

A study on the classification of ancient glassware based on unsupervised learning

Renjie Wu *

School of Mathematics and Statistics, Yunnan University, Yunnan, Kunming

* Corresponding Author Email: weilai@mail.ynu.edu.cn

Abstract. In the process of weathering, ancient glass objects have changed their composition ratio due to the large exchange of internal elements with external environmental elements, which affects the correct judgment of their categories. This paper establishes a relevant model for the classification and research of glass artifacts. The relationship between the surface weathering of these glass artifacts and the type, decoration, and color of the glass is analyzed; the statistical pattern of the content of chemical components with and without weathering on the surface of the artifacts is analyzed in conjunction with the type of glass, and the content of their chemical components before weathering is predicted based on the data of the weathering points. To study the classification pattern of high potassium glass and lead-barium glass; to select the appropriate chemical composition for each category for sub-category classification, to give specific division methods and division results, and to analyze the reasonableness and sensitivity of the classification results. Identify the categories to which unknown categories of glass artifacts belong, and analyze the sensitivity of the classification results. To analyze the correlations between the chemical components of glass artifact samples of different categories and compare the differences in the correlations between the chemical component contents of different categories.

Keywords: Ancient glass identification; composition analysis; mean prediction; weighted KNN.

1. Introduction

The Silk Road was a channel of cultural exchange between China and the West in ancient times, of which glass was a valuable witness to its early trade exchanges. Our native glass system is dominated by $PbO - SiO_2$, $PbO - BaO - SiO_2$, which is highly fragile and difficult to withstand more pronounced changes in heat and cold [1]. For this reason, the development of ancient indigenous glass was severely limited and could not be well integrated into the lives of ordinary people. The exotic technology that came with the Silk Road brought a qualitative change to the glass-making process in ancient China.

The main raw material of glass is quartz sand, whose main chemical composition is silicon dioxide. The main chemical composition of the glass made differs depending on the co-solvent added during calcination. It is now known that glass is divided into lead-barium glass and high-potassium glass because of the different content of the main chemical components, and it is generally believed that the former is the ancient native glass (high content of lead and barium), and the latter was introduced by the Silk Road and popular in the Lingnan region of China and Southeast Asia [2,3].

2. Methods

First, based on the relationship between weathering and glass type, decoration, and color on the surface of glass artifacts, we analyze the statistical pattern of the content of chemical composition with and without weathering on the surface of artifacts and predict the content of chemical composition before weathering at weathering sites. Columnar table analysis, through the columnar table and cross-plot, to study the frequency and percentage of different categories. Variance analysis and correlation analysis among the four variables were conducted. For the former, the weathering condition, glass type, decoration, and color were considered definite category data, and the chi-square test was used. For the four variables, a two-by-two comparison was performed, and based on a P

value less than 0.01 and combined with the Phi and other indices of the quantitative analysis of effects, it was concluded that there was a significant difference between surface weathering and type, color and ornamentation, color and type, and type and ornamentation; for the latter, considering the inapplicability of Pearson's correlation coefficient, we coded the fixed-class variables i.e., discrete data continuous. The latter, considering the inapplicability of Pearson's correlation coefficient, we coded the definite class variables, i.e., discrete data continuously, and performed Spearman's correlation test, concluding that weathering was significantly correlated with type at the 0.01 level (two-tailed) and the coefficient was significantly different from zero. Descriptive statistics, frequency histogram statistics, face plot analysis, normal distribution test, and mean t-tests between different samples were performed for four types of samples: high potassium weathered, high potassium unweathered, lead-barium weathered and lead-barium unweathered. For the three severely weathered sites, we screened samples of the same kind as them, i.e., the lead-barium C category, established the weights using gray correlation [4], calculated the mean values of different components before and after weathering, and used the weighted mean prediction method to derive the content of components before weathering at the three severely weathered sites.

The dummy variables of whether weathering and ornamentation were established. A binary logistics regression model was established with glass type as the dependent variable and 14 chemical composition contents, weathering, and ornamentation as independent variables. The independent variables were screened for significance based on the p-values. Based on the above classification by glass type, subclassifications were made for high potassium glass and lead-barium glass, respectively. First, we used K-means clustering [5, 6], a poor model based on the clustering evaluation rules of small within-group variance and large between-group variance, as well as classification accuracy. Considering the sensitivity of the model to initial values, we used systematic clustering and drew a line graph based on the aggregation coefficients, and combined with the elbow rule, we divided the data into 2-3 classes optimally. Finally, divided into four classes of samples, we calculated the mean and variance of different samples, compared the coefficient of variation, and selected chemical components with large coefficients of variation using box-line plots as the basis for classification. The content of different chemical components fluctuated up and down by 5%, and again the data were imported into SPSS, where we performed a sensitivity analysis of the systematic clustering model and gave the evaluation of the model.

Using machine learning with higher accuracy. After data preprocessing, we use samples from known categories as input and unknown categories that need to be identified as output. Importing the data into MATLAB, we invoked the classification toolbox, using ten-fold cross-validation, and invoking all machine learning models, we arrived at the weighted KNN model with the highest classification accuracy and the largest AUC value [7,8]. And based on this model we identified unknown categories of artifacts, including information on color, type, and ornamentation. The data fluctuated up and down by about 5%, and then the data were imported into the MATLAB machine learning toolbox to test the stability and sensitivity of the model, and the results of the weighted KNN model were stable.

The glass was first divided into high potassium glass and lead-barium glass by glass type, and the correlations between their different components were analyzed by matrix scatter plots as well as correlation coefficients. For their correlation coefficients, there existed negative and positive correlations between the components, and a p-value greater than 0.8 was a strong positive correlation. The closer to 1 indicates the more correlation. The two classified glass types were further subclassified according to whether they were weathered or not, i.e., weathered, into four categories: high potassium weathered, high potassium unweathered, lead-barium weathered, and lead-barium unweathered, further developing the correlation elaboration of the different chemical components of the four categories [9,10].

3. Results

3.1. Glass Weathering

For ornamentation, ornamentation A and C have a balanced sample distribution on weathering, while ornamentation B is all weathered; for glass type, high potassium is more resistant to weathering than lead barium, considering the higher potassium content of high potassium glass versus the high lead barium composition of lead barium glass, we guess that the potassium composition is more resistant to weathering than lead barium; for color, the black category is all weathered (excluding purple), indicating that the more severe weathering may be, the darker the color.

From the results of the chi-square test, we can conclude that: surface weathering and ornamentation, significant P-value is 0.084*, the level does not present significance, accept the original hypothesis, so there is no significant difference; surface weathering and type, significant P-value is 0.009***, the level presents significance, reject the original hypothesis, so there is a significant difference; surface weathering and color, significant P-value is 0.307, the level does not present significance, accept the original hypothesis, so there is a significant difference. level does not present significance, accepting the original hypothesis, so there is no significant difference. The effect quantification analysis was performed on its basis.

From the results of the weathering-effects quantification analysis, it can be concluded that: the value of ornamentation is 0.292, the degree of difference between ornamentation and surface weathering is moderate; the value of type is 0.344, and the degree of difference between type and surface weathering is moderate; the value of color is 0.403, so the degree of difference between color and surface weathering is moderate.

From the *phi* coefficient, it can be concluded that: the *phi* value of ornamentation and weathering is less than 0.3, which indicates a weak correlation; the values of type, color, and weathering are between 0.3 and 0.6, which indicates a moderate correlation.

From the coefficient, it can be concluded that it is not feasible to predict weathering using only ornamentation, type, and color.

For the four definite classes of data, it does not apply to the Pearson correlation test, so we coded the discrete data, such as whether weathering was coded 0-1. In this way, we imported the processed data into SPSS with MATLAB and performed the Spearman correlation test. The specific Spearman coefficients and Spearman correlation coefficients to test whether they are different from zero are shown in Figure 1.

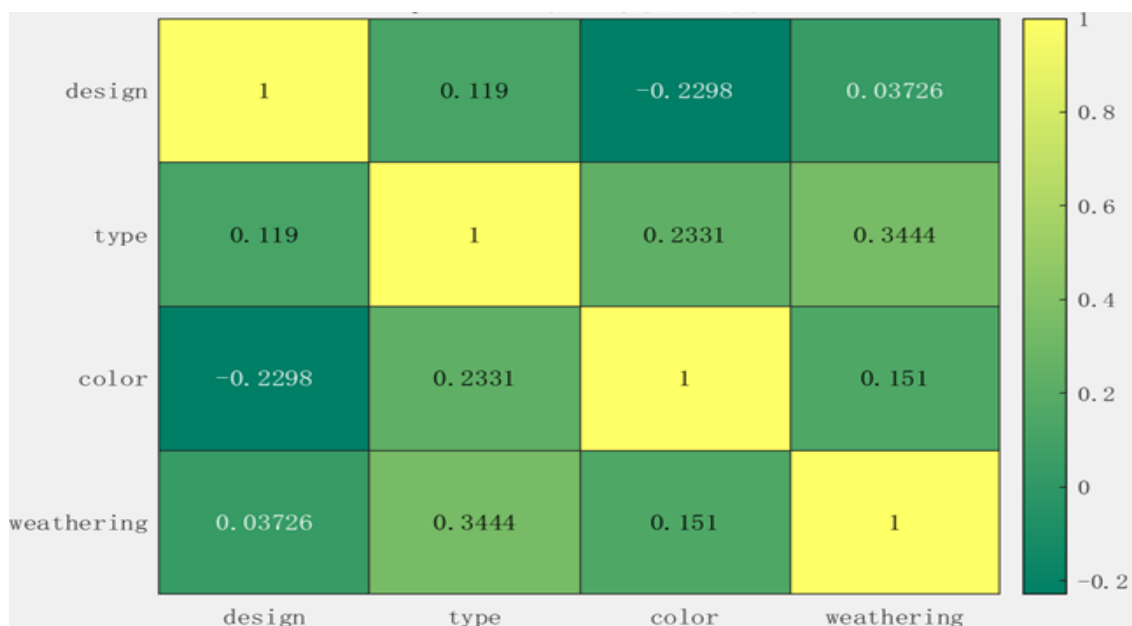


Figure 1. Spearman's correlation coefficient.

The correlation coefficient between color and ornamentation is 0.2298, which shows that color is negatively correlated with ornamentation; the correlation coefficient between type and weathering is 0.3444, which shows that type is strongly correlated with weathering. From the results of Spearman's coefficient different from the zero tests, it can be concluded that: color is negatively correlated with ornamentation, and the correlation coefficient is significantly different from zero; type is most strongly correlated with weathering, and the correlation coefficient is significantly different from zero.

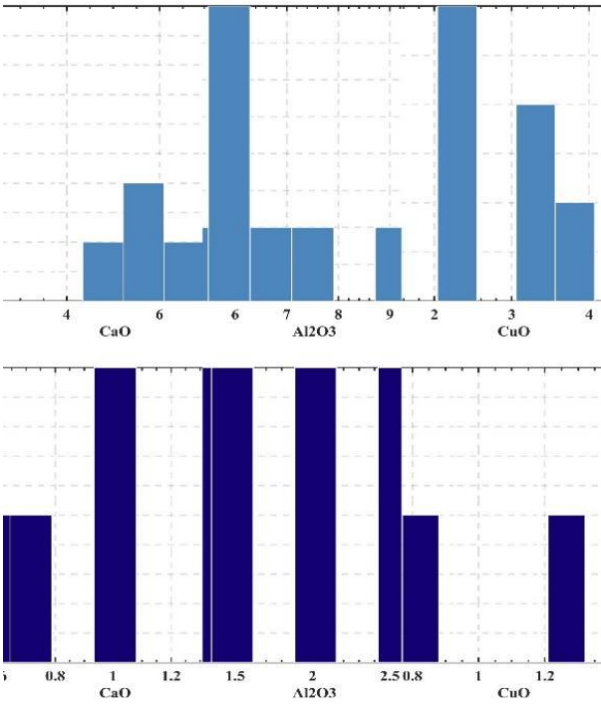


Figure 2. Comparison before and after high potassium weathering.

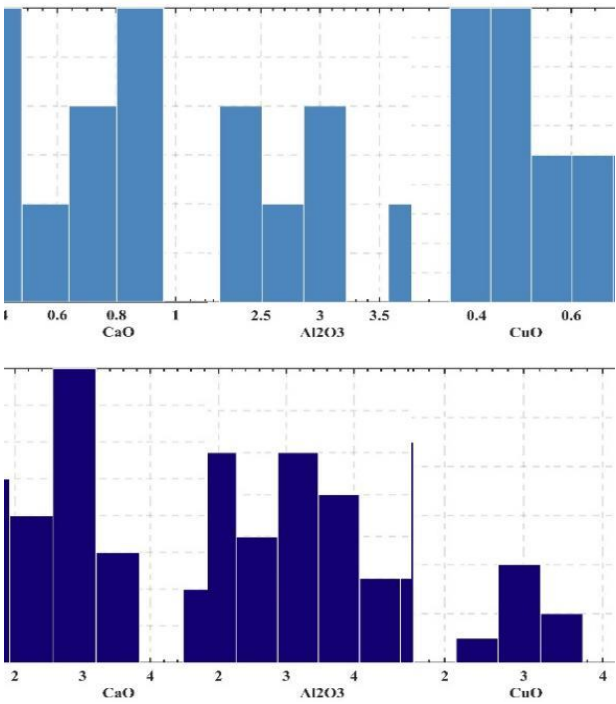


Figure 3. Comparison of lead and barium before and after weathering.

From Fig. 2 Fig. 3, it can be seen that the content of several major chemical components increased after weathering of high potassium and decreased after weathering of lead barium. Specifically, for x , the content of x decreases after weathering of high-potassium glass and increases after weathering of lead-barium glass; for y , the content of y decreases after weathering of high-potassium glass and increases after weathering of lead-barium glass. The different trends of increase and decrease in the content of the main chemical components of high potassium glass and lead-barium glass before and after weathering may be caused by their different compositions.

About comparing the differences between the different categories, we first performed a normal test for the main components, followed by a t-test for the contents of those satisfying the normal distribution. Considering the small sample using the Shapiro test, we tested whether the four components meet the normal distribution as well as the t-test in the R environment.

Using MATLAB software, normalization was performed after inputting the data of the corresponding parent and child series. The results can be obtained as shown in Table 1.

Table 1. Normalization of correlation analysis results.

Relevance	0.0688	0.0692	0.0695	...	0.0727	0.0727	0.0732	0.0742
Relevance Ranking	a_2	a_{14}	a_3	...	a_6	a_4	a_{12}	a_9

Based on the similarity of the correlation between the 14 chemical components and whether they are weathered, we consider the weights to be the same for simplicity. For the 3 severely weathered points, i.e., 08 severely weathered, 26 severely weathered, and 54 severely weathered, we used the weighted mean prediction method, where the weights are the same for each component. The predicted values regarding the 3 numbers are shown in Table 2.

Table 2. Pre-weathering chemical content of the three severely weathered predictions.

Chemical composition	Interval 8	Mean value 8	Interval 26	Mean value 26	Interval 54	Mean value 54
SiO ₂	[35.37,44.92]	40.14	[34.48,44.03]	39.25	[47.87,57.42]	52.64
Na ₂ O	[-0.26,0.39]	0.06	[-0.26,0.39]	0.06	[-0.26,0.39]	0.06
K ₂ O	[-0.08,0.06]	-0.01	[0.32,0.46]	0.39	[-0.08,0.06]	-0.01
CaO	[0.69,2.29]	1.49	[0.51,2.11]	1.31	[-2.5,-0.9]	-1.7
MgO	[-0.58,0.06]	-0.26	[-0.58,0.06]	-0.26	[0.53,1.17]	0.85
Al ₂ O ₃	[0.47,1.59]	1.03	[0.54,1.66]	1.1	[3.01,4.13]	3.57
Fe ₂ O ₃	[0.14,0.65]	0.39	[0.14,0.65]	0.39	[0.14,0.65]	0.39
CuO	[1.24,2.52]	1.88	[1.7,2.98]	2.34	[-0.56,0.72]	0.08
PbO	[4.87,15.44]	10.16	[2.34,12.91]	7.63	[30.88,41.45]	36.17
BaO	[22.13,31.32]	26.73	[26.96,36.15]	31.56	[-8.49,0.7]	-3.89
P ₂ O ₅	[1.41,5.25]	3.33	[-0.11,3.73]	1.81	[7.98,11.82]	9.9
Sr	[0.2,0.45]	0.32	[0.29,0.54]	0.41	[0.79,1.04]	0.91
SnO ₂	[-0.07,0.02]	-0.02	[-0.07,0.02]	-0.02	[-0.07,0.02]	-0.02
SO ₂	[11.92,15.95]	13.94	[12.84,16.87]	14.86	[-3.11,0.92]	-1.09
Total content	[97.89,100.48]	99.19	[99.54,102.13]	100.84	[96.57,99.16]	97.87

From Table 2 it can be seen that for the prediction of three severely weathered points, some negative values can be found, and considering that the treatment of missing values is treated as zero, we consider that the content of components with negative values is zero before weathering.

3.2. Glass classification law

Table 3. Logistic regression significant factors.

Independent variable	Score	Degree of freedom	Significance	Independent variable	Score	Degree of freedom	Significance
SiO ₂	27.772	1	0	B	14.71	1	0
K ₂ O	31.706	1	0		56.494	17	0
CaO	12.256	1	0	Al ₂ O ₃	9.447	1	0.002
PbO	37.265	1	0	weathering	8.589	1	0.003
BaO	28.908	1	0	CuO	5.792	1	0.016
Sr	19.927	1	0	Fe ₂ O ₃	4.352	1	0.037

From the results in Table 3, it can be concluded that has a significant effect on whether weathering is present. Considering that the composition of lead-barium glass is dominated by and the composition of high potassium glass is dominated by K_2O , it is possible to discriminate between high potassium glass and lead-barium glass based on the above chemical composition content.

We clustered the samples into two classes and calculated the classification accuracy and F1 score under this algorithm. About the variability of different fields under the K-means clustering algorithm and the ANOVA table is shown in Table 4.

Table 4. K-means clustering ANOVA table.

Ingredients	Clustering mean square	Degree of freedom	Error mean square	Degree of freedom	F	Significance
SiO ₂	20864.487	1	148.246	59	140.742	0
Na ₂ O	13.145	1	2.854	59	4.607	0.036
K ₂ O	166.376	1	14.194	59	11.722	0.001
CaO	2.412	1	6.903	59	0.349	0.557
MgO	1.014	1	0.389	59	2.609	0.112
Al ₂ O ₃	38.06	1	4.588	59	8.295	0.006
Fe ₂ O ₃	0.601	1	1.443	59	0.416	0.521
CuO	0.436	1	1.914	59	0.228	0.635
PbO	17321.834	1	97.545	59	177.578	0
BaO	464.679	1	22.429	59	20.718	0
P ₂ O ₅	170.875	1	7.994	59	21.376	0
Sr	0.722	1	0.043	59	16.823	0
SnO ₂	0.046	1	0.126	59	0.367	0.547
SO ₂	0.271	1	0.224	59	1.208	0.276

The results of the systematic clustering analysis showed that there were 19 items predicted to be high potassium glass when it was true; 22 items predicted to be lead-barium glass when it was true, and 20 items predicted to be wrongly classified as high potassium glass when it was true.

We calculate the classification accuracy and score as 67% and 0.66 respectively, and the process is as follows.

$$P = \frac{19 + 22}{39 + 22} = 0.672131$$

$$F_1 = \frac{2 * 19}{2 * 19 + 0 + 20} = 0.655172$$
(1)

Combining the classification accuracy of the above K-means clustering model and hierarchical clustering model, the accuracy of both models in discriminating high potassium glass from lead-barium glass is about 70%, which is better.

Table 5. Subclassification results.

Lead barium1	39,40,43 Part 1
Lead barium2	2,11,19,24,30 Part 12,34,36,38,41,43 Part 2,49,50,51 Part 12,52,56,57,58
High Potassium 1	7,9,10,12,18,21,,22,27
High Potassium 2	1,03 Part 2,4,5,06 Part 12,13,14,15,16,17...

The dispersion of the data from the whole was examined using box-line plots, and the chemical components with larger standard deviations or larger coefficients of variation were selected as the main classification components. According to the formula

$$CV = \frac{std}{mean} \quad (2)$$

For lead-barium 1 and 2, component x was selected as the basis of classification; for high potassium 1 and 2, component y was selected as the basis of classification.

To test the clustering model we established, firstly we made the content of 14 chemical components fluctuate up and down by 5%, and secondly, we imported the data into the model, we got better stability of the dichotomous model of lead-barium and high potassium, and the specific differences between the subclassification models of lead-barium and high potassium are shown in Table 6.

Table 6. Sensitivity analysis of classification model.

Lead barium1*	36,38,52
Lead barium2*	2,11,19,24,30 Part 12,34,39,40,41,43 Part 12,49,50,51 Part 12,56,57,58
High Potassium 1*	7,9,10,12,18,21,,22,27
High Potassium 2*	1,03 Part 2,4,5,06 Part 12,13,14,15,16,17...

After 5% perturbation from the data in Table 6, the high potassium subclassification model is more stable and the lead-barium subclassification model is more unbalanced. However, we found that Pb-Ba 36,38,52 and 39,40,43 site 1 are Pb-Ba weathered C. Considering the existence of uncertainty factors and inappropriate treatment of missing value outliers, we consider the model more stable.

3.3. Sensitivity analysis

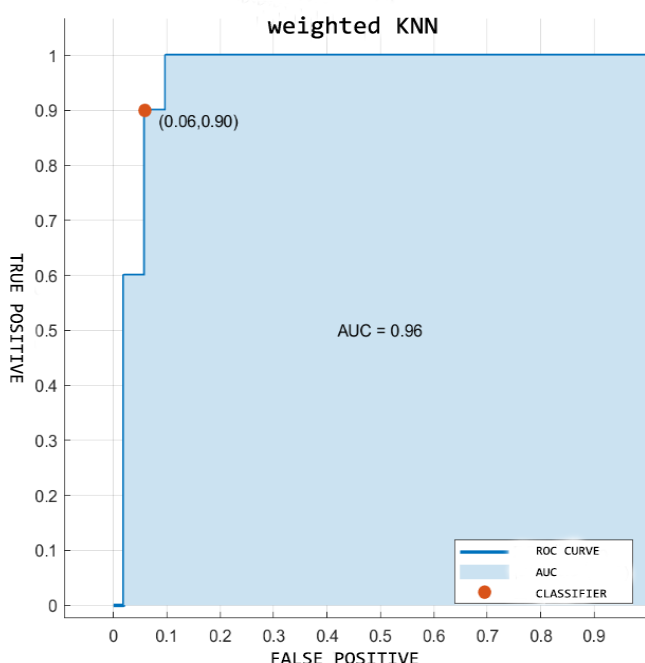


Figure 4. Fluctuation-free AUC.

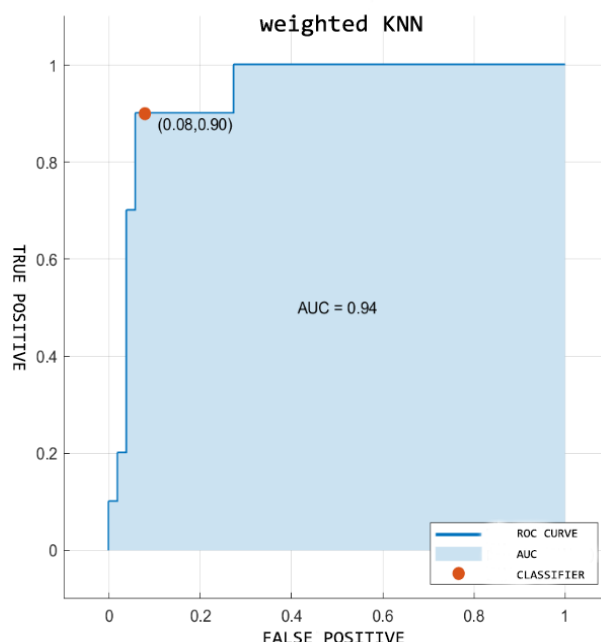


Figure 5. Fluctuating AUC.

From the results in the above figure, it can be concluded that the AUC before and after data perturbation remains around 0.95, so the weighted KNN has a good prediction effect and stability.

Our 14 chemical components fluctuate up and down by 5%, and the resulting classification results are still considerable. This indicates that the stability of our established weighted KNN model is good and the sensitivity to the input variable data is low, i.e., a small change in the content of the

independent variable chemical composition will not make the dependent variable discriminate the results change too much.

The samples were divided into two categories of high potassium glass and lead-barium glass by glass type and sub-classified by weathering on their basis, and the chemical compositions with strong linear correlation were identified by matrix scatter plots. The details are shown in Table 7.

Table 7. Correlation coefficients for different combinations of chemical components.

Category	Combination and correlation coefficient	
High Potassium	$SiO_2 - CaO$	$MgO - Al_2O_3$
	-0.84	0.75
Lead Barium	$SiO_2 - PbO$	$CaO - P_2O_5$
	-0.88	0.8
High potassium unweathered	$SiO_2 - CaO$	$Fe_2O_3 - P_2O_5$
	-0.68	0.71
High potassium weathering	$SiO_2 - CaO$	$CaO - Al_2O_3$
	-0.89	0.96
Lead barium unweathered	$SiO_2 - PbO$	$MgO - Al_2O_3$
	-0.87	0.75
Lead barium weathering	$SiO_2 - PbO$	$CaO - P_2O_5$
	-0.88	0.76

For lead-barium glass, $SiO_2 - PbO$ shows a highly negative correlation; for high potassium glass, $SiO_2 - CaO$ is highly negatively correlated and highly positively correlated. The main component of quartz sand is SiO_2 , so the content in lead-barium and high-potassium glass is not too high; high-potassium glass can be subclassified according to the component, and also according to the content.

For high potassium glass, the correlation coefficient after weathering is larger in absolute value and shows a negative correlation, which shows that the relative loss after weathering is more; for lead-barium glass, the correlation coefficient before and after weathering is unchanged, which shows that the relative trend of content after weathering does not change much.

4. Conclusions

This paper establishes a relevant model for classifying and researching glass artifacts. The relationship between the surface weathering of glass artifacts and glass is analyzed; the statistical law of the content of chemical composition with and without weathering on the surface of artifacts is analyzed, and the content of its chemical composition before weathering is predicted based on the data of weathering points. Study the classification pattern of high potassium glass and lead-barium glass; select appropriate chemical composition for each category for sub-category classification, give specific division methods and division results, and analyze the reasonableness and sensitivity of the classification results. Identify the categories to which unknown categories of glass artifacts belong, and analyze the sensitivity of the classification results. To analyze the correlations between the chemical components of glass artifact samples of different categories and compare the differences in the correlations between the chemical component contents of different categories.

References

- [1] Chen Shuyu, Hou Zhili. An analysis of the ancient Silk Road and ancient Chinese glass [J]. China Ethnic Expo, 2019, (05): 87-88.
- [2] Gong Yanlu, Lv Jia. A collaborative training method combining semi-supervised clustering and weighted KNN [J]. Computer Engineering and Applications, 2019, 55(22): 114-118.

- [3] Wang Jie, Li Mo, Ma Qinglin, Zhang Zhiguo, Zhang Meifang, Wang Julin. A weathering study of an octagonal lead-barium glass vessel from the Warring States period [J]. *Glass and Enamel*, 2014, 42(02): 6-13.
- [4] Shoukui Si, Xijing Sun. *Mathematical modeling and algorithms* [M]. Beijing: National Defense Industry Press, 2011: 358-360.
- [5] Wang Chengguang. User portraits based on optimized K-means clustering algorithm [J]. *Science and Technology Innovation and Application*, 2022, 12(18): 18-21.
- [6] Nie Zhen. A library patron segmentation model based on systematic clustering [J]. *Modern Intelligence*, 2009, 29(09): 158-161.
- [7] Yan, Xiao-Ming. A weighted KNN classification algorithm based on category-averaged distance [J]. *Computer System Applications*, 2014, 23(02): 128-132.
- [8] Liu Wanying. Application of feature-weighted KNN classification algorithm in cross-border e-commerce talent training [J]. *Microcomputer Applications*, 2020, 36(09): 44-46.
- [9] Feng Bailing, Cao Shiyuan. Rumination on Chinese Warring Han Dynasty lead-barium glassware [J]. *Luoyang Archaeology*, 2021, (01): 47-54.
- [10] Liu Song, Lv Liangbo, Li Qinghui, Xiong Zhaoming. Glass beadwork excavated from Han tombs in Lingnan and Sino-foreign exchanges along the Han Dynasty Maritime Silk Road [J]. *Heritage Conservation and Archaeological Science*, 2019, 31(04): 18-29.