

Study on properties and composition of ancient glass based on PCA and gradient lifting tree model

Hengtian Zhao ^{1,2}, Junjia Du ³, Yayue Xu ², Yujie Li ⁴, Yingai Jin ^{1,2,*}

¹ State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun 130022, China

² College of Automotive Engineering, Jilin University, Changchun 130022, China

³ College of Computer Science and Technology, Jilin University, Changchun 130022, China

⁴ College of Journalism & Communication, Jilin University, Changchun 130022, China

* Corresponding Author Email: jinya@jlu.edu.cn

Abstract. In order to help archaeologists better study ancient glass, uncover the mystery of the early Silk Road, and also better protect these cultural heritage. In this paper, the properties and composition of ancient glass are studied. All types of glass samples were analyzed by PCA principal component analysis and weighted summation, and the type characteristic equation was obtained. After calculating the candidate points iteratively in MATLAB, the optimal critical eigenvalue was set to 5, and the type judgment model was established. According to get the type of the characteristic equation, the selection is related components as high potassium and the composition of class division, choose negative correlation components as lead, barium and the composition of class division, through the PCA principal component analysis of two types of glass data are reduced to 3 d in the class we USES Kmeans clustering algorithm for two types of glass, For the sensitivity analysis of subclass partitioning, we used SOBOL global sensitivity analysis method, and obtained that principal component 1 would have the greatest impact on the model output.

Keywords: PCA, cluster, k-means, sensitivity analysis

1. Introduction

In ancient times, glass is the name of jade. Generally, glass is made of quartz sand, combined with a variety of fluxes such as plant ash and natural natarine, melted at high temperature, formed and made after cooling. Although the final glass products look similar [1], but because the flux added in the process of refining is different, its main chemical composition is not the same, they can be simply divided into lead barium glass, high potassium glass and other types [2].

When experts studied ancient glass, they found that it was highly susceptible to weathering caused by the burial environment. In the process of weathering, the proportion of chemical composition always changes, which seriously affects the correct judgment of cultural relic experts on its category. Archeologists generally classify ancient glass into two types: high-potassium glass and lead-barium glass based on its chemical composition and other detection methods. However, it is difficult to solve the problems of the relationship between the surface weathering of glass relics and the type, decoration and color of glass, the statistical rule of the weathering chemical composition content on the surface of cultural relics, and the correlation between the chemical composition of cultural relics [3]. Therefore, it is necessary for the team to use mathematical expertise to analyze and model the problem and put forward a reasonable plan to help the development of cultural relics.

2. Establishment of glass type judgment model

Firstly, all types of glass samples were processed by PCA principal component analysis and weighted summation, and the glass type characteristic equation was obtained as follows:

$$\begin{aligned} \beta = & 0.0506X_1 + 0.0258X_2 + 0.2203X_3 + 0.1950X_4 + 0.1307X_5 + 0.1793X_6 + 0.1248X_7 \\ & - 0.0207X_8 - 0.1482X_9 - 0.0635X_{10} - 0.0065X_{11} - 0.0771X_{12} + 0.0961X_{13} + 0.0772X_{14} \end{aligned}$$

Set the high potassium code to 0 and the lead barium code to 1. The type eigenvalue of each sample was taken as the abscissa, and the type code was taken as the ordinate, and the points were generated in MATLAB as shown in Figure 1.

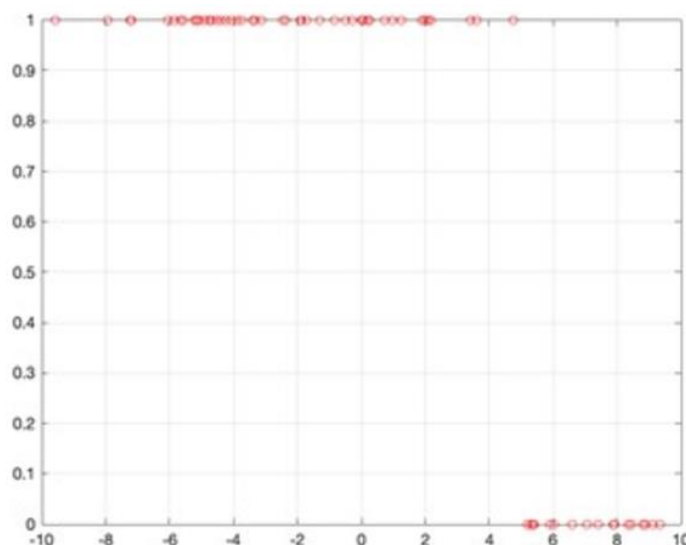


Figure 1. Diagram of the relationship between type encoding value and type eigenvalue

Glass subclass division components is shown as table1, which 1 for high potassium glass, 2 for lead and barium glass

Table 1. Glass subclass division components

	SiO ₂	Na ₂ O	K ₂ O	CaO	MgO	Al ₂ O ₃	Fe ₂ O ₃	CuO	PbO	BaO	P ₂ O ₅	SrO	SnO	SO ₂
1	√	√	√	√	√	√	√	√	√					
2										√	√	√	√	√

It has been observed that the type-coded value is obviously segmented with respect to the type-coded eigenvalue. We take the midpoint absciss of the largest absciss coded as 1 and the smallest absciss coded as 0 as the critical eigenvalue: $L_s = 2$

That is, samples with type eigenvalue greater than 5.0 are high potassium samples, and samples with type eigenvalue less than 5.0 are lead and barium samples. At this point, the glass type judgment model was completely established [4].

3. Subclass division

Before this paper got the glass type characteristic equation, we get a sample of each component of sample types of correlation, because of the type of characteristic value, the greater the the strength of its "high potassium, type characteristic value is smaller, the" lead barium "stronger, therefore we select are related components as high potassium glass composition of class division, negatively related to the selection of components as lead barium glass composition of class division [5].

3.1. Dimension reduction of components for subclass division of two types

We selected the partitioning components of different types of glass, but the dimensions were too high. There were nine indicators of high potassium and five indicators of lead and barium. Before clustering, dimension reduction should be carried out first [6].

After MATLAB calculation and comparison, we chose to reduce both of them to three dimensions. The contribution rate of principal components after dimensionality reduction and the cumulative contribution rate of each type of glass are shown in Figure 2 and Figure 3[7].

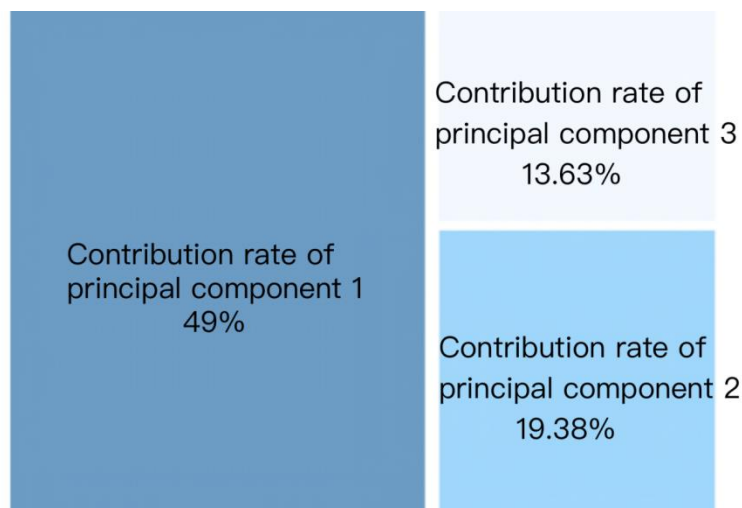


Figure 2. PCA dimensionality reduction results of high-potassium glasses

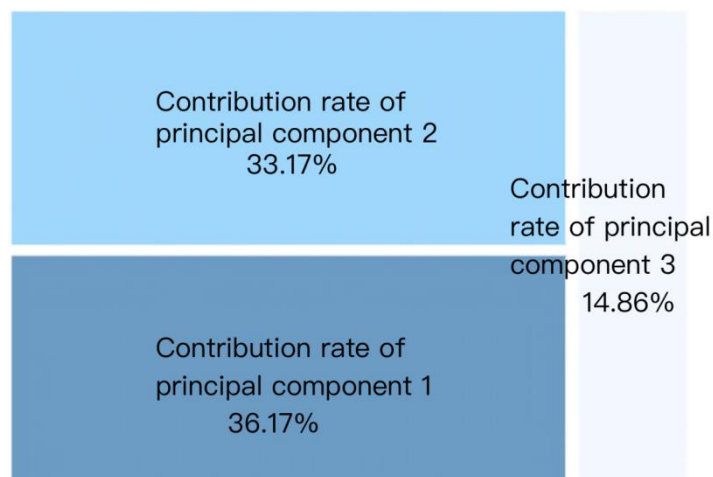


Figure 3. PCA dimensionality reduction results of lead-barium glasses

3.2. Cluster partitioning with Kmeans clustering algorithm

K-means clustering algorithm (Kmeans clustering algorithm) is an iterative clustering analysis algorithm. Its step is to randomly select k objects as the initial cluster center, then calculate the distance between each object and each sub-cluster center, and assign each object to the nearest cluster center.

The algorithm steps are as follows:

Appropriately select the initial center of c classes;

In k iterations, for any sample, the distance to each center of c is obtained, and the sample is assigned to the class of the center with shorter distance.

The center value of the class is updated by means, etc [8].

For all c cluster centers, if the value remains unchanged after iterative update, the iteration ends, otherwise, the iteration continues.

In the rationality analysis, we respectively compared the cluster evaluation index values of glass subclass divided into 2, 3 and 4 categories. By comparing the index values, it can be seen that the optimal number of subclass division is 2 for the high potassium category, while the optimal number of subclass division is 3 for the lead-barium glass.

Figure 4 and Figure 5 show the partition results of high potassium class and lead-barium class respectively by python programming.

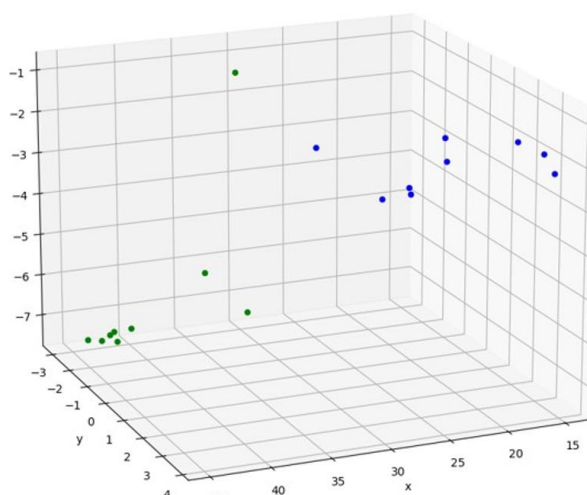


Figure 4. Subclass division of high potassium class

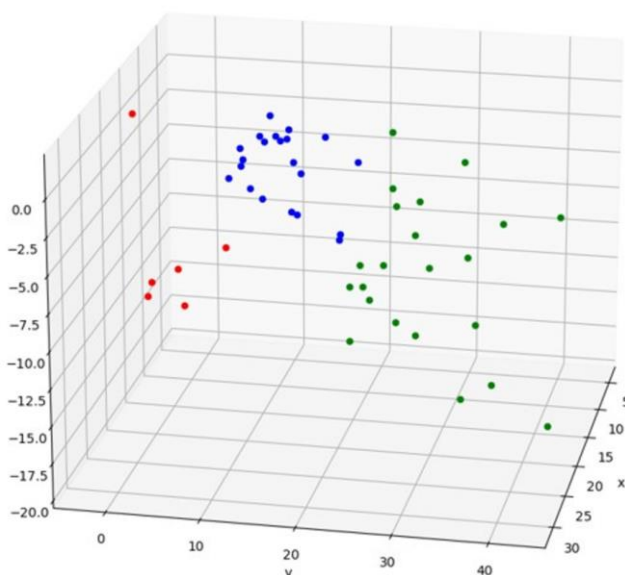


Figure 5. Subclass division of lead and barium

3.3. Rationality analysis of division

For the rationality analysis of clustering effect, three values, namely Silhouette Coefficient, CH Score (Calinski Harabasz Score) and Davidsonburg Deen index (DBI), were selected as evaluation indicators. Silhouette follows class compactness. The Silhouette value is used to describe the similarity between a target and other clusters for the target cluster. The range is (-1, 1). A higher value indicates a higher degree of matching between the target and its own cluster, and a lower degree of matching between the target and other clusters. If the value is higher, the clustering result is better. If the value is too small or negative, it may be caused by too many or too few clusters [9].

The CH score score S is defined as the ratio of between-cluster dispersion to within-cluster dispersion, and is calculated by evaluating between-class variance and within-class variance. A higher score indicates a better clustering effect

The Davisenberg index (DBI), also known as the classification accuracy index, calculates the sum of the mean intra-class distance (CP) of any two categories divided by the center distance of the two clusters to obtain the maximum value. A smaller DB means a smaller intra-class distance and a larger inter-class distance. A smaller DB value means that the clustering results are close within the same cluster, and different clusters are far apart. That is, the smaller the intra-class distance, the larger the inter-class distance. The smaller the DB value, the better the clustering effect. The rationality index is shown in Table 2.

Table 2. Reasonableness Indicators

The class number	Outline of the coefficient of	CH scores	Davidson Burdin index
High potassium type glass class 2	0.6983886934494303	84.96084590595432	0.38167007704847367
High potassium class glass 3	0.6505596592429352	94.10208816186515	0.47737659849300035
High potassium type glass 4 classes	0.5224499307043895	101.79904616743114	0.6343795143772748
Lead and barium glass class 2	0.46425242325214305	54.29956727847498	0.8229705617592876
Lead and barium glass class 3	0.519743945734751	60.66659755606723	0.62280896391525
4 classes of lead and barium glass	0.4756429942558888	66.00633919592762	0.7730878055063998

When the number of subclasses of lead-barium glass and high-potassium glass is 2, 3, and 4, the three evaluation index values can be calculated directly by sklearn package in python. For high-potassium glass, when the subclass partition number is 2, the contour coefficient and Davidsonburg index are the best in three cases, while when the subclass partition number is 3 or 4, only one index is the best in three cases. Therefore, for high-potassium glass, the subclass partition number is 2. For lead-barium glass, when the subclass partition number is 3, the contour coefficient and Davidsonburg index are the best among the three, and when the subclass partition number is 2 or 4, only one index is the best among the three, so the subclass partition number is 3 for lead-barium glass.

4. Sensitivity analysis of partitioning

The sensitivity of the model to the three principal components was analyzed by SOBOL global sensitivity analysis. The Sobol sensitivity index quantifies how much variance each uncertain parameter contributes to the model output. If a parameter has a low sensitivity index, the change of this parameter will lead to a relatively small change in the final model output. On the other hand, if a parameter has a high sensitivity index, a change in that parameter will result in a large change in the model output. The first-order sensitivity of the parameters can be obtained by calculating the formula. The first-order Sobol sensitivity index S measures the direct effect of each parameter on the model variance as follows [10]:

$$S_i = \frac{V[E[Y]Q_i]}{V[Y]} \quad (1)$$

$E[Y]Q_i$ herei Denotes parameter Q_i The expected value of the output Y when fixed.

The first-order Sobol sensitivity exponent tells us when we fix the parameter Q_i When, the expected variance of the model decreases. The total Sobol sensitivity index ST_i includes the sensitivity of the first-order effect and the sensitivity of the interaction (covariance) between a given parameter Q_i and all other parameters (Homma and Saltelli, 1996), which can be calculated from Equation (1):

$$S_{T_i} = 1 - \frac{V[E[Y]Q_{i-1}]}{V[Y]} \quad (2)$$

The Q_{i-1} Said in addition to the Q_i , all uncertain parameters except q . The sum of Sobol sensitivity indices is equal to or greater than 1. If there are no higher-order interactions, the sum of the first-order and total-order sobol indices is 1. The Salib package in python was used to calculate the respective models for the high potassium class and the lead and barium class. The sensitivities are shown in Tables 3 and 4. It can be observed from the table that principal component 1 has the largest impact on the model output in both total and first-order sensitivities.

Table 3. Sensitivity analysis of high-potassium models

High k class	ST	STconf
X_1	0.999601	0.020335
X_2	0.000200	0.000382
X_3	0.000000	0.000000
High k class	S1	S1conf
X_1	0.999601	0.020337
X_2	0.000200	0.000382
X_3	0.000000	0.000000

Table 4. Sensitivity analysis of lead and barium models

Lead barium class	ST	STconf
X_1	1.000092	0.031067
X_2	0.454437	0.010169
X_3	0.001099	0.001268
Lead barium class	S1	S1conf
X_1	0.549500	0.020964
X_2	0.005127	0.016372
X_3	0.000366	0.000670

5. Conclusion

The properties and composition of ancient glass were studied in this paper. For all types of glass sample PCA principal component analysis and weighted summation, get type characteristic equations, iterative candidate in MATLAB calculation, set the optimal critical characteristic value to 5, according to the type of characteristic equation, positively related to the selection of components as high potassium and the composition of class division, choose negative correlation components as lead, barium and the composition of class division, The data of the two types of glass were reduced to three dimensions by PCA PCA. In the subclass division, Kmeans clustering algorithm was used to divide the two types of glass cultural relics. In the sensitivity analysis of subclass division, SOBOL global sensitivity analysis was used. The results show that the principal component 1 will have the greatest impact on the model output.

References

- [1] Pierro Marco, Gentili Damiano, Liolli Fabio Romano, Cornaro Cristina, Moser David, Betti Alessandro, Moschella Michela, Collino Elena, Ronzio Dario, van der Meer Dennis. Progress in regional PV power forecasting: A sensitivity analysis on the Italian case study [J]. Renewable Energy, 2022, 189,13.
- [2] Bucherie Agathe, Hultquist Carolynne, Adamo Susana, Neely Colleen, Ayala Fernanda, Bazo Juan, Kruczkiewicz Andrew. A comparison of social vulnerability indices specific to flooding in Ecuador: principal component analysis (PCA) and expert knowledge [J]. International Journal of Disaster Risk Reduction, 2022,73,90-100.
- [3] Miao Rui, Guo Peng, Huang Wenjie, Li Qi, Zhang Bo. Profit model for electric vehicle rental service: [3] differential pricing strategy and differential analysis [J]. Energy,2022,249.
- [4] Huang Tao, Xiong Qingwen, Du Peng, Qian Libo, Ding Shuhua, Li Zhongchun, Deng Jian. Study on parameter importance ranking [4]. Progress in Nuclear. A study on the OECD/NEA PKL4 test i2.2 with efficient moment-independent sensitivity analysis [J] Energy, 2022, 14,7 (in Chinese).
- [5] Herran Alonso Juan, Saul Luis Angel, Perea Luque Juan Rafael, Lopez Gonzalez M. Angeles. A Principal Components Analysis [5]. Journal of Constructivist Psychology, 2002,35(1). (In Chinese)
- [6] Liu Xiaojing. Labor Market Resource Allocation Optimization Based on Principal Component Analysis [J]. Journal of Mathematics, 2022,202, 2.

- [7] Riccardo De Flaviis, Giampiero Sacchetti, Dino Mastrocola. Wheat classification according to its origin by an [7] implemented volatile organic compounds analysis [J]. Journal of Food Chemistry, 2021, 341 (Pt 1).
- [8] Ali Dabba, Abdelkamel Tari, Samy Meftali, Rabah Mokhtari. Gene selection and classification of microarray data method [8] based on mutual information and moth flame algorithm [J]. Expert Systems with Applications, 2021, 166 (prepublish).
- [9] Performance Analysis of KMeans and KMediods Algorithms in Air Pollution Prediction [J]. International Journal of Recent Technology and Engineering, 2020, 8 (5).
- [10] S. Suganya, T. Meyyappan, S. Santhosh Kumar. Performance Analysis of KMeans and KMediods Algorithms in Air Pollution Prediction of air pollution [J]. International Journal of Recent Technology and Engineering (IJRTE), 2020, 8(5).