Glass classification and identification based on logistic regression analysis and K-means++ clustering algorithm

Xiao Chen*

Chongqing University-University of Cincinnati Joint Co-op Institute, Chongqing University, Chongqing, China

*Corresponding author: chen3x6@mail.uc.edu

Abstract: During the weathering process, a large number of internal elements are exchanged with environmental elements, resulting in a change in the proportion of glass composition, which affects the correct judgement of glass categories. By comparing the chemical composition of high potassium glass and lead-barium glass, this paper finds that the proportion of silicon dioxide and potassium oxide is significantly higher in the high potassium glass category, while the proportion of lead oxide and barium oxide is high in the lead-barium glass category, and the above-mentioned substances with large differences in content are defined as classified substances. A logistic regression analysis was carried out to derive a regression equation between the classified substances and the glass type, which was used as a basis for determining the glass type. Based on this, a K-means++ clustering algorithm was used to perform subclass classification, resulting in three subclasses for each type. Finally, the glass of unknown composition was identified according to the established model.

Keywords: Glass composition identification; logistic regression analysis; K-Means++.

1. Introduction

Glass circulation through the Silk Road[1], is a valuable physical evidence of cultural exchange between China and the West[2-4]. The main raw material for glass is quartz sand, and in order to reduce the melting temperature of pure quartz sand during the refining process, ancient craftsmen often resorted to such things as grass wood ash, natural bubble soda[5], saltpeter and lead ore as fluxes, and often limestone as a stabilizer[6].

Ancient Chinese glass compositions are made up of a variety of compositional systems[7]. Our high lead glass is unique in the world of ancient glass compositions. In particular, PbO-BaO-SiO2 system glass. is only found in China in ancient ancient glass compositions.

Differences in fluxes caused the main chemical composition of the glass to differ as well. For example, the glass obtained by fluxing with lead ore is called lead-barium glass, which has a high content of lead oxide (PbO) and barium oxide (BaO)[8]; the glass made by fluxing with substances containing high amounts of potassium, such as grass ash, is called potassium glass. Ancient glass is highly susceptible to weathering, and during the process of weathering, a large number of internal elements are exchanged with environmental elements, resulting in changes in the proportions of the glass composition and affecting the correct determination of the glass type[9-10].

2. Modelling glass classification

2.1 Glass classification rules

Firstly, the data was visualized to produce a histogram of the difference between the mean and average values for each component of high potassium glass and lead barium glass, the results of which are shown below.

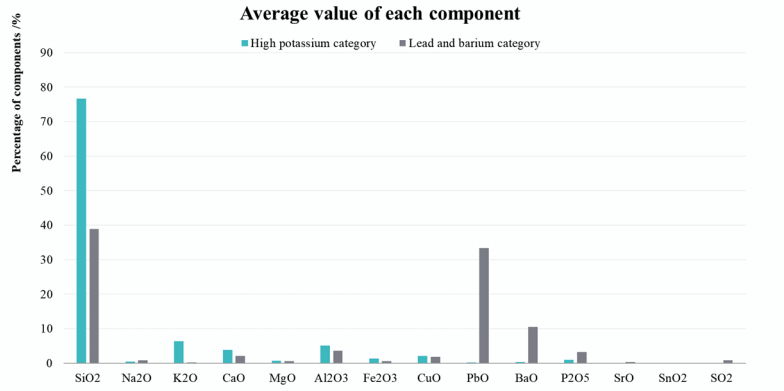

Figure 1: Histogram of the mean values of the components of the high potassium and lead-barium glasses

As can be seen from the figure 1, the two components of silicon dioxide and potassium oxide are significantly higher than those of lead-barium glass in the high potassium glass category, while the two components of lead oxide and barium oxide are significantly less than those of lead-barium glass. Therefore, we infer that the main components of high potassium glass are silicon dioxide and potassium oxide, while the main components of lead-barium glass are silicon dioxide, lead oxide and barium oxide. The results are shown in the figure 2.

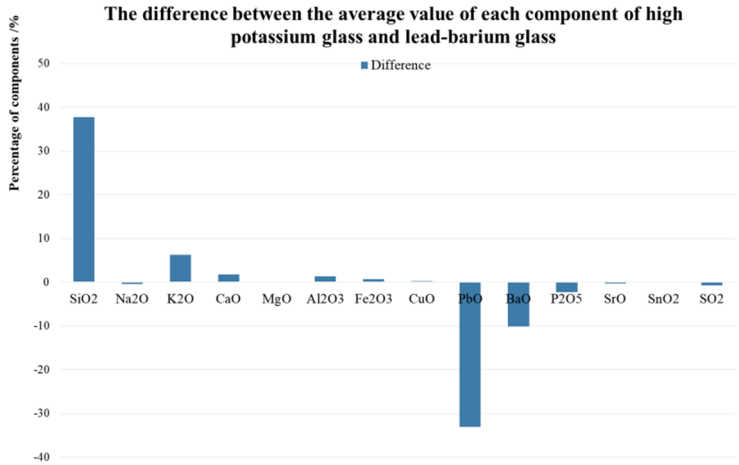

Figure 2: Histogram of the difference between the mean values of the components of high potassium glasses and lead-barium glasses

We defined these widely varying components as categorical components and first conducted simple sample statistics on the categorical components with the following table 1:

Table 1: Statistical table describing the content of components that differ significantly between the high potassium and lead-barium glass types

Descriptive statistics					
	N	Minimum	Maximum	Mean	Standard deviation
Silicon dioxide	67	3.72	96.77	49.02	24.32
Potassium oxide	67	0.00	14.52	1.85	3.88
Lead oxide	67	0.00	70.21	24.46	19.51
Barium oxide	67	0.00	35.45	7.78	8.42

Next, we defined the high potassium category as 1 and the lead-barium category as 0. We standardized the classification components and then performed a linear regression analysis, the results of which are shown in the table 2:

Table 2: Table of regression analysis coefficients for the content of components that differ significantly between high potassium and lead-barium glasses and their types

Coefficients							
Models	Unstandardized factor		standardized factor	t	Significance	Covariance statistics	
	B	Standard errors	Beta			tolerance	VIF
(Constant)	0.27	0.03		9.75	0.00		
Zscore:Silicon dioxide (SiO2)	0.14	0.09	0.31	1.47	0.15	0.09	11.41
Zscore:Potassium oxide (K2O)	0.21	0.05	0.47	4.49	0.00	0.36	2.81
Zscore:Lead oxide (PbO)	-0.09	0.09	-0.21	-1.06	0.29	0.10	10.18
Zscore:Barium oxide (BaO)	-0.04	0.05	-0.10	-0.82	0.41	0.28	3.60

The B values of the regression coefficients show that silicon dioxide and potassium oxide are positively correlated with the type codes, and lead oxide and barium oxide are negatively correlated with the type codes, in general agreement with the observations in the visual data plots. The regression equation for the type codes can be derived as:

$$\text{Type codes} = 0.269 + 0.0138 \times \text{Z.SiO2} + 0.209 \times \text{Z.K2O} - 0.094 \times \text{Z.PbO} - 0.043 \times \text{Z.BaO} \quad (1)$$

The regression equation was next used to validate the data and some of the data validation results are shown in the table 3:

Table 3: Categorization prediction of sample types using regression equations

Type	**SiO2**	**K2O**	**PbO**	**BaO**	**Predicted value**	**Forecast**	**Accurate or not**
High Potassium	0.84	2.10	-1.25	-0.92	2.75	Type	Accurate
Lead Barium	-0.52	-0.21	1.18	-0.92	0.19	High Potassium	Accurate
High Potassium	1.56	0.86	-1.24	-0.92	1.52	Lead Barium	Accurate
High Potassium	0.52	2.71	-1.18	-0.58	3.33	High Potassium	Accurate
High Potassium	0.69	2.02	-1.25	-0.92	2.66	High Potassium	Accurate
High Potassium	0.52	2.35	-1.25	-0.92	2.99	High Potassium	Accurate
High Potassium	0.77	1.42	-1.24	-0.76	2.06	High Potassium	Accurate
High Potassium	0.44	1.50	-1.24	-0.81	2.14	High Potassium	Accurate
High Potassium	1.79	-0.48	-1.25	-0.92	0.18	High Potassium	Inaccurate
Barium lead	-1.19	-0.48	0.22	2.78	-0.15	Barium lead	Accurate
Barium lead	-1.83	-0.48	0.41	2.71	-0.18	Barium lead	Accurate

From the predicted results for all the data, it can be seen that the accuracy is 90% and therefore the classification can be considered as a valid one. The ability to classify unknown types of glass using this regression equation.

2.2 Subclass classification based on K-means clustering algorithm

This question uses the K-means clustering algorithm for subclassification. After the clustering process, we are able to estimate, analyze and predict more accurately within each class individually using statistical models; we can also study the differences between classes, which is very convenient for the next subclass classification. The flow chart of the algorithm is as figure 3:

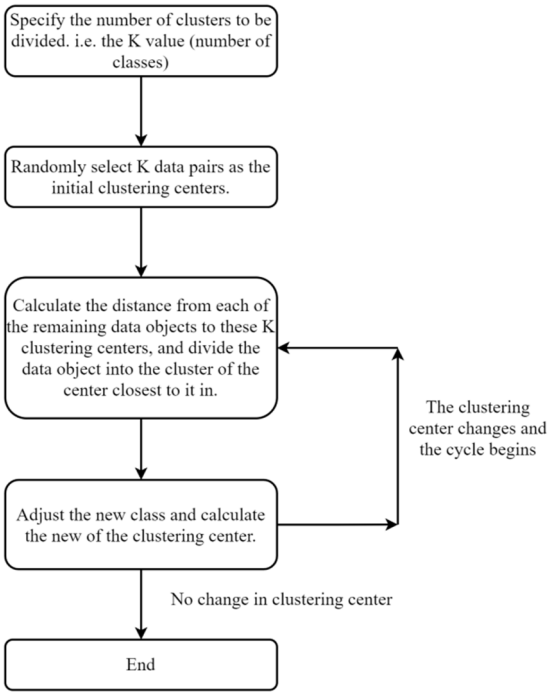

Figure 3: Flow chart for subclassing based on K-means clustering algorithm

We used the K-means++ clustering algorithm to iteratively analyze the chemical components in the table above and obtained the corresponding subclass divisions, with the results divided into the high potassium and lead-barium classes shown in table 4:

1）High Potassium

Table 4: Statistical table of the various types of artefacts numbered in the potassium category

Statistics on the numbering of various types of artefacts in the High Potassium category		
Category 1	Category 2	Category 3
03 Part 2	03 Part 1	01
06 Part 1	07	04
06 Part 2	09	05
21	10	13
	12	14
	18	16
	22	
	27	

The final selected clustering centers are as follow table 5:

Table 5: Table of final cluster center values calculated by the K-means++ clustering algorithm for different subclasses of artefact numbers in the high potassium class

Final Clustering Center			
	Clustering		
	1	2	3
Zscore: Silicon dioxide (SiO2)	-0.70	1.01	-0.88
Zscore: Sodium oxide (Na2O)	-0.43	-0.43	0.85
Zscore: Potassium oxide (K2O)	0.09	-0.79	0.99
Zscore: Calcium oxide (CaO)	0.05	-0.89	1.15
Zscore: Magnesium oxide (MgO)	1.02	-0.63	0.16
Zscore: Aluminium oxide (Al2O3)	1.03	-0.88	0.49

Zscore: Iron oxide (Fe2O3)	1.19	-0.75	0.21
Zscore: Copper oxide (CuO)	0.74	-0.59	0.30
Zscore: Lead oxide (PbO)	0.91	-0.47	0.03
Zscore: Barium oxide (BaO)	1.66	-0.47	-0.47
Zscore: Phosphorus pentoxide (P2O5)	1.24	-0.44	-0.24
Zscore: Strontium oxide (SrO)	1.25	-0.43	-0.25
Zscore: Tin oxide (SnO2)	-0.24	0.29	-0.24
Zscore: Sulphur dioxide (SO2)	-0.43	-0.43	0.86

Table 6: Table of final cluster centroid distance values calculated by the K-means++ clustering algorithm for different subclasses of artefact numbers in the high potassium class

Distance between final clustering centers			
Clustering	Category 1	Category 2	Category 3
Category 1	0	5.38	4.17
Category 2	5.38	0	4.36
Category 3	4.17	4.36	0

Table 7: Evaluation table of the different subclasses of artefacts in the high potassium class calculated by the K-means++ clustering algorithm by F-test analysis

ANOVA						
	Clustering		Error			
	Mean Square	Degree of Freedom	Mean Square	Degree of Freedom	F	Significance
Zscore: Silicon dioxide (SiO2)	7.41	2.00	0.15	15.00	51.16	0.00
Zscore: Sodium oxide (Na2O)	3.26	2.00	0.70	15.00	4.67	0.03
Zscore: Potassium oxide (K2O)	5.42	2.00	0.41	15.00	13.21	0.00
Zscore: Calcium oxide (CaO)	7.17	2.00	0.18	15.00	40.31	0.00
Zscore: Magnesium oxide (MgO)	3.72	2.00	0.64	15.00	5.84	0.01
Zscore: Aluminium oxide (Al2O3)	5.97	2.00	0.34	15.00	17.74	0.00
Zscore: Iron oxide (Fe2O3)	5.23	2.00	0.44	15.00	11.97	0.00
Zscore: Copper oxide (CuO)	2.78	2.00	0.76	15.00	3.65	0.05
Zscore: Lead oxide (PbO)	2.54	2.00	0.80	15.00	3.19	0.07
Zscore: Barium oxide (BaO)	7.08	2.00	0.19	15.00	37.26	0.00
Zscore: Phosphorus pentoxide (P2O5)	4.04	2.00	0.59	15.00	6.80	0.01
Zscore: Strontium oxide (SrO)	4.06	2.00	0.59	15.00	6.85	0.01
Zscore: Tin oxide (SnO2)	0.63	2.00	1.05	15.00	0.60	0.56
Zscore: Sulphur dioxide (SO2)	3.35	2.00	0.69	15.00	4.87	0.02

As clusters have been selected to maximize the differences between cases in different clusters, the F-test should only be used for descriptive purposes. The measured significance levels have not been corrected for this and so cannot be interpreted as a test of the hypothesis that the 'cluster means are equal'.

As can be seen from the table 5、6、7, the significance statistics for all chemical compositions are less than 0.05, except for copper oxide (CuO) and tin oxide (SnO2), which indicates that the clustering algorithm is highly correlated with chemical composition and a good method of subclassing.

The modelling approach for lead-barium glass is consistent with that for high-potassium glass and is therefore not repeated.

Statistics on the number of classified cases is shown as figure 4.

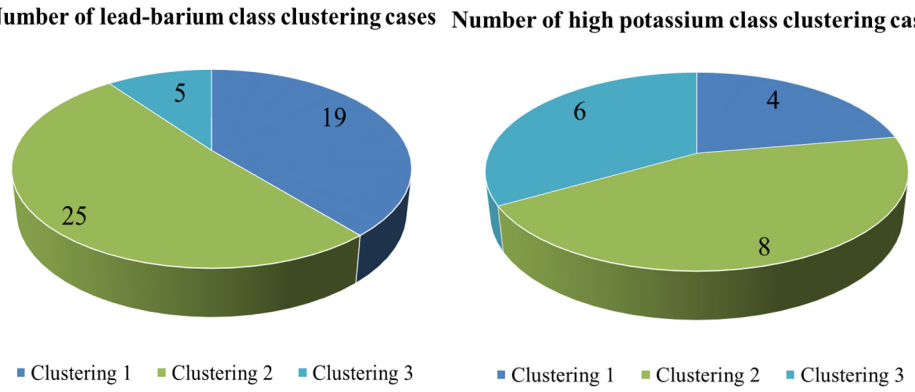

Figure 4: Statistical pie of the number of classified cases

Rationality and sensitivity analysis of sub-categorization:

Indicators to specifically evaluate the rationality and sensitivity of the sub-category split are also given below.

Reasonableness:

Compactness d is an evaluation indicator of classification rationality and is used to measure the compactness of all points within a category, expressed as the average distance of all points within a category to the center of the category, and is calculated as follows.

$$d = \frac{1}{n}\sqrt{\sum_{k=1}^{14}(x_k - x_{Dk})^2}^{\frac{1}{n}} \quad (2)$$

x_{DK}: Coordinates of category centroids; x_K: Coordinates of the point; n: Number of samples within category

1. High Potassium

The results of the calculations obtained were.

The average distance for category 1 was p1 = 2.9174; for category 2, p2 = 1.5545; and for category 3, p3 = 2.4143. All three values were small, with the artefacts of each type being closer to their respective centroids, which indicates that all three fit well and were classified from the same class. This means that the use of the clustering algorithm model for sub-classing has been successful.

2. Lead and barium

The results of the calculations are.

The average distance for category 1 is p1 = 3.3162; the average distance for category 2 is p2 = 2.5418; the average distance for category 3 is p3 = 2.8908.

The small mean values for the three classes are consistent with the compactness results for the subclass delineation of high potassium glass above, i.e. each type of artefact is at a small distance from the centroid of its construction. This suggests that the subclass delineation under the model of the clustering algorithm is relatively successful and that its delineation criteria are quite reasonable.

Sensitivity analysis:

Sensitivity analysis was carried out by first normalizing the data, followed by random scaling of the perturbed data within the range (-130%, +130%) and observing how the results differed from the unperturbed data, and the results are shown in the following table.8:

Table 8: Classification accuracy after perturbation

Classification accuracy after perturbation						
	Scaling 1%	Scaling2%	Scaling5%	Scaling10%	Scaling20%	Scaling30%
Accuracy	100%	100%	100%	100%	100%	100%

It can be seen that the model is 100% accurate after receiving perturbations, and still maintains good robustness and stability, thus proving that the clustering model used in this question is universal and resistant to disturbances.

3. Type determination based on chemical composition analysis

Next, we used the regression equation above (Type code = 0.269 + 0.0138*Z.silicon dioxide + 0.209 + Z.potassium oxide - 0.094*Z.lead oxide - 0.043*Z.barium oxide) to analyze and predict the standardized data, the results of which are shown in the table 9:

Table 9: Table of predicted results after raw data

Heritage number	SiO_2	K_2O	PbO	BaO	Predicted value	**Type of prediction**
A1	0.72	-1.13	-1.04	-0.75	-0.51	Barium Lead
A2	-0.91	-1.13	1.12	-0.75	-0.74	Barium lead
A3	-1.14	1.29	1.46	0.31	1.60	High Potassium
A4	-1.00	0.28	0.49	1.13	0.64	High Potassium
A5	0.16	-0.47	-0.27	-0.26	0.05	Barium lead
A6	1.31	1.27	-1.04	-0.75	1.90	High Potassium
A7	1.22	0.61	-1.04	-0.75	1.24	high potassium
A8	-0.37	-0.72	0.30	1.82	-0.35	Barium lead

Based on the above classification results, we will further subdivide the unknown artefacts into subclasses of the respective artefacts' categories. First we apply the K-means algorithm to calculate the distance of unknown artefacts A1-A7 from the three clustering centers derived from the second question, and then compare their sizes before grouping A1-A7 into the one closest to the three clustering centers. The results of the subclassification are shown in the following table 10:

Table 10: Table of results of further subclassification of high potassium glass products

Heritage number	Clustering 1	Clustering 2	Clustering 3	Classification
A3	3.63	5.45	3.80	1
A4	1.81	4.51	3.54	1
A6	5.90	2.37	4.23	2
A7	5.61	2.00	3.92	2

Artefacts numbered A3 and A4 are all in the high potassium category 1 subclass, while artefacts numbered A6 and A7 are all in the high potassium category 2 subclass.

As can be seen from Table 10, artefact number A2 is in subclass 1 of lead and barium, artefact numbers A1 and A5 are both in subclass 2 of lead and barium, while artefact number A8 is in subclass 3 of lead and barium.

Sensitivity Analysis:

Sensitivity tests were conducted on the classification results based on the categories assigned in the first question. The perturbed treated data were randomly scaled within the range (-130%, +130%) to observe how the results differed from the unperturbed data for sensitivity analysis, and the results are shown in the table 11:

Table 11: Table of sensitivity analysis tests on classification results

	Classification results after perturbation						Classification results	Accuracy
Heritage number	Rank1	Rank2	Rank5	Rank10	Rank20	Rank30	Rank0	/
A1	2	2	2	2	2	2	2	100%
A2	2	2	2	2	2	2	2	100%
A3	1	1	1	1	1	1	1	100%
A4	1	1	1	1	1	1	1	100%
A5	2	2	2	2	2	2	2	100%
A6	2	2	2	2	2	2	2	100%
A7	2	2	2	2	2	2	2	100%
A8	3	3	3	3	3	3	3	100%

It can be seen that the accuracy is up to 100%, so the model has good robustness, generalizability and resistance to interference.

4. Conclusions

In order to classify the glass, this paper compares and analyzes the chemical composition of high potassium glass and lead-barium glass, and finds that the percentage of silicon dioxide and potassium oxide in high potassium glass is significantly higher, while the percentage of lead oxide and barium oxide in lead-barium glass is very high, and we define the above-mentioned substances with large differences in content as classified substances. A logistic regression analysis was carried out to derive a regression equation between the classified substances and the type of glass, which was used as a basis for determining the type of glass. Based on this, we used the K-means++ clustering algorithm to perform subclass classification, and three subclasses were divided under each glass type. Finally, the compactness, P-value and F-value were selected to measure the reasonableness of the subclass delineation, and the data were perturbed during the sensitivity analysis, resulting in an accuracy of 100% for the model when the perturbation range was 30%.

For the identification of the glass, the classification was carried out using the model developed to give the classification results. On this basis, the data was perturbed and the variable values were scaled randomly within the range (-130%, +130%) and the results were observed in comparison to the results without perturbation, resulting in an accuracy of 100% after perturbation, which demonstrates the robustness and generalizability of the model.

References

[1] Wang Dong, Wen Rui, Zhu Yingpei, Hu Xingjun, Li Wenying. Study on glass beads with metal foil layers excavated from Yingpan Cemetery, Yuli County, Xinjiang[J]. Archaeology and Cultural Relics, 2022(04):117-122.

[2] Ji Luoyuan, PJ Cherian. Peacock blue-glazed pottery specimens excavated from Kerala, India [J]. Journal of the Palace Museum, 2022(06):55-67+148. DOI:10.16319/j.cnki.0452-7402.2022.06.004.

[3] Do Khai Ly, Su Miao, Yang Yuanyuan. Propagation and communion - the influence of Chinese culture on decorative arts along the Silk Road [J]. Journal of Donghua University (Social Science Edition), 2022,22(02):37-46. DOI:10.19883/j.1009-9034.2021.0319.

[4] Wang Liyuan. The foreign exchange of Qi from the unearthed artifacts in and around Linzi, the capital of Qi[J]. Hai Dai Journal, 2021(02):148-159.

[5] Pan Ling, Tan Wenyu. The cultural factors of the west in the Xianbei remains of Hulunbuir--and the "grassland silk road" in the Han period[J]. Archaeology, 2022(05):110-120+2.

[6] Ma Qian,Pollard A. Mark,Yu Yifan,Li Zhuanjie,Liao Linling,Wang Long,Li Man,Cai Luwu,Ping Li,Wen Rui. Laser ablation inductively coupled plasma mass spectrometry analysis of potash and m-Na-Al glasses in China- using Kernel methods for trace element analysis[J]. Heritage Science, 2022, 10(1).

[7] Shang Yue. A grassland silk road connects east and west [N]. Liaoning Daily, 2022-01-05(012). doi:10.28534/n.cnki.nlnrb.2022.000037.

[8] Jiang Bo. Ports, shipwrecks and trade goods: archaeological findings and research on the Maritime Silk Road[J].Studies in the history of maritime transport, 2021(04):8-22. DOI:10.16674/j.cnki.cn35-1066/u.2021.04.002.

[9] Huang Qiaohao. From the local to the exotic, experiencing the "lucidity" of sea silk glassware[J]. Collection. Auction, 2021(06):56-61.

[10] An Jiayao. The Silk Road and Glassware[J]. Cultural Heritage, 2021(12):87-92.