A study on the chemical composition of ancient glass products based on gray correlation analysis

Xian Li[1, *] and Ruoqi Wang[1]

[1] School of Mathematics and Science, Chengdu University of Technology, Yibin, China

* Corresponding Author Email: 1572750732@qq.com

Abstract. The ancient high potassium and lead-barium glass made locally at the end of the Warring States period in China not only represents the trade interactions along the Silk Road in China, but also is the crystallization of science and art of the ancient people, and has high research value. In this paper, we first adopt the systematic hierarchical clustering based on Euclidean distance, by dividing different degrees of Euclidean distance to obtain the statistical law of glass type classification. Based on this, a hierarchical clustering model based on PCA dimensionality reduction is used to classify the chemical composition into subclasses. Then, the unknown glass-types were identified according to the classification law, and a classification model based on decision tree cross-validation was established. To test the reliability of the model, we used a comparative analysis with the mean line graph of each component before and after known weathering, and the results were finally found to be reliable. Finally, we used gray correlation analysis to discuss the classification of the two types of glass artifacts, solved the correlation degree of fourteen elements in the two types of glass separately, and analyzed the differences between them.

Keywords: Ancient Glass; PCA; Decision tree; Grey relational analysis.

1. Introduction

As early as the end of the Warring States period, indigenous homemade lead-barium silicate glass appeared in China [1], and ancient glass products are not only susceptible to weathering due to being buried in the ground, but their degree of weathering is also related to a variety of factors. This is because these artifacts usually reflect the historical style and culture of the same period and have a greater authenticity. Therefore, the study of the weathering influencing factors and chemical composition analysis of the native glass objects excavated in China is beneficial to the preservation of ancient glass objects and the identification of glass artifacts excavated.

In this paper, the chemical composition of high potassium glass and lead-barium glass before and after weathering were analyzed by data, and the classification rules of both were summarized. Then PCA was used to reduce the dimensionality of the data to derive the chemical composition to be used. Then the clustering analysis of Euclidean distance was used to classify the type of points close to the data, the basis of subclass classification of the two glasses was studied, and the classes of eight unknown glass artifacts were identified. Finally, gray correlation analysis was used to analyze the association between the chemical composition of high potassium glass and lead-barium glass, and the differences in the association.

2. Classification law of high potassium and lead barium glass

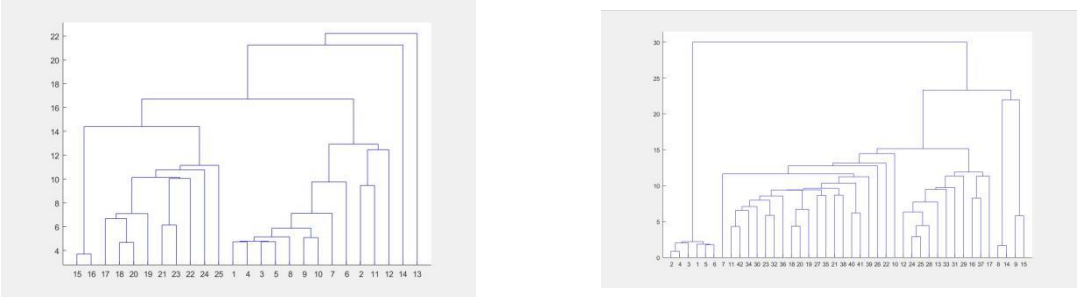
2.1. Classification Law

We take the content of 14 elements of each artifact as the independent variables, and these independent variables can be classified into high potassium glass and lead-barium glass by the glass hierarchical clustering analysis based on Euclidean distance.

Euclidean distance, which calculates the distance between two points in Euclidean space, is the most easily understood distance calculation method, and its Euclidean distance is calculated as follows.

$$dicted = \sqrt{\sum_{k-1}^{m}(x_{ik} - x_{jk})^2} \quad (1)$$

The distances between all the weathered artifacts were calculated, and those with similar distances were grouped into categories to analyze the specific high potassium and lead-barium glass. This is shown in the figure below.



(a) Unweathered (b) Weathered

Fig. 1 Spectra of high potassium and lead-barium glasses before and after weathering

2.2. Subclass division

When the data is reduced from high-dimensional space to low-dimensional, similar features will be merged because of the variance, so the data will be reduced and the number of features will be decreased, which is helpful to prevent the overfitting phenomenon.

So we thought of using principal component analysis (PCA) to eliminate some element categories to reduce the dimensionality and then perform the clustering algorithm. This also allows the algorithm to run faster [2].

(1) Specific steps of PCA

Suppose there are m n-dimensional data

Step 1: Form the original data into an n-row and m-column matrix X by column.

Step 2: Zero-mean for each row of X, i.e., subtract the mean value of this row.

Step3: Covariance matrix

Step4: Find out the eigenvalues of the covariance matrix and the corresponding eigenvectors

Step5: Arrange the eigenvectors into a matrix according to the corresponding eigenvalues from top to bottom, and take the first K rows to form the matrix P

Step6: Y = PX is the data after dimensionality reduction to K dimension

(2) Clustering

Using the same cluster analysis as in the first question, similarities were identified and the same Euclidean distance classification of weathered and unweathered was performed for high potassium.

$$d_{ij} = \sqrt{|x_{1i} - x_{1j}|^2 + |x_{2i} - x_{2j}|^2 + |x_{3i} - x_{2j}|^2 + |x_{4i} - x_{4j}|^2} \quad (2)$$

The error rate (Erate) is also calculated by dividing the amount of variance (magd) by the total sample size (tsas), and the error rate is calculated as follows.

$$Erate = \frac{magd}{tsas} \quad (3)$$

The same is true for lead barium glass.

Based on the data already calculated, draw the lineage diagrams separately.

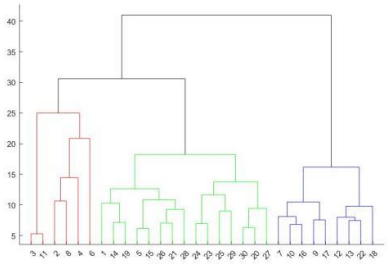


Fig. 2 High potassium glass spectrum

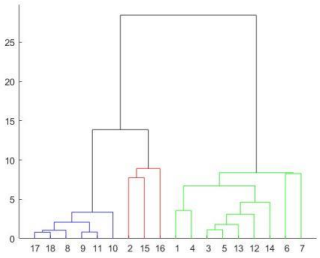


Fig. 3 Lead-barium glass spectrum

3. Glass-type identification based on the decision tree model

We applied a decision tree model [3], named Modletree, and used library functions from Matlab to solve the problem for analysis. The predict function predicts the Modletree and compares it with the test set to find the labels and put the labels into an array. Here, 1 is set as the label of high potassium and 2 is set as the label of lead and barium. The numbers in the test set are transformed into labels and the labels of the test set and the labels of the predicted set are compared for a loop.

The values were perturbed by 1 %,10 %, 15 %, and 20 %, and the training process was simulated with the viewable view, and the classification results obtained were exactly the same. That is, A1, A6, and A7 are high potassium glass, and A2, A3, A4, A5, and A8 are lead-barium glass. We regard the up and down of 80%-120% as correct, then we can analyze that this model has a high correct rate.

The following figure shows the comparison results and the model training process.

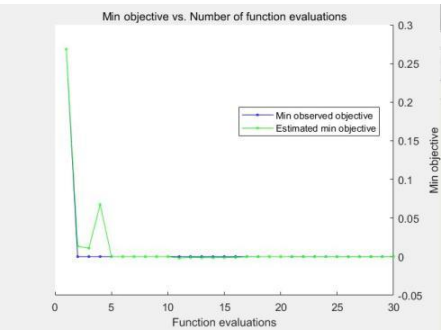


Fig. 4 Comparison of test set and prediction set

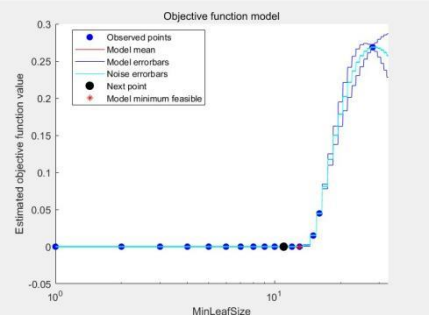


Fig. 5 Simulation training process

4. Analysis of the chemical composition of glass based on grey relational

Grey correlation analysis has obvious advantages for multi-objective response problems and is based on the principle of measuring the degree of correlation between data series based on the fit of the curve geometry [4].

The steps of gray correlation are as follows.

Step 1: Determine the parameter series. Firstly, the data series reflecting the behavioral characteristics of the system should be selected, which can also be understood as finding the mapping quantity of the system.

Step 2: Process the raw numbers, because each factor has different units of measurement, thus the raw data have differences in magnitude and order of magnitude, and the different magnitudes and orders of magnitude do not facilitate comparison, or it is difficult to draw correct conclusions when comparing. Therefore, the raw data are usually dimensionless before calculating the correlation degree.

The behavior sequence of factor Xi is as follows.

$$X_i = (X_{i1}, X_{i2}, \dots, X_{i4}) \tag{4}$$

Step 3: Calculate the correlation coefficient. Set the reference series after data processing as follows.

$$X_0' = \{x_0'(1), x_0'(2), \dots, x_0'(n)\} \tag{5}$$

The comparative numbers are listed below.

$$X_i' = \{x_i'(1), x_i'(2), \dots, x_i'(n)\}, \quad i = 1, 2, \dots, m \tag{6}$$

From the geometric point of view, the degree of correlation is essentially the degree of similarity between the shape of the curve of the reference series and the comparison series. Where the shape of the curve between the comparison series and the reference series is close, the degree of correlation between the two is greater; conversely, if the shape of the curve differs significantly, the degree of correlation between the two is smaller. Therefore, the size of the difference between the curves can be used as a measure of correlation. Then $\Delta_i(k) = |x_0'(k) - x_i'(k)|, k = 1, 2, \dots, n$, the maximum difference and the minimum difference between the two poles are as follows.

$$\Delta(max) = max\,max\Delta_i(k), \Delta(min) = min\,min\Delta_i(k), i = 1, 2\cdots, \mathrm{m} \tag{7}$$

Then the number of correlation coefficients is as follows.

$$\gamma_{0i}(k) = \frac{\Delta(min) + \rho\Delta(max)}{\Delta_i(k) + \rho\Delta(max)}, \rho \in (0,1), k = 1,2, \dots, n;\ i = 1,2, \dots \tag{8}$$

Step 4: Calculation and comparison of correlation degree. Since the correlation degree between each comparison series and the reference series is reflected by n correlation coefficients, the correlation information is scattered and not easy to compare as a whole. Therefore, it is necessary to centralize the correlation information. And the average is a way to centralize the information. That is, the average of the correlation coefficients of the comparison series and the reference series in each period to quantitatively reflect the degree of correlation between the two series, the formula is as follows.

$$\gamma_{0i} = \frac{1}{n}\sum_{k=1}^{n} \gamma_{0i}(k), i = 1,2, \dots, m \tag{9}$$

Therefore, the chemical composition between lead-barium glass and high-potassium glass is correlated in the following table.

Table 1. Relevance Tabler

Category	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO	BaO	P_2O_5	SrO	SnO_2	SO
High potassium	0.893	0.972	0.921	0.981	0.98	0.982	0.951	0.97	0.962	0.986	0.968	0.962	0.962
Lead barium	0.899	0.977	0.969	0.975	0.667	0.973	0.978	0.991	0.987	0.949	0.987	0.959	0.932

The highest correlations were found with PbO for both glass types when BaO or SnO was used as the parent sequence, respectively; among them, P2O5 and SiO2 had the highest correlation with each other for high-potassium glass, and SiO2 and P2O had the highest correlation with each other for lead-barium glass.

Based on solving the correlations among the chemical components, the correlation values of different glass types were compared. The correlation coefficients of the remaining thirteen evaluation terms of lead-barium glass were much more discrete than those of high-potassium glass, regardless of which chemical composition was selected as the parent series, and the correlation coefficients of high-potassium glass were mostly stable in the interval [0.9, 1], while the correlation coefficients of lead-barium glass were more fluctuating. Among them, the gray correlation values of Al2O3 with PbO or CuO as the parent sequence in lead-barium glass are all the lowest values, i.e., the lowest evaluations, and on the contrary, for Al2O3 with PbO or CuO as the parent sequence in high-potassium glass are all the highest values, i.e., the highest evaluations.

5. Conclusions

In this paper, we first used principal component analysis to reduce the dimensionality of the data and then used cluster analysis to analyze the subclassification of high potassium glass and lead barium. However, the criterion of PCA dimensionality reduction is to select the principal component that causes the greatest variance of the original data on the new axis. However, features with small variances are not necessarily unimportant, and such a unique criterion may lose some important information.

In discussing the association and differences in chemical composition between lead-barium glass and high-potassium glass, a gray correlation model is used to classify the two types of glass artifacts in this paper. The gray correlation method is well thought out and can largely reduce the losses due to information asymmetry. However, this method requires the current determination of the optimal value of each index, which is too subjective; moreover, the optimal value of some indexes of the gray correlation method is difficult to determine, which is subject to subsequent improvement.

References

[1] Chen Shuyu. The origin and development of ancient glass in China[J]. Heritage Identification and Appreciation,2019(4):44-45.

[2] Fu Yalong. Research on compositional data processing methods-Example of 1:200,000 rock chip geochemical data in Jianshan-Pingkouxia area, Gansu Province[D]. Shaanxi:Chang'an University,2019.

[3] Quinlan, J. R. (1986). Induction of decision trees[J]. Machine learning, 1(1), 81-106.

[4] Wu, B., Cui, Y. C., Meng, G. C., et al. Structural optimization of drug-type covers based on empowered gray correlation analysis[J]. Energy Containing Materials,2023,31(1):83-91.