A study on the classification of ancient glassware based on K-means clustering algorithm

Yiming Liu [1,*], Chenxu Li [2]

[1] School of Traffic and Transportation, Shijiazhuang Tiedao University, Shijiazhuang, China, 050043

[2] School of Safety Engineering and Emergency Management, Shijiazhuang Tiedao University, Shijiazhuang, China, 050043

* Corresponding Author Email: 1421214509@qq.com

Abstract. In this paper, a mathematical model based on cluster analysis is developed for the composition of ancient glass products, and the classification rules of high potassium glass and lead-barium glass as well as the method of subclassing them are given. In this paper, scatter plots were used to screen out the chemical constituents in high potassium and lead-barium glasses that varied considerably at different sampling points. The contents of these chemical components were systematically clustered separately and the aggregation coefficient line graphs were plotted to obtain the number of classes of clusters. The K-means algorithm was then used to obtain a classification of high potassium glass into three categories using alumina content and lead-barium glass into two categories using barium oxide content. Finally, the model was subjected to reasonableness and sensitivity analysis, and the results showed that the sensitivity of the model was low.

Keywords: K-means, Cluster analysis, Glass classification.

1. Introduction

Glass was valuable physical evidence of trade during the Silk Road period, and its production process originated from the West Asian and Egyptian regions, but the chemical composition of the products made from glass can vary due to the origin of the raw materials [1-2].

Glass in the refining in order to reduce the melting temperature, often add grass ash, natural alkali, saltpeter and lead ore and other fluxes, and with the stabilizer limestone used together. The chemical composition of the final product varies depending on the fluxes added when the glass is fired [3]. Ancient glass is subject to weathering as a result of the environment in which it is buried [4]. During the weathering process, a large number of internal elements are exchanged with environmental elements, and the proportions of the components change, leading to a compromise in the correct determination of their category [5-6]. For artefacts with no weathering on the surface, shallower local weathering cannot be excluded; for weathered artefacts, there may also be unweathered areas on the surface [7]. There is a body of data relating to ancient glasswork in China, divided into two types: high potassium glass and lead-barium glass [8]. The sum of the proportions of the components in the original data should be 100%, but due to testing methods and other factors, it is possible that the sum of the components may not be 100% [9]. The classification of high potassium glass and lead-barium glass is analyzed; for each category, the appropriate chemical composition is chosen to classify the glass into subcategories, the specific classification method and results are given, and the reasonableness and sensitivity of the classification results are analyzed [10].

2. The basic fundamental of single-objective optimization models

The chi-square test is used to determine whether there is a correlation between two categorical variables based on how well the actual frequencies match the theoretical frequencies. The original hypothesis of the chi-square test H_0 is that the observed frequencies are not significantly different from the expected frequencies. The chi-square statistic is calculated by the formula

$$\chi^2 = \sum_{i=1}^{k} \frac{(A_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(A_i - np_i)}{np_i} \tag{1}$$

Where A_i indicates the observed frequency of class i , E_i indicates the desired frequency of class i , n indicates the total frequency and p_i indicates the desired frequency of class i .

Conditions for the use of the results of the chi-square test of the column table.

(1) Analysis using Pearson's chi-square when the sample size $n \geqslant 40$, all expected frequencies $T \geqslant 1$ and the number of cells at $T < 5$ does not exceed 20%.

(2) All cases not meeting the above are analyzed using the $Fisher$ precision test.

The sample size was greater than 40 and all expected frequencies were $T \geqslant 5$, so Pearson's chi-square was used for significance analysis. p-values were less than 0.1, indicating a strong correlation between surface weathering and category at the 10% confidence level, i.e. category has a significant effect on whether weathering is present. Further tau-y coefficients for Goodman and Guru ska can be calculated as the correlation coefficient between surface weathering and category. A chi-square test for the relationship between surface weathering and ornamentation was conducted using SPSS. $T < 5$ has a grid of 33.3% , which has exceeded 20% , and therefore an exact test of significance was conducted using $Fisher$. The p-value is less than 0.1, indicating a strong correlation between surface weathering and ornamentation at a 10% confidence level, i.e., ornamentation has a significant effect on whether or not weathering is present. Further tau-y coefficients for Goodman and Galuski can be calculated, which are the correlation coefficients between surface weathering and ornamentation. A chi-square test for the relationship between surface weathering and ornamentation was conducted using SPSS and the results were as follows Expected frequencies exist $T < 1$, so an exact test was used for significance analysis using $Fisher$. A p-value greater than 0.1 indicates that there is no correlation between surface weathering and color at the 10% confidence level, i.e., color has no significant effect on whether or not weathering is present.

The Mann-Whitney U test is a non-parametric test that can be used to test both the significance of data that fit a normal distribution and the significance of data that do not fit a normal distribution. The original hypothesis of the Mann-Whitney U test H_0 is that the means of the two aggregates are not significantly different. The formula for its statistic is.

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 \\ U_2 &= n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2 \end{aligned} \tag{2}$$

Where n_1 is the weathered sample size, n_2 is the unweathered sample size, W_1 is the sum of the ranks of the weathered samples and W_2 is the sum of the ranks of the unweathered samples.

The results of the Mann-Whitney U test for high potassium glass are shown in Table 1.

Table 1. Mann-Whitney U test results for high potassium glass

Chemical composition	P-value	Significance	Chemical composition	P-value	Significance
Silicon dioxide	0.000	Significant	Copper oxide	0.291	Not significant
Sodium oxide	0.437	Not significant	Lead oxide	0.053	Not significant

Potassium oxide	0.002	Significant	Barium oxide	0.291	Not significant
Calcium oxide	0.024	Significant	Phosphorus pentoxide	0.010	Significant
Magnesium oxide	0.013	Significant	Strontium oxide	0.102	Not significant
Aluminium oxide	0.000	Significant	Tin oxide	0.820	Not significant
Iron oxide	0.024	Significant	Sulphur dioxide	0.437	Not significant

The results of the Mann-Whitney U test for lead-barium glass are shown in Table 2.

Table 2. Mann-Whitney U test results for lead-barium glass

Chemical composition	P-value	Significance	Chemical composition	P-value	Significance
Silicon dioxide	0.000	Significant	Copper oxide	0.052	Not significant
Sodium oxide	0.011	Significant	Lead oxide	0.000	Significant
Potassium oxide	0.143	Not significant	Barium oxide	0.521	Not significant
Calcium oxide	0.002	Significant	Phosphorus pentoxide	0.000	Significant
Magnesium oxide	0.853	Not significant	Strontium oxide	0.031	Significant
Aluminium oxide	0.024	Significant	Tin oxide	0.620	Not significant
Iron oxide	0.745	Not significant	Sulphur dioxide	0.206	Not significant

The chemical components that did not change significantly before and after weathering were removed, and descriptive statistics were performed on the mean values of the remaining chemical components to find the change in the content of each chemical component after weathering compared to before weathering. The mean statistics for the high potash glass are shown in Table 3.

Table 3. Statistical results for mean values of high potassium glass

Chemical composition	Unweathered content	Content after weathering	Value of change in content after weathering
Silicon dioxide	67.984	93.963	25.979
Potassium oxide	9.331	0.543	-8.788
Calcium oxide	5.333	0.870	-4.463
Magnesium oxide	1.079	0.197	-0.883
Aluminium oxide	6.620	1.930	-2.000
Iron oxide	1.932	0.265	-1.475
Phosphorus pentoxide	1.403	0.280	-1.123

As can be seen from the above table, the silica content of the high potassium glass is significantly higher after weathering than before weathering, and the content of the rest of the chemical components has decreased. The mean statistics for lead-barium glass are shown in Table 4.

Table 4. Statistical results for mean values of lead barium glass

Chemical composition	Unweathered content	Content after weathering	Value of change in content after weathering
Silicon dioxide	54.660	24.913	-29.747
Sodium oxide	1.683	0.216	-1.466
Calcium oxide	1.320	2.695	1.375
Aluminium oxide	4.456	2.970	-1.486
Lead oxide	22.085	43.314	4.228
Phosphorus pentoxide	1.049	5.277	0.150
Strontium oxide	0.268	0.419	-0.152

As can be seen from the table above, the content of silica, sodium oxide, aluminum oxide and strontium oxide decreased after weathering compared to before weathering, while the content of all other chemical components increased to varying degrees.

3. Results

3.1. The establishment of simulation model

In high potassium glass, the contents of the three chemical components silica, potassium oxide and alumina vary considerably at different sampling points, and in lead barium glass, the contents of the three chemical components silica, lead oxide and barium oxide vary considerably at different sampling points, so we systematically clustered these three chemical components in high potassium and lead barium glass one by one respectively, and then obtained the optimum by means of the aggregation coefficient fold diagram The optimal number of clusters was then obtained by means of the aggregation coefficient line graph. The aggregation coefficients were calculated as follows.

(1) Let a total of samples be aggregated into K classes, then the degree of distortion of the k class is

$$J_K = \sum_{i \in C_k} |x_i - u_k|^2 \tag{3}$$

Where indicates the C_k k class and u_k indicates the location of the center of gravity of the k class.

(2) The sum of the distortion levels of the K classes are

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} |x_i - u_k|^2 \tag{4}$$

3.2. Analysis of experimental results

We found that alumina clustered best among the three chemical compositions of high potassium glass and barium oxide clustered best among the three chemical compositions of lead-barium glass by k-means++ clustering using SPSS with the criterion that the distance between groups was as large as possible and the distance within groups was as small as possible. The first cluster center of alumina is 3.04, the second cluster center is 6.21 and the third cluster center is 10.54, while the first cluster center of barium oxide is 7.78 and the second cluster center is 29.89. The clustering effect is better when the magnitude of data change is larger, so it is reasonable to choose the chemical composition

with larger change as the data for our clustering. rationality. The larger the intergroup distance and the smaller the intra-group distance, the more obvious the clustering effect, so it is reasonable to use the largest possible inter-group distance and the smallest possible intra-group distance as our criteria. Finally, sensitivity analyses were carried out for high potassium glass and lead-barium glass, and the results are shown in Tables 5 and 6 respectively.

(1) High Potassium Glass

Table 5. Sensitivity analysis of changes in alumina content in high potassium glass

Percentage of change	Number of artefacts whose classification results have changed	Percentage of change	Number of artefacts whose classification results have changed
-6%	0	6%	0
-12%	3	12%	2
-18%	3	18%	2

(2) Barium lead glass

Similarly, the results of the sensitivity analysis for changes in the barium oxide content of lead barium glass can be obtained as follows

Table 6. Sensitivity analysis of changes in barium oxide content in lead barium glass

Percentage of change	Number of artefacts whose classification results have changed	Percentage of change	Number of artefacts whose classification results have changed
-6%	0	6%	0
-12%	0	12%	1
-18%	0	18%	1

As can be seen from the two tables above, the number of artefacts whose categories changed when the content of the chemical composition changed was small, indicating that the clustering model was less sensitive and more robust.

4. Conclusions

This paper describes the classification process of high potassium glass and lead-barium glass with a certain degree of accuracy and ingenuity. A mathematical model was developed using cluster analysis to give the classification pattern of high potassium glass and lead-barium glass and the method of subclassifying them respectively, and the correlations between the chemical components in the different glass categories were obtained by Spearman correlation analysis. Finally, the model was analyzed for reasonableness and sensitivity, and the results showed that the sensitivity of the model was low.

References

[1] Chi, Shengchao, Zhang, Zhenyuan. Statistical and predictive analysis of drainage water quality characteristics of drainage tunnels based on normal test[J]. Northern Transportation,2021(02):88-91+94.

[2] Huang Y. Research on the application of distance discrimination method in the classification of highway tunnel surrounding rock[J]. Resource Information and Engineering,2019,34(06):47-49.

[3] PENG Xiaofeng, ZHANG Xuan, PENG Yan, HUANG Shunli, LIU Suhua, Aniyer Hailili, LI Rong. Screening of new quality lines of medium-length staple land cotton based on systematic clustering method[J]. Cotton Science,2022,44(04):26-32.

[4] Liang Zhengyou,Wang Lu,Li Xuanang,Yang Feng. Multi-view point cloud alignment based on K-means ++[J]. Computers and Modernization,2022(02):97-101.

[5] Yu Qun, Huo Xiaodong, He Jian, Li Lin, Zhang Jianxin, Feng Yuyao. Trend prediction of power grid outage in China based on Spearman's correlation coefficient and system inertia[J/OL]. Chinese Journal of Electrical Engineering:1-12.

[6] Yan Shaoqiang,Liu Weidong,Yang Ping,Wu Fengxuan,Yan Zhe. A variety of cluster sparrow search algorithms based on K-means clustering[J]. Journal of Beijing University of Aeronautics and Astronautics:1-13.

[7] Zhang Yukun. Research on customer segmentation of e-commerce students based on K-Means cluster analysis[J]. Mall Modernization,2022,(08):33-35.

[8] Wu B., Yu D.. Deep convolutional neural network-based classification and detection of mobile phone glass cover surface defects[J]. Software Engineering,2021,24(12):6-10.

[9] Zhang Jian, Hao Nannan. Gemological classification and identification of meteorite glass[J]. Shandong Chemical Industry,2021,50(17):160-161.

[10] Liu Z. X., Deng P. F., Pan S. H., Luo W. D., Zhang Shilin, Li Tao Ming. Classification and research status of fire resistant glass[J]. Guangzhou Chemical Industry,2021,49(15):16-18.