A logistic regression-based model for classification and evaluation of chemical composition of silicate glass products

Yaozheng Zhang [#, *], Yang Zhi [#]

Chongqing University of Posts and Telecommunications, Chongqing, China, 400065

* Corresponding author: 15922913680@163.com

[#] These authors contributed equally.

Abstract. Glass is one of the earliest human-made materials, and it has played an irreplaceable role in the development of human civilization. In ancient China, glass was an important commodity on the Silk Road, and it became a valuable witness to early trade and cultural exchanges between China and the West. Due to manufacturing processes, ancient glass is easily affected by burial environments and undergoes weathering. During the weathering process, there is a significant exchange of elements between the glass and the environment, causing changes in the proportions of the glass's chemical composition, which can affect the correct determination of the glass type. In this paper, we studied multiple sets of weathered and unweathered silicate glasses, extracted data on the chemical composition, color, and patterns of the sampling points, and first found a significant difference between the six chemical composition elements in the unweathered parts for the two silicate glass categories. Then, we introduced logistic coefficients and established a chemical composition evaluation model based on logistic regression to obtain the classification patterns of the two glass types. To further subdivide the glass into subcategories, we selected the top five chemical composition contents with the highest standard deviation for clustering under each of the two glass types and divided them into three subcategories. We obtained specific partitioning methods and results and analyzed the rationality of the results.

Keywords: Logistic Regression, System clustering method, Glass Classification.

1. Introduction

Glass has a long history of manufacture and has played an irreplaceable role in human society for thousands of years. In various industries today, it still has its unique advantages. Ancient glass is a valuable witness to cultural exchanges between East and West civilizations [1]. As cultural relics, ancient silicate glasses are easily affected by environmental factors due to their manufacturing processes [2], leading to weathering phenomena and making it difficult for professionals to make correct judgments about the glass type, which affects their analysis and research. However, current research on ancient glass is mostly focused on glass composition and cultural development [3] [4], and there is little research on glass classification. This paper aims to study the classification patterns of two types of ancient glass to address the issue of sub-classification of the two types of glass and propose a preliminary plan for modern research on the classification of ancient glass.

Currently, there are several known methods for glass classification research. As early as the 1980s, Shi Meiguang and others analyzed the chemical composition of a batch of potassium glass from Chinese Han tombs and classified the glass [5]. Zhou Liangzhi used modern research methods to discuss the weathering mechanism of silicate glass, identified weathering products, and provided reference for the processing of the chemical composition data in this paper [6]. Zhao Fengyan and others analyzed the chemical composition of several glass objects excavated from Xi'an using a high-performance portable energy dispersive X-ray fluorescence spectroscopy analyzer [7]. However, current chemical research methods cannot accurately classify glass more precisely according to its composition. Kan Yinghao used instruments to detect the chemical composition of ancient glass from a glass workshop in Boshan, Shandong, China, dating back to the late Yuan and early Ming dynasties, and observed crystal distribution phenomena using a scanning electron microscope. By judging the category and content of the chemical composition in the glass according to the color and shape of the crystals [8], the analysis method for the rationality of the classification results in this paper is of reference significance. Wang Zihao and others used the decision tree method to solve the rough classification problem of ancient glass and achieved good results [9].

This paper is based on existing publicly available data and multiple sets of weathered and unweathered high-potassium and lead-barium glass data to analyze the classification patterns of the two types of glass. It can be clearly seen from the data that the classification patterns are related to the proportion of 14 chemical composition elements. By introducing a classification method and selecting the significant differences in chemical composition as the main classification basis for the two types of glass, accurate classification patterns for the two types of glass can be obtained. Secondly, to subdivide the two types of glass into subcategories, suitable chemical composition elements need to be selected for the two types of glass categories and analyzed for weights and contribution rates of the 14 chemical composition elements, followed by dimensionality reduction processing and system clustering. Using the results of dimensionality reduction, the clustering results were systematically clustered in the two categories of glass, and the clustering results were summarized inductively to obtain the basic features of different categories as specific subdivision methods. Finally, a rationality analysis was conducted to explain the relationship between the sub-categories and the location of the sampling points based on the relationship between the color of the sampling points and their chemical composition

2. Establishment and solution of glass classification model.

2.1. Establishment of Model

2.1.1. Sample collection and data processing

The research data used in this study was obtained from a batch of related data on pre- and post-weathering ancient glass provided by the official website of the 2022 China Undergraduate Mathematical Modeling Competition [10]. Archaeologists have provided the chemical composition, color, and glass type of the cultural relic’s samples through professional techniques. As there are missing data in the chemical composition of the samples where the chemical component is not detected, it is considered to use a value of zero to supplement it. In addition, due to the detection methods and other reasons, the sum of component proportions may not add up to 100%. Accordingly, data where the sum of the component proportions ranges between 85% to 105% shall be considered valid. Therefore, it is necessary to calculate the total chemical composition of each sampling point to eliminate invalid data prior to analysis.

After eliminating the invalid data, the remaining valid data is categorized into two types according to the type of glass. The mean values of the chemical composition of the unweathered parts of the two types of glass are calculated. By comparing these mean values, the differences in the chemical composition between the two types of glass can be roughly understood, providing a reference for accurately classifying the glass in the future.

2.1.2. Chemical component evaluation model based on logistic regression

The logistic model is based on probability theory, which maps the logits values calculated between the independent variable and the dependent variable to a range of probability values from 0 to 1. By establishing the probability relationship between the dependent variable and the independent variable, classification can be achieved [11] [12].

For the glass classification problem, the classification rules of glass types are analyzed first. There are two types of glass, high-potassium glass and lead-barium glass, which are independent and do not interfere with each other. Therefore, a binary logistic model can be established to solve this problem.

The high-potassium glass is coded as 0, and the lead-barium glass is coded as 1, and the mathematical model is established as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \tag{1}$$

$$P = \frac{1}{1-exp[-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m)]} \tag{2}$$

Among them, β_i represents the coefficient of each variable, and β_0 is the intercept term. y represents the type of glass as the dependent variable, p represents the probability of lead-barium glass, and 1-p represents the probability of high-potassium glass.

Considering the positive relationship between probability and odds, odds can replace probability to describe the likelihood of random events occurring. Equation (1) can be transformed into equation (3) by using odds to describe the likelihood:

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{3}$$

To analyze the chemical composition that may affect the type of glass, the elements SnO_2 and SO_2 were not considered due to their excessive missing values in the dataset, and the remaining elements were used as factors that could affect the determination of the type of glass to establish a logistic model.

2.1.3. Subclass division model based on system clustering

In order to conduct subcategory division, appropriate chemical compositions were first selected for subcategory division. The merged data was screened and classified, and weathered data was selected, and divided into high-potassium glass and lead-barium glass according to the type of glass. The standard deviation of each chemical composition was calculated, and the formula for calculating the standard deviation is as follows:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{4}$$

Here, s represents the standard deviation, s^2 represents the variance, x_i represents the chemical composition content of the monitoring point with the number i of the cultural relic, and $\bar{x}$ represents the mean of the chemical composition content.

The obtained standard deviation data was subjected to standardization, and the formula for standardization is as follows:

$$z = \frac{x-minx}{max\,x-min\,x} \tag{5}$$

Here, z represents the standardized data.

A larger standard deviation indicates that the chemical composition contributes more to the classification. Chemical compositions with larger standard deviations were selected as the criteria for system clustering under the two glass categories, and the subcategory method and results were obtained.

In system clustering, each sample to be studied is first regarded as a separate category. Then, the two categories with the highest similarity between the samples are merged, followed by the merging of the two categories with the highest similarity among the merged categories. This pairwise merging process is repeated until all samples are merged into one category [13] [14].

Finally, the rationality of the classification results was analyzed. The characteristics of the classified samples were summarized based on the results obtained from the subcategory model based on system clustering. To demonstrate the rationality of the classification results, the classified features were re-entered into the known dataset, and the cultural relic scanning points in the dataset were classified and counted using the new features. The results were compared with the existing ones to obtain rationality.

2.2. Solution of model

2.2.1. Data preprocessing result

The chemical composition results obtained from each cultural relic sampling point were accumulated. It was found that the cumulative chemical composition value of the cultural relic sampling point numbered 15 was 79.47%, and that of the cultural relic sampling point numbered 17 was 71.89%. Therefore, data from the cultural relic sampling points 15 and 17 should be excluded in the subsequent data processing.

To divide the remaining valid data into two categories, according to the type of glass, and sorted according to whether they were weathered or not. The chemical composition content data of the unweathered parts were selected. The mean values were calculated and compared for the two types of glass. The results are shown in Table 1:

Table 1. Comparison of mean values of chemical constituents of two types of unweathered glass

Chemical Component	High Potassium Glass	Lead Brium Glass	Rate (Larger/Smaller)
SiO_2	67.98416667	53.44384615	1.272067255
Na_2O	2.78	3.343333	1.202638
K_2O	10.17909	0.42	24.23593
CaO	6.399	1.455455	4.396565
MgO	1.295	0.914286	1.416406
Al_2O_3	6.62	3.194615	2.072237
Fe_2O_3	2.318	2.021667	1.146579
CuO	2.675455	1.84	1.454051
PbO	0.705714	23.59385	33.43258
BaO	1.436	10.49923	7.311442
P_2O_5	1.53	1.068181818	1.432340426
SrO	0.083333	0.4825	5.79
SnO_2	2.36	0.42	5.619048
SO_2	0.406667	3.66	9

It can be seen that the chemical compositions with a proportion greater than 2 are PbO, K_2O, SO_2 ,BaO, SrO, SnO_2, CaO, Al_2O_3. Due to the excessive missing values of SnO_2 and SO_2, their impact on classification was not considered. Therefore, the classification patterns of high-potassium glass and lead-barium glass may be mainly related to the six elements of PbO, K_2O, BaO, SrO, CaO and Al_2O_3.

2.2.2. Chemical composition evaluation results based on logistic regression

The results obtained from the chemical composition evaluation model based on logistic regression [15] [16] are as follows, as shown in Table 2:

Table 2. The variables in the logistic equation

	B	S.E.	Wald	Degree of Freedom	Significance Level	Exp(B)
SiO_2	-0.115	0.053	4.618	1	0.032	0.891
Na_2O	1.115	1.131	0.972	1	0.324	3.051
K_2O	-9.118	2064.451	0	1	0.996	0.000
CaO	-1.005	0.412	6.547	1	0.11	0.348
MgO	-1.861	1.302	2.042	1	0.153	0.156
Al_2O_3	-0.927	0.366	6.411	1	0.011	0.396
Fe_2O_3	-0.042	0.357	0.014	1	0.907	0.959
CuO	-0.156	0.217	0.518	1	0.472	0.856
BaO	50.089	2801.867	0	1	0.986	5.663E+21
PbO	4.411	789.343	0	1	0.996	82.387
P_2O_5	-0.125	0.291	0.185	1	0.667	0.882
SrO	306.287	57605.203	0	1	0.996	1.0438E+133

B is the logistic regression coefficient, which represents the change in the natural logarithm value of the odds ratio when the value of the independent variable increases by one unit while the values of other independent variables remain constant. When the probability is relatively low, it can be understood that the probability will increase several times from the original probability. This is similar to the maximum likelihood estimation. In the logistic regression model, Exp(B) is used to explain the degree of increase in the outcome. When Exp(B) is greater than 1, it indicates that the factor promotes the outcome; when Exp(B) is less than 1, it indicates that the factor hinders the outcome; when Exp(B) is equal to 1, it indicates that the factor has no effect on the outcome.[15] [16]

By observing the values in the rightmost column, it can be found that the values of Exp(B) for *SrO*, *BaO*, and *PbO* are 1.0438E+133, 5.6663E+21, and 82. 387.All these values are greater than 1 and much larger than other values, indicating that *SrO*, *BaO*, and *PbO* all promote the encoding of lead-barium glass with a value of 1. Thus, it can be considered that *SrO*, *BaO*, and *PbO* can be used as the classification basis for lead-barium glass. The corresponding Exp(B) values for K_2O,MgO,CaO,and Al_2O_3 are 0.000, 0.156, 0.348, and 0.396. There is a significant quantitative difference between these values and other values. This suggests that K_2O ,MgO, CaO ,and Al_2O_3 can be used as the classification basis for high-potassium glass.

2.2.3. Subclass division results based on system clustering

The standard deviation of each chemical composition was calculated by standardizing each chemical composition of high potassium glass and lead-barium glass. The following results were obtained, as shown in Table 3.

Table 3. Standard deviation of chemical composition content after standardization of two glasses

	High Potassium Glass		Lead barium glass
SrO	0.386151077	SrO	0.38251122
SO_2	0.377905058	SnO_2	0.344862322
Na_2O	0.364535182	Al_2O_3	0.331894473
PbO	0.348094089	SiO_2	0.321666207
CaO	0.340324403	Na_2O	0.317158878
BaO	0.328773088	MgO	0.313813889
MgO	0.326949236	CaO	0.311918829
CuO	0.312243451	Fe_2O_3	0.302491914
P_2O_5	0.305091398	PbO	0.29203244
SiO_2	0.29894325	BaO	0.292473126
Al_2O_3	0.294498576	CuO	0.281595664
SnO_2	0.276385399	K_2O	0.271275739
Fe_2O_3	0.264195074	SO_2	0.266469355
K_2O	0.258498892	P_2O_5	0.262428819

The largest 5 standardized chemical composition contents with the highest standard deviation were selected for system clustering for each of the two types of glass, and the clustering results are shown in Figure 1:

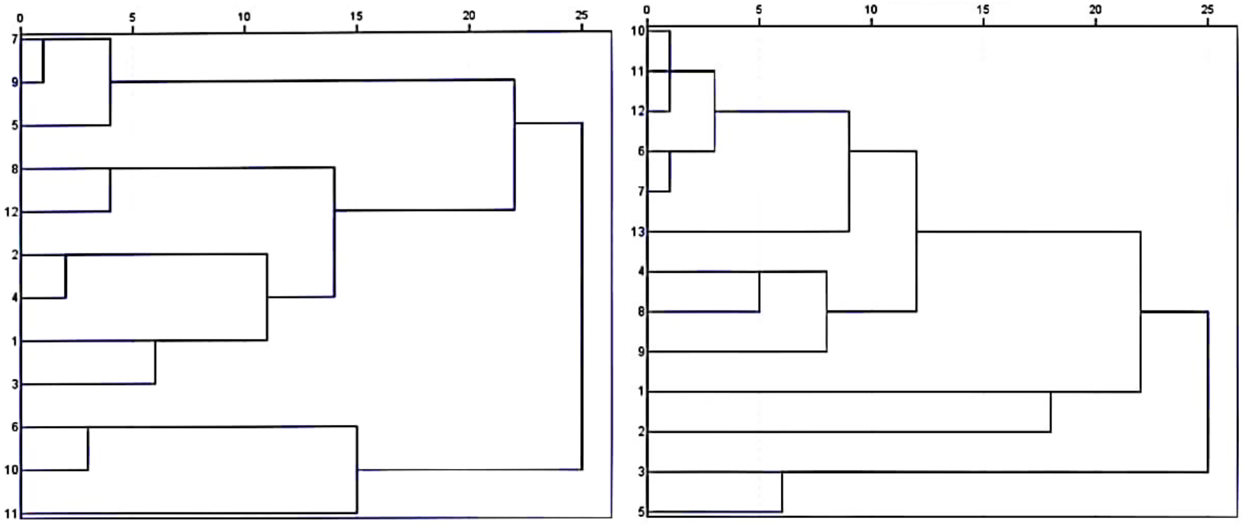

Figure 1. High potassium glass (left) and lead-barium glass (right) clustering results

According to the dendrogram obtained from clustering, it can be found that when the adjusted relative distance is 20, the clustering results have more obvious 3 classes.

When the glass type is high potassium glass, the sampling point division yields the results shown in Table 4:

Table 4. High potassium glass sampling point division

	Na_2O	CaO	PbO	SrO	SO_2	Heritage Sampling Sites
Class 1	0	6.9	0	0.02	0.4067	1 4 5
Class 2	0	3	0.535	0.0667	0	3 6 18 21
Class 3	2.78	8.4	0.5767	0.0133	0	13 14 16

When the glass type is lead-barium glass, the sampling point division results are shown in Table 5:

Table 5. Lead barium glass sampling point division

	SO_2	Al_2O_3	SrO	SnO_2	Na_2O	Heritage Sampling Sites
Class 1	63.292	20.984	0.122	0	0.542	55 35 33 37 31
Class 2	47.466	3.978	0.3775	0	1.464	20 24 45 46 47
Class 3	35.635	0	0.705	4.365	0.745	30

Rationalization of the results: the samples in the classification results are corresponded to the colors and sampling points in the known data set, as shown in Table 6:

Table 6. High potassium glass sampling points - color correspondence table

category	sampling point	Color
1: K-SO_2	01	bluish green
	04	bluish green
	05	bluish green
2: K-BaO	03-Part1	bluish green
	21	bluish green
	06-Part1	bluish green
	18	dark blue
	06-Part2	bluish green
	03-Part1	bluish green
3: K-Na_2O	16	Cambridge blue
	13	Cambridge blue
	14	bottle green

Analysis of Table 6 and known data, for high potassium glasses:

Based on chemical composition measurement: only category 1 artifacts contain SO_2, only category 2 artifacts contain BaO, only category 3 artifacts contain Na_2O, so the chemical composition to identify the artifact material, classification is reasonable.

Based on color measurement: categories 1, 2 are mainly blue-green, and category 3 is mainly light blue. Since the color of potassium glass itself is blue-green, most of the artifacts presented in the analysis of high potassium glass are similar in color to blue-green [17], which is in accordance with objective facts, which can justify the classification.

Based on the sampling point measurement: both sampling points of sample 3 were classified in one category, and two sampling points of sample 6 were also classified in one category. The reasonableness of the classification can be illustrated.

Table 7. Lead barium glass sampling point-color correspondence table

category	sampling point	Color
1: $Pb\text{-}Ba\text{-}SO_2$	33	bottle green
	35	light green
	55	green
	37	bottle green
	31	purple
2: $Pb\text{-}Ba\text{-}Na_2O$	46	Cambridge blue
	45	Cambridge blue
	20	Cambridge blue
	24	purple
	47	Cambridge blue
3: $Pb\text{-}Ba\text{-}SnO_2$	30-Part1	dark blue
	30-Part2	dark blue

Analysis of Table 7 and known data, for lead-barium glass:

Based on chemical composition measurement: only category 1 of the artifacts containing SO_2, category 2 of the artifacts containing Na_2O accounted for the most, only category 3 of the artifacts containing SnO_2, so you can identify the artifacts based on the chemical composition of the material, classification is reasonable.

Based on color measurement: category 1 in most of the cultural relics show green, category 2 in most of the cultural relics show light blue, category 3 in the cultural relics show dark blue, you can judge the category based on color, classification is reasonable.

Based on the sampling point measurement: both sampling points of sample 30 were classified in one category, which can justify the classification.

In summary, each category is strongly differentiated, in accordance with the classification principle, and the model effect is excellent.

3. Conclusion

This paper utilized an available dataset to preprocess the data and obtain the mean values of each chemical composition without weathering from two types of glass. Through a ratio comparison of the mean values, it was discovered that six chemical components of unweathered glass were significantly different under the two glass types. Subsequently, logistic coefficients were introduced to establish a logistic regression-based chemical composition assessment model, effectively providing the classification patterns of the two glass types. Following this, the max-min method was employed to normalize the chemical composition, with the standard deviation of each chemical composition under the two glass types calculated separately. The top five chemical compositions with the largest to smallest standard deviation were selected as indicators for subclass classification under the two types of glasses. Systematic clustering methods were then used to cluster the two glass types into three distinct classes, with specific division methods and results obtained. Finally, the classification characteristics of the division were summarized with color, sampling point, and chemical composition being used as critical indicators to verify the rationality of the results. This verification proved the rationality of the clustering division, while the classification results demonstrated a high degree of reliability. The methodology employed in this study was rigorously applied and provided insightful findings.

References

[1] Gan Fuxi. The ancient Silk Road and ancient Chinese glass [J]. Journal of Nature, 2006 (05): 253 - 260+246.

[2] Bo Yimeng, Huang Xiaolei, Wei Meili. Scientific analysis and research of glassware excavated from the Liao tomb of Xiaolama Gou [J]. Journal of Liaoning Provincial Museum, 2014 (00): 244 - 252.

[3] Chen Shuyu. The origin and development of ancient glass in China [J]. Antiquities Identification and Appreciation,2019, No.151 (04): 44 - 45.

[4] Cheng Zhuhai. A pilot study on the development of ancient glass in China[J]. Journal of Silicates, 1981 (01): 96 - 104+124 - 126.

[5] Shi M.K., He O.L., Zhou F. Z. Study of potassium glass from a group of Chinese Han tombs [J]. Journal of Silicates, 1986 (03): 307 - 313.

[6] Zhou, Liang-Chi. The main factors affecting the weathering of silicate glass [J]. Journal of Dalian Light Industry College, 1984 (01): 34 - 44.

[7] Zhao Fengyan, Chen Bin, Chai Yi et al. pXRF analysis of several glass wares excavated in Xi'an and discussion of related issues[J]. Archaeology and Cultural Relics, 2015, No.210 (04): 111 - 119.

[8] Kan Yinghao. Analysis of glass technology excavated from glass workshops in late Yuan and early Ming dynasties in Boshan, Shandong [D]. Shandong University, 2022.

[9] Wang Zhahao, Zhao Xiang, Li Zhiqun et al. A machine learning-based method for predicting and subclassifying the original composition of weathered silicate glasses [J]. Journal of Silicates, 2023, 51 (02): 416 - 426.

[10] Chinese Society of Industrial and Applied Mathematics [EB/OL]. [2022-09-15]. http://www.mcm.edu.cn./html_cn/node/5267fe3e6a512b ec793d71f2b2061497.html.

[11] Wei Chengjing, Li Guodong. Video image encryption Design Based on Hyperchaotic System and Logistic Mapping [J]. Computer Engineering, 2022, 48 (5): 263 -- 271.

[12] Jie Zhang, Mengmeng Zhang, Hongyan Li. Analysis of urban road traffic accident severity based on binary logistic model [J]. Traffic Information and Safety, 2022, 40 (05): 70 - 79.

[13] Hu, Lei-Fang. Five common systematic cluster analysis methods and their comparison [J]. Zhejiang Statistics,2007, No. 270 (04): 11 - 13.

[14] Sun Yihao, Li Qiuyan, Ding Yan et al. A comprehensive evaluation method for county power grids based on principal component analysis and system clustering [J]. Power System Protection and Control, 2017, 45 (08): 30 - 36.

[15] Shi Yanlong. Analysis of imported food safety risk perception and its influencing factors based on factor analysis and logistic regression [J]. Journal of Food Safety and Quality Inspection, 2022, 13 (03): 978 - 985.

[16] Chen, X. X., Guo, H. T. The joint use of factor analysis and logistic regression in financial early warning of agricultural listed companies [J]. Mathematical Statistics and Management, 2022, 41 (01): 11 - 24.

[17] Sichin Bilig, Li Qinghui, Gan Fuxi. Analysis of ancient Chinese potassium glass fractions by laser exfoliation-inductively coupled plasma-atomic emission spectrometry/mass spectrometry [J]. Analytical Chemistry, 2013, 41 (09): 1328 - 1333.