Subclass classification of ancient glassware based on K-Means and GMM

Ke Xu *, Yifan Chen, Jiayi Nie

Sichuan University Pittsburgh Institute, Sichuan University, Chengdu, China

* Corresponding Author Email: sitdownkevin@stu.scu.edu.cn

Abstract. This paper focuses on the composition analysis and type identification of ancient glass objects. This paper establishes a judgment model based on the surface characteristics of cultural relics samples, and analyzes the classification laws of high-potassium glass and lead-barium glass; then, the samples of the same category are classified into subclasses according to K-Means method, Gaussian mixture method (GMM) and Mini-Batch K-Means method, and according to each evaluation method, the K-Means method is chosen comprehensively to establish The K-Means method was used to establish a clustering model for glass samples, and the results of subclass classification under each category were obtained. Based on the composition and surface characteristics of the samples, descriptive statistics, scatter plots, S-W test and Pearson correlation coefficient were used to analyze the correlations of chemical compositions among samples of different categories; then, based on the subclass classification made in Problem 2, the differences of the correlations among samples of different categories of artifacts were analyzed using the methods of correlation analysis and difference analysis based on different ranges of data sets.

Keywords: Ancient glass; K-Means; GMM; subclass classification; S-W test.

1. Introduction

Glass, whose main raw material is quartz sand, the main chemical composition of silicon dioxide (SiO2). As one of the important goods carried on the Silk Road in ancient times, it was introduced to our country from the West Asian and Egyptian regions, and then modified by our ancient people in situ to have a different chemical composition [1]. In fact, the differences in chemical composition depend to a large extent on the fluxes added during refining to lower the melting temperature [2]. Among the various glasses of different compositions, lead-barium glass as well as potassium glass were widely popular in China in ancient times [3].

There is a body of data on ancient glass products in China, and archaeologists have analyzed the chemical composition of the samples and classified them into two types [4]: high potassium glass and lead-barium glass. In addition, archaeologists have provided three basic data based on the information obtained from the tests, recording the basic information of glass artifacts, the percentage of chemical composition of classified glass artifacts and the percentage of chemical composition of unclassified glass artifacts, respectively [5,6].

In this paper, we will analyze the classification rules of high potassium glass and lead-barium glass; select the appropriate chemical composition for each category for subcategory division, and give the specific division method as well as the division results, and at the same time, analyze the rationality and sensitivity of the classification results.

Based on the above issues, the correlations between the chemical compositions of glass artifact samples of different categories will be analyzed, and the correlations between the chemical compositions of different categories will be compared and their differences will be analyzed.

2. Model assumptions and notation

2.1. Assumptions [7]

1. Assume that each variable in the sample conforms to a normal distribution

2. Assume that all data measurements are true and valid, without considering the effects of human error

3. It is assumed that the small amount of data is also representative

4. the significance level in this paper is 0.05.

2.2. Notations

Important notations used in this paper are listed in Table 1.

Table 1. Notations

Symbols	Meaning
x_{i1}	The decoration of the i-th sample
x_{i2}	Type of sample i
x_{i3}	Color of sample i
x_{i4}	Surface weathering of sample i
$y_{ij}\ j = 1, 2, \ldots, 14$	Chemical composition of sample i
w_{ij}	Weighting weights of sample i
m	Number of clusters in the clustering model

3. Model construction and solving

3.1. Analysis of classification laws for two types of glass

(1) Direct judgment by color

For black, green, light green and glass samples with missing color, they are all types of lead-barium glass, so they can be judged directly by color.

(2) Judgment by color and ornamentation

(1) Among all samples, ornament B is only found on blue-green high-potassium glass; conversely, if ornament A or C is present and the surface is weathered, the glass can be judged directly by color.

or C and the surface is weathered, the sample is a lead-barium glass.

(2) For all light blue samples, if they have motif A, they are lead-barium glass.

(3) Judgment by color, ornamentation and surface weathering

1. for the blue-green sample, if it has ornament A or C and the surface is not weathered, it is high potassium glass; conversely, if it has ornament A or C and the surface is weathered, it is lead-barium glass; 2. for the light blue sample, if it has ornament A or C and the surface is weathered, it is lead-barium glass

2. for the light blue sample, if it has ornament C and the surface is not weathered, it is high potassium glass; conversely, if it has ornament C and the surface is weathered, it is lead-barium glass.

(4) Judgment by color, ornamentation and chemical composition content

1. for the dark blue sample, if it has ornament A and does not contain iron oxide (Fe_2O_3), lead oxide (PbO) and barium oxide (BaO), it is high potassium glass; on the contrary, if it has ornament A and contains iron oxide (Fe_2O_3), lead oxide (PbO) and barium oxide (BaO), it is lead-barium glass.

2. for light blue samples with ornament C and a high proportion of potassium oxide (K_2O) in all chemical components, high potassium glass; conversely, if the proportion of potassium oxide (K_2O) in all chemical components is very low, lead-barium glass.

3. For the dark green sample, if it has ornamentation C and high content of potassium oxide (K_2O), calcium oxide (CaO) and aluminum oxide (Al_2O_3), and does not contain barium oxide (BaO), it is a high potassium glass; on the contrary, if it has ornamentation C and low content of potassium oxide (K_2O), calcium oxide (CaO) and aluminum oxide (Al_2O_3), and contains barium oxide (BaO), it is a lead-barium glass.

According to the above judgment method, the overall classification law of the two types of glass samples can be summarized.

3.2. Subclass classification based on cluster analysis method

Considering the large variability of the original data, it is difficult to distinguish the data characteristics of some samples by only dividing all data into 2 clusters, and both evaluation methods give high s values for clusters 2 and 3. The K-Means method was used to cluster the data into clusters of 3[8,9].

To further demonstrate the clustering effect of K-Means with images, the data were dimensionalized using the PCA principal component analysis method.

Considering the intuitiveness of two-dimensional images and the higher interpretability of the first two eigenvalues, it was decided to use two-dimensional data for visualizing the clustering results, which are shown in Figure 1 below.

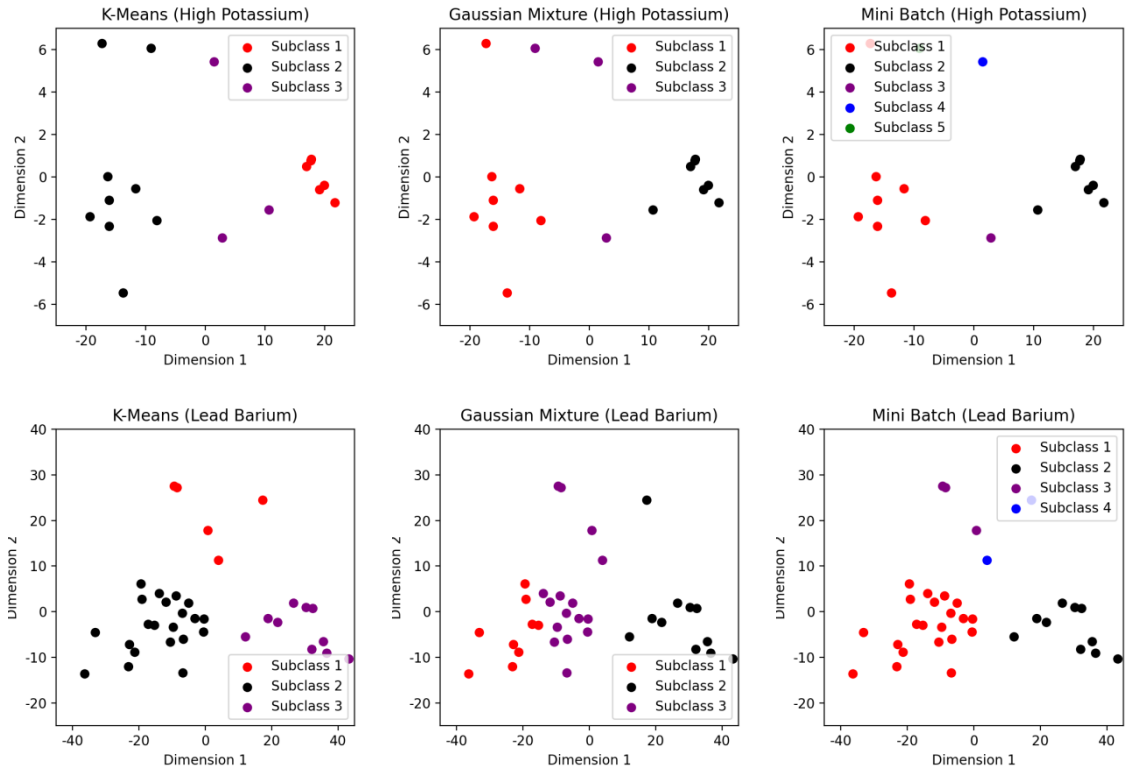

Figure 1. Visualization of clustering results

With the K-Means method and the Gaussian Mixture Method (GMM), all data points are concentrated more regularly in three groups, and there is some bias in the clustering results of the GMM method in the presentation of data for the high-potassium glass samples. Therefore, the data visualization results support the choice of the K-Means method and the good results of this method in clustering.

According to the K-Means method, the clustering of the 67 sets of base data was divided into 3 subclasses, and the subclasses were clearly distinguished from each other with significant characteristics.

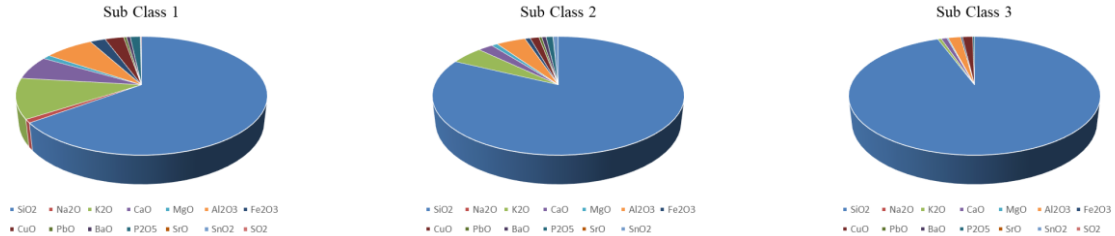

Figure 2. Classification of high potassium glasses under K-Means method

In order to further demonstrate the commonality of each subcategory and the individual differences between different subcategories, all data points within each subcategory were averaged to form the proportional characteristics of the subcategory's own components, and the results of different subcategories were compared, and the results are shown in Figure 2 below.

Among the three subclasses of high-potassium glasses, subclass 1 has a much higher content of non-silica (SiO_2) than the other two subclasses, and the content of potassium oxide (K_2O) in subclass 1 is much higher than that in subclass 1 and subclass 3.

The content of silicon dioxide (SiO_2) in subclass 3 is much higher than that in the other two subclasses; the content of silicon dioxide (SiO_2) in subclass 2 is between subclass 1 and subclass 3, and the content of potassium oxide (K_2O) and aluminum oxide (Al_2O_3) is closer.

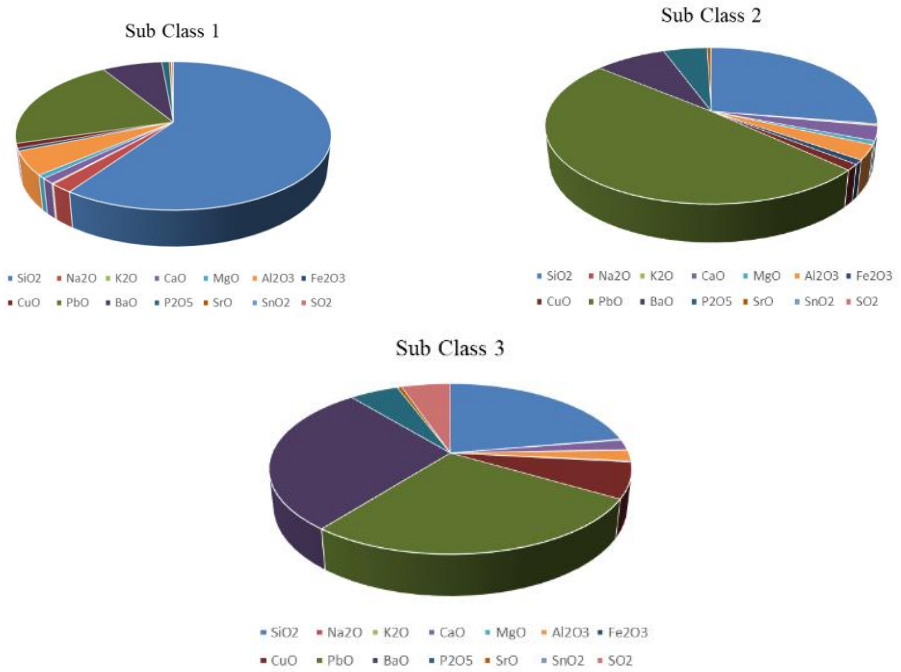

Figure 3. Classification of lead barium glass under K-Means method

As shown in Figure 3, among the three subclasses of lead-barium glass, the content of silica (SiO_2) in subclass 1 is much higher than the other two subclasses, but significantly lower than all subclasses of high-potassium glass; the content of lead oxide (PbO) in subclass 2 is much higher than any other subclasses, reaching nearly 50%; the content of lead oxide (PbO) and barium oxide (BaO) in subclass 3 are both higher, both exceeding 25%. Based on the above six subclasses, all the artifact samples were sequentially numbered into the subclasses to which they belonged, and the results are shown in Table 2 below.

Table 2. Subclassification results for glass samples

Glass Type	High Potassium Glass			Lead barium glass		
Subclass	1	2	3	1	2	3
Artifact number	1,3,4 5,6,13 14,16	7,9,10 12,22,27 12	18,21	1,19,30 34,36,38 39,40,41 43,49,50 51,52,54 56,57,58	23,25,28 29,31,32 33,35,37 42,44,45 46,47,48 53,55	8,11,20 24,26

3.3. Correlation analysis on the composition of different types of glass samples

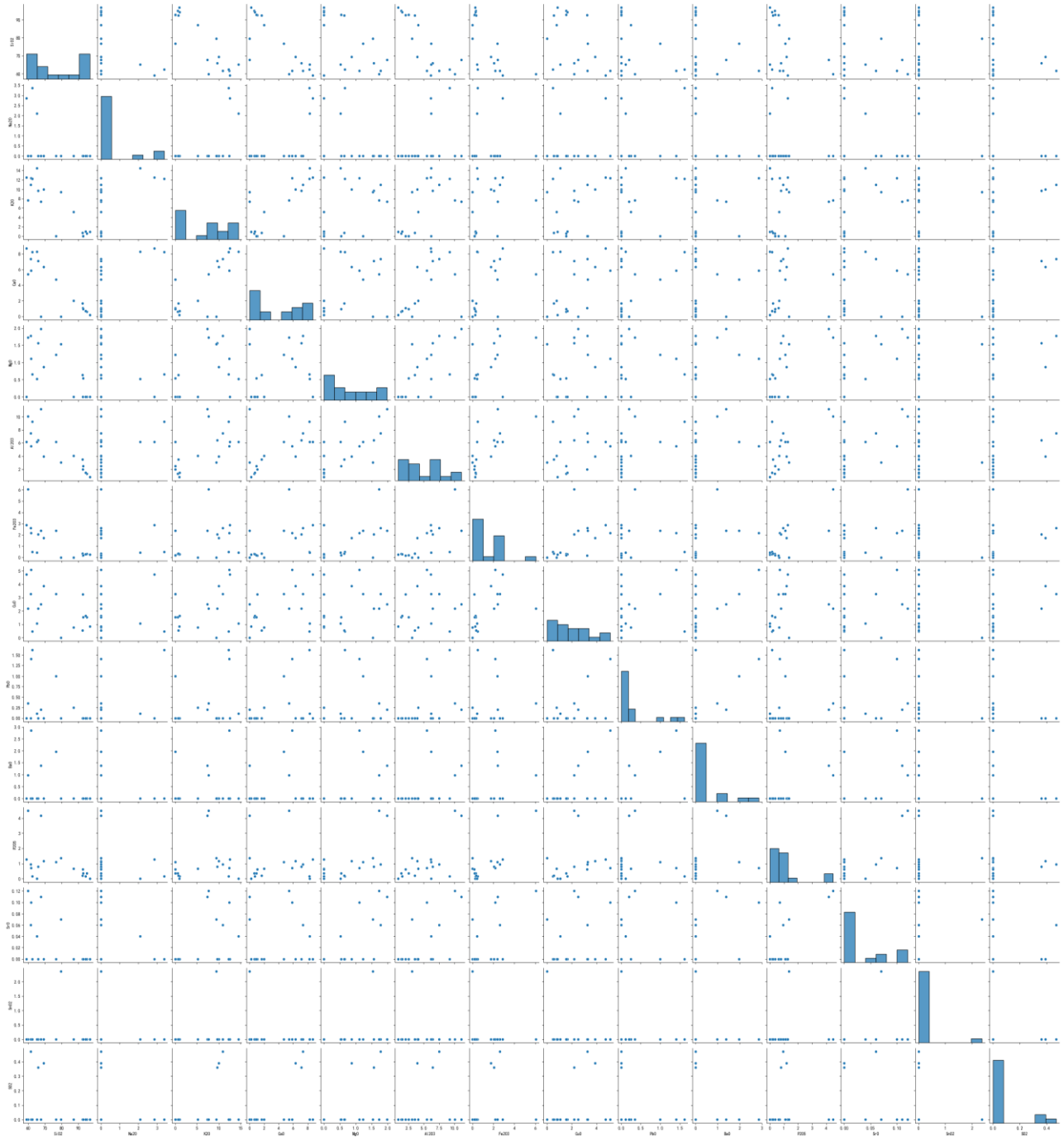

Figure 4. Scatter diagram of chemical composition of high potassium glass and lead-barium glass

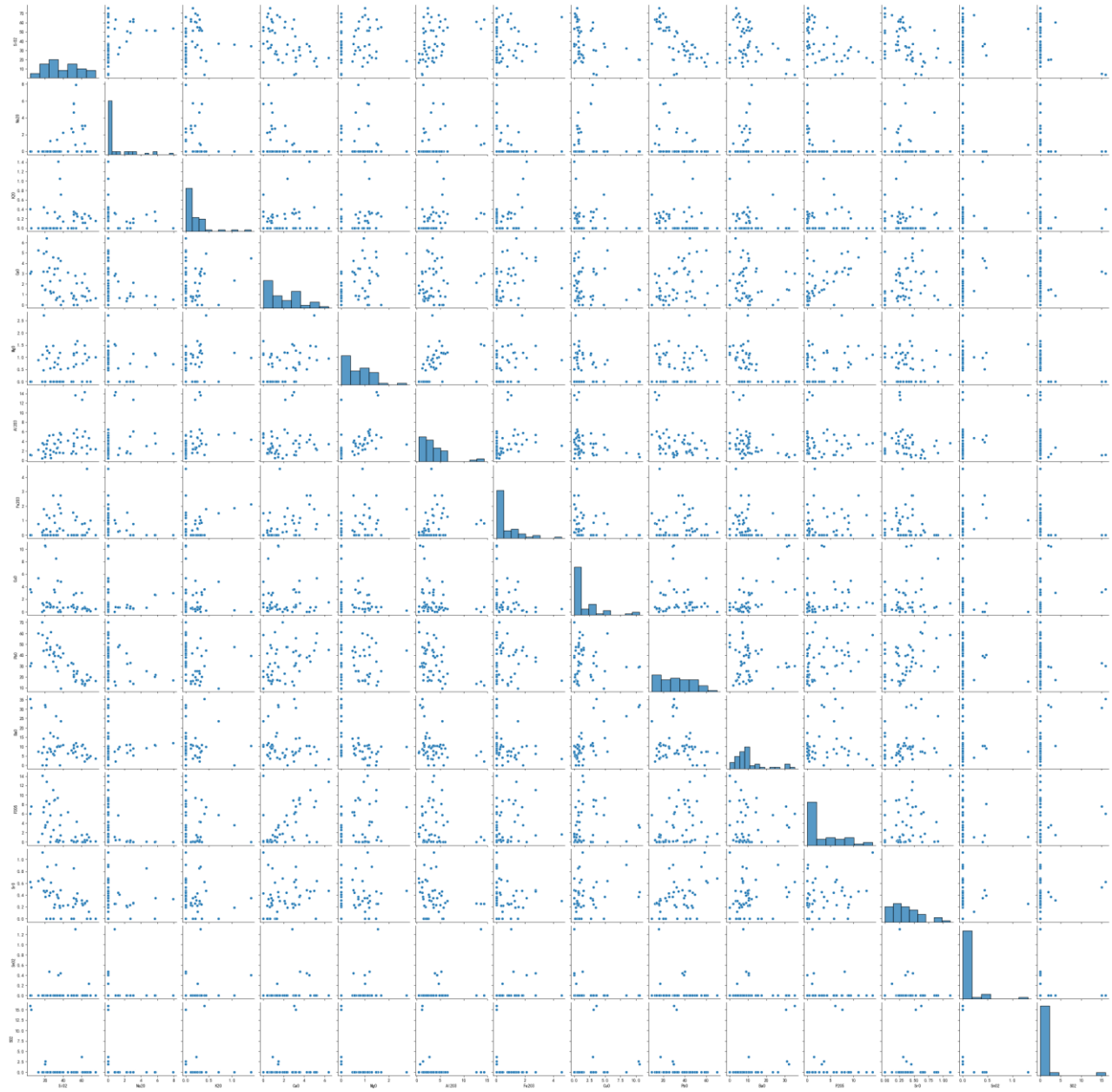

Figure 5. Scatter diagram of chemical composition of high potassium glass and lead-barium glass

For the intrinsic chemical compositions of different types of glass artifact samples, the data were firstly statistically classified and analyzed by using SPSS to obtain the values of the maximum value, mean, standard deviation, median, variance, kurtosis and skewness of the data for high potassium glass and lead-barium glass, respectively.

Based on the available data, a scatter plot was first drawn to determine the linear relationship between the variables, as shown in Figure 4 and Figure 5.

From the scatter plot, we can see that there is a linear relationship between the chemical components.

The results of the S-W test (Shapiro-Wilk test) for the normality of the data were obtained from the table of descriptive statistics of the data obtained according to the above steps [10].

The S-W test is applied to small samples (sample size ≤ 5000) and shows the normality of silica (SiO_2) and alumina (Al_2O_3) in the available data. (SiO_2) and alumina (Al_2O_3) satisfy the normal distribution. The rest of the chemical composition variables also approximately satisfy the normal distribution in the QQ plot, so the model assumptions are reasonable and the data can be further analyzed.

3.4. Analysis on the variability of chemical composition association of different categories of samples

Based on the correlations obtained in the above process, the variability between different chemical compositions was further analyzed.

(i) Subclassification of high-potassium glasses 1

1. the largest positive correlation of this classification is between iron oxide and phosphorus pentoxide, with a correlation coefficient of 0.78; the largest negative correlation is between potassium oxide and phosphorus pentoxide, with a correlation coefficient of -0.86.

2. silicon dioxide showed negative correlation with all other chemical components, while the rest of the chemical components showed positive and negative correlations.

Subcategory 2

1. the largest positive correlation in this classification is between calcium oxide and alumina, with a correlation coefficient of 0.98; the largest negative correlation is between silica and calcium oxide, with a correlation coefficient of -0.93.

2. there were some chemical components missing which made correlation analysis impossible.

3. all chemical components have positive and negative correlations with other components.

Subclassification 3

1. this classification has only perfect correlation and perfect non-correlation among the remaining chemical components, with correlation coefficients of only 1

and -1.

(ii) Lead barium glass sub-category 1

1. the largest positive correlation of this classification is between iron oxide and aluminum oxide, with a correlation coefficient of 0.69; the largest negative correlation is between iron oxide and lead oxide, with a correlation coefficient of -0.58.

2. sodium oxide shows negative correlation except for positive correlation with silica and barium oxide; all chemical components have positive and negative correlation with other components.

Subcategory 2

1. the largest positive correlation of this classification is between calcium oxide and aluminum oxide with a correlation coefficient of 0.67; the largest negative correlation is between silica and lead oxide with a correlation coefficient of -0.65.

2. strontium oxide shows positive correlation except for negative correlation with silica and tin oxide; all chemical components have positive and negative correlation with other components.

Subcategory 3

1. the largest positive correlation of this classification is between iron oxide and alumina with a correlation coefficient of 0.93; the largest negative correlation is between alumina and lead oxide with a correlation coefficient of -0.97.

2. there were some chemical components missing which prevented correlation analysis.

3. all chemical components have positive and negative correlations with other components.

4. Conclusion

There are different statistical classification laws among different glass categories, and at the same time, there are commonalities and individualities in chemical composition within the same category of samples, making several different subcategories within the same category of glass samples. By analyzing the chemical composition of different samples of the same category, a specific classification method can be established and the results of each subcategory can be obtained. In this paper, three different methods were used to cluster the samples of high potassium glass and lead-barium glass, and the clustering results were visualized by PCA principal component analysis. The mean and standard deviation of the chemical composition of each sample in the same subclass were calculated to further analyze their compositional characteristics and their own differences. The K-Means method can better classify the target data into subclasses with a cluster number of 3. Therefore,

it was determined that the method was used and the clustering results of the method and the effect analysis of the results were output.

There are some correlations between the chemical components contained in the different classes of glass artifact samples. The correlations were also different due to the differences in the proportional content of the chemical components of the samples. The SPSS software was used to analyze the relationships and the differences. Based on the data characteristics of the available data, descriptive statistics were performed on the total data set, and a number of statistics were obtained, including the maximum value, mean, variance, kurtosis and skewness. Then, the linear relationships in the data were further analyzed based on the plotted scatter plots. Meanwhile, the normality of each data variable was analyzed using the S-W test results obtained from the descriptive statistics of the data. Based on this, the correlations among the variables were obtained for the correlations among the different chemical components of the heritage samples using corrcoef and Pearson correlation coefficients, and the heat map was plotted.

References

[1] Takuya Iwanaga, William Usher, and Jonathan Herman. Toward SALib 2.0: Ad- vancing the accessibility and interpretability of global sensitivity analyses. Socio- Environmental Systems Modelling, 4:18155, May 2022.

[2] Jon Herman and Will Usher. SALib: An open-source python library for sensitivity analysis. The Journal of Open Source Software, 2(9), jan 2017.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blon- del, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[4] Organizing Committee of Gaojiao Society. 2022 National University Student Mathematical Modeling Competition. The National Mathematical Modeling Competition for University Students, 2022.

[5] Wang Jie; Li Mo; Ma Qinglin; Zhang Zhiguo; Zhang Meifang; Wang Julin;. Weathering of an octagonal lead-barium glass vessel from the Warring States period. Institute of Metallurgy and Materials History, University of Science and Technology Beijing; Beijing Key Laboratory of Electrochemical Processes and Technologies, Beijing University of Chemical Technology; Chinese Academy of Cultural Heritage, 2014.

[6] Liu, Song, Qinghui Li, and Fuxi Gan. "Chemical analyses of potash–lime silicate glass artifacts from the Warring States period in China." Spectroscopy Letters 48.4 (2015): 302-309.

[7] Bjelland, Tor Kristian. "Classification: assumptions and implications for conceptual modeling." (2005).

[8] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the royal statistical society. series c (applied statistics) 28.1 (1979): 100-108.

[9] Pham, Duc Truong, Stefan S. Dimov, and Chi D. Nguyen. "Selection of K in K-means clustering." Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 219.1 (2005): 103-119.

[10] Hanusz, Zofia, Joanna Tarasinska, and Wojciech Zielinski. "Shapiro–Wilk test with known mean." REVSTAT-Statistical Journal 14.1 (2016): 89-100.