Development and application of comprehensive evaluation model based on correlation analysis algorithm, multiple logistic regression and clustering algorithm - Taking glass composition analysis and identification as an example

Sichang Yang [1, *, #], Luyao Dai [2, #], Changwen Lu [1, #]

[1] School of IoT Engineering, Jiangnan University, Jiangsu, China, 214122

[2] School of food science and Engineering, Jiangnan University, Jiangsu, China, 214122

* Corresponding author: 2089196948@qq.com

[#]These authors contributed equally.

Abstract. Take glass as an example. In ancient times, when firing glass, in addition to the raw material quartz sand, chemical components such as stabilizer and flux need to be added. During the storage process of glass, the internal elements will exchange with the environmental elements in a large amount, resulting in the weathering of the glass surface. In this paper, the composition of ancient glass is analyzed and identified by establishing correlation analysis model, multiple logistic regression and cluster model. The results show that: (1) There is a high correlation between glass type and surface weathering. Lead barium glass is easy to be weathered, and high potassium glass is not easy to be weathered. (2) High-potassium glass and lead-barium glass have the highest classification accuracy when they are divided into two sub-categories. Therefore, glass can be divided into four categories: high-potassium I, high-potassium II, lead-barium I, and lead-barium II.

Keywords: Glass, Correlation analysis model, Multiple logistic regression, Clustering model.

1. Introduction

The main raw material of glass is quartz sand, and the main chemical composition is silicon dioxide (SiO_2). Due to the high melting point of pure quartz sand, in order to reduce the melting temperature, it is necessary to add flux during smelting. The commonly used fluxes in ancient times include plant ash, natural caustic soda, nitrate and lead ore, and limestone is added as stabilizer. After calcination, limestone is converted into calcium oxide (CaO). The main chemical composition of the added fluxes is also different. At the same time, ancient glass is easily weathered by the influence of burial environment. In the process of weathering, the internal elements exchange with the environmental elements in a large amount, resulting in the change of their composition proportion, thus affecting the correct judgment of their categories. This has brought misery to archaeological work.

According to the chemical composition and other detection methods of these cultural relics, archaeologists have divided them into two types: high-potassium glass and lead-barium glass. The characteristics of these data are compositional, that is, the cumulative sum of the proportion of each component should be 100%, but the cumulative sum of the proportion of its components may not be 100% due to detection methods and other reasons. In this question, the data of the cumulative sum of component proportions between 85% and 105% is considered as valid data.

Zhang [1] introduced seven simulation methods of glass composition properties, including addition method, phase diagram method, Priven method, topological binding theory, molecular dynamics simulation, machine learning and mathematical statistics simulation, and summarized the main theoretical basis, simulation process and application status of each simulation method. Zhou [2] prepared the standard material for laser glass composition analysis by high-temperature melting method. The results show that the developed reference material has good uniformity and stability. He [3] took the composition of ordinary Na_2O-CaO-SiO_2 glass as an example, formulated the quantitative index of material property evaluation, and explored the influence of relevant components on material property index under different displacement adjustment modes.

However, there is no relevant research involving the identification of the composition of ancient glass. Based on this, this paper proposes a comprehensive evaluation model of glass composition analysis and identification to provide guidance for archaeological research.

2. Whether there is statistical law of weathering chemical composition content on the surface

This section analyzes the relationship between glass surface weathering and type, color and decoration, and can use the chi-square test to test its correlation and calculate the correlation coefficient [4-6].

There are only "weathering" and "no weathering" on the glass surface; The glass decoration can be divided into three types: A, B, and C. There is no difference between high and low levels, and it is classified data with parallel relationship; There are two types of glass: high-potassium glass and lead-barium glass, which are also classified data with parallel relationship; There are eight kinds of glass colors with a gradual trend. In order to simplify the model, convert them into numerical data by color rendering wavelength and sort them by wavelength size.

The joint tables of surface weathering and glass type, decoration and color are listed respectively, as shown in Table 1, Table 2 and Table 3:

Table 1. List of glass types and surface weathering

Type/surface weathering	Weathering	No weathering	Sum
Lead barium glass	28	12	40
High potassium glass	6	12	18
Sum	34	24	58

Table 2. List of glass decoration and surface weathering

Type/surface weathering	Weathering	No weathering	Sum
A	11	11	22
B	6	0	6
C	17	13	30
Sum	34	24	58

Table 3. List of glass color and surface weathering

Type/surface weathering	Weathering	No weathering	Sum
Black	2	0	2
Blue-green	9	6	15
Green	0	1	1
Light blue	12	8	20
Light green	1	2	3
Deongaree	0	2	2
Dark green	4	3	7
Purple	2	2	4
Sum	30	24	54

Assuming that surface weathering is independent of glass type, chi-square statistics are calculated:

$$\chi^2 = \sum_{i}^{2}\sum_{j}^{2} \frac{n_{ij} - \frac{n_i * n_j}{n}}{\frac{n_i * n_j}{n}} \tag{1}$$

n_{ij} is the statistical value in the table, n_i or n_j is the value in the total item, i=1,2; j=1,2; n=58.

During the weathering process, the proportion of silicon dioxide and iron oxide decreases, while the proportion of other components increases. According to the literature, the internal elements of lead-barium glass exchange with the environmental elements in a large amount during the weathering process, the S_i element gradually loses, and the Pb element and the Ba element generate $PbCO_3$ and $BaSO_4$ respectively and accumulate in the outer layer of the weathering surface.

During the weathering process, the proportion of silicon dioxide in the high-potassium glass increases, and the proportion of other components decreases, and the loss of potassium oxide is serious. Secondly, predict the chemical composition content of the weathering point before weathering.

Because there are many kinds of chemical components to be predicted and there is a correlation between the components, the regression prediction method is selected, and the weathering condition is taken as the dependent variable. However, weathering is classified as a variable, which violates the default condition of linear relationship between dependent variable and independent variable in linear regression analysis, so we turn to look for the relationship between probability and independent variable when the dependent variable takes a certain classification value, namely:

$$p = p(x) \tag{2}$$

So the logical regression algorithm that meets the conditions is selected to build the model. Due to the difference of main chemical composition, the high-potassium glass and lead-barium glass are classified and discussed.

$$\log it(p) = \ln(\frac{p_m}{1-p_m}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots \beta_k x_k \tag{3}$$

Where m is the dependent variable category code, k is the number of independent variable categories, and pm is the probability that the dependent variable is of category m, β is the regression coefficient, which can reflect the influence of the corresponding independent variable on the dependent variable.

For the classification results of M dependent variables, M-1 regression models are trained. If M is selected as the main category, then:

$$\ln \frac{p(Y_i = 1)}{p(Y_i = M)} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki} \tag{4}$$

$$\ln \frac{p(Y_i = 2)}{p(Y_i = M)} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki} \tag{5}$$

……

$$\ln \frac{p(Y_i = M\text{-}1)}{p(Y_i = M)} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki} \tag{6}$$

The left and right of the formula are indexed to obtain:

$$p(Y_i = m) = p(Y_i = M)\exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki}) \tag{7}$$

Since the final probability p is added to 1, we can get:

$$p(Y_i = M) = 1 - \sum_{m=1}^{M} p(Y_i = M) \exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki}) \tag{8}$$

$$p(Y_i = M) = \frac{1}{1 + \exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki})} \tag{9}$$

Bring formula (5) into formula (3) to get:

$$p(Y_i = m) = \frac{\exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki})}{1 + \exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki})} \tag{10}$$

For regression coefficient β_k. The maximum likelihood method is used for estimation. If you want to use the regression coefficient to reflect the impact of the corresponding independent variable on the dependent variable, you need to conduct isometric processing on the data. Normalization is adopted here:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{11}$$

Set:

$$p(Y = m_1) = h(x_1) \tag{12}$$

$$p(Y = m_2) = h(x_2) \quad \ldots\ldots \tag{13}$$

$$p(Y = M) = 1 - h(x_i) \tag{14}$$

Then the likelihood function is:

$$\prod_{j=1}^{N} [h(x_{i\,j})^{y_i}][1 - h(x_{ij})]^{1-y_i} \tag{15}$$

Take the logarithm to get:

$$L(\beta) = \sum_{j}^{N} (y_j(\beta^{\mathrm{T}} x_{ij}) - \log(1 + \exp(\beta^{\mathrm{T}} x_{ij}))) \tag{16}$$

For lead-barium glass, the dependent variable is silicon dioxide content, and the silicon dioxide content of the weathering lead-barium glass is divided into three categories: the content in the interval (30, 49] is category 1, the content in the interval (49, 61] is category 2, and the content in the interval (62, 76] is category 3. The content of lead oxide, barium oxide, copper oxide, and aluminum oxide is calculated by using Matlab β The absolute values are - 96, - 90, 88 and 87, respectively. Take the logistic regression model and solve it with Matlab to get the prediction table.

Due to the high chemical composition of lead oxide and barium oxide, the chemical content data of lead oxide and barium oxide are divided into two parts, so that the lead oxide with the content of

(15.5,22.5) is Class 1, (22.5,40) is Class 2, the barium oxide with the content of less than 20 is Class 1, and the barium oxide with the content of more than 20 is Class 2. The following prediction table (Table 4) can be obtained:

Table 4. Prediction Table

PbO	BaO	SiO_2	MgO	CaO	Calculate
1	2	1	0.00	0.00	100.00%
2	2	1	0.00	0.10	100.00%
2	1	2	0.00	0.25	100.00%
2	1	1	0.31	0.94	100.00%
2	1	2	0.37	0.19	100.00%
2	1	1	0.59	1.00	100.00%
2	1	2	1.00	0.00	100.00%
1	1	3	0.6	0.14	85.40%
1	1	3	0.44	0.19	74.60%
1	1	3	0.00	0.08	68.90%
1	1	3	0.00	0.10	66.00%
1	1	3	0.53	0.36	54.60%
1	1	3	0.00	0.20	50.50%
1	1	2	0.00	0.20	49.50%
1	1	2	0.53	0.36	45.40%
1	1	2	0.00	0.10	34.00%
1	1	2	0.00	0.08	31.10%
1	1	2	0.44	0.19	25.40%
1	1	2	0.60	0.14	14.60%

The accuracy of the predictable part is close to 100%, but the prediction range is not comprehensive. Later, other chemical components can be added or calcium oxide can also be processed in sections.

3. Sub-classification method of high-potassium glass and lead-barium glass

The cluster classification method is used to classify the chemical composition of each sampling point. If the glass can be successfully divided into lead barium glass and high potassium glass according to the chemical composition, and the error is small compared with the actual situation, the classification rule can be obtained [7].

Clustering is carried out based on the similarity degree of chemical component content, and the similarity degree is reflected by the quantitative method. Let the cultural relics to be classified have p variables, so each sampling point can be regarded as a point in the p-dimensional space, and the similarity between the sample points can be measured by distance.

If the content of 14 kinds of chemical components is taken as a variable, the variables with weak connection with the classification results may be taken into consideration, which will increase the difficulty of solving and will not benefit the accuracy of solving. Therefore, it is necessary to screen the variables and leave the chemical components with strong correlation with the glass type for analysis. Because the glass type has a significant correlation with surface weathering, the variables that have a significant correlation with weathering are selected, and the correlation between these variables and glass type is not bad.

The Spielberg correlation coefficient between the content of 14 chemical components and weathering is calculated. The correlation coefficient greater than 0.8 is highly correlated, the correlation coefficient in the interval (0.5,0.8) is moderately correlated, the correlation coefficient in the interval (0.3,0.5) is low correlated, and the correlation coefficient less than 0.3 is irrelevant. Thus, five chemical components with the strongest correlation with weathering are selected: silicon dioxide, potassium oxide, lead oxide, barium oxide and strontium oxide. Since the chemical properties of lead

oxide and barium oxide are similar, they are combined into one variable. After the above work, the cluster space dimension p is successfully reduced to 4. Ω is the set of sampling points, dij is the distance between two sampling points, which meets the following conditions:

$$d_{ij} \geq 0, i, j E \Omega \tag{17}$$

$$d_{ij} = 0, i = j \tag{18}$$

$$d_{ij} = d_{ji}, i, j E \Omega \tag{19}$$

$$d_{ij} \leq d_{ik} + d_{kj}, i, j, k E \Omega \tag{20}$$

d_{ij} satisfies positive qualitative, symmetric and triangular inequalities. Min's distance commonly used in cluster analysis:

$$d_{(q)_{ij}} = \left[\sum_{k=1}^{p} |x_k - y_k|^q \right]^{\frac{1}{q}}, q > 0 \tag{21}$$

Euclidean distance has the widest applicability among Mint distances, with a q value of 2, that is: Shortest distance method:

$$D(G_1, G_2) = \min_{\substack{xi=G1 \\ yi=G2}} \{d(x_i, y_j)\} \tag{22}$$

Longest distance method:

$$D(G_1, G_2) = \max_{\substack{xi=G1 \\ yi=G2}} \{d(x_i, y_j)\} \tag{23}$$

Barycenter method:

$$D(G_1, G_2) = d(\bar{x}_i, \bar{y}_j) \tag{24}$$

Class average method:

$$D(G_1, G_2) = \frac{1}{n_1 n_2} \sum_{x_i \in G_1} \sum_{x_j \in G_2} d(x_i, y_j) \tag{25}$$

When the distance value is about 22, it will be divided into two categories. According to the mark, the two categories are lead-barium glass and high-potassium glass. The cluster centers are as follows (Table 5):

Table 5. Final cluster center of lead barium glass and high potassium glass

Final cluster center		
	Cluster	
	High potassium glass	Lead barium glass
SiO2	2.94	1.33
K2O	1.42	0.04
SrO	0.03	0.5
BaO, PbO	0.03	2.3

Considering the correlation between glass type and weathering, the sampling points are divided into weathered points and unweathered points for clustering again. The strongly correlated chemical composition data were screened out by Spielberg correlation coefficient. Silica, potassium oxide, calcium oxide, magnesium oxide, aluminum oxide, lead oxide and barium oxide were extracted without weathering; Silica, lead oxide, barium oxide and strontium oxide are obtained by weathering.

The cluster results of non-weathering glass types are obtained (Table 6):

Table 6. Clustering results of non-weathered glass types

Final cluster center		
	Cluster	
	High potassium glass	Lead barium glass
SiO_2	64.81	41.1
Na_2O	0.38	1.47
K_2O	0.14	0.34
CaO	0.69	1.37
MgO	0.61	0.27
Al_2O_3	3.11	3.13
CuO	0.78	2.95
Fe_2O_3	0.82	0.79
PbO	19	27.05
BaO	6.98	15.46
P_2O_5	0.57	1.41
SrO	0.18	0.43

The clustering results fit well with the actual situation. on the contrary, for the chemical composition content data of the weathering group, after many parameter changes and weight adjustments, it is still unable to obtain the clustering results with high accuracy. It can be seen that weathering will increase the error of clustering. In this regard, after the glass types are classified, the two types of glass are divided into sub-categories. Since there are two or three subclasses of high-potassium glass and lead-barium glass, it is tentatively necessary to divide them into six categories through data characteristics. Since the number of clusters has been determined, the *k*-means clustering method is selected.

Let the number of sample points be n, and randomly select K initial cluster centers $z_j(I)$, where I represents the number of iterations, j=1,2,..., K; Calculate the Euclidean distance between the sample point and the initial cluster center when:

$$D(x_i, z_j(I)) = \min\left\{(x_i, z_j(I))i = 1,2,\ldots\ldots,n,\ j = 1,2,\ldots\ldots,K\right\}$$
$$x_i \in w_j \tag{26}$$

Then calculate the error square sum criterion function:

$$Jc = \sum_{j=1}^{K} \sum_{i=1}^{K} \| x_i^{(j)} - z_j(I) \|^2 \tag{27}$$

Where x_i (j) is the sample in class j, and Z_j (I) is the cluster center of class j;
If:

$$| Jc(I+1) - Jc(I) | < \varepsilon \tag{28}$$

Then the algorithm ends; Otherwise I=I+1, K new cluster centers are calculated:

$$Z_j = \frac{1}{n} \sum_{i,j=1}^{K} X_i^{(j)} \tag{29}$$

The distance between the cluster center and each sample to the cluster center is obtained by using Matlab. The subcategories of high-potassium glass and lead-barium glass are divided into 2,3,4,5 categories, and their accuracy is tested by training. The test accuracy is the highest when both kinds of glass are divided into two categories. The high potassium glass can be divided into high potassium I glass and high potassium II glass; Lead barium glass is divided into lead barium I glass and lead barium II glass.

Use Matlab to add noise data on the basis of the original data, observe the change of distance from each point to the cluster center and the change of subclass category, and test the rationality of the model.

The distance from the material point to the cluster center before and after adding noise data and its classification are shown in Table 7 and 8:

Table 7. High potassium

Before adding noise		After adding noise	
Subclass	Distance to cluster center	Subclass	Distance to cluster center
1	7.203	1	7.242
1	5.357	1	5.265
1	4.571	1	4.584
1	3.203	1	3.268
1	3.617	1	3.647
1	2.923	1	2.875
1	13.365	1	13.323
1	14.683	1	14.72
2	10.17	2	10.211
2	5.412	2	5.374
2	2.042	2	1.987
2	4.983	2	5.025
2	7.727	2	7.693
2	5.743	2	5.726
2	4.125	2	4.141
2	7.495	2	7.455

Table 8. Lead and barium

Before adding noise		Before adding noise	
Subclass	Distance to cluster center	Subclass	Distance to cluster center
2	22.58013	2	22.59132
2	15.76852	2	15.77052
2	13.3388	2	13.34087
1	11.06345	2	11.05089
1	6.03478	1	6.04643
1	5.55564	1	5.54767
1	11.67424	1	11.68529
1	4.06388	1	4.07259
1	6.59459	1	6.6078
1	6.86936	1	6.87625
1	12.08599	1	12.09404
2	13.7571	2	13.76424
2	12.20585	2	12.21225

Through the work in Table 8, we can obtain the cluster center points of two subclasses of high-potassium glass and lead-barium glass, which are composed of the exact chemical composition content. If the content of a chemical component in the cluster center is changed, the cluster center will also change, and the distance from the sample point to the cluster center will also change. The degree of change can be used as an indicator of the sensitivity of the response model.

Through the fitting test, it is found that the distance between the data points and their distance will change greatly if the chemical composition content of the cluster center is changed arbitrarily. Therefore, the sensitivity of the model is high.

4. Identification of types of glass relics

The classification of glass types solved by Matlab is shown in Table 9:

Table 9. Classification of glass types

Cultural relic number	Cultural relics subclass	Surface weathering
A1	High potassium class I	No weathering
A2	Lead and barium II	Weathering
A3	Lead and barium II	No weathering
A4	Lead and barium II	No weathering
A5	High potassium class II	Weathering
A6	High potassium class I	Weathering
A7	High potassium class I	Weathering
A8	Lead and barium I	No weathering

5. Conclusion

(1) There is a high correlation between glass type and surface weathering. Lead barium glass is easy to be weathered, and high potassium glass is not easy to be weathered

(2) High-potassium glass and lead-barium glass have the highest classification accuracy when they are divided into two sub-categories. Therefore, glass can be divided into four categories: high-potassium I, high-potassium II, lead-barium I, and lead-barium II.

References

[1] Zhang Liyan, Li Hong, Chen Shubin, Li Zhongdi, Ruan Zhizhi, Xue Tianfeng, Qian Min, Fan Sijun. Simulation method for composition and properties of glass [J]. Journal of Silicate, 2022, 50 (08): 2338 - 2350.

[2] Zhou Tong, Ji Kejian, Wang Xueqin, Deng Weihua, Hua Lan, Zhao Xiaogang. Development of reference materials for laser glass composition analysis [J]. Chemical Analysis and Metrology, 2012, 21 (04): 7 - 9.

[3] He Xuyuan. Effect of glass composition on the properties of molding materials [J]. Glass enamel and glasses, 2022, 50 (06): 31 - 37.

[4] Yu Fei, Zhao Yanhong. Analysis of influencing factors of rural e-commerce development in Hebei Province based on chi-square test [J]. Journal of Hebei Software Vocational and Technical College, 2021, 23 (02): 17 - 20.

[5] Zhang Hongliang, Ke Bailing, Dai Xiangzhu. Analysis of the influencing factors and countermeasures of college counselors' job burnout based on the chi-square test [J]. Research on Ideological and Political Education, 2020,36 (03): 147 - 151.

[6] Zhang Hao, Xiao Yong, Yang Zhaoxu, Zhang Rui, Xu Bin. Integrated navigation system based on dual-state chi-square fault detection [J]. Journal of Aviation, 2020, 41 (S2): 53 - 60.

[7] Liu Zijian, Wang Yong, Liu Yuanni, Zhou Yousheng. Efficient short text stream clustering algorithm based on plot memory [J/OL]. Computer engineering: 1-11 [2023-02-24].