International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 11 ǁ November 2023 ǁ PP. 14-25

A Novel Approach of Fusion Relevance Analysis towards Real Taxation Data Sets

Daifang Gou[1, 2] • Linsong Xiao[1, 3] • Chengyu Hou[1, 3] • Chang Liu[4] • Yue Wang[5] • Wenzao Li[1, 3]

[1] *College of Communication Engineering, Chengdu University of Information Technology, Chengdu, China*
[2] *State Administration of Taxation Yibin Sanjiang New District Taxation Bureau, Yibing, China*
[3] *Meteorological information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes of Chengdu University of Information Technology, Chengdu, China*
4 Chengdu Civil Affairs Bureau Office, Chengdu, China
[5] *Education Informationization and Big Data Center, Education Department of Sichuan Province, Chengdu, China*

Abstract: *Nowadays, tax data is increasingly important for understanding the rationale of relevant industries and the feedback of corporate economic cycles to the market. Therefore, the reasonable and timely intelligent analysis and processing of tax data have become a hot research filed. However, in some cases where the data has fewer dimensions, there are fewer means to mine deep-level information. Meanwhile, there is a lag in intelligent analysis and application, resulting in weak support for the industry from relevant data. In addition, there is a lack of effective detection capabilities for abnormal data. To address these issues, in this paper, we considered the accuracy of various algorithms and took real tax revenue data from XX city as the research object. We proposed a fusion relevance analysis algorithm that combines the Pearson algorithm and cosine similarity algorithm Cosine Similarity Algorithm (CSA). Firstly, we used the real tax data set of two companies between 2019 to 2021 for lime. Then, using this algorithm, we conducted correlation analysis on the relevant data of different years. The analysis results to some extent revealed some deep-level information, further enhancing the feasibility of automated data analysis. Subsequently, the results were compared with those of Chi-square Algorithm (CA) and Manhattan Algorithm (MA). Through corresponding numerical tables, it was evident that under this model condition (based on real taxable amount data), The simulation results show that compared with cosine similarity, the accuracy of companies A and B has increased by an average of 28.66% and 42.92% respectively.*
Keywords: *Comprehensive analysis, Data mining, Data relevance analysis, Tax data.*

Date of Submission: 20-10-2023 Date of acceptance: 03-11-2023

I. Introduction

1.1 Background and motivation

In today's data-driven era, in-depth understanding of the rationality of related industrial structures, market feedback on cyclical goods, and the assessment of different resilience of similar enterprises to similar situations can be achieved through the analysis of tax-related data. These insights not only have a significant impact on policymakers, industry stakeholders, and economists but also raise crucial questions about economic development and industrial optimization. However, despite the abundance of data available to managers today, there is a significant lag in the intelligent analysis and application of tax-related data, leading to a decline in the ability of data to support industries [[1]].

1.2 Challenges and solutions

The value of tax data goes far beyond its surface. They carry crucial information about economic activities, including corporate profits, market demand, and macroeconomic conditions. However, fully leveraging this data requires effective analysis methods and tools to uncover the underlying conditions and demands. Additionally, the detection of anomalous data remains a challenge, especially in large-scale tax datasets [[1]].

To address these issues, this study introduces an innovative fusion data relevance analysis method. Compared to traditional analysis algorithms such as chi-square analysis and Manhattan algorithm, the proposed algorithm in this paper has more advantages in this model, providing more accurate judgments on relevance and greater sensitivity in detecting anomalous data [0，[3]]. Using real industry tax revenue data from XX city, this paper first conducts a correlation analysis of the taxable amounts of two companies for different years. Subsequently, the tax data for these two companies from 2019 to 2021 is used to construct line charts, allowing

for an intuitive observation of trends and changes. To further explore data information and improve accuracy, various data relevance analysis methods are employed, including the proposed fusion relevance analysis algorithm, Manhattan algorithm, and chi-square analysis. These algorithms not only help detect correlations between data but also identify potential anomalous data points.

This paper will present detailed data analysis results and provide in-depth discussion and interpretation of these results. This will help reveal data patterns and trends under specific conditions and provide insights for real-world application decisions. The comprehensive nature of this study lies not only in its focus on intelligent analysis of tax data but also in the detection and interpretation of anomalous data. By constructing relevance analysis and anomaly detection models, this paper aims to provide a comprehensive approach for businesses, policymakers, and researchers to better understand the potential value of tax data and provide more reliable data for industry and decision support. In the following chapters, this paper will detail the research methods, experimental results, and discussions, hoping to contribute useful insights for further exploration in the fields of data analysis and economic research.

1.3 Contributions and organization

1) Establishment of correlation analysis among similar companies

In today's highly competitive business environment, understanding the correlation between different companies within the same industry becomes crucial. This correlation can include partnerships, market competition, supply chain connections, etc. Therefore, this study aims to establish an effective correlation analysis method to explore potential connections between similar enterprises. By analyzing these correlations, this paper can better understand interactions and influences within the industry, providing strong support for companies in formulating strategies and decisions.

2) Establishment of a data anomaly detection model based on fusion relevance analysis algorithm, chi-square analysis, and Manhattan distance

Data anomaly detection has always been a key issue in data analysis. In this study, we not only focus on correlation analysis but also on the detection and handling of anomalous data. This paper proposes a series of anomaly detection models based on different algorithms, including the fusion relevance analysis algorithm, chi-square analysis, and Manhattan distance. These algorithms can help us quickly and accurately identify anomalies in data, which is particularly important for large-scale datasets like tax data.

3) Analysis and discussion of the correlation of the same period, year, and futures based on the above models

In the study, this paper will use the aforementioned algorithm models to conduct correlation analysis of data from the same period, year, and futures. Specifically, the paper will focus on the correlation between companies in taxable amount data and whether anomalies exist.

II. Related Work

In this section, we will comprehensively discuss three key aspects: common data analysis methods, current issues in data analysis processing, and the correlation analysis algorithm proposed in this paper.

2.1 Common Data Analysis Methods

Common descriptive statistics use methods such as mean, median, mode, variance, standard deviation, and percentiles to summarize the central tendency, dispersion, and distribution of numerical data. Hypothesis testing uses methods such as t-tests, chi-square tests, analysis of variance, and z-tests to assess significant differences or relationships between variables or groups. Regression analysis models the relationship and makes predictions for numerical data using methods such as linear regression, logistic regression, and polynomial regression. Cluster analysis involves k-means clustering and hierarchical clustering, grouping similar data points based on numerical or categorical data. Classification uses methods like decision trees, random forests, and support vector machines to classify data points into predefined categories, typically using classification or labeled data [[4]].

Taking the example of research on decision tree algorithms domestically and abroad, an earlier decision tree algorithm is the ID3 algorithm, which has poor timeliness in data classification. In response to the shortcomings of the ID3 algorithm, some domestic experts and scholars have conducted a series of studies. For example, Wang Tao and others designed and optimized an improved fuzzy decision tree classification algorithm, which significantly reduces noise data in data classification and improves the efficiency and effectiveness of data processing. Wang Xizhao analyzed the impact of the number of attribute values on decision tree inductive learning, using the idea of selecting an appropriate branch merging strategy to merge branches of the classification tree to improve the tree's interpretability and generalization ability. Zhai Junhai and others proposed a novel fuzzy decision tree fusion method, which can effectively improve the timeliness of data classification and significantly enhance the accuracy of algorithm data processing. In addition, some scholars have proposed corresponding decision tree improvement algorithms based on the characteristics of the data. For instance, Xu Peng and others

adopted a C4.5 decision tree classification method based on the characteristics of traffic data.

Chart 2.1

data mining algorithm	Related description
Decision tree algorithm [[5]]	An algorithm that performs tree structure classification or regression based on the attributes of the data.
random forest algorithm [[6]]	Ensemble algorithm composed of multiple decision trees for classification and regression.
K-means clustering algorithm [[8]]	Divide the data points into K clusters, each cluster having similar characteristics.
DBSCAN clustering algorithm [**Error! Reference source not found.**]	Density-based clustering algorithm capable of identifying irregularly shaped clusters.
Support vector machine algorithm [[10]]	Supervised learning algorithms for classification and regression to find optimal segmentation hyperplanes.
Naive Bayes algorithm [[11]]	Classification algorithm based on Bayes' theorem, suitable for tasks such as text classification.
Neural network algorithm [[12]]	Algorithms based on artificial neuron simulations for deep learning and pattern recognition.

In foreign research, to meet the needs of processing large-scale datasets, scholars have proposed numerous improved models for decision tree algorithms. For instance, Breiman and others introduced a binary tree classification method that uses the Gini coefficient as a measure of attribute importance and can handle continuous data [[13]]. Mehta and others addressed the difficulty of handling large-scale data by traditional methods such as ID3 and C4.5, proposing the SLIQ algorithm, which can rapidly and dynamically select classification nodes, making it suitable for large-scale data processing. Rajeev and others integrated the building and pruning phases into the PUBLIC algorithm, significantly improving efficiency compared to the original algorithm. Kema and others, based on C4.5, presented a novel hybrid classification system and verified its classification capabilities using cross-validation.

2.2 Existing Issues

The current commonly used methods are typically optimized for multivariate datasets. If applied to single-variable data under the conditions discussed in this paper, it may lead to increased computational costs and extended processing times, impacting real-time or interactive analysis tasks that require responsiveness [[13]].

Standard data analysis techniques often lack growth rate analysis, focusing predominantly on static data snapshots. This can result in enterprises and researchers overlooking crucial insights into data evolution and changes over time, affecting their decision-making abilities regarding trends, patterns, and emerging opportunities or issues.

Moreover, limited correlation analysis methods often simplify and fail to capture complex nonlinear relationships or hidden correlations. This may result in incomplete or inaccurate analyses, potentially leading to misleading decisions, missed opportunities, or the inability to detect underlying factors influencing the data.

2.3 The Proposed Algorithm in This Paper

The algorithm introduced in this paper is a fusion correlation analysis algorithm that analyzes correlation by weighted calculation of Pearson's coefficient and cosine similarity [[15]，[15]]. The algorithm is compared with chi-square analysis and the Manhattan distance algorithm using real tax revenue data from XX city. Experimental results indicate that, under this model, the fusion correlation algorithm proposed in this paper outperforms the two traditional algorithms in terms of accuracy and intuitiveness. Additionally, it addresses the issue of being unable to detect anomalous situations.

III. Tax Revenue Data Modeling

By understanding the fundamental characteristics and importance of tax revenue data, we will delve into specific methods and strategies for analyzing this data in detail. We will conduct an in-depth examination of the structure of tax revenue data, with a particular focus on key indicators that can reveal a company's financial condition and economic behavior, laying the groundwork for more in-depth and detailed data analysis. In the following sections, we will explore seasonal analysis, correlation analysis, and the development of a final association analysis model in more detail.

3.1 Discussion on Tax Revenue Data based on Fusion Algorithm

Tax revenue data, as a reflection of a country's economic activities, possesses unique characteristics distinct from other forms of financial data. Understanding these unique features is crucial for effectively analyzing and modeling phenomena related to taxation. Tax revenue data typically includes various financial indicators such as taxable income, deductions, and exemptions, collectively providing insights into the financial health of

taxpayers. Additionally, it encompasses a time dimension, allowing us to examine tax behaviors and trends across different years, quarters, and fiscal periods.

To ensure the applicability and quality of tax revenue data in subsequent analyses, we have undertaken a series of preprocessing steps. Data preprocessing involves tasks such as data cleaning, transformation, and normalization. We provide detailed explanations of the techniques employed to handle missing values, outliers, and data inconsistencies, thereby enhancing the reliability of the dataset.

3.2 Discussion About Seasonality in Data

We will delve into the cyclical trends present in tax revenue data. Our investigation will include an examination of market responses to cyclical goods and a study of how different fiscal years affect the risk resilience of enterprises. Using time series analysis methods, our goal is to reveal inherent seasonal and cyclical patterns in the data and elucidate their impact on businesses and industries.

In tax revenue data, market responses to cyclical goods often exhibit clear cyclic patterns, including seasonal fluctuations in sales and variations in sales volumes during different fiscal quarters. Beyond the cyclical nature of market responses to cyclical goods, our investigation also covers how individual enterprises adjust their performance and formulate strategies to cope with increased economic uncertainty during cyclical fluctuations.

To provide a comprehensive perspective, we will analyze tax revenue data from various industry sectors, enabling us to identify performance differences during cyclical fluctuations and assess the sensitivity of each industry sector to market cycles. This analysis contributes to a better understanding of the complex interrelationships and dependencies between different sectors in the economic landscape.

3.3 Discussion of Data Correlation

To unveil the intricate network of data correlations, we have carefully selected analytical methods suitable for tax revenue data analysis. Specifically, we employ the Pearson correlation coefficient and chi-square analysis, both of which demonstrate excellent statistical techniques for measuring the strength and direction of associations between variables. One focus of this section is the comparative analysis of corresponding tax amounts, breaking down the data by carefully examining tax amount variations across different fiscal quarters and years. Detailed graphical representations, including line graphs and bar charts, will be presented to intuitively capture patterns and fluctuations in tax amounts. Additionally, statistical tests will be employed to assess the significance of the observed correlations. To ensure the robustness of the analysis, various data processing techniques such as the Manhattan distance algorithm and cosine similarity analysis are employed. These methods are chosen because they can handle complex datasets and reveal hidden relationships that may evade detection by conventional statistical methods.

3.4 Association Model Formulation

This article introduces an association analysis model based on two commonly used correlation algorithms. By assigning weights to different parameters and performing calculations, we derive the formula for this model as equation (1):

$$W = \lambda_1 P + \lambda_2 C \tag{1}$$

Among them: W is the final reference value of the model, P represents the value of Pearson correlation coefficient, C represents the value of cosine similarity, and λ_1 and λ_2 represent the weight coefficients. By assigning weights, we combine multiple correlation algorithms designed for low-dimensional data, thus obtaining a more precise method to represent the relationships between tax data.

IV. Tax Data Dimension Discussion Model

4.1 Pearson Correlation Coefficient

The Pearson correlation coefficient, commonly referred to as the linear correlation coefficient, is a statistical indicator used to measure the strength of linear dependence between two continuous variables. In various fields, especially in data mining, psychometrics, and bioinformatics, this coefficient is often used as an initial tool for correlation estimation.

Let two sets of random variables be denoted as X and Y, with their observed values as X_i and Y_i, and their respective means as $\overline{X}$ and $\overline{Y}$. In this context, the Pearson correlation coefficient r is defined as (2):

$$r = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2 \sum(Y_i-\overline{Y})^2}} \tag{2}$$

The structure of this formula elucidates its essence: in the numerator, we measure the joint fluctuation or covariance of two variables, while the denominator standardizes these fluctuations, ensuring that the result falls within the range of -1 and 1.

A noteworthy characteristic of the Pearson coefficient is its symmetry. That is to say, the Pearson coefficient between variables X and Y is the same as that between Y and X This property ensures that

regardless of how the two variables are chosen or labeled, the measurement of correlation is consistent.

However, the Pearson coefficient has its limitations which like many statistical methods. Most notably, it can only measure linear relationships for two curves of taxing data. Complex non-linear relationships may not be adequately captured. Additionally, its sensitivity to outliers means that the presence of outliers can significantly skew the coefficient.

4.2 Manhattan Distance

Manhattan distance often referred to as the L1 norm or taxi distance, originates from normed vector spaces and serves as a key tool in various fields, including machine learning, computational geometry, and data mining. Due to its simplicity and practicality in a variety of applications, especially in grid-based patterns or lattice structures, it stands out.

Mathematically, the Manhattan distance between two n-dimensional points with coordinates $(p_1, p_2, \ldots, p_n)$ and $(q_1, q_2, \ldots, q_n)$ is defined equation (3) as follows:

$$D_M(P, Q) = \sum_{i=1}^{n} |p_i - q_i| \quad (3)$$

This formula captures the essence of its name. Just like in a city with grid patterns, such as Manhattan, where taxis travel along vertical and horizontal routes to reach their destination, Manhattan distance calculates the sum of the vertical and horizontal distances between two points in a grid. Unlike the Euclidean distance, which measures the shortest straight-line distance between two points, Manhattan distance is fundamentally grid-based.

Manhattan distance has several notable characteristics:

1. Non-negativity: for any points P and Q, $D_M(P, Q) \geq 0$. Furthermore, if only $D_M(P, Q) = 0$, $P = Q$.
2. Symmetry: For any two points a and b, the Manhattan distance from a to b is the same as the distance from P to Q, $D_M(P, Q) = D_M(Q, P)$.
3. Triangle Inequality: For any points P, Q and R, $D_M(P, R) \leq D_M(P, R) + D_M(Q, R)$.

In the context of data mining and machine learning, Manhattan distance is often used in algorithms that require distance measurements, such as the K-nearest neighbors algorithm, especially when the feature dimensions are inherently grid-aligned or when the L1 norm is more suitable due to the nature of the data. Its computational simplicity, combined with its robustness to changes in data distribution, often makes it a preferred choice in high-dimensional spaces, where the curse of dimensionality can render other distance metrics ineffective or computationally expensive.

In conclusion, Manhattan distance provides a fundamental, grid-centered distance metric in a wide range of metric spaces. Its straightforward properties, coupled with its alignment with grid structures, make it a robust and reliable tool for addressing various academic and applied challenges in data mining and other fields.

4.3 Chi-Square Analysis

Chi-square analysis is primarily based on the chi-square statistic, a classic method used to test differences between observed and expected frequencies in categorical data. It finds extensive application in various statistical investigations, particularly in the fields of social sciences, biology, and medical research.

Built upon the foundation of chi-square tests, it is commonly employed to determine the independence between two categorical variables. The essence of this test lies in comparing observed frequencies (actual data) with expected frequencies (the frequencies expected under the assumption of independence between variables).

The mathematical definition of the chi-square statistic is:

$$x^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

In this context, O_{ij} represents the observed frequency in the *i*-th row and j-th column, while represents E_{ij} the corresponding expected frequency. The expected frequency is calculated under the assumption of complete independence between the two variables.

A larger value of the chi-square statistic implies a significant difference between the observed and expected frequencies, providing sufficient evidence to reject the null hypothesis (the E_{ij} variable is independent). Conversely, a smaller chi-square value indicates that the difference between observation and expectation can be attributed to random error.

It is important to understand that the chi-square test has its own set of assumptions and limitations. One major prerequisite is that the expected counts for each cell should be sufficiently large, typically considered as 5 or greater. When this condition is not met, the results of the chi-square test may no longer be accurate.

In the context of data mining, chi-square analysis is often used for feature selection, especially when selecting classification features most relevant to the target variable. Additionally, it is frequently employed in association rule mining to determine the strength of associations between items.

4.4 Cosine Similarity

Cosine similarity is a measurement standard that captures the cosine value of the angle between two non-zero vectors, making it an intuitive representation of direction and alignment. It originates from the geometric interpretation of multi-dimensional spaces and is a measure of distinguishing vector direction consistency without considering their magnitudes. This property makes cosine similarity very useful in many applications, especially in text mining and information retrieval, where the size of vectors (e.g., raw term frequencies) may be less relevant than the direction they represent (e.g., context or meaning).

Given two vectors, A and B, their cosine similarity is computed as follows:

$$sim(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (5)$$

In this context, $A \cdot B$ represents the dot product of vectors. $\|A\|$ and $\|B\|$ respectively denote the magnitudes (or Euclidean norms) of vectors A and B The resulting value of this calculation will always lie between -1 and 1. The value of 1 signifies complete similarity (vector alignment), 0 indicates orthogonality (no similarity), and -1 denotes complete dissimilarity (vectors are completely opposed).

A fundamental characteristic of cosine similarity is its invariance to vector magnitude. This means that if we scale a vector proportionally (increase its magnitude) without changing its direction, its cosine similarity with other vectors remains unchanged. This property is invaluable when dealing with raw data because the pure frequency of terms can mask the contextual importance of terms.

In the field of data mining, cosine similarity plays a crucial role in document clustering, collaborative filtering, and content-based recommendation systems. By treating documents, users, or items as vectors (usually in term frequency or TF-IDF space), cosine similarity helps identify patterns of similarity or content relevance, assisting systems in making informed recommendations or clustering items.

In summary, cosine similarity serves as an important, direction-centric measure in multidimensional space. Its insensitivity to magnitude, coupled with its geometric intuition, makes it a versatile and powerful metric in the toolkit of data scientists and researchers across various domains.

V. Results and Discussion based on Fusion Algorithm

In this chapter, we will conduct data analysis and experiments based on Company A and Company B. We will begin by analyzing data using four commonly used correlation algorithms, providing examples to illustrate the data relationships they reflect. Finally, we will perform data analysis using the proposed data analysis algorithm in this paper to demonstrate the advantages of the proposed algorithm.

5.1 Chi-Square Analysis Based on Company Tax Data

Using publicly available taxable amount data from Company A and Company B for the years 2019 to 2021, sourced from the XX City Tax Bureau, we created line graphs depicting the taxable amounts for each month for these two companies. The following charts were derived from the experimental analysis, where we conducted chi-square analyses for each quarter between two consecutive years and visualized the results in bar graphs. For instance, we calculated the chi-square analysis results for the taxable amounts between the first quarters of 2019 and 2020 for Company A, drawing relevant conclusions.

Initially, we plotted line graphs for the taxable amounts of both companies from 2019 to 2021, providing a visual representation of their general relationship for subsequent experimental comparisons.

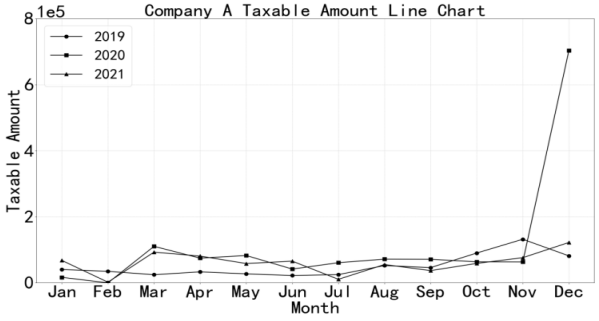

Fig1 company A Taxable Amount Line

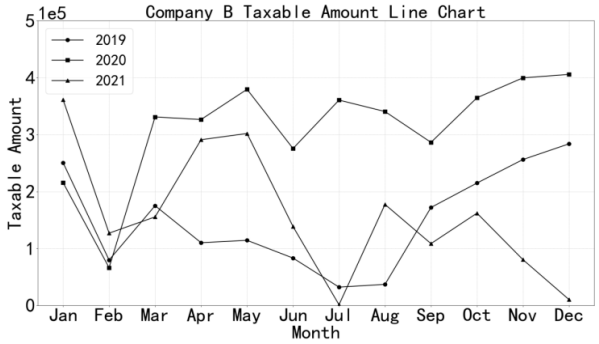

Fig2 company B Taxable Amount Line

In the above graph, it can be observed that Company A experienced relatively minor fluctuations over the three years, with a sudden spike in the last month of 2020 due to uncontrollable factors. The relationships among the three lines representing Company B are not very clear. At this point, utilizing relevant data correlation analysis algorithms can help showcase the corresponding data relationships and assist in analyzing the tax data.In this section, chi-square analysis was employed for data analysis, yielding two key values: the T-statistic and the P-value.

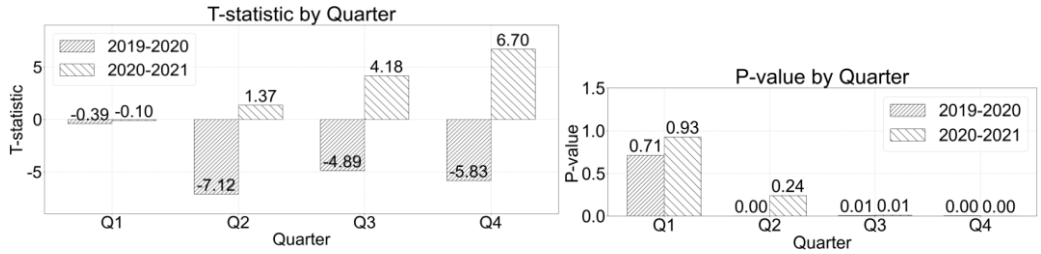

Fig3 T-statistic and P-statistic by Quarter

As shown in the above graph, the T-statistic and P-values for Company A are represented using a bar chart. For example, in the first quarter of 2019 and 2020, the T-statistic for the company is -0.28, and the P-value is 0.81. Since the T-statistic serves as an indicator of significant differences between data, and the P-value represents the probability of obtaining a statistic as extreme as, or more extreme than, the one observed under the assumption that the null hypothesis is true.

Analyzing these values, we can conclude that the difference between the taxable amounts for Company A in the first quarter of 2019 and 2020 is relatively low. This indicates that there is no significant difference, and there are no abnormal values in the taxable amounts for Company A during these two years in the first quarter.

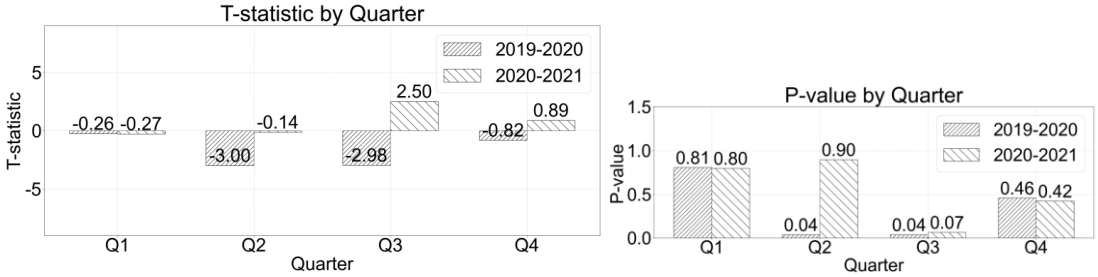

Fig4 T-statistic and P-statistic by Quarter

The same analysis can be applied to Company B. In Figure 4, for the fourth quarter of 2020 and 2021, the T-statistic value is 6.70, and the P-value is 0. Therefore, we can conclude that there is a significant difference in the taxable amounts for Company B in the fourth quarter of 2020 and 2021. This suggests the presence of potentially anomalous values in the taxable amounts for Company B during these two years in the fourth quarter.

5.2 Manhattan Distance Correlation Discussion Based on Company Tax Revenue Data

This section primarily focuses on analyzing the correlation between the two companies using the Manhattan distance algorithm. We employ this algorithm to analyze the data and create corresponding bar charts to showcase the observed data correlation in this dimension. The experimental results are depicted in the following

graph. We continue to analyze the data relationships on a quarterly basis for each pair of consecutive years, allowing us to observe whether there are any anomalous values.

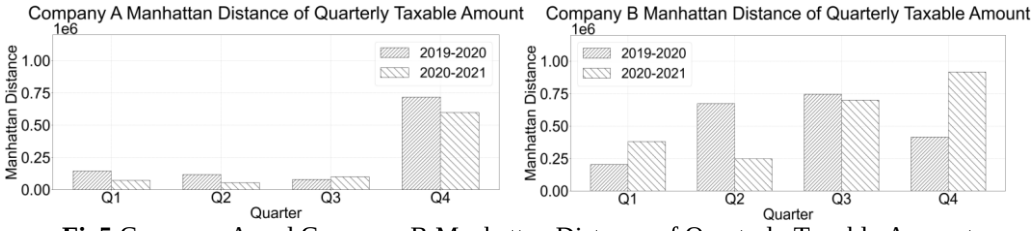

Fig5 Company A and Company B Manhattan Distance of Quarterly Taxable Amount

The Manhattan distance formula is used to demonstrate the degree of separation between data, in this experiment, specifically the differences in taxable amounts between consecutive quarters for each company over two years. This aids in detecting anomalous tax data and identifying the range of abnormal data. For instance, in Figure 4, the Manhattan distance value for Company A's fourth quarter of 2019-2020 is larger compared to other quarters, indicating a significant difference in taxable amounts between those two years during that quarter, highlighting the time period when anomalous data occurred. In contrast, Company B's Manhattan distance value for the second quarter of 2020-2021 is relatively smaller, suggesting similarity in taxable amounts between those two years during that quarter.

5.3 Discussion of Cosine Similarity in Company Tax Data

Cosine similarity is a crucial method in the process of analyzing data correlation. It primarily focuses on the angles between data vectors, representing the direction and trend of the data. The value of cosine similarity can help determine whether two sets of data exhibit positive or negative correlation. Additionally, it can reveal periodic information about the taxable amounts of these companies.

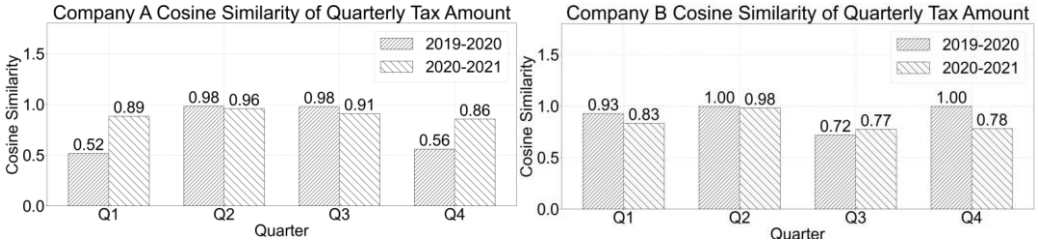

Fig6 Company A and Company B Cosine Similarity of Quarterly Taxable Amount

Regarding the results of cosine similarity, it can also roughly indicate the trends of increase or decrease between data sets. For instance, the cosine similarity for Company A between the first quarters of 2019 and 2020 is only 0.52, indicating inconsistent trends in their increase or decrease, suggesting a significant difference in direction. On the other hand, for Company B in the second quarter of 2019 and 2020, the cosine similarity is 1, suggesting that the data trends in those quarters for the two years are consistent, with essentially no anomalies.

5.4 Pearson Correlation Coefficient Discussion of Company Tax Data

The Pearson correlation coefficient is a statistical method that measures the degree of linear relationship between two sets of data. Like cosine similarity, Pearson correlation coefficient is an important tool for exploring correlations in data. It not only reveals linear correlations between data sets but also provides a quantitative analysis of data trends.

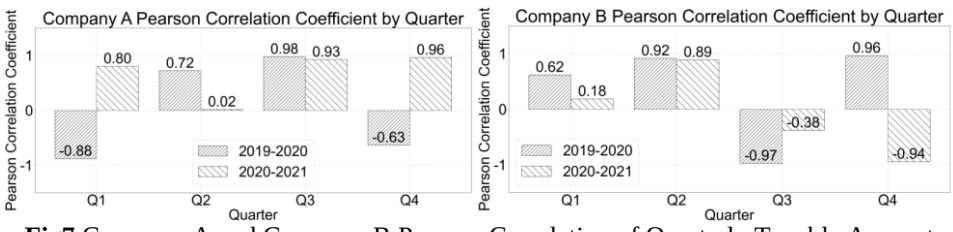

Fig7 Company A and Company B Pearson Correlation of Quarterly Taxable Amount

Take companies A and B as an example:In Figure 7, the Pearson coefficient for Company A between the first quarters of 2019 and 2020 is -0.88. This indicates that during that quarter, there is a roughly negative correlation in the taxable amounts for the two years, suggesting a significant difference in trends. In contrast, for Company B, the Pearson coefficient for the first quarters of 2020 and 2021 is 0.18, indicating a lack of significant correlation between the two lines during that quarter—there is neither a clear positive nor negative correlation. However, in Company B's fourth quarters of 2019 and 2020, the Pearson coefficient reaches 0.96, indicating a clear positive correlation during that period.

5.5 Correlation Discussion between Futures and Taxable Amounts

The data for this study is based on the taxable amounts of two companies engaged in limestone production. These amounts are derived from products made from limestone (cement, quicklime, and hydrated lime), which, to some extent, also reflect correlations. For instance, as the prices of these commodities fluctuate, taxable amounts may fluctuate correspondingly. The data in the following charts have been standardized to provide a more intuitive visual representation of the relationship between taxable amounts and futures.

Fig8 Company A Relationship Between cement and Tax Amount in 2019

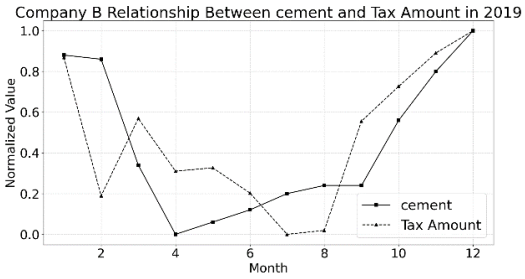

Fig9 Company B Relationship Between cement and Tax Amount in 2019

By comparing the annual taxable amounts and corresponding cement prices for the two companies in 2019, after normalizing the data, the overall trends of these two variables appear to be very similar. Although certain months may experience variations due to uncontrollable factors such as the COVID-19 pandemic, the overall trends are remarkably close.

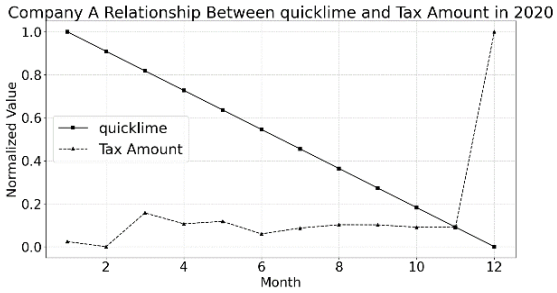

Fig10 Company A Relationship Between quicklime and Tax Amount in 2020

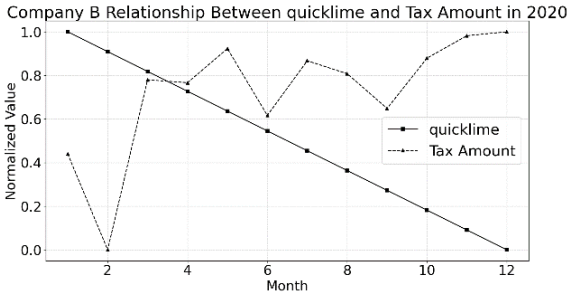

Fig11 Company B Relationship Between quicklime and Tax Amount in 2020

By comparing the annual taxable amounts and corresponding quicklime prices for the two companies in 2020, after normalizing the data, there appears to be a certain degree of negative correlation. This suggests that on the international market, as the price of quicklime, a derivative product of limestone, gradually decreases, there is an increase in the usage of quicklime, leading to an increase in taxable amounts for both companies in this regard.

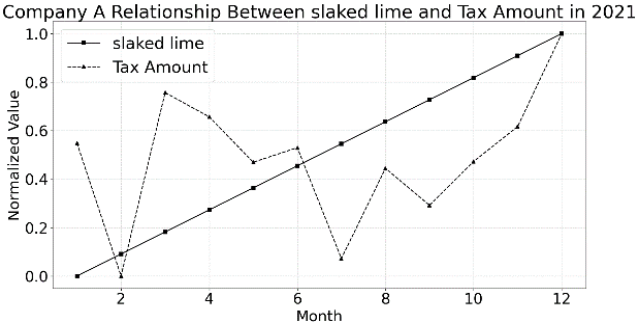

Fig12 Company A Relationship Between slaked lime and Tax Amount in 2021

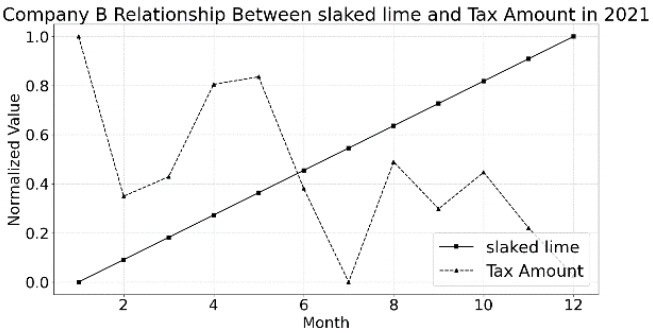

Fig13 Company B Relationship Between slaked lime and Tax Amount in 2021

By comparing the annual taxable amounts and corresponding hydrated lime prices for the two companies in 2021, after normalizing the data, their trends seem to exhibit a certain degree of correlation.

For Company A, as the hydrated lime prices gradually increased month by month, the taxable amounts experienced significant fluctuations in the first half of the year, possibly due to internal or external factors. However, in the second half of the year, the taxable amounts gradually increased as the hydrated lime prices rose.

On the other hand, for Company B, there seems to be a negative correlation with hydrated lime prices throughout the year. This suggests that as the prices rose, the company's usage of limestone, the raw material for producing hydrated lime, decreased.

5.6 Application of the Association Model on Tax Data

For the proposed relationship analysis model algorithm in this paper, which combines the Pearson correlation coefficient with a weighted distribution pattern based on cosine similarity, experimental analysis has been conducted. The analysis is still based on the data for each consecutive two years for Companies A and B, analyzed on a quarterly basis (with λ_1 and λ_2 in Equation 1.1 set as 0.5 each), resulting in Figure 14.

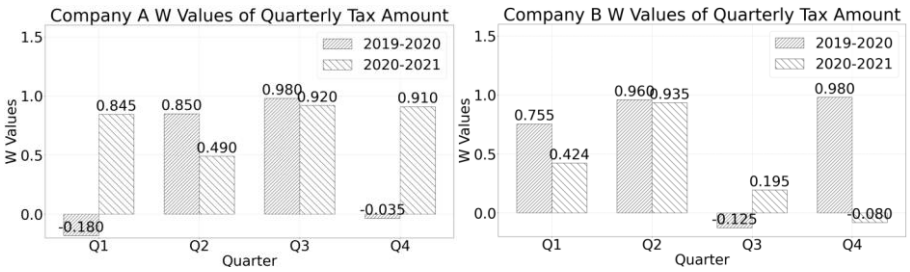

Fig14 W Value of Quarterly Tax Amount by Company A and Company B

Results obtained from the computational analysis are plotted in the form of a bar chart. For example, in the first quarter of 2019 and 2020 for Company A, the correlation relationship is not very obvious, indicating a lack of clear correlation. However, in the second quarter of 2020 and 2021 for Company B, the W value reaches 0.96, indicating a significant positive correlation during that quarter, with the lines showing a close resemblance in direction and trend.

VI. Conclusion

Through multidimensional analysis and comparison of taxable amounts data for Company A and Company B from 2019 to 2021, this study delved into the application and value of various analysis methods in handling real tax data. The research proposed a comprehensive association analysis model that combines cosine similarity and Pearson correlation coefficient, achieving a comprehensive analysis of enterprise tax data.

Initially, this study explored the differences and significance of the data using chi-square analysis and t-statistics, helping us understand the tax performance and fluctuations of the two companies in different quarters. Manhattan distance further revealed the degree of tax data dispersion and differences between quarters, while cosine similarity and Pearson correlation coefficient provided in-depth insights into the directionality and linear relationships within the tax data of the two companies. Additionally, we investigated the correlation between the taxable amounts for both companies and the futures prices of their main products. This analysis demonstrated a connection between taxable amounts and futures prices, reflecting the impact of market price fluctuations on the companies' operational performance and profitability. It also underscored the close relationship between business operations and the broader economic market.

The association analysis model proposed in this paper offers a more flexible and comprehensive analytical tool. It allows us to harness the strengths of various analytical methods and more accurately reveal the inherent connections and patterns of change in corporate tax data. This model offers better precision compared to traditional correlation analysis algorithms and provides a comprehensive assessment from multiple angles, making anomalous data more apparent.

Acknowledgements

We would like to express our heartfelt gratitude to all those who have contributed to this research. Firstly, we would like to thank our supervisor for his/her valuable guidance and insightful comments throughout the entire research process. We also extend our sincere appreciation to all the participants who have generously devoted their time and effort to provide us with valuable data and insights. Finally, we would like to thank our families and friends for their unwavering support and encouragement. This research would not have been possible without their support. Also, we thank the lab funding from “Meteorological Information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes of Chengdu University of Information Technology”; “Integrated Computing and Chip Security Sichuan Collaborative Innovation Center of Chengdu University of Information Technology”. The Open Project of National Intelligent Society Governance Testing Area (NO. ZNZL2023A04). the fund of the Scientific and Technological Activities for Overseas Students of Sichuan Province 2022(30) and funds from the Sichuan Provincial Department of Human Resources and Social Welfare” Researches on Key Issues of Edge Computing Server Deployment and Computing task Offloading”. Thank you for the funding, named "Research on the Construction and Management of Smart Civil Affairs Shared Data System" of Chengdu Civil Affairs Bureau. And Sichuan Science and Technology Program, Soft Science Project (No.2022JDR0076).

References

[1]. Saragih, A. H., Reyhani, Q., Setyowati, M. S., & Hendrawan, A. (2023). The potential of an artificial intelligence (AI) application for the tax administration system’s modernization: the case of Indonesia. Artificial Intelligence and Law, 31(3), 491-514.

[2]. Wu, Q., Zhang, X., & Zhao, B. (2023). A novel adaptive kernel-guided multi-condition abnormal data detection method. Measurement, 206, 112257.

[3]. Li, J., Qureshi, M., Gupta, A., Anderson, S. W., Soto, J., & Li, B. (2019). Quantification of degree of liver fibrosis using fibrosis area fraction based on statistical chi-square analysis of heterogeneity of liver tissue texture on routine ultrasound images. Academic

Radiology, 26(8), 1001-1007.
[4]. Wang, Y. C., Xing, Y., & Zhang, J. (2023). Voronoi treemap in Manhattan distance and Chebyshev distance. Information Visualization, 22(3), 246-264.
[5]. Liu, C., Lin, B., Lai, J., & Miao, D. (2022). An improved decision tree algorithm based on variable precision neighborhood similarity. Information Sciences, 615, 152-166.
[6]. Yang, W., & Yang, J. (2022). Construction of College Physical Education MOOCS Teaching Model Based on Fuzzy Decision Tree Algorithm. Mathematical Problems in Engineering, 2022.
[7]. Wang, Q., & Chen, H. (2020). Optimization of parallel random forest algorithm based on distance weight. Journal of Intelligent & Fuzzy Systems, 39(2), 1951-1963.
[8]. Anam, S., Fitriah, Z., Hidayat, N., & Maulana, M. H. A. A. (2023). Classification Model for Diabetes Mellitus Diagnosis based on K-Means Clustering Algorithm Optimized with Bat Algorithm. International Journal of Advanced Computer Science and Applications, 14(1).
[9]. Xue, X., Huang, S., Xie, J., Ma, J., & Li, N. (2021). Resolvable cluster target tracking based on the DBSCAN clustering algorithm and labeled RFS. IEEE Access, 9, 43364-43377.
[10]. Wang, C., Dong, Y., Xia, Y., Li, G., Martínez, O. S., & Crespo, R. G. (2022). Management and entrepreneurship management mechanism of college students based on support vector machine algorithm. Computational Intelligence, 38(3), 842-854.
[11]. Zhao, Y., & Tian, S. (2021). Identification of hidden disaster causing factors in coal mine based on Naive Bayes algorithm. Journal of Intelligent & Fuzzy Systems, 41(2), 2823-2831.
[12]. Li, Y. (2022). Research on neural network algorithm in artificial intelligence recognition. Sustainable Energy Technologies and Assessments, 53, 102691.
[13]. Bertoli-Barsotti, L. (2023). Equivalent Gini coefficient, not shape parameter!. Scientometrics, 128(1), 867-870.
[14]. Sheluhin, O. I., & Kazhemskiy, M. A. (2020). Influence of fractal dimension on network anomalies binary classification quality using machine learning methods. Automatic Control and Computer Sciences, 54, 216-228.
[15]. Liu, Y., Mu, Y., Chen, K., Li, Y., & Guo, J. (2020). Daily activity feature selection in smart homes based on pearson correlation coefficient. Neural Processing Letters, 51, 1771-1787.
[16]. Chen, J., Guo, Z., & Hu, J. (2021). Ring-regularized cosine similarity learning for fine-grained face verification. Pattern Recognition Letters, 148, 68-74.