Construction of High-quality Economic Development Indicator System Based on Unsupervised Learning

Liwen Tang[1], Liuyang Bian[2], Zhiang Ma[3], Guangxia Zhao[4], Qi Wang[*]

tang-liwen@foxmail.com [1], bianliuyang1@163.com [2], 18055372340@163.com [3,]
zhaoguangxia@mail.ustc.edu.cn [4], wangqi@ipp.ac.cn[*]

Institutes of Physical Science and information technology, Anhui University, Hefei, China [1,2,3,4]
Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China*

Abstract. This research mainly studies the problems currently encountered in the construction of the indicator system for high-quality economic development in China. The indicator system for high-quality economic development in China is not sound enough. Most of the research is based on the relevant concepts of the new development stage and the report of the 19th National Congress of the Communist Party of China. To construct an evaluation indicator system, there is a lack of quantitative method. This paper proposes a method of quantitatively constructing an indicator system based on the indicator systems constructed by other scholars, using spectral clustering combined with the unsupervised learning method of Laplacian score. The results were tested and preliminary research results were obtained. This method can be used to conduct deeper analysis and obtain more instructive conclusions.

Keywords: High-quality economic development; unsupervised learning; spectral clusteri ng; Laplacian.

1 Introduction

The theme of the new development stage is to promote high-quality economic development, and the strategic formulation of the development direction has key practical significance for social development. Since entering the new era, China's economic development attaches great importance to high-quality development, and promoting high-quality economic development is currently the primary goal of China. In order to solve the series of tests of China's social and economic development after building a new socialist power and grasp the new trend given by the key development strategic opportunity period, on May 14, 2020, President Xi Jinping clearly put forward the concept of "domestic and international double-cycle mutual promotion" for the first time in the new development stage[1]. The establishment of a high-quality development indicator system has important guiding significance for evaluating China's overall development status.

In this era, many experts have published many relevant papers on the scientific research of the new development mode of "dual circulation". Constructing an effective indicator system is the main issue in measuring high-quality economic development, and it is also a relatively complex project. If a single indicator value is selected, it will be difficult to reflect the multi-dimensional characteristics of high-quality development. Therefore, a practical indicator

BDEDM 2024, January 12-14, Ningbo, People's Republic of China

DOI 10.4108/eai.12-1-2024.2347186

system is necessary to measure high-quality economic development. In summary, this article intends to explore a new indicator system for constructing statistical measures of the new development stage and provide a new way of thinking for evaluating high-quality economic development in my country.

Total factor productivity is a key definition in the economy and a key tool for analyzing economic development. Total factor productivity reflects the ability and diligence of a country or region to get rid of poverty, backwardness, and develop the economy within a certain stage [2]. Based on the effectiveness and contribution of increasing total factor productivity to economic growth, it can be clarified whether national economic policies should focus on increasing total supply or adjusting industrial structure. Therefore, many scholars use total factor productivity as an indicator to measure high-quality development.With the deepening of research, some scholars believe that the indicator system based on total factor productivity cannot comprehensively measure the level of high-quality development[3]. Therefore, academic and political circles are gradually trying to build a comprehensive indicator system that reflects high-quality development levels, but at this stage, unified standards and norms have not been established. For example, based on the new development concept, Jianhua Cheng and Lijun Zhang[4] constructed 5 first-level indicators and 23 second-level indicators of economic vitality, social stability, people's livelihood, innovation drive, and ecological civilization from a local perspective; Yongwei Su and Chibo Chen [5] Based on the connotation and goals of high-quality development, a high-quality development evaluation indicator system has been established with 6 first-level indicators and 37 second-level indicators, including quality benefit improvement, structural optimization, and kinetic energy conversion.

There are two common problems in related research when constructing an indicator system. First, most scholars construct the evaluation indicator system based on relevant concepts of the new development stage and the report of the 19th National Congress of the Communist Party of China. There is no specific method to screen evaluation indicators and no method to measure the quality of the indicator system. Second, the indicator systems constructed by researchers based on their respective perspectives are very different, making it impossible to make a good horizontal comparison of evaluation indicators. This research refers to 14 articles on the construction and measurement of indicators for high-quality economic development under the new development stage and proposes a method to quantitatively construct an indicator system based on the indicator system constructed by other scholars, using spectral clustering combined with the unsupervised learning method of Laplacian score.

2 Data and Methods

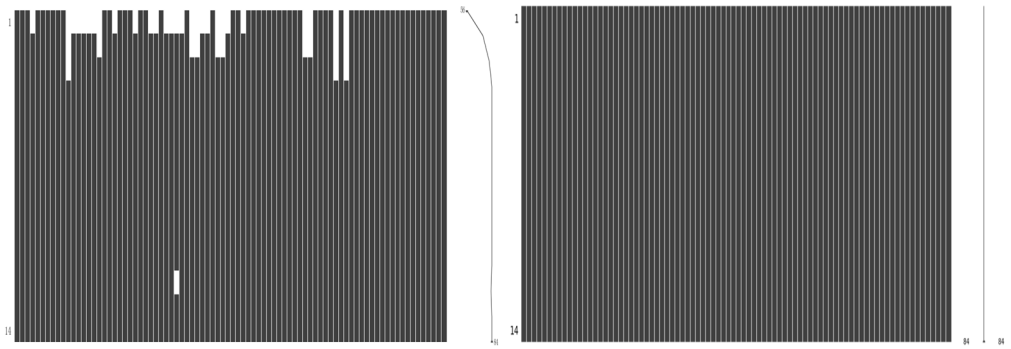
2.1 Data collection

Based on the indicator systems studied by different scholars, they are integrated and summarized as the indicator set of the indicator system of this study [6], which to a certain extent is conducive to ensuring the scientific nature of the indicators. Specifically, it includes 83 economic indicators including the Engel coefficient of urban households from 2008 to 2022, total freight volume, urban population water penetration rate, urban per capita public green space area, and experimental development expenditures of scientific research

institutions. The measurement data of these economic indicators are all from the China Statistical Yearbook[7], which ensures the validity of the data to a certain extent.

2.2 Data preprocessing

When exploring the data, we found that some indicators have missing values through the visualization method of missing values. As shown in the figure1-a, since the indicator data selected in this paper are all continuous variables and have certain relations, we chose the linear regression method to process. [8]. The figure1-b is the result after the missing value is processed.



a)Visualization with missing values b) Visualization with filling values

Fig. 1. Missing value handling visualization.

If the dimensions of each indicator are not unified, it may affect the modeling results, so we use the following formula (1) to standardize the method.

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

2.3 Feature selection

Removing redundant features can better assist in modeling. The strategy used in this study is low-variance filtering [9], which is to measure the variance of all the indicator data information in the indicator set. If the variance is relatively large, it means that the internal changes of the indicators are large and can be used as a classification. If the variance is relatively small, it means that there is no obvious change in the indicator and it is not suitable for classification. When the variance is 0, it means that the indicator has no basic difference in the evaluation indicator set.

2.4 Build Models

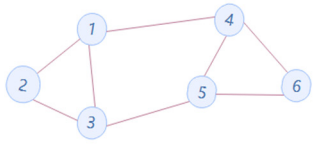
2.4.1 Spectral clustering

Spectral clustering is a widely used clustering optimization algorithm. Compared with the traditional K-MEANS clustering optimization algorithm, spectral clustering has stronger adaptability to data distribution, and the actual effect of clustering is also better. At the same time, compared with other clustering algorithms, spectral clustering requires less computation

and simple algorithm implementation [10]. It treats each sample in the dataset to be clustered as a vertex in the graph. These vertices are connected together, and the edges connected have weights. The size of the weights indicates the degree of similarity between these samples. Vertices of the same class are very similar. In graph theory, the weight of the edges connecting them in the vertices of the same class is very large, and the weight of the edges not connecting them in the vertices of the same class is very small. Therefore, the ultimate goal of spectral clustering is to find a way to cut graphs, so that each subgraph after cutting has a larger weight inside and a smaller weight between different subgraphs. The steps of spectral clustering algorithm in this study can be roughly divided into three stages:

1) Building a Laplace Matrix L for a Dataset

There are generally three methods: ε-neighborhood, k-nearest neighbor method and fully connected layer. The first two methods can construct sparse matrices suitable for large samples, while the third composition method is the opposite, which is more suitable for small samples. Since the index set established in this paper is more inclined to small samples, the third composition method is adopted: fully connected layer. As shown in Figure 2, we process the data through the fully connected layer and construct the Laplace matrix L [11]



PRE-PROCESSING

0	1	2	3	4	5	6
1	3	-1	-1	0	-1	0
2	-1	2	-1	0	0	0
3	-1	-1	3	-1	0	0
4	0	0	-1	3	-1	-1
5	-1	0	0	-1	3	-1
6	0	0	0	-1	-1	2

Fig. 2. Calculate the Laplace matrix L of the data.

2) Matrix factorization

As shown in Figure 3, by calculating the eigenvalue λ of the Laplace matrix L and the eigenvector X, the corresponding eigenvector space is obtained.

DECOMPOSITION

λ =

0.0
1.0
3.0
3.0
4.0
5.0

X =

0.4	0.3	-0.5	-0.2	-0.4	-0.5
0.4	0.6	0.4	-0.4	0.4	0.0
0.4	0.3	0.1	0.6	-0.4	0.5
0.4	-0.3	0.1	0.6	0.4	-0.5
0.4	-0.3	-0.5	-0.2	0.4	-0.5
0.4	-0.3	0.4	-0.4	-0.4	0.0

Fig. 3. Eigenvector Space.

3) Clustering

Take the K-MEANS method to cluster the feature vectors in the figure above, and the results are shown in Figure 4. The features of the nodes are represented and sorted and split.

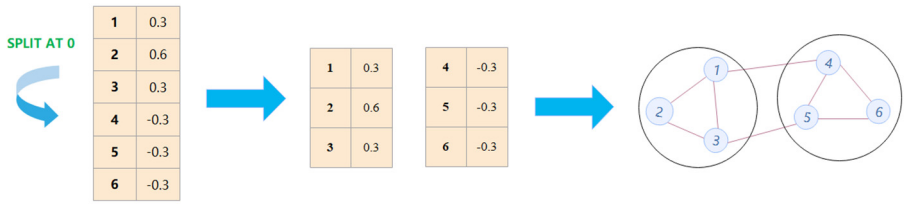


Fig. 4. Clustering flow chart.

2.4.2 T-SNE

In order to see the clustering effect more intuitively, this paper adopts the T-SNE method for data lake visualization analysis. T-SNE is a visual analysis algorithm, which is often applied to datasets in 2D or 3D space. By reducing the dimensionality of high-dimensional data, its feature space is reduced, and most of the initial information of the data is finally retained [12].

2.4.3 Metrics

Clustering metrics can be briefly divided into two categories: external metrics and internal metrics. Among them, external metrics refer to the use of real categories of data for evaluation, and in practice, expert knowledge can be used to assist in screening; internal metrics refer to the use of sample points and cluster centers to judge the merits of clustering without the assistance of external models. Since the indicator system construction method studied in this paper does not have real categories, this paper mainly considers using internal metrics as evaluation functions, using the following three methods:

a) Silhouette coefficient[13]: The value of the profile coefficient is between [-1, 1]. The larger the value, the closer the similar samples are, the farther apart the different samples are, and the better the clustering effect.

b) Calinski-Harabaz Index(CH Index)[14]: The CH Index must first calculate the degree of separation and compactness. The CH index is obtained by the ratio of the degree of separation and compactness. The larger the ratio, the better the clustering effect.

c) Davidson Bodine Index (DBI)[15]: The minimum value of DBI is 0, and the smaller the value, the better the clustering effect.

2.4.4 Preliminary construct indicator system

After obtaining the classification situation through clustering, in order to make the measurement more scientific and concise, it is necessary to select excellent indicators to construct the indicator system. This paper adopts the Laplace scoring algorithm[16] as a strategy to evaluate whether the index is excellent. The essence of the Laplace scoring algorithm is a similarity-based method, which seeks to maximize the dispersion of the index data while maintaining the ability of similarity combined with the local characteristics of the data.

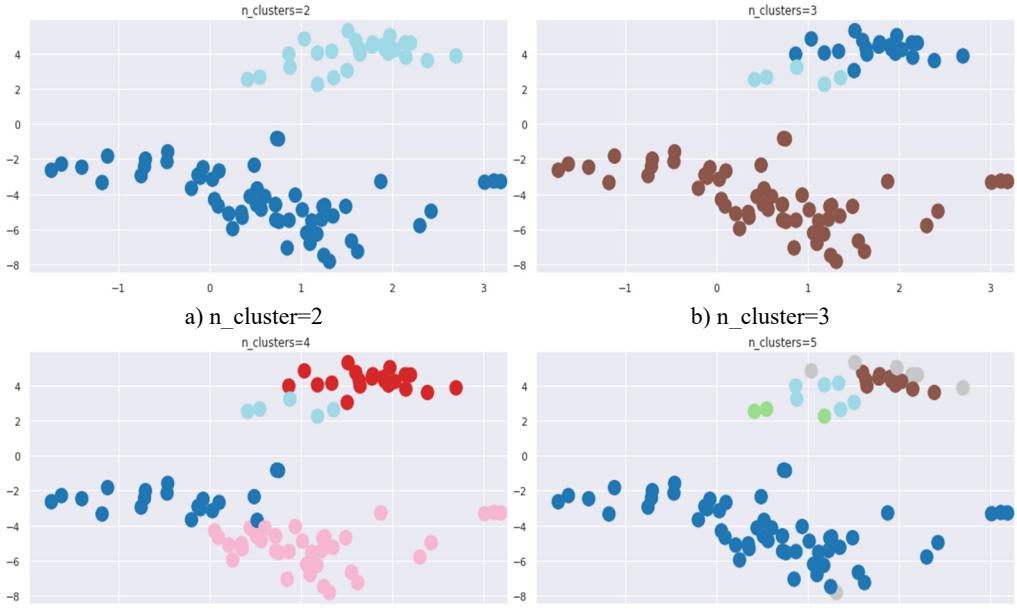
2.4.5 Construct indicator system and calculate comprehensive indicator by weighting

Select the top half of the Laplace scores in each category to form a new indicator system, and then use the new indicator system to calculate the Laplace score to weight each indicator to calculate the economic comprehensive indicator. And use the comprehensive economic indicator to test for China's economic development from 2008 to 2022 to see if the system can more objectively reflect the trend of China's economy.

3 Results and Discussions

3.1 Clustering result visualization

The (a)(b)(c)(d) subgraphs in Figure 5 respectively represent the clustering renderings when the T-SNE visualization method is used to take 2, 3, 4, and 5 of the clustering categories n, respectively. It can be more intuitively observed that samples of different distances are clustered in different areas. The overall observation is still not very good to determine which n is better, but the visualization method provides us with a rough range of n.



a) n_cluster=2 b) n_cluster=3

c) n_cluster=4 d) n_cluster=5

Fig. 5. Clustering flow chart

3.2 Spectral clustering results

Table 1 gives the index situation of taking different number of clusters after spectral clustering. It can be seen that the contour coefficient and CH score are the largest when n_cluster = 2, and the scores are 0.5988 and 132.7179 respectively. The Davidson Bodine index has the smallest score when n_cluster = 2. According to the principle that the contour coefficient and CH score

are larger, the smaller the DBI, the better the metric. It can be seen that the clustering effect is best when n_cluster = 2. Based on the clustering effect diagram described in Section 3.1, this paper decides to cluster the samples into two categories.

Table 1. clustering different metrics results.

n_cluster	Silhouette coefficient	CH Index	DBI
2	0.5988	132.7179	0.6659
3	0.5662	94.9708	0.9264
4	0.5489	67.6827	0.8366
5	0.3060	61.4235	1.0893

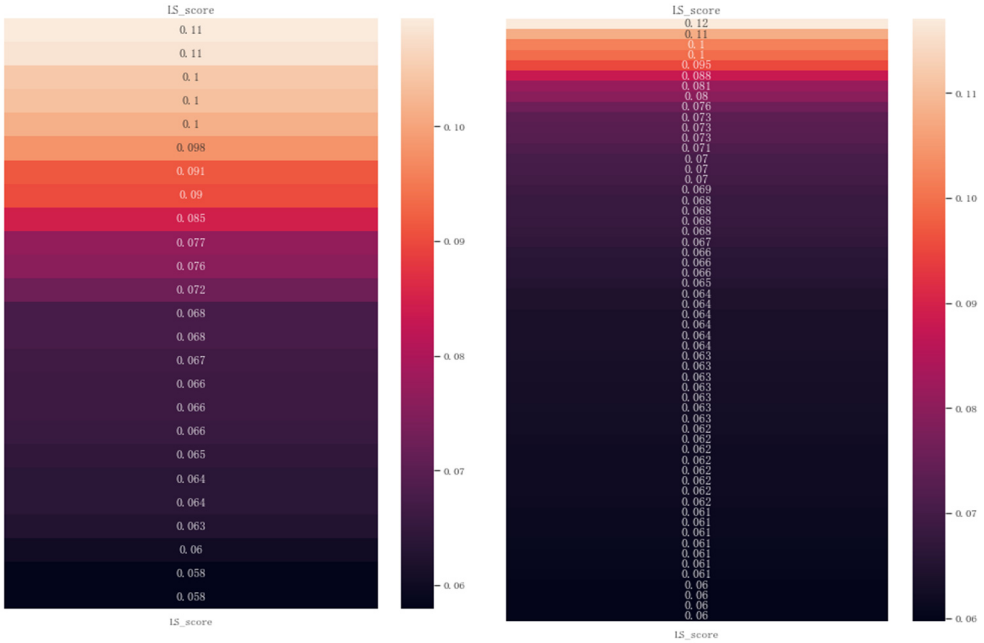
Table 2 shows the indicatos for different categories after clustering

Table 2. indicator system after clustering.

catagory	indicator
0	● Average annual consumption expenditure per person in urban households ● Urban household average per capita disposable income ● There are public transportation vehicles per 10,000 people in cities ● The number of domestic rural residents visiting ● ….
1	● Growth rate of national fiscal final account expenditure ● State fiscal final accounts revenue _ current year-on-year growth rate _ ● The number of rural residents with minimum living security ● Minimum living security for urban residents ● …

3.3 Preliminary construct indicator system results

After processing the two types of indicator subsets obtained by spectral clustering by the Laplace scoring algorithm, the score heat map of the indicator subset shown in Figure 6 is obtained, where a represents the score heat map of category 0, and b represents the score heat map of category 1. The lighter the color, the higher the score.



a) category 0 indicators heatmap b) category 1 indicators heatmap

Fig. 6. Different categories score heat map.

3.4 The results of constructing indicator system and calculating comprehensive indicator by weighting

Figure 7 shows the weighted results of the final selected indicators, among which the higher weight indicators are Final Consumption Expenditure Contribution Rate, Nature reserve area, The contribution rate of the primary industry and other indicators among which the highest weight exceeds 0.125.

In order to test the applicability of the high-quality economic development indicator system constructed in this paper, we used the data from 2008 to 2022 to substitute the comprehensive economic indicator value calculated by the system, and used the result to evaluate China's overall economy. The results are shown in Figure8 below.

It can be seen from the figure8 that China's economic development overall showed a steady upward trend from 2008 to 2022, rising from 1.360977 in 2008 to 2.425635 in 2022. The overall trend is consistent with the calculations of scholars such as Bo Shi [17], Jun Han [18], Min Wei [19]. Further analysis shows that the comprehensive score showed a short-term downward trend in 2018 and 2020 respectively, and it was learned through the international form that in January 2018, the United States and China started a new round of Sino-US trade war, which had a certain impact on China's overall economic situation in 2018. However, the Chinese government achieved the actual effect of economic recovery in 2019 by rapidly adjusting and taking countermeasures such as transforming the industrial layout. This is in line with the increase indicator in 2019 on the figure. In early 2020, due to the impact of the COVID-19 epidemic, China's overall economic development declined in 2020. After 2021 to

2022, the impact of COVID-19 on China's economy has been effectively alleviated. The above situation is consistent with the manifestation on the figure8. Therefore, the new indicator system constructed in this study has certain applicability.

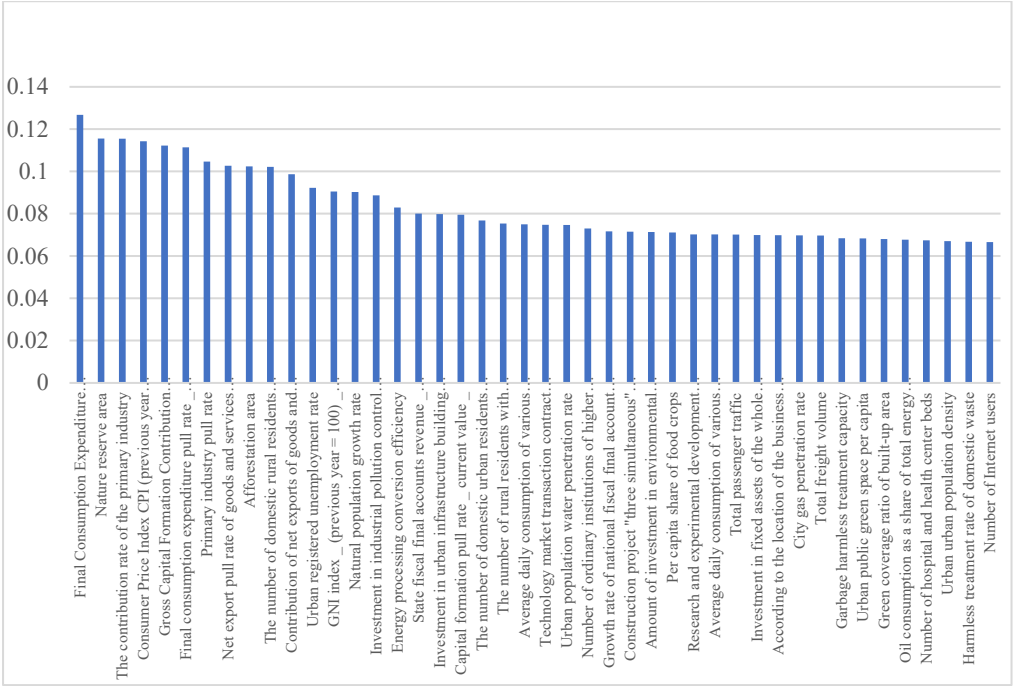


Fig. 7. Weight of each indicator.

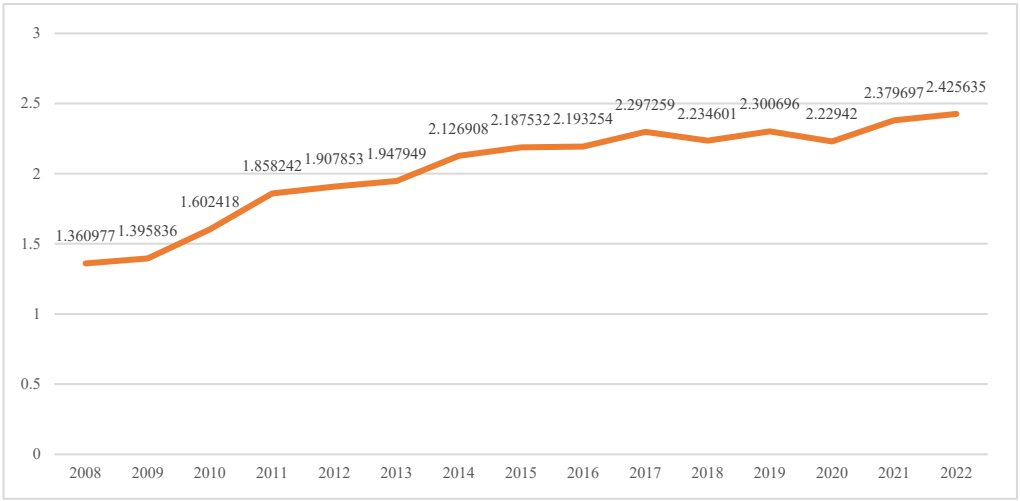


Fig. 8. The comprehensive economic indicator from 2008 to 2022.

4 Conclusions

This paper focuses on the establishment of an indicator system for high-quality economic development under the new development stage. By learning from the indicator system established by various scholars to form an indicator set, combined with the availability of data, a total of 83 indicators from 2008 to 2022 were collected from the "China Statistical Yearbook", and the spectral clustering algorithm was selected to cluster the indicators. After that, the T-SNE method was used to visualize the clustering effect, and then the contour coefficient, CH score and Davidson-Boulding index were used to select the classification of the best clustering effect. Then, the Laplace score algorithm was used to evaluate each subset of indicators separately, and the best index was selected as the new indicator system. After that, a comprehensive indicator of China's economy was weighted and calculated. By comparing the research of other scholars with the international situation, the system made a accurate evaluation of China's economic development from 2008 to 2022. Later, this method can be used to conduct regional statistical analysis of different provinces or regions, and the indicator system calculated by the model can be classified by field to improve interpretability.

References

[1] Xi, J. (2021). Grasp the new stage of development, implement the new development concept, and build a new development pattern. Contemporary Communist Party Members, (10), 3-9.

[2] Liu, Z. (2019). On China's economic high-quality development from the perspective of total factor productivity. Guide to Economic Research, (32), 5-7+13.

[3] Li, Q. (2021). Construction and measurement of economic high-quality development evaluation index system. Statistics & Decision, 37(15).

[4] Cheng, J. & Zhang, L. (2022). Research on evaluation, influencing factors and trend forecast of high-quality economic development in Anhui Province. Journal of Bengbu University, 11(03).

[5] Su, Y. & Chen, C. (2019). Construction and empirical analysis of economic high-quality development evaluation index system. Statistics & Decision, 35(24).

[6] Su, L. & Ma, X. (2022). Construction of economic high-quality development evaluation index system. Statistics & Decision, 38(02).

[7] National Bureau of Statistics of China. (2022). China statistical yearbook 2022. Retrieved from https://www.stats.gov.cn/sj/ndsj/

[8] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594-604), 309-368.

[9] T. Hastie, R. Tibshirani and J. Friedman, “Elements of Statistical Learning”, Springer, 2009.

[10] Damle, A., Minden, V., & Ying, L. (2019). Simple, direct, and efficient multi-way spectral clustering. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1486-1495).

[11] Chung, F. R. K. (1997). Spectral graph theory. American Mathematical Soc.

[12] L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms.Journal of Machine Learning Research 15(Oct):3221-3245, 2014.

[13] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[14] Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-Simulation and Computation, 3(1), 1-27.
[15] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.
[16] Stigler, S. M. (1986). Laplace's 1774 memoir on inverse probability. Statistical Science, 1(3), 359-363.
[17] Shi, B. & Zhang, B. (2019). Measurement and analysis of high-quality economic development in prefecture-level cities in China. Social Science Research, (03), 19-27.
[18] Han, J. & Zhang, H. (2019). Measurement of regional energy consumption in China under the background of high-quality economic development. Technoeconomics & Management Research, 36(07).
[19] Wei, M. & Li, S. (2018). Measurement research on the level of high-quality economic development in China in the new era. Technoeconomics & Management Research, 35(11).