K-means++ clustering-based composition analysis and identification model for glass products

Zhuoyuan Li *, Beier Kang, Ruijie Zhang

School of Automation, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi, 710121

* Corresponding Author Email: hhuoya@163.com

Abstract. Ancient glass is susceptible to weathering by environmental influences, causing changes in its compositional proportions and thus affecting the correct judgment of its elemental categories. In this paper, the study on the analysis and identification of ancient glass composition applies gray correlation analysis, K-means++ cluster analysis and principal component analysis to analyze the relationship between weathering on the surface of cultural relics and their glass types, ornamentation, and color; the content pattern of chemical composition with and without weathering on the surface of cultural relics is statistically analyzed by glass type, and the content of its chemical composition before weathering is predicted based on the weathering point detection data. Will be high potassium glass, and lead barium glass classification laws for analysis; appropriate chemical composition of each category for subcategory division, giving the division method and division results and analysis of its reasonableness and sensitivity. Analysis of the chemical composition of unknown categories of artifacts, identification of the types, and the sensitivity of the results are analyzed. The correlations between the chemical composition of glass artifacts of different categories are analyzed and the differences in their chemical composition correlations are compared.

Keywords: Gray correlation analysis; Normal distribution; K-means++; Cluster analysis; Principal component analysis

1. Introduction

The early glass was introduced to China through the Silk Road, and our ancient glass was made by absorbing its technology and using our native material, quartz sand, with co-solvents. Ancient glass is susceptible to weathering due to the burial environment and the resulting changes in internal elements, which affects the correct determination of its category.

The color of glass has a certain relationship with its chemical composition. Some metallic oxides or salts mixed with ancient glass during manufacture can give the glass a different color, and these metallic oxides can have an effect on the degree of weathering of the glass in the burial site because of the different degrees of oxidation. It can be seen that there is a correlation between the color of glass and its chemical composition, and both of them have a certain influence on the weathering of glass[1]. According to the available data related to ancient glass products in China, the chemical composition of the artifact samples and other testing means will be divided into high potassium glass and lead-barium glass. The production of ancient glass, with ore, also known as silica, as the main raw material, and the choice of different co-solvents lead to different chemical compositions, and among them, the use of metal substances such as copper powder, iron shavings, and Dann lead as coloring agents, can make the glass present different colors[2].

In this paper, based on the classification information of cultural relics and the proportion of the main components to analyze the weathering of the surface of cultural relics and its glass type, decoration, and color relationship; through the type of glass, statistical analysis of the surface of cultural relics with and without weathering chemical composition content pattern, according to the weathering point detection data, the prediction of its weathering before the chemical composition content[3]. Will be high potassium glass, and lead barium glass classification laws for analysis; appropriate chemical composition of each category for sub-category division[4], giving the division method and division results and analysis of its reasonableness and sensitivity[5]. Analyze the chemical composition of artifacts of unknown categories, identify the types, and analyze the sensitivity of the results[6].Analysis of the correlation between the chemical composition of glass

artifacts of different categories and comparison of the differences in their chemical composition correlations[7].

2. Method

2.1. Correlation coefficient

The reference series was first selected. We selected the number of artifacts to be studied as the reference series.

$$x_0 = \{x_0(k)|k = 1,2,\dots,n\} = [x_0(1), x_0(2), \dots, x_0(58)] \tag{1}$$

where k denotes the moment[8].There are three factors influencing whether the surface is weathered or not, so there are three comparison series

$$x_0 = \{x_0(k)|k = 1,2,\dots,n\} = \{x_i(1), x_i(2), \dots, x_i(n); i = 1,2,3 \tag{2}$$

According to the definition of correlation coefficient, the correlation coefficient of the comparison series xi to the reference series x0 at time k is

$$\xi_i = \frac{min_s min_t |x_0(t) - x_s(t)| + \rho max_s max_t |x_0(t) - x_s(t)|}{|x_0(k) - x_i(k)| + \rho max_s max_t |x_0(t) - x_s(t)|} \tag{3}$$

where ρ is the discrimination coefficient, and in this paper, we let ρ = 0.5.

The correlation degree is the concentration of the correlation coefficients at each moment into an average value, and likewise the concentration of information that is too scattered. Using the concept of correlation degree, we can factorize various problems. According to the definition of correlation, the correlation of the series xi concerning the reference series x0 is:

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k) \tag{4}$$

where n = 58

Due to the small amount of data thea on chemical composition and the large number of 0 values, we chose the weighted average method to predict the chemical composition before weathering. Depending on the type, to predict the chemical composition before weathering for high potassium glass and lead-barium glass, respectively.

2.2. K-means++ Systematic Clustering Model

K-means++ clustering algorithm process: first you need to specify the number of classes K, then randomly select K data objects as the initial cluster center, then calculate the distance of the remaining data objects to the K initial cluster center, the data object into the nearest center of the cluster class, then adjust the new class and recalculate the center of the new class, this process will be repeated to see if the center converges, if convergence or reach the number of iterations then stop the cycle[9].

2.3. Principal component analysis

There is a strong correlation between variables and variables, and principal component analysis is used to transform multiple indicators into a few principal components as a way to reduce the dimensionality of the data, and for the subcategory classification of glass types[10].

3. Results and Discussion

3.1. Weathering

From Table 1, it can be seen that whether the surface of glass artifacts is weathered or not is more closely related to its type.

Table 1. Calculation results of correlation degree.

Evaluation items	Relevance	Ranking
Type	0.938	1
Ornamentation	0.883	2
Color	0.764	3

Using the same gray correlation analysis, the relationship coefficients for each component of high potassium glass and lead-barium glass were solved separately as shown in Table 2 and Table 3.

Table 2. Relationship coefficient of each component of high potassium glass.

Ingredients	Number of relevant contacts	Ingredients	Number of relevant contacts
(SiO_2)	0.9531	(CuO)	0.9434
(Na_2O)	0.8555	(PbO)	0.8828
(K_2O)	0.9422	(BaO)	0.8632
(CaO)	0.9357	(P_2O_5)	0.9317
(MgO)	0.9305	(SrO)	0.8802
(Al_2O_3)	0.9637	(SnO_2)	0.8644
(Fe_2O_3)	0.9242	(SO_2)	0.8550

Table 3. The relationship coefficient of each component of lead barium glass

Ingredients	Number of relevant contacts	Ingredients	Number of relevant contacts
(SiO_2)	0.9697	(CuO)	0.9248
(Na_2O)	0.8915	(PbO)	0.9518
(K_2O)	0.9166	(BaO)	0.949
(CaO)	0.9335	(P_2O_5)	0.9089
(MgO)	0.9263	(SrO)	0.9443
(Al_2O_3)	0.9478	(SnO_2)	0.8858
(Fe_2O_3)	0.9141	(SO_2)	0.8884

For high potassium glass, the silica content before and after weathering is on the rise, potassium oxide and other content is on the decline, and some of the content such as iron oxide has a small difference in change, no obvious trend; for lead-barium glass, the silica content before and after weathering is on the decline, lead oxide, barium oxide, and other content is on the rise, and the same part of the content as high potassium glass such as tin oxide has no obvious trend.

3.2. Glass types

Table 4. Comparison of clustering results and actual values for high potassium glass.

Artifact Number	Clustering results	Actual value
Weathering	7 9 10 12 22 26 43	7 9 10 12 22 27
Unweathered	1 3 3 4 5 6 6 13 14 20 16 18 21	1 3 3 4 5 6 6 13 14 20 16 18 21

Table 5. Comparison of clustering results and actual values of lead barium glass.

Artifact Number	Clustering results	Actual value
Weathering	2 8 8 11 19 26 26 27 34 36 38 39 40 41 48 49 50 51 51 52 54 54 56 57 58	2 8 8 11 19 26 26 34 36 38 39 40 41 43 43 48 49 50 51 51 52 54
Unweathered	23 24 25 28 29 30 30 31 32 33 35 37 42 42 44 45 46 47 49 50 53 55	23 24 25 28 29 30 30 31 32 33 35 37 42 42 44 45 46 47 49 50 53 55

The comparison of Table 4 and Table 5 shows that the correct rate of clustering results for high potassium glass is 94.7%, and the correct rate of clustering results for lead-barium glass is 95.8%.

Table 6. Table of results of principal component analysis.

Eigenvector	x_1	x_2	x_3	.....	$x14$
Eigenvalue	4.026102	3.436356	2.044539	.....	$-5.382857e^{-16}$
Contribution rate	0.287578	0.245454	0.146038	.....	$-3.844898e^{-17}$
Cumulative contribution rate	0.287578	0.533032	0.679071	.....	1

Table 6 is table of results of principal component analysis. The principal components corresponding to the eigenvalues with a cumulative contribution of more than eighty percent are generally taken. Based on the existing types, i.e., high potassium unweathered, high potassium weathered, lead-barium unweathered, and lead-barium weathered as one class, we selected the appropriate chemical components for subclass classification (two-class classification) by principal component analysis.

The results of the two-class division of chemical components were examined using K-means++ cluster analysis, and the chemical component content was clustered by SPSSPRO using high-potassium unweathered as an example.

where category 1 and category 2 include the results of mean ± standard deviation, F indicates the test result, and the P value is the significant result. P value is the significant result. The analysis of whether the P value is less than 0.05, if it is less, it is significant and can indicate that there is a significant difference between the two groups so that the sample categories can be classified according to the form of mean + standard deviation. From the above table, it can be seen that the significance of silica is less than 0.05 and the P-value presents the lowest, so silica was selected as a classification criterion for subclassification[11], and Figures 1-4 show the classification results of the visual analysis. Table 7 is the cluster analysis results.

Table 7. Cluster analysis results.

Chemical composition	Category 1 (n = 9)	Category 2 (n = 3)	F	P
(SiO_2)	63.624±3.558	81.063±5.368	43.059	0.000***
(Na_2O)	0.927±1.427	0.0±0.0	1.186	0.302
(K_2O)	10.818±2.37	4.87±4.718	8.897	0.014**
(SO_2)	0.136±0.205	0.0±0.0	1.226	0.294
(SnO_2)	0.0±0.0	0.787±1.363	3.75	0.082*
(SEO)	0.048±0.051	0.023±0.04	0.551	0.475
(BaO)	0.579±1.001	0.657±1.137	0.013	0.912
(P_2O_5)	1.523±1.652	1.04±0.354	0.238	0.636
(PbO)	0.41±0.64	0.417±0.52	0	0.987
(CuO)	2.819±1.565	1.353±1.714	1.897	0.198
(Fe_2O_3)	2.312±1.643	0.79±1.368	2.057	0.182
(Al_2O_3)	7.349±2.346	4.433±1.603	3.891	0.077*
(MgO)	1.133±0.672	0.917±0.809	0.215	0.653
(CaO)	6.363±2.64	2.24±2.363	5.714	0.038**

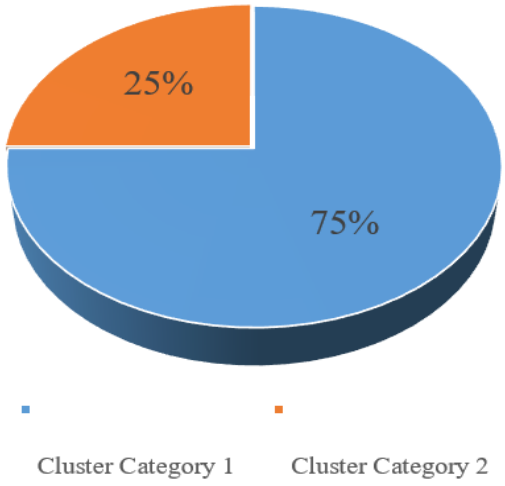

Figure 1. High potassium unweathered.

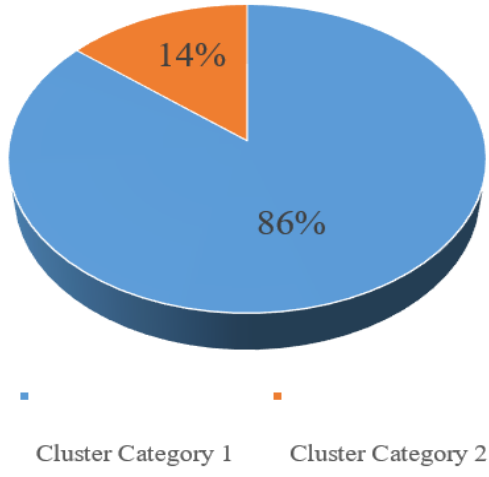

Figure 2. High potassium weathering.

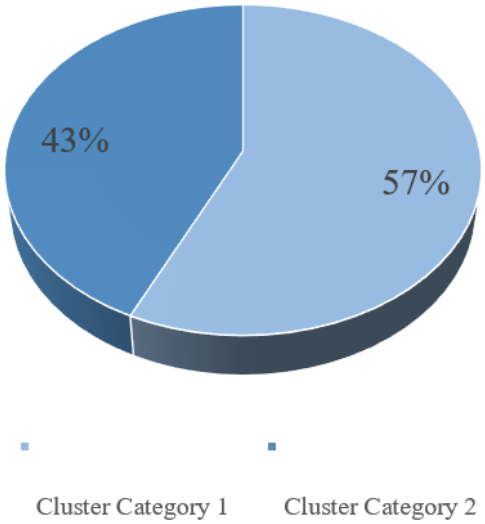

Figure 3. Unweathered lead and barium.

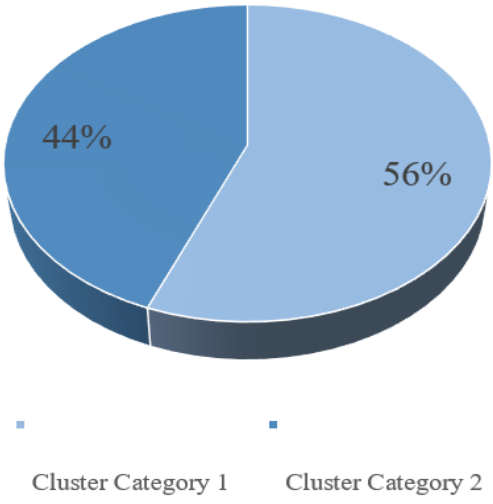

Figure 4. Weathering of lead and barium.

From the above classification criteria, combined with the visualization graphs, the subclasses were classified using silica, with cluster category 1 being the multi-silica category and cluster category 2 being the less-silica category. Finally, sensitivity analysis was performed for the above data, and data perturbation was performed. The sensitivity values were set to 1%, 2%, 5%, 10%, 20%, and 30% to observe the changes in the model results, and the final calculation results showed that the accuracy of the model was 91.3043% when the perturbation range was 30%.

The accuracy of the model was still 100% when the perturbation value was less than 30%, which indicated that the model had a good clustering effect. The correct rate is the same as that of the high potassium class glass when the lead-barium class glass is weathered, both are 100%, while for the lead-barium class glass without differentiation, the model accuracy is 91.3043% when the perturbation effect is 30%, which also reflects the stability of the model to some extent.

3.3. Sensitivity Analysis

The sensitivity analysis and data perturbation were performed by observing the results compared with the results before the changes were made. The values of sensitivity changes were set to 1%, 2%, 5%, 10%, 20%, and 30%, respectively, to observe the changes in the model results with the control of other parameters unchanged, as shown in Table 8.

Table 8. Glass artifact inference table.

Artifact Number	1% Perturbation	2% Perturbation	.....	30% Perturbation
A1	100%	100%	.....	100%
A2	100%	100%	.....	100%
A3	100%	100%	.....	100%
A4	100%	100%	.....	100%
A5	100%	100%	.....	100%
A6	100%	100%	.....	100%
A7	100%	100%	.....	100%
A8	100%	100%	.....	100%

The final calculation results show that the perturbation range is from 1% to 30%, and the accuracy of the result is 100%, which shows that the model has good stability and adaptability.

3.4. Variability

Table 9. Glass artifact inference table.

Evaluation items	Relevance	Ranking
(Al_2O_3)	0.942	1
(K_2O)	0.937	2
MgO)	0.915	3
(CaO)	0.911	4
(CuO)	0.908	5
(Fe_2O_3)	0.9	6
(P_2O_5)	0.892	7
(SEO)	0.846	8
(PbO)	0.837	9
(BaO)	0.811	10
(SnO_2)	0.804	11
(SO_2)	0.793	12
(Na_2O)	0.791	13

Table 9 is the glass artifact inference table. The degree of correlation between the content of different chemical components and silica varies, and the greater the correlation, the stronger the correlation between the parent sequence and the child sequence, from which the correlation between different chemical components can be analyzed.

Table 10. Standard deviation of correlation coefficients for different species.

Category	The standard deviation of the relationship coefficient
High potassium unweathered	0.0974
High potassium weathering	0.1792
Lead barium unweathered	0.0714
Lead barium weathering	0.0688

Table 10 is the standard deviation of correlation coefficients for different species. The standard deviation of the correlation coefficient of 0.1792 for the high potassium weathering and 0.0974 for the unweathered category are higher than that of the weathered and unweathered categories of lead-barium glass, which leads to the conclusion that the chemical composition of lead-barium glass is more relevant than that of high potassium glass.

3.5. Discussion

To observe more intuitively whether the weathering change pattern, visualization plots were used for the analysis to observe the change in the data of different types of glass. The use of gray correlation analysis does not require a large sample size and is equally applicable in the case of samples with or without regularity. Moreover, its calculation is small and easy to use, and there is no discrepancy between quantitative results and qualitative analysis results.

K-means++ clustering analysis is easy to understand, ensures better scalability when dealing with large data sets, and its algorithm complexity is low, resulting in better subclassification results.

The accuracy of the results obtained by weighted average prediction is higher.

In the subclassification process, only chemical components are used for classification, and multiple classification problems such as color and ornamentation are not considered.

Principal component analysis cannot contain all the original data after dimensionality reduction, and generally carries ambiguity.

4. Conclusions

In this paper, we analyze the weathering of the surface of cultural relics and its glass type, decoration, and color relationship; through the glass type, statistical analysis of the content pattern of the chemical composition of the surface of cultural relics with and without weathering, according to the weathering point detection data, the prediction of its chemical composition content before weathering. will be high potassium glass, lead barium glass classification laws for analysis; appropriate chemical composition of each category for sub-category division, giving the division method and division results, and analysis of its reasonableness and sensitivity. Analyze the chemical composition of artifacts of unknown categories of annexes, identify the types, and analyze the sensitivity of their results. Analysis of the correlation between the chemical composition of glass artifacts of different categories and comparison of the differences in their chemical composition correlations.

References

[1] Strugaj Gentiana, Herrmann Andreas, Rädlein Edda. AES and EDX surface analysis of weathered float glass exposed in different environmental conditions[J]. Journal of Non-Crystalline Solids,2021,572.

[2] Cheng Qian, Wang Bo, Guo Jinlong, Cui Jianfeng. Study on the composition characteristics of glassware excavated from the ancient cemeteries of the Silk Road and the Ancient Kingdom [J]. Glass and Enamel, 2012,40(02):21-29.

[3] Guo Weimin. The Combination of Grey Theory and Weighted Average Method to Predict the Sales Volume of Household Air Conditioning Products [J]. Inner Mongolia Science and Technology and Economy, 2021 (15): 61-62.

[4] Luo Chaoxi, He Lifang, Yang Changzhou, Yang Shaping, Huang Bin, Ma Guangyu, Huang Hongwei. K-means based particle size measurement of metallic mineral optical flakes [J]. Nonferrous Metal Engineering,2022,12(07):125-131.

[5] Yin Junqing, He Yalong, Liu Shan, Chen Yongkang. Sensitivity analysis of multi-round hole auxiliary nozzle spray hole structure parameters [J]. Silk,2022,59(08):54-61.

[6] Nie Zuoxing, Yu Deji, Li Rong, Yao Lingyun. Global sensitivity analysis of body noise transfer function based on Sobol's method[J]. China Mechanical Engineering,2012,23(14):1753-1757.

[7] Xing Tao, Yong Yi, Hou Jiang, Wu Yi, Wu Di, Liu Hengbo. Comprehensive assessment of water quality based on principal component analysis and hierarchical cluster analysis [J]. Sichuan Environment, 2022,41 (04): 131-139. DOI: 10.14034/j.cnki.schj.2022.04.019.

[8] Huang Renquan. Spatial and temporal evolution and gray correlation analysis of high-quality development level in the Yellow River Basin: empirical evidence based on 2000-2018 [J]. Ecological Economics,2022,38(09):62-70.

[9] Yang Junchuang, Zhao Chao. Overview of K-Means Clustering Algorithm [J]. Computer Engineering and Application, 2019, 55 (23): 7-14+63.

[10] Wei Xueqin, Li Wen, Li Bingguo, Geng Lufei, Lin Yushi, Pang J. Evaluation of the quality of 12 plant enzymes based on principal component and cluster analysis [J]. Food Research and Development,2022,43(17):41-48.

[11] Ouyang Ziqiang, Hao Peng, Liu Muyang. Study on crystallization of glass silica and countermeasures [J]. Glass, 2022,49 (08): 39-44.