**DOI: 10.23977/ferm.2022.050406**
**ISSN 2523-2576 Vol. 5 Num. 4**

# *Research on bank efficiency evaluation based on principal component analysis and cluster analysis*

**Minghui Fan, Shixing Han**

*College of Engineering, Tibet University, Lhasa 850000, China*

***Keywords:*** Principal component analysis, Cluster analysis, Gravel plot, Heat map

***Abstract:*** As the trend of global economic integration intensifies, the challenges faced by the financial sector are becoming more and more evident. Many banks are experiencing a surge in non-performing loans and bad debts, leading to bank indebtedness and even bankruptcy, for which it is crucial to evaluate the efficiency and analyze bank failures of banks in other countries so as to avoid losses to our national economy. This paper establishes a mathematical model to evaluate each bank and conducts an in-depth analysis of bank failures. This paper firstly preprocesses the basic data and eliminates the banks with high percentage of missing values. A bank efficiency evaluation model based on principal component analysis and cluster analysis is established, and 13 main indicators are obtained and evaluated for bank efficiency by using gravel diagram, two-dimensional distribution diagram of factor load quadrant and heat map, and then the line connecting the central values of the 13 indicators is used as the dividing line of bank failure efficiency by using cluster analysis.

## 1. Introduction

### 1.1 Background

Banks play an important role not only in the process of economic and social development of the country, but also in the life of the people and the development of the society[1]. The state of bank operations determines the efficiency of economic resource allocation and the condition with the real economy, while also affecting the state of society, and even the country's financial system. With the rapid development of society, the financial sector has been subject to tremendous impact, and many international banks are facing yearly increase in the frequency of failure compared to domestic ones. Domestic and foreign scholars, members of the financial community, and domestic and foreign government officials have paid considerable attention to this event of international bank failures, which has triggered a series of widespread concerns[2].

Among the failed banks, in addition to a large number of small and medium-sized banks, some large bank pages have suffered an existential crisis because of unsecured business and other reasons In order to prevent domestic banks from facing this situation, it is particularly important to collect experience and take targeted measures to cope with it by studying bank efficiency and evaluating it, analyzing the causes of bank failures[3].

## 1.2 The main work of this paper

Collating appropriate input-output data is beneficial for understanding the bank's operations and thus making better analysis. Firstly, we pre-process the data and eliminate the bank data with many missing values. The problem dealt with in this paper belongs to the problem of extracting data, for solving this kind of problem, we establish the bank efficiency evaluation model based on principal component analysis and cluster analysis method, use the gravel diagram, factor load quadrant two-dimensional distribution diagram and heat map to analyze to extract the main indicators and evaluate the bank efficiency, then use the cluster analysis, divided into two categories, to solve the cut-off line of bank failure efficiency, the specific idea is shown in Figure 1.

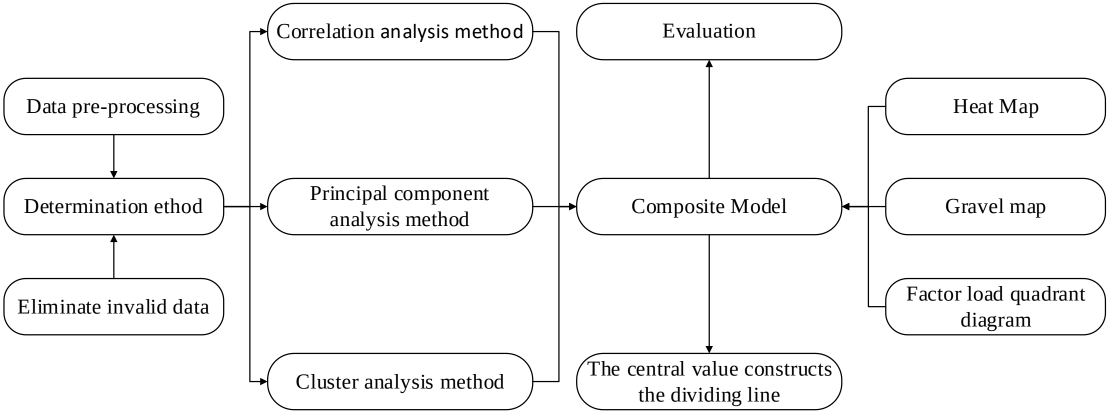

Correlation analysis method
Evaluation
Data pre-processing
Heat Map
Determination ethod
Principal component analysis method
Composite Model
Gravel map
Eliminate invalid data
Factor load quadrant diagram
Cluster analysis method
The central value constructs the dividing line

Figure 1 Flowchart of ideas

## 2. Model Assumptions and Notation

### 2.1 Model Assumptions

1) assuming that the data underlying the 64 indicators of the bank are true and reliable.

2) assuming that the factors of bank failure depend on the given 64 indicators and that no other factors influence them.

3) assuming that the data excluded by data processing do not affect the analysis of the problem

4) assuming that the most representative banks have common characteristics.

### 2.2 Definition and Symbol Description

| Symbol Definition | Symbol Description |
|---|---|
| $\bar{x}_j$ | Mean of the jth indicator in the sample used |
| $D$ | Indicator distance from the bank failure efficiency cutoff |
| $D$ | Number of input parameters |
| $n$ | Number of fuzzy subsets |
| $W$ | Bank indicator weighting values |

## 3. Model building and solving

### 3.1 Development of a bank efficiency analysis model based on principal component analysis and cluster analysis

In the study of the efficiency evaluation of each bank in Poland, it is necessary to classify 64 types of indicators and dig out the appropriate input and output indicators, and the principal component analysis and cluster analysis just achieve the requirements by comprehensive analysis, so we establish the bank efficiency evaluation model based on principal component analysis and cluster analysis to study[4].

### 3.1.1 Principal component analysis method

Since the given data lacked certain rationality, and the principal component analysis method was to transform the original indicators into a few comprehensive indicators for analysis through dimensionality reduction processing, sending out uncertainty. In addition, the original indicators in this question reach 64, with a large number of indicators and complex correlations, and should be processed by dimensionality reduction on the basis of retaining the information of the original indicators to the maximum extent, so as to obtain representative indicators.

(1) Standardization of indicators

The 64 bank failure indicators are standardized so that the mean of each indicator is 0 and the variance is 1, and the differences in magnitude and order of magnitude of each indicator value are cleared by the formula.

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{v\left(x_j\right)}}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, p \tag{1}$$

where the normalization matrix is.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{2}$$

(2) Correlation Analysis

Correlation coefficient matrix.

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{np} \end{pmatrix} \tag{3}$$

The correlation coefficient, calculated as.

$$r_{ij} = \frac{\sum_{i=1}^{n} \left|x_{ki} - \bar{\bar{x}}_i\right| \left|x_{kj} - \bar{\bar{x}}_j\right|}{\sqrt{\sum_{k=1}^{n} \left(x_{ki} - \bar{\bar{x}}_i\right)^2 \left(x_{kj} - \bar{\bar{x}}_j\right)^2}} \tag{4}$$

(3) Calculate the characteristic root

Make $R - \lambda_p I = 0$.

$$\begin{pmatrix} r_{11}-\lambda_1 & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22}-\lambda_2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{np}-\lambda_p \end{pmatrix} = 0 \tag{5}$$

According to the above equation, the eigenvalues of each index and the corresponding eigenvectors are derived, and the eigenvalues are sorted from largest to smallest.

(4) Calculate contribution margin

The variance contribution of the principal component $y_i$.

$$\varepsilon_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{6}$$

Then, the cumulative contribution of the mth principal component is.

$$\sum_{i=1}^{m} \varepsilon_i = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{7}$$

In the above equation, the representative ability of $y_i$ to $x_p$ information is proportional to the value of $\varepsilon_i$ and the cumulative contribution is not less than 86%.

### 3.1.2 Cluster analysis methods

Cluster analysis is a common data analysis method that classifies indicators according to their characteristics, thus reducing the number of indicators studied, has a strong scientific nature, and avoids the influence of perceived factors on qualitative analysis[5].

(1) Calculating Euclidean distance

In order to analyze more reasonable data on bank input and output indicators, we obtained four principal component base data based on principal component analysis, and used Euclidean distance to characterize the distance D between regional samples for division. According to the magnitude of the sample distance, the dividing line of bank failure efficiency is obtained. The calculation formula is.

$$\text{Distance}\left(O_i, O_j\right) = \sqrt{\sum_{k=1}^{n} \left(O_{ik} - O_{jk}\right)^2} \tag{8}$$

One of the methods with the relative efficiency between the units of input and output, the basic principle is as follows.

$$\begin{cases} \min \theta_0^t \\ s.t \\ \sum_j \lambda_j x_{i,j}^t \le x_{i,0}^t \theta_0^t, i = 1, 2, \cdots, m \\ \sum_j \lambda_j y_{r,j}^t \le y_{r,0}^t, r = 1, 2, \cdots, l \\ \lambda_j \ge 0 \\ j = 1, 2, 3, \cdots, N \end{cases} \tag{9}$$

Productivity changes are further analyzed by decomposition indices, but also the efficiency changes in different stages of the decision unit can be evaluated. The specific calculation formula is.

$$M_{t,t+1} = \left[ \frac{D^t\left(x^{t+1}, y^{t+1}\right)}{D^t\left(x^t, y^t\right)} \times \frac{D^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D^{t+1}\left(x^t, y^t\right)} \right]^{0.5} \tag{10}$$

Where $M_{(t,t+1)}>1$, then it can be stated that the productivity of the evaluated decision unit is increased from stage t to stage t+1; $M_{(t,t+1)}<1$, then the productivity is decreased. The model is calculated as.

$$\begin{cases} D^t\left(x^{t+1}, y^{t+1}\right) = \min \theta_0^t \\ \text{s.t.} \\ \sum_j \lambda_j x_{i,j}^t \le x_{i,0}^{t+1} \theta_0, i = 1, 2, \cdots, m \\ \sum_j \lambda_j y_{r,j}^t \le y_{r,0}^{t+1}, r = 1, 2, \cdots, l \\ \lambda_j \ge 0 \\ j = 1, 2, 3, \cdots, N \end{cases} \tag{11}$$

(2) Preliminary breakdown of indicators

The Malmqist Index can be divided into the following two specific components.

$$M_{t,t+1} = \frac{D^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D^t\left(x^t, y^t\right)} \left[ \frac{D^t\left(x^{t+1}, y^{t+1}\right)}{D^{t+1}\left(x^{t+1}, y^{t+1}\right)} \times \frac{D^t\left(x^t, y^t\right)}{D^{t+1}\left(x^t, y^t\right)} \right]^{0.5} \tag{12}$$

(3) Further decomposition of FS

Since the above FS description is the average frontier surface change, and the actual change is multi-faceted, from t+1 to t, the basic law of the frontier surface is partly rising and partly falling, if only a part of it is described, it seems simple, so the FS is divided into two parts, respectively.

$$FS_1 = \frac{D^t\left(x^t, y^t\right)}{D^{t+1}\left(x^t, y^t\right)} \tag{13}$$

$$FS_2 = \frac{D^t\left(x^{t+1}, y^{t+1}\right)}{D^{t+1}\left(x^{t+1}, y^{t+1}\right)} \tag{14}$$

Using (5) and (6) to represent the evaluated decision unit at moments t and t+1, respectively, these possible changes with respect to phases t and t+1 are.

$$FS = \left[ \frac{D^{t}\left(x^{t}, y^{t}\right)}{D^{t+1}\left(x^{t}, y^{t}\right)} \times \frac{D^{t}\left(x^{t+1}, y^{t+1}\right)}{D^{t+1}\left(x^{t+1}, y^{t+1}\right)} \right]^{0.5} = \left(FS_{1} \times FS_{2}\right)^{0.5} \tag{15}$$

### 3.2 Solution of bank efficiency analysis model based on principal component analysis and cluster analysis

Based on the model developed above, the solution was performed using SPSSPRO software, and the solution steps were.

1) Perform KMO and Bartlett's test to determine whether principal component analysis can be performed. For the KMO value: 0.8 on is very suitable for principal component analysis, between 0.7-0.8 is generally suitable, between 0.6-0.7 is less suitable, between 0.5-0.6 means poor, and under 0.5 means extremely unsuitable. For Bartlett's test ($p < 0.05$,strictly speaking $p < 0.01$), if the significance is less than 0.05 or 0.01 and the original hypothesis is rejected, it indicates that principal component analysis can be done, and if the original hypothesis is not rejected, it indicates that these variables may provide some information independently and are not suitable for principal component analysis;

2) By analyzing the variance interpretation table and the gravel plot, the number of principal components is determined. The variance interpretation table mainly looks at the contribution of the principal components to the explanation of the variables (which can be interpreted as how many principal components are needed to express the variables as 100%), and if it is too low (e.g., below 60%), the principal component data need to be adjusted; the role of the gravel plot is to confirm the number of principal components to be selected according to the slope of the decline of the eigenvalues The combination of these two can be used to confirm or adjust the number of principal components;

3) The importance of the hidden variables in each principal component can be analyzed by analyzing the principal component loading coefficients and the heat map;

4) Based on the principal component loadings, the spatial distribution of principal components is presented by means of quadrant plots by reducing the dimensionality of multiple principal components into two or three principal components. If 2 principal components are extracted, the 3D loading principal component scatter plot cannot be presented, and if 1 principal component is extracted, the principal component quadrant plot cannot be displayed;

5) Derive the principal component composition formula and weights by analyzing the component matrix;

6) Output the comprehensive score of principal component analysis method.

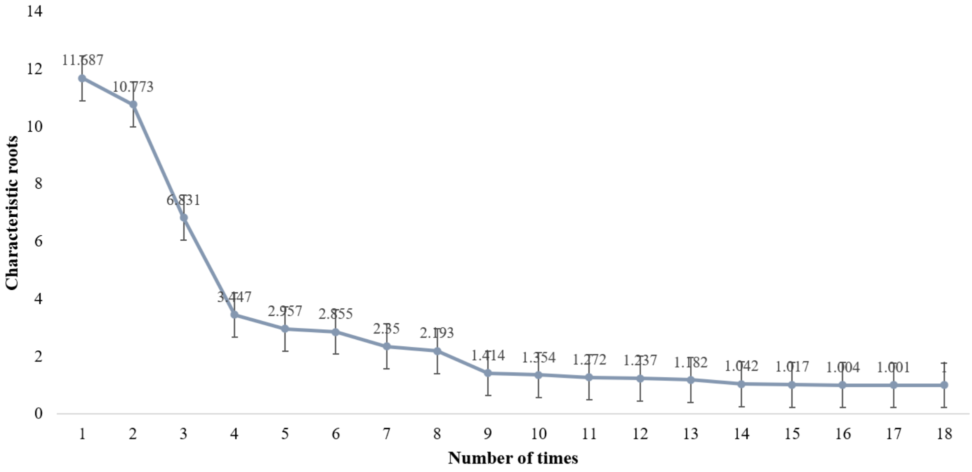


Figure 2 Gravel plot situation of 64 indicator eigenvalues

From Figure 2, it can be clearly found that the importance analysis of eigenvalues by principal component analysis method, the eigenvalues of the top eight indicators are obviously larger than the eigenvalues of other indicators, which are 11.687, 10.773, 6.831, 3.447, 2.957, 2.855, 2.350, and 2.193, and the same can be obtained according to the slope of the decline of eigenvalues that need to be The number of selected principal components is 8. Since the first four principal components have more significant eigenvalues compared to the others, the data indicators of bank inputs and outputs in this question are four principal components.

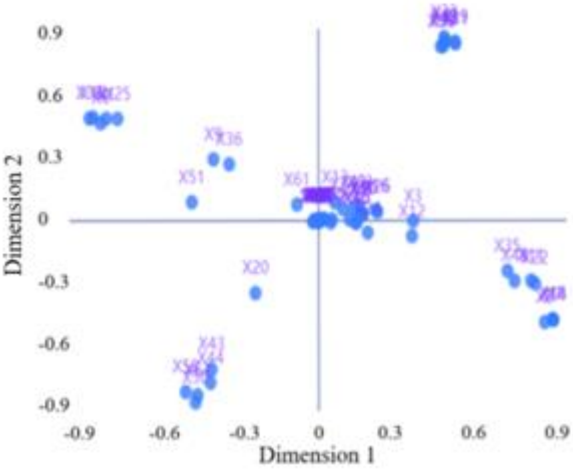


Figure 3 Two-dimensional distribution of factor load quadrants

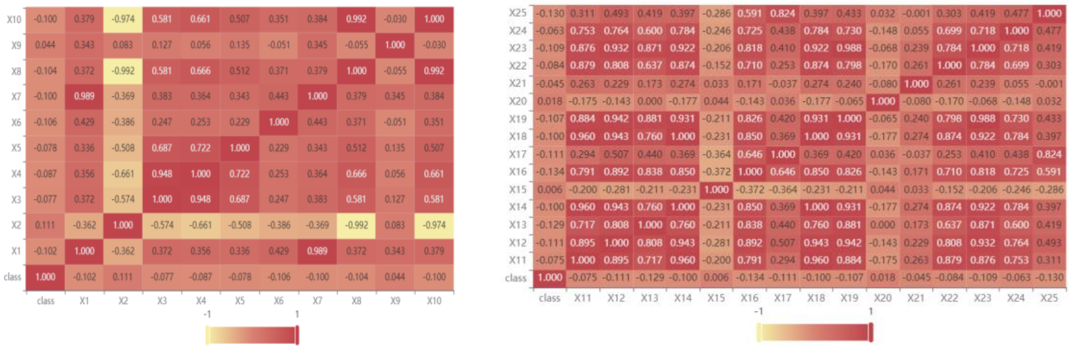
According to Figure 3, the spatial distribution of principal components is presented by means of quadrant plots. Since we extracted four principal components, we can directly present a two-dimensional loading principal component scatter plot, which results in the highest number of indicators in the first quadrant of the distribution, followed by the second and fourth quadrants, and the least in the third quadrant.

In order to extract indicators more precisely, we conducted correlation analysis based on principal component analysis to extract indicators again, with the following steps.

1) first test whether there is a statistically significant relationship between XY (p-value less than 0.05 or 0.01, strictly 0.01, not strictly 0.05).

2) Analyze the positive and negative direction of the correlation coefficient for and the degree of correlation.

3) To summarize the results of the analysis.



| | class | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| X10 | -0.100 | 0.379 | -0.974 | 0.581 | 0.661 | 0.507 | 0.351 | 0.384 | 0.992 | -0.030 | 1.000 |
| X9 | 0.044 | 0.343 | 0.083 | 0.127 | 0.056 | 0.135 | -0.051 | 0.345 | -0.055 | 1.000 | -0.030 |
| X8 | -0.104 | 0.372 | -0.992 | 0.581 | 0.666 | 0.512 | 0.371 | 0.379 | 1.000 | -0.055 | 0.992 |
| X7 | -0.100 | 0.989 | -0.369 | 0.383 | 0.364 | 0.343 | 0.443 | 1.000 | 0.379 | 0.345 | 0.384 |
| X6 | -0.106 | 0.429 | -0.386 | 0.247 | 0.253 | 0.229 | 1.000 | 0.443 | 0.371 | -0.051 | 0.351 |
| X5 | -0.078 | 0.336 | -0.508 | 0.687 | 0.722 | 1.000 | 0.229 | 0.343 | 0.512 | 0.135 | 0.507 |
| X4 | -0.087 | 0.356 | -0.661 | 0.948 | 1.000 | 0.722 | 0.253 | 0.364 | 0.666 | 0.056 | 0.661 |
| X3 | -0.077 | 0.372 | -0.574 | 1.000 | 0.948 | 0.687 | 0.247 | 0.383 | 0.581 | 0.127 | 0.581 |
| X2 | 0.111 | -0.362 | 1.000 | -0.574 | -0.661 | -0.508 | -0.386 | -0.369 | -0.992 | 0.083 | -0.974 |
| X1 | -0.102 | 1.000 | -0.362 | 0.372 | 0.356 | 0.336 | 0.429 | 0.989 | 0.372 | 0.343 | 0.379 |
| class | 1.000 | -0.102 | 0.111 | -0.077 | -0.087 | -0.078 | -0.106 | -0.100 | -0.104 | 0.044 | -0.100 |

| | class | X11 | X12 | X13 | X14 | X15 | X16 | X17 | X18 | X19 | X20 | X21 | X22 | X23 | X24 | X25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X25 | -0.130 | 0.311 | 0.493 | 0.419 | 0.397 | -0.286 | 0.591 | 0.824 | 0.397 | 0.433 | 0.032 | -0.001 | 0.303 | 0.419 | 0.477 | 1.000 |
| X24 | -0.063 | 0.753 | 0.764 | 0.600 | 0.784 | -0.246 | 0.725 | 0.438 | 0.784 | 0.730 | -0.148 | 0.055 | 0.699 | 0.718 | 1.000 | 0.477 |
| X23 | -0.109 | 0.876 | 0.932 | 0.871 | 0.922 | -0.206 | 0.818 | 0.410 | 0.922 | 0.988 | -0.068 | 0.239 | 0.784 | 1.000 | 0.718 | 0.419 |
| X22 | -0.084 | 0.879 | 0.808 | 0.637 | 0.874 | -0.152 | 0.710 | 0.253 | 0.874 | 0.798 | -0.170 | 0.261 | 1.000 | 0.784 | 0.699 | 0.303 |
| X21 | -0.045 | 0.263 | 0.229 | 0.173 | 0.274 | 0.033 | 0.171 | -0.037 | 0.274 | 0.240 | -0.080 | 1.000 | 0.261 | 0.239 | 0.055 | -0.001 |
| X20 | 0.018 | -0.175 | -0.143 | 0.000 | -0.177 | 0.044 | -0.143 | 0.036 | -0.177 | -0.065 | 1.000 | -0.080 | -0.170 | -0.068 | -0.148 | 0.032 |
| X19 | -0.107 | 0.884 | 0.942 | 0.881 | 0.931 | -0.211 | 0.826 | 0.420 | 0.931 | 1.000 | -0.065 | 0.240 | 0.798 | 0.988 | 0.730 | 0.433 |
| X18 | -0.100 | 0.960 | 0.943 | 0.760 | 1.000 | -0.231 | 0.850 | 0.369 | 1.000 | 0.931 | -0.177 | 0.274 | 0.874 | 0.922 | 0.784 | 0.397 |
| X17 | -0.111 | 0.294 | 0.507 | 0.440 | 0.369 | -0.364 | 0.646 | 1.000 | 0.369 | 0.420 | 0.036 | -0.037 | 0.253 | 0.410 | 0.438 | 0.824 |
| X16 | -0.134 | 0.791 | 0.892 | 0.838 | 0.850 | -0.372 | 1.000 | 0.646 | 0.850 | 0.826 | -0.143 | 0.171 | 0.710 | 0.818 | 0.725 | 0.591 |
| X15 | 0.006 | -0.200 | -0.281 | -0.211 | -0.231 | 1.000 | -0.372 | -0.364 | -0.231 | -0.211 | 0.044 | 0.033 | -0.152 | -0.206 | -0.246 | -0.286 |
| X14 | -0.100 | 0.960 | 0.943 | 0.760 | 1.000 | -0.231 | 0.850 | 0.369 | 1.000 | 0.931 | -0.177 | 0.274 | 0.874 | 0.922 | 0.784 | 0.397 |
| X13 | -0.129 | 0.717 | 0.808 | 1.000 | 0.760 | -0.211 | 0.838 | 0.440 | 0.760 | 0.881 | 0.000 | 0.173 | 0.637 | 0.871 | 0.600 | 0.419 |
| X12 | -0.111 | 0.895 | 1.000 | 0.808 | 0.943 | -0.281 | 0.892 | 0.507 | 0.943 | 0.942 | -0.143 | 0.229 | 0.808 | 0.932 | 0.764 | 0.493 |
| X11 | -0.075 | 1.000 | 0.895 | 0.717 | 0.960 | -0.200 | 0.791 | 0.294 | 0.960 | 0.884 | -0.175 | 0.263 | 0.879 | 0.876 | 0.753 | 0.311 |
| class | 1.000 | -0.075 | -0.111 | -0.129 | -0.100 | 0.006 | -0.134 | -0.111 | -0.100 | -0.107 | 0.018 | -0.045 | -0.084 | -0.109 | -0.063 | -0.130 |

Figure 4 Heat map of correlation between indicators and bank failures

According to the correlation of the heat map as Figure 4, we can find that X1, X2, X6, X8, X12, X13, X16, X17, X19, X23, X25, X39 and X46, a total of 13 items have a strong correlation with whether the bank fails. The analysis steps are.

1) Cluster category variability analysis according to the fields;

2) Analysis of the frequency of each clustering category according to the clustering summary;

3) According to the data set clustering annotation can be known to which category each sample data is classified;

4) The cluster center coordinates can be used to analyze the distance of each sample from the center point;

5) An overview of the analysis is performed.

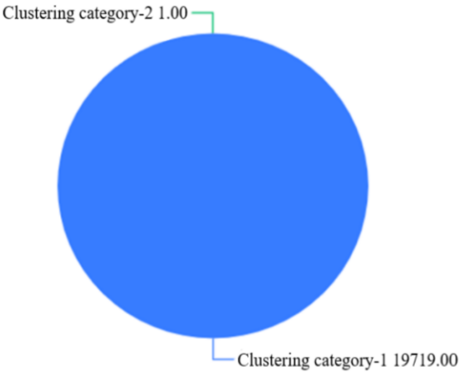


Figure 5 Clustering Summary Chart

From Figure 5, we can learn that the clustering results are divided into 2 categories, the frequency of clustering category 1 is 19719, accounting for 99.995%; the frequency of clustering category 2 is 1, accounting for 0.005%, and the coordinates of the cluster centroids obtained are shown in Table 1.

Table 1 Clustering centroid coordinates

| Category | X1 | X2 | X6 | X8 | X12 | X39 | X46 |
|---|---|---|---|---|---|---|---|
| 1 | 0.0552 | 0.5279 | 0.0654 | 1.4300 | 0.3158 | 0.0374 | 1.2469 |
| 2 | 0.0048 | 0.5681 | -0.0147 | 0.6303 | 0.0261 | 0.1437 | -0.0344 |
| Category | X13 | X17 | X25 | X23 | X19 | X16 | |
| 1 | 0.0616 | 2.5055 | 0.3423 | 0.0090 | 0.0159 | 0.3653 | |
| 2 | 2340.2000 | 0.0064 | 0.5105 | 0.3581 | 1.1664 | 0.0064 | |

## 4. Conclusion

In this paper, after analyzing the input-output data of banks, we use principal component analysis and correlation analysis to process the data of 64 indicators that affect bank failure, and we can find that X1, X2, X6, X8, X12, X13, X16, X17, X19, X23, X25, X39 and X46, a total of 13 items have a strong correlation with whether the bank fails or not, by the heat map. After that, we performed cluster analysis on the 13 indicators and they were divided into two categories. The cut-off line of bank failure efficiency is obtained based on the coordinates of the cluster centroids, which provides some reference value for the normal operation of domestic banks.

## References

*[1] Wen Ke. Dynamic parametric neural network for investment banking risk prediction model[J]. Science and Technology Bulletin, 2015,31(09):192-195.*

*[2] Wu Jianfei, Kang Yinhong, Song Xin, Liang Youpeng. Reference crop evapotranspiration prediction based on NARX*

*model[J]. Journal of drainage and irrigation machinery engineering, 2021, 39(05):533-540.*

*[3] Wu Haiyan, Xu Zhiliang. Study on the origin traceability of Guangdi Long based on principal component analysis and discriminant analysis[J]. Journal of Pharmaceutical Analysis, 2022, 42(03):387-393.*

*[4] Zheng Meiling, Liu Qianjin, Yang Jinchu, Li Ruili, Xu Kejing, Li Yaoguang, Du Jia, Zhang Junsong. Comprehensive evaluation of aroma quality of different sweet potato infusions based on principal component analysis and cluster analysis[J]. China Food Additives, 2022, 33(03):196-206.*

*[5] Gao Qianqian, Fan Hong. Research on systemic risk and investment strategy based on bank-asset bilateral network model[J]. China Management Science, 2021, 29(07):1-12.*