Research on the difference of Digital inclusive Finance--based on Multi-index Panel data clustering

Jinmin Zhang^{1,*}, Xufang Li^{1,2} and Dijun Fan¹

¹School of Management, Shanghai University of Engineering Science, Shanghai, China

²Business School, University of Shanghai for Science and Technology, Shanghai, China

Abstract. Objective: to understand the development level of digital inclusive finance in 31 provinces in China in recent years, so that the areas with poor development level can speed up the development. Methods: the data of Peking University Digital inclusive Financial Index from 2012 to 2018 were collected, and the optimal clustering number was determined, and then the cross-sectional data and multi-index panel data were clustered respectively. Conclusion: the level of digital inclusive finance in China shows an upward trend as a whole, but there are great differences in the development level of digital inclusive finance among 31 provinces in China, in which East China is the best, South China and Central China have a better overall development level, and North China, Northwest, Southwest and Northeast are poor in overall development level, but have provincial differences.

1 Introduction

The Fourteenth five-year Plan in 2020 proposes to speed up the digital planning and construction of digital China. Digital inclusive finance is to digitize inclusive finance and provide more convenient financial services for the people. The digital inclusive financial index is a comprehensive index reflecting the development level of digital inclusive finance. From 2012 to 2018, the level of digital inclusive finance in China is on the rise as a whole. In order to speed up the development level of digital inclusive finance, this paper focuses on the regional differences of the development level of our country, in the hope that the government can take more measures to make up for the shortcomings, reduce the regional differences, and improve the development level of digital inclusive finance as a whole².

However, the research on digital inclusive finance mainly focuses on the development process of digital inclusive finance, the measurement of indicators and the impact on economy, innovation, enterprises and so on. Few scholars have studied the regional differences of digital inclusive finance, but understanding regional differences can play a positive role in the development of digital inclusive finance. Most of the existing researches on the differences of digital inclusive finance in different provinces in China use the data of a certain year for cluster analysis, although they can reflect the differences in the level of digital inclusive finance in different provinces in that year, but they are lack of time dimension. It does not reflect the comprehensive level difference of the development of digital inclusive finance in China in recent years. Multi-index panel data includes three dimensions: time, sample and index, which can better and

more comprehensively reflect the development of digital inclusive finance in our provinces in recent years. Therefore, this paper applies multi-index panel data clustering to digital inclusive finance.

2 Data structure and characteristics of Multi-index Panel data

2.1 Multi-index panel data structure

Panel data is a data type that integrates cross-sectional data and time series data. It has two dimensions: time series and cross-section, which is divided into single-index panel data and multi-index panel data. The single index panel data is similar to the cross-section data and can be represented by a two-dimensional table. The structure of multi-index panel data is more complex than that of single index panel data, and it is a kind of three-dimensional structure data, which can be transformed into multiple two-dimensional data representations on the plane. The representation of multi-indicator panel data is shown in Table 1:

Table 1. Multi-indicator panel data

Sample Time	1	2	3		k		m
1	$x_{1,1,n}$	$x_{1,2,n}$	$x_{1,3,n}$		$x_{1,k,n}$		$x_{1,m,n}$
2	$x_{2,1,n}$	$x_{2,2,n}$	$x_{2,3,n}$		$x_{2,k,n}$		$x_{2,m,n}$
3	$x_{3,1,n}$	$x_{3,2,n}$	$x_{3,3,n}$		$x_{3,k,n}$		$x_{3,m,n}$
4	$x_{4,1,n}$	$x_{4,2,n}$	$x_{4,3,n}$		$x_{4,k,n}$		$x_{4,m,n}$
.....							
t	$x_{t,1,n}$	$x_{t,2,n}$	$x_{t,3,n}$		$x_{t,k,n}$		$x_{t,m,n}$

* Corresponding author: xiaoxiao720809@163.com

$x_{i,j,n}$ for $i=1,2,3\dots t$ $j=1,2,3\dots k\dots m$
 $n=1,2,3\dots s$
 where i,j and n represent time, sample and index respectively

2.2 Digital characteristics of multi-index panel data

(1) The mean of sample k in time i is

$$\bar{x}_{i,k} = \frac{1}{s} \sum_{n=1}^s x_{i,k,n} \quad (1)$$

The variance is

$$Var(x_{i,k}) = \frac{1}{s-1} (x_{i,k,n} - \bar{x}_{i,k})^2 \quad (2)$$

(2) At time i , the mean of all the samples for the n th index is

$$\bar{x}_{i,n} = \frac{1}{m} \sum_{j=1}^m x_{i,j,n} \quad (3)$$

The variance is

$$Var(x_{i,n}) = \frac{1}{m-1} \sum_{j=1}^m (x_{i,j,n} - \bar{x}_{i,n})^2 \quad (4)$$

(3) The mean of the n th index in the k th sample is

$$\bar{x}_{k,n} = \frac{1}{t} \sum_{i=1}^t x_{i,k,n} \quad (5)$$

The variance is

$$Var(x_{k,n}) = \frac{1}{t-1} \sum_{i=1}^t (x_{i,k,n} - \bar{x}_{k,n})^2 \quad (6)$$

(4) At time I , the mean value of the overall sample is

$$\bar{x}_I = \frac{1}{ms} \sum_{j=1}^m \sum_{n=1}^s x_{i,j,n} \quad (7)$$

The variance is

$$Var(x_i) = \frac{1}{ms-1} \sum_{j=1}^m \sum_{n=1}^s (x_{i,j,n} - \bar{x}_I)^2 \quad (8)$$

(5) The mean of sample j is

$$\bar{x}_j = \frac{1}{ts} \sum_{i=1}^t \sum_{n=1}^s x_{i,j,n} \quad (9)$$

The variance is

$$Var(x_j) = \frac{1}{ts-1} \sum_{i=1}^t \sum_{n=1}^s (x_{i,j,n} - \bar{x}_j)^2 \quad (10)$$

(6) The mean of index n is

$$\bar{x}_n = \frac{1}{tm} \sum_{i=1}^t \sum_{j=1}^m x_{i,j,n} \quad (11)$$

The variance is

$$Var(x_n) = \frac{1}{tm-1} \sum_{i=1}^t \sum_{j=1}^m (x_{i,j,n} - \bar{x}_n)^2 \quad (12)$$

3 Clustering method

3.1 Determination of clustering number K value

Clustering algorithm belongs to unsupervised learning, and the clustering number needs to be determined artificially. Too much or too little clustering number will have a negative impact on the clustering results. The optimal K value can be determined by the elbow method. When the number of clusters is less than the optimal number of clusters, increasing the value of k will greatly reduce the SSE; when the number of clusters is greater than the optimal number of clusters, although the more the number of clusters, the degree of aggregation will increase, but the extent of the decrease of SSE value will decrease sharply and gradually tend to be parallel.

Sum of squares of error: $SSE = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2$
 for $i=1,2,3\dots k$ $j=1,2,3\dots n$

Fig. 1. Clustering k number

From Figure 1, we can see that the digital inclusive financial differences in different regions can be best divided into four categories.

3.2 Systematic clustering method

Clustering methods include K-means clustering, fuzzy clustering, dynamic clustering and clustering combined with intelligent algorithms⁵. Systematic clustering is the most widely used clustering analysis method, which mainly includes the distance between points and the distance between classes. The distance between points includes Euclidean distance, Chebyshev distance, Ming's distance, absolute distance, and the distance between classes includes the shortest distance method, the longest distance method, the center of gravity method, and the sum of square deviation method. Considering the characteristics and distribution of the data, the Euclidean distance is selected to describe the distance between samples, and the Ward method is used to measure the distance between classes.

(1) Euclidean distance: $d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$

(2) Ward method:

Suppose it is divided into n samples into $\{L_1, L_2, \dots, L_k\}$

x_{it} represents the i sample in class L_t , n_t represents the number of samples in class L_t , $\bar{x}_t$ is the center of gravity of L_t . The sums of squared deviations of L_t is

$$S_t = \sum_{i=1}^{n_t} (x_{it} - \bar{x}_t)' (x_{it} - \bar{x}_t) \quad (13)$$

The sums of squared deviations within the whole class is

$$S = \sum_{t=1}^k L_t \quad (14)$$

3.3 Multi-index panel data clustering

With the addition of the time dimension, the distance between the points becomes the distance between the cross section and the cross section, and the Euclidean distance is extended, and the distance between the faces is also the sum of the distances of all the points in the face. At the same time, it is necessary to consider the change speed of the sample index value with time and the fluctuation degree of the sample index value with time. The distance formulas used are the Absolute Quantity Euclidean Distance(AQED), the Increment Speed Euclidean Distance(ISED), the Variation Coefficient Euclidean Distance(VCED), and the Comprehensive Variation Euclidean Distance(CED).

AQED between sample j_1 and j_2 is

$$d_1(j_1, j_2) = \left[\sum_{i=1}^t \sum_{n=1}^s (x_{i,j_1,n} - x_{i,j_2,n})^2 \right]^{\frac{1}{2}} \quad (15)$$

ISED between sample j_1 and j_2 is

$$d_2(j_1, j_2) = \left[\sum_{i=1}^t \sum_{n=1}^s \left(\frac{\Delta x_{i,j_1,n}}{x_{i-1,j_1,n}} - \frac{\Delta x_{i,j_2,n}}{x_{i-1,j_2,n}} \right)^2 \right]^{\frac{1}{2}} \quad (16)$$

$$\Delta x_{i,j_1,n} = x_{i,j_1,n} - x_{i-1,j_1,n} \quad \Delta x_{i,j_2,n} = x_{i,j_2,n} - x_{i-1,j_2,n}$$

VCED between sample j_1 and j_2 is

$$d_3(j_1, j_2) = \left[\sum_{i=1}^t \left(\frac{\overline{x_{i,j_1}}}{S_{i,j_1}} - \frac{\overline{x_{i,j_2}}}{S_{i,j_2}} \right)^2 \right]^{\frac{1}{2}} \quad (17)$$

$$\overline{x_{i,j_1}} = \frac{1}{s} \sum_{n=1}^s x_{i,j_1,n} \quad \overline{x_{i,j_2}} = \frac{1}{s} \sum_{n=1}^s x_{i,j_2,n}$$

$$S_{i,j_1} = \frac{1}{s-1} \sum_{n=1}^s (x_{i,j_1,n} - \overline{x_{i,j_1}})^2$$

$$S_{i,j_2} = \frac{1}{s-1} \sum_{n=1}^s (x_{i,j_2,n} - \overline{x_{i,j_2}})^2$$

CED between sample j_1 and j_2 is

$$d(j_1, j_2) = \alpha d_1(j_1, j_2) + \beta d_2(j_1, j_2) + \gamma d_3(j_1, j_2) \quad (18)$$

4 Clustering analysis

This paper uses the data of digital inclusive financial index of Peking University, and the selected indicators are "breadth of digital financial coverage", "depth of use of digital finance" and "degree of digitalization of inclusive finance". The index is the first-level index calculated by the Institute of Digital Finance Peking University based on 24 specific indicators. Based on the digital inclusive financial index data of each province from 2012 to 2018, the annual digital inclusive financial level of 31 provinces in China is clustered by systematic clustering method, and

then the digital inclusive financial level of each province for 7 years is comprehensively considered. the multi-index panel data clustering method is used for clustering. In this paper, the software used for the clustering of cross-sectional data is python, multi-index panel data clustering, and the software used is R. The clustering results are shown in Table 2 and Figure 2 below.

Table 2. Classification results of 31 provinces by two clustering methods

		First	Second	Third	Fourth
Cross-sectional Data sets	2012	3, 25, 27, 28	2, 4, 6, 7, 15, 19, 23, 24, 26, 29, 30	5, 11, 13, 14, 16, 17, 20, 21, 22	0, 1, 8, 9, 10, 12, 18
	2013	2, 3, 4, 6, 7, 15, 19, 23, 24, 25, 26, 27, 28, 29, 30	5, 11, 13, 14, 16, 17, 20, 21, 22	1, 9, 12, 18	0, 8, 10
	2014	4, 23, 25, 27, 28	2, 3, 6, 7, 15, 19, 24, 26, 29, 30	1, 5, 9, 11, 12, 13, 14, 16, 17, 18, 20, 21, 22	0, 8, 10
	2015	4, 23, 27, 28, 29	2, 3, 6, 7, 15, 19, 24, 25, 26, 30	5, 11, 13, 14, 16, 17, 20, 21, 22	0, 1, 8, 9, 10, 12, 18
	2016	2, 6, 17, 25, 27, 28, 29, 30	3, 4, 7, 15, 19, 23, 24	1, 5, 11, 13, 14, 16, 20, 21, 22, 26	0, 8, 9, 10, 12, 18
	2017	3, 4, 23, 25, 27, 28, 29, 30	2, 5, 6, 7, 11, 13, 14, 15, 17, 19, 20, 21, 22, 24, 26	1, 9, 12, 16, 18	0, 8, 10
	2018	3, 4, 6, 7, 23, 25, 27, 28, 29, 30	2, 5, 11, 13, 14, 15, 17, 19, 20, 21, 22, 24, 26	1, 9, 12, 16, 18	0, 8, 10
Multi-indicator panel data		2, 3, 4, 6, 7, 15, 19, 23, 24, 25, 26, 27, 28, 29, 30	5, 11, 13, 14, 17, 20, 21, 22	1, 9, 12, 16, 18	0, 8, 10

0 Beijing 1 Tianjin 2 Hebei 3 Shanxi 4 Inner Mongolia
 5 Liaoning 6 Jilin 7 Heilongjiang 8 Shanghai 9 Jiangsu
 10 Zhejiang 11 Anhui 12 Fujian 13 Jiangxi 14 Shandong
 15 Henan 16 Hubei 17 Hunan 18 Guangdong 19 Guangxi
 20 Hainan 21 Chongqing 22 Sichuan 23 Guizhou 24 Yunnan
 25 Tibet 26 Shaanxi 27 Gansu 28 Qinghai 29 Ningxia
 30 Xinjiang

As shown in figure 2, the digital inclusive finance of 31 provinces in China can be divided into four categories. The first category is that the areas with poor development level of digital inclusive finance are Tibet, Qinghai, Gansu, Guizhou, Xinjiang, Hebei, Jilin, Yunnan, Henan, Heilongjiang, Guangxi, Inner Mongolia, Ningxia, Shanxi, Shaanxi. The second category of digital inclusive financial development level of general areas are Hainan, Chongqing, Liaoning, Shandong, Anhui, Jiangxi, Hunan, Sichuan; the third category of digital inclusive financial development level of Tianjin, Hubei, Jiangsu, Fujian, Guangdong; the fourth category of digital inclusive financial development level of the best areas are Beijing, Shanghai, Zhejiang. As can be seen from Table 2, the clustering results of cross-sectional data show that individual regions belong to different categories

indifferent years, resulting in a lack of stability. Multi-index panel data clustering considers all years comprehensively, which excludes not only the chance of clustering results in a single year, but also the volatility of clustering results in individual cities, so as to make the

results more reliable and stable. in the short term, corresponding policies and measures can be formulated according to the clustering results to reduce the regional differences in the development level of digital inclusive finance in China.

Clustering

Fig. 2. Clustering result graph of multi-index panel data

5 Conclusion and suggestion

In this paper, the development level of digital inclusive finance in 31 provinces is basically divided into four categories according to the breadth of digital financial coverage, the depth of digital financial use and the degree of digitalization of inclusive finance. It is not difficult to see that the regional differences of digital inclusive finance in China are large, and there are few well-developed provinces, and most provinces, especially in the western region, have a low level of development.

How to improve the overall level of digital inclusive financial development in China according to the influencing factors such as the level of economic development, population density and Internet coverage in these four types of areas, reducing regional differences is an important content that our government, enterprises and academic circles should pay attention to. This paper hopes that through the cluster analysis of the development level of digital inclusive finance in various regions of the country, each region can have a clear understanding of its own positioning, and make the connections and characteristics of similar areas more clearer. Let people from all walks of life pay attention to these areas. At the same time, the government should also make great efforts to make up for its shortcomings so as to speed up the development of areas with a poor level of digital inclusive financial development.

Acknowledgments

We are grateful to the editors and the two anonymous referees. Their comments and suggestions have greatly helped us to improve this paper. This research has been

supported by China Postdoctoral Science Foundation (2019M651540) and General Subject of Philosophy and Social Science Planning in Shanghai(2019BGL020).

References

1. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.
2. Corrado, G., & Corrado, L. (2017). Inclusive finance for inclusive growth and development. *Current opinion in environmental sustainability*, 24, 19-23.
3. Guo, F., Wang, J.Y., Wang, F., Kong, T. and Cheng, Z.Y. (2020) Measuring the Development of China's Digital Inclusive Finance: Index Compilation and Spatial Characteristics. *Economics*, No. 4, 1401-1418.
4. Xiaoling, S. (2017). Empirical Analysis of Digital Inclusive Finance Bridging the Urban-rural Residents' Income Gap. *Finance & Economics*, 06.
5. Bing-yun, Z. H. E. N. G. (2008). The clustering analysis of multivariable panel data and its application. *Application of Statistics and Management*, 3, 265-270.
6. Nezhadshahmohammad, F., Pourrahimian, Y., & Aghababaei, H. (2018). Presentation of a multi-index clustering technique for the mathematical programming of block-cave scheduling. *International Journal of Mining Science and Technology*, 28(6), 941-950.
7. Wang, Y., & Zuo, W. (2018). Research on Multi-index Grey Clustering Approach Based on Adaptive Weight for Panel Data. *Journal of Grey System*, 30(3).