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

Zijie Ji *

College of Power and Energy Engineering, Harbin Engineering University, Harbin, Heilongjiang, 150001

* Corresponding author: jizijie2022@163.com

Abstract. In order to investigate the effect of chemical composition on weathering, all post-weathering data were transformed into pre-weathering, for which the basis for the classification of high potassium and lead-barium glasses was inferred using a support vector machine algorithm. Afterwards, a detailed subclass differentiation method was determined for the two main classes of glass based on the K-means cluster analysis method. Further classification is then carried out on the basis of what is known about the classification. To identify unknown glasses, logistic regression was also carried out on the composition of weathered samples to predict the pre-weathering content.

Keywords: Logistic regression model, K-means cluster analysis, Glass identification.

1. Introduction

The Silk Road is an important part of China's and the world's history, with many important early international trades taking place there and greatly contributing to early international exchanges [1-2]. It was the Silk Road that brought China closer to the rest of the world. Today, with the construction of the Belt and Road project, the expansion of international cooperation and foreign trade markets [3-4], and the numerous archaeological discoveries, the history of the Silk Road is even more important to mankind. Valuable physical evidence of the early trade is glass [5].

Glass was first introduced to China from Western Asia. Since its introduction to China, its production techniques have been greatly improved and its composition has become different from that of foreign ones [6]. In essence, however, glass is still made by firing quartz (SiO2), which has a high melting point, with a variety of fluxes at high temperatures, and is a complex silicate system, but in general it can all be seen as a combination of SiO2 and a variety of oxides, i.e., satisfying the general formula:

$$m\mathrm{SiO}_2 \cdot n_1\mathrm{M}_{1,x_1}\mathrm{O}_{y_1} \cdots\cdots n_i\mathrm{M}_{i,x_i}\mathrm{O}_{y_i} \cdot p\mathrm{H}_2\mathrm{O} \tag{1}$$

Where each Mi in most cases indicates a metallic element, sometimes it also refers to a non-metallic element, which has an important influence on the properties of the glass [7-8]. Ancient glass, on the other hand, undergoes slow chemical reactions over the course of its long history, and varies greatly in whether it will weather, or even in its appearance after weathering, depending on the geographical context of its origin, method of manufacture and burial site. Although a variety of methods for classifying glass have been developed in China, however, ambiguities exist between the methods, leading to ambiguous classification and making it difficult to ensure the reliability of the results. Here, in order to explore the influence of chemical composition on weathering, to provide a detailed classification of glass [9], and to improve the ability to discriminate between unknown glasses, we have modelled some ancient Chinese glass objects for analysis [10].

2. Glass classification model building and solving

2.1. Linear differentiation of the type of glass to which it belongs

The two main types of glass in ancient China are high potassium glass and lead-barium glass, representing two different Chinese cultures. In order to facilitate the study of glass artefacts, the type of glass must first be determined.

For the determination of composition, we used the data treated in Problem 1, solving for the content of substances prior to weathering, for the analysis of the 52 types of glass tested as normal, taking into account the exclusion of any interference from factors affecting the physical and chemical properties of the glass. By definition, the broad glass classes are clearly determined by lead, barium and potassium, which are present in the oxides. Let the corresponding oxide content be x,y,z respectively (in %) For any glass, if the content of these three components can be determined by comparing them, the broad category to which they belong can be determined.

A binary linear regression analysis is carried out, based on the principle that the regression plane derived from the predicted observations for each of the two classes of glass is set as $z = ax + by + c$, The unknown constants are calculated as follows.

$$
\begin{aligned}
Q &= \sum_{i=1}^{n}\left(z_i - ax_i - by_i - c\right)^2 \\
\frac{\partial Q}{\partial a} &= -2\sum_{i=1}^{n} x_i\left(z_i - ax_i - by_i - c\right) = 0 \\
\frac{\partial Q}{\partial b} &= -2\sum_{i=1}^{n} y_i\left(z_i - ax_i - by_i - c\right) = 0 \\
\frac{\partial Q}{\partial c} &= -2\sum_{i=1}^{n}\left(z_i - ax_i - by_i - c\right) = 0
\end{aligned}
\tag{2}
$$

Where n denotes the number of points sampled for each broad category and applies to both broad categories. The solution is solved using Python, i.e. by substituting the coordinates of the predicted points (x_i,y_i,z_i) and finding the regression plane for the two categories of scattered points. Then find the plane *f* such that the points on *f* are equally distant from both regression planes and the scatter points can be well distributed on both sides of *f*. The equation for *f* is obtained as z=0.654-0.309*x*+5.680*y*.

Figure 1 is analysed as follows:

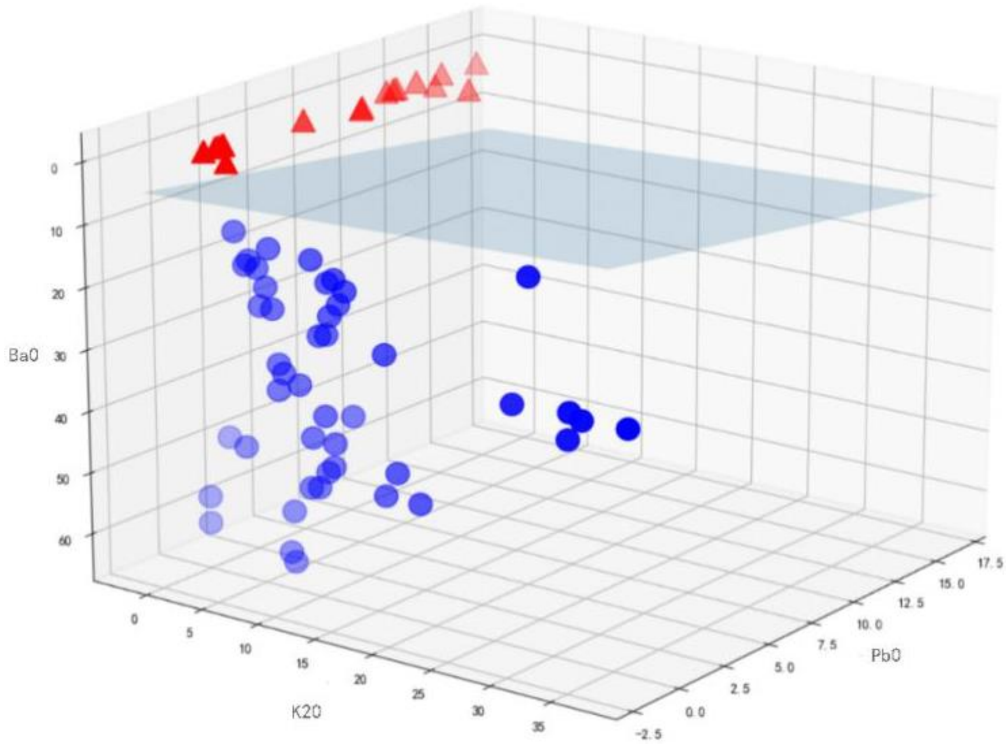

Figure 1. Scatter and discriminant planes for two types of glass data

Obviously, if $z > 0.654-0.309x+5.680y$, it is high potassium glass; if $z < 0.654-0.309x+5.680y$, it is lead-barium glass. If $z = 0.654-0.309x+5.680y$, another test point should be taken and re-analysed. If enough points are analysed, the glass type can be determined.

2.2. Specific divisions of glass

For each category the appropriate chemical composition is chosen to sub-classify it, giving the specific classification method and the results of the classification. We first solved for a reasonable number of classifications for both types of artefacts based on aggregation coefficients.

We adopted three classification methods, and for the clustering model, the clustering results in closer data distance between the same categories and further data distance between different categories. The ratio of between-group dispersion to within-group dispersion is denoted by *F*.

The *F* values for the three classification methods are shown in Table 1 below:

Table 1. Ratio F values of between-group dispersion to within-group dispersion for the three classification methods

Classification method	KMEANS	BIRCH	DBSCAN
F-value	96.5802	93.8338	143.3043

It is clear that the DBSCAN algorithm has the best classification results, indicating that it is reasonable among the above three categories. Therefore, we finally selected the DBSCAN cluster analysis algorithm for classification.

The specific results for the lead-barium type classification are shown in Table 2:

Table 2. Results of DBSCAN cluster analysis of lead-barium type glass artefacts

Chemical composition	Clustering categories (mean ± standard deviation)				F	P
	Category 1(n=19)	Category 2(n=11)	Category 3(n=10)	Category 4(n=5)		
(SiO_2)	27.732±7.604	50.176±6.452	66.254±4.794	16.144±12.003	83.739	0.001
(Na_2O)	0.262±0.646	2.714±3.014	0.981±1.4	0.0±0.0	5.614	0.003
(K_2O)	0.194±0.4	0.217±0.239	0.158±0.103	0.08±0.179	0.274	0.844
(CaO)	3.036±2.048	1.187±1.281	1.217±0.829	1.932±1.169	4.446	0.009
(MgO)	0.813±0.762	0.716±0.562	0.632±0.576	0.0±0.0	2.192	0.104
(Al_2O_3)	3.128±1.522	3.837±1.852	5.529±4.494	1.193±0.332	3.76	0.018
(Fe_2O_3)	0.908±0.977	0.418±0.722	0.786±1.417	0.0±0.0	1.4	0.257
(CuO)	1.098±1.204	2.168±1.813	0.68±0.839	7.275±3.629	19.944	0.001
(PbO)	48.798±7.745	24.669±6.971	16.676±3.105	30.154±1.669	67.279	0.001
(P_2O_5)	4.877±4.87	2.566±3.591	0.531±0.627	4.123±2.897	3.07	0.038
(SrO)	0.437±0.301	0.286±0.253	0.176±0.132	0.58±0.212	3.957	0.014
(BaO)	8.647±5.274	11.046±5.991	5.991±3.139	31.356±3.215	33.51	0.001
(SnO_2)	0.071±0.169	0.0±0.0	0.023±0.074	0.0±0.0	1.106	0.358
(SO_2)	0.0±0.0	0.0±0.0	0.366±1.158	7.163±7.795	11.961	0.001

When the number of categories is 4, the graph line tends to fall more slowly, so we have chosen to classify the lead-barium type glass artefacts into 4 categories.

The high-potassium type classification is given below, as shown in Table 3:

Table 3. K-means clustering analysis results for high potassium type glass artefacts

Chemical composition	Clustering categories (mean ± standard deviation)		F	P
	Category 1(n=9)	Category 2(n=9)		
(SiO_2)	64.91±3.836	90.286±6.627	98.839	0.001
(Na_2O)	0.94±1.447	0.0±0.0	3.801	0.069
(K_2O)	11.034±2.4	2.016±3.297	44.013	0.001
(CaO)	6.499±2.697	1.337±1.44	25.661	0.001
(MgO)	1.158±0.688	0.444±0.606	5.446	0.033
(Al_2O_3)	7.487±2.35	2.79±1.694	23.661	0.001
(Fe_2O_3)	2.358±1.663	0.445±0.746	9.917	0.006
(CuO)	2.877±1.592	1.502±1.148	4.416	0.052
(PbO)	0.414±0.646	0.141±0.338	1.269	0.277
(P_2O_5)	0.585±1.012	0.222±0.667	0.809	0.382
(SrO)	1.549±1.668	0.54±0.462	3.053	0.100
(BaO)	0.049±0.052	0.008±0.024	4.496	0.050
(SnO_2)	0.0±0.0	0.27±0.809	1	0.332
(SO_2)	0.14±0.212	0.0±0.0	3.923	0.065

When the number of categories is 2, the graph line tends to decline more slowly, so we chose to classify the high potassium glass artefacts into 2 categories.

For the sensitivity analysis of the classification results, a random out range was set, the results were brought back to the K-Means algorithm for reclassification and the *F*-value was calculated, and the accuracy of the model was calculated by comparing it with the true value, setting the sensitivity changes to 1%, 2%, 5%, 10%, 20% and 30%, and finally concluding that when the perturbation range was 30%, the accuracy of the model could still be guaranteed. The specific results are shown in Table 4 below:

Table 4. Perturbation error accuracy table

Category	1%	2%	5%	10%	20%	30%
Pre-weathered barium lead glass	100%	100%	100%	100%	100%	100%
Post-weathering barium lead glass	100%	100%	100%	100%	100%	100%
High potassium glass before weathering	100%	100%	100%	100%	100%	100%
High potassium glass after weathering	100%	100%	100%	100%	100%	97%

From the above table it can be seen that for high potassium glass adding random perturbations, the accuracy of the model remains 100% when the perturbation value is less than 30%, indicating good clustering and robustness of the model; for lead-barium type glass, the accuracy of the model is 91% only when the perturbation effect is 30%, the rest of the cases are correct at 100%, considering that the error of measurement at 30% is almost impossible to occur. It is therefore considered that it will not be affected. The accuracy and stability of our model is then reflected to some extent.

For the sensitivity analysis of the classification results, a range of random numbers was set and the results were brought back to the DBSCAN algorithm for reclassification and the specific classification category and test statistic F-value were calculated and compared with the true values to calculate the accuracy of the model, setting the sensitivity changes at 1%, 2%, 5%, 10%, 20% and 30% respectively, as shown in Table 5 below:

Table 5. Perturbation error accuracy table

Category	1%	2%	5%	10%	20%	30%
High Potassium	99.6%	98.2%	98.4%	96.2%	92.9%	90.4%
Lead & Barium	99.5%	97.6%	95.4%	91.6%	86.5%	82.0%

The above table shows that for high potassium glass artefacts with random perturbations, the accuracy of the model is still higher than 90% when the perturbation value is less than 30%, indicating good clustering and robustness of the model; for lead-barium glass artefacts, the accuracy of the model is lower than 85% only when the perturbation effect is 30%, thus reflecting to some extent the accuracy and stability of our model.

3. Development and solution of a glass identification model

3.1. Analysis of unknown glass types

We normalized the variable for the unknown glass to 100% and presumably the pre-weathering portion of the data is shown in Table 6 below.

Table 6. Significant constituent content of unknown glass samples

Heritage number	Surface weathering	K_2O	PbO	BaO
A1	No weathering	0	0	0
A2	Weathering	0	35.6252491	0
A3	No weathering	1.3740216	39.9880698	4.7383539
A4	No weathering	0.8229193	25.2917476	8.6562777
A5	Weathering	0.37141118	12.27664522	2.16823824
A6	Weathering	1.3622607	0	0
A7	Weathering	0.98354074	0	0
A8	No weathering	0.23	21.24	11.34

Substituting into the regression model obtained in the previous section, the types of glass are therefore shown in Table 7, respectively:

Table 7. Classification of unknown glass

Number	Type	Number	Type
A1	High Potassium	A5	Lead Barium
A2	Lead Barium	A6	High Potassium
A3	Barium lead	A7	High Potassium
A4	Barium lead	A8	Lead barium

3.2. Model analysis

Integration learning is a very popular machine learning strategy. The basic starting point is to bring together algorithms and various strategies that can be used for both classification and regression problems. The idea is easy to understand and the most representative algorithm is the random forest.

The aim of the SVM is to draw a line that "best" distinguishes between these two types of points, so that if new points are added later, the line will also be able to make a good classification. The key to this is finding the right classification plane, and support vector machines propose the idea of maximizing the classification spacing.

This model uses a random forest as a crossover decision to train the feature recursive elimination RFE algorithm to obtain suitable destination feature variables that optimally describe the type of glass artefact to which they belong. The trained RFE model was able to obtain the most helpful few components for classification among the 10 and 14 chemical components, as shown in Figure 2 below.

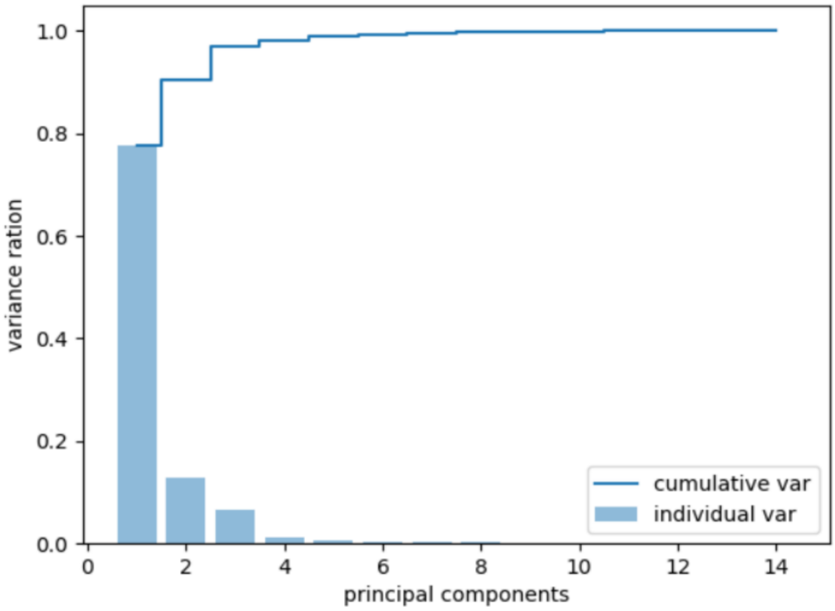

Figure 2. Amount of information contained in the feature vector

The feature variables obtained in this way are able to solve the dichotomous problem for glass artefacts very well. Conversely, if the four feature variables with the best results are removed and the relatively good variables #5 and #6 are selected, the dichotomous classification problem cannot be solved well.

After removing the 4 main components, the remaining 10 main chemical components were then trained using random forest for cross-validation, and the results were re-obtained as shown in Figure 3 below. The best classification results are obtained when two feature variables are selected, and the effect decreases when the number of feature variables continues to be increased.

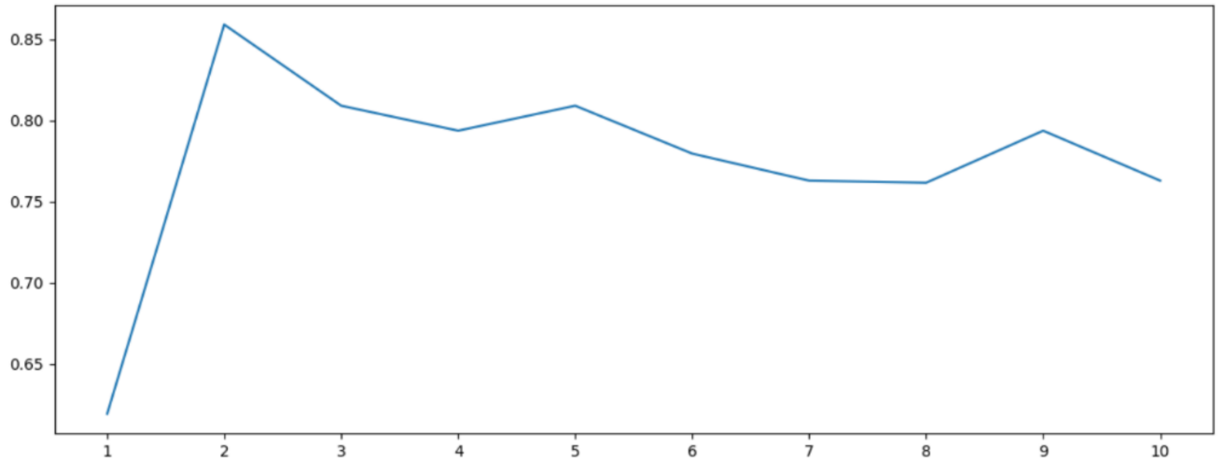

Figure 3. RFE effect after removing the 4 main variables

3.3. Sensitivity analysis

We know that there is an error in the measured sampling point data, and as the total can be estimated from known samples, the quantification of the error is determined from the pre-weathering data derived in the previous section.

Considering the measurement error, which varies by component content of the glass, but the error is independent of the type of glass, only samples with a certain type of component (i. e. not zero in content) in all samples are considered, and the solution mean and variance are:

$$\mu_i = \frac{1}{n_i} \sum_{\substack{j=1 \\ x_{ij} \neq 0}}^{63} \left(x_{ij} - \overline{x_j} \right)$$
$$\sigma_i^2 = S_i^2 = \frac{1}{n_i - 1} \sum_{\substack{j=1 \\ x_{ij} \neq 0}}^{63} \left(x_{ij} - \overline{x_j} \right)^2 \tag{3}$$

Therefore, the content values for each type of ingredient follow a normal distribution. For the determination error values, the distribution is normal. For this reason, it can be considered that a certain component contains a certain amount of error and the error does not represent data utilization, so the rest of the data is no longer varied. Substituting the errors for the analysis, the results are shown in Table 8 below at different levels of significance α:

Table 8. Perturbation error accuracy table

Category	1%	2%	5%	10%	20%	30%
Pre-weathered barium lead glass	100%	100%	100%	100%	100%	100%
Post-weathering barium lead glass	100%	100%	100%	100%	100%	100%
High potassium glass before weathering	100%	100%	100%	100%	100%	100%
High potassium glass after weathering	100%	100%	100%	100%	100%	100%

When the solution is completed, the classification of the glass remains unchanged, indicating that the classification is not easily disturbed by chance errors. The classification is therefore reasonable and resists chance interference well.

4. Conclusions

For the glass classification problem, a known logistic regression model was used to solve the problem, converting all post-weathering data to pre-weathering, for which the basis for the classification of high potassium glass and lead-barium glass was inferred using a support vector machine algorithm. Afterwards, a detailed subclass differentiation method for the two major glass classes was determined based on K-means cluster analysis. In addition, we performed model tests within a 30% deviation to determine the superiority of K-means cluster analysis.

For the glass identification problem, again the glass composition given by the test data in the samples to be tested was first transformed and equiprimordially normalized to 100%, while logistic regression was also performed on the composition of the weathered samples to predict the pre-weathering content. The identification of the unknown glass was then completed by content values based on the main components utilised in the classification method in Section 2.

References

[1] Wu Changfeng. Isotope archaeology traces ancient Silk Road glass production and distribution patterns [N]. Science and Technology Daily, 2022 - 11 - 25 (003).

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

[3] An exhibition of ancient glassware from the Silk Road Museum of Art at Dunhuang [J]. Cultural Identification and Appreciation, 2021 (15): 102.

[4] Wang Xiaofu. The fragrant silk road: early exchanges between Oman and China: an answer to the question of the "Silk Road"[J]. Journal of Tsinghua University (Philosophy and Social Science Edition), 2020, 35 (04): 1 - 14+211. DOI: 10.13613/j.cnki.qhdz.002956.

[5] Tan X. The transmission and influence of Islamic glass on the Silk Road (8th-16th centuries) [D]. Jinan University, 2020. DOI: 10.27167/d.cnki.gjinu.2020.000412.

[6] Chen Shuyu, Hou Zhili. An analysis of the ancient Silk Road and ancient Chinese glass [J]. China Ethnic Expo, 2019 (05): 87 - 88.

[7] Li Fang. Research on Han Dynasty imitation glass bronze vessels[D]. Hunan University, 2019. doi: 10.27135/d.cnki.ghudu.2019.001711.

[8] Tian Ye. The "Silk Road on Ice" sets sail from Yamal [J]. China Oil Enterprise, 2018 (08): 27 - 32+2.

[9] Yang Hong. Western artworks introduced by the Silk Road [J]. Art Market, 2018 (06): 62 - 65.

[10] Xiong Zhaoming. Archaeological findings from the Hepu Port on the Han Dynasty Maritime Silk Road [J]. Democracy and Science, 2018 (01): 25 - 28.