ISSN 2616-5902 Vol. 3, Issue 10: 98-104, DOI: 10.25236/AJBM.2021.031018

Application of Clustering Analysis of Panel Data in Economic and Social Research Based on R Software

Weifeng Wang[1], Yugui Lu[2]

[1]School of Business, Hechi University, Yizhou Guangxi, China
[2]School of Mathematics and Physics, Hechi University, Yizhou Guangxi 546300, China

Abstract: *Panel clustering analysis has been widely used in the study of difference categories in the economic and social fields due to its advantages of analyzing data in both individual and time dimensions. After elaborating on the theory of Univariate panel data and Multivariable panel data clustering method, taking the 2008-2018 provincial panel data of Chinese residents' consumption index as an example, the R software implementation methods, procedures and specific steps of the single index and multiple indices of panel data clustering was investigated in this paper, hoping to provide a reference for the workers and researchers in economic and social statistics analysis.*

Keywords: *Univariate panel data, Multivariable panel data, Clustering analysis, R Software*

1. Introduction

The research objects of traditional cluster analysis are mostly cross-sectional data of different samples in a fixed period, and its research methods are basically mature, however, with the development of economy and society, static cluster analysis based on cross-sectional data can no longer meet the needs of economic and social analysis. Since the panel data contains data features in two dimensions, space and time, cluster analysis of panel data can analyze the category of the sample both statically and dynamically. Based on this, scholars have put forward a variety of methods for cluster analysis of panel data on the basis of cluster analysis of cross-sectional data [1-8]. However, at present, the results of cluster analysis of panel data mainly focus on theoretical research, while the research on software implementation is less. As a free, open and powerful statistical software, R software has been favored by more and more statisticians and researchers. However, at present, the application of R software in cluster analysis is still limited to sectional data. Based on this, in order to provide some reference for the researchers of economic and social statistical analysis, this paper studies the realization method of R software for cluster analysis of Univariate panel data and Multivariable panel data based on the principle of panel data clustering method.

2. Clustering method of panel data

Panel data can be classified into single-indicator panel data and multi-indicator panel data. For the cluster analysis of Univariate panel data, the Univariate panel data can be transformed into a two-dimensional data table, and then the cluster analysis can be completed directly by referring to the cluster analysis method of cross-section data [1]. However, the cluster analysis method of Multivariable panel data is more complex because of its complexity of data form. The following article will briefly introduce the clustering methods of single indicator panel data and multi-indicator panel data.

2.1. Clustering method of Univariate panel data

When cluster analysis is carried out on single index panel data, the time dimension can be regarded as the index dimension of cross-section data, and then cluster the single index panel data according to the cluster analysis method of cross-section data [1]. If the total number of samples is *N*, and the time length is *T*, then $X_i(t), i = 1, \cdots, N,\ t = 1, \cdots, T$ means the observed value of the first sample at the time, and its data matrix form is as follows:

$$\begin{bmatrix} X_1(1) & X_1(2) & \cdots & X_1(T) \\ X_2(1) & X_2(2) & \cdots & X_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ X_N(1) & X_N(2) & \cdots & X_N(T) \end{bmatrix}$$

Steps of Univariate panel data clustering:

(1) The single index panel data is converted into the form of cross-section data, that is, the whole time *T* is regarded as the cross-section data *T* indexes;

(2) Cluster analysis of Univariate panel data;

(3) Determine the number of classifications according to the cluster tree diagram;

(4) Give the clustering results.

2.2. Clustering method of Multivariable panel data

Multivariable panel data contains three dimensions of sample (space), index, and time at the same time. It is a three-dimensional dynamic data. When performing sample clustering, the relevant definitions of distance between samples and distance between classes in multivariate statistics cannot be directly applied. Therefore, Zelei Xiao (et al.) [2] proposed to use principal component analysis to reduce the dimension of the original panel data, that is, to carry out principal component analysis on the cross-sectional data at each time, and then calculate the comprehensive score of each sample at each time with the variance contribution rate as the weight, thus obtaining the Univariate panel data containing only one index (comprehensive score), and finally using the Univariate panel clustering method to realize clustering. However, when dimension reduction is carried out separately, different cross-section data have different hyperplanes. Juan Ren [3] puts forward an improved method, which expands the number of samples from *N* to *NT*, that is, expands the three-dimensional data matrix of panel data into a two-dimensional data matrix according to the index order, and then carries out factor dimension reduction on the expanded data only once. The cluster analysis method of Multivariable panel data studied in this paper is the Multivariable panel data cluster analysis method proposed by Juan Ren in Literature [3].

Suppose there are *N* samples and *p* indexes in the whole population, and the time length is *T*. The observed value of the *i* index of the *j* sample at time *t* is $X_{ij}(t)$. The three-dimensional data matrix X^m is as follows.

$$\begin{bmatrix} X_{11}(1) & X_{12}(1) & \cdots & X_{1p}(1) & X_{11}(2) & X_{12}(2) & \cdots & X_{1p}(2) & \cdots & X_{11}(T) & X_{12}(T) & \cdots & X_{Np}(T) \\ X_{21}(1) & X_{22}(1) & \cdots & X_{2p}(1) & X_{21}(2) & X_{22}(2) & \cdots & X_{2p}(2) & \cdots & X_{21}(T) & X_{22}(T) & \cdots & X_{Np}(T) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ X_{N1}(1) & X_{N2}(1) & \cdots & X_{Np}(1) & X_{N1}(2) & X_{N2}(2) & \cdots & X_{Np}(2) & \cdots & X_{N1}(T) & X_{N2}(T) & \cdots & X_{Np}(T) \end{bmatrix}$$

The three-dimensional data matrix X^m is expanded into a two-dimensional data matrix X^V according to the index dimension as follows:

$$X^V = \left[X_1, X_2, \cdots, X_p \right]$$

Among them, $X_j = \left[X_{1j}(1) \quad \cdots \quad X_{1j}(T) \quad \cdots \quad X_{Nj}(1) \quad \cdots \quad X_{Nj}(T) \right]$.

The specific thinking and calculation steps of Multivariable panel data are as follows:

(1) Expanding the number of samples from *N* to *NT*, that is, transforming the data matrix X^m into X^V;

(2) Standardize the data, i.e.

$$Z_{ij}(t) = \frac{X_{ij}(t) - \overline{X}_j}{\mathrm{var}(X_j)}$$

Among them, $\overline{X}_j = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{ij}(t)$, $\mathrm{var}(X_j) = \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{ij}(t) - \overline{X}_j)^2$

(3) Applicability test of factor analysis, that is, through KMO test, analyze the correlation between variables to verify whether the data is suitable for factor analysis;

(4) Determination of common factors, that is, solving the factor load matrix by principal component method and determining the number of common factors;

(5) Factor rotation, i.e. orthogonal rotation of the common factor selected in (4) according to the maximum variance;

(6) Calculate the comprehensive factor score, that is, after calculating the common factor score after rotation, take the contribution rate of variance after rotation as the weight to calculate the comprehensive score of each sample at each time;

(7) Complete the cluster analysis of Multivariable panel data according to the clustering method of the Univariate panel data for the comprehensive scores obtained.

(8) According to the clustering tree, the appropriate classification number is determined and the classification result is given.

3. R software implementation of cluster analysis of Univariate panel data.

3.1. Data sources

In this paper, the annual data of per capita consumption expenditure of education and culture in 31 provinces (autonomous regions and municipalities) in China from 2008 to 2018 are selected to study the R software implementation of Univariate panel data clustering. The data comes from the column of per capita consumption expenditure of residents in different areas of people's life in China Statistical Yearbook in each year.

3.2. Software implementation

The R software implementation steps of Univariate panel data clustering are as follows:

(1) Convert the Univariate panel data into the form of cross-sectional data, that is, convert the panel data of per capita expenditure on education and culture in each region of China from 2008 to 2018 into cross-sectional data with each year as the column and each region as the row, and store the data as "Univariate panel data. csv".

(2) Import Univariate panel data.

```
>setwd("F:/Rdata") # Set work path

>data=read.csv("Univariate panel data.csv") # Import data

>mydata=as.matrix(data[,-1],31,11)# Drop column 1(region)

# Use column 1 of dataset "data" as the row name of the new dataset "myData"

>rownames(mydata)=data[,1]
```

(3) Clustering of Univariate panel data. The distance matrix is calculated by the function "dist()",and then cluster analysis of Univariate panel data is realized by the function "hclust()".

```
# The distance matrix is calculated by Euclidean distance.

> distance=dist(mydata,method="euclidean")

#Class distance is calculated by class average method "Average".

>hc=hclust(distance,method="average")
```

(4) Output the results of cluster analysis. Determine the classification number k according to the drawn tree map (Figure 1 shows that k=4 is more appropriate), and then use rect.hcluster () to mark the classification results in the tree map so as to better read the classification results. Finally, cutree() is used to output the clustering results

```
# When the parameter "hang" is set to -1, the tree is drawn from the bottom

>plot(hc,hang=-1)
```

```
# "k " is the number of classes, and " border" can set the color of the rectangle
>re=rect.hclust(hc,k=4,border="red") #The results are shown in Figure 1
# Using "cutree()" to cut the clustering result from the tree
> H= cutree(hc,k=4);H #The results are shown in Table 1
```

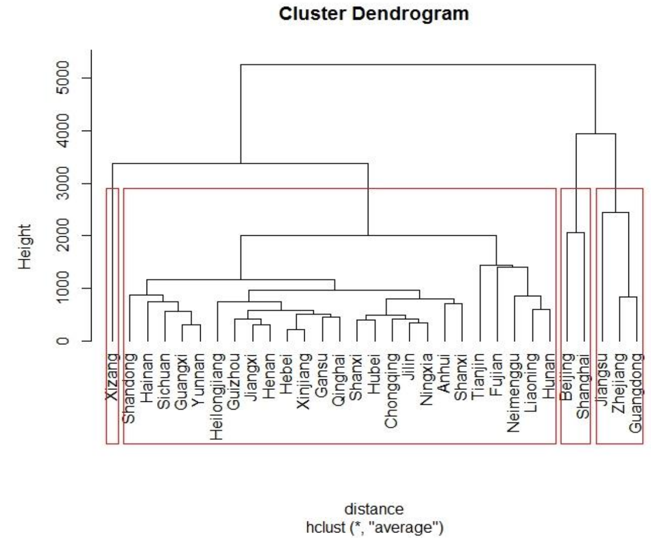

Figure 1: Cluster tree diagram of Univariate panel data

Table 1: Cluster results of Univariate panel data

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Beijing, Shanghai	Tianjin, hebei, Shanxi, Neimenggu, Liaoning, Jilin, Heilongjiang, Anhui, Fujian, Jiangxi, Shandong, Henan, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Guizhou, Yunnan, Shaanxi, Gansu, Qinghai, Ningxia, Xinjiang	Jiangsu, Zhejiang, Guangdong	Xizang

(5) Difference analysis of classes. In order to analyze the differences in educational and cultural consumption expenditure, using"aggregate()" to calculate the class centers of each class.

```
# Merge the classification results H into the original dataset "mydata"
>mydata1=cbind(mydata,H)
# sort dataset "mydata1" by classification result H
>mydata1[order(mydata1$H),]
# calculate the average of the classes
>aggregate(mydata1[,1:11],by=list(mydata1[,12]),FUN=mean)
```

Table 2: Cluster Centers

Cluster	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Cluster1	2833.2	3133.3	3748.7	3526.6	3709.9	4053.5	3289.9	3676.4	3930.6	4301.3	4524.4
Cluster2	1084.6	1296.6	1549.7	1456.4	1597.2	1912.6	1373.2	1542.9	1726.2	1882.1	2046.2
Cluster3	2124.6	2664.1	2930.7	2719.8	3009.5	3120.4	2124.1	2323.1	2586.7	2737.6	2788.3
Cluster4	966.7	1062.8	1230.9	514.4	550.5	1551.3	266.7	314.1	370.1	441.6	609.3

Table 2 shows that cluster 1(Beijing, Shanghai) has the largest per capita expenditure on education and culture, followed by cluster 2 and cluster 3, the least is the cluster 4(Xizang).

4. R software realization of cluster analysis of Multivariable panel data

4.1. Data sources

This paper selects the annual data of eight major consumption expenditures of residents in 31 provinces (autonomous regions and municipalities) in China from 2008 to 2018, in order to summarize the R realization process of Multivariable panel data clustering analysis. The data comes from *China Statistical Yearbook* of each year.

4.2. Software implementation

The R software implementation steps of Multivariable panel data clustering are as follows:

(1) Download and load the psych package.

```
>install.packages("psych")
>library(psych)
```

(2) Calculate the value of KMO to determine that the data is suitable for factor analysis.

```
>setwd("F:/Rdata")
>data2=read.csv("Multivariable panel data.csv")
# Drop column 1 (year) and column 2 (region).
>mydata2=data2[,3:10]
>N=31# Set the number of samples to N.
>T=11# Set the length of time to T.
>p=8#Set the index number to 8
> KMO(mydata2)# Calculate the value of KMO
```

Note: In the KMO output result, the value displayed in the Overall MSA column is the KMO value. The results show that the KMO value of the Multivariable panel data is 0.83, indicating that it is suitable for factor analysis.

(3) Determine the number of principal components. Using “principal()” in psych package to realize principal component analysis of Multivariable panel data.

```
>pc=principal(mydata2,nfactors = p,rotate='none',score=TRUE)
#Extract the factor load and cumulative contribution rate of pc.
> pc$loadings
```

Note: Principal component analysis does not perform factor rotation, so the factor rotation parameter rotate is set to "None". In the output of Loadings, "SS loadings" column is the eigenvalues of common factors, "Proportival Var" column is the variance contribution rate of each eigenvalue, and "Cumulative Var" column is the cumulative contribution rate. The results showed that the eigenvalues of the principal components are 5.441, 0.986, 0.703, 0.382, 0.173, 0.12, 0.11, 0.09, and the cumulative contribution rates are 0.68, 0.80, 0.89, 0.94, 0.96, 0.98, 0.99, 1.00, respectively. When extracting the first principal component and the second principal component, the cumulative contribution rate is 80%, and the second characteristic value (0.986) is close to 1, meeting the criterion of eigenvalue greater than 1 and the cumulative contribution rate greater than 80%, so it is appropriate to extract two principal component.

(4) Factor rotation. Using “principal()” to realize factor analysis based on principal component method.

```
#Factor rotation by " varimax "
>rc=principal(mydata2,nfactors=2,rotate='varimax',score=TRUE)
#Extract the eigenvalues of factor analysis
>rc$loadings
```

The output shows that the eigenvalues of the common factors after the factor rotation are 3.515 and 2.912.

(5) Calculate the score of comprehensive factor. Using the formula $F=\sum_{i=1}^{m}\lambda_i \mathrm{F}_i \Big/ \sum_{i=1}^{m}\lambda_i$ to calculate the comprehensive factor score (λ_i is the characteristic value of the i common factor and F_i is the i common factor score)

```
# Convert the common factor score to a matrix
>rc.score=as.matrix(rc$score,N,2)
# Define eigen values for common factors
>rc.loading=c(3.515,2.912)
# Calculate the score of comprehensive factor
>RC=rc.score%*%(rc.loading/sum(rc.loading))
```

(6) Cluster the comprehensive factor score data matrix RC by using the Univariate panel clustering method, and complete the clustering of Multivariable panel data.

```
>RCC=matrix(RC,N,T)
>row.names(RCC)= data2[1:N,2]
>distance1=dist(RCC,method="euclidean")
>hc1=hclust(distance1,method="average")
>plot(hc1,hang=-1) # Cluster diagram
# Cluster diagram shows that it is suitable to divide into five categories.
>re=rect.hclust(hc1,k=5)
>HM=cutree(hc1,k=5);HM
```

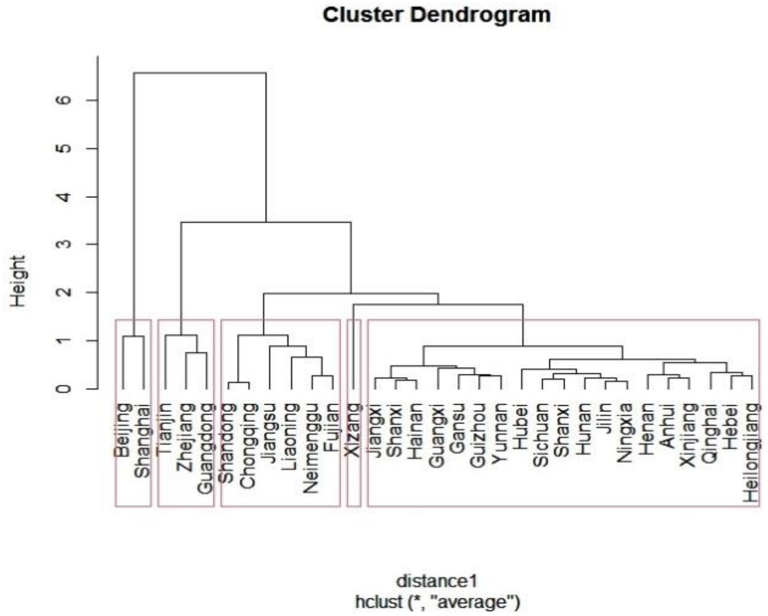

Figure 2: Cluster tree diagram of Multivariable panel data

(7) Difference analysis of classes.

Table 3: Cluster Centers

Cluster	Food	Clothing	Housing	Living	Transportation	Education	Medical	Other
Cluster 1	8089.6	2028.7	6476.5	1757.1	3413.7	3702.5	2335.6	1093.2
Cluster 2	7198.4	1590.7	3598.2	1245.2	2965.1	2597.8	1656.1	699.1
Cluster 3	4321.2	1215.5	1929.6	807.6	1527.0	1488.6	1123.2	371.1
Cluster 4	5439.8	1627.3	2548.0	1086.1	2061.8	1927.9	1366.7	549.7
Cluster 5	4656.6	1105.9	1156.0	456.6	914.3	716.2	389.9	384.4

Table 3 shows that the consumption level of cluster1 {Beijing, Shanghai} is the highest, followed by cluster 2 {Tianjin, Zhejiang, Guangdong}, the least is the cluster 5(Xizang).and the consumption level of cluster1 and cluster2 are significantly higher than the other three clusters.

5. Conclusion

Panel data contains data characteristics in two dimensions: space and time. Cluster analysis of panel data can analyze the category of samples from both static and dynamic aspects. Therefore, the clustering results of panel data are more realistic than cross-sectional data. Therefore, this paper summarizes the method of cluster analysis of panel data with R software through two examples, and explains the commonly used packages and functions to improve the readability of the code. From the perspective of the R implementation process of cluster analysis, Multivariable panel data is more complicated than Univariate panel data, but it is also more practical.

References

[1] Bingyun Zheng. (2008). The Clustering Analysis of Multivariable Panel Data and Its Application. Application Statistics and Management, 27(02), 265-270.
[2] Zelei Xiao, Bangyi Li, Sifeng Liu. (2009). The Discussion on the Clustering Way Based on the Multi-dimensional Panel Data and Empirical Analysis. Application Statistics and Management, 28(05), 831-838.
[3] Juan Ren. (2013). Fusion Clustering Analysis of Multivariable panel data. Application Statistics and Management, 32(01), 57-67.
[4] Jianping Zhu, Minken Chen. (2007). The Cluster Analysis of Panel Data and Its Application. Statistical Research, 24(04), 11-14.
[5] Yinguo Li, Xiaoqun He. (2010). Panel Data Clustering Method and Application. Statistical Research, 27(09), 73-79.
[6] Zedong Wang, Guangming Deng. (2019). Discussion on Panel Data Clustering Based on Trend Distance. Statistics and Decision, 35(08), 35-38.