A study on the discrimination and identification of ancient glassware types based on cluster analysis

Liyang Zheng [1], Shuangyang Tao [2], Hua Zhang [1], Siyang Liu [1], Xuan Dong [1], Jingyuan Li [1, *]

[1] School of Environment Science and Engineering, Nanjing University of Information Science & Technology, Nanjing, China, 210044

[2] History Culture and Tourism College, Sichuan Normal University, Sichuan, China, 610066

* Corresponding author: jingyuanli@nuist.edu.cn

Abstract. The classification and identification of ancient glass artifacts is complicated due to the exchange of internal elements, which can be influenced by a multiplicity of different factors, such as weathering, erosion, and so on. To classify and identify the types of glass artifacts, this paper uses cluster analysis (K-Means) to study the main chemical composition of high potassium glass and lead-barium glass. It is found that the chemical composition of different glass (potassium oxide, barium oxide, lead oxide) and subclasses types are significantly different, and the content of lead oxide in lead-barium glass also has a large difference. Using the clustering centers obtained by cluster analysis (K-Means), the eight unknown types of glass were first determined visually and briefly based on the different proportions of potassium oxide, barium oxide, and lead oxide. As for accurate classification, the Euclidean distances between the unknown components and the sample centroids were calculated separately, and the shortest Euclidean distances were identified by comparison, and then categorized for identification.

Keywords: Cluster Analysis, Ancient Glass Artifacts, Euclidean Distance, Classification and Identification.

1. Introduction

Ancient glass is highly susceptible to weathering by the burial environment. During the weathering process, the internal elements are exchanged with the environmental elements in large quantities, resulting in changes in their composition ratios and thus affecting the correct determination of their categories [1-2]. The analysis and identification of ancient glass artifacts have attracted extensive attention from many researchers [3]. To classify the types of glass artifacts, this paper attempts to find the classification rules of two glass types by counting the different chemical compositions of the two types of glass, giving the division method of subclass division [4-5], and judging the rationality of the division method. In order to identify the types of glass artifacts, the content of chemical substances in unknown glass artifacts is an important criterion for classification, and the unknown glass is classified by the significant differences in the chemical composition of different types of glass. Then a sensitivity analysis is performed by changing a parameter of the type of glass [6].

2. Glass classification model building and solving

In this study, the data of the chemical composition of ancient glass artifacts first has been preprocessed to study the classification law of two types of glass. Secondly, the K-means clustering model is used to classify the data, and then perform the clustering summary to find out the significant chemical composition that can classify the glass [7]. Finally, judge the reasonableness of the k value taken by calculating the profile value. The flow of the K-means clustering model is shown in Figure 1.

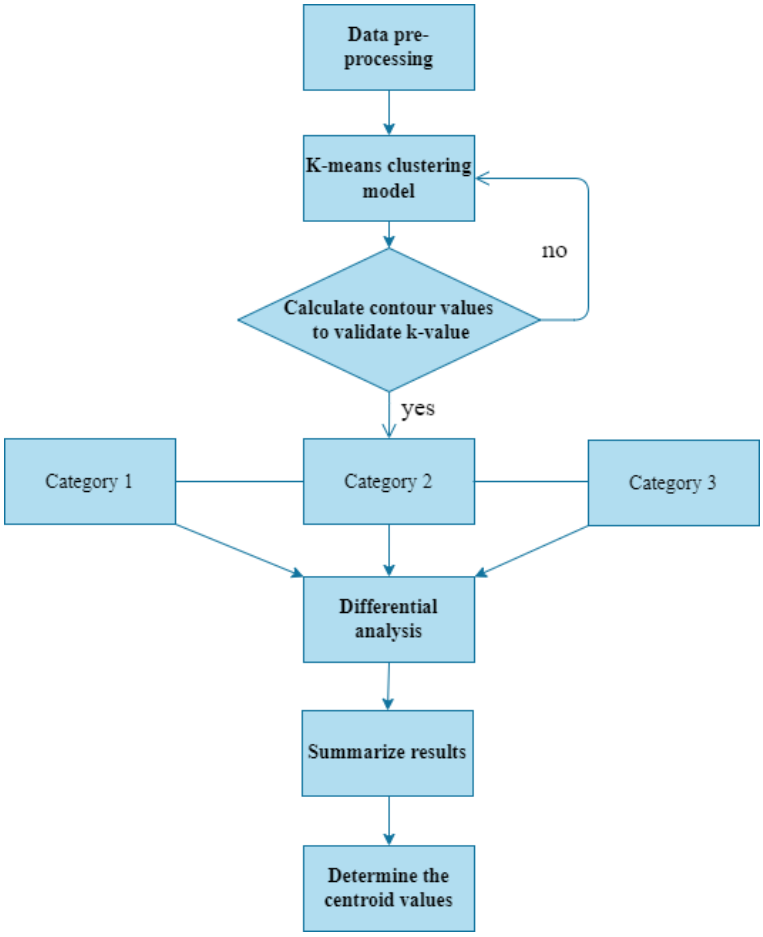


Figure 1. Flow chart of K-means clustering model

2.1. Data source

The dataset used in this study was obtained from the 2022 Chinese National College Student Mathematical Modeling Competition, specifically from Problem C. The data concerns ancient Chinese glassware, which has been classified by archaeologists into two categories based on their chemical composition and other testing methods: high-potassium glass and lead-barium glass.

The dataset includes information on the proportions of various chemical components of the glassware, along with their respective classifications. The data is characterized by its compositional nature, whereby the sum of the proportions of each component should be 100%. However, due to limitations of the testing methods and other factors, some of the data may be less than 100%. To account for this, we considered data with the sum of component proportions between 85% and 105% as valid data.

The data has been preprocessed, and the resulting dataset is available on GitHub. This dataset can be accessed using the following link: [https://github.com/magemaa/2022-C].

2.2. Data pre-processing

Use the Vlookup function in Excel for merging to facilitate statistics. Screen the chemical elements that have a strong influence on the type of glass Based on the title Lead Barium Glass with a high content of lead oxide (PbO) and barium oxide (BaO) Potassium glass is made by firing substances with high potassium content can be known as potassium oxide, lead oxide, and barium oxide. Make a table of these three elements [8].

2.3. Model Building

To explore the distribution pattern of high potassium glass and lead-barium glass and analyze the rationality and sensitivity of the classification model. The clustering category variability analysis is

performed based on the fields and then clustering summary analysis is performed to analyze the frequency of each clustering category.

According to the data set clustering annotation can know to which category each sample data is classified, the cluster center coordinates can be used to analyze the distance between each sample and the center point, which reflects the correlation degree between a point and the center point, the correlation is strong then classified into one category, if the correlation is weak then continue to iterate the analysis, and finally, the analysis results are reviewed to observe the significant differences of chemical substances of various glasses after classification [9-10].

2.4. Systematic clustering model for glass species (k-means)

(1) The data are clustered with k-means, and then the cluster regions are plotted, as shown in Figure 2:

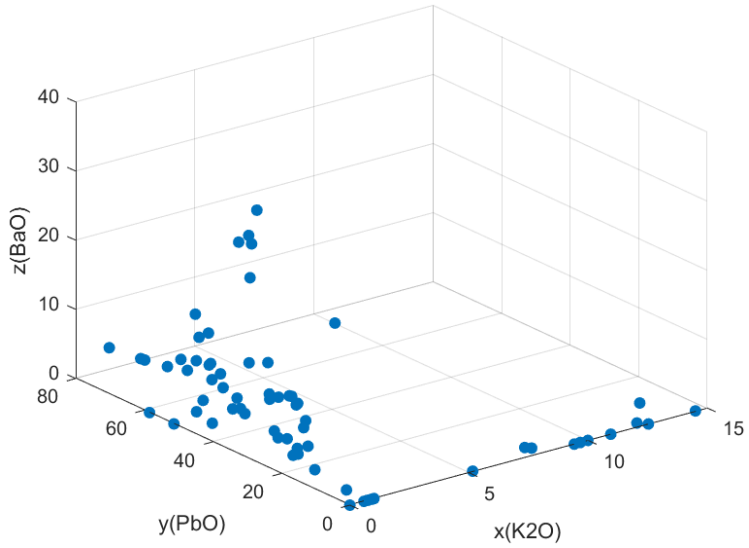


Figure 2. K-means clustering

The larger clusters seem to be divided into three regions to cluster the data. Specify k=3 clusters.

(2) Verify the reasonableness of the value of K taken and calculate the contour value. The contour values are shown in Figure 3.

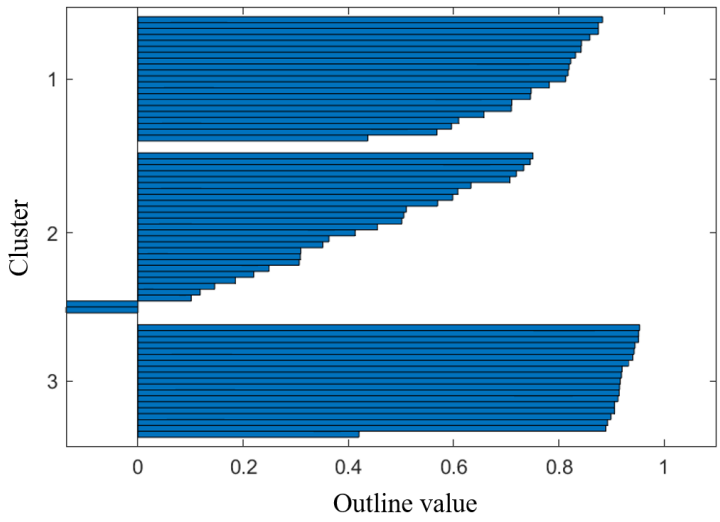


Figure 3. Contour map

The profile value is close to 1, so it is more reasonable to make k=3.

(3) Cluster analysis results

The results of the field variability analysis are shown in Table 1.

Table 1. Field Variance Analysis Table

	Clustering categories (mean ± standard deviation)			F	P
	Category 2(n=27)	Category 1(n=21)	Category 3(n=19)		
Type	2.0±0.0	2.0±0.0	1.053±0.229	412.657	0.000***
Potassium oxide	0.157±0.173	0.188±0.377	6.081±5.345	29.379	0.000***
Barium oxide	12.664±9.955	8.098±4.669	0.485±0.899	17.522	0.000***
Lead oxide	22.761±7.081	47.965±8.558	0.908±2.806	244.309	0.000***

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

The results of the ANOVA showed that: for the variable potassium oxide (K2O), the significance P-value was 0.000*** (1%), showing significance at the level and rejecting the original hypothesis, indicating that the variable potassium oxide (K2O) was significantly different between the categories classified by the cluster analysis.

For the variable barium oxide (BaO), the significant p-value of 0.000*** (1%), shows significance at the level, rejecting the original hypothesis, indicating that the variable barium oxide (BaO) is significantly different between the categories divided by cluster analysis.

For the variable lead oxide (PbO), the significance p-value is 0.000*** (1%), which presents significance at the level and rejects the original hypothesis, indicating that the variable lead oxide (PbO) is significantly different between the categories delineated by the cluster analysis.

For the analyzed objects, the clustering summary was performed, and the summary graph was drawn, as shown in Figure 4. it can be seen to be more evenly divided into three categories. The rationality of clustering was further verified.

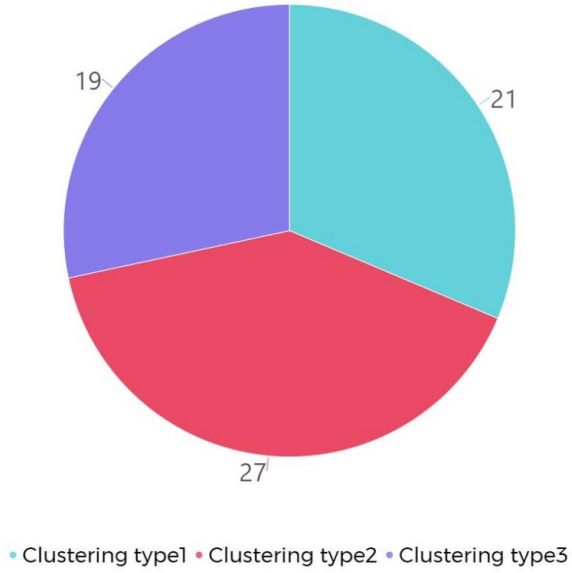


Figure 4. Clustering Summary

The dataset clusters are labeled as shown in Table 2:

Table 2. Data clustering table

Clustering type	Type	Potassium oxide	Barium oxide	Lead oxide
1	2	1.05	0	47.43
3	1	5.19	0	0.25
3	1	12.37	2.86	1.41
3	1	9.67	0	0
3	1	10.95	0	0
3	1	7.37	1.38	0.2
3	1	7.68	0.97	0.35
3	1	0	0	0
2	2	0	31.23	28.68

The above table shows the partial data clustering annotation of the model clustering results, it is a preview result, and only the first 9 entries of the integrated ranking are shown.

The coordinates of clustering centroids are shown in Table 3:

Table 3. Table of clustering centers

Clustering type	Center Value_Type	Central value_potassium oxide	Central value_Barium oxide	Central Value_Lead Oxide
1	2	0.1880	8.0980	47.9647
2	2	0.1574	12.6644	22.7607
3	1.0526	6.0805	0.4847	0.9078

As can be seen from the table, cluster (1) and cluster (2) contain lower levels of potassium oxide and higher levels of lead oxide and barium oxide, while cluster (3) is the opposite. Cluster (1) clearly contains more lead oxide and less barium oxide compared to cluster (2).

2.5. Sensitivity analysis of clustering analysis results

The chemical compositions of various glasses were added 2%, 5%, and 10% using MATLAB, brought into the K-means clustering model to re-form the clustering centers, and the differences were observed by comparing these clustering centers with the initial clustering centers. The differences in the center values of potassium oxide, lead oxide, and barium oxide were obtained for the 3 categories at each degree of disturbance.

The absolute value of the difference between each center value and the original clustering center after comparing the different levels of disturbance and performing cluster analysis was less than 1% (the unit of the component on the way is the percentage of the total substance), which proves that the established K-means clustering model has good robustness and will not be distorted by small fluctuations of the data.

3. Establishment and solution of the glass identification model

In order to see the main chemical composition of different artifacts, a line graph was drawn based on the main components of the eight unknown glasses, and then the Euclidean distance method was used to discern the type of artifacts belonging to the artifacts and to determine the sensitivity of the results, a small perturbation was made to a parameter of the glass, and the results before and after the perturbation were compared. The glass identification process is shown in Figure 5.

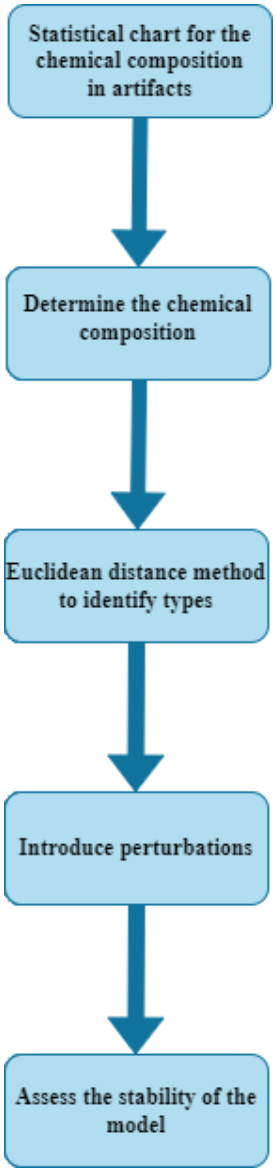


Figure 5. Glass identification flow chart

3.1. Analysis of the chemical composition of unknown glass artifacts

Using the judging criteria above, i.e., the content of potassium oxide, barium oxide, and lead oxide to determine the classification of unknown glass artifacts, we made bar charts of the major chemical content of the eight unknown artifacts in order to visualize the data and present more clearly the differences in chemical content among the different artifacts. As shown in Figure 6:

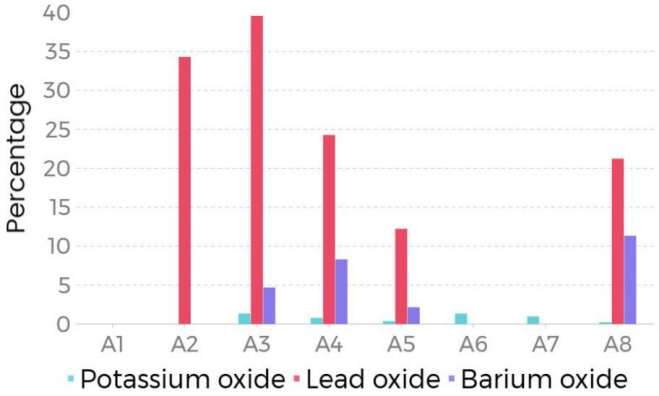


Figure 6. Statistical chart of chemical composition content

It is observed that the lead oxide content is higher in glass artifacts A2, A3, A4, A5, and A8, and the potassium oxide content is higher in A3, A4, A6, and A7. It can be found that the lead oxide and potassium oxide content is higher in A3, and A4, and it is impossible to judge their types directly.

3.2. Identification of unknown glass artifacts

We used to calculate the Euclidean distance between the vectors and the cluster centroids and selected the minimum value by using the Euclidean distance formula.

$$d(i,j)=\sqrt{\left[\left(x_{i1}-x_{j1}\right)^2+\left(x_{i2}-x_{j2}\right)^2+\left(x_{in}-x_{jn}\right)^2\right]} \tag{1}$$

Let the chemical composition of the A_1 glass artifact be $X_{1\,j}$, such as $\{X_1,\ X_2, X_3\}$ as vector A1, calculate the Euclidean distance between vector A1 and each clustering center, take the minimum value of the three distances, and group them into one category.

And so on, the minimum Euclidean distance of the vector is obtained, and all the unknown glass artifacts in Table 3 are classified. The classification results are shown in Table 4.

Table 4. Classification results

Cultural relics Number	Potassium oxide	Lead oxide	Barium oxide	Identify Type
A1	0	0	0	High potassium
A2	0	34.3	0	lead barium
A3	1.36	39.58	4.69	lead barium
A4	0.79	24.28	8.31	lead barium
A5	0.37	12.23	2.16	High potassium
A6	1.35	0	0	high potassium
A7	0.98	0	0	high potassium
A8	0.23	21.24	11.34	lead barium

3.3. Sensitivity analysis of classification results

To determine the effect of each component on the system as a whole to evaluate the stability of the established clustering model, we have to perform a sensitivity analysis. The data of the chemical element content of a certain glass is gradually increased proportionally to perturb the system as a whole (e.g., potassium oxide content data are each increased by -5%, -10%, 5%, 10%, etc.) brought into the k-means model in Problem 2 to observe whether the classification criteria change.

Using MATLAB, we increased potassium oxide and barium oxide by 5% to 10% respectively, in the type of glass after cluster analysis. It was found that by increasing potassium oxide and barium oxide by 5% to 10%, respectively, the obtained artifact categories did not change, and the clustering model was judged to have good stability.

A summary of this research problem: By solving the Euclidean distance between the unknown artifacts and the clustering center, we obtained A1, A5, A6, and A7 as high potassium glass and A2, A3, A4, and A8 as lead-barium glass, and found that the results obtained before and after the perturbation were consistent with those before the perturbation, which proved its stability.

4. Conclusion

In summary, this study demonstrates the effectiveness that a relatively simple K-Means algorithm is used to establish an ancient glass classification model. The main chemical composition of high potassium glass and lead-barium glass was analyzed by cluster analysis (K-Means), and the content of lead oxide in lead-barium glass was divided into three categories. By analyzing the main chemical composition, this model can accurately categorize different types of ancient glass. As to identify the

types of unknown glasses, the analysis was performed to determine the categories based on the clustering centers obtained from the cluster analysis (K-Means). To obtain accurate identification, the Euclidean distance between the unknown components and the center point of the sample is calculated, respectively, and then the shortest Euclidean distance was selected to categorize for identification. The unknown artifacts A1, A5, A6, and A7 were obtained as high potassium glass, and A2, A3, A4, and A8 were lead-barium glass. The successful identification of eight unknown glasses in this study indicates the practicality of this classification model. Additionally, the sensitivity analysis conducted in this study proves the stability of the cluster analysis and provides a method for further improving the accuracy of the classification model. This study provides a valuable contribution to the field of ancient glass analysis and offers a promising approach for future research in this area.

References

[1] Wang Zhihao, Zhao Xiangdri, Li Zhiqun, et al. A machine learning-based method for prediction and subclassification of the original composition of weathered silicate glass [J]. Journal of Silicates, 2023, 51 (02): 416 - 426.

[2] Yin Y. L. Composition analysis of ancient glass products by correlation prediction [J]. Contemporary Chemical Research, 2023, (01): 122 - 126.

[3] He Yilin, Tao Xiao. Machine vision-based defect detection and sorting system for cell phone glass cover [J]. Industrial Control Computer, 2022, 35 (12): 74 - 75.

[4] Xu Lingling, Chi Dongxiang, Huang Yanqian, et al. An improved k-GSMOTE method for classification of imbalanced data [J]. Journal of Shanghai Institute of Electrical Engineering, 2022, 25 (02): 88 - 94.

[5] Sun Yi, Zhao Ying Hun, Wang Tian Qi, et al. Research and improvement of a K-mean optimization algorithm [J]. Automation Technology and Applications, 2021, 40 (09): 1 - 5.

[6] Zhang Dongmei, Li Min, Xu Dachuan, et al. A review of theory and algorithms for k-means problems [J]. China Science: Mathematics, 2020, 50 (09): 1387 - 1404.

[7] Xiong Honglin, Fan Chongjun, Zhao Shan, et al. A multi-scale convolutional neural network-based method for glass surface defect detection [J]. Computer Integrated Manufacturing Systems, 2020, 26 (04): 900 - 909.

[8] Liu L, Hu F, Tian Y. P. Application of K-means clustering analysis in PM (2 .5) source analysis technology [C]//Professional Committee of Aerosol of Chinese Particle Society. Proceedings of the 8th National Workshop on Atmospheric Fine and Ultrafine Particle Technology and PM (2. 5) Source Spectrum Exchange, 2015: 16.

[9] Hou Zili. Qinhuangdao glass museum Tang and Song glass artifacts specific dating testimony - talk about the evolution and development of Sui, Tang, and Song glass artifacts [J]. Popular Literature and Arts, 2015 (13): 47.

[10] Cao Caixia, Guo Hong. Classification of disease types of stone, ceramic and glass relics in collections using fuzzy mathematical knowledge [J]. Journal of Beijing Union University (Natural Science Edition), 2009, 23 (04): 58 - 60.