Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 5, No. 3, 2023

Research and Application of System-based Clustering and Principal Component Analysis Algorithms

Guowei Li, Shanwei Yang, Sai Li, Nan Wang, Juan Li*

School of information Engineering, Wuhan Business University, Wuhan 430056, China
* **Corresponding author**: looj@wbu.edu.cn

Abstract: The Silk Road was a channel of cultural exchange between China and the West in ancient times, in which glass was a valuable physical evidence of early trade, and the early glass in China was made by absorbing some foreign technology, which also led to a different chemical composition. Nowadays, most of the glass artifacts are roughly divided into lead-barium glass and high-potassium glass, and each of the different parts of the artifacts will be observed and sampled for analysis in the study of the artifacts, and the identification of their composition types has been hampered by natural weathering over thousands of years. Therefore, in view of such problems and the large number and complexity of chemical components, we propose to sub-classify the different types of glass artifacts through systematic clustering and principal component analysis algorithm model, and the basis of classification is the chemical composition, classification of the content of the different artifact sampling points, that is, the artifact number, and finally through sensitivity analysis to evaluate and test the classification results. The classification results can greatly reduce the workload of analyzing and identifying the types of artifacts, and provide a reference basis and methodological guidance for the problem of identifying and classifying artifacts.

Keywords: Glass artifacts, Systematic clustering; Principal component analysis; Sensitivity analysis.

1. Introduction

Principal component analysis is a technique used to explore the structure of high-dimensional data, while systematic cluster analysis is a technique to find the intrinsic structure between data, and both are mostly applied to data analysis type of topics, which also leads to many domestic and international scholars today to conduct research based on this model algorithm as well. Based on this, Qin Liyue et al. conducted an in-depth analysis and study on the comprehensive quality of roasted macadamia nuts kernels, which improves the theoretical basis and scientific basis for the processing and development of such foods. Chao Zhonghao et al. gave their model for the evaluation of volatile flavor substances of butter hot pot base, which enabled them to effectively evaluate the flavor of hot pot base. Therefore such as nowadays such models are widely used and researched, we intend to apply systematic clustering and principal component analysis algorithms to the problem of identification analysis of cultural relics, so that its work of cultural relic identification can be optimized to a certain extent, and refine various analysis algorithms and identification results testing so that it can effectively ensure the solution of such model algorithms for this problem.

2. Problem Description

The main raw material of glass is quartz sand, the main chemical composition of which is SiO2, and due to the high melting point of pure quartz sand, fluxes are added during refining in order to lower the melting temperature. The fluxes commonly used in ancient times are grass ash, natural alkali, saltpeter and lead ore, etc., and add limestone as a stabilizer, limestone calcination after conversion to CaO. added fluxes are different, its main chemical composition is also different. For example, lead-barium glass, which has a high content of PbO and BaO when lead ore is added as a flux in the firing process, is usually regarded as a glass variety invented by China itself, and the glass of Chu culture is dominated by lead-barium glass. In this context, we intend to analyze the relevant data of an existing batch of ancient glass artifacts, in which the chemical composition content varies among different artifacts and also the sampling points are different, and on this basis for each category choose the appropriate chemical composition to classify them into subcategories, give the specific classification method and analyze the rationality and sensitivity of the classification results.

3. Model Building and Solving

3.1. Problem 1 model building and solving

3.1.1. Establishment of the principal component clustering model [1-3]

(1) Principal component analysis is a dimensionality reduction algorithm that transforms multiple indicators into a few principal components, i.e., replacing old data with fewer new data, and this these principal components all satisfy the linear relationship of the initial variables and do not correlate with each other, but reflect the characteristic results described by the full set of variables to the greatest extent. This problem can be started with a principal component analysis, which is performed as follows:

Step1. First of all, it needs to be standardized.

Annex 2 of this question involves 69 heritage sampling points, that is, there are 69 sample points, which involves 14 indicators, so that the value of the i indicator of the j sample is k_{ij} , followed by the following standardization of each indicator is F_{ij} :

$$F_{ij} = \frac{k_{ij} - \widehat{F}_i}{h_i} \tag{1}$$

Where $\widehat{F}_i$ is the mean of the i indicator and h_i is the standard deviation of the i indicator. The purpose of standardization here is mainly to resolve errors and mistakes in the data results due to different magnitudes.

Step2. Calculate the correlation coefficient matrix of the sample matrix x.

By the following equation:

$$R=\frac{\sum_{k=1}^{n}\left(F_{ki}-\widehat{F}_i\right)\left(F_{kj}-\widehat{F}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(F_{ki}-\widehat{F}_i\right)^2\sum_{k=1}^{n}\left(F_{kj}-\widehat{F}_j\right)^2}} \tag{2}$$

The correlation coefficient matrix can be obtained from $R=(r_{il})_{14\times 69}$, where r_{il} is the correlation coefficient between the indicator of i and the indicator of l , and according to the correlation coefficient matrix, $r_{ii}=1$, $r_{ii}=r_{il}$, and where :

$$r_{il}=\frac{\sum_{k=1}^{69}F_{ik}F_{lk}}{67}(i,l=1,2,3\ldots 14) \tag{3}$$

Step3. Calculate the R eigenvalues and eigenvectors by MATLAB.

First calculate the eigenvalues of the correlation coefficient matrix R : $\lambda_1\geqslant\lambda_2\geqslant\lambda_3\geqslant\cdots\geqslant\lambda_{14}$,where the eigenvectors are $a_1,a_2,a_3,\ldots\ldots,a_{14}$,where $a_1=(t_{1i},t_{2i},t_{3i},\ldots,t_{14i})^T$.

Step4.Calculate the principal component contribution rate and cumulative contribution rate.

Contribution rate:

$$\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}(i=1,2,\cdots,p) \tag{4}$$

Cumulative contribution rate:

$$\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}(i=1,2,\cdots,p) \tag{5}$$

Step5.The principal component analysis was performed by matlab code, and then the first, second and third .... corresponding to the eigenvalues whose cumulative contribution of these two categories exceeded 85% were taken respectively The first $m\,(m\leqslant p)$ principal components, and then after determining the p principal components, the descriptive and statistical analyses of key variables and data were then performed.

2) After performing principal component analysis, the two classes are then subclassified by using the p principal component features as the basis for systematic clustering. Systematic clustering firstly divides the samples belonging to one class, and then always calculates the distance between subclasses, and soon then step by step in the conclusion, its will be merged into one big class.

3) Determine the value of K :elbow rule

The elbow rule[5] is to roughly estimate the optimal number of clusters by the graph. One of the locations where the effect of improvement by the degree of distortion of the graph has the greatest decrease in effect is the elbow, and the degree of distortion is generally used to determine the optimal value.

The systematic clustering can be started by dividing the new data obtained by principal component analysis for high potassium glass and lead-barium glass into K_1 and K_2 class clusters, respectively, where $K_1,K_2=\{1,2,3,\ldots,\beta\}$, where ($\beta<n\,n$ is the atomic level), and where β is the inflection point value where the slope in the elbow function shows a significant decrease.

3.1.2. Solution of the model

1) According to the above established model and steps, this question first through matlab respectively for these two types of principal component analysis, can get the two types of glass after the principal component analysis of key variables including eigenvalues and cumulative contribution rate, and then through excel to process the data, where the high potassium class of glass part of the principal component information is as follows:

Table 1. Information on some of the main results obtained from principal component analysis of high potassium glass

Eigenvector	a1	a2	a3	a4	a5
Eigenvalue	5.4002	2.4303	1.7154	1.6488	0.9485
Contribution Value	0.3857	0.1736	0.1225	0.1178	0.0678
Cumulative contribution value	0.3857	0.5593	0.6818	0.7996	0.8674

The cumulative contribution has reached about 87% at principal component 5, and the first five principal components were taken as the combined indicators after screening according to the requirements. It is also easy to see that principal component 1 and principal component 2 contribute the most to the cumulative contribution among them, i.e., it can be seen that silica and sodium oxide may be important influencing factors and indicators for the classification of high potassium glass subclasses.

From the information in Table 2 below, it can be seen that the cumulative contribution of principal component 8 has reached about 88%, i.e., the composite index corresponding to the seven principal components mentioned above is taken according to the requirements. The contribution values of the

first three components are higher and the difference is not very large, and it can be seen that silica, sodium oxide and potassium oxide may be more important indicators for the later subcategory of lead-barium glass classification.

Table 2. Information on some of the main results obtained from the principal component analysis of lead barium glass

Eigenvector	a1	a2	a3	a4	a5	a6	a7
Eigenvalue	3.5756	2.9487	1.6646	1.0847	0.9076	0.8406	0.7473
Contribution Value	0.2554	0.2106	0.1189	0.0775	0.0648	0.0600	0.0534
Cumulative contribution value	0.2554	0.4660	0.5849	0.6624	0.7272	0.7873	0.8407

Based on the above analysis, the final results of the combined index matrix of the screened lead-barium glass and high potassium glass were set to $f_1, f_2, \ldots, f_8$ and $F_1, F_2, .., F_5$ respectively.

(2) Then the combined index matrix data obtained by principal component analysis of the two were imported into SPSS for systematic aggregation, and a spectrum chart could be obtained separately, and the number of categories was determined after their division.

(3) The aggregation coefficients obtained from the above SPSS were processed in excel and the aggregation coefficients were plotted in descending order:

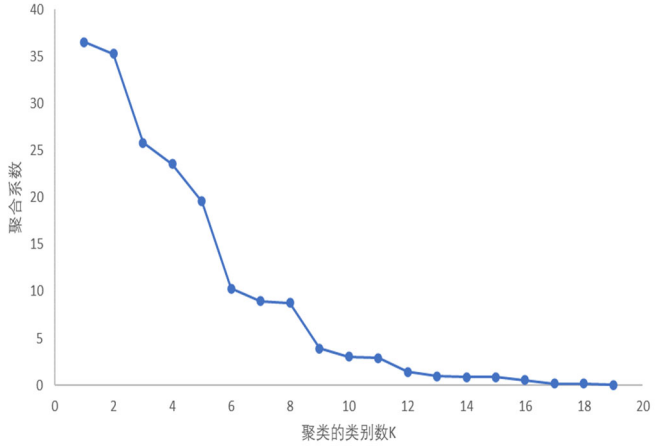

Figure 1. Folding line of polymerization coefficient of high potassium glass

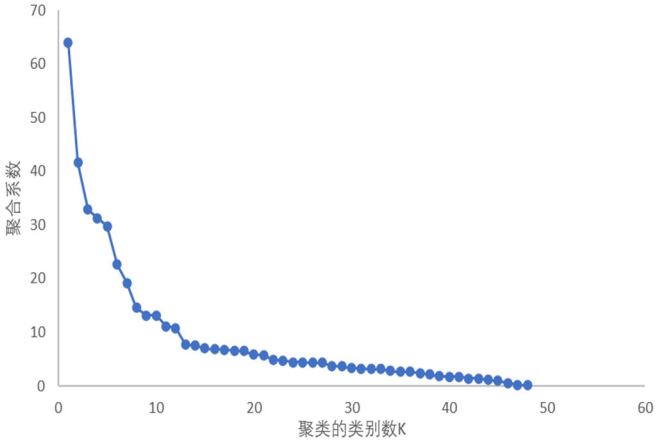

Figure 2. Folding line of polymerization coefficient of lead-barium glass

According to the aggregation coefficient line graph, when $K_1 =$ 6, the decreasing trend of the line tends to slow down, according to the elbow rule so the subcategory of high potassium glass can be divided into the category number K_1 .

According to the aggregation coefficient folding graph, it can be seen that when $K_2 =$ 7, the decreasing trend of the folding line tends to slow down, according to the elbow rule so the subcategory of lead barium glass can be classified as the category number K_2 .

Then the number of categories corresponding to the division of the two categories was divided on the spectrogram generated by SPSS, and the subcategories of the two were obtained by observing the divided spectrogram.

3.1.3. Sensitivity analysis of the model

From the analysis of the above data results, it can be concluded that at the confidence level of 95%, the component content of silica and potassium oxide has a significant influence on the presence or absence of weathering on the surface, so it is reasonable to judge that it may cause overlap in classification and unreasonable classification categories on the final classification results, so we can use reducing the component content of both of them to fit the number of categories by elbow rule and folding diagram of clustering coefficient again to see Whether the two have a greater

influence on the subcategory classification results of high potassium glass and lead barium glass, and if the difference in influence is large, we will classify them as a sensitivity factor set. We followed the above steps to plot the aggregation coefficient line graphs for both raw metadata without principal component analysis by excel and SPSS respectively for the results after reducing the content of treated components, and the results after removing potassium oxide are as follows:

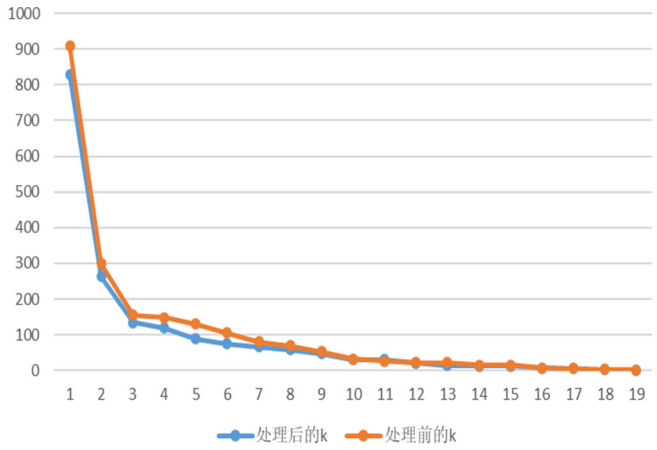

Figure 3. Comparison of polymerization coefficients of high potassium glass before and after removal

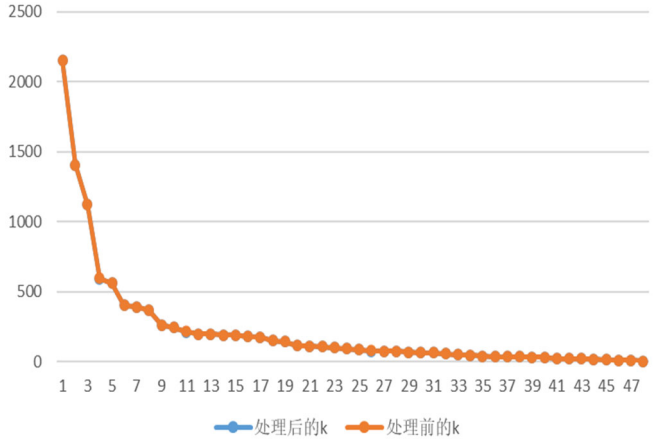

Figure 4. Comparison of polymerization coefficients of lead-barium glass before and after removal

It can be seen by Figures 4 and 5 that the polymerization coefficient fold lines for high potassium and lead-barium glasses before and after treatment almost exactly overlap, indicating that the removal of potassium oxide does not have a significant effect on the final subclassification of both of them, i.e. indicating that potassium oxide is not a sensitive set of factors for the samples.

And the results after removing silica are as follows:

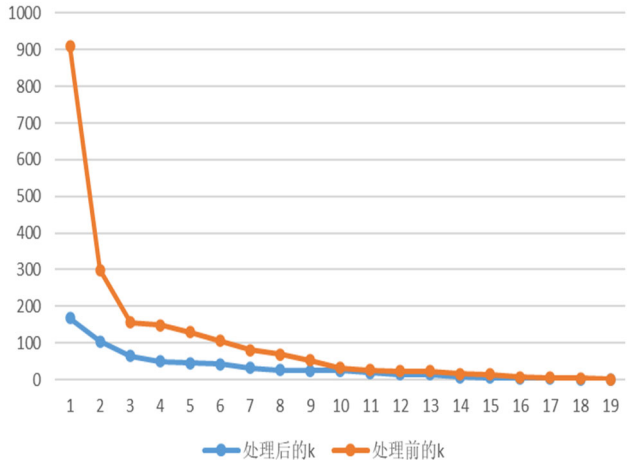

Figure 5. Comparison of polymerization coefficients of high potassium glass before and after removal

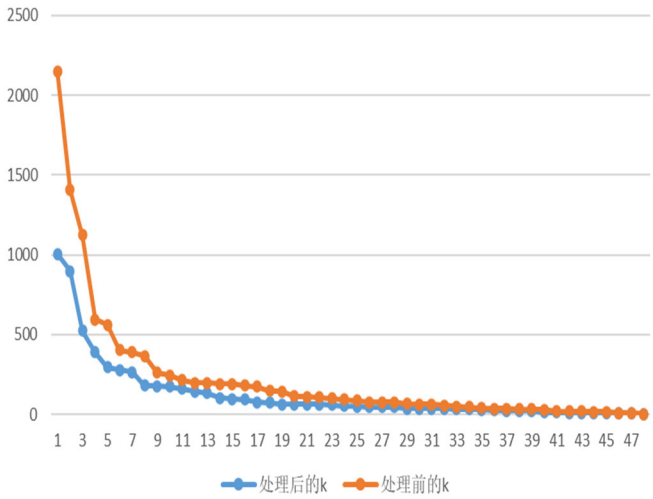

Figure 6. Comparison of polymerization coefficients of lead-barium glass before and after removal

Figure 6 shows that the removal of silica has a significant effect on the classification of the subclasses of high potassium glass, with a large difference in the degree of trend; Figure 7 shows that although the folds do not overlap or are close before and after the treatment, the degree of trend of the two folds is similar, and then the elbow rule shows that the classification results of the two are not significantly different. Therefore, it can be seen that the reduction data do not have a large impact on the final classification results of lead-barium glass, while for high potassium glass, the composition content of silica can be classified as a sensitive set of factors for its results.

4. Conclusion

This paper focuses on the application of cluster analysis and principal component analysis algorithm models. In the study we positioned the experimental object to the identification of the category of glass artifacts, and in the identification work we mainly through the above model algorithm for its different artifacts sub-classification, the glass artifacts into six categories, lead barium glass into seven categories, and the category content is based on the analysis of chemical composition of the artifacts sampling points, and then through sensitivity analysis of its classification results for a certain evaluation and test, the results obtained the classification results are less affected by some chemical composition. The results obtained the classification results of the method is less affected by some of the chemical composition, the type of cultural relics for the identification of the problem brings a valuable reference to improve the accuracy and efficiency of identification.

Acknowledgment

The authors gratefully acknowledge the financial support from Innovation and Entrepreneurship Training Program of Wuhan Business University (202211654165), Teaching reform Research Project of Wuhan Business University (2022Y009).

References

[1] GUO M, Li D, Li Y, et al. Comprehensive evaluation of Jihua peanut varieties with high oleic acid based on principal component and cluster analysis[J]. CHINESE JOURNAL OF OIL CROP SCIENCES, 2022, 44(6): 1210.

[2] LEI Yue, GONG Yanlong, DENG Ruyue, et al. Comprehensive Evaluation of Quality Characteristics of Parboiled Rice Based on Principal Component Analysis and Cluster Analysis[J]. Science and Technology of Food Industry, 2021, 42(7): 258-267.

[3] XU Qingyu, YU Jing, ZHU Dawei, ZHENG Xiaolong, MENG Lingqi, ZHU Zhiwei, SHAO Yafang. Nutritional Quality Evaluation of Different Rice Varieties Based on Principal Component Analysis and Cluster Analysis[J]. China Rice, 2022, 28(6): 1-8.