Component Analysis and Identification of Glass Relics based on the Random Forest Model

Yingkun Zhang *, Yuliang Chen, Zinuo Zhang, Changyu Liu, Zhipeng Yu

Shijiazhuang Tiedao University, Shijiazhuang, China

* Corresponding Author Email: 1540838594@qq.com

Abstract. In the process of weathering, a large amount of internal elements and environmental elements were exchanged, resulting in changes in the composition proportion of ancient glass, which would affect the determination of its category. In this paper, the composition analysis and category identification of ancient glass were carried out by statistical analysis of the attached data and random forest classification model.

Keywords: Random Forest Classification Model; Glass of Cultural Relics; Composition Analysis.

1. Restatement of the Problem

The Silk Road played an important role in promoting the development of glass products in ancient China. Because fluxes were needed to reduce the melting temperature in the refining of glass, different fluxes had different chemical compositions. In the process of glass weathering, a large amount of exchange between interior elements and environmental elements leads to changes in the composition proportion of glass products. Due to the different buried environment, the degree of weathering of unearthed glass products is also different, which has a great influence on the judgment of categories. There are a number of relevant data of ancient glass products in China, and archeologists have divided them into two types according to their chemical composition and other inspection methods. Annex Form 1 gives the classification information of these cultural relics, and Annex form 2 gives the proportion of the corresponding main components.

The problems to be solved in this paper are as follows:

Question 1: Analyze the relationship between surface weathering of glass cultural relics and glass types, ornamentations and colors, and analyze the statistical rule of weathering chemical composition content on the surface of cultural relics, and finally predict the chemical composition content before weathering based on weathering point detection data.

Question 2: Analyze the classification rules of high potassium glass and lead-barium glass; For each category, the appropriate chemical fractions were selected to divide the subclasses, and the specific classification methods and results were given, and the rationality and sensitivity of the classification results were analyzed.

Question 3: For different types of glass cultural relics, analyze the correlation between their chemical components, and compare the differences of the correlation between the chemical components of different types.

2. Establishment and Solution of Model

2.1 Establishment and Solution of the Model of Problem 2

2.1.1 Data Preprocessing

Through preliminary analysis of the attachment 1 and the attachment 2, we found that four cultural relics samples numbered 19, 40, 48 and 58 in Form I were missing color information, and the data of two cultural relics sampling points numbered 15 and 17 in the attachment 2 were invalid.

(1) Processing of missing values

First find out the missing color information of cultural relics samples, other numbered samples with the same characteristics are screened from the attachment 1(including glass type, decoration, whether wind):

Table 1. Missing value treatment group 1

	serial number	grain	type	color	surface weathering
Samples to be processed	19	A	lead barium	?	weathering
	48	A	lead barium	?	weathering
A sample with the same characteristics	23	A	lead barium	blue-green	weathering
	28	A	lead barium	light blue	weathering
	29	A	lead barium	light blue	weathering
	42	A	lead barium	light blue	weathering
	44	A	lead barium	light blue	weathering
	49	A	lead barium	black	weathering
	50	A	lead barium	black	weathering
	53	A	lead barium	light blue	weathering

Table 2. Missing value treatment group 2

	serial number	grain	type	color	surface weathering
Samples to be processed	40	A	lead barium	?	weathering
	58	A	lead barium	?	weathering
A sample with the same characteristics	38	A	lead barium	dark green	weathering
	39	A	lead barium	dark green	weathering
	41	A	lead barium	light green	weathering
	43	A	lead barium	light blue	weathering
	51	A	lead barium	light blue	weathering
	52	A	lead barium	light blue	weathering
	54	A	lead barium	light blue	weathering
	56	A	lead barium	blue-green	weathering
	57	A	lead barium	blue-green	weathering

According to the analysis of the above two tables and considering the need to ensure the integrity of data, we choose the color that appears most frequently in each sample with the same feature as the missing value as the replacement value. Excel was used to make the frequency bar chart as follows:

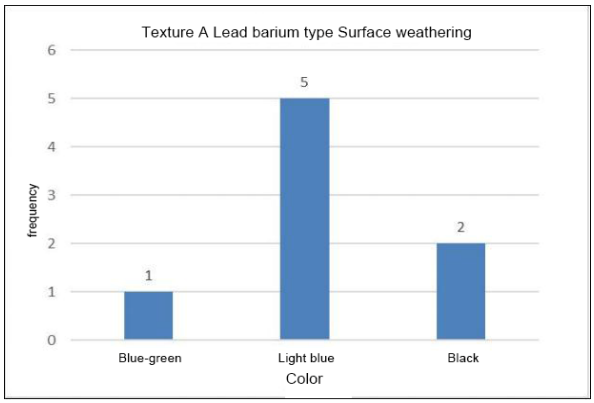


Fig 1. Color frequency diagram (a)

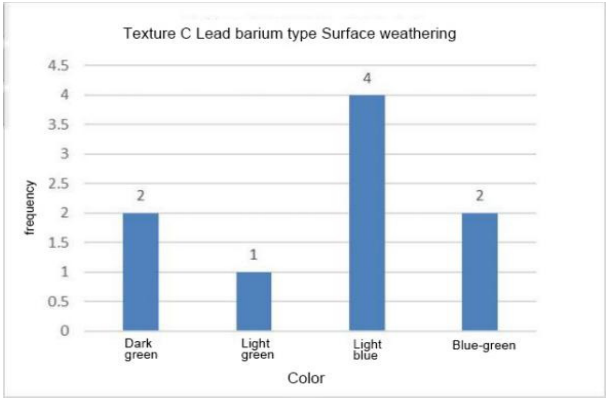


Fig 2. Color frequency diagram (b)

According to the above analysis results, the color with the highest frequency in other samples with the same characteristics as the missing sample is selected as the replacement value of the missing value. The replacement results are as follows:

Table 3. Supplementary results of missing values

serial number	grain	type	color	surface weathering
19	A	lead barium	light blue	weathering
48	A	lead barium	light blue	weathering
40	C	lead barium	light blue	weathering
58	C	lead barium	light blue	weathering

(2) Eliminate invalid samples

After accumulating the composition proportions of the cultural relics samples in Forms II and III, observe whether the sum is between 85% and 105%, delete the samples that are not within the range of valid data, and use Excel to process the results as follows:

Table 4. Summation of sample components

Heritage sampling points	Sum of components
15	79.47
17	71.89

Therefore, the data of cultural relics samples 15 and 17 are deleted.

2.1.2 Model Establishment

(1) Analysis of correlation

In order to explore the relationship between the surface weathering of glass relics and their glass types, patterns and colors, we first made a simple analysis of the data in Table 1, and found that the surface weathering and their glass types, patterns and colors are categorical variables. For the correlation analysis between categorical variables and categorical variables, we used chi square test to make a specific analysis.

The basic formula of [1] for chi-square test is:

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

In the above equation, A is the actual number, and T is the theoretical number inferred from hypothesis testing.

According to SPSS software, the surface weathering and glass type, ornamentation and color were respectively tested by chi-square test, and the cross table and test results were as follows:

Based on the chi-square test (cross analysis) results in the above table, the following conclusions can be drawn: 1. In the Chi-square test of surface weathering and glass type, the significance value is less than 0.05, and the null hypothesis cannot be rejected. There is a significant difference between surface weathering and glass type, that is, surface weathering and glass type are significantly correlated; 2. In the Chi-square test of surface weathering and ornamentation, the significance value is not less than 0.05, so the null hypothesis is rejected. There is a significant difference between surface weathering and ornamentation, that is, there is no obvious correlation between surface weathering and ornamentation; 3. In the chi-square test of surface weathering and color, the significance value is not less than 0.05, so the null hypothesis is rejected. There is a significant difference between surface weathering and color, that is, there is no obvious correlation between surface weathering and color.

In conclusion, the surface weathering is correlated with the type of glass, and the surface weathering of glass articles containing lead and barium is more likely to be weathered than that of

glass articles containing high potassium. The analysis results show that the surface weathering has no obvious correlation with ornamentations and colors.

Table 5. Chi-square test results

Feature	Division	Surface weathering		Total	χ^2	P
		Weathering	No weathering			
Type	Hyperkalemia	6	12	18	7.658	0. 009***
		28	12	40		
Total		34	24	58		
Grain	A	11	11	22	5.214	0. 071*
	B	6	0	6		
	C	17	13	30		
Total		34	24	58		
Colour	blue-green	8	8	16	8.691	0. 457
	light blue	11	4	15		
	purple	2	0	2		
	dark green	3	4	7		
	dark blue	3	4	7		
	black	0	2	2		
	light green	6	2	8		
	green	1	0	1		
Total		34	24	58		

(2) Analysis of statistical laws

Glass types based on analysis of cultural relics of the sample surface had weathered statistical regularity of chemical component content, the first thing you need to glass type and the presence of weathering class variables in combination group, the two groups according to the result of grouping and chemical composition analysis, the combination group is divided into four groups: high potassium + weathering, high potassium + no weathering, lead barium + weathering, lead barium + no weathering.

The data in the attachment 2 are screened by the above grouping method. Since the severely weathered points are from the regolith and do not belong to the sample surface, they should be screened out here. The specific screening results are as follows.

Table 6. Screening of regolith data

screen out requirements	data sample point
samples taken from the regolith	08 severe weathering point
	26 severe weathering point
	54 severe weathering point

After sifting out the above weathering data, the chemical composition content of each sample in the combination group was statistically analyzed, the average value was calculated, and the relative proportion of chemical composition before and after weathering of cultural relics under different glass types was compared.

In order to facilitate the analysis of the relative change relationship between the unweathered and weathered chemical compositions, the relative increment (after weathering versus before weathering) is defined as follows:

$$\alpha_{ij} = \frac{x_{ij}^{(1)} - x_{ij}^{(0)}}{x_{ij}^{(0)}} \times 100\%$$

In the above formula, α represents the average relative increment of the composition j after weathering of the glass of type i, $x_{ij}^{(0)}$ represents the average content of the composition j before weathering of the glass of type i, $x_{ij}^{(1)}$ represents the average content of the composition j after weathering of the glass of type i, $i = 1,2$ represent high potassium glass and lead barium glass, $j = 1,2...,14$ represent 14 chemical components of silicon dioxide, sodium oxide and sulfur dioxide respectively etc.

The average contents of the four types of chemical components are statistically calculated by Excel, and the average relative incremental changes of the chemical components before and after weathering of high potassium glass and lead barium glass are calculated. The results of the above two treatments are summarized as follows:

Table 7. Comparison of average chemical composition of high potassium glass before and after weathering

Chemical composition	Unweathered	After weathering	The relative increment	Chemical composition	Unweathered	After weathering	The relative increment
SiO2	67.984	93.963	38.21%	CuO	2.453	1.562	-36.32%
Na2O	0.695	0.000	-100.00%	PbO	0.412	0.000	-100.00%
K2O	9.331	0.543	-94.18%	BaO	0.598	0.000	-100.00%
CaO	5.333	0.870	-83.68%	P2O5	1.403	0.280	-80.04%
MgO	1.079	0.197	-81.78%	SrO	0.042	0.000	-100.00%
Al2O3	6.620	1.930	-70.85%	SnO2	0.197	0.000	-100.00%
Fe2O3	1.932	0.265	-86.28%	SO2	0.102	0.000	-100.00%

Table 8. Comparison of average chemical composition content before and after weathering of lead-barium glass

Chemical composition	Unweathered	After weathering	The relative increment	Chemical composition	Unweathered	After weathering	The relative increment
SiO2	53.444	35.900	-32.83%	CuO	1.557	1.933	24.14%
Na2O	0.772	1.039	34.68%	PbO	23.594	36.562	54.97%
K2O	0.258	0.144	-44.43%	BaO	10.499	9.438	-10.10%
CaO	1.232	2.371	92.52%	P2O5	0.904	3.693	308.56%
MgO	0.492	0.731	48.47%	SrO	0.297	0.331	11.45%
Al2O3	3.195	4.007	25.44%	SnO2	0.065	0.061	-5.74%
Fe2O3	0.933	0.606	-35.01%	SO2	0.282	0.138	-51.13%

The percentage accumulation histogram of each chemical component in high potassium glass and lead barium glass before and after weathering is made by Excel, the relative changes of each component of the two types of glass before and after weathering can be seen intuitively, as shown in the following figure:

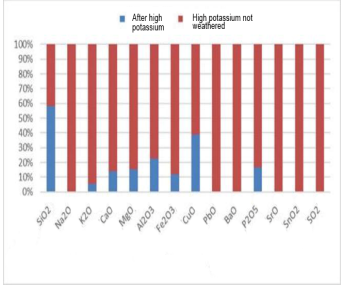


Fig 3. Relative proportion of componentsof high potassium glass before and after weathering

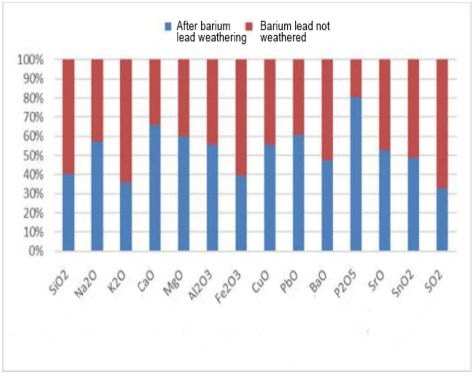


Fig 4. Relative proportion of components before and after weathering of lead barium glass

Based on the above two tables and figures, it can be intuitively concluded that the statistical rules of weathering chemical composition content on the surface of cultural relics samples are as follows:

① After weathering of high-potassium glass [5], the average content of SiO2 will be reduced, and the average content of other chemical components will be significantly decreased, most of the decrease range in 70~ 100%. The contents of Na2O, PbO, BaO, SrO, SnO2 and SO2 in weathered samples are all zero.

②After weathering, the content of a considerable part of the chemical components of lead-barium glass will increase, for example, the content of phosphorus pentoxide will increase greatly after weathering, the increase reaches 308.56%; The content of other chemical components increased by about 20~50%.

In general, the chemical composition of different glass relics will change greatly after weathering, and different types of glass relics have different changes, which may be related to the chemical composition of the soil in the buried environment during weathering.

(3) Prediction of chemical composition content

According to the detection data of weathering points, the chemical composition content before weathering is predicted. The relative increment before and after weathering obtained in Table 7 and Table 8 is taken as the known condition. According to the definition of relative increment, we transform it to obtain the prediction formula that can be used to predict the content of various chemical components before weathering:

$$\hat{x}_{ij}^{(0)} = \frac{x_{ij}^{(1)}}{1+\alpha_{ij}}$$

Where, is the predicted value of $\hat{x}_{ij}^{(0)}$, represents the predicted average content of the j component of the i type glass without weathering, $x_{ij}^{(0)}$ represents the content of the component to be predicted after weathering, α_{ij} is the average relative increment of the j component of the i type glass after weathering obtained in Tables 7 and 8.

2.1.3 Solving the Model

Table 9. Prediction results for part of the problem

type	Serial number	The sampling point	x_{i1}	...	x_{i8}	x_{i9}	...	x_{i14}
High potassium	07	The surface sampling	67.08632		2.418342	0		0
High potassium	09	The surface sampling	68.75045		2.434045	0		0
High potassium	10	The surface sampling	70.01664		1.319095	0		0
High potassium	12	The surface sampling	68.22227		2.59108	0		0
High potassium	22	The surface sampling	66.81861		0.863693	0		0
High potassium	27	The surface sampling	67.08632		2.418342	0		0
...		...		...			...	
Lead barium	02	The surface sampling	54.0122		0.20944	30.3059		0

Lead barium	08	The surface sampling	29.9836		8.38569	18.5068		5.27931
Lead barium	43	The surface sampling	76.3138		2.1508	0		0
Lead barium	43	The surface sampling	49.0249		0.58805	31.8191		0
Lead barium	50	The surface sampling	67.024		0.91026	28.3296		0
Lead barium	57	The surface sampling	37.8443		0.93443	29.1204		0
Lead barium	58	The surface sampling	45.2434		2.52134	25.3921		0

MATLAB programming was used to substitute the data of weathering points into Equation (3) to obtain it. Partial prediction results of weathering points are shown in the following table (see Appendix for complete prediction results).

2.2 Establishment and Solution of the Model of Problem 2

In order to explore the classification rules of glass types, it is necessary to determine exactly what chemical composition plays a 'key' role in distinguishing the two types of glass, that is, to find the classification characteristics. In this paper, the classification rules of high-potassium glass and lead-barium glass are analyzed by studying the random forest model in the process of characteristic importance evaluation. Then the two glass types are subclassified by systematic clustering.

2.2.1 Random Forest Seeks Classification Characteristic Indexes

The essence of feature screening in random forest [2] is to judge how much 'contribution' each feature makes on each tree in random forest, then take an average value, and finally compare the contributions between features.

The 'contribution' here can usually be measured by the Gini index.

Firstly, VIM is used to represent the importance score of variables, and the Gini index is represented by GI. If there are J features, which are $X_1, X_2, \ldots, X_J$ decision trees and C categories, it is necessary to calculate the Gini index score of each feature (VIM_j^{Gimi}), namely the average change of node splitting impurity of the so-called J feature in all decision trees of RF.

The Gin index of the node q of the i decision tree can be calculated by the following formula:

$$\mathrm{GI}_{\mathrm{q}}^{(\mathrm{i})} = \sum_{\mathrm{c}=1}^{|\mathrm{C}|} \sum_{\mathrm{c}' \neq \mathrm{c}} \mathrm{p}_{\mathrm{qc}}^{(\mathrm{i})} \mathrm{p}_{\mathrm{q}'}^{(\mathrm{i})} = 1 - \sum_{\mathrm{c}=1}^{|\mathrm{C}|} (\mathrm{p}_{\mathrm{qc}}^{(\mathrm{i})})^2$$

In the above equation, C represents C categories, and p_{qc} represents the proportion of category c in node q.

The Gini index change before and after node q branches can represent the importance of eigenvalue X_j at node q in the i tree. Specifically expressed as:

$$\mathrm{VIM}_{\mathrm{jq}}^{(\mathrm{Gini})(\mathrm{i})} = \mathrm{GI}_{\mathrm{q}}^{(\mathrm{i})} - \mathrm{GI}_{1}^{(\mathrm{i})} - \mathrm{GI}_{\mathrm{r}}^{(\mathrm{i})}$$

In the above formula,$GI_1^{(i)}$and $GI_r^{(i)}$ respectively represent the Gini index of the two new nodes after the branch.

Furthermore, if the feature X_j is q at the node of the i tree, then the importance of q in the i tree is:

$$VIM_j^{(Gini)(i)} = \sum_{q \in Q} VIM_{jq}^{(Gini)(i)}$$

Assuming that there are I trees in RF, then

$$VIM_j^{(Gini)} = \sum_{i=1}^{I} VIM_j^{(Gini)(i)}$$

Finally, all the obtained importance scores were normalized to get

2.2.2 Solving Random Forest Classification Model

With the content of 14 chemical components as independent variables and two types of glass (high-potassium glass, lead-barium glass) as dependent variables, a random forest classification model was established by training and training data set through MATLAB programming. After calculating the importance of features, the following statistical graph is obtained:

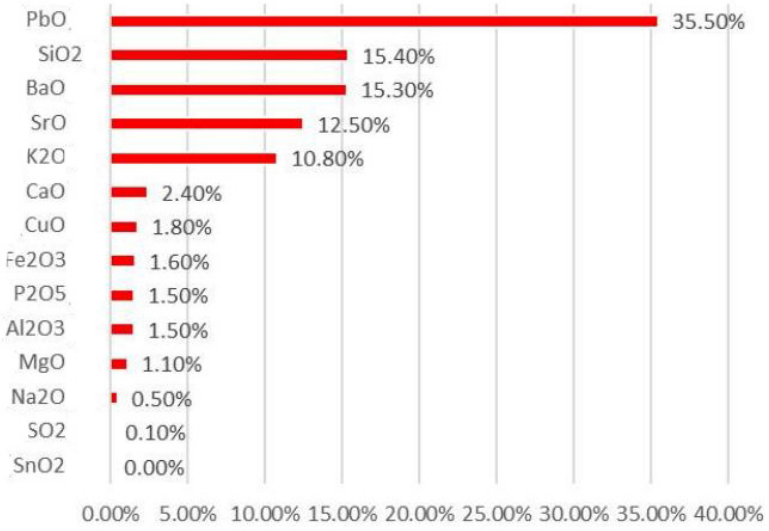


Fig 5. Calculation Results of Characteristic Importance

It can be seen from the above figure that the important components that distinguish these two types of glass are sorted by their influence. The top 5 components are PbO > SiO_2 >BaO > SrO >K_2O.

Combined with the attached data, the average relative contents of each chemical component in the four groups of samples with high potassium weathering, high potassium unweathering, lead-barium weathering, and lead-barium unweathering were analyzed and compared as follows:

Table 10. Comparison of characteristic components between high potassium glass and lead-barium glass

group	type	PbO	SiO_2	BaO	SrO	K_2O
weathering group	high potassium	0.00	93.96	0.00	0.00	0.54
	lead barium	36.56	35.90	9.44	0.33	0.14
no weathering group	high potassium	0.41	67.98	0.60	0.04	9.33
	lead barium	23.59	53.44	10.50	0.30	0.26

Through comprehensive analysis of the above table and the characteristic components obtained by random forest, it can be seen that the classification rules of high-potassium glass and lead-barium glass are as follows:

1. when the content of PbO and barium oxide BaO of a glass relic is lower than 1%, it is very likely to belong to the high-potassium type glass.

2. When the SiO2 content of a glass relic is higher than 90%, it is highly likely to be high-potassium glass after weathering.

3. When the SrO content of a glass cultural relic is about 0.3%, the glass shall belong to lead and barium glass.

2.2.3 Subclass Division Method based on Standard Deviation

(1) Model preparation

According to the need of solving this problem, we first make a statistical concept explanation for the following terms: variance: the mean of the squared discrete difference between the values of each variable and its mean.

Standard deviation: square root of variance.

Boxplot [3]: It is drawn by the five feature values of the maximum value, minimum value, median value and two quartiles of a group of data. It is mainly used to reflect the features of the original data distribution, and can also be used to compare the distribution features of multiple groups of data.

Interquartile range (IQR): The quartile spread, which is the distance between the third and first quartiles.

Outlier: A value in a data index that is significantly different from one or more other values. Outliers can be identified by IQR.

(2) Selection of classification indicators

Based on the above explanation of the classification rules of glass types, in order to make it possible, we first need to make interactive comparison and analysis of the chemical composition of high-potassium glass and lead-barium glass respectively, and then carry out sub-classification of high-potassium glass and lead-barium glass on this basis.

We first calculated the standard deviation of the content of various chemical components of the two glass types before and after weathering. The standard deviation is represented by symbol σ, which is often used as the measurement basis for the degree of distribution in statistics and can reflect the degree of dispersion of a certain data set. The standard deviation calculation formula is as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

The standard deviation of four types of sample data (high potassium unweathered, high potassium weathered, lead-barium unweathered, lead-barium weathered) was calculated by MATLAB programming, and the calculation results are as follows:

Table 11. Standard deviation of each chemical component in the four types of sample data

Chemical composition	High potassium weathering	High potassium is not weathered	Lead barium weathering	Lead barium is not weathering
SiO_2	1.734	8.755	9.005	14.587
Na_2O	0.000	1.287	0.587	1.538
K_2O	0.445	3.920	0.246	0.398
CaO	0.488	3.092	1.667	1.458
MgO	0.306	0.676	0.719	0.545
Al_2O_3	0.964	2.492	2.747	1.385
Fe_2O_3	0.069	1.667	0.751	1.445
CuO	0.935	1.660	2.981	2.491
PbO	0.000	0.589	12.077	9.094
BaO	0.000	0.982	7.966	6.950
P_2O_5	0.210	1.434	3.989	1.571
SrO	0.000	0.048	0.23	0.314
SnO_2	0.000	0.681	0.286	0.158
SO_2	0.000	0.186	0.661	1.015

The smaller the standard deviation, the smaller the difference of the chemical composition in different glass contents, so it is difficult to be used as an indicator to distinguish different kinds of glass. The larger the standard deviation, the greater the difference of the chemical composition in different glass, so it can be used as a characteristic index to distinguish different kinds of glass.

In order to more intuitive comparison with different chemical composition content is the size of the standard deviation, we map a box to observe the distribution of data as a whole, through the median, quantile, 25% 75% quantile, edge and bottom edge on statistics generated a box body figure, acceptable to the normal data, the enclosure on the outside of the edge and bottom edge is abnormal data, as shown below:

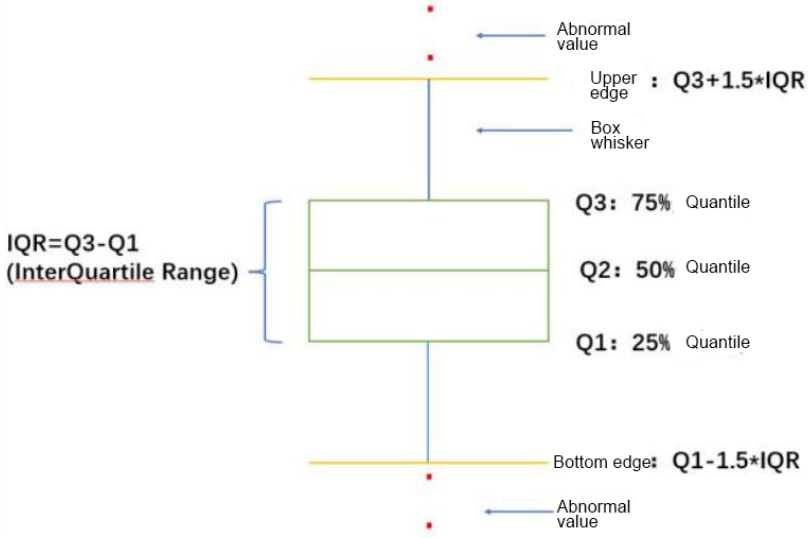


Fig 6. Box type diagram

According to the standard deviation of various samples obtained in the above table, we drew a box chart of standard deviation through SPSS software, and arranged the standard deviation data of each group from small to large. Q1 is equal to the fourth value of the sample from small to large, Q2 is the median of the number sequence, Q3 is equal to the third value of the sample from small to large, and interquartile spacing (IQR)=Q3-Q1.

The upper box indicates that the overall data is too large, and the lower box indicates that the overall data is too small, as shown in the following figure:

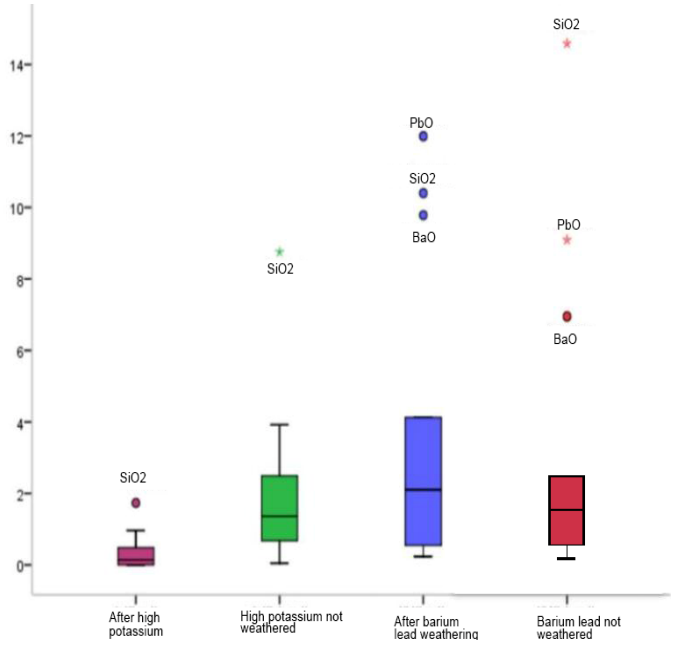


Fig 7. Analysis of standard deviation box shape

Consider the rationality of the glass and the class division, we choose difference obvious indicators, doesn't mean all the more obvious difference can be the same kind of difference, the greater the more help to correct and reasonable division of the class, by above can see intuitive, box figure said the outliers of points on the standard deviation of the difference of 0 index, index of this kind of hard to reflect the difference between the data, Therefore, it is not conducive to the subclass division of glass, which should be eliminated, that is, other indicators should be selected during the subclass division.

(3) Partition method -- k-means ++ clustering

①When dividing each category into subclasses, the partition method chosen in this paper is K-means ++ clustering, and the specific flow chart of the algorithm is as follows:

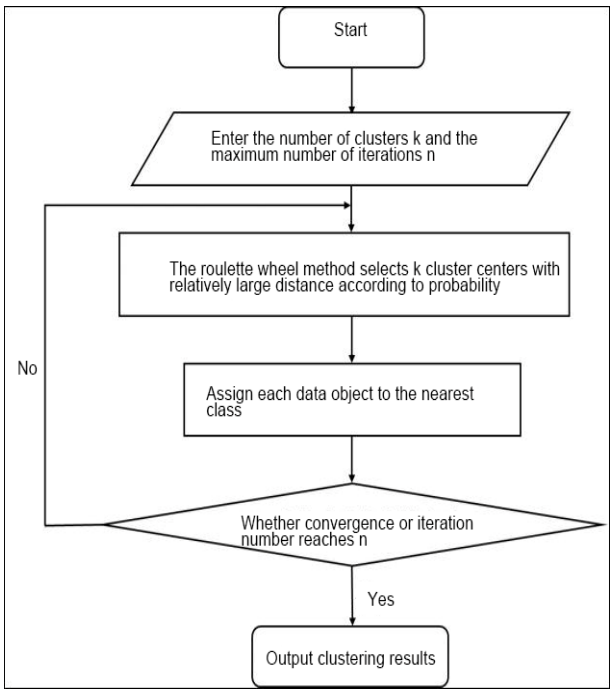


Fig 8. K - mean++clustering flow chart

②Quantitative index difference analysis

SPSS software was used to analyze the differences of selected appropriate chemical components, and four groups of data including high potassium weathering, high potassium weathering without weathering, lead-barium weathering, lead-barium weathering without weathering were analyzed respectively.

The specific evaluation indexes of difference analysis are F-test results and significance P value, which mainly analyzes whether the P value of each index is significant, that is, $P<0.05$.

Table 12. Difference analysis of high potassium weathering group

Chemical composition	F	P
Na_2O	1.892	0.199
K_2O	18.357	0.002***
CaO	4.321	0.064
MgO	0.289	0.603
Al_2O_3	1.569	0.239
Fe_2O_3	1.311	0.279
CuO	0.141	0.715
PbO	0.023	0.882
BaO	1.510	0.247
P_2O_5	6.307	0.031*
SrO	0.620	0.449
SnO_2	0.476	0.506
SO_2	1.959	0.192

Since K-means ++ clustering is adopted in this paper, this clustering method can maximize the difference between individual cases in different clusters, so F-test should only be used for descriptive

purposes. If the measured significance level P<0.05, the null hypothesis is rejected and significant differences exist in the obtained data, indicating that this index (chemical composition) plays a key role in clustering results (subclass division). On the contrary, it indicates that the data do not show differences, that is, the corresponding index has a slightly weak influence on the clustering results.

The Table 12 only shows the difference analysis results of the high-potassium weathering group.

According to the difference analysis results in the above table, we can see that the P values of K2O and P2O5 marked red are less than 0.05, so these two chemical components are suitable as indicators for subclass division.

(4) Result of subclass division

According to the above K-means ++ clustering method [4], we used SPSS to classify high-potassium glass and lead-barium glass into subclasses, and the results of the classification were visualized as follows:

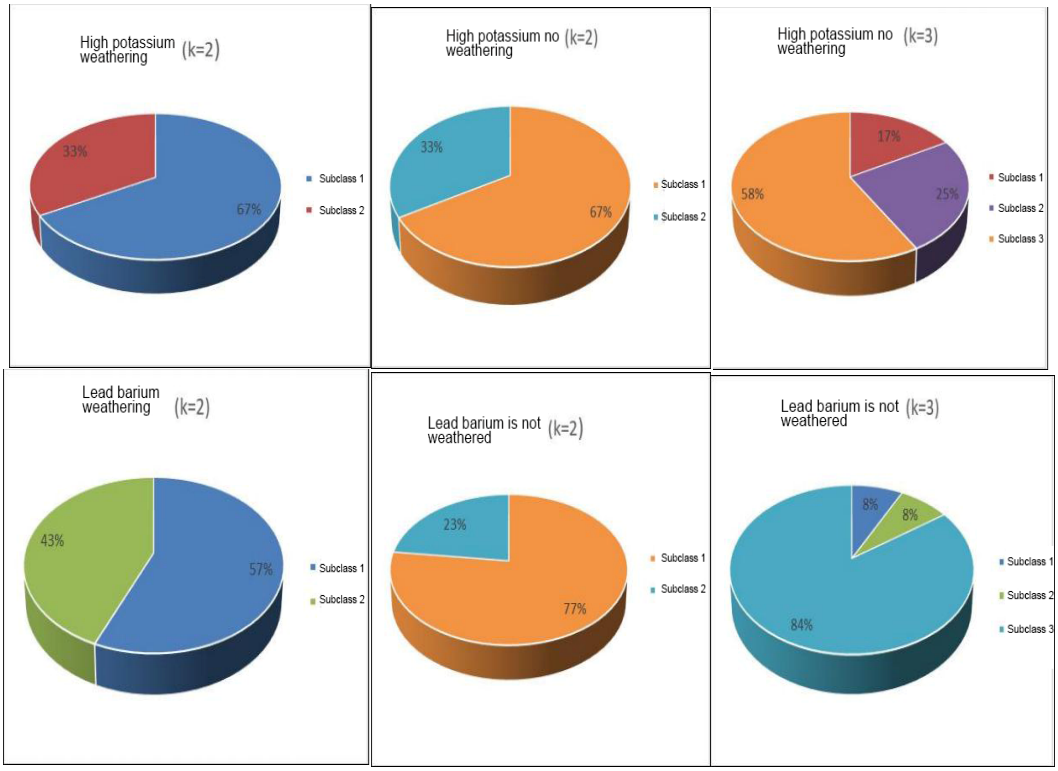


Fig 9. Visualization results of subclass division

(5) Sensitivity analysis

By adjusting the value of the variable, it was randomly scaled within the range (-125%, +125%). The comparison between the results and the results without disturbance was observed and the sensitivity analysis was conducted. The results are as follows:

Table 13. Sensitivity analysis results

category	5%	10%	15%	20%	25%
high potassium weathering 2	100%	100%	100%	100%	100%
high potassium is not weathered 2	100%	100%	100%	99%	95%
lead and barium weathering 2	100%	100%	100%	100%	100%
lead barium is not weathered 2	100%	100%	100%	100%	90%

As can be seen from the sensitivity analysis results, when the data disturbance is greater than 20%, the prediction accuracy of high potassium unweathered class begins to decline. When the disturbance

is more than 25%, the clustering accuracy of Pb and barium unweathered classes starts to decrease. However, in the overall range of (-120%, + 120%), the model has good robustness and universality.

2.3 Establishment and Solution of the Model of Problem 4

Question 4 requires the analysis of the correlation between the chemical composition of different categories of glass cultural relics samples and the difference between the chemical composition correlation of different categories.

2.3.1 Pearson Correlation Coefficient

Both high-potassium glass and lead-barium glass are composed of 14 chemical components, which are continuous quantitative variables. We first thought of using Pearson correlation coefficient to characterize the relationship between these chemical components.

Pearson's correlation coefficient is a measure of the degree of linear correlation between two variables, assuming that there are two groups of data $X:\{X_1, X_2, \ldots, X_n\}$ and $Y:\{Y_1, Y_2, \ldots, Y_n\}$, the population mean is:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

The sample covariance is:

$$\mathrm{Cov}\,(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})\,(Y_i - \overline{Y})}{n-1}$$

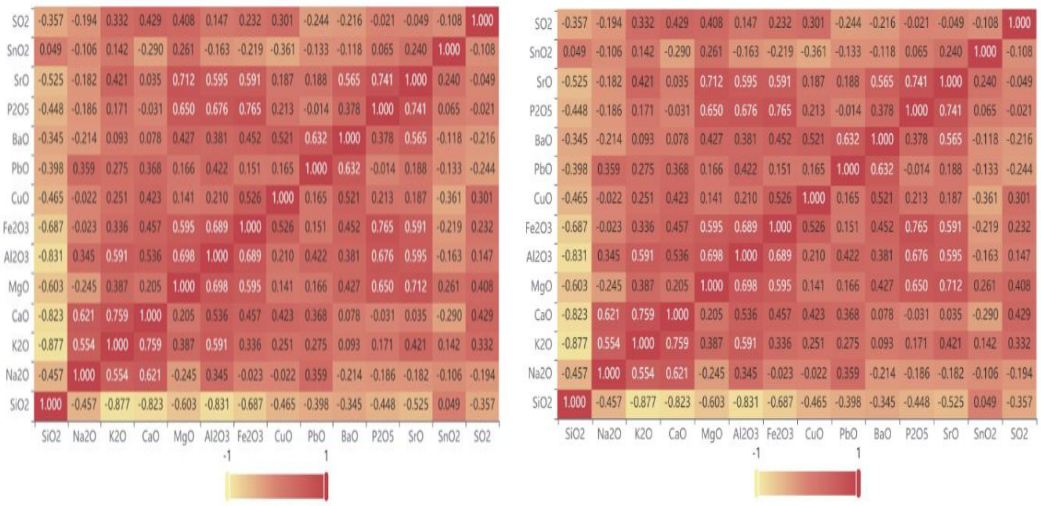
The Pearson correlation coefficient of the sample is:

$$r_{XY} = \frac{\mathrm{Cov}\,(X, Y)}{S_X S_Y}$$

Where, S_X is the sample standard deviation of X:

$$S_X = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2)}{n-1}}, S_Y = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2)}{n-1}}$$

MATLAB was used to calculate the Pearson correlation coefficient between the chemical components of two kinds of glass cultural relics sample 14, and draw the correlation coefficient thermogram, as shown below:



(a)High potassium glass **(b)Lead barium glass**

Fig 10. Thermal Diagram of Correlation Coefficient

The Pearson correlation coefficient of different variables can be more intuitively analyzed by the thermal map in the form of color depth.

Pearson's correlation coefficient can only be used to measure the correlation degree of linear variables. We first need to determine that there is a linear relationship between two variables, so that Pearson's correlation coefficient can reflect their correlation degree. In other words, a large Pearson correlation coefficient does not necessarily mean a high degree of linear correlation between two variables. Therefore, it is necessary to carry out the hypothesis test. Pearson correlation coefficient hypothesis test requires that the experimental data come from a normally distributed population. We test whether the data are normally distributed by combining *Shapiro-Wilk* test and PP diagram, and some PP diagrams are as follows:

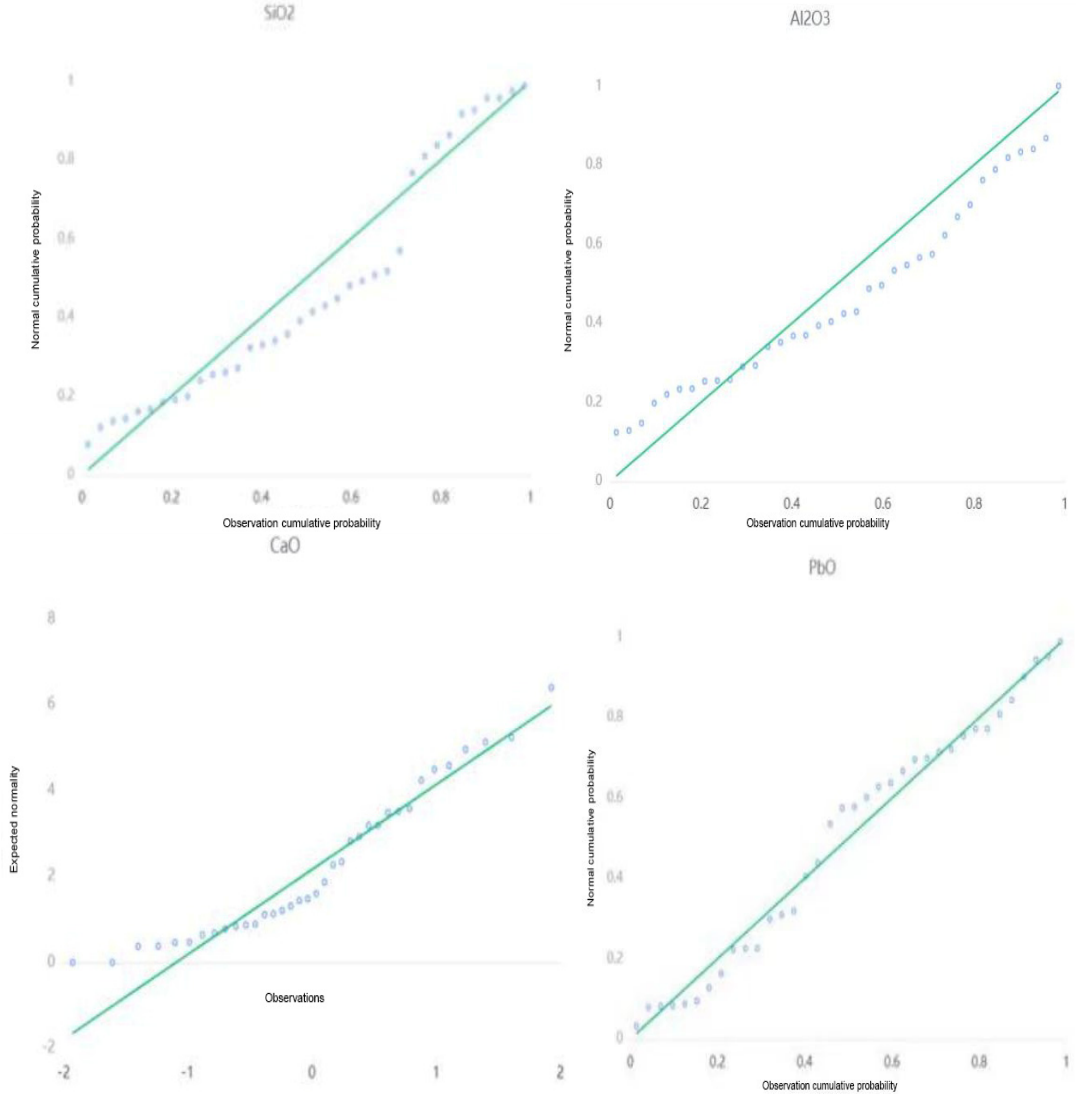


Fig 11. P-P Diagram

According to the PP figure, we can think that the experimental data basically conform to the normal distribution.

After confirming that the experimental data were in line with normal distribution, we conducted a significance test between variables and found that the results were significant.

2.3.2 Grey Correlation Analysis

Grey correlation analysis [5] is a systematic analysis method. Different from regression analysis and principal component analysis, grey correlation analysis has a fast calculation speed and is equally applicable to irregular samples and sample size, and there will be no inconsistency between qualitative analysis results and quantitative results. The basic idea is to judge whether the reference data column is closely related by determining the degree of geometric similarity between the

reference data column and several relative data columns, which reflects the degree of correlation between curves.

For the purpose of this problem, the chemical composition of two kinds of glass, although there are 14 but every kind of chemical composition, the proportion of total composition is different, this leads to some of the chemical composition of very small though has also been detected, but is its impact on the overall properties and appearance of glass, study the component and other components of relationship has no meaning.

Taking high-potassium glass as an example, lead and barium glass was treated in the same way. The proportions of 14 chemical components in 18 high-potassium glasses were averaged and the chemical components with very small proportions were excluded. The remaining components were shown in the following table:

Table 14. High potassium glass

SiO_2	K_2O	CaO	Al_2O_3	Fe_2O_3	CuO	P_2O_5	SrO
76.64	6.4	3.85	5.06	1.38	2.16	1.03	0.03

Table 15. Lead and barium glass

SiO_2	CaO	Al_2O_3	CuO	PbO	BaO	P_2O_5	SrO
34.42308	2.207436	3.044872	2.036154	36.74051	11.37128	3.819487	0.377949

Only analysis of total accounted for 96.53% of SiO_2, K_2O, CaO, Al_2O_3, Fe_2O_3, CuO, P_2O_5, and SrO relationship in the 8 kinds of chemical composition, using grey correlation analysis of the correlation coefficient is analyzed, the chemical composition of the selection of relatively large as the dependent variable of analysis (mother sequence), The other suitable variables were selected as independent variables (sub-sequences), and the grey correlation analysis model was established to calculate the grey correlation degree.

(1) Determining the analytical sequence

In grey correlation analysis, the data series that can reflect the behavior characteristics of the system is called parent series (parent index), which is similar to the dependent variable. The data series composed of factors that affect the behavior of the system are called sub-series (sub-indicators), similar to independent variables. In this case, the high potassium glass, combined with system clustering results with chemical composition proportion, silica, potassium oxide and alumina as high potassium three important ingredients in the glass, and its proportion and reached 88%, so I have to only three kinds of chemical components in sequence, to analyze its relationship with the other remaining ingredients.

Take SiO_2 as the parent sequence for example:

$$X_0 = (X_0(1), X_0(2), \cdots, X_0(n))^T$$

Subsequence is

$$X_1 = (X_1(1), X_1(2), \cdots, X_1(n))^T$$
$$X_2 = (X_2(1), X_2(2), \cdots, X_2(n))^T$$
$$X_m = (X_m(1), X_m(2), \cdots, X_m(n))^T$$

(2) preprocessing of variables

First, the mean value of each index is calculated, and then each element of the index is divided by its mean value, so as to achieve the purpose of removing dimensional influence, narrowing the range of variables and simplifying calculation.

Initial value processing:$x_i(k) = \frac{x_i(k)}{x_i(1)}, k = 1,2 \cdots n; i = 0,1,2 \cdots m$

Mean processing:$x_i(k) = \frac{x_i(k)}{x_i}, k = 1,2 \cdots n; i = 0,1,2 \cdots m$

(3) Calculate the correlation coefficient between each index in the subseries and the parent series

$$y\{x_0(k), x_i(k)\} = \frac{a + \rho b}{|x_0(k) - x_i(k)| + \rho b}$$

In the above equation, ρ ∈ (0,1)is the discrimination coefficient, the smaller ρ is, the greater the resolution, generally 0.5, a is the maximum difference between the two levels, b is the maximum difference between the two levels.

$$a = \min_i \min_k |X_0(k) - X_i(k)|, b = \max_i \max_k |X_0(k) - X_i(k)|$$

(4) Calculate the grey correlation degree

The correlation degree represents the degree of similar correlation between each child sequence and its parent sequence, which is calculated by the average value of the correlation coefficient. The correlation degree value is between 0 and 1, and the larger the value is, the stronger the correlation between the child sequence and its parent sequence, and the higher the correlation degree is, the closer the relationship between the child sequence and its parent sequence is. The calculation formula is:

$$y(x_0, x_i) = \frac{1}{n}\sum_{i=1}^{n} y\{x_0(k), x_i(k)\}$$

High-potassium glass takes SiO_2, K_2O and Al_2O_3 as the parent sequence, and other components as the sub-sequence to establish the gray correlation model; lead-barium glass takes SiO_2, PbO and BaO as the parent sequence, and other components as the sub-sequence to establish the gray correlation model. MATLAB is used to solve the problem, and the results are as follows:

Table 16. Grey correlation analysis of high potassium glass

Evaluation of item	correlation	ranking	Evaluation of item	correlation	ranking	Evaluation of item	correlation	ranking
Al_2O_3	0.775	1	CaO	0 83	1	CaO	0 755	1
CuO	0.773	2	Al2O3	0.772	2	P2O5	0.735	2
P_2O_5	0.76	3	Fe2O3	0.765	3	Fe2O3	0.722	3
K_2O	0.703	4	CuO	0.737	4	CuO	0.681	4
CaO	0.689	5	P2O5	0.703	5	SrO	0.6	5
Fe_2O_3	0.683	6	SrO	0.696	6			
SrO	0.608	7						

Table 17. Gray correlation analysis of lead and barium glass

Evaluation of item	correlation	ranking	Evaluation of item	correlation	ranking	Evaluation of item	correlation	ranking
Al_2O_3	0.822	1	SrO	0 826	1	SrO	0 755	1
PbO	0.791	2	Al_2O_3	0.805	2	CuO	0.735	2
BaO	0.782	3	CaO	0.8	3	Al_2O_3	0.722	3
SrO	0.743	4	BaO	0.787	4	CaO	0.681	4
CaO	0.732	5	P_2O_5	0.741	5	P_2O_5	0.6	5
CuO	0.72	6	CuO	0.718	6			
P_2O_5	0.684	7						

Taking the first table as an example for analysis, the correlation degree value is between 0 and 1, the larger the value is, the stronger the correlation between the evaluation item and the parent sequence, and the higher the correlation degree is. Based on the correlation degree value, the ranking of all evaluation items is obtained. As can be seen from the above table, for the 7 evaluation items, alumina (Al2O3) has the highest evaluation (correlation degree: 0.775), and the strongest correlation with silica, followed by copper oxide (CuO) (correlation degree: 0.773).

2.3.3 Analysis of Variance

This problem also need to compare between different categories of chemical composition of the difference of correlation in this paper, the different categories Glass of grey correlation coefficient analysis of variance to observe the chemical composition of high potassium and lead barium glass association, the difference of the types of glass of a class variables, each kind of glass corresponds to the 18 gray close contact number, The 36 grey correlation coefficients were defined as quantitative fields.

(1) Test for normality

SPSS was used to test whether the overall distribution of data was normal, and the results were as follows:

Table 18. Normality test results

Variable name	Sample size	Median	Mean	Standard deviation	Partial degrees	Kurtosis	S - W test
Correlation	36	0.739	0.738	0.054	-0.541	0.489	0.966(0.326)

Shapiro-Wilk test was used for the sample. If the data were significant ($p<0.05$ or 0.01), it meant that the null hypothesis (the data were in line with the normal distribution) was rejected; otherwise, it meant that the data were in line with the normal distribution. The P value of significance of this test was 0.326. The null hypothesis could not be rejected, so the data were fully normally distributed.

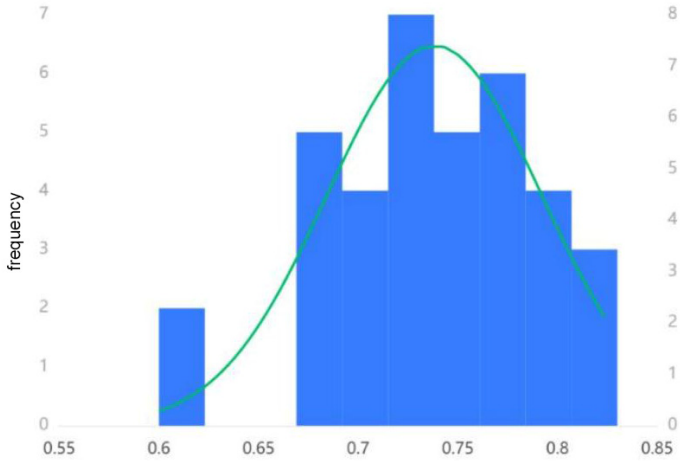


Fig 12. Histogram of normality test

As can be seen from the above figure, the data can be considered as normally distributed.

(2) Homogeneous test of variance

The solution results of SPSS are as follows:

Table 19. Homogeneous test of variance

	Type(Standard deviation)		F	P
Correlation	High potassium(n=18) (0.058)	Lead barium(n=18) (0.045)	0.699	0.409

As can be seen from the table, the P value of significance is 0.409, which does not show significance at the level, and the null hypothesis cannot be rejected, so the data meet the homogeneity of variance.

References

[1] Fang Xiangzhong. Chi Square Distribution and Chi Square Test. China Statistics, 2022 (05): 29-31.

[2] Wang Jingli, Yan Shicheng, Li Hao, Liang Xuepeng. Research on specific surface feature recognition of functional area based on random forest model. Mapping and Spatial Geographic Information, 2022,45 (08): 5-8.

[3] Wang Haotian, Li Xiaorun, Zhao Liaoying Cable segmentation method based on box graph and threshold boundary of break point.Computer Application and Software, 2021,38 (09): 244-249.

[4] Zhong Xi, Sun Xiang'e. Research on naive Bayesian integration method based on Kmeans++clustering. Computer Science, 2019, 46 (S1): 439-441+451.

[5] Guo Hong, Chen Yongming. Some Properties of Data Preprocessing Operator of Grey Relational Analysis Model.Journal of Chengdu University of Information Engineering, 2020,35 (02): 235-238.

[6] Zhu Longyin, Zhang Qifeng, Yan Danhai, Fan Linxiu. Research on the application of systematic clustering method in pattern recognition of traditional Chinese medicine. Journal of Gannan Normal University, 2018, 39 (03): 90-93.