Research on glass type classification based on Logit model and K-Means clustering algorithm

Jiayao Li [#], Cheng Lin [#], Liwei Zhuang [#], Zhilong Xu [#], Zhenlin Liang [#], Yuning Gao [*]

College of Computer Engineering, Guangzhou Huali College, guangzhou, guangdong, 511325

* Corresponding author: 2016120038@jou.edu.cn

[#]These authors contributed equally.

Abstract. Glass artifacts have been important in human history and are often studied by archaeologists and art historians to understand the development of society, technological progress, and cultural exchange. However, the systematic classification of glass artifacts is a major challenge to their study, because glass artifacts excavated from archaeological excavations are often highly weathered, making it difficult to classify them, so a scientific and reliable method to analyze and systematically classify glass artifacts based on their detected chemical composition is of great importance to the study of human history and culture. In this paper, two methods, logit model classification and cluster analysis, were used to determine the key to distinguishing high-potassium glass from lead-barium glass by the content of PbO, SrO, SnO2, and CaO components. Next, a comparison of the three cluster analyses was used to determine the use of the k-means algorithm to further subdivide the high potassium class and the lead-barium glass artifacts into two subclasses each: high Al-Fe and high Cu-Zn; and high PbP and high Na-Zn. Finally, the sensitivity analysis of the model and the robustness of the model and the reasonableness of the results were analyzed using Pearson correlation coefficients.

Keywords: Logit model, K-Means++ clustering algorithm, Pearson correlation coefficient, glass products.

1. Introduction

The aim of this study is to classify different types of glass artifacts through elemental content analysis. Glass artifacts are one of the precious cultural heritage left by ancient civilizations [1], with exquisite production techniques, diverse styles and rich culture [2]. The study of the classification of glass artifacts has been an important issue in the fields of archaeology, conservation and material science. For the study of glass artifacts, some researchers have used, for example, X-ray fluorescence analysis [3] with Raman spectroscopy studies [4] and other laser techniques for chemical composition analysis [5], and some have used statistical analysis studies such as fuzzy mathematics [6]. In recent years, machine learning methods have also been widely used in the classification and analysis of glass artifacts. In this paper, we choose two methods, Logit model and K-means++ algorithm, to use common elements as the basis of classification, and learn the dataset of glass artifacts of known types by Logit model method to predict the types of new samples and explore the classification laws; by cluster analysis method, high potassium glass and lead-barium glass are each divided into two clusters and sub-classified into different types of glass artifacts. This study helps to distinguish more accurately as well as refine the classification of high potassium glass and lead-barium glass artifacts, which can provide some reference for cultural relic identification and cultural research.

2. Study Design

Classification was explored for high potassium glass and lead-barium glass artefacts.This paper uses both Logit model classification and cluster analysis, as shown in Figure 1. The glass artefacts were analysed by the Logit model algorithm to derive the significant chemical composition, and then the clustering algorithm was used to classify the data into different clusters based on similarity, and

multiple clustering algorithms were combined to compare the results in order to select the optimal result and improve the classification stability and accuracy. Finally, the clustering results were presented by PCA algorithm to reduce the dimensionality, and the final subclasses were classified by combining the clustering centres and chemical composition differences. And sensitivity analysis and Pearson correlation analysis were used to assess the robustness of the model and the reasonableness of the results.

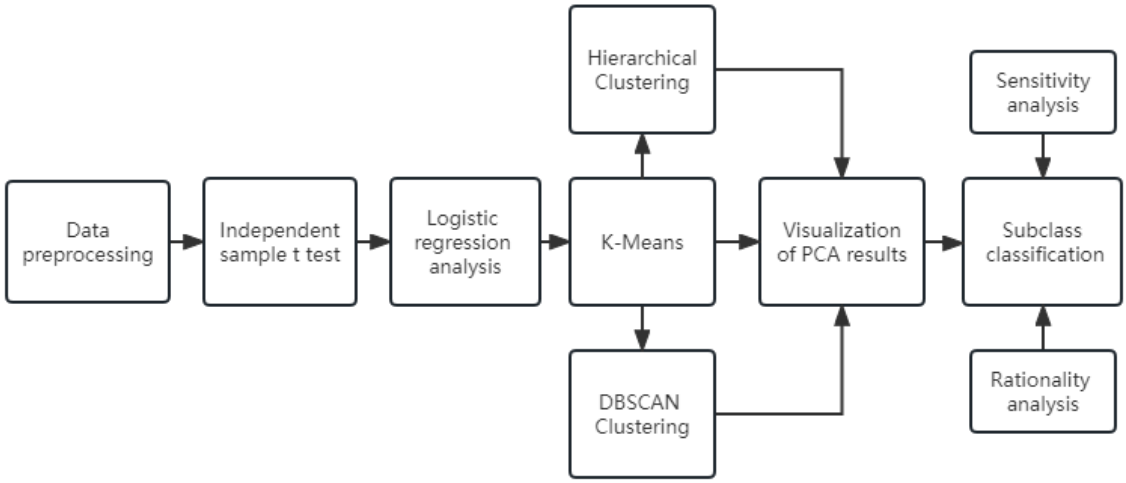

Figure 1. Research design drawings

2.1. Descriptive statistics and independent samples t-test

Table 1. Descriptive statistics and independent samples t-test results

Variable Name	Variable Value	Sample size	Average value	Standard deviation	P	Mean Difference
(SiO_2)	High Potassium	18	76.644	14.467	0.000(***)	37.768
	Lead Barium	49	38.876	18.646		
(Na_2O)	High Potassium	18	0.463	1.089	0.337	0.441
	Lead Barium	49	0.904	1.813		
(K_2O)	High Potassium	18	6.402	5.308	0.000(***)	6.229
	Lead Barium	49	0.173	0.276		
(CaO)	High Potassium	18	3.845	3.308	0.004(***)	1.795
	Lead Barium	49	2.05	1.635		
(MgO)	High Potassium	18	0.785	0.712	0.441	0.139
	Lead Barium	49	0.646	0.63		
(Al_2O_3)	High Potassium	18	5.057	3.077	0.101	1.389
	Lead Barium	49	3.668	3.009		
(Fe_2O_3)	High Potassium	18	1.376	1.566	0.025 (**)	0.72
	Lead Barium	49	0.656	0.948		
(CuO)	High Potassium	18	2.156	1.492	0.659	0.276
	Lead Barium	49	1.88	2.47		
(PbO)	High Potassium	18	0.274	0.514	0.000 (***)	33.075
	Lead Barium	49	33.349	14.947		
(BaO)	High Potassium	18	0.399	0.842	0.000 (***)	10.091
	Lead Barium	49	10.49	8.331		
(P_2O_5)	High Potassium	18	1.028	1.281	0.019 (**)	2.265
	Lead Barium	49	3.293	3.909		
(SrO)	High Potassium	18	0.028	0.044	0.000 (***)	0.32
	Lead Barium	49	0.348	0.264		
(SnO2)	High Potassium	18	0.131	0.556	0.437	0.073
	Lead Barium	49	0.058	0.213		
(SO2)	High Potassium	18	0.068	0.157	0.329	0.732
	Lead Barium	49	0.8	3.139		

We first cleaned the data by removing invalid data where the sum of components was not between 85% and 105%, and assigned a value of 0 to missing values considered as undetected components.To initially analyze the significant differences between the sample and the two types of artifacts, descriptive statistics with independent samples t-test were used for the data in this paper [7].

From Table 1, it is clear that the high potassium and lead-barium samples differed significantly in a variety of components that may be key factors in distinguishing the two types of samples.

First, the mean values of SiO2 composition and K2O composition were significantly higher in the high-potassium samples than in the lead-barium samples, 76.644 and 38.876, and 6.402 and 0.173, respectively; therefore, SiO2 and K2O composition may be the key factors to distinguish the two types of samples. In addition, the mean values of CaO and K2O compositions in the high-potassium samples were also higher than those in the lead-barium samples, while the mean values of BaO, P2O5 and PbO compositions in the lead-barium samples were higher than those in the high-potassium samples.

Therefore, based on these data, it can be tentatively inferred that the two types of samples, high K and PbB, can be effectively classified by their chemical composition, where SiO2, K2O, CaO and PbO are probably the most important factors in distinguishing the two types of samples.

2.2. Logit model modeling

The Logit model [8] is a binary machine learning algorithm that predicts the likelihood of a sample belonging to a category by mapping the results of linear regression to a logistic function that yields a probability value between and 1. In the Logit model, we define the following logistic functions (also known as sigmoid functions):

$$g(z) = \frac{1}{1+e^{-z}} \tag{1}$$

Where $z = \boldsymbol{\omega}^T \mathbf{x}$ denotes the linear combination of the input feature value $\mathbf{x}$ with the corresponding weight parameter $\boldsymbol{\omega}$. g(z) takes values in the range of 0 to 1, we usually label the sample as positive class when $g(z) \geq 0.5$. When $g(z) < 0.5$, we then mark the sample as a negative class. Logit models usually use maximum likelihood estimation to solve the model parameters, specifically, to find the optimal weight parameter $\boldsymbol{\omega}$ by maximizing the similarity between the actual and predicted markers of the sample. The specific solution process can be expressed by the following equation:

$$L(\boldsymbol{\omega}) = \prod_{i=1}^{m} \left(g(\boldsymbol{\omega}^T \boldsymbol{x}_i)\right)^{y_i} \left(1 - g(\boldsymbol{\omega}^T \boldsymbol{x}_i)\right)^{1-y_i} \tag{2}$$

Where $y_i \in 0,1$ denotes the actual marker of the i-th sample(1 indicates lead-barium class, 0 indicates high potassium class), m denotes the total number of samples. We can take $L(\boldsymbol{\omega})$ to its logarithmic form:

$$l(\boldsymbol{\omega}) = \sum_{i=1}^{m} \left[y_i log\left(g(\boldsymbol{\omega}^T \boldsymbol{x}_i)\right) + (1 - y_i) log\left(1 - g(\boldsymbol{\omega}^T \boldsymbol{x}_i)\right) \right] \tag{3}$$

The final optimization objective is to maximize $l(\boldsymbol{\omega})$:

$$\omega^* = argmax_{\omega} l(\boldsymbol{\omega}) \tag{4}$$

To avoid model overfitting, a regularization term is usually introduced into the objective function:

$$\boldsymbol{\omega}^* = argmax_{\omega} \left[l(\boldsymbol{\omega}) - \frac{\lambda}{2} \| \boldsymbol{\omega} \|_2^2 \right] \tag{5}$$

Where λ is the regularization factor and $\| \boldsymbol{\omega} \|_2^2$ denotes the L2 parametric square of the weight parameter.

2.3. Logit model model results analysis

Table 2. Regression coefficient table

Feature	Coefs	p-values
(PbO)	1.142	0.000(***)
(BaO)	0.808	0.337
(SrO)	0.510	0.000(***)
(Na_2O)	0.254	0.004(**)
($Al2O_3$)	0.231	0.441
(SO_2)	0.149	0.101
(P_2O_5)	0.059	0.025(*)
(MgO)	-0.078	0.659
(SnO_2)	-0.082	0.000(***)
(CaO)	-0.366	0.000(***)
(Fe_2O_3)	-0.375	0.019(*)
(CuO)	-0.527	0.000(***)
(K_2O)	-0.885	0.437
(SiO_2)	-1.070	0.329

As can be seen from Table 2, the coefficients of PbO, SrO and Na2O are positive and have low p-values, which implies that they have a significant effect on the classification of the glass as lead-barium glass. On the contrary, the coefficients of Al2O3, SO2, P2O5, MgO, CaO, Fe2O3, CuO and K2O are negative, which indicates that they have a significant influence on the classification of the glass as a high potassium glass. Among them, CuO has the strongest influence, while MgO and K2O have the weakest influence.

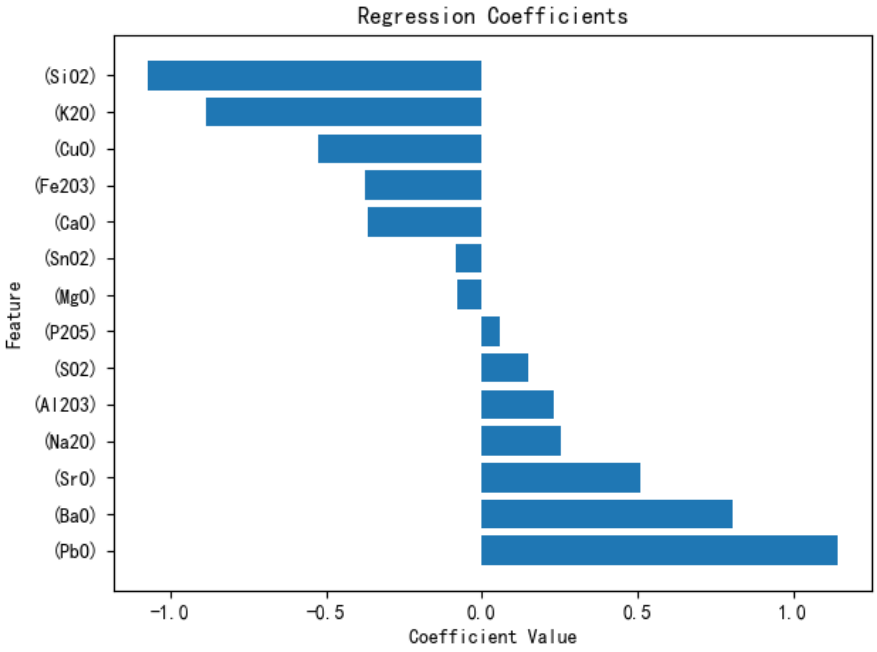

Figure 2. Regression coefficient graph

According to the results of descriptive statistics and independent sample t-test, there are significant differences between high potassium glass and lead-barium glass in different chemical compositions. Among them, the mean values of lead-barium glass are higher than those of high-potassium glass in chemical compositions such as PbO, BaO and SrO, while the mean values of CaO, SnO2 and CuO are lower than those of high-potassium glass in chemical compositions.

Through Figure.2, it is obvious that the chemical composition plays a great role in the classification of high potassium glass and lead-barium glass.PbO, SrO, and CuO play the most significant role in differentiating high potassium glass from lead-barium glass, followed by CaO, SnO2, P2O5, and Fe2O3. while SiO2, K2O, Al2O3, SO2, MgO, Na2O, BaO, etc. chemical components play a relatively small role in the classification of high-potassium glass and lead-barium glass.

In summary, the content of chemical components such as PbO, SrO and CuO may be the basis for the division of high potassium glass and lead-barium glass. When the content of PbO and SrO is high, it belongs to the lead-barium category; on the contrary, when the content of SnO2, CaO and Fe2O3 is high, it belongs to the high potassium category.

2.4. K-Means clustering algorithm

In this paper, the K-means++ algorithm [9], which is an improvement of the K-means algorithm, is used, and its core idea is to improve the way of selecting the initial centroids to make them more representative and dispersed. The specific algorithm is as follows:

1. Select a random sample point from the data set X as the center of mass of the first cluster:

$$C_1 = X_{i_1} \tag{6}$$

2. For each sample point x_i in the dataset, calculate its nearest center-of-mass distance d_i from the selected cluster:

$$d_i = \min_{j=1,2,\ldots,k} \|x_i - C_j\|^2 \tag{7}$$

3. To select a new cluster core, follow these steps:

a. Select sample x_i as a new cluster center of mass with probability $\frac{d_i^2}{\sum_{i=1}^{n} d_i^2}$

b. Repeat step a until k cluster cores are selected.

4. Continue the steps of the K-means algorithm: assign each sample point to the cluster closest to it and recalculate the center of mass of each cluster until the center of mass no longer changes or the maximum number of iterations is reached. Selecting the initial cluster center of mass by the K-means++ algorithm can make the clustering results more accurate and stable.

2.5. Determination of model parameters

For the number of subclasses of the model, we used the elbow rule to confirm that the number of subclasses of both high potassium glass and lead-barium glass were both two, as shown in Figure 3:

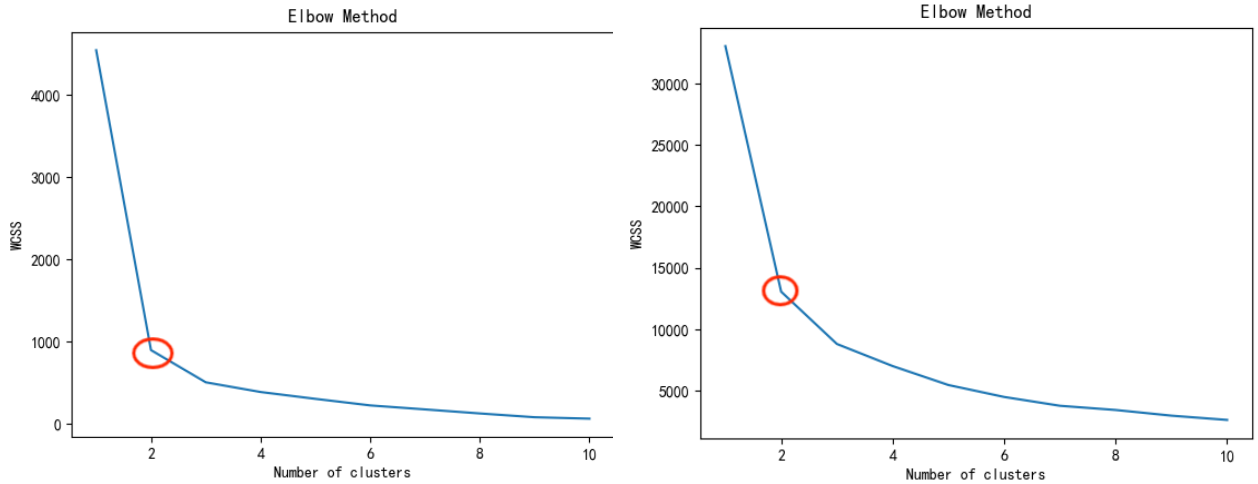

Figure 3. High potassium class and lead-barium class elbow analysis chart

2.6. Analysis of model results

To ensure the reliability of the results, this paper uses both K-Means clustering algorithm, hierarchical clustering algorithm [10] and DBSCAN algorithm [11] to cluster and analyze the data, and uses PCA to reduce the dimensionality of the clustering results as shown in Figure 4 and Figure 5:

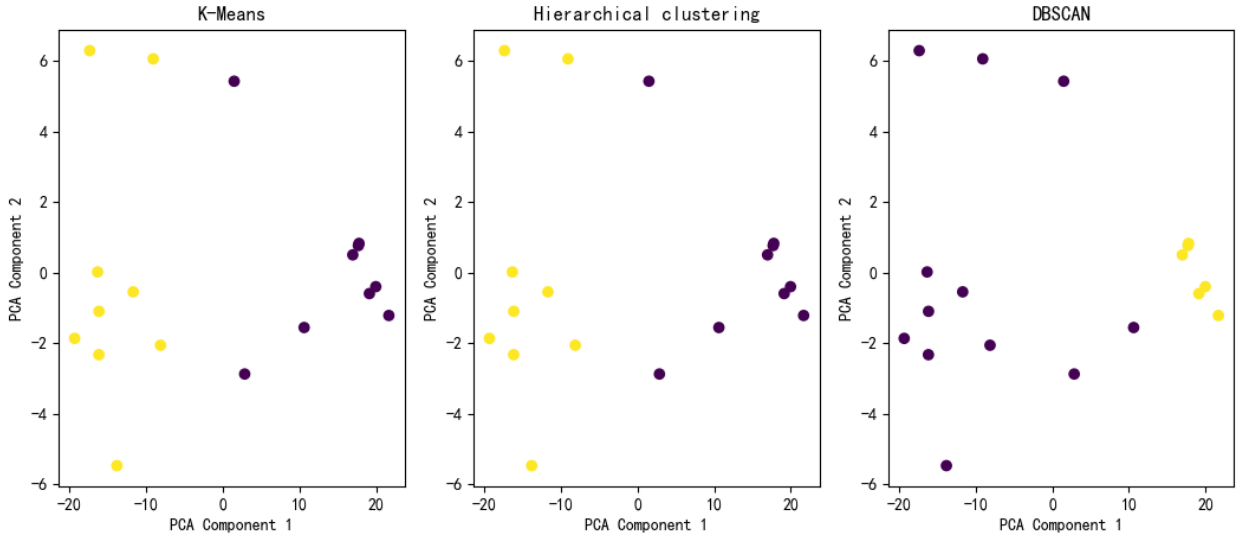

Figure 4. High potassium class

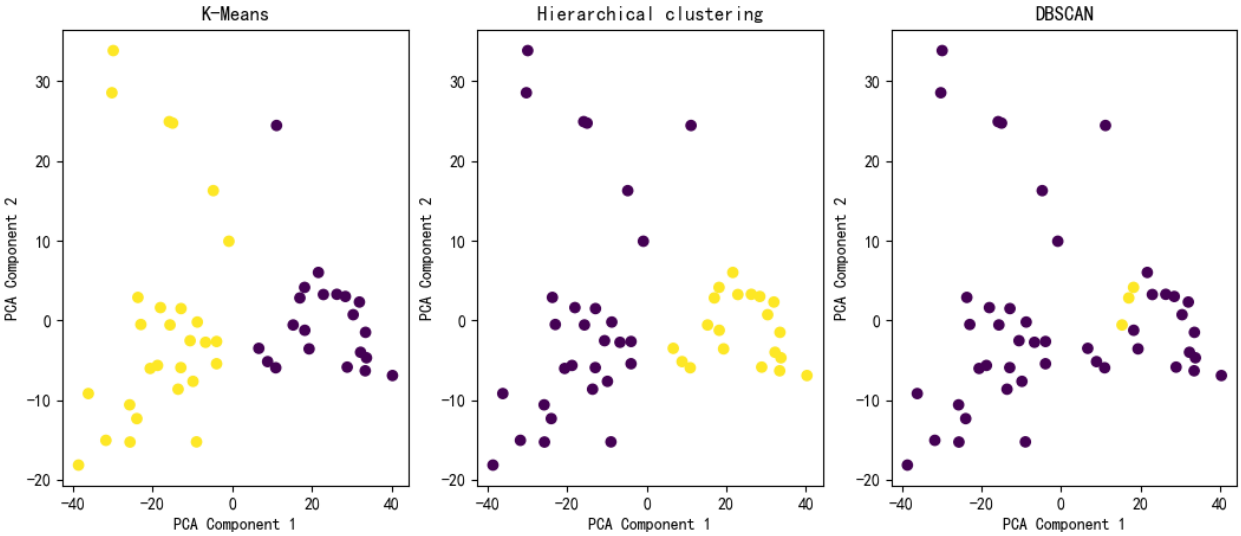

Figure 5. Lead and barium

The three clustering methods were evaluated using contour coefficients and the results are as follows:

Table 3. High potassium class profile coefficient table

	K-Means	Hierarchical Clustering	DBSCAN
Silhouette Coefficient	0.651	0.651	0.587

Table 4. Lead-barium class profile coefficient table

	K-Means	Hierarchical Clustering	DBSCAN
Silhouette Coefficient	0.512	0.502	-0.088

According to Table 3 and Table 4, the contour coefficients of both K-Means and hierarchical clustering are 0.651 in the high potassium class, while the contour coefficient of the DBSCAN algorithm is 0.587. This indicates that the K-Means and hierarchical clustering algorithms have

similar performance and outperform the DBSCAN algorithm in the clustering of the high potassium class. In contrast, in the lead-barium class, the contour coefficients of K-Means and hierarchical clustering were slightly different, 0.512 and 0.502, respectively. but the contour coefficient of DBSCAN algorithm was -0.088, which indicated that the algorithm could not delineate the lead-barium class well.

Therefore, for the subclass classification of the high potassium class and the lead-barium class, we used the results of the K-Means algorithm as our final classification results. The clustering centers are shown in Table 5 and Table 6.

Table 5. High Potassium Class Clustering Center

	(SiO_2)	(Na_2O)	(K_2O)	(CaO)	(MgO)	(Al_2O_3)	(Fe_2O_3)
Cluster1	63.624	0.927	10.818	6.363	1.133	7.349	2.312
Cluster2	89.663	0.000	1.986	1.327	0.437	2.764	0.440
	(CuO)	(PbO)	(BaO)	(P_2O_5)	(SrO)	(SnO_2)	(SO_2)
Cluster1	2.819	0.410	0.579	1.523	0.048	0.000	0.136
Cluster2	1.492	0.139	0.219	0.533	0.008	0.262	0.000

Table 6. Lead-barium class clustering center

	(SiO_2)	(Na_2O)	(K_2O)	(CaO)	(MgO)	(Al_2O_3)	(Fe_2O_3)
Cluster1	24.915	0.172	0.163	2.731	0.602	2.620	0.680
Cluster2	57.490	1.881	0.187	1.142	0.704	5.064	0.624
	(CuO)	(PbO)	(BaO)	(P_2O_5)	(SrO)	(SnO_2)	(SO_2)
Cluster1	2.415	43.448	12.377	4.916	0.448	0.047	1.269
Cluster2	1.165	19.884	7.975	1.128	0.229	0.073	0.174

The content of SiO2, K2O and CaO in Cluster1 is relatively high in the high potassium category, while Cluster2 has the highest content of SiO2 and also contains high levels of Al2O3 and Fe2O3. and Fe2O3 are also relatively high. Therefore, the high potassium category was subdivided into high Al-Fe and high Cu-Zn categories, and the Pb-Ba category was subdivided into high Pb-Pb and high Na-Zn categories based on the significantly different components in chemical content.

In order to check the robustness of the model, the sensitivity analysis was performed using the contour coefficient as the evaluation criterion, and the results are shown in Figure 6.

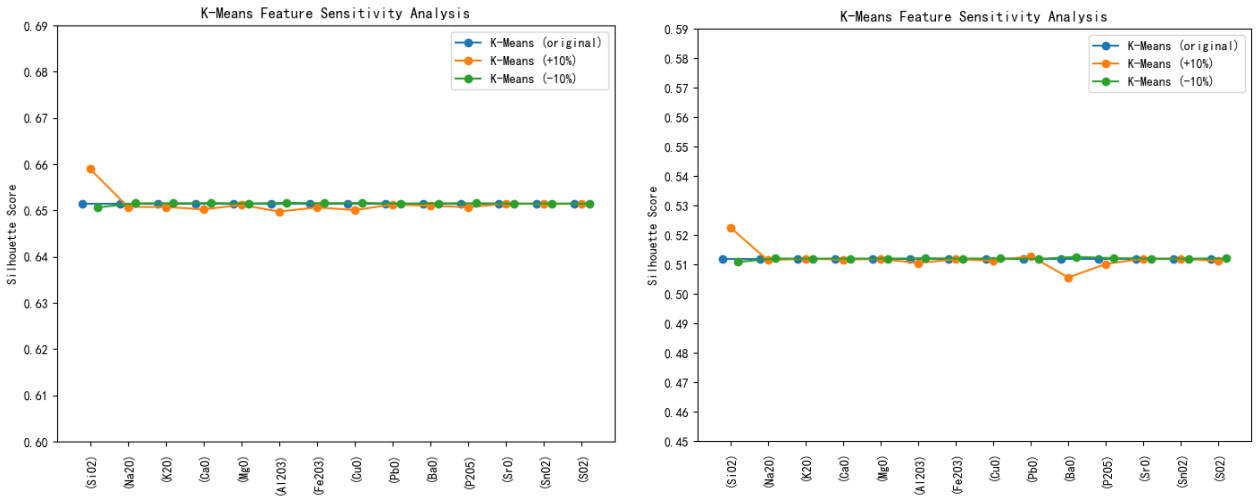

Figure 6. Sensitivity analysis of high potassium class and lead-barium class models

To explore the reasonableness of the clustering results, this paper uses Pearson correlation coefficients[12]to analyze within each cluster as shown in Figure.7 and Figure.8.

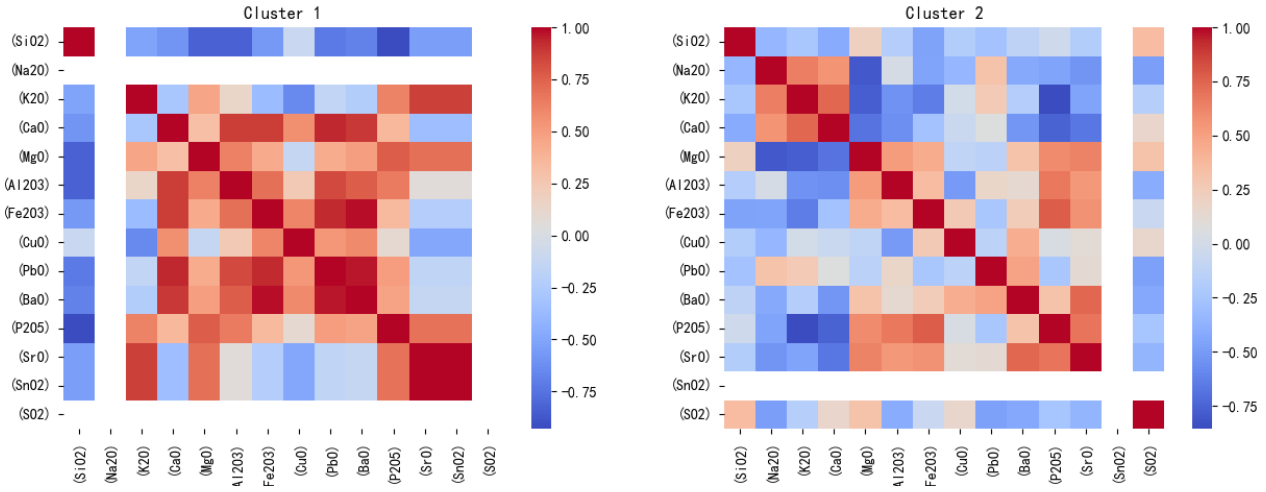

Figure 7. High potassium class

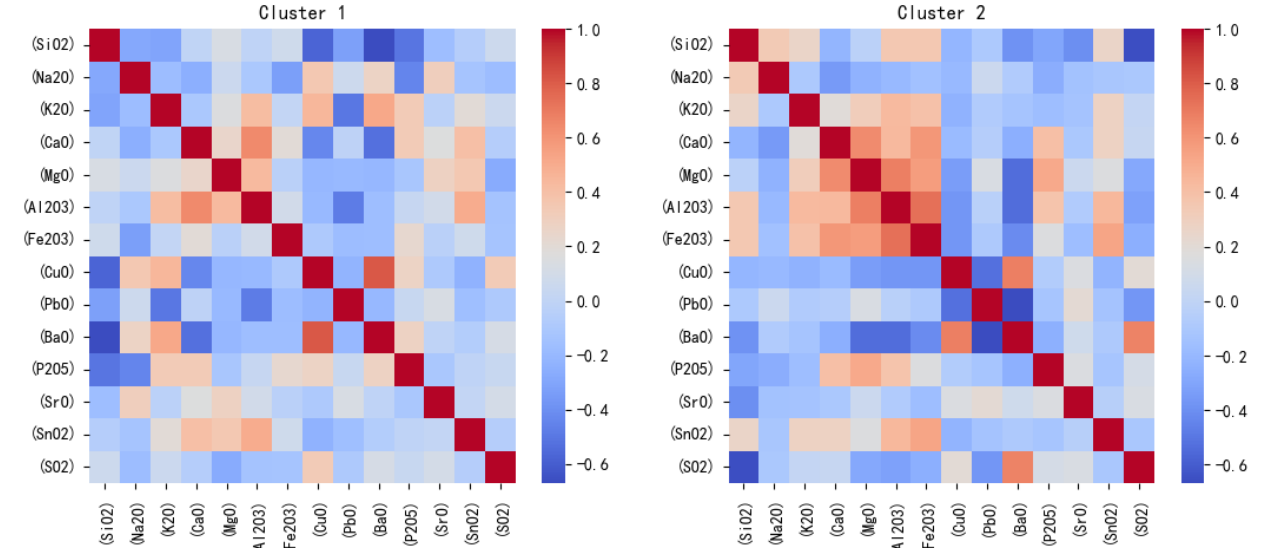

Figure 8. Lead and barium

The figure shows that the data points within the same cluster divided by subclasses all have a high correlation with each other, which proves that the clustering results are reasonable.

3. Conclusion

In this paper, the characteristics of glass artifacts were studied and analyzed using two methods: logit model and cluster analysis. More accurate classification results and classification laws were obtained through the Logit model i.e. when the artifacts have high contents of PbO, BaO, SrO, they belong to the PbB class; when the contents of K2O, SnO2, CaO, Fe2O3 are high, they belong to the high potassium class. This study also successfully divided the glass artifacts into two clusters each by using the K-means++ algorithm, and subdivided the high potassium category into high Al-Fe and high Cu-Zn categories by combining the cluster centers with the chemical composition content, and the Pb-Ba category into high Pb-Pr and high Sn-Zn categories.

In summary, this study successfully studied and classified glass artifacts by analyzing their characteristics using two methods: logit model and cluster analysis, which provides an important reference basis for the identification and conservation of glass artifacts. Meanwhile, the research methods and ideas of this study can also provide references for research in related fields.

References

[1] Gan Fuxi. The road of glass and jade--Another discussion on the cultural and technological exchange between China and foreign countries of pre-Christian silicate artifacts [J]. Journal of Guangxi University for Nationalities (Natural Science Edition), 2009, 15 (04): 6 - 17. DOI: 10.16177/j.cnki.gxmzzk.2009.04.012.

[2] Liu Shuna. Research on glassware of nomads in ancient northern China [D]. Inner Mongolia Normal University, 2022. DOI: 10.27230/d.cnki.gnmsu.2022.000949.

[3] Chen Shuyu. The origin and development of ancient glass in China [J]. Antiquities Identification and Appreciation, 2019, No.151 (04): 44 - 45.

[4] Zhang Yu, Jia Cui, Sun Ou et al. Scientific analysis of inlaid glass ornaments from the Qing Dynasty hanging screens in the Forbidden City Collection [J]. Conservation and Archaeological Science, 2021, 33 (01): 73 - 80. doi: 10.16334/j.cnki.cn31 - 1652/k.20191001594.

[5] Wang, Chenlu, et al. "Application of laser technology in the conservation of cultural relics." Laser & Optoelectronics Progress 59.17 (2022): 1700003.

[6] Cao, Caixia, and Guo, Hong." Classification of disease types of stone, ceramic and glass relics in collections using fuzzy mathematical knowledge." Journal of Beijing Union University (Natural Science Edition) 23.04 (2009): 58-60. doi: 10.16255/j.cnki.ldxbz.2009.04.024.

[7] Gao Yixiang, Yang Minhong,and Li Lanhui." Excel and SPSS analysis of independent sample t-test." Livestock and Feed Science 39.10 (2018): 79 - 82. doi: 10.16003/j.cnki.issn1672 - 5190.2018.10.019.

[8] Zeng Yanbing, Wang Lixia, Zhang Liangwen et al. A principal component logistic regression study of factors influencing life satisfaction of elderly people in nursing institutions [J]. China Health Statistics, 2018, 35 (05): 699 - 702+706.

[9] Yang, J. B., Zhao, Chao, C. A review of research on K-Means clustering algorithm [J]. Computer Engineering and Applications, 2019, 55 (23): 7 - 14+63.

[10] Tao Yang, et al. "A hierarchical clustering algorithm based on DTW distance metric." Computer Engineering and Design 40.01 (2019):116-121. doi:10.16208/j. issn1000 - 7024.2019.01.019.

[11] Li, W. J., et al. "An algorithmic study on adaptive determination of DBSCAN algorithm parameters." Computer Engineering and Applications 55.05(2019):1-7+148.

[12] Li, Hongbin, He, Guangzhong, and Guo, Qiu-ting." A method for similarity retrieval of organic mass spectra based on Pearson correlation coefficients." Chemical Analysis Metrology 24.03 (2015): 33 - 37.