Research on Urbanization of China Based on Factor Analysis and Cluster Analysis

Wenting Feng

School of Statistics , Shanxi University of Finance and Economics, Taiyuan 030006, China;

2507166300@qq.com

Abstract. With the convening of the 19th National People's Congress of the Communist Party of China, building a modern economic system received extensive attention. However, urbanization is one of the indispensable key factors. So this paper will focus on the urbanization level of 31 provincial capital cities and 5 municipalities with independent planning status in China. First，12 indicators are selected as the basis for judging the urbanization. Then, the four factors are extracted based on factor analysis，which are science and technology culture, population, public life, industry, and environmental economy, and the index system is established. Next, the 36 cities are clustered to classify the level of urbanization in each region based on the cluster analysis. In the end, the 36 cities are clustered into 3 categories, the result and analysis are shown at the end of this paper.

Keywords: Descriptive statistics Factor analysis Cluster analysis Urbanization

1. Introduction

1.1 Background

The goal of building a modern economic system has been put forward in the 19th National People's Congress of the Communist Party of China. However, different regions have different levels of development and different levels of urbanization. Therefore, in formulating policies, it is necessary to adjust measures to local conditions and adopt economic policies that are consistent with development. So it is inevitable to study the level of urbanization in each region.

Urbanization refers to the population shift from rural to urban areas, "the gradual increase in the proportion of people living in urban areas", and the ways in which each society adapts to the change. [1] With the development of economy, urbanization has become an inevitable trend of social development. However, there are many differences in the economic base, population situation, people's living standard and resource conditions in different regions, resulting in uneven development and different urbanization processes. Therefore, it is necessary to evaluate the level of urbanization effectively and find out the reasons for the difference.

1.2 What the paper do

To research on the urbanization, we should construct a scientific and reasonable index system firstly. According to multiple inspections and considerations，we have selected five first-level indicators, and twelve secondary indicators to construct the index system. At the same time, we use X1-X12 to represent these twelve indicators. (**Table1**.)

Table1 The Index system

Economic development level	Real GDP per capita(X1) Share of tertiary industry in GDP(X2)
Population development level	The proportion of employees in the tertiary industry to all employees(X3) Population density(X4) Natural population growth rate(X5)
Social development level	The number of buses per 10,000 people(X6) The number of doctors per 10,000 people(X7) The number of fixed telephones per 10,000 people(X8)
Scientific and cultural development level	The number of college students per 10,000 people(X9) The number of Internet broadband access users(X10) The number of books in the public library per 100 people(X11)
Environmental development level	Garbage disposal rate(X12)

Next, we collected data according to the index system. In this paper, we choose 31 provincial capital cities and 5 municipalities with independent planning status in China. And all the statistics are drawn from China Statistical Yearbook 2016[3] and China City Statistical Yearbook 2016[4].

2. Descriptive statistics

First, we preprocess the data. We find that missing values existed in 36 cities because the data in Lasa were incomplete. After repeated searches, the latest data were not available, so the city's data was excluded. Next we have a new descriptive statistical.

According to the statistical description, we find that Shenzhen's GDP per capita, natural population growth, the number of buses per 10,000 people, and the number of fixed telephones per 10,000 people are the maximum, and Beijing has the biggest Share of tertiary industry in GDP, and the highest proportion of employees in the tertiary industry to all employees; Shanghai has the highest population density and 10,000 people have the largest number of fixed telephones.

3. Factor analysis empirical process

3.1 Applicability judgment of factor analysis

At first, we calculate the correlation coefficient matrix between 12 variables. Then we conduct a test to see if it is suitable for factor analysis.

Table2. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.611
Bartlett's Test of Sphericity	Approx. Chi-Square	224.029
	df	66
	Sig.	.000

From the above table, it can be seen that the KMO value reaches 0.611, indicating that it is suitable for factor analysis; at the same time, the P value of Bartlett’s sphericity test is close to 0, which is far less than 0.05, which also shows that factor analysis can be performed.

3.2 Extraction factor

(1) Communalities. From the communalities, we find that the common factors cover almost 70% of most variables.

(2) Total variance explained. From the Total Variance Explained table, it can be seen that the

cumulative contribution rate of the eigenvalues of the first four factors reaches 76.313%, which is satisfactory, so the first four factors are retained.

(3) Factor rotation. From the component matrix, we find that there is no obvious difference in the load of the common factor on some of the original variables, so it is necessary to rotate the factors. The factor load matrix obtained after rotation is as follow (**Table3.**)

Table3. Rotated Component Matrixa

	Component			
	1	2	3	4
Real GDP per capita	.492	.362	.046	.574
Share of tertiary industry in GDP	.040	.242	.892	.073
The proportion of employees in the tertiary industry to all employees	-.041	-.311	.830	-.154
Population density	.611	.405	.063	.362
natural population growth rate	-.122	.848	-.142	-.183
the number of buses per 10,000 people	.357	.802	.045	.129
the number of doctors per 10,000 people	.102	.647	.530	.296
the number of fixed telephones per 10,000 people	.886	.173	.305	.112
the number of college students per 10,000 people	-.615	.187	.236	.070
the number of Internet broadband access users	.880	.111	.021	-.006
the number of books in the public library per 100 people	.293	-.196	.384	.736
Garbage disposal rate	-.205	.047	-.246	.799

From the rotated component matrix, we can draw the following conclusions.

a) The first public factor has a relatively large load on the four variables: the number of fixed telephones per 10,000 people, the number of Internet broadband access users, the number of college students per 10,000 people, and population density. So they fall into one category and are named as science and technology culture and population factors;
b) The second public factor has a relatively large load on the three variables: natural population growth rate, the number of buses per 10,000 people and the number of doctors per 10,000 people. This shows that these three variables have a good correlation, and it can be seen that this is related to people's public life. , it is named as public life factors;
c) The third public factor has a relatively large load on the two variables: share of tertiary industry in GDP and the proportion of employees in the tertiary industry to all employees. It can be seen that both are related to the tertiary industry and are therefore named as industrial factors;
d) The fourth public factor has a relatively large load on the three variables: garbage disposal rate, the number of books in the public library per 100 people and Real GDP per capita. It can be seen as related to the environmental economy. Therefore, it is named as environmental economic factors.

(4) Factor score

The factor score expression for the four factors can be determine as follows from the Component Score Coefficient Matrix:

$$Y_1 = 0.094X_1^* - 0.070X_2^* - 0.013X_3^* + 0.166X_4^* - 0.109X_5^* + 0.048X_6^* - 0.093X_7^* + 0.324X_8^* - 0.302X_9^* + 0.359X_{10}^* + 0.028X_{11}^* + 0.094X_{12}^*$$

$$Y_2 = 0.074X_1^* + 0.083X_2^* - 0.154X_3^* + 0.099X_4^* + 0.439X_5^* + 0.336X_6^* + 0.262X_7^* - 0.027X_8^* + 0.142X_9^* - 0.039X_{10}^* - 0.195X_{11}^* - 0.015X_{12}^*$$

$$Y_3 = -0.047X_1^* + 0.423X_2^* + 0.424X_3^* - 0.035X_4^* - 0.075X_5^* - 0.027X_5^* - 0.220X_6^* + 0.087X_7^* + 0.142X_8^* - 0.043X_9^* + 0.136X_{10}^* - 0.153X_{11}^*$$

$$Y_4 = 0.266X_1^* - 0.027X_2^* - 0.111X_3^* + 0.113X_4^* - 0.161X_5^* - 0.033X_6^* + 0.092X_7^* - 0.073X_8^* + 0.094X_9^* - 0.126X_{10}^* + 0.421X_{11}^* + 0.538X_{12}^*$$

The X_1^*~X_{12}^* in the expression is the normalized data of the original data.

For further comprehensive evaluation we also need to calculate the composite score by weighting the two common factors as their weighted contribution to the cumulative contribution rate.

$$\text{the composite score} = \left(\frac{23.635}{76.313}Z_1 + \frac{19.560}{76.313}Z_2 + \frac{17.904}{76.313}Z_3 + \frac{15.214}{76.313}Z_4\right)$$

Among them, Z_1, Z_2, Z_3 and $Z_4$are the scores of the 36 cities on the common factor.

4. Cluster analysis

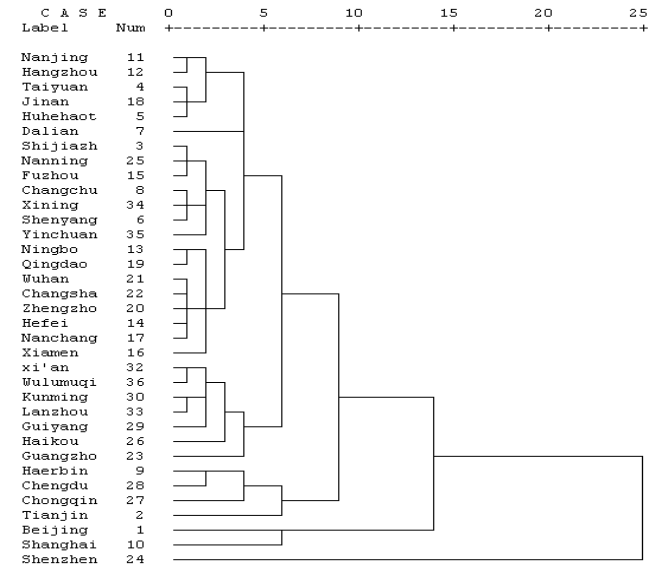
Based on the four factors obtained from the above factor analysis, a cluster analysis was conducted on 30 provincial capital cities and 5 municipalities with independent planning status in China to classify their urbanization levels.

(1) Check the data

There are no missing values in the data we have chosen and they can all be effectively classified.

(2) Systematic clustering

The following hierarchical diagram is obtained by clustering analysis.



It can be seen from the hierarchical diagram that the three categories are suitable. Among them, Shenzhen is a class, Beijing and Shanghai are one class, and the rest is a class. This clustering result conforms to the present situation of China. Shenzhen is a coastal developed area, its science and technology culture, public life, industrial development, environmental economy and other levels all in the national leading position; as the capital of our country, Beijing is the center of political and economic culture. Shanghai, also known as an international metropolis, has a better development level. The rest of the world is more mundane.

5. Conclusion

Through factor analysis, China's urbanization index system was established and four factors were proposed. Then cluster analysis was performed and 35 cities were classified into three categories. The highest level of urbanization was Shenzhen, followed by Beijing, Shanghai, and the rest. This is consistent with the development status of China. From this paper we can see that there are still great differences in the level of urbanization in various regions, and the government needs to implement corresponding policies to encourage development and promote coordinated development among regions. For cities with a relatively high level of urbanization, the government can focus on technological culture and environmental development. For other places, while carrying out economic construction, the government needs to increase support and give policy support. At the same time, it must also pay attention to environmental protection.

References

[1] "Urbanization". MeSH browser. National Library of Medicine. Retrieved 5 November 2014. The process whereby a society changes from a rural to an urban way of life. It refers also to the gradual increase in the proportion of people living in urban areas.

[2] Ligang Wang .Comprehensive Evaluation of Urbanization Levels in Cities and Cities in Guizhou Provinces Based on Principal Component Analysis, Clustering and GIS Analysis Methods[J].Journal of Northwest University Nationalities(Philosophy and Social Sciences),2011(02):96-103.

[3] National Bureau of Statistics of Urban Socioeconomic Survey. China Statistical Yearbook 2016. China Statistics Press.

[4] National Bureau of Statistics of Urban Socioeconomic Survey. China City Statistical Yearbook 2016. China Statistics Press.