Research on optimization of index system design and its inspection method: data quality diagnosis, index classification and stratification

Received 18 May 2022
Revised 19 September 2022
Accepted 1 October 2022

Kedong Yin
Marine Development Studies Institute of Ocean University of China, Qingdao, China and
Institute of Marine Economy and Management, Shandong University of Finance and Economics, Jinan, China
Yun Cao and Shiwei Zhou
Marine Development Studies Institute of Ocean University of China, Qingdao, China and
School of Economics, Ocean University of China, Qingdao, China, and
Xinman Lv
School of Economics, Ocean University of China, Qingdao, China

Abstract

Purpose – The purposes of this research are to study the theory and method of multi-attribute index system design and establish a set of systematic, standardized, scientific index systems for the design optimization and inspection process. The research may form the basis for a rational, comprehensive evaluation and provide the most effective way of improving the quality of management decision-making. It is of practical significance to improve the rationality and reliability of the index system and provide standardized, scientific reference standards and theoretical guidance for the design and construction of the index system.

Design/methodology/approach – Using modern methods such as complex networks and machine learning, a system for the quality diagnosis of index data and the classification and stratification of index systems is designed. This guarantees the quality of the index data, realizes the scientific classification and stratification of the index system, reduces the subjectivity and randomness of the design of the index system, enhances its objectivity and rationality and lays a solid foundation for the optimal design of the index system.

Findings – Based on the ideas of statistics, system theory, machine learning and data mining, the focus in the present research is on "data quality diagnosis" and "index classification and stratification" and clarifying the classification standards and data quality characteristics of index data; a data-quality diagnosis system of "data review – data cleaning – data conversion – data inspection" is established. Using a decision tree, explanatory structural model, cluster analysis, K-means clustering and other methods, classification and hierarchical method system of indicators is designed to reduce the redundancy of indicator data and improve the quality of the data used. Finally, the scientific and standardized classification and hierarchical design of the index system can be realized.

Funding: This paper is supported by the National Social Science Found Major Projects of China (14ZDB151), Humanities and social science research project of the Education Ministry of China (21YJAZHO45), and the Fundamental Research Funds for the Central Universities (202161047).

Marine Economics and Management
Vol. 5 No. 2, 2022
pp. 105-146
Emerald Publishing Limited
2516-158X
DOI 10.1108/MAEM-05-2022-0005

Originality/value – The innovative contributions and research value of the paper are reflected in three aspects. First, a method system for index data quality diagnosis is designed, and multi-source data fusion technology is adopted to ensure the quality of multi-source, heterogeneous and mixed-frequency data of the index system. The second is to design a systematic quality-inspection process for missing data based on the systematic thinking of the whole and the individual. Aiming at the accuracy, reliability, and feasibility of the patched data, a quality-inspection method of patched data based on inversion thought and a unified representation method of data fusion based on a tensor model are proposed. The third is to use the modern method of unsupervised learning to classify and stratify the index system, which reduces the subjectivity and randomness of the design of the index system and enhances its objectivity and rationality.

Keywords Index system design, Index data quality diagnosis, Index classification and stratification

Paper type Research paper

1. Introduction

Index system design is a systematic project with complex characteristics, and it is one of the important methods for the evaluation subject to recognize the representation of the evaluation object, understand the connotation of the evaluation object and analyze the mechanism of the evaluation object. As a scientific decision-making process for management cognition and comprehensive evaluation, index system design has wide application and practical needs in many fields, such as economic system, social system, ecological system, cultural system, science and technology system, education system, etc. The complexity of index system design (Peng *et al.*, 2017) is manifested in five aspects: the complexity of the research object system, the diversity of the research object system (Yin *et al.*, 2017), the diversity of evaluation objectives, the abstraction of index definition and selection and the dynamic nature of evaluation process (Jiang *et al.*, 2019) (Table 1).

At present, according to the domestic and foreign research literature on the index system construction process, the qualitative research design stage, the quantitative index screening index stage and the index quality inspection stage are the three basic stages of the index system design and construction process.

The qualitative research design stage is mainly based on the experience and knowledge of experts in the relevant fields of the evaluation object and the comprehensive evaluation index sample information already possessed, using expert consultation and opinion solicitation

Complexity feature	Specific performance description
Complexity of the research object	Any research object becomes a complex and dynamic system whole, which is composed of many elements and shows a certain complex relationship structure of elements and the overall function of the system
Diversity of the research object	The structure, function and characteristics of any research object will have changes and differences in time and space, and the construction of the index system must take into account the diverse and changing characteristics of this system
Multidimensionality of evaluation goals	Many evaluation problems in economic and social management are multi-objective decision-making problems. How to balance and synthesize the needs of multi-objective or multi-decision-making subjects is a complex problem often faced in the construction of index system
Abstraction of index definition	It is an abstract process to describe the systematic characteristics of the research object by indexing, and it is difficult to describe and define the systematic characteristics of the research object accurately and completely with "nouns" or "terms"
Dynamic nature of the evaluation process	With the changes and development of economy, society and environment, the system characteristics and evaluation objectives of the research objects will also change, which puts forward new requirements for the index system

Table 1. Complexity characteristics of index system design

method (Delphi method), grounded theory and other related qualitative research theories. Then, the systematic characteristics of the research object are excavated and analyzed according to the evaluation purpose. Finally, we comprehensively design and construct the comprehensive evaluation index system of the research object system (Hai-Min *et al.*, 2020; Gu *et al.*, 2015; Laura, 2020).

In the stage of quantitative screening of indicators, it is mainly based on the actual data of the comprehensive evaluation index system, using multiple statistical methods such as multiple regression analysis, correlation analysis, discriminant analysis, factor analysis and quantitative analysis methods such as information entropy, distribution entropy and uncertainty theory (Song and Jiang, 2015; Liu *et al.*, 2019; Zhao and Liang, 2016; Alcantud *et al.*, 2019). Based on the comprehensive evaluation indicators and hierarchical structure constructed in the qualitative research and design stage, quantitative screening and testing are carried out to reduce the redundancy between indicators and further improve the independence, hierarchy and integrity of the comprehensive evaluation indicator system.

The index quality inspection stage is mainly based on traditional mathematical models, combined with modern methods and theories such as machine learning and data fusion technology, to carry out rationality inspection, reliability inspection and other related inspections for the comprehensive evaluation index system after the quantitative screening index stage (Wang *et al.*, 2018; Chen *et al.*, 2015; Duan, 2020). In order to solve the comprehensiveness and reliability of index selection and reduce randomness and subjectivity, Yin *et al.* (2019) standardized, scientized and systematized the selection process of indicators through the identification of complete sets of indicators, expert assignment and quality inspection. It provides a certain theoretical support and reference standard for the construction of the index system (Yin *et al.*, 2019).

Data quality is an important basis for quantitative screening indicators and is the basis for data analysis and mining, indicator system design and decision support (Wang *et al.*, 2014). Wang *et al.* (1993) believe that the important reasons for "rich data and lack of information" include the lack of an effective data quality diagnosis system and low data quality, such as data inconsistency, incomplete data and data duplication (Wang *et al.*, 1993). Lee *et al.* (2015) pointed out that low-quality sample data will not only lead to low quality of comprehensive evaluation and decision-making, but also cause unmeasurable losses and unimaginable problems, so the guarantee of data quality is imminent (Lee *et al.*, 2015). Janssen *et al.* (2017) pointed out that different data sources have different data quality, and data scale magnifies problems in accuracy and diversity (Janssen *et al.*, 2017).

Data quality diagnosis is an important basis for data quality management and control. Data quality diagnosis is based on traditional statistical analysis techniques, combined with data mining techniques and data cleaning methods, to explore the causes of data disturbances, examine data distribution and test data authenticity, mutation and deletion to improve data quality. Missing data, data redundancy, data anomalies and data errors are common types of data noise (Mo, 2018; Shuang *et al.*, 2018). In the data quality diagnosis, a more significant problem is that the indicator data is incomplete, that is, there is a problem of missing indicator data. The lack of indicator data refers to the incompleteness of the indicator dataset when one or some indicator data is missing; the traditional data analysis theory and its statistical model analysis are subjective method systems for utilizing and processing indicator data. However, the premise is that it is based on fully observable index data samples. Therefore, the direct application of classic standard statistical analysis models to missing data analysis lacks adaptation, and inappropriate data missing analysis methods will lead to the accuracy and correctness of data quality diagnosis (Giorgio and Alessandro, 2016; Forghani and Yazdi, 2014). Domestic and foreign studies have given many similar definitions of data anomaly from different knowledge fields, such as anomaly, singularity, outlier (Ye, 2019), deviation (Escobar *et al.*, 2017) and so on. Data anomaly is defined as a data object that does not obey a specific

data distribution and is far away from other data points; it also refers to a data object that deviates greatly from a dataset of the same category in data classification (Ding *et al.*, 2020).

The index system structure, that is, the hierarchical structure relationship of the index system and the relationship within each layer of indicators; it includes the content of the index system in multiple dimensions such as index level, index quantity and index membership. At present, domestic and foreign research on the hierarchical division of the index system have not yet formed a normative method. Most of the literature directly determines the level division of the index system based on subjective methods such as empirical judgment and expert discussion, and its scientificity, normativeness, accuracy and feasibility are generally questioned. The scientific design index system structure is used, and the complex system is decomposed into several elements by using the ISM model (interpretation structure model), and the multi-level low-order decomposition of the system structure is realized based on practical experience and modern science and technology (Li *et al.*, 2021). With the wide application of modern methods such as machine learning, more and more scholars use classification decision tree (Luo *et al.*, 2020), cluster analysis hierarchy process (Wang *et al.*, 2009), K-means dynamic clustering, etc. for the index system layering. They effectively solve the information duplication and uncertainty of the index system in multi-attribute decision-making and improve the credibility and scientificity of the index system stratification.

At present, most of the existing literature focus on subjective experience or refer to the literature to construct the index system. Its subjectivity, randomness and blind obedience are strong, and it lacks not only the data quality diagnosis of the index system, but also the design process of the index system based on systematic, normative, reliable and scientific. This paper aims at the problems of data quality diagnosis and classification and stratification in the current index design at home and abroad. This paper uses the latest causal analysis, optimization algorithms, machine learning and other methods at home and abroad, and a standardized and scientific index system design and inspection process has been systematically designed.

2. Index data and quality

Data is the carrier of information, the basic condition for research work and the twin brother of indicators. Reliable, credible, correct and standardized high-quality data is an important guarantee for accurate evaluation and scientific grasp of the status and dynamic changes of the evaluation object. In the 1950s, foreign scholars realized the problem of data quality, and the warning slogan "Garbage In, Garbage Out" was put forward immediately. The "Guidelines for Data Governance of Banking Financial Institutions" was issued by the China Banking and Insurance Regulatory Commission in May 2018, which emphasizes the "importance of high-quality data in exerting the value of data". Incomplete data, inconsistent data and duplication of data cannot effectively utilize relevant data information (Yuan *et al.*, 2018; Xu and Li, 2020), which may lead to false conclusions and even serious decision-making mistakes.

2.1 Index data classification

Data is the embodiment of facts, which is a record of reality. Generally, data can be represented as categorical data, ordinal data and numerical data according to the different measurement scales used. According to the relationship between objective subjects and time, data can be divided into cross-sectional data, time series data and panel data. According to different collection methods, data can be divided into statistical data, survey data and experimental data. According to different data structures, data can be divided into structured data, semi-structured data and unstructured data. According to different collection methods, data can be divided into remote sensing data, aerial photography data, sonar data and instrument record data. Therefore, data collections such as multi-source, heterogeneous and frequency mixing have appeared at present (Table 2).

Division standard	Data types	Connotation	Features
Measurement scale	Categorical data	According to the classification or grouping of certain attributes of objective phenomena, data reflecting the type of things	There is a juxtaposed relationship between each category, and there is no need to distinguish the pros and cons or the size, and the order can be changed arbitrarily; for example, according to the nature of the economy, enterprises are divided into state-owned, collective, private and other economies
	Sequential data	Non-numeric data measured for the level difference or order difference between objective objects	The categories are ordered and can be represented by numerical codes; for example, the test scores can be divided into excellent, good, medium, pass and fail
	Numeric data	The observed value of the objective object measured according to the digital scale, which is expressed as a specific value	Indicates the quantitative characteristics of the objective object; for example, income 1000 yuan, age 20 years, weight 50 kg, etc.
The relationship between objective subject and time	Section data	Observation data of the same indicator of multiple objective objects at the same time point	Obtained in different spaces to describe the changes of objective objects at a certain moment
	Time series data	Observation data of an indicator of an objective object at different time points	Acquired in chronological order, describing how an objective phenomenon changes over time
	Panel data	Sample data composed of multiple cross-sectional sample observations on a time series	Also called "parallel data" or Panel Data, it is an m*n data matrix, which records a certain indicator data of m objects on n time nodes
Collection method	Statistical data	Data on the characteristics, scale, structure, level and other indicators of natural elements in a certain area	Using a certain measurement scale to measure the results of things, different measurement scales get different types of statistical data
	Survey data	Indicator data collected through surveys or observations without external interference	Statistical indicator data about socio-economic phenomena are basically survey observation data
	Experimental data	Metric data collected from control subjects	Much of the data in the natural sciences come from experimental data
Data structure	Structured data	Data with rigorous logical and physical structure stored in the database	It is expressed as a two-dimensional structure, the data is in row units, and the attributes of each row of data are the same
	Semi-structured data	Data with multiple types of textual, irregular, incomplete or implicit structures	Entities belonging to the same class can have different properties; for example, storing employee resumes
	Unstructured data	It can only be stored in the form of various types of files, and the structure is not fixed	Generally stored in binary data format; for example, various documents, pictures, videos, etc.

Table 2.
Indicator data types, connotations and characteristics

2.2 Index data quality

Since the 1950s, "quality" has gradually become a word widely used and researched. The word was initially mainly used in material products and then expanded to the field of services and data research (Huang and Chen, 2021). Since the 1980s, with the rise of information technology, data quality has become a research hotspot. Wang and Strong (1996) proposed that data quality refers to the use suitable for data consumers, and the quality of data depends on the individual who uses the data, and there are differences in the "use suitability" of different users in different environments (Wang and Strong, 1996). The National Institute of Statistical Sciences puts forward seven viewpoints on data quality research, namely, (1) Data is a product; (2) the data has quality, and this quality stems from the process of generating the data; (3) data quality can be measured and improved in principle; (4) data quality is related to the environment; (5) data quality is multi-dimensional; (6) data quality has multi-scale characteristics; and (7) human factors are the core. Yuan *et al.* (2004) and Shamim *et al.* (2019) proposed that data quality can be described by correctness, accuracy, non-contradiction, consistency and completeness (Yuan *et al.*, 2004; Shamim *et al.*, 2019).

Data quality is that the index data itself is consistent with the analysis purpose of the index data, which can meet the explicit needs or implicit needs of the index system. In different indicator systems, data consumers have different requirements for data quality. Some consumers mainly focus on the accuracy and consistency of data, while others pay attention to the real-time and relevance of data. Therefore, when the user can successfully perform data analysis on the data, the data quality can be considered to meet the requirements. High-quality index data can effectively reduce the error of research conclusions and improve the scientificity of index system design (Table 3).

Data quality is the life of data. For the statistical survey index data, considering the negligence of work in the survey process, the casual attitude of the respondents, the design of the questionnaire structure and the selection of sampling methods, data bias may occur. Using the information system to collect the data of the micro-subject business activities, in the process of data entry, conversion and database connection, conversion, problems such as field errors, data duplication, and missing records may occur. Some data published by the statistics department may have quality problems due to experience judgment, subjective consciousness, human interference, statistical system and other reasons. Data source channels are generally divided into a single data source and multiple data sources. In the

Table 3.
Meaning of data quality characteristics

Nature	Meaning
Normative	The extent to which data conforms to data standards, data models, business rules, metadata or authoritative reference data
Accuracy	The index data can truly reflect the real situation of the objective phenomenon, the error between it and the true value is small, and the data is unique and not repeated
Timeliness	The first publication of indicator data can reflect the described socio-economic phenomenon for the first time, with a very short time lag
Consistency	Indicator data from different sources are highly consistent in terms of indicator meaning, time, scope, caliber, comparability and other attributes
Integrity	The indicator data can fully reflect the various attributes of the indicator, and there are no missing records or missing attributes
Reliability	The sources and channels of indicator data are transparent and authoritative, and the statistical methods for data collection are standardized, feasible and scientific
Acquired	How easy it is for users of the indicator data to obtain their statistics and related auxiliary information
Cohesiveness	Users can correctly understand the meaning of the data, and the indicator data and its auxiliary information are highly correlated with the data source and other related data

process of schema design of indicator data and actual data processing, data quality problems are often caused (Liu and Lin, 2019) (see Table 4).

2.3 Factors affecting data quality

The factors that affect data quality include information, technology, process and management (Table 5).

3. Index data quality diagnosis

Data quality diagnosis is a method to explore the basic characteristics of data, test the reliability of data, find abnormal data, misentered data or artificial data in data series, eliminate or correct wrong data and avoid bias or even wrong research conclusions. In a broad sense, it is to test and judge whether the sample data has the quality characteristics of standardization, accuracy, cohesion, integrity, reliability, consistency and timeliness (Bao *et al.*, 2016), ensure the objectivity, accuracy and authenticity of the empirical analysis results, improve the credibility and authority of the index data, give full play to the information value of data resources and provide a solid guarantee for the construction of a scientific and reasonable index system. Therefore, before the classification, stratification and optimization inspection of the index system, it is necessary to carry out data quality diagnosis, fill in the missing data, eliminate similar data and deal with abnormal data or inconsistent data, so as to effectively ensure the data quality.

Index data quality diagnosis usually uses mathematical statistics, relevant theories or empirical judgment to analyze the quality of data, and tests and judges the reliability and rationality of data according to the internal relevance of indicators in real economic and social phenomena. Combined with data mining, data preprocessing and other methods, considering

Data sources	Source of the problem	Quality issues
Single source of truth	Pattern design issues	Lack of integrity or uniqueness constraints, unreasonable schema design
	Data processing issues	Data entry, misspellings, similar, duplicate, missing records, contradicting data
Multiple data sources	Pattern design issues	Incomplete and non-standard design of source heterogeneous data models and schemas, naming conflicts and structural conflicts
	Data processing issues	Data redundancy, contradiction or inconsistency, inconsistent samples, inconsistent timing, invalid data

Table 4. Classification of data quality Issues

Influencing factors	Specific description
Information factor	Metadata description and misunderstandings, various properties of data metrics; for example, the data source specifications are not uniform, cannot be guaranteed, and the frequency of changes is not appropriate
Technical factors	Data quality problems caused by abnormal technical aspects of specific data processing; it mainly includes data creation, data acquisition, data transmission, data loading, data use, data maintenance and other aspects
Process factor	Data quality problems caused by improper system operation procedures and manual operation procedures; mainly from the system data creation process, transfer process, loading process, use process, maintenance process and audit process and other links
Management factors	Data quality problems caused by personnel quality and management mechanism, such as personnel training, personnel management, training or management deficiencies or management defects caused by improper reward and punishment measures

Table 5. Factors affecting data quality

the characteristics of different types of data, the system carries out data processing and quality inspection and diagnoses such as data review, data cleaning, data conversion and data inspection (Cheng, 2010) (Figure 1).

3.1 Index data review

Index data review refers to the process of checking whether the sample data meets the basic characteristics of data quality (Table 3). Specifically, data review can be used to diagnose whether the sample size of data meets the basic requirements of index system design and optimization, whether it has accuracy, integrity and consistency, and whether there are missing values, abnormal values and other suspicious data.

Index data review usually uses descriptive statistical analysis methods to describe data characteristics through tabulation, classification, graphics and other general analyses. Through the frequency characteristic analysis, centralized quantity analysis, dispersion degree analysis and distribution characteristic analysis of the index data, it can determine whether there is "dirty" data in the dataset.

For the indicator X_i, its dataset can be expressed as $X_i = \{x_{ij}, j = 1, 2, \ldots\ldots, n\}$.

3.1.1 Central tendency analysis. Central tendency analysis generally uses three indicators of mean, median and mode to reflect the general level and degree of concentration of data.

(1) Mean, which is a measure of the central tendency of a dataset, used to measure the general level and central tendency of quantitative index data; it mainly includes two forms: arithmetic mean and geometric mean.

The formula for calculating the arithmetic mean $\overline{X_i^1}$ is:

$$\overline{X_i^1} = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \tag{3-1}$$

The formula for the geometric mean $\overline{X_i^2}$ of the indicator X_i is follows:

$$\overline{X_i^2} = \sqrt[n]{x_{i1} * x_{i2} * \ldots\ldots * x_{in}} \tag{3-2}$$

(2) The median is the value in the middle half of the dataset in ascending order. It can divide the value set into two equal parts and is not affected by the maximum or

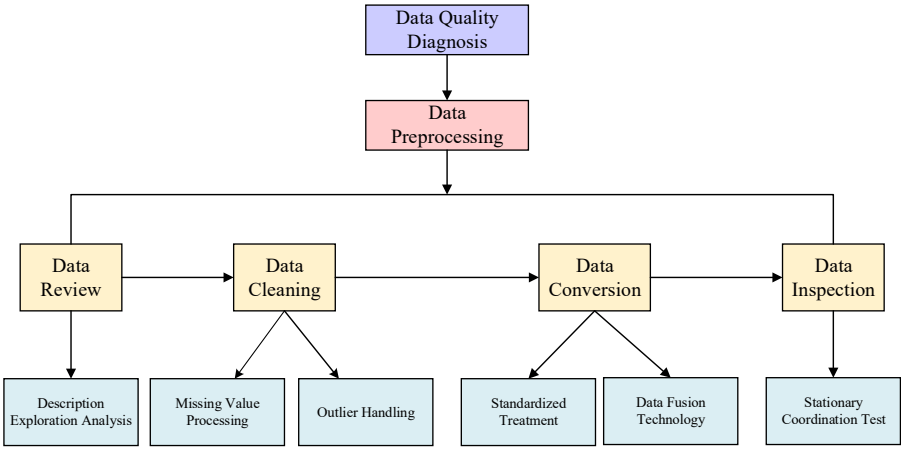

Figure 1.
Index data quality diagnosis process based on data preprocessing

minimum value. When the individual data in a dataset varies greatly, the median is often used to describe the central tendency of this set of data.

Arranged in ascending order $X_i = \{x_{i1}, x_{i2}, \ldots\ldots, x_{in}\}$, if $X_{(i)} = \{x_{(i1)}, x_{(i2)}, \ldots\ldots, x_{(in)}\}$, then

When n is odd, the median $m_{i_{0.5}}$ is:

$$m_{i_{0.5}} = x_{\left(\frac{n+1}{2}\right)} \tag{3-3}$$

When n is even, the median $m_{i_{0.5}}$ is follows:

$$m_{i_{0.5}} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \tag{3-4}$$

(3) Mode is the value with the most occurrences and the highest frequency in the indicator dataset. It can reflect the obvious central trend point of the data set and represent the general level of the data. The mode can be expressed as $Mode_i$.

In reality, the probability of each data appearing multiple times for most indicators is very small. Therefore, it is necessary to process the similarity distribution of the data (taking each data and the data within its $\pm 5\%$ or $\pm 3\%$ as similar data) to better analyze and find the distribution of the mode (similar data).

3.1.2 Data frequency analysis. The central tendency analysis of indicator data is simple and intuitive, but there are also many obvious shortcomings, and in-depth and detailed data analysis cannot be carried out. Frequency analysis is to clarify the distribution of index data at different levels through the frequency distribution table or frequency distribution histogram of each data and to identify the numerical characteristics and internal structure of the index. Frequency analysis and cross-frequency analysis are possible to directly identify whether there is an outlier problem and to check whether the range of data values is consistent with the theoretical distribution range.

The frequency distribution table consists of four elements: frequency, percentage, effective percentage and cumulative percentage. Frequency is the number of times that the index data falls within a certain data set interval. Percentage refers to the percentage of each frequency to the total number of samples. Valid percentage refers to the percentage of each frequency to the total number of valid samples (total samples minus the number of missing samples). Cumulative percentage refers to the result of accumulating each percentage step-by-step.

Also, due to the fact that the probability of each data appearing multiple times for most indicators is very small; therefore, it can be used to better find the distribution of similar data by processing the similarity distribution of the data (with each data and the data within its $\pm 5\%$ or $\pm 3\%$ range as similar data) and then performing frequency analysis on similar data of indicators.

3.1.3 Dispersion degree analysis. Dispersion degree analysis refers to the use of discrete indicators such as range, standard deviation and coefficient of variation to reflect the range of variation and degree of difference in index data.

(1) Range is the interval span (or dispersion) between the maximum value and the minimum value in the dataset. Range is the simplest measure of variation, and it measures the maximum range of variation in a set of indicator data.

The range of the indicator X_i is follows:

$$Range_i = \underset{j}{Max}\, x_{ij} - \underset{j}{Min}\, x_{ij} \tag{3-5}$$

(2) The standard deviation is a reflection of the average degree of dispersion of a dataset. The arithmetic mean of the indicator dataset and the sum of squared deviations of each sample value are calculated first, and then the square root is calculated.

The sample standard deviation $S.D._i$ of the indicator X_i is follows:

$$S.D._i = \sqrt{\frac{\sum_{j=1}^{n}\left(x_{ij} - \overline{X_i^1}\right)^2}{n-1}} \tag{3-6}$$

(3) The coefficient of variation, for a data set, is the value obtained by comparing the standard deviation with the arithmetic mean. As a relative statistic, the smaller the coefficient of variation of the dataset is, the more obvious the central tendency of the data set is, and the smaller the degree of dispersion is.

The coefficient of variation V_i of the indicator X_i is follows:

$$V_i = \frac{S.D._i}{X_i^1} \tag{3-7}$$

3.1.4 Distribution characteristic analysis. Distribution characteristic analysis is a method to explore the symmetry and steepness of the dataset distribution by calculating the skewness coefficient and the kurtosis coefficient and to check whether there are abnormal values in the data through the 3σ principle.

(1) The skewness coefficient is a characteristic statistic that describes the degree of deviation between the data distribution and the symmetrical distribution and can reflect the degree of deviation of the index data distribution. The larger the absolute value of the skewness coefficient is, the higher the degree of deviation is. When the skewness coefficient is greater than 0, the data distribution is said to be right-skewed, otherwise, it is left-skewed.

The skewness coefficient $Skew_i$ of the indicator X_i is follows:

$$Skew_i = \frac{\sum_{j=1}^{n}\left(x_{ij} - \overline{X_i^1}\right)^3 \Big/ n}{S.D._i} \tag{3-8}$$

(2) The kurtosis coefficient is a characteristic statistic that describes the peak height of the distribution function curve at the mean value, which can reflect the steepness of the index data distribution. Generally, if the kurtosis coefficient is greater than 3, it means that there are multiple abnormal values in the index dataset, and there are other data concentrated around the mode, so the distribution must be steeper.

The kurtosis coefficient $Kurt_i$ of the indicator X_i is follows:

$$Kurt_i = \frac{\sum_{j=1}^{n}\left(x_{ij} - \overline{X_i^1}\right)^4 \Big/ n}{S.D._i^2} - 3 \tag{3-9}$$

(3) The 3σ principle means that if the deviation of a certain value from the arithmetic mean in the dataset is within three times the standard deviation, it can be regarded as no abnormality in the data. Because if $p(|x_{ij} - \overline{X}_i| > 3\sigma)$ is less than 0.003, the occurrence of this value is a small probability event, so it can be determined that this value is an outlier.

Generally, the data distribution with a skewness of 0 and a kurtosis of 3 is a standard normal distribution, when the kurtosis is greater than 3, the tails on both sides of the data distribution image are thicker than the normal distribution, and the peaks are sharper.

According to the analysis results of the central tendency, the degree of dispersion and the distribution characteristics, the extreme outliers in the index data can be basically determined.

Another category of economic statistics should also be outliers if they show regular changes. For example, if the rate of change of a set of adjacent data remains roughly the same ($10 \pm 1\%$), there is also a high probability that these data will be manipulated artificially.

3.2 Index data cleaning

Indicator data cleaning refers to the process of improving data quality by analyzing suspicious data and its causes and existing forms and transforming "dirty data" into data that meets data quality characteristics. It mainly includes the processing of missing data, abnormal data and inconsistent data of indicators.

3.2.1 Missing data handling. The problem of missing values is a common phenomenon in datasets. The reasons for the lack of data may be that the definition of the indicator is not standardized, the data classification at the practical operation level is inconsistent and inaccurate, or the cost of obtaining the indicator data information is too high, or the subjective factors of the data collectors lead to the omission of data information, or information is lost because the data storage medium is faulty.

Regarding the handling of missing data, the easiest way is to ignore indicators with missing values, but it may waste a considerable amount of data, or lose an important indicator. The most common method is to estimate missing values based on the intrinsic correlation between the indicator data and fill in the missing values by interpolation; or to estimate missing values using data from other indicators based on the external correlation between multiple indicators and fill in missing values by interpolation. With the wide application of machine learning methods, scholars at home and abroad have begun to explore the use of support vector machines, random forests, cluster analysis and other machine learning methods to fill in missing data. Thus, the problem of missing data estimation can be better solved and the close relationship between indicators can be maintained.

Ignore means that if a record of an indicator has missing attribute data values, the indicator will be excluded from data analysis. This method is simple and easy to implement, but if the indicator is very important, it can easily lead to serious bias, and it is only suitable for unimportant indicators with large amounts of missing data that cannot be repaired.

Interpolation refers to the use of a certain method to determine a reasonable substitute value for the missing data of the indicator and interpolate it to the position of the original missing data, thereby reducing the estimator deviation that may be caused by missing data. Interpolation value is an effective mean to deal with missing data and its distribution. The basic idea is to use auxiliary information to find a substitute value for each missing data. Interpolation methods can be divided into two categories: single interpolation method and multiple interpolation methods.

3.2.1.1 Single interpolation. The single interpolation method is to fill in the missing values by constructing a single substitute value, so that the original dataset containing the missing value forms a complete data set. Among them, the mean interpolation method and the regression interpolation method are two commonly used single interpolation methods.

(1) Mean interpolation method. The mean value interpolation method is to use the mean value of the existing observation data of the indicator as a substitute value for the missing value. If the missing value is non-numeric indicator data, the missing value can be filled with the value with the highest frequency of the indicator in all other samples according to the principle of mode. If the missing values are numerical index data, unconditional mean interpolation and stratified mean interpolation can be used to fill the data. Mean interpolation is only suitable for simple point estimates, not for more complex analyses that require variance estimates.

- Mode interpolation method. The mode interpolation method is to analyze the data mode in the indicator and use it as the filling value for all the missing items of the indicator. The method is characterized by easy operation, simple process, insensitivity to extreme values in variables and good robustness.

For example, from January 1 to January 7, 2021, the average temperature data of a city on January 5 is missing. According to the mode interpolation method, the mode of this group of data is −1, so the average temperature on January 5 is −1℃ (Table 6).

- Unconditional mean interpolation method. The unconditional mean interpolation method uses the arithmetic mean of all observed data as the interpolation value; the advantage of this method is that it is easy to operate and can effectively reduce the deviation of univariate parameter point estimates such as mean and total. The disadvantage is that the interpolation results will distort the distribution of indicators within the sample interval, and the interpolation results will lead to an underestimation of variance in mean and aggregate estimates.

For example, in Table 7, sample No.2 of indicator 2 has missing values. According to the unconditional mean interpolation method, the arithmetic mean of samples No. 1, 3, 4, 5 is calculated as $\frac{16+31+28+29}{4} = 26$, then the surrogate value of sample No. 2 for indicator 2 is 26.

The obvious problem with the unconditional mean is that all known data are equally weighted, whereas the reality should be that closely adjacent data are more informative. Therefore, it is necessary to modify the weights of different data according to the positional relationship between missing data and adjacent data. A very simple approach is that the weight of the known data furthest away from the missing data should be at least 1, the data the next farthest should have a weight of 2 and so on. Then the unconditional mean interpolation value of the missing data of No. 2 sample of indicator 2 should be the weighted value: $(3*16 + 3*31 + 2*28 + 1*29)/9 = 25.11 \approx 25$.

- Hierarchical mean interpolation method. The hierarchical mean interpolation method stratifies the samples according to an auxiliary index and then interpolates the missing values of each layer with the average value of the known observed data of that layer.

For example, in Table 7, with indicator 3 as the auxiliary indicator, samples 1, 2, 4 are in the same layer; the substitute value of the missing value of sample No. 2 of indicator 2 is the arithmetic mean 22 of No. 1 and No. 4 of indicator 2. However, the obvious problem of the hierarchical mean interpolation method is that it does not consider the correlation between

Table 6. Average temperature from January 1st to January 7th

Date	January 1	January 2	January 3	January 4	January 5	January 6	January 7
Average temperature (℃)	−1	−2	−1	−1	/	−2	−3

Table 7. Indicator sample data sheet

Sample number	1	2	3	4	5
Indicator 1	633.56	535.76	489.73	519.03	768.99
Indicator 2	16	/	31	28	29
Indicator 3	0	0	1	0	1

indicators. If there is no correlation or a small correlation between indicator 2 and indicator 3, this method will produce large errors.

(2) Regression interpolation method. The basic idea of regression interpolation is to use the correlation between the indicators to establish a regression model, take the missing value indicator as the explained variable and use the information of other known indicator data to estimate the missing.

Suppose there is a set of $h-1 < m$ observable single indicator dataset $\{X_1, X_2, ..., X_{h-1}\}$. The indicator with missing data is then h-th indicator, which has r missing data and n-r known observation data.

Using the indicators of n-r data and the corresponding indicator $(x_{1j}, x_{2j}, ..., x_{h-1,j})$, perform multiple linear regression through the least squares method and take the regression prediction value as the missing value.

$$\widehat{x}_{hj} = \widehat{\beta}_0 + \sum_{i=1}^{h-1} \widehat{\beta}_i x_{ij}, j = 1, 2, ..., n-r \tag{3-10}$$

Use the above regression formula and judge the regression effect by the goodness of fit R^2 and the residual root mean square. The simple regression interpolation method is widely used and simple to operate. However, if the volatility and randomness of data changes are considered, multiple interpolation methods can be used to fill in missing index data.

(3) Grey prediction interpolation method. The grey prediction interpolation method is to use the index of missing values as the explained variable for small sample data and use the index of the known data as the explanatory variable to establish a multivariate grey prediction model GM (1, N) to fill the missing data. The grey prediction interpolation method can effectively deal with the defect that the regression interpolation method is not suitable for small sample data.

Suppose the original dataset of $h-1 < m$ observable single metrics is $\{X_1^{(0)}, X_2^{(0)}, ..., X_{h-1}^{(0)}\}$. There are h-th indicators $X_h^{(0)}$ with missing data, there are r missing data, and n-r known observation data. The differential equation of GM(1,N) is:

$$\frac{dx_h^{(1)}}{dt} + ax_h^{(1)} = \sum_{i=1}^{h-1} b_i x_i^{(1)} \tag{3-11}$$

Among them, $X_i^{(1)} (i = 1, 2, ..., h-1)$ is a cumulative generation of the original data sequence $X_i^{(0)}$.

Based on the least squares method, the estimated value of the parameter to be estimated $[a, b_1, ..., b_{h-1}]$ can be obtained, namely:

$$\left[\widehat{a}, \widehat{b}_1, ..., \widehat{b}_{h-1}\right]^T = \left(B^T B\right)^{-1} B^T X_h^{(0)} \tag{3-12}$$

where, $B = \begin{bmatrix} -\frac{1}{2}\left[x_h^{(1)}(1) + x_h^{(1)}(2)\right] & x_1^{(1)}(2) & ... & x_{h-1}^{(1)}(2) \\ -\frac{1}{2}\left[x_h^{(1)}(2) + x_h^{(1)}(3)\right] & x_1^{(1)}(3) & ... & x_{h-1}^{(1)}(3) \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{2}\left[x_h^{(1)}(n-r-1) + x_h^{(1)}(n-r)\right] & x_1^{(1)}(n-r) & ... & x_{h-1}^{(1)}(n-r) \end{bmatrix}$.

3.2.1.2 Multiple interpolation-random forest method. Multiple interpolation method is a relatively complex method of missing data interpolation first proposed by Rubin in 1978. Its basic principle is to generate a series of possible filling values for each missing value, so as to form several complete datasets, and then analyze each filled dataset by the method of analyzing the complete dataset and finally analyze the obtained dataset. A number of results are combined to obtain the final result. The characteristic of the multiple interpolation method is to introduce suitable random errors in the filling process, so that the estimates of all parameters are unbiased, and a better estimate of the standard error can be obtained from multiple interpolations. Compared with the single interpolation method, the multiple interpolation method can make up for the distortion caused by the missing data, thereby improving the estimation efficiency and reliability, and the multiple interpolation method can better fill in the indicators of the missing data due to unexpected events. Commonly used multiple interpolation methods include the random forest method and Markov Chain Monte Carlo (MCMC) simulation method.

(1) Random Forest. Random forest is an ensemble learning theory. Its basic principle is that for a training set containing n samples, multiple sampling sets are obtained by carrying out n multiple sampling from the initial sample, and each sampling sample is modeled based on the idea of decision tree, and several decision trees are combined to predict, and the prediction results are obtained based on the maximum probability idea. The process of filling missing data with random forest is as follows:

Assuming that $X = (X_1, X_2, ..., X_m)$ is an index data matrix of order $n \times m$, for X_h, record its missing dataset as $j_h^{mis} \subseteq \{1, 2, ..., n\}$. Then, the data matrix X can be divided into four parts: y_h^{obs} is the known observation value in the index X_h, y_h^{mis} is the corresponding missing value, and the data of the remaining $m-1$ index columns corresponding to the row of $j_h^{obs} \subseteq \{1, 2, ..., n\} \backslash j_h^{mis}$ is recorded as x_h^{obs}, the data corresponding to the row of j_h^{mis} is recorded as x_h^{mis}.

First, use simple interpolation methods such as mean interpolation or regression interpolation to initially fill X.

Second, the indicator set of missing data in X is recorded as ML, and the indicators are sorted according to the missing data rate from small to large.

Third, denote the existing data after filling as X_{old}^{imp}; when the stopping criterion γ is not satisfied, for $h \in M$, it classifies y_h^{obs} and x_h^{obs} based on the basic principles and methods of the random forest; based on the classification results, x_h^{mis} is used to predict y_h^{mis}, and the predicted value y_h^{mis} is used to update and fill the matrix, denoted as X_{new}^{imp}.

The stopping criterion γ means that the loop is stopped when the difference between the updated padding matrix and the pre-update padding matrix increases. Among them, the formula for calculating the degree of difference in filling the matrix is as follows:

$$\Delta_N = \left[\sum_{i=1}^{m} \sum_{j=1}^{n} \left(x_{ij}^{new} - x_{ij}^{old} \right)^2 \right] \Big/ \left[\sum_{j^{mis}} \sum_{i=1}^{m} \left(x_{ij}^{old} \right)^2 \right] \tag{3-13}$$

Fourth, the final padding matrix X^{imp} is obtained when the stopping criterion γ is satisfied.

(2) Expectation maximization estimation algorithm. The expectation maximization algorithm is based on the idea of maximum likelihood estimation and fills in the missing data in an iterative manner. Assuming that X is a data matrix, when there are random missing data, the maximum likelihood estimates of unknown parameters are obtained from the marginal distribution of known observed data. The EM algorithm

is an iterative operation that includes two processes of prediction (E) and estimation (M). Among them, step E refers to predicting missing data when the estimated values of unknown parameters are determined. Step M, based on the predicted missing data, calculates the parameter-corrected value of the maximum likelihood estimate.

The central idea of the expectation maximization estimation method is to regard the missing data appearing in the likelihood function $L(\theta|X)$ as a function of $(\theta|X_{obs})$, use the missing data to replace the conditional expectation and obtain the parameter estimates that satisfy the convergence through multiple iterations.

Each iteration is a two-step iteration. The first step uses the existing information of the data to find the expected value of the missing data, which is called E. The second step is to do the maximum likelihood estimation based on the replacement of the missing value, which is called M. This is done until convergence, and the expected value of the final missing data is used as its estimate.

The most intuitive application of the expectation-maximization estimation algorithm is K-Means dynamic clustering. Regarding the prediction E of the EM algorithm, due to the concealment of the centroid of each cluster, it can be assumed that there are K-initialized centroids. For the estimation M of the EM algorithm, the distance between each index sample and each centroid is calculated, and the index samples are assigned to the nearest centroid. Repeat the two steps E and M, when the centroid no longer changes, K-Means dynamic clustering is completed.

3.2.1.3 Multiple interpolation method –Markovian Monte Carlo method. MCMC method is a method for exploring the posterior distribution based on Bayesian theory. The MCMC method is based on the original dataset and the initial interpolation value obtained by the EM algorithm and uses the DA algorithm (data amplification algorithm) to perform two steps of data interpolation and posteriori. The basic idea and specific process of the MCMC method are as follows:

The first step is an interpolation. Let the mean vector of the indicators be $\mu = [\mu_{obs}, \mu_{mis}]$, the covariance matrix be $\Sigma = \begin{vmatrix} \Sigma_{obs} & \Sigma_{obs,mis} \\ \Sigma_{obs,mis} & \Sigma_{mis} \end{vmatrix}$ and the imputed values are drawn for missing values from the conditional distribution $p(X^{mis}|X^{obs}, \varphi)$.

The second step is the posterior. During each loop operation, the parameter φ is simulated with the mean vector and covariance matrix obtained from the previous interpolation.

Looping k times of interpolation and posterior can generate a Markov chain $[(X_1^{mis}, \varphi_1), (X_2^{obs}, \varphi_2), k]$ of sufficient length; when the Markov chain is clustered in a stable distribution $p(X^{mis}, \varphi|X^{obs})$, interpolation values can be extracted for missing values approximately independently.

3.2.2 Abnormal data handling. Abnormal data, also known as outliers, are data objects in a dataset that violate basic characteristics. Abnormal data usually includes two situations: one is that the abnormal data is obviously inconsistent with the overall data and the probability of occurrence is small; the other is that the abnormal data is regarded as an impurity point. Abnormal data may be caused by quality problems of the data itself, such as statistical errors, equipment failures resulting in abnormal results. However, it may also be a true reflection of the development and changes in phenomena, such as the global economic indicator data in 2020. Therefore, after detecting outliers, we need to judge whether they are outliers in line with the actual situation and characteristics of the dataset.

The processing methods of outliers can be summarized as direct deletions and as missing values. However, if the outliers are real data caused by special events, they should not be handled arbitrarily and should be used as singular points as dummy variable data when

constructing an econometric model. For example, data smoothing technology is to eliminate abnormal data by building a functional model; it includes the moving average method and exponential smoothing method.

(1) Moving average method. The moving average method refers to calculating a time-series average including a certain lag period at a time according to the gradual passage of data over time. Therefore, when the values of the time series have abnormal values due to random fluctuations, the moving average method is an applicable smoothing method. Also, the moving average method can be divided into two types: simple average and weighted average according to the difference in weight distribution in different periods.

- Simple moving average method. Since the weight of the data in each period of the i-th indicator is equal, that is:

$$x_{it}^{*} = (x_{i,t-k} + x_{i,t-k-1} + \ldots + x_{i,t-1} + x_{i,t+1} + x_{i,t+2} + \ldots + x_{i,t+s}) \big/ (k+s) \tag{3-14}$$

where x_{it}^{*} is the smoothed forecast and $k + s$ is the number of moving average periods.

- Weighted moving average method. Considering the difference in the degree of data effect in different periods, the data close to the outliers have a larger weight value. The weight design is best handled by the correlation coefficient between adjacent data. Specifically, the rolling window method can be used to decompose the original data sequence into multiple new sequences and then perform the correlation coefficient analysis. The weighted moving average method is as follows:

$$x_{it}^{*} = (\omega_{t-k} x_{i,t-k} + \omega_{t-k-1} x_{i,t-k-1} + \ldots + \omega_{t-1} x_{i,t-1} + \omega_{t+1} x_{i,t+1} + \omega_{t+2} x_{i,t+2} + \ldots + \omega_{t+s} x_{i,t+s}) \tag{3-15}$$

Among them, ω_i is the weight of the i-th period, $\sum_i \omega_i = 1$.

(2) Exponential smoothing method. Exponential smoothing method means that the exponential smoothing value of the current period is equal to the weighted average of the actual observation value of the current period and the exponential smoothing value of the previous period. The recurrence relation under the first index is as follows:

$$x_{it}^{*} = \alpha x_{it} + (1-\alpha) x_{it-1}^{*} \tag{3-16}$$

Among them, α is the smoothing coefficient, and when α approaches 0, it means that the actual value of the current period has little effect, while the exponential smoothing value of the previous period has a great influence, that is, the effect of the forward actual value on the current exponential smoothing value decreases slowly.

Generally, the appearance of outliers will cause significant changes in the data series. One method is empirical value, and the value range is usually 0.6–0.8. The other method requires scientific judgment. The first one is to design according to the change law of the data itself, such as the multi-order autocorrelation coefficient of the data. The second one is, on the basis of the autocorrelation coefficient, combined with the correlation between the variable data and other closely related variable data, a comprehensive design is carried out.

3.2.3 Missing data quality check. 3.2.3.1 Overall quality check of missing data itself. Missing data can pose considerable challenges to the analysis and interpretation of evaluation results and can undermine the validity and reliability of the results and conclusions. Subject to the risk of bias, different types of missing data have three types of missing data quality problems: non-random, random and completely random and different missing data quality inspection methods are needed. Among them, under the random missing mechanism, the parameters of the complete dataset and the missing dataset are independent and irrelevant, so the missing data caused by the random missing (including completely random missing) mechanism can be ignored (Table 8).

For the indicator Y, the dataset matrix is $Y = \{Y_{obs}, Y_{mis}\}$, where Y_{obs} is the observed dataset of the indicator Y, and Y_{mis} is the missing dataset; R is defined as the indicator variable of the indicator Y, when $R = 1$, it indicates that Y is not missing, and when $R = 0$, then Y is missing; given the observed value and missing value of the index Y, the probability of missing is $p(R|Y_{obs}, Y_{mis})$.

For the completely random missing mechanism (MCAR mechanism), there is $p(R|Y_{obs}, Y_{mis}) = p(R)$, that is, whether the observed data of the indicator Y is missing is not directly related to the datasets Y_{obs} and Y_{mis}. For the random missing mechanism (MAR mechanism), there is $p(R|Y_{obs}, Y_{mis}) = p(R|Y_{obs})$. That is, whether the observed data of the indicator Y is missing is associated with the dataset Y_{obs}. For the non-random missing mechanism (NMAR mechanism), there is $p(R|Y_{obs}, Y_{mis}) = p(R|Y_{mis})$ or $p(R|Y_{obs}, Y_{mis})$, that is, whether the observed data of the indicator Y is missing is correlated with the dataset Y_{obs} or Y_{mis}.

(1) Completely random missing mechanism test.

Under the completely random missing mechanism, the two datasets Y_{obs}, Y_{mis} of the indicator, Y have the same distribution as their own distribution $L(Y|R = 0) = L(Y|R = 1) = L(Y)$. In fact, due to the existence of missing data, the distribution of $L(Y|R = 0)$ and $L(Y)$ is unknown, so the test of the completely random missing mechanism (MCAR) is verified by introducing a substitute indicator X for indicator Y. The null and alternative hypotheses for this test are as follows:

H0. Missing data mechanism is MCAR

H1. Missing data mechanism is not MCAR

Assume that the data sequence of the indicator Y is $y = (y_1, y_2, \ldots, y_{n_1}, y_{n_1+1}, \ldots, y_n)$, where the observation sequence and the missing sequence are $y_{obs} = (y_1, y_2, \ldots, y_{n_1})$ and $y_{mis} = (y_{n_1+1}, \ldots, y_n)$, respectively. The data sequence of the substitute indicator X is$x = (x_1, x_2, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_n)$, correspondingly, $x_{obs} = (x_1, x_2, \ldots, x_{n_1})$, $x_{mis} = (x_{n_1+1}, \ldots, x_n)$; and $(y_1, x_1), \ldots, (y_n, x_n)$ has an independent and identically distributed relationship.

When the distribution of the surrogate index x is known and obeys the normal distribution, first test the homogeneity of variance by judging whether the mean and variance of the population of the surrogate index x are the same, and then perform the analysis of variance test.

Data missing mechanism	Test method
Missing Completely at Random (MCAR)	By judging the consistency of descriptive statistical features such as mean and variance
Missing at Random (MAR)	Logit model is used to describe the distribution of missing variables and test the significance
Not Missing at Random(NMAR)	Analyze the patterns and causes of missing data

Table 8. Data missing mechanism and its test method

Test statistic K^2 for homogeneity of variances:

$$K^2 = \frac{2.3026}{c}\left[(n-r)\ln S^2 - \sum_{i=1}^{r}(n_i-1)\ln S_i^2\right] \sim \chi^2(r-1) \tag{3-17}$$

where $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot})^2$, $S^2 = \frac{1}{n-r}\sum_{i=1}^{r}(n_i-1)S_i^2$, $c = 1+\frac{1}{3(r-1)}\left[\sum_{i=1}^{r}(n_i-1)^{-1} - (n-r)^{-1}\right]$, $n = n_1+n_2+\ldots+n_r$. For $K^2 < \chi_\alpha^2(r-1)$, it shows that the index dataset satisfies the homogeneity of variance in the statistical sense, and then it is verified by variance analysis whether the mean of the dataset is basically indifference. The F-test statistic for ANOVA is follows:

$$F = \frac{S_A/(r-1)}{S_E/(n-r)} \sim F_\alpha(r-1, n-r) \tag{3-18}$$

where $S_A = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\overline{x}_{i\cdot}-\overline{x})^2$, $S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot})^2$, $\overline{x}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij}$, $\overline{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}$.

When $F < F_\alpha(r-1, n-r)$, it indicates that there is no difference between the mean values of the predicted dataset and the missing data set, that is, the mean values are equal, so the null hypothesis is accepted, and the missing mechanism of the data set is considered to be a completely random missing mechanism (MCAR).

When the distribution of the surrogate index x is unknown or its distribution is skewed, the Kruskal–Wallis test is used to explore whether the overall distribution of the predicted data set and the missing data set is significantly different, then the K-W statistic is given by the following equation:

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i\left(\overline{R}_i - \overline{R}\right)^2 \sim \chi^2(k-1) \tag{3-19}$$

Among them, N and n_i represent the total capacity of the mixed sample and the capacity of the i-th group of samples, respectively, and $\overline{R}_i$ and $\overline{R}$ represent the average rank of the i-th group of samples and the overall average rank, respectively.

If $H < \chi_\alpha^2(k-1)$, it is considered that there is no difference in the distribution of the predicted data set and the missing dataset, that is, the data missing mechanism is a completely random missing mechanism (MCAR).

(2) Missing at random mechanism test.

For the indicator variable R of the indicator Y, the logit model is used to explore its distribution characteristics, and then the test of the random missing mechanism is equivalent to the coefficient significance test of the surrogate variable X, namely:

$$P(R|Y,X) = P(R|X) = \text{logit}^{-1}(f(X)) \tag{3-20}$$

If the coefficient of the surrogate variable X is significant, it means that there is a correlation between the missing data and the observed surrogate variable X, so the missing data mechanism is the random missing mechanism (MAR). On the contrary, if the coefficient of the surrogate variable X is not significant, the missing mechanism is non-missing at random mechanism (NMAR).

(3) Non-random missing mechanism test.

The non-random missing mechanism (NMAR) is the most common form of missing data, and various statistical processing methods will obtain biased estimates, so statistical tests cannot identify this missing mechanism. In general, NMAR mechanisms can be identified by discerning patterns and causes of missing data.

3.2.3.2 Overall quality inspection after missing data repair based on statistical features. For indicator Y, the original dataset matrix is $Y = \{Y_{obs}, Y_{mis}\}$. The patched dataset is $Y = \{Y_{obs}, Y_{add}\}$, where the patched dataset of missing dataset Y_{mis} is Y_{add}, then $Y_{add} = (y'_{n_1+1}, \ldots, y'_n)$. Based on the mean and sample variance of the two datasets Y_{obs} and Y_{add}, the independent sample t-test (Independent Sample t-Test) is used to test the mean difference between the two datasets. The null hypothesis of the t-test is H_0: the mean of the two datasets is not significantly different, and the alternative hypothesis is H_1: the means of the two datasets are significantly different.

The independent sample t-test statistic is as follows:

$$t = \frac{\overline{Y}_{obs} - \overline{Y}_{add}}{\sqrt{\frac{S^2_{Y_{obs}}}{n_1 - 1} + \frac{S^2_{Y_{add}}}{n_2 - 1}}} \sim t_\alpha(n-2), n_2 = n - n_1 \tag{3-21}$$

If $t > t_\alpha(n-2)$, the null hypothesis H_0 is accepted, indicating that the means of the two datasets are not significantly different; therefore, the missing data after repair is considered to be reliable in a statistical sense.

3.2.3.3 Overall quality inspection after missing data repair based on regression analysis. For the observation dataset Y_{obs} with a sample size of n_1, it is divided into two sub-datasets $Y^1_{obs} = [y_1, y_1, \ldots, y_{n_{11}}]$ and $Y^2_{obs} = [y_{n_{11}+1}, y_{n_{11}+2}, \ldots, y_{n_1}]$ on average. The observation data subset Y^2_{obs} and the missing data repaired set Y_{add} are combined into a regression experimental set, through the autoregressive model (AR model) inverts and estimates the observed data to predict the subset $\widehat{Y}^1_{obs}$, and use the t-test to measure the difference between the two datasets Y^1_{obs} and $\widehat{Y}^1_{obs}$. The specific steps are follows:

(1) Build an autoregressive model based on Y^2_{obs} and Y_{add}: $\widehat{y}_i = c + \sum_{j=n_{11}+1}^{n_1} \varphi_j y_{n-j}$;

(2) Use the autoregressive model of Step (1) to obtain the predicted subset of observation data $\widehat{Y}^1_{obs}$;

(3) Record the mean and variance of the two datasets Y^1_{obs} and $\widehat{Y}^1_{obs}$ as $\overline{Y}^1_{obs}$ and $\overline{\widehat{Y}}^1_{obs}$, $S^1_{Y_{obs}}$ and $S^1_{\widehat{Y}_{obs}}$, respectively, then the t-test statistic is as follows:

$$t_{re} = \frac{\overline{Y}^1_{obs} - \overline{\widehat{Y}}^1_{obs}}{\sqrt{\frac{S^1_{Y_{obs}} + S^1_{\widehat{Y}_{obs}}}{n_{11} - 1}}} \sim t_\alpha(2n_{11} - 2) \tag{3-22}$$

If $t_{re} > t_\alpha(n-2)$, the means of the two datasets Y^1_{obs} and $\widehat{Y}^1_{obs}$ are not significantly different; therefore, the missing data after patching are considered to be reliable in a statistical sense.

3.3 Index data transformation

If the comparability between the index data is weak or basically not comparable, it is easy to cause large differences in the classification and stratification of the indexes and the

comprehensive evaluation results. Therefore, the index data needs to be transformed. Generally, there are basic transformation, dimensionless, normalization, data fusion and other transformation methods.

3.3.1 Data basic transformation method. There are four basic data transformation methods: logarithmic transformation, square transformation, square root transformation and reciprocal transformation. Basic transformations can meet general requirements for data analysis (Table 9).

3.3.2 Data normalization transformation method. Data normalization is the process of standardizing data. Data normalization can transform the original value into a dimensionless value. Normalization will change the variance of the original data but will not change the relationship between the data. The variance of all standardized variables is equal, and the relationship information between the standardized variables is consistent with the original variables.

The function of data normalization is to effectively prevent unreasonable index weights due to differences in initial value ranges and enhance data comparability.

(1) Extreme value transformation method. The original data matrix and the transformed data matrix of the index are $X = \{x_{ij}\}$ and $Z = \{z_{ij}\}$, respectively. The maximum value and the minimum value of the absolute value of the j-th column in $X = \{x_{ij}\}$ are $|x_j|^{\max}$ and $|x_j|^{\min}$, respectively.

For the positive index, the extreme value transformation formula is as follows:

$$z_{ij} = \frac{x_{ij}}{|x_j|^{\max}} \tag{3-23}$$

For the negative index, there are two extreme value transformation formulas:

$$z_{ij} = 1 - \frac{x_{ij}}{|x_j|^{\max}} \tag{3-24}$$

$$z_{ij} = \frac{|x_j|^{\min}}{x_{ij}} \tag{3-25}$$

The data after extreme value transformation, its interval range is $[-1, 1]$.

(2) Standardized transformation method. Standardization transformation is also known as Z-score standardization. The mean and variance of the j-th index of the original data matrix $X = \{x_{ij}\}$ are $\overline{x}_j$ and σ_j^2 respectively, then the standardization formula is as follows:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j} \tag{3-26}$$

Transformation method	Logarithmic transformation	Square transformation	Reciprocal transformation	Square root transformation
Type of data	Equal ratio data or the ratio of each group of values to the mean is not much different	The variance is inversely proportional to the mean squared or left skewed	The variance is proportional to the mean squared and does not tend to or less than zero	Poisson distributed or slightly skewed, or the sample variance is positively related to the mean

Table 9.
Basic transformation method of index data

All z_{ij} have similar variances, and their actual maximum and minimum values depend on the j-th index.

(3) Interval transformation method. Let $[a,b]$ be the target interval for the value of z_{ij}, then:

$$z_{ij} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min}}(b-a) + a \tag{3-27}$$

In practical applications, the value is generally required to be between 0 and 1, that is, a is equal to 0 and b is equal to 1, and the interval transformation is the maximum-minimum transformation.

(4) Subsection transformation method. For neither the benefit index nor the cost index, subsection transformation can be performed. Assuming that the optimal attribute interval is $[x_j^0, x_j^*]$, x_j' is the lowest limit value, and x_j'' is the highest limit value; then the trapezoidal transformation formula is as follows:

$$z_{ij} = \begin{cases} 1 - \dfrac{x_j^0 - x_{ij}}{x_j^0 - x_{ij}'}, x_{ij}' \le x_{ij} \le x_j^0 \\ 1, x_j^0 \le x_{ij} \le x_j^* \\ 1 - \dfrac{x_{ij} - x_j^*}{x_j'' - x_j^*}, x_j^* \le x_{ij} \le x_j'' \\ 0, else \end{cases} \tag{3-28}$$

Specially, if the upper and lower limits of the optimal attribution interval of the index attribute are equal, the transformation function image degenerates into triangle.

(5) Decimal scaling normalization. Decimal scaling normalization refers to shifting the decimal point position of the index data X_i to the left or right. The moving digits and direction of the decimal point are related to the maximum absolute value of the index data X_i, namely:

$$z_{ij} = \frac{x_{ij}}{10^k}, j = 1, 2, ..., n \tag{3-29}$$

where k is the smallest integer that makes $Max(|z_{ij}|) < 1$.

For example, if the value range of X_i is $[-85, 21]$, then the maximum absolute value of the possible values of X_i is 85, then use $10^2 (k = 2)$ to divide each value, so that the value range of the index X_i becomes $[-0.85, 0.21]$.

3.3.3 Multi-source data fusion technology. Multi-source data fusion technology refers to a method that uses all the information to jointly reveal the characteristics of the evaluation object by comprehensively analyzing the data of different types of information sources or affiliation through physical models, parameter classification and cognitive models.

In index data transformation, multi-source data fusion technology focuses on the heterogeneous transformation and unified representation of multi-source heterogeneous data. In the big data structure, there are three types of data: structured, semi-structured and unstructured. However, the diversification of data sources will lead to problems such as data

heterogeneity and confusion and inconsistency in dimensions, which will greatly increase the difficulty of the unified representation of data. Therefore, by introducing the tensor space model of high-order dimension, a unified representation and analysis framework of index data based on the tensor model is designed, that is, low-order sub-tensors of structured, semi-structured and unstructured data structures are firstly constructed, respectively. Tensor expansion operator is used as a tool to expand data fusion of three low-order subtensor spaces, and then it is used to fuse the data of three kinds of structures into the tensor space of different order by using tensor expansion operator, so as to achieve effective fusion and unified representation of multi-source heterogeneous data.

According to the basic theoretical method of the tensor model, different transformation function models based on multi-source heterogeneous data of different dimensions are constructed. It takes time, space and user as the three basis spaces of tensor order, maps various features of heterogeneous data to each dimension of tensor basis space and expounds the sub-tensor representation methods and specific processes of unstructured, semi-structured and structured data based on video data, XML document data and database tables.

(1) Background knowledge of tensor model

Denote the P order tensor as $T \in R^{I_1 \times I_2 \times \ldots \times I_p}$, and the k modular expansion matrix of the tensor T is $T_{(k)} \in R^{I_k \times (I_{k+1} I_{k+2} \ldots I_k I_1 I_2 \ldots I_{k-1})}$, where element $e_{i_1 i_2 \ldots i_k i_{k+1} \ldots i_k}$ has i_k as the row coordinate and $(i_{k+1}-1)I_{k+2} \ldots I_k I_1 \ldots I_{k-1} + (i_{k+2}-1)I_{k+3} \ldots I_k I_1 \ldots I_{k-1} + (i_2-1)I_3 I_4 \ldots I_{k-1} + \ldots + i_{k-1}$ the column coordinate.

- The single-modular product operation of tensor T and matrix U means that tensor T is multiplied with matrix U along the p order to obtain a new tensor T^{new}, namely $T^{new} = (T \times_p U)_{i_1 i_2 \ldots i_{p-1} j_p i_{p+1} \ldots i_p} = \sum_{i_p=1}^{I_p} (e_{i_1 i_2 \ldots i_{p-1} i_p i_{p+1} \ldots i_p} \times u_{j_p i_p})$.
- The multi-modular product operation of the tensor T and the tensor Q means that the tensor T is multiplied by the tensor Q along some specified orders to obtain a new tensor $T^{new'}$, namely: $T^{new'} = (T \times_N Q)_{i_1 \ldots i_N k_{N+1} \ldots k_M} = \sum_{k_1 \ldots k_N} t_{i_1 \ldots i_N k_1 \ldots k_N} q_{k_1 \ldots k_N k_{N+1} \ldots k_M}$.

where $T \in R^{I_1 \times \ldots \times I_N \times K_1 \times \ldots \times K_N}$, $Q \in R^{K_1 \times \ldots \times K_N \times K_{N+1} \times \ldots \times K_M}$.

- The core tensor S and the approximate tensor $\widehat{T}$ are obtained by decomposing the high-order singular value of the P order tensor $T \in R^{I_1 \times I_2 \times \ldots \times I_p}$. The calculation process is as follows:

$$\begin{aligned} S &= T \times_1 U_1^T \times_2 U_2^T \ldots \times_p U_p^T \\ \widehat{T} &= S \times_1 U_1 \times_2 U_2 \ldots \times_p U_p \end{aligned} \tag{3-30}$$

where the matrix $U_1, \ldots, U_p$ corresponds to the expansion matrix of each module of the tensor, and the truncated left singular vector matrix obtained by singular value decomposition, that is, the truncated orthonormal basis space of the tensor T.

- For two third-order tensors $T^1 \in R^{I_t \times I_s \times I_u \times I_1}$ and $T^2 \in R^{I_t \times I_s \times I_u \times I_2}$, the tensor expansion operator f_{ex} is as follows:

$$f_{ex}: T^1 \overrightarrow{\times} T^2 \to T^E, T^E \in R^{I_t \times I_s \times I_u \times I_1 \times I_2} \tag{3-31}$$

- Tensor order and tensor dimension.

The tensor order refers to the order of multi-source heterogeneous data converted into tensors after encoding. It is fused into the base tensor space based on the tensor expansion operator, and then a high-order unified tensor model is established.

The tensor dimension means that when the new multi-source heterogeneous data is fused, if the feature order in the new data already exists in the original tensor space. The most fine-grained method is used to expand the dimension of the feature order.

(2) Multi-source heterogeneous data tensorization

- Sub-tensorization method of unstructured data.

The video in Apple-m4v format can be regarded as a fourth-order subtensor $T_{m4v} \in R^{I_f \times I_w \times I_h \times I_c}$, where I_f, I_w, I_h, I_c represents the four tensor spaces of video frame, image width, image height and color of Apple-m4v format video, respectively. For example, a 1080p video in Apple-m4v format with an image width and height of 768×576 and a color space including red, yellow and blue can be represented as a fourth-order tensor $T_{m4v} \in R^{1080 \times 768 \times 576 \times 3}$.

In order to uniformly represent the fourth-order sub-tensor as a third-order sub-tensor, the video can be converted from color to grey, and the color conversion formula is as follows:

$$Grey = 0.299red + 0.587green + 0.114blue, Grey \in [0, 255] \tag{3-32}$$

According to the color conversion formula, the fourth-order tensor $T_{m4v} \in R^{1080 \times 768 \times 576 \times 3}$ can be converted into a third-order tensor $T'_{m4v} \in R^{1080 \times 768 \times 576}$.

- Sub-tensorization method of semi-structured data.

The Extensible Markup Language is an XML document with a hierarchical tree structure, which can be represented as a third-order sub-tensor T_{XML}, namely, $T_{XML} \in R^{I_{er} \times I_{ec} \times I_{en}}$, where the rows, columns and ASCII codes of the XML document identification matrix are I_{ec}, I_{er}, I_{en}, respectively.

Figure 2(a-d) shows the specific process of converting from Extensible Markup Language (XML document) to a third-order tensor model. Figure 2(a-c) represent the Extensible Markup Language and its parse tree and simple dendrogram, respectively.

The data elements of the 12 rectangular boxes in Figure 2(b) correspond to the 12 circles in Figure 2(c), and the length of element *Marine Economics and Management* is 28, so the third-order tensor with I_{er}、I_{ec}、I_{en} as the tensor order is $T_{XML} \in R^{12 \times 12 \times 28}$.

- Sub-tensorization method of structured data.

Structured data is generally stored in relational databases. Generally, for database tables containing simple types, a field can be represented by numbers or characters, and then a matrix can be used to represent the database table. However, for fields of complex types, it is necessary to add a new tensor order to represent.

The two database tables in Figure 3 are used to represent the student's trajectory information and personal information and contain four fields and two fields respectively. According to the method of structured data tensorization, it can be obtained a fourth-order tensor $I_{TAB-1} \in R^{I_t \times I_x \times I_y \times I_{id}}$ and a second-order tensor $I_{TAB-2} \in R^{I_{id} \times I_{name}}$. Based on the tensor expansion operator f, a fifth-order tensor is obtained by fusion, namely:

$$I_{TAB-3} = R^{I_t \times I_x \times I_y \times I_{id}} \overrightarrow{\times} R^{I_{id} \times I_{name}} \tag{3-33}$$

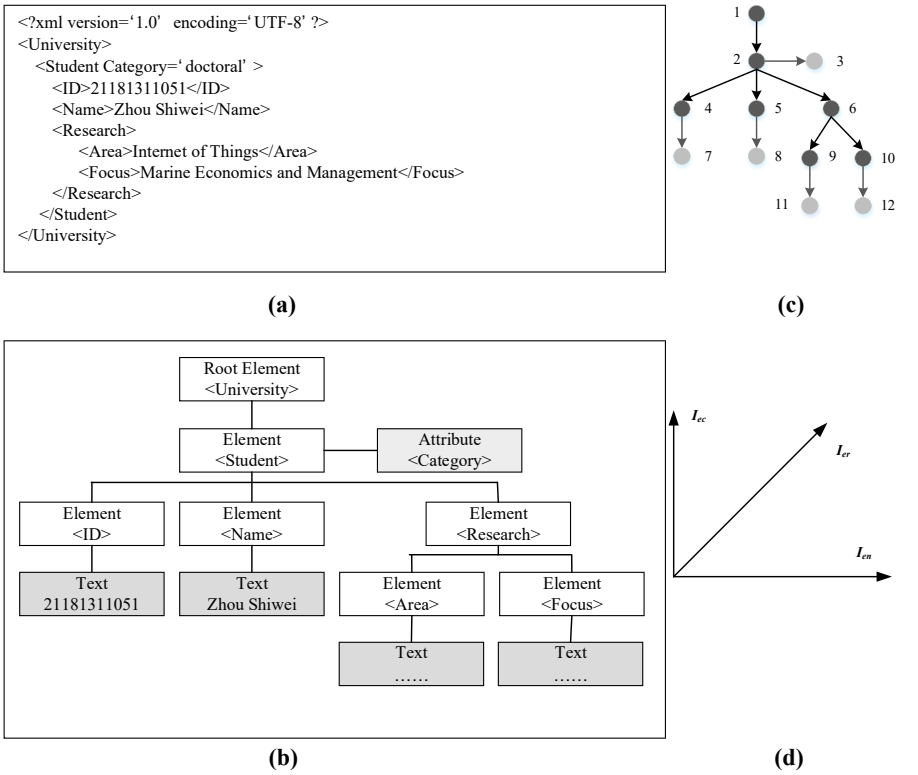

Figure 2.
The process of converting an Extensible Markup Language XML document into a third-order tensor

(3) Low-order tensor space fusion method

In order to uniformly represent the multi-source heterogeneous data of the low-order sub-tensor space, the low-order tensor space is fused into a high-order tensor space through the tensor fusion operator f_{co}, namely:

$$f_{co}: (d_u \cup d_{semi} \cup d_s) \rightarrow T_u \cup T_{semi} \cup T_s \tag{3-34}$$

where d_u, d_{semi}, d_s represent unstructured data, semi-structured data and structured data, respectively.

3.4 Index data test

Index data test is to verify the stability, coordination and consistency of data by using simple chart analysis, unit root test and correlation number test. The purpose of data test is to evaluate and judge whether the index data meets the needs of analysis and ensure that wrong and biased data are eliminated. Graphic judgment and unit root test are usually used to verify the stability and coordination of data, and relevant logic is used to verify the consistency of data.

3.4.1 Graphic judgment method. Typically, trend curve is used to test the stationarity of time series, mainly including a time series diagram and autocorrelation function diagram.

Record	StudentID	Longitude	Latitude	Time
1	D21181311038	123.4889122	52.72339765	10-15 22:13:16
2	D21181311051	123.4887399	52.72336882	10-15 22:13:26
3	D21181311056	123.4879912	52.72338776	10-15 22:13:36
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Record	StudentID	StudentName
1	D21181311038	Zhang Qi
.	.	.
.	.	.
.	.	.

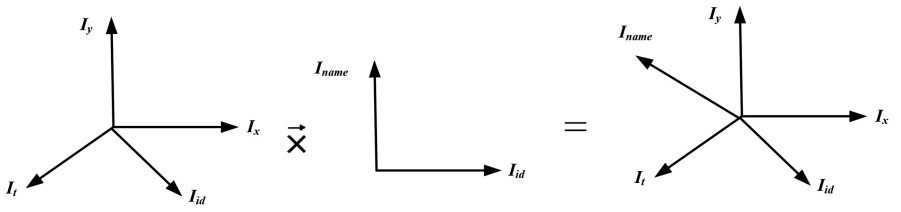

Figure 3.
The tensorized representation process of structured database tables

(1) Time series diagram. Time series diagram refers to the scatter diagram of time series with time as the horizontal axis and index value as the vertical axis. The judgment criterion is: if the time series shows a continuous fluctuation process around a constant mean value, it is a stable time series; on the contrary, it is a non-stationary sequence.

(2) Autocorrelation function diagram. Autocorrelation function is used to measure the degree of correlation between the values of index data at different times. Generally, the autocorrelation function of a stationary sequence will quickly converge to 0.

In practice, the sample autocorrelation function γ_k is used to measure the sequence autocorrelation coefficient, which is defined as follows:

$$\gamma_k = \frac{\sum_{t=1}^{n-k} (X_t - \overline{X})(X_{t+k} - \overline{X})}{\sum_{t=1}^{n} (X_t - \overline{X})}, k = 1, 2, \ldots, n \tag{3-35}$$

The significance test of sample autocorrelation function includes the following two ways:

- Significance test of whether the sample autocorrelation function γ_k is close to 0 enough.

Bartlett proved that for a time series generated by white noise process, for all $k > 0$, the sample autocorrelation coefficient approximately obeys the distribution $N(0, \frac{1}{n})$, and n is the number of samples.

- Test the original hypothesis H_0 of stationarity, that is, for all $k > 0$, the autocorrelation coefficient is 0. The test is conducted by constructing Q_{LB} statistic:

$$Q_{LB} = n(n+2)\sum_{k=1}^{m}\left(\frac{r_k^2}{n-k}\right) \sim \chi^2(m) \tag{3-36}$$

where m is lag length. At the significance level of $\alpha = 0.01$, if $Q_{LB} > \chi^2_{0.01}(m)$, we reject the original hypothesis, indicating that it is a non-stationary series.

3.4.2 Unit root test. For time series such as macroeconomic data and monetary and financial data, unit root test is an effective method to test time series stationarity. Generally, the existence of unit roots in time series is expressed in the continuous volatility of the series, and the unit root can be eliminated by the difference method.

Augmented Dickey-Fuller (ADF) unit root test is the most basic and general unit root test method, whose basic idea is to judge the stability of the sequence according to the absolute value of the characteristic root by constructing the autoregressive model with lag operator. The specific process is as follows.

Original hypothesis of hypothesis test H_0: $\delta = 0$, the time series contains a unit root.

ADF unit root test is done through the following three models:

$$Model\,1: \Delta X_t = \delta X_{t-1} + \sum_{i=1}^{m}\beta_i \Delta X_{t-i} + \varepsilon_t$$

$$Model\,2: \Delta X_t = \alpha + \delta X_{t-1} + \sum_{i=1}^{m}\beta_i \Delta X_{t-i} + \varepsilon_t$$

$$Model\,3: \Delta X_t = \alpha + \beta t + \delta X_{t-1} + \sum_{i=1}^{m}\beta_i \Delta X_{t-i} + \varepsilon_t$$

In Model 3, the time variable t represents the changing trend of the series over time, and constant terms and trend terms differentiate Model 1 from other two models.

In practice, ADF unit root test starts with Model 3, and then tests Model 2 and Model 1 in turn. According to the critical value table of ADF unit root distribution, when the original hypothesis H_0 is rejected, it indicates that the time series is a stationary series. However, when the test results of all three models accept the null hypothesis, the time series is considered to be non-stationary.

Most non-stationary time series can generally become stationary through one or more differences.

(1) If the test results show that there is a unit root in the time series and the parameter before the time variable is significantly 0, the series has the randomness of fluctuation trend. The randomness trend can be eliminated by using the difference method.

(2) If the test results show that there is no unit root in the time series and the parameter of the time variable is significantly not 0, it indicates that there is a deterministic trend in the series, and deterministic trends can only be eliminated by removing trend items.

3.4.3 Correlation logic test. Correlation logic test is a rough test of the reliability of index data based on the logical relationship between indicators. Its logical relationship is generally a highly correlated relationship between indicators determined by closely related objective phenomena. There are two main forms: one is the stable proportional relationship of the total amount index itself. Second, there is a certain degree of positive and negative consistency between changing trends of aggregate indicators and relevant indicators, such as the relationship between the inflation rate and unemployment rate. The correlation logic test can be verified by calculating the

correlation coefficient, and Pearson correlation coefficient is statistically applicable to datasets with normal distribution characteristics, while Spearman correlation coefficient is applicable to datasets with skew distribution characteristics (Table 10).

(1) Pearson correlation coefficient. Because it is difficult to obtain the overall correlation coefficient ρ, a specific sample correlation coefficient r can be used to replace it.

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\sum x_i \cdot y_i}{\left[\sum x_{i^2} \cdot \sum y_{i^2}\right]^{\frac{1}{2}}} \tag{3-37}$$

(2) Correlation coefficient significance test. In order to judge whether the correlation coefficient is statistically significant, that is, whether there is a significant linear correlation between series, the corresponding significance test, referred to as the correlation test, must be carried out. Relevant inspection steps are as follows.

Firstly, the sample correlation coefficient r is calculated;

Secondly, according to the sample size n and significance level α ($\alpha = 0.01$), check the correlation coefficient table to obtain the critical value r_α (with a degree of freedom $n-2$).

Finally, test and judge: when $|r| > r_\alpha$, there is a significant linear correlation between X and Y, otherwise, the linear correlation is not significant.

(3) Spearman correlation coefficient. Spearman correlation coefficient is also known as Spearman rank correlation coefficient. "Rank" refers to order or ranking. It uses the rank of two indicators for linear correlation analysis and does not require the distribution of original variables. The correlation coefficient is nonparametric.

Assuming that the sample size of both two indicators X and Y is n. X_i and Y_i, respectively, represent the ith values of the two random variables. Arrange X and Y in ascending or descending order to get an ordered set of two elements, say X' and Y'. Among them, the rank of X_i in X and the rank of Y_i in Y are represented by x_i and y_i, respectively. The elements in the set of x and y are correspondingly subtracted to obtain a rank difference set d, where $d_i = x_i - y_i$. Spearman correlation coefficient is expressed as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2 - 1\right)} \tag{3-38}$$

Spearman correlation coefficient test steps is listed as follows:

- Calculate Spearman correlation coefficient ρ;
- According to the sample size n and significance level α ($\alpha = 0.01$), check the correlation coefficient table to obtain the critical value r_s (with a degree of freedom $n-2$);
- Test and judgment: when $|\rho| > r_s$, there is a significant linear correlation between X and Y, otherwise, the correlation is not significant

Table 10. Comparison of application range of two correlation coefficients

Correlation coefficients	Pearson correlation coefficient	Spearman correlation coefficient
Application range	Continuous data, normal distribution, linear relationship	No requirement for the distribution of raw data

4. Index classification and stratification

The index system is a high generalization and systematic abstraction of the element structure of the studied object, and the process of index classification and stratification is an important link in building a systematic, complete and feasible index architecture. Index system structure is a hierarchical order in which indexes are interrelated and affect each other, the characteristics of which include stability, hierarchy, openness, relativity, etc.

In terms of hierarchical structure decomposition, the target layer is decomposed into primary index, secondary index, tertiary index, terminal index and so on through hierarchical structure relationship. According to different specific evaluation tasks, the target layer is divided into several control layers, criterion layers and scheme layers through a hierarchical structure relationship. The indicator data information between levels is only transmitted vertically between the related upper- and lower-level indicators, and the upper information has priority over the lower information. The indicators at the upper level are a comprehensive summary of the indicators at the lower level and different indicators at the same level are independent and incompatible with each other. According to the principles of feasibility and scientificity of index design, the classification and stratification of index system and the number of end indicators should not be too much. Therefore, index classification and stratification have an important impact on the quality of index system.

At present, the methods used for index classification and stratification mainly include classification decision tree, interpretation structure model, naive Bayesian classification, K-means dynamic clustering and binary K-means clustering, etc.

4.1 Classification decision tree

As an important classification method in data mining technology, decision tree originated from Concept Learning System (CLS). The ID3 algorithm uses the degree of information uncertainty reduction as the criterion for the selection of index attributes and is a relatively representative decision tree algorithm (Yang *et al.*, 2018). When constructing a decision tree, the classification and regression tree CART algorithm keeps part or all of the training set in staying in memory, continuous and discrete metric data can be handled well (Lin and Luo, 2019). Based on the information gain selection properties, the C4.5 algorithm can process continuous data and overcome some shortcomings of the ID3 algorithm. The improved C4.5 algorithm no longer uses the binary search method instead of the traditional linear search method, making the search more thorough and the target easier to search (Ruggieri, 2002). There may be variables with the same attributes but different classifications in the training set of decision tree, and the new method of minimum distance classification and attribute reduction can effectively improve the classification accuracy. Hosseini *et al.* (2021) constructed a classification decision tree model based on the GA algorithm, which improved the classification performance of the model for single-output and multi-output datasets (Hosseini *et al.*, 2021). The classification decision tree has high classification accuracy, simple operation, good robustness to noisy data (data with errors or anomalies) and can also deal with missing sample attribute values, so it has become a practical and popular classification algorithm, while ID3 algorithm, C4.5 algorithm and CART algorithm are the most representative decision tree algorithms.

Generally, a disadvantage of machine learning algorithms is that they are difficult to interpret; however, CART is clear and easy to interpret, and CART is also a powerful tool for building expert systems. In the investment field, CART is often used to enhance the identification of financial reporting fraud, provide equity and fixed income products selection support and simplify the presentation of investment strategies to clients (Table 11).

4.1.1 CART classification regression tree. The supervised learning algorithm uses the labeled data (known output results) to train the model, establishes the mapping relationship between the input variable X (feature) and the output variable Y (target/label) and uses the trained model to predict label for new data.

Since the CART algorithm divides the feature space into binary, the essence of the decision tree is a binary tree that can be used for both classification and regression. The classification decision tree and regression decision tree correspond to the discrete data and continuous data of the prediction result, respectively. It performs binary segmentation on each feature by recursively and then predicts the attribute type of the input sample according to the input eigenvalues. The CART algorithm is not only capable of dealing with irrelevant features, but also produces feasible and well-performing results on large datasets in a relatively short period of time.

If the predicted sample falls to a certain leaf node, the classification decision tree of discrete data will output the most attributes of all sample attribute categories, while the regression decision tree of continuous data will output the mean value of all samples in the leaf node.

In a decision tree, the root node and each decision node represent a combination of attribute features and their cutoff values. The selection criteria for attribute features and cutoff values at each node is to minimize the classification error (e.g. mean squared error). Splitting process ends at a node when further splitting fails to significantly improve the grouping error within the dataset, and this node is called the terminal node. The same feature can appear multiple times in node classification, such as root nodes and decision nodes, but there should be corresponding combinations of different cutoff values.

4.1.2 CART classification tree-discrete data splitting. For a given sample dataset $X = (x_1, x_2, ..., x_m)$ whose feature set is $T = (t_1, t_2, ..., t_k)$, that is, there are k attribute categories. Its Gini coefficient can be expressed as follows:

$$Gini(X) = 1 - \sum_{k} p_k^2 \tag{4-1}$$

In the above formula, p_k refers to the frequency of occurrence of kth feature (or attribute) category in the dataset X. Gini coefficient refers to the degree of uncertainty that the sample in X belong to a certain feature (or attribute), and the smaller the Gini coefficient is, the lower uncertainty of the attribute category of the sample is.

For a sample set X containing n samples, according to the attribute feature t_i, the sample data set X is divided into two subsamples X_1 and X_2, then the Gini coefficient is expressed as follows:

$$Gini_{t_i}(X) = \frac{N(X_1)}{N(X)} Gini(X_1) + \frac{N(X_2)}{N(X)} Gini(X_2) \tag{4-2}$$

Algorithm	Support model	Tree structure	Feature selection	Continuous value processing	Missing value handling	Pruning
ID3	Classification	Polytree	Information gain	Not support	Not support	Not support
C4.5	Classification	Polytree	Information gain ratio	Support	Support	Support
CART	Classification, regression	Binary tree	Gini coefficient, mean square error	Support	Support	Support

Table 11. Comparison of three typical decision tree algorithms

where $N(X)$, $N(X_1)$ and $N(X_2)$ represent the sample size of dataset X, X_1 and X_2, respectively.

For the sample dataset X, calculate the Gini coefficient after dividing the sample dataset X into two subsamples under each attribute feature. The optimal dichotomy scheme is to divide the subsample corresponding to the minimum value of the Gini coefficient.

For the sample dataset X, according to all attribute categories of the feature set and their optimal bisection scheme, select the subsample with minimum Gini coefficient for segmentation as the optimal bisection scheme of the sample dataset X, and the attribute category is the optimal splitting attribute and optimal splitting attribute value of the sample dataset.

4.1.3 CART regression tree – continuous data splitting. The index to evaluate the splitting attribute of the regression decision tree is the minimum variance σ: the smaller the variance σ is, the less difference among subsamples are, indicating that the effect of selecting this attribute as the evaluation splitting attribute is more significant.

If the sample set X contains continuous prediction results, the formula for calculating the total variance is $\sigma(X) = \left[\sum(y_k - \mu)^2\right]^{\frac{1}{2}}$. In the formula, μ represents the mean of the prediction results in the sample set X, and y_k represents the prediction result of the kth sample.

For the sample set X containing n samples, according to the *ith* attribute value of the attribute T, the dataset X is divided into two sub-sample sets X_1 and X_2, then the divided variance σ is $Gain\_\sigma(X) = \sigma(X_1) + \sigma(X_2)$.

For attribute T, calculate the variance σ after dividing the sample data set into two different sub-sample sets X_1 and X_2 under each attribute type i separately, and the optimal bisection scheme of attribute T is the sub-sample corresponding to the smallest variance σ value.

The optimal bisection scheme of the sample set X is the minimum value of the variance corresponding to the optimal bisection scheme of all attributes T.

4.1.4 Pruning of CART classification and regression trees. The pruning of the CART classification and regression tree refers to the process of simplifying the complexity of the model by limiting the maximum height and depth of the tree and limiting the total number of nodes in order to prevent overfitting. Overfitting means that if the sample dataset is infinitely divided, the decision tree can achieve a complete classification of the sample dataset. However, for test samples, such decision trees have the characteristics of large size and complex structure, which in turn result in a high classification error rate.

Decision tree pruning is the process of testing, correcting and revising the classification tree (or regression tree) generated in the previous stage, that is, using the data in the test dataset to check whether the generation process and production results of the decision tree are optimal, and pruning branches with large prediction errors. In general, the pruning algorithm for CART trees depends on the complexity parameter α (α decreases as the complexity of the decision tree increases) and if the change in the accuracy of the classification result is less than α times the change in the complexity of the decision tree, prune the node. The appropriate α is usually determined through the cross-certification method, and then the optimal decision tree is obtained.

There are two types of pruning methods: pre-pruning and post-pruning. Pre-pruning is to terminate the growth of the decision tree in advance in the process of building the decision tree, so as to avoid generating too many nodes. Although the pre-pruning method is simple, it is not practical because it is difficult to accurately judge when to terminate the growth of the tree.

Post-pruning is to replace the sub-node tree whose confidence is not up to standard with a leaf node based on the constructed classification decision tree (or regression decision tree), and the class label of the leaf node is marked as the class with the highest frequency in the

sub-node tree. Common post-pruning methods include REP (Reduced Error Pruning), PEP (Pessimistic Error Pruning) and CCP (Cost Complexity Pruning) (Table 12).

For the CCP cost complexity pruning method, the main process is to generate a sub-node tree sequence $\{T_0, T_1, \ldots, T_n\}$ based on the original decision tree T_0, and then determine the optimal decision tree based on the true error estimate of the decision tree.

(1) Determine the parameter set $\{T_0, T_1, \ldots, T_n\}$

The basic idea of generating a sub-node tree sequence $\{T_0, T_1, \ldots, T_n\}$ is that starting from T_0. It is obtained by trimming the branch in the decision tree A that minimizes the increase in the error of the training sample dataset at time T_t. Considering the complexity of the decision tree, the error increase rate α of the pruned classification decision tree (regression decision tree) branch is as follows:

$$\alpha = \frac{R(t) - R(T_t)}{|L(T_t) - 1|} \tag{4-3}$$

In the above formula, $R(t)$ and $R(T_t)$ are the errors of the child node tree of node t after being pruned and before being pruned. $R(t) = r(t) * p(t)$, and $r(t)$ represents the error rate of node t, and $p(t)$ represents the ratio of the number of samples of node t to the number of samples in the training set. The number of leaves of the decision tree is reduced to $L(T_t) - 1$.

For example, in Figure 4, A and B represents two attribute categories in the training set, and there are corresponding demarcation values of falling nodes belonging to Category A and Category B.

For node t_4, there are 46 samples of class A and 4 samples of class B, respectively, then the samples of node t_4 belong to class A. The error of subtrees (t_8 and t_9) of node t_4 before pruning is $R(T_{t_4})$: $\frac{1}{45} * \frac{45}{80} + \frac{2}{5} * \frac{5}{80} = \frac{3}{80}$. The error $R(t_4)$ after clipping the two subtrees (t_8 and t_9) of node t_4 is $\frac{4}{50} * \frac{50}{80} = \frac{4}{80}$, so the error increase rate is $\alpha(t_4) = \frac{\frac{4}{80} - \frac{3}{80}}{2-1} = \frac{1}{80}$.

	REP	PEP	CCP
Pruning method	Bottom-up	Top-down	Bottom-up
Error estimation	Error estimation on pruning set	Continuous correction	Standard error

Table 12. Comparison of three typical post-pruning methods

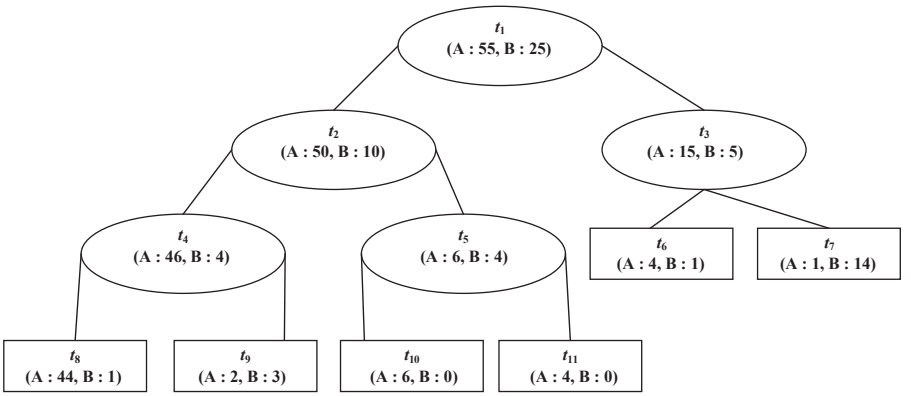

Figure 4. Example of a classification decision tree for two attribute categories

The error increase rates of all nodes are sequentially obtained by cyclically performing the above process, as shown in Table 13. Since in line T_0 of the original tree, $\alpha(t_4)$ has the smallest error increase rate among all non-leaf nodes, the branch of node t_4 of T_0 is cut to obtain T_1; in line T_1, although the error increase rates of t_2 and t_3 are the same, but since t_2 has two-level branches, a smaller decision tree can be obtained by pruning the branches of t_2.

(2) Determine the best tree T_i

Based on the sub-node tree sequence $\{T_0, T_1, \ldots, T_n\}$ generated in (1), two methods of cross-validation and independent pruning datasets are used to measure the error rate and use this as the criterion for selecting the best tree.

- Optimal tree selection based on cross-validation.

Divide the training set X into V subsets $R^1, R^2, \ldots, R^V$ and build the decision tree $T^1, T^2, \ldots, T^V$ based on the $R-R^1, R-R^2, \ldots, R-R^V$, correspondingly there are V different parameter families $T^1(\sigma), T^2(\sigma), \ldots, T^V(\sigma)$. For each cross-validation training, calculate the geometric mean $\sigma_{i^{av}}$ of the atomic node tree sequence σ_i and σ_{i+1}, and the best decision tree selected has the characteristic of being smaller than the maximum σ_j value in the geometric mean $\sigma_{i^{av}}$.

- Best tree selection based on independent pruning dataset method.

The independent pruning dataset method means that each decision tree T^i is classified according to the pruning set; in the sub-node tree sequence, E' and N' represent the minimum number of misclassifications of any decision tree on the pruning set and the number of misclassifications in the pruning set, respectively. The final pruning tree is the decision tree corresponding to the number of misclassifications of the pruning set is less than (or equal to) the minimum value of $E' + SE(E')$.

The formula for standard error $SE(E')$ is:

$$SE(E') = \sqrt{\frac{E' \times (N' - E')}{N'}} \tag{4-4}$$

4.2 Interpretive structure model

In 1973, Professor Warfelt proposed the interpretation structure model (ISM). The ISM method mainly uses people's practical experience and professional knowledge to construct the system and its elements into a multi-level hierarchical structural model, so as to transform the ambiguous problems or elements into an intuitive model with good structural relationships and has strong practicability and pertinence in index classification and hierarchical design. In the ISM method, there are only two element relationships in the relationship matrix, 0 and 1, and introducing the concept of the fuzzy matrix in fuzzy mathematics can improve the accuracy of element relationships (Zou *et al.*, 2018). In 2019, Lin *et al.* combined gray association analysis (GRA) with ISM, first used GRA to determine the relationship between variables, and then used ISM for classification and stratification (Lin *et al.*, 2019).

Table 13. Error increase rate for all nodes

T_0	$\alpha(t_4) = 0.0125$	$\alpha(t_5) = 0.0500$	$\alpha(t_2) = 0.0250$	$\alpha(t_3) = 0.0375$
T_1	$\alpha(t_5) = 0.0500$	$\alpha(t_2) = 0.0375$	$\alpha(t_3) = 0.0375$	
T_2	$\alpha(t_3) = 0.0375$			

The construction process of the explanatory structural model is as follows:

(1) Establish a system element relationship table. According to the existing relevant theories and actual situations, determine the elements (indicators) in the system and relationships among elements, which are described through tables.

(2) According to the element relationship table in (1), make a directed relationship graph description, establish an adjacency matrix $A_{m\times m}$, and m is the number of indicators. Matrix element a_{ij} is 1 if index X_i has influence on index X_j otherwise a_{ij} is 0, that is:

$$a_{ij} = \begin{cases} 1, \text{when } X_i \text{ has an influence on } X_j; \\ 0, \text{when } X_i \text{ has no influence on } X_j. \end{cases} \tag{4-5}$$

The judgment of whether the index X_i has an influence on the index X_j can be determined either by the objective method of the correlation coefficient or by the subjective method of expert judgment. However, the relationship between X_i and X_j may not be a complete binary relationship, or it may be between 0 and 1, so the results of subjective and objective judgments may be inconsistent.

The Kappa coefficient can test the consistency of the two methods and clarify whether the index X_i has an influence on the index X_i.

As an index to measure the classification accuracy, the Kappa coefficient can evaluate the consistency and reliability of the classification results or two sets of data, e.g. to compare whether the evaluation levels of two experts are consistent, etc. The formula for calculating the Kappa coefficient is as follows:

$$kappa = \frac{p_0 - p_e}{1 - p_e} \tag{4-6}$$

Among them, p_0 is the consistency probability of the two judgment results, and p_e is the random consistency probability of the two judgment results. In actual research, the test standard of Kappa coefficient is shown in Table 14.

For example, two methods of correlation coefficient and expert judgment are used to judge the mutual influence relationship of each index, and the obtained results are shown in Table 15. Then $p_0 = \frac{a+d}{n^2}$, $p_e = \frac{(a+c)\times(a+b)+(b+d)\times(c+d)}{n^2\times n^2}$.

(1) The reachable matrix M can be obtained by Boolean operation. If a matrix only has 1 on its diagonal element and all other elements are 0, such a matrix is called an identity matrix and is represented by the letter I. According to the Boolean matrix algorithm, it can be proved that:

Kappa coefficient	0.00~0.02	0.02~0.20	0.20~0.40	0.40~0.60	0.60~0.80	0.80~1.00
Degree of consistency	Very low	low	Fair	Moderate	High	Very high

Table 14. Kappa coefficient test standard

Feature relationship value		Correlation coefficient method 1	0
Expert judgement	1	a	B
	0	c	D

Note(s): $a+b+c+d=n^2$

Table 15. Example of Kappa coefficient

$$(A+I)^2 = I + A + A^2$$

The same can be proved:

$$(A+I)^k = I + A + A^2 + \ldots + A^k$$

If the adjacency matrix satisfies the condition:

$$(A+I)^{k-1} \neq (A+I)^k = (A+I)^{k+1} = M$$

Then M is called the reachability matrix of the adjacency matrix A; the reachability matrix indicates whether there is a connecting path from one element to another.

(2) The division of levels and relationships

Define reachability set $R(X_i)$: In reachability matrix M, the index set whose row corresponds to index X_i contains 1 in the column corresponding to the matrix element is the reachable index set.

Define antecedent set $Q(X_i)$: In the reachability matrix M, the index set of the row corresponding to the matrix element whose column corresponds to index X_i contains 1 is the leading index set.

Define common set $C(X_i)$: $C(X_i) = R(X_i) \cap Q(X_i)$.

If X_i is the top-level indicator, $R(X_i) \cap Q(X_i) = R(X_i)$ must be satisfied, so the first-level indicator set $L = \{X_i\}$ is obtained. When the first-level index set is obtained, cross out the row and column where the index is located in the reachability matrix, and the second-level and third-level index sets can be obtained by the same method, and so on to get the entire hierarchy substructure.

4.3 Naive Bayesian Classification

Naive Bayesian Classification (NBC) originates from the Bayesian theory proposed by British scholar R.T. Bayes in 1763. The tree extended naive Bayesian classifier (TAN) represents the relationship between all variables as a tree structure based on the interdependence between variables (Friedman *et al.*, 1997). In addition, the performance of the Weighted Naive Bayesian classification model (WNB) is significantly improved. With the increase in data volume, a single classifier is difficult to meet the needs of research, so many scholars combine NBC with other methods (Kouziokas, 2020). Ding and Wang (2019) proposed a weighted naive Bayesian classification algorithm with Jensen-Shannon divergence (Ding and Wang, 2019).

Naive Bayesian classification algorithm is a probabilistic classification algorithm based on the Bayesian classification algorithm. It is a group of very simple and fast classification algorithms. Compared with the original algorithm, naive Bayes classification algorithm is efficient and has a wide range of applications. It is usually suitable for data sets with very high dimensions, fast speed and few adjustable parameters. It is very suitable to provide a fast and rough basic scheme for classification problems. For a given item index to be classified, naive Bayesian classification assumes that each index has an independent relationship and solves the probability of other remaining indexes when this index appears. The category corresponding to the maximum probability is the category of the item to be classified. Naive Bayes classification steps are as follows:

The first step is to divide each index appropriately according to the comprehensive evaluation object and form a set of training samples by subjectively classifying the indexes. The index to be classified is the input content, and the characteristic attributes and training samples are the output content.

In the second step, the classifier is constructed by calculating the conditional probability estimation of each feature attribute category for each index.

(1) Let D be the set of all indicators, and each indicator X_i is represented by an n-dimensional vector $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$;

(2) Suppose there are k attribute classes $C_1, C_2, \ldots, C_k$ and given the index vector X_i, it belongs to the attribute class of the maximum *a posteriori* probability. That is, the prediction index vector X_i of naive Bayesian classification belongs to the attribute class C_p if and only if:

$$P(C_p|X_i) > P(C_q|X_i),\ 1 \le p, q \le k, p \ne q \tag{4-7}$$

where $P(C_p|X_i) = \frac{P(X_i|C_p)P(C_p)}{P(X_i)}$; at this time, the maximum $P(C_p|X_i)$ is called the maximum *a posteriori* probability.

(3) Since $P(X_i)$ is constant for all attribute classes, only $P(X_i|C_p)P(C_p)$ is required to be the largest. If the prior probability of attribute classes is unknown, it is usually assumed that these attribute classes are equal probability, that is, $P(C_1) = P(C_2) = \ldots = P(C_k) = \frac{1}{k}$, and $P(C_p|X_i)$ is maximized accordingly, otherwise, $P(X_i|C_p)P(C_p)$ is maximized.

(4) X_i is divided by attribute categories. If $P(C_p|X_i) = Max\{P(C_1|X_i),\ P(C_2|X_i), \ldots, P(C_k|X_i)\}$ is met, there will be $X_i \in C_p$.

In the third step, the classifier is used to classify the indicators. The input is the classifier and the indicators to be classified, and the output is the corresponding relationship between the indicators to be classified and the categories.

4.4 Cluster analytic hierarchy process

Cluster analysis is a multivariate statistical method to classify indicators or samples. This method is to classify scientifically and reasonably according to the characteristics of the research object without prior knowledge. Cluster analysis includes the system clustering method and K-means clustering method. The systematic clustering method, also known as the hierarchical clustering method, initially takes each index or sample as a class and aggregates the closest (least distant) index or sample into small classes, then merges the aggregated small classes according to the distance between classes, continues and finally aggregates all small classes into a large class. K-means clustering method is an iterative clustering analysis algorithm, which randomly selects K objects as the initial seed clustering center, then calculates the distance between each object and K seed clustering centers and assigns each object to the nearest cluster center.

The earliest clustering algorithm is the hierarchical clustering algorithm. Its main idea is to divide the given data to be clustered into levels. The purpose is to divide or aggregate the data in a hierarchical structure according to certain link rules and finally form the clustering results. Hierarchical clustering has high temporal and spatial complexity, low clustering efficiency and large error. The fast hierarchical clustering HAC algorithm reduces the space-time complexity and computation (Teymourian *et al.*, 2022). The improved split hierarchical clustering algorithm DHCA has better time efficiency than a dichotomy (Xia *et al.*, 2014).

4.4.1 Cluster hierarchy analysis process.

(1) If there are m indicators or samples, and each indicator has k sample values, m initial classes containing only one indicator will be constructed;

(2) Calculate the distance d_{ij} between two classes (indicators) of m and record it as $D = \{d_{ij}\}$, where $d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$;

(3) Merge the two nearest indicators (classes) into a new class;

(4) Calculate the distance between the new category and other indicators (categories);

(5) Repeat steps (3) and (4) until all indicators (classes) are merged into one class and the operation ends

(6) Draw the clustering pedigree;

(7) Determine the number of categories.

4.4.2 R-cluster analysis steps. There are different clustering methods according to different research objects. When classifying samples, distance is often used to measure the affinity between samples, and Q-type clustering is used. When classifying variables or indicators, the similarity coefficient is often used to measure the similarity between variables or indicators, and R-type clustering is selected.

R-type clustering analysis is a clustering analysis method for classifying and layering variables or indicators, that is, the similarity coefficient between variables or indicators is selected as clustering statistics. The specific steps are as follows:

The degree of similarity between two variables or indicators is expressed by the similarity coefficient. The closer the absolute value of the similarity coefficient is to 1, the closer the relationship between variables or indicators is. The closer the absolute value is to 0, the farther the relationship between variables or indicators is. The most commonly used similarity coefficients are as follows:

(1) Correlation coefficient. The sample correlation coefficient formula is usually used.

$$\lambda_{ij} = \frac{\sum_{l=1}^{n}(x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{\sqrt{\sum_{l=1}^{n}(x_{il} - \overline{x}_i)^2}\sqrt{\sum_{l=1}^{n}(x_{jl} - \overline{x}_j)^2}} \tag{4-8}$$

where $X_i = (x_{i1}, x_{i2}, ..., x_{in}), i = 1, 2, ..., m$ is the data vector of the i th variable. Obviously, there is $|\lambda_{ij}| \leq 1$.

(2) Included angle cosine. If any two indexes X_i and X_j are regarded as two vectors in n-dimensional space, and the cosine value of the angle between the two vectors is expressed by $\cos\theta_{ij}$.

$$r_{ij} = \cos\theta_{ij} = \frac{\sum_{l=1}^{n} x_{il}x_{jl}}{\sqrt{\sum_{l=1}^{n} x_{il}^2}\sqrt{\sum_{l=1}^{n} x_{jl}^2}} \tag{4-9}$$

where $X_i = (x_{i1}, x_{i2}, ..., x_{in}), i = 1, 2, ..., m$ is the data vector of the i-th index.

Obviously, $|\cos\theta_{ij}| \leq 1$. If $\cos\theta_{ij} = 0$, it means that the two indicators are independent of each other and can be divided into two relatively independent classes. At the same time, it can

be seen that when the mean $\overline{x}_i$ of each index is 0, the included angle cosine is a special form of the correlation coefficient.

(3) As a tool to express the approximation of variables, the included angle cosine and correlation coefficient are uniformly recorded as c_{ij}, which obviously has $|c_{ij}| \leq 1$. When $|c_{ij}| = 1$, it indicates that variable X_i is completely related to variable X_j. When $|c_{ij}| = 0$, it indicates that variable X_i is completely independent of variable X_j. The larger $|c_{ij}|$, the stronger the correlation between variable X_i and variable X_j. Accordingly, the variables with strong correlation are grouped into one category and the variables with less correlation are grouped into another category. A critical value r ($0 \leq r \leq 1$) is usually set according to the demand. When $|c_{ij}| \geq r$, variable X_i and variable X_j are grouped into one class.

In practical clustering, in order to facilitate understanding, the measurement formula of the similarity relationship between variables is usually transformed:

$$d_{ij} = 1 - |c_{ij}| \text{ or } d_{ij}^2 = 1 - c_{ij}^2 \tag{4-10}$$

d_{ij} represents the distance between variable X_i and variable X_j, and $|d_{ij}| \leq r$ classifies variable X_i and variable X_j.

4.5 K-means dynamic clustering

The system clustering method requires calculating and comparing the distance between different indicators. At the same time, the "cluster between classes" needs to be calculated at each step of category aggregation. Especially when the sample size is large, the computational workload is relatively large and the tree graph is very complex, which is inconvenient for analysis. K-means dynamic clustering method refers to the initial rough classification through many expert consultations and demonstrations and then correcting the inaccurate and unreasonable classification according to some optimal principle until the appropriate and appropriate classification results (Yang and Zhao, 2019).

4.5.1 K-means common distance formula.

(1) Absolute distance (block distance): $d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$

(2) Euclidean distance: $d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right]^{\frac{1}{2}}$

(3) Mahalanobis distance: $d_{ij} = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$

where $X_i = (x_{i1}, x_{i2}, \ldots, x_{in}), i = 1, 2, \ldots, m$ is the data vector of the i-th variable and $\boldsymbol{S}$ is the corresponding covariance matrix.

4.5.2 K-means initial classification method.

(1) If there are indicators or variables clustered into classes, empirically select indicators as the initial clustering center (condensation point).

(2) According to the principle of nearest distance, a new condensation point classification is determined. Recluster each condensing point and merge the indicators or variables closest to the condensing point to form a new condensing point class.

(3) Calculate the centroids of all new condensing points to replace the original condensing points and repeat steps (2) and (3) until all indicators or variables are classified.

Among them, the centroid $X_a = \frac{1}{n_a}\sum_{X_i \in G_a} X_i$ of the new condensation point class G_a, $X_i = (x_{i1}, x_{i2}, ..., x_{in}), i = 1, 2, ..., m$ is the data vector of the i-th variable.

K-means initial classification method has the advantages of small calculation volume and fast calculation speed. However, the final clustering result is closely related to the initial condensation point.

4.5.3 K-means dynamic clustering steps.

(1) Carry out expert consultation and demonstration many times and divide all objects (indicators or samples) into initial condensation points.

(2) According to the principle of nearest distance and based on the calculation results of Euclidean distance, each object (index or sample) is classified into the category closest to the condensation point.

(3) Recalculate the condensation points of various types for the latest objects and classes.

(4) Repeat steps (2) and (3) until the classification is stable.

As a representative and typical point, the selection of the condensation point is directly related to the classification and stratification of the index system and has a great impact on the hierarchical structure of the final index system. There are three common methods:

- Empirical selection method. According to the characteristics and requirements of the classification and stratification of the index system, the original classification is preliminarily set based on historical experience. The condensation point of each category is selected based on representative and typical indicators. This method has certain subjectivity.
- Minimum-Maximum distance method. First, the sum of two index datasets with the largest distance (or the largest representativeness and independence) is selected as two initial condensation points. Second, the distance between and each other index is measured step by step, and the index classification is determined based on the principle of minimum distance.
- Select a new condensation point X_{h3}. X_{h3} should be the minimum distance from X_{h1} and X_{h2}, which is significantly greater than the minimum distance from all other indicators to X_{h1} and X_{h2}. At this point, return to the previous step. Otherwise, the calculation step of the condensation point ends.

4.6 Binary K-means clustering

K-means dynamic clustering is sensitive to the initially selected K-aggregation point. If the K-aggregation point is not selected properly, it is difficult to obtain the global optimal solution. Qiu and Zhang (2012) combined decision tree, SVM and Binary K-means Algorithm to effectively reduce the classification time of high-dimensional data (Qiu and Zhang, 2012). Issa (2016) proposed a parallel Binary K-means algorithm based on Hadoop distributed platform, which can obtain an ideal speedup ratio (Issa, 2016).

Binary K-means clustering algorithm is an improvement of the K-means clustering algorithm in selecting cluster centroid and improving clustering time efficiency. Its core idea is to first take all clustering data as a cluster and divide it into two. Then the cluster with the largest sum of squares of residuals is divided into two clusters. Repeat the dichotomy of the clusters with the maximum sum of squares of residuals until the number of clusters is equal to K. Every time the binary K-means algorithm updates the cluster to which the clustering data belongs, it reduces the calculation of the centroid similarity between the clustering data and the clustering cluster and improves the time efficiency of the clustering algorithm.

where the sum of squares of residuals is $SSE = \sum_{i=1}^{n}(X_i - \overline{X})^2$, $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, $i = 1, 2, \ldots, n$, $\overline{X}$ is the mean of variable X_i.

The steps of binary K-means clustering algorithm are as follows:

(1) Initially, all indexes are taken as a cluster. By randomly selecting two indexes as the centroid of the initial cluster, the Euclidean distance between each index in the cluster and the centroid of the two clusters is calculated, and each index is divided into its closest cluster.

(2) After these indexes are divided into clusters, the centroid of each cluster is calculated and updated with all indexes in the cluster.

(3) Repeat steps (1) and (2) for multiple iterations until the cluster class to which all indicators belong in the previous and subsequent iterations do not change.

(4) Calculate the sum of squares of residuals of each cluster, and then use the K-means clustering algorithm to divide the cluster with the largest sum of squares of residuals into two families. At this time, the number of current clusters increases by one.

(5) Step (4) is repeated until the number of clusters is K, which means K clusters are obtained. It makes the indexes in each cluster have good local structure similarity and stability.

5. Conclusion

Based on the scientific primary selection of index systems from "theoretical basis analysis, relevant factor analysis and process mechanism description" to "system structure analysis, system hierarchical decomposition, complete set identification of indicators" and "expert assignment quality inspection," this paper designs the data quality diagnosis system from the perspective of data preprocessing, which expounds the index data review, index data cleaning, index data conversion and index data inspection process to ensure the quality of index data. For index classification and stratification, the hierarchical structure classification design of the index system is carried out based on machine learning methods such as Classification Decision Tree, Interpretation Structure Model, Hierarchical Cluster Analysis and K-Means Dynamic Clustering.

Through modern methods such as complex networks, machine learning and data fusion technology, on the one hand, this paper systematically designs the quality diagnosis of index data and the classification and stratification of index system, effectively ensuring the quality of index data. On the other hand, we realize the scientific classification and stratification of index system, reduce the subjectivity and randomness of index system design, enhance objectivity and scientificity and lay a solid foundation for the optimal design of index system.

References

Alcantud, J.C.R., Feng, F. and Yager, R.R. (2019), "An *N*-soft set approach to rough sets", *IEEE Transactions on Fuzzy Systems*, Vol. 28 No. 11, pp. 2996-3007.

Bao, X., Gao, H. and Hu, L. (2016), "A comparative study of multiple interpolation methods in longitudinal missing data", *China Health Statistics*, Vol. 33 No. 1, pp. 45-48.

Chen, J.F., Hsieh, H.N. and Do, Q.H. (2015), "Evaluating teaching performance based on fuzzy AHP and comprehensive evaluation approach", *Applied Soft Computing*, Vol. 28, pp. 100-108.

Cheng, K. (2010), *Research on Statistical Data Quality Diagnosis and Management*, Vol. 12, Zhejiang Gongshang University Press, Hangzhou.

Ding, Y. and Wang, X. (2019), "Naive Bayesian classification algorithm based on improved feature weighting", *Research on Computer Application*, Vol. 36 No. 12, pp. 3597-3600+3627.

Ding, J., Huang, T., Xu, J. and Wang, J. (2020), "Anomaly detection method for multidimensional distance clustering based on time series", *Computer Engineering and Design*, Vol. 41 No. 7, pp. 1935-1940.

Duan, C. (2020), "Construction of online lending credit evaluation index system based on K-S test and distance correlation analysis", *Technical Economy*, Vol. 39 No. 5, pp. 35-47+59.

Escobar, N., Márquez, I.A.V., Quiroga, J.A., Trujillo, T., González, F., Aguilar, M. and Escobar-Pérez, J. (2017), "Using Positive Deviance in the prevention and control of MRSA infections in a Colombian hospital: a time-series analysis", *Epidemiology and Infection*, Vol. 145 No. 5, pp. 981-989.

Forghani, Y. and Yazdi, H.S. (2014), "Robust support vector machine-trained fuzzy system", *Neural Networks*, Vol. 50, pp. 154-165.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997), "Bayesian network classifiers", *Machine Learning*, Vol. 29 No. 2, pp. 131-163.

Giorgio, G. and Alessandro, R. (2016), "A Robust approach for the background subtraction based on multi-layered self-organizing maps", *IEEE Transactions on Image Processing : A Publication of the IEEE Signal Processing Society*, Vol. 25 No. 11, pp. 5239-5251.

Gu, J., Ren, P. and Xu, Z. (2015), "Research on performance evaluation method of venture capital guiding fund based on intuitive fuzzy analytic hierarchy process", *Chinese Management Science*, Vol. 23 No. 9, pp. 124-131.

Hai-Min, L., Wan-Huan, Z., Shui-Long, S. and An-Nan, Z. (2020), "Inundation risk assessment of metro system using AHP and TFN-AHP in Shenzhen", *Sustainable Cities and Society*, Vol. 56, 102103.

Hosseini, S.M.J., Arasteh, B., Isazadeh, A., Mohsenzadeh, M. and Mirzarezaee, M. (2021), "An error-propagation aware method to reduce the software mutation cost using genetic algorithm", *Data Technologies and Applications*, Vol. 55 No. 1, pp. 118-148.

Huang, G. and Chen, L. (2021), "A comparative study of foreign scientific data quality assessment frameworks", *Library and Information*, No. 1, pp. 97-107.

Issa, J. (2016), "Performance characterization and analysis for Hadoop K-means iteration", *Journal of Cloud Computing*, Vol. 5 No. 1, pp. 1-15.

Janssen, M., Haiko, V. and Wahyudi, A. (2017), "Factors influencing big data decision-making quality", *Journal of Business Research*, Vol. 70, pp. 338-345.

Jiang, H., Ma, C., Xu, X. and Lan, Q. (2019), "Multivariate missing data interpolation algorithm imitating EM and its application in credit evaluation", *China Management Science*, Vol. 27 No. 3, pp. 11-19.

Kouziokas, G.N. (2020), "A new W-SVM kernel combining PSO-neural network transformed vector and Bayesian optimized SVM in GDP forecasting", *Engineering Applications of Artificial Intelligence*, Vol. 92, 103650.

Laura, K.N. (2020), "Computational grounded theory: a methodological framework", *Sociological Methods and Research*, Vol. 49 No. 1, pp. 3-42.

Lee, Y.W., Pipino, L.L., Funk, J.D. *et al.* (2015), *The Journey of Data Quality*, Higher Education Press, Beijing, Vol. 7, Complied by Huang, Wei, Wang, Jiayin, Su, Qin et al.

Li, S., Wang, M., Chai, J., Zhao, Z. and Zhang, L. (2021), "Research on influencing factors of power grid green development based on DEMATEL-ISM model", *Practice and Understanding of Mathematics*, Vol. 51 No. 8, pp. 283-294.

Lin, S. and Luo, W. (2019), "A new multilevel CART algorithm for multilevel data with binary outcomes", *Multivariate Behavioral Research*, Vol. 54 No. 4, pp. 578-592.

Lin, X., Cui, S., Han, Y., Geng, Z. and Zhong, Y. (2019), "An improved ISM method based on GRA for hierarchical analyzing the influencing factors of food safety", *Food Control*, Vol. 99, pp. 48-56.

Liu, B. and Lin, P. (2019), "Review on the quality of big data at home and abroad", *Journal of Information Science*, Vol. 38 No. 2, pp. 217-226.

Liu, Z., Zhan, Q. and Tian, G. (2019), "A review of factor Analysis comprehensive evaluation research", *Statistics and Decision-Making*, Vol. 35 No. 19, pp. 68-73.

Luo, S., Sun, Z. and Pan, L. (2020), "Soft clustering node split hierarchy model", *Journal of Beijing Institute of Technology*, Vol. 40 No. 3, pp. 305-309.

Mo, Z. (2018), "Construction of big data quality measurement model", *Intelligence Theory and Practice*, Vol. 41 No. 3, pp. 11-15.

Peng, Z., Zhang, A., Wang, S. and Bai, Y. (2017), "Design principles and construction process of comprehensive evaluation index system", *Scientific Research Management*, Vol. 38 No. S1, pp. 209-215.

Qiu, G. and Zhang, J. (2012), "Adaptive classification method of SVM decision tree based on binary K-means", *Research on Computer Application*, Vol. 29 No. 10, pp. 3685-3687+3709.

Ruggieri, S. (2002), "Efficient C4. 5 [classification algorithm]", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 14 No. 2, pp. 438-444.

Shamim, S., Zeng, J., Shariq, S.M. and Khan, Z. (2019), "Role of big data management in enhancing big data decision-making capability and quality among Chinese firms: a dynamic capabilities view", *Information and Management*, Vol. 56, 103135.

Shuang, H., Li, G., Feng, J. and Wang, N. (2018), "Review of structured data cleaning technology", *Journal of Tsinghua University (Natural Science Edition)*, Vol. 58 No. 12, pp. 1037-1050.

Song, X. and Jiang, S. (2015), "Research on ecological performance evaluation of eco-industrial parks based on principal component analysis and set-pair analysis: taking Shandong eco-industrial park as an example", *Resource Science*, Vol. 37 No. 3, pp. 546-554.

Teymourian, N., Izadkhah, H. and Isazadeh, A. (2022), "A fast clustering algorithm for modularization of large-scale software systems", *IEEE Transactions on Software Engineering*, Vol. 48 No. 4, pp. 1451-1462.

Wang, R.Y. and Strong, D.M. (1996), "Beyond accuracy: what data quality means to data consumers", *Journal of Management Information Systems*, Vol. 12 No. 4, pp. 5-33.

Wang, Y.R., Kon, H.B. and Madnick, S.E. (1993), "Data quality requirements analysis and modeling", *International Conference on Data Engineering*, IEEE, Vienna.

Wang, N., Du, H. and Wang, S. (2009), "Hierarchical clustering algorithm based on modular index optimization", *World Science and Technology Research and Development*, Vol. 31 No. 5, pp. 824-826+807.

Wang, Y., Huang, W. and Zu, Z. (2014), "Some preliminary thoughts on big data industry and management issues", *Science and Technology Promotes Development*, No. 1, pp. 15-19.

Wang, D., Zhu, J., Liu, X. and He, L. (2018), "A review and prospect of research on functional data clustering analysis", *Mathematical Statistics and Management*, Vol. 37 No. 1, pp. 51-63.

Xia, Q., Li, X., Song, Y. and Zhang, B. (2014), "Aircraft grouping based on improved divisive hierarchical clustering algorithm", *Journal of Air Transport Management*, Vol. 40, pp. 157-162.

Xu, L. and Li, M. (2020), "Research on data preprocessing methods in dynamic comprehensive evaluation", *Chinese Management Science*, Vol. 28 No. 1, pp. 162-169.

Yang, J. and Zhao, C. (2019), "A review of K-Means clustering algorithm research", *Computer Engineering and Application*, Vol. 55 No. 23, pp. 7-14+63.

Yang, S., Guo, J.Z. and Jin, J.W. (2018), "An improved Id3 algorithm for medical data classification", *Computers and Electrical Engineering*, Vol. 65, pp. 474-487.

Ye, F. (2019), "Research on clustering algorithm of uncertain data stream based on outlier detection", *Journal of Chinese Academy of Electronic Sciences*, Vol. 14 No. 10, pp. 1094-1099.

Yin, J., Wang, N., Ge, S. and Liu, W. (2017), "Validity verification of information system complexity measurement method based on data complexity", *Journal of Management*, Vol. 14 No. 4, pp. 590-599.

Yin, K., Zhou, S. and Xu, T. (2019), "Research on optimization of index system design and its inspection method: indicator design and expert assessment quality inspection", *Marine Economics and Management*, Vol. 2 No. 1, pp. 1-28.

Yuan, C., Luo, L. and Sheng, X. (2004), "Research on data quality in information system", *Chinese Library Journal*, No. 1, pp. 50-52.

Yuan, M., Liu, F., Zeng, C. and Xie, L. (2018), "A review of data quality dimensions and frameworks", *Journal of Jilin University (Information Science Edition)*, Vol. 36 No. 4, pp. 444-451.

Zhao, X. and Liang, J. (2016), "A weighted clustering algorithm for mixed data attributes based on information entropy", *Computer Research and Development*, Vol. 53 No. 5, pp. 1018-1028.

Zou, J., Zhu, Q. and Liu, P. (2018), "Research on the vulnerability factors of traditional tourism villages based on interpretation structure model", *Economic Geography*, Vol. 38 No. 12, pp. 219-225.

Corresponding author
Shiwei Zhou can be contacted at: 523887825@qq.com