Scientific Research Publishing

Open Journal of Applied Sciences, 2023, 13, 2149-2171
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

The Establishment of Mathematical Models for the Composition Analysis and Identification of Ancient Glass Products

Jenny Zhang*, Ding Li, Yu Xie, Junfeng Xiang

School of Energy and Power Engineering, Jiangsu University, Zhenjiang, China
Email: *2582137156@qq.com

How to cite this paper: Zhang, J., Li, D., Xie, Y. and Xiang, J.F. (2023) The Establishment of Mathematical Models for the Composition Analysis and Identification of Ancient Glass Products. *Open Journal of Applied Sciences*, **13**, 2149-2171.
https://doi.org/10.4236/ojapps.2023.1311167

Received: October 10, 2023
Accepted: November 26, 2023
Published: November 29, 2023

Abstract

Glass is the precious material evidence of the trade of the early Silk Road. The ancient glass was easily affected by the environmental impact and weathering, and the change of composition ratios affected the correct judgment of its category. In this paper, mathematical models and methods such as Chi-square test, weighted average method, principal component analysis, cluster analysis, binary classification model and grey correlation analysis were used comprehensively to analyze the data of sample glass products combined with their categories. The results showed that the weathered high-potassium glass could be divided into 12, 9, 10 and 27, 7, 22 and so on.

Keywords

Principal Component Analysis, System Clustering, Sensitivity Analysis, Binary Classification Model, Logistic Regression Analysis, Grey Correlation Analysis

1. The Problem Restatement

1.1. Problem Background

The Silk Road was the channel of cultural exchange between China and the West in ancient times. The exotic customs brought by the Silk Road had a subtle influence on the style of Chinese glass art. The Silk Road running through West Asia, East Asia, Europe and other places connected various styles along the road well [1].

The main raw material of glass is quartz sand; the main chemical composition is silica (SiO_2). Considering the high melting point of pure quartz sand, in order to reduce its melting temperature, the auxiliary use of flux is needed in refining.

Ancient commonly used fluxes are grass and wood ash, natural natrofoam, saltpeter and lead ore, and add limestone as a stabilizer, limestone after calcination into calcium oxide (CaO). Due to the addition of different flux, its main chemical composition is also different.

1.2. Problems to Be Solved

Now we have a batch of data about ancient Chinese glass products. Based on the chemical composition of these relics samples and other detection means, archaeologists have divided them into two types: high-potassium glass and lead barium glass. The classification information of these artifacts is given in Annex Form 1, and the proportion of the corresponding major component is given in Annex Form 2 (blank space indicates that the component is not detected). Your team is requested to conduct analysis and modeling according to the relevant data in the attachment to solve the following problems:

Question 1: The relationship between the surface weathering of these glass relics and their glass types, patterns and colors was analyzed. Combined with the type of glass, the statistical rule of whether there is weathering chemical composition content on the surface of cultural relics samples is analyzed, and the chemical composition content before weathering is predicted according to the detection data of weathering point.

Question 2: Analyze the classification rules of high-potassium glass and lead-barium glass according to the attached data. Appropriate chemical components were selected for each category and subcategorized. Specific classification methods and results were given, and the rationality and sensitivity of classification results were analyzed.

Question 3: Analyze the chemical composition of glass relics of unknown category in attached Form 3, identify their type, and analyze the sensitivity of classification results.

Question 4: Analyze the correlation between chemical components of glass cultural relics samples of different categories, and compare the differences of the correlation between chemical components of different categories.

2. Method

2.1. Analysis of Problem 1

Question one is divided into three small questions. The first part requires the analysis of the relationship between the surface weathering of these glass relics and their glass types, patterns and colors. First of all, after preprocessing the problem data, we should use Excel to visualize the data through images, have a preliminary fuzzy impression of the data, and compare and analyze whether the surface is weathered or not with the glass type, decoration and color respectively. Since all the data analyzed were qualitative data types, Chi-square test was applied to analyze the difference of qualitative data types to explore whether the glass type, ornamentation and color had a significant impact on weathering.

The second question is to analyze whether there is statistical rule of weathering chemical composition on the surface of cultural relics samples based on the type of glass. To solve this problem, we need to analyze the glass types by classifying them into lead barium and high potassium, and make statistical analysis based on the content of their chemical components.

The third question requires predicting the chemical composition content before weathering according to the weathering point detection data. After classifying the data in Annex 2 according to the type of glass and whether the glass is weathered or not, and eliminating the relevant problem data, we analyzed the data according to the problem. Since the number of relevant samples is not too large, and some chemical components in the data are not detected and replaced by 0 value, and there are many kinds of chemical components to be predicted, we used the weighted average method for prediction. Firstly, normal function is used to determine the weight of all component data, and weighted average is used to determine the average trend of each index, so as to predict the components before weathering.

2.2. Analysis of Problem 2

Question two is divided into two questions. The first part requires analyzing the classification rules of high-potassium glass and lead-barium glass according to the attached data. According to the average results obtained from Question 1 and Question 3, we use the drawing to carry out data visualization processing. Through the observation of the image, we could obtain the significance of the difference in the content of different types of glass in whether there is weathering or not, so as to determine the classification rule.

The second question is to select appropriate chemical components for each category, subdivide them, and analyze the rationality and sensitivity of the classification results. Considering the change and influence of chemical composition after weathering, we only select glass relics without weathering for analysis and classification in this case. Since there are many types of chemical components, principal component analysis is firstly carried out to determine several indexes of major components, and then cluster analysis is carried out on the numbering of glass relics according to these indexes, so as to obtain specific classification results. Finally, the rationality and sensitivity analysis were carried out by adjusting the main chemical components screened out and referring to relevant literature.

2.3. Analysis of Problem 3

Question 3 requires analysis of the chemical composition of glass relics of unknown category in Form 3, identification of their type, and analysis of the sensitivity of the classification results. This problem requires the establishment of a binary classification model, which converts the qualitative indicators of high potassium and lead barium into 0 - 1 variables. Based on the results of principal component analysis obtained in Question 2, linear discrimination is conducted

by using logistic regression to obtain the results of identification types. When dealing with this problem, we plan to divide all glass relics of different types into weathering and non-weathering parts, and carry out the application and treatment of classification model twice, so as to obtain more accurate and detailed results. After that, we analyzed the sensitivity of the results by adjusting the content of the main components of the glass relics of unknown category.

2.4. Analysis of Problem 4

Question 4 is divided into two mini-questions. The first part requires the analysis of the correlation between the chemical components of different types of glass relics samples. Since various chemical components in the data given in form 2 are blank and not detected, we choose to establish a grey correlation analysis model, and select the chemical component with a large proportion as the dependent variable (parent sequence), and other related main components as the independent variable (subsequence), and carry out grey correlation analysis on the two types of high potassium and lead barium respectively. The close degree between each index is determined, and the grey correlation coefficient is obtained for analysis.

The second question is to compare the differences in chemical associations between different classes. Based on the grey correlation coefficients of the two categories obtained in the first question, the results can first be compared and analyzed through literature review, and then the variance test can be conducted on the two groups of coefficients to compare the significant differences between them.

3. Model

3.1. Model Assumptions

1) Assume that the data given in the attachment are true and accurate;

2) Assume that the sample glass products are only affected by environmental weathering, without considering the existence of damage caused by other human factors;

3) It is assumed that the contents of each chemical component in glass products meet normal distribution;

4) Suppose that the cultural relics with shallow local weathering but still visible color patterns of cultural relics are marked as non-weathering on the surface;

5) It is assumed that the sum of the proportion of detected components in Form 2 and the data between 85% and 105% are regarded as valid data.

3.2. Symbol Description

Symbol	Description	Unit
A	Actual observed value	1
T	Theoretical inferential value	1

Continued

r_{ij}	Coefficient of correlation between variable x_i and x_j	1
ρ	Resolution coefficient	1
r_i	Comparison of correlation between sequence and reference sequence	1
χ^2	The amount of difference between variables of a given class	1
x_i	Chemical component i content	1
y_i	The content normalized by normal distribution is used	1

4. Establishment and Solution of the Model

4.1. Establishment and Solution of Problem 1 Model

4.1.1. Data Preprocessing

1) Elimination of invalid data

According to the title, Form 2 gives the proportion of the corresponding main components, and the sum of the proportion of each component should be 100%, but the sum of the proportion of components may not be 100% due to the means of detection and other reasons. In this question, the sum of component proportions and the data between 85% and 105% are regarded as valid data, so the data will be eliminated outside this range. The results are shown in Table 1.

2) Processing of vacancy data

In the Form 1, the colors of some cultural relics cannot be given, so we cannot determine them according to the existing information. In order to simplify the model, we eliminate these vacant data when analyzing the data in the Form 1. The results are shown in Table 2.

4.1.2. Establishment and Solution of the First Question Model

This problem requires an analysis of the relationship between the surface weathering of these glass relics and their glass types, patterns and colors. Through the analysis, the data were visualized and then Chi-square test was applied to analyze the difference of the data.

1) Data visualization processing

Excel was used to map and analyze the surface weathering of glass relics in terms of glass types, patterns and colors. The images displayed in Figure 1 are as follows.

From the above chart in Figure 1, we might conclude that:

a) The glass relics of sample B are not weathered, while about half of the glass relics of sample A and C are weathered.

b) For glass types, most lead-barium glass is weathered, and the weathering rate of high-potassium glass is relatively small, and the difference between lead-barium and high-potassium glass is large.

c) The coloring of ancient glass generally depends on the addition of transition metal elements such as Fe, Co and Mn. These elements exist in ionic state in

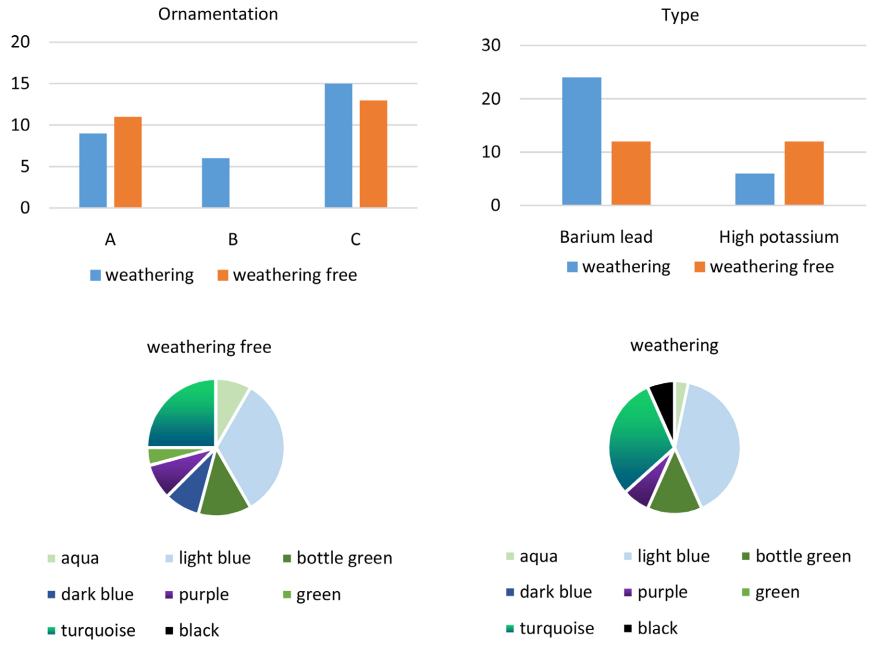

Figure 1. The relationship between the surface weathering of glass relics and its glass type, ornamentation and color.

Table 1. Excluded data from Form 2.

Cultural relic number	Heritage sampling point	Total percentage
17	17	71.89
15	15	79.47

Table 2. Excluded data from Form 1.

Cultural relic number	Ornamentation	Type	Color	Surface weathering
19	A	Lead Barium	-	Weathering
40	C	Lead Barium	-	Weathering
48	A	Lead Barium	-	Weathering
58	C	Lead Barium	-	Weathering

glass, and the coloring of glass is determined by the valence state of ions [2]. The proportion and number of light blue, dark green, purple and light green in weathering and non-weathering are roughly the same, and the proportion of black and blue green in weathering is larger. The difference of other colors is not obvious, so it can be inferred that whether the surface is weathered or not has little correlation with color.

2) Chi-square test

Chi-square test is mainly used to compare the difference analysis among certain variables and to calculate the deviation degree between the actual observed

value and the theoretical inferred value of the sample [3]. The calculation formula of chi-square test is as follows:

$$\chi^2 = \sum \frac{(A-T)^2}{T} \tag{1}$$

where, A is the actual observed value and T is the theoretical inferred value.

We used SPSS software to conduct chi-square test on the data, and the results are shown in Table 3.

According to the data in the above table, for surface weathering, the significance P value of ornamentation is 0.056*, the significance P value of glass type is 0.020**, and the significance P value of color data is 0.507. It can be concluded that the significant difference between glass type and surface weathering is the most obvious and has the greatest influence on its weathering, while the color has the least influence on surface weathering. Through chi-square distribution test, the results obtained by us are roughly the same as the data visualization analysis, which makes the results better tested.

4.1.3. Establishment and Solution of the Second Question Model

This problem requires the analysis of the statistical rule of weathering chemical composition on the surface of cultural relics samples based on the type of glass. After classifying glass types, SPSS software was used to conduct statistical analysis on the contents of each chemical component, and the mean value, standard deviation, median value, variance, kurtosis and coefficient of variation (CV) were calculated.

Table 3. Chi-square distribution test results.

Title	Name	Surface weathering		Total	X^2	Correction X^2	P
		Weathering free	Weathering				
Ornamentation	A	11	9	20	5.747	5.747	0.056*
	B	0	6	6			
	C	13	15	28			
Type of glass	Lead barium	12	24	36	5.4	4.134	0.020**
	High potassium	12	6	18			
Color	Light green	2	1	3	6.287	6.287	0.507
	Light blue	8	12	20			
	Dark green	3	4	7			
	Dark blue	2	0	2			
	The purple	2	2	4			
	Green	1	0	1			
	Blue-green	6	9	15			
	Black	0	2	2			

Note: ***, ** and * represent the significance level of 1%, 5% and 10% respectively.

4.1.4. Establishment and Solution of the Third Minor Question Model

This problem requires predicting the chemical composition content before weathering according to the weathering point detection data. According to the problem analysis, because the number of relevant samples is not large, and some chemical components in the data are not detected and replaced with 0 value, and there are many kinds of chemical components to be predicted, the weighted average method were used for prediction [4].

The weighted average method uses several observed values of the same variable arranged in chronological order in the past and takes the occurrence times of the chronological variable as the weight to calculate the weighted arithmetic average of the observed values. This number is used as a trend prediction method to predict the predicted value of the variable in the future period. The averages are calculated to take into account changes in long-term trends.

1) Data processing and calibration

Based on the assumption that the composition statistics of all chemical substances are in normal distribution, namely

$$X \sim N\left(\mu,\sigma^2\right), Y = \frac{X-\mu}{\sigma} \sim N(0,1) \tag{2}$$

Convert the data to a standard normal distribution and, in the case of a practical problem, change Y to a positive value. Since the data in column 14 has a large range of 0 s, in order to simplify programming calculation, the prediction in column 14 is omitted and 0 is directly taken as the prediction result.

Define the chemical composition of statistics, 13 chemical variables of $x_1, x_2, \cdots, x_{13}$, standardized postscript for $y_1, y_2, \cdots, y_{13}$, weight of w_k, k for sample size.

2) Determine the weight of each chemical component

Standard normal distribution function is used to determine the weight:

$$f(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \tag{3}$$

3) Calculate the weighted arithmetic mean

$$\overline{y}_i = \frac{y_1 w_1 + y_2 w_2 + y_3 w_3 + \cdots + y_k w_k}{\sum_1^k w_i} \tag{4}$$

After weighted average of 13 indexes before and after weathering, $\overline{y}_i$ and $\overline{y}_i'$ are obtained by using the above formula.

$$\alpha_i = \frac{\overline{y}_i}{\overline{y}_i'} \tag{5}$$

Define α_i for the proportion of the ith, an index before and after weathering factor, the calculation of weathering before each content.

With weathering finally point data of each chemical indicator on proportion factor α_i, so as to forecast the chemical composition of the weathering front.

4) Prediction calculation results

Based on the calculation of the above average weighting method and the prin-

ciple of MATLAB programming, we obtained the prediction results of the chemical composition content of each weathering point before weathering.

4.2. Establishment and Solution of Problem 2 Model

4.2.1. Establishment and Solution of the First Question Model

This problem requires the classification law of high potassium glass and lead barium glass to be analyzed according to the attached data. We divided the glass into weathering and non-weathering two parts respectively to analyze and compare the classification rules.

According to the comparison of data in the figure above, it can be roughly concluded that:

1) Among non-weathered glass products, high potassium glass and lead barium glass can be distinguished mainly according to the components potassium oxide, lead oxide and barium oxide. The potassium oxide in high potassium glass is much higher than that in lead barium glass. The composition of barium oxide and lead oxide in lead-barium glass is much higher than that in high-potassium glass.

2) In weathered glass products, the distinction between high potassium glass and lead barium glass can be mainly based on potassium oxide, phosphorus pentoxide, barium oxide, lead oxide, sulfur dioxide and other chemical components. The potassium oxide in high potassium glass is much higher than that in lead barium glass. However, unlike non-weathered glass products, lead-barium glass contains sulfur dioxide, tin oxide, barium oxide, lead oxide, and sodium oxide, which are not detected in high-potassium glass.

To sum up, potassium oxide and barium oxide can be used to distinguish high potassium glass from lead barium glass. High potassium oxide content can be divided into high potassium glass, barium oxide content can be divided into lead barium glass, and vice versa.

4.2.2. Establishment and Solution of the Second Question Model

This problem requires the selection of appropriate chemical components for each category, subclassification, and analysis of the rationality and sensitivity of the classification results. Considering the change and influence of chemical composition after weathering, we divided each category into pre-weathering and post-weathering subcategories respectively, and established a systematic clustering model based on principal component analysis and solved the problem.

1) Principal component analysis selecting major chemical components Principal component analysis is a dimensionality reduction algorithm that converts multiple indices into a small number of principal components that are linear combinations of the original variables and are unrelated to each other. Because there are many kinds of chemical constituents in this problem, principal component analysis is used to reduce the dimension of chemical constituents to simplify the model. The steps of the principal component analysis algorithm are as follows:

Step 1: Calculate the correlation coefficient matrix

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \tag{6}$$

In Formula (2), $r_{ij}\left(i,j=1,2,\cdots,p\right)$ is the correlation coefficient between the original variables x_i and x_j, and the calculation formula is

$$r_{ij}=\frac{\sum_{k-1}^{n}\left(x_{ki}-\overline{x}_i\right)\left(x_{kj}-\overline{x}_j\right)}{\sqrt{\sum_{k-1}^{n}\left(x_{ki}-\overline{x}_i\right)^2\sum\left(x_{kj}-\overline{x}_j\right)^2}} \tag{7}$$

Since R is a real symmetric matrix (*i.e.* $r_{ij}=r_{ji}$), you only need to calculate its upper or lower triangular elements.

Step 2: Calculate the eigenvalues and eigenvectors

First characteristic equation solution is $\left|\lambda I-R=0\right|$ and the characteristic value of $\lambda_i\left(i=1,2,\cdots,p\right)$, and make it in order of magnitude, that is, $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$; Then, the eigenvector $e_i\left(i=1,2,\cdots,p\right)$. The eigenvalue and the contribution rate and cumulative contribution rate of each principal component are calculated from the correlation coefficient matrix.

z_i contribution rate of principal component:

$$r_i\Big/\sum_{k=1}^{p}\gamma_k,\left(i=1,2,\cdots,p\right) \tag{8}$$

Cumulative contribution rate:

$$\sum_{k}^{m}\gamma_k\Big/\sum_{k}^{p}\gamma_k \tag{9}$$

Step 3: Calculate the principal component load matrix

$$p\left(z_k,x_i\right)=\sqrt{\gamma_k}e_{ki},\left(i,k=1,2,\cdots,p\right) \tag{10}$$

Thus, the principal component score can be further calculated:

$$Z=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{p1} & z_{p2} & \cdots & z_{pp} \end{bmatrix} \tag{11}$$

Five principal components were extracted by this method, which were silicon dioxide, sodium oxide, potassium oxide, calcium oxide and magnesium oxide. Similarly, principal components of other categories can be obtained from this. The results are shown in **Table 4**.

2) Establish a systematic clustering model for classification

The algorithm flow of the system clustering model is shown in **Figure 2**.

In this problem, the clustering number k was set as 2, and SPSS was used to conduct cluster analysis on the four categories (before high potassium weathering,

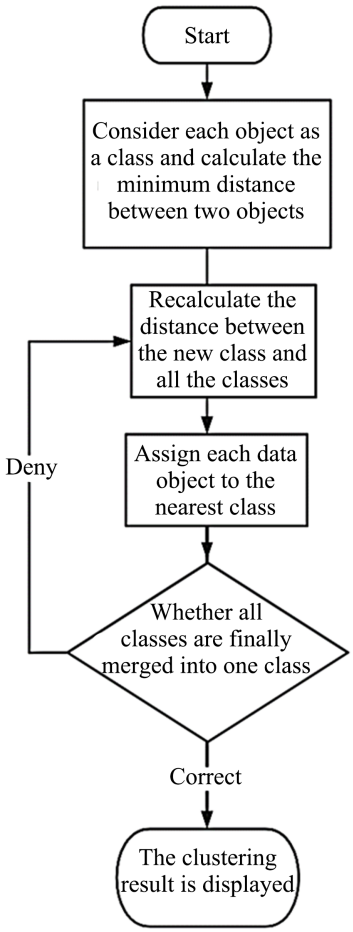

Figure 2. Flow chart of system cluster analysis.

Table 4. Results of principal component analysis.

Category	Principal component of chemical substance
Lead barium is not weathered	Silica, sodium oxide, potassium oxide, calcium oxide, magnesium oxide
Lead barium weathering	Silica, sodium oxide, potassium oxide, calcium oxide, magnesium oxide
High potassium	Silica, sodium oxide, potassium oxide, calcium oxide, magnesium oxide
High potassium weathering	Silicon dioxide, potassium oxide

after high potassium weathering, before lead barium weathering and after lead barium weathering) respectively according to a few principal component indexes selected in the above principal component analysis, and the clustering spectrum was obtained, thus obtaining subclass results. The results are shown in **Figure 3**.

Through the mathematical method of combining principal component analysis and clustering model [5], the classification and sub-classification results of glass cultural relics were finally obtained in **Figure 4**.

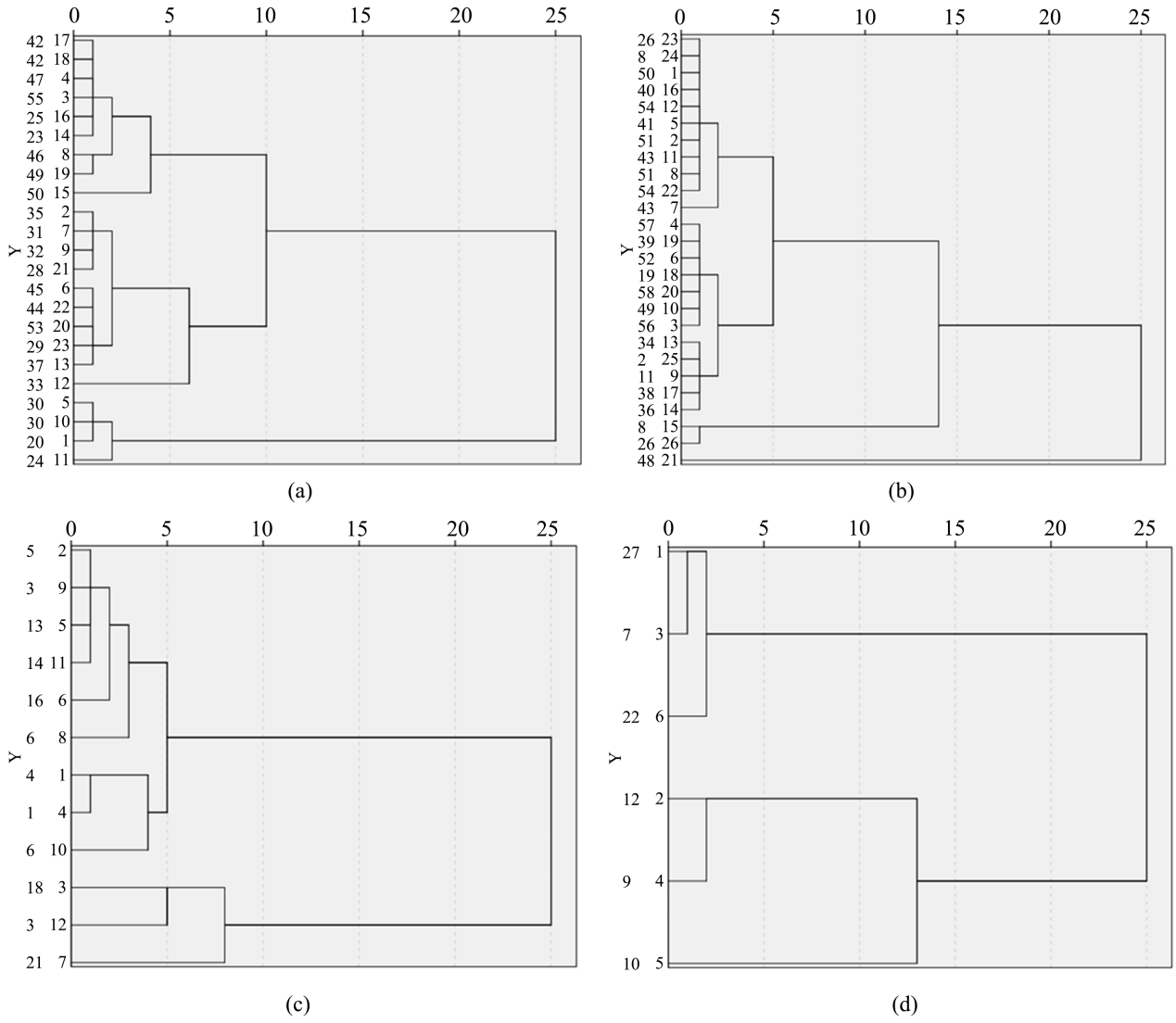

Figure 3. Cluster analysis pedigree diagram. (a) Lead barium is not weathered; (b) Lead barium weathering; (c) High potassium without weathering; (d) High potassium weathering.

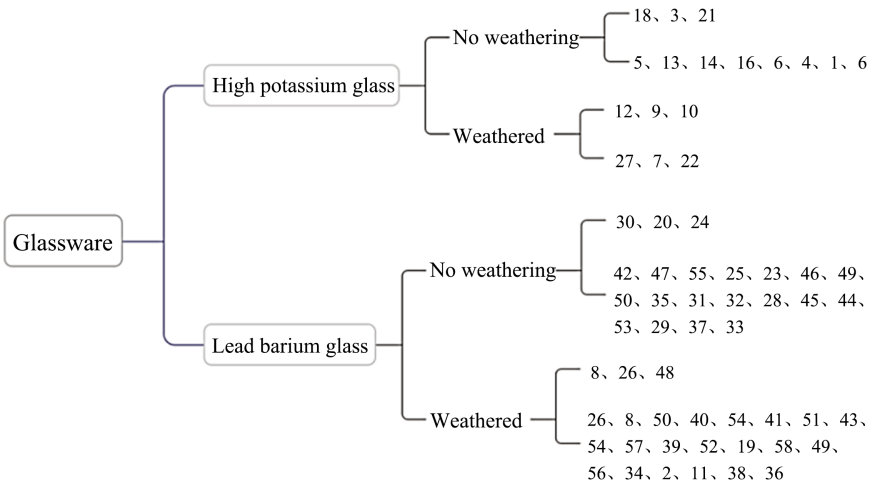

Figure 4. Classification results of glass relics.

4.2.3. Results and Discussion

1) Result test and analysis

a) Rationality analysis

Principal component analysis uses the idea of dimensionality reduction to transform multiple indicators into fewer comprehensive indicators on the premise of retaining the original data features, avoiding the interference of redundant information and reducing the influence of subjective factors. Cluster analysis takes the results of each principal component as a new index of evaluation, and uses the square Euclidean distance and the average join to construct the system cluster graph, so as to measure the advantages and disadvantages of various classes.

It is found that the species with better results in the chemical composition of glass products cluster together in the same group, which provides a certain scientific basis and indirectly verifies the rationality of the results of principal component analysis and cluster analysis. By combining the two methods, this paper not only simplifies the index in a certain level, avoids the repetition of data, but also objectively classifies different types of glass according to weathering or not.

b) Sensitivity analysis

This problem requires sensitivity analysis of classification results. Based on the results of principal component analysis, we up-regulated and down-regulated the values of main chemical contents, and compared the changes of the results before and after. If the results changed after a chemical value was adjusted by 10%, it indicated that the sensitivity of the data to the chemical value was high; otherwise, if there was no difference before and after the change, it indicated that the sensitivity of the data to the chemical value was low.

Specific practices: On the basis of the classification of weathering or not, the main chemical components were increased by 10%, 20%, 30% in turn, and decreased by 10%, 20%, 30% to compare the changes in the results before and after.

Similarly, by using SPSS to increase the disturbance by 20% and 30% respectively, it is found that the results are changed, and the accuracy of the results is 100%.

Through the above results, we conclude that increasing and decreasing the content of each main chemical component has no significant effect on the classification result, so the sensitivity of the classification result is not high.

4.2.4. Improvement of the Second Minor Question Model

Since the above systematic clustering model based on principal component analysis is difficult to establish quantitative indicators of known data through SPSS for rationality analysis, and considering the large errors in the above results, the accuracy and rationality of the results are somewhat deficient, so we establish the binary K-means clustering model to improve this problem.

The algorithm steps of binary K-means clustering model are shown in **Figure 5**.

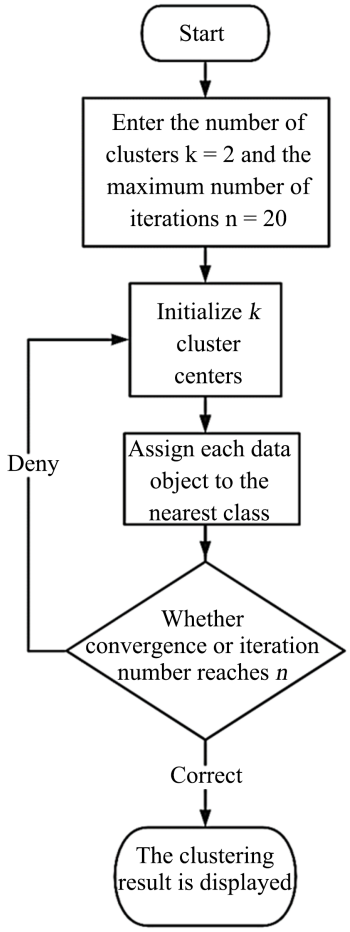

Figure 5. Flowchart of binary *K*-means clustering model algorithm.

This paper takes lead barium weathering as an example to analyze.

1) Significant difference analysis to determine the clustering basis index

We used SPSS software to analyze the significance of each index before clustering, and selected two main indicators as the basis for clustering.

2) Cluster analysis results

SPSS was used for cluster analysis, and the classification results were shown in **Figure 6**.

3) Rationality analysis

In the rationality analysis, we use tightness as an index to measure rationality. The tightness is the average distance from the sample points in each cluster to the cluster center. For clustering results, it is necessary to use the mean value of all cluster tightness to measure the quality of clustering results. The smaller the value of tightness is; the highest similarity of samples in the cluster is. The formula is as follows:

$$CP_c = \frac{1}{n}\sum_{i=1} n\left\|x_i - C\right\| \tag{12}$$

Combined with the analysis results of SPSS software, the tightness of each sample value can be obtained in **Table 5** and **Figure 7**.

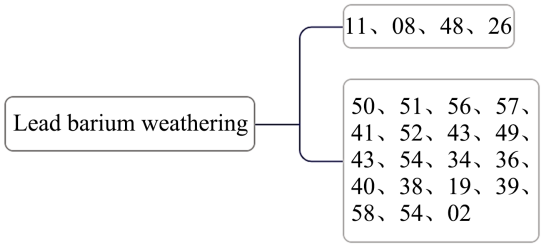

Figure 6. Cluster analysis results of lead barium weathering.

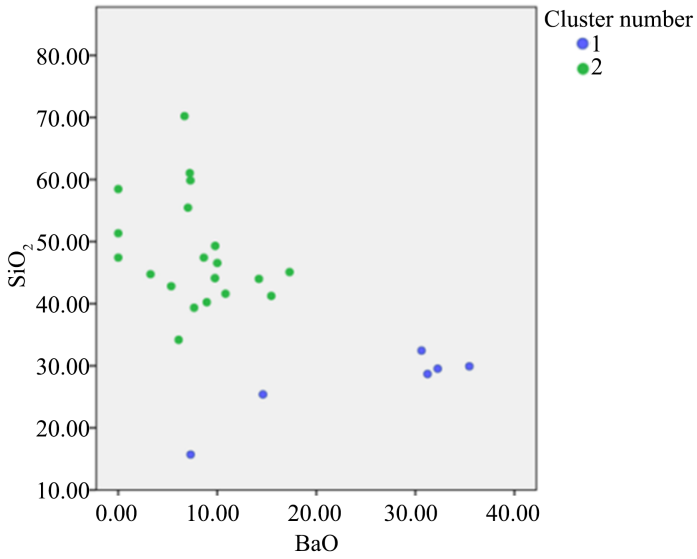

Figure 7. Visualization of Lead-Barium weathering tightness.

Table 5. Cluster tightness.

	Number	Category	Distance from cluster center
Lead Barium weathering	50	2	10.96112
	51 position 2	2	10.35684
	56	2	11.51126
	57	2	11.51387
	…	…	…
	08	1	9.41552
	02	2	13.77817
	26 Severe weathering point	1	24.02632
	Cluster tightness		14.23245

Based on tightness visualization and data analysis, the results of subclass classification can be well analyzed. Similarly, the average density of 17 clusters in the four categories of lead barium weathering, lead barium unweathering, high potassium weathering and high potassium unweathering were 14.23, 10.91, 1.19 and 6.04, respectively. The density of the cluster is within 15, and the value of

high potassium weathering is about 1. According to the smaller the density, the higher the degree of similarity within the cluster, it can be concluded that the model has a higher degree of similarity within the cluster, and the clustering results are reasonable and accurate.

The conclusion and sensitivity analysis of this model can be followed by the method of the original model, so as to get the relevant results.

4) Comparison of models

Compared with the system clustering model based on principal component analysis, the dichotomy K-means clustering model can better analyze the rational utilization of the tightness index. At the same time, the algorithm of this model is simple and express, which can carry out typical significance analysis for different categories. At the same time, the algorithm is more efficient and the results are more accurate.

4.3. Establishment and Solution of Problem 3 Model

This problem requires analysis of the chemical composition of glass relics of unknown category in Form 3, identification of their type, and analysis of the sensitivity of classification results. This problem requires the establishment of a binary model, and the basic ideas are shown in **Figure 8** [6].

4.3.1. Establishment of Binary Model

1) Establish linear probability model (LPM)

In order to simplify the model, this linear probability model directly uses regression model for regression:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i \tag{13}$$

Convert to vector product form:

$$y_i = x_i'\beta + \mu_i \tag{14}$$

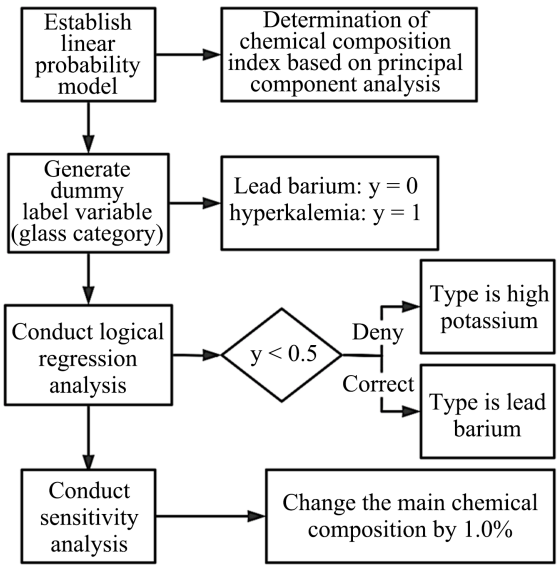

Figure 8. Thinking of problem 3 model.

2) Select the connection function

Sigmoid function $F(x, \beta)$ is introduced as the connection function to connect the explained variable x with the explained variable y:

$$F(x,\beta)=S(x_i'\beta)=\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \tag{15}$$

Conclusion by maximum likelihood estimate beta $\hat{\beta}$ into the formula

$$\hat{y}_i=S(x_i'\hat{\beta})=\frac{\exp(x_i'\hat{\beta})}{1+\exp(x_i'\hat{\beta})}=\frac{e^{\hat{\beta}_0+\hat{\beta}_1x_i+\hat{\beta}_2x_{2i}+\cdots+\hat{\beta}_kx_{ik}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1x_i+\hat{\beta}_2x_i+\cdots+\hat{\beta}_kx_{ki}}} \tag{16}$$

3) Logical regression analysis for classification

In the case of a given x, consider y distribution probability of two points

$$\begin{cases}P(y=1\mid x)=F(x,\beta)\\P(y=0\mid x)=1-F(x,\beta)\end{cases} \tag{17}$$

Because $E(y\mid x)=1\times P(y=1\mid x)+0\times P(y=0\mid x)=P(y=1\mid x)$. So we can be $\hat{y}$ interpreted as "$y=1$" probability

$$\hat{y}_i=P(y_i=1\mid x)=S(x_i'\hat{\beta})=\frac{\exp(x_i'\hat{\beta})}{1+\exp(x_i'\hat{\beta})}=\frac{e^{\hat{\beta}_0+\hat{\beta}_1x_i+\hat{\beta}_2x_{2i}+\cdots+\hat{\beta}_kx_{ik}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1x_i+\hat{\beta}_2x_i+\cdots+\hat{\beta}_kx_{ki}}} \tag{18}$$

If $\hat{y}\geq 0.5$, is that its predictions $y=1$; Otherwise think its predictions $y=0$.

We used SPSS software to conduct logistic regression analysis of weathered and unweathered samples of unknown types according to the existing data, and based on the main chemical components generated in Question 2. In the case, we followed [7]. The final classification results are shown in **Table 6**.

4.3.2. Sensitivity Analysis

This problem requires sensitivity analysis of classification results. Based on the results of principal component analysis in question 2, we up-regulated and down-regulated the values of main chemical contents, increased the disturbance rate, and compared the changes in the results before and after. If the result changes

Table 6. Results of question 3.

Type	Cultural relic number	Surface weathering
High Potassium	A1	Weathering free
Lead Barium	A2	Weathering
Lead Barium	A3	Weathering free
Lead Barium	A4	Weathering free
Lead Barium	A5	Weathering
High Potassium	A6	Weathering
High Potassium	A7	Weathering
High Potassium	A8	Weathering free

after a chemical value is adjusted by 10%, it indicates that the sensitivity of the data to the chemical value is high. On the contrary, if there is no difference before and after modification, it indicates that the sensitivity of the data to the chemical value is low.

Specific practices: On the basis of the classification of weathering or not, the main chemical components were increased by 10%, 20%, 30%, and decreased by 10%, 20%, 30% in turn, and the changes of the results before and after were compared.

Similarly, by using SPSS to increase the disturbance by 20% and 30% respectively, it is found that the results are changed, and the accuracy of the results is 100%.

Through the above results, we conclude that increasing and decreasing the content of each main chemical component has no significant effect on the classification result, so the sensitivity of the classification result is not high.

4.4. Establishment and Solution of Problem 4 Model

4.4.1. Establishment and Solution of the First Question Model

This problem requires the analysis of the correlation between the chemical constituents of different types of glass relics samples. According to the problem analysis, the model of grey correlation analysis is selected [2].

The basic idea of grey correlation analysis is to judge whether the relationship is close according to the similarity degree of geometric shapes of sequence curves. The closer the curves are, the greater the correlation degree among corresponding sequences will be; otherwise, the smaller it will be. The calculation steps are as follows. Step P1: Determine the analysis sequence

Determine the reference sequence that reflects the behavior characteristics of the system and the comparative sequence that influences its behavior. Mother a reference sequence (sequence) for $Y=\left\{Y\left(k\right)|k=1,2,\cdots,n\right\}$; Compare sequence (a sequence) for $X_i=\left\{X_i\left(k\right)|k=1,2,\cdots,n\right\},i=1,2,\cdots,m$ [8].

According to the background of the problem, the main raw material of glass is quartz sand, and the main chemical component is silica (SiO_2). Therefore, we choose the content of silica as the reference sequence (parent sequence), and other components as the comparison sequence (sub-sequence), respectively, to conduct the correlation analysis between the chemical components of high-potassium glass and lead-barium glass.

Step 2: Dimensionless variable

Because the data in each factor column in the system may be different in dimension, it is not easy to compare or get the correct conclusion when comparing. Therefore, when grey relational degree analysis is performed, dimensionless data processing is generally required. Since the chemical composition content of each substance in this question is percentage, dimension has been unified, so there is no need to carry out dimensionless treatment.

Step 3: Calculate the correlation coefficient

$x_0\left(k\right)$ and $x_i\left(k\right)$ correlation coefficient:

$$\zeta_i(k)=\frac{\min\limits_i\min\limits_k\Delta_i(k)+\rho\max\limits_i\max\limits_k\Delta_i(k)}{\Delta_i(k)+\max\limits_i\max\limits_k\Delta_i(k)} \tag{19}$$

Which $\Delta_i(k)=\left|y(k)-x_i(k)\right|$.

$\rho\ э\ (0,\infty)$, called distinguish coefficient. ρ is smaller, the greater the resolution.

Step 4: Calculate the correlation degree

Since the correlation coefficient is the correlation degree value between the comparison series and the reference series at each moment (that is, each point in the curve), it has more than one number, and the information is too scattered to facilitate the overall comparison. Therefore, it is necessary to gather the correlation coefficient at each moment into one value, that is, to calculate its average value, which is expressed as the quantity of correlation degree between comparison series and reference series. The correlation degree r_i formula is as follows:

$$r_i=\frac{1}{n}\sum_{k=1}^{n}\zeta_i(k),k=1,2,\cdots,n \tag{20}$$

Step5: Sort the correlation degree

If $r_1 < r_2$, then the reference series y is more similar to the comparison series x_2.

Calculate the average of various correlation coefficients

SPSS software was used to carry out grey correlation analysis on high-potassium glass and lead-barium glass respectively, and the correlation degree was obtained in Table 7.

Table 7. Results of grey correlation analysis.

Lead Barium correlation		High Potassium correlation		Ranking
Evaluation item	Correlation degree	Evaluation item	Correlation degree	
Alumina (Al_2O_3)	0.953	Alumina (Al_2O_3)	0.932	1
Barium oxide (BaO)	0.936	Copper oxide (CuO)	0.931	2
Lead oxide (PbO)	0.934	Phosphorus pentoxide (P_2O_5)	0.917	3
Magnesium oxide (MgO)	0.931	Magnesium oxide (MgO)	0.911	4
Strontium oxide (SrO)	0.924	Potassium oxide (K_2O)	0.907	5
Calcium oxide (CaO)	0.921	Calcium oxide (CaO)	0.903	6
Potassium oxide (K_2O)	0.92	Iron oxide (Fe_2O_3)	0.897	7
Copper oxide (CuO)	0.915	Lead oxide (PbO)	0.873	8
Iron oxide (Fe_2O_3)	0.911	Tin oxide (SnO_2)	0.866	9
Sodium oxide (Na_2O)	0.9	Strontium oxide (SrO)	0.864	10
Phosphorus pentoxide (P_2O_5)	0.896	Barium oxide (BaO)	0.855	11
Tin oxide (SnO_2)	0.884	Sulfur dioxide (SO_2)	0.85	12
Sulfur dioxide (SO_2)	0.882	Sodium oxide (Na_2O)	0.849	13

4.4.2. Establishment and Solution of the Second Question Model

The problem requires a comparison of the differences in chemical composition associations among different classes.

1) Data visualization comparison

First of all, based on the gray correlation degree of the first question, high potassium and lead barium are compared. The image is **Figure 9**.

According to the results of the figure above, it can be found that the correlation degree of most chemical components is not much different, and the differences are mainly reflected in barium oxide, lead oxide, sulfur dioxide and other substances. According to the background of the problem, the content of lead oxide and barium oxide in lead-barium glass is high because lead ore is used as flux. There is high potassium glass with plant ash as the flux which contains high potassium contents. According to the correlation ranking of the first question, in addition to the correlation of alumina content ranked the first, the corresponding chemical content of different categories has also been well verified.

2) Paired sample T test

The paired sample T-test belongs to the double population test, which is to test whether the difference between the average of two samples and the population they represent is significant.

Construction statistics:

$$t = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}} \tag{21}$$

Among them, $i = 1, \cdots, n$, $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ pair sample difference of average of $s_d = \sqrt{\frac{\sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2}{n-1}}$ for matching the sample standard deviation of this difference, n for matching sample.

This statistic follows a t distribution of $n - 1$ degrees of freedom under the null hypothesis that $\mu = \mu_0$ is true.

The above table shows the results of model test, including mean value, standard deviation, T-value, degree of freedom, P-value of significance, etc:

1) To analyze whether the P-value of each group was significant (P < 0.05 or <0.01);

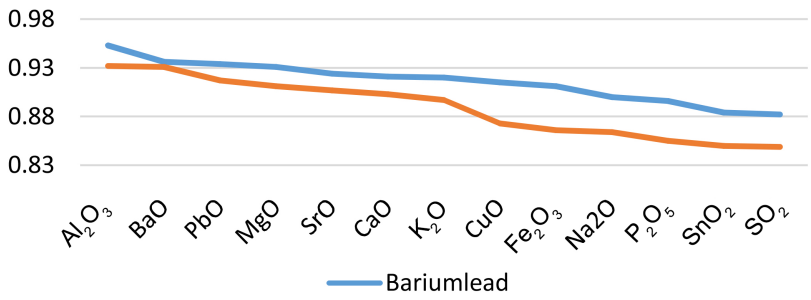

Figure 9. Correlation degree comparison.

2) If it is significant, the null hypothesis is rejected, indicating that there is a difference between the samples of each pairing; otherwise, it means that there is no significant difference between the samples of each pairing.

3) Cohen's d value: the effect size is below 0.20: the effect is too small; 0.20 - 0.50: the effect is small; 0.50 - 0.80: the effect is large; and above 0.80: the effect is large.

The T-test results of paired samples showed that based on the field high potassium paired lead barium, the significance P value was 0.006***, showing significance at the level, rejecting the null hypothesis, so there was a significant difference between the words high potassium paired lead barium. Cohen's d value of the difference range was 0.914, indicating a very large difference range.

5. Conclusions and Prospects

5.1. Advantages of the Model

1) For problem 1, the relationship between indicators can be described more simply and clearly through data visualization, and the chi-square test is used to represent the deviation degree between the actual value and the theoretical value of statistics to reflect the difference, which is convenient and efficient. The weighted average method is used to predict the trend of the prediction problem, and the prediction problem with small number of samples but large variation of indicators is simplified, and the method is relatively novel.

2) For problem 2, we use principal component analysis to synthesize and simplify the preprocessed data, so as to objectively determine the contribution of different chemical components. Meanwhile, combined with the clustering model, the calculation is efficient and has certain expansibility. After that, we built a dichotomous K-means clustering model to improve the reasonable analysis, making the model easier to understand and more efficient to calculate. Finally, through the analysis of the rationality and sensitivity of the model, the advantages of the model can be verified.

3) For problem 3, logistic regression is used to directly consider the types of chemical components without assuming data distribution to avoid problems caused by inaccurate distribution. Then, binary classification is carried out based on the linear probability model, which is easy to understand and can solve the problem well.

4) For problem 4, grey correlation analysis is adopted, which is reliable and reasonable according to the development trend of data without the requirement restriction on sample size; moreover, the paired sample T test was used to objectively analyze the diversity of different types, and the results were more in line with expectations.

5.2. Disadvantages of the Model

1) The system clustering model is greatly affected by abnormal values, and the results may not be ideal and reasonable enough.

2) Features cannot be screened by logistic regression itself, resulting in overfitting and difficult to deal with uneven data.

5.3. Model Improvement

1) When using the system clustering model, the model can be further optimized and innovated, that is, using the K-Means ++ clustering model or combining it with the dichotomy method, so that the classification results can be more reliable and the rationality and sensitivity can be tested better.

2) In case of overfitting in logistic regression, cross-validation can be used to eliminate the influence of chance, and finally an average accuracy rate can be obtained for each model to further improve the accuracy of classification results.

3) For the sensitivity analysis in questions 2 and 3, the analysis result of changing the content of each chemical component as the disturbance term is not very significant, so the disturbance term can be changed to continue the sensitivity analysis. For example, in problem 2, the clustering center of the binary K-means model can be disturbed by 10% or more of the circumference, the Euclidean distance from the coordinates of the original clustering center point can be calculated, and the accuracy of the model can be calculated and its sensitivity can be analyzed by comparing with the real value.

5.4. Model Promotion

Based on data processing and analysis, this model analyzes the relationship between the surface weathering of glass relics and its glass types and patterns, and predicts the chemical composition before weathering as well as the classification rules of different glasses, which can test the rationality sensitivity from the data analysis results. At the same time, the establishment of grey correlation analysis method adopted by this model can effectively analyze the correlation difference of different components, and can also be applied to other glass products, and to predict and infer their chemical components, so as to promote the development of archaeology and other disciplines.

Funding

Funding Open Access funding provided thanks to the university level innovation project of Jiangsu University agreement with Springer Nature.

Conflicts of Interest

The authors declare they have no financial interests.

References

[1] Zhao, Z. (2020) Detection and Analysis of Unearthed Glass Products from the Maicha Warring States Cemetery in Liye, Hunan. *Hunan Archaeology Bulletin*, 288-301. https://kns.cnki.net/kcms2/article/abstract?v=vCcGnC-OR21aoCndRK0RaUhQrD

O2EFHYDnkZYpPjUmCFKdDh6ZClhUnOpV1KwiVLpDf5CDPH2OiaRlsRzRLCiff9mB6q_E-X1jMmU6ya8U1PMpky8HBW58-LAuEnVsIbbOUdAgtis3w=&uniplatform=NZKPT&language=CHS

[2] Chen, S. and Hou, Z.L. (2019) Analyze the Ancient Silk Road and Ancient Chinese Glass into the Chest. *National Exhibition of China*, **5**, 87-88.

[3] Xie, Y., Ren, J., Huang, H. and Zhang, X. (2020) Grid Independence Analysis of Computational Fluid Dynamics Based on Chi Square Test. *Science Technology and Engineering*, **20**, 123-127.

[4] Wu, P. (2022) Application of Weighted Average Method in Calculation of Aging Time of Liupao Tea. *China Tea*, **44**, 50-53.

[5] Li, Y., Wang, D., Qi, L., Zhang, Q., Niu, R., Hou, D. and Shi, J. (2022) Comprehensive Evaluation on Yield and Quality of Cucumber under Different Nitrogen Application Rates Based on Principal Component Analysis and Cluster Analysis. *Journal of Northeast Agricultural Sciences*, **47**, 110-114+149. (in Chinese)

[6] Xu, X.-N. (2022) Research on Electronic Archives Management Technology Based on Binary Classification Model. *Journal of Microcomputer Applications*, **38**, 159-163.

[7] Zhang, Z., Li, X., Lu, X., Xu, J. and Wei, Y. (2018) Prediction of Soil Cu Pollution Risk Based on Binary Logistic Regression Model. *Chinese Journal of Soil Science*, **21**, 1418-1426.

[8] Yang, J., Ren, H., Zhao, X., Shen, J., Wang, A. and Wang, Z. (2021) Application of Grey Correlation Analysis and Cluster Analysis in Maize Variety Comprehensive Evaluation. *Seed*, **40**, 107-115. (in Chinese)

Appendix: Supplementary Materials

Supplementary Materials presents the design, calculation process and methods of software and programs.