Analysis and identification of glass products based on fisher linear discriminant analysis and logistic regression for dichotomous classification

Weifeng Ni *, Tianqiu Pu, Shubing Liu

Hehai University, Changzhou, China

* Corresponding Author Email: niweifeng202212@163.com

Abstract. This paper focuses on establishing multiple linear regression models, correlation analysis, fisher linear discriminant analysis, logistic regression to deal with dichotomous classification, sensitivity analysis, systematic clustering and gray correlation analysis to solve the problem of compositional analysis and identification of ancient glass products. The subclasses of each category of glass are divided, and the division results are derived, so that the rationality and sensitivity of the classification results can be analyzed, and finally the classification laws of high potassium glass and lead-barium glass are derived. Analysis The chemical composition of unknown glass artifacts, predict their glass types, and perform sensitivity analysis on the predicted results.

Keywords: Glass product composition; Correlation analysis; fisher linear discriminant analysis.

1. Introduction

The main raw material of glass is quartz sand, and the main chemical composition is silicon dioxide. In order to reduce its melting temperature, often add fluxes in the glass, different types of fluxes added to different glasses, so the main chemical composition in different glasses are different [1]. For example, lead barium glass with the addition of lead ore as a flux, its high content of lead oxide and barium oxide; potassium glass is fired with a high content of potassium substances such as grass wood ash as a flux [2]. Ancient glass is highly susceptible to weathering by the burial environment, and weathering leads to changes in its composition ratio, while affecting the color and decoration of the artifacts, thus affecting the correct judgment of its category [3].

Archaeologists, based on the chemical composition of artifact samples and other testing techniques, now classify a group of ancient glass objects into two types: high potassium glass and lead-barium glass.

This paper will derive the results of the division, so as to analyze the rationality and sensitivity of the classification results, and finally derive the classification rules for high potassium glass and lead-barium glass. The chemical composition of the glass artifacts will also be analyzed to predict their glass types, and sensitivity analysis will be conducted on the predicted results.

2. Model assumptions and notation

2.1. Assumptions [4]

1. Assume that the outliers and tolerance rate of weathering point detection data are less than 10%.
2. Assume that the chemical composition data of the artifacts conform to the gray system model.
3. Artifact sampling points of the proportion of each component of the cumulative sum between 85% - 105% of the data is considered valid data.
4. Assume that the decoration, color and glass type and glass artifact surface weathering into a linear relationship.

2.2. Notations

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

Table 1. Notations

Symbols	Description
A_i	Three types of decoration
B_i	Two types of glass
C_i	Eight colors of the artifacts
z_i	Elements within set variable 1
U	Typical variable 1
V	Typical variable 2
Y	Dummy variables
x_i	Each chemical composition in the artifact
c_k	The kth class
u_k	Position of the center of gravity of the kth class
J	Coefficient of aggregation
K	Number of classes of clustering
y_i	Gate value

3. Model construction and solving

3.1. Fisher linear discriminant analysis and logistic regression for dichotomous identification of unknown artifact types

DA (Linear Discriminant Analysis) is a classical linear discriminant method, also known as Fisher Discriminant Analysis, which first uses fisher's linear discriminant analysis to select the appropriate chemical components for each class to subclass. The idea of this method is relatively simple: given a training set of samples, we try to project the samples onto a one-dimensional line, so that the projection points of similar samples are as close and dense as possible, and the projection points of dissimilar samples are as far away as possible [5]. All sample data were imported into SPSS first, and then weathered and unweathered dummy variables were created using 0-1 logistic regression [6].

From the data, it can be seen that the prediction rate is 100%. From the table of the canonical discriminant function coefficients to know the classification relationship of various chemical components on the glass. This gives

Y=0.043*X1+0.32*X2-0.188*X3+0.023*X4+0.204*X5+0.209*X6+0.098*X7-0.182*X8+0.118*X9+0.216*X10+0.095*X11-0.608*X12+0.015*X13-0.102* X14-6.377 (Y>0) means the glass type is lead barium, Y<0 means the glass type is high potassium, X1 represents the content of silica, X2 represents the content of sodium oxide, X3 represents the content of potassium oxide, X4 represents the content of calcium oxide, X5 represents the content of magnesium oxide, X6 represents the content of aluminum oxide, X7 represents the content of iron oxide, X8 represents the content of copper oxide, X9 represents the content of for lead oxide, X10 for barium oxide, X11 for phosphorus pentoxide, X12 for strontium oxide, X13 for tin oxide, and X14 for sulfur dioxide).

Then the results and then logistic regression processing of the second classification of the unknown type of cultural relics for classification test, first the cultural relics sampling points for random sampling, of which eighty percent as the experimental group, of which twenty percent as the detection group, so selected 21, 22, 27, 01, 03 parts 1, 26 severely weathered, 28 unweathered points, 51 parts 1, 54 severely weathered and 57 of these ten data as the detection group, and the results obtained by dichotomizing the experimental groups with logistic regression can be obtained in Table 2.

Table 2. Variables in the equation

		B Standard Error	Wald	Degree of freedom	Significance
Silicon dioxide(SiO2)	.591	7852.766	.000	1	1.000
Sodium oxide(Na2O)	1.075	13199.425	.000	1	1.000
Potassium oxide(K2O)	-3.497	9846.718	.000	1	1.000
Calcium oxide(CaO)	1.836	27413.467	.000	1	1.000
Magnesium oxide(MgO)	8.913	13098.736	.000	1	.999
Aluminum oxide(Al2O3)	1.402	7019.304	.000	1	1.000
Iron oxide(Fe2O3)	2.045	9779.827	.000	1	1.000
Copper oxide(CuO)	-4.332	11736.838	.000	1	1.000
Lead oxide(PbO)	1.294	7786.303	.000	1	1.000
Barium oxide(BaO)	3.451	10892.708	.000	1	1.000
Phosphorus pentoxide(P2O5)	-.101	12167.169	.000	1	1.000
Strontium oxide(SrO)	-21.304	58611.806	.000	1	1.000
Tin oxide (SnO2)	6.293	118009.828	.000	1	1.000
Sulfur dioxide(SO2)	2.852	8266.939	.000	1	1.000
Constants	-70.061	780490.598	.000	1	1.000

From the results, it can be seen that the prediction success rate of 100 percent is considered to be due to the overfitting phenomenon of the data due to too many indicators of the dependent variable (chemical composition), and further relationships of the generated variables were obtained[7].

$$y_i = \frac{e^{-70.061+0.591SiO_2+1.075Na_2O-3.497K_2O+1.836CaO+8.913MgO+1.402Al_2O_3+2.04Fe_2O_3-4.33CuO+1.294PbO+3.45BaO-0.101P_2O_5-21.304SrO+6.293SnO_2+2.852SO_2}}{1+e^{-70.061+0.591SiO_2+1.075Na_2O-3.497K_2O+1.836CaO+8.913MgO+1.402Al_2O_3+2.04Fe_2O_3-4.33CuO+1.294PbO+3.45BaO-0.101P_2O_5-21.304SrO+6.293SnO_2+2.852SO_2}} \quad (1)$$

If y>=0.5 the results are judged to be of the lead-barium type and vice versa for the high potassium type. Using y to predict the detection group, it was statistically found that the correct prediction rate of the detection group reached 80%, indicating that the overfitting situation was not significant and basically satisfied the classification criteria.

3.2. fisher linear discriminant analysis model optimization

Because the prediction results presented 100 percent success, it was intended that the model would be further investigated [8]. Firstly, random sampling was taken, and eighty percent of the samples were first taken as the experimental group, and twenty percent of the samples were taken as the test group. So first ten experimental groups were randomly sampled, namely, 06 site 1, 06 site 2, 07, 09, 16, 25 unweathered sites, 30 site 1, 34, 37, 50. Then Fisher discriminant analysis was performed on the remaining data.

After data processing, the predicted experimental composition power remained at 100 percent, and the relationship between the classification of various chemical compositions on the glass was known from the canonical discriminant function coefficient table. This gives Y=0.051*X1+0.338*X2-0.120*X3-0.222*X4+0.466*X5+0.288*X6+0.447*X7-0.158*X8+0.140*X9+0.238*X10+0.146*X11-0.966*X12-0.408*X13-0.087 *X14-9.071(Y>0 means the glass type is lead barium, Y<0 means the glass type is high potassium, X1 represents the

content of silica, X2 represents the content of sodium oxide, X3 represents the content of potassium oxide, X4 represents the content of calcium oxide, X5 represents the content of magnesium oxide, X6 represents the content of aluminum oxide, X7 represents the content of iron oxide, X8 represents the content of copper oxide, X9 represents the content of for lead oxide, X10 for barium oxide, X11 for phosphorus pentoxide, X12 for strontium oxide, X13 for tin oxide, and X14 for sulfur dioxide). The coefficients of the model before optimization changed less than the magnitude, and it can be seen that the original model was established slightly more reasonably. However, when the data from the experimental group were predicted, the following phenomena appeared as in Table 3.

Table 3. Data Prediction Phenomenon

	Original value	Predicted value		
06 Part 1	0	1	0.00673	0.99327
06 Part 2	0	1	0.08358	0.91642
7	0	0	1	0
9	0	0	1	0
16	0	0	1	0
25 Unweathered spots	1	1	0	1
30 Part 1	1	1	0	1
34	1	1	0	1
37	1	1	0.21626	0.78374
50	1	1	0	1
			Correct Rate	80%

It is known that the algorithm is flawed, for example, part 06, where the original data for part 06 is a high potassium class, but the prediction model predicts its species to be lead-barium. The guess is that because of the influence of outliers, the above two equations appear slightly different, but the success rate reaches eighty percent and the surface prediction equation barely passes the test.

3.3. Sensitivity analysis of fisher's linear discriminant analysis model

This is followed by sensitivity analysis, which is a method of analyzing how sensitive the state or input data of a system or model is to the parameters of the system. The usual procedure is to change the value of an important parameter in the model, while controlling for other parameters, and then observe the change in the model results. First, fix the value of the chemical component as the average of all data, and then change the value of one chemical component at a time by a factor of 0.9-1.1 (continuous variation) and keep the content of other chemical components constant [9,10]. The change in Y-value was observed and the sensitivity of each variable to Y-value was obtained from the trend of change.

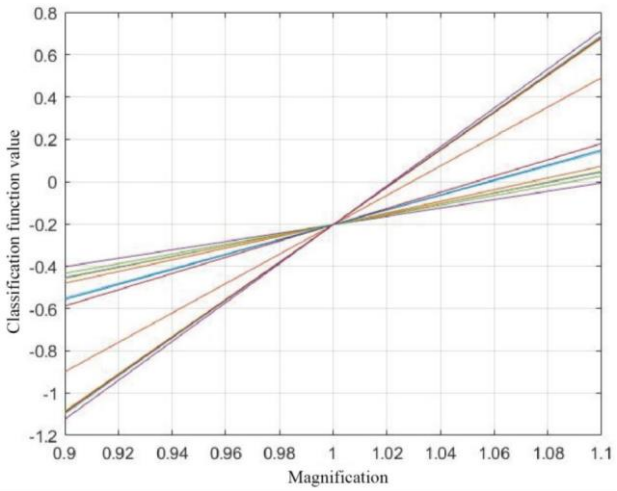


Figure 1. Variation of Y-value

Sensitivity analysis was performed using Matlab by drawing a line graph for each change in Y value and presenting all the cases in the same graph as in Figure 1.

From the graphs, it can be seen that the change in the multiplicity of each variable produces a linear change in the Y value, and as the multiplicity increases, the Y value also shows an increasing trend and a significant change. This shows that the sensitivity of the model is high.

3.4. Test of clustering models on subclassification results

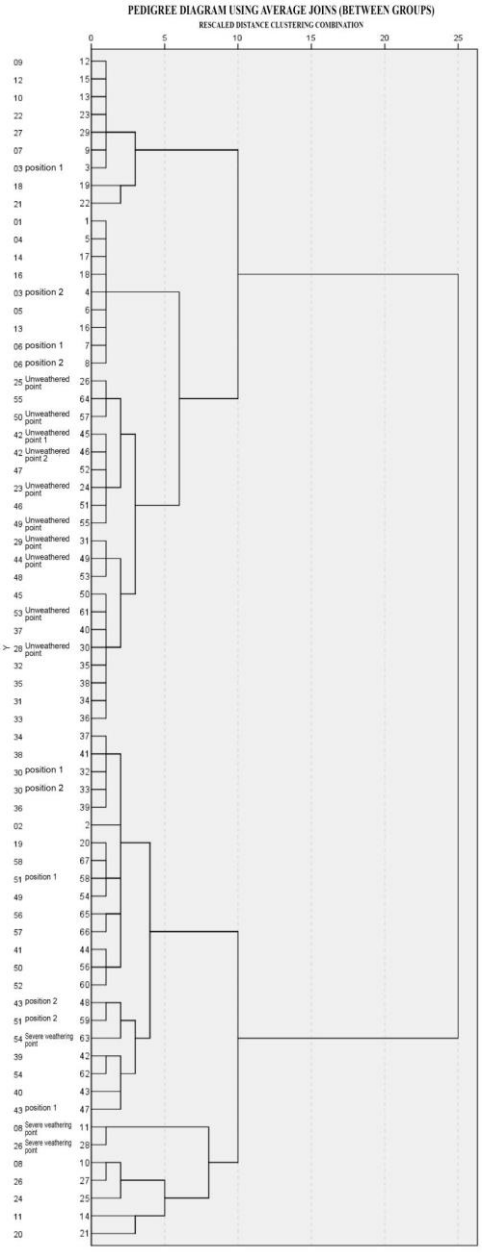


Figure 2. Genealogy Chart

Because the classification relationships of various chemical components to glass have been analyzed, the initial clustering of the artifact sampling points is performed first. Because the K-meansS algorithm has the following disadvantages: (1) it requires the user to give the number of clusters K to be generated in advance (2) it is sensitive to initial values (3) it is sensitive to isolated point data. Then, the initial clustering is performed using the systematic clustering method

The degree of distortion varies the most when the value of K goes from 1 to 2. After exceeding 2, the change in the degree of distortion decreases significantly. Thus the elbow is K=2, so it is reasonable to classify the data collected at the heritage sampling sites into two categories (corresponding to the two types of high potassium and lead-barium, respectively).

From the elbow rule, the optimal number of clusters can be roughly estimated by the graph. The sum of the distortion of each class: The sum of the distortion of each class: The distortion of each class is equal to the squared sum of the distance between the center of gravity of the class and the position of its internal members; Suppose a total of n samples are divided into K classes (K≤n-1, i.e., at least two elements in one class) denote the kth class (k=1,2,. ,K), and the position of the center of gravity of the class is denoted as Then the degree of distortion of the kth class is: $\sum_{i \in c_k} \left| x_i - u_k \right|^2$ (the meaning of the absolute value notation here is the distance, which can be considered as a generalized notation). Define the total degree of aberration for all classes: $J = \sum_{k=1}^{K} \sum_{i \in c_k} \left| x_i - u_k \right|^2$. In some multivariate statistical textbooks, J is also referred to as the coefficient of aggregation.

Further, the value of K is determined as 2, generating a spectral plot as in Figure 2.

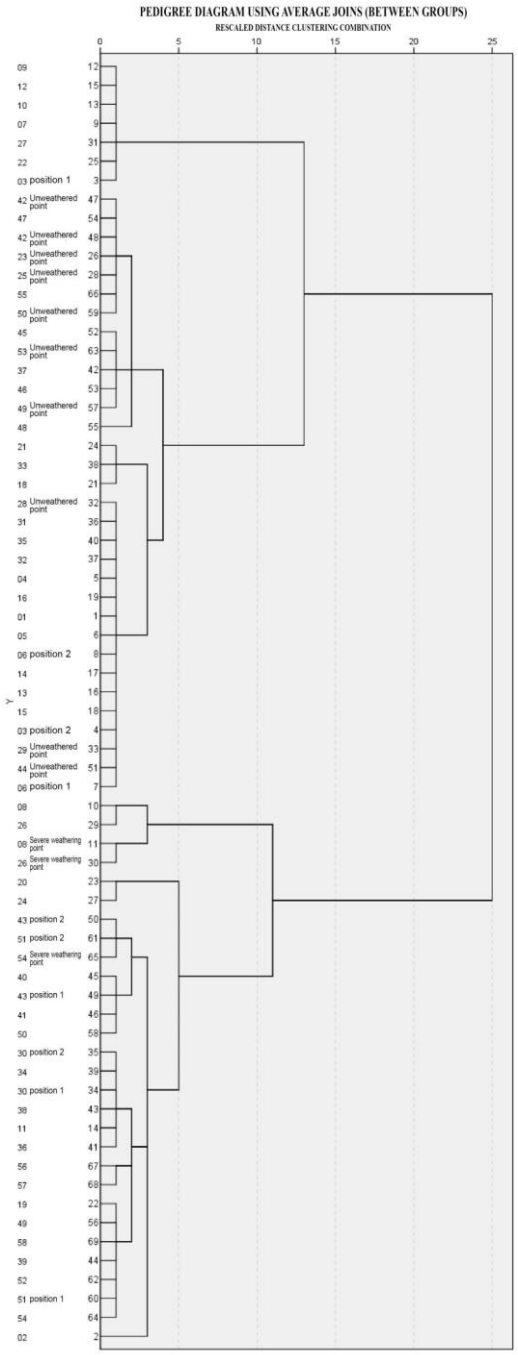


Figure 3. Genealogy Chart

The comparison with the original data of the heritage sampling points yielded a correct rate of systematic clustering up to 70.15 percent, which basically satisfied the clustering criteria.

Then the typical correlation function coefficient to take the absolute value of the chemical composition (Na2O, MgO, Al2O3, BaO, SrO) and the main components (SiO2) as the basis for sub-classification. Preliminary clustering was performed using the systematic clustering method

The K value has the greatest variation in distortion from 1 to 4, and the variation in the degree of distortion decreases significantly after exceeding 4. Because of the need to classify the artifact sampling sites into high potassium and lead-barium categories, but as indicated by the aggregation coefficient line graph, a K value of 4 is the best value, probably because of the significant influence of weathering and unweathering on the clustering results. However, the K value result was still determined to be 2 as the cluster for systematic clustering.

The generated spectral plots are as Figure 3.

Comparing the classification results with the original data, it was statistically found that the correct clustering rate reached 69.56% with the correct clustering rate of all the chemical components was not much different, so the expected results were basically achieved, indicating that the subclassification results were basically qualified.

3.5. Model-based application and in-depth exploration of boundary values

Because the previous paper used fisher linear discriminant analysis and logistic regression to process the dichotomous classification to identify the type of unknown artifacts to derive classification models, the original models were applied to analyze the chemical composition of the unknown category of glass artifacts to identify the type to which they belong.

From the previous section, the following two classification models were derived using fisher:
Y1=0.043*X1+0.32*X2-0.188*X3+0.023*X4+0.204*X5+0.209*X6+0.098*X7-
0.182*X8+0.118*X9+0.216*X10+0.095*X11- 0.608*X12+0.015*X13-0.102*X14-6.377

Y2=0.051*X1+0.338*X2-0.120*X3-0.222*X4+0.466*X5+0.288*X6+0.447*X7-
0.158*X8+0.140*X9+0.238*X10+0.146*X11-0.966*X12-0.408*X13-0.087* X14-9.071

It is concluded that A2, A3, A4, and A5 are lead-barium types, A6, A7, and A8 are high potassium types, and A1 is a boundary value to be determined. Therefore, further analysis of the chemical content contained in different types of glass should be carried out to investigate the attribution of A1 as the boundary value.

Because the optimization model and the known data have differences in the boundary values, the statistics of Y1 and Y2 are greater than 2 for lead-barium, less than -1 for high potassium, and fluctuate between high potassium and lead-barium when greater than -1 and less than 2. Therefore, further data analysis should be performed several times with the Monte Carlo model, picking out 80% of the chemical composition of the unknown category of glass artifacts as the experimental group, and 20% of the samples as the test group for several iterations. The final results were obtained: there was an 89.9% probability of high potassium and a 10.1% probability of lead-barium. This led to the conclusion that the glass type of A1 should be high potassium.

After classifying the eight unknown artifact types, a sensitivity analysis was performed on the results. The data of all chemical compositions of these eight artifacts were averaged and the number3Code file code was used to process these eight averages. The 8 unknown variables were set as the average of these 8 values, and then the values were varied continuously from 0.9 times to 1.1 times of the original values. The change of the dummy variable Y value is obtained, and it is observed that Y value has a strong linear relationship with each chemical composition. The sensitivity of the classification model is high as the Y-value changes positively with the change of the chemical composition.

4. Conclusion

In order to analyze the classification pattern of high potassium and lead barium, this paper first used fisher linear discriminant analysis to obtain the canonical discriminant function coefficients, and established a linear model based on the canonical discriminant function coefficients. Then, the classification results were tested by logistic regression dichotomy, and the sampling points of the artifacts were randomly sampled, of which 80% were taken as the experimental group and the remaining 20% were taken as the test group for verification. Then the fisher linear discriminant analysis model was optimized and sensitivity analysis was performed. Finally, the results of subclassification were tested and justified by systematic clustering.

In order to identify the types of unknown types of glass artifacts that belong to them, two classification models derived from fisher linear discriminant analysis were applied to classify eight unknown types of glass, and one of the glass types was explored in depth with boundary values, and Monte Carlo simulation was used several times for simulation analysis, and the final conclusion was finally reached.

References

[1] He Xiaoqun. Cluster analysis-systematic clustering method. Multivariate Statistical Analysis.2019,6.

[2] Yang J-B, Huang Y-S. Value of logistic regression and Fisher's linear discriminant analysis model in the differential diagnosis of benign and malignant ovarian tumors.2018.06.025.

[3] Hou Z-X,Zhe-Ri, Zhang C-J. Study on the weathering mechanism of sandstone quality cultural relics a case study of Yungang Grottoes. School of Engineering and Technology, China University of Geosciences (Beijing), Beijing 100083.2019.0313

[4] Lin Zhaoyu, Li Jinzhan, Chen Sengsen. Application of Poisson distribution chi-square test based on Excel function and charting tools. Fujian Taoyuan State Forestry, Dada, Fujian 366101.2019.21.071-02

[5] Nana Zhou and Zhijian Rao. Analysis of factors influencing grain yield in Yunnan Province based on gray correlation analysis.2017.08025.

[6] Cui Shangshu, Yang Lian, Li Ting. Research on air quality prediction algorithm based on principal component analysis and multiple linear regression.2022-06-0097-04.

[7] Liu QJ. Tumor subtype classification based on BP network sensitivity analysis for feature gene selection.2008.1005

[8] Zhang Hanxia. Micro expression recognition model based on multi classification logistic regression. 2022.10.001

[9] Li Xueyang, Shao Xigao. A study on the influencing factors of the quality of college students based on multiple linear regression and Lasso regression 2022,38(04)

[10] Qin Qiusheng. Based on multiple linear regression analysis, the prediction model of the number of college students is constructed. 2022, (22)