Economic Data Mining Research based on Complex Networks and factor analysis

LI Yajie[1,a], HUANG Tao[2,b] and MA Liwen[3,c]

[1] Department of Mathematics, School of Science, Beijing University of Posts and Telecommunications, Haidian District, Beijing, 100876, China

[2] Department of Mathematics, School of Science, Beijing University of Posts and Telecommunications, Haidian District, Beijing, 100876, China

[3] Department of Mathematics, School of Science, Beijing University of Posts and Telecommunications, Haidian District, Beijing, 100876, China

[a] lyj7712@163.com, [b]532087392@qq.com, [c]maliwenmath@sina.com

Keywords: Complex Network; Economic Development Level; Factor Analysis

Abstract. Economical data mining is practical and meaningful. Economic development levels of different regions in China are varied. We use complex network analyzing method to make network graphs and did nodal analysis and community structure detection of these regions to indicate the difference of the economic development levels of each province in China. At the same time, we did the modeling analysis of the development levels of China's provinces by multivariate statistics factor analyzing. We made the comprehensive score model and did the evaluation of the provincial economic development level. The result shows that there are regional differences between provinces in China and the regions with similar economic development levels are closely connected or related to each other.

1. Introduction

With the implementation of the reforming policies of China, the Chinese economy has been increasing rapidly. The developing speed varies in difference regions due to economical, social, natural and historical reasons. The Chinese government has approved the planning documents of the Yangtze River delta, the Pearl River Delta (PRD or Chu Chiang Delta), the West Coast Economic Zone of Fujian Province, Tianjin Binhai Hi-Tec Industrial Development Zone, Guanzhong - Tianshui Economic Zone in Shanxi province and Gansu province, Tumen River Delta in Jilin province in east China, Poyang lake ecological economic zone, Zhuhai Hengqin Zew Zone and the Wanjiang City group in Anhui province. The main purpose of our study is to find the regional similarities and shorten the economical differences between the regions thus to make the economic development sustainable and coordinate. Therefore it is very important to objectively evaluate the economic viability and search for the reason that lead to these differences to realize sustainable and coordinate regional economic development .

Lack of balance in provincial economic development has become a problem in China. To intuitively present the regional economic development feature in China we made the network graphs with complex network method. In the graph we can intuitively indicate the regional differences between the provinces through the relations of the node points and community detection. To make the evaluation of the provincial economic development level measureable, we applied factor analyzing method on the comprehensive assessment of the 31 Chinese provinces economic development level.

2. The Complex Network of Economic Development Level

Studying method. We can use networks to describe large amount of complex systems. A typical network is combined by many nodes and edges that connect the nodes together. In the graph, the nodes stand for different individuals in the real world and the edges refer to the relations between the individuals. Usually when there are some certain relations between two nodes there will be an edge

linking them together. If not, the nodes are not linked. If two nodes connected by an edge in the network, they are referred to as the adjacent. For example, the nervous system can be thought of as a large number of nerve cells linked by nerve fibers to form a interconnected network. And computer networks can be regarded as an independent computer linked with other computes via communication media such as fiber, twisted-pair cable or coaxial cable, etc to form the network [1, 2]. And electric power networks, social networks, the traffic systems are similar.

Researches on system networks have experienced three stages. The first stage, the scientists believed that in the real system the relationship between each element can be expressed with some regular structures such as Euclid two-dimensional grid, a grid pattern similar to the grid lines on our T-shirts. It is like the network formed by rings next to each other or like a group of girls armed together to form a circle and dance around the bonfire. The second stage is in the late 1950s. Mathematicians believed that the edges linking the nodes are randomly formed. Under this theory, the networks made by the mathematicians were random networks. Random networks has been considered to be the most appropriate to describe real network system afterward. The third stage is in recent years. Due to the rapid development of computer data processing and computing power, the scientists found that a large number of network is neither networks formed with certian rules nor randomly formed. These are different from the above both networks. Among these networks, small world effect (1998 Watts and Strogatz 'work announced small world network - WS network) and networks of scale-free properties (1999, Albert Barabasi and Albert provide BA scale-free network model –BA model) are referred to as complex networks.

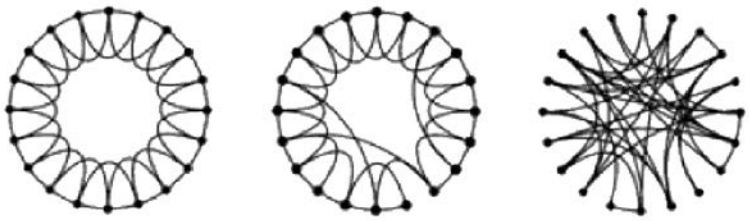


Fig. 1 Regular Network, Small World Network, Random Network

Below are some methods for complex networks analysis:

1. The Boost Graph Library (BGL) is a member of the C++ Boost libraries. It is characterized by flexibility and high processing efficiency. It works with a template supplied by the library. You can create a basic type, such as custom node class, on the basis of the template. All sorts of figure algorithm are of very high execution efficiency in this library. The shortcoming of BGL is that complex network analysis algorithms are not supplied.

2. Quick Graph is a NET component Library written with C# programming language. The algorithm provided in Quick Graph is similar to that provided in BGL. It can be seen as a Boost Graph Library implemented on the NET platform. The efficiency is not high enough for processing large computing problems because it is run under the virtual machine.

3. Igraph - a complex network analysis library written with C# programming language.Igraph is an open source C program library used to establish and operate undirected graphs and directed graphs. It contains both some traditional algorithms like minimum support tree and network flow and also some new emerging algorithms (such as community structure search). Igraph is written with C# programming language and is supplied with access to R language and Python R language which makes it suitable for researchers to use.

4. NetworkX is a Python pack used to establish and operate complex networks and study their structures, functions and dynamics. It supplies the most popular complex networks analyzing algorithms including the classic graph algorithm. Currently NetworkX can only be used in Python language context. The documents of NetworkX are clear and highly readable. NetworkX has good program structures but most of time its processing efficiency is lower than igraph so it is not suitable for large scale network analysis.

5. Gephi is a free open source cross-platform network analysis software based on JVM. It is mainly used as visualization and research tools in various networks, complex system, dynamic and

hierarchical graph. It can be used to do exploratory data analysis, social network analysis, dynamic analysis, biological network analysis, network community analysis and classification, etc. It is highly visualized but is weak in diversity of community discovery algorithms.

Study Process. We selected 12 indexes reflecting the economic and financial development levels of 12 provinces. These indexes are : X1- gross domestic product (hundred million Yuan); X2- social fixed assets investment (hundred million Yuan); X3- local fiscal revenue (general budget revenue, hundred million Yuan); X4- local fiscal expenditure (general budget expenditure, hundred million Yuan); X5- total retail sales of social consumer goods (hundred million Yuan); X6- the consumer price index (suppose last years' index number as 100); X7- import and export amount (hundred million dollar); Xx8- deposit of banks and financial institutions (hundred million Yuan); X9- loan of banks and financial institutions (hundred million Yuan); X10- amount of listed companies in China; X11- insurance premium income (hundred million Yuan); X12- insurance density of all institutions (Yuan per capital). The data are quoted from China's financial statistics yearbook.

Suppose the indicators of each province are: $Y=\{Y_{ij},i=1,2,\ldots I;\ j=1,2,\ldots J\}$,Among which Y_{ij} refers to the score that province i gets for indicator j. Here I=31, J=12. To make it comparable, we standardized the score and get the standard scroe matrix is :$X=\{X_{ij},i=1,2,\ldots I;\ j=1,2,\ldots J\}$,refer with:Eq.1.

$$X_{ij}=\frac{Y_{ij}}{\sqrt{(\sum_{i=1}^{I}Y_{ij}^{2})/I}} \tag{1}$$

We use angle cosine distance to indicate the provincial economic development level differences. We define the angle distance between province r and province s as: P_{rs} ,refer with:Eq.2.

$$P_{rs}=1-\frac{\sum_{j=1}^{J}X_{rj}X_{sj}}{\sum_{j=1}^{J}X_{rj}^{2}\sum_{j=1}^{J}X_{sj}^{2}} \tag{2}$$

Namely the economic development level difference between province r and province s is cosine of the included angle vector quantity formed by the vector quantity derivated from the above 12 indicators. All indicators are positive numbers between [0,1]. The distance indicates the similarity between two provinces economic development levels.

In a connected graph G with multiple fixed points, if graph A contains all vertexes and some of the edges of G, and at the same time there is no circuit in graph, then A can be named as the spanning tree of G. The minimum spanning tree is a spanning tree with the fewest points and edges. Here minimum means the weight of the edges is the smallest. In this article, we use each provinces and regions as the points and the cosine distance between the points as the weight of the edges. The minimum spanning tree we get here shows the main structure elements of the complex system formed by the relationship between the provinces. This main structure indicates the regional economic development situation in China and the difference between these regions and the provinces with powerful economy.

We use Kruskal algorithm to generate a minimum spanning tree. Specific steps are as follows:

(1) Suppose every province is an initial isolated point;

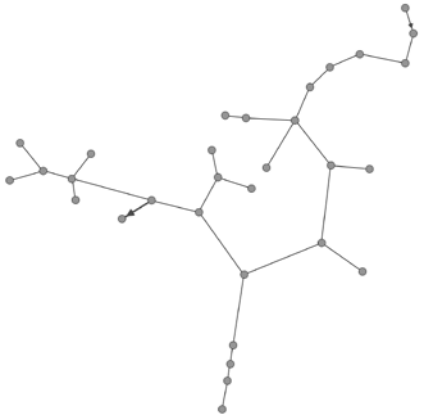
(2) Find the two nearest provinces and connect them to form a minimum spanning tree;

(3)Find the provinces that have the smallest distance to the spanning tree in the in the rest of the provinces and link them together. Make sure there wont be any circle in the graph. A new spanning tree is created when we finish doing this;

(4) Repeat steps (3), until all the provinces have been connected.

We can create provincial economic relations complex networks by using the provinces for the nodes and the relationship between the provinces for the edges. There are many methods to make complex networks on the bases of the node matrixes, for example threshold method. In this method we fix a threshold value through the remaining topological property, the network connectivity, the minimum spanning tree and the joint events. Then to make an edge between nodes between which the distance is smaller than the fixed threshold value. Another method is to link the nodes to their nearest

nodes [3, 4]. To make it simple in practice, we apply the second method. We link each province to its nearest provinces (we chose 3 for calculation). In this way we make the minimum spanning tree as a reference to judge the economic development level of a certain province in China.



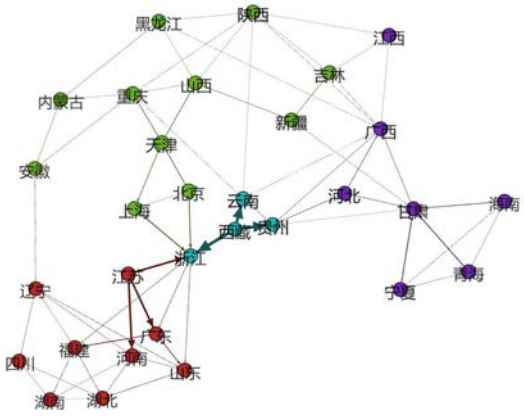


Fig. 2 Minimum Spanning Tree Fig. 3 Complex Network

Figure 2 shows minimum spanning tree which adopted the cosine distance of each province as the edges. Figure 3 shows the network adopted the cosine distance of each province. Figure 3 is made by gephi. Gephi modular functions are used to do the community detection.

Conclusion. We can find by the complex network diagram that economic levels are similar between provinces and cities that are closely related to each other. For example, Beijing, Shanghai, Zhejiang, Jiangsu, Guangdong and Shandong have closer relations and these provinces and cities are well developed in economy. Yunnan, Guizhou, Tibet, Ningxia, Qinghai,Gansu and Hainan are closely related and these provinces and cities are relatively poorly developed.

Economy developed levels of Chinese provinces can be ranked into three levels: Provinces and cities in east China are developed best, economy situations of provinces and cities in the central regions are on the second stage; the western areas are the worst. From the aspect of community detection, the eastern coastal provinces and cities with higher economic levels and the central areas are in a same community. The four municipality directly under the Central Government (Beijing, Shanghai, Tianjin and Chongqing) and the northern provinces and cities are in a same community. And the western provinces are divided into to communities. In one community the less developed provinces adopted the economy developing models of the well developed provinces to develop their own economy. But provinces in the other community cannot develop their economy because they are located in the hinterland of China and with poor transportation system. These provinces should re-position their economy development strategy under the China's west development policies and sustainable development strategies in order to maintain and develop their advantages and formulate corresponding policies to make up the shortfall at the same time. Through doing this the regional economic vitality can be comprehensively enhanced.

3. Factor analysis of the economy development level

Factor analysis is an important method of multivariate statistics. It is mainly used for evaluation, dimension reduction, ranking, clustering, etc. It was originated form Pearson and Charles Spearmen's research on intelligence tests in the early 20th century. We use SPSS17.0 program to do factor analysis in this article. We use the principal component method to extract the common factor to get the variance contribution rate. The variance contribution rate indicates the overall variance contribution of the common factor supplied by the variances. It is a measurement to the importance of the common factor. We get the variance contribution rate of the three common factor are separately 68.561%, 11.530% and 8.607%. Information extraction quantity of the first three common factor of is

more than 88% and the first three characteristic root is larger than1, as shown in table 1. So the first three common factor extraction is scientific.

Table 1 Variance Contribution Rate

Component	Initial Characteristic Value			Rotated Quadratic Sum		
	Sum	Percentage of variance	Accumulation (Percentage)	Sum	Percentage of variance	Accumulation (Percentage)
1	8.227	68.561	68.561	4.908	40.902	40.902
2	1.384	11.530	80.090	4.629	38.572	79.474
3	1.033	8.607	88.697	1.107	9.223	88.697
4	.719	5.995	94.692			
…	…	…	…			
12	.003	.024	100.000			

Because the factor analysis result did not answer our questions well, we adopted the maximum variance factor rotation method and got the factor loading matrix in table 2. The statistical significance of the factor loading lies in the relation between the variable Xi and common factor Fj. It shows how much Xi relies on Fj. In statistics it is called weight, but in psychology it is called load. Weight means the functions of variable Xi on common factor Fj. It reflexes the importance of the variable Xi for common factor Fj. [5]

Table 2 Rotated Factor Loading Matrix

	Common factor		
	1	2	3
X1	.489	.837	.144
X2	.268	.896	.195
X3	.741	.600	.172
…	…	…	…
X12	.893	-.178	-.037

From the rotated factor loading matrix we can find that the common factor has high load on X3 (local fiscal revenue), X7(import and export amount), X8(deposit of banks and financial institutions), X9 (loan of banks and financial institutions), X10(amount of listed companies in China), X12(insurance density of all institutions). X3 (local fiscal revenue) is the government's income as the country manager and the owner of the state-owned assets. It reflects the development level of local economy. X7 (import and export amount) is used to observe the total scale of a country's foreign trade. X8 (deposit of banks and financial institutions) and X9 (loan of banks and financial institutions) show the bank credit scale. The banks had great influences on local financial and economics. X10 (quantity of listed companies)is also an important index of local financial. Public companies are first and foremost companies. They are the entities to produce a product or services. Then they become listed companies. The listed companies are important source of modern economic growth. It is the trading objects in the stock market and the cornerstone of the securities market. They influences the modern economic development through the capital market.X12 (insurance density of all institutions) is the annual insurance fee per capital. Insurance can effectively reduce the risk of financial industry and sustain the financial stability and healthy development.

In conclusion, the common factor F1 reflects the development of the market economy. The higher the score it gets, the better the economy development condition is in the area. Common factor F2 reflects the government's macroeconomic regulation and control. Score on this factor reflects the strength of a regional government macroeconomic regulation and control policy. Common factor F3 only has relatively high load on X6 (the consumer price index). CPI is a reflection of the price variation of households to buy products and services. So F3 is the common factor reflecting the price

level.Take factor loading as the coefficient we can get model which original variable X_i relies on common factor F_j as follow: X1, X2. Others are similar. [6]Refer with:Eq.3.

$$X1=0.489\cdot F1-0.837\cdot F2+0.039\cdot F3 \quad , X2=0.268\cdot F1+0.896\cdot F2+0.195\cdot F3. \tag{3}$$

Then we calculate the factor scores. We use the proportion of the three factors' variance contribution rates out of total variance contribution rate as the weight to do the weighting. With this method we get the synthesis score F of each province: F, refer with Eq.4.

$$F=(40.902\cdot F1+38.572\cdot F2+9.223\cdot F3)/88.697. \tag{4}$$

The synthesis result and the comprehensive assessment is displayed as in table 3. We can take the score of factor F1 as the lateral axis and the score of factor F3 as the vertical axis to sketch the factor score scatter diagram. The scatter diagram is not shown in the article.

Table 3 Comprehensive Assessment

City	F1	F2	F3	F	Comprehensive Assessment
Beijing	2.65354	-1.24121	-0.2729	0.66	4
Tianjin	0.3347	-0.72536	-0.13082	-0.17	8
Jiangsu	-0.25057	3.70766	-1.40979	1.35	2
Zhejiang	1.23303	0.66784	-0.00559	0.86	3
…	…	…	…	…	…
Guangdong	2.44499	1.65669	0.04836	1.85	1
Hainan	-0.61796	-0.84377	-0.38794	-0.69	15
Chongqing	-0.29781	-0.29414	-0.22186	-0.29	11
Sichuan	-0.05971	0.5391	0.04626	0.21	6

Note: We displayed part of the results in the table to make sure it won't take too much space.

Combining the provinces' common factor score and their synthesis score we can see that the provinces that have better economy development levels are Guangdong, Jiangsu, Shandong, Shanghai and Zhejiang; the worst ones are Tibet, Qinghai, Ningxia, Hainan and Guizhou. Guangdong's market economy is better developed. The Guangdong government's macroeconomic regulation and control policy has stronger influences and the price level is relatively low. Guangdong's economic development level is higher than other provinces and cities in China. The Jiangsu government's macroeconomic regulation and control policy works well but the market economy development and its price level are less good, more attention should be paid on this. The provinces and cities with lower synthesis score usually has low score in market economy development and government's macroeconomic regulation and control policy. This indicates that we need to do more in these two aspects. We should make combined efforts of the market developemnt and the government's control to ensure the regional economy growth.

We can see from the factor scores that Shanghai, Beijing, Guangdong, Zhejiang are of higher market economy development level. Of which Shanghai is much better than the others. Qinghai and Tibet has low market economy development level. Jiangsu, Shandong Guangzhou are doing well in the regional economy development due to good government policy and control.

Seeing from the aspect of geographical position, the market economy development of the coastal provinces and cities are higher than those in the hinterland. Cities and provinces with the best economy performance are located at the Pearl River delta economic zone, the Yangtze River delta economic zone and Bohai Sea economic zone. The central region lies the second best provinces and the western region lies the regions with worst performance. How to speed up the economic development of the vulnerable provinces and cities to drive the progress of their surrounding areas is an important task that influences the overall economic development of China.

ACKNOWLEDGEMENTS

This research is supported by the Fundamental Research Funds for the Central Universities (BUPT2015TS01).

References

[1] D.J. Watts and S.H. Strogatz:*Collective dynamics of "small world" networks.* Nature, Vol. 393 (1998), p. 440-442.

[2] M.Faloutsos, P.Faloutsos and C.Faloutsos:*On power-law relationships of the Internet topology*.ACM SIGCOMM Computer Communication Review, Vol. 29 (1999),p. 251-262.

[3] Y.Yang, and H.Yang:*Complex network-based time series analysis.* Physica A Vol. 387(2008) , p. 1381-1386.

[4] M.E.J.Newman:*The structure and function of complex networks*. SIAM Rev, Vol. 45(2003),p.167-256.

[5]X.Q.He:*Multivariate statistical analysis*. (Beijing: China renmin university press2012).

[6]G.H.Ruan:*The data statistics and analysis - SPSS application tutorial.* (Beijing: Peking University press2005).