Classification and identification of glass based on principal component and systematic cluster analysis

Wenjie Liu *, Kuntan Zhang, Xinran Ma

College of Mechanical and Electrical Engineering, Northeast Forestry University, Heilongjiang, China

* Corresponding Author Email: 18846757526@163.com

Abstract. The composition of glass relics in different regions is not completely the same. Meanwhile, due to the large amount of internal elements exchanged in the process of glass burial and weathering, in order to determine and find a method to classify and identify glass relics, principal component analysis and systematic clustering are used to select the appropriate indexes for clustering by selecting the composition indexes of existing glass samples. The types of unknown glass artifacts are determined by classification from the types of known glass artifacts. Finally, after multiple cluster analysis, two unknown types of glass relics were identified as high potassium and lead barium. After the identification, the glass chemical composition data were fine-tuned up and down, and it was found that the identification type remained unchanged and the identification effect was stable and accurate.

Keywords: principal component analysis, system clustering, k-means algorithm, glass identification.

1. Introduction

With the continuous progress of archaeological excavations, the number of ancient glass artifacts unearthed in China, along the Silk Road and overseas is gradually increasing, and the surface weathering degree of the unearthed glass relics is different. In addition, after testing the chemical composition, it is found that the chemical composition of glass artifacts in China and overseas is very different. [1-2] All kinds of glass ware can be classified according to their chemical composition, and the production age and other related information can be predicted according to the degree of weathering on the surface of the ware and the chemical composition before weathering. [3-4] This study can also provide more scientific and in-depth evidence for our archaeological work.

2. Classification of glass

2.1. Data processing of glass classification research

1. Data analysis combines the data of glass relics with unweathered surface in the existing forecast data.
2. Data of severe weathering points shall be excluded in data processing.
3. The chemical composition content data of unweathered sites can be regarded as the chemical composition content data of unweathered cultural relics.

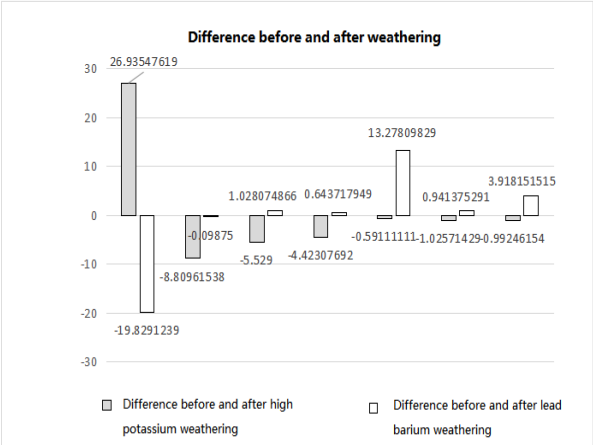
2.2. Glass classification research classification rules

According to the data in the samples, we carried out independent T-test of the chemical composition of the weathered glass samples, calculated the difference of the components before and after weathering, and made a boxplot of the chemical composition content. The independent T-test results are shown in Table 1.

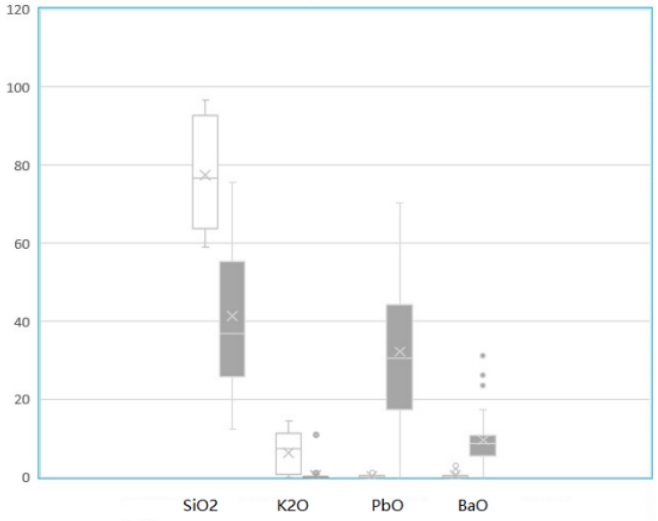
Table 1. Results of independence T-test

component	Silicon dioxide	Sodium oxide	Potassium oxide	Calcium oxide	Magnesium oxide	Alumina	Ferric oxide	Copper oxide
	Lead oxide	Barium oxide	Phosphorus pentoxide	Strontium oxide	The oxide	Sulfur dioxide		
p	0.000	0.889	0.001	0.026	0.371	0.042	0.093	0.927
	0.000	0.000	0.001	0.000	0.501	0.416		

From the following table, we can know that the chemical components that passed the t test are: silica, potassium oxide, calcium oxide, alumina, lead oxide, barium oxide, strontium oxide, sulfur dioxide. Among them, $p < 0.01$, there are six chemical components, which mainly distinguish high potassium and lead barium chemical components. The chemical composition changes of weathered and unweathered glass relics of two types are shown in Figure 1.



(1) Difference of seven components before and after weathering



(2) Boxplot of four components

Figure 1. Weathering and unweathering chemical composition of two types of glass relics

Through the above analysis, we obtained the classification law of high potassium and lead barium:

1. In glass relics of high potassium type, the content of silica is mainly distributed in 63.8-92.7%, while in lead-barium type, the content of silica is 25.7%-55.2%

2. The content of potassium oxide in high-potassium cultural relics is about 0.665%-11.1%, which is far higher than that in lead-barium cultural relics, as can be seen from the boxplot.

3. The contents of lead oxide and barium oxide in lead-barium glass cultural relics are distributed between 17.2%-44.1% and 5.4%-10.8% respectively. It can be seen from the boxplot that the contents of these two components in lead-barium cultural relics are much higher than those in high-potassium cultural relics.

4. After weathering the glass relics of high potassium type, the composition of silica increases significantly, while that of lead barium decreases.

5. After weathering the glass relics of high potassium type, the content of potassium oxide is greatly reduced, and the type of lead and barium is almost unchanged.

6. For the three components of calcium oxide, alumina and barium oxide, the contents of high-potassium type all decrease after weathering, while the contents of lead-barium type increase.

7. For phosphorus pentoxide and lead oxide, the content of high-value types decreased slightly after weathering, while the content of lead barium type increased significantly.

2.3. Methods and results of subclass division

We need to classify the proportion according to the glass components, but too many types of glass components will undoubtedly increase the difficulty and complexity of the analysis problem. In addition, for the classification of glass, although each component is different, different components may be correlated with each other, which is caused by potential factors that play a dominant role in glass classification. In order to select the appropriate components for the subclassification of glass, we first use the principal component analysis, which is a method of dimensionality reduction, which can convert multiple pound-dependent variables into a few pound-dependent variables. The dimensionality of these components was reduced by principal component analysis, as follows:

Step 1: Standardize processing:

$$\tilde{F}_{ij} = \frac{F_{ij} - \bar{F}_i}{s_i} \tag{1}$$

$\bar{F}_i$ and s_i are the mean and standard deviation of the index respectively.

Step 2: Calculate the correlation coefficient matrix of standardized data, and find out the eigenvalues and eigenvectors of the correlation coefficient matrix.

Mark the correlation coefficient between the i index and the i' index as $r_{ii'}$, and the formula is as follows:

$$r_{ii'} = \frac{\sum_{k=1}^{14} \tilde{F}_{ik} \tilde{F}_{i'k}}{14-1}, \quad i, \quad i' = 1,2,...14 \tag{2}$$

Then the correlation coefficient matrix is: $R = (r_{ii'})_{14\times14}$

Step 3: Calculate the eigenvalues of the correlation coefficient matrix λ 1 ≥λ2 ≥ · · · ≥ λ 14 ⊗,and their corresponding eigenvectorsξ1 , ξ2 , · · · , ξ2 ,ξ1 = (μ1i, μ2i, · · ·, μ27i)T consists of 14 new indicator variables from the eigenvector.

$$\begin{aligned} \{Y_1 &= \mu_{1,1}\tilde{F}_1 + \mu_{2,1}F_2 + \ldots + \mu_{14,1}\tilde{F}_{14} \\ \{Y_2 &= \mu_{1,2}\tilde{F}_1 + \mu_{2,2}F_2 + \ldots + \mu_{14,2}\tilde{F}_{14} \\ \{Y_{14} &= \mu_{1,14}\tilde{F}_1 + \mu_{2,14}F_2 + \ldots + \mu_{14,14}\tilde{F}_{14} \end{aligned} \tag{3}$$

Among them, Yi is the ith principal component for statistical analysis.

Step 4: According to the above steps, this paper uses SPSS statistical software, principal component analysis can transform multiple indicators into a few unrelated comprehensive indicators.

The eigenvalues and cumulative contribution rates of corresponding principal components are shown in the following table. The cumulative contribution rate of high potassium principal components is shown in Table 2.The cumulative contribution rate of principal components of lead barium is shown in Table 3.

Table 2. Cumulative contribution rate of principal components of high potassium

Principal	1	2	3	4	5
Cumulative contribution rate	0.2258	0.4388	0.5826	0.7110	0.8160

Table 3. Cumulative contribution rate of principal components of lead barium

Principal	1	2	3	4	5	6
Cumulative contribution rate	0.2534	0.4366	0.5833	0.6775	0.7515	0.8180

Firstly, the cumulative contribution of high price glass is analyzed. The first four principal components provide 81.60% information, which satisfies the principle of principal component analysis. Moreover, the contribution rate of principal component 1 and principal component 2 is relatively large, which can explain that principal component 1 and principal component 2 May be the most important indexes for the classification of high potassium glass. According to the linear coefficient of principal component analysis, principal component 1 is mainly the combination of SiO2, CaO and SnO2, principal component 2 is the combination of K2 O and P2 O5, and principal component 3 is the combination of PbO and SO2. Principal component 4 is the combination of MgO and CuO, and principal component 5 is the combination of K2 O, MgO and SrO. These combinations indicate that SiO2, MgO, K2 O, PbO and SO2 play an important role in the classification of high potassium glass and are more suitable components for high potassium glass than other components.

In the analysis of the cumulative contribution of lead barium glass, the first six principal components provide 81.80% of the information, which meets the principle of principal component analysis. Moreover, the contribution rate of principal component 1 and 2 is relatively large, which can explain that principal component 1 and principal component 2 May be the most important indexes for the classification of lead-barium glass. According to the linear coefficient of principal component analysis, among the physicochemical indexes of lead-barium glass, principal component 1 is mainly the combination of SiO2, MgO, AlO3 and FeO3. Principal component 2 is the combination of SiO2, Na2O, CaO, MgO, P2O5 and SrO. Principal component 3 is the combination of PbO and P2O5. Principal component 4 is the combination of Al2O3, SrO and SnO2. Principal component 5 is the combination of Na2O and SO2, and principal component 6 is the combination of K2 O and CaO. These combinations indicate that SiO2, Na2O, MgO, CaO, P2O5, SrO and SO2 play an important role in the classification of lead-barium glass. Is more suitable for lead barium glass than other components.

In the next step, we will use the appropriate components selected above to carry out systematic clustering and divide the two types of glass relics into subclasses. Systematic clustering, also known as hierarchical clustering, is based on the principle that samples are first regarded as one class, the closest two or several classes are selected after calculating the distance, and then the new classes are used to continue the calculation until all samples are combined into a big class. The flow chart of clustering model is shown in Figure 2.

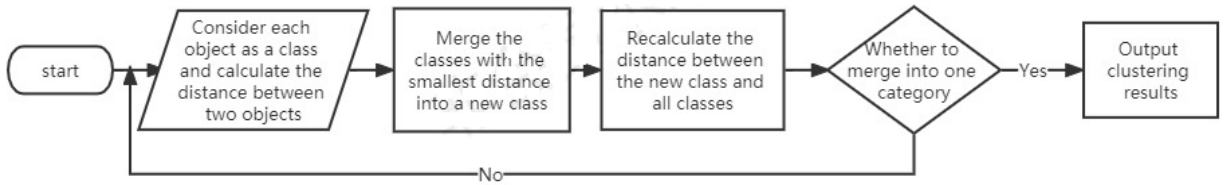


Figure 2. Flow chart of clustering model

Spss was used for calculation, and the following line charts and pedigree charts were obtained respectively (see appendix), and the clustering category k was analyzed. The broken line diagram of clustering coefficient is shown in Figure 3.

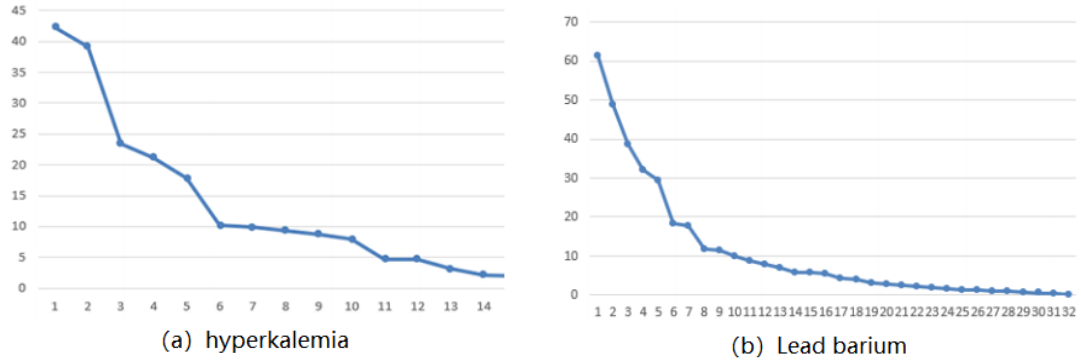


Figure 3. Broken line diagram of clustering coefficient

As can be seen from the broken line chart of clustering coefficient of high-potassium glass, when the classification number exceeds three, the change trend of clustering coefficient slows down. Therefore, high-potassium glass cultural relics are divided into three sub-categories. Similarly, lead-barium glass cultural relics are divided into five sub-categories. According to the genealogy chart, the chemical content data of the unweathered relics in the list, and the predicted data of the weathered relics in the first question, the cultural relics samples were divided into seven sub-categories.

• High potassium type glass is divided into the following subclasses [5-7]:

1. Standard high potassium glass: the overall composition is more uniform, and the composition distribution of most high potassium glass samples is roughly the same.

2. SnO2 high potassium glass: high potassium glass containing SnO2 in its composition.

3. P2-o5-fe2-o3 high potassium glass: the content of p2-O5 and Fe2-O3 in the composition is higher than that of most samples.

• Lead barium glass is divided into the following subclasses:

1. Standard lead barium glass: The overall composition is relatively uniform, and the composition distribution is roughly the same as that of most lead barium glass samples.

2. SO lead barium glass: lead barium glass relics containing a certain amount of sulfur dioxide.

3. P2O5 lead-barium glass: phosphorus pentoxide content is higher than other lead-barium glass relics.

4. SnO2 lead barium glass: lead barium glass containing tin oxide.

The partitioning results are shown in Table 4.

Table 4. Division results

Partition result		Number
High potassium	Standard high potassium glass	01 , 03 , 04 , 07 , 09, 10, 12, 13, 14, 16 , 21 , 22 , 27
	SnO2 high potassium glass	18
	P2O5-Fe2O3 high potassium glass	06
Lead barium	Standard lead barium glass	02, 08, 11, 19 , 24 , 26 , 31-36 , 38-43 , 45-47 , 49-52 , 54-58
	SO2 lead barium glass	37
	P2O5 lead barium glass	20
	SnO2 lead barium glass	48

2.4. Rationality and sensitivity analysis

• Rationality analysis: Since the systematic clustering method is to calculate the distance between classes and subclassify the two kinds of glass samples based on them, we transform the distance

measurement method, choose intra-group distance and centroid distance as the basis for classification and compare with intra-group distance. It is found that the statistical graph of clustering coefficient of all distance methods shows the same trend. The categories of subcategories are basically the same and can be reasonably explained.

• Sensitivity analysis: We use random data disturbance to affect the classification process to test whether the classification result is significantly affected by disturbance. We randomly assigned 10% value fluctuation to some data, and then carried out clustering. The wave of the broken line graph of clustering coefficient was obtained as shown in the figure below. The comparison line graph before and after the random fluctuation of clustering coefficient is shown in Figure 4.

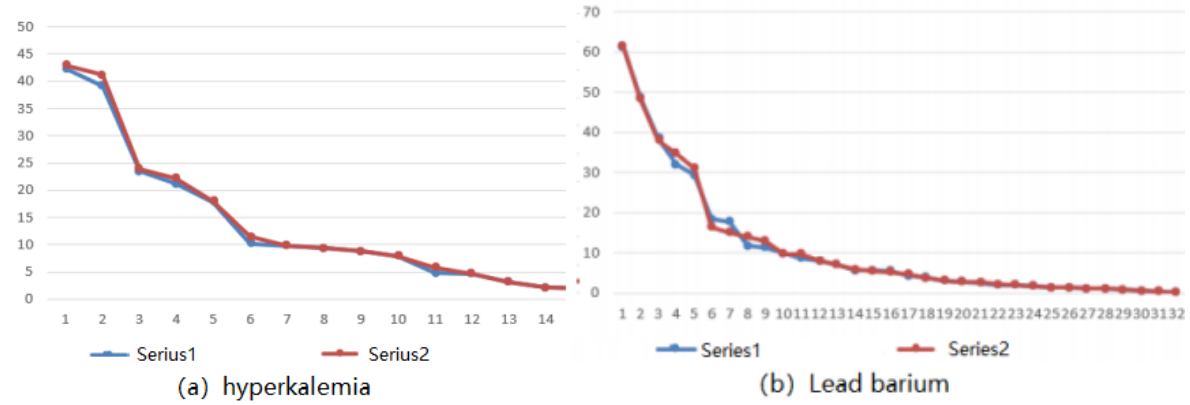


Figure 4. Comparison of cluster coefficient before and after random fluctuation

As can be seen from the figure, no matter for the subclassification of high-priced glass or the classification of the ratio of lead to barium, the values and trends of the broken line graphs of the clustering coefficients after the addition of disturbance data are basically consistent with those before the addition of disturbance. And the classification results did not change accordingly. In conclusion, the model has high stability.

3. Research on glass identification

3.1. Predict the types of unknown cultural relics

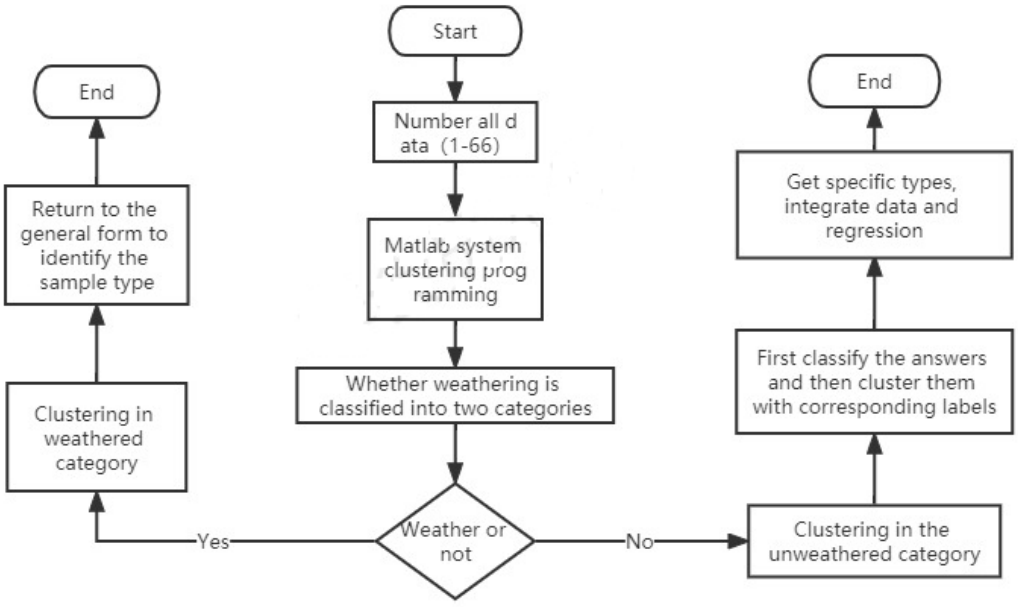


Figure 5. Flow chart of K-means model

The flow chart of K-means model is shown in Figure 5.

Kmeans algorithm is one of the most commonly used clustering algorithms, and its main ideas are: In the case of given K value and K initial class cluster center points, each point (i.e., data record) is assigned to the class cluster represented by the nearest class cluster center point. After all points are allocated, the class cluster center point is recalculated according to all points in a class cluster (take the average value). Then, the steps of allocating points and updating the center point of class cluster are iterated until the change of the center point of class cluster is small or the specified number of iterations is reached. Suppose a given data sample X contains n objects, X = {X1, X2, X3,.. Xn}, where each object has m dimension attributes. The goal of Kmeans algorithm is to gather n objects into the specified k class clusters according to the similarity between objects, and each object belongs to and only belongs to a class cluster whose distance from the center of the class cluster is the smallest. For Kmeans, it is necessary to initialize k clustering centers {C1, C2, C3,...., Ck}, $1 \le k \le n$, and then calculate Euclidean distance from each object to each clustering center, as shown in the following formula:

$$\text{dis}\ (X_i,\ C_i) = \sqrt{\sum_{t=1}^{m} (X_{it} - C_{jt})^2} \tag{4}$$

In the above equation,Xi represents the i-th object, $1 \le i \le n$, Cj represents $1 \le j \le k$ of the JTH cluster center; Xit represents the T-th attribute of the ith object; $1 \le t \le m$; Cjt represents the T-th attribute of the JTH cluster center. Compare the distance between each object and each cluster center in turn, assign the object to the cluster cluster closest to the cluster center, and get k cluster {S1, S2, S3,.., Sk} Kmeans algorithm defines the prototype of the cluster with the center. The class cluster center is the mean value of all objects in the class cluster in each dimension, and its calculation formula is as follows:

$$C_t = \frac{\sum X_i}{|S_i|} \tag{5}$$

Where C1 represents the center of the L-th cluster, $1 \le i \le k$, |S1 | represents the number of objects in the first class cluster, Xi represents the ith object in the first class cluster,$1 \le i \le |S1|$. [8-10]

Table 5. Related information prediction of glass relics of unknown types

Cultural relic number	The number of the artifact whose composition is closest	ornamentation	type	color	Weathering or not	k
A1	21	A	High potassium	blue-green	unweathered	k=15
A2	36	C	Lead barium	Dark green	weathered	k=25
A3	24	C	Lead barium	The purple	unweathered	k=15
A4	30	A	Lead barium	Dark blue	unweathered	k=15
A5	29	A	Lead barium	Light blue	weathered	k=25
A6	7.12.22.27	B	High potassium	blue-green	weathered	k=25
A7	7.22.27	B	High potassium	blue-green	weathered	k=25
A8	46,47	A	Lead barium	Light blue	unweathered	k=15

According to the results calculated by matlab based on the above principles, the classification of unknown glass relics is obtained, and the results are shown in Table 5. The relevant information prediction of glass relics of unknown types is shown in Table 5.

For high-potassium glass, if the k value is 15 or above, it can be concluded that the sample glass is of the same type as 21, 36, 24, 30, 29, etc. Due to the large k value, the number of its centroid positions is large, and the classification is fine. Therefore, when the number is small and the k value is large, Therefore, in this case, the pattern, type and color are all the same. According to the subclass analysis of glass classification study, the samples can be divided into corresponding subclasses.

However, the difference is that for the unknown eighth sample, although its component content clustering is the same or similar to that of 45,46,47. However, it has chemical composition that No. 45.46.47 test sample does not have, and its sulfur dioxide content is extremely high. Therefore, through the observation of chemical component data and the running of MATLAB program, it is found that when k is 4-8, the test sample No.37 and the eighth unknown sample are in the same category. Through data and actual analysis, the eighth unknown sample is classified into lead barium sulfur dioxide sample.

To sum up, the cultural relics numbered A1, A6 and A7 belong to standard high-potassium glass, the cultural relics numbered A2, A3, A4 and A5 belong to standard lead-barium glass, and the cultural relics numbered A8 belong to SO2 lead-barium glass.

3.2. Sensitivity analysis

We give 5% disturbance and conduct K-means clustering analysis again. See the appendix for specific results:

Above, under the perturbation of 5% or 10%, the subclasses of unknown cultural relics are predicted, and the patterns, types and colors do not change. It can be concluded that the model is stable.

4. Conclusions

Glass classification and identification based on principal component and systematic cluster analysis through data analysis and data visualization It is concluded that there is a great difference in the content of silica in high-potassium and lead-barium glass, potassium oxide is mainly distributed in high-potassium glass, lead oxide and barium oxide are mainly distributed in lead-barium glass, and there is a great difference in the composition of the two kinds of glass before and after weathering. Then, the principal component analysis method was used to select the component indexes, select the appropriate indexes and conduct systematic cluster analysis with spss. Finally, according to the chemical composition content, the high-potassium glass was divided into 3 subclasses, and the lead barium glass into 4 subclasses.

Meanwhile, MATLAB system was used to analyze all chemical composition data of weathered glass and unweathered glass, and k-means algorithm was applied to cluster analysis of all data. Parameter clustering of unweathered glass was adopted with k=15, and the type of unknown glass samples was inferred by referring to the chemical composition data and type of known sample glass in the same type. The weathered glass was identified by parameter clustering with k=25 and the same method was used to identify its type. After the identification, the chemical composition data of 10% unknown glass samples were up-regulated and down-regulated respectively. The K-means algorithm was used for clustering analysis again, and it was found that the clustering remained unchanged after the data changes and the type identification was extremely stable.

References

[1] Analysis of the Ancient Silk Road and ancient Chinese glass [J]. Chen Shuyu, Hou Z L. National Museum of China.2019, (05)

[2] Characteristics of glass in early Ancient China [J]. Wang Zifan. Tiangong.2020, (03)

[3] Study on the composition and origin of ancient glass materials [J]. Hu Zhizhong; Li Pei; Jiang Luman; Wang Tongyang; Du Gu. Rock Ore Test. 2020,39(04)

[4] Several Problems on the Research of Ancient Chinese Glass Art.[J]. Gu Liangliang; Zhou Jing. Art and Design (Theory). 2021,2(04)

[5] Weathering of silicate glass [J]. Wang Chengyu, TAO Ying. Journal of the Chinese Ceramic Society,2003(01):78-85

[6] pXRF analysis of some glassware unearthed in Xi 'an and discussion on related problems [J]. Zhao Fengyan; Chen Bin; Chai Yi; Dong Junqing; Li Qinghui. Archaeology and Cultural Relics. 2015, (04)

[7] Research and Application of reference materials for glass composition analysis [J]. Chen Songlin. Petroleum and Chemical Standards and Quality in China. 2022,42(23)

[8] A weighted k-means clustering algorithm and its application [J]. Shen Xiujuan, Xue Shuo. Journal of Qujing Normal University. 2022,41(03)

[9] HE Xuansen, He Fan, Xu Li, Fan Yueping. Determination of Optimal Clustering Number by k-Means Algorithm [J]. Journal of University of Electronic Science and Technology of China. 2022,51(06)

[10] Wang Sen, LIU Chen, XING Shuaijie. k-means Clustering Algorithm Research Review [J]. Journal of East China Jiaotong University. 2022,39(05)