

Glass classification study based on principal component analysis and systematic clustering

Junyang Shao, Zimin Wang *, Kangqi Wang

School of Electrical and Mechanical Engineering, Guangdong University of Technology,
Guangzhou, Guangdong, China, 510006

* Corresponding author: wzm13650832995@163.com

Abstract. Ancient Chinese glass is similar in appearance to Western glass, but the chemical composition and its content are very different. In order to properly explore the classification laws of different categories of glass and further study their subclassification, this paper starts from the sample data and uses systematic clustering and principal component analysis to develop a glass classification study based on the index Euclidean distance.

Keywords: principal component analysis, systematic clustering, glass classification, subclassification.

1. Introduction

With the prosperity of the Silk Road, glass was introduced to China from abroad and had a significant impact on our cultural life, and became one of the symbols of the exchange of multiple civilizations. Early glass was often made into bead-shaped ornaments in the West Asian and Egyptian regions and introduced into China. Our ancient glass lead barium glass and potassium glass but very different from the early glass in the West. Although the appearance of the two are similar, but the chemical composition and its content is very different. The main raw material of glass is quartz sand, that is, the main chemical composition of silicon dioxide. In the glassmaking need to add flux, the purpose of which is to reduce the temperature required to melt pure quartz sand. The type of flux added during production also affects the chemical composition of the resulting glass. The burial process is also very likely to cause changes in the chemical composition of the glass. When buried, some of the glass is affected by weathering, and a large number of elements from the internal and external environments of the glass are exchanged, resulting in changes in the composition ratio of the glass, which has a great impact on the correct determination of the glass category.

In response to the above problem, the classification laws of high potassium glass and lead-barium glass need to be analyzed based on the sample data, in which the glass artifacts in the samples have category labels, and according to the labels, the overall sample data are divided, and the classification laws between them can be analyzed and compared separately by the charts. However, a simple analysis of the graphical data alone is not deep enough.

To address these issues, Mohammad Jamal El-Khatib et al. used neural networks to train and validate the model for glass classification and could predict the glass type by 96.7% [1]; Farahnaz Kolehini et al. used pXRF and Raman spectroscopy for compositional analysis in classifying glasses and proposed some [2]; Bulat N. Galimzyanov et al. developed a proportional viscosity model to classify glasses according to molten glass droplets [3]; Wang Deqiang et al. classified low melting point glasses according to glass composition [4]; Wang Yingzhu et al. classified ancient glass phase artifact types based on the study of glass chemical composition and microstructure [5].

In order to dig deeper into the data information distribution law, this paper does not pay attention to the label information it contains, and establishes a self-identification classification model - using the method of systematic clustering, through the shortest Euclidean distance of the spatial dimension between sample points, the sample points are categorized, and their categorization is completely dependent on the data itself, and by digging deeper into the classification results, the results are used to infer the distribution law of different glass types of chemical composition content distribution law; in order to analyze the type sub-classification method and extract the appropriate chemical

composition, this paper further analyzes the data: in the sample, there are 14 indicators of chemical composition of each glass artifact, the number of indicators is too many, at the same time, after the correlation analysis between indicators, it can be found that there is a large correlation between indicators, which is not suitable for linear analysis and other methods.

In order to analyze the number of indicators and reduce the correlation between indicators, this paper uses the method of principal component analysis [6] and analyzes and extracts the main chemical components represented by each major component as the subsequent subclassification indicators according to the linear expression of principal components. Since the relevant labels for subclassification [7] cannot be determined at present, a self-identifying clustering method, systematic clustering [8], is used to subclassify each glass artifact. Finally, the classification results are analyzed and correlation tests are performed for their rationality and sensitivity. (Flowchart can be added later)

2. Exploration of classification laws

To address the research question, it was necessary to identify the classification patterns of high potassium glass and lead-barium glass based on the data in the sample. In this paper, the problem is reduced to finding objects with similar characteristics that are distinguished into multiple class schemes, and the data are subjected to "systematic cluster analysis" using cluster analysis in multivariate statistical analysis.

2.1. Systematic clustering analysis

Step1: Calculate the minimum distance between objects

Calculate the shortest Euclidean distance $D(x)$ between each sample and the cluster center c_i , and choose the shortest distance among the sample points.

$$D(x) = \sqrt{\sum_{t=1}^m (X_{it} - C_{jt})^2} \quad (1)$$

Where X_{it} is the t th attribute of the i th sample object and C_{jt} is the t th attribute of the j th sample object.

Step2: Merge new classes

Among all the calculated distances, find the two objects with the smallest distance and merge them into a new category.

Step3: Recalculate the minimum distance

Recalculate the Euclidean distance between the new class and all classes by putting the new class.

Step4: Recursive merging

Repeat steps 2 and 3 until all objects are finally combined into one class.

In this paper, the Form II normalized data are divided into data for surface weathering characteristics or not, and systematic clustering is performed separately using SPSS software.

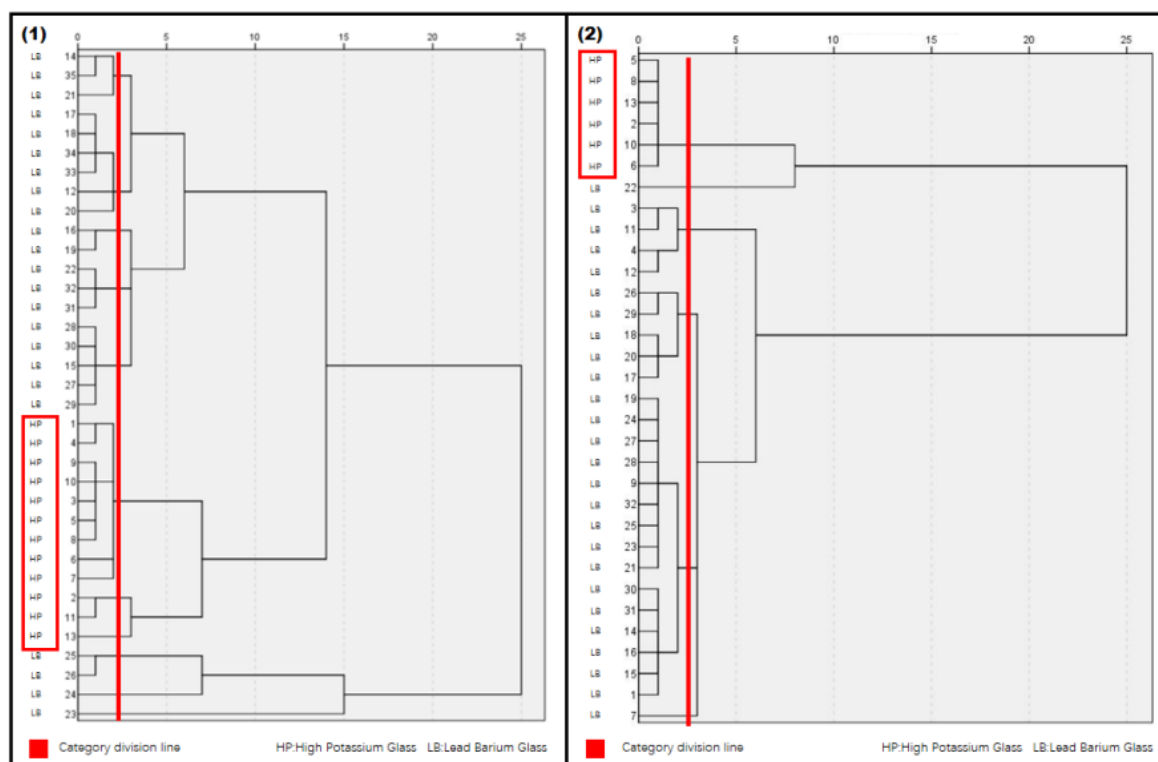


Figure 1. (a) shows the results of unweathered classification (b) shows the results of weathered classification

In Figure 1 above, the location of high potassium is shown in the red box in the figure. The red line is the classification division line, indicating that high potassium and lead acid continue to merge into one category to the right at the location shown on this line, i. e. the characteristics of these two categories have the smallest distance here of all the terms. The analysis of the two classes of glass compositions truncated on the classification division line can be derived by observing the sample data points without weathering and after weathering. Among all the chemical compositions, the two types of glass with large differences between high potassium and lead barium are: K_2O 、 SiO_2 、 PbO 、 BaO 、 CaO . Among the unweathered data, the percentage of chemical composition of CaO for high potassium glass is about 100 times higher than that of lead-barium glass. Also, high potassium glass SiO_2 chemical composition ratio is generally higher than lead barium glass. However, the chemical composition of lead-barium glass PbO 、 BaO is 10 times higher than that of high-potassium glass, and there are some high-potassium glasses that do not contain PbO 、 BaO . In the weathering data, the proportion of SiO_2 in high potassium glass is 40%-70% higher than that in lead-barium glass, and the high potassium glass no longer contains PbO 、 BaO . In contrast, the percentage of PbO 、 BaO in lead-barium glass is still 50% and 10%.

Combining the spectrogram with the attached data, it can be concluded that the condition for determining the distribution pattern of high potassium glass and lead-barium glass is the high or low chemical ratio of K_2O 、 SiO_2 、 PbO 、 BaO , as shown in Figure 2.

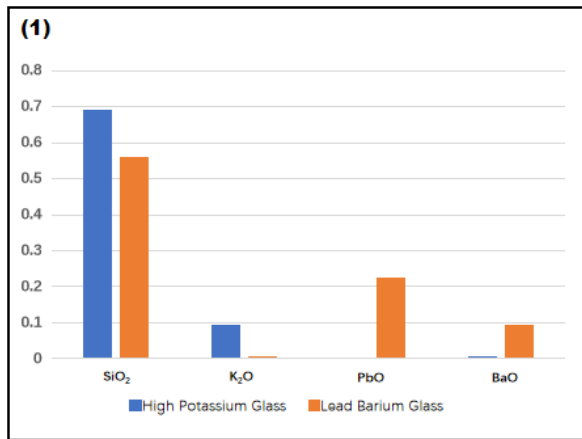


Figure 2 (a) Comparative histogram of glass patterns without weathering

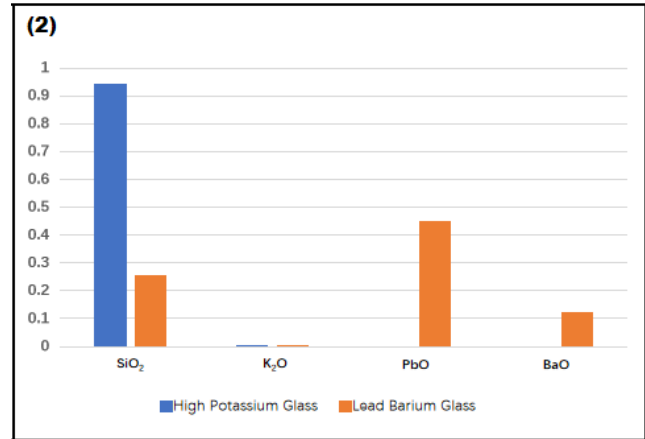


Figure 2 (b) Comparative histogram of glass patterns with weathering

2.2. Subclass classification method and result analysis

First, the data were analyzed, and the chemical composition indicators were selected according to the characteristics of the indicator data, and the subclassifications of the glass samples under the main major categories of high potassium and lead-barium were calculated for the data with unknown classification results using the self-identification clustering method. The steps are as follows.

Correlation test of indicators

Since the chemical composition ratios of glass artifacts recorded in the sample data are mostly discrete data, the correlation tests of the chemical composition indexes of high potassium glass and lead-barium glass were conducted separately by using Spearman correlation analysis.

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$$S_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$S_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} \quad (2)$$

Where, r_{xy} is the correlation coefficient between the indicator and the x y S_x indicator. And S_y are the sample standard deviations of the indicators x and y .

The data of high potassium glass and lead-barium glass in the form were screened and the different types of glass artifacts were imported into the SPSSPRO system separately for Spearman correlation analysis using SPSSPRO [9], and the results of the analysis were thermos graphed as follow figure 3 and figure 4.

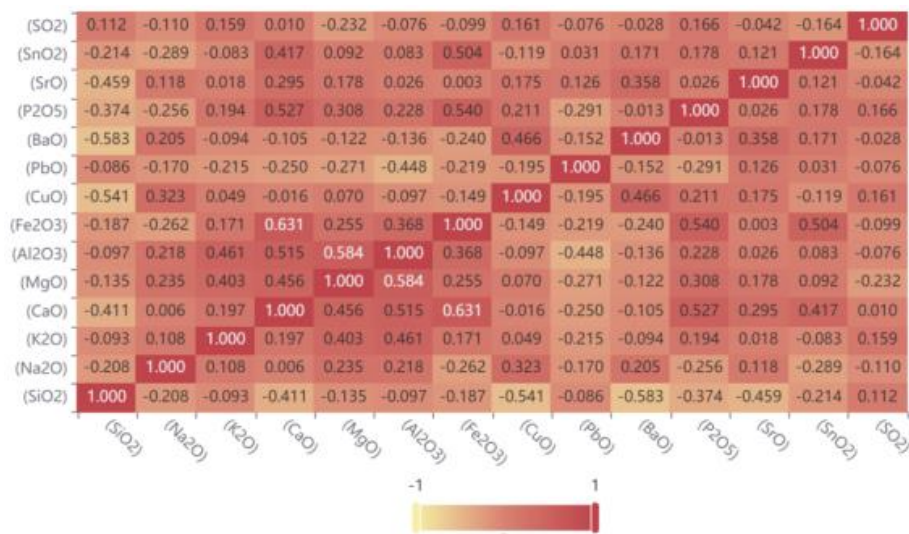


Figure 3. Thermal diagram of chemical composition correlation coefficient of high potassium glass

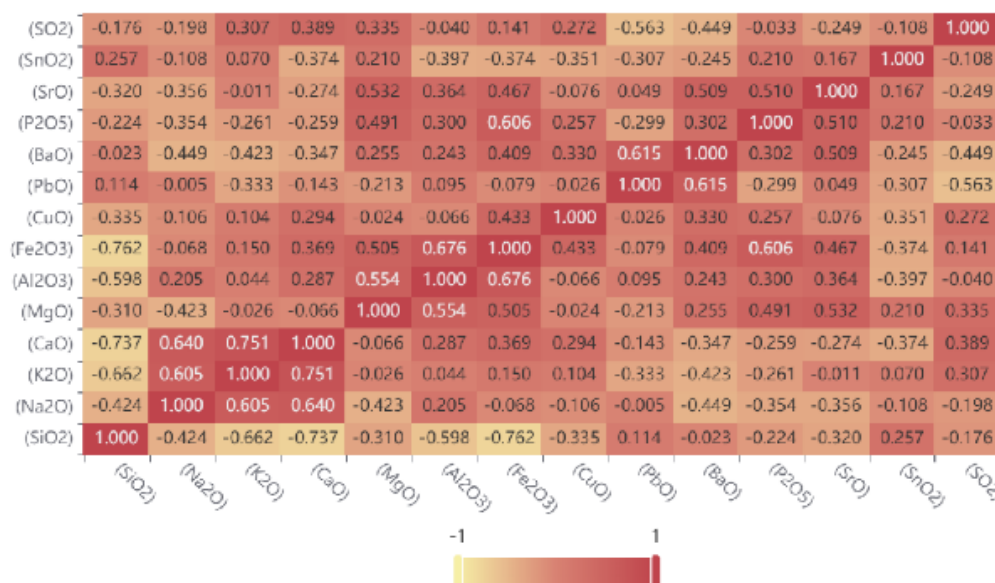


Figure 4. Thermal diagram of chemical composition correlation coefficient of lead barium glass

It is obvious that the chemical composition of both high potassium glass and lead-barium glass has a large correlation with each other.

In addition, the number of indicators of glass substances in the sample data is too large, and in order to reduce the subcategorization dimension and the number of analyzed indicators and their correlation, this paper uses the method of principal component analysis to extract the main indicators used for subcategorization of glass categories.

2.3. Glass artifacts sub-category modeling

Step1: Selecting reasonable chemical composition indicators by principal component analysis

Principal component analysis is a dimensionality reduction algorithm that transforms variable indicators with correlation into principal component indicators that are uncorrelated with each other; combines multiple variable indicators into a smaller number of principal component indicators; and transforms indicators with a single meaning into principal component indicators with a combined meaning.

The amount of data in the annex is limited, and at the same time, there is a certain correlation between indicators, this paper uses principal component analysis to reduce the correlation between indicators while determining the comprehensive impact indicators.

In this paper, the fourteen chemical components of the glass were characterized by principal component analysis to recompose a new set of mutually independent comprehensive combined indicators to replace the original indicators, from which the principal components can be defined as

$$\begin{aligned} F_1 &= a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \cdots + a_{113}X_{13} + a_{114}X_{14} \\ F_2 &= a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \cdots + a_{213}X_{13} + a_{214}X_{14} \\ &\vdots \\ F_{14} &= a_{141}X_1 + a_{142}X_2 + a_{143}X_3 + \cdots + a_{1413}X_{13} + a_{1414}X_{14} \end{aligned} \quad (3)$$

Where F_i is the main component and X_i is the indicator vector. The coefficient a_{ij} is determined as follows.

1. a_{ij} needs to meet

$$a_{i1} + a_{i2} + \cdots + a_{i13} + a_{i14} = 1 \quad (i = 1, 2, \cdots, 14) \quad (4)$$

F_i is uncorrelated with F_j and F_1 is the largest variance of all linear combinations of $X_1, X_2, \cdots, X_{14}$, F_2 is the largest variance except for F_1 , and so on.

$F_1, F_2, \cdots, F_{14}$ are the first, second, ..., and fourteenth principal components of the supplier, respectively.

Calculation of the contribution of the principal component F_i .

$$d_i = \frac{\lambda_i}{\sum_{i=1}^{14} \lambda_i} \quad (5)$$

Cumulative contribution of principal component indicators.

$$\sum_{i=1}^n d_i \quad (6)$$

Among them, the larger the value of d_i , the stronger the information of $X_1, X_2, \cdots, X_{14}$ combined by this principal component. In this paper, the top n principal component indicators corresponding to the eigenvalues with a cumulative contribution of more than 80% are selected and are considered to capture most of the information of the original indicator variables.

Since the contribution rate of the principal component indicator represents the amount of information of the original data covered by the indicator, i.e., the larger the contribution rate, the more information content of the principal component indicator and the more important the indicator is. Therefore, in this paper, the weights can be calculated according to the contribution rate of the selected indicators as follows.

$$w_i = \frac{d_i}{\sum_{i=1}^n d_i} \quad (7)$$

The results of solving the principal component model using Matlab software are as follows, with the following principal component solving indicators for high potassium glass artifacts.

Table 1. Principal component solving index for high potassium glass

Eigenvector	a1	a2	a3	a4	a5	...
(SiO ₂)	0.338	-0.388	-0.016	-0.085	-0.128	...
(Na ₂ O)	0.110	0.416	-0.130	0.408	-0.086	...
(K ₂ O)	-0.026	0.397	0.194	0.274	0.398	...
(CaO)	-0.003	0.525	-0.007	-0.108	0.014	...
(MgO)	-0.326	-0.136	0.276	-0.036	0.271	...
(Al ₂ O ₃)	-0.384	0.110	-0.059	0.199	-0.319	...
(Fe ₂ O ₃)	-0.438	0.054	0.040	-0.060	-0.171	...
(CuO)	-0.130	0.192	-0.125	-0.533	0.211	...
(PbO)	0.001	0.049	-0.580	0.082	0.273	...
(BaO)	-0.235	-0.164	-0.461	-0.198	0.360	...
(P ₂ O ₅)	-0.419	-0.163	0.083	0.097	-0.260	...
(SrO)	-0.410	-0.152	0.033	0.194	0.232	...
(SnO ₂)	0.106	-0.245	0.295	0.342	0.484	...
(SO ₂)	0.005	0.187	0.452	-0.449	0.099	...
Eigenvalue	4.188	3.234	2.094	1.616	1.073	...
Contribution rate	0.299	0.231	0.150	0.115	0.077	...
Cumulative contribution rate	0.299	0.530	0.680	0.795	0.872	...

From the above Table 1, it can be seen that the cumulative contribution of the first five principal components reached 87.2%, so the first four principal components are considered to be better reflecting the indicator information. In the first principal component F_1 , the coefficient of SiO_2 indicator has the largest positive loading indicator, which is also an important distinguishing indicator between high potassium glass and lead-barium glass, so the first principal component can be called SiO_2 principal component. The second principal component F_2 , CaO 、 Na_2O 、 K_2O , has the highest positive loadings and reflects the distribution of chemical components in high potassium glass, so the second principal component can be called the proportional principal component of characteristic chemical components. In the third principal component, PbO 、 BaO has a large loading value, and these two indicator variables are also more important indicators to distinguish high potassium glass from

lead-barium glass, the smaller the value is, the more the glass conforms to the characteristics of high potassium glass, so the third principal component is a proportional principal component of non-high potassium chemical components. The fourth and fifth principal components have limited contribution and relatively evenly distributed loading values, both of which can be considered as principal components representing the proportion of trace element chemical composition.

The solution of the principal component model for lead-barium glass is similar to that for high potassium glass and is not described in detail here.

According to the method of principal component analysis, the glass chemical elements that can reflect the overall characteristics of each principal component were classified and organized in this paper, as shown in Table 2.

Table 2. Principal component chemical element labels

Glass Types	Principal Components	Representative Components
High Potassium Glass	First principal component	SiO_2
	Second principal component	CaO 、 Na_2O 、 K_2O
	Third principal component	PbO 、 BaO
	Fourth principal component	CuO
	Fifth Principal Component	SnO_2
Lead barium glass	First principal component	SiO_2 、 CaO
	Second principal component	BaO
	Third principal component	PbO
	Fourth principal component	K_2O
	Fifth Principal Component	Na_2O 、 SO_2
	Sixth principal component	SrO

In this paper, the principal components of each glass type are used as indicators for the subsequent study. Since this glass type subcategory does not have a classification label, this paper classifies the glass type subcategories by the self-identification classification method - systematic clustering, which is done by the calculation of Euclidean distance in the dimensional space.

Step2: Subclassification of glass types

The data of each principal component indicator is used as the appropriate chemical component indicator data for each category, which is embedded in the systematic clustering model, and the subclass division of glass artifacts is further solved through the continuous aggregation of the sample points with the minimum Euclidean distance in space [10], which finally completes the overall clustering of samples and analyzes the division results.

Using SPSS software, this paper can be analyzed to conclude that High potassium glass spectrum chart as Figure 5.

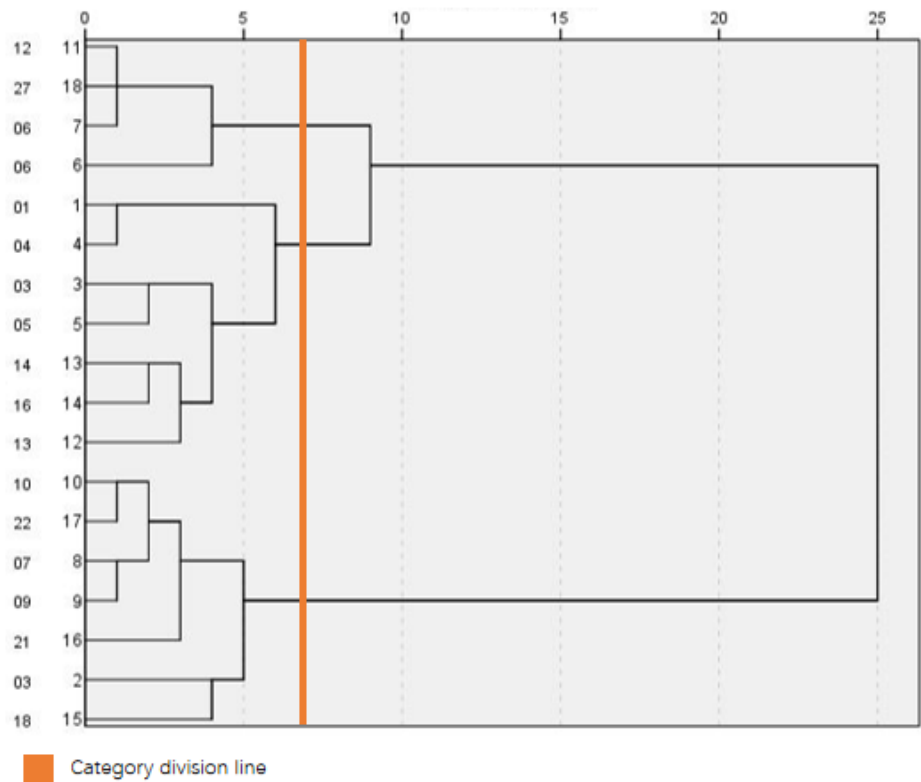


Figure 5. High potassium class division spectrum map

Annex data can be learned, this paper can be divided into three subcategories of high potassium glass: the first category of CaO content greater than 5%, Al_2O_3 content of 4%-7%, Na_2O content of 2%-3%, recorded as high calcium subcategory, the proportion of this subcategory of all samples is 35%; the second category of CaO content of 4%-5%, Al_2O_3 content greater than 8%, and P_2O_5 content greater than 4%, recorded as high phosphorus aluminum subcategory, the proportion of this subcategory of all samples is 23%; the third category of CaO content are less than 3%, Al_2O_3 content of 3%-6%, SiO_2 content greater than 70%, recorded as low calcium subcategory. This subcategory accounts for 23% of all samples; the third category, with CaO content less than 3%, Al_2O_3 content 3%-6%, and SiO_2 content greater than 70%, is recorded as the low calcium subcategory, which accounts for 41% of all samples. According to the spectrum chart of the division results after cluster analysis of high phosphorus glass derived from the analysis, the division method corresponds well with the division results as summarized by the results divided at the vertical line.

The method of classification of lead-barium glass is approximately the same as that of high potassium glass, but the characteristic elements and the number of elements is different. According to the attached data, after traversing all chemical elements. Since the chemical contents are not highly distinguishable except for SiO_2 and CuO contents which can be distinguished into four categories of high and low contents, respectively, these two characteristic contents are chosen as the basis for classification in this paper. In addition, there is a high degree of overlap between the types of samples with high and low SiO_2 and CuO contents, so this paper divides the lead barium glass into two categories: the first category: SiO_2 content of 16%-42%, CuO content of 3%-10%, recorded as the low silicon high copper subcategory, the proportion of this subcategory is 10% of all samples; the second category: SiO_2 content of 38%-72%, CuO content of less than 3%, recorded as the high silicon low Cu subclass, which accounts for 90% of all samples.

The calculation will result in the classification of lead-barium glass spectral chart is highly similar to the above division method, and it can be concluded that the division method of lead-barium glass has some reasonableness.

3. Sensitivity and rationality analysis of classification

Based on the classification results, this paper performs sensitivity analysis on the results of the model without weathering. The larger distance indicates that the chance of classification error is also greatly increased when the difference between the classifications is grouped together when they are larger. Therefore, this paper argues that the Euclidean distance between classifications can be used to measure the accuracy of classification and judge the model sensitivity by it. In this paper, the sensitivity analysis of the classification model and subclassification model in this question is conducted separately, as shown in Figure 6 and Figure 7.

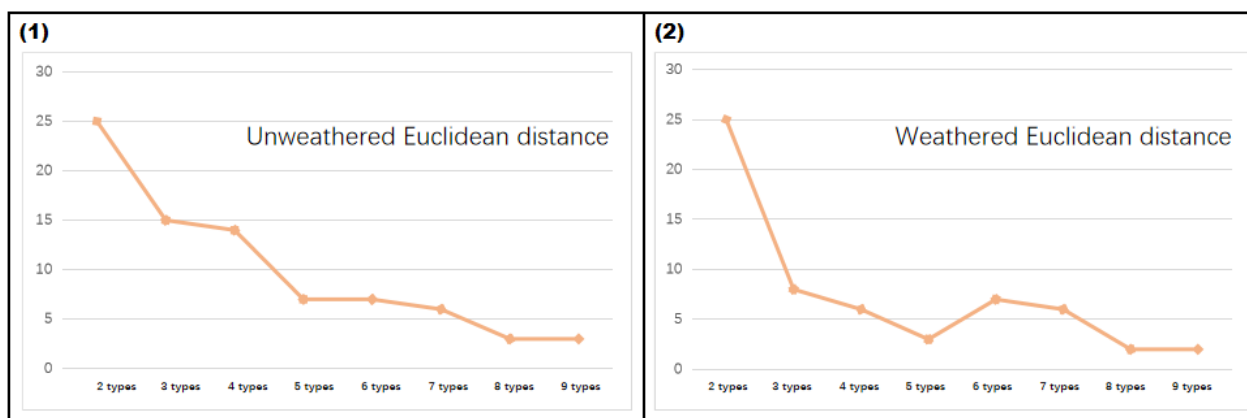


Figure 6. Unweathered and weathered European distance

From the sensitivity analysis graph of the weather-free classification model, it can be seen that the Euclidean distance between clusters of this system fails to show an obvious linear relationship with the number of classifications, but its overall trend has a more obvious decrease in the Euclidean distance between categories as the classification categories increase. Among them, when the classification is two categories three categories between the linear slope is larger, this paper considers this model has a higher sensitivity.

From the sensitivity analysis graph of the weathering classification model, it can be seen that when the systematic clustering divides the types into two categories, the Euclidean distance between categories is the largest, i.e., the difference between the two categories is the largest. With the increase of the number of classification categories (when it is greater than three categories), the change of Euclidean distance gradually flattens out, the difference distance between categories gradually decreases, and the classification accuracy is improving, and the model is more stable.

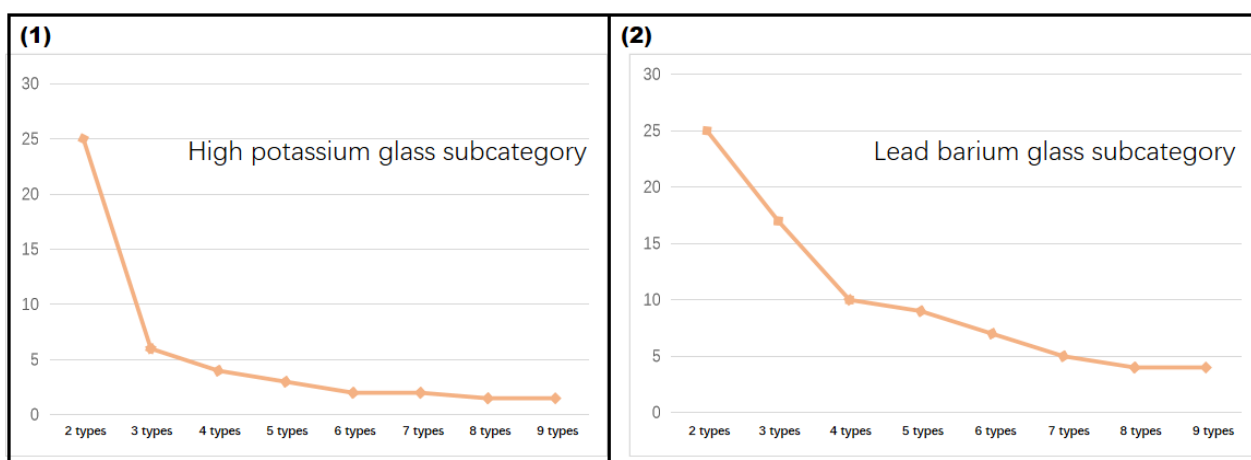


Figure 7. Subclassification sensitivity analysis of high potassium glass and lead-barium glass

From the sensitivity analysis graph of the subcategorization model of high potassium glass, it is clear that the Euclidean distance is maximum when the system clustering category is two, i.e., the difference between the two categories is significant. As the number of classification categories increases (when it is greater than three), the number of classifications is linearly related to the Euclidean distance, and the slope is low, and the model is stable.

From the sensitivity analysis graph of the subcategorization model of lead barium glass, it can be seen that when the systematic clustering classifies the types into two or three categories, the Euclidean distance between categories is larger, and the difference between its categories is large. However, as the number of classification categories increases, the Euclidean distance decreases at a slower rate, and the difference between categories is not large, and the model is more stable when the classification categories are larger than four categories.

4. Conclusion

To address the problem of glass species classification and subclassification, this paper performs systematic clustering among samples by calculating the Euclidean distance among sample points from the chemical composition data of glass samples. Based on the clustering results, the classification rules of glass artifacts are further understood. At the same time, in the process of data analysis, for the problem of too many indicators and large correlation, this paper uses the method of principal component analysis to effectively extract the corresponding chemical composition as sub-classification indicators while reducing the correlation between indicators by dimensionality reduction. Systematic clustering was performed again to classify different types of glasses into subclasses, and the classification results were analyzed and tested for reasonableness and sensitivity. Among them, the subclass of high potassium glass can be divided into high calcium subclass, high

phosphorus and aluminum subclass and low calcium subclass; the subclass of lead and barium glass can be divided into low silicon and high copper subclass and high silicon and low copper subclass.

References

- [1] Mohammad Jamal El-Khatib, Bassem S. Abu-Nasser, Samy S. Glass Classification Using Artificial Neural Network [J]. 2019.
- [2] Farahnaz Koleini, Philippe Colombari, Innocent Pikirayi. Post-15th century European glass beads in southern Africa: composition and classification using pXRF and Raman spectroscopy [J]. Journal of Archaeological Science: Reports, 2020, 29.
- [3] Bulat N, Galimzyanov, Anatolii V. Mokshin A novel view on classification of glass-forming liquids and empirical viscosity model [J]. 2021, 570: 121009.
- [4] Wang De Qiang, Pan Wei, He Li An, et al. The Review of Glass Frit Compositions for Glazing [J]. 2006.
- [5] Wang Yingzhu, Wang Lele, Ma Qinglin, et al. Chemical composition and microstructure of two Warring States period papilliform vessels [J]. 2015, (6): 8.
- [6] Herve Abdi Lynne J. Williams. Principal Component Analysis [J]. 2002, 87(4): 513.
- [7] Maurice Auslander, Idun Reiten. Applications of contravariantly finite subcategories [J]. 1991, 86 (1): 111 - 52.
- [8] Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui et al. Clustering with Deep Learning: Taxonomy and New Methods [J]. 2018.
- [9] Dominique Makowski, Mattan S. Ben-Shachar, Indrajeet Patil, et al. Methods and Algorithms for Correlation Analysis in R [J]. 2020, 5 (51): 2306.
- [10] Diego P.P. Mesquita, João P.P. Gomes et al. Euclidean distance estimation in incomplete datasets [J]. 2017.