developed some heuristic optimization algorithms (also known as intelligent optimization algorithms), such as genetic algorithm, ant colony algorithm, simulated annealing algorithm, artificial neural network, tabu search, particle swarm algorithm, fruit fly optimization algorithm, krill swarm algorithm, grey wolf algorithm, dolphin echolocation algorithm, pigeon swarm optimization algorithm, etc. These algorithms simulate some phenomena in nature, and their ideas and content involve mathematics, physics, biological evolution, artificial intelligence, neuroscience, statistical mechanics and etc., providing new ideas and means for solving complex problems.

5.1.2 Development of Intelligent Optimization Algorithms

In the early 1980s, scientists conducted in-depth research on the models, theories and application technologies of genetic algorithms, artificial neural networks, simulated annealing, and tabu search. These algorithms were collectively referred to as intelligent optimization algorithms algorithms, also known as modern heuristic algorithms. In the field of management science, commonly used algorithms are as follows:

Genetic Algorithm (GA) is a random search and optimization algorithm. It was first proposed in 1975 by Professor Holland and his students of the University of Michigan. It is mainly based on Darwin's theory of evolution and Mendel's genetics. Biological heredity, mutation and selection play an important role in the evolution of organisms. It not only enables organisms to maintain their own inherent characteristics, but also enables organisms to constantly change themselves to adapt to the new living environment. Genetic algorithm is a calculation model based on group evolution. It exchanges information through methods such as reproduction, mutation, and competition among individuals in the group to survive the fittest, thereby approaching the optimal solution of the problem step by step.

Artificial Neural Network (ANN) is based on the basic principles of biological neuron networks. Its essence is to imitate the structure and function of the brain, and an information processing system that processes large-scale information through the application of computers. The multilayer forward neuron network proposed by Ainsley and Papert is currently the most commonly used network structure.

Simulated Annealing (SA) is a general random search algorithm, proposed by N. Metropolis et al. in 1953. Its basic idea is to compare the solution process of a certain type of optimization problem with the thermal balance problem in statistical thermodynamics, and try to find the global optimal solution or approximate global optimal solution of the optimization problem by simulating the annealing process of high-temperature objects.

Tabu Search Algorithm (TS) is a global domain search algorithm, first proposed by Glover in 1986, and it is a deterministic iterative optimization algorithm. It is mainly proposed for the shortcomings of the general descent algorithm. When the general descent algorithm searches for a local optimal solution, the algorithm can easily stop automatically, and the tabu search adopts the tabu criterion to avoid the searched objects as much as possible, thereby ensuring that the difference The exploration of the search path guides the algorithm to jump out of the local optimal solution and find the global optimal solution.

Particle Swarm Optimization (PSO) algorithm is a global random search algorithm based on swarm intelligence proposed by Kennedy and Eberhart, inspired by the research results of artificial life, by simulating the migration and swarming behavior of birds in the foraging process. In 1995, they published a paper entitled "Particle Swarm Optimization" at the IEEE International Neural Network Academic Conference, which marked the birth of the PSO algorithm. PSO is also based on the concepts of "population" and "evolution". By simulating the cooperation and competition between individuals, it realizes the search for the optimal solution in a complex space. However, particle swarm optimization is not completely like other evolutionary algorithms. PSO does not perform evolutionary operator operations such as crossover, mutation, and selection on individuals, but treats each individual in the group as having no quality in the D-dimensional search space. A particle in the same volume, each particle moves in the solution space at a certain speed, and has been gathered to the best position in its own history and the best position in the history of the neighborhood to optimize the candidate solution. The PSO algorithm is relatively easy to understand, has fewer parameters, and is easy to implement. It has strong global search capabilities for nonlinear and multi-peak problems, and has received extensive attention in scientific research and practice.

5.2 PSO (Particle Swarm Optimization)

5.2.1 Algorithm Principle

Imagine a scenario: a group of birds are randomly distributed in an area, and there is only one piece of food in this area. Every bird does not know where the food is, but they all know the distance between their current location and the food. So what is the optimal strategy for finding food? The easiest and most effective way is to look for the bird closest to the food in your field of vision. If we regard food as the optimal solution and the distance between the bird and the food as the fitness of the function, then the process of finding food by the bird can be regarded as a process of function optimization.

In the PSO algorithm, the potential solution of each optimization problem can be regarded as a bird in the search space, which is called a "particle". Each particle has a fitness value, which is determined by the optimized function, and each particle also has a speed, which determines the direction and distance of their flight. Initialize it as a bunch of random particles at the beginning of the optimization. Then, all particles follow the current optimal particle to search in the solution space, and then find the optimal solution through iteration. In each iteration, the particle updates itself by tracking two extreme values. The first extreme value is the optimal solution currently found by the entire group. This extreme value is called the global extreme value. In addition, instead of using the entire population, only a part of it is used as the neighbors of the particle. The extremum among all neighbors is called the local extremum. The second extreme value is the optimal solution found by the particle itself, and this extreme value is called the individual extreme value.

5.2.2 Algorithm Model

At the beginning of the algorithm, the position and velocity of the randomly initialized particles constitute the initial population, which is uniformly distributed in the solution space. Suppose the population size of the particle swarm is M and the decision space is D-dimensional, where the coordinate position of the particle i at time t can be expressed as $X_{\mathrm{i}}^t = \left(x_{i1}^t, x_{i2}^t, \ldots, x_{id}^t\right), i = 1, 2, \ldots, M$, and the velocity of the particle i is defined as the distance the particle moves in each iteration. With v,$V_{\mathrm{i}}^t = \left(\mathrm{v}_{i1}^t, v_{i2}^t, \ldots, v_{id}^t\right), i = 1, 2, \ldots, M$ the particle i is the flight speed and position in the d-th dimension subspace at time t are adjusted according to the following formula:

$$v_{id}^{t} = \omega v_{id}^{t-1} + c_1 r_1 \left(p_{id} - x_{id}^{t-1}\right) + c_2 r_2 \left(p_{gd} - x_{id}^{t-1}\right) \tag{5.1}$$

$$v_{id}^{t} = \begin{cases} v_{\max}, v_{id}^{t} > v_{\max} \\ -v_{\max}, v_{id}^{t} < -v_{\max} \end{cases} \tag{5.2}$$

$$x_{id}^{t} = x_{id}^{t-1} + v_{id}^{t} \tag{5.3}$$

Among them, ω is the weight of inertia; c_1 and c_2 are learning factors, also called acceleration constants; r_1 and r_2 are random numbers between [0, 1]; p_{gd} represents the most historical location record in the entire group, that is, the global extreme value $gbest$; p_{id} represents the historical optimal position record of the current particle, the individual extreme value $pbest$; v_{id} is the velocity of the particle, $v_{id} \in [-v_{\max}, v_{\max}]$, $v_{\max}$ are constants, set by the user to limit the velocity of the particle. The algorithm at this time is the standard particle swarm algorithm. When $\omega = 1$, it is called the original particle swarm algorithm.

The right side of Formula (5.1) consists of three parts:

The first part is the "inertia", which represents the influence of the particle's previous speed on the particle's trajectory, and represents the particle's trust in the current state of its own motion. It is mainly through the speed at the particle's previous moment is multiplied by a control factor called "inertia weight" to realize the process of inertial motion.

The second part is the "cognition", which represents the influence of the particle's own experience on the particle's trajectory, that is, the particle's own thinking, which represents the tendency of the particle to approach the best position in its own history.

The third part is the "society", which represents the influence of group experience on the trajectory of particles, represents the group historical experience of collaboration and knowledge sharing between particles, and shows that particles have a tendency to approach the best position in the group or neighborhood history.

5.2.3 Algorithm Flow

The particle swarm optimization algorithm has the characteristics of simple programming and easy implementation. The specific steps for its realization are given below:

Step 1: Initialize the particle swarm, initialize the swarm size N, particle position x_i and velocity V_i.

Step 2: Calculate the fitness value $F_{it}[i]$ of each particle.

Step 3: For each particle, compare its fitness value $F_{it}[i]$ with the individual extreme value $p_{best}(i)$. If $F_{it}[i] > p_{best}(i)$, replace $p_{best}(i)$ with $Fit[i]$.

Step 4: For each particle, compare its fitness value $Fit[i]$ with the global extreme value g_{best}, if $F_{it}[i] > p_{best}(i)$, use $F_{it}[i]$ instead of $g_{best.}$

Step 5: Update the velocity v_i and position x_i of the particle according to formulas (5.1) and (5.3).

Step 6: If the end condition is met (the error is good enough or the maximum number of cycles is reached) exit, otherwise return to Step 2.

5.2.4 Parameter Analysis and Setting

There are some explicit and implicit parameters in the PSO algorithm, and their values can be adjusted to change the way the algorithm searches the problem space. In the basic PSO algorithm, the parameters that need to be adjusted mainly include population size, maximum speed, inertia weight, and acceleration factor.

1. **Population Size**

Generally speaking, a population size in the range of 20–40 can ensure a comprehensive search of the solution space. For most problems, a population size of 10 can achieve better results, but for some specific categories or a more difficult problem, the population size sometimes needs to be in the range of 100–200.

2. **Maximum Speed**

The maximum velocity $V_{\max}$ determines the resolution (or accuracy) of the area between the particle's current position and the best position. If $V_{\max}$ is too large, the particle is likely to miss a good solution; if $V_{\max}$ is too small, the particle cannot search sufficiently outside the local area, and it is easy to fall into the local optimal value. Generally speaking, $V_{\max}$ is usually set to the range width of the particle. For example, if the particle x (x_1, x_2, x_3), x_1 belong to [−10, 10], then the size of $V_{\max}$ is 10.

3. **Inertia Weight**

The inertia weight ω is a digital quantity used to indicate the degree of influence of the particle's previous velocity on the current velocity. The setting of the inertia weight will affect the balance between the global search ability and the local search ability of the particle.

Using a larger inertia weight, the algorithm has a strong global search capability. Because the inertia weight determines the degree to which the speed at the previous moment is retained, a larger value can enhance the ability to search for areas that have not been reached before, which is conducive to enhancing the algorithm's global search ability and jumping out of local extreme points.

The adjustment strategies of the inertia weight ω mainly include linear change, fuzzy self-adaptation and random change, among which the linear decreasing strategy is the most used.

4. **Acceleration Factor**

The acceleration factors c_1 and c_2 are a very important set of parameters, used to adjust the particle's own experience and group experience, which will affect the particle's trajectory. If the value of c_1 is 0, then only the group experience will affect the movement of the particles. At this time, its convergence speed is faster, but for some complex problems it is easy to fall into local convergence; if the value of c_2 is 0, only its own experience affects the particles If there is no information interaction ability between particles and particles, then a group of size M is equivalent to running a single particle M times, which completely loses the characteristics of the swarm intelligence algorithm and it is difficult to obtain the optimal solution; if Both c_1 and c_2 are 0, the particle does not contain any empirical information and can only search a limited area, so it is difficult to find a better solution.

The acceleration factors c_1 and c_2 can be regarded as a control parameter, set $\phi = c_1 + c_2$. If $\phi = 0$, the coordinate value X of the particle simply increases linearly. If ϕ is very small, the control of particle speed is very small, so the movement trajectory of the group changes very slowly. When ϕ is larger, the change frequency of the particle's spatial position increases, and the particle change step size increases accordingly. Generally speaking, when $\phi = 4.1$, it has a better convergence effect.

5.2.5 Case Analysis

Example 5.1 Find the minimum value of the Sphere function.

Sphere function: $f(x) = \sum_{i=1}^{n} x_i^2$ is a commonly used test function of PSO algorithm, which can be used to evaluate the optimization efficiency of PSO algorithm. Here it selects the two-dimensional Sphere function, uses MATLAB software to write particle swarm algorithm program to solve, verify the effectiveness of particle swarm algorithm in solving the function optimization problem, and analyze through simplified model, which helps to understand the principle and principle of particle swarm optimization algorithm. The specific calculation process.

The function selected here is $f = x_1^2 + x_2^2$, and find its minimum value.

Solution:

1. **Parameter Assignment**

Population size N = 40;

The maximum number of iterations M = 1000 times;

Inertia weight $w = 0.6$;

Acceleration factor $c_1 = c_2 = 2$;

Particle dimension D = 2。

2. **Calculation Process**

The first step is to initialize the particle swarm. For ease of understanding, we assume that the group size is 2, and the initial position of particle 1 is $X_1 = (2, 1)$, the initial velocity is $V_1 = (-1, 1)$, and the initial position of particle 2 is $X_2 = (-1, 1)$, the initial speed is $V_2 = (1, -2)$.

The second step is to initialize the individual extreme value of the particle. Calculate the fitness value fitness (X_i) of each particle and compare it with the individual extreme value p(i). At the beginning of the optimization, the initial position of all particles is its best position, so p(1) = 5, p(2) = 2, $Y_1 = X_1 = (2, 1)$, $Y_2 = X_2 = (-1, 1)$.

The third step is to initialize the global extremum. By comparing the individual extrema of all particles, we can get the global extremum pg = x(2, :) = (−1, 1), fitness(pg) = 2.

The fourth step is to update the velocity v_i and position x_i of the particle according to Formulas (5.1) and (5.3), and then update the individual extreme value of the particle.

v(i, :) = w*v(i, :) + c1*rand1*(y(i, :)-x(i, :)) + c2*rand2*(pg-x(i, :));

x(i, :) = x(i, :) + v(i, :);

Among them w = 0.6, c1 = c2 = 2, rand is a random number between [0,1], suppose rand1 = 0.4, rand2 = 0.7.

Through calculation, the velocity of particle 1 after the first update is v(1, :) = (−4.8, 0.6), and then according to the particle position update formula, the position of particle 1 after the first optimization is x(1, :) = (−2.8, 1.6), since the updated p(1) = 10.4, the individual extreme value of particle 1 is still y(1, :) = (2, 1). In the same way, the velocity of particle 2 after the first optimization is v(2, :) = (0.6, −1.2), and the position is x(2, :) = (−0.4, −0.2), because at this time p(2) = 0.2, the individual extremum is optimized as y(2, :) = (−0.4, −0.2).

The fifth step is to update the global extremum of the particle swarm. Compare the individual extrema of all particles after the first optimization with the global extremum before optimization. At this time, the global extremum is updated to pg = (−0.4, −0.2), fitness(pg) = 0.2.

This is true for every subsequent cycle. Through the first cycle, we can see that the position of particle 2 has been optimized, and the global optimal position has also been optimized. After many cycles, the optimal solution is finally obtained.

3. **Programming**

The program is mainly composed of two files, among which pso.m is the interface file and fitness.m is the function to be solved. Different problems can be solved by modifying the objective function in fitness.m.

(1) pso.m file

function [xm, fv] = PSO(fitness, N, c1, c2, w, M, D)

Among them, %fitness- is the objective function, N is the population number, c1, c2 are the learning factors, w is the inertia weight, M is the number of iterations, and D is the particle dimension.

```
format long；
% Initial population
for i=1：N
  for j=1：D
        x(i, j)=randn；   % Initialize particle position randomly
        v(i, j)=randn；   % Initialize particle speed randomly
    end
end
```

%Here calculate the fitness of each particle itself, and initialize the individual extreme value and global extreme value of the particle

```
for i=1：N
   p(i)=fitness(x(i, ：))；
   y(i, ：)=x(i, ：)；
end
```

```
pg = x(N, :);
for i=1:(N-1)
    if fitness(x(i, :))<fitness(pg)
        pg=x(i, :);                %pg is the global optimal
    end
end
%Perform PSO algorithm steps
for t=1:M
    for i=1:N
        v(i, :)=w*v(i, :)+c1*rand*(y(i, :)-x(i, :))+c2*rand*(pg-x(i, :));
        x(i, :)=x(i, :)+v(i, :);
        if fitness(x(i, :))<p(i)
            p(i)=fitness(x(i, :));
            y(i, :)=x(i, :);
        end
        if p(i)<fitness(pg)
            pg=y(i, :);
        end
    end
 Pbest(t)=fitness(pg);
end
xm = pg';
fv = fitness(pg);
```

(2)Objective function fitness.m file

```
function f=fitness(x)
f=x(1).^2+x(2).^2；
End
```

(3) Enter “call m file” on the command line

Enter [xm, fv] = PSO(@fitness, 40, 2, 2, 0.6, 1000, 2) in the command line, run the program, and get the result.

4. Result Analysis

The optimization results of objective function $f = x_1^2 + x_2^2$ are as follows:

xm = 1.0e−055*

x(1) = 0.48647304974433

x(2) = 0.11225025578588

fv = 2.492561480515418e−111

Among them, fv is the optimal value, and xm is the value of the independent variable corresponding to the optimal value.

It should be noted that the minimum value of the objective function should be zero. However, due to the random selection of the initial value and the limitation of the number of iterations, the minimum value of the final solution of the function is not 0, but very close to zero. Through experiments, it is found that when the number of iterations reaches 4000, the minimum value is 0.

5.2.6 Advantages and Disadvantages of the Algorithm

In recent years, PSO has developed rapidly and has been widely used in many fields. Among them, applications in typical theoretical problems include: combinatorial optimization, constrained optimization, multi-objective optimization, and dynamic system optimization. Applications in the actual production areas include: power systems, automatic control, filter design, pattern recognition and image processing, data clustering, machinery, chemical engineering, communications, robotics, biological information, economics, medicine, task distribution, TSP, etc.

1. **Advantages of the Algorithm**
   (1) The algorithm rules are simple, easy to implement, and widely used in engineering applications;
   (2) The convergence speed is fast, and there are many measures to avoid falling into local optimum;

(3) There are fewer parameters to be adjusted, and there are relatively mature theoretical research results to the selection of parameters at present.

2. **The Disadvantages of the Algorithm**

(1) The optimization ability of PSO algorithm mainly depends on the interaction and mutual influence between particles. If the interaction and mutual influence between particles are removed from the algorithm, the optimization ability of the PSO algorithm becomes very limited. Since standard PSO algorithm optimization relies on competition and cooperation between groups, the particles themselves lack a mutation mechanism. Once a single particle is constrained by a certain local extreme value, it is difficult for it to escape this constraint. At this time, it needs to rely on other particles. Successfully discovered. In the initial stage of the algorithm operation, the convergence speed is relatively fast, and the motion trajectory swings in a sine wave, but after a period of operation, the speed starts to slow down or even stagnate. When the speed of all particles is almost zero, the particle swarm loses the ability to further evolve, and the algorithm execution can be considered to have converged. In many cases (such as the optimization of high-dimensional complex functions), the algorithm does not converge to the global extremum, or even the local extremum. This phenomenon is called premature convergence or stagnation. When this phenomenon occurs, the particle swarm is highly clustered, and the diversity is severely lacking. The particle swarm will not jump out of the gathering point for a long time or never. Therefore, a large number of improvements to the particle swarm optimization algorithm are focused on increasing the diversity of the particle swarm, so that the particle swarm can maintain the ability of further optimization throughout the iteration process.

(2) The initialization process is random. In most cases, the initial solution group can be guaranteed to be evenly distributed, but the quality of the individual cannot be guaranteed, and part of the solution group is far from the optimal solution. If the initial solution group is better, it will help the efficiency of the solution and the quality of the solution.

5.2.7 Improvement of Particle Swarm Algorithm

Because the particles in PSO always gather in the best position in their own history, neighborhood or group history, the rapid convergence effect of the particle population is formed, and the phenomenon of falling into a local extreme, premature convergence or stagnation is easy to occur. At the same time, the performance of PSO also depends on algorithm parameters. In order to overcome the above shortcomings, researchers from various countries have successively proposed various improvement measures. This paper divides these improvements into three categories: particle swarm initialization, neighborhood topology, and hybrid strategy.

1. **Particle Swarm Initialization**

Research shows that particle swarm initialization will have a certain impact on the performance of the algorithm. In order to make the initial population evenly cover the entire search space as much as possible, thereby improving the global search ability, Richard and Ventura proposed a population initialization method based on centroidal voronoi tessellations (CVTs); Xue Mingzhi et al. used orthogonal design methods to initialize the population; Campana et al. rewritten the standard PSO iterative formula into a linear dynamic system, and based on this, studied the initial positions of the particle swarms so that they have orthogonal trajectories.

2. **Neighborhood Topology**

The particle swarm algorithm can be divided into global model *gbest* and local model *pbest* according to whether the particle neighborhood is the entire group. In the *gbest* model, all particles in the group can exchange information, and each particle moves to the best position in the history of all particles. Although the *gbest* model has a faster convergence rate, it is easy to fall into the local extremum. In order to overcome the shortcomings of the *gbest* model, the researchers limited the information exchange range of each particle and proposed various local models *pbest*. For example, in the initial stage of search, Suganthan defines the domain of each particle as itself; as the number of iterations continues to increase, the neighborhood will gradually expand to the entire population. For example, Veeramachaneni uses the fitness distance ratio to select particles adjacent to the particle. One of the most commonly studied methods is to divide the neighborhood of social relations according to the index number of the particle storage array, which mainly include: ring topology, wheel topology, tower topology, star topology, von Neumann topology, and random topology. For different optimization problems, the performance of these topologies is different, but overall, random topology can show better performance for most problems, followed by von Neumann topology.

In addition, in the standard PSO algorithm, all particles only move to the best historical position of themselves and the neighborhood, without learning from other excellent particles in the neighborhood, which will cause waste of information resources and even fall into local extremes; In view of these problems, Kennedy et al. proposed a fully informed particle swarm (FIPS). In FIPS, in addition to the best historical position of itself and its neighborhood, each particle can also learn other outstanding particles in the neighborhood.

3. **Mixed Strategy**

Hybrid strategy is the application of other technologies or traditional optimization algorithms or other evolutionary algorithms to PSO to improve particle diversity, improve local development capabilities, enhance particle global exploration capabilities, convergence speed and accuracy. There are generally two ways of such a combination: one is to use other optimization techniques to adjust parameters, such as shrinkage factor, inertia weight, acceleration constant, etc.; the other is to combine PSO with other evolutionary algorithms or other technologies. For example, Clerc combined ant colony algorithm and PSO algorithm to solve discrete optimization problems; Robinson and Juan et al. combined GA and PSO for antenna optimization design and recurrent neural network design, respectively; Angeline introduced tournament selection into PSO algorithm, based on individual current position fitness, compare each individual with several other individuals, and then sort the entire group according to the comparison result, replace the worst half of the position and speed with the best half of the current position and speed in the particle swarm, while retaining The best position of the individual remembered by each individual; Higashi introduced Gaussian mutation into the PSO; El-Dib et al. cross-operated the particle position and velocity; Miranda et al. used multiple operations of mutation, selection, and reproduction at the same time. Adapt to determine the best position in the neighborhood, inertia weight and acceleration constant in the velocity update formula; Zhang Wenjun et al. in Tsinghua University use differential evolution (DE) operation to select the best position of the particle in the velocity update formula; while Kannan et al. use DE to optimize The inertia weight and acceleration constant of PSO.

In addition, there are other hybrid PSOs:

(1) Gaussian PSO: Secrest et al. introduced Gaussian function into the PSO algorithm to guide the movement of particles; Gaussian PSO no longer needs inertial weights, and the acceleration constant is generated by random numbers that obey the Gaussian distribution.
(2) Chaos particle swarm optimization: According to the ergodicity of chaotic motion, a chaotic sequence generates based on the historical best position of the particle swarm, and randomly replaces the position of a particle in the particle swarm with the optimal position in this sequence, Eberhart proposed chaos particle swarm optimization (CPSO).
(3) Immune particle swarm optimization: The biological immune system is a highly robust, distributed, self-adaptive nonlinear system with strong recognition capabilities, learning and memory capabilities. Hu X introduced the immune information processing mechanism of the immune system (antibody diversity, immune memory, immune self-regulation, etc.) into the PSO, and respectively proposed an immune PSO based on vaccination and an immune PSO based on immune memory.
(4) Quantum particle swarm optimization: Poli R. uses quantum individuals to propose discrete PSO, while Jelmer uses quantum behavior to update particle positions.

5.3 GA (Genetic Algorithm)

5.3.1 Overview of Genetic Algorithm

1. **The Concept of Genetic Algorithm**

In the mid-nineteenth century, Darwin created the theory of biochemical evolution, the core of which is survival of the fittest and survival of the fittest. At the same time, in the evolution of individuals or populations, it is always accompanied by crossover and mutation of genes to produce new individuals and populations, thereby realizing biological evolution. Professor John Holland of the University of Michigan in the United States was inspired by Darwin's theory of evolutionary biology and invented an optimization algorithm based on biological genetics and evolutionary mechanisms for intelligent and comprehensive search, Genetic Algorithm (GA). GA embodies and realizes the replication, crossover, mutation, dominance and other processes and elements in the genetic process through operators, and optimizes them through a limited number of iterations.

GA mainly realizes the optimization of complex problems. At the same time, due to its inherent heuristic and operational independence, it can be applied to many fields, including: function optimization, combinatorial optimization, production scheduling, automatic control, image processing, machine learning, Robotics, data mining, etc.

2. **Characteristics of Genetic Algorithm**

The advantages of GA are as follows:

(1) Feasible solutions can be expressed broadly
(2) Optimization by group
(3) No auxiliary information required
(4) Intrinsic heuristic random search
(5) It is not easy to fall into local optimum
(6) It has inherent parallel computing capability
(7) It has good scalability and can be combined with other technologies.

GA has the following limitations:

(1) Irregular and inaccurate coding may occur
(2) The characteristics of a single code of genetic algorithm may make it unable to express the constraints of the optimization problem
(3) The search efficiency is low
(4) It may appear premature convergence

Table 5.1 Terms and explanations in genetic algorithm

Terms	Explanations
Chromosomes	Solution encoding (encoding form)
Gene	The characteristics of the components in the coding of the solution
Allele	Eigenvalue of coded components
Locus	The position of the gene in the string
Genotype	Solution structure
Phenotype	Decoding parameters, decoding methods, excellent individuals
Individual	Solution or its corresponding code
Survival of the fittest	Choosing according to the size of each individual fitness value
Group	Solution set composed of many individuals
Reproduction	A set of excellent solutions selected according to the fitness value
Crossover	The selected individual forms a set of new solutions through crossover
Mutations	The process of changing one or some components in an individual code

3. **Terms in Genetic Algorithm**

GA draws on the ideas and terminology in genetics. Table 5.1 lists the main terms and gives an explanation of these terms in the genetic algorithm.

5.3.2 The Basic Process of Genetic Algorithm

1. **The Operation Process of Genetic Algorithm**

The main process of GA is to simulate and operate mathematically the selection, duplication, crossover, mutation and other phenomena in natural genetics, starting from any initial random population, and selecting by calculating the fitness of each individual in the initial population operating. Subsequently, the selected individuals are regarded as the parents, and genetic crossover and mutation are performed between the parents to produce new individuals and populations, and then operations such as selection, crossover, and mutation are performed on the new populations, and so on. When the number of iterations reaches the set number, the genetic algorithm will terminate. It should be noted that certain calculation processes involve the situation that the number of individuals in the population remains unchanged, and at this time, a replication process should be added to the genetic algorithm process. The genetic algorithm operation process is shown in Fig. 5.1.

2. **Basic Operation of Genetic Algorithm**

From the above basic process, we can know that the main operations in genetic algorithms are selection, crossover and mutation.

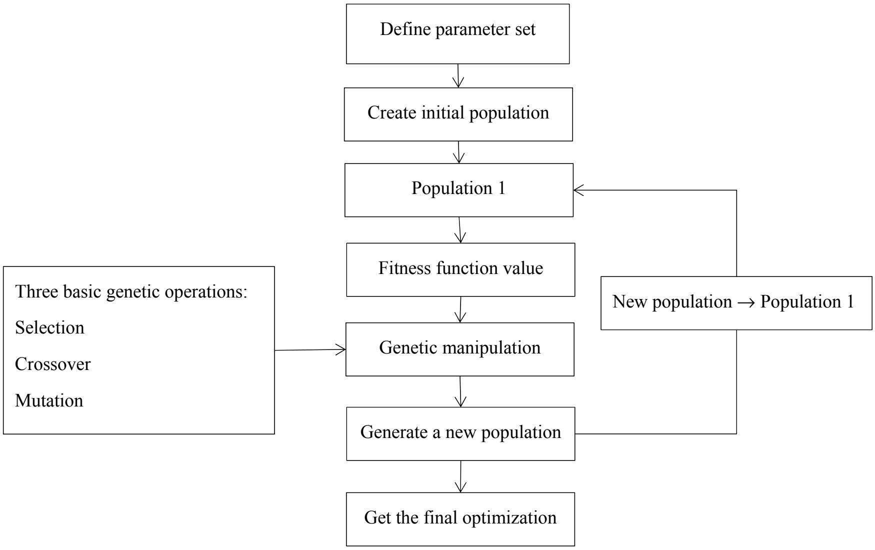


Fig. 5.1 Operation flow chart of genetic algorithm

(1) Selection. According to the fitness function value, many individuals in the population are screened, good individuals are retained, and weak individuals are eliminated. The selected individuals will also serve as parents to breed new individuals or populations. The principle of selection follows that the larger the fitness function value, the higher the probability of being selected. We use a simple example to illustrate the selection process.

    Assuming that the fitness function values of U1, U2, U3, and U4 are 1, 2, 3, and 4, respectively, the total fitness of the group is F = 10, and their respective selection probabilities Pk = Uk/F = 0.1, 0.2, 0.3, 0.4. It can be seen that U4 with the largest fitness function value has the largest probability of being selected, while U1 with the smallest fitness value has the smallest probability of being selected.

(2) Crossover. After completing the selection, a crossover process is to be carried out. The crossover process can be said to be the most important process in genetic algorithms, because a new generation of individuals can be produced through crossover operations, and at the same time, the new generation of individuals retain the characteristics of excellent parents. Combine the selected individuals in pairs, and perform crossover operations in each pair according to a certain probability.

    For example, there are currently 4 individuals, which are A = (1, 0, 1, 0, 1), B = (0, 0, 1, 0, 1), C = (1, 1, 1, 0, 1), D = (1, 1, 1, 0, 0), divide A and B into a pair, and C and D into a pair. Assuming that the position where the crossover occurs is 4 and the probability is 0.5, then the result after the crossover may be: A = (1, 0, 1, 0, 1), B = (0, 0, 1, 0, 1), C = (1, 1, 1, 0, 0), D = (1, 1, 1, 0, 1).

(3) Mutation. Mutation is the operation of changing genes in selected individuals with a certain probability. The probability of each gene mutation = the set probability value/the length of individual chromosomes. In general, individual codes are in binary form, so the length of the chromosome is usually the length of the string. In binary, the mutation of a gene is that if the gene here is 1, then the gene will be 0 after the mutation.

For example, the existing individual X = (1, 1, 0, 1, 0, 0, 0, 1, 0). Assuming that the probability of mutation is 0.7, the possible mutation result is X = (0, 1, 0, 1, 0, 0, 0, 1, 0).

5.3.3 The Realization of Genetic Algorithm in MATLAB

As a tool for implementing GA, MATLAB has a powerful computing capability. Below, the process of selection, crossover, and mutation in GA is introduced through MATLAB.

1. **Implementation of Selection**

First, we randomly create a 5×5 binary matrix, Chrom, and calculate the fitness function value of each individual, FitnV. Then with 0.6 as the generation gap (the percentage of the parent being eliminated), the random traversal sampling principle is used for selection operation, and a new population SelCh is generated. The realization in MATLAB is shown in Fig. 5.2.

2. **Implementation of Crossover**

First, create a 2×8 matrix, Chrom, perform a single-point crossover operation on Chrom, and set the crossover probability to 0.5. After the crossover operation, a new matrix, NChrom is obtained. The realization in MATLAB is shown in Fig. 5.3.

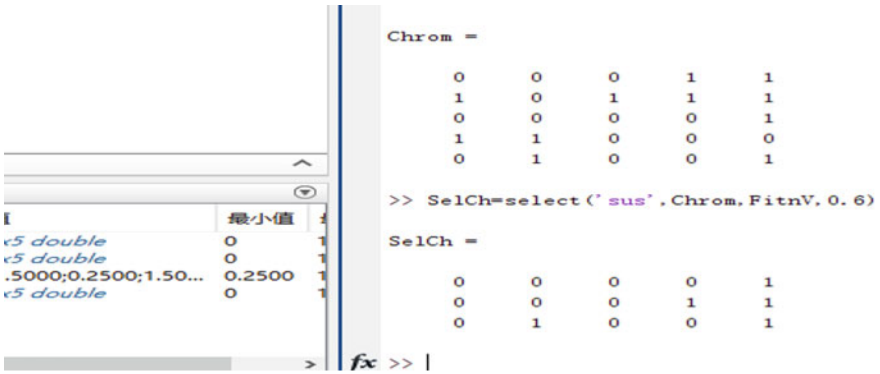


Fig. 5.2 MATLAB selection operation

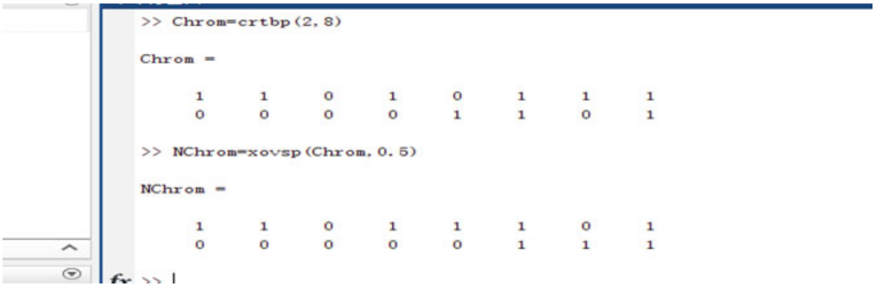


Fig. 5.3 MATLAB crossover operation

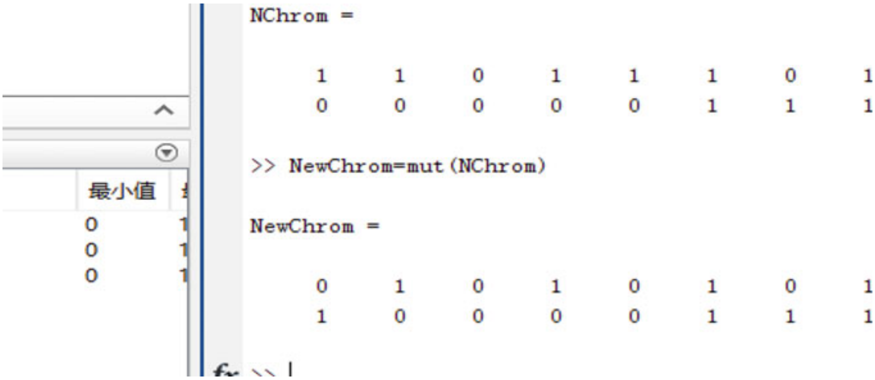


Fig. 5.4 MATLAB mutation operation

3. **Implementation of Mutation**

According to the new matrix generated by the crossover operation, NChrom performs discrete mutation operation, and its overall mutation probability is the default value (0.7). After the mutation operation, a new matrix, NewChrom is generated. The implementation in MATLAB is shown in Fig. 5.4.

5.4 FOA (Fruit Fly Optimization Algorithm)

5.4.1 Algorithm Principle and Model

Fruit fly is an insect widely found in temperate and tropical regions. It has better olfactory and visual sensitivity than other species. In the process of foraging, individual fruit flies will first use their own olfactory organs to seek food, while at the same time sending to or receiving from the other fruit flies around the odor information. Next, the fruit flies will use their visual organs to compare and find the position of the fruit flies in the current group that has collected the best odor information. Other fruit flies in the group will fly to this position and continue to search.

Fruit Fly Optimization Algorithm (FOA) is a swarm intelligence heuristic optimization method proposed by Chinese scholar Pan Wenchao (2011), which simulates the entire intelligent process of fruit fly foraging activities. FOA is suitable for solving optimization problems with complex constraints and strong correlation between the component elements of the solution, and has been widely used in many classical combinatorial optimization fields, such as gasification parameter optimization, resource constraint scheduling problems, and so on.

FOA is an emerging group intelligent optimization algorithm based on the bionics principle of fruit flies foraging behavior. FOA simulates the predation process of Drosophila using its keen sense of smell and vision, and FOA realizes a group iterative search of the solution space. The FOA principle is easy to understand, simple to operate, easy to implement, and has strong local search capabilities. FOA simulates the foraging behavior of fruit fly populations and uses a mechanism based on fruit fly group collaboration to perform optimization operations. The FOA mechanism is simple. The entire algorithm only includes two parts: olfactory search and visual search. The key parameters are only the population number and the maximum iterative search times. FOA adopts population-based global search strategy, group collaboration, information sharing, and has good global optimization capabilities. As a general-purpose algorithm, FOA does not depend on the specific information to solve the problem, and is suitable for mixing with other algorithms, and it is easy to obtain a hybrid algorithm with better performance.

Since it was proposed, FOA has attracted the attention of many scholars, and has been successfully applied to such things as financial crisis early warning modeling, multi-dimensional knapsack problem, power forecasting, neural network parameter optimization, supply chain location allocation problems, online auctions, logistics services and many other fields.

This section summarizes the steps of FOA as follows:

STEP 1: Initialize the parameters, set the fruit fly population size $sizepop$, the number of iterations $maxgen$, and initialize the fruit fly population position X_axis, Y_xis;

Step 2: Initialize the flight direction and distance of the individual fruit fly randomly, where $Random\ Value$ is the search distance;

$$X_i = X\_axis + Random\ Value$$
$$Y_i = Y\_axis + Random\ Value$$

Step 3: Estimate the $Dist_i$ between the individual fruit fly and the origin, and calculate the taste concentration judgment value S_i;

$$Dist_i = \sqrt{x_i^2 + y_i^2}$$
$$S_i = 1/Dist_i$$

Step 4: Calculate the taste concentration of individual fruit flies (fitness value)

$$Smell_i = function(S_i)$$

Step 5: Find the fruit fly individuals with the best taste concentration, and let the fruit fly population fly to the optimal position.

$$\begin{aligned} [bestSmell\ bestindex] &= \min(Smell_i) \\ Smellbest &= bestSmell \\ x\_axis &= X(bestindex) \\ y\_axis &= Y(bestindex) \end{aligned}$$

Step 6: Iterative optimization, repeat Steps 2–5 until the maximum number of iterations is reached.

From the above calculation steps of FOA, we can see that the standard FOA adopts a global random search strategy based on the population, and guides the next search of the population by tracking the information of the current optimal solution, so that the population can carry out a local random search centered on the current optimal solution, searching in a better direction.

5.4.2 Algorithm Evaluation and Improvement

FOA has strong global optimization capability and high convergence accuracy, and has great research value in the engineering field. However, the fixed search step size adopted by the original FOA cannot take into account the global search capability and the local search capability at the same time, which affects the convergence speed and the optimization capability. And the algorithm is easy to fall into the local optimal and difficult to jump out, leading to premature convergence. This section focuses on these shortcomings of the algorithm, and improves the algorithm itself to make it better adapted to the solution of the location model.

This section proposes an improved fruit fly algorithm (IFOA), which dynamically adjusts the search step size of the algorithm by introducing a weight function. The purpose of selecting dynamic adjustment is to divide the search into two parts: first, use a larger search step to perform a global search in a larger area to ensure the ability of the global search; then use a smaller step to search to refine the search in a small area, to ensure local search capabilities. The random search expression for individual fruit flies is as follows:

$$\begin{aligned} X_i &= X\_axis + \psi \cdot rand() \\ Y_i &= Y\_axis + \psi \cdot rand() \end{aligned}$$

Among them, ψ is the weight function of the dynamic search step, and the formula is as follows:

$$\psi = \exp(-\left|\frac{Smell_i - Smell_{i-1}}{Smell_i}\right|^{\zeta})$$

Among them, $Smell_i$ is the optimal fitness value of the i-th generation, and is the optimal fitness value of the i-th generation. When the ratio changes greatly, it indicates that the fruit fly population is searching in a new space, and the increase of the weight function is conducive to the global search; when the ratio change rate is small, it means that the fruit fly population is in the local search, and the weight function reduction is beneficial to the local search. Refined search for search. Among them, $\zeta > 0$ and with the increase of ζ, the search ability of fruit flies increases, and it is not easy to fall into the local optimum, which can be adjusted according to actual problems.

Except for the random search expressions, the expressions of the other IFOA algorithm steps are consistent with the basic FOA principle.

In order to test and analyze the cruise performance of the improved algorithm, this section selects the following two classic test functions to test FOA and IFOA respectively. The specific function form, search interval, theoretical extreme value and function type are shown in the following table (Table 5.2)

The specific initialization parameters are set to the population size $sizepop = 20$, the maximum iteration number $maxgen = 100$, and the position of the fruit fly population is randomly initialized. The result comparison of function test is shown in the following table (Fig. 5.5, Table 5.3)

It can be seen from the test results that IFOA has stronger global and local search capabilities and higher convergence accuracy than FOA. Under the same conditions, no matter the average value or the standard deviation, the IFOA results are better than FOA. Under the same convergence accuracy conditions, the effective number of iterations of IFOA is less than FOA, indicating that IFOA

Table 5.2 Test function

Test function	Range	Optimal	Peak
$\min f_1 = -\exp(-0.5 \cdot \sum_{i=1}^{n} x_i^2)$	[−1, 1]	−1	Single
$\min f_2 = \frac{\sin^2(\sqrt{\sum_{i=1}^{n} x_i^2}) - 0.5}{(1 + 0.001(\sum_{i=1}^{n} x_i^2))^2} + 0.5$	[−100, 100]	0	Multiple

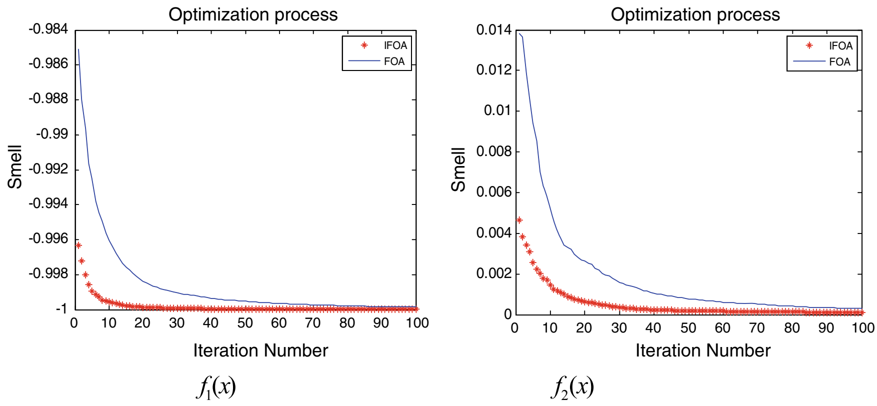


Fig. 5.5 Test result

Table 5.3 Mean and Standard Deviation of function test

Function	Result	FOA	IFOA
$f_1(x)$	Mean	3.0546×10^{-6}	-1
$f_2(x)$	Std	7.5382×10^{-6}	0
	Mean	3.1266×10^{-6}	3.1266×10^{-6}
	Std	8.6307×10^{-6}	8.6307×10^{-6}

converges faster. Therefore, we can know that the convergence accuracy and convergence speed of the IFOA proposed in this section have been improved, and the applicability is more extensive, proving that it is effective in improving FOA.

5.5 WOA (Whale Optimization Algorithm)

5.5.1 Algorithm Principle and Implementation

This section will introduce a new meta-heuristic optimization algorithm—Whale Optimization Algorithm (WOA), which imitates the hunting behavior of humpback whales. The optimization results show that compared with the most advanced optimization methods, WOA is highly competitive. The whale optimization algorithm is a new meta-heuristic group optimization algorithm proposed by Australian scholars Mirjalili et al. (2016). It is inspired by the predation behavior of humpback whales with bubble nets. It is carried out through the predation process of humpback whales constantly approaching their prey. Mathematical simulation to find the optimal solution to the problem to be solved. Whale optimization algorithm is also a meta-heuristic optimization algorithm. The main difference between WAO and other swarm optimization algorithms is that the former uses random or best search agents to simulate hunting behavior, and uses spirals to simulate the bubble net attack mechanism of humpback whales.

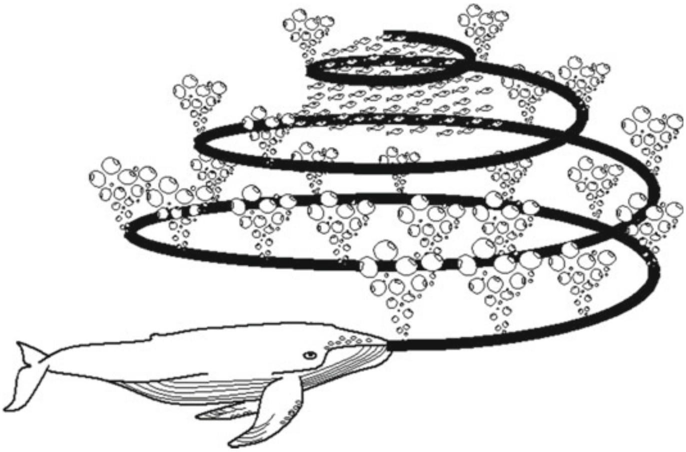


Fig. 5.6 Bubble net foraging behavior of humpback whales

The swarm intelligence optimization algorithm has better global search ability and convergence speed than traditional optimization methods. FOA, Dolphin Echolocation algorithm (DE), etc. are widely used in actual engineering. Whales are large mammals in the ocean. Adult humpback whales can reach up to 18 m in length and weigh up to 30 tons. Humpback whales are social animals and feed mainly on small marine organisms such as krill and scaly fish. Humpback whales have a special predatory behavior, the bubble net foraging method, as shown in Fig. 5.6:

Bubble net predation behavior can be divided into two stages, the first is the upward spiral stage; the second is the double cycle stage. In the upward spiral phase, the whale first dives and swims under 12 m to find food, and then determines the predator target, and then begins to form bubbles around the prey in a spiral trajectory similar to the number "9" and go upstream; the double cycle phase includes the coral cycle, and tapping the water surface with tail leaves and capture loops. Based on the bubble-net foraging behavior of humpback whales, Mirjalili and Lewis derived WOA.

WOA simulates the humpback whale's hunting bubble net behavior, and after multiple iterations, the search position is updated and the optimal solution is finally obtained. The algorithm mainly includes three stages: encircling prey, attacking with bubble net, and seeking food.

1. **Prey Phase**

In this stage, the artificially bred whale population has no prior knowledge at all, and does not know the location of the prey. The algorithm uses random individual positions in the population to locate and navigate and find the optimal solution.

The artificially bred whale closest to the food is equivalent to a current local optimal solution, and other artificially bred whales will approach this position and gradually surround the prey. The mathematical model of this process is as follows:

$$D = \left|CX^*(t) - X(t)\right| \tag{5.4}$$

$$X(t+1) = X^*(t) - A \cdot D \tag{5.5}$$

Among them, X* is the local optimal solution, X is the position vector, t is the current iteration number, and A and C are coefficients. A is a random number on the interval [−2, 2], and C is a random number on the interval [0, 2]. The mathematical expressions of A and C are as follows:

$$A = 2a \cdot r - a \tag{5.6}$$

$$C = 2 \cdot r \tag{5.7}$$

Among them, *a* linearly decreases from 2 to 0, and its mathematical expression a = 2-2t/tmax, and tmax is the maximum number of iterations. *r* is a random number on the interval [0, 1].

2. **Bubble Net Attack Stage**

In this stage, Mirjalili and Lewis imitated the predation behavior of the humpback whale bubble net, and achieved the purpose of finding the local optimal solution through the encircling contraction and the spiral update position.

(1) Enveloping and Shrinking

The whale group updates its position according to Formula (5.5). When −1 < A < 1, the artificially bred whale will approach the artificially bred whale with the best current position, shrinking the predation range.

(2) Spiral Update Position

Artificially bred whales continuously approach their prey in a spiral motion, and the mathematical model of attacking the prey is as follows:

$$X(t+1) = D' \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t) \tag{5.8}$$

Among them, $D' = |X^*(t) - X(t)|$ represents the distance vector between the artificially bred whale and the best solution obtained so far, b is a constant coefficient used to define the shape of the logarithmic spiral, and a random number in $l \in [-1, 1]$. When $l = -1$, the distance between artificially bred whale and the food is the closest; when l = 1, the distance is the farthest. At the same time, it should be noted that when the artificially bred whale approaches its prey in a logarithmic spiral shape, it shrinks and surrounds the prey. In order to realize the simultaneous progress of the enveloping contraction and the spiral update position, Mirjalili assumes that a 50% probability is selected between the contraction encircling mechanism or the spiral update position model to update the position of the whale. The specific mathematical model is as follows:

$$X(t+1) = \begin{cases} X^*(t) - A \cdot D, & \text{if } p < 0.5 \\ D' \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t), & \text{if } p \geq 0.5 \end{cases} \tag{5.9}$$

Among them, p is a random number belonging to [0, 1].

3. **Seeking Food Stage**

The whale optimization algorithm controls the artificially bred whale to swim to obtain food by controlling the parameter A, so as to obtain local optimization and global search capabilities. When |A|>1, the artificially bred whale will be forced to update its position to the randomly selected reference whale Xrand. This method enables the algorithm to perform a global search to obtain the global optimal solution. The mathematical model is as follows:

$$D = |C \cdot X_{\text{rand}} - X(t)| \tag{5.10}$$

$$X(t+1) = X_{\text{rand}} - A \cdot D \tag{5.11}$$

Among them, Xrand is the position vector of the reference whale obtained randomly.

5.5.2 Algorithm Evaluation and Improvement

WOA has the characteristics of few parameters, simple algorithm principle, and strong operability. It has been proved that its convergence accuracy and convergence speed are superior to particle swarm optimization, gravity search algorithm, differential evolution algorithm, etc. However, the standard WOA is prone to premature maturity due to its random search mechanism, and there are problems such as slow convergence and easy to fall into local optimal solutions when solving large-scale optimization problems. In view of the shortcomings of basic whale optimization algorithms, such as early maturity and slow convergence. This section mainly introduces the following improved adaptive whale optimization algorithms WGPWOA. The specific improvement measures are as follows:

1. **Adaptive Probability**

WOA searches for the optimal solution while using the shrinking encirclement, and also uses the logarithmic spiral method to update the position. In order to realize the synchronization of these two mechanisms, Mirjalili chooses $P_0 = 50\%$ probability to update the position of the artificially bred whale and its mathematical model is as shown in the formula. That is, when $p < 0.5$, the artificially bred whale shrinks and surrounds, and the position of the artificially bred whale is controlled by the control parameter A. When $|A| < 1$, the artificially bred whale will approach the best whale position at the current position, shrinking the predation range; when |When A| > 1, the individual whale is forced to update its position to the randomly selected reference whale, so as to perform a global search to obtain the global optimal solution. When $P \geq 0.5$, the artificially bred whale performs logarithmic spiral update. The size of the probability P_0 affects the local convergence speed and global search ability of the algorithm to a certain extent. When the value of P_0 is small, it is beneficial to accelerate the local convergence, when the value of P_0 is large, it is beneficial to jump out of the local optimal solution and perform a global search. In this paper, through a large number of simulation experiments, the probability P_0 is set to a linearly increasing value from 0.3 to 0.7, so that the algorithm has a strong local search ability in the early stage, and can enhance the ability to jump out of the local optimal solution in the later stage of the algorithm, and obtain the global optimal solution. So that the algorithm gains adaptive ability. The specific expression of P_0 is:

$$P_0 = 0.3 + 0.4 \cdot t/t_{\max}$$

Among them, t is the current iteration number, and t_{max} is the maximum iteration number.

2. **Adaptive Inertia Weight**

In 1998, Y. H. Shi proposed an improved particle swarm algorithm with inertial weights. The inertial weight w was dynamically adjusted during the search process. As the search progresses, w can be linearly reduced to ensure that the algorithm can start with larger velocity steps can detect better areas in the global scope; while the smaller w value at the later stage of the search ensures that the particles can do a fine search around the extreme points, so that the algorithm has a greater probability of going to the global The optimal solution position converges. Since this algorithm can guarantee a better convergence effect, it is defaulted to the standard particle swarm algorithm. In this paper, inspired by Y. H. Shi, inertia weight is introduced in the whale optimization algorithm to improve the algorithm. Thereby improving the convergence accuracy of WOA and speeding up the convergence speed. The expression for introducing the inertia weight w is as follows:

$$W = w_{\max} - \frac{(w_{\max} - w_{\min}) \cdot t}{t_{\max}}$$

Among them, t_{max} represents the maximum number of iterations, w_{max} and w_{min} represent the maximum inertia weight and minimum inertia weight respectively, and t represents the current iteration number. w gradually decreases as the number of iterations increases, thereby further balancing the development and exploration capabilities of WOA. In this paper, the maximum inertia weight w_{max} = 0.1 and the minimum inertia weight w_{min} = 0.01.

$$X(t+1) = \begin{cases} wX^*(t) - A \cdot D, & \text{if } p < p_0 \\ D' \cdot e^{bl} \cdot \cos(2\pi l) + wX^*(t), & \text{if } p \geq p_0 \end{cases} \tag{5.12}$$

3. **Gaussian Mutation Operator**

As the number of iterations of the basic whale optimization algorithm increases, more and more artificially bred whales will gather at the local optimal solution, leading to premature algorithm maturity and reduced global search capabilities. Gaussian distribution is an important type of probability distribution, which has been widely used in many fields. In this paper, Gaussian mutation is introduced into the basic whale optimization algorithm, which produces a disturbance term that obeys the Gaussian distribution to the artificially bred whale position, expands its search range, and avoids the algorithm from falling into a local optimal solution. The probability density function of Gaussian distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Among them: σ is the variance of the Gaussian distribution, and μ is the expectation. The Gaussian variation perturbation formula for the artificially bred whale position is as follows:

$$X'(t) = X'(t) \cdot \left(1 + k \cdot \text{Gauss}\left(\mu, \sigma^2\right)\right)$$

Among them, k is a random number on (0, 1), and X'(t) is the new position generated by the current optimal artificially bred whale after enveloping and contracting, and spirally updating the position. Here we choose μ = 0 and σ2 = 1.

4. **Algorithm Description**

The improved WGPWOA algorithm process proposed in this section is as follows:

(1) Initialization: Initialize the artificially bred whale group Xi (i = 1, 2, …, NP); the initialization parameters are the artificially bred whale population size NP, the maximum number of iterations t_{max}, the maximum inertial weight w_{max}, and the minimum inertial weight w_{min}.

(2) Calculate the fitness of each individual artificially bred whale, and make the position of the individual with the best fitness value the most optimal position.
(3) Update the values of parameters A, C, l and P.
(4) If $P < P_0$, if $|A|>1$, randomly select an individual (Xrand) in the artificially bred whale population, and update the position according to Eq. (5.11); if $|A| < 1$, follow Eq. (5.12) to update the location.
(5) If $P > P_0$, update the position according to formula (5.12).
(6) Perturb the current individual position with Gaussian mutation operator.
(7) Judge whether the termination condition is reached, if so, output the optimal individual position, which is the optimal solution; otherwise, return to step (2).

The pseudo code of WGPWOA proposed is shown in the following table (Table 5.4):

5.6 GWO (Grey Wolf Optimization)

5.6.1 Algorithm Principle and Model

Grey Wolf Optimization (GWO) algorithm is inspired by the creature "grey wolf", which belongs to a type of swarm intelligence algorithms. Like all other heuristic algorithms constructed based on the behavior and living habits of a certain type of biological group in nature, the grey wolf optimization algorithm also operates by simulating the command hierarchy and hunting mechanism of the grey wolf population.

Grey wolf (scientific name: *Canis lupus*; nickname: wolf, plain wolf, timber wolf, etc.) is the largest carnivorous canine in existence, and is currently widely considered to be at the top of the food chain. Grey wolves like to live in groups, with a population of about five to twelve. The grey wolf is famous for its strict command hierarchy system, as shown in the figure (Fig. 5.7).

The grey wolf population has four different command structure levels, which can be arranged in order from high to low: α wolves, β wolves, δ wolves, and ω wolves. αwolf is at the top of the command hierarchy and is the leader wolf (the wolf king) in the grey wolf group. As a leading wolf, αwolf not only needs to complete its social behavior as an individual wolf in daily life, but also needs to control the overall situation and make corresponding decisions to control the time and place of predation and rest in its group. β wolves are located in the second echelon of the command hierarchy. They are deputy wolves that assist α wolves in making decisions or organizing other group activities. As decision-makers and disciplinary committees in the wolves, β wolves must abide by α wolves' command on the one hand, and on the other aspect assists them in issuing orders to other lower-level wolves, and provides timely feedback to them while ensuring that the orders of α wolves are issued smoothly in the wolves. δ wolves, also known as subordinate wolves in the pack, are located at the third level of the command hierarchy. They are allowed to issue orders to the lowest level ω wolves in the

Table 5.4 Pseudo code of WGPWOA

```
Algorithm WGPWOA
    Initialize the artificially bred whale swarm{Xi, i=1, 2, … , NP};
    Initialization parameters, population size NP, maximum number of iterations
tmax, maximum inertial weight wmax, and minimum inertial weight wmin;
  While (t<tmax) do
    Calculate the fitness of each individual artificially bred whale{f(Xi), i=1,
2, … , NP};
    The position of the individual with the best fitness value is the most optimal
position;
    for i =1 to NP do
      Update the values of parameters A, C, l and P;
      if (P<P0) do
        if |A|≥1 do
          Randomly select an individual (Xrand) in the artificially bred
whale population, and update the position according to Formula (5.11)
        else if |A|<1 do
          Update position according to Formula (5.12)
        end
      else if (P≥P0) do
        Update position according to Formula (5.12)
      end
    end
  Perform Gaussian mutation operator perturbation on the current individual whale
position;
  t=t+1
  end
```

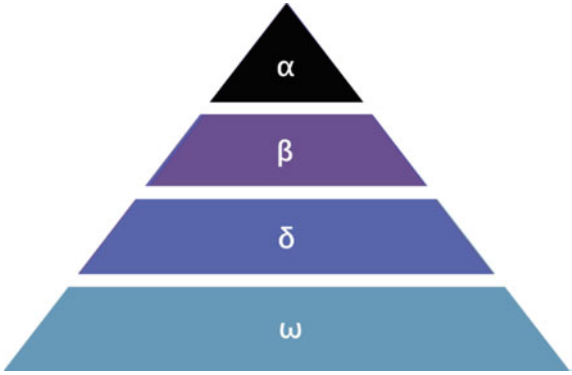


Fig. 5.7 Schematic diagram of grey wolf hierarchy

pack while obeying the orders of α wolves and β wolves. ω Wolf is at the bottom of the command hierarchy and plays the role of scapegoat.

In addition to the social hierarchy of wolves, another social behavior-orderly group hunting activities also provides modeling ideas for GWOA. Hunting activities can be divided into the following main stages-first, tracking, chasing and sneaking close to the prey; then rounding up and continuing to interfere with the prey until it stops moving; finally, attacking the prey. The above steps are shown in the figure (Fig. 5.8):

GWO is mathematically modeled and calculated based on the command hierarchy and hunting behavior of the grey wolf group. Then this section will be divided into four small parts to introduce the algorithm's mathematical modeling process

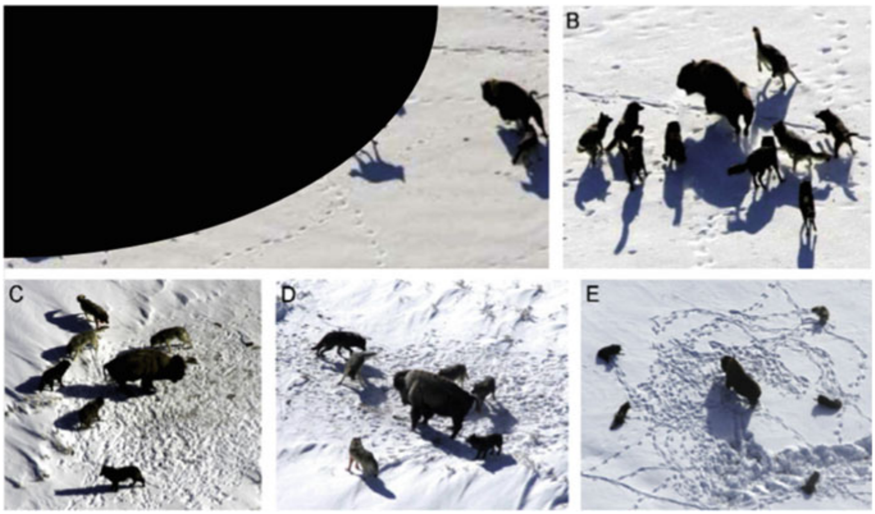


Fig. 5.8 Hunting process of grey wolf

for the four stages of grey wolf hunting behavior: encircling, hunting, attacking the current prey and searching for other prey.

1. **Surround Prey**

First, in order to establish a mathematical model for the grey wolf population to round up prey, we simulate the level distribution of the grey wolf population in nature, and arrange the solutions into α, β, δ, and ω according to their advantages and disadvantages. Therefore, the model when surrounding prey can be expressed as:

$$D = |C \cdot XP(t) - X(t)| \tag{5.13}$$

$$X(t+1) = X\,P(t) - A. \tag{5.14}$$

Equations (5.13) and (5.14) respectively define the distance between the grey wolf and its prey and the way the grey wolf changes its position in the iterative process. The symbol t represents the current iteration of the algorithm, XP represents the position vector of the prey, X represents the position vector of the grey wolf, and A and C are the parameter vectors that affect the direction of the grey wolf in the formula.

The calculation formula of parameter vector A and C is:

$$\mathrm{A} = 2\mathrm{a} \cdot \mathrm{r1} - \mathrm{a} \tag{5.15}$$

$$\mathrm{C} = 2 \cdot \mathrm{r2} \tag{5.16}$$

In the iterative process, the value of the control parameter a decreases linearly from the value 2 to the value 0; r1 and r2 are random vectors in the interval [0, 1].

The process of encircling prey represented by the above formula can be clearly represented by the following figure. As shown in the figure, in a two-dimensional space, a grey wolf with coordinates (X, Y) will change its position accordingly due to the movement of the prey at (X*, Y*). In other words, by adjusting the values of the vectors A and C, the grey wolf can reach any position in the specified space, that is, we can obtain the optimal solution through continuous iteration according to the current position of each solution (Fig. 5.9).

Figure 5.10 shows the location of grey wolves in three-dimensional space. Similarly, by adjusting the values of the random vectors r_1 and r_2, the grey wolf pack can reach any position shown in the figure.

The same concept is also applicable to *n*-dimensional space. In the hypercube, the wolf pack still moves around the optimal solution obtained so far until the optimal solution so far is obtained.

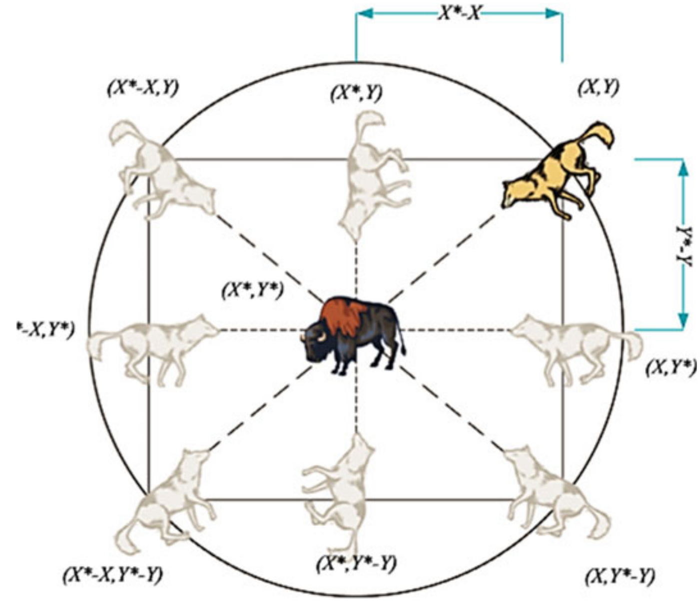


Fig. 5.9 The position vector of grey wolf in two-dimensional space

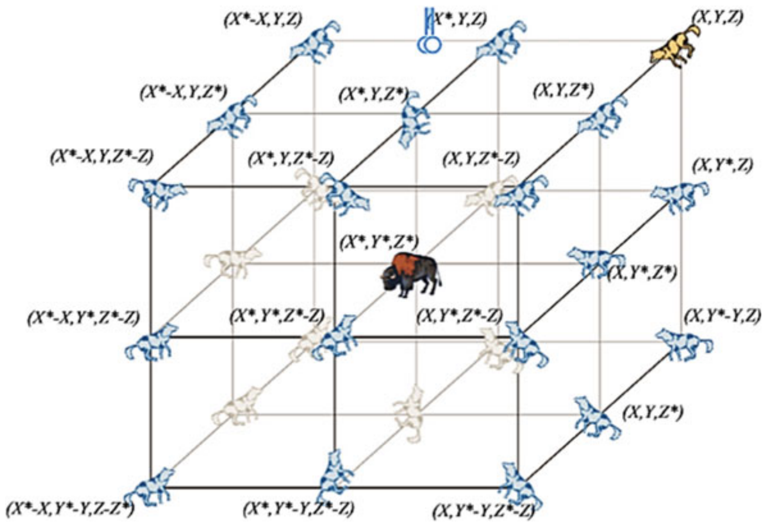


Fig. 5.10 The position distribution of grey wolves in three-dimensional space

2. **Hunt Down Prey**

During the hunting process, the grey wolves need to locate and round up their prey to prepare for the next attack. The hunting behavior is usually dominated by α wolves, and β wolves and δ wolves will also participate from time to time. Since the problems we need to solve in daily life are usually larger than two-dimensional or three-dimensional, and the search space is more complicated, so in order to build a mathematical model and predict the location of wolves and prey better and more accurately, we need to be in the algorithm space record the first three optimal solutions α, β, and δ collected so far, and use their positions as the calculation basis to update the positions of all other search agents. The update process can be expressed by the following formulas:

$$D_\alpha = |C_1 \cdot X_a(\mathrm{t}) - X(\mathrm{t})|$$

$$D_\beta = \left|C_2 \cdot X_\beta(\mathrm{t}) - X(\mathrm{t})\right|$$

$$D_\delta = |C_3 \cdot X_\delta(\mathrm{t}) - X(\mathrm{t})|$$

$$X_1(\mathrm{t}+1) = X_a(\mathrm{t}) - A_1 \cdot D_a$$

$$X_2(\mathrm{t}+1) = X_\beta(\mathrm{t}) - A_2 \cdot D_\beta$$

$$X_3(\mathrm{t}+1) = X_\delta(\mathrm{t}) - A_3 \cdot D_\delta$$

$$X_P(\mathrm{t}+1) = (X_1 + X_2 + X_3)/3$$

Figure 5.11 shows this update process in a more intuitive way-we can observe that α wolves, β wolves and δ wolves update their positions according to the above formula, and their new positions must be one around their current positions. Inside the circle. Simply put, α wolves, β wolves, and δ wolves estimate the possible range of their prey based on their positions, while other wolves in the pack randomly update their positions according to the position of their prey.

3. **Attack Current Prey**

In the third stage of the hunting process, the grey wolves wait for an opportunity to attack the prey that has stopped moving. In order to model the process of approaching the prey more accurately, we began to reduce the value of the control parameter *a*. When the control parameter *a* decreases linearly from 2 to 0 in the iterative process, *A* takes a random value within the interval of [−2a, 2a]. When the value of *A* is within [−1, 1], the search agent will choose between its current

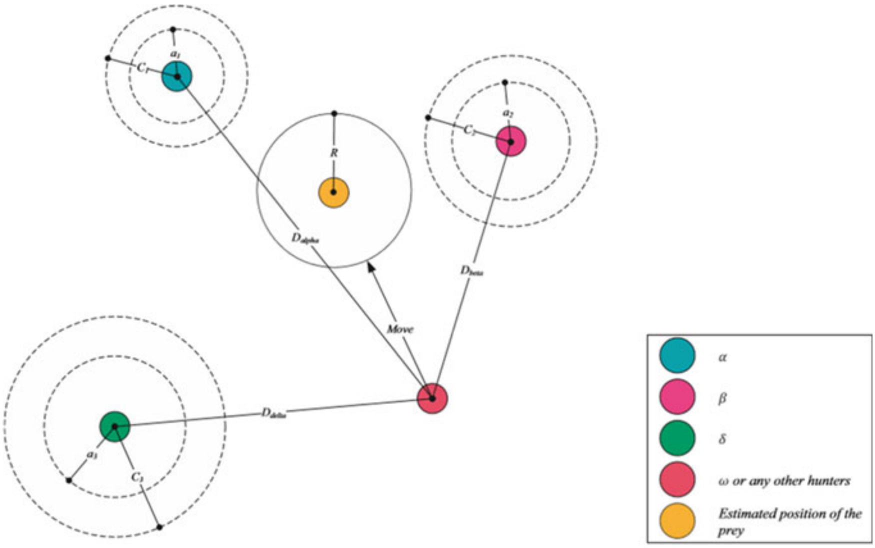


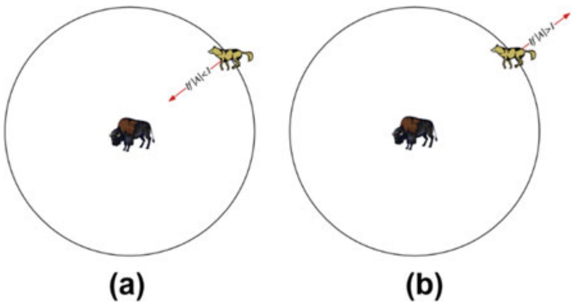
Fig. 5.11 Grey Wolf location update process

location and the location of the prey, and randomly select a location for location update operations. At this time, as shown in the figure (a) below, the wolves approach and attack the prey (Fig. 5.12)

4. **Search for Other Prey**

Different from the previous part, when |A| > 1, it means that when the grey wolves are hunting, the rest of the wolves except α wolves, β wolves and δ wolves will be scattered around the area. Search for locations where prey may appear, and explore whether there are other "better" prey (as shown in the figure (b) above). This helps the grey wolf optimization algorithm to conduct a more extensive and in-depth exploration of feasible solutions in the entire search space, and improve the possibility of finding the global optimal solution as soon as possible. In addition to

Fig. 5.12 Attacking prey and searching for (other) prey



A, there is another element in the grey wolf optimization algorithm that can affect its global search capability-vector C. This vector is randomly selected in [0, 2], and its function is to randomly assign a value (ie weight) to the influence of the prey on the determined distance value. When $C > 1$, the influence becomes larger; and when $C < 1$, the influence become smaller. The existence of C makes GWO show stronger randomness in the search process, and enhances the probability of the wolf pack to explore other prey.

5.6.2 Algorithm Flow

Step 1: Parameter initialization, including grey wolf population size N, search dimension D, maximum number of iterations tmax, upper and lower boundaries ub and lb of the vector space, parameters a, A, C, etc.

Step 2: In the search space, randomly initialize the upper and lower boundaries of the variables to generate the position of each individual in the grey wolf population.

Step 3: Calculate the fitness value of each individual wolf in the grey wolf pack.

Step 4: Sort according to the fitness value from high to low, find the top three individual wolves, which are defined as α, β and δ wolves, and save their positions as Xα, Xβ and Xδ in turn.

Step 5: Calculate the distance between other grey wolf individuals in the population and Xα, Xβ and Xδ according to the above formula, and update the position of each grey wolf individual.

Step 6: Update the parameters a, A, C and the position of the grey wolf pack.

Step 7: The number of iterations $t = t + 1$.

Step 8: Determine whether the algorithm has reached the maximum number of iterations tmax. If yes, end the calculation and output the optimal solution; otherwise, go to Step 3.

Chapter Summary

The optimization algorithm is suitable for solving microscopic optimization problems with relatively sufficient information. The effectiveness and value of the algorithm must be fully reflected in practical applications. The main points of this chapter are particle swarm algorithm and genetic algorithm. The Particle Swarm Optimization algorithm is a heuristic global random search algorithm based on swarm intelligence, which is easy to understand, easy to implement, and has strong global search capabilities. In this chapter, the basic principles, models, and processes of PSO are elaborated, and the advantages and disadvantages of particle swarm optimization, improved forms and application fields are reviewed. Genetic algorithm is also a widely used and mature optimization algorithm. It is a method of searching for the optimal solution by simulating the natural evolution process. This chapter focuses on the basic operation of genetic algorithm.

Important Concepts and Terms

optimization
algorithm
optimization algorithm
Particle Swarm Optimization (PSO)
intelligent optimization algorithm
population
fitness
selection
reproduction
crossover
mutation.

Questions and Exercises

1. Please describe the principle of particle swarm (PSO) algorithm.
2. Introduction: What are the improved methods of particle swarm optimization algorithm?
3. Calculation: Apply PSO algorithm to solve the minimum value of function $f = x(1)^2 + x(2)$.
4. Calculation: Use genetic algorithm to calculate the maximum value of the following function:

$$f(x) = x\sin(10\pi \cdot x) + 2.0, x \in [-1, 2]$$

5. Choose binary coding to create an initial population of 40 individuals and 20 length. The generation gap is 0.9, the random ergodic sampling method is used for selection, single-point crossover and discrete mutation are used for recombination, and the elite strategy is used to insert outstanding individuals from the parent into the offspring. The maximum genetic generation number is 50.

Case Study 1

Optimal allocation of resources

A resource with a value of W is allocated to the project set P, P contains n projects, and $P_i(i = 1, 2, \ldots, n)$ is the i project. To allocate resources to n projects, the average rate of return of allocating resources W_i to project P_i is b_i, and the risk loss rate of project P_i is r_i. The allocation of resource W_i to project P_i requires a cost, and the expense rate is α_i, and when the value of the allocated resource does not exceed the given value μ_i, the cost is calculated according to the resource allocation μ_i. The proportion of the resource W_i allocated to the project

P_i to the total resource value W is x_i, and the resource allocation cost is:

$$c_i(x_i) = \begin{cases} 0,\ x_i = 0 \\ \alpha_i \mu_i,\ 0 < x_i < (1+\alpha_i)\frac{\mu_i}{W} \\ \frac{\mu_i W}{1+\alpha_i} x_i\ \ (1+\alpha_i)\frac{\mu_i}{W} < x_i < 1 \end{cases} \tag{5.17}$$

The goal of resource allocation is to make the net income as large as possible and the overall risk as small as possible, so it is a dual objective function. You can use multiplication and division to simplify the establishment of the following multi-objective decision-making model, set the income as $I(X)$ and the risk as $R(X)$.

$$\begin{cases} \max I(X) \\ \min R(X) \\ I(X) > 0, R(X) > 0, X \in D \end{cases} \tag{5.18}$$

Maximize the total benefit of the project after resource allocation:

$$\max Q(X) = \frac{I(X)}{R(X)}$$

$$\min F(X) = \min \frac{1}{Q(X)}$$

The benefits of the project P_i are: $I_i(x_i) = b_i M x_i - b_i c_i$; The risks of the project P_i are: $R_i(x_i) = r_i M x_i - r_i c_i$; The required resource value of the project P_i is: $\phi_i = M x_i$; Mission achievement rate is λ_i; The resource allocation combination is: $x = (x_1, x_2, \cdots, x_n)$; Total benefits of project P collection: $I(x) = \sum_{i=1}^{n} \lambda_i I_i(x_i)$; Overall risk is: $R(x) = \sum_{i=1}^{n} R_i(x_i)$; Resource constraints: $\phi(x) = \sum_{i=1}^{n} \phi_i(x_i)$; A collection of related items for resource allocation: $P_{i_1} \cap P_{i_2} = \varphi,\ i_1 \neq i_2$; There are no missing resources in the allocation process: $\bigcup_{i=1}^{n} M_i = M,\ M_{i_1} \cap M_{i_2} = \varphi,\ i_1 \neq i_2.$

On the basis of the above model, suppose the total value of a certain resource is 70 million yuan, which is allocated to 6 projects, the number of particles is 60, the maximum number of iterations is 500, the learning factor $c_1 = 2.5$, $c_2 = 1.5$, and the project income weight is 0.7, the risk weight is 0.3, and the optimization data of resource allocation structure is shown in the following table.

Please use standard PSO and original PSO to optimize the resource allocation model, and compare and analyze the optimization effect (Table 5.5)

Table 5.5 Data value

$\phi_i(x_i)$	b_i	r_i	α_i	μ_i	λ_i
$\phi_1(x_i)$	25	5.0	5	120	100
$\phi_2(x_i)$	36	3.5	4	150	98
$\phi_3(x_i)$	31	2.0	4.5	180	100
$\phi_4(x_i)$	19	1.5	5.5	100	95
$\phi_5(x_i)$	44	6.0	6	180	96
$\phi_6(x_i)$	51	4.5	4	200	95

Case Study 2

Optimization of the Distribution Route of the Commercial Vehicle
In recent years, the production layout of auto companies has become more dispersed, and logistics outlets have increased significantly, which has increased the difficulty of logistics services accordingly. In addition, the coordinated transportation and resource integration of multi-brand vehicles have also brought new challenges to vehicle logistics. Therefore, if the choice of transportation mode and route relies solely on the experience of planners and familiarity with network operations, it is difficult to adapt to changes in external resources and demands. It is necessary to connect the service outlets scattered in different areas to build a fast-response logistics network to achieve the lowest possible logistics cost under the premise of ensuring a certain service level.

1. **Model Building**

The choice of transportation mode and route can be described as: define an undirected graph G = (V, E), V represents all nodes in the network (warehouse, transfer station or city, etc.); E represents edge set, including transportation arc and the reloading arc between modes; the transportation arc refers to the connection between different nodes of the same transportation mode, and the arc weight indicates the transportation cost of goods between the two nodes; the reloading arc refers to the connecting line between the same node and different transportation modes, and the arc weight represents the cost of reloading the goods at the node. If a certain mode of transportation is not provided between two nodes or no replacement occurs, the weight of the arc (edge) is infinite.

(1) Basic assumptions

① There are at most two alternative modes of transportation (highway, waterway) for the transportation of goods between two cities.

② There is a good connection in the transfer process, and the goods are instantly reloaded at the node, and there is no inventory.

③ The transportation volume is inseparable, that is, only one transportation mode can be selected between a certain two specific cities.

④ The restrictions on transportation price and transportation time by transportation volume are not considered.

⑤ The reshipment of goods can only occur at nodes, and at most each node can be reshipped once.

(2) Symbol definition

N indicates the number of all node cities;

M_i represents the set of transportation modes available in the ith city $\begin{cases} 1.\ \text{highway} \\ 2.\ \text{waterway} \end{cases}$

Q indicates the total amount of transportation in multimodal transportation;

$C^k_{i,j}$ represents the cost of cargo transportation from node i to j with mode k, $i, j \in \{1, 2, \ldots, N\}, k \in M_i$;

d^{kl}_i represents the transfer cost when the transportation mode is changed from k to l at node i,$i, j \in \{1, 2, \ldots, N\}, k, l \in M_i$;

t^{kl}_i represents the transit time when the transportation mode is changed from k to l at node i,$i \in \{1, 2, \ldots, N\}, k, l \in M_i$.

$v^k_{i,j}$ represents the transportation speed of the kth transportation vehicle selected from node i to node j,$i, j \in \{1, 2, \cdots, N\}, k \in M_i$;

$s^k_{i,j}$ represents the transportation distance from node i to node j to select the k-th transportation means,$i, j \in \{1, 2, \ldots, N\}, k \in M_i$;

$X^k_i = \begin{cases} 1, city : i; vehicle : k \\ 0, others \end{cases} \quad i \in \{1, 2, \ldots, N\}, k \in M;$

$r^{kl}_i = \begin{cases} 1, city : i; vehicle : k \to l \\ 0, others \end{cases} \quad i \in \{1, 2, \ldots, N\}, k, l \in M;$

P_t indicates the penalty cost for exceeding the contract period.

(3) Build the Model

The model goal of multimodal transportation mode selection is to minimize the total transportation cost. The objective function is composed of transportation cost, transit cost and penalty cost.

$$\min Z = \sum\nolimits_i \sum\nolimits_k X^k_{i,i+1} C^k_{i,i+1} S^k_{i,i+1} Q^k_{i,i+1} + \sum\nolimits_i \sum\nolimits_k \sum\nolimits_l r^{kl}_i d^{kl}_i + P_t$$

Restrictions:

$$\sum_k X^k_{i,i+1} = 1, \forall i \in \{1, 2, \ldots, N-1\} \tag{5.19}$$

$$\sum_k \sum_l r^{kl}_i = 1, \forall i \in \{1, 2, \ldots, N\} \tag{5.20}$$

$$X^k_{i-1,i} + X^l_{i,i+1} \geq 2r^{kl}_i, \forall i \in \{2, 3, \ldots, N-1\}, \forall k, l \in M \tag{5.21}$$

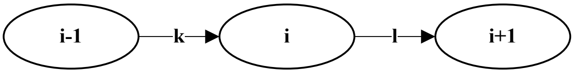


Fig. 5.13 Transportation mode conversion

$$t_{\min} \leq \sum_i \sum_k \frac{S_{i,i+1}^k X_{i,i+1}^k}{V_{i,i+1}^k X_{i,i+1}^k} + \sum_i \sum_k r_i^{kl} t_i^{kl} \leq t_{\max} \tag{5.22}$$

$$r_i^{kl} \in \{0, 1\}, \forall i \in \{2, 3, \ldots, N-1\}, \forall k, l \in M \tag{5.23}$$

$$X_{i,i+1}^k \in \{0, 1\}, \forall i \in \{1, 2, \ldots, N-1\}, \forall k \in M \tag{5.24}$$

$$P_t = \begin{cases} 0, t < T \\ (t-T)p, t \geq T \end{cases} \tag{5.25}$$

Formula (1) means that only one means of transportation can be selected for the goods from node i to node $i + 1$; formula (2) means that goods will only be transferred once in city i; formula (3) means that to ensure the continuity of transportation, in the city The cargo i is converted from the k-th transportation mode to the l-th transportation mode, that is, select k from i-1 to i, and select l from i to $i + 1$, as shown in Fig. 5.13.

The discussion is divided into three situations:

① If the decision variables are $X_{i-1,i}^k$ and $X_{i,i+1}^l$ are both equal to 1, then the decision variable r_i^{kl} of the corresponding transit transportation mode is also 1, which is $X_{i-1,i}^k + X_{i,i+1}^l = 2r_i^{kl}$ in this case;

② If one of the decision variables are $X_{i-1,i}^k$, and $X_{i,i+1}^l$ has a value of 0 and the other is 1, then r_i^{kl} must be 0, at this time, $X_{i-1,i}^k + X_{i,i+1}^l > 2r_i^{kl}$;

③ If the decision variables are $X_{i-1,i}^k$ and $X_{i,i+1}^l$ are both equal to 0, r_i^{kl} is also 0 at this time, and $X_{i-1,i}^k + X_{i,i+1}^l = 2r_i^{kl}$ at this time.

In this way, the continuity of the transportation process is ensured, and the relationship between $X_{i,j}^k$ and r_i^{kl} is established.

Formula (4) indicates that the time spent in the transportation of goods must be within the time range $[t_{\min}, t_{\max}]$ required by the carrier. The first part is the transportation time of the goods, and the second part is the transit time for the conversion of the transportation mode of the goods; formula (5) represents the decision variable r_i^{kl}; $X_{i,i+1}^k$ in formula (6) is a variable from 0 to 1; T in formula (7) is the maximum time period allowed from the starting point to the end point, and p is the penalty value per unit time.

2. Algorithm Implementation

Since genetic algorithm has the characteristics of global search, there is an optimized solution generation in each generation of population, combined with the characteristics of the transportation mode route selection model, the genetic algorithm is used to solve the problem.

(1) Chromosome Coding Method

In the transportation mode route combination optimization problem, the transportation mode and the transit node in the transportation route are involved. Therefore, these two different factors should be used as the genes in the chromosome when encoding the chromosome. Among them, the odd-numbered digits in the sequence represent the information of the nodes passed by in a transportation plan, and the even-numbered digits represent the information of the transportation mode used when transporting between two nodes before and after the even-numbered digits in the transportation scheme.

The specific coding process is as follows. The first gene encoded by the chromosome is the source node S of the goods, starting from node S, and the third position gene is randomly selected from other nodes that are connected to the source node in some way. The second position gene represents the transportation method used when transporting between the first and third nodes, so this process is repeated until the destination node is reached. You can get an initial population.

(2) Definition of Fitness Function

According to the objective function, the objective function value Z_h of the feasible solution is obtained. If the chromosome corresponds to the infeasible solution, a large integer M is assigned to Z_h, and the fitness function is $f_h = \frac{1}{Z_h}$. The larger the f_h, the better the performance of the chromosome. The closer the solution is to the optimal solution.

(3) Design of Genetic Operator
① Selection Operator

A selection strategy combining survival of the fittest and proportional selection is adopted. Sort the k chromosomes of each generation population according to the f_h value, and reproduce one chromosome with a large f_h value directly into the next generation. The remaining $k-1$ chromosomes in the next-generation population are generated by proportional selection, that is, the probability of each individual entering the next generation is equal to the ratio of its fitness value to the sum of individual fitness values in the entire population, $(\frac{f_i}{\sum f_i})$. The higher the fitness value, the greater the probability of being selected and the greater the probability of entering the next generation.

② Crossover Operator

Use the single-point crossover method. First, the population is randomly paired, secondly, the intersection position is randomly generated according to a uniform distribution, and finally, some genes between the paired chromosomes are exchanged to generate a new population.

③ Mutation Operator

First, a gene is randomly selected from the chromosome as the mutation point, keeping the gene from the source node to the mutation point unchanged, and the genes after the mutation point are randomly selected from the nodes connected to the mutation point.

④ Termination Criteria

The condition for the termination of the iteration is to determine whether the algebra of the iteration reaches the specified value, if it reaches the specified value, stop the evolution, and find the feasible solution corresponding to the chromosome with the largest fitness.

3. Case Analysis

Taking Datang's vehicle logistics as an example, the analysis only considers the choice of road and water transportation. First construct a transportation network diagram, as shown in Fig. 5.14.

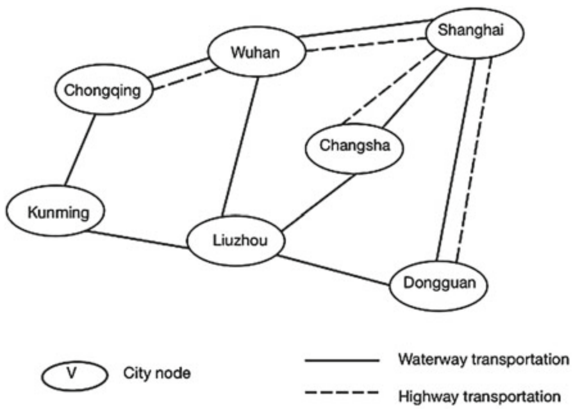


Fig. 5.14 Transportation network diagram

Table 5.6 Transportation distance and transportation unit price of different transportation modes among cities

	Highway		Waterway	
	Transport distance (km)	Transportation costs (Unit price: RMB/vehicle)	Transport distance (km)	Transportation costs (Unit price: RMB/vehicle)
Shanghai → Wuhan	811	73,822.09	514.04	35,990.28
Shanghai → Changsha	1173	106,773.5	899.54	62,985.79
Shanghai → Guangzhou	1619	147,371.09	874.93	61,262.59
Wuhan → Chongqing	1099	100,037.57	1187.1	83,120.74
Wuhan → Liuzhou	990	90,115.74	∞	∞
Changsha → Liuzhou	724	65,902.82	∞	∞
Guangzhou → Liuzhou	914	83,197.76	627.7	63,998.28
Chongqing → Kunming	1237	112,599.16	∞	∞
Liuzhou → Kunming	1083	98,581.16	∞	∞

Table 5.7 Unit transfer costs and transfer time between different modes of transportation

	Highway	Waterway
Highway	0/0	200/13.5
Waterway	200/13.5	0/0

Note The denominator is the unit transfer cost (yuan/vehicle), and the numerator is the transfer time (vehicle/day)

Assume that the carbon emission tax is 10 yuan/ton, diesel is 7.78 yuan/liter, and a commercial vehicle is 1.5 tons. Table 5.2 shows the distance and unit price of highway and waterway transportation between cities, and Table 5.6 shows the unit transfer cost and transfer time between different modes of transportation (Table 5.7).

Output the result through the Matlab program, the optimal path is:

Shanghai $\xrightarrow{Waterway}$ Wuhan $\xrightarrow{\text{Highway}}$ Chongqing $\xrightarrow{Highway}$ Kunming.

The optimal cost is 112,720.66 yuan, and the optimal time is 7 days. Through route optimization, the proportion of water transportation has been increased, the cost of the enterprise has been reduced, the requirements of customers have been met more safely and efficiently, and the social responsibility of creating a green and low-carbon logistics has been implemented.

Further Readings

1. Ji Zhen, Liao Huilian, Wu Qinghua. *Particle Swarm Algorithm and Its Application* [M]. Beijing: Science Press, 2009.
2. Li Li, Niu Ben. *Particle Swarm Optimization Algorithm* [M]. Beijing: Metallurgical Industry Press, 2009.
3. Pan Feng, Li Weixing, Gao Qi, et al. *Particle Swarm Optimization Algorithm and Multi-Objective Optimization* [M]. Beijing: Beijing Institute of Technology Press, 2013.
4. Guo Wenzhong, Chen Guolong. *Discrete Particle Swarm Optimization Algorithm and Its Application* [M]. Beijing: Tsinghua University Press, 2012.
5. Wen Zheng. *Proficient in MATLAB Intelligent Algorithm* [M]. Beijing: Tsinghua University Press, 2015.
6. Lei Yingjie, Zhang Shanwen, Li Xuwu, Zhou Chuangming. *MATLAB Genetic Algorithm Toolbox and Its Application* [M]. Xi'an: Xidian University Press, 2005.
7. Guo Yaohuang. A heuristic algorithm for freight truck scheduling [J]. *System Engineering*, 1989, 7(1): 47–53.
8. Lu Lin, Tan Qingmei. Research on a kind of random VRP hybrid particle swarm algorithm [J]. *Systems Engineering and Electronic Technology*, 2006, 28(2): 244–247.
9. Liang Xueling, Jin Wenzhou. Research on the model and algorithm of transportation mode selection [J]. *Transportation and Computer*, 2008, 3 (26): 38–39.

System Reliability

6

Hao Zhang

Learning Target

1. Understanding the basic concepts of reliability, system reliability, network reliability, failure rate, etc.
2. Understanding the calculation method of system reliability and failure rate.
3. Understanding the reliability calculation principle of series system, parallel system and hybrid system.
4. Getting familiar with Fault Tree Analysis, and will apply it to deal with practical problems.

Introduction

The Vanishing *Columbia*

In April 1969, NASA proposed a plan to build a reusable space vehicle. In January 1972, the United States formally planned the development of a space transportation system for space shuttles, defining the design of the shuttle, consisting of a solid rocket booster that can be recycled and reused, two external fuel tanks that are not recycled and three parts of a reusable orbiter. There are more than 2.6 million parts in the shuttle, which is a typical complex system.

The United States has developed five types of space shuttles: Space Shuttle *Columbia*, Space Shuttle *Challenger*, Space Shuttle *Discovery*, Space Shuttle *Atlantis* and Space Shuttle *Endeavour*. Space Shuttle *Columbia* was developed in 1981 as the first space shuttle to transport astronauts and equipment between space

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_6

and ground. The Columbia has a total length of about 56 m, a wingspan of about 24 m, a take-off weight of about 2,040 tons, a total take-off thrust of 2,800 tons and a maximum payload of 29.5 tons.

On January 16, 2003, *Columbia* carried out its 28th flight, and on February 1, the explosion broke down when *Columbia* returned, killing all seven astronauts on board and scattering debris in Texas and Louisiana. It was investigated that the foam insulation material of the outer fuel tank of the shuttle hit the left wing during the launch, resulting in a large area of damage to the thermal insulation protective layer on the body surface, and finally disintegrated completely due to the combustion caused by ultra-high temperature air friction during the return flight. This tragedy once again comes to a conclusion that as the artificial system becomes larger and larger, it becomes more and more difficult to ensure the reliable operation of the system. There are 2.6 million parts in the shuttle, and each defect may lead to fatal consequences.

There are many high-risk products, projects or programs in the socio-economic field, such as the storage and transportation of dangerous chemicals, the operation of high-speed rail, and the damage to people's lives and property in the event of accidents. Therefore, the reliability management of such high-risk complex systems is crucial. So what is the reliability of the system? How to analyze the reliability of the system? How to apply system reliability in the field of management science?

Source: http://news.sohu.com/20130204/n365455821.shtml (accessed on 09/09/2021).

6.1 Overview of Reliability

Reliability is an important engineering discipline, and the birth of the discipline cannot be separated from the needs of society, the development of science and technology. The relevant research on reliability can be traced to World War II, when the complexity of military technical equipment led to the high failure rate of equipment. Although the research of reliability engineering first began from the military field, with its continuous development, reliability engineering is widely used in many other fields and plays an important role in improving the economic benefits of enterprises. The research of reliability engineering in China started late, and there is still a big gap of research level between China and the West, although the speed of our development is fast.

6.1.1 Basic Concepts

1. **Reliability**

According to the Chinese national standard GB-6583, reliability refers to the ability of the product to complete the specified function within the specified time under the specified conditions. The contents of reliability research include not only equipment, components, but also systems. The definition of reliability points out that the three elements of reliability are “specified conditions”, “specified time”, “specified functions”. The important tools of reliability engineering are probability theory and mathematical statistics.

The “specified conditions” include the use of the product or system, environmental conditions and working conditions at the time of operation, including climatic conditions, physical conditions, etc. For example, the reliability of the same type of car on highways and on rugged mountain roads is not the same, so the specified conditions must be specified when studying reliability.

“The specified time” refers to the task time specified by the product or system. With the increase of product task time, the probability of product failure will increase and the reliability of product will decrease gradually. Therefore, the task time should be clearly defined in the study of reliability. For example, a car that has been used for two years is significantly more likely to fail than a car that has been used for 10 years.

“Specified function” refers to the functions and technical specifications that a product or system must possess. The number of functions and the technical index directly affect the reliability index. For example, the technical index of automobile includes power index, fuel economy index, control stability index and ride comfort index, etc. When the specified function is considered differently, the reliability of the vehicle is very different.

2. **System Reliability**

Because the types of systems are diverse and the relevant disciplines have their own characteristics, the definition of system reliability also has different connotations in different disciplines. But overall system reliability is defined as the ability of the system to complete the specified functions under specified conditions and within a specified time. Reliability research was originally widely used in the reliability evaluation of single electronic components. Until the 1960s, reliability research gradually extended to the reliability of general products and the reliability of more complex associated systems. However, complex systems may contain multiple subsystems or basic elements, and the functions of each subsystem and basic elements and their related technical indicators are different. The reliability of each subsystem or basic element is the most difficult part.

3. **Network Reliability**

The network is often used to describe the structure of the system. For example, the transportation network in the traffic system is composed of the station and the road. The station can be abstracted into nodes and the road can be abstracted into arcs. Assuming a system structure consisting of nodes and arcs becomes a network S. Let x be any arc in the network S, and the “system normal” event is expressed as S, "arc x failure event" as $\overline{x}$. The probability $P_x = P(x)$ of a segment of arc x, T normal operation at a fixed time is known in a given network S. The reliability of the network can be defined as the probability that the system S at any T normal time, that is $R_T = P(S)$. Of course, the real traffic network, supply chain network, logistics network all belong to the entity network. And social network, information network these belong to virtual network. There are also systems that contain both physical and virtual networks, and the reliability of such systems will be more complex.

6.1.2 Reliability Research Principle and Contents

Reliability, from another point of view, is the failure rate of the structure. Failure rate refers to the probability of failure in a unit time after a product that has not failed at a certain time. The higher the failure rate is, the lower the reliability is; conversely, the higher the reliability is. The basic methods for solving loss efficiency include direct integral method, numerical simulation method and non-probabilistic uncertainty.

Direct integral method: a method of solving integrals by algebraic or triangular identity transformations based on basic integral tables and basic algorithms. This method is suitable for solving simple system.

Numerical simulation method: by means of computer, through numerical calculation and image display, the purpose of studying engineering and physical problems and even all kinds of problems in nature is achieved. The numerical simulation method is simple in principle and reliable in analysis.

Non-probabilistic uncertainty: this is a new approach for failure rate analysis in systems that are large, expensive and have insufficient data samples. Through the set model, the uncertainty factors are input to obtain the range of uncertainty.

In addition, the methods of solving the loss efficiency include fuzzy uncertainty, approximate analytic method, numerical integral method and so on.

As a new subject, reliability has its own system, method and technology, and has three research branches.

1. **Reliability Engineering**

Reliability engineering refers to the technical scheme and organizational management measures to ensure that the product or system achieves the predetermined reliability function in the process of design and operation. Reliability technology runs through the life cycle of the product and the whole process of system operation.

Reliability design. Through the early design to lay the reliability foundation of the system. The reliability model is established to predict the reliability of the product or system, and the failure mechanism is analyzed.

Reliability test. The reliability of the product or system is tested by simulation. Study the reliability of the product or system under certain time and cost constraints, find out the weak links, and improve these links.

Operational phase reliability. Maintain system reliability during system operation, carry out early troubleshooting and defect handling research. High reliability is maintained through reliability monitoring and diagnosis of system operation.

2. **Reliability Physics**

The research of reliability physics began in the 1960s, when semiconductor devices developed rapidly and often failed. The causes of failure are closely related to physics and failure physics. Reliability physics is mainly studied from the aspect of product mechanism, which provides the basis for the development of products.

3. **Reliability Mathematics**

The statistical law of product failure is studied, which mainly involves mathematical statistical methods of reliability prediction, analysis, design and evaluation of products or systems. For example, the space shuttle is composed of orbiter, booster, outer storage box and so on, and each part consists of hundreds, thousands, or even tens of thousands of basic components. Any component problem can cause serious accidents. For example, problems in a certain link in the urban dangerous goods logistics system may lead to the collapse of the whole logistics system and affect the basic production and life of the city.

6.1.3 Significance of Reliability Research

(1) **Reliability studies can reduce the probability of accidents**.
The reliability of products and systems not only affects the development of enterprises, but also poses a threat to personal safety. In 1996, after a flight of about 22 s, China's launch vehicle *Long March 3 B* hit on the hillside less than 2 km from the launch frame, and then brought a violent explosion with casualties. Later, the cause of the failure was identified as the failure of an electronic component. By studying the reliability of products and systems, we can find out the weak links and reduce the probability of failure.

(2) **Reliability research can effectively improve economic efficiency.**
In order to improve the reliability of the product or system, it is necessary to invest a lot of money in the early stage to carry out environmental simulation analysis, reliability prediction, analysis, experiment and so on. However, the reliability of the product or system is greatly improved after reliability research. This can effectively reduce maintenance costs and downtime costs, avoid unnecessary economic losses, thereby reducing the total cost and

improving economic efficiency. Although the early investment cost of reliability engineering is high, considering the total cost and service efficiency, it is necessary to invest in reliability engineering. For example, every holiday, the online merchants' various promotional activities lead to a surge in online shopping volume. If the system service capacity of the online shopping platform is high, it will win enterprises more customers' trust and more trading volume. In the long run, it will bring more economic benefits to enterprises.

(3) **Reliability studies can improve system utilization.**
The improvement of product reliability means the reduction of failure rate, which plays an important role in improving system utilization.

6.2 Reliability Eigenvector

Reliability eigenvector is a general term of various reliability indexes used to express the overall reliability of the system. Reliability, failure rate and other indicators are reliability eigenvectors.

1. **Reliability**

Reliability refers to the probability that a product or system completes a specified function under specified conditions and within a specified time. Degree is usually expressed in R. The reliability function of the system is described as:

$$R(t) = 1 - F(t) = \frac{N - n(t)}{N} \tag{6.1}$$

In terms of probability distribution, it represents the percentage of the part of the work that completes the specified function without failure in the specified conditions and time.

2. **Failure rate**

The failure rate is a product that has not failed at the t of a certain time. The probability of failure in the next unit time after the t of that time is recorded as $\lambda(t)$. An estimate of the failure rate refers to the ratio of the number of products that fail at the next unit of a certain time t to the number of products that have not failed at that time, that is recorded as $\hat{\lambda}(t)$.

There is a N product, starting from time $t = 0$, to the time t, the number of product failures are $n(t)$, and to the $(t + \Delta t)$ time, the number of product failures are $n(t + \Delta t)$, that is, if there are $\Delta n(t) = n(t + \Delta t) - n(t)$ product failure in the time interval $[t, t + \Delta t]$, the average failure rate of the product in the time interval

$[t, t + \Delta t]$ is defined as:

$$\overline{\lambda}(t) = \frac{n(t + \Delta t) - n(t)}{[N - n(t)] \cdot \Delta t} = \frac{\Delta n(t)}{[N - n(t)] \cdot \Delta t} \tag{6.2}$$

The estimated value $\hat{\lambda}(t)$ of the failure rate is:

$$\hat{\lambda}(t) = \frac{\text{Number of failed products per unit time within (t, t + }\Delta\text{t)}}{\text{Number of working products at t}} = \frac{\Delta n(t)}{(N - n(t))\Delta t} \tag{6.3}$$

6.3 System Reliability and Calculation

1. **Reliability of Series Systems**

In all the components of the system, if any unit fails, it will cause the whole system to fail. Such a system is called a series system. The structure of the series system is shown in Fig. 6.1.

The system S with n units, "System S reliable" is recorded as Ω_S, "Unit i reliable" is recorded as Ω_i, "unit failure" is recorded as $\overline{\Omega_i}$, the system normal working event is the intersection of Ω_i, namely: $\Omega_S = \bigcap_{i=1}^{n} \Omega_j$.

The reliability of the series system is:

$$R_S = P[\bigcap_{i=1} \Omega_j] = P\{\Omega_1\}P\{\Omega_2|\Omega_1\}\dots P\{\Omega_1|\Omega_{i-1}, \Omega_{i-2}, \dots \Omega_1\} \tag{6.4}$$

When $P\{\Omega_1|\Omega_{i-1}, \Omega_{i-2}, \dots \Omega_1\}$ represents that unit 1 to *i*-1 are valid, the conditional probability of Unit *i* is also valid.

The series system is simple, easy to coordinate and has high control degree. But problems at any one of them can paralyze the system. The more units in series, the lower the reliability of the system. Therefore, in order to improve the reliability of series system, one is to simplify the system design and reduce the number of units in the system, the other is to improve the reliability of weak units in the system.

2. **Reliability of Parallel Systems**

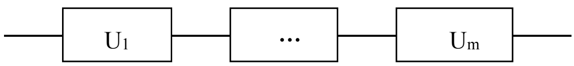


Fig. 6.1 Structure of series system

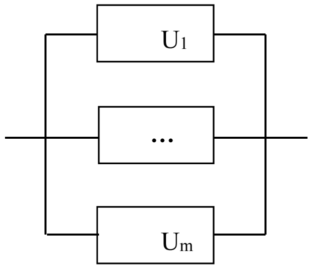


Fig. 6.2 Structure of parallel system

Only when all the units that make up the system fail, the system will fail. Such a system is called a parallel system. The structure of the parallel system is shown in Fig. 6.2.

The failure event of the system is the intersection of the failure events $\overline{\Omega_i}$ of each unit. When the failure events $\overline{\Omega_i}(i = 1, 2, \ldots, m)$ of a single unit are independent of each other, the reliability mathematical model of the system is as follows:

$$R_s(t) = 1 - \prod_{t}^{n} [1 - R_s(T)] \tag{6.5}$$

The parallel structure greatly reduces the dependence on the composition unit and makes the whole system more flexible. However, the resource allocation of parallel structure is large and the cost is high.

3. **Reliability of Mixed Systems**

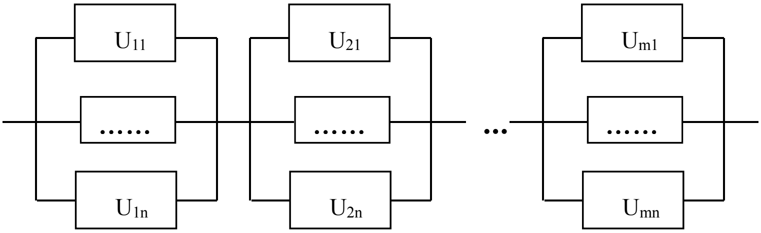
When there are both series and parallel structures in the system, it is called hybrid system. The hybrid system is divided into series–parallel system and parallel-series system, as shown in Fig. 6.3.

A series–parallel model is composed of m subsystems in series, and the internal components of each subsystem are parallel. Suppose i subsystem has the same n parts in parallel. The function of the system is that at least one part of the system works normally and the system can work normally.

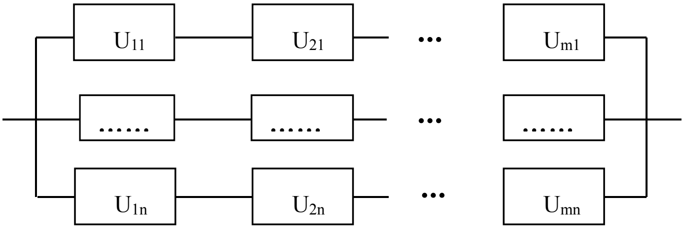
If the reliability of each component is: $R_{ij}(t)(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$, then the reliability of the series–parallel model is as follows:

$$R(t) = \prod_{i=1}^{m} \{1 - \prod_{j=1}^{n} [1 - R_{ij}(t)]\} \tag{6.6}$$

Parallel-series system is composed of m subsystems in parallel, each subsystem internal components are in series. The function of the system is that at least one subsystem is normal and the system can work properly.



(a) Series-parallel System



(b) Parallel-series System

Fig. 6.3 Structure of the hybrid system

The reliability of the parallel-series model is:

$$R(t) = 1 - \prod_{i=1}^{m} \{1 - \prod_{j=1}^{n} [1 - R_{ij}(t)]\} \tag{6.7}$$

The reliability analysis of the hybrid system needs to be decomposed into several modules to calculate the reliability of each module. Then the overall reliability is calculated according to the structural characteristics between the modules.

4. **Reliability of Network Systems**

According to node failure, network system can be divided into two models: node failure and node non-failure.

(1) Node non-failure model

The network system consists of arcs between nodes. It is assumed that the arcs are independent of each other and only normal and invalid two states, when the node is normal, the reliability of the node is 1. And the problem can be described that if the G is a network and r_i, r_j two nodes in the network, the probability of r_i to r_j can be obtained.

(2) Node failure model

The model of node failure is based on graph theory. The system is composed of nodes and edges. Both nodes and edges have the probability of working normally. In general, the reliability of node failure network includes network survivability and invulnerability.

The survivability of the network. It refers to the probability that the network with certain fault probability for nodes or links can keep the network connected under the action of random failure. The influence of network topology and random failure on the reliability of the whole network is studied.

The invulnerability of the network. In a network with completely determined topology, the ability of network to maintain connectivity under the influence of deterministic destruction is studied. It is necessary to destroy at least the number of nodes or links in the network when interrupting the connection between some nodes. The invulnerability is only related to the network topology.

Example 6.1 Supply Chain Network Reliability. According to the spatial combination model of supply chain, it can be divided into three structural models: linear structure, chain structure and network structure. Linear structure refers to the existence of only one supplier, one manufacturer and one seller throughout the supply chain. Chain structure and linear structure belong to series mode, but increase the number of node enterprises. In a chain structure, the raw materials needed by suppliers for production and operation cannot be provided by themselves, but need to be provided by other suppliers. Similarly, the seller does not sell the product directly to the consumer, but through the first-level or multi-level distributor to the consumer, which forms a chain structure model with multi-level suppliers and multi-level sellers. Network structure refers to the supply chain, including multiple levels, multiple suppliers, manufacturers and sellers. Because the supply chain network is mostly a complex network composed of many suppliers, manufacturers, distributors, retailers and end users, the supply chain network is a complex network structure, and each enterprise on the chain is the node that makes up the network system. Because each node enterprise is a cooperative relationship, each node has certain uncertainty and affects the reliability of the whole network, so the cooperation between node enterprises constitutes the basis of the overall reliability of the network. It plays a direct and key role in maintaining the reliability and effectiveness of the network and the overall synergy of the network.

The supply chain network consists of multiple suppliers ($\mathrm{B}_1, \mathrm{B}_2 \cdots \mathrm{B}_\mathrm{n}$), multiple manufacturers ($\mathrm{C}_1, C_2 \cdots C_\mathrm{n}$), multiple distributors ($\mathrm{D}_1, D_2 \cdots D_\mathrm{n}$), and multiple vendors ($\mathrm{E}_1, E_2 \cdots E_\mathrm{n}$), as shown in Fig. 6.4. In the supply chain network, the important products are provided by more than two suppliers, which improves the range of manufacturers' choice and improves the reliability accordingly.

In the supply chain network structure, each chain is composed of nodes and arcs. Nodes represent the production and operation activities within the enterprise, and arcs represent the transmission process of materials or information between

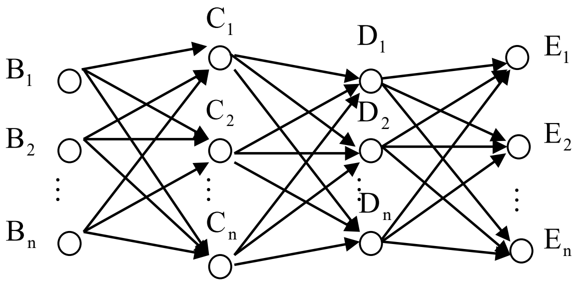


Fig. 6.4 Supply chain network structure

enterprises. Any two directly connected node enterprises (i and j) form a unit chain, and the reliability of the unit chain is reflected as the reliability of the internal reliability (r_i and r_j) of the two node enterprises and the reliability of the material or information transmission between the two nodes, that is the set of arc reliability ($r_{i,j}$), recorded as R_{ij}. Since the nodes in the middle level are both the end of the previous arc and the starting point of the next arc, in order to avoid repeated calculation, the reliability of the supply chain network can be regarded as the reliability of the unit chain

$$R_{ij} = r_i r_{i,j} \tag{6.8}$$

Because the final node of the supply chain network has no continuous arc, the reliability of the final node is the reliability of the unit chain at the last level.

To determine the reliability of supply chain network, it is necessary to calculate the reliability of all nodes and arcs in the network. According to the actual situation of supply chain network operation, if any two directly connected node enterprises cannot realize cooperation, the unit chain of the two nodes will fail. $u_{i,j}$ can represents the validity of collaboration of any single transmission unit chain, by Boolean variable

$$u_{i,j} = \begin{cases} 1 & unit\ chain\ transfer\ is\ valid \\ 0 & unit\ chain\ transfer\ is\ invalid \end{cases} \tag{6.9}$$

$$r_{i,j} = p(u_{i,j}) \tag{6.10}$$

If the probability of effective transmission of unit chain material or information is represented as $r_{i,j}$, then R_{ij} is the probability of effective cooperation between two points, $\left(1 - R_{ij}\right)$ the probability of failure of the chain. The path from the node enterprise i along the transmission direction to the end point has m path, then the reliability of the supply chain network is as follows:

$$R = 1 - \prod_{1}^{m} \left(1 - \prod R_{ij}\right) \tag{6.11}$$

In the network structure, if any level of supply chain fails, the supply chain network will fail as a whole. The statistical inference of the characteristic quantity of the supply chain network reliability is carried out by using the probability statistics method, the purpose of which is to effectively use all kinds of existing information to evaluate it, and according to the reliability evaluation, before the risk accumulates to the critical failure of supply chain system, the supply chain network is optimized and effective coordination measures are taken to improve the reliability level of supply chain network.

6.4 System Reliability Fault Analysis

There are many factors that cause system failure, so system failure analysis becomes an important part of system reliability.

6.4.1 Fault Tree Analysis

1. **Construction of Fault Trees**

Fault Tree Analysis (FTA) technology, developed by Bell Telephone Laboratory in 1962 is one of the important methods of system reliability analysis. FTA is to take the most pessimistic fault modes of the analyzed system as the goal of the analysis, and then find out all the factors that directly lead to the occurrence of this fault, and then continue to find out all the direct factors that affect the next level of events. The main purpose of qualitative analysis is to find out all the fault modes that cause the system failure, that is, to find all the fault modes that cause the top event. The main purpose of quantitative analysis is to find out the probability of top event and other quantitative indexes when the probability of occurrence of the underlying event or the basic event is given. FTA can identify the risk of various systems, not only can analyze the direct cause of the fault, but also reveal the potential cause of the fault.

The fault tree includes the following contents: (1) the possible disaster failure of the system, that is, the determination of the top event; (2) the inherent or potential risk factors within the system, including due to manual misoperation; (3) the connection and restriction relationship between each subsystem and each element, that is, the logical relationship between input (cause) and output (result), and marked with special symbols.

FTA is manifested in the form of a fault tree diagram (also known as a negative analysis tree). The fault tree diagram is a logical causality diagram that displays the state of the system (top event) according to the state of the element (basic event). The method of graphical "model" path is used to represent a system's "model" in which the basic events are connected by logical symbols from the failure of a predictable or unpredictable fault event.

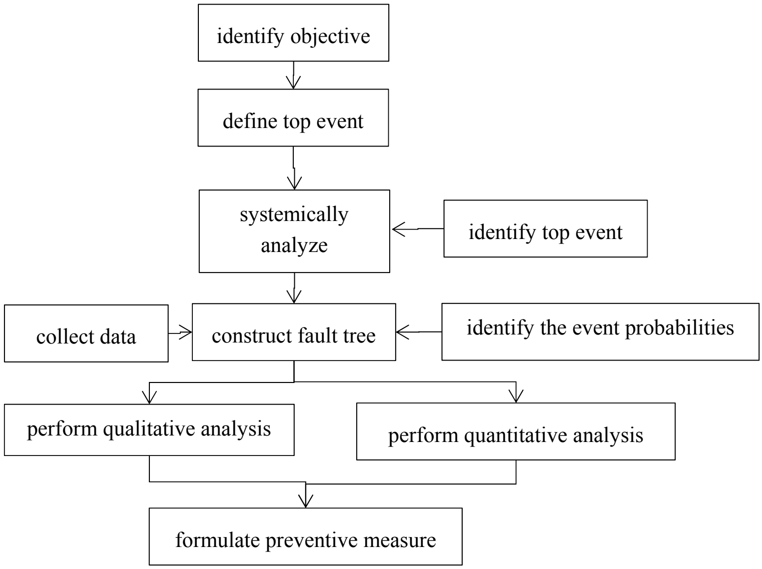


Fig. 6.5 Steps for fault tree analysis

FTA is divided into three stages: compiling fault tree, quantitative and qualitative analysis, formulating preventive countermeasures and improving system. As shown in Fig. 6.5.

FTA reveals the direct influencing factors of the existing faults and potential faults by comprehensively and vividly describing the various factors leading to the disaster faults and their logical relations. Through quantitative and qualitative analysis, it is useful to find out the potential dangers in the system, to find out the defects of the system, and to provide the basis for improving the safety design, formulating preventive measures and taking management countermeasures.

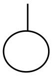
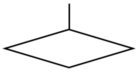
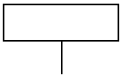
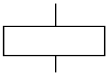
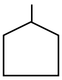
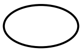
2. **Events and Logic Gates**

Event refers to the occurrence of conditions or actions, that is, direct or indirect causes of failure. It is mainly divided into three categories: bottom event, result event and special event (Table 6.1).

The logic gate only describes the logical causality between events in fault tree analysis, that is, when the input event satisfies some conditions, it will lead to the output event. The following table details the logical relationship between various input and output events (Table 6.2).

Example 6.2 High inventory is a common problem in manufacturing enterprises, which can be divided into internal and external reasons. The fault tree analysis method is used to analyze the possible causes of "enterprise inventory", to find out the key

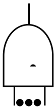
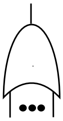
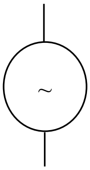
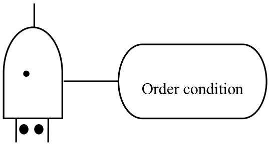
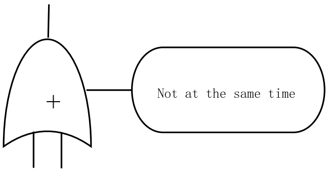
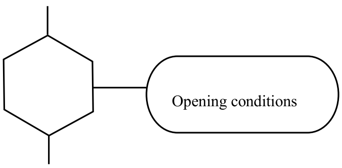
Table 6.1 Events and symbols

Events and symbols			Definition
Event symbol	Bottom events The only cause event in the fault tree that causes other events	 basic event	The circular symbol represents the basic event in the fault tree and is the bottom event that does not need to be identified
		 Undeveloped event	Represents the undiscovered event in the fault tree, which should be proved but temporarily unable or need not be found out. Such events generally indicate occurrence but small probability
	 Result events The result event in the fault tree can be a top event or an intermediate event. The lower end is connected to the logic gate, representing an input to the logic gate of the event	 Top event	The top event is the result of the joint action of all events. Is the source of downlink analysis in the fault tree, located at the top of the fault tree
		 Intermediate event	The intermediate event is the result event between the top event and the bottom event
	Special events Special symbols are commonly used in fault trees to indicate their particularity or attention-causing attribute	 External event	External events are special events that must or must not occur under normal working conditions. According to the fault requirement, it may be a normal event or a fault event
		 Conditioning event	Conditioning events describe events that are specifically restricted by the logic gate

problem in many reasons, and to take the most effective measures to reduce the risk of the factors leading to excessive inventory.

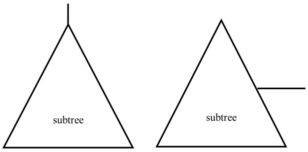
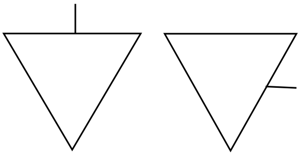
As shown in Fig. 6.6, in order to avoid excessive inventory, enterprises must find out the bottom events that affect inventory. A cut set, is a set of basic events that lead to a top event, that is, a set of basic events in a fault tree can cause a top event. The minimum set of basic events that cause the top event is called the minimum cut set. The logic sum of several event logic products is obtained by Boolean algebraic simplification method, in which each logical product is the minimum cut set. This example uses Boolean algebra method to solve the minimum cut set of the fault tree. The relationship between M1 and M2 is a union (or gate), the same M3, M4, M5 is a union, X1 to X16 is also a union, so the equation is obtained:

Table 6.2 Logical gates and symbols

Name of symbol		Definition
Logic symbol	 AND gate	It represents that if and only if all input events occur, the output event occurs
	 OR gate	It indicates that the output event occurs when at least one input event of all events occurs
	 NOT gate	It indicates that the output event is the opposite event of the input event
	 NAND gate	It indicates that output events occur only when input events occur in order
	 XOR gate	It indicates that an output event occurs if and only if a single input event
	 VOTING gate	It indicates that the output event occurs only when r or more of the n input events occur
	 INHIBIT gate	It indicates that if and only if the conditional event occurs, the input event will cause the output event to occur

(continued)

Table 6.2 (continued)

Name of symbol		Definition
	 Same transfer symbols	The same transfer symbol refers to the same position, direction and code of the subtree. Transfer-in symbol refers to going to the subtree referred to by alphanumeric code. Transfer-out symbol refers to going here with the same alphanumeric code
	 Similar transfer symbols	The similar transfer symbols indicate that the subtree position, transfer direction and subtree code are the same, and the event symbol is different. Similar transfer-in symbol refers to going to a subtree with a similar structure and different event labels. Similar transfer-out symbol refers to going here with the same alphanumeric code and different event codes

A = M1 + M2 = (M3 + M4 + M5) + M2 = [(X1 + X2 + X3 + X4) + (X5 + X6 + X7 + X8) + (X9 + X10 + X11 + X12)] + (X13 + X14 + X15 + X16) = X1 + X2 + … + X16, so the minimum cut set is obtained by Boolean algebraic simplification method are respectively {X1}, {X2}, {X3} …… {X16}.

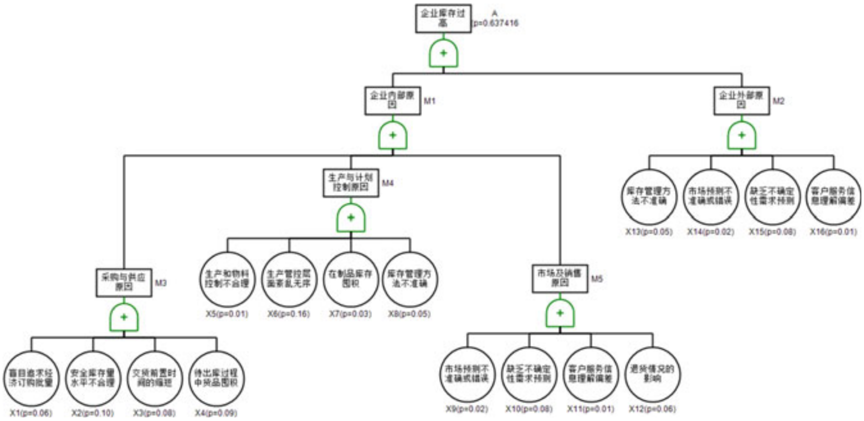
Assume a minimum reliability requirement is 0.8, then the maximum unreliability is 0.2 (i.e. the top event probability), Fig. 6.6 shows that, the causes of M3 failure are {X1, X2, X3, X4}, similarly M4 there are {X5, X6, X7, X8}, the causes of failure M5 reason are {X9, X10, X11, X12}, and M2 are { X13, X14, X15, X16}, M1 are { M3, M4, M5}, A are {M1, M2}.

Generally, the key importance of all factors is sorted from large to small, the top is the key factor, the next is the sub-key factor, and the last is the non-key factor. It is generally believed that the relatively large probability of bottom events is the key factor. P (X6) = 0.16, P (X2) = 0.1, P (X4)0.09, P (X3) = P(X10) = 0.08, so the probability of bottom event X6, X2, X4, X3, X10 is relatively large, which can be considered as a key factor.

6.4.2 Bayesian Network

1. **Fundamental Principle**

Bayesian network is a probabilistic uncertainty based reasoning network proposed by Judea Pearl in 1988, also known as belief network, causal probability network and so on. Bayesian network is a combination of Bayesian method and graphic theory, which plays an important role in the study of uncertainty. It is based on



企业库存过高 Overstocking 内部原因 internal reasons
生产与计划控制 Production and planning control
生产和物料控制不合理 Unreasonable production and material control
生产管控层面紊乱无序 Disorderly production control
在制品库存堆积 Work-in-process inventory backlog
库存管理方法不准确 Inaccurate Inventory management methods

采购与供应原因 Procurement and supply
盲目追求经济订购批量 Blind pursuit of economic ordering volume
安全库存量水平不合理 Unreasonable safety stock
交货前置时间的缩短 Reduced lead time
待出库过程中货品囤积 To-be-delivered inventory backlog

市场及销售原因 Marketing and sales
市场预测不准确或错误 Inaccurate or wrong market forecasts
缺乏不确定性需求预测 Lack of uncertainty in demand forecasting
客户服务信息理解偏差 Misunderstanding of customer service information
退货情况的影响 Impact of returns

外部原因 External causes
库存管理方法不准确 Inaccurate inventory management methods
市场预测不准确或错误 Inaccurate or wrong market forecasts
缺乏不确定性需求预测 Lack of uncertainty in demand forecasting
客户服务信息理解偏差 Misunderstanding of customer service information

Fig. 6.6 Fault tree diagram of excessive stock in enterprises probability and statistical theory and has the advantages of strong reasoning ability and convenient decision making.

The algorithm of Bayesian network is mainly based on Bayesian formula.

(1) Bayesian formula

Let the sample space of the experimental E be Ω,B as the E event, the event $A_1, A_2 \ldots A_n$ are incompatible, and $A_1, A_2 \ldots A_n$ is a complete event group, that is $\bigcup_{i=1}^{n} A_i = \Omega, A_i A_j = \emptyset, P(A_i) > 0$.

Then according to the multiplication theorem and conditional probability:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \tag{6.12}$$

The upper formula is the Bayesian formula used by Bayesian network in reasoning. Among them, the prior probability and the posterior probability are represented as $P(A_i)$ and $P(A_i|B)$. Based on the prior probability, the posterior probability is derived by Bayesian formula.

(2) Joint probability distribution

Suppose n nodes in the Bayesian network are represented as $X_{1,}$ $X_{2,}$ …. X_n respectively. The value of the node is expressed in lowercase letters, such as x_i, the value of the node X_i. For Bayesian networks with n nodes, the joint probability distribution can be obtained by chain rule as follows:

$$P(X_1, X_2 \ldots X_n) = \prod_{i=1}^{n} P(X_i|X_1, X_2, \ldots X_{i-1}) \tag{6.13}$$

If let the parent nodes set of Xi as parent (Xi), then the conditional probability of node Xi is:

$$P(X_i|X_1, X_2, \ldots.X_{i-1}) = P(X_i|\text{parent}(X_i)) \tag{6.14}$$

Therefore, the upper formula can be reduced to

$$P(X_1, X_2 \ldots X_n) = \prod_{i=1}^{n} p(X_i|\text{parent}(X_i)) \tag{6.15}$$

Formula (6.14) is the joint probability distribution of Bayesian network obtained by simplification according to conditional independence. The calculation of probability distribution can be simplified by this method.

2. **Parameter Learning**

The parameter learning of Bayesian network is essentially to learn the probability distribution table of each node under the condition of known network structure. The probability distribution of early Bayesian networks is mainly determined by expert judgment, but this method, which is analyzed only by expert experience, will lead to a large deviation from the observed data. The current popular method is to learn the probability distribution of these parameters from the data and then do further analysis and calculation. An important feature of Bayesian networks is the direct description of the real world rather than the process of its reasoning.

Bayesian network modeling is mainly divided into two parts: qualitative and quantitative. Qualitative part: according to the expert knowledge and experience, the content and number of network nodes and the master–slave relationship of each node are listed, and the Bayesian network topology is finally determined. Quantitative part: according to the statistical data, the conditional probability values of each child node and parent node can be obtained.

6.5 Application Case

O2O Risk Identification Model of Supply Chain Failure

This application case takes fresh agricultural products as an example to study an O2O risk identification model of fresh agricultural products supply chain failure is studied. According to the Ellie consulting statistics, China's e-commerce market reached 12.3 trillion yuan in 2014, 21.3% higher than in 2013. Among them, online shopping increased by 48.7%, China's living service O2O market is 248.01 billion yuan in 2014, the O2O of local living services increased by 42.8%; And in the second quarter of 2015, China's e-commerce market as a whole traded at 3.75 trillion yuan. Up 22.1% year on year, the month-on-month increase was 7.8%. With the rapid development of mobile Internet, O2O platforms that meet people's daily needs emerge. Throughout the people's food, clothing, housing and entertainment and other aspects.

1. **Determining Bayesian Network Nodes and Ranges**

A model of supply chain failure risk identification in fresh agricultural products O2O mode is established through Bayesian network, 28 factors are determined, a five-level hierarchy is formed, and its influencing factors are divided into 28 network nodes. According to the O2O model of fresh agricultural products, the supply chain model combines the manufacturers, suppliers, service providers and customers in the supply chain, and takes the final customer satisfaction as the measure index. The model can fully reflect the influencing factors in the failure risk of the supply chain.

(1) Product quality factors

As an important index to measure the supply chain the O2O mode of fresh agricultural products, product quality directly affects the evaluation of customer satisfaction, which is mainly reflected in five aspects: picking technology, storage technology level, packaging technology level, breakage rate of fresh agricultural products and fresh agricultural products preservation period:

①The picking technology of agricultural products is mainly reflected in the use of tools in the process of picking agricultural products and the treatment technology after picking, the selection, grading and dressing of agricultural products, which makes the agricultural products easy to transport, store and sell in the process of reducing the loss of quality.

②Storage technology level refers to the cold storage technology and preservation technology after picking and treating agricultural products, so as to reduce the loss of agricultural products.

③Packaging technology level refers to fresh agricultural products' packaging treatment in order to reduce moisture loss, enhance product value, promote product sales, including packaging materials, packaging methods, packaging equipment and so on.

④The effect of damage rate of fresh agricultural products on product quality is manifested in the loss of fresh agricultural products and the breakage of sales process.

⑤The high moisture content and perishable characteristics of fresh agricultural products determine the preservation period of fresh agricultural products.

(2) Product price factors

And the effect of product price on customer satisfaction under the O2O mode of fresh agricultural products is mainly manifested in four aspects: purchasing cost, reverse cost, holding cost and transportation cost.

①The influencing factors of purchasing cost are mainly reflected in the purchasing times, purchasing quantity, credit level and bargaining power of purchasing enterprises of fresh agricultural products.

②Reverse cost refers to the recovery of packaging materials, the collection and feedback of sales information and the return of agricultural products.

③The cost of holding is reflected in the safety inventory prepared by the seller of agricultural products to cope with the uncertain changes in the market, as well as the cost of turnover inventory to meet the daily needs of consumers. The uniqueness of fresh agricultural products determines the flexibility of inventory and should avoid excessive holding costs.

④Transportation cost refers to the loading and unloading and handling of fresh agricultural products from sellers to final consumers. For some fresh agricultural products with special requirements, special packaging is needed, and cold chain logistics is needed for distribution.

(3) Service level factors

The influence of service level on customer satisfaction is reflected in logistics level, external environmental factors, while logistics level is affected by distribution ability, distribution efficiency and utilization of logistics resources, and external environmental factors are reflected in capital level, information level, personnel ability and policies and regulations.

①Logistics level refers to the online order processing ability and distribution efficiency of fresh agricultural products in O2O mode, and the utilization efficiency of logistics resources such as technical equipment and infrastructure.

②The external environment is manifested in the proportion of knowledge structure, the personnel ability of practitioners' skill level, the accuracy of market prediction, the information level of the timeliness of fresh agricultural products information, as well as the policies and regulations, the level of capital and so on.

The serial number and range of each node in Bayesian network are shown in Table 6.3.

2. **Bayesian Network Construction**

There are three main ways to determine the Bayesian network structure model: ① The topology of the network model is established manually by expert knowledge; ② Through the analysis of the database, the modeling method of Bayesian network structure can be automatically obtained; ③ A modeling approach combining the two stages, that is, first establishing Bayesian network structure manually by expert knowledge, then obtaining the Bayesian network model before the database modified by the analysis of the database.

(1) Calculation of conditional probability in Bayesian networks

Let the set of parent nodes X_i be *parent*(X_i), a total of n, using x_i to represent the value of the X_i, B_i to represent the vector which is consist of the parent node variable, the value of the vector b_i represents the value of vector B_i, then the conditional probability of the node X_i is:

$$P(X_i|\text{parent}(X_i)) = \frac{P(X_i, \text{parent}(X_i))}{P(\text{parent}(X_i))} = \frac{P(X_i = x_i, B_i = b_i)}{P(B_i = b_i)}$$

Table 6.3 Range of Bayesian networks

Order	Nodes	Range
A	Customer satisfaction	(0, 1)
A_1	Product quality	(0, 1)
A_2	Service level	(0, 1)
A_3	Product price	(0, 1)
B_1	Technical level	(0, 1)
B_2	Damage rate of fresh agricultural products	(0, 1)
B_3	Fresh agricultural products preservation period	(0, 1)
B_4	Logistics level	(0, 1)
B_5	External environment	(0, 1)
B_6	Transportation cost	(0, 1)
B_7	Holding cost	(0, 1)
B_8	Reverse cost	(0, 1)
B_9	Procurement cost	(0, 1)
C_1	Distribution capacity	(0, 1)
C_2	Distribution efficiency	(0, 1)
C_3	Utilization of logistics resources	(0, 1)
C_4	Personnel capacity	(0, 1)
C_5	Level of informationization	(0, 1)
C_6	Capital level	(0, 1)
C_7	Technical level of packaging	(0, 1)
C_8	Storage technology level	(0, 1)
C_9	Agricultural picking technology	(0, 1)
D_1	Utilization of technical equipment	(0, 1)
D_2	Infrastructure utilization	(0, 1)
D_3	Proportion of knowledge structure	(0, 1)
D_4	Skills level of practitioners	(0, 1)
D_5	Market forecasting accuracy	(0, 1)
D_6	Timeliness of information on fresh agricultural products	(0, 1)

The conditional probability is calculated as follows for X_i with three parent nodes. When the parent node is in State0, the conditional probability value of the node X_i in State0 is:

$$\begin{aligned} & P(X_i = \text{State0} | B_1 = \text{State0}, B_2 = \text{State0}, B_3 = \text{State0}) \\ & = \frac{P(X_i = \text{State0}, B_1 = \text{State0}, B_2 = \text{State0}, B_3 = \text{State0})}{P(B_1 = \text{State0}, B_2 = \text{State0}, B_3 = \text{State0})} \end{aligned}$$

Among them, the parameter State0 represents the state in which the node is in, and can be described as the state of the event is poor or better. From the above formula, it can be seen that the calculation of conditional probability requires a large number of sample data to meet the requirements of different values of each node, and with the increase of the number of parent nodes, the calculation amount is also increasing. In the case of unable to obtain accurate probability, it is necessary to use the concept of group decision to judge according to the experience of experts. Experts are asked about the conditional probability of nodes by questionnaire survey and triangular fuzzy number method is used to process relevant data.

IPCC (Intergovernmental Panel on Climate Change) uses seven-file hierarchical language variables to describe the probability values. Each language variable and its corresponding probability values and triangular fuzzy numbers are shown in Table 6.4.

The conditional probability table of nodes is obtained by questionnaire survey, and can be converted into triangular fuzzy number according to Table 6.4

$$\tilde{P}_{ij}^{k} = (a_{ij}^{k}, m_{ij}^{k}, b_{ij}^{k})(k = 1, 2, \ldots, q)$$

The average fuzzy probability value of node X_i in j state is

$$\tilde{P}'_{ij} = \frac{\tilde{P}_{ij}^{1} \oplus \tilde{P}_{ij}^{2} \oplus \ldots \oplus \tilde{P}_{ij}^{q}}{q} = (a'_{ij}, m'_{ij}, b'_{ij})$$

The exact probability of node X_i in j state is further calculated by means of mean area method:

$$P'_{ij} = \frac{a'_{ij} + 2m'_{ij} + b'_{ij}}{4}$$

The conditional probability value of the node is obtained by normalization:

$$P = \frac{P'_{ij}}{\sum P'_{ij}}$$

Table 6.4 Meaning values and corresponding triangular modulus of event occurrence probability

Probabilistic range (%)	Triangular fuzzy numbers	Statement
<1	(0.0, 0.0, 0.1)	Very low
1–10	(0.0, 0.1, 0.3)	Low
10–33	(0.1, 0.3, 0.5)	Lower
33–66	(0.3, 0.5, 0.7)	Medium
66–90	(0.5, 0.7, 0.9)	Higher
90–99	(0.7, 0.9, 1.0)	High
>99	(0.9, 1.0, 1.0)	Very high

(2) Determination of conditional probability tables

A complete Bayesian network model contains network topology and parameters in the model. Model parameters refer to the probability distribution table on each node. After establishing the network topology, it is necessary to quantitatively describe the probability relationship between each node, which is the basis of reasoning using Bayesian network. A suitable conditional probability table is introduced for each node (Conditional Probability Table, CPT). The following is an example of a node's conditional probability table, as shown in Table 6.5.

3. **Numerical Validation**

Through the analysis and arrangement of the questionnaire data and input it into Bayesian simulation software, the probability values of each node can be obtained, in which "State0" indicates that the state is poor, "State1" indicates that the state is good.

According to the reverse reasoning of Bayesian network, when the customer is not satisfied, that is $P(A = \text{State0}) = 1$, it can be considered that the failure risk of supply chain occurs at this time, and the influence of service level, product price and product quality on customer dissatisfaction is weakened in turn.

According to the simulation reasoning of Bayesian network by simulation software, the service level has the greatest influence on the failure risk of supply chain, the corresponding probability value is 37%, followed by product price and product quality. The most approximate cause of supply chain failure risk is {infrastructure utilization →logistics resource utilization →logistics level →service level →customer satisfaction}.

According to the actual evaluation results and the questionnaire survey of influencing factors, customers first attach importance to the quality of service level, followed by product price and product quality. Fresh agricultural products e-commerce consumer groups are mainly young people, especially young office workers fast-paced lifestyle more in line with fresh agricultural products e-commerce "fast", "keeping indoors" and other selling points. Their income level is relatively high, and the purchasing power is strong, so for them the product

Table 6.5 Conditional probability of logistics resource utilization

Condition		Triangular fuzzy numbers		Probability	
State0	State1	State0	State1	State0	State1
D_1, D_2	——	(0.10,0.23,0.43)	(0.37,0.57,0.73)	0.31	0.69
D_1	D_2	(0.17,0.37,0.57)	(0.43,0.63,0.83)	0.37	0.63
D_2	D_1	(0.23,0.43,0.63)	(0.37,0.57,0.70)	0.45	0.55
——	D_1、D_2	(0.13,0.30,0.50)	(0.63,0.83,0.97)	0.27	0.73

price is a secondary factor compared with punctual and efficient delivery of products. Fresh agricultural products market competition is increasingly fierce, due to the low threshold of the industry, a large number O2O fresh agricultural products e-commerce enterprises emerged. In order to prevent the failure risk of fresh agricultural products supply chain under the O2O mode, e-commerce enterprises should pay special attention to the improvement of service level, improve distribution efficiency and distribution ability, and provide customers with high quality service experience.

Summary

The introduction of reliability knowledge into management science is a hot spot in recent years. Many research results have been obtained in many fields, such as supply chain reliability and logistics reliability. With the promotion of big data analysis, I believe that in the future in the field of management reliability research more specific and in-depth. The main points of this chapter include: the basic concept and principle of reliability, the reliability calculation of feature vector, reliability, failure rate, etc., the reliability calculation of series system, parallel system and hybrid system, and fault tree analysis.

Important concepts and terminology

reliability

reliability degree

system reliability

network reliability

reliability design

failure rate

series system

parallel system

....

Fault Tree Analysis, FTA

Questions and exercises

1. What are the differences and connections between reliability management and quality management?
2. Take supply chain reliability as an example to elaborate the principle of system failure analysis using fault tree.

3. Calculation: a system accords with the exponential distribution of parameters λ, its average life is 100 h, what is the reliability of 10 h, 50 h, 100 h of continuous operation?

Case analysis

Applying multi-layer Bayes estimation method to study the reliability of supply chain network, assuming that the data of 24 supply chain networks in a certain industry are shown in Table 6.6, that is, the sample starts from the effective operation time to the monitoring time. No failure time is (unit: month). According to the distribution characteristics of the sample without failure operation period, the data are divided into 12 groups ($i = 1, 2, \cdots, 12$). The i group of data indicates that there is a n_i supply chain network without failure operation time t_i at the monitoring time.

According to the analysis of the data, the failure-free operation period of the supply chain network can be regarded as the system life, and the exponential distribution can be approximated. The application of multi-layer Bayes estimation method can improve the accuracy of the prior distribution parameters to a certain extent.

Let the life T of the empirical sample (that is, no failure operation period) follow the exponential distribution, and its density function is:

$$f(t) = \lambda \exp(-t\lambda),\ t > 0 \tag{6.16}$$

Let the non-failure data carry out m fixed time truncated test separately, the end time is $t_i(t_1 < t_2 \cdots < t_m)$(unit: month), the number of test samples is $n_i(n_i$ is the sum of the number of t_i eligible individuals in the sample), if the result of the test is that all individuals have no failure, $(t_i, n_i)(i = 1, 2, \cdots, m)$ is called no failure data.

If the λ prior density kernel is λ^{a-1}, $0 < \lambda < \lambda_0$, $0 < a < 1$ and a is a constant, but cannot determine the specific value of a, the uniform distribution on (0,1) is taken as the prior distribution of a, and its density function is $\pi(a) = 1,\ (0 < a < 1)$. Then the prior density of λ is:

$$\pi(\lambda|a) = \frac{a}{\lambda_0^a}\lambda^{a-1},\ 0 < \lambda < \lambda_0 \tag{6.17}$$

Table 6.6 No failure data for supply chain networks

i	1	2	3	4	5	6	7	8	9	10	11	12
t_i	6	9	12	15	18	21	24	27	30	33	36	39
n_i	3	2	3	2	3	1	3	1	2	2	1	1

The multilayer prior density of λ is:

$$\pi(\lambda) = \int_0^1 \pi(\lambda|a)\pi(a)da = \int_0^1 \frac{a}{\lambda_0^a}\lambda^{a-1}da,\ 0 < \lambda < \lambda_0 \tag{6.18}$$

Theorem m fixed time truncated tests were conducted on samples with exponential lifetime distribution (1), as a result, none of the individuals failed, non-failure data obtained are (t_i, n_i), $(i = 1, 2, \cdots, m)$, if the multilayer prior density $\pi(\lambda)$ of λ is given by (3), the multilayer Bayes of λ is estimated to be:

$$\hat{\lambda} = \frac{1}{T}\frac{\int_0^1 \frac{a\Gamma(a+1)}{(T\lambda_0)^a} I_{T\lambda_0}(a+1)da}{\int_0^1 \frac{a\Gamma(a)}{(T\lambda_0)^a} I_{T\lambda_0}(a)da} \tag{6.19}$$

and $T = \sum_{i=1}^{m} n_i t_i$, $I_x(a) = \frac{1}{\Gamma(a)}\int_0^x t^{a-1}\exp(-t)dt$, $(0 < x < \infty)$ is incomplete Gamma function, $\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)dt$ is Gamma function.

Apply multilayer Bayes estimation method to calculate supply chain network reliability. According to the information of example 6.1, the knowledge of system reliability is applied to analyze the measures to improve the reliability of supply chain network.

Further Readings

1. Li Shouze, Yu Jianjun, Sun Shudong. Identification and evaluation of supply chain failure risk [J]. *Application Research of Computers*, 2010, 27(12): 4568–4570
2. Li Enping, Glan, Dong Guohui, et al. Analysis of supply chain reliability diagnosis based on Bayesian network [J]. *Logistics Technology*, 2010, 29(17): 96–99.
3. Hu Wei, Cheng Youming. DEMATEL improvement and its application in supply chain failure analysis [J] *Statistics & Decision*, 2013, (6): 83–86.
4. Zhang Hao, Yang Haoxiong, Guo Jinlong. Multi-level Bayes estimation model of supply chain network reliability [J] *Journal of Systems Science and Mathematical Sciences*, 2012, 32(1): 45–52.
5. Lu Ningyun, He Kelei, Jiang Bin, et al. A fault prediction method based on Bayesian network [J] Journal of Southeast University, 2012, 42(1): 87–91
6. [6] Zhou Guohua, Peng Bo. Risk factor analysis of construction project quality management based on Bayesian network: Taking Beijing-Shanghai high-speed railway construction project as an example [J]. *China Soft Science*, 2009, (9): 99-106.
7. Analysys. Special research report on O2O market of living services in China, 2015[DB/OL].] https://www.analysys.cn/article/detail/9811. (Accessed 09/09/2015).
8. Bi Yuping. *Study on Logistics Supply Chain Model of Fresh Agricultural Products in China: Taking Shandong Fresh Agricultural Products as an Example* [M]. Beijing: China Social Sciences Press, 2014: 130–152
9. Wang Tao, Liao Binchao, Ma Xin, et al. Construction safety risk probability assessment method based on Bayesian network [J] *China Civil Engineering Journal*, 2010, 43(18): 51–61.
10. [10] Fan Houming, Wen Wenhua, Zhang Enying, et al. Evaluation of oil spill risk based on Bayesian network in ship's port-to-port handling operations [J]. *Journal of Mathematics in Practice and Theory*, 2015, 45 (1): 51-61.

11. Ma Dezhong, Zhou Zhen, Yu Xiaoyang, et al. Reliability analysis of multi-state Bayesian network based on fuzzy probability [J] *Systems Engineering and Electronics*, 34(12): 2607–2611.
12. Lu Mingyin, Xu Renping. *System Reliability* [M]. Beijing: Mechanical Industry Press, 2008.
13. Wang Xianpei. *System Reliability Theory* [M]. Wuhan: Wuhan University Press, 2012.
14. Zhang Weifang, Tang Qingyun, et al. *New Progress in System Reliability Research* [M]. Beijing: National Defence Industry Press, 2014.

Game Theory

7

Hao Zhang

Learning Target

1. Mastering the basic concepts of game theory and be able to construct basic game models.
2. Understanding the development process and representatives of game theory.
3. Mastering typical analysis methods of non-cooperative games and cooperative games.
4. Mastering the analytical thinking of evolutionary games.

Introduction

The advertising game between Coca-Cola and Pepsi

1. The creative game between Coca-Cola and Pepsi in online advertising

(1) Sports competition. Coca-Cola was awarded the designated beverage qualification for the Olympic Games. The online advertisement pour in about the Olympic Games. The red color of Coca-Cola brand and the Olympic flame matched very well. The establishment of a volunteer registration website greatly increased the support rate of the Internet. Meanwhile, Pepsi uses the NBA and the American Baseball League to reach a balance. Pepsi's online advertisements are livelier, and both the composition of the picture and the use of animation convey a "cool" feeling. From the NBA to baseball, from Oscars to *Tomb Raider* (game and movie), Pepsi's online advertising can always capture the interests and concerns of teenagers.

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_7

(2) Music competition. Pepsi has worked hard in this regard. Richi Martain, Britney Spears and the Weezer band have shown a smile to consumers for Pepsi on the Internet. Coca-Cola ads hiring Aguilera as its spokesperson on the Internet are also all the rage.
(3) Event competition. Coca-Cola and Pepsi are very good at creating opportunities for themselves to attract attention. Pepsi mobilized netizens to vote for "Pepsi-Cola's Best TV Commercial" on the Internet, and Coca-Cola held an online music competition.

David Ogilvy once said: "Unless your advertising contains a big idea, it will pass like a ship in the night." In fact, creativity refers to the way of conveying information. There are almost no successful brands that do not have core creativity in their advertisements. Creativity sparkles advertisements. Creativity makes a product brand emerge from many brands. If the advertising theme is a dragon, then creativity is the eyes. Two major company advertising competition did not cause any loss to any party, but increased their popularity. In terms of online advertising creativity, the game between the two cola giants can be described as evenly matched.

2. The characteristics of the advertising strategy game between Coca-Cola and Pepsi.

(1) Advertising form: "following" but "not pursuing the same"

The game between Coca-Cola and Pepsi's advertising strategy is intense and exciting. Their respective strategic goals are clear, their response strategies are timely and effective, and they are innovative. They are always paying attention to the other side's dynamics, finding the right time, and launching an offensive. Both in strategy and decision-making, they have their own advantages.

Pepsi first announced a comprehensive network promotion cooperation with Yahoo, while Coca-Cola signed a huge dollar network promotion contract with AOL. Coca-Cola gained the upper hand by being the exclusive sponsor of the Chinese Olympic Games, while Pepsi took the initiative to open up new ways and fully utilized the world theme of the Olympics to carry out public welfare activities, and won a good social response.

In order to highlight the Olympic theme, Coca-Cola's TV advertising spokespersons selected stars from different sporting events, such as Liu Xiang and Yao Ming. This poses a difficult problem in the choice of Pepsi's spokesperson, as champions are exclusive. However, as we mentioned above, the advertising strategies of the two run parallel well. Pepsi uses the "grass-root hero" strategy, that is, they elected a spokesperson from consumers. Pepsi's "grass-root hero" and Coca-Cola's "celebrity endorsements" are so well-matched.

Flat sticker is a form of advertising that Coca-Cola is good at. During the Olympics, Coca advertised much on public facilities, such as bus bodies, handlebars, and commercial centers, making a feeling that Coca-Cola is everywhere and the Olympics are around. Pepsi has always paid great attention to their consumer groups, that is, "youths". When the website Xiaonei was very popular among students, Pepsi took it as a stage to call on Chinese to take action with the theme of the Olympic Games. The large-scale consumer interactive activity of "Dancing China Together", all ordinary people who are willing to cheer for the Chinese team may become the protagonists on the new Pepsi "Photo Can". The image of these ordinary people will appear on the more than 100 million Pepsi bottles and cans, side by side with the heroes of the Chinese team cheering together, leading the fashion once again.

(2) The advertising slogan is "struggling" but "non-relative".

To a good advertisement, the content is the main body, and the slogan is soul. A good slogan is the eyes of the brand, which are significant for people to understand the connotation of the brand and build brand loyalty. Successful slogans are impressed and unforgettable. So what are the characteristics of successful advertising slogans? They are tasteful, novel and thematic. Company is equal to brand, and brand is equal to commodities, so the commodity ads are building the image of the company, like Coca-Cola, one of its slogan "Always Coca-Cola", while Pepsi "Ask for more" and "Dare for more". The advertising slogans of Pepsi and Coca-Cola are changeable with time, bringing happiness and hope to consumers in their own unique way.

Source: "Analysis of Coca-Cola and Pepsi's advertising strategies from the perspective of game theory".

7.1 Overview of Game Theory

7.1.1 Related Concepts of Game Theory

1. **Definition of Game Theory**

Game theory is a new branch of modern mathematics and an important subject of operations research. It refers to the process by which individuals or organizations, under certain conditions or rules, choose their own behaviors or strategies based on the information they have, and obtain corresponding benefits or results.

Game here in Chinese is *boyi*. In ancient China, *bo* and *yi* respectively referred to two games. *Bo* is interpreted as meaning gambling. "Yi" is interpreted as the meaning of playing *weiqi* or Chinese chess. The word *boyi* comes from the Chinese classics *The Book of Han Dynasty*.

At present, game theory has become one of the standard analysis tools of economics, and it has a wide range of applications in management, biology, international relations, computer science, political science, military strategy and many other disciplines.

2. **The basic elements of game theory**

Generally speaking, a complete game usually contains 4 elements.

(1) Player. Players, namely game participants, also known as game parties, refer to individuals or organizations that can independently make decisions and actions and bear the results of decisions in the game. A game with two players participating is called a "two-player game", and a game with more than two players is called a "multiplayer game".
(2) Strategy and strategy set. The strategy of the game refers to the action plan that the players can take. The set of all the strategies of a player is called the strategy set or action space of the player. If in the game, the strategy set of each player is a finite set, then the game is called a "limited game", otherwise it is called an "infinite game".
(3) Payoff and payoff function. Payoff refers to the payoff earned by each player in the game. This benefit is not only related to the strategy chosen by the players themselves, but also related to the set of strategies adopted by all players. The payoff function means that the "profit" of each player is a function of a set of strategies adopted by all players. The value of the payment function may be a positive or negative number, it may be a certain amount, such as profit, output, etc., or it may be a quantified utility such as satisfaction.
(4) Equilibrium. The equilibrium of the game is the set of actions formed by the optimal strategy of all players in the game. In economics, equilibrium means that the relevant quantity is a stable value.

3. **Classification of Game Theory**

The content of the game is very extensive. According to the different characteristics of the game, it can be divided into the following categories:

(1) Cooperative game and non-cooperative game. According to whether the players in the game can reach a binding agreement for mutual cooperation, the game can be divided into a cooperative game and a non-cooperative game. The difference between the two lies in whether there is a binding agreement between the players that interact with each other. If there is, it is a cooperative game, otherwise it is a non-cooperative game. Cooperative games emphasize group rationality, while non-cooperative games emphasize individual rationality and individual optimal decision-making.
(2) Complete information game and incomplete information game. According to the players' understanding of other players, it can be divided into complete information game and incomplete information game. The complete information game means that in the game process, each player has accurate information about the characteristics, strategy space and revenue function of other players. Incomplete information game refers to a game played when

players in the game do not have accurate information about the characteristics, strategy space and profit function of other players.

(3) Static game and dynamic game. According to the sequence of actions, it can be divided into static game and dynamic game. The static game means that in the game, the actions of the players are carried out at the same time, or although they are not selected at the same time, the later actors do not know what specific actions the first actor took. Dynamic game means that in the game, the actions of the players in the game have a sequence, and the later actors can observe the actions of the first actors.

In a static game, if the players in the game fully understand the characteristics of the game, the strategy space, and the return function, it is called a static game with complete information; on the contrary, if the players in the game are not fully understand the characteristics, strategy space and profit function, it is called static game with incomplete information. Similarly, dynamic games can be divided into dynamic game with complete information and dynamic game with incomplete information. The above four games are all non-cooperative games.

(4) Zero-sum game, constant-sum game and variable-sum game. From the perspective of benefit distribution, it can be divided into zero-sum game, constant-sum game and variable-sum game.

Zero-sum game: also called "strictly competitive game." The interests of the players are always opposed, and the preferences are usually different. The parties participating in the game, under strict competition, the gains of one party inevitably mean the losses of the other party, and the sum of the gains and losses of all parties in the game is always "zero". For example, guessing coins, playing chess, Tian Ji's strategy on horse racing.

Constant-sum game: The sum of the interests between the players is a constant. The interests of the players are antagonistic and competitive, and there is a possibility of a win–win situation. For example, the distribution of a fixed amount of bonus.

Variable-sum game: all games except zero-sum game and constant sum game. The sum of the interests of the game parties under different strategy combinations (results) is often different in the variable sum game. For example, prisoners' dilemma, institutional issues, etc.

In addition, game theory has many other forms of classification. For example, according to the number or duration of the game, it can be divided into limited games and infinite games; according to the form of expression, it can also be divided into general games or extended games; according to the different hypothesis of complete rationality and bounded rationality used in games, it can be divided into traditional game and evolutionary game; according to whether the players in the game are aligned or not, it can also be divided into aligned game and non-aligned game.

7.1.2 The Development of Game Theory

1. **Early thoughts of game theory**

The game theory can be traced back to ancient China. For example, in the Spring and Autumn Period, the military thoughts mentioned in *The Art of War* reflected the thoughts of game theory, and this is the earliest book on game theory.

In 1838, French economist Antoine Augustin Cournot put forward the "Cournot model" on output competition, and then in 1883 put forward the "Bertrand model" on price competition.

The famous mathematician John von Neumann proved the minimum and maximum theorem in 1928. After that, this theorem was called the essence of game theory, and since then it also announced the official birth of game theory.

2. **The formation of the game theory system**

The mid-1940s to the early 1950s marked the establishment and formation of a modern game theory system. *Theory of Game and Economic Behavior* published in 1944 by the mathematician John von Neumann and the economist Oskar Morgenstern, is considered to be the foundation stone of game theory. It also marks the formal formation of the game theory system.

3. **The boom of game theory**

Since the 1950s, it has entered a period of vigorous development of game theory.

In 1951, John Nash used the fixed point theorem to prove the existence of game equilibrium points, and proposed the concept of "Nash equilibrium", which laid a solid foundation for the cooperative and non-cooperative theories of game theory, and provided the later scholars the important theoretical guidance.

German economist Reinhard Selten and American economist John Harsanyi introduced incomplete information theory into the study of game theory in the 1970s, which played a role in promoting the development of game theory.

American economist Professor David M. Kreps and Professor Robert Wilson extended the subgame perfect equilibrium in 1982. American statistician J. Neyman studied repeated games in 1985. During this period, many research works on game theory came out, and more and more scholars applied game theory to the field of economics. Until the end of the 1980s, game theory had formed a complete system. In the 1990s, game theory was used in statistics, biology, military and many other fields. After that, game theory has become a part of mainstream economics and has had a significant impact on economics. Since game theory has entered a period of vigorous development, it has aroused widespread social attention to game theory, and this field has also produced many Nobel Prize winners in economics.

7.1.3 Game Theory and the Nobel Prize in Economics

Since 1994, many Nobel Prize winners in economics have studied fields related to game theory.

In 1994, the Nobel Prize in Economics was awarded to John Hessani of the University of California at Berkeley, John Nash of Princeton University and Reinhard Selten of the University of Bonn in Germany. Their contribution is to study the equilibrium analysis theory of non-cooperative games, which has had a great impact on the development of economics and game theory (Fig. 7.1).

In 1996, the Nobel Prize in Economics was awarded to James A. Mirrlees of Cambridge University in the United Kingdom and William Vickrey of Columbia University in the United States. Their contributions are respectively in the field of information economics theory and game theory (Fig. 7.2).

In 2001, the Nobel Prize in Economics was awarded to A. Michael Spence from Stanford University, George A. Akerlof from the University of California, Berkeley, and Joseph E. Stiglitz from Columbia University. They used game theory to study economic incentives under the conditions of asymmetric information in

Fig. 7.1 John Harsanyi, John Nash and Reinhard Selten (from left to right)

Fig. 7.2 James A. Mirrlees and William Vickrey (from left to right)

Fig. 7.3 Michael Spence, George A. Akerlov and Joseph E. Stiglitz (from left to right)

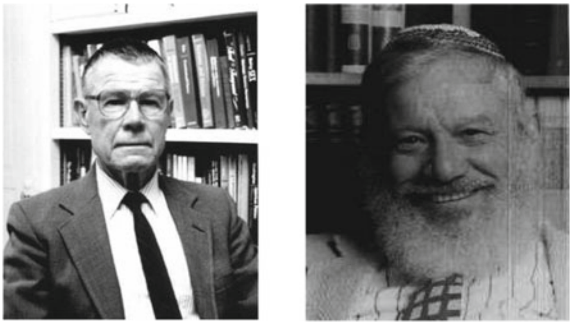


Fig. 7.4 Thomas Crombie Schelling and Robert John Aumann (from left to right)

the real economy, and laid the foundation for the general theory of asymmetric information markets (Fig. 7.3).

In 2005, the Prize was awarded to Thomas Crombie Schelling of the University of Maryland and Robert John Aumann of the Hebrew University of Jerusalem. Their contribution is to analyze the mechanism of conflict and cooperation through game theory, which is used to study different types of conflicts, trade disputes, price disputes and long-term cooperation modes in society (Fig. 7.4).

In 2007, they were awarded to Leonid Hurwicz from the University of Minnesota, Eric S. Maskin from Princeton University and Roger B. Myerson from the University of Chicago. (Roger B. Myerson) Nobel Prize in Economics. The research of these three people laid the foundation for mechanism design theory (Fig. 7.5).

In 2012, the Nobel Prize in Economics was awarded to American economists Alvin E. Roth and Lloyd S. Shapley. Their main contribution is the creation of the theory of “stable distribution” and the practice of “market design”. Through game theory, Shapley studied and compared multiple pairing models, and proposed a Gale-Shapley algorithm that can theoretically obtain the optimal solution. Roth applied Shapley’s results to the real market and launched a series of empirical studies (Fig. 7.6).

Fig. 7.5 Leonid Hewicz, Eric S. Maskin and Roger B. Myerson (from left to right)

Fig. 7.6 Alvin E. Roth and Lloyd Shapley (from left to right)

Fig. 7.7 French economist Jean Tirole

In 2014, the French economist Jean Tirole was awarded the Nobel Prize in Economics. His contribution lies in not only solving the problems of industrial organization theory and conspiracy through the use of game theory, but also innovating in regulatory theory (Fig. 7.7).

7.2 Non-cooperative Game

7.2.1 Complete Information Game

1. **Complete information static game**

A static game with complete information means that players in each game make decisions at the same time, or not at the same time, the latter actors do not know what specific actions the first actor took, and each player in the game has complete information of the characteristics, the strategy space and revenue function of the other.

(1) Payoff matrix. The payoff matrix is a matrix that includes players, strategy sets, and revenue values. Table 7.1 shows the payoff matrix of the classic game "Prisoner's Dilemma".

Suspects A and B are the two players in the game. Suspect A can choose two strategies, namely, confession and non-confession, and these two strategies are vertically on the left side of the payoff matrix; Suspect B can also choose two strategies, confession and non-confession, and these two strategies are horizontally in the matrix. The numbers in parentheses in the matrix represent the revenue value of the game, which represents the number of years sentenced in this game. The number before the comma in the brackets represents the payoff of Suspect A, and the number after the comma in the brackets represents the payoff of Suspect B.

(2) Nash equilibrium. Nash Equilibrium refers to it when the players participating in the game have no incentive to change their strategy. In the "Prisoner's Dilemma" game, when Suspect A chooses a confession strategy, it is the best strategy for Suspect B to choose confession; when Suspect B chooses a confession strategy, confession is also Suspect A's optimal strategy. For these two players, when they have no motivation to change their strategy, Suspect A chooses to confess, and Suspect B also chooses to confess, which reaches Nash equilibrium of the "Prisoner's Dilemma".

(3) The approach to game equilibrium

① Horizontal line method. The horizontal line method means that when the two numbers in the brackets of the payoff matrix have horizontal lines below the two numbers, the strategy corresponding to the payoff is the equilibrium of the game. The game equilibrium of solving the "Prisoner's Dilemma" by the method of marking is shown in Table 7.2. The equilibrium of the "Prisoner's Dilemma" can be found by drawing a horizontal line as confession, confession.

Table 7.1 Payoff Matrix of the "Prisoner's Dilemma" Game

		Suspect B	
		Confess	Not confess
Suspect A	Confess	(3, 3)	(0, 5)
	Not confess	(5, 0)	(1, 1)

Table 7.2 The horizontal line method of the Prisoner's Dilemma game

		Suspect B	
		Confess	Not confess
Suspect A	Confess	(3, 3)	(0, 5)
	Not confess	(5, 0)	(1, 1)

Table 7.3 The payoff matrix of the Palm or Back game

		B	
		Palm	Back
A	Palm	(1, −1)	(−1, 1)
	Back	(−1, 1)	(1, −1)

② The mixed strategy equilibrium. When the game equilibrium cannot be found by the horizontal line method, it can be solved by a mixed strategy. Mixed strategy equilibrium means that the players in the game choose strategies based on a certain probability, and make no difference when other players choose each strategy. Take the Palm or Back game as an example. The rules are as follows: when A and B choose palm or back at the same time, A wins, A gets 1, and B gets −1; when one chooses palm and the other chooses back, B wins and gets 1, and A gets a gain of −1. The payoff matrix of this game is shown in Table 7.3 to find the equilibrium solution of this game.

Example 7.1

The Palm or Back game cannot find an equilibrium solution through the "horizontal line method". The following is solved by a mixed strategy.

Solution:

Suppose that the probability that A chooses "palm" is $p_1^{(palm)}$, the probability of choosing "back of hand" is $p_1^{(back)}$, and $p_1^{(palm)} + p_1^{(back)} = 1$. Then,

The payoff of B choosing the palm is: $(-1)*p_1^{(palm)} + 1*p_1^{(back)}$;

The payoff of B choosing the back is: $1*p_1^{(palm)} + (-1)*p_1^{(back)}$.

In order to make no difference in the payoff of B when choosing the palm or the back, combine the two formulas to obtain.

$$(-1) * p_1^{(palm)} + 1 * p_1^{(back)} = 1 * p_1^{(palm)} + (-1) * p_1^{(back)}$$
$$p_1^{(palm)} + p_1^{(back)} = 1$$

Solve the equations to get.

$$p_1^{(palm)} = p_1^{(back)} = 1/2.$$

Similarly, suppose that the probability that B chooses “palm” is $p_2^{(palm)}$, The probability of choosing “back of hand” is $p_2^{(back\ of\ the\ hand)}$, and $p_2^{(palm)} + p_2^{(back\ of\ the\ hand)} = 1$. Then:

The payoff of A’s choice of palm is: $1*p_2^{(palm)} + (-1)*p_2^{(back\ of\ the\ hand)}$;

The payoff of A’s choice of the back of the hand is: $(-1)*p_2^{(palm)} + 1*p_2^{(back\ of\ the\ hand)}$.

In order to make no difference in the payoff of A when choosing the palm and the back of the hand, the two methods are combined to obtain.

$$1 * p_2^{(palm)} + (-1) * p_2^{(back\ of\ the\ hand)} = (-1) * p_2^{(palm)} + 1 * p_2^{(back\ of\ the\ hand)}$$
$$p_2^{(palm)} + p_2^{(back\ of\ the\ hand)} = 1.$$

Solve the equations to get.

$$p_2^{(palm)} = p_2^{(back)} = 1/2.$$

Therefore, the mixed strategy equilibrium of the Palm or Back game is: A and B both randomly choose the two strategies “palm” and “back” with the probability of each.

③ Dominant strategy and equilibrium. Dominant strategy refers to the fact that when players in the game choose a strategy, the return value of a strategy is better than other strategies. Conversely, no matter what kind of strategy other players use, the profit of a player using a certain strategy is always less than or equal to the profit obtained when choosing other strategies. The strategy with small profit is called the dominant strategy of the game.

In the “Prisoner’s Dilemma”: No matter what strategy Suspect B chooses, the best strategy for Suspect A is confession. Therefore, “confession” is the dominant strategy of Suspect A. The result of the game caused by Suspect A’s choice of “confessing” strategy is separated from the game payoff matrix, as shown in Table 7.4. In the same way, the optimal strategy for Suspect B is also confession. Therefore, Suspect B’s choice of “confession” strategy is separated from Table 7.4, and the final result is shown in Table 7.5. The equilibrium result is that Suspect A confessed and Suspect B confessed.

Table 7.4 Payoff Matrix (1)

		Suspect B	
		Confess	Not confess
Suspect A	Confess	(3, 3)	(0, 5)

Table 7.5 Payoff Matrix (2)

		Suspect B
		Confess
Suspect A	Confess	(3,3)

Dynamic games with complete information A dynamic game with complete information means that the players in the game have a complete understanding of the game information, and there is a sequence of their actions.

(1) Game tree

Game tree is a representation method of dynamic game, usually the game represented by game tree is called Extensive Form Game. The game tree has 5 components:

① Players.
② Order of action. In a dynamic game, there is a sequence of actions of players in the game.
③ Strategy set: all possible strategies that players can take.
④ Payoff function. The relationship between the strategy adopted by the player and the payoff.
⑤ Information Set: The set of information that players understand during the game.

(2) The construction method of the game tree. Since there is a sequence of actions of players in the game, the order of nodes in the game tree from top to bottom is used to indicate the order in which players make decisions. It should be noted that when constructing the game tree, only a path from top to bottom can be used, and a cycle cannot be formed. The following will take the "real estate development" game as an example to construct a game tree.

Example 7.2 Suppose two real estate developers A and B decide to develop an office building in the same area. If they both develop, the two will compete with each other, which will result in a part of their vacant houses, so they will both lose 5 million; when there is only one developer, the office building can be sold all, with a profit of 5 million, and the other has 0; when the two neither develop, both have 0. Assume that Developer A makes a decision first, and Developer B makes a decision after seeing A's decision. The game tree structure is shown in Fig. 7.8.

Solution:

(3) Sub-game. Sub-game is a branch of the original game tree and can independently form a game. Take the "real estate development" game as an example. The part framed by a dotted line in Fig. 7.9 is called a sub-game. It should be noted that the starting node of a sub-game must be one and cannot be the starting node of the original game tree; some games do not have sub-games and some games contain multiple sub-games.

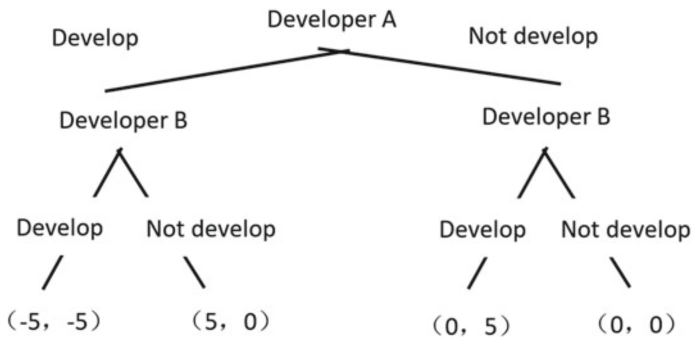


Fig. 7.8 Game tree of "real estate development" game

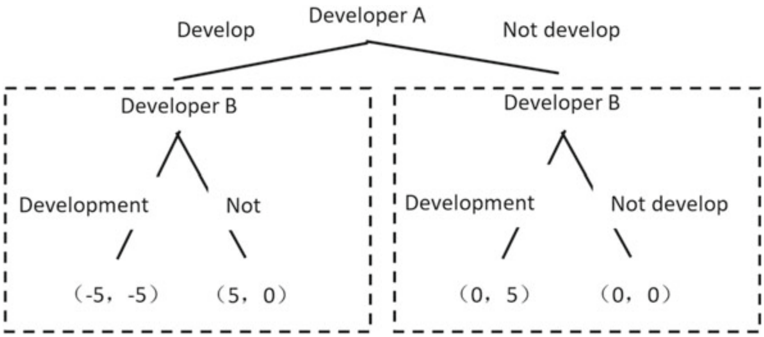


Fig. 7.9 The sub-game of the "real estate development" game

(4) Reverse induction method. Reverse induction is a method of solving dynamic games. The following will take the "real estate development" game as an example to use reverse induction to find the equilibrium solution of the dynamic game.

Example 7.3 Step 1: Start the analysis from the last subgame.

When Developer A chooses to develop, Developer B chooses to develop and get the payoff of −5, So B chooses not to develop. In the same way, when Developer A chooses not to develop, B chooses to develop with payoff 5, not to develop with payoff 0, so Developer B chooses to develop. As shown in Fig. 7.10a.

Step 2: Continue to analyze the actions of the players in the next stage.

At this time, we will continue to analyze the actions of Developer A. When A chooses to develop and B chooses not, A's payoff is 5; when A chooses not and B chooses to develop, A's payoff is 0. Therefore, Developer A chooses to develop. As shown in Fig. 7.10b.

Step 3: Find the equilibrium of the game.

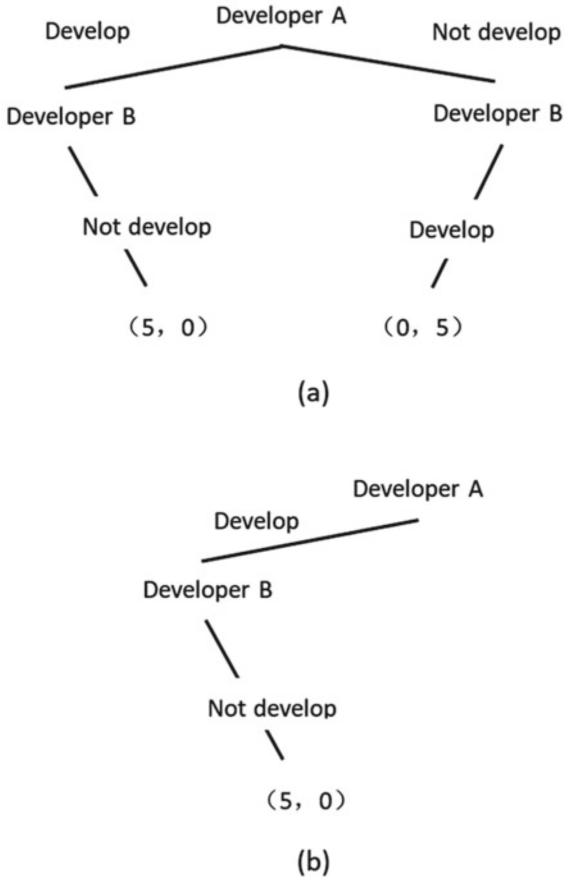


Fig. 7.10 Diagram of reverse induction

According to the first two steps, the equilibrium of the game is obtained: A chooses to develop, and B chooses not to develop.

7.2.2 Incomplete Information Game

1. **Static Game with Incomplete Information**
   (1) Cournot oligarch game with incomplete information. The Cournot oligarch game is a type of static game, which assumes that the players in the two games clearly know the cost function of themselves and the opponent. However, the Cournot oligopoly game with incomplete information assumes that player 1 clearly knows his cost function, but not know the cost function of player 2 except its probability. Player 2 clearly knows his own cost function and player 1's cost function. The following will take

the Cournot oligarch game with incomplete information as an example to understand the static game with incomplete information and its solution process.

Example 7.4 In the Cournot oligopoly game, assume that Players 1 and 2 in the two games are Manufacturers 1 and 2, and they compete for output in the market. The total output on the market is $Q = q_1 + q_2$, the demand function of the market is $P = A - Q = A - (q_1 + q_2)$. Here Q is the total output, q_1 is the output of Manufacturer 1, q_2 is the output of Manufacturer 2, P is the market price, and A is the exogenous constant. Suppose the cost function of Manufacturer 1 is $C(q_1) = cq_1$, where c is the exogenous constant. Assume that the cost function of Manufacturer 2 may be $C(q_2) = c_A q_2$, or maybe $C(q_2) = c_B q_2$. Here c_A and c_B are exogenous constants, and $c_A > c_B > 0$. The cost function of Manufacturer 2 is $C(q_2) = c_A q_2$, the probability is θ, the cost function of Manufacturer 2 is $C(q_2) = c_B q_2$, the probability is $1 - \theta$. Solve the optimal output of the two manufacturers.

Solution:

Since Manufacturer 2 knows the cost function of itself and Manufacturer 1, the decision-making process of Manufacturer 2 is consistent with the Cournot game solution method with complete information.

First, Manufacturer 2 regards the output q_1 as a given, and calculates the profit of Manufacturer 2 according to the two uncertain cost functions of Manufacturer 2 respectively.

When the cost function of Manufacturer 2 is $C(q_2) = c_A q_2$, the profit function of Manufacturer 2 is

$$\pi_2^A = Pq_2 - C(q_2) = (A - q_1 - q_2)q_2 - c_A q_2 \tag{7.1}$$

When the cost function of Manufacturer 2 is $C(q_2) = c_B q_2$, the profit function of Manufacturer 2 is

$$\pi_2^B = Pq_2 - C(q_2) = (A - q_1 - q_2)q_2 - c_B q_2 \tag{7.2}$$

Respectively calculate the partial derivative of Eqs. (7.1) and (7.2) with respect to q_2 and set it to zero to obtain

$$\frac{\partial \pi_2^A}{\partial q_2} = -q_2 + (A - q_1 - q_2) - c_A = 0 \tag{7.3}$$

$$\frac{\partial \pi_2^B}{\partial q_2} = -q_2 + (A - q_1 - q_2) - c_B = 0 \tag{7.4}$$

From Eqs. (7.3, 7.4), we can get

$$q_2^A = (A - c_A - q_1)\big/ 2 \tag{7.5}$$

$$q_2^B = (A - c_B - q_1)\big/ 2 \tag{7.6}$$

Secondly, analyze the profit function of Manufacturer 1. When in a complete information game, the profit function of Manufacturer 1 is

$$\pi_1 = Pq_1 - C(q_1) = (A - q_1 - q_2)q_1 - cq_1$$

However, since Manufacturer 1 does not know the exact information of Manufacturer 2, he can only make decisions according to the expected cost function of Manufacturer 2. Because Manufacturer 1 knows that the probability of Manufacturer 2 selecting output q_2^A is θ, the probability of selecting output q_2^B is 1 $1-\theta$. Therefore, in the game of incomplete information, the profit function of Manufacturer 1 is

$$\pi_1 = \theta\left[\left(A - q_1 - q_2^A\right)q_1 - cq_1\right] + (1-\theta)\left[\left(A - q_1 - q_2^B\right)q_1 - cq_1\right] \tag{7.7}$$

Taking the partial derivative of Eq. (7.7) with respect to q_1 and making it zero, we get

$$q_1 = \left[A - \theta q_2^A - (1-\theta)q_2^B - c\right]\big/ 2 \tag{7.8}$$

Finally, the Eqs. (7.5), (7.6) and (7.8) are combined to solve the linear equation of three variables, and finally the optimal output is obtained

$$q_1 = [A - 2c + \theta c_A + (1-\theta)c_B]\big/ 3$$

$$q_2^A = [2A + 2c - (3+\theta)c_A - (1-\theta)c_B]\big/ 6$$

$$q_2^B = [2A + 2c - \theta c_A - (4-\theta)c_B]\big/ 6$$

(2) Harsanyi transformation. Harsanyi transformation is a method of solving games with incomplete information. He finds an equilibrium solution by transforming the game into a game tree. The steps of Harsanyi transformation are as follows:

Step 1: Introduce "nature" as the first player in the game to make a decision. "Nature" determines the type of player i in each game.

Step 2: Each player i in the game chooses a strategy from its own strategy space $S_i (i = 1, 2, \ldots, n)$.

Step 3: Determine the expected return of each game participant at the end of the game tree according to the revenue function.

The following will use an example to introduce the Harsanyi transformation.

Table 7.6 Payoff matrix when Enterprise 2 is an inefficient enterprise

		Enterprise 2	
		Strategy U	Strategy V
Enterprise 1	Strategy L	(−10, 10)	(5, 5)
	Strategy R	(0, 20)	(0,15)

Table 7.7 Payoff matrix when Enterprise 2 is an inefficient enterprise

		Enterprise 2	
		Strategy U	Strategy V
Enterprise 1	Strategy L	(−10, −10)	(5, 5)
	Strategy R	(0, 10)	(0,15)

Example 7.5 Suppose there are two enterprises in the market, Enterprises 1 and 2. Enterprise 2 may be an "efficient" or "inefficient" enterprise. Enterprise 2 knows its type, but Enterprise 1 does not know what type of Enterprise 2 is, and only knows that the probability of Enterprise 2 being an "efficient" enterprise is p, the probability of an "inefficient" enterprise is $1 - p$. The payoff matrix for Enterprise 2 being an efficient enterprise or an inefficient enterprise is shown in Table 7.6 and Table 7.7, respectively.

In accordance with the steps of Harsanyi transformation, the analysis is as follows:

Solution:

Step 1: The introduction of "nature" can be understood as "nature" acts first, and then Enterprises 1 and 2 act simultaneously, as shown in Fig. 7.11.

Step 2: As shown in Fig. 7.11, Enterprise 1 has one information set and Enterprise 2 has two information sets. For Enterprise 1, he can choose Strategy L or Strategy R. For Enterprise 2, when Enterprise 2 is an efficient enterprise, it can choose the

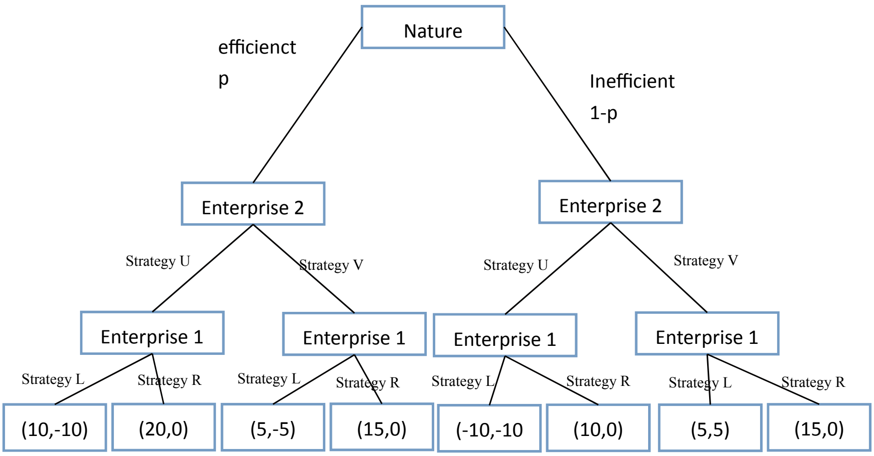


Fig. 7.11 Game tree

Strategy $U_{Efficienct}$ or the Strategy $V_{Efficienct}$; When Enterprise 2 is an inefficient enterprise, it can choose Strategy $U_{Inefficient}$ or Strategy $V_{Inefficient}$. So there are eight possible game outcomes:

((Strategy $U_{Efficienct}$, Strategy $U_{Inefficient}$), Strategy L).
((Strategy $U_{Efficienct}$, Strategy $U_{Inefficient}$), Strategy R).
((Strategy $U_{Efficienct}$, Strategy $V_{Inefficient}$), Strategy L).
((Strategy $U_{Efficienct}$, Strategy $V_{Inefficient}$), Strategy R).
((Strategy $V_{Efficienct}$, Strategy $U_{Inefficient}$), Strategy L).
((Strategy $V_{Efficienct}$, Strategy $U_{Inefficient}$), Strategy R).
((Strategy $V_{Efficienct}$, Strategy $V_{Inefficient}$), Strategy L).
((Strategy $V_{Efficienct}$, Strategy $V_{Inefficient}$), Strategy R).

According to the comparison of the payoff, it is found that when Enterprise 2 is an efficient enterprise, the return of choosing Strategy $U_{Efficienct}$ (when Enterprise 1 chooses Strategy L, the payoff is 10; when Enterprise 1 chooses Strategy R, the payoff is 20) is higher than the payoff of choosing Strategy $V_{Efficienct}$ (when Enterprise 1 chooses Strategy L, the payoff is 5; when Enterprise 1 chooses Strategy R, the payoff is 15); When Enterprise 2 is an inefficient enterprise, the payoff of choosing Strategy $V_{Inefficient}$ (when Enterprise 1 chooses Strategy L, the payoff is 5; when Enterprise 1 chooses Strategy R, the payoff is 15) is higher than the payoff of choosing Strategy $U_{Inefficient}$ (when Enterprise 1 chooses the payoff for Strategy L is −10; when Enterprise 1 chooses Strategy R, the payoff is 10). Therefore, Enterprise 2 will only choose (Strategy $U_{Efficienct}$, Strategy $V_{Inefficient}$) as its strategy.

Step 3: Calculate the expected payoff. Because Enterprise 1 does not know whether Enterprise 2 chooses Strategy $U_{Efficienct}$ or Strategy $V_{Inefficient}$, Enterprise 1 can only choose prior probability (or prior belief, a common concept in Bayesian statistics) to calculate the expected payoff.

When Enterprise 1 chooses a Strategy L, the expected payoff is $p \times (-10) + (1-p) \times 5 = 5 - 15p$;

When Enterprise 1 chooses Strategy R, the expected payoff is $p \times 0 + (1-p) \times 0 = 0$.

Therefore, the two expected payoff equations are combined $5 - 15p = 0$ to obtain $p = 1/3$. The final game result can be obtained, which is the Bayesian-Nash equilibrium. The results are as follows:

When $p < 1/3$, the expected payoff of Enterprise 1 choosing Strategy L is greater than the expected payoff of choosing Strategy R; the final game result is ((Strategy $U_{Efficienct}$, Strategy $V_{Inefficient}$), Strategy L).

When $p > 1/3$, The expected payoff of Enterprise 1 choosing Strategy L is less than the expected payoff of choosing Strategy R; the final game result is ((Strategy $U_{Efficienct}$, Strategy $V_{Inefficient}$), Strategy R).

When $p = 1/3$, The expected payoff of Enterprise 1 choosing Strategy L is equal to the expected payoff of choosing Strategy R. The final game result is ((Strategy $U_{Efficienct}$, Strategy $V_{Inefficient}$), Strategy L or Strategy R).

2. **Dynamic game with incomplete information**

(1) Equilibrium of dynamic game with incomplete information. The dynamic game with incomplete information is similar to the static game with incomplete information, and the equilibrium solution of the game can also be obtained through the Harsanyi transformation.

The dynamic game equilibrium with incomplete information should have the following characteristics:

① The players participating in the game have subjective beliefs at each game node.
② The equilibrium must satisfy the decision of the players in the game on each information set of the game, and players' decisions are optimal.

Example 7.6 Two enterprises participate in the game, and Enterprise 1 acts first, Enterprise 2 acts later. Enterprise 1 may choose Strategy L or Strategy R. Enterprise 2 does not know the choice of Enterprise 1, and only has a priori probability about the choice of Enterprise 1. Enterprise 2 thinks that the probability of Enterprise 1 choosing Strategy *L* is *p*, and the probability of choosing Strategy *R* is *1-p*. The payoff matrix is shown in Table 7.8.

Solution:

In order to facilitate analysis, draw the game tree of the game according to the payoff matrix, as shown in Fig. 7.12. According to the prior probability of Enterprise 2,

The expected payoff of Enterprise 2 choosing Strategy U is $p \times 4 + (1 - p) \times 5 = 5 - p$;

Table 7.8 Payoff matrix

		Enterprise 2	
		Strategy U	Strategy V
Enterprise 1	Strategy L	(3, 4)	(0, 2)
	Strategy R	(1, 5)	(0, 3)

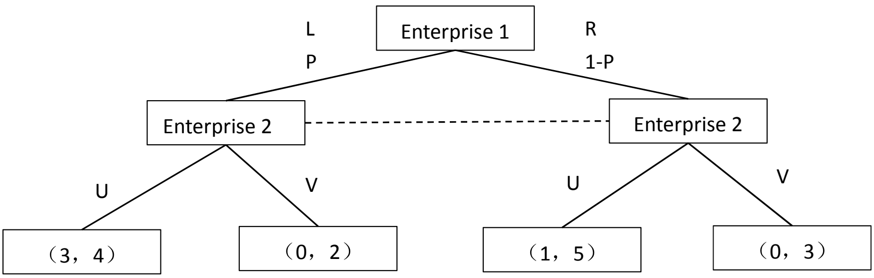


Fig. 7.12 Game tree

The expected payoff of Enterprise 2 choosing *Strategy V* is $p \times 2 + (1 - p) \times 3 = 3 - p$。

It can be seen that for any $0 \leq p \leq 1$, when choosing Strategy U, the expected payoff of Enterprise 2 is higher, so Enterprise 2 will choose Strategy U. And Enterprise 1 anticipates the choice logic of Enterprise 2, so rational Enterprise 1 will choose Strategy L, because Enterprise 1 has the largest payoff (3), so the equilibrium is (Strategy L, Strategy U).

(2) Refined Bayesian Nash Equilibrium (BNE). Refined BNE, as the name suggests, is the refinement of BNE, which eliminates the "empty threats" in the BNE.

Example 7.7 Continuing the analysis following the previous example, add Strategy M of Enterprise 1 to this example. The payoff matrix is shown in Table 7.9, and the game tree is shown in Fig. 7.13.

Solution:

For any $0 \leq p \leq 1$, Enterprise 2 will choose Strategy U. Therefore, Enterprise 1 knows that if it does not choose Strategy M, Enterprise 2 must choose Strategy U, and Enterprise 1 will choose Strategy L accordingly.

Table 7.9 Payoff matrix of stereo 7.7

		Enterprise 2	
		Strategy U	Strategy V
Enterprise 1	Strategy L	(3, 4)	(0, 2)
	Strategy R	(1, 5)	(0, 3)
	Strategy M	(1, 20)	(1, 20)

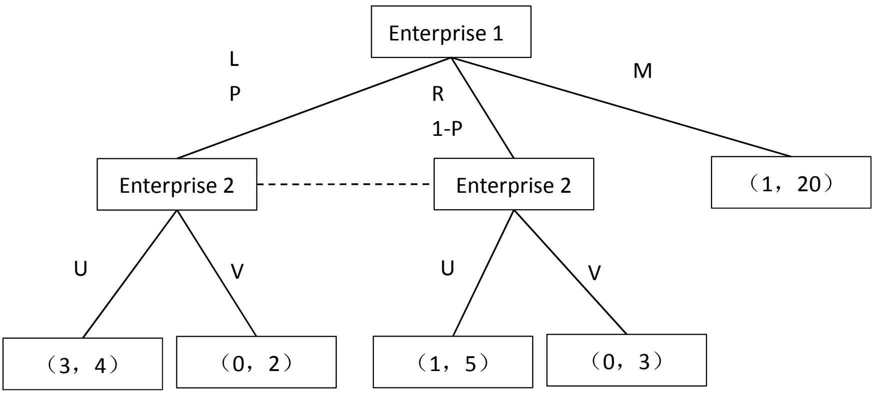


Fig. 7.13 The game tree of Example 7.7(1)

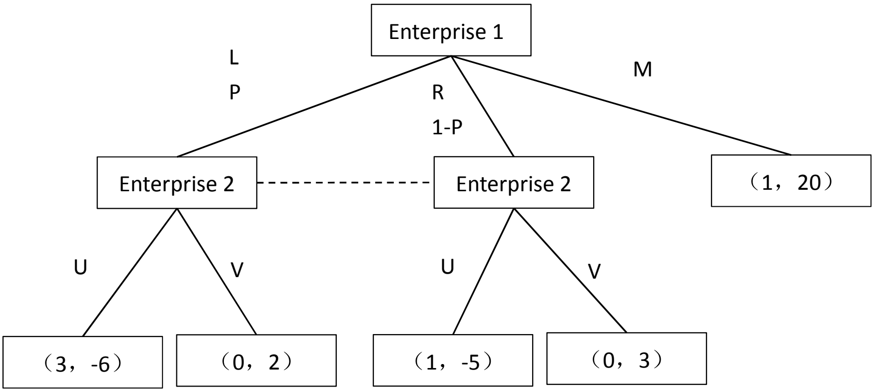


Fig. 7.14 The game tree of Example 7.7(2)

And the most desired result of Enterprise 2 is that Enterprise 1 chooses Strategy M, and it can get payoff 20, thus ending the game. In order to achieve its goal, Enterprise 2 finds an impartial organization and makes a promise: if Enterprise 1 does not choose Strategy M, Enterprise 2 will donate 10 if it chooses Strategy U. At this time, the game payoff has changed, as shown in Fig. 7.14. According to the prior probability of Enterprise 2,

The expected payoff of Enterprise 2 choosing Strategy U is $p \times (-6) + (1-p) \times (-5) = -5 - p$;

The expected payoff of Enterprise 2 choosing Strategy V is $p \times 2 + (1-p) \times 3 = 3 - p$。

It can be seen that for any $0 \leq p \leq 1$, when choosing Strategy U, the expected payoff of Enterprise 2 is higher, so Enterprise 2 chooses Strategy V. At this time, Enterprise 1 predicts the choice of Enterprise 2, and Enterprise 2 turns the threat of its "inevitable choice of Strategy V" into a credible threat. Therefore, Enterprise 1 will choose Strategy M instead of Strategy L (because the payoff of Enterprise 1 is 1, greater than the benefits of the other two strategies). So the final equilibrium is (Strategy M, Strategy V), and this Nash equilibrium is also a refined BNE.

7.3 Cooperative Game

7.3.1 Basic Concept

Cooperative game refers to a game in which players can jointly reach a binding and enforceable agreement. The cooperative game mainly emphasizes efficiency, justice, fairness and collective rationality.

1. **Concepts of player, coalition, characteristic function**

(1) Players

Players refer to the parties involved in decision-making in the game. Usually, all players in the game gather as $N = \{1, 2, \ldots, n\}$, where i represents the i-th player in the game($i \in N$).

(2) Coalition

A coalition refers to a collection of some or all players in the game. If the set consisting of all subsets of N is $\mathrm{P}(N)$, then any element in $\mathrm{P}(N)$ is an alliance.

(3) Characteristic function

The form of the characteristic function is defined as a two-tuple, (N, v), Characteristic Function $v : \mathrm{P}(N) \to \mathbb{R}$, satisfying $v(\emptyset) = 0$。

2. **The premise of the cooperative game**

Definition 7.1

If a cooperative Strategy (N, v), satisfies.

$v(S \cup T) \geq v(S) + v(T), \forall S, T \in \mathrm{P}(N), S \cap T = \emptyset.$

It is said that (N, v) satisfies the super-additivity.

Super-additivity shows that when two disjoint coalitions cooperate, the benefits obtained are greater than the sum of benefits when acting alone. Super-additivity is also the premise of cooperative games. If the cooperative game does not satisfy super-additivity, the basis of cooperation will be lost.

7.3.2 Distribution of Cooperative Games

The distribution of the cooperative game refers to how the players participating in the cooperation distribute the final profit, that is, the solution of the cooperative game.

1. **Definition of allocation**

Definition 7.2 For cooperative games $(N, v) \in G^N$, if the n-dimensional vector $x = (x_1, x_2, \ldots, x_n)$ meets the following two conditions.

(1) $x_i \geq v(\{i\}), i = 1, 2, \cdots, n$
(2) $\sum\limits_{i \in N} x_i = v(N)$

Then call x the distribution of (N, v), where x_i represents the distribution of player i in the game, $i = 1, 2, \ldots, n$.

Condition (1) is called the individual rationality condition, that is, for each player i, if the final distribution of payoff is less than the payoff when acting alone, the player will not accept this distribution. Condition (2) is called the group rationality condition. First, all $v(N)$ in the coalition must be distributed; secondly, all distributions cannot exceed the total payoff $v(N)$, otherwise the players in the game will not accept this distribution.

2. **Solving the cooperative game**

At present, there are generally two methods of solving classical cooperative games: dominant solution and valuation solution. The concepts of dominant solution mainly include core and stable set, and the concepts of valuation solution mainly include Shapley value and Banzhaf-Coleman force index. The following will focus on the core and Shapley values.

(1) Core

Definition 7.3 Let x and y be two distributions of cooperative game (N, v), and $S \in \mathrm{P}(N)$ is an coalition. If it satisfies.

- $x_i > y_i, \forall i \in S$
- $\sum_{i \in S} x_i \leq v(S)$,

It is called x dominates y on S, denoted as: $x \succ_S y$. If there is a certain $S \in \mathrm{P}(N)$ for x and y such that $x \succ_S y$, then x is said to dominate y, which is recorded as $x \succ y$.

Definition 7.4 In the cooperative strategy (N, v), all those that are allocated and not dominated are called the core of (N, v), denoted as $C(N, v)$.

Theorem 7.1 Let (N, v) be a cooperative game with n people, $x \in \mathbb{R}^n$, $x \in C(N, v)$, if and only if the following two formulas hold.

- $\sum_{i \in S} x_i \geq v(S), \forall S \in \mathrm{P}(N)$
- $\sum_{i \in N} x_i = v(N)$.

The core is defined from the perspective of distribution dominance. The concept is intuitive and simple. It not only satisfies individual rationality, but also satisfies collective rationality. If the core is not empty, no coalition can raise ambiguities about the distribution in the core, so the core has a strong stability. The core

$C(N, v)$ can be expressed as.

$$C(N, v) = \left\{ x \in \mathbb{R}_+^n \mid \sum_{i \in S} x_i \geq v(S), \sum_{i \in N} x_i = v(N), \forall S \in \mathrm{P}(N) \right\}$$

Example 7.8 The hypothetical United Nations Security Council's five permanent members voted, and it passed if more than two votes are in favor. The characteristic function of the game is.

$v(1, 2, 3) = v(1, 2, 4) = v(1, 2, 5) = v(1, 2, 3, 4) = v(1, 2, 3, 5) = v(1, 2, 4, 5) = v(1, 2, 3, 4, 5) = 1$, and for all other S, $v(S) = 0$. Solve the core of the game.

Solution:

Apply the theorem, there are $\sum_{i=1}^{5} x_i = 1$, for each coalition $x_i \geq 0$, $i = 1, 2, 3, 4, 5$, $x_1 + x_2 + x_3 \geq 1$, $x_1 + x_2 + x_4 \geq 1$, $x_1 + x_2 + x_5 \geq 1$, so the core is $C(v) = \{(a, 1 - a, 0, 0, 0) : 0 \leq a \leq 1\}$。

(2) Shapley value

The Shapley value is an important valuation solution for cooperative games. This value has good operability and analytical rationality in mathematical calculations. Therefore, the solution of Shapley value is regarded as the most useful solution for cooperative games.

In 1953, Shapley put forward three reasonable hypotheses.

Definition 7.5 Assume that $(N, v) \in G^N$, if the n-dimensional vector $\varphi(N, v)$ satisfies the following three axioms, then the vector is called the Shapley value $\varphi(N, v) = (\varphi_1(N, v), \varphi_2(N, v), \cdots \varphi_n(N, v))$ of the cooperative game (N, v) on G^N, or the Shapley function of upper.

Axiom 1 (validity) If $T \in C(N|v)$, then $\sum_{i \in T} \varphi_i(N, v) = v(T)$,

Axiom 2 (symmetry) Let π be any permutation of N, then $\varphi_i(N, \pi v) = \varphi_i(N, v)$,

Axiom 3 (additivity) Arbitrary Strategy (N, v_1) and (N, v_2),any coalition $S \subseteq N$, the countermeasure $(N, v_1 + v_2)$ is defined as $(v_1 + v_2)(S) = v_1(S) + v_2(S)$, then there is $\varphi_i(N, v_1 + v_2) = \varphi_i(N, v_1) + \varphi_i(N, v_2)$ for all $i \in N$.

Theorem 7.2 Arbitrary cooperative game $(N, v) \in G^N$, there is the only Shapley value that satisfies the three axioms $Sh(N, v) = (Sh_1(N, v), Sh_2(N, v), \cdots Sh_n(N, v))$, the specific form is.

$$Sh_i(N, v) = \sum_{S \subseteq N \backslash \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[v(S \cup \{i\}) - v(S)], \forall i \in N \quad (7.9)$$

In the above formula, $v(S \cup \{i\}) - v(S)$ represents the marginal contribution of player i to coalition S, and $\gamma = \frac{|S|!(|N|-|S|-1)!}{|N|!}$ represents the probability of appearance of coalition S. The Shapley value is to distribute the payoff of the coalition according to the marginal contribution. When the characteristic function of the classical cooperative game satisfies the convexity, this cooperative game is also called convex game, and the cooperative game must satisfy super-additivity. The Shapley value is included in the core, which reflects the stability of the distribution.

Example 7.9 Suppose there are three Enterprises A, B, and C cooperating, and their payoff functions are respectively $v(1)$, $v(2)$ and $v(3)$. If three enterprises operate separately, they can make a profit $v(1) = v(2) = v(3) = 1$; A, B coalition can be profitable $v(1, 2) = 4$; A, C coalition can be profitable $v(1, 3) = 3$; B, C coalition can be profitable $v(2, 3) = 2$; A, B, C work together to make profits $v(1, 2, 3) = 5$Please use the Shapley value method to distribute the profits of the three enterprises.

Solution:
Enterprise A.
First, find the value of γ. Using the formula $\gamma = \frac{|S|!(|N|-|S|-1)!}{|N|!}$, $|S|$ is the number of enterprises in coalition S, and $|N|$ is the number of all enterprises participating in the cooperation, the γ value corresponding to different coalition S can be obtained. The result of γ value is shown in Table 7.10.

Secondly, according to formula 7.9, calculate the profit distributed by Enterprise A. It can be seen from Table 7.11 that the profit distributed by Enterprise A is $\varphi_1 = 1 \times 1/3 + 3 \times 1/6 + 2 \times 1/6 + 3 \times 1/3 = 13/6$.

(2) Enterprise B

Similarly, from Table 7.12, the profit distributed by Enterprise B can be obtained as $\varphi_2 = 1 \times 1/3 + 3 \times 1/6 + 1 \times 1/6 + 2 \times 1/3 = 5/3$。

(3) Enterprise C

Table 7.10 The value of γ under different coalitions of Enterprise A

S	$\|S\|$	$\|N\|$	γ
$\{\phi\}$	0	3	$0!(3-0-1)!/3! = 1/3$
$\{2\}$	1	3	$1!(3-1-1)!/3! = 1/6$
$\{3\}$	1	3	$1!(3-1-1)!/3! = 1/6$
$\{2, 3\}$	2	3	$2!(3-2-1)!/3! = 1/3$

Similarly, from Table 7.13, the profit distributed by Enterprise C can be obtained as $\varphi_3 = 1 \times 1/3 + 2 \times 1/6 + 1 \times 1/6 + 1 \times 1/3 = 7/6$.

Therefore, Companies A, B, and C respectively get benefits $13/6$,$5/3$ and $7/6$.

3. **Cooperative Game of Coalition Structure**

(1) Basic concepts

In a game with a coalition structure, the payoff distribution of each player is not only related to the coalition he is in, but also related to other coalitions. The definition of the coalition structure is given below.

Definition 7.6 A coalition structure on the set N of a limited game is a two-tuple (N, Γ), where $\Gamma = \{B_1, B_2, \ldots B_m\}$, $1 \leq m \leq n$, satisfy.

- $\cup_{k=1}^{m} B_k = N$
- $B_k \cap B_l = \emptyset, \forall k, l \in \{1, 2, \ldots, m\}, k \neq l$, then B_k is called a structural coalition.
- the value of Owen

Definition 7.7 The game of Owen coalition structure is generally represented by a triple $\langle N, v, \Gamma \rangle$, where N is the set of participants, (N, v) is the cooperative game, and (N, Γ) is the coalition structure.

Owen value can be regarded as the promotion of Shapley value in the situation of cooperative game with coalition structure. First, the value of the characteristic function of the largest coalition is allocated among the coalitions; secondly, on

Table 7.11 The marginal contribution of Enterprise A under different coalitions

γ	S	$v(s \cup \{i\}) - v(s)$
$1/3$	$\{\phi\}$	$1 - 0 = 1$
$1/6$	$\{2\}$	$4 - 1 = 3$
$1/6$	$\{3\}$	$3 - 1 = 2$
$1/3$	$\{2, 3\}$	$5 - 2 = 3$

Table 7.12 Different Coalitions of Enterprise B

γ	S	$v(s \cup \{i\}) - v(s)$
$1/3$	$\{\phi\}$	$1 - 0 = 1$
$1/6$	$\{1\}$	$4 - 1 = 3$
$1/6$	$\{3\}$	$2 - 1 = 1$
$1/3$	$\{1, 3\}$	$5 - 3 = 2$

Table 7.13 Marginal contribution of Enterprise C under different coalitions

γ	S	$v(s \cup \{i\}) - v(s)$
$1/3$	$\{\phi\}$	$1 - 0 = 1$
$1/6$	$\{1\}$	$3 - 1 = 2$
$1/6$	$\{2\}$	$2 - 1 = 1$
$1/3$	$\{1, 2\}$	$5 - 4 = 1$

the basis of the first stage, the allocation value of each coalition structure is allocated twice within the coalition. In 1977, Owen proposed the definition of Owen coalition structure value.

Definition 7.8 Known cooperative game with coalition structure $\langle N, v, \Gamma\rangle$, Among them, $\Gamma = \{B_1, B_2, \ldots, B_m\}$, for any $i \in B_k \in \Gamma$, the Owen coalition structure value of player i is defined as:

$$\varphi_i(N, v, \Gamma) = \sum_{L \subset id(N), k \notin L} \sum_{i \in S \subseteq B_k} \frac{|L|!(m - |L| - 1)!}{m!} \times \frac{||S| - 1|!(|B_k| - |S|)!}{|B_k|!} \times [v(S \cup \Gamma(L)) - v((S \backslash \{i\}) \cup \Gamma(L))],$$

$\Gamma(L) = \cup_{j \in L} B_j.$

7.3.3 Matching Game

Matching game first began with Gale and Shapley (1962) in a study on marriage matching and university admissions. In 2012, the Nobel Prize in Economics was awarded to Alvin Roth and Lloyd Shapley, the founders and promoters of matching game theory, in order to reward the two scholars' contribution on the "stable allocation theory and the market design practice", which won much attention to the matching game.

1. **The basic model of matching game theory**

(1) One-to-one matching. Among the one-to-one matching problems, the most typical one is the marriage matching problem.

Assume that two finite sets $M = \{m_1, m_2, \ldots, m_n\}$ and $W = \{w_1, w_2, \ldots, w_n\}$ respectively represent the set of all unmarried men and unmarried women, $M \cap W = \emptyset$. Every man in M has a strict preference for every woman in W, and every woman in W has a strict preference for every man in M. Participant i's preference relationship can be represented by the symbol "$\succ$" , $x \succ y$ means that for participant i, x dominates y. The preference set of all participants is $P = \{P(m_1), P(m_2), \ldots, P(m_n); P(w_1), P(w_2), \ldots, P(w_p)\}$, then the marriage market can be expressed as $(M, W; P)$.

(2) Many-to-one matching. Many-to-one matching is a generalization of one-to-one matching, which is closer to reality. For example, the labor market, university admissions, etc.

Suppose the hospital and medical students belong to two disjoint sets: $H = \{h_1, h_2, \ldots, h_n\}$ and $S = \{s_1, s_2, \ldots, s_m\}$, There are q_i internship positions in the hospital h_i. Hospital h has strict preference $P(h)$ on set $S \cup \{u\}$, and student s has strict preference $P(s)$ on set $H \cup \{u\}$, where u is the set of matched participants. $h_i P(s) h_k$ indicates that student s strictly prefers hospital h_i, and $h_i R(s) h_k$ indicates that student s either strictly prefers hospital h_i or thinks that there is no difference between hospital h_i and h_k. Because preference is strict preference, only when $i = k$ is indistinguishable. Therefore, the preference vector $P = \{P(h_1), P(h_2), \ldots, P(h_n); P(s_1), P(s_2), \ldots, P(s_m)\}$. The matching result is represented by $\mu : H \cup S \to H \cup S \cup \{u\}$, and $|\mu(s)| = 1, \forall s \in S, |\mu(h_i)| = q_i, \forall h_i \in H$. For any $h \in H, s \in S, \mu(s) = h$ if and only if $s \in \mu(h)$. Each student can only be employed by one hospital at most, and each hospital cannot recruit more students than the number of internship positions.

2. **Deferred-acceptance algorithm (GS algorithm)**

The core task of the matching game is to study whether there is a stable match in the market. Stable matching means that two parties that prefer each other are matched, and the matching objects of each participant are acceptable, and there is no pair of participants who have not matched. In the research of matching stability, the rejection algorithm is an effective method.

Let's take a job search as an example to explain in detail the procedure for rejecting the algorithm. Assuming a given preference P, the steps of the rejection algorithm are as follows:

Step 1: Every job applicant submits a job application to his most preferred company; each company leaves its most preferred application among all the applications it receives and rejects all the others.

Step 2: Every job seeker submits a job application to his most preferred company that has not yet rejected him; each company leaves his most preferred application among all the applications he receives and rejects all other applications.

Step 3: Repeat Step 2 until no rejection occurs, and the algorithm ends.

Step 4: The match produced at the end of the algorithm is the final stable match.

Example 7.10 An enterprise is going to recruit 3 employees to be operations manager, sales manager and purchasing manager. After preliminary screening, 5 applicants are left as manager candidates. After the final round of interviews, these 5 people can voluntarily choose to run for each post and list their priority order for the post. Suppose A1, A2, and A3 respectively represent three positions of operation manager, sales manager and purchasing manager, and B1, B2, B3, B4, and B5 respectively represent 5 candidates. The preference order of companies and candidates is shown in Tables 7.14 and 7.15.

Solution:

(1) Step 1: According to the candidates' preferences, B1 prefers the position of operation manager; B2 and B3 prefer the position of sales manager; B4 and B5 prefer the position of purchasing manager.

According to the company's preference for candidates, the operation manager position prefers B2, followed by B4 and B1. However, since the operation manager is ranked third in the preference order of B2 and B4, therefore, in step 1, the operation manager position is temporarily selected as B1. Similarly, the position of sales manager prefers B3, and the position of purchasing manager prefers B5.

The first round of temporary matching: (A1, B1), (A2, B3), (A3, B5), B2, B4 temporarily lost.

(2) Step 2: Each rejected candidate applies to his second favorite position, B2 applies to A3, B4 applies to A2, and each position chooses the best among the candidates in the second round and the candidates in the previous round.

A1 and B1 still match; A2 chooses between B3 and B4. Because A2 never prefers B4, A2 matches B3; A3 chooses between B2 and B5. Because A3 prefers B2, A3 and B2 match. B2 replaces B5 in step 1.

The second round of temporary matching: (A1, B1), (A2, B3), (A3, B2), and B4, B5 temporarily are lost.

Table 7.14 Order of preference of the enterprise

Post	A1 (operation manager)	A2 (sales manager)	A3 (purchasing manager)
Personnel order	B2	B5	B2
	B4	B3	B5
	B1	B2	B4
	–	B1	–

Table 7.15 The preference order of candidates

Candidate	B1	B2	B3	B4	B5
Job order	A1	A2	A2	A3	A3
	A3	A3	A1	A2	A2
	–	A1	–	A1	–

(3) Step 3: Candidates B4 and B5 who were rejected in the last round apply to their preferred positions in the next round, B4 applies to A1, and B5 applies to A2.

In the same way, A1 chooses B4 from B1 and B4; A2 chooses B5 from B3 and B5.

The third round of temporary matching: (A1, B4), (A2, B5), (A3, B2), and B1, B3 temporarily are lost.

Step 4: B1 applies to A3, and B3 applies to A1.

A1 chooses B4 from B3 and B4; A3 chooses B2 from B2 and B1.

At this time, the rejected candidates B1 and B3 have no chance to apply again, and the procedure ends.

Final match (stable match): (A1, B4), (A2, B5), (A3, B2). See Table 7.16.

7.4 Evolutionary Game

7.4.1 Overview

Evolutionary game theory is a new theory that combines game theory and dynamic evolution process analysis theory. In reality, most people are bounded rationally due to the influence of cognitive ability. Therefore, the evolutionary game is another game in which the assumption of completely rational behavior is extended to the assumption of bounded rational behavior. The two most important concepts in evolutionary games are evolutionary stable strategies and replication dynamics.

John Maynard Smith's "The theory of games and the evolution of animal conflict" published in 1973 marked the birth of evolutionary game theory. Later, in 1974, Smith and Price first proposed the basic concept of evolutionary stability strategy. Until now, many scholars have done a lot of research on evolutionary games. Zhu Qinghua and Dou Yijie (2007) used evolutionary games to study the game relationship between the government and core enterprises. Huang Minmei (2010) used evolutionary game to study the cooperative mechanism of supply chain collaborative product development. Xu Minli et al. (2012) constructed an evolutionary model of food quality input from suppliers and manufacturers in response to frequent food safety incidents. Wang Xiangbing et al. (2012) applied evolutionary games to the study of regional innovation systems, and analyzed the process of cooperative innovation between enterprises and scientific research institutions and the evolution of cooperative innovation models.

Table 7.16 Stable matching

Post	A1 (operation manager)	A2 (sales manager)	A3 (purchasing manager)
Candidate	B4	B5	B2

7.4.2 Evolutionary Stability Strategy (ESS)

Assume that a small group of mutants in the population are prescribed to adopt the same (mixed or pure) original Strategy $x \in \Delta$. Assuming that all mutants are prescribed to use some other (pure or mixed) mutation Strategy $y \in \Delta$. Denote the proportion of mutators in the population (after entering) as ε, and $\varepsilon \in (0, 1)$. From this two-state (there are two different strategies), pairs of people are randomly selected repeatedly to play the game, and the probability of each drawing is the same. Therefore, if one person is drawn to participate in the game, then the probability of his opponent adopting the mutation Strategy y is ε, and the probability of the opponent adopting the existing strategy is $1 - \varepsilon$. The benefits of matching in this dual-state population are the same as the benefits of matching players with mixed strategies $w = \varepsilon y + (1 - \varepsilon)x \in \Delta$. Therefore, the post-entry payoff of the existing strategy is $u(x, w)$, and the post-entry payoff of the mutation strategy is $u(y, w)$.

Biology tells us that evolutionary forces will not choose mutation strategies, if and only if its payoff (adaptability) is lower than the payoff of existing strategies, that is

$$u\left[x, \varepsilon y + (1 - \varepsilon)x\right] > u\left[y, \varepsilon y + (1 - \varepsilon)x\right]$$

Strategy $x \in \Delta$ is known as evolutionarily stable, and this inequality is valid for any "mutation" Strategy $y \neq x$ if the proportion of mutations in the population is small enough.

Definition 7.9 If for any Strategy $y \neq x$, there is some $\bar{\varepsilon}_y \in (0, 1)$ such that the above inequality holds for all $\varepsilon \in \left(0, \bar{\varepsilon}_y\right)$, then $x \in \Delta$ is an evolutionary stable strategy (ESS).

7.4.3 Replicator Dynamic

Replicator dynamics refers to the repeated game of large groups composed of members with slow learning speed.

The following will take the "market scramble" game as an example to analyze the dynamics of replication.

Example 7.11 In the evolutionary game, the essence of each game is played by a member of the former group and a member of the latter group. In this example, the potential entrant group and the incumbent group are playing games. The payoff matrix is shown in Table 7.17. Please analyze the replicator dynamics and evolutionary stability strategies of the two group members separately.

Solution:

Table 7.17 "Market scramble" game payoff matrix

		Incumbent	
		Struggle	Acquiescence
Potential entrant	Entry	(0, 0)	(2, 2)
	No-entry	(1, 3)	(1, 3)

Assume that the proportion of potential entrants who adopt the "entry" strategy is x among the incumbents, the proportion of the "struggle" strategy is y.

Payoff of potential entrants:

$$u_{1e} = y \times 0 + (1 - y) \times 2 = 2(1 - y)$$

$$u_{1n} = y \times 1 + (1 - y) \times 1 = 1$$

$$\overline{u}_1 = x \times u_{1e} + (1 - x)u_{1n} = 2x(1 - y) + (1 - x)$$

Among them, u_{1e} is the payoff when the potential entrant adopts the "entry" strategy; u_{1n} is the payoff when the potential entrant adopts the "no-entry" strategy; $\overline{u}_1$ is the expected payoff of the potential entrant.

Payoff of the incumbent:

$$u_{2s} = x \times 0 + (1 - x) \times 3 = 3 - 3x$$

$$u_{2n} = x \times 2 + (1 - x) \times 3 = 3 - x$$

$$\overline{u}_2 = y \times u_{2s} + (1 - y)u_{2n} = 3 - 2xy - x$$

Among them, u_{2e} is the payoff when the incumbent adopts the "struggle" strategy; u_{2n} is the payoff when the incumbent adopts the "acquiescence" strategy; $\overline{u}_2$ is the expected payoff of the incumbent.

Potential entrants copy the dynamic equation

$$\frac{dx}{dt} = x[u_{1e} - \overline{u}_1] = x(1 - x)(1 - 2y)$$

Analyze the replicator dynamic equation of potential entrants: when $y = 1\big/2$, $dx\big/dy$ is always 0, and x is in a stable state; when $y \neq 1\big/2$, $x^* = 0$ and $x^* = 1$ are two stable states, where when $y > 1\big/2$, $x^* = 0$ It is ESS (the point corresponding to ESS is the point whose tangent slope is less than 0), when $y < 1\big/2$, $x^* = 1$ is ESS, as shown in Fig. 7.15.

(1) $y = 1/2$, $x^* \in [0, 1]$, as shown in Fig. 7.15a.

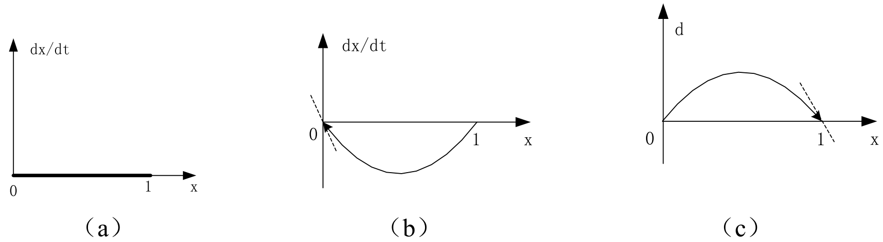


Fig. 7.15 Dynamic phase diagram of group replication of potential entrants

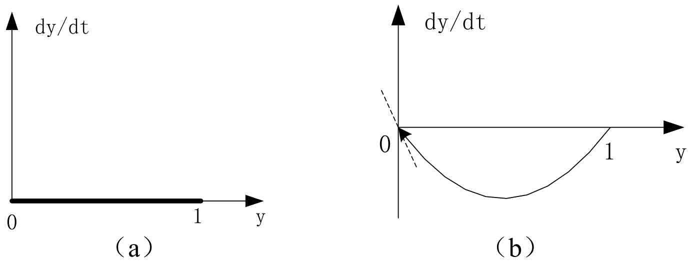
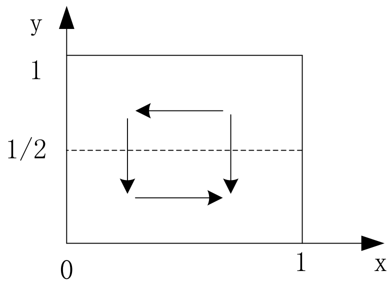


Fig. 7.16 Dynamic phase diagram of group replication of incumbents

Fig. 7.17 Stable relationship between potential competitors and incumbents



(2) $y > 1/2$, ESS: $x^* = 0$, as shown in Fig. 7.15b.
(3) $y < 1/2$, ESS: $x^* = 1$, as shown in Fig. 7.15c.

Incumbents and potential entrants replicator dynamic equations

$$\frac{dy}{dt} = y[u_{2s} - \overline{u}_2] = y(1-y)(-2x)$$

The incumbent's replicator dynamic equation is analyzed: when $x = 0$, then dx/dy is always 0, and x is in a stable state; when $x \neq 0$, then $y^* = 0$, $y^* = 1$, where $y^* = 0$ is ESS, as shown in Fig. 7.16.

(1) $x = 0$, $y^* \in [0, 1]$, as shown in Fig. 7.16a.

(2) $x \neq 0$, ESS: $y^* = 0$, as shown in Fig. 7.16b.

Finally, the relationship between potential competitors and incumbents is summarized in a picture (Fig. 7.17).

Chapter Summary

This chapter introduces the relevant theoretical knowledge of game theory. Through the study of this chapter, readers can have a preliminary understanding of game theory, and can solve simple practical problems with the theory. This chapter first gives the basic concepts of game theory, introduces its development and achievements; secondly, it focuses on the related concepts, properties and solutions of non-cooperative games and cooperative games; finally, it introduces evolutionary game. The main points of this chapter include: an overview of game theory, non-cooperative games, cooperative games, and evolutionary games.

Important Concepts and Terms

game theory
non-cooperative game
cooperative game
incomplete information game
complete information game
static game
dynamic game
zero-sum game
constant-sum game
variable-sum game
Nash equilibrium
Harsanyi transformation
matching game
Gale-Shapley algorithm
evolutionary game theory

Thinking Questions and Practice Questions

1. What are the basic elements of game theory?
2. What is the Nash equilibrium? Use the horizontal line method to solve the Nash equilibrium of the game shown in the table below.
3. What are preconditions of cooperative games? What are the differences between cooperative games and non-cooperative games?
4. Which three axioms need to be satisfied by the Shapley value method? What is its specific form?
5. What is the coalition structure?

		Participant 2	
		Strategy X	Strategy Y
Participant 1	Strategy A	(1, 4)	(2, 3)
	Strategy B	(0, 9)	(8, 7)

Case Study

The Application of Stackberg Game in the Design of Wholesale Price Contract.

The Stackberg complete information dynamic game is an output leadership model. As the leader, one party will anticipate the impact of the output that it decides on the follower and fully understand the follower's response function. That is, the leader first decides an output, the follower can observe it, and decide his own based on it. The idea of Stackberg game is often used in contract design and its feasibility judgment. Take the wholesale price contract design between Manufacturer Datang and Retailer Borui as an example.

Assuming that Datang is the price leader and Borui is the follower. Datang and Borui sign a wholesale price contract with each other. Borui determines the order quantity according to market demand and wholesale price, and Datang according to Borui's order Mass organization of production, Borui company bears all losses and market risks of unsold products. X is the market demand; $F(x)$ is the distribution function of demand X; $f(x)$ is the probability density function of demand x; μ is the expected value of market demand x, $\mu = E(X) = \int_0^\infty xf(x)d(x)$; Q is the quantity of products ordered by Borui to Datang before the sales season; p is the unit retail price of the product; v is the unit price at which Borui will process and sell stock products after the sales season; c_e is the unit inventory cost; c_u is the cost caused by the unit product shortage; w is the wholesale price of the unit product from Datang to Borui; c is the production cost of the unit product. Then Borui Company:

Expected sales volume: $S(Q) = Q - \int_0^Q F(x)dx$.
Expected inventory:$I(Q) = Q - S(Q)$.
Expected shortage:$L(Q) = \mu - S(Q)$.
Expected profit:

$$\pi_R = pS(Q) + vI(Q) - wQ - c_eI(Q) - c_uL(Q) = (p + c_e + c_u - v)S(Q) - (w + c_e - v)Q - c_u\mu$$

Optimal order quantity: $Q_R^* = \arg\max \prod_R$. Taking the partial derivative of p π_R with respect to Q and making it equal to 0, we get:

$$Q_R^* = F^{-1}\left(\frac{p + c_u - w}{p + c_e + c_u - v}\right)$$

Datang expects profits:

$$\pi_S = (w - c)Q$$

The overall profit of the two manufacturers:

$$\begin{aligned}\pi_T &= \pi_R + \pi_S = pS(Q) + vI(Q) - c_e I(Q) - c_u L(Q) - cQ \\ &= (p + c_e + c_u - v)S(Q) - (c + c_e - v)Q - c_u \mu\end{aligned}$$

Take the partial derivative of π_T with respect to Q and set it equal to 0 to obtain the function of the equilibrium production volume of the two Enterprises:

$$F\left(Q^*\right) = \frac{p + c_u - c}{p + c_e + c_u - v}$$

Thus the overall equilibrium output of the two manufacturers:

$$Q^* = F^{-1}\left(\frac{p + c_u - c}{p + c_e + c_u - v}\right)$$

In order to maximize the overall profit, it must meet $Q_R^* = Q^*$, which is $w = c$, at this time, Datang will not make any profits.

Assume that the market demand X of Borui's A product obeys a uniform distribution on $[a, b]$ $(a(b))$. The order quantity of A product before the sales season is Q, the retail price is 50, the processing price of the unit inventory product after the sales season is 40, the unit inventory cost of the product is 5, the unit stock-out cost is 6, the unit wholesale price is 20, and the unit production The cost is 15. Then there are:

$$F(X) = \ (X - a)/(b - a)X \in [a, b]$$

$$f(x) = 1/(b - a)\times \in (a, b)$$

$$\mu = \ (a + b)/2$$

$$I(Q) = Q - S(Q) = Q^2/2(b - a) - aQ/(b - a)$$

$$L(Q) = \mu - S(Q) = \ (a + b)/2 - bQ/(b - a) + Q^2/2(b - a)$$

$$\pi_R = \ (36b - 15a)Q/(b - a) - 21Q^2/2(b - a) - 3(a + b)$$

$$\text{Derivative}\, Q_R^* = \ (36b - 15a)/21$$

$$\pi_T = (41b - 20a)Q/(b - a) - 21Q^2/2(b - a) - 3(a + b)$$

$$\text{Derivative } Q^* = (41b - 20a)/21$$

Let $Q_R^* = Q^*$, deduct a = b. It contradicts to a < b. The profit maximization of Borui and the common profit maximization of the two manufacturers cannot be realized at the same time. In reality, this “bilateral effect” is inevitable, and the optimal decision of each member is not necessarily optimal for the joint system.

Question: If you consider Borui’s favorable and unfair aversion behavior, and its reference value for the favorable and unfair aversion is Datang’s profit, is it possible that the wholesale price contract model makes Borui’s profit and the two manufacturers’ profit? Maximize your profits at the same time?

Further Readings

1. Drew Fudenberg, Jean Tirole. *Game Theory* [M]. Beijing: Renmin University of China Press, 2010.
2. John F. Nash, Lloyd S. Shapley, John C. Harsanyi et al. *Game Theory Classics* [M]. Beijing: Renmin University of China Press, 2013.

Management Simulation

8

Hao Zhang

Learning Target

1. Understanding the connotation and type of simulation, know the common simulation technology and software.
2. Getting familiar with the principle of system dynamics, will use relevant software to build system dynamics model, understand the model analysis method.

Introduction

The Club of Rome and World Model

In April 1968, Italian economist Aurelio Peccei gathered in Rome 30 scientists, economists, writers, educators and entrepreneurs from 10 countries to discuss the problems and dilemmas facing mankind in the future. And this meeting led to the establishment of The Club of Rome. Its purpose is to promote human awareness of the various independent and changing components of the global system in which it lives (social, economic, political and natural); to enable decision-makers on Earth to take note of these interconnected and interacting relationships, thereby to generate a new understanding of the state and future of the world, and to promote the birth and implementation of new policies.

In the early 1970s, The Club of Rome, with 75 scientists and scholars from 25 countries, was puzzled by a series of world problems. Given that some of the usual research methods and tools are based on single factors, they cannot understand the overall nature of the system which is larger than the sum of the parts, and they are helpless in the face of nonlinear, high-order and multiple feedback systems, so it's really difficult for them to answer the question of this complex and huge system. The

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_8

Club of Rome finally turned to the system dynamics that had emerged in the 1950s and 1960s in applied research on industrial enterprise management and urban issues. System dynamics draws the essence of feedback theory and information theory, and forms a new subject of analyzing and studying information feedback system and understanding to solve system problems on the basis of system theory.

In the summer of 1970, all members of The Club of Rome attended a two-week seminar on system dynamics at Massachusetts Institute of Technology, at the invitation of J. W. Professor Forrester, founder of system dynamics. Forrester explained the theory and application of system dynamics to the participants, and introduced the prototype (WORLD1) of the pre-designed world model to the club members. The model clearly analyzes the interaction and relationship of the important factors in the world today. Encouraged by the possibility that system dynamics might solve the doubts of the future world, The Club of Rome decided to commission the MIT Systems Dynamics Group to further study the future world's prospects. At MIT, an international research group was set up, headed by Professor Dennis Meadows, a student of Forrester. On the basis of WORLD1 model development, Forrester published *World Dynamics*, and under his guidance, the research group published *The Limits to Growth* and other works, explaining their views and research results on the development of the future world, and introducing the WORLD1 model.

After completing the book *World Dynamics*, Professor Forrester believes that the world's existing series of puzzling problems cannot be solved at the world level; second, the United States has serious national problems, such as inflation, economic growth recession and so on, because the state is unable to solve these serious national problems, has led to widespread public dissatisfaction. Professor Forrester also took the initiative to meet the new task of studying American national model. It took 12 years and more than $6 million to complete the modeling work.

The main purpose of the system dynamics national model in the United States is to combine the main social and economic activities of the country into a whole and become a system dynamics model with a strict organizational structure, which is convenient to analyze and study the different functions of each component of the system.

The system dynamics has wide application, such as to analyze the development of the world, a region or a country, also to analyze the environmental problems and enterprise management problems. System dynamics is a branch of simulation. In recent years, the rapid development of simulation technology based on automation and computer has provided us with a lot of technology and tools to simulate and study management science problems, thus making a colorful research on management.

Source: Wang Qifan. From World Model to National Model [J]. *Science Journal*, 1985, (02):16–22.

8.1 Overview of Simulation

8.1.1 Concept of Simulation

The actual system is established into a physical model or a mathematical model, and then the results of the experimental study of the model are applied to the actual system. This method is called analog simulation research, or simulation for short.

Simulation research was more and more used in defense industry, military aviation, aerospace engineering, civil engineering, marine industry, machinery manufacturing and other fields in the early stage, but it is gradually playing an important role in the analysis and research of management science, such as in traffic policy impact analysis, municipal solid waste treatment, bullwhip effect of supply chain, environmental sustainable development analysis and so on.

8.1.2 Classification of Simulation

According to the model division used in the simulation, the simulation can be divided into two categories: mathematical simulation and physical simulation.

Mathematical simulation is the process of abstracting the actual system and describing its characteristics with mathematical relations to obtain the mathematical model of the system. The advantages of mathematical simulation are convenience, flexibility and economy, and its disadvantage is limited by system modeling technology, that is, system mathematical model is not easy to establish.

Physical simulation is the process of constructing the physical model of the system according to the physical properties of the real system and carrying out experiments on the physical model. It can add the actual running data of the system to the physical model, the non-linear factors and interference factors that cannot be predicted in the mathematical simulation, so the physical simulation is more intuitive and visual. Its disadvantage is that the model is difficult to change, the experiment is limited and the investment is large.

8.1.3 Role of Simulation

(1) The simulation process can be seen as a process of running experiments, which also provides a comprehensive and dynamic collection, aggregation, collation and analysis of information.
(2) The problems of prediction, analysis and evaluation in object systems which are difficult to establish physical and mathematical models can be solved efficiently by simulation.
(3) In order to facilitate the analysis of problems, the simulation process usually reduces a complex system to several relatively simple subsystems.

(4) Through simulation, the hidden problems in the original system can be found and solved in time, and the researchers can be inspired to produce new research ideas on the problems.
(5) In many cases, simulations are more cost-effective than actual exercises.

8.1.4 Modern Modeling and Simulation Technology

Modern modeling and simulation technology is a multidisciplinary and comprehensive technology based on similar principles, model theory, system technology, information technology and related professional technologies in the field of modeling and simulation applications, using computer systems, application-related physical effects equipment and simulators as tools, and using models to participate in existing or envisaged systems for research, analysis, design, processing, production, testing, operation, evaluation, maintenance and end-of-life (life-cycle) activities.

Modern modeling and simulation technology has been widely used under the demand of the development of various industries and the promotion of science and technology. The hot spots such as "intelligence, networking, collaboration, universality, virtualization" are the development direction of modern modeling and simulation technology. Combined with high performance computing, modeling and simulation technology will become an important means of management science research.

Macroscopically, the technical system of modern modeling and simulation consists of modeling technology, modeling and simulation support system technology and simulation application technology, as shown in Fig. 8.1.

8.1.5 Common Techniques for Management Simulation

Management simulation refers to the use of computer software technology to highly refine the specific work in enterprise management, to display specific practical situations as realistically as possible through certain means and forms, and to create a simulation environment for certain business management activities. In order to consolidate knowledge, exercise knowledge ability and improve their comprehensive quality, people experience real business process. At present, in the field of management science research, management simulation is widely used in supply chain logistics system operation simulation, resource planning and allocation, production scheduling and scheduling, information dissemination, market prediction analysis, management policy impact analysis and so on. The common management simulation techniques are: Multi-Agent technology, social network, Bayesian learning, system dynamics and the software such as Flexsim, MATLAB, Anylogic and Simulink.

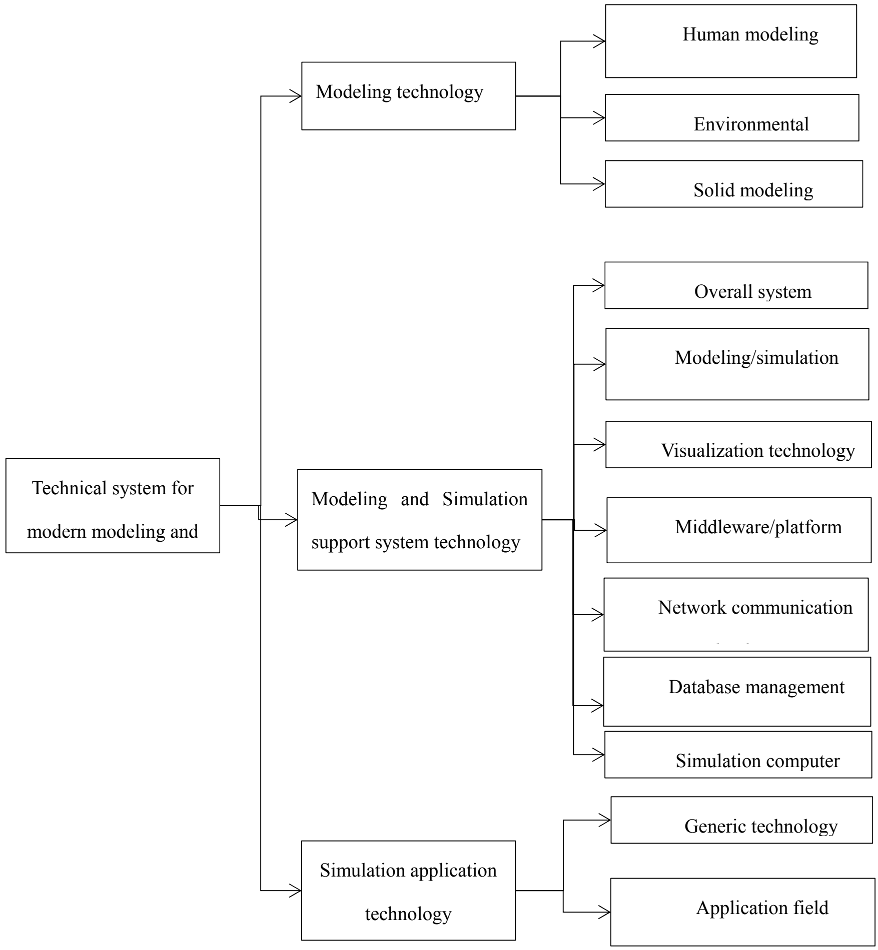


Fig. 8.1 Technical system for modern modeling and simulation

(1) Flexsim

Flexsim is a three-dimensional logistics simulation software in the United States, which can be used in system modeling, simulation and business process visualization. The object parameters in the Flexsim can represent all existing physical objects, such as machine equipment, operators, conveyor, forklifts, warehouses, containers, etc. At the same time, the data information can be represented by the rich model library of Flexsim. Flexsim is also an object-oriented open software, in which objects, windows, graphical user interfaces, menu lists, object parameters and so on are very intuitive. These objects can be exchanged between different users, libraries and models, and combined with the highly customizable of objects, the speed of modeling can be greatly improved.

(2) MATLAB

MATLAB is a commercial mathematics software produced by American MathWorks Company for advanced technical computing language and interactive environment for algorithm development, data visualization, data analysis and numerical calculation. MATLAB can perform matrix operations, draw functions and data, implement algorithms, create user interfaces, connect programs in other programming languages, etc. By using the functions of optimization toolbox in MATLAB, such as linear programming and quadratic programming, finding the maximum and minimum values of functions, multi-objective optimization, constraint optimization, nonlinear equation solving and so on, many practical problems in modern management can be solved conveniently.

(3) Anylogic

Anylogic is a simulation technology widely used in modern management, which can model and simulate discrete, continuous and mixed systems. By using Anylogic, we can easily complete the modeling and simulation of transportation, logistics, supply, policy influence, information dissemination and other systems in modern management.

8.2 System Dynamics

8.2.1 What is System Dynamics?

System dynamics is a branch of modern modeling and simulation technology. It is based on system theory and absorbs the essence of cybernetics and information theory. It is a comprehensive subject that not only analyzes and studies information feedback system, but also understands and solves system problems. In terms of system methodology, the method of system dynamics is the unity of structural method, functional method and historical method.

System dynamics is studied under the condition that the structure of the system determines the behavior of the system. It believes that there are many variables in the system and causality in the feedback loop of their interaction. These causal connections constitute the structure of the system, which becomes the fundamental determinant of the behavior of the system.

When people solve problems, they want to obtain better solutions. Therefore, the process of system dynamics solving problems is essentially an optimization process to obtain better system behavior. Because in the system dynamics, the system structure determines the behavior of the system, so it analyzes the function and behavior of the system from the point of view of the system structure, and obtains the better system behavior by finding the better structure of the system.

In the process of system dynamics, the whole system is regarded as a complete mechanism, which has multiple information causality feedback. Therefore, when the system dynamics method is used to analyze the whole system and get a lot of valuable and interrelated information, people usually set up the causality diagram of the system, then analyze and simplify the causality diagram in detail, and make it change into the system flow diagram. Finally, the simulation of the real system is completed by using various types of simulation software on the computer.

Then to build the structure of the system. Policy analysis or optimization is usually used, including parameter optimization, structural optimization, boundary optimization. Using parameter optimization to find better system behavior is mainly realized by changing several more sensitive parameters in the system; using structure optimization to find better system behavior is mainly realized by increasing or reducing horizontal variables and rate variables in the model; using boundary optimization to find better system behavior is mainly realized by changing system boundary and boundary conditions.

8.2.2 Development and Prospect of System Dynamics

The development of system dynamics can be divided into three stages:

1. Birth of System Dynamics—1950–1960s

The early stage of system dynamics is also called industrial dynamics. During this period, Professor Forrester's "Industrial Dynamics", published in *Harvard Business Review*, was regarded as the foundation of system dynamics. Then he elaborated the methodology and principle of system dynamics and the basic principle of system dynamic behavior. Later, Professor Forrester studied the development of the city in depth, and put forward the system dynamics model of the city.

2. System dynamics developed—1970–1980s

The main achievement of this stage is the success of the system dynamics world model and the American national model. These two models have successfully provided a new research idea for the economic community, so the system dynamics method has attracted the wide attention of scholars all over the world, and has greatly promoted the dissemination and development of system dynamics in the world. And established its position in the study of social and economic problems.

3. Extensive application of system dynamics—1990s to Present

At this stage, system dynamics is widely spread and developed worldwide. Many scholars have adopted systematic dynamics methods to study the social and economic problems, involving a wide range of fields, such as economy, energy, transportation, environment, ecology, biology, medicine, industry, city and so on. The scholars regard system dynamics as a discipline and are strengthening its links with control theory, system science, mutation theory, dissipative structure and bifurcation, structural stability analysis, sensitivity analysis, statistical analysis, parameter estimation, optimization techniques, generic structures, expert systems and so on.

System dynamics is based on the theory of nonlinear dynamics. From the future research trend in the field of management science, on the one hand, the application of system dynamics will be more extensive; on the other hand, there will be more in-depth research in several fields, such as nonlinear dynamics and complex systems, mental models based on subject modeling, dynamic decision-making and learning models, organizational and social evolution models.

8.2.3 System Dynamics Application Steps

1. Identification of problem

The strategy of the system is always closely related to the system objectives. Therefore, when using the system dynamics method to analyze the problem, we should first master the system objectives, that is, the purpose of modeling and simulation, so that the boundary of the system, the relationship between variables and the parameters needed for the system operation can be determined according to the objectives.

2. Determination of system boundaries

The determination of system boundary in system dynamics refers to the determination of internal variables. System boundaries should be as accurate as possible. Factors that do not affect the state of the system or have nothing to do with the purpose of the study should be excluded from the system and not placed within the system boundary.

3. Determining causation

The determination of system causality is an important link in system dynamics simulation. To determine causality is to analyze the interaction and feedback between various variables within the system, and then express this causality in the form of feedback loop, so as to have a whole grasp of the structure of the system. Master the interaction of system variables and the role of system and external environment.

4. Establishing System Dynamics Model

After analyzing the causality of system elements, the causality diagram and stock flow diagram of the system can be made by using the special software of system dynamics. The special software of system dynamics includes the representation method and symbol of various variables. Vensim and Anylogic are commonly used simulation software for system dynamics. After the causality diagram and stock flow diagram are determined, the function relation between variables can be input in the software to simulate the system.

5. Model testing and debugging

Any model that fully meets the conditions and can carry out high-quality simulation cannot be established one-time, it needs to be constantly debugged, refined, and modified. The process of this cycle is to study the relationship between the structure, mechanism and simulation behavior of the system, so as to achieve the real purpose of modeling. When testing the structure of the model, we can consider whether the dimension is correct or not, and whether the model is still meaningful under boundary conditions. The structural sensitivity, parameter sensitivity and so on can be considered when testing the simulation behavior of the model. In addition, you can also use "appearance" test, parameter re-test, strange behavior test, extreme test and many other methods.

6. Optimization of simulation

Optimization of simulation means that by simulating the model, simulating the running state of the system, finding the influencing factors closely related to the operation of the system, analyzing the measures to improve the operation of the system, and finding a more satisfactory strategy. In the process of optimization simulation, the optimal state of the system can be sought by changing the value of one or some indexes to observe its influence on the simulation results.

8.2.4 Basic Concept of System Dynamics

The basic concepts of system dynamics are as follows:

(1) System: a set or whole of interrelated, distinct, interacting elements with a specific function.
(2) Feedback: the relationship between the output and the input of the same unit or sub-module system. For a complete system, the output and input of feedback correspond to the input of the system itself and the external environment.
(3) Feedback system: a system that contains the interaction between feedback links and each feedback link.
(4) Feedback loop: a closed path consisting of a chain of causality and interaction.
(5) Causality loop diagram (CLD): it contains multiple variables, and the connections between the variables are represented by arrows, as shown in Fig. 8.2, so CLD is an important tool to represent the feedback structure of the system.

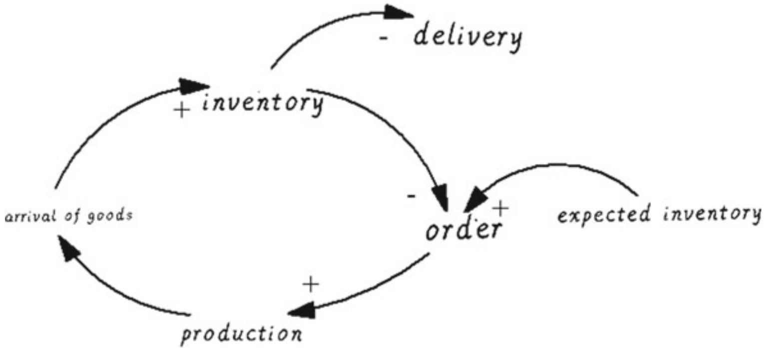


Fig. 8.2 Inventory ordering causal loop

(6) Causality chain polarity: In CLD, each causal chain can be represented by positive (+) or negative (−) polarity. Polarity means that when the arrow end variable changes, the corresponding arrow end variable will also change. When the change trend of arrow tail and arrow end variable is the same, the polarity is positive, and when the change trend of arrow end and arrow end variable is opposite, the polarity is negative. As shown in Fig. 8.2, when the arrival of the goods increases, the inventory increases, and the polarity of the causality chain is positive.
(7) The polarity of the feedback loop: the polarity of the feedback loop is also divided into positive feedback and negative feedback, which is positive feedback when the deviation of the variables in the loop is enhanced, and negative feedback when the variables in the control loop tend to be stable. If the feedback loop contains even negative causality chains, the feedback loop is positive feedback, and if the feedback loop contains odd negative causality chains, the feedback loop is negative feedback.
(8) Delay: It usually takes a certain amount of time for any decision to be implemented and to produce an effect, which is called delay. There are delays everywhere in the circulation channel of the system. There are two kinds of solution: one is to ignore the shorter delay under comparison, the other is to simply merge the delay of multiple series on the channel to form a delay.
(9) Stock flow chart: a graphical model representing the interrelated forms of horizontal and rate variables in the feedback loop and the relationships between the loops in the feedback system. The stock flow chart should include at least one horizontal variable and one rate variable. As shown in Fig. 8.3 inventory system inventory flow chart.
(10) Horizontal variable: also known as a state variable or flow rate, it is a cumulative variable whose numerical magnitude represents the condition of a system variable at a given time, equal to the net difference between the inflow rate and the outflow rate.
(11) Rate variable: also known as rate of change, it increases or decreases the value of a horizontal variable over time. The rate variable represents the speed

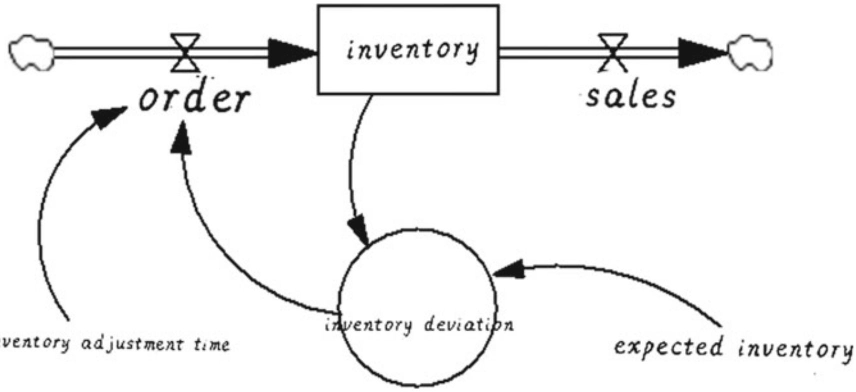
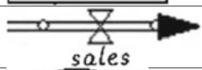


Fig. 8.3 Inventory system inventory flow

at which a horizontal variable changes. The horizontal variable is represented by a rectangle, and the specific symbol should include a streamline describing the flow rate of input and output, the name of the variable, etc. the rate variable is represented by a valve symbol and shall include the variable name, the streamline of the flow controlled by the rate variable and the amount of information input on which it depends.

(12) Auxiliary variables: auxiliary variables are variables set in the information channel between state variables and rate variables. The symbolic representation of variables and other components in the inventory flow chart of the inventory system is shown in Table 8.1.

Table 8.1 Variables and component symbols in inventory flow chart

Name of variable or component	Symbol
Horizontal variables	
Rate variables	
Auxiliary variables (or constants, exogenous variables)	
Logistics chain/information chain	
Source or leak	

8.2.5 Testing of the Model

After the model is established, the test and analysis test are more important. Any model cannot fully reflect the reality, there will always be ideal assumptions, then the model needs to be constantly tested, adjusted, has achieved its usability in the reality, to achieve the purpose of simulation. Several important model testing methods are as follows:

(1) Testing of system boundaries

At the beginning of the model, the boundary of the system should be determined first. There are many important variables in the model. Whether these variables are endogenous or exogenous, and whether the behavior of the system model is sensitive to the change of boundary setting is large or small, all of which belong to the test of system boundary. In addition to drawing advice from experts, this test can also make some changes to certain variables: increase or decrease, and then observe whether a closed loop can be formed to test the system boundary.

(2) Testing of system model validity

The validity test of the model can also be called the test of its reliability and accuracy. The basic requirement of the model is to meet the demand of its user. Secondly, it is not the most important problem for the system dynamics model to be complex or simple, but whether it can solve the problem of actual users is. Therefore, whether it is practical and applicable is one of the measures of model validity test. For the simulation and comparison of historical data, the accuracy of the model can be tested to a certain extent, so as to test its effectiveness.

(3) Testing of dimensional consistency

Dimension is the unit of measurement of variables in the model, which represents the size or quantity of physical quantities, such as m, Kg, yuan, tons and so on. Dimensional consistency test is a simple test, with the purpose to ensure that the dimension has certain practical significance, and to maintain internal unity in the formula.

(4) Testing of parameter estimates

The purpose of this test is to verify whether the parameter value is consistent with the descriptive situation given by the system and whether there are corresponding indicators or factors in reality. The basic parameter estimation methods include: statistical method, expert opinion method, single equation or multiple equation methods of estimation, subsystem or partial model test method.

(5) Sensitivity testing

Sensitivity testing have some correlation with parameter estimation tests because sensitivity analysis can help us to find parameters that are extremely sensitive to the effects of system behavior and focus on more accurate estimation of these parameters. For less sensitive parameters, we only need a rough estimate of the value. The sensitivity test can also judge the behavioral sensitivity of the model structure and the sensitivity of some policy parameters.

(6) Testing in extreme cases

To test whether the model is still running normally and whether it can still simulate the actual change law when several typical extreme cases appear, this can test the stability and reliability of the system equation. There are usually two methods for testing extreme cases: one is to manually verify the equations in the model one by one to consider whether they contain extreme cases; the other is to give extreme values to some parameters in simulation, observe the simulation results and changes, and make analysis.

8.3 Multi-agent System

8.3.1 Overview of Multi-agent System

Since John McCarthy put forward the concept of “artificial intelligence” at the Dartmouth seminar in 1956, the concept of “agent” has begun to rise. An agent can be regarded as a computing entity in a certain environment, which can play a role independently, perceive the environment through the sensor, and act on the environment through the effector. There are two definitions of agent: in a broad sense, it is a network-based, cooperative and self-controlled entity; in a narrow sense it is an entity with the ability of intelligence, feeling, understanding and emotion similar to human beings.

In the early 1990s, the research of multi-agent system developed rapidly. Multi-agent system is a collection of many computable physical entities or abstract intelligences, each of which acts on itself and environment, and intercommunicates information with other intelligences. The emergence of multi-agent promotes the development of many software technologies and is the latest development direction of AI. First, multi-agent can develop new programming or solving methods to deal with incomplete, uncertain knowledge through communication between agents; second, collaboration between agents not only improves the basic capabilities of each agent, but also further understands social behavior from the interaction of agents; finally, systems can be organized in modular form.

Multi-agent technology has autonomy, distribution, coordination, self-organization ability, learning ability and reasoning ability. Multi-agent technology has unparalleled expressiveness for complex systems, which is due to the heterogeneity of each agent in the same multi-agent system. Therefore, it provides a unified model for various practical systems. Thus, it provides a unified framework for the research of various practical systems, and its application field is very broad. The three significant advantages of multi-agent simulation over other simulation methods are (1) applicability, (2) intelligence and interactivity, and (3) distribution and mobility. The multi-agent system simulation technology is the inheritance and development of the traditional module-based and object-oriented simulation technology, which makes the multi-agent simulation system widely applicable. The multi-agent system can make up for the deficiency of single intelligence, make the whole system more intelligent and interactive than any single intelligence. The agent based on distributed computing is distributed and mobile, which is more suitable for the needs of group decision making.

8.3.2 Multi-agent Modeling

Multi-agent modeling mainly includes three elements: agent, intelligence and interaction. It can make the agent intelligent, such as emotional, reasonable and so on, according to the needs of research. Interaction is achieved through the environment or language between the agents, and the agents can be influenced by the agents (or people and environment), leading to changes in behavior.

By using the bottom-up research method, through the study of the individual characteristics and behaviors of the system, the model of individual characteristics and behaviors is established. The structure and function of the system are studied by mapping the individual behavior into the behavior of the agent.

1. Selection of agent system architecture

The architecture of agent mainly includes the following three kinds, which need to choose the intelligent architecture according to the characteristics of the research contents.

(1) Stratified architecture

The hierarchical architecture is a centralized system, which follows the hierarchical organization of traditional manufacturing enterprises. The multi-agent hierarchical architecture commonly used in intelligent systems is shown in Fig. 8.4.

(2) Federal architecture

Because of the centralization of power, the hierarchical architecture centrally stores and shares the domain information, which brings many problems to the system construction. People gradually realize that the federal architecture is more suitable for application systems, especially the large application systems. The federated agent system has no centralized storage of shared data, and all data is stored in

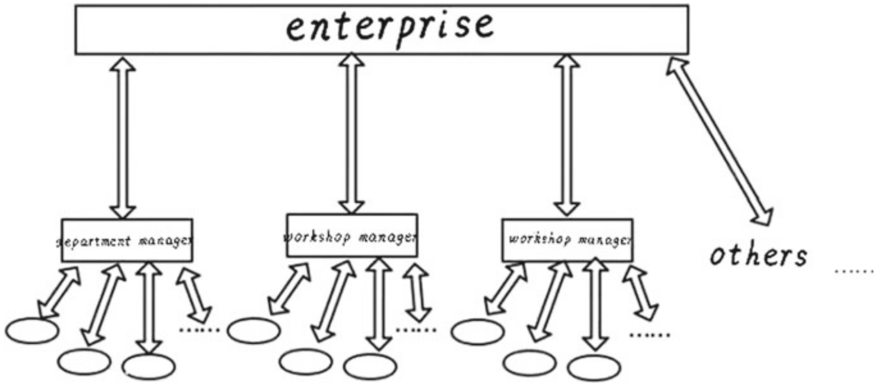


Fig. 8.4 Stratified architecture

the local database of the distributed agent and the distributed data is updated by the messaging of the agent.

Facilitator method is a widely accepted method in federal architecture. Facilitator combines several agents into a group as shown in Fig. 8.5, communication between the agents (within and between groups) is always done through an interface called Facilitator. Facilitator provides a reliable network communication layer and sends messages between the agents according to the message content, and is responsible for coordinating the control of multi-agent activities.

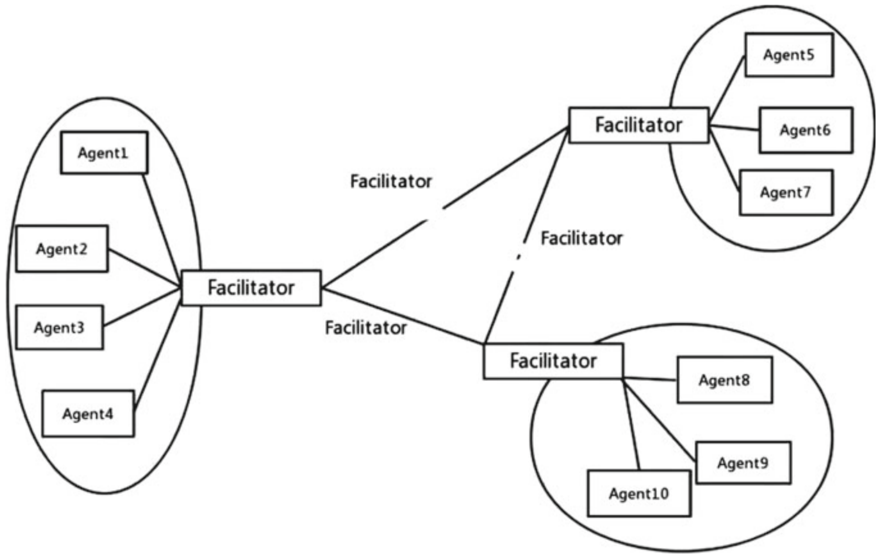


Fig. 8.5 Facilitator Federal multi-agent architecture of the method

(3) Architecture of autonomous intelligent system

The autonomous agent method is also called the agent network method, and its communication and state are not unified. In the autonomous intelligent system, each agent must know when to send the message to where; whether other agents are available; and what capabilities they have. The autonomous intelligent body method is suitable for large granularity multi-agent application system, which has simple structure and easy management and control.

For the development of individual intelligent system for the purpose of learning, it is recommended to use autonomous intelligent system, because it contains fewer agents and has a typical and pure system structure. Federal architecture is often used in large-scale application development, and different types and numbers of agents can be selected to design intelligent systems according to different needs.

2. Communication, coordination and cooperation of agents

Communication, coordination and cooperation are three key technologies in multi-agent system.

(1) Communication

Communication provides support for information exchange and activity coordination, which is the most important method and way to cooperate and coordinate between agents. The agent communicates with each other by sending messages. In the multi-agent system, the intelligent body can write messages or post some results to the shared data warehouse, and can also obtain information from it for data exchange to achieve the purpose of communication.

(2) Coordination

The purpose of the coordination process is to ensure that all necessary components of the whole problem are included in the activities of at least one agent, so that the agent interactively integrates all activities into an integrated overall solution, while also ensuring that team members work in a purposeful and coordinated manner, and finally ensuring that all these goals can be achieved within limited computing power and resources.

(3) Cooperation

Cooperation is often one of the main features of multi-agent systems that are different from distributed computing, object-oriented systems and expert systems. Typical levels of cooperation are: full cooperation, partial cooperation, and hostility. Fully cooperative agents can solve non-independent problems, but communication costs are high. In order to ensure consistency and cooperation, this intelligence can change its goals to meet the needs of other agents. Hostile agents, on the contrary, do not cooperate or even hinder each other's goals, such a system communication costs are low. Traditional systems are usually located between the two, they have no clear hostile goals, and at least some agents can have cooperation.

(4) Relationship between communication, coordination and cooperation

Coordination may require cooperation, but cooperation between a set of agents may not necessarily lead to coordination, and may actually lead to inconsistent behavior. Because if an agent wants to work together successfully, it must maintain each other's models and develop and maintain models for future interaction; adding a wrong agent is likely to produce inconsistent behavior. Coordination may also occur without cooperation. Similarly, lack of cooperation between agents does not necessarily lead to inconsistency. Competition is a form of coordination of hostile agents.

In order to facilitate coordination or cooperation, an agent usually needs to communicate with each other. The key to promoting coordination through communication is that an agent must let other agents understand its goals, intentions, results and states.

3. Modeling steps for multi-agents

The modeling steps of multi-agent systems are generally divided into the following four steps:

(1) The division of agents

The total task or goal of the system is decomposed into several sub-tasks or sub-goals. At the same time, the function of the system is decomposed and assigned to the corresponding intelligent body. Different functions require different smart body types. According to the situation of sub-task and sub-goal, as well as the decomposition of system function, we can determine the type and quantity of required agents and their functions.

(2) Static structure modeling of agents

According to the system, the basic structure of each intelligent body can be designed as reaction type, careful thinking type or mixed type. At the same time, according to the decomposition of the functions and functions of the system, determine the functions that each intelligence needs to achieve, and abstract the goals, knowledge and capabilities that should be possessed.

(3) Determination of the dynamic behavior of agents

The dynamic behavior of agents usually has external interaction behavior and internal thinking state. Internal state is generally described by reasoning mechanism. External interaction behaviors include communication, conflict resolution and cooperation. The dynamic behavior of these two aspects is essentially unified.

(4) Integration of multi-agent systems

The whole multi-system modeling process adopts the bottom-up design method: first to define each agent, then to study how to complete the task solution of one or more entities.

8.3.3 AnyLogic-Based Multi-agent Models

The function of multi-agent technology must be realized and applied through the development platform. At present, the commonly used multi-agent modeling development platform is: Anylogic, Swarm, Ascape, REAPST and so on. AnyLogic is an original simulation software used to design complex systems including discrete, continuous, intelligent and mixed behavior. AnyLogic, based on the latest complex system design methodology, is the first tool to introduce UML language into the field of model simulation and the only commercial software to support hybrid state machines, a language that can effectively describe discrete and continuous behavior. AnyLogic can quickly build the simulation model and hardware environment of the design system, such as physical equipment and operators.

AnyLogic has such characteristics as follows:

(1) AnyLogic supports Java languages

AnyLogic not only includes graphical modeling language, but also allows users to extend simulation models in Java programming language. Java is the native language of the AnyLogic. It cannot only customize the model by writing Java program, but also use Java Applets to generate the model. Any standard web browser can open the model. Models of these Java Applets can be placed on the site and published to users as a basic tool for decision support.

(2) AnyLogic supports r multiple approaches

AnyLogic can build discrete event (DE) models, system dynamics (SD) models, and agent body (AB) models. AnyLogic allows these simulation modeling methods to be all combined in the same model without a fixed hierarchy. This hybrid modeling method can deal with the modeling problems of various complex systems.

(3) AnyLogic has built-in multiple model libraries

AnyLogic includes enterprise library, pedestrian library, track library, road traffic library, and can define such elements as human, cars, walls, areas, paths, nodes, and such behaviors of the elements as in-and-out, movement, information exchange, and so on. It provides the model basis for multi-agent simulation and can effectively visualize the simulation process to verify and display the simulation model.

Example 8.1

Simulation of Traffic Signal Light Model at Multiple Intersection

(1) Model background

Suppose there are three intersections in a section of the area, a crossroad, a crosswalk, and a T intersection, and each is controlled by a set of traffic lights. Each car in the area is a separate agent, which drives in five directions from the area at different flows, passing through the area and then driving out. In this network, the behavior of small and medium-sized cars will be controlled by traffic lights and strictly abide by traffic rules. Through the simulation of the model, the influence of traffic signal combination under different traffic flow on the traffic network in this area is observed, and the best combination of traffic signal duration under different traffic flow is found out. Minimize the waiting time of vehicles in the road network.

(2) Model building

The road network of this section area is constructed in Fig. 8.6 in the AnyLogic, a total of 6 roads are set, respectively road, road1, road2, road3, road4, road5, and the number of lanes is set as two-way 4 lanes according to the road conditions. The stop line position of each road, vehicle driving path and other model elements are shown in Fig. 8.6.

Define vehicle agent properties according to Table 8.2.

The definition of the behavior of a car, an agent, is shown in Fig. 8.7. The carSource represents the car's source, the agent car is generated from the car source, and the carSource1 generated car starts from the road and goes to the road2, road3, road4, road5, respectively. Among them, to go to road3, road4, road5 need to pass road1; cars generated by carSource2 from the road2, respectively go to the road, road3, road4, road5, of which go to road3, road4, road5 need to pass road1; cars generated by carSource3 come from the road3, respectively to the road, road2, road4, road5, of which to go to road or road2 needs to pass road1; cars generated by carSource4 from the road4, respectively go to the road, road2, road3, road5, of

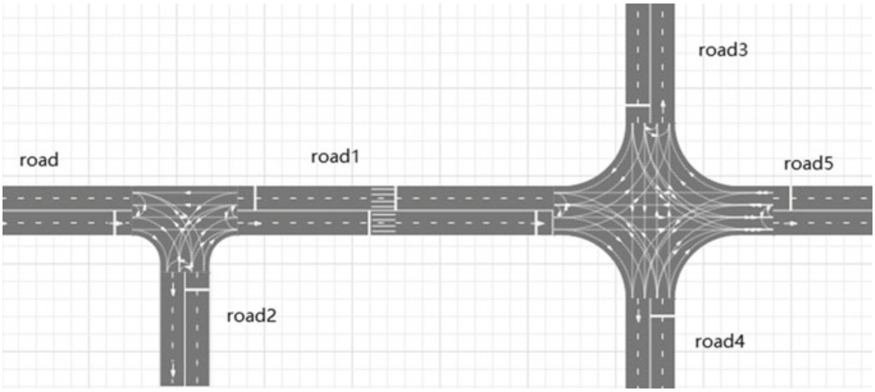


Fig. 8.6 AnyLogic intersection model

Table 8.2 Vehicle agent properties

Car source	Starting road	Hourly traffic flow	Vehicle length (m)	Initial velocity (km/h)	Preferred velocity (km/h)	Maximal acceleration (km/h)	Maximum deceleration (km/h)
carSource1	road	500	4	60	60	2	5
carSource2	road2	500	4	60	60	2	5
carSource3	road3	500	4	60	60	2	5
carSource4	road4	500	4	60	60	2	5
carSource5	road5	500	4	60	60	2	5

which to go to road, road2 needs to pass road1; cars generated by carSource5 from the road5, respectively go to the road, road2, road3, road4, of which to go to road, road2 needs to pass road1. The target probability of each car source generated is shown in Table 8.3.

Defining the duration of the trafficLight, trafficLight1, trafficLight2 is shown in Table 8.4, in which the duration of each direction of each traffic light is set a parameter identification, which is convenient to debug and change during the simulation process. The initial value of the traffic light duration parameter is set as shown in Table 8.5.

(3) Simulation results and analysis

The simulation results of the regional road network are shown in Fig. 8.8, and the histogram of the probability density distribution of the vehicle passing time in the regional road network is shown in Fig. 8.9, where the horizontal axis represents the time distribution of the vehicle passing through the road network. The red line represents the average time of the vehicle passing through the road network.

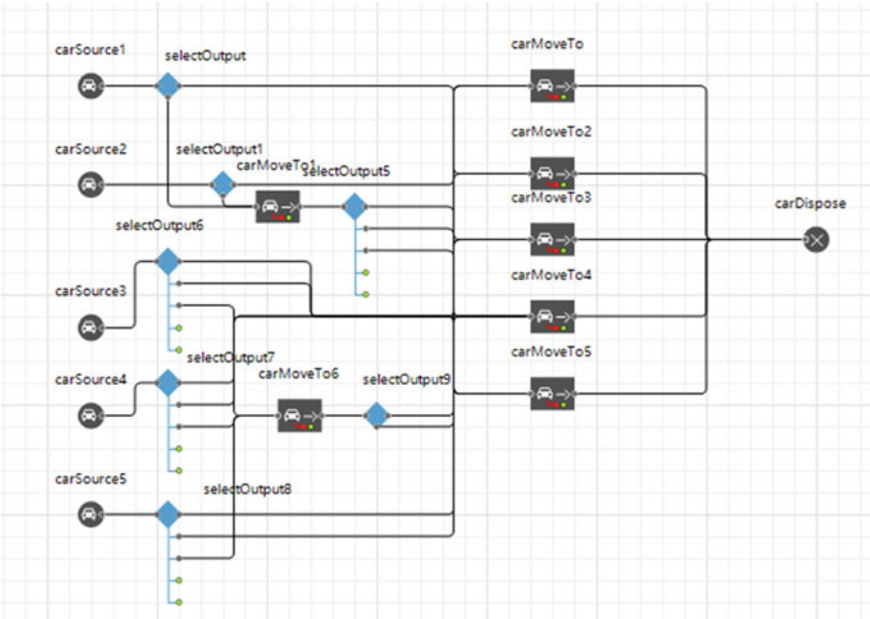


Fig. 8.7 Car agent behavior

Table 8.3 Destination probability of car source

	road	road2	road3	road4	road5
carSource1	–	0.5	0.15	0.2	0.15
carSource2	0.5	–	0.15	0.2	0.15
carSource3	0.2	0.2	–	0.3	0.3
carSource4	0.15	0.15	0.3	–	0.4
carSource5	0.2	0.2	0.3	0.3	–

Table 8.4 Duration parameters of traffic signal

	East–west green light	East–west red light
trafficLight	IntersectionWE	IntersectionS
trafficLight1	Intersection1Drive	Intersection1Stop
trafficLight2	Intersection2WE	Intersection2SN

Through this model, we can intuitively see the traffic flow of the road network in this area, move the sliders of different sections of the road, change the traffic flow from each direction into the area, and move the sliders of different junctions.

Table 8.5 Initial value of traffic signal duration parameters

Parameter name	Initial value	Minimum value	Maximum value
IntersectionWE	20	0	60
IntersectionS	20	0	60
Intersection1Drive	20	0	60
Intersection1Stop	20	0	60
Intersection2WE	20	0	60
Intersection2SN	20	0	60

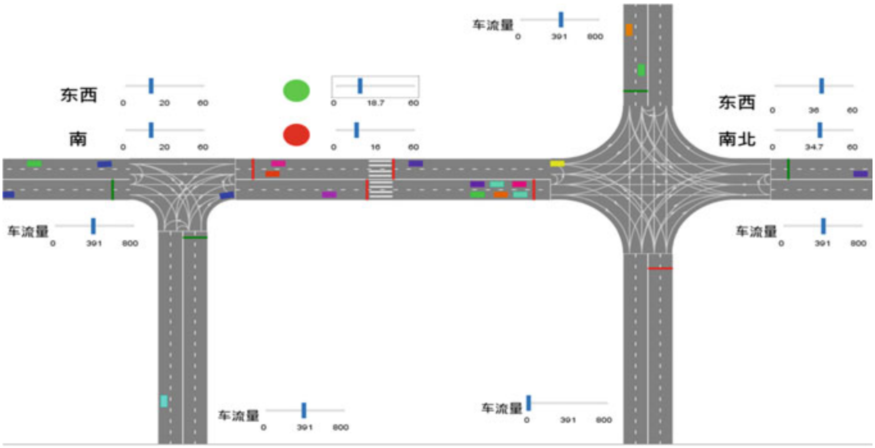


Fig. 8.8 Regional road network simulation results

Finding out the best combination of traffic light duration can minimize the waiting time of vehicles in the road network and reduce traffic congestion.

Summary

Simulation often involves hardware equipment, computer software, management measures, human behaviors, and so on. It is a comprehensive technical field, so managers need to have more comprehensive professional knowledge, at least need to be familiar with the principle of simulation. Based on systematic dynamics, the chapter leading students to understand and explore management simulation, with the main points: the concept and classification of simulation, the common simulation technology, the principle and application methods of system dynamics, and so on.

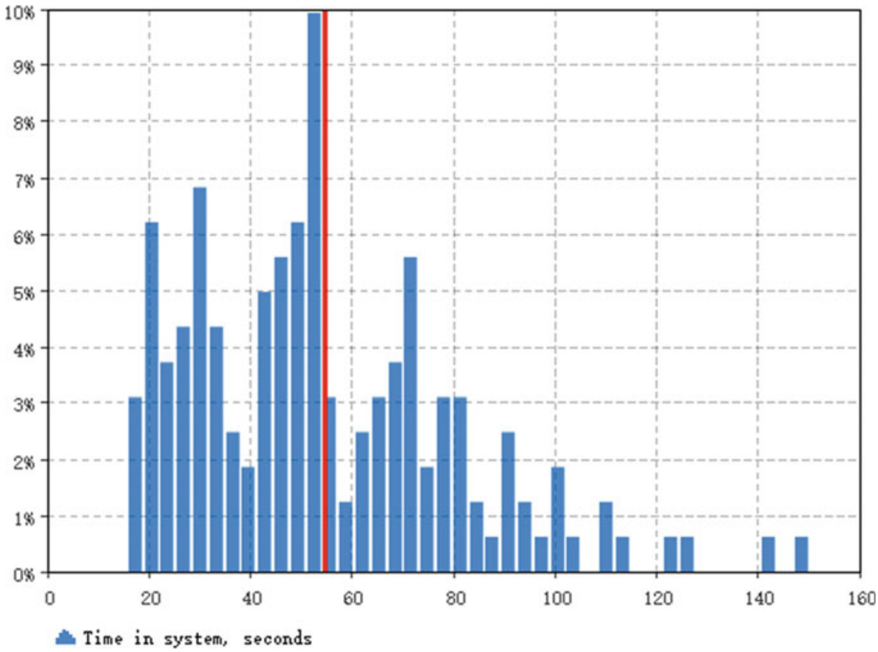


Fig. 8.9 Histogram of probability density distribution of vehicle passing time

Important Concepts and Terminology

simulation
system dynamics
feedback loop
causal loop diagram
level variable
rate variable
auxiliary variable
constant
sensibility analysis

Questions and Exercises

1. What are the commonly used management simulation techniques?
2. Please briefly describe the application steps of system dynamics.
3. What are the types of system dynamics model testing methods? What is the purpose of different test methods?
4. Analysis of quality and safety supervision of fresh agricultural products.

As one of the most fresh and unique products in the category of agricultural products, fresh agricultural products are the necessities of human survival and are closely related to people's daily life. Their quality and safety is related to the national economy and people's livelihood, and attracts the general concern of the people.

Problem identification and system boundary determination: at present, most vegetables in the market enter the market directly through logistics transportation from vegetable origin (including vegetable transit centers, vegetable markets, large supermarkets, etc.), and rarely go through deep processing. Therefore, vegetable quality and safety control focuses on the management of production links at the source and the management of transportation and distribution links. In the production link, the agricultural residue mainly comes from the use of pesticides and fertilizers in the production process. Once the sum of the two exceeds the reasonable use, it will inevitably lead to the increase of the final agricultural residue exceeding the standard rate. In reality, a large number of unqualified agricultural products are produced in the production link because of the excessive amount of agricultural residues. Due to the lack of effective supervision or loose management, these unqualified agricultural products always enter the market through logistics transportation. In addition, this phenomenon occurs in the transportation process, that is, some agricultural products contaminated by logistics transportation flow into the consumer markets, which results in a large number of unqualified agricultural products. In transportation and circulation, transportation distance and transportation time are often too long, and transportation equipment (such as cold chain transportation) cannot keep up with the biological pollution caused by the transportation process. Many illegal vendors add preservatives or other illegal additives to vegetables, which threatens the quality and safety of vegetables.

From the point of view of the supply chain of agricultural products, taking the quality and safety supervision of fresh agricultural products as an example, this problem analyzes the influence of each link in the whole supply chain on the quality and safety of fresh agricultural products, so as to make readers understand more clearly how to use system dynamics to solve practical problems.

Please use the tool to draw a causal relationship diagram for the quality and safety of fresh agricultural products, choose a causal chain arbitrarily, point out its polarity, and explain how the arrow-end variable will change when the variable at the end of the arrow of this causal chain is changed.

The system flow diagram of the quality and safety of fresh agricultural products, find or virtualize some data, give analysis ideas, conduct simulations, and analyze and discuss the results.

Case Analysis

Evolution of American Electronic Information Industry

(1) Research assumptions

The system dynamics is used to simulate the evolution path of American electronic information industry. The evolution model follows the following assumptions:

Assumption 1: the electronic information industry is regarded as the main body of evolution, focusing on the influence of government behavior on the electronic information industry, the competition and cooperation among enterprises in the industry, and the evolution of upstream and downstream industries.

Assumption 2: the personnel engaged in the field of electronic information research and development include enterprise R&D personnel, government-owned scientific research institutions developers, institutions of higher learning related researchers, electronic information technology enthusiasts. In the construction of the model, the four are unified as the core technical personnel, do not distinguish.

Assumption 3: in reality, inflation, exchange rate and other factors will have an impact on the scale, pattern, import and export of the electronic information industry. In the function fitting process of the model, these factors are difficult to quantify accurately. The effect of inflation, exchange rate and other factors on the evolution of electronic information industry is not directly reflected, which will be indirectly reflected by GDP, industry added value, personal consumption and other factors.

Assumption 4: taking the number of patents as the evaluation index of industrial technological innovation ability, because the electronic information industry belongs to the technology-driven industry, major technological breakthroughs often make the industry transition. However, the time node and influence degree and scope are difficult to predict and measure.

(2) Model building

According to the evolution motivation of the electronic information industry in the United States, the causal cycle diagram of the evolution mechanism of the electronic information industry is constructed, as shown in the Fig. 8.1. The electronic information industry evolution mechanism consists of four aspects: technology invention, the achievement transformation, the market demand and the government behavior. The four aspects are interrelated and affect each other. Technological innovation begins with a new idea, which must combine and integrate social needs with technical feasibility and form technological invention after research and development. Not all technological inventions can be commercialized. The decisive factor from technological innovation to industrialization lies in the market, and the market demand is the fundamental motive force of technological invention industrialization. If we cannot achieve high income through the market, the industrialization of technological inventions cannot be carried out. The market plays an important role in technological innovation and diffusion through the law of value and the mechanism of resource allocation. From the new product contacting market to the formation of a certain market scale, for the manufacturer, it is necessary to go through a stage in which the input is greater than the return, that is, the stage of the transformation of the results. At this stage, we should continue to improve the function of the new product to make it more suitable for the needs of consumers, at the same time, improve the production process, reduce the manufacturing cost, reduce the sales price, and form mass production. In the early stage of commercialization of technological innovation results, a large amount of capital, equipment and human resources need to be invested. At this time, the market potential of new products is not expected to be clear, and the return from new products is very small, which brings greater commercial risks to enterprises. Therefore, for technological innovation with market potential, in the early stage of commercialization, it is necessary for the government to give certain support, integrate all aspects of resources, promote the industrialization process of technological innovation through economic and political means such as policies, regulations, systems, etc., so that technological innovation can pass through the early stage of market development.

(3) Issues

① In order to further characterize the evolution mechanism of the electronic information industry, please draw a causal cycle diagram, and construct a flow diagram of the evolution mechanism of the electronic information industry based on the causal cycle diagram. In the diagram, the technological invention is represented by the annual growth of patents, the market demand is represented by per capita GDP and personal consumption, the industrial influence and achievement transformation is represented by the industrial added value, industrial R&D investment, electronic information industry practitioners and core technicians, the government behavior is represented by government revenue, government expenditure, government science and technology investment, government procurement, and the economic environment is represented by GDP, the population and employees.

② Please inquire the data, train model, conduct sensitivity analysis, simulate the future development of the American electronic information industry, and analyze the results.

Further Readings

1. Zhong Yongguang, Jia Xiaojing, Qian Ying, et al. *System Dynamics* (2nd Edition) [M], Beijing: Science Press, 2013.

2. Jia Ren'an et al. *Dynamics of Organizational Management Systems* [M]. Beijing: Science Press, 2014.

Complexity Science

9

Hao Zhang

Learning Target

1. Understanding the scope and development of complexity science.
2. Understanding the characteristics of complex systems.
3. Understanding the theoretical connotation and application scope of dissipative structure, synergy, chaos theory, catastrophe theory, hypercycle theory, fractal theory, self-organized criticality.
4. Applying the basic models of several typical theories to scientific research.

Introduction

Fractal music

There contains much wonderful fractal art in beautiful music, especially in the classics, in which the pitch and rhythm have fractal features. The noise of different frequencies in the pitch is mainly divided into white noise, red noise and pink noise. White noise sounds basically the same before and after, with the strongest predictability and the weakest randomness; the red noise has no rules to follow, no beauty at all, the strongest randomness and the weakest predictability; in terms of predictability and there is a balance between randomness, called "1/f noise", also known as "pink noise". From the time domain waveform, the sound wave image of pink noise has the fractal structure, and the analysis of a large number of music scores has long confirmed that most of the world's famous songs are pink noise. There are beats in music, and some types of music have a strong sense of rhythm, and the rules that appear will have a certain repetitiveness, showing a kind of self-similarity, but not exactly the same, making the rhythm of music in a random.

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_9

Fig. 9.1 Harlan J. Brothers and B. B. Mandelbrot (right)

Considering other musical elements such as pitch and rhythm, algorithm artists can use computer programming to design specific algorithms to generate fractal music through multiple iterations, and use the principle of self-similarity to construct some synthetic music with self-similar segments, with themes repeated in the repeating cycle of minor tunes again and again, and some random changes can be added to the rhythm. The effects created by fractal music can realistically imitate real music on both the macro and micro levels. Instead of sounding grandiose, the fractal music sounds interesting. It can even turn into music the famous Mandelbrot set, which is the most bizarre and magnificent geometric figure ever made by mankind, once known as the "fingerprint of God". Scanning on the set, the data obtained is converted into the tones on the piano keyboard, so as to express the structure of the Mandelbrot in a musical way.

In fact, fractal music has become the most exciting new field of music research. B. B. Mandelbrot, the father of fractal geometry, also believes that the inspiration for creating beautiful music can be found from the perspective of fractals. Harlan J. Brothers describe his original research on fractal music in the chapter "The Nature of Fractal Music" in the book *Benoit Mandelbrot: A Life in Many Dimensions*. We can access to some interesting fractal music on his personal website (Fig. 9.1).

Fractal music is a perfect combination of rational mathematics and perceptual music. It is worth more exploration of scientists and artists for the mysterious charms hidden behind beautiful things.

Source from:

[1] Fractal Music. BROTHER TECHNOLOGY [EB/OL]. http://www.brotherstechnology.com/math/fractal-music.html.
[2] Fractal Art in Rhythm [EB/OL]. http://www.guokr.com/article/118941/.
[3] Daniel J. Levitin, Parag Chordia and Vinod Minon: Musical rhythm spectra from Bach to Joplin obey a 1/f power law, *Proceedings of the National Academy of Sciences*, 2012, 109(10): 3716–3720.

9.1 Overview of Complexity Science

9.1.1 The Evolution of Complexity Science

After the 1970s, the non-linear science, represented by dissipative structure theory, synergetics, catastrophe theory, chaos, fractal theory, hypercycle theory, etc., has made major achievements, shaking the determinism which ruled physics, geometry and mathematics for years. These non-linear theories constitute an important methodology for studying complexity science. Complexity science is a science that explores the source and performance of the complexity of a complex system, and studies the mechanism and general laws of the evolution of the complex system that generate overall emergence under the interaction of many subsystems and the environment. Complexity science involves many fields and disciplines such as nature, physics, biology, society, economy, environment, ecology, engineering, etc. It breaks the traditional thinking mode of linearity, equilibrium, and simple reductionism in classical system science, and establishes a new thinking model of non-linearity, non-equilibrium and holism. Complexity is the core of system science research. It refers to the complex dynamic behaviors and integrality of open complex systems due to multiple subsystems, types, hierarchical structures, and uncertain factors, which leading to the interaction of the system with the environment during the evolution process. Complexity is the behavioral characteristic of complex systems, which are diversity, multi-level, integrality, openness, non-equilibrium, nonlinearity, dynamics, uncertainty, and self-organization. Prigogine pointed out: "Complexity exists at all levels. The complexity of different levels is not only different, but also unified. The physical level already has the minimum complexity. Complexity is the product of self-organization, and the system evolves into dissipation. The structure produces the complexity of the physical level, and on this basis, a higher level of complexity is produced through higher forms of self-organization." Holland pointed out: "The population has complexity in the evolution process, and it continues to improve its adaptation to the environment. The individuals in a population can be evaluated through adaptation, that is, adaptation creates complexity."

Complexity science research mainly goes through three stages: Edgar Morin's theory, Prigogine's Brussels School, and SFI (Santa Fe Institute)'s theory. Edgar Morin was the first to put forward complexity research and proposed the concept and theory of "complexity paradigm", discussing complexity science from the perspective of social science and philosophy, distinguishing the relationship between system and whole, and proposing general 13 ways and principles to "challenge complexity". In 1969, Belgian scientist Prigogine proposed the theory of dissipative structure, and put forward the concept of "complexity science" in *Order out of Chaos* published by him and Isabelle Stengers. The self-organization theory school represented by him has become the mainstream of complexity scientific research at this stage. The characteristics of this school's research are the large number of system elements, and the elements have their own and independent movements. Self-organization theory has become the core theory for studying system complexity, and the theories such as synergetics, catastrophe theory, and hypercycle theory have studied system complexity from different angles. SFI was established in New Mexico, USA in 1984, bringing together experts in different fields and disciplines to carry out interdisciplinary research and explore the commonalities between different complex systems. SFI's theory believes that the complexity of things develops from simplicity and emerges in the process of adapting to the environment. All complex systems have the ability to achieve a certain balance between order and chaos. SFI's theory calls the economy, ecology, immune system, nervous system, and computer network complex adaptive systems, and believes that there are certain general laws that control the behavior of these complex adaptive systems. The academic leader of SFI Murray Gell-Mann believes that the focus of complexity research is not the complexity of the object or the environment, but the complexity of the subject itself, that is, the subject's complex resilience and the corresponding complex structure.

Many famous scholars have emerged in the research of complexity science in China, such as Qian Xuesen, Cheng Siwei, Dai Ruwei, Wei Yiming, Zhang Yutai, Wei Yu, Yu Jingyuan, etc. They have made many achievements in the field of complexity science. The core of China's complexity research is Qian Xuesen. As early as the mid-1980s, he had insights into the importance of this new scientific direction. He gathered a group of forces through systematic seminars and started with the open theory of complex giant systems. The pioneer of complexity research in China is called Qian Xuesen School. Qian Xuesen approached complexity research from two aspects. One is to solve major national practical problems, such as major economic decision-making issues in army building and national system reform; the other is to establish basic science-level systems theory, that is, the work of systems science. Its concept formation has gone through three steps: giant system, complex giant system, open complex giant system. Cheng Siwei believes that complexity is mainly manifested in five aspects: (1) the network connection between the units, (2) the multi-level and multi-functional structure, (3) the ability to learn to reorganize and improve itself, (4) open to the environment and the interaction with it, and (5) dynamics and some degree of self-predictability." The physics-affair-humanity methodology advocated by Gu Jifa and Zhu Zhichang to solve complex problems has Chinese cultural characteristics. Xie Nanbin uses entropy and dissipative structure theory to analyze the evolution of the higher education system and proposes. In order to improve the order of the system by controlling the entropy change, it provides theoretical and decision-making thinking for the management of our higher education system. Huang Xinrong believes that complexity science has laid a solid scientific foundation for the birth of big data technology, and the big data technology is the continuation and technical realization of complexity. Zhang Chenhong and others analyzed and explained the 2008 U.S. financial crisis as a manifestation of the sudden change of the system in the self-organization process of the U.S. financial system based on the theory of system self-organization criticality; based on organizational criticality theory, catastrophe and chaos theory, they conducted fractal test, statistical analysis and fractal dimension calculation of the logarithmic return series of the US S&P 500 index from January 2006 to April 2009, revealing three differences before and after the outbreak of the US financial crisis. Hu Yingyue discussed the system crisis detection method from the brittleness theory of complex systems and boundary shell theory, combining extenics, frame pair analysis and other knowledge.

Management is complex, and its complexity is manifested in the complexity and uncertainty of the environment, the complexity of the multi-level, multi-unit, multi-function and multi-objective of the organizational system, the nonlinearity and non-determinism of prediction, decision-making, and control. The complexity of self-organization and adaptation process that interact with various factors and interacts with complex environments. In the actual operation of the enterprise, there are the complexity of the operation law of the management object, the complexity of the management structure, the complexity of the concept, understanding, psychology and behavior of the management subject, the complexity of nonlinearity and chaos in management, the complexity of management control, and the management philosophy, the complexity of the evolution process of material-based management, human-based management and knowledge-based management, and the complexity of the integration of management rationality and human nature. The development trend of management science, with systemization, socialization, internationalization, flexibility, virtualization, and intelligence etc., all makes its unprecedented challenges.

9.1.2 Characteristics of Complex Systems

1. Simplicity and complexity

Complexity and simplicity are relative concepts, which are high-level generalizations of two different attributes of things or laws. Simplicity means that things or laws have universal, basic, and invariable common attributes, that is, commonality; complexity refers to things or laws that have special, diverse, changing, and individual attributes, that is, individuality. Simple things or laws can be transformed into complex things or laws when certain conditions are met. With further changes in conditions, they may become simple things or laws again.

2. Definition of complex system

Relative to the simple system and the nonlinear system, The complex system is composed of many interacting subsystems. The overall function or characteristics of the system cannot be obtained from the functions or characteristics of the subsystems. The so-called interaction refers to the interaction between the subsystems in countless possible ways, so that the complex system emerges the overall function that all the subsystems do not have. The research objects of complex systems include natural phenomena, physical phenomena, biological phenomena, life phenomena, ecological phenomena, social phenomena, economic phenomena, etc. The components of a complex system are generally non-homogeneous and have a multi-level structure. Not only do various parts interact with each other, but also complex nonlinear effects exist between subsystems and levels.

3. Characteristics of complex systems

(1) Nonlinearity

Part or all of the components that make up a complex system must be of nonlinearity. The essence of non-linearity refers to the mutual influence, mutual restriction and interdependence of things, rather than unilateral influence. It is the nonlinear effect that causes the bifurcation, catastrophe, and chaos of the system.

(2) Multi-level

Complex systems present the characteristics of multi-level and multi-function in structure, and each level constitutes a component of the previous level. Generally, there is no superposition principle among the multiple levels of a complex system. Each time a new level is formed, new properties will emerge. Under normal circumstances, the more complex the system, the more levels, and the level of hierarchy is a basic characteristic describing the complexity of the system.

(3) Emergence

Emergence is an integral characteristic of the evolution and evolution of complex systems. The so-called emergence refers to the interaction between the components of a complex system to form a complex structure. While expressing the characteristics of the components, it also transmits the newly generated characteristics as a whole. That is to say, once many parts form a system in a certain way, it will produce attributes, characteristics, behaviors and functions that the system as a whole has but parts or the sum of parts do not have, and once the whole is reduced to irrelevant parts, then these attributes, characteristics, behaviors and functions no longer exist. We call the emergence of complex systems that these high-levels possess but no longer exist after being restored to low-levels.

(4) Adaptation

Complex systems have evolutionary characteristics. Systematic evolution means that the components, scale, structure, or function of the system will self-adjust and adapt to changes in internal and external environments over time in a direction that is conducive to its own existence. In the process of constantly adapting to the environment, the system tends to become more and more complex.

(5) Openness

The complex system is open, which refers to the exchange of matter, energy, information, etc. between the system itself and the system around it, so that the components of the system, as well as the system itself and the environment, maintain interaction, and promote the system to continuously adapt to changes in the environment.

(6) Dynamics

Complex systems are dynamic, in constant evolution, and have adaptive and evolutionary capabilities. In a complex system, a large number of nonlinear effects produce a wealth of variable dynamic phenomena, and tend to highlight a certain behavior among many different behaviors, or a certain behavior dominates, or produce a complex overall emergence. In a dynamic system, when the parameter changes pass a certain critical value, a so-called "crisis" occurs. When this kind of catastrophe occurs, the chaotic attractor may be destroyed, new attractors may appear, multiple attractors may appear, and fixed point may occur. The system is in a complex state due to "crisis" at certain critical parameters, and constantly undergoes this dynamic evolution and becomes a path to chaos.

9.1.3 Dissipative Structure

Dissipative structure theory, proposed by Belgian physicist Prigogine in 1969 and developed rapidly, is to study the structure and characteristics of non-equilibrium open systems. It is hailed as "one of the brilliant achievements" in the 1970s. At present, this theory has been applied to many aspects of natural and social research, and a series of important results have been achieved. Dissipative structure theory studies the mechanism, conditions and laws of an open system's transformation from chaos to order. The theory believes that an open system far from the equilibrium state, when the external conditions or a certain parameter of the system changes to a certain critical value, through the fluctuations, that is, the non-equilibrium phase transition, it is possible to change from the original chaos and disorder. The state transforms into a new state that is ordered in time, space, or function. This kind of macro-ordered structure formed far away from the equilibrium non-linear region requires constant exchange of material and energy with the outside world to form or maintain a new stable structure. Prigogine's dissipative structure theory breaks through the traditional physics framework of closed systems and balanced structures as the main research objects. It takes open systems far from the balance as the research objects. The research system exchanges energy and matter with the outside world to maintain a stabilized macro system structure, that is, the self-organization phenomenon of the macro system in an non-equilibrium state. Dissipative structure theory unifies the second law of thermodynamics and Darwinian evolution in a broader and more general theory.

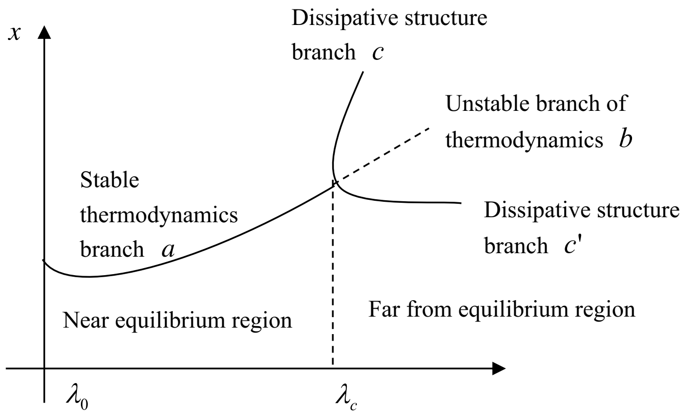


Fig. 9.2 System bifurcation phenomenon

In thermodynamics, the control parameter λ is used to describe the degree to which the system deviates from the equilibrium state, and the state parameter x is used to describe the state of the system. As shown in Fig. 9.2, λ_0, λ_c are two threshold points. When $\lambda < \lambda_0$, the system corresponds to the equilibrium state; when $\lambda_0 < \lambda < \lambda_c$, the system corresponds to the near equilibrium state, which is in the stable part of the thermodynamic branch; when $\lambda > \lambda_c$, the system corresponds to the far from equilibrium state, and the thermodynamic branch loses stability. In the unstable branch b, the influence of fluctuations on the system is continuously amplified through nonlinear coherence and chain reactions. The system may undergo abrupt changes, jumping from the unstable branch b to the branches c and c', and these two branches with orderly structure are called the dissipative structure branches, and maintain stability by exchanging matter and energy with the outside world, that is, the dynamic non-equilibrium structures.

The formation of the dissipative structure requires certain conditions, as follows:

(1) Being an open system. The second law of thermodynamics states that the entropy of an isolated system cannot be reduced, and the result of its evolution must be to reach the equilibrium state of maximum entropy. Therefore, the necessary condition for the system to become orderly is that the system is openness. Only when an open system exchanges material energy with the outside world and introduces negative entropy flow from the outside world to offset its own increase in entropy, can the total entropy of the system gradually decrease, and it is possible to move from chaos to order. Isolated systems cannot produce dissipative structures. According to the second law of thermodynamics, the entropy change of the system can be divided into two parts: $dS = d_iS + d_eS$, and here d_iS represents the part that changes the entropy of the system due to internal reasons in the system, and has nothing to do with the outside world. For an open system, as long as the external influence on

the entropy change of the system is not considered, only the influence of the internal mechanism of the system on the entropy change is analyzed. No matter what external effect, there is still $d_i S > 0$. $d_e S$ represents the part of the system whose entropy changes due to the connection between the system and the outside world. If there is no connection with the outside world, $d_e S = 0$, otherwise $d_e S \neq 0$. The total entropy change of the system is the sum of two parts. In an open system, only $d_e S < 0$ and $|d_e S| > d_i S$, then $dS < 0$. The entropy of the entire system decreases, and the system transforms from chaos to order.

(2) Far from equilibrium. Prigogine's principle of minimum entropy proposes that only when the system is far from the equilibrium state and in the nonlinear region of force and flow, can it evolve into an ordered structure and produce a sufficiently large negative entropy flow. The open system leaves the equilibrium state under the influence of the outside world, and gradually increases the degree of openness. The outside influence on the system is strengthened, and the system is gradually pushed from the near equilibrium state to the far from equilibrium state. In this case, it is possible for the system to form an ordered structure, that is, being far from the equilibrium state is a necessary condition for the system to form a dissipative structure.

(3) Of nonlinear interactions among the various subsystems in the system. Through the non-linear interaction between the subsystems, the synergy and coherence effects between the subsystems are generated, and new properties emerge, and the system can change from chaos to order. Only when the nonlinear term is saturated, can it stabilize to the dissipative structure state.

(4) Order out of fluctuations. Fluctuation refers to the deviation of a certain variable and behavior in the system from the average value, which causes the system to leave its original state or track. When the system is in a stable state, the fluctuation is a kind of interference, and the system has anti-interference ability and maintains the original stable state; when the system is in an unstable critical state, the fluctuation will cause the system to transition from an unstable state to a new orderly state. The macro order that appears after instability is determined by the fastest growing fluctuations.

9.1.4 Synergetics Theory

1. Principles of Synergetics

The term synergetics is derived from the Greek, which means "the study of coordination and cooperation". It reveals the common principle of the macro-variation of the structure of many complex systems. The founder of synergetics is German physicist Hermann Harken, whose main ideas stem from his research on laser theory. He discovered that many atoms that emit light independently of each other and the photoelectric fields they generate can produce monochromatic light-laser with the same phase and direction under certain constraints. Harken obtained general principles from his research on lasers, which were used to explain other self-organization phenomena. He found that the evolution of self-organizing systems from chaos to order, whether they belong to a natural system or a social system, is the result of the synergy between a large number of subsystems, and can be carried out with similar theoretical schemes and certain mathematician models.

Harken pointed out: "The reason of calling the discipline synergetics is that, on the one hand, the object it studies is the combination of many subsystems (usually of the same type or of several different types) to produce a macro Structure and function on a scale; on the other hand, many different disciplines collaborate and discover the general principles governing self-organizing systems". Synergy refers to the coordination, cooperation or synchronization of many subsystems. Synergetics studies that how the parts of a complex system compete and cooperate to form an overall self-organizing behavior, and explores the general principles governing the synergy of subsystems near the turning point where the macroscopic state of the system undergoes a qualitative change.

In natural science, order is the determination of the relationship between two elements. Being of order refers to the regular connection or transformation between the various elements within a thing and the thing; that is, there is a mathematical partial order relationship between the subsystems in the system. Being out of order refers to the chaotic and irregular combination of various elements within things or between things, and the irregularity in the transformation of movement. Order parameter is a concept introduced by the physicist Lev Davidovich Landau to describe the continuous phase transition. It is used to indicate the emergence of new structures and to distinguish the type and degree of order of the continuous phase transition and certain phase transition ordered structures. Haken borrows the concept of order parameter instead of entropy as a criterion to measure whether the system is orderly. Order parameter and dominance principle are two central concepts of synergetics. The order parameter is produced by the cooperation of subsystems. It is the parameter of the order degree of the macroscopic overall mode of the collective movement of a large number of subsystems. It is the macroscopic parameter introduced to describe the overall behavior of the system. On the one hand, the order parameter is the product of a large number of subsystems in the system. On the other hand, once the order parameter is formed, it will dominate the subsystem and dominate the overall evolution of the system. The order parameter is not only the representation and measurement of the cooperation effect of the subsystem, but also the measurement of the overall order of the system. The size of the order parameter can be used to mark the degree of macro order. When the system is out of order, the order parameter is zero. When the external conditions change, the order parameter also changes. When the critical point is reached, the order parameter increases to the maximum. At this time, a macro-orderly organized structure appears. Haken believes that no matter what system, if a parameter changes from chaos to order during the evolution of the system, and can indicate the formation of a new structure and reflect the degree of order of the new structure, it is an order parameter.

In the critical process of the evolution of the system from chaos to order, the relative change speed of different parameters is different, often by several orders of magnitude. Based on this, the various subsystems or parameters in the system are divided into two kinds of variables, namely fast variables and slow variables. Slow variables dominate fast variables and determine the evolution of the system. Slow variables and fast variables cannot exist independently. Slow variables make the system break away from the old structure and tend to the new structure; while fast variables make the system stable on the new structure. Along with the orderly evolution of the system structure, the two types of variables are interrelated and interacting, showing a synergic movement. Order parameters are often evolved from one or several slow variables.

The principle of dominance refers to the role of dominating or servicing subsystems after the formation of order parameters, and dominating the overall evolution of the system. The system changes from chaos to order and from order to a more complex ordering process, that is, in the process of repeatedly forming new self-organization, the order parameter always dominates other stable modes to form a certain structure or order, and the order parameter plays a leading role. If there is no dominating center of order parameters, the system will be in a state of chaos.

2. **Order parameter model**

In the evolution of the system from disorder to order, or from one order to another, the evolution behavior of different state variables near the critical point is significantly different. Some state variables are greatly damped, decay quickly, and have large numbers, which have no obvious impact on the evolution process; some state variables appear critically undamped, decay slowly, and have very few numbers, which play a major role in the evolution process. At the critical point of the evolution of the system, there are many state variables and different states. A large number of microscopic state variables in the system interact, and the resulting chaos and synergy make up and offset each other, ensuring the evolutionary trend of the overall macroscopic state. This stable evolutionary trend requires continuous dissipation of matter, energy and information to maintain.

Suppose the internal parameters of the system are: $q_1, q_2, q_3, \ldots, q_n$, it can be represented by *n*-tuple vector $\vec{q}$, then

$$\vec{q} = \{q_1, q_2, q_3, \ldots, q_n\}$$

The state change of $\vec{q}$ reflects the evolution of the state of the system, and the generalized Langevin equation is used to express the movement of $\vec{q}$ as

$$\dot{q}_i = K_i(\vec{q}) + F_i(t) \quad (i = 1, 2, \ldots, n) \tag{9.1}$$

$K_i(\vec{q})$ describes the nonlinear effect of system parameters, and $F_i(t)$ is the random fluctuation force formed by various small disturbances. When $F_i(t)$ is neglected, under the interaction of n parameters, the state equation of the system can be expressed as

$$\left.\begin{aligned} \dot{q}_1 &= -\gamma_1 q_1 + f_1(q_1, q_2, \ldots, q_n) \\ \dot{q}_2 &= -\gamma_2 q_2 + f_2(q_1, q_2, \ldots, q_n) \\ &\vdots \\ \dot{q}_n &= -\gamma_n q_n + f_n(q_1, q_2, \ldots, q_n) \end{aligned}\right\} \tag{9.2}$$

In Formula (9.2), $\{\gamma_i\}$ is the damping coefficient of each parameter, and $\gamma_i > 0$ plays a role of dissipation in the system. The variable set $\{q_i\}$ contains the order parameter u and other slow variables and a large number of fast variables. When the system approaches the critical point, the order parameter's ability to dominate other variables gradually strengthens, and other variables gradually obey the order parameter's control. The evolution of the system is dominated by the order parameter u, which dominates the movement of other parameters. Assuming that the internal nonlinear effect of the system finally produces the main sequence parameter u. Let $u = q_1$, then Formula (9.2) can be rewritten as

$$\left.\begin{aligned} \dot{u} &= -\gamma_1 u + f_1(u, q_2, \ldots, q_n) \\ \dot{q}_2 &= -\gamma_2 q_2 + f_2(u, q_2, \ldots, q_n) \\ &\vdots \\ \dot{q}_n &= -\gamma_n q_n + f_n(u, q_2, \ldots, q_n) \end{aligned}\right\} \tag{9.3}$$

When the system evolves to near the critical point, according to the principle of synergy, in Formula (9.3), $\gamma_1 \to 0$, let

$$\dot{q}_2 = \dot{q}_3 = \dot{q}_4 = \cdots = \dot{q}_n = 0$$

Formula (9.3) is transformed as

$$\left.\begin{aligned} 0 &= -\gamma_2 q_2 + f_2(u, q_2, \ldots, q_n) \\ 0 &= -\gamma_3 q_3 + f_3(u, q_2, \ldots, q_n) \\ &\vdots \\ 0 &= -\gamma_n q_n + f_n(u, q_2, \ldots, q_n) \end{aligned}\right\} \tag{9.4}$$

Solve Formula (9.4) and express other parameters in terms of order parameters u, and we get

$$q_i = g_i(u) \quad (i = 2, 3, \ldots, n) \tag{9.5}$$

When all the parameters are expressed by the order parameter, the expression is brought into the order parameter expression in Formula (9.3), and the order parameter equation is as follows

$$\begin{aligned} \dot{u} &= -\gamma_1 u + f_1(u, q_2, \cdots, q_n) = -\gamma_1 u + f_1(u, g_2(u), g_3(u), \cdots g_n(u)) \\ &= -\gamma_1 u + f_1(u) = \phi(u) \end{aligned} \tag{9.6}$$

Assume that multiple order parameters are generated in the system, and that the macro-order structure of the system is jointly determined by the cooperation of these order parameters. The order parameters compete and merge with each other. Once the influence of the external environment reaches a certain threshold, it may appear that only one order parameter alone dominates the evolution of the system, and the other order parameters evolve into servo quantities, and the structure determined by the main sequence parameters becomes the systemetic macrostate. From Formula (9.6), suppose the order parameter equation is as follows

$$\dot{u} = \phi(u) = (A - C)u - u^3 \tag{9.7}$$

The parameter A in Formula (9.7) represents the influence of the external environment on the system and is called the control parameter. Assume that the control parameters do not change in the critical area of the evolution of the system, then the evolution trend of the system will not change; once the control parameters are changed, the evolution trend and process of the system will change accordingly. C is a fixed state value, a critical value for the system to evolve under a certain environment and a certain stage of development. The structure of the enterprise undergoes fundamental changes near this critical value. u^3 is a nonlinear term. Let $\dot{u} = 0$, Formula (9.7) is changed to

$$0 = (A - C)u - u^3 \tag{9.8}$$

Solve Formula (9.8), and we get

$$u = 0 \text{ and } u = \pm\sqrt{A - C}$$

(1) When $A < C$, $u = \pm\sqrt{A - C}$ is an imaginary number, which has no meaning for the social system. And the only branch of the stationary solution is $u = 0$, which is stable. But for the system, the order parameter zero is meaningless for the system evolution.
(2) When $A > C$ and $u = \pm\sqrt{A - C}$ is a real number, Formula (9.7) is a nonlinear equation describing the order parameter, which is linearized for analysis

$$\frac{d\Delta u}{dt} = \phi'(u_0)\Delta u = -2(A - C)\Delta u \tag{9.9}$$

The solution of Formula (9.9) is

$$\Delta u = \Delta u_0 e^{-2(A-C)t}$$

$u = 0$ is still the stationary solution of the order parameter equation, but the solution is unstable. As shown in Formula (9.9), $(A - C)$ is positive, and any slight deviation from $u = 0$ will lead to a large deviation in the enterprise system. Under the action of the nonlinear term $-u^3$, the order parameter will be limited to a finite value not zero. The order parameter is the characterization of the degree of order of the system. The improvement of the order parameter contributes to the evolution of the system. Therefore, the solution of $u = 0$ cannot play a positive role in the transition of the system from one ordered state to another.

When $t \to \infty$, $\Delta u \to 0$, all $u = \pm\sqrt{A - C}$ solutions are stable. Under the action of the non-linear term $-u^3$, the sequence parameters after the instability of the system evolution branch do not diverge to infinity, but converge on the dissipative structure branch where the sequence parameter is zero.

(3) Once a certain element becomes the order parameter, the other state parameters of the servo are evolved with the order parameter as the core, and the behavior characteristics of all state parameters will more or less show the behavior characteristics of the order parameter.

3. Application of synergetics

Synergetics has a wide range of applications. It has achieved important achievements in many aspects such as physics, chemistry, biology, astronomy, economics, sociology and management science. Aiming at cooperation effects and organizational phenomena, synergetics can solve some systemic complexity problems, and it can be used to establish a coordinated organization system to achieve work goals.

In the field of business management, the first to propose the concept of synergy was the American strategic theory researcher H. Igor Ansoff. In his book *Corporate Strategy* published in 1965, he regarded collaboration as the one of the four elements of corporate strategy (product market scope, development direction, competitive advantage and synergy), and explain how the strategy based on the concept of synergy can link the company's diversified businesses like a "link", so that the company can make full use of its existing advantages to open up new development space. Synergetics can provide a unified theoretical basis for business management at a higher theoretical level. Using the ideas and methods of synergy, systematic research is conducted based on the cooperative law of management objects. The purpose is to achieve management synergy, that is, to form the overall emergence of the system, so that the result of the system is greater than the sum of the input elements. The synergy mechanism of modern enterprises is manifested in the coordination, cooperation and mutual promotion of various subsystems, thereby forming an orderly structure of the system. The synergy mechanism of enterprises is partially reflected in such aspects as strategy, supply chain, technological innovation, corporate culture, and logistics coordination. The strategic synergy is the concentrated manifestation of the coordination of other subsystems at the macro level, and the manifestation of the overall emergent synergy effect of the enterprise. Only on the basis of the optimization of strategic synergy can the synergy of other parts of the enterprise exert its effect.

In addition, in the field of natural sciences, synergetics is mainly used in physics, chemistry, biology and ecology. For example, in the aspect of ecology, the relationship between the growth and decline of predators and prey groups is studied; the application of synergy to study static face recognition; the synergy analysis method of the synthesis reaction kinetics of ethyl tert-amyl ether is proposed; the synergetics research on the law of concrete acoustic emission and visual simulation; Lorenz model and population dynamics in biological synergy, etc. In the field of social sciences, it is mainly used in sociology, economics, psychology and behavioral sciences. For example, the stochastic model of the formation of public opinion; the selection of supply chain partnership evaluation indicators; the independent innovation of regional manufacturing; the scientific development model of enterprises; the cultural integration of multinational mergers and acquisitions. In the field of engineering technology, it is mainly used in electrical engineering, mechanical engineering and civil engineering. For example, models and methods for rational allocation of water resources in river basins based on synergy theory; grid cascading reaction fault prevention model based on synergy theory; single-phase grounding fault line selection method for small current grounding systems based on synergy theory; the research on process of naval ship CGF (computer generated force) anti-submarine state identification based on synergetics.

9.2 Chaos Theory

9.2.1 The Development of Chaotic Dynamics

In the mid-nineteenth century, natural scientists studied the equilibrium state in thermodynamics, Brownian motion, Tyndall phenomenon, and the random collision of reaction groups in the reaction system, and found that these states are in chaotic and disordered states.

At the end of the nineteenth century and the beginning of the twentieth century, the famous French mathematician and physicist Poincaré discovered chaos when studying celestial mechanics, especially when studying the three-body problem, becoming the first scholar to discover chaos in modern science. In 1903, Poincaré proposed the "Poincaré Conjecture" in his book *Science and Method*. He organically combined topology and dynamics system and pointed out that there are periodic orbits in the three-body motion of the solar system. The gravitational interaction of the three bodies can produce surprisingly complex behaviors, and certain solutions of deterministic equations are unpredictable (Fig. 9.3).

Since the twentieth century, thermodynamics and statistical physics have experienced a development stage of equilibrium-near equilibrium-far from equilibrium.

Fig. 9.3 Poincaré, French mathematician and physicist

In the initial development of the theory of dissipative structure and synergetics, it focused on the study of how the system developed from chaos to order, and found some mechanisms and conditions for the system to develop from chaos to order. That is, the system should be open, in a state far from equilibrium, with nonlinear interactions within it, and state transitions must be realized through fluctuations. Based on this, Prigogine once came to the conclusion that "order out of chaos".

In the 1950s and 1960s, American meteorologist Edward Norton Lorenz discovered the regular behavior of a deterministic system when studying the orbits of weather changes, but at the same time discovered that the same system can appear nonperiodic under certain conditions. This is inconsistent with the authoritative views of the meteorological community at that time. In 1963, Lorentz simplified the complicated meteorological changes in the famous paper "Determinism Aperiodic Flow" and obtained a differential equation system, namely the famous Lorentz equation:

$$\begin{cases} \dot{x} = -\sigma(x - y) \\ \dot{y} = -xz + rx - y \\ \dot{z} = xy - bz \end{cases} \tag{9.10}$$

It is a completely definite third-order ordinary differential equation system. The three parameters in the equation are σ, r and b. He found that under the slight interference, the orbital changes in a certain space are extremely strong, which is mathematically called sensitive to initial conditions, that is, unpredictability of the topology, which means on the established orbit, under the slight interference, the motion orbit will have a huge deviation. In order to describe the extreme sensitivity of the chaotic complex system, Lorenz made an analogy: A small change in

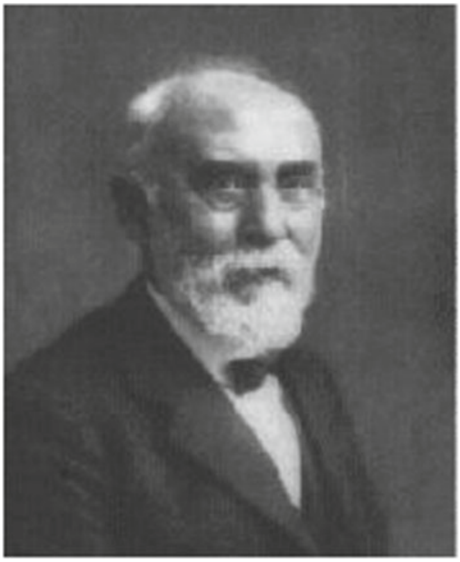


Fig. 9.4 Edward Lorenz, American meteorologist

airflow caused by a butterfly accidentally flapping its wings in a certain place in the southern hemisphere may turn into a tornado in a certain place in the northern hemisphere in a few weeks. This is the famous "butterfly effect". Lorentz's contribution to the study of chaos is that he reveals the basic characteristics of a series of chaotic motions, such as deterministic non-periodicity, sensitive dependence on initial conditions, and unpredictability of long-term behavior. He also discovered the first strange attractor in chaos research—the Lorenz attractor. He provided an important model for chaos research and was the first to conduct specific research using numerical calculation methods on a computer (Fig. 9.4).

In 1964, French astronomer Henon M. was inspired by the study of globular star clusters and Lorenz attractors, and proposed the Henon map:

$$\begin{cases} x_{k+1} = 1 + by_k - ax_k^2 \\ y_{k+1} = x_k \end{cases} \tag{9.11}$$

This equation is a non-integrable Hamiltonian system with 2 degrees of freedom, but when the parameter $b = 0.3$ and the parameter a is changed, it is found that the distribution of the system motion trajectory in the phase space becomes more and more random. Henon M. obtains the simplest kind of attractor called Henon attractor.

In 1971, French physicist D. Ruelle and Dutch mathematician F. Takens jointly published the famous paper "On the essence of turbulence", which was the first in academia to use chaos to describe a new perspective on the formation mechanism of turbulence to reveal the nature of turbulence. After analysis, they found that the dynamic system has a particularly complex new attractor, named Strange attractor, and introduced a dissipative system, proved that the motion related to this attractor is chaotic motion, leading the first path towards chaos.

Fig. 9.5 Tien-Yien Li

Fig. 9.6 James A. Yorke

In 1975, Chinese American scholar Tien-Yien Li and his mentor, mathematician James A. Yorke, published a famous article "Period Three implies chaos" in *The American Mathematical Monthly*, which deeply revealed the evolution process from order to chaos. Also born was the famous Li-Yorke theorem, describing the mathematical characteristics of chaos. Tien-Yien Li and Yorke took the lead in introducing the term "chaos" in dynamic research, establishing a central concept for this research field (Figs. 9.5 and 9.6).

In the field of biological research, population ecologists in particular have made special contributions to the establishment of chaos. In 1976, the American mathematical ecologist May R. published an article titled “Simple mathematical model with complex dynamics process” in the American journal *Nature*, in which he proposed the famous population (insect-population) equation, i.e. the Logistic model:

$$x_{k+1} = \mu x_k(1 - x_k) \tag{9.12}$$

The one-dimensional mapping model seems simple and is deterministic, but when the parameters change within a certain range, it has extremely complex dynamic behavior. May used numerical calculations to study the insect mouth model, and saw both regular cycle doubling and bifurcation phenomena and irregular “strange” phenomena. At the same time, he also found that stable periodic motion appeared in random motion. The establishment of Logistic model greatly promoted the development of chaos.

In 1978, the American physicist Feigenbaum M. J. published an article titled “Quantitative universality for class of nonlinear transformation” based on May’s research, and discovered the scale and universal constants in the period-doubling bifurcation phenomenon with his colleagues, when they were studying a unimodal map represented by Logistic map. This means that chaos has a solid theoretical foundation in modern science.

In the 1980s, scholars focused on studying how the system evolved from order to chaos, and the properties and characteristics of chaos, which led to the further development of chaos science. In 1981, F. Takens et al. proposed an experimental method to determine singular attractors based on Whitney’s topological embedding theorem. In 1983, Canadian physicist L. Glass published a famous article “Calculating the singularity of strange attractors” in the journal *Physics*, and proposed a method to calculate the fractal dimension of singular attractors in experimental systems, namely the G-P algorithm, creating a worldwide upsurge in calculating the dimension of time series. In 1983, the famous Chinese scientist Hao Bolin published an article titled “Bifurcation, chaos, strange attractors and others” (Fig. 9.7). In 1984, he edited and published the book *Chaos*. In 1986, the first China Chaos Research Conference was held in Guilin, Guangxi Province. The research of chaos science in our country received more and more attention. In 1987, P. Grassber et al. proposed the theory and method of reconstructing dynamical systems. By extracting chaotic characteristic quantities such as fractal dimension and Lyapunov exponent from time series, the research of chaos theory entered the stage of practical application.

Since the 1990s, more and more attention has been paid to the application of chaos science. Chaos and other disciplines penetrate and promote each other, such as mathematics, physics, astronomy, chemistry, engineering, electronics, information science, meteorology, cosmology, Geology, economics, biological sciences, etc. The study of chaos has not only promoted the development of other disciplines, but the development of other disciplines has promoted the in-depth study

Fig. 9.7 Hao Bolin

of chaos. Among them, the research on the control and synchronization of chaos has made significant progress in the 1990s. In 1990, Ott, Grebogi and Yorke, three physicists from the University of Maryland, jointly published the paper "Controlling chaos". In the paper, they put forward a comparative system based on infinitely many unstable periodic orbits embedded in chaotic attractors. And a rigorous parameter perturbation method (called OGY method), which only makes small perturbations to the system parameters and feeds them back to the system, realizing the stability of the system's orbit on a specific orbit expected in an infinite number of unstable periodic orbits. Some scholars have promoted the OGY method and verified its effectiveness with various experiments. For example, Ditto and Roy et al. used the OGY method to complete the chaos control experiment in the amorphous magneto-elastic strip system and the circuit system, and for the first time successfully realized the control of a fixed point in a physical system. Carrol and Pecore of NRL (United States Naval Research Laboratory) put forward the idea of chaos synchronization and realized the experiment of using chaos synchronization for secure communication. In addition, chaos has made a series of significant progress in the application research of intelligent information processing, computer graphics, signal detection, medical and biotechnology, and social economy. Chaos optimization is another important content of chaos research developed in recent years, which mainly includes two aspects of application in function optimization and combinatorial optimization.

9.2.2 Definition of Chaos

In 1975, Tien-Yien Li and Yorke first proposed the concept of "chaos" in the sense of modern science and gave a mathematical definition of chaos. The definition is defined from the perspective of interval map, based on the Li-Yorke theorem.

Li-Yorke theorem: Let $f(x)$ be a continuous self-map on $[a, b]$, if $f(x)$ has 3 periodic points, then for any positive integer n, $f(x)$ has n periodic points.

Li-Yorke chaos: If the continuous self-map f on $[a,\ b]$ satisfies the following conditions, it is called chaotic.

(1) The period of the period point of f has no upper bound.
(2) There is an uncountable subset $S \subset [a,\ b]$, there is no periodic point in S, and the following conditions are satisfied:
A. For any $x, y \in S$, we have

$$\lim_{n\to\infty} \inf\left|f^n(x) - f^n(y)\right| = 0; \tag{9.13}$$

B. For any $x, y \in S, x \neq y$, we have

$$\lim_{n\to\infty} \sup\left|f^n(x) - f^n(y)\right| > 0; \tag{9.14}$$

C. For any periodic point y of any $x \in S$ and f, we have

$$\lim_{n\to\infty} \sup\left|f^n(x) - f^n(y)\right| > 0 \tag{9.15}$$

In Li-Yorke chaos, the two limits in A and B indicate that the points x and y in the subset are quite scattered and concentrated, and the limit in C indicates that the subset does not approach any periodic point. The definition shows that for the continuous self-map f on $[a, b]$, if there is a periodic point with a period of 3, there must be a periodic point with a period of any positive integer, then chaotic phenomena must occur, and chaotic motion has three important characteristics: ① There are countable infinitely many stable periodic orbits; ② There is an uncountable infinitely many stable non-periodic orbits; ③ There is at least one unstable aperiodic orbit. The Li-Yorke definition is one of the more widely spread mathematical definitions of chaos, but the definition only shows the "existence" of chaos mathematics, and does not describe their measurement and stability.

In 1989, Devaney chaos was more intuitive, easier to understand, and had a wider impact. The definition is as follows:

Let f be the map on the metric space V, if $f : V \to V$ satisfies the following conditions, then f is chaotic on V.

(1) f has a sensitive dependence on initial conditions, there is $\delta > 0$, for any $\varepsilon > 0$ and any $x \in V$, there are y and a natural number n in the I field of ε of x, so that $d(f^n(x), f^n(y)) > \delta$;
(2) f has topological transitivity, for any open set X, Y on V, there are $k > 0$, $f^k(X) \cap Y \neq \Phi$;
(3) The periodic point set of f is dense in V.

The sensitive dependence on initial conditions is also called unpredictability, which means that no matter how close the distance between x and y is, the distance d

between the two will reach greater than δ under the multiple actions of f, and such y exists in any small neighborhood of x, that is, any small initial error will cause a large enough difference in actual results after multiple iterations. Topological transitivity means that the neighborhood of any point will cover the metric space V under multiple actions of f. The dense periodic point set in V means that the chaotic system has regular components, that is, there are dense periodic points.

9.2.3 Characteristics of Chaos

Chaotic motion is a complex motion state unique to deterministic nonlinear dynamic systems. It is a non-periodic and bounded dynamic behavior sensitive to initial conditions. It appears in some dissipative systems, non-integrable Hamiltonian systems and nonlinear discrete mapping systems. Chaos is a unique attribute of deterministic nonlinear systems. Chaotic motion is a random process that appears in deterministic systems. Non-linearity is the minimum condition for dynamic systems to generate chaos. No linear system can produce chaos. Chaos has the following main characteristics:

1. **Boundedness**

Chaos is bounded, and its trajectory is always limited to a certain area, which is called the chaotic domain of attraction (DA). No matter how unstable the chaotic system is, its trajectory will never go out of the chaotic domain of attraction. Therefore, from the overall perspective, the chaotic system is bounded and stable.

2. **Linear**

For a system with a dissipative structure, when the nonlinearity is further enhanced, chaos will generally appear. This means that nonlinearity leads to chaos, and nonlinearity is the most basic condition for chaotic phenomena in dynamic systems. At the same time, chaos is a non-periodic and bounded dynamic behavior in a nonlinear dynamic system. For some parameter values, a non-periodic dynamic process will occur under almost all initial conditions.

3. **Internal randomness**

Chaotic motion is an uncertain motion generated by a deterministic nonlinear system. Generally, a deterministic nonlinear system can only produce random output if random input is applied. The chaotic system produces a state of motion similar to random after applying deterministic input. This is obviously generated spontaneously within the system, which is called internal randomness, which reflects the local instability of the chaotic system. Local instability means that the behavior of certain aspects of the system's motion (such as certain dimensions) strongly depends on the initial conditions. Although the law of the system is deterministic, the resulting chaotic behavior is difficult to determine. The probability distribution function of any region in the attractor is not zero, and it has internal randomness.

4. **Ergodicity**

Ergodicity means that the chaotic signal can traverse all the states in a finite time without repeating its own rules in the chaotic domain of attraction, and pass through each state point.

5. **Fractal dimension**

Fractal dimension is used to describe the behavior characteristics of the system motion orbit in phase space. The motion trajectory of the chaotic system in space is folded infinitely in a certain limited area to form a special curve. The dimension of this kind of curve is not an integer, but a fraction, called fractal dimension. The fractal dimension of chaos means that a point set in n-dimensional space has an infinitely fine structure, and has self-similar and overall similar properties at any scale, that is, chaotic attractors have self-similar properties. The structure of various attractors is characterized by fractal dimensions, that is, it has a non-integer dimension less than the dimension n of the space.

6. **Sensitivity dependence on initial conditons**

The nature of the large difference in output due to the small difference in the input initial conditions is called the sensitive dependence on initial conditions, which is the cause of chaos. In other words, when the system is in a chaotic state, if the initial conditions change slightly, it will cause a huge difference in its motion behavior. In other words, even if starting from an arbitrarily close initial state or two arbitrarily close initial conditions, after a finite time, their orbits will be exponentially separated from each other, that is, the long-term behavior of a chaotic system is unpredictable. This characteristic reflects the extreme sensitive dependence of the chaotic orbit on initial conditions.

7. **Universality**

Universality means that different systems will show some common characteristics when they tend to be chaotic, and they will not change according to specific system equations or system parameters. It usually includes the universality of the structure and the universality of the measurement. The universality is mainly embodied in several universal constants of chaos (like Feigenbaum constant), which is a manifestation of the inherent regularity of chaos. The Feigenbaum constant is obtained through the study of the Logistic equation, which reflects a general dynamic invariance of the system when it tends to chaos. For example, in a class of nonlinear map where period-doubling bifurcation enters chaos, the bifurcation speed and height have Feigenbaum δ constant and Feigenbaum α constant respectively. These constants are the universal numerical characteristics of the system going to chaos through period-doubling bifurcation.

8. **Regularity**

The chaotic motion may seem chaotic, but in fact it has its own internal laws, and the orbit traverses all states without repetition in a certain range on its own terms.

9.2.4 Lyapunov Exponent

Lyapunov exponent is an important index describing the stability of a dynamic system in a steady state. The spectrum of Lyapunov exponents (LEs) is a set of average quantities that reflect the shrinkage and diffusion characteristics of the initial phase trajectory in different directions over time. It is used to measure the exponential attraction or separation of two adjacent trajectories over time when the initial conditions in the phase space are different. Generally speaking, the value of the Lyapunov exponent can be used as a basis for judging the type of stable state behavior of the system, such as equilibrium point, limit cycle, chaos, or hyperchaos.

For one-dimensional map

$$x_{n+1} = f(x_n) \tag{9.16}$$

Assuming that the initial point is x_0 and the adjacent point is $x_0 + \delta x_0$, the distance between them after n iterations

$$\delta x_n = \left| f^{(n)}(x_0 + \delta x_0) - f^{(n)}(x_0) \right| = \frac{df^{(n)}(x_0)}{dx} \delta x_0 \tag{9.17}$$

When $\left|\frac{df}{dx}\right| > 1$, the initial point x_0 is separated from the adjacent point $x_0 + \delta x_0$ after n iterations; when $\left|\frac{df}{dx}\right| < 1$, the initial point x_0 is close to the adjacent point $x_0 + \delta x_0$ after n iterations. In chaotic motion, the trajectories of the system are separated and close to each other, so the value of $\left|\frac{df}{dx}\right|$ is constantly changing. In order to observe the degree of separation or closeness of the two trajectories as a whole, the iterative process needs to be averaged. Suppose the exponent in the exponential separation caused by each iteration is λ, then the distance between the two points that are separated by ε after n iterations is

$$\left| f^{(n)}(x_0 + \varepsilon) - f^{(n)}(x_0) \right| = \varepsilon e^{n\lambda(x_0)} \tag{9.18}$$

$$\lambda(x_0) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ln\left|\frac{df^{(n)}(x)}{dx}\right|_{x=x_i} \tag{9.19}$$

The $\lambda(x)$ in the formula is called Lyapunov exponent, which represents the tendency of exponential speed separation or closeness between adjacent discrete points caused by each iteration during multiple iterations. Lyapunov exponent, as the result of long-term average along the track, is an overall characteristic, and its value is a real number. One-dimensional map has only one Lyapunov exponent, which may be greater than, equal to, or less than zero. For example, for a stable fixed point, there is $\left|\frac{df}{dx}\right| < 1$ or for a stable period n, then $\lambda < 0$; for period-doubling bifurcation points, there is $\left|\frac{df}{dx}\right| = 1$, then $\lambda = 0$; for chaotic motion, due to the sensitive dependence on initial conditions, then $\lambda > 0$. The positive Lyapunov exponent ($\lambda > 0$) indicates that the motion orbit is unstable in every part, the adjacent orbital indices are separated rapidly, and the orbit is folded repeatedly under the effect of the overall stability factor, which proves that the system is in a chaotic state. The negative Lyapunov exponent ($\lambda < 0$) indicates that the phase volume shrinks; the orbit is locally stable and insensitive to initial conditions, and corresponds to periodic orbital motion. When $\lambda = 0$, it means that the system corresponds to the stable boundary.

9.2.5 Logistic Map and Tent Map

1. **Logistic map and its parameter characteristics**

Logistic map was proposed in 1976 by the mathematical ecologist R. May in a later influential review, called the insect population model. Later, after Feigenbaum's research, it is concluded that once the period-doubling bifurcation occurs in the system, it will inevitably lead to the occurrence of chaos.

The expression of Logistic map is as follows:

$$x_{n+1} = f(x_n, \mu) = \mu x_n(1 - x_n) \tag{9.20}$$

Formula (9.20) is a unimodal map and the most representative one-dimensional non-linear map. It describes the change of the number of biological groups with generations, where n represents time, x_n represents the number of births in the nth generation, and x_{n+1} represents the $n+1$th birth number of the generation. $x \in [0, 1]$, μ are the control parameters, $0 \le \mu \le 4$ (Fig. 9.8).

The Lyapunov exponent of Logistic map is

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left|f'(x)\right| \tag{9.21}$$

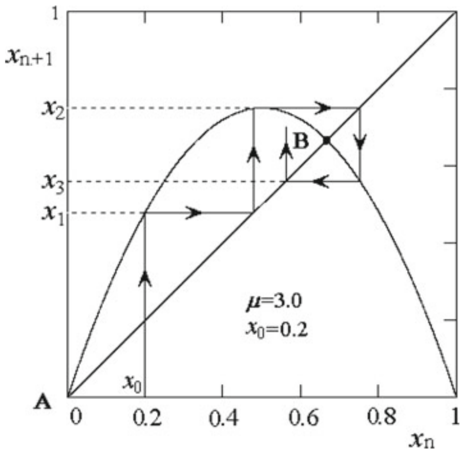


Fig. 9.8 Logistic map

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \ln|\mu - 2\mu x_i| \tag{9.22}$$

(1) When $\mu \in (0, 1]$, λ tends to 0 from a negative value, and the value of x converges to a fixed point.
(2) When $\mu \in (1, 3]$, $\lambda < 0$ and x converges to $(1 - \frac{1}{\mu})$. There are two fixed points, as shown in A and B in Fig. 9.8. When $\mu = 2$, there is a super stable point.
(3) At $\mu \in (3,\ 3.571448]$, the system state is a period-doubling bifurcation, and λ is in a cyclic process from 0 to $-\infty$ and then to 0. Period-doubling bifurcation is a special form of bifurcation, which is the process of continuously dividing into two: the unstable period-1 burstling is changed into period-2, then changed into period-4, then changed into period-8…; when period-2^{n-1} is unstable, it is changed into period-2^n. This process is called period-doubling bifurcation. Such infinite bifurcation will inevitably lead to the emergence of chaotic orbits, which is one of the most typical ways to chaos (Fig. 9.10).
(4) At $\mu \in (3.571448,\ 3.82842]$, the system is in a state of intermittent chaos. At this time, $\lambda > 0$, in some windows, $\lambda < 0$ (Fig. 9.11).
(5) At $\mu \in (3.9,\ 4]$, the system is in a chaotic state, $\lambda > 0$. When $\mu = 4$, λ reaches the maximum value of 0.69 (Fig. 9.12).

2. **Tent map and its parameter characteristics**

Tent map is a piecewise linear one-dimensional map with uniform probability density, power spectral density and ideal correlation characteristics. The mathematical expression is

$$x_{n+1} = \alpha - 1 - \alpha|x_n|,\ \ \alpha \in (1,\ 2] \tag{9.23}$$

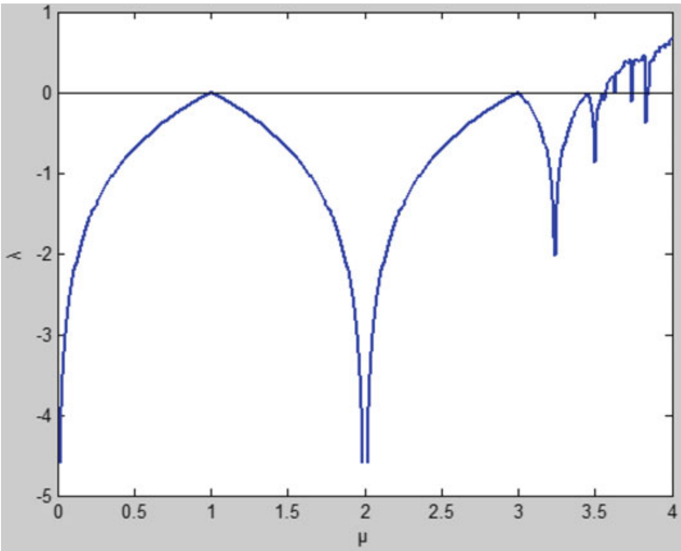


Fig. 9.9 Calculation results of Lyapunov exponent with parameter μ

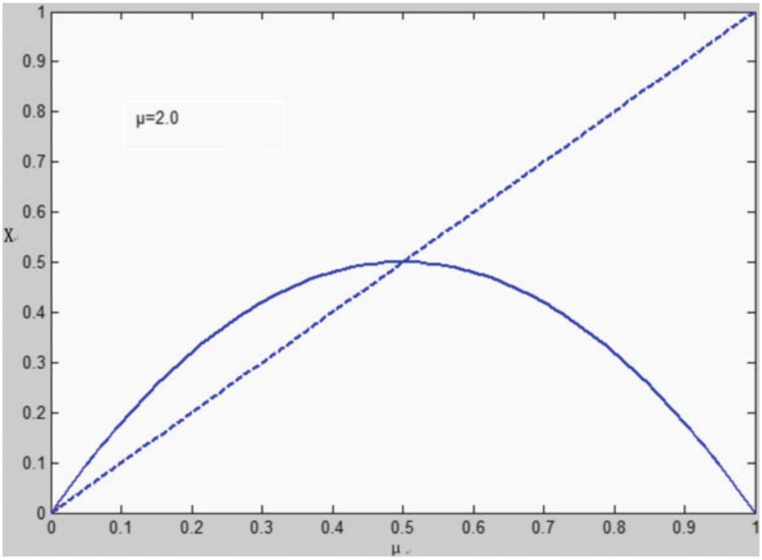


Fig. 9.10 Logistic iteration at $\mu = 2$

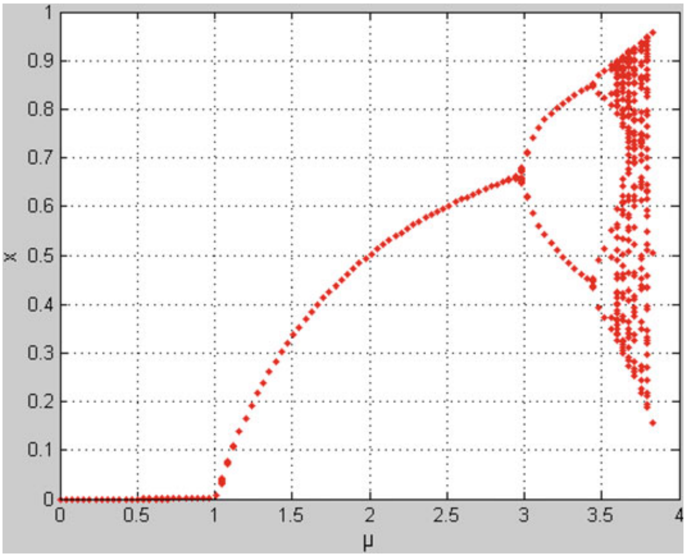


Fig. 9.11 Logistic bifurcation diagram at $\mu = 3.83$

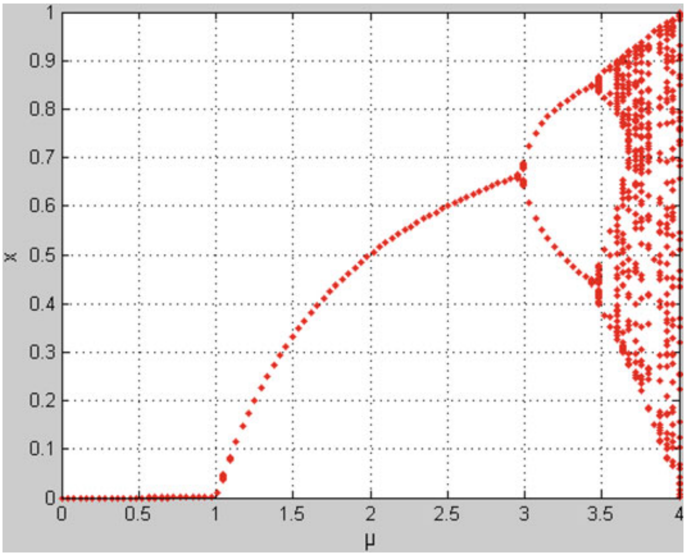


Fig. 9.12 Logistic bifurcation diagram at $\mu = 4$

The Lyapunov exponent of Tent map is

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \ln\left|\frac{dx_{n+1}}{dx_n}\right| = \ln\alpha \tag{9.24}$$

When $\alpha \le 1$, $\lambda \le 0$, the system is in a stable state; when $\alpha > 1$, $\lambda > 0$, the system is in a chaotic state; when $\alpha = 2$, $\lambda_{\max} = \ln 2$, is the central Tent map, and its mathematical expression is

$$x_{k+1} = \begin{cases} 2x_k, & 0 \le x_k \le 0.5 \\ 2(1 - x_k), & 0.5 < x_k \le 1 \end{cases} \tag{9.25}$$

Tent map has a simple structure and good traversal uniformity. It is more suitable for arithmetic processing of large-scale data sequences, and the iteration speed is faster than Logistic map. Tent map has two fixed points: 0 and 2/3. Because of the slope $|f'(x)| = 2 > 1$ of each point, these two fixed points are both unstable. When $x_k = 0, 1/2, 1$ is taken in formula (9.25), x_{k+1} is 0, 1, 0 respectively. This shows that the original value of x_k is in the two intervals from 0 to 1/2 and from 1/2 to 1. After one iteration, the value of x_{k+1} becomes the interval from 0 to 1, which is equivalent to lengthening the image. This iterative process continues, and its elongation characteristic eventually leads to the exponential split of adjacent points, which produces a sensitive dependence on initial conditions, while the folding process keeps the track bounded.

9.2.6 Chaos Prediction

The chaotic sequence is not completely unpredictable. Due to the non-linearity, boundedness, regularity and other characteristics of the chaotic system, it shows order and regularity in a certain area. Therefore, the chaotic phenomenon is possible in the short term. Forecast, but not suitable for long-term forecasting.

1. Chaotic time series

Chaotic time series is a common chaotic discrete situation in chaotic systems. It is a time series with chaotic characteristics generated by a chaotic model. Since the reconstruction of phase space proposed by Packard et al. (1980), the approach of studying chaos from time series has been widely recognized, and the prediction of chaotic time series has gradually received attention. In chaotic time series forecasting, the commonly used forecasting methods mainly include: global method, local method, weighted zero-order local method, weighted first-order local method, time series prediction method based on Lyapunov exponent, and time series based on neural network Forecast methods, etc. At present, the prediction of chaotic time series is mainly used in power system short-term load forecasting, stock market forecasting, rotor remaining life forecasting, weather forecasting, hydrological forecasting, slope displacement forecasting, etc. Its application prospects are very broad.

2. Reconstruct phase space

According to Takens' theorem, find an appropriate embedding dimension $m \geq 2d + 1$ for the time series $x(t),\ t = 1,\ 2, \ldots$ (d is the dimension in the motive force system), and the time delay τ, then the trajectory in the reconstructed phase space R^m and the motive force system maintain the same differential Embryo. The reconstructed phase space is

$$Y(t) = (x(t),\ x(t+\tau),\ \ldots,\ x(t+(m-1)\tau)) \in R^m,\ (t = 1,\ 2,\ \ldots,\ N)$$

Therefore, there is a smooth map $\hat{f} : R^m \to R^m$, making

$$Y(t+1) = f(Y(t)),\ t = 1,\ 2,\ \ldots$$

Then there is

$$(x(t+\tau),\ x(t+2\tau),\ \ldots,\ x(t+m\tau)) = f(x(t),\ x(t+\tau),\ \ldots,\ x(t+(m-1)\tau))$$

3. Global prediction

The global method is to use all points in the phase space as the fitting object to fit the fitting trend line $f(\bullet)$ that can express its law, and predict the direction of the trajectory from this. Generally, according to the characteristics of the data in the time series, there are three common simple fitted trend lines: straight line, parabola, and exponential curve. There are usually two ways to determine the fitted trend line. The first is to make a graph and judge according to the data distribution characteristics of its phase space; the second is to compare the growth of the data, if the first-level growth of the time series (periodical growth) is roughly the same, the fitted trend line is a straight line; the secondary growth of the time series (period-to-period growth) is roughly the same, then the fitted trend line is a parabola; the chain-to-month development speed of the time series is roughly the same, then the fitted trend line is an exponential curve.

Since the phase space complexity of the original chaotic system is unknown and the form of its dynamic equation is unknown, we cannot find its true map f from the actual limited data. Therefore, the usual approach is to construct the map $\hat{f}$ according to the given data so that $\hat{f}$ approximates the theoretical f, namely

$$\sum_{t=0}^{N} \left[Y(t+1) - \hat{f}(Y(t)) \right]^2$$

Table 9.1 Total retail sales of consumer goods from 1996 to 2015 (100 million yuan)

Years	1996	1997	1998	1999	2000	2001	2002
Total	28,360.2	31,252.9	33,378.1	35,647.9	39,105.7	43,055.4	48,135.9
Years	2003	2004	2005	2006	2007	2008	2009
Total	52,516.3	59,501.0	68,352.6	79,145.2	93,571.6	114,830.1	133,048.2
Years	2010	2011	2012	2013	2014	2015	
Total	158,008.0	187,205.8	214,432.7	242,842.8	271,896.1	300,930.8	

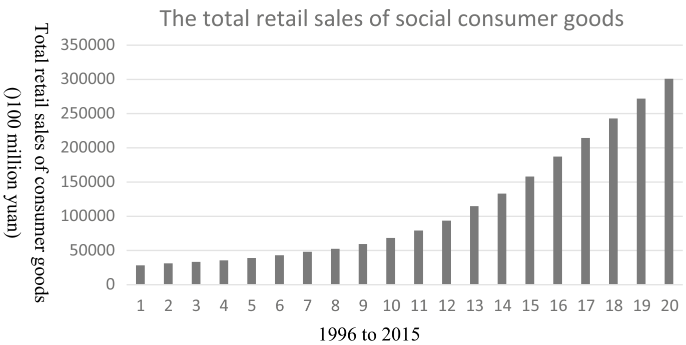


Fig. 9.13 Total retail sales of consumer goods from 1996 to 2015 (100 million yuan). (*Data source* National Bureau of Statistics: http://data.stats.gov.cn/easyquery.htm?cn=C01)

The $\hat{f} : R^m \to R^m$ reaches the minimum value, where the specific form of $\hat{f}$ must be specified in the specific calculation.

Generally, when the dimension d is low, higher-order polynomials, rational expressions, etc. are used for global approximation; when d is higher, the higher-order polynomials will make the prediction accuracy of the prediction algorithm drop rapidly, so the typical linear regression analysis method is used, that is, a first-order linear model.

4. Examples

Example 9.1 Retail sales of consumer goods are affected by many factors, such as social development, population growth, climate change, and seasonal changes. Table 9.1 shows the total retail sales of consumer goods from 1996 to 2015 (100 million yuan). Try to use this data as the basis to predict the total retail sales of consumer goods in 2016 (Fig. 9.13).

Solution 1:

Step 1: Denote the time series of total retail sales of social consumer goods as $x(t)$, $t = 1, \ldots, 20$, then the reconstructed phase space is $Y(t) =$

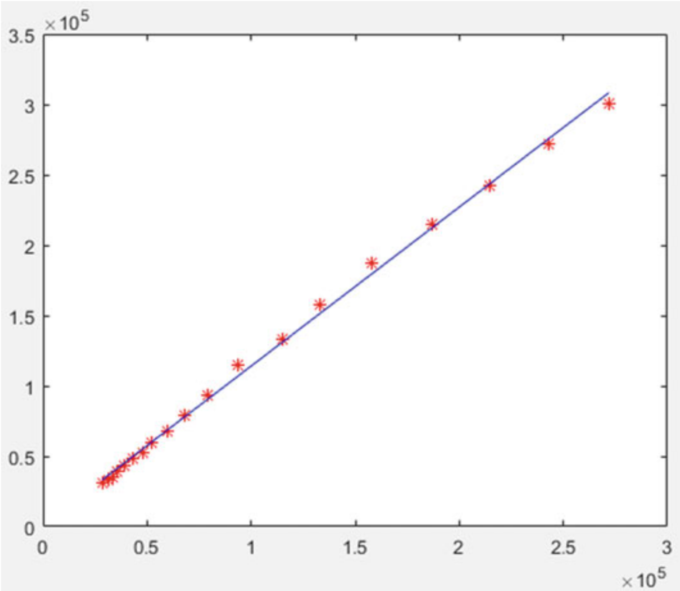


Fig. 9.14 Discrete time series phase space reconstruction image of total retail sales of consumer goods

$(x(t),\ x(t+\tau),\ \ldots,\ x(t+(m-1)\tau)) \in R^m$, $(t = 1,\ 2,\ \ldots,\ 20)$, and there is a smooth map $\hat{f} : R^m \to R^m$, that is, $Y(t+1) = f(Y(t))$, $t = 1,\ 2,\ \ldots 20$. Assume that the embedding dimension $m = 1$ and the time delay $\tau = 1$ of the phase space are reconstructed.

Step 2: Use MATLAB to make a discrete time series phase space reconstruction image of the total retail sales of consumer goods (as shown in Fig. 9.14).

Step 3: From Fig. 9.14, it can be seen that the function $\hat{f} : x(t+1) = \hat{f}(x(t))$ is the closest to a linear function. Therefore, assuming $x(t+1) = f(x(t)) = a + bx(t)$, its matrix form is $A = BC$, where

$$A = \begin{bmatrix} x(2) \\ x(3) \\ \vdots \\ x(20) \end{bmatrix},\ B = \begin{bmatrix} 1 & x(1) \\ 1 & x(2) \\ \vdots & \vdots \\ 1 & x(19) \end{bmatrix},\ C = \begin{bmatrix} a \\ b \end{bmatrix}$$

For any linear equation $A = BC$, there is always $B^T A = B^T BC$. When the matrix $B^T B$ is a non-singular matrix, C has a unique solution, namely $C = (B^T B)^{-1} B^T A$.

Step 4: Use MATLAB calculation (least squares method), put A, B into the above formula to get $a = -784.390$, $b = 0.8832$. Therefore, the forecast model for this time series is $x(t+1) = 1133.2 + 1.1298x(t)$, and the forecast result is

Table 9.2 Linear regression analysis of retail sales of consumer goods (1)

ANOVA[a]						
Model		Sum of squares	Degree of freedom	Mean square	F	Significance
1	Regression	138,769,904,897.416	1	138,769,904,897.416	7978.111	0.000[b]
	Residual	295,695,107.741	17	17,393,829.867		
	Total	139,065,600,005.157	18			

[a] Dependent variable: Xt2
[b] Predictor: (constant), Xt1

Table 9.3 Linear regression analysis of retail sales of consumer goods (2)

Coefficient[a]						
Model		Unstandardized coefficient		Standardization factor	t	Significance
		B	Standard error	Beta		
1	(Constant)	1133.171	1604.251		0.706	0.490
	Xt1	1.130	0.013	0.999	89.320	0.000

[a] Dependent variable: Xt2

$x(21) = 341124.818$ (100 million yuan). That is, the predicted value of total retail sales of consumer goods in 2016 is 34,112,481,800 million.

Solution 2:

Steps 1–2 are the same as Solution 1.

Step 3: From Fig. 9.14, it can be seen that the function $\hat{f} : x(t+1) = \hat{f}(x(t))$ is the closest to a linear function. Therefore, assume $x(t+1) = f[x(t)] = a+bx(t)$, use the analysis tool of the SPSS software and use regression to perform linear fitting, you can get Tables 9.2 and 9.3, as shown below.

In Table 9.3, the constant is a, the variable Xt1 is the variable $x(t)$, and Xt2 is the dependent variable $x(t + 1)$. In Table 9.2, the observed value of the F statistic for the significance test of the regression equation is 7978.111, and the corresponding probability P-value is approximately 0. Under the condition of a significance level of 0.05, because the P-value is less than the significance level, the null hypothesis of the significance test of the regression equation is rejected (the regression coefficient is not 0 when the regression coefficient is different, and there is no significant linear relationship between $x(t)$ and $x(t + 1)$), namely it is reasonable to choose a linear model. According to Table 9.3, SPSS regression linear fitting function $x(t+1) = 1133.171+1.13x(t)$, then $x(21) = 341184.975$ (100 million yuan), that is, the predicted value of the total retail sales of consumer goods in 2016 is 34,118,497,500 million yuan. Note: Under the condition of the significance level of 0.05, the significance level of the variable is less than 0.05, and the

```
x=[28360.2 31252.9 33378.1 35647.9 39105.7 43055.4 48135.9 52516.3 59501
68352.6 79145.2 93571.6 114830.1 133048.2 158008 187205.8 214432.7 242842.8
271896.1 300930.8]
plot(x(1:19),x(2:20))

MATLAB program for calculating fitting function and drawing:
clear all
close all
a=[28360.2 31252.9 33378.1 35647.9 39105.7 43055.4 48135.9 52516.3 59501
68352.6 79145.2 ...
   93571.6 114830.1 133048.2 158008 187205.8 214432.7 242842.8 271896.1
300930.8];
[~,sizea]=size(a);

A=[a(1:end-1)]';
B=[a(2:end)]';
[P,S]=polyfit(A,B,1);
Bfit=polyval(P,A);
plot(A,B,'r*',A,Bfit,'b-'); %Drawing;
```

Fig. 9.15 MATLAB program 1

significance level of the constant 0.490 is greater than 0.05, which indicates that the coefficient of the linear function fitted by SPSS for this question is reasonable, and there exists some difference between the constant and the null hypothesis. The above two problem-solving ideas use the global prediction method, but due to the different calculation tools used, there may be a certain error in the prediction results of the two. Therefore, in the actual forecasting work, we should try our best to choose the calculation tools and methods that can maximize the forecasting accuracy.

9.3 Catastrophe Theory

Since Newton and Leibniz founded the calculus and calculus equations, many continuous, smooth, and gradual changes in nature can be quantitatively described and explained by the calculus method, and we have successfully established various mathematical and physical models, such as Newtonian motion mechanics model, Maxwell's electromagnetic field model, and Einstein's field equations. However, there are still many abrupt changes and leaps in our social lives or nature. The complexity and discontinuity of these processes make the system's behavior space indifferentiable. For example, water boiling and ice melting, volcanic earthquake, rock rupture, financial crisis, war eruption, stock market volatility, price movement and so on. Catastrophe theory, as an emerging discipline that studies change or leap processes, together with dissipative structure theory and synergy theory, links the formation, structure, and development of systems. It is an important branch of systems science and can be used to understand and predict complex system behaviors.

9.3.1 Theory Introduction

What we call catastrophe theory now originated in the late 1960s, and it is regarded as part of Chaos Theory. The catastrophe theory was proposed by the famous French mathematician Professor Rene Thom in 1968 to explain the embryonic process in embryology. In 1972, Rene Thom published the book *Structural Stability and Morphogenesis* and made an independent and systematic exposition of this theory. This exposition also officially announced the birth of the catastrophe theory, a new branch of mathematics. Because of Rene Thom's great contribution in this area, he won the Fields Medal, the highest award in contemporary international mathematics.

The emergence of catastrophe theory has not only attracted attention from all aspects, but also aroused fierce controversy. For decades, it has not only been used in the fields of mathematics, physics, mechanics and other natural sciences, but also widely used in social sciences and biology. The famous British mathematician Professor Chiman called catastrophe theory "an intellectual revolution in mathematics—the most important discovery after calculus."

9.3.2 Fundamental Contents

Catastrophe theory uses the study of the stability of the structure of things to reveal the law of changes in things. It uses topology as a tool and is based on structural stability theory. It is based on mathematical theories such as topological dynamics, singularity theory, structural stability and calculus, and describes the state of the system at critical points to specialize in the theories of discontinuous and abrupt changes in multiple forms, structures and social economic activities of nature.

Catastrophe theory studies the phenomena and laws of transition from one stable configuration to another. It points out that any state of motion is in a stable state or an unsteady state. The stable state can still maintain the original state under the action of external forces, while the unsteady state quickly leaves the original state under the action of external force and enters another state range. The nonlinear system changes from one equilibrium state to another in the form of abrupt changes.

Catastrophe theory believes that the phase transition of a system, that is, the evolution from a stable state to another stable state with different qualities, can be achieved through nonlinear catastrophes or continuous gradual changes. The specific method depends on the specific conditions. If the intermediate transition state of the phase transition is an unstable state, the phase transition process is an abrupt change; if the intermediate transition state is a stable state, the phase transition process is a gradual change. In principle, the phase change mode of the system can be controlled by changing the control conditions.

Rene Thom's catastrophe theory is to use mathematical tools to describe the abrupt changes of the system state, and give the parameter area where the system is in a stable state. When the parameter changes, the system state changes. When the parameter passes through a specific position, the system state will come to catastrophe.

9.3.3 The Mathematical Description of Catastrophe Theory

1. **Basic concept of model**

The potential system is the core content of the elementary catastrophe theory. It is determined by the interaction between the various components of the existing system and the relative relationship between the system and the environment. Therefore, the potential in the system can be determined by the state variables (behavior variables) and the external control parameters.

The state space and the control space can be realized under the collective conditions of various internal state variables and external control parameters that may change. To this end, we denote the n-dimensional state space (behavior space) composed of n state variables as R^{n}, and the m-dimensional control space composed of m control parameters as R^{m}, then $\mathrm{R}^{\mathrm{n+m}}$ represents the integrated space or high-dimensional space. Mathematically, the state space and the control space also constitute the high-dimensional state surface of the hypersurface. In the study of catastrophe theory, the high-dimensional surface space $\mathrm{R}^{\mathrm{n+m}}$ is usually projected onto the control space R^{m} to study the changes of system state parameters. When m is small (especially when $m \leq 3$), the complexity of the problem will be greatly reduced.

2. **Mathematical analysis of R. Thom catastrophe**

The mathematical model described in Rene Thom's catastrophe theory is a gradient model. The mathematical prototype of the model can be described as follows:

Let the smooth function f be a continuous function of the partial derivatives:$R^{m+n} \to R$.

We denote the coordinate of state variable space R^n as $X = (x_1, x_2, \cdots, x_n)$ 2, and the coordinate of control parameter variable space $\mathrm{R^m}$ as $Y = (y_1, y_2, \cdots, y_m)$, where m is the number of variables of the control parameter.

From the high-dimensional surface space R^{m+n} R to the control parameter space projection, for any point on $\mathrm{y} \in R^m$, the gradient of R^{m+n} is:

$$grad_x f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

The abrupt manifold of derived function f is:

$$M_f = \left\{ (x, y) \middle| grad_x f = 0 \in R^n \right\}$$

Then the catastrophe of f is mapped to X_f:$M_f \to R^m$.

The Whitney topology is introduced into the projection space F. Here we use a set of basic neighborhood $\{V(\varepsilon, k)\}$. For all $k \in N$ and positive and continuous functions, ε: $R^{n+m} \to R_+$, where $V(\varepsilon, k)$ is the function $f \in V(\varepsilon, k)$ if and only if $|D_\alpha f(x, y)| < \varepsilon(x)$, $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_\mathrm{n} < k$, and

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial X_1^{\alpha_1} \cdot \partial X_2^{\alpha_2} \cdot \cdots \cdot \partial X_n^{\alpha_n}}$$

Rene Thom proved in his theory: when $m \leq 4$, that is, the number of control parameters is not more than 4, and the value of n is $n \leq 2$, the resulting potential function has only seven different types, namely topological structures (see Table 9.1 for details).

3. **Primary catastrophe types and their mathematical models**

Rene Thom pointed out that as long as the variable $m \leq 4$ is controlled, there are only 7 types of elementary catastrophes. They are cusp, swallowtail, fold, butterfly, elliptical, hyperbolic, and parabolic umbilicus. These types can form a certain geometric structure, which constitutes what Thom calls the "universal unfolding" of the singularity S. However, the topological characteristics of catastrophes caused by multi-dimensional attractors are very difficult at this stage, so there is little research in the academic circle.

(1) Cusp catastrophe: An attractor is divided into two attractors that do not communicate with each other. This type is one of the most used catastrophes in academia. There are two control variables in the model, denoted by α_1, α_2; and there is one state variable, denoted by x (Fig. 9.16). For a dynamic system, its potential energy (potential function) is

$$V(x) = x^4 + \alpha_1 \cdot x^2 + \alpha_2 \cdot x \tag{9.26}$$

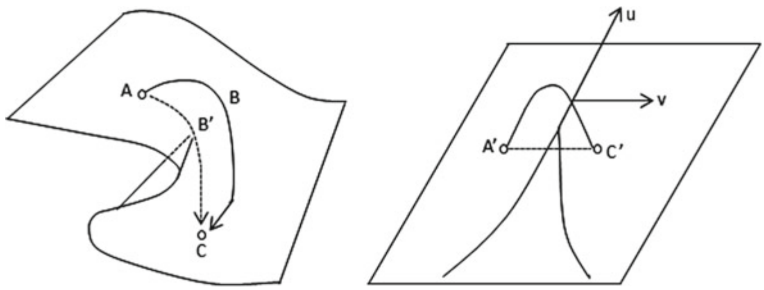


Fig. 9.16 Cusp catastrophe model. (*Note* The picture on the right is a plan view after the projection of the left picture)

The function of the balance surface is determined by $V'(x) = 0$, namely

$$4x^3 + 2\alpha_1 \cdot x + \alpha_2 = 0 \tag{9.27}$$

Derivation of Eq. (9.27), that is, $V''(x) = 0$ is the singular point set Eq. (9.3)

$$12x^2 + 2\alpha_1 = 0 \tag{9.28}$$

Combine Eq. (9.27) with Eq. (9.28) and eliminate the state variable x in the equation to obtain the Eq. (9.29) of the control space as follows. The curve determined by the equation is the branch point set or the branch set, and the catastrophe theory shows that when variables satisfy the bifurcation set equation system, catastrophe will occur, which is the core of catastrophe theory. The divergence point set equation for cusp catastrophe is

$$8\alpha_1^3 + 27\alpha_2^2 = 0 \tag{9.29}$$

$\alpha_1 = -6x^2, \alpha_2 = 8x^3$The decomposition form of the equation of the set of bifurcation points can be written as

$$\alpha_1 = -6x^2, \quad \alpha_2 = 8x^3$$

(2) Fold catastrophe: One attractor ruptures and is captured by another attractor with a smaller potential. This kind of catastrophe is relatively simple in all catastrophe types. At this stage, it is widely used to study rock dynamic instability, fault earthquakes, etc. The control variable in its mathematical model is only 1 α_1, and the state variable is only 1 x. The potential function of the folding catastrophe is

$$V(x) = x^3 + \alpha_1 \cdot x$$

Its equilibrium surface function is $3x^2 + \alpha_1 = 0$, and its equilibrium surface is shown in Fig. 9.17:

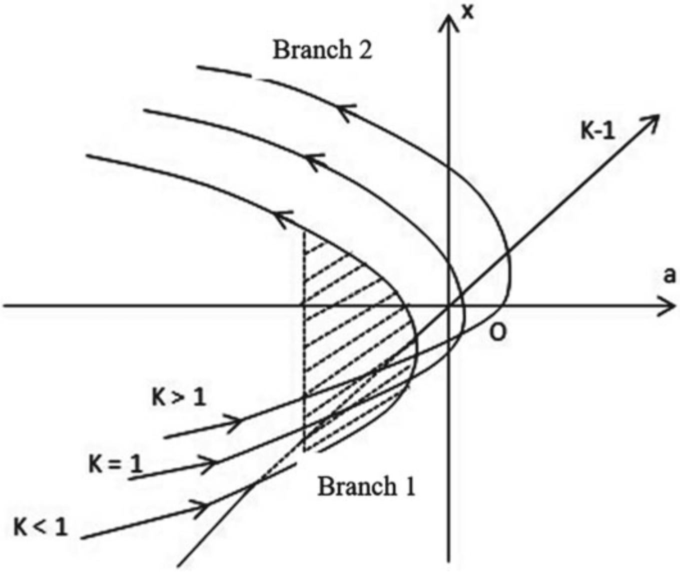


Fig. 9.17 The equilibrium surface of the folded catastrophe model

(3) Swallowtail catastrophe: A groove is cut off from a wavefront surface, and the bottom of the groove is the edge of the shock wave. There are 3 control variables in the swallowtail catastrophe model, represented by α_1, α_2 and α_3, the state variable is x, and its potential function is

$$V(x) = x^5 + \alpha_1 \cdot x^3 + \alpha_2 \cdot x^2 + \alpha_3 \cdot x$$

Its equilibrium surface is a hypersurface, and its function is

$$5x^4 + 3\alpha_1 \cdot x^2 + 2\alpha_2 \cdot x + \alpha_3 = 0$$

The set of singularities is

$$20x^3 + 6\alpha_1 \cdot x + 2\alpha_2 = 0$$

Combine the equilibrium surface function and the singularity set function to eliminate the state variable x to get the divergence set.

(4) Butterfly catastrophe: the splitting (or "swelling") of the free side shock wave. The butterfly catastrophe model has 4 control variables, α_1, α_2, α_3, and α_4, and 1 state variable is x, and its potential function is

$$V(x) = x^6 + \alpha_1 \cdot x^4 + \alpha_2 \cdot x^3 + \alpha_3 \cdot x^2 + \alpha_4 \cdot x$$

Its equilibrium surface is a hypersurface, and the surface function is

$$6x^5 + 4\alpha_1 \cdot x^3 + 3\alpha_2 \cdot x^2 + 2\alpha_3 \cdot x + \alpha_4 = 0$$

The set of singularities is

$$30x^4 + 12\alpha_1 \cdot x^2 + 6\alpha_2 \cdot x + 2\alpha_3 = 0$$

(5) Elliptical catastrophe: the sharp point of the picket is the singularity. There are three control variables in the elliptical catastrophe model, namely α_1, α_2 and α_3; there are two state variables, namely x and y, and the potential function is

$$V(x, y) = \frac{1}{3}x^3 - x \cdot y^2 + \alpha_1(x^2 + y^2) - \alpha_2 \cdot x + \alpha_3 \cdot y$$

Its balance surface is a hypersurface, and its hypersurface function is

$$\begin{cases} x^2 - y^2 + 2\alpha_1 \cdot x - \alpha_2 = 0 \\ -2x \cdot y + 2\alpha_1 \cdot y + \alpha_3 = 0 \end{cases}$$

(6) Hyperbolic catastrophe: When a wave ruptures, the wave crest is the singularity. The hyperbolic umbilical point catastrophe model has 3 control variables, α_1, α_2 and α_3; there are 2 state variables, x, y, and its potential function is

$$V(x, y) = x^3 + y^3 + \alpha_1 \cdot x \cdot y - \alpha_2 \cdot x - \alpha_3 y$$

Its hypersurface function is

$$\begin{cases} 3x^2 + \alpha_1 \cdot y - \alpha_2 = 0 \\ 3y^2 + \alpha_1 \cdot x - \alpha_3 = 0 \end{cases}$$

(7) Parabolic catastrophe: between the ellipse and the hyperbolic, this singularity can be seen when the jet is interrupted. The parabolic catastrophe model has 4 control variables, α_1, α_2, α_3 and α_4; there are 2 state variables, x, y, and the potential function is

$$V(x, y) = y^4 + x^2 \cdot y + \alpha_1 \cdot x^2 + \alpha_2 \cdot y^2 - \alpha_3 \cdot x - \alpha_4 \cdot y$$

The equilibrium surface function is

$$\begin{cases} 2x \cdot y + 2\alpha_1 \cdot x - \alpha_3 = 0 \\ x^2 + 4 \cdot y^3 + 2\alpha_2 \cdot y - \alpha_4 = 0 \end{cases}$$

In summary, the functional forms of the seven elementary catastrophe models are as follows (Table 9.4).

Table 9.4 Seven types of elementary catastrophe functions

Name	State dimension	Control dimension	Potential function expression	Equilibrium surface expression
Cusp	1	2	$V(x) = x^4 + \alpha_1 x^2 + \alpha_2 x$	$4x^3 + 2\alpha_1 x + \alpha_2 = 0$
Fold	1	1	$V(x) = x^3 + \alpha_1 x$	$3x^2 + \alpha_1 = 0$
Swallowtail	1	3	$V(x) = x^5 + \alpha_1 x^3 + \alpha_2 x^2 + \alpha_3 x$	$5x^4 + 3\alpha_1 x^2 + 2\alpha_2 x + \alpha_3 = 0$
Butterfly	1	4	$V(x) = x^6 + \alpha_1 x^4 + \alpha_2 x^3 + \alpha_3 x^2 + \alpha_4 x$	$6x^5 + 4\alpha_1 x^3 + 3\alpha_2 x^2 + 2\alpha_3 x + \alpha_4 = 0$
Elliptical	2	3	$V(x, y) = \frac{1}{3}x^3 - xy^2 + \alpha_1(x^2 + y^2) - \alpha_2 x + \alpha_3 y$	$\begin{cases} x^2 - y^2 + 2\alpha_1 x - \alpha_2 = 0 \\ -2xy + 2\alpha_1 y + \alpha_3 = 0 \end{cases}$
Hyperbolic umbilical	2	3	$V(x, y) = x^3 + y^3 + \alpha_1 xy - \alpha_2 x - \alpha_3 y$	$\begin{cases} 3x^2 + \alpha_1 y - \alpha_2 = 0 \\ 3y^2 + \alpha_1 x - \alpha_3 = 0 \end{cases}$
Parabolic umbilical	2	4	$V(x, y) = y^4 + x^2 y + \alpha_1 x^2 + \alpha_2 y^2 - \alpha_3 x - \alpha_4 y$	$\begin{cases} 2xy + 2\alpha_1 x - \alpha_3 = 0 \\ x^2 + 4y^3 + 2\alpha_2 y - \alpha_4 = 0 \end{cases}$

4. **Model and theoretical application**

The seven elementary catastrophe models mentioned above are relatively intuitive and concise, and they are widely used in the fields of natural sciences and social sciences, such as physics, transportation, biology, management, and economics. In physics, cusp catastrophes can be used to study optical caustics, phase transitions, bending states of elastic beams, rock and soil mechanics, etc., and butterfly catastrophes can be used to explain the phenomenon of iced water and water vapor; in transportation, the swallowtail catastrophe can be used to discuss traffic flow forecasting, vehicle derailment and other related research. Catastrophe theory can well explain the relationship between road vehicle occupancy, vehicle operating speed and traffic flow, and can provide a scientific basis for the management and operation of expressways. In biology, catastrophes can explain the formation of natural organisms, gastrulation, and the formation of stimuli in nerves, etc. In social sciences, catastrophes can be used to explain conflict outbreaks, stock collapses, economic crises, etc. It is mainly determined by the state variables and control variables of the system for the catastrophe models apply to what research objects.

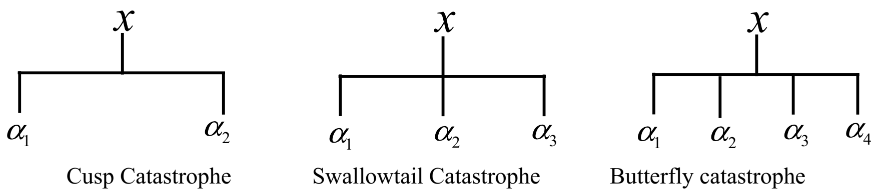


Fig. 9.18 Schematic diagram of each catastrophe system model in catastrophe progression method

9.3.4 Catastrophe Progression Method

Catastrophe progression method, also known as catastrophe fuzzy subordinate function method, is an important application branch of catastrophe theory. It is based on catastrophe theory, combining catastrophe model and fuzzy mathematics theory, using normalized formulas derived from bifurcation set equations to synthesize targets. Compared with traditional evaluation methods (such as fuzzy comprehensive evaluation method), the catastrophe progression method does not need to subjectively determine the weight of each index, nor does it need to select the membership function. It only needs to sort the importance of each index, simple and easy to operate.

1. The principle of catastrophe progression method

In the process of comprehensive evaluation of catastrophe progression, each index factor is regarded as a control variable that affects the target of the system, and the different qualities of each index factor (control variable) in the system are transformed into the same quality state, that is, the same state variable. The fuzzy subordinate function of the corresponding dimension gets the final evaluation result. In the catastrophe progression method at this stage, cusp catastrophe, swallowtail catastrophe and butterfly catastrophe are the three most widely used catastrophe types. The catastrophe system model is as follows (Fig. 9.18).

2. Derivation of normalization

Through the potential function of the system, the plane surface equation and the singular point set equation of the system can be derived, thereby obtaining the bifurcation point set equation. The bifurcation point set equations of the common three catastrophic system model decomposition forms are as follows (Table 9.5):

The decomposition form of the bifurcation set equation of the cusp catastrophe model is

$$\alpha_1 = -6x^2, \ \ \alpha_2 = 8x^3$$

Table 9.5 Normalization formulas of common catastrophe models

Catastrophe type	Control dimension	State dimension	Normalization formula
Cusp catastrophe	2	1	$x_{\alpha_1} = \sqrt{\alpha_1}, x_{\alpha_2} = \sqrt[3]{\alpha_2}$
Swallowtail catastrophe	3	1	$x_{\alpha_1} = \sqrt{\alpha_1}$, $x_{\alpha_2} = \sqrt[3]{\alpha_2}, x_{\alpha_3} = \sqrt[4]{\alpha_3}$
Butterfly catastrophe	4	1	$x_{\alpha_1} = \sqrt{\alpha_1}, x_{\alpha_2} = \sqrt[3]{\alpha_2}$, $x_{\alpha_3} = \sqrt[4]{\alpha_3}, x_{\alpha_4} = \sqrt[5]{\alpha_4}$
Shed catastrophe *	5	1	$x_{\alpha_1} = \sqrt{\alpha_1}, x_{\alpha_2} = \sqrt[3]{\alpha_2}$, $x_{\alpha_3} = \sqrt[4]{\alpha_3}$, $x_{\alpha_4} = \sqrt[5]{\alpha_4}, x_{\alpha_5} = \sqrt[6]{\alpha_5}$

The decomposition form of the bifurcation set equation of the swallowtail catastrophe model is

$$\alpha_1 = -6x^2,\ \alpha_2 = 8x^3,\ \alpha_3 = -3x^4$$

The decomposition form of the bifurcation set equation of the butterfly catastrophe model is

$$\alpha_1 = -10x^2,\ \alpha_2 = 20x^3,\ \alpha_3 = -15x^4,\ \alpha_4 = 4x^5$$

Taking the cusp catastrophe model as an example, the corresponding normalization formula is derived: the divergence point set equation of the catastrophe model is transformed to obtain

$$x_a = \sqrt{-\frac{\alpha_1}{6}},\ x_b = \sqrt[3]{\frac{\alpha_2}{8}}$$

According to the membership function of fuzzy mathematics, let $\alpha_1 = 6\alpha_1'$ (absolute value), $\alpha_2 = 8\alpha_2'$, get $x_{\alpha_1} = \sqrt{\alpha_1'}$, $x_{\alpha_2} = \sqrt[3]{\alpha_2'}$; thus, the values of α_1', α_2', and x are all limited to 0–1. Therefore, the normalized formula for the cusp catastrophe is

$$x_{\alpha_1} = \sqrt{\alpha_1},\quad x_{\alpha_2} = \sqrt[3]{\alpha_2}.$$

Similarly, the normalized formula of the swallowtail catastrophe model is

$$x_{\alpha_1} = \sqrt{\alpha_1},\ x_{\alpha_2} = \sqrt[3]{\alpha_2},\ x_{\alpha_3} = \sqrt[4]{\alpha_3}.$$

Similarly, the normalized formula of the butterfly catastrophe model is

$$x_{\alpha_1} = \sqrt{\alpha_1},\ x_{\alpha_2} = \sqrt[3]{\alpha_2},\ x_{\alpha_3} = \sqrt[4]{\alpha_3},\ x_{\alpha_4} = \sqrt[5]{\alpha_4}$$

3. The general steps of the catastrophe progression method

The general steps of using the catastrophe progression method to comprehensively evaluate system targets are as follows:

(1) According to the research purpose, the system indicators are decomposed layer by layer to establish a hierarchical, primary and secondary evaluation indicator system. Generally, the number of secondary indicators for each indicator does not exceed 5;
(2) Based on the establishment of a hierarchical evaluation index system with a hierarchical, primary and secondary structure, determine the corresponding catastrophe system model for each target index (except for the underlying index). The catastrophe system models commonly used in catastrophe progression method mainly include cusp type, swallowtail type, butterfly type and shed type (Indian hut type);
(3) Perform data standardization processing on the underlying indicators to make the indicators comparable. The commonly used data standardization processing method is the maximum and minimum value standardization processing;
(4) Use the normalization formula of a catastrophe system model to calculate each control variable to obtain each abrupt change level value;
(5) According to the direction of each control variable acting upon the state variables, determine the decision-making principle. If the control variables are related to each other, it is the principle of complementarity (take the average value of each control variable), if the control variables are not related, it is the principle of non-complementarity (take the minimum value of each control variable);
(6) Repeat steps ④ and ⑤ to perform hierarchical calculation on the index system to obtain the total catastrophe level value, and then comprehensively evaluate and analyze the target according to the total catastrophe level value. At this stage, many research results have introduced the catastrophe progression evaluation method into related research fields, such as applying it to maritime traffic risk prediction research, agricultural cold-chain logistics risk assessment model research, and financial credit risk research.

9.3.5 Theoretical Meaning

The main advantages of catastrophe theory:

(1) Catastrophe theory is based on modern mathematics such as topology, sets, manifolds, and focuses on the scientific method of studying the complexity of discontinuous catastrophe phenomena under continuous action. Its theoretical model is particularly suitable for studying systems whose internal structure is not yet clear..

(2) The proposal of catastrophe theory helps us to understand the true face of change management and understand the essence of chaos theory. It can understand the process of organization or change system from a new perspective.
(3) Catastrophe theory enables many discontinuous catastrophe phenomena in nature to be described and explained by mathematical methods. As an important branch of complexity systems science, it has greatly enriched the theoretical research system of this discipline.

Limitations of catastrophe theory:

(1) Due to the considerable complexity, the related research and analysis of catastrophe theory rarely involves complex systems involving multiple (more than 5) control variables. Therefore, the current catastrophe theory for complex behavioral systems still exists big limitations.
(2) Catastrophe theory is a young branch of mathematics and scientific research method. At this stage, its research depth is limited. Even if it is to predict the simplest system behavior, its research is still challenging.
(3) For the mathematical model of catastrophe theory, the choice of control variables and system state variables is the key to whether the model can be successfully used. However, it is difficult to determine when the academic community has quite a lot of this problem.

The catastrophe theory has brought great promotion and development to the fields of society, science, technology, economy, management, etc. Its views have universal guiding significance and guide people to look at problems from the perspective of catastrophes in a wide range of fields. However, its development is only a few decades, and the relevant research on catastrophe theory at this stage is not deep enough. As a fairly young theoretical model and method, catastrophe theory still has many problems to be solved in its application, and has much space to be studied and promoted.

9.4 Hypercycle Theory

With sunrise and sunset, the moon surging and declining, nature moves forward in a continuous cycle; as spring passes and autumn comes, flowers bloom and fall, life gradually evolves in the cycle of reproduction. Circulation is the most common phenomenon in nature, so studying the specific laws of the circulation of various things is an important goal of continuous exploration in the field of scientific research. Over the past few decades, with the emergence of the three new theories of system science, people's understanding of system cycles has gradually changed, from equilibrium to non-equilibrium, linear to nonlinear, organization to self-organization, and simple to complex. The emergence of the hypercyclic theory, together with the dissipative structure theory and the synergy theory, formed the self-organizing theoretical system of the non-equilibrium system, and jointly revealed the various mysteries of the cycle evolution of the objective world.

Fig. 9.19 Manfred Eigen

9.4.1 Theory Introduction

In the early 1970s, the famous German biophysicist Manfred Eigen, based on the theoretical foundations of molecular biology, evolution, information theory, self-organization theory, and non-equilibrium thermodynamics, believed that the origin of life information was a form of hypercycle process of self-organization of molecules, and proposed a theory about the self-organization phenomenon of non-equilibrium systems—the hypercycle theory, which is used to study cell biochemical systems, molecular systems and information evolution. Hypercycle theory is a new system theory based on biochemistry and molecular biology. It further advances Bertalanfi's ecosystem and organ system theory to the cellular and molecular level, thus creating molecular systems biology study field (Fig. 9.19).

There are many cyclic processes promoted by the catalytic action of enzymes in various life phenomena in nature, and the basic cycle constitutes a higher level cycle, that is, the hypercycle. The so-called hypercycle refers to a kind of cycle in which the catalytic cycle is functionally linked by cycle coupling, and is also called the catalytic hypercycle. Hypercyclic system refers to a system that connects autocatalytic or self-replicating units through a cyclic connection. In this system, each replication unit can guide its own replication and catalyze the production of the next intermediate. At present, domestic research on this theory mainly focuses on organizational learning and innovation, strategic alliances and network evolution.

9.4.2 Theoretical Principle

The development process of life can be divided into chemical evolution stage and biological evolution stage. Biological evolution depends on heredity and catastrophe. Nucleic acid and protein are the two most important biological macromolecules in the process of heredity and catastrophe. Since all living organisms use a unified genetic code and basically the same decoding method, the realization of the decoding process requires the close cooperation of hundreds or thousands of molecules, so in the process of life, these hundreds of thousands species cannot be formed together and tightly organized. Therefore, there must be a self-organization stage with biological macromolecules between the chemical evolution stage and the biological evolution stage, and this form of molecular self-organization is the hypercycle. For example, a nucleic acid is used as a self-replicating template, and the self-replication process of its sequence is generally not performed directly, but the protein encoded by it affects the self-replication of another nucleic acid. This structure is a kind of hypercyclic structure.

9.4.3 Hierarchical Levels

The hypercycle theory proposed by M. Eigen focuses on the study of cycle phenomena in biochemistry. He divides cycle phenomena into three different hierarchical levels:

The first level is the transformation reaction cycle, which is a process of self-regeneration as a whole. In the circulation process, through the action of the catalyst E, the reactant S generates the final product P, and the related catalyst can be separated from the product to form a circulation system in which ES and EP are intermediate products. In biochemistry, the cycle of enzymes is a cycle of transformation reactions. The cycle process is shown in the Fig. 9.20.

The second level is the catalytic reaction cycle, which is a self-replicating process as a whole. It is a higher level than the transformation reaction cycle, requiring at least one intermediate product in the reaction cycle process to be a self-catalyst, that is, the intermediate product itself acts as a catalyst to accelerate or delay the transformation of the reactants. Simple self-replication unit cycle, such as DNA self-replication, is a kind of catalytic reaction cycle. The cycle process is shown in the Fig. 9.21.

The third level is the hypercycle, that is, the cycle formed by the cycle. According to M. Eigen's hypercycle theory, each component in the hypercycle system

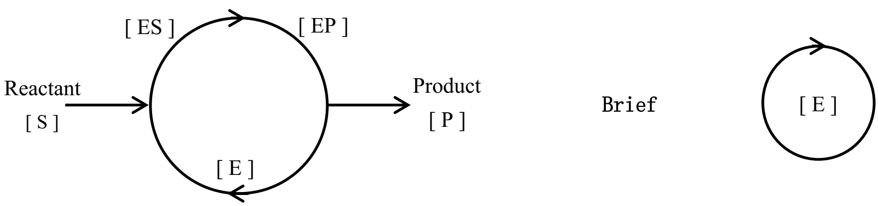


Fig. 9.20 Transformation reaction cycle

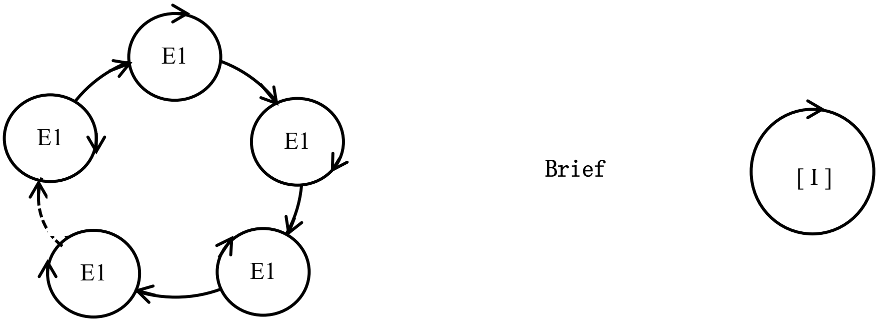
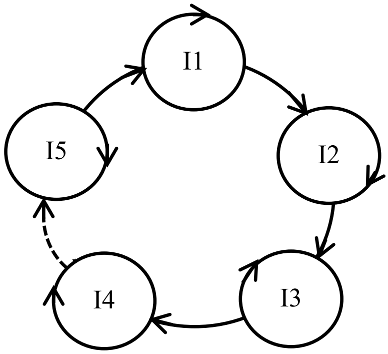


Fig. 9.21 The catalytic reaction cycle

Fig. 9.22 Hyperloop



can not only self-replicate, but also catalyze the self-replication of the next component. However, in the actual hypercycle organization, it is not required that all components play the role of self-catalyst. Generally, only one link in the cycle is required to be a self-replicating unit, then the cycle can show the characteristics of the hypercycle, namely self-regeneration, self-replication, self-selection, self-optimization and evolution to a higher state. The catalytic hypercycle process is shown in the Fig. 9.22.

In summary, the transformation reaction cycle is a self-regeneration process, that is, the catalyst is regenerated after a cycle ends; the catalytic reaction cycle is a self-replication process, that is, the intermediate product itself acts as a catalyst and guides the transformation of reactants into products; in addition to self-regeneration and self-replication, the hypercycle can also perform self-selection and optimization, thereby evolving to a higher level of complexity. Regarding the origin of life information, M. Eigen put forward the concept of hypercycle theory, and pointed out that the close cooperation between nucleic acid and protein is realized through complex compound hypercycle, so as to promote the origin and evolution of life.

9.4.4 Important Nature

1. Characteristics of hypercycle

(1) Irreversibility: The result of a hypercycle is not to return to the origin without change, but to reach a new starting point to start a new cycle. Therefore, the hypercycle is a kind of development, ascending, evolution and deep cycle. This cycle is an irreversible process.
(2) Openness: Hypercycle is not in a closed system, but in an open system. Through the continuous exchange of logistics, energy and information with the external environment, it can output the surplus and consume what it needs, in order to maintain the normal operation of the cycle.
(3) Catalysis: The essence of catalysis is to greatly accelerate the rate of chemical reaction by adding a small amount of objects, but it does not cause loss. The key to the supercycle process is to realize self-cycle and other cycles through catalysis.

2. Important properties of hypercycle

(1) The hypercycle can be enlarged or reduced to a certain extent, as long as this change has the advantage of competitive choice.
(2) The hypercycle can make the multiple units linked by the cycle coexist in a stable and controlled manner and can grow together, but the replication units that are not part of the cycle can compete with each other.
(3) Once the hypercycle appears, it can be stably maintained. The organization of hypercycle is the basis for the formation and evolution of biological macromolecules. The circulatory organization is both stable and mutable, thus establishing relevant codes and forming corresponding cells.

3. Mathematical description of hypercycle evolution

Hypercycle theory points out that the evolution of biological macromolecules with hypercycle structure can be described by mathematical models (differential equations). Assuming that $x_i (i = 1, 2, 3, \ldots, n)$ is a state variable, where n is the number of species, they will have errors in the process of continuous replication, and these errors will easily lead to the variation of dominant species. The mathematical model describing this evolutionary process is:

$$\frac{dx_i}{dt} = (A_i Q_i - D_i) \cdot x_i + \sum_j x_j \phi_{ji} + x_i \Omega \ (i, \ j = 1, \ 2, \ 3, \ldots, n)$$

Among them, A_i is the replication rate that is the production rate, Q_i is the quality factor of the replication template, that is, the correct replication ability, D_i

is the decomposition rate, ϕ_{ji} is the missynthesis coefficient from species j to species i, that is, the coefficient of variation, and Ω is the environmental impact factor. Therefore, this mathematical model (system of equations) contains various influencing factors of species variation. Through analysis of equations, numerical calculations and related discussions, the trends of species evolution can be obtained. It is worth noting that when the missynthesis coefficient ϕ_{ji} is very small, the initial state of a species dominance will change to another species dominance; when the missynthesis coefficient ϕ_{ji} is large, a state where multiple species coexist will eventually be formed. The emergence of this system of equations provides a new mathematical tool for evolutionary problems.

9.4.5 Theoretical Significance

The hypercycle theory has the following significances:

(1) It provides a new model for the formation and evolution of macromolecular organisms. This supercycle model summarized from biomolecules is of great significance to the analysis of general complex systems.
(2) As an important branch of complexity systems science, it has a profound impact on the study of system evolution laws, system self-organization methods, and complex system processing.
(3) It not only clarifies the origin and evolution of pre-life cells, but also puts forward the theory and method of self-organization and development of non-equilibrium systems based on the hypercycle of biological systems, which greatly enriches the systemic science theory.
(4) It is not only conducive to people's understanding and transformation of nature and society, and promoting social and economic development, but also conducive to the virtuous circle of mankind, and eliminate the vicious circle that is not conducive to mankind.

9.4.6 Fractal Theory

The world we live in is rich and colorful, with uninterrupted mountains and rivers, intertwined coasts, unpredictable clouds, turbulent waves, and winding rivers. They are all in very complex and irregular forms in nature. They no longer have the continuous and smooth nature of the mathematical analysis that we have already known. Over the years, facing these ever-changing and irregular forms, people are used to describing them with traditional Euclidean geometry and analyzing them with analytic geometry, projective geometry, and differential geometry. The facts have proved that the results obtained by using geometric theories describing rules to explain irregular geometric forms are not scientific and reasonable. Due to traditional mathematics can no longer describe them, a new branch of modern mathematics, fractal theory, came into being.

Fig. 9.23 B. B. Mandelbrot

9.4.7 Theory Introduction

Fractal theory is a very popular and active new subject and theory at this stage. The theory was born in the mid-1970s and was proposed by the famous American mathematician B. B. Mandelbrot. In 1967, Mandelbrot published a famous paper titled “How Long is the British Coastline?” in the authoritative American magazine *Science*, in which he first proposed the concept of fractal. In 1982, he published the foundational work of fractal theory, *Fractal Geometry of Nature*, and proposed the “fractal theory” that shocked the academic world. The theory attempts to reveal and describe the laws and physical nature hidden deep in the chaotic phenomenon, thus opening up a new field of physics and mathematics, which has aroused great interest from many scholars (Fig. 9.23).

The biggest feature of fractal theory lies in the use of fractional dimensions and mathematical methods to describe and analyze objective things, revealing that under certain conditions, certain processes or aspects of the world (such as form, structure, function, and energy) can show the overall similarity, and it believes that the change of the spatial dimension can be either discrete or continuous. Since its inception, fractal theory has been widely used in various fields of natural science and social science, thus forming many new disciplines and becoming an important frontier branch of nonlinear science.

9.4.8 Definition

The fractal refers to a phenomenon, figure, or physical process with self-similar characteristics, that is, a geometric form in which the components are similar to the whole in some form. In brief, it means that the system has “self-similarity”

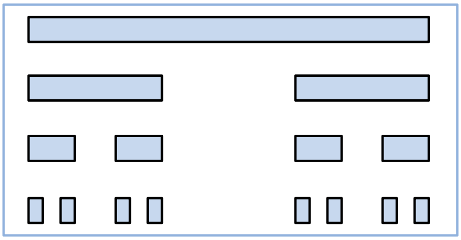


Fig. 9.24 Three-part Cantor set construction process

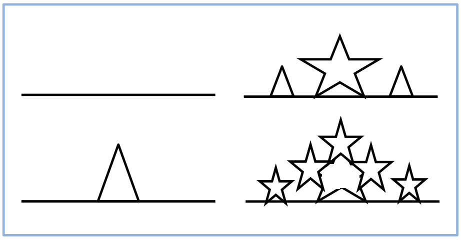


Fig. 9.25 The generation process of Koch curve

and "scale invariance", which are its main characteristics. Their specific contents are as follows:

(1) Self-similarity. As the core of fractal theory, it means that the characteristics of a certain structure or process are similar from different spatial scales or time scales, or that the part or part of the target object is enlarged at the same scale in all directions. The shape is similar to the overall shape. It is worth noting that there will also be self-similarity between the whole and the whole or between the part and the part. Normally, the manifestation of self-similarity is more complicated, and it is not a simple overlap with the whole after a certain amount of magnification. For example, a tree usually consists of many branches. If you compare a certain branch with a large tree, you will find that they are similar in composition and form, and the branches also show obvious self-similarity (Fig. 9.24).

(2) Scale invariance. The scale invariance refers to arbitrarily selecting a part of the area in the fractal and magnifying it, and the resulting graph will show the morphological characteristics of the original image. Because in fractal, zooming in or out of the target will not change its shape, complexity and irregularity, so the scale invariance in fractal is also called expansion invariance. For the category of fractal, according to the degree of self-similarity, it can be divided into regular fractal and random fractal. Regular fractal refers to specific strict self-similarity, such as Cantor set, Sierpinski carpet, Koch curve, etc.; random fractal refers to self-similarity in a statistical sense. Similarities, such as rough surfaces, coastlines, clouds, etc., fractals in the real world mostly exist in the form of random self-similarity (Fig. 9.25).

9.4.9 Fractal Dimension Measurement Method

1. The concept of fractal dimension

Normally, people regard a single point as zero-dimensional, line one-dimensional, surface two-dimensional, and ordinary space three-dimensional. Einstein in the theory of relativity proposes a four-dimensional space–time that introduces the time dimension into the traditional three-dimensional space. However, traditional mathematics is difficult to quantitatively describe the dimensionality of the structure of some complex natural phenomena, such as coastlines, snowflakes, branches, and clouds. With people's continuous research on fractal phenomena, the concept of fractal dimension has gradually developed. Its emergence has made a quantitative description of the complex configuration that traditional mathematics cannot explain. The fractal dimension, also known as fractal dimension or fractal dimension, is a basic concept of geometry and space theory, and is regarded as the quantitative representation and basic parameter of fractal in fractal theory. The value of the fractal dimension can be either an integer or a decimal. It reflects the degree of irregularity of the fractal body. The larger the fractal dimension, the more complex the distribution, and vice versa.

2. Measurement of fractal dimension

For different research purposes and research methods, fractal dimensions have different definitions and their calculation methods are different. Here we introduce three more commonly used fractal dimension measurement methods, they are topological dimension, Hausdorff dimension and box dimension.

(1) Topological dimension. For a geometric body, its topological dimension is equal to the number of independent coordinates required to determine the specific location of a point. Generally, a small box with a scale of r is used to completely cover a d-dimensional geometric object (set the total scale as unit 1), and the required number of $N(r)$ and the used scale of r satisfy the following relationship

$$N(r) = \frac{1}{r^d}$$

Obtain by further deformation

$$d = \frac{\ln[N(r)]}{\ln(1/r)}$$

where d is defined as the topological dimension.

Taking a two-dimensional square as an example, with side length as the unit length, assuming that a small square with a scale size of $r = 1/2$ is used to cover the big one, the relationship between the number $N(r)$ required for complete coverage and the scale r is as follows

$$N(1/2) = 4 = \frac{1}{(1/2)^2}$$

It can be found that the same applies when $r = 1/k$, which proves the correctness of the above relationship.

(2) Hausdorff dimension. From the above introduction to topological dimensions, it can be seen that it has two obvious characteristics: one is that its dimension is an integer; the other is that for a geometric body with a strictly determined dimension, if it is measured with the same dimension as the scale, it can be measured. A certain value $N(r)$ is obtained, but if it is measured with a scale lower than its dimension, the result is infinity; if it is measured with a scale higher than its dimension, the result is zero. For example, if the unit length line segment is used to measure the unit area square, the result is infinity, indicating that the scale used is too thin; on the contrary, the unit square is used to measure the unit line segment, and the result is zero, indicating that the scale used is too thick. Therefore, for fractal geometric objects, the definition of topological dimension needs to be extended to fractal dimension. Its mathematical expression is

$$N(r) = r^{-D_H}$$

Take the natural logarithm of both sides to get

$$D_H = \frac{\ln[N(r)]}{\ln(1/r)}$$

The D_H in the formula is the Hausdorff fractal dimension, referred to as fractal dimension for short. It is a fractal dimension based on the Hausdorff measure, and it is the foundation of the fractal geometric dimension theory. It can be an integer or a fraction. For a specific fractal, its D_H is greater than its topological dimension (the fractal dimension of all fractals is greater than its topological dimension). It is worth noting that the calculation of the Hausdorff fractal dimension is usually quite complicated.

(3) Box dimension. Box dimension, also known as box counting dimension, is the most widely used fractal dimension at this stage. It is relatively simple to calculate the box dimension of a fractal. It can be defined as: Use a small box (square or round) with a side length of r to completely contain and cover the fractal A. Let $N(r)$ be the minimum number of boxes required to contain A, then

$$D_0 = \lim_{r \to 0}\left(\frac{\log N(r)}{\log(1/r)}\right)$$

where D_0 is the box dimension of fractal A. Although the calculation of the box dimension seems simple, for high-dimensional problems, the amount of calculation will also increase rapidly, and it is difficult to obtain a convergence result.

In addition to the above three fractal dimensions, there are others: entropy dimension, information dimension, logarithmic dimension, capacity dimension and Kolmogorov entropy.

9.4.10 Application of Fractal Theory

Fractal theory has developed rapidly in the past 30 years, and its appearance has a huge impact on various disciplines. Fractal theory is widely used in the fields of physics, biology, astronomy, chemistry, materials science, computer graphics technology, information science and economics. For example, in image processing, fractal theory can be applied to image segmentation, target recognition, image compression, image analysis and synthesis; natural scenery simulation, animation production, etc. can be realized on a computer; strange peaks and valleys, and unique scences can be generated in film and television. In economics, it can be applied to stock price changes, income distribution and futures price behavior, etc. It can also be used for house decoration design, fashion design, etc. In recent years, it is also a rising branch of management science to use the fractal theory to manage cities, as the urban architecture, commercial network layout, infrastructure construction, highway information construction, etc. meet the theoretical research category of fractal structure to a certain extent. It can be seen that in today's knowledge economy society, fractal theory will become the foundation of management science.

Fractal theory is an important branch of complexity science. Although its history is only a few decades, it is widely used in almost all fields of natural science and social science. It is a hot subject and frontier subject in domestic and foreign research at this stage. As a new concept and method, its enables people to describe many problems in the nonlinear world with new concepts and methods. It can not only reveal complex phenomena and related laws hidden behind things, but also opened up a new direction to do so. The fractal geometry in fractal theory not only allows us to enjoy the beauty brought by the integration of art and science, and the unity of art and mathematics, it also has a profound significance of scientific methods.

Fractal theory is currently undergoing rapid development. Although some important results have been achieved in the research of fractal and the subject has taken shape in theory, there is a long way to go for the researchers of this field.

9.5 Self-organized Criticality

The concept of self-organized criticality is an important concept proposed by Per Bak and his colleagues in 1987, which has the characteristics of space–time evolution, and is an important concept of space–time freedom in complex dynamic systems. The competition and cooperation among a large number of components of this complex dynamic system make the system evolve to a critical state, where small events can cause chain reaction accidents and affect some components in the system. The chain reaction throughout the whole is the essence of the dynamic behavior of the system. In terms of macro performance, there are more minor incidents than major incidents. The power spectrum of a system with self-organized critical characteristics has a typical 1/f noise property, that is, a power-law relationship between scale and frequency is satisfied. This relationship is considered as a mathematical characterization of self-organized critical characteristics.

In the 1980s, Santa Fe Institute (SFI) put forward the concept of the edge of chaos in the study of artificial life. In the computer simulation of the artificial life model, it was found that there was a phase transition between order and chaos. At this level under certain conditions, the behavior of the system will have complex phenomena of adaptation and self-organization. This phenomenon has shown a certain universality in models of connectionism, such as cellular automata and gene networks. The phenomena described by the concept of the edge of chaos are also manifested in life, ecology, and social systems. The edge of chaos shows a certain critical state. The system reaches the edge of chaos through its own adaptation and self-organization. The state of self-organized criticality is considered to be on the edge of chaos. The edge of chaos refers to a very narrow area between the periodic area and the chaotic area. This area is located at the transition point (or critical point) between order and chaos. For this reason Packard and Kauffman call it the edge of chaos. But Bak calls it weak chaos, and thinks it is caused by self-organized criticality. Computer simulation shows that self-organized criticality and the edge of chaos are connected, and the system in the state of self-organized criticality is just on the edge of chaos. Kauffman's research further pointed out that the extensive dissipative dynamical system that is far from equilibrium is self-coordinating, spontaneously evolves through the self-organization process to finally balances and settles on the edge of chaos, and self-organized criticality emerges.

There is a clear difference between weak chaos and complete chaos: chaotic systems are very sensitively dependent on initial conditions, and a small initial disturbance will increase exponentially over time. Therefore, this uncertainty makes long-term predictions for large systems impossible. The uncertainty of the behavior of a weak chaotic system grows with time, but the growth rate is much slower than that of a complete chaotic system. It follows a power law rather than an exponential growth like a chaotic system. The system is on the edge of chaos. For a system with weak chaotic phenomena and self-organized criticality characteristics, it does not have such a time scale. There is a long-term correlation in time between events, and it has memory ability and therefore is predictable. Although self-organized criticality is also sensitive to disturbances, it is not sensitive to all disturbances. Only the individual components in the critical state will have long-range correlations, and the system state is not affected by initial conditions, so Bak calls self-organizing criticality is weak chaos, but it is not equivalent to chaos. Weak chaos is the early region where the system enters a completely chaotic state. It just means that it has chaotic properties in some aspects, such as orderliness and violent disturbances.

The economic system is neither in a stable subcritical state nor in a chaotic supercritical state, but a self-organized critical state in between. The enterprise system contains a large number of elements, and the elements have complex nonlinear interactions with each other. And the enterprise is an open system, the external environment continuously provides information and materials to the system, and the enterprise also dissipates materials and information to the external environment. There are also a large number of phenomena that obey power distribution in enterprises, such as commodity price fluctuations. The enterprise system can spontaneously evolve toward a critical state under the complex effects of internal and external environments. A small disturbance in the critical area will cause the macroscopic fluctuations or sudden changes of the system, and the enterprise is prone to crisis. If there is a synergistic effect between the subsystems, the system can effectively prevent and eliminate the adverse effects of self-organized criticality, and expand the beneficial effects.

Appendix

1. **Logistic chaotic map**

```
clear all
a=input('a=');
x=input('x0=');
n=input('n=');
x1=linspace(0,1,100);
plot(x1,a*x1.*(1-x1),'-b',x1,x1,'-g')
hold on
for i=1:n;
    y=a*x*(1-x);
    pause
    line([x x],[x y],'color',[1 0 0]);
    pause
    line([x y],[y y],'color',[1 0 0]);
    x=y;
end
```

2. **Lyapunov exponent under Logistic map**

```
clear all;clc;
i=0;
for u=0:0.01:4
i=i+1;
x=0.05;
y=0;
for j=1:100
x=u*x*(1-x);
end
for k=1:200
x=u*x*(1-x);
y=y+log(abs(1-2*x));
end
Le(i,1)=u;
Le(i,2)=log(u)+y/200;
end
plot(Le(:,1),Le(:,2),'linewidth',2);
hold on;
n=0:0.001:4;
plot(n,0,'k','linewidth',3)';
xlabel('¦Ì');ylabel('¦Ë');
```

```
for k=1:200
x=u*x*(1-x);
y=y+log(abs(1-2*x));
end
Le(i,1)=u;
Le(i,2)=log(u)+y/200;
end
plot(Le(:,1),Le(:,2),'linewidth',2);
hold on;
n=0:0.001:4;
plot(n,0,'k','linewidth',3)';
xlabel('¦Ì');ylabel('¦Ë');
```

Summary
For students majoring in economics, management, or other liberal arts, they often find it relatively abstract and difficult to understand when they are new to complexity science. However, they would find that it is actually very interesting as learning goes on, for the complexity science can show us a distinctive management thinking, which is important for students in all majors, even they do not master the mathematical derivation methods. The main points of this chapter include: the characteristics of complex systems, dissipative structures, synergy, chaos theory, catastrophe theory, hypercycle theory, fractal theory, etc.

Important Concepts and Terms

Complexity
Complex system
Emergence
Adaptation
Dissipative structure
Fluctuation
Synergetics theory
Order parameter
Period three implies chaos
Sensitive dependence on initial conditions
Lyapunov exponents

Logistic map
Tent map
Catastrophe theory
Catastrophe progression method
Hypercycle theory
Fractal theory
Self-organized criticality.

Questions and Exercises

1. What is the emergence of complex systems?
2. For an enterprise system, what conditions are required to form a dissipative structure?
3. Briefly describe the formation process of order parameters.
4. Please use the sensitive dependence on initial conditions to explain the principle of the butterfly effect.
5. What are the main characteristics of fractals? Where can fractals be applied?

Case Study

Internet public opinion crisis

With the popularization of internet products or services, including online news information, search engines, Moments of Friends, official accounts, online videos, online communities, BBS forums, e-commerce, online games, online marketing and other forms, the spread of internet public opinion (IPO) spreads fast and widely. Negative IPO may produce a crisis, which may mislead or harm the majority of netizens. Events are easily amplified and spread, and topics are difficult to control, thus affecting the healthy development of the entire internet industry. The reasons for the IPO crisis include the following: natural factors, spreading rumors, unfair competition behavior, internal management errors, emergencies, attacks by competitors, negative speculation, etc.

The fundamental purpose of crisis management is to avoid and deal with crises in a timely manner by establishing an effective crisis prevention and handling mechanism, thereby reducing possible losses. The principle of dealing with the crisis of online public opinion is particularly important. Only by adhering to correct public relations principles can we avoid the further deterioration of the IPO crisis, gradually reduce the impact of crisis events, and avoid the expansion of the scope of public opinion. A complete IPO crisis handling mechanism should be established, and IPO should be guided to a positive side by promptly and impartially publishing the true causes and latest developments of crisis events. Through the crisis early warning mechanism, relevant online public opinion can be effectively monitored, guided and managed, and the crisis can be controlled in the bud.

Please comprehensively use the theoretical knowledge of complexity science to analyze the development process of IPO crisis events, refine the characteristics of the critical state of crisis events, and briefly analyze the prevention and handling methods of network public opinion crisis.

Further Readings

1. Wu Tong. *Research on Self-organization Methodology* [M]. Beijing: Tsinghua University Press, 2001.
2. Miao Dongsheng. *Complexity Science Research* [M]. Beijing: China Book Publishing House, 2015.
3. Yan Zexian, Fan Dongping, Zhang Huaxia. *Introduction to System Science: Exploration of Complexity* [M]. Beijing: People's Publishing House, 2006.
4. Lv Jinhu, Lu Jun'an, Chen Shihua. *Chaotic Time Series Analysis and Its Application* [M]. Wuhan: Wuhan University Press, 2002.

Structural Equation Modeling

10

Hao Zhang

Learning Target

1. Understanding the principles and applicable conditions of the structural equation modeling (SEM).
2. Constructing SEM for specific management issues.
3. Being familiar with a software that deals with SEM and analyzing the results of calculations.

Introduction

Academic achievements analysis

In the analysis of college students' academic achievement, it is difficult to directly explain the influencing factors through one or two variables. When there are many variables and the relationship between the variables is complicated, an effective quantitative tool is needed. Compared with traditional statistical methods, structural equation modeling (SEM) is a measurement research technique that can integrate measurement and analysis. It can simultaneously estimate multiple measurement indicators and latent variables in the model.

The "Study on the influence of college students' academic goals and autonomy on academic achievements" analyzes the mechanism of the influence of college students' autonomy in the two development areas of academic and friendship on academic goals and academic achievements. The study randomly selected 580 students from Wuhan's universities directly under the Ministry of Education as subjects,

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_10

and used questionnaires to conduct random sampling surveys. The establishment of SEM for autonomy and college students' academic goals, GPA, and academic ability shows that academic autonomy has a strong predictive effect on academic goals and self-aware academic ability, and it can also positively predict GPA; the academic external control can also predict these three. It shows that for this study, the degree of determination of academic goals is not only affected by the individual's subjective wishes (studying hard to enter university, etc.), but also by external factors (such as family pressure, etc.). The study also found that a good level of friendship autonomy at the university level can have a positive effect on the formation of its academic goals, but it cannot affect its evaluation of its own academic ability; on the contrary, it may improve the evaluation of its own academic ability due to external factors, for example, the social opinion that college students should have good academic ability, etc.

Source:

[1] Qiu Haozheng. *Structural Equation Modeling: Theory, Technology and Applications of LISREL* (2nd edition) [M]. Taipei: Shuangye Bookstore, 2011.
[2] Zhang Zhao. Study on the influence of college students' academic goals and autonomy on academic achievements [J]. *School Party Building and Ideological Education*, 2013, 03:42–44.

10.1 Overview of Structural Equation Modeling

1. Basic concepts of structural equation modeling

Structural equation modeling (SEM) was also called Linear Structural Relationships (LISREL) in the early days. It is a commonly used empirical research method in sociological research, which is widely used in data analysis and model testing. It originated from the path analysis invented by geneticist Eswall Wrihgt in the 1920s. Structural equation is a research methodology based on statistical analysis technology, used to deal with the inquiry and analysis of complex multivariate research data. Generally speaking, SEM is classified as advanced statistics, a part of multivariate statistics. SEM effectively integrates the two mainstream techniques of statistics "factor analysis" and "path analysis". Swedish statistician Karl Jöreskog put forward the relevant concept in 1970s and developed the analysis tool LISREL software. The advantages and technical development made SEM a trend, and it has generally become one of the necessary skills for social and behavioral science researchers. SEM began to be used in psychology, sociology and other fields in the 1970s, and was closely connected with econometrics in the early 1980s. Now SEM technology has been widely used in many disciplines.

SEM is a multivariate statistical technique based on the covariance matrix. It is used to detect and verify the hypothetical relationships between the manifest variable and the latent variable, and between the latent variables in the theoretical model. It is a statistical analysis technique that uses the corresponding linear equation system to express the causal theory on the basis of the existing causal theory. Its purpose is to explore and formulate the causal relationship in the form of cause-and-effect pattern and path diagram. Different from the traditional exploratory factor analysis, in SEM a specific factor structure can be proposed and tested whether it is consistent with the data. In addition, through multi-group analysis of structural equations, it is also possible to understand whether the relationship between the variables in different groups remains unchanged, and whether the mean values of all factors are significantly different. SEM can replace multiple regression, path analysis, factor analysis, covariance analysis and other methods.

2. Advantages of SEM

(1) SEM can handle multiple dependent variables simultaneously

SEM can consider many dependent variables simultaneously and explore the relationship between many dependent variables. It is a significant improvement compared with the traditional statistical method which can only explore the relationship between one-to-one variables.

(2) Measurement errors in the independent variables and the dependent variables in SEM can be allowed

The independent variables and the dependent variables in the structural equation will inevitably have more or less errors in the statistical process, and SEM allows such errors.

(3) SEM allows simultaneous estimation of factor structure and factor relationship

To understand the correlation between the latent variables, each latent variable is measured with multiple indicators or items. The common method is: firstly, to use factor analysis to calculate the relationship between each latent variable (i.e. factor) and the item (i.e. factor load), and the factor score is used as the observed value of the latent variable; secondly, to calculate the correlation coefficient of the factor scores as the correlation between the latent variables. These two steps are carried out simultaneously, that is, the relationships between the factor and the topic and between the factors are considered at the same time.

(4) Compared with traditional methods, SEM can use flexible measurement models

Traditional factor analysis is difficult to deal with the situation where an indicator belongs to multiple factors, while SEM can be used to build more complex models.

(5) SEM can design the relationship between the latent variables and estimate the fitting degree

Traditional path analysis only estimates the strength of each path (relationship between the variables). Besides this, SEM can also calculate the overall fitting degree of different models to the same is closer to the true relationship presented by the data.

3. Disadvantages of SEM

(1) SEM is a verificational technology rather than an exploratory technology

SEM can neither deeply explore the inner essence of events, nor be used to develop new theories. Therefore, SEM can only explain the relationship between the variables on the premise of theoretical conception, but cannot discover the causal relationship between things. For example, in model analysis, if a certain model is finally selected, this model is only the best model revised under a certain theory, but it cannot be guaranteed that it is the best among all models.

(2) SEM cannot verify the correctness of the model

SEM passing the test does not mean that the model is completely correct, but means that there is no reason to reject the model, which is reasonable and effective to a certain extent. Moreover, the model must be established on the basis of the analysis of multiple sets of sample data, and pass the test of multiple sets of sample data, and the given model may produce unexplainable results.

(3) There are many uncontrollable factors in SEM

Samples, whether too large or too small, will affect the subsequent model fitting, so samples of a relatively appropriate number are necessary. Also, the questionnaire measurement process has many uncontrollable factors due to man-made influence.

10.2 Composition of Structural Equation Modeling

1. **Measured model**

The measurement model has observed variables and latent variables. Observed variables are sometimes called indicator variables, explicit variables or explicit variables. They are data that can be obtained by scales or measurement tools, such as income, satisfaction, etc.; the latent variables cannot be directly measured, so data measured by observed variables is required, such as corporate strategy, psychological changes, etc. The latent variables can be constructed via observed variables. The observed variables are usually represented by rectangular or square symbols, while the latent variables are usually represented by oval or circular symbols.

Exogenous variable means that as a variable, it only has the function of explanatory variable. It is not affected by other variables in the system or model, and affects other variables. In the path diagram, the exogenous variable does not have arrows pointing to it, but arrows pointing to other variables from it.

Endogenous variables refer to variables to be determined by the model or variables to be studied by the model. In the path diagram, as long as there is an arrow pointing to it, it is an endogenous variable.

Latent variables include exogenous latent variables and endogenous latent variables.

The regression equation of the measurement model is as follows:

$$x = \Lambda_x \xi + \delta$$
$$y = \Lambda_y \eta + \varepsilon$$

- x a vector of exogenous indicators.
- y a vector of endogenous indicators.
- Λ_x the relationship between exogenous indicators and exogenous latent variables.
- Λ_y the relationship between endogenous indicators and endogenous latent variables.
- δ error term of exogenous index x.
- ε error term of endogenous index y.
- η endogenous latent variable.
- ξ exogenous latent variable.

2. **Structural model**

The structural model is an explanation of the causal relationship model between the latent variables. The latent variable as the cause is called the external latent variable (or exogenous latent variable, latent independent variable, and extrinsic latent variable); the latent variable as the result is called the internal cause latent variable (or endogenous latent variable, latent dependent variable, and endogenous latent variable).

The regression equation of the structural model is as follows:

$$\eta = B\eta + \Gamma\xi + \zeta$$

Structural equations are a set of equations that express the relationship between the latent variables.

Equation description:

B path coefficient, indicating the relationship between the endogenous latent variables.

Γ path coefficient, indicating the influence of the exogenous latent variables on the endogenous latent variables.

ζ residual term of structural equation.

Model assumptions:

(1) The mean value of the error termε andδin the measurement equation is zero;
(2) The mean value of the residual term ζ of the structural equation is zero;
(3) The error terms ε and δ are not correlated with the factors η and ξ, and ε and δ are not correlated;
(4) There is no correlation between the error term ζ and the factors ξ, ε, δ.

3. **SEM analysis process**

SEM is a very general and important linear statistical modeling technique, and it is usually confirmed by empirical studies. The basic idea of the structural equation model is: firstly, based on previous theories and existing knowledge, through inferences and assumptions, a model of the relationship between a set of variables is formed; then a set of observed variable data is obtained through questionnaire surveys, and the resulting covariance matrix (also called the sample matrix) based on the data. SEM is to verify the fitting degree between the hypothetical model and the sample matrix. If the hypothetical model can fit the objective sample data, the model is valid; otherwise, the model needs to be modified. If the modification still does not meet the requirements of the fitting index, it is necessary to negate the hypothetical model. The SEM analysis process is shown in Fig. 10.1.

4. **The premise of the SEM**

In order to obtain good estimation results, the data of the structural equation model must meet the relevant premise assumptions.

(1) Requirements of sample content

When analyzing through SEM, in order to obtain reliable and stable data results, the number of samples must reach a certain level, and various fitting indicators, distributions, tests and their efficacy are meaningful, and then the model can be reasonably evaluated.

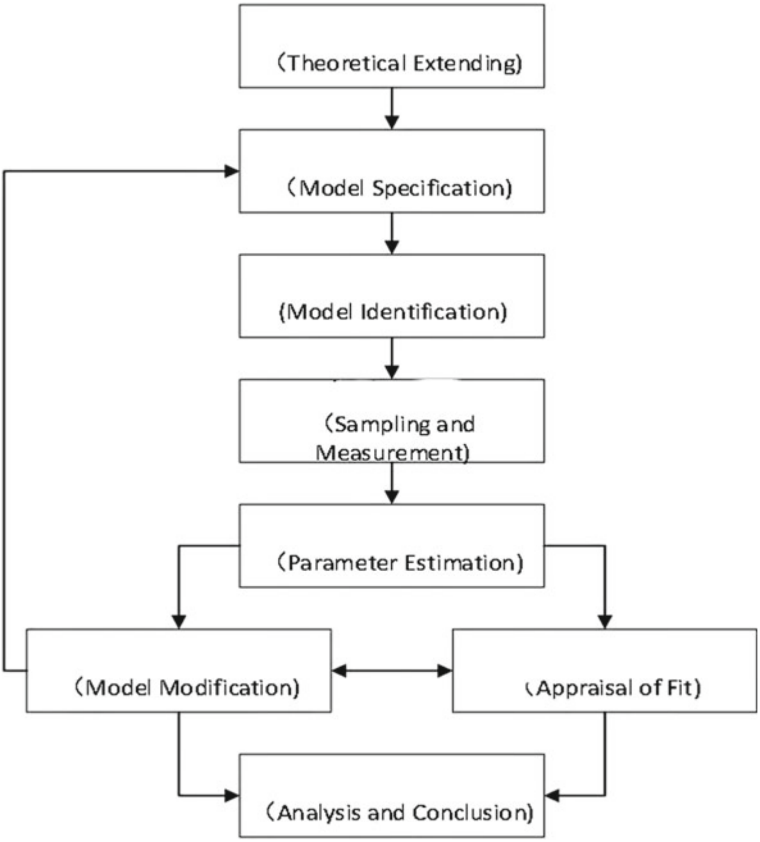


Fig. 10.1 The basic procedure of SEM analysis

(2) Distribution of data

SEM analysis requires that the data should obey the multivariate normal distribution.

(3) Non-linearity and interaction effects

Although in actual research and application, the research hotspot has always been the linear model, but under many research conditions, researchers have found that latent variables are likely to have a nonlinear relationship, and these data cannot be fully explained by the linear model. In order to effectively explain the nonlinear relationship of latent variables, the linear structural equation model can be transformed into a nonlinear structural equation model.

5. **Selection of overall model evaluation criteria**

Obtaining parameter estimates means that a specific theoretical model will be obtained. The model needs to be evaluated to determine how well this particular model fits the actual data. At least two aspects need to be evaluated:

(1) Test whether the parameters contained in the model are statistically significant;
(2) Evaluate the overall fitting degree of the model.

Among them, the fitting index is the main index to evaluate the overall fitting degree of the model. At the same time, there are many fitting indexes, and the specific meaning and calculation methods of each index are different.

Based on the fitting function, most of the fitting indices can be calculated. The degree of freedom reflects the complexity of the model. The simpler the model, the more degrees of freedom; conversely, the more complex the model, the fewer degrees of freedom. The index that can most intuitively reflect the degree of model and data fit is the value χ^2. The larger the value χ^2, the lower the fit between the model and the data. But the value χ^2 is easily affected by the sample content, that is, when N is large, the value χ^2 will be large; when N is small, the value χ^2 will be small. Therefore, many scholars have successively designed dozens of fitting indices. These dozens of fitting indexes can be roughly divided into: parsimony fit index, comparative fit index, absolute fit index, information criterion index (Table 10.1).

10.3 Application Case of Structural Equation Modeling

The relationship model of urban distribution service quality, supplier, customer relationship and customer satisfaction.

1. **Case background**

Urban distribution refers to the distribution business activities carried out within a city. It has become an important means to support the economic development of the entire city. In order to better play the role of urban distribution, its service level should be improved and perfected from the perspective of meeting customer requirements. To this end, the first step is to analyze the impact of urban distribution service quality on customer satisfaction and the dimensions included in urban distribution service quality in order to find out the weak links in urban distribution services and to make corresponding improvements.

As an integral part of urban logistics, urban distribution has always been regarded as a necessary link between production and consumption. In the 1990s, the research on logistics began to analyze the quality of logistics distribution, and the concepts of customer satisfaction and customer loyalty in logistics services

Table 10.1 Evaluation indexes and evaluation criteria of SEM fit

Index name		Evaluation standard
Absolute fit index	χ^2 (chi-square distribution)	The smaller the better
	GFI (goodness of fit index)	Greater than 0.9
	AGFI (adjusted goodness of fit index)	Be greater than 0.9
	RMR (root mean square residual)	Less than 0.05, the smaller the better
	SRMR (standard root mean square residual)	Less than 0.05, the smaller the better
	RMSEA (approximate root mean square residual)	Less than 0.05, the smaller the better
Comparative fit index	NFI (normed fit index)	Greater than 0.9, the closer to 1, the better
	TLI (Tucker-Lewis index)	Greater than 0.9, the closer to 1, the better
	CFI (comparative fit index)	Greater than 0.9, the closer to 1, the better
	IFI (incremental fit index)	Greater than 0.9, the closer to 1, the better
Information criteria index	AIC (Akaike information criterion)	As small as possible
	CAIC (consistent AIC)	As small as possible

were born from this. The logistics industry is a typical service-based industry. As an important part of the logistics industry, urban distribution has obvious service characteristics. However, there are still few research results on the quality of urban distribution and its relationship with supplier-customer relationship (SCR) and customer satisfaction (CS), while CS and SCR will directly affect corporate performance, such as market share, revenue and profit, etc. Understanding the current situation of urban distribution, especially service quality (SQ) and customer satisfaction, and analyzing its internal relationship, can provide effective management strategies for urban distribution service companies to improve service quality and enhance customer satisfaction, thereby promoting further development of urban distribution.

2. **Research hypotheses**

Combining existing theories and practices, we find that SQ of urban distribution services has an important impact on CS and SCR, and there is also a mutual influence between CS and SCR. Therefore, this study puts forward the corresponding hypothetical relationship.

According to the existing research results, SQ will directly affect CS. Therefore, the first hypothesis in this research is:

H1: SQ has a significant impact on CS.

From the perspective of relationship marketing, the success of a company depends on the ability to establish and develop long-term SCR in an increasingly competitive market. Good SQ can reduce customers' uncertainty about transactions, and can determine the probability of customers and sellers to continue transactions. American quality professional Philip Crosby et al. concluded that good SQ is a basic prerequisite for a good relationship. SQ can attract and retain customers in a relatively long period of time, thus helping to establish a good SCR. Therefore, the second hypothesis proposed is:

H2: SQ has a significant impact on SCR.

A large number of studies and practices have shown that CS is an important factor in SCR, vice versa. Therefore, this study believes that the SCR and CS are mutually influential. The hypotheses made on the relationship between the two are as follows:

H3: CS has a significant impact on SCR.

H4: The SCR has a significant impact on CS.

The relationship between SQ and the five dimensions (*see below*) is represented by the following five hypotheses:

H5: Reliability has a significant impact on SQ.

H6: Responsiveness has a significant impact on SQ.

H7: Assurance has a significant impact on SQ.

H8: Empathy has a significant impact on SQ.

H9: Tangibility has a significant impact on SQ.

See below for details about dimensions.

3. **Variable definition and structure model**

As an "experiential" product, city delivery service is intangible, and it can only be evaluated after service. Customer experience has a high degree of uncertainty and must be quantitatively evaluated through specific quality signals or service characteristics. According to the characteristics of the urban distribution industry, the measurement indicators in the SERVQUAL model for evaluating SQ are appropriately adjusted, and SQ is comprehensively evaluated from the following five dimensions:

(1) Reliability: The ability of urban distribution companies completing the promised service reliably and accurately.
(2) Responsiveness: The ability of urban distribution companies providing customers with fast service proactively.
(3) Assurance: The knowledge and politeness of urban distribution service personnel and other abilities of winning customers' trust.

(4) Empathy: The care and personalized service provided by urban distribution companies to customers, and convenience of service, communication with and understanding of customers are included.
(5) Tangibility: The physical facilities, equipment, internal environment of the store and the appearance of the staff of the urban distribution companies.

Customer satisfaction in this study is the overall feelings of customers about the service provided by the distribution companies. The basis of customer satisfaction theory is the "Disconfirmation Mode" in psychology. According to this model, the level of customer satisfaction is directly affected by two factors: the perceived quality and the customer expectations. According to the method suggested by Carman (1990), this research directly asks the difference between the customer's perception and expectation of the received urban delivery service, and regards "the actual service received meeting their own expectations (y_1)" as an observed variable of customer satisfaction. As a highly subjective indicator, customer satisfaction can also be reflected by the customer complaint rate to a certain extent. According to the survey, many companies ignore the dissatisfaction of a large number of customers. The American market research company Technical Assistance Research Program Institute (TARP) has conducted further research on the level of customer dissatisfaction. The results of the study show that only 4% of dissatisfied customers will complain to the company. Therefore, "no complaints about the quality of delivery service (y_2)" is another observed variable of customer satisfaction.

The concept of relationship marketing provides new ideas for the development of the logistics industry. Relationship marketing emphasizes the long-term relationship between customers and enterprises, and regards value creation as the result of the joint efforts of both parties to the transaction. Since relationship marketing focuses on how to establish long-term relationships with customers, this article takes "cooperation time with distribution companies (y_3)" as an observation variable to measure the supplier-customer relationship. In addition, whether customers are willing to continue to cooperate with the distribution company reflects the relationship between the two from a dynamic perspective. Therefore, this study takes "willing to continue cooperation with the companies in the future (y_4)" as another observed variable of the supplier-customer relationship. This study uses 17 observed variables to reflect each structural variable, and the meaning of each variable is shown in Table 10.2.

The study adopts SEM, which can not only study the observed variables, but also the relationship of hidden variables; not only the direct effect of variables, but also the indirect effect between them, and the latter can be taken as the methodological basis of urban distribution service quality evaluation and the influence relationship between various factors.

The above variables are calculated by AMOS software. The operation interface of AMOS software is shown in Fig. 10.2.

The meanings of various functions are shown in Table 10.3.

Table 10.2 The meaning of each variable

Structure variable	Observed variable
Reliability (ξ_1)	High accuracy of delivery time (x_1) No loss, wrong or damage (x_2) Claims after loss or damage (x_3)
Responsiveness (ξ_2)	Convenience of processing (x_4) Speed of complaint resolution (x_5)
Assurance (ξ_3)	Customers' need met well (x_6) Staff being warm and friendly (x_7)
Empathy (ξ_4)	Proactively informing customers of the required information, business process and time required (x_8) Making some adjustments for customers (x_9) Considering customers in the service process (x_{10})
Tangibility (ξ_5)	Advanced and fast means of transportation (x_{11}) Appropriate dressing of service personnel (x_{12}) Easy access to the website and support system (x_{13})
City delivery customer satisfaction (CS)	The actual service received meeting their expectations (y_1) No complaints about the delivery service quality (y_2)
Supplier-customer relationship (SCR)	Cooperation time with distribution companies (y_3) Willing to continue cooperation with the companies in the future (y_4)

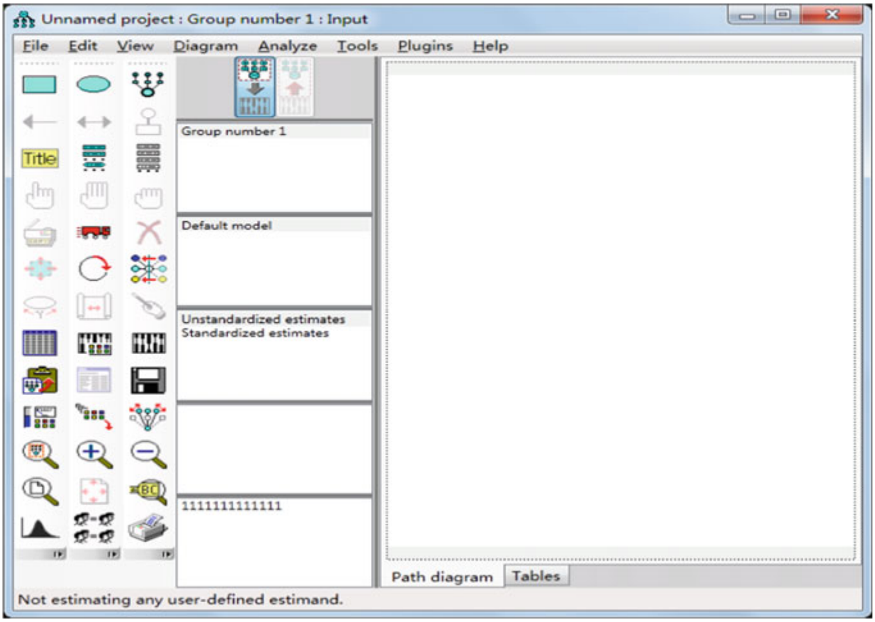


Fig. 10.2 Operation interface of Amos

Table 10.3 Functions of the main buttons in AMOS operation

	Plot observed variables		Move objects		Object properties settings
	Plot latent variables		Delete object		Object transfer
	Plotting variable indicators		Change the shape of an object		Keep symmetry
	Map causality		Rotating latent variable indicator		Select area to zoom in
	Draw correlation		Indicator variable mirror rotation		Zoom in
	Increase residual variables		Move parameter value		Shrink objects
	Title		Redeployment road map		Show full page
	List variables in the pattern		Variable fine-tuning		Adjust the road map
	List the variables in the data		Select database		Enlarge the partial path map
	Select a single object		Analysis property setting		Multi-group analysis
	Select all objects		Calculate estimate		Print path map
	Remove all selected objects		Copy path map to cut version		Recovery
	Copy objects		Storage path diagram		Pattern-defined search

The specific SEM constructed is shown in Fig. 10.3. It shows that SQ has both direct and indirect influence on CS and SCR, because CS and SCR also influence each other.

The observed variables in this study (except Y) were scored using the Likert 9-point scale, 1 means "strongly disagree", and 9 means "strongly agree". In order to maintain dimensional consistency with other observed variables, this study divides the cooperation time into 9 intervals, and different intervals correspond to different scores. The specific relationship is shown in Table 10.4. The study investigates the composition of each factor, tests the reliability and validity of each factor, performs confirmatory factor analysis and structural equation analysis, visually displays the relationship between variables through the path diagram, and decomposes the correlation coefficient to examine the direct and indirect effects of one variable on another, explores the relationship between variables, and verifies the rationality of this structural relationship.

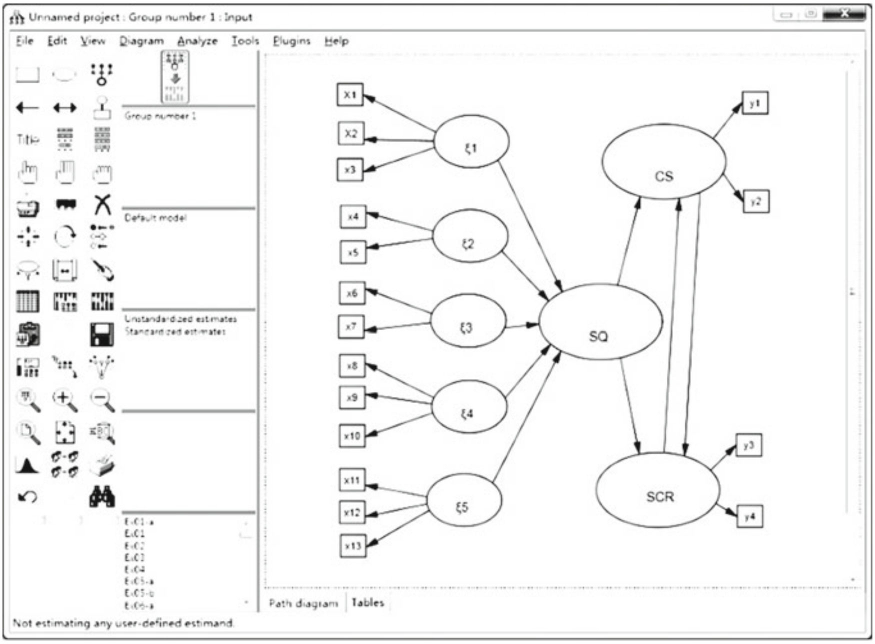


Fig. 10.3 Structural equation modeling

Table 10.4 Correlation coefficient

Points	1	2	3	4	5	6	7	8	9
Cooperation time y_3 (month)	≤3	4–6	7–9	10–12	13–15	16–18	19–21	21–24	≥25

4. **Analysis result**

This study selects four distribution companies as the object of empirical analysis. The sample data come from a questionnaire survey on 350 customers from four companies. After testing, all Cronbach α exceeds 0.7, indicating that the research scale has good internal consistency reliability. Exploratory factor analysis of 17 measurement items shows that the KMO value was 0.831 (>0.8), and the significance level of Bartlett's spheroid test also reached the standard ($sig = 0.000 < 0.05$), so CFA analysis can be performed.

The final fitting status of the SEM is shown in Table 10.5.

Table 10.5 shows that in the default model after correction, the values of CFI, GFI, NFI, IFI, RFI are good, reflecting a good fitting degree between the data and the model. The adjusted Root Mean Square of Approximation (RMSEA) is 0.054, indicating that the structural equation model meets or exceeds the recommended level. The ratio of statistics (χ^2) to degrees of freedom (df) 1.78 also shows that

Table 10.5 The final fitting status of the structural equation model

Model	χ^2	df	ρ	χ^2/df	CFI	GFI	AGFI	NFI	IFI	RFI	RMSEA
Default model	256.72	144	0.746	1.78	0.95	0.84	0.95	0.94	0.95	0.93	0.054

Table 10.6 The standardized path coefficients of the structural relationships among various endogenous variables

Hypothesis	Path relationship	Direct utility	Indirect utility	Total utility	Test result
H1	SQ → CS	0.41^b	0.08	0.49	Support
H2	SQ → SCR	0.31^a	0.26	0.57	Support
H3	CS → SCR	0.63^c			Support
H4	SCR → CS	0.26^a			Support
H5	Reliability → SQ	0.53^b			Support
H6	Responsiveness → SQ	0.72^c			Support
H7	Assurance → SQ	0.43^b			Support
H8	Empathy → SQ	0.67^c			Support
H9	Tangibility → SQ	0.45^b			Support

Note The path coefficient is a standardized value, ^{a} indicates the significance level of 0.05, ^{b} indicates 0.01, ^{c} indicates 0.001

an appropriate fitting degree. The analysis results show that the SEM proposed in this study fits the data well and has high validity. The standardized path coefficients reflecting the structural relationship among the various endogenous variables are shown in Table 10.6. Since the P values are all less than 0.05, it means that the 9 hypothetical paths have passed the significance test, and the corresponding hypotheses are supported by empirical analysis. Both are statistically significant. SQ significantly affects SCR and CS; SCR and CS are also significantly positively correlated.

5. **Conclusion**

(1) SQ has a significant impact on both SCR and CS, indicating that the hypotheses H1 and H2 have been verified. In addition, SQ also has an indirect impact on CS through SCR, and an indirect impact on SCR through CS. The total influence coefficient of SQ on CS is 0.49, with the direct influence coefficient 0.41 and the indirect influence coefficient 0.08. The total influence coefficient of SQ on SCR is 0.57, with the direct influence coefficient 0.31 and the indirect influence coefficient 0.26. The direct impact of SQ on CS is greater than the direct impact on SCR, but the indirect impact of SQ on SCR is greater. It shows that improving SQ is very important for maintaining a good SCR.

(2) The influence coefficient of CS on SCR is 0.63, and the influence coefficient of SCR on CS is 0.26. The influence relationship between the two is direct and

significant, showing that the hypotheses H3 and H4 are verified. The effect of CS on SCR is slightly stronger. Therefore, in order to establish a successful SCR and to maintain it as long as possible, the distribution company needs to improve the level of CS. On the other hand, maintaining a close and long-term SCR can enable companies to grasp customer needs more accurately and deeply, so that actual SQ can meet or exceed customer needs and expectations, which is conducive to improving CS. Therefore, the two are complementary and can form a virtuous circle.

(3) Among the five dimensions of SQ, the most significant dimension is responsiveness, with an impact coefficient of up to 0.72; the second is empathy; the least significant is tangibility. Therefore, distribution companies should distinguish the priority when improving service quality, focusing on the improvement of responsiveness and empathy, while avoiding spending too much money on modern vehicles or network construction. By strengthening personnel training and emphasizing communication with customers, the companies continuously improve SQ and increase the efficiency of urban distribution and the sustainability of profitability.

(4) The direct impact of SQ on SCR is not very large. The reason is that in addition to the quality of delivery service, there are many other factors that affect SCR, such as price and brand. But this study did not involve these factors. Therefore, in order to establish a long-term cooperative relationship with customers, the companies cannot rely solely on the delivery SQ, although improving SQ helps to establish and maintain a good SCR to a certain extent. In the future, we can conduct further research on the factors affecting SCR of the urban distribution. The companies strengthen communication with customers, and establish a long-term strategic partnership based on common interests through accurate grasp of customer needs, so as to achieve long-term stability and development of the companies.

Summary

This chapter mainly introduces SEM. Firstly, it introduces its basic concepts and the history of its development. Secondly, it elaborates the structure of SEM, explains the analysis process of its application, and gives the evaluation criteria. Finally, taking the urban distribution problem in Beijing as a case, it illustrates the application method of SEM in detail. The main points of this chapter include: the measurement model of SEM, structural model, process and evaluation criteria.

Important Concepts and Terms

structural equation modeling, SEM.
manifest variable, or observed variable.
measured variable.
latent variable.
exogenous variable.

endogenous variables.
absolute fit index, AFI.
comparative fit index, CFI.
information criteria index, ICI.

Questions and Exercises

1. What is structural equation modeling? Under what circumstances is it appropriate to use structural equation models?
2. Briefly describe the application steps of SEM.
3. Briefly describe the differences among endogenous latent variables, exogenous latent variables, and explicit variables. Point out the endogenous latent variables, exogenous latent variables, and explicit variables in Fig. 10.4.
4. What are the types of fit indices of SEM? What are the main evaluation indicators in each category?

Case study

Here it establishes a structural equation relationship model for the customer satisfaction of Datang restaurant, and analyzes the impact of hotel staff service, price perception, quality perception, and loyalty on customer satisfaction. The survey uses questionnaires, and the subjects are the customers who have been served in

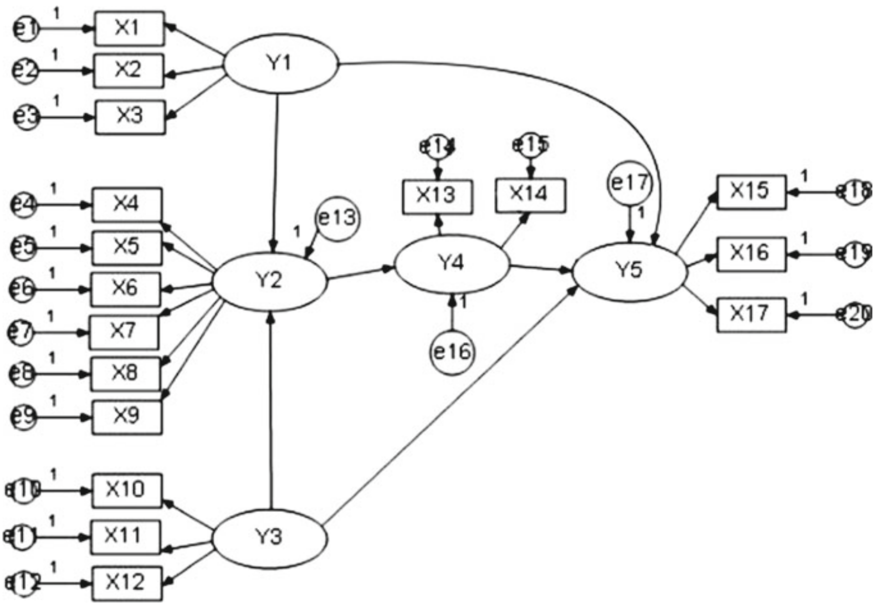


Fig. 10.4 Structural equation modeling

Table 10.7 Meaning of variables

Latent variable	Explicit variable
Staff service ξ_1	Staff service attitude x_1
	Level of facilities x_2
	Level of average serving speed x_3
Price perception ξ_2	Reasonability of pricing x_4
	Pricing compared to peers x_5
Quality perception ξ_3	Richness of dishes x_6
	Freshness of dishes x_7
	Taste of dishes x_8
Satisfaction ξ_4	Overall satisfaction x_9
	Satisfaction with the restaurant's improvement x_{10}
Loyalty ξ_5	Dining frequency x_{11}
	Recommendation rate x_{12}

this restaurant. According to the constructed theoretical model, a suitable questionnaire survey is designed to obtain data on the significant variables of customer satisfaction with the restaurant, with fitting, correction and explanation of the model proposed in the article. The questionnaire includes 5 latent variables and 12 explicit variables, in Likert 10-level measurement. The latent variables and explicit variables included in the model are shown in Table 10.7.

Research hypotheses:

H1: Employee service has a significant impact on customer satisfaction.

H2: Price perception has a significant impact on customer satisfaction.

H3: Quality perception has a significant impact on customer satisfaction.

H4: Loyalty has a significant impact on customer satisfaction.

H5: Employee service has a significant impact on price perception.

H6: Quality perception has a significant impact on price perception.

H7: Customer satisfaction has a significant impact on loyalty.

The structural equation model is shown in Fig. 10.5.

Questions:

1. Design a Likert 10-level questionnaire of the observed variables and conduct a survey on more than 200 respondents.
2. Analyze the questionnaire data, examine the composition of each latent variable, use SPSS to test the reliability and validity of each latent variable, and appropriately delete and improve the latent variables.
3. Use AMOS to decompose the correlation coefficient, study the relationship between the latent variables, and verify the correctness of the model and its assumptions. The fitting correction is performed when it's incorrect, with a conclusion drawn.

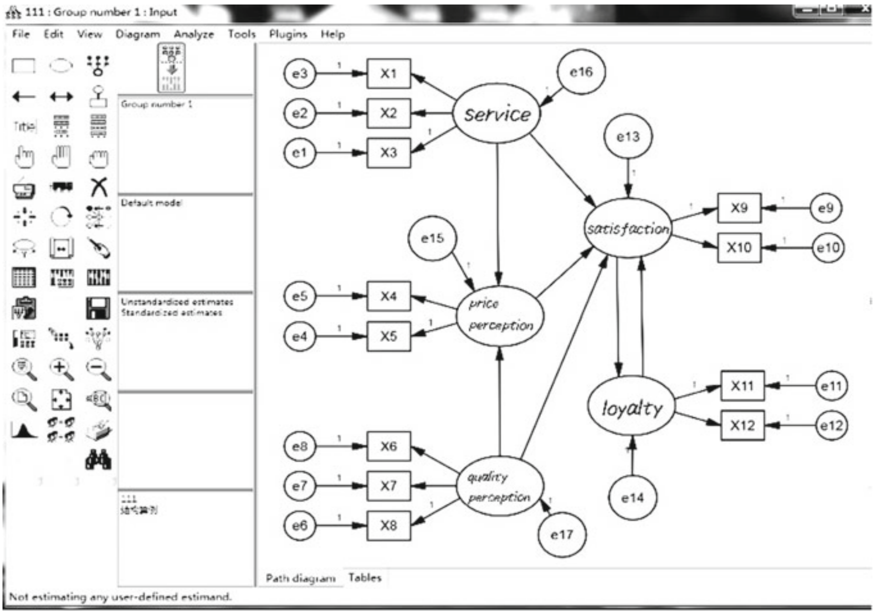


Fig. 10.5 Structural equation model

Further Readings

1. Wu Minglong. *Structural Equation Model: Advanced Amos Practice* [M]. Chongqing: Chongqing University Press, 2013.
2. Li Jianning. *Introduction to Structural Equation Modeling* [M]. Hefei: Anhui University Press, 2004.
3. Hou Jietai. *Structural Equation Model and Its Application* [M]. Beijing: Economic Science Press, 2004.
4. Huang Fangming. *Structural Equation Model: Theory and Application* [M]. Beijing: China Taxation Press, 2005.
5. Wang Jichuan. *Structural Equation Model: Method and Application* [M]. Beijing: Higher Education Press, 2011.
6. Wang Weidong. *Principles and Applications of Structural Equation Modeling* [M]. Beijing: Renmin University of China Press, 2010.

Markov Chain

11

Hao Zhang

Learning Goals

1. Understand the Markov process and the basic concept of Markov chain.
2. Understand the classification of Markov models, the connotation of continuous-time Markov chain and hidden Markov models.
3. Clarify the application scope and application methods of Markov model, and use Markov model to make predictions.

Introduction
Study Notes from a Project Manager

Due to the needs of the project, I checked the literature about marketing application of Markov chain and made some attempts.

Markov chain is a predictive tool. Given that the choice space faced by customers can be divided into n mutually exclusive states, the long-term trend of customers can be described by their transitions between different states. Transferring has two important characteristics: randomness and being of no aftereffect (or being memoryless). It's like a walk, metaphorically speaking, without a set goal. Each step depends only on where the previous step went, and there are several possibilities. This feature enables the prediction of customers' long-term behavior to be divided into several independent units, and the state of each time point is determined by the state of the previous moment and the transition probability matrix representing all possibilities, which is very flexible.

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_11

Markov chain fits the description of many economic phenomena. The most typical example is the stock market. When you are buying wealth management products, funds for example, you will notice the company's statement that the past performance is not necessarily indicative of future results, which is a live embodiment of Markov chain. The stock market is irregular. Studies have proved that the accuracy of using historical data to predict stocks or the stock market trends is not better than that of flipping a coin. In the field of marketing, such as online analysis, we often use path analysis to find out how customers use the website. However, path analysis assumes that the customer's browsing process has certain rules to follow, and the browsing behavior of the customer is actually more consistent with the Markov process. From one page to another, it is completely stochastic, so it is more accurate to describe it with the Markov model. Are all problems applicable to Markov process? A Markov chain requires tracking an object (a customer, for example) as it moves between different states over time, and its behavior is repetitive. The customer can only choose one state at a time: either to stay in the current state, or to enter another state. Some researchers use Markov models to predict the direction of the real estate market. However, most property purchases are one-off without continuity; the customers who have purchased multiple estates are more likely to own more than one estates, and their purchase behaviors are highly related with previous purchase experiences. Thus it is appropriately to describe the latter by other models.

The establishment of Markov chain model itself is not complicated, with three steps necessarily: ① Set the state; ② Calculate the transition probability matrix; ③ Calculate the result of the transition. The status may be given, such as different brands and web pages; or it needs to be divided according to the results of data analysis. In theory, the more states, the more accurate the prediction results. However, too many classifications will lose marketing significance and cause difficulties in use. RFM (Recency Frequency Monetary) is a common state division method to predict customer lifetime value using Markov chain. RFM can express customer's transaction status and reflect the relationship between customers and companies. There is a problem of optimization when using RFM to distinguish customer status. If there are many variables, cluster analysis or decision trees can be considered. The transition probability matrix can be constructed by using observations directly or by assigning values based on expert opinions. It is available to use models such as multi-state logistic regression, decision trees, neural networks, or stochastic function models. The advantage of models is that the noise in the observed data can be eliminated, and thus the transition probability can be refined to the individual. Whether the transition probability is stable is a problem that needs special attention. Markov model assumes that the transition probability of customers between different states is constant and does not change with time. Therefore, the estimation of the customer's state at each time point is carried out through the iteration of a single transition matrix. But for a longer time frame, especially in industries with long-term relationships with customers, such as banking and insurance, this assumption needs to be revised. Some important events, such as marriage, childbirth, retirement, etc., will inevitably change the transition probability of customers.

Whether the application of Markov chain can properly predict the market and the value of customers, like other models, needs to be tested by practice. The performance of the model after landing is the final judgement. However, validation with data is a necessary step in the modeling process.

Source http://blog.sina.com.cn/s/blog_6520908501017qv9.html (accessed 09/01/2013).

11.1 Markov Process

Markov process is a kind of stochastic process. Its original model is the Markov chain, which was proposed by the Russian mathematician A. A. Markov in 1906.

Markov process is a typical stochastic process. The theory studies the state of a system and its transfer. It determines the change trend of the state by studying the initial probability of different states and the transition probability between states, so as to predict the future.

Markov process has two basic characteristics (Markov property). One is "no aftereffect", that is, the future state of a thing and the probability of its occurrence only depends on the state of the thing now, and has nothing to do with the state of the previous time. In other words, it does not depend on its past evolution. The other is "ergodicity", that is, no matter what state things are in, the Markov process gradually tends to be stable over a long period of time, and it has nothing to do with the initial state. In the practice, many processes are Markov processes, such as Brownian motion caused by particles in liquids, the number of people infected with infectious diseases, inventory problems in stores, and the queues at banks, etc.

Markov processes are expressed mathematical as:

Definition 11.1 Let $X(t), t \in T$ be a stochastic process. If $X(t)$ is observed at the time of $t_1, t_2, \cdots, t_{n-1}, t_n (t_1 < t_2 < \cdots < t_{n-1} < t_n \in T)$, the corresponding observed value $x_1, x_2, \cdots, x_{n-1}, x_n$ satisfied the condition.

$$\begin{aligned} &P\{X(t_n) \le x_n \,|X(t_{n-1}) = x_{n-1}, X(t_{n-2}) = x_{n-2}, \ldots, X(t_1) = x_1\} \\ &\quad = P\{X(t_n) \le x_n \,|X(t_{n-1}) = x_{n-1}\} \end{aligned} \tag{11.1}$$

or

$$F_X(x_n; t_n|x_{n-1}, x_{n-2}, \ldots, x_2, x_1; t_{n-1}, t_{n-2}, \ldots, t_2, t_1) = F_X(x_n; t_n|x_{n-1}; t_{n-1}) \tag{11.2}$$

such a process is called the process with Markov properties or Markov process.

where

$$F_X(x_n; t_n | x_{n-1}, x_{n-2}, \ldots, x_2, x_1; t_{n-1}, t_{n-2}, \ldots, t_2, t_1)$$

represents the conditional distribution function when the value of time $X(t_n)$ is x_n under the condition of $X(t_{n-1}) = x_{n-1}, X(t_{n-2}) = x_{n-2}, \cdots, X(t_1) = x_1$.

If time t_{n-1} is regarded as "present", because $t_1 < t_2 < \cdots < t_{n-1} < t_n$, then t_n can be regarded as "future", and $t_1, t_2, \cdots, t_{n-2}$ as "past". Therefore, the above definition can be expressed as that the value of $X(t_n)$ in the future is independent of the value of $X(t_1), X(t_2), \cdots, X(t_{n-2})$ in the past, given the value of $X(t_{n-1})$ in the present state is x_{n-1}.

11.2 Markov Chain

11.2.1 Definition

Markov chain refers to a Markov process in which time and state parameters are discrete. It is the simplest Markov process.

The time studied in general Markov process is infinite, and it is a continuous variable with continuous numerical values. Two adjacent values can be divided infinitely, and the states of study are infinite. The time parameters of Markov chain are discrete values. In economic forecast, the general dates are days, months, seasons and years. At the same time, the state of the Markov chain is finite. For example, the market sales state can be "salable" and "unsalable". The future state of the market is only related to the current state, and not to the previous state (no aftereffect is established).

It is described mathematical as:

Definition 11.2 If the stochastic process $X(n), n \in T$ satisfies the following conditions:

(1) The time set is taken as a non-negative integer set $T = \{0, 1, 2, \cdots\}$ corresponding to each moment. The state space is a discrete set, denoted as $E = \{E_0, E_1, E_2, \cdots\}$, that is, $X(n)$ is a discrete time state.
(2) For any integer $n \in T$, the conditional probability satisfies:

$$\begin{aligned} P\{X(n+1) = & E_{n+1} | X(n) = E_n, X(n-1) = E_{n-1}, \ldots, X(0) = E_0\} \\ = & P\{X(n+1) = E_n | X(n) = E_n\} \end{aligned} \quad (11.3)$$

then call $X(n), n \in T$ a Markov chain, and denote that

$$P_{ij}^{(k)} = P\{X(m+k) = E_j | X(m) = E_i\}, E_i, E_j \in E \tag{11.4}$$

represent the probability that the system is in state E_i at time m and the system is in the state E_j at time $m+k$.

The conditional probability equation, that is, the probability of the state $X(m+k) = E_j$ of $X(n)$ at time $X(m+k) = E_j$ is only related to the state $X(m) = E_i$ at time m, and is independent of the state before m. It is one of the mathematical expressions of Markov property (no aftereffect). No aftereffect means that once the state of a certain stage is determined, the evolution of the subsequent process will no longer be affected by various previous states and decisions.

11.2.2 Relevant Concepts

1. **State and state variables**

State: A condition in which an objective thing may appear or exist. For example, goods may be salable or may be unsalable; the machine may ran normally or may not work properly.

Different states of the same thing must be mutually independent: two states cannot exist at the same time. The state of objective things is not fixed. When conditions change, the state often changes. For example, a product is originally unsalable in the market, but by reason of promotion and other factors, it may become a best-selling one.

State variables are generally used to represent the state: $X_t = i\begin{pmatrix} i = 1, 2, \cdots, N \\ t = 1, 2, \cdots \end{pmatrix}$, which represents a stochastic motion system, and at time $t(t = 1, 2, \cdots)$, the state is $i(i = 1, 2, \cdots N)$.

2. **State transition probability and its transition probability matrix**

(1) One-step transition probability matrix. Suppose the state space of the system is $E = (E_1, E_2, \ldots, E_n)$, and each time the system can only be in one of these states, so each state has n turns (including turns to itself), that is

$$E_i \to E_1, E_i \to E_2, \ldots, E_i \to E_i, \ldots, E_i \to E_n$$

Under the condition that the system is in state E_i at time m, the conditional probability of the system in state E_j at time $m+k$ can be expressed as:

$$P_{ij}^{(k)} = P\{X(m+k) = E_j | X(m) = E_i\}, E_i, E_j \in E \tag{11.5}$$

In particular, when $k = 1$,

$$p_{ij} = P\{X(m+1) = E_j | X(m) = E_i\}, E_i, E_j \in E$$

that is, when the system is in state E_i at time m, the conditional probability that the system is in state E_j at time $m + 1$ is called the transition probability from state E_i to state E_j through one transition. The matrix formed by the set of one-step transition probabilities of all the states of the system is called one-step state transition probability matrix. Its form is as follows:

$$P = \begin{array}{c} \\ E_1 \\ E_2 \\ \vdots \\ E_n \end{array} \begin{array}{c} \begin{array}{cccc} E_1 & E_2 & \cdots & E_n \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \end{array} \tag{11.6}$$

This matrix has the following two properties:

Non-negative: $p_{ij} \geq 0, i, j = 1, 2, \cdots, n$

The sum of row elements is 1, that is $\sum_{j=1}^{n} p_{ij} = 1, i = 1, 2, \cdots, n$

Example 11.1 There are three garment factories A, B and C producing the same kind of clothing, and there are 1000 customers. It is assumed that during the study period, no new users join and no old users quit, only some customers are transferred. It is known that in April, 500 are customers to factory A; 400 customers to B; 100 customers to C. In May, A had 400 original customers left, with 50 transferring to B and 50 to C. B had 300 original customers left, with 20 transferring to A and 80 to C. C had 80 original customers left, with 10 transferring to A and 10 to B.

Calculate its state transition probability.

Solution:

The customer transfer in May is shown in Table 11.1.

Table. 11.1 Customer transfer in May

	A	B	C	Total
A	300	50	50	400
B	120	300	80	500
C	20	20	60	100
Total	440	370	190	1000

$P_{11} = 400/500 = 0.8 \quad P_{12} = 50/500 = 0.1 \quad P_{13} = 50/500 = 0.1$

$P_{21} = 20/400 = 0.05 \quad P_{22} = 300/400 = 0.75 \quad P_{23} = 80/400 = 0.2$

$P_{31} = 10/100 = 0.1 \quad P_{32} = 10/100 = 0.1 \quad P_{33} = 80/100 = 0.8$

State transition probability matrix:

$$\mathrm{P} = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.05 & 0.75 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

(2) k-step transition probability matrix. According to the definition of one-step transition probability, k-step transition probability is the probability of the system's transition from state E_i to state E_j through k times, which can be expressed as

$$P_{ij}^{(k)} = P\{X(m+k) = E_j | X(m) = E_i\}, E_i, E_j \in E$$

Therefore, the k-step transition probability matrix of the system is a matrix composed of the k-step transition probability sets of all states. The form is as follows:

$$P^{(k)} = \begin{array}{c} \\ E_1 \\ E_2 \\ \vdots \\ E_n \end{array} \begin{array}{c} \begin{array}{cccc} E_1 & E_2 & \cdots & E_n \end{array} \\ \begin{pmatrix} p_{11}^{(k)} & p_{12}^{(k)} & \cdots & p_{1n}^{(k)} \\ p_{21}^{(k)} & p_{22}^{(k)} & \cdots & p_{2n}^{(k)} \\ \vdots & \vdots & & \vdots \\ p_{n1}^{(k)} & p_{n2}^{(k)} & \cdots & p_{nn}^{(k)} \end{pmatrix} \end{array} \tag{11.8}$$

This matrix has the following three properties:

Non-negative: $p_{ij}^{(k)} \geq 0, i, j = 1, 2, \ldots, n$

The sum of row elements is 1, that is $\sum_{j=1}^{n} p_{ij}^{(k)} = 1, i = 1, 2, \ldots, n$

$P^{(n)} = P^{(n-1)}P = P^n$

Example 11.2 The market of Borui Company has three states: E1, E2 and E3 (that is salable, ordinary and unsalable). The market transfer of the company is shown in Table 11.2. Try to find the two-step state transition probability matrix of the company's market.

Table. 11.2 Company market states transition

The current state of the company's market	The next state of the company's market		
	E1	E2	E3
E1	21	7	14
E2	16	8	12
E3	10	8	2

Solution:

First write down the one-step transition probability matrix

$$P^{(1)} = \begin{bmatrix} 0.500 & 0.167 & 0.333 \\ 0.444 & 0.222 & 0.334 \\ 0.500 & 0.400 & 0.100 \end{bmatrix}$$

The two-step state transition probability matrix can be calculated from the one-step transition probability matrix by the formula $P^{(n)} = P^n$:

$$P^{(2)} = P^2 = \begin{bmatrix} 0.500 & 0.167 & 0.333 \\ 0.444 & 0.222 & 0.334 \\ 0.500 & 0.400 & 0.100 \end{bmatrix}^2 = \begin{bmatrix} 0.491 & 0.254 & 0.255 \\ 0.488 & 0.257 & 0.255 \\ 0.478 & 0.212 & 0.310 \end{bmatrix}$$

(3) **Steady-State Probability**

The state probability is the steady-state probability when the Markov chain reaches a steady state. Under certain conditions, the Markov chain will reach a stable state after k-step transfer.

(1) Conditions of stable state. If the one-step transition probability matrix is a normal probability matrix, the Markov chain can reach a stable state.
(2) Solving the steady-state probability. According to the definition of steady state of Markov chain, when in a stable state, there is $S^{(k+1)} = S^{(k)}$, that is $S^{(k+1)} = S^{(k)}P = S^{(k)}$.

Assume $\begin{cases} S^{(k)} = (x_1, x_2, \cdots, x_n) \\ S^{(k+1)} = S^{(k)} \cdot P = S^{(k)} \end{cases}$, and $\sum_{i=1}^{n} x_i = 1$ is the state vector after k-step transition. The one-step transition probability matrix is

$$P = \begin{bmatrix} P_{11} & \ldots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \ldots & P_{nn} \end{bmatrix}$$

According to $S^{(k+1)} = S^{(k)} \cdot P = S^{(k)}$, it is expanded to

$$(x_1, x_2, \ldots, x_n)\begin{bmatrix} P_{11} & \ldots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \ldots & P_{nn} \end{bmatrix} = S^{(k)} = (x_1, x_2, \ldots, x_n) \tag{11.9}$$

By calculation, the following equation set is obtained:

$$\begin{cases} P_{11}x_1 + P_{21}x_2 + \ldots + P_{n1}x_n = x_1 \\ P_{12}x_1 + P_{22}x_2 + \ldots + P_{n2}x_n = x_2 \\ \vdots \\ P_{1n}x_1 + P_{2n}x_2 + \ldots + P_{nn}x_n = x_n \\ x_1 + x_2 + \ldots + x_n = 1 \end{cases} \tag{11.10}$$

The shift term is transformed into

$$\begin{cases} (P_{11} - 1)x_1 + P_{21}x_2 + \ldots + P_{n1}x_n = 0 \\ P_{12}x_1 + (P_{22} - 1)x_2 + \ldots + P_{n2}x_n = 0 \\ \vdots \\ P_{1n}x_1 + P_{2n}x_2 + \ldots + (P_{nn} - 1)x_n = 0 \\ x_1 + x_2 + \ldots + x_n = 1 \end{cases} \tag{11.11}$$

There are n variables in Eq. (11.11), but there are $n + 1$ equations, indicating that one of the equations is not independent and the nth equation needs to be eliminated:

$$\begin{bmatrix} (P_{11} - 1) & P_{21} & \ldots & P_{n1} \\ P_{12} & (P_{22} - 1) & \ldots & P_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{11.12}$$

let

$$P_1 = \begin{bmatrix} (P_{11} - 1) & P_{21} & \ldots & P_{n1} \\ P_{12} & (P_{22} - 1) & \ldots & P_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}, X^{((n))} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

then

$$P_1 X^{(n)} = B$$

$$X^{(n)} = P_1^{-1} B$$

that is, $X^{(n)}$ is the steady-state probability of the Markov chain.

11.3 Classification of Markov Chain Models

11.3.1 Continuous-Time Markov Chains

Definition 11.3 Suppose the stochastic process $\{X(t), t \geqslant 0\}$, the state space $I = \{i_n, n \geqslant 0\}$, if for any and $0 \leqslant t_1 < t_2 < \ldots < t_{n+1}$ $i_1, i_2, \ldots, i_{n+1} \in I$, there is.

$$\begin{aligned} &P\{X(t_{n+1}) = i_{n+1} \,|X(t_1) = i_1, X(t_2) = i_2, \ldots, X(t_n) = i_n\} \\ &\quad = P\{X(t_{n+1}) = i_{n+1} \,|X(t_n) = i_n\} \end{aligned} \tag{11.13}$$

then call $\{X(t), t \geqslant 0\}$ as a continuous-time Markov chain.

In the above formula, the conditional probability is expressed as

$$P\{X(s+t) = j \,|X(s) = i\} = p_{ij}(s, t)$$

Definition: If the transition probability of $p_{ij}(s, t)$ is independent of s, then it is said that the continuous-time Markov chain has a stationary or homogeneous transition probability, and then the transition probability is abbreviated as

$$p_{ij}(s, t) = p_{ij}(t)$$

its transition probability matrix is abbreviated as

$$P(t) = \left(p_{ij}(t)\right)$$

A continuous-time Markov chain, whenever it enters state i, has the following properties:

(1) The time in state i before moving to another state follows an exponential distribution with parameter v_i;
(2) When the process leaves state i, it then enters state j with probability p_{ij}, $\sum_{j \neq i} p_{ij} = 1$.

When $v_i = \infty$, state i is called an instantaneous state;

When $v_i = 0$, state i is called the absorbed state.

A continuous-time Markov chain transfers from one state to another according to a discrete time Markov chain. But before transferring to the next state, the time it stays in each state obeys an exponential distribution. In addition, the staying time in the state i process and the next arrival state must be independent stochastic variables.

11.3.2 Hidden Markov Model

Hidden Markov model (HMM) is a kind of Markov chain, whose states cannot be observed directly, but can be observed through a sequence of observation vectors. Each observation vector is expressed in various states through some probability density distributions. Each observation vector is generated by a state sequence with corresponding probability density distribution. Therefore, HMM is a double stochastic process, a hidden Markov chain with a certain number of states and a set of display stochastic functions. Since the twentieth century, HMM has been applied to speech recognition, computer character recognition, mobile communication core technology "multi-user detection", bioinformatics science, fault diagnosis and other fields.

HMM can be described by five elements, including two state sets and three probability matrices.

(1) The hidden state S. These states satisfy the Markov property among them and are the actually implied in the Markov model. These states are usually not available by direct observation (such as S_1, S_2, S_3, etc.).
(2) The observable state O. It is associated with the hidden state in the model and can be obtained by direct observation. (such as O_1, O_2, O_3, etc., the number of observable states is necessarily not the same as the number of hidden states.)
(3) The initial state probability matrix $\boldsymbol{\pi}$. Represents the probability matrix of the hidden state at the initial time $t = 1$. For example, when $t = 1, P(S_1) = p_1$, $P(S_1) = p_2$, $P(S_1) = p_3$, then the initial state probability matrix $\pi = \left[\, p_1\ p_2\ p_3 \,\right]$.
(4) Hidden state transition probability matrix A. It describes the transition probability between states in the hidden Markov model. Where $A_{ij} = P\left(S_j|\mathrm{S_i}\right)$, $1 \leqslant i, j \leqslant N$ is the probability that the state is S_j at time $t + 1$ under the condition that the state is S_i at time t.
(5) Observation state transition probability matrix B. Let N represent the number of hidden states and M represent the number of hidden states, then

$$B_{ij} = P(O_i|S_j), 1 \leq i \leq M, 1 \leq j \leq N$$

represents the probability that the observation state is O_i at time t and the hidden condition state is S_j.

HMM can be succinctly represented by $\lambda = (A, B, \pi)$ triples. HMM adds the set of observable states and the probability relationship between these states, which is actually an extension of the standard Markov model.

11.4 Application of Markov Chain Models

Markov analysis, also known as Markov transition matrix method, refers to a prediction method that predicts future changes of stochastic variables by analyzing the current changes of these variables under the assumption of Markov process.

The simplest type of Markov chain prediction method is to predict the most likely state in the next period. Here are the steps:

Step 1: Divide the states of the predicted object. Starting from the prediction purposes, consider the decision making needs to classify the state of the phenomenon.

Step 2: Calculate the initial probability. The state probability obtained by analyzing historical data of practical problems is called the initial probability.

Step 3: Calculate the state transition probability.

Step 4: Make prediction according to transition probability.

From the state transition probability matrix P, if the prediction object is currently in state E_i, then P_{ij} describes the possibility that the current state E_i will change to state $E_j(j = 1, 2, \ldots, N)$ in the future. According to the maximum possibility as the selection principle: choose the largest of P_{j1}, P_{j2}, …, P_{jN} as the prediction result.

1. **Calculate market share**

Example 11.3 Guangzhou, Shenzhen and Macau Special Administrative Region of the People's Republic of China produce and sell certain food ingredients. It is necessary to predict the market share in the next few months. The specific steps are as follows:

Step 1: Conduct market survey

(1) Current market share (the proportion of customers purchasing food ingredients from Guangzhou, Shenzhen and Macau).

Results: the customers who bought Guangzhou ingredients accounted for 40%, those to Shenzhen and Macau accounted for 30% respectively, and (40%, 30%, 30%) is called the current market share distribution or the initial distribution.

(2) Investigate the flow of customers.

The flow condition is:

① 40% of the customers who bought food ingredients from Guangzhou last month remain this month, and 30% of them transferred to Shenzhen and Macau respectively.

Table. 11.3 The flow of ingredients purchased by customers

	Guangzhou (%)	Shenzhen (%)	Macau (%)
Guangzhou	40	30	30
Shenzhen	60	30	10
Macau	60	10	30

② 60% of the customers who bought the ingredients from Shenzhen last month transferred to Guangzhou this month, 30% remain, and 10% to Macau.
③ 60% of the customers who bought food ingredients from Macau transferred to Guangzhou, 10% to Shenzhen, and 30% remain.

Step 2: Establish a mathematical model.

For the convenience of calculation, 1, 2 and 3 represent the food ingredients in Guangzhou, Shenzhen and Macau respectively. According to the results of market survey, the flow of customers' purchase of food ingredients is shown in Table 11.3.

$$P = \begin{pmatrix} P11 & P12 & P13 \\ P21 & P22 & P23 \\ P31 & P32 & P33 \end{pmatrix} = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.6 & 0.3 & 0.1 \\ 0.6 & 0.1 & 0.3 \end{pmatrix}$$

Step 3: Make market forecasts.

Suppose the initial market share distribution is $(P_1, P_2, P_3) = (0.4, 0.3, 0.3)$, and the market share distribution after three months is $(P_1(3), P_2(3), P_3(3))$.

If the trend of customer flow is stable for a long time, the market share will reach a stable equilibrium after a period of time.

$$(P_1(n), P_2(n), P_3(n)) = (P_1, P_2, P_3) \begin{pmatrix} P_{11}(n) & P_{12}(n) & P_{13}(n) \\ P_{21}(n) & P_{22}(n) & P_{23}(n) \\ P_{31}(n) & P_{32}(n) & P_{33}(n) \end{pmatrix}$$
$$= (P_1, P_2, P_3) \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix}$$

A stable market equilibrium means that the number of customers lost to each product is offset by the number of new customers gained during the flow of customers.

Step 4: Forecast long-term market share.

Since the one-step transition probability matrix P is a normal probability matrix, the long-term market share is the market share under equilibrium state, that is, the stationary distribution of the Markov chain.

Let the long-term market share be

$$X = (x_1, x_2, x_3)$$

then

$$\begin{cases} (x_1, x_2, x_3)\begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.6 & 0.3 & 0.1 \\ 0.6 & 0.1 & 0.3 \end{bmatrix} = (x_1, x_2, x_3) \\ x_1 + x_2 + x_3 = 1 \end{cases}$$

such that

$$X = (x_1, x_2, x_3) = (0.5, 0.25, 0.25)$$

2. Human resources forecast

Example 11.4 The employees of Borui Company are divided into five categories: intern, ordinary staff, director, general manager, and former staff. The current status (550 employees) is expressed as:

$$P(0) = (135, 240, 115, 60, 0)$$

The Company's previous record is

$$P = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0.6 & 0.25 & 0 & 0.15 \\ 0 & 0 & 0.55 & 0.21 & 0.24 \\ 0 & 0 & 0 & 0.8 & 0.2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Try to analyze the structure of employees after three years and how many new employees should be recruited into the workforce while keeping the distribution of employees unchanged (550) in three years.

Solution:

Distribution of employees after one year:

$$(1) = P(0) \cdot P$$

$$= (135, 240, 115, 60, 0)\begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0.6 & 0.25 & 0 & 0.15 \\ 0 & 0 & 0.55 & 0.21 & 0.24 \\ 0 & 0 & 0 & 0.8 & 0.2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$= (81, 198, 123, 72, 76)$$

To maintain the total number of 550, 76 have left, so 76 new employees should be recruited in the first year:

$$P'(1) = (81 + 76, 198, 123, 72, 0)$$

Distribution of employees after the second year:

$$\begin{aligned} P(2) &= P'(1) \cdot P \\ &= (157, 198, 123, 72, 0) \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0.6 & 0.25 & 0 & 0.15 \\ 0 & 0 & 0.55 & 0.21 & 0.24 \\ 0 & 0 & 0 & 0.8 & 0.2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\ &= (94, 182, 117, 83, 74) \end{aligned}$$

To keep the total number of people unchanged, 74 employees should be added:

$$P'(2) = (94 + 74, 182, 117, 83, 0)$$

Distribution of employees after the third year:

$$\begin{aligned} P(3) &= P'(2) \cdot P \\ &= (168, 182, 117, 83, 0) \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0.6 & 0.25 & 0 & 0.15 \\ 0 & 0 & 0.55 & 0.21 & 0.24 \\ 0 & 0 & 0 & 0.8 & 0.2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\ &= (101, 176, 111, 91, 72) \end{aligned}$$

72 employees should be added. At the end of the third year, the staff structure is

$$P'(3) = (173, 176, 111, 91, 0)$$

3. Profit forecast

The state in the nth period is represented by X_n:

$$X_n = \begin{cases} 1, \ products\ of\ the\ n-th\ period\ are\ salable \\ 2, \ products\ of\ the\ n-th\ period\ are\ unsalable \end{cases}$$

Let $\{X_n\}$ be a homogeneous Markov chain with a state space of $S = \{1, 2, \ldots, N\}$, and its transition matrix is $P = \left(P_{ij}\right)_{N\times N}$.

Let $r(i)$ denote the profit obtained when the system is in state i during a certain period. Such a Markov chain is called profitable.

When $r(i) > 0$, it is called profit; when $r(i) < 0$, it is called expense.

(1) Total expected profit for a limited period. Let $vk(i)$ denote the expected total profit ($k \geqslant 1, i \in S$) obtained before the state transition at of the kth step under the condition that the initial state is:

$$\begin{aligned} vk(i) &= \sum_{n=0}^{k-1} \text{expected profit for the nth period} \\ &= \sum_{n=0}^{k-1} E\{r(X_n)|X_0 = i\} \\ &= \sum_{n=0}^{k-1}(\sum_{j=1}^{N} p_{ij}^{(n)} r(j)) \end{aligned} \tag{11.14}$$

Example 11.5 shows that $k = 4$, the current month is recorded as the first month, and find the expected profit $v_4(1)$ obtained before the fourth step of the state (that is, the first four months).

Let $r(i)$ denote the profit obtained when the system is in state i in a certain period, let 1 denote "salable", and 2 denote "unsalable". Then

$$v_4(1) = r(1) + \sum_{n=1}^{4-1}\left[p_{11}^{(n)} r(1) + p_{12}^{(n)} r(2)\right]$$

$$v_4(1) = r(2) + \sum_{n=1}^{4-1}\left[p_{21}^{(n)} r(1) + p_{22}^{(n)} r(2)\right]$$

$$v_4 = (v_4(1), v_4(2))^T$$

$$P^{(n)} = \begin{bmatrix} p_{11}^{(n)} & p_{12}^{(n)} \\ p_{21}^{(n)} & p_{22}^{(n)} \end{bmatrix}$$

$$r = (r(1), r(2))^T$$

$$\begin{aligned} v_4 &= r + \sum_{n=1}^{4-1} p^{(n)} r = \sum_{n=0}^{4-1} p^{(n)} r \\ &= \left(\sum_{n=0}^{4-1} p^{(n)}\right) r = (E + P + P^2 + P^3) r \end{aligned} \tag{11.15}$$

Table. 11.4 Sales records for the past 24 months

Month	1	2	3	4	5	6
Sales status	Salable	Salable	Unsalable	Salable	Unsalable	Unsalable
Month	7	8	9	10	11	12
Sales status	Salable	Salable	Unsalable	Salable	Salable	Unsalable
Month	13	14	15	16	17	18
Sales status	Salable	Salable	Unsalable	Unsalable	Salable	Salable
Month	19	20	21	22	23	24
Sales status	Unsalable	Salable	Salable	Unsalable	Salable	Salable

Example 11.5 The electronic products produced by Bo Rui Company have two kinds of monthly market conditions: salable and unsalable. If the product is salable, it will make a profit of 500,000 yuan; if the product is unsalable, it will result in a loss of 300,000 yuan. The survey recorded sales over the past 24 months as shown in Table 11.4.

Question: If the products are salable in the current month, take the current month as the first month, and find the total expected profit before the fourth step of the state transition (that is, the first four months).

Solution:

Let 1 for "salable" and 2 for "unsalable". Given r

$$r = \begin{bmatrix} r(1) \\ r(2) \end{bmatrix} = \begin{bmatrix} 50 \\ -30 \end{bmatrix}$$

$i = 1$, V_4 has three forms of formula:

$$v_4(1) = r(1) + \sum_{n=1}^{4-1} \left[p_{11}^{(n)} r(1) + p_{12}^{(n)} r(2) \right]$$

$$v_4 = \left(\sum_{n=0}^{4-1} p^{(n)} \right) r = (E + P + P^2 + P^3) r$$

$$v_4(i) = r(i) + \sum_{j=1}^{2} p_{ij} v_3(j), i = 1, 2$$

$$v_0(i) = 0, i = 1, 2, \ldots, N$$

Find the state transition probability matrix P.

Estimate the state transition matrix P, and estimate the probability of continuously salable with statistical frequency.

$$p_{11} = \frac{7}{15 - 1} = 50\%$$

The numerator 7 is the number of continuously "salable" occurrences in the Table 11.4. The denominator 15 is the number of times "salable" appears in the Table 11.4. Since the 24th month was "salable" and there was no subsequent record, so it is reduced by 1.

$$p_{12} = \frac{7}{15-1} = 50\%,\ p_{21} = \frac{7}{9} = 78\%,\ p_{22} = \frac{2}{9} \approx 22\%$$

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \\ 0.78 & 0.22 \end{bmatrix}$$

$$r = \begin{bmatrix} r(1) \\ r(2) \end{bmatrix} = \begin{bmatrix} 50 \\ -30 \end{bmatrix},\ P = \begin{bmatrix} 0.5 & 0.5 \\ 0.78 & 0.22 \end{bmatrix}$$

$$v_4 = \left(\sum_{n=0}^{4-1} p^{(n)})r = (E + P + P^2 + P^3\right)r$$

$$v_4 = \begin{bmatrix} 1.875 & 0.875 \\ 1.86295 & 0.27905 \end{bmatrix}\begin{bmatrix} 50 \\ -30 \end{bmatrix} = \begin{bmatrix} 67.5 \\ 54.776 \end{bmatrix}$$

$$v_4(1) = 67.5$$

The result is: If the products are salable in the current month, the total expected profit obtained in the first four months is 675,000 yuan.

(2) Average profit per unit time in unlimited period

For $i \in S$, the average profit per unit time for an unlimited period with the initial state of i is defined as

$$v(i) = \lim_{k\to\infty} \frac{V_k(i)}{k} \tag{11.16}$$

Denote

$$v = [v(1)v(2)\ldots v(N)]^T,\ V_k = [v_k(1), v_k(2), \ldots, v_k(N)]^T$$

$$V_k = \left(\sum_{n=0}^{k-1} P^n\right)r = (E + P + P^2 + \cdots + P^{k-1})r$$

then

$$V = \lim_{k\to\infty} \frac{V_k}{k} = \lim_{k\to\infty} \frac{\left(E + P + P^2 + \cdots + P^{k-1}\right)r}{K}$$

If the Markov chain considered has a stationary distribution:

$$P^m = \begin{bmatrix} p_{11}^{(m)} & p_{12}^{(m)} & \cdots & p_{1N}^{(m)} \\ p_{21}^{(m)} & p_{22}^{(m)} & \cdots & p_{2N}^{(m)} \\ \vdots & \vdots & & \vdots \\ p_{N1}^{(m)} & p_{N2}^{(m)} & \cdots & p_{NN}^{(m)} \end{bmatrix} \to \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_N \\ \pi_1 & \pi_2 & \cdots & \pi_N \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_N \end{bmatrix}$$

It can be proved that:

$$v = \lim_{k\to\infty} \frac{V_k}{k} = \lim_{k\to\infty} \frac{\left(E + P + P^2 + \cdots + P^{k-1}\right) r}{K} = \lim_{k\to\infty} P^k r$$

$$= \begin{bmatrix} \pi_1 & \pi_2 & \ldots & \pi_N \\ \pi_1 & \pi_2 & \ldots & \pi_N \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \ldots & \pi_N \end{bmatrix} \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(N) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{N} \pi_j r(j) \\ \sum_{j=1}^{N} \pi_j r(j) \\ \vdots \\ \sum_{j=1}^{N} \pi_j r(j) \end{bmatrix} \quad (11.17)$$

The average profit per unit time of unlimited period has nothing to do with the initial state, which is expressed as

$$v(i) = \sum_{j=1}^{N} \pi_j r(j)$$

Chapter Summary

Markov chain is often used in the prediction or evaluation of queuing problems, and coding technology, bioinformatics, hydrological resources and other fields. The premise of its application is that the research object should be discrete events with Markov properties. When it is applied to the study management science, we should first understand the characteristics of the research object. The main points of this chapter include the basic concepts and application methods of Markov process, Markov chain, continuous-time Markov and hidden Markov model.

Key Concepts and Terms

Markov process
Markov chain
Stochastic process
State transition probability
Steady-state probability
Continuous-time Markov chain, CTMC

Table. 11.5 Air states in the past month

Date	1	2	3	4	5	6	7	8	9	10	11	12
Air condition	2	2	3	1	1	2	5	5	4	5	5	4
Date	13	14	15	16	17	18	19	20	21	22	23	24
Air condition	1	2	2	4	3	4	2	2	5	5	0	1
Date	25	26	27	28	29	30	31					
Air condition	1	3	2	4	4	1	1					

Hidden Markov model, HMM

Questions and Exercises

(1) Suppose there are six states of air quality: non-polluted, excellent, good, lightly polluted, heavy polluted, and severely polluted, which are represented by state variables $X_n = 0, 1, 2, 3, 4, 5$. The air conditions for the past month are shown in Table 11.5. Try to find the state transition probability.

(2) Suppose that an institution's investment income in a stock has three states, namely 1, 2, and 3. When the market is in state 1, the annual return is -4%. When the market is in state 2, the annual return is 30%. When the market is in state 3, the annual return is 10%. Suppose the state transition probability matrix P is applied to the weekly state transition of this stock:

$$P = \begin{pmatrix} 0.8 & 0.04 & 0.16 \\ 0.05 & 0.8 & 0.15 \\ 0.1 & 0.15 & 0.75 \end{pmatrix}$$

(1) Try to find the steady-state distribution of this stock in the market.
(2) Assume that 1 million yuan is invested in this stock for 6 years, and find the expected total profit.

3. Briefly describe the properties of no aftereffect and state transition probability of Markov chains.

Further Readings

1. Sun Qinghua & Sun Hao. *Content, Method and Technique of Stochastic Process* [M]. Wuhan: Huazhong University of Science and Technology Press, 2004.
2. William J. Stewart. *Probability Theory, Markov Chains, Queuing and Simulation* [M]. Beijing: World Publishing Corporation, 2013.
3. Cheng Weiqi, Huang Ximin *et al*. *Markov Chains: Models, Algorithms, and Applications* [M]. (Translated by Chen Xi). Beijing: Tsinghua University Press, 2015.

Grey Systems Theory 12

Hao Zhang

Learning Goals

1. Understanding the basic concept and application scope of grey systems.
2. Mastering the method of grey correlation analysis.
3. Mastering the principle of GM (1,1), and using it to build grey prediction model.

Introduction

Deng Julong and Grey Systems Theory

Professor Deng Julong, born in 1933 in Lianyuan City, Hunan Province, graduated from the Department of Electrical Engineering, Huazhong College of Engineering (the predecessor of Huazhong University of Science and Technology) in 1955. In the 1960s, Professor Deng Julong was mainly engaged in control theory research and devoted himself to the establishment of dynamic correction methods and theories for multivariable systems. In the late 1970s, when he was engaged in the research of control theory, he realized that since information and materials were flowing and operating, they must have their own law. To explore and master the law in this unknown world, he asked some questions and even boldly put forward some fantastic ideas.

In the practice of teaching and scientific research, questions flashed through his mind, and his thoughts on these questions inspired him to write many academic papers. In 1982, the international journal *Systems & Control Letters* published his paper "The Control Problems of Grey Systems", announcing the birth of

H. Zhang (✉)
School of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China
e-mail: zhhaozhhao@126.com

H. Zhang (ed.), *Models and Methods for Management Science*,
Management for Professionals, https://doi.org/10.1007/978-981-19-1614-4_12

Grey Systems Theory. Later, he successively published 20 monographs on grey systems, such as *Grey Systems (Social Economy)*, *Grey Control Systems* and *Multi-dimensional Grey Planning*.

Black indicates lack of information, white indicates complete information, and grey is called to indicate incomplete information. A system with incomplete information is a grey system. The grey systems theory challenges the previous one that hold that systems are insoluble without complete information.

The informatization of social systems and upsizing of technical systems, this trend has led to the difficulty in obtaining information and the complexity of systems. Based on the current technical means, it is difficult to obtain complete and correct information. This fact has impacted the concept of completeness, and general system theory cannot adapt to it. Grey systems theory can solve these problems.

In the establishment of grey system theory, Professor Deng Julong made bold assumptions. For example, he created grey system generation method after seeing many shortcomings of traditional mathematical statistics methods. What kind of thinking is this? Let's take a look at Deng Julong's thinking track: Although the objective world is complex and the data describing its behavior characteristics may be chaotic, it must be ordered, have some function and cause-and-effect relationship, or any system itself has its own internal law. But these rules are obscured by the confusion of the phenomena and bewildered by the disorganized appearance of data. And the generation of the behavior characteristic data of the system is an attempt to discover the inner laws from the disordered phenomena.

Nowadays, the application scope of grey system theory has been extended to many fields such as industry, agriculture, society, economy, energy, transportation, petroleum, geology, water conservancy, meteorology, ecology, environment, medicine, education, sports, military, law, finance, etc. It has successfully solved many practical problems. For example, grey comprehensive prediction of oil and gas traps, grey prediction of estuary terrain evolution, grey prediction of debris flow trend, grey diagnosis system of traditional Chinese and western medicine, etc. The establishment of grey system theory gave birth to a batch of new interdisciplinary subjects such as grey hydrology, grey geology, grey statistics, grey breeding and grey control, which promoted the development of science.

Source: Dongyan Mao. Exploring in the "grey" world: An interview of Professor Deng Julong [J]. *Statistics and Decision*. 1995, (02): pp 6–9.

12.1 Basic Concepts of Grey Systems

In 1982, Chinese scholar Professor Deng Julong published the paper "The Control Problems of Grey Systems" in the journal *Systems & Control Letters*. In the same year, Professor Deng's another paper "Grey Control Systems" was published in the *Journal of Huazhong University of Science and Technology*. The publication of these two papers marked the formal birth of grey systems theory.

The concept of grey systems is derived from black systems and white systems. A white system means that the information of the system is completely sufficient and the internal characteristics of the system are completely known. A black system means that the outside world is ignorant of the internal information of the system and can only observe and study through the connection between the system and the outside world. The grey system is between the two, which refers to a "small sample" and "poor information" uncertain system with some information is known and some information is unknown. After more than 20 years of development, the influence of grey systems theory is expanding. At present, the main content includes grey correlation analysis, grey model (GM) and grey prediction model. It is a technical system with system analysis, evaluation, modeling, prediction, decision making, control and optimization as the main body.

The knowledge of grey systems in this book comes from the classical grey systems theory proposed by Professor Deng Julong, and does not involve other modified grey systems.

12.2 Grey Correlation Analysis

Correlation degree analysis method is to measure the degree of correlation between factors according to the degree of similarity or difference in the development of factors. In fact, correlation degree analysis is a comparison of geometric relationships of statistical data in various periods of the system, that is, a quantitative comparative analysis of the development trend of dynamic process. It reveals the characteristics and degree of the dynamic correlation between things. A measure of the correlation between two systems or factors is called the correlation degree. If the relative changes between the two are basically the same in the process of system development, they are considered to be highly correlated. On the contrary, the correlation degree is small. The correlation degree analysis method has no requirement on the size of sample size, less computation, and does not need a typical distribution law. Moreover, there will be no inconsistency between the quantitative results of the correlation degree and the qualitative analysis.

12.2.1 Definition of Correlation Coefficient

A reference sequence is given as

$$Y_j = \left(y_j(1), y_j(2), y_j(3), \cdots, y_j(\mathrm{k}), \cdots, y_j(\mathrm{n})\right), j = 1, 2, 3, \ldots, s$$

A comparison sequence is given as

$$X_i = (x_i(1), x_i(2), x_i(3)_2 \cdots, x_i(\mathrm{k}), \cdots, x_i(\mathrm{n})), i = 1, 2, 3, \ldots, t$$

then the correlation coefficient is defined as follows

$$\xi_{ji}(k) = \frac{\text{minmin}\left|y_j(k) - x_i(k)\right| + \rho\text{maxmax}\left|y_j(k) - x_i(k)\right|}{\left|y_j(k) - x_i(k)\right| + \rho\text{maxmax}\left|y_j(k) - x_i(k)\right|}$$

In the calculation we take $\rho = 0.5$, where the value of ρ does not change the relative strength of the correlation.

Due to $\xi_{ji}(k)$ can only reflect the correlation between points, the correlation information is scattered, and it is inconvenient to describe the correlation between the sequences. It is necessary to integrate $\xi_{ji}(k)$. Using γ_{ji} as defined to represent the final calculated correlation degree. Based on the practical background $\left|\gamma_{ji}\right|$ greater than 0.7 called strong correlation, less than 0.3 called weak correlation. The correlation degree is given as

$$\gamma_{ji} = \frac{\sum_{k=1}^{n} \xi_{ji}(k)}{n}$$

In practical applications, there are often multiple reference sequences and multiple comparison sequences. Let r_{ij} represent the correlation degree of the comparison sequence X_i to the reference sequence Y_j, and the correlation degree matrix R can be constructed as follows

$$R = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mn} \end{bmatrix}$$

According to the size of each element of the matrix R, we can analyze and judge which factors have the major influence and which have the minor influence. The factors that have the major influence are called dominant factors. When the elements of a column are larger than those of other columns, the sub-factors corresponding to the column are called the dominant sub-factors. If the elements in a row are larger than the elements in other rows, the parent elements corresponding to the row are called the dominant parent factors.

12.2.2 An Example of Grey Correlation Analysis

Through the China Economic Information Network Statistical Database, we have found the gross national product (GNP), the added value of primary industry, secondary industry and tertiary industry of a city for five consecutive years. By analyzing the correlation between this city's GNP and the added value of the three industries, this case studies the factors that have the greatest impact on the city's GDP growth and puts forward constructive suggestions (see Table 12.1). (It can also be realized by the mathematical software MATLAB, the code is shown in the appendix).

Table. 12.1 GNP and the added value of the three industries (billion)

Indicators	Year 5	Year 4	Year 3	Year 2	Year 1
The added value of primary industry	159	159.64	150.2	136.27	124.36
The added value of secondary industry	4545.51	4292.56	4059.27	3752.48	3388.38
The added value of tertiary industry	16,626.32	15,348.61	13,669.93	12,363.18	10,600.84
GNP	21,330.83	19,800.81	17,879.4	16,251.93	14,113.58

Table. 12.2 Data standardization processing (1) (billion)

Indicators	Year 5	Year 4	Year 3	Year 2	Year 1
The added value of primary industry	1	1.004	0.945	0.857	0.782
The added value of secondary industry	1	0.944	0.893	0.826	0.745
The added value of tertiary industry	1	0.923	0.822	0.744	0.638
GNP	1	0.928	0.838	0.762	0.662

Table. 12.3 Data standardization processing (2) (billion)

Indicators	Year 5	Year 4	Year 3	Year 2	Year 1
The added value of primary industry	0	0.076	0.107	0.095	0.12
The added value of secondary industry	0	0.016	0.055	0.064	0.083
The added value of tertiary industry	0	0.005	0.016	0.018	0.024

Table. 12.4 Grey correlation coefficient

Indicators	Grey correlation coefficient
The added value of primary industry	0.504
The added value of secondary industry	0.643
The added value of tertiary industry	0.839

(1) Data standardization, that is, the data in each column are divided by the data in the first column. The results are shown in Table 12.2.
(2) Calculate the absolute value difference by subtracting the reference sequence from the three sets of comparison sequences respectively, and taking the absolute value. The results are shown in Table 12.3. The minimum value is 0 and the maximum value is 0.12.
(3) Calculate the correlation coefficient. Select $\rho = 0.5$. The correlation degree was obtained as shown in Table 12.4.

According to the calculation results, the correlation between the added value of third industry and the city's GNP is the largest, indicating that increasing investment in tertiary industry can significantly promote economic development. The correlation coefficient between the added value of primary industry and GDP development is the smallest. The above analysis results are consistent with the practical conclusions, which shows that correlation analysis can be well applied to economic analysis.

12.3 Gm (1,1)

The characteristic of the grey prediction method is: at first, the discrete data are treated as the discrete value of the continuous variable in the corresponding changing process, so that the differential equation can be used for data processing, and the original data are not directly used but its cumulative generated number can be used. The generated sequence is processed by the differential equation model. GM (1,1) is suitable for sequences with strong exponential law and can only describe monotone changes. The practical application is the same as the practical application of grey prediction.

1. **Definition of GM (1,1)**

The differential equation of grey system theory is called GM (1,1). G represents grey, M represents model, and GM (1,1) represents a first-order, one-variable differential equation model.

Let

$$X^{(0)} = \left(x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\right), X^{(1)} = \left(x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)\right),$$

then $x^{(0)}(k) + ax^{(1)}(k) = b$ is called the original form of GM (1,1).

Let

$$Z^{(1)} = \left(z^{(1)}(1), z^{(1)}(2), \cdots, z^{(1)}(n)\right),$$

where

$$z^{(1)}(k) = \frac{1}{2}\left(x^{(1)}(k) + x^{(1)}(k-1)\right), k = 1, 2, \cdots, n$$

then $x^{(0)}(k) + az^{(1)}(k) = b$ is called the basic form of GM (1,1).

Let $X^{(0)}$ be a non-negative sequence: $X^{(0)}=\left(x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\right)$; $X^{(1)}$ is the 1-Ago (one time accumulation) sequence of $X^{(0)}$:$X^{(1)} = \left(x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)\right)$, where $x^{(1)}(k)=\sum_{i=1}^{k} x^{(0)}(i), k = 1, 2, \cdots, n$;;

$Z^{(1)}$ is the generated sequence immediate adjacent to the mean value of $X^{(1)}$,$Z^{(1)} =$

$$\left(z^{(1)}(1), z^{(1)}(2), \cdots, z^{(1)}(n)\right),$$

where

$$z^{(1)}(k) = \frac{1}{2}\left(x^{(1)}(k) + x^{(1)}(k-1)\right), k = 1, 2, \cdots, n$$

If $\hat{a} = [a, b]^T$ is a parametric column and $Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}, B = \begin{bmatrix} -Z^{(0)}(2) & 1 \\ -Z^{(0)}(3) & 1 \\ \vdots & \vdots \\ -Z^{(0)}(n) & 1 \end{bmatrix}$, then the least square estimation parameter column of GM (1,1)$x^{(0)}(k) + az^{(1)}(k) = b$ satisfies $\hat{a} = [a, b]^T = \left(B^T B\right)^{-1} B^T Y$.

Let $X^{(0)}$ be a non-negative sequence, $X^{(1)}$ is the 1-Ago (one time accumulation) sequence of $X^{(0)}$, $Z^{(1)}$ is the generated sequence immediate adjacent to the mean value of $X^{(1)}$, then $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$ is white differential equation of GM (1,1) model of $x^{(0)}(k) + az^{(1)}(k) = b$, also called shadow equation.

$B, Y, \hat{a}$ satisfy

$$\hat{a} = [a, b]^T = \left(B^T B\right)^{-1} B^T Y,$$

then

(1) The solution of the white differential equation $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$ (also known as the time response function) is

$$x^{(1)}(t) = \left(x^{(1)}(1) - \frac{b}{a}\right)e^{-at} + \frac{b}{a}.$$

(2) The time response function sequence of GM (1,1) $x^{(0)}(k) + az^{(1)}(k) = b$ is

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a}\right)e^{-ak} + \frac{b}{a}, k = 1, 2, \ldots, n$$

(3) Reducing value is

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) = \left(1 - e^{a}\right)\left(x^{(0)}(1) - \frac{b}{a}\right)e^{-ak}, k = 1, 2, \ldots, n$$

Generally, the parameter $-a$ is called the development coefficient, and b is the grey action. $-a$ reflects the development trend of $\hat{x}_1$ and $\hat{x}_0$. The grey action b in GM (1,1) model is the data mined from the background value. It reflects the relationship between data changes and is a concrete manifestation of the extension of connotation.

After the establishment of GM (1,1), it is necessary to conduct model test to check whether the stability and accuracy of the model meet the standard until it can reach the expected standard. The test method include relative residual test, variance test and confidence P value test.

2. **An Example of GM (1,1) Model**

Let original sequence

$$\begin{aligned} X^{(0)} &= \left(x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), x^{(0)}(4), x^{(0)}(5)\right) \\ &= (2.87, 3.28, 3.34, 3.39, 3.68) \end{aligned}$$

Solution:

Step 1: After one accumulation process on the original sequence

$$\begin{aligned} X^{(1)} &= \left(x^{(1)}(1), x^{(1)}(2), x^{(1)}(3), x^{(1)}(4), x^{(1)}(5)\right) \\ &= (2.87, 6.15, 9.49, 12.88, 16.56) \end{aligned}$$

Step 2: Test the smoothness of $X^{(0)}$.

From $\alpha(k) = \frac{x^{(0)}(k)}{x^{(1)}(k-1)}$, we get

$$\begin{aligned} \alpha^{(1)}(3) &= \frac{x^{(0)}(3)}{x^{(1)}(2)} = \frac{3.34}{6.15} \approx 0.54 \\ \alpha^{(1)}(4) &= \frac{x^{(0)}(4)}{x^{(1)}(3)} = \frac{3.39}{9.49} \approx 0.31 < 0.5 \\ \alpha^{(1)}(5) &= \frac{x^{(0)}(5)}{x^{(1)}(4)} = \frac{3.68}{12.88} \approx 0.29 < 0.5 \end{aligned}$$

When $k > 3$, the quasi-smooth condition is satisfied.

Step 3: Check whether $X^{(1)}$ has the quasi-exponential law:

From $\beta^{(1)}(k) = \frac{x^{(1)}(k)}{x^{(1)}(k-1)}$, we get

$$\beta^{(1)}(3) = \frac{x^{(1)}(3)}{x^{(1)}(2)} = \frac{9.49}{6.15} \approx 1.54$$
$$\beta^{(1)}(4) = \frac{x^{(1)}(4)}{x^{(1)}(3)} = \frac{12.88}{9.49} \approx 1.36$$
$$\beta^{(1)}(5) = \frac{x^{(1)}(5)}{x^{(1)}(4)} = \frac{16.56}{12.88} \approx 1.29$$

When $k > 3$, $\beta^{(1)}(k) \in [1, 1.5]$, $\delta = 0.5$, the quasi-exponential law is satisfied and GM (1,1) can be established for $X^{(1)}$.

Step 4: To generate sequence immediate adjacent to the mean value of $X^{(1)}$.

Let $Z^{(1)} = \frac{1}{2}\left(x^{(1)}(k) + x^{(1)}(k-1)\right)$, we get

$$Z^{(1)} = (2.87, 4.51, 8, 11.19, 14.72)$$

then

$$B = \begin{bmatrix} -4.51 & 1 \\ -8 & 1 \\ -11.91 & 1 \\ -14.72 & 1 \end{bmatrix}, Y = \begin{bmatrix} 3.28 \\ 3.34 \\ 3.39 \\ 3.68 \end{bmatrix}$$

$$B^T B = \begin{bmatrix} 442.8666 & -39.1400 \\ -39.1400 & 4.0000 \end{bmatrix}$$

thus

$$\left(B^T B\right)^{-1} = \begin{bmatrix} 0.0167 & 0.1634 \\ 0.1634 & 1.8489 \end{bmatrix}$$

Step 5: The least square estimation of parameter column $\hat{a} = [a, b]^T = \left(B^T B\right)^{-1} B^T Y$ is carried out, we get

$$\left(B^T B\right)^{-1} B^T Y = \begin{bmatrix} -0.0351 \\ 3.0792 \end{bmatrix}$$

Step 6: Determine model

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = b$$

and the time response function is

$$X^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a}\right)e^{-ak} + \frac{b}{a} = 90.5956e^{0..0351k} - 87.7256.$$

then bring in the solution obtained in Step 5.

Step 7: Calculate the simulated value of $X^{(1)}$.

$$X^{(1)} = (2.87, 6.1064, 9.4584, 12.9301, 16.5258)$$

Step 8: Calculate the simulated value of $X^{(0)}$.

$$X^{(0)} = (2.87, 3.2364, 3.352, 3.4717, 3.5957)$$

Step 9: Test error.

By calculating the residual (the actual value minus the simulated value): $\varepsilon(k) = x^{(0)}(k) - \hat{x}^{(0)}(k)$, the final residual value is 0.0158, which has a high accuracy.

12.4 Grey Prediction Model

12.4.1 Definition of Grey Prediction

Grey prediction refers to the prediction of the development and change of system behavior characteristic values using GM (1,1). Estimation of the time when outliers in behavioral characteristic values occur is called distortion prediction. Topology prediction is the overall research on the future situation and waveform of disordered waveforms. A set of interrelated grey prediction models are established for the system behavior characteristic indicators, and the changes in the coordination relationship between the prediction variables are called system predictions. These applications all use GM (1,1) to process data, and regard “random process” as “grey process” and “random variable” as “grey variable”.

12.4.2 An Example of Grey Prediction

It is known that the CO_2 content in the global atmosphere from 2006 to 2014 is shown in Table 12.5 and the CO_2 content in the global atmosphere in the next few years is predicted according to the data (Data source: http://co2now.org/).

Table. 12.5 CO_2 Content in the Global Atmosphere from 2006 to 2014 (10^{-6})

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014
Content	381.9	383.76	385.59	387.37	389.85	391.63	393.82	396.48	398.61

Table. 12.6 Predicted CO_2 Content in the Global Atmosphere from 2006 to 2023 (10^{-6})

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014
Content	381.9	383.45	385.55	387.66	389.79	391.93	394.07	396.24	398.41
Year	2015	2016	2017	2018	2019	2020	2021	2022	2023
Content	400.60	402.79	405.00	407.23	409.46	411.71	413.96	416.23	418.52

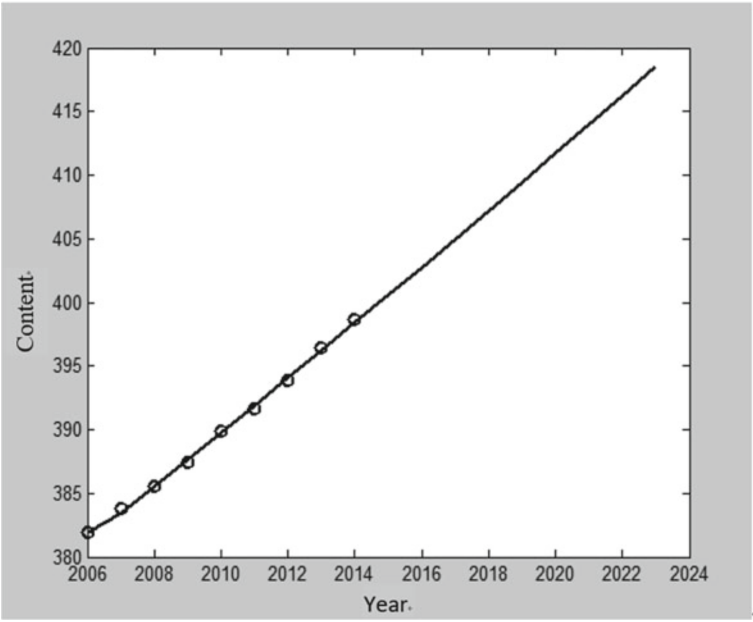


Fig. 12.1 Comparison of the original data and the predicted data of CO_2 content in the global atmosphere (the straight line is the predicted data)

The grey prediction model is essentially the development and application of the GM (1,1). The specific implementation process of the example is calculated by MATLAB software, and the detailed calculation steps can refer to the classic GM (1,1). (See the appendix for specific code).

Run the MATLAB calculation program, and the predicted data are shown in Table 12.6.

The comparison between the original data and the predicted data as shown in Fig. 12.1.

Appendix

1. **Grey correlation degree matrix program**

```
clear all;
format short;
x = [159 159.64 150.2 136.27 124.36.
4545.51 4292.56 4059.27 3752.48 3388.38.
16,626.32 15,348.61 13,669.93 12,363.18 10,600.84.
21,330.83 19,800.81 17,879.4 16,251.93 14,113.58];
n1 = size(x, 1);
for i = 1:n1.
x(i,:) = x(i,:)/x(i, 1);
end.
data = x;
consult = data(4:n1,:);
m1 = size(consult, 1);
compare = data(1: 3,:);
m2 = size(compare, 1);
for i = 1: m1.
for j = 1: m2.
t(j,:) = compare(j,:) - consult(i,:);
end.
min_min = min(min(abs(t')));
max_max = max(max(abs(t')));
resolution = 0.5;
coefficient = (min_min + resolution*max_max)./(abs(t) + resolu-
tion*max_max);
corr_degree = sum(coefficient')/size(coefficient, 2);
r(i,:) = corr_degree;
end.
r.
```

2. **GM (1,1) model program**

```
clear all;
×0 = [2.87 3.28 3.34 3.39 3.68];
n = length(×0);
lamda = ×0(1: n-1). / ×0(2: n);
range = minmax(lamda);
if range(1, 1) < exp(- (2/ (n + 2))) | range(1, 2) > exp(2/ (n + 2)).
error('The grade ratio does not fall within the range of the gray
model');
```

```
else.
% Blank line output.
disp(' ');
disp('Can be modeled with G(1, 1)');
end.
 ×1 = cumsum(×0);
for i = 2: n
z(i) = 0.5*(×1(i) + ×1(i -1));
end.
B = [-z(2: n)', ones(n - 1, 1)];
Y = ×0(2: n)';
u = B\Y;
x = dsolve('Dx + a*x = b', 'x(0) = ×0');
x = subs(x, {'a', 'b', '×0'}, {u(1), u(2), ×1(1)});
forecast1 = subs(x, 't', [0: n-1]);
y = vpa(x);
forecast11 = double(forecast1);
exchange = diff(forecast11);
forecast = [×0(1), exchange].
epsilon = ×0 - forecast;
delta = abs(epsilon./×0);
Q = mean(delta).
S1 = Std(×0,1);
S1 _new = S1*0.6745;
temp_P = find(abs(epsilon - mean(epsilon)) < S1_new);
P = length(temp_P)/ n.
```

3. Grey prediction program

```
clear.
syms a b;
c = [a b]';
A = [381.9, 383.76, 385.59, 387.37, 389.85, 391.63, 393.82, 396.48,
398.61];
B = cumsum(A);
n = length(A);
for i = 1: (n - 1).
C(i) = (B(i) + B(i + 1))/ 2;
end.
D = A; D(1) = [];
D = D';
E = [- C; ones(1, n - 1)];
c = inv(E*E')*E*D;
c = c';
```

```
a = c(1); b = c(2);
F = []; F(1) = A(1);
for i = 2: (n + 9).
F(i) = (A(1) - b/a)/ exp(a*(i - 1)) + b/a;
end.
M = []; M(1) = A(1);
for i = 2: (n + 9).
M(i) = F(i) - F(i - 1);
end.
t1 = 2006: 2014;
t2 = 2006: 2023;
M.
plot(t1, A, 'ko', 'LineWidth', 2).
hold on.
plot(t2, M, 'k', 'LineWidth', 2).
xlabel('years', 'fontsize', 11).
ylabel('content', 'fontsize', 11).
```

Summary
Grey systems theory is a great contribution made by Chinese scholars to the development of system science. Grey systems are common in the field of economy and management. For such systems, managers often have some information, but the information is incomplete. It is an effective method to apply the grey systems theory to solve the problems with grey nature in reality. The main points of this chapter include: the concept of grey systems, grey correlation analysis, classical GM (1,1), grey prediction method and technical implementation.

Key Concepts and Terms
grey system.

correlation degree.

grey correlation analysis.

grey prediction.

grey model, GM.

Questions and Exercises

1. Briefly describe the concept and characteristics of grey systems.
2. Briefly describe the principle of GM (1,1).
3. Let original sequence

$$X^{(0)} = \left(x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), x^{(0)}(4), x^{(0)}(5)\right)$$

Table. 12.7 Incomes of an Economic Sector (%)

Year	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Agriculture	39.1	41.6	43.9	44.9	45.3	45.8
Industry	45.8	43.4	42.3	41.9	41.6	41.1
Logistics industry	3.4	3.3	3.5	3.5	3.6	3.7
Financial industry	6.7	6.8	5.4	4.7	4.5	4.2

$$= (2.81, 3.36, 3.52, 3.65, 3.78),$$

determine whether GM (1, 1) model can be used. If it is available, find the sequence $X^{(1)}$.

4. Grey correlation analysis: Assuming that the incomes of an economic sector is shown in Table 1. Grey correlation method is used to calculate and analyze the relationship between agricultural development and the other three Table 12.7.

Further Readings

1. Julong Deng. *Grey Systems Theory* [M]. Wuhan: Huazhong University of Science and Technology Press, 1991.
2. Julong Deng. *Basic Approaches to Grey Systems Theory* [M]. Wuhan: Huazhong University of Science and Technology Press, 2004.
3. Sifeng Liu & Naiming Xie. *Grey Systems Theory and Its Applications* [M]. Beijing: Science Press, 2013.