

Divide China's Economic Regions in 2019 Based on Cluster Analysis and Principal Component Analysis

Zhichao Zhan^{*}, Yongquan Jin, Meihua Dong

Mathematics Department, Yanbian University, Yanji, P. R. China

Email address:

zzc010324@163.com (Zhichao Zhan), 976684141@qq.com (Yongquan Jin), sxdmh@ybu.edu.cn (Meihua Dong)

^{*}Corresponding author

To cite this article:

Zhichao Zhan, Yongquan Jin, Meihua Dong. Divide China's Economic Regions in 2019 Based on Cluster Analysis and Principal Component Analysis. *International Journal of Statistical Distributions and Applications*. Vol. 7, No. 4, 2021, pp. 83-88. doi: 10.11648/j.ijstd.20210704.11

Received: July 5, 2021; **Accepted:** July 26, 2021; **Published:** November 5, 2021

Abstract: In recent years, China's economic development has been very rapid. While China is developing rapidly, each province has contributed its share, but in different regions, economic development is different. Different regions must have advantages in different aspects, so in order to divide China's 31 provinces into different categories. In order to get the ranking of the provinces that have the greatest impact on China's economy. We first adopt the method of principal component analysis to reduce the dimensions of 11 variables that affect the economic factors of each province, and obtain two principal components to reflect all sample information. Then, perform dimensionality reduction and cluster analysis on the obtained data, and use the sum of squared variance (WARD) method to perform cluster analysis on the two principal components. Finally, the social development of 31 provinces in my country is divided into 4 categories. It is concluded that Beijing and Shanghai are first-level developed provinces, Jiangsu and Guangdong are second-level developed provinces, Hebei, Sichuan, Hunan, Shandong, Henan, Shanxi, and Hubei are third-level developed provinces, Tianjin, Hainan, Tibet, Qinghai, Ningxia, Inner Mongolia, Jilin, Gansu, Xinjiang, Fujian, Chongqing, Liaoning, Anhui, Shaanxi, Jiangxi, Guizhou, Yunnan, Heilongjiang, and Guangxi are four-tier developed provinces. I hope our results can help relevant departments.

Keywords: Principal Component Analysis, Cluster Analysis, Economy, Squared Deviations Method

1. Introduction

With the rapid economic development in China today, there are certain economic differences among various provinces in China, and each province has its own special economic source.

In order to better study the differences between provinces, we selected the primary industry (referring to agriculture), secondary industry (referring to manual manufacturing), tertiary industry (modern service industry and commerce), agriculture, forestry and animal husbandry, The total value-added of the five economic benefits of the construction industry in 2019 (unit: 100 million yuan). Since economic growth is also closely related to the flow of people and the permanent population between provinces, we have studied each The province's permanent population at the end of 2019, urban registered unemployed population, and passenger traffic will affect the province's economy (unit: ten thousand people). At the same time, people's pursuit of scenery and

attitudes towards the people's growing material and cultural needs The yearning for food is also a major factor. Many provinces also rely on tourism and catering to increase their province's fiscal revenue. Therefore, we have considered the passenger turnover of each province (unit: 100 million person-kilometers) and the headquarters of chain catering companies. Number (unit: number). In the fiscal revenue of each province, many staff will pay a certain tax, so we are considering the economic status of each province. We can use local fiscal tax revenue (100 million yuan) to represent the total income level of the staff in each province. In order to comprehensively study the differences between provinces, we use variance analysis, correlation analysis and principal component analysis to study the relationship between the economy of each province and the eleven factors, and then use two principal components to compare the eleven factors All the information is extracted, and then these two principal components are analyzed, and the provinces are divided into different regions through cluster analysis. After division, we

can get that the 31 provinces across After analyzing the obtained variables, we can think that only the first two principal components can reflect the information of all variables. At this time, we study the first two principal components and obtain the coefficients and distributions of the first two principal components as follows: the country can be divided into 4 categories.

2. Algorithm Introduction

2.1. Introduction to Principles of Principal Component Analysis

Principal component analysis (PCA) is a statistical method that expresses the idea of dimensionality reduction. It uses orthogonal transformation to transform linearly related values into a set of unrelated values through the analysis of multiple variables. Without losing all the information of the sample, the dimensionality of the variable is reduced. Principal component analysis uses the sample data matrix to obtain the eigenvalues of the sample data matrix when analyzing p variables. Through the eigenvalues, we analyze the contribution rate of each principal component [1]:

$$\rho_j = \sum_{i=1}^p \frac{\lambda_j}{\lambda_i}$$

$$\chi_j = \sum_{i=1}^j \rho_j$$

$$\lambda_1 > \lambda_2 > \dots > \lambda_p$$

Where ρ_j represents the contribution of the j principal component, χ_j represents the cumulative contribution of the j principal component, and λ_j represents the eigenvalue of the j principal component. Through the expression, we can see that the eigenvalue of the principal component is monotonically decreasing Trend, so you can get the trend of decreasing contribution. Generally speaking, when the cumulative contribution reaches more than 85%, we can consider it to contain all the sample information.

In principal component analysis [5], in order to make each variable have a connection with the final result we get, we put forward the theory of contribution rate. The contribution rate is the proportion of each variable coefficient in each principal component equation, which is expressed as follows:

$$\eta_j = \sum_{i=1}^p \frac{\alpha_j}{\alpha_i}$$

$$\gamma_j = \sum_{i=1}^j \eta_j$$

Among them, η_j is the contribution rate of the j variable, and γ_j is the cumulative contribution rate of the j variable. In the analysis, we try not to make the cumulative contribution rate of the variable γ_j close to zero [6].

The purpose of principal component analysis is to simplify the data structure and replace the original variables with as few principal components as possible $Z_1 \dots Z_m (m < p)$ In this way, we get the purpose of dimensionality reduction, and

the information contained in the data is basically the same as the original information, and the obtained principal component part can explain the meaning of the data.

2.2. Cluster Analysis

The Cluster analysis is to group data objects according to the information between the objects described in the data and their relationships. It can classify objects with the same attributes into one class, while objects with very different attributes are assigned to different classes. That is to say, objects with the same attributes in a group will be clustered together, while objects with different attributes will be grouped together. In another group. According to different classification objects, cluster analysis can be divided into two categories. One is sample clustering, also known as Q-type clustering [2], which is equivalent to classifying the observation matrix by rows. One type is variable clustering, also known as R-type clustering [3], which is equivalent to classifying the observation number rectangle by column. In this analysis, we are practically Q-type classification. At the same time, in the process of cluster analysis [4], we have many ways to judge the distance between two samples. Since the sample value is usually multi-dimensional, we generally use the Mahalanobis distance to analyze the Mahalanobis distance of the samples X_i and X_j :

$$d_{ij}(M) = (X_i - X_j)'S^{-1}(X_i - X_j)$$

We can cluster the samples through Mahalanobis distance to determine the relationship between the samples. There are 8 classification methods in cluster analysis. In this analysis, we use the sum of squared deviation method (WARD) [6]. The main difference between the classification methods is that the distance between the samples is different, so the results obtained are slightly different.

3. Summary of Results

3.1. Use Principal Component Analysis to Process Data

First, we process part of the data. For the convenience of research, we write the variables as $\chi_1 \dots \chi_{11}$. Input the data into SAS statistical software [2] and analyze the data. The analysis results are as follows:

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	7.47040154	5.37186950	0.6791	0.6791
2	2.09853203	1.65881449	0.1908	0.8699
3	0.43971754	0.08128779	0.0400	0.9099
4	0.35842975	0.07150579	0.0326	0.9425
5	0.28692395	0.06840489	0.0281	0.9685
6	0.21851906	0.15647472	0.0199	0.9884
7	0.06204434	0.02554317	0.0056	0.9941
8	0.03650117	0.01742349	0.0033	0.9974
9	0.01907768	0.01272594	0.0017	0.9991
10	0.00635174	0.00285053	0.0006	0.9997
11	0.00350121		0.0003	1.0000

Figure 1. Eigenvalues.

From the figure below, we can see that $\lambda_1 \approx 7.47, \lambda_2 \approx$

2.10, $\lambda_3 \approx 0.44$, we can see that $\lambda_2 \gg \lambda_3$, at this time, the cumulative contribution rate is 86.99%, which is already satisfied to determine the value of m , and the contribution of the principal component Z_3 As $0.04 \approx 0$, we can think that

the first two principal components can well reflect the information of all variables. Next, we test the cumulative contribution of the principal components [7]. We can get the coefficient of the principal component as:

The PRINCOMP Procedure

Eigenvectors

		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11
x1	x1	0.299388	-0.355924	0.049370	-0.028525	0.380542	0.276951	-0.174617	0.033286	0.107784	0.512929	-0.502600
x2	x2	0.344620	0.105537	-0.270351	-0.239389	0.065688	-0.317202	-0.050295	0.476466	-0.613352	0.155762	0.068977
x3	x3	0.328441	0.288348	0.009837	-0.160161	0.130621	-0.085965	-0.098705	-0.027950	0.108755	-0.652623	-0.557237
x4	x4	0.273544	-0.270638	0.494865	-0.335573	-0.684672	0.096297	0.023941	0.108031	-0.029201	-0.006311	-0.056944
x5	x5	0.275754	0.414233	0.217131	-0.168416	0.064071	-0.331101	-0.407035	-0.084669	0.452918	0.300747	0.307731
x6	x6	0.345244	-0.158443	0.186874	0.070053	0.165740	-0.205544	0.136612	-0.741792	-0.396829	-0.023721	0.138566
x7	x7	0.294955	-0.366724	0.087880	-0.053042	0.383419	0.268975	-0.031571	0.267969	0.156543	-0.399631	0.540005
x8	x8	0.304043	0.178707	-0.569784	-0.357827	-0.178219	0.380419	0.344399	-0.244780	0.200939	0.115746	0.118935
x9	x9	0.321683	0.019485	-0.265180	0.610798	-0.361761	0.226142	-0.503436	-0.038301	-0.092871	-0.067314	0.051574
x10	x10	0.333682	-0.125955	-0.039551	0.453311	-0.083065	-0.421744	0.552979	0.214586	0.345545	0.066418	-0.072960
x11	x11	0.139293	0.571341	0.439494	0.246836	0.130069	0.456145	0.305749	0.150853	-0.209725	0.116744	0.035297

Figure 2. Coefficient graph.

We calculate the contribution of the first two principal component coefficients in all principal components through software calculations, and get the following results [8].

Table 1. Contribution rate.

variable	χ_1	χ_2	χ_3	χ_4	χ_5		
contribution	0.247	0.170	0.191	0.231	0.232		
variable	χ_6	χ_7	χ_8	χ_9	χ_{10}	χ_{11}	
contribution	0.194	0.231	0.167	0.135	0.168	0.250	

Through the cumulative contribution rate of each variable in the first two principal components, we can see that the contribution rate of all variables exceeds 13%, so we can think that the data we get is valid [9].

After analyzing the obtained variables, we can think that

only the first two principal components can reflect the information of all variables. At this time, we study the first two principal components and obtain the coefficients and distributions of the first two principal components as follows [10]:

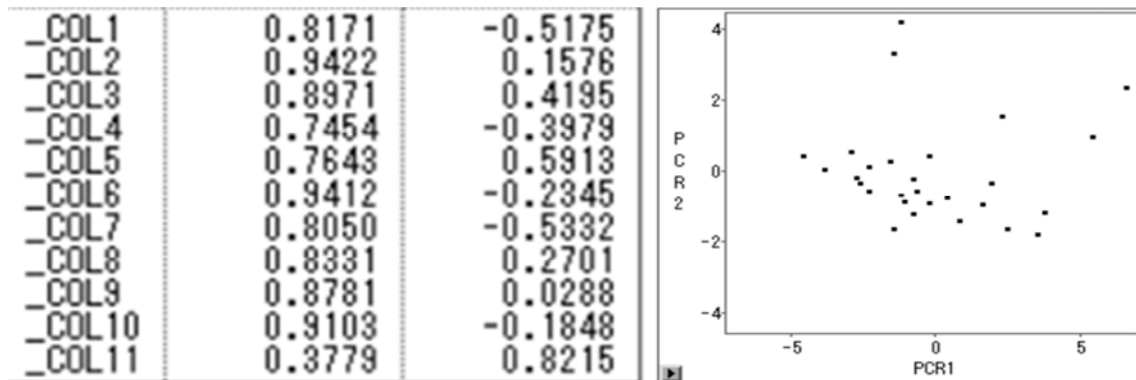


Figure 3. Coefficient and distribution map.

Through software calculation, we get the first two principal component expressions as:

$$\begin{cases} F_1 = 0.81X_1 + 0.94X_2 + 0.89X_3 + 0.74X_4 + 0.76X_5 + 0.94X_6 + 0.80X_7 + 0.83X_8 + 0.87X_9 + 0.91X_{10} + 0.38X_{11} \\ F_2 = -0.51X_1 + 0.15X_2 + 0.41X_3 - 0.39X_4 + 0.59X_5 - 0.23X_6 - 0.53X_7 + 0.27X_8 + 0.02X_9 - 0.18X_{10} + 0.82X_{11} \end{cases} \quad (3)$$

For the first principal component, except for X_{11} (the total number of chain restaurants), the proportions of all variables are above 0.7, so we can think that the first principal component mainly reflects the information of the first ten variables. The proportion of the X_{11} variable of the second principal component is 0.82, so the second principal

component is mainly explained by the variable X_{11} . Therefore, we believe that the first two principal components can well reflect all the information of the sample data [11], and we can also perform dimensionality reduction processing on the sample data, and then we can classify the data after processing.

3.2. Cluster Analysis Process

First, we input the data obtained after principal component

dimensionality reduction into statistical software, analyze the data, and obtain the following results after analysis:

Eigenvalues of the Correlation Matrix							
	Eigenvalue	Difference	Proportion	Cumulative			
1	1.01235721	0.02471443	0.5062	0.5062			
2	0.98764279		0.4938	1.0000			
The data have been standardized to mean 0 and variance 1							
Root-Mean-Square Total-Sample Standard Deviation				1			
Root-Mean-Square Distance Between Observations				2			
Cluster History							
NCL	--Clusters Joined--		FREQ	SPRSQ	RSQ	PSF	PST2
30	OB29	OB30	2	0.0000	1.00	5653	.
29	OB5	OB7	2	0.0001	1.00	1000	.
28	OB28	OB31	2	0.0001	1.00	657	.
27	OB24	OB25	2	0.0002	1.00	444	.
26	OB26	CL30	3	0.0002	.999	363	33.4
25	OB6	OB12	2	0.0005	.999	233	.
24	OB14	OB27	2	0.0006	.998	186	.
23	CL29	CL28	4	0.0009	.997	144	10.9
22	OB4	OB17	2	0.0011	.996	118	.
21	OB8	OB20	2	0.0011	.995	104	.
20	OB21	CL26	4	0.0012	.994	95.9	11.7
19	OB3	OB23	2	0.0014	.993	88.9	.
18	CL24	CL27	4	0.0017	.991	82.8	4.6
17	OB15	OB16	2	0.0020	.989	77.7	.
16	CL19	OB18	3	0.0020	.987	75.1	1.4
15	OB13	OB22	2	0.0020	.985	74.3	.
14	CL25	CL18	6	0.0035	.981	68.9	4.7
13	OB2	CL20	5	0.0036	.978	66.0	7.5
12	OB1	OB9	2	0.0038	.974	64.8	.
11	CL14	CL21	8	0.0081	.966	56.7	6.4
10	OB10	OB19	2	0.0111	.955	49.4	.
9	CL16	CL17	5	0.0131	.942	44.5	7.2
8	CL23	CL15	6	0.0147	.927	41.8	19.3
7	CL22	OB11	3	0.0158	.911	41.1	14.3
6	CL8	CL11	14	0.0266	.885	38.4	9.5
5	CL9	CL7	8	0.0544	.830	31.8	9.2
4	CL13	CL6	19	0.0828	.747	26.6	21.6
3	CL5	CL10	10	0.1306	.617	22.5	10.4
2	CL12	CL4	21	0.2732	.344	15.2	34.2
1	CL2	CL3	31	0.3436	.000	.	15.2

Figure 4. Separation square sum method.

Among them, NCL represents the number of categories, and represents the total number of categories after the formation of the new category. It can be seen that the cluster analysis finally divides all the samples into one category. Clusters joined indicates the merged category, which indicates which two categories are merged [12]. FREQ indicates how many samples are contained in the class obtained by this merger. PSF represents the pseudo F statistic, PST2 is the pseudo t^2 statistic, and Tie indicates whether there are more pairs of candidate classes with the smallest distance.

Analyzing the results of the three methods [13], we can see that all PSF, PST2 and Tie are the same, and Tie is not in this analysis. PSF statistics are used to evaluate the effect of clustering into G. If the clustering is the effect is relatively good. The sum of squared deviations between classes is

larger than the sum of squared deviations within classes. Through analysis, we can see that it has a gradually increasing trend, so it is difficult to judge the number of classifications. At this time, we can consider the PST2 statistics. The PST2 statistics are used to evaluate the effect of the combination of this step. That is to say, when the value of PST2 is large, it means that the combination of the two classes is very good. In this analysis, we can see that when the PST2 [14] statistics are the largest and the second largest, NCL=3 and 27. At this time, we can think that it is appropriate to divide into 4 categories or 28 categories based on the PST2 statistics, because the number of categories generally does not exceed 6, so we can choose to be divided into 4 categories more appropriate.

3.3. Result

Through the processing of data and the judgment of the

analysis results, we divided the country into 4 regions [3], and then we obtained the pedigree diagrams of the respective clustering methods through statistical software:

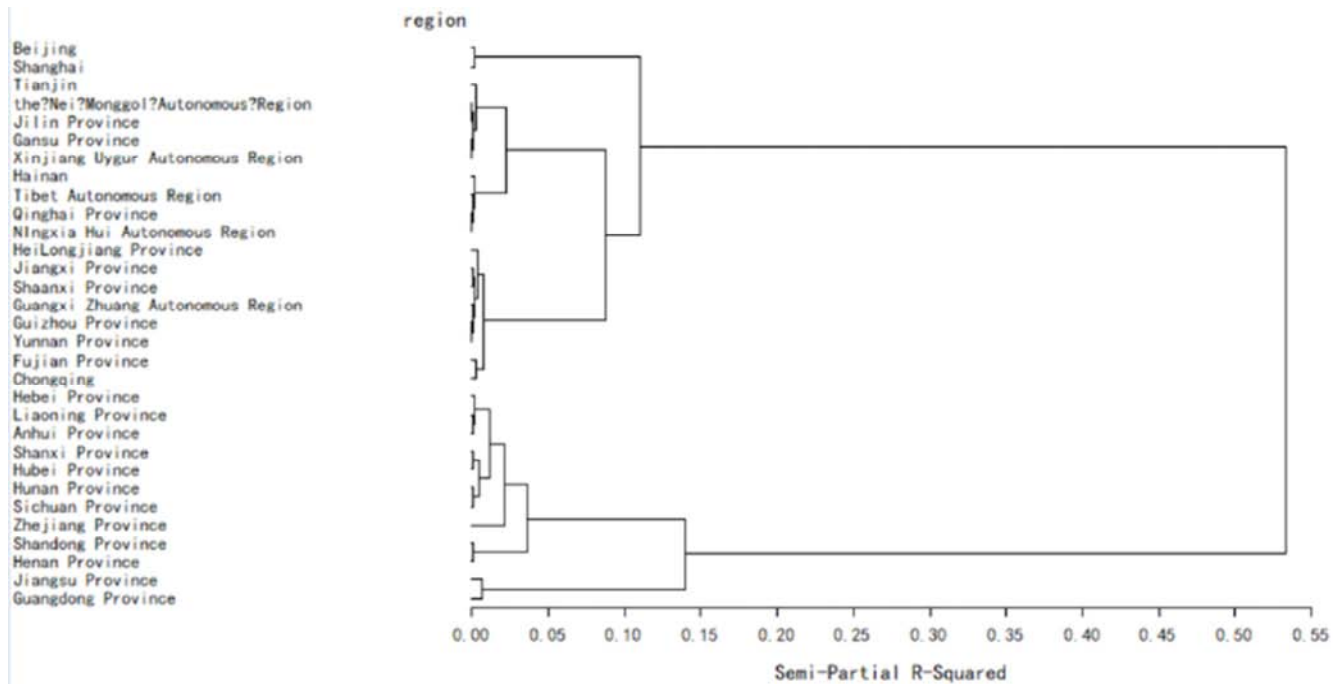


Figure 5. WARD method.

By analyzing the pedigree diagram, we can see the results of the classification. The WARD method is divided into categories (for ease of writing, only the first two characters of each province, except Heilongjiang) are $G_1 = \{\text{Beijing, Shanghai}\}$, $G_2 = \{\text{Jiangsu, Guangdong}\}$, $G_3 = \{\text{Hebei, Sichuan, Hunan, Shandong, Henan, Shanxi, Hubei, Zhejiang}\}$, $G_4 = \{\text{Tianjin, Hainan, Tibet, Qinghai, Ningxia, Inner Mongolia, Jilin, Gansu, Xinjiang, Fujian, Chongqing, Liaoning, Anhui, Shaanxi, Jiangxi, Guizhou, Yunnan, Heilongjiang, Guangxi}\}$. We can see that Beijing and Shanghai belong to the same category. It is not difficult to understand that they are the two provinces with the strongest economic capacity in China [15]. They rank among the top two in China's GDP list every year. They play a vital role in China's economic development. Similarly, Beijing and Shanghai are both Chinese politics. The center of economy, science, culture and international exchanges, the comprehensive strength development of these two provinces is obviously ranked at the highest level in China. Jiangsu and Guangdong are in the second category. Jiangsu and Guangzhou are both coastal provinces. Jiangsu is dominated by heavy industry development and has many foreign-funded enterprises. Foreign-funded enterprises have brought a lot of GDP to Jiangsu. Guangdong's maritime trade is smooth, and it has developed well in both light and heavy industries. Therefore, the comprehensive level of these two provinces is relatively high. The third category is Hebei, Hebei, Sichuan, Hunan, Shandong, Henan, Shanxi, Hubei, Zhejiang. The fourth category is Tianjin, Hainan, Tibet, Qinghai, Ningxia, Inner Mongolia, Jilin, Gansu, Xinjiang, Fujian, Chongqing, Liaoning, Anhui, Shaanxi, Jiangxi, Guizhou, Yunnan, Heilongjiang, Guangxi. These provinces are mainly distributed in inland areas,

mainly based on agriculture, so it is divided into the fourth category [16].

4. Conclusion

The combination of principal component analysis method and clustering method reduces the workload to a certain extent. Through the principal component analysis method, we integrate the information of multiple indicators into two indicators, and then use the two indicators to perform cluster analysis, which greatly reduces the workload of cluster analysis. However, this method also has shortcomings. First, when we use principal component analysis to analyze the original sample data array, the information extraction is not complete, and important information may be missed, and the samples are standardized, which reduces the number of samples. The difference between the different indicators does not well reflect that different indicators have different effects on the overall. Second, in the process of using cluster analysis, we must first know the number of classifications. In this case, classify all samples first, and then determine the number of classifications. If the number of classifications is unreasonable, we need to recalculate at this time.

References

- [1] Data Sources: <http://www.stats.gov.cn/>.
- [2] Edited by Gao Huixuan. Applied Multivariate Statistical Analysis. Beijing: Peking University Press. 2005.01.

- [3] SAS Software and Statistics Application Course/Wang Yuanzheng, Editor-in-Chief Xu Yajing. Beijing: Machinery Industry Press, 2007.1.
- [4] Li Na. Evaluation of the coordinated development of China's regional economy based on AHP cluster analysis [J]. Microcomputer Application, 2021, 37 (04): 151-153.
- [5] Chen Jiaqi, Xiang Guangxin. Research on the Comprehensive Evaluation of Rural Regional Economic Development Differences in Hunan Province Based on Principal Component Analysis [J]. China Economic and Trade Guide (Secondary), 2021 (08): 71-74.
- [6] Wu C X. Research on urban residents' income based on principal component analysis and cluster analysis [J]. Journal of Huangshan University, 201, 23 (03): 7-10.
- [7] Meng Q. Comprehensive evaluation of economic development quality in Anhui Province based on principal component analysis [J]. Journal of Xichang University (Natural Science Edition), 201, 35 (02): 43-48.
- [8] Ji Xionghua, Bai Zongming, Song Zhufang. Evaluation of government economic governance capability based on Principal Component Analysis [J]. Journal of Yan'an Vocational and Technical College, 201, 35 (03): 1-4+13.
- [9] Fan Yameng. Evaluation of comprehensive economic strength of Counties in Chongqing based on principal component analysis [J]. Guangxi Quality Supervision Review, 2021 (05): 73-74.
- [10] Liu Yao, XIONG Jianping. Comparison of tourism development level in Hubei province based on principal component analysis and cluster analysis [J]. Tourism Survey, 2021 (10): 153-158.
- [11] Hu Jiangxia, Luo Zhigao, Wen Chuanhao. Statistics and Decision, 201, 37 (12): 82-85.
- [12] Shan N. County economic pattern and its difference evolution and mechanism analysis in Jiangsu Province [J]. Journal of Inner Mongolia University of Finance and Economics, 20119 (03): 104-107.
- [13] Jiang Yonghong, Zhang Yindan. Competitiveness evaluation and Promotion countermeasures of prefecture-level cities in Shaanxi Province: Based on factor analysis and cluster analysis [J]. Journal of Shaanxi University of Administration, 201, 35 (02): 9-14.
- [14] Yang J K. Research on the structure of Chinese residents' consumption expenditure -- based on factor analysis and cluster analysis [J]. Modern Business, 2021 (14): 7-9.
- [15] TAN Y. Analysis of regional economic differences based on cluster analysis model -- a case study of Sichuan Province. China Market, 2021 (14): 4-7.
- [16] Gao Huixian, Xue Yulian, Fang Zhong. Journal of Fujian Normal University, 2021, 39 (02): 189-195.