Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080214
ISSN 2371-8838 Vol. 8 Num. 2

A study of ancient glass subclassification based on K-means algorithm

Xuan Yang[1,*], Pengao Tian[2], Jiahe Weng[1]

[1]*School of Surveying, Mapping and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, China*
[2]*School of Science, Beijing University of Civil Engineering and Architecture, Beijing, China*
[*]*Corresponding author: 1944641063@qq.com*

Keywords: K-means, high-potassium glass, lead-barium glass, random forests

Abstract: The chemical composition of glass artifacts has an important impact on ancient glass artifacts, this paper explores the various chemical compositions of the artifacts' surfaces by studying the changing law of the chemical composition ratio of the two glass artifacts' surfaces after being affected by weathering and classifying the ancient glass into subclasses. This paper firstly adopts the random forest classification method to explore how to distinguish the chemical composition of high-potassium glass and lead-barium glass to a greater extent under different weathering situations, and finds that SrO2 is the largest determinant for distinguishing the two kinds of glass after weathering, and BaO is the main indicator for determining the category before weathering. In addition, box plots were drawn in the overall dimension to preliminarily screen out reasonable chemical compositions for subclassification. Finally, the k-means clustering method was applied to establish the subclass division model, in which the k values of the model were all taken as 2. BaO was taken for subclass division in lead-barium glass, and Al2O3 was taken for subclass division in high-potassium glass, respectively.

1. Introduction

Glass was an important trading object on the Silk Road and has important cultural value. Quartz sand is the main raw material of glass in ancient China, and its main chemical composition is silica. Refining the different fluxes added, the chemical composition of the glass will produce differences. Such as lead ore as a flux and made of lead-barium glass, its lead oxide, barium oxide content will increase; grass ash as a flux of potassium glass, its potassium content is higher. The stabilizer, limestone, needs to be added during the refining process, and will be converted into calcium oxide when made, which becomes one of the chemical components of the finished glass. As glass artifacts will be weathered in the process of burial, resulting in changes in the proportion of their chemical composition, it is difficult to distinguish between the categories. Among them, the ancient glass with no weathering on the surface may still show the color and decoration of the artifacts, but there may still be shallow weathering in the local area; on the contrary, the ancient glass with large weathering on the surface may also have unweathered areas [1] [2].

There is a batch of data related to the ancient glass products in China, which is divided into two

categories of high-potassium glass and lead-barium glass, and listed the classification information of the artifacts and the proportion of the main components, the cumulative sum of the proportions of the components should be 100%, but due to certain reasons that may lead to the cumulative sum of the proportions of the components and the non-100% of the situation, the study considers the data between 85%~105% of the data as the valid data.

2. Random Forest Classification Methods

Due to the small number of collection points of the studied surface artifact samples, it is necessary to select the more important components among multiple chemical components to serve as an important basis for distinguishing between high-potassium glass and lead-barium glass, and the relationship between each chemical component should be nonlinear because of the transformation and robbing of the elements between each chemical component during the process of glass manufacturing and being weathered later on. In view of the complexity of the data, we choose the random forest classification algorithm to find out the classification law of high potassium glass and lead-barium glass in different states (due to the large difference in the proportion of the content of each chemical component before and after weathering, so we discuss two different periods of glass state before and after weathering).

Random forest is a classification algorithm suitable for the classification algorithm with few samples, high dimensionality and its data are all nonlinear, because the algorithm is composed of multiple decision trees and there is no association between each tree, the calculation is judged by each decision tree and the final decision of the result. Its training set is large, which can reduce the overfitting situation, which is reflected in this problem in the reasonableness of the selection of the classification interval for high potassium glass and lead-barium glass is better than other algorithms.

Steps for determining the classification laws of the random forest algorithm for SPSSPRO web pages:

Step1: Import the support material unweathered and weathered.xlsx into the SPSSPRO web page;

Step2: Determine the various parameters of the model and use a portion of the data to train and build a random forest classification model;

Step3: Calculate the feature importance by the obtained classification model;

Step4: use the remaining data for test data;

Step5: obtain the classification evaluation results of the model, and derive the classification law of high potassium glass and lead-barium glass.

The CART algorithm is one of the quantitative methods of feature selection for decision trees and is based on the following principle:

Select an independent variable from all quantitative independent variables of a certain category of glass, and then select a threshold value to divide the space into two parts, one part of the samples are satisfied greater than this threshold value, while the other part is the opposite, that is, less than the threshold value [3].

Threshold selection: For a variable attribute, its division point is the midpoint of a pair of consecutive variable attribute values. The division of each attribute is ordered according to the amount of impurity that can be reduced. The impurity metric is generally measured here in analogy to the Gini coefficient in economics. The Gini impurity of a node can be defined as:

$$Gini（A）=1-\sum_{i=1}^{C} p_i^2 \tag{1}$$

Where pi is the probability that the sample belongs to class i.

The two sets of data obtained are then recursively processed until the entire 14-dimensional space (14 chemical components) is fully partitioned. The segmentation process is as follows:

Constantly try each attribute of a certain class of glass and the corresponding split point of each attribute, try to find a division with the largest impurity variable, and the subtree divided by that attribute is the optimal branch. If all the samples of the current node do not belong to the same category or there is only one sample left, then this node is a non-leaf node.

For each attribute of the glass is divided to determine the Gini coefficient to get the decision tree, after further optimization to get the optimal result that is the best classification at this time.

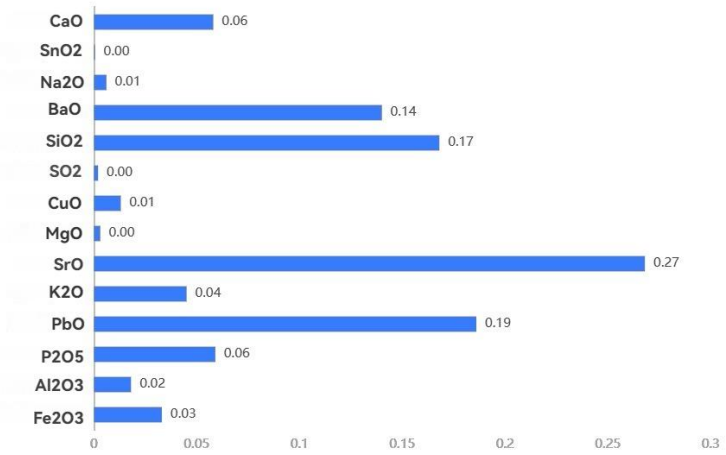

Figure 1: Characteristic importance of each chemical component of the weathered state of the surface.

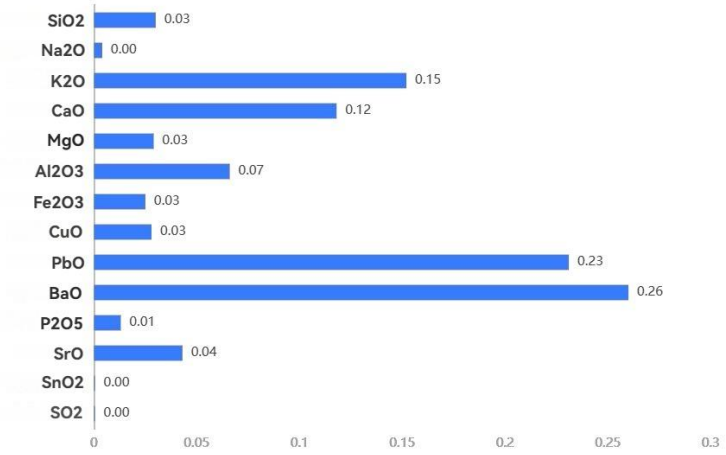

Figure 2: Characteristic importance of each chemical constituent in the unweathered state.

In this paper, the chemical composition corresponding to the characteristic importance greater than 0.1 is taken as an important basis for the classification of glass. As can be seen from the Figures 1 and 2, the classification laws of high potassium glass and lead-barium glass in unweathered and weathered state are as follows:

a. After weathering.

BaO, SiO2, SrO2, PbO play a more important role in the classification of high-potassium glass, lead-barium glass, of which SrO2 has the greatest degree of importance.

b. Unweathered:

BaO, K2O, CaO, PbO play a more important role in the classification of the two types of glass,

with BaO having the greatest degree of importance.

c. In both states, BaO and PbO play more important roles in the judgment of glass types.

3. Subclassification of ancient glass

3.1. Subclassification model based on K-means algorithm

The creation steps are as follows:

Step1: Utilize Excel to select elements that may be used as subclassification indicators;

Step2: Use the evalclusters function in MATLAB to select the chemical molecules that have the most tendency to be clustered by using the k-value calculated by the three clustering internal evaluation methods that come with the function, as well as to determine the k-value selected;

Step3: Use k-means clustering method for subclassification;

Step4: Output the classification situation, the clustering center division ends.

The principle of k-means is as follows:

1) First select K points as the initial clustering center;

2) And then calculate the distance from each single element ratio to the K clustering centers (this question takes Archimedes distance algorithm), to find the closest to the point of the clustering center, it will be attributed to the corresponding clusters;

3) After all the points are attributed, all the points are divided into K clusters. After that, recalculate the average distance from the center of each cluster, and designate it as the new "clustering center";

4) Iterate 2 - 3 steps until the abort condition is reached.

Abort conditions are generally the number of iterations, the minimum squared error MSE, the rate of change of the cluster center, the termination conditions for this question is the number of iterations [4].

There are the following indicators that determine the size of the k-value and determine the order of magnitude of the subclasses assigned.

a.SC: For the purposes of this question is the average of the contour coefficients of the set of all chemical element occupancies. The range of values is [-1,1], with samples of the same category being close together and samples of different categories being far apart and having high scores.

b.CH: Essentially it is the ratio of inter-cluster distances to intra-cluster distances, which is also referred to as the variance ratio criterion. It measures the closeness of the class (intraclass distance) by calculating the sum of the squares of the distances between the points within the class and the class center, and the sum of the squares of the distances between the points at the center of each class and the center point of the data set to measure the separation of the data set (interclass distance).

c. DB: Calculate the average sum of the intra-class distances of any two classes divided by the distance between the centers of the two classes, and find the maximum value. Smaller DB means smaller intra-class distances and larger inter-class distances.

The larger the indicator CH, SC, the better; DB: the smaller the better. Combine the three indicators to determine the size of the k value.

(Due to ch evaluation index in the value of large when the effect is more prominent, the question of the sample is less ch as an evaluation standard is not appropriate, so the following table ch value to determine the optimal k value is for reference only) [5].

Due to the weathering of unweathered different chemical ratio content needs to be classified and discussed.

Step1:

Lead-barium is divided into weathered and unweathered two cases for overall comparison. The

box plots are shown in Figures 3 and 4:

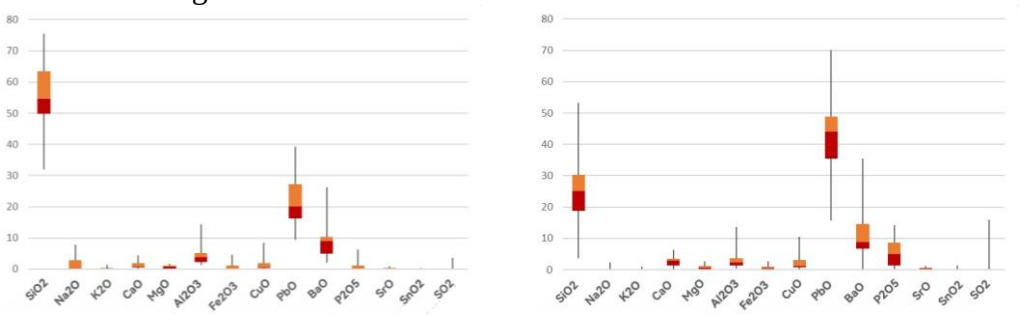

Figure 3: Summary of the proportion of each chemical composition of lead-barium glass before and after weathering.

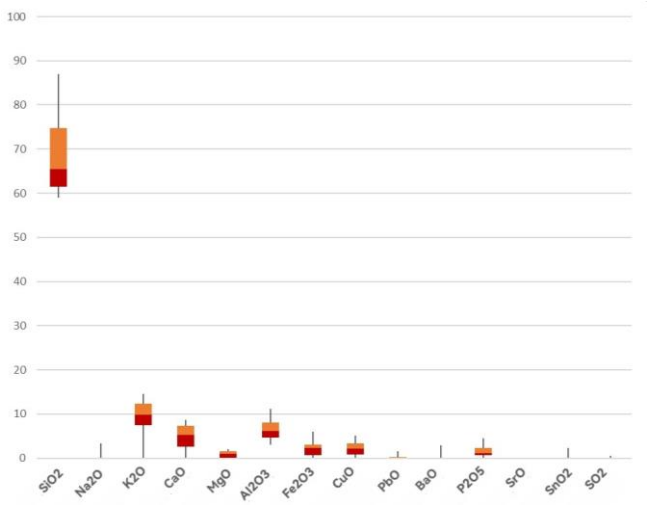

Figure 4: The data are superimposed to show the percentage of each chemical composition in weathered high-potassium glass.

Based on the box plots three to four indicators can be initially screened to determine that they have potential that can be applied to subclassification.

Conditions for screening indicators:

Try to select those with a large interval span, those that account for a large proportion of glass chemical molecules, and those with more valid data, i.e., no 0% and less 0% chemical molecule categories.

In summary can be summarized as follows:

Here lead-barium glass for primary screening, select barium oxide, lead oxide, silica 3 indicators, high potassium glass for primary screening, select silica, potassium oxide, calcium oxide, aluminum oxide 4 indicators.

Step2:

The preliminary identification through the Excel box diagram can be a large reduction in this step in the arithmetic cost. This step in the use of MATLAB functions were taken k = 1-5 and then brought to the DB, SC, CH three types of evaluation indicators function to calculate the Tables 1-3.

Table 1: Evaluation results related to the selection of different k values for barium oxide in weathered lead-barium glass.

K	2	3	4	5
DBI	0.2082	0.5774	0.2980	0.3146
SCI	0.9183	0.7332	0.8702	0.8543
CH	98.6841	118.1064	286.8005	371.9074

Table 2: Evaluation results related to the selection of different k values for barium oxide in unweathered lead-barium glass.

K	2	3	4	5
DBI	0.2358	0.2788	0.3882	0.4223
SCI	0.9366	0.8844	0.7860	0.7886
CH	59.9830	216.4182	244.7418	279.6973

Table 3: Evaluation results related to the selection of different k values for alumina in high-potassium glasses.

K	2	3	4	5
DBI	0.5256	0.4302	0.4485	0.4731
SCI	0.7862	0.7715	0.7173	0.6986
CH	42.4185	69.2026	74.7828	92.5335

According to the nature of the three evaluation indexes, the optimal k-value under the index function can be selected.

Based on the table below, the chemical components to be analyzed in the subcategories should be selected, and the k-values calculated in the previous step are used here.

On the basis of the k value to select the subclassification standard chemical components

1) Due to the small sample data, it is better to choose the clusters with smaller k value to be more meaningful.

2) It is best to select the weathered and unweathered k value is the same in favor of the subsequent merger of subclasses.

3) The selection of the optimal k-value derived from different evaluation methods with small differences can indicate that the k-value is more reliable.

After determining the standard chemical composition of subclassification, the k value is then determined

The results of the previous operation are shown in the Tables 4-6:

Table 4: Lead-barium glass on barium oxide subclassification evaluation of optimal k-value.

	DBI	SC1	CH
Weathered	2	2	5
Unweathered	2	2	5

Table 5: Lead-barium glass on barium oxide subclassification evaluation of optimal k-value.

	DBI	SC1	CH
Weathered	4	5	5
Unweathered	5	2	5

In summary, it is concluded that the subclassification of lead-barium glass by the proportion of barium oxide composition is more obvious, and the internal evaluation index is higher, and it is determined that lead oxide is used as the evaluation criterion for subclassification.

And the value of k is determined to be 2.

Table 6: Lead-barium glass on barium oxide subclassification evaluation of optimal k-value.

	DBI	SC1	CH
Weathered	3	3	4
Unweathered	5	5	5

Table 7: Glass on silica subclassification evaluation of optimal k-values.

	DBI	SC1	CH
SiO2	4	4	5
K2O	5	2	5
GaO	4	4	5
Al2O3	3	2	5

As shown in Table 7, it is more appropriate to choose aluminum oxide as the sub-classification standard for the index of lead-barium glass.

3.2. Subclassification results

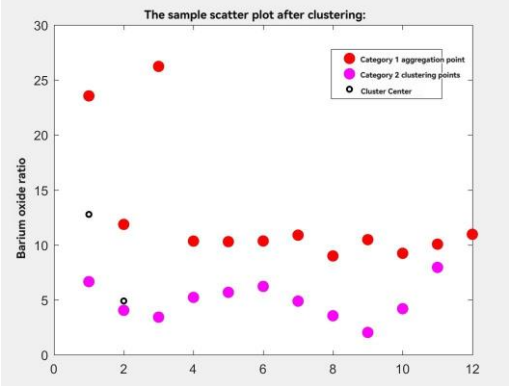

Figure 5: Two-dimensional expansion of unweathered lead-barium glass at k=2.

From the Figure 5, the clustering centroids of barium oxide are 12.7675 and 4.8936, i.e., two subclasses are divided around these two points.

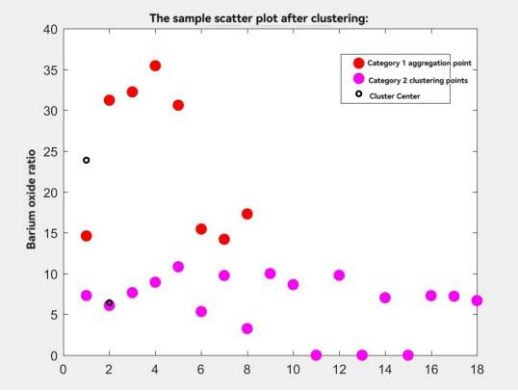

Figure 6: Two-dimensional expansion of barium oxide.

From the Figure 6, the clustering centroids for barium oxide were obtained as 6.4378, 23.8887 respectively. i.e. the two subclasses were classified around these two points for classification.

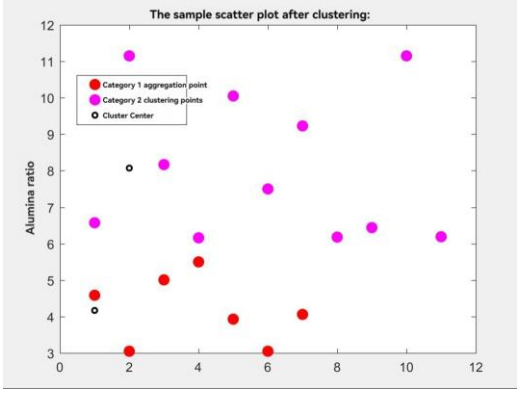

Figure 7: Aluminum Oxide 2D Expanded View.

From the Figure 7, the clustering centroids for alumina were derived as 4.1690, 8.0720. i.e., the two subclasses are clustered around these two points for categorization.

4. Conclusions

In this paper, the sampling points of each artifact surface are firstly divided into two groups of unweathered and weathered. We cannot study the classification law of high-potassium glass and lead-barium glass purely from the number of constituents accounted for, so the extraction can distinguish the main components of the two kinds of glass to a greater extent, and then study the classification law of the two kinds of glass before and after weathering.

References

[1] Jinyun Cao, Tianyu Xu, Yong Liu, et al. Composition prediction and classification of ancient glassware based on RBF and SVM [J]. Science and Technology Innovation and Application, 2023, 13(18):41-43, 48.

[2] Wang Jie, Li Mo, Ma Qinglin, et al. Weathering study of an octagonal columnar lead-barium glass vessel from the Warring States period [J]. Glass and Enamel, 2014, 42(2):6-13.

[3] Xiao Jinjuan, Pang Jinxiang, Chen Wenzhuo. Principal component identification and classification of ancient glass based on random forest model [J]. Science and Technology Innovation, 2023(14):37-40.

[4] Lin Xiaoqing. K-means clustering algorithm applied to online learning behavior research in big data era[J]. Electronic Design Engineering, 2021, 29(18):181-184, 193.

[5] T.J. Xu, P.S. Wang, X. B. Yang. Three-branch k-means clustering algorithm based on artificial bee colony[J]. Computer Science, 2023, 50(6):116-121.