Clustering analysis of the distribution characteristics of unobserved economic regions in China ——Based on multi-indicator panel data

Yang Fan^{1,*}

¹Beijing Jiaotong University, School of Economics and Management, Haidian, Beijing, China

Abstract: The existence of unobserved economy is one of the important factors affecting GDP calculation. This paper uses the provincial panel data from 2010 to 2019 in China, and adopts the method of principal component feature extraction to carry out cluster analysis on the multi-indicator panel data. This method preserves the dynamic characteristics of the panel data, calculates the comprehensive score of each eigenvalue, and gives weight to the eigenvalue by using the entropy method, so as to optimize the clustering results representing the eight indicators of the unobserved economy. Through the analysis, it is found that the regional development of China's unobserved economy is obviously different, and each type has different influencing factors. This result has important practical significance for different regions in China to formulate differentiated unobserved economic governance policies. This also helps to make better use of resources and develop an energy-saving economy.

1 introduction

Gross domestic product is a core indicator that reflects the economic development of a country. However, as for the calculation results of GDP, there are controversies in the society that GDP is "overestimated" or "underestimated," and GDP growth does not match the current level of economic development. The existence of the unobserved economy has affected the comprehensiveness of GDP accounting and has had an impact on all aspects of our social life. Research on the regional distribution characteristics of the unobserved economy will help to form a differentiated unobserved economy governance policy.

This paper uses the method of principal component feature extraction to process multi-indicator panel data by reading related literature, retains the dynamic characteristics of the panel data, calculates a comprehensive score for each feature, and then uses the entropy method to perform systematic clustering, and analyzes the representative unobserved economy. The eight indicators are clustered and the corresponding conclusions are drawn.

2 Feature extraction of panel data

2.1 Characteristics of multi-indicator panel data

Panel data (panel data) is also called time series-cross-section data, which refers to the data obtained by the same cross-sectional units (such as households and companies)

in different periods. It also has the characteristics of time and space dimensions, and can reveal the dynamic characteristics of the research object.

The basic idea of cluster analysis is to classify a batch of samples or variables according to their characteristics without prior knowledge. The individual characteristics within these categories are similar, and the individual differences between different categories are very large.

Multi-index panel data has many observations in the three dimensions of sample size, index, and time. Suppose there are q samples in the multi-index panel data, each sample has p indicators, and each individual recording time is T , then each data point is used, where $i=1,2,\dots,q$; $j=1,2,\dots,p$; $t=1,2,\dots,T$.

2.2 Feature extraction of multi-index panel data

2.2.1 Standardization of panel data

In order to eliminate the impact brought by the order of magnitude, the panel data is first standardized. Because the data in this article are statistical data, the average method is used to process the data [1].

$$Z_{ij}(t) = \frac{x_{ij}(t)}{\bar{x}_j} \quad (1)$$

Among them, $\bar{x}_j = \frac{\sum_{i=1}^N \sum_{t=1}^T x_{ij}(t)}{NT}$. The mean of each variable after the change is 1, Standard deviation is $Var(X_j^*) = \frac{1}{NT-1} \sum_{i=1}^N \sum_{t=1}^T \left[\frac{x_{ij}(t)}{\bar{x}_j} - 1 \right]^2 = \frac{Var(X_j)}{\bar{x}_j^2} =$

* Corresponding author: 529603660@qq.com

$$\left(\frac{\sigma_j}{\bar{x}_j}\right)^2.$$

For ease of presentation, the representative is still used $x_{ij}(t)$ to $z_{ij}(t)$.

Since this article is clustering between different provinces, the denominator of the averaging process is the mean of all samples of the same indicator, so that the differences between the samples can be preserved.

2.2.2 Feature value extraction of panel data indicators

For panel data $\{x_{ij}(t)\}$, Suppose there are q samples, each sample has p indicators, and the recording time of each individual is T [2].

Definition 1: The "absolute quantity" characteristic of the j -th indicator of sample i in the time dimension is denoted as:

$$AQF(F_{ij}) = \frac{\sum_{t=1}^T x_{ij}(t)}{T} \quad (2)$$

Definition2: The fluctuation characteristics of the j -th indicator of sample i in the time dimension are recorded as:

$$VF(F_{ij}) = \left[\frac{\sum_{t=1}^T (X_{ij}^*(t) - \bar{X}_{ij}^*)^2}{T-1} \right]^{\frac{1}{2}} \quad (3)$$

Among them, $\bar{X}_{ij}^* = \frac{\sum_{t=1}^T X_{ij}^*(t)}{T}$.

Definition 3: The "skewness" characteristic of the j -th indicator of sample i in the time dimension is recorded as:

$$SCF(F_{ij}) = \frac{(X_{ij}^*(t) - \bar{X}_{ij}^*)^3}{T(\sigma_{ij}^*)^3} \quad (4)$$

Among them $\sigma_{ij}^* = \left[\frac{\sum_{t=1}^T (X_{ij}^*(t) - \bar{X}_{ij}^*)^2}{N} \right]^{\frac{1}{2}}$. Represents

the standard deviation of the j -th index of individual i in the time dimension; $SCF(F_{ij})$ measures the symmetry of the index value of the sample i over the entire period. If it is greater than 0, it means that the value of this indicator is distributed to the right; otherwise, it is distributed to the left.

Definition 4: The "skewness" feature of the j -th indicator of sample i in the time dimension is denoted as

$$KCF(F_{ij}) = \frac{\sum_{t=1}^T (X_{ij}^*(t) - \bar{X}_{ij}^*)^4}{T(\sigma_{ij}^*)^4} - 3 \quad (5)$$

Definition 5: The "trend" characteristic of the j -th indicator of sample i in the time dimension is denoted as:

$$TF(F_{ij}) = \frac{\sum_{t=1}^T (X_{ij}^*(t) - \bar{X}_{ij}^*) \left(t - \frac{T+1}{2}\right)}{\sum_{t=1}^T \left(t - \frac{T+1}{2}\right)^2} \quad (6)$$

2.2.3 The secondary extraction of feature quantities

In the cluster analysis, in order to avoid the correlation of the feature value of each indicator, this paper uses

principal component analysis to reduce the dimension of the same feature value of different indicators to obtain the comprehensive score of each feature value.

Definition 6: $F_1, F_2, \dots, F_p$ is p dimensional index vector $AQF(F_i)$ extracted principal components, the variance contribution rate of the principal component F_k is $\alpha_k (k = 1, 2, \dots, p)$. Then the comprehensive score of the "absolute quantity" feature $AQF(F_{ij})$ after principal component dimensionality reduction is:

$$F_AQF(F_i) = \sum_k \alpha_k F_k$$

In the same way, the comprehensive scores of "volatility", "skewness", "kurtosis" and "trend" can be defined separately as

$$F_VF(F_i), F_SCF(F_i), F_KCF(F_i), F_TF(F_i).$$

2.2.4 The weighting of eigenvalues

In this paper, the eigenvalues after principal component analysis has different influences on individual differences, and the entropy method is used to assign corresponding weights $w_j (j = 1, 2, \dots, 5)$. Calculating the corresponding weight of each principal component features w_1, w_2, w_3, w_4, w_5 .

3 Empirical analysis

3.1 Index selection

The key point of the research in this paper is to select appropriate indicators to properly characterize the unobserved economy at the provincial level in China. On the basis of the previous research, this article further describes the unobserved economy in detail by reading domestic and foreign literature and selecting appropriate indicators. The explanation is as follows:

3.2 Data sources and descriptive statistics

This paper selects 31 provinces across the country as the cross-sectional unit, and collects 10 years of data for each province from 2010 to 2019. It should be noted that the National Bureau of Statistics of China has not released the government final consumption data of some provinces and cities in 2018 and 2019. Here we use the common "time series forecasting method" to supplement the government final consumption rate data.

1. Tax burden (X1). Scholars generally believe that the tax burden has a significant positive impact on the unobserved economy. This paper selects the proportion of total tax revenue to local GDP as a measure of the actual burden of the provincial-level economy [3].

2. Urban unemployment rate (X2). Domestic and foreign studies have shown that the unemployment rate is an important factor affecting the unobserved economy, and a higher unemployment rate also means to a certain extent that certain problems in the national real economy. It is worth noting that although the unemployment rate is the best indicator to measure the level of unemployed, China's official statistics currently only publish the "urban registered unemployment rate". In the absence of a more

suitable indicator, this article adopts the "urban registered unemployment rate". "Unemployment rate" is used as a measure of unemployment rate.

3. Government control (X3). The scale of the unobserved economy is closely related to the degree of government regulation, but the relationship between government regulation and the unobserved economy is uncertain. This article uses the indicator "government actual consumption/GDP" to measure the impact of government regulation on the unobserved economy.

4. Inflation rate (X4). The severity of inflation is measured by the inflation rate, which reflects the extent to which commodity prices continue to rise in a certain period of time. This article chooses the Consumer Price Index (CPI), which measures changes in the price levels of consumer goods and services, to measure inflation.

5. Number of self-employed persons (X5). The article by Bordignon and Zanardi pointed out that self-employed persons can easily avoid taxes by reducing their tax base and underreporting their personal income [4]. The number of private enterprises and self-employed persons in urban and rural areas registered over the years is used as a proxy for the proportion of self-employed persons.

6. Personal disposable income (X6). At present, China's income distribution is in a "gourd" pattern, with a large income distribution gap. For low-income groups, raising wages will reduce the frequency of their participation in microeconomic activities to a certain extent. Here we choose the disposable income of urban residents as an indicator to measure income.

7. The proportion of social security expenditure in GDP (X7). Dell'Anno believes that in areas with high tax rates, if the level of social security is higher, it may indirectly promote the improvement of social security. Unobserved economic growth [5]. This article chooses the proportion of local government fiscal social security expenditures to GDP to measure this indicator.

8. The level of power consumption (X8). Electricity consumption is a key factor that reflects the real economy [6]. Those engaged in unobserved economies also need electricity consumption. The higher power level also reflects the increase in the scale of the underground economy to a certain extent. This paper selects the power consumption of each province as a key indicator to express the level of power consumption.

3.3 Cluster analysis

Clustering the indicators of the unobserved economy, the weight of each principal component feature extracted by the entropy method is 0.239, 0.279, 0.168, 0.09, 0.224, respectively. It can be seen that the "absolute quantity" level, "volatility" level and "trend" level of the period have a relatively large contribution to the difference between samples, while the contribution of "skewness" and "kurtosis" is relatively small. The annual data has a larger growth trend as a whole, resulting in "skewness" and "kurtosis" having little effect on sample differences.

Perform Z-Score standardization on the "absolute quantity" feature, "fluctuation" feature, "skewness" feature, "kurtosis" feature and "trend" feature calculated

by the principal component analysis method to eliminate the magnitude of influence, and then multiply by Corresponding weights w_1, w_2, w_3, w_4, w_5 to obtain a new data set for analysis Q-type clustering using hierarchical clustering. According to the output of the software, a scatter plot of the aggregation coefficient changing with the number of classifications is drawn. When the number of classifications is 7, the curve becomes more stable.

Table 1 Panel data clustering results

Category	Clustering result
The first sort	Guangdong
Second category	Jiangsu, Zhejiang, Shandong
The third category	Tibet, Qinghai
The fourth category	Liaoning, Xinjiang, Heilongjiang, Ningxia, Jilin, Guizhou, Tianjin, Jiangxi, Guangxi, Shaanxi
The fifth category	Hebei, Henan, Hubei, Chongqing, Anhui, Fujian
Sixth category	Beijing Shanghai
Seventh category	Shanxi, Hunan, Sichuan, Inner Mongolia, Yunnan, Hainan, Gansu

From Table 1 we can see the category attribution of the provinces, cities, and autonomous regions across the country. On this basis, this article analyzes the performance differences of various categories in unobserved economically related variables.

According to the software output results, we know that the number of self-employed persons and power consumption indicators of category 1 are in the bottom, and the underground economy of this category has developed more serious; the indicators of category 2 are relatively evenly ranked, indicating a relative degree of development Relatively light; the third category of government control indicators, inflation rate indicators, and social security indicators ranked last, comprehensively judging that the level of development is relatively serious; the fourth category of taxation indicators, self-employed personnel indicators, and electricity consumption level indicators are at the upstream level. The degree of development is relatively low; most of the indicators in category 5 are ranked in the upper middle and upper reaches, indicating that the degree of development is relatively low; most indicators of category 6 are ranked relatively late, and the degree of development is judged to be relatively serious; the indicators of category 7 The ranking is at the mid-stream level, and it is judged that its unobserved economic degree is at the mid-stream level in other categories.

To sum up, the third category composed of Tibet and Qinghai is the place with the most severe unobserved economic development in all provinces, municipalities and autonomous regions in China. Shandong, Zhejiang and Jiangsu constitute the second category, Liaoning, etc. constitute the fourth category, and Hebei The 6 provinces, municipalities and autonomous regions (category 5) represented by Henan have relatively little unobserved economic development. The level of underground economic development in the 6th category composed of

Beijing and Shanghai and the 17 provinces, municipalities and autonomous regions represented by the three provinces of Yunnan, Guizhou, and Northeast China (category 3) and the 4 provinces, municipalities and autonomous regions represented by Gansu and Qinghai (category 5) are relatively low. weight. The unobserved economic development of the 7 provinces, municipalities and autonomous regions represented by Shanxi and Hunan are in the middle reaches of other categories. On the whole, the regional distribution of unobserved economic development in China still presents obvious unbalanced characteristics, and the underground economic development of most provinces, municipalities and autonomous regions in the past 10 years has not been significantly curbed.

3.4 Robustness test

The results of cluster analysis are often sensitive to changes in clustering methods, so this paper uses "average connection between groups" instead of "ward" to re-cluster. From the clustering results, the difference between the two methods is not significant. The characteristic of the "ward" method is that the sum of squared deviations between the same class is small, while the sum of squared deviations between different classes is large. The average connection method between groups contains more cases in individual classes, which may affect the significance of clustering.

4 Conclusion

This paper starts from the basic characteristic indicators that affect the development of unobserved economy, and adopts a clustering method suitable for multi-indicator panel data. This method fully considers the five characteristics of "absolute value", "volatility", "skewness", "kurtosis" and "trend" in the time dimension to measure the dynamic characteristics of panel data, and uses entropy The value method solves the above-mentioned feature weight problem. Through eight indicators, including taxation indicators and urban unemployment rate indicators, the relative level of underground economic development in China's provinces, municipalities and autonomous regions from 2010 to 2019 is described, and cluster analysis is performed. Through analysis, we can know that the development of China's unobserved economy presents obvious regional differences. These differences are driven by 8 different indicators, and the influencing factors of each type of region are different. This conclusion can enable us to better analyze the unobserved regional differences in the economy and provide different governance solutions for different regions, thereby promoting the rational allocation of resources and energy conservation and emission reduction.

References

1. Dai Dayang, Deng Guangming. Panel data clustering method based on principal component feature extraction [J]. Statistics and Decision, 2018, 34(21): 72-76.
2. Hou Diqing. Clustering method and application research based on panel data [D]. Nanjing University of Aeronautics and Astronautics, 2016.
3. Wang Yongxing, Jing Weimin. Regional development differentiation of China's underground economy: cluster analysis test based on multi-index panel data[J]. Nankai Economic Research, 2014 (06): 44-57.
4. Bordignon, M., Zanardi, A., 1997, "Tax Evasion in Italy", *Giornale Degli Economisti e Annali di Economia*, Vol.56, pp.169—210.
5. Dell'Anno, Roberto. THE SHADOW ECONOMY IN PORTUGAL: AN ANALYSIS WITH THE MIMIC APPROACH.[J]. Journal of Applied Economics, 2007, 10 (2):253-277.
6. Wang Yongxing. Has China's economic scale been overestimated-Based on the counter-evidence of the underground economy [J]. Nankai Economic Research, 2018 (06): 74-92.