Clustering model research based on composition analysis and identification of glass products

Lei Lv [1, *], Suqing Duan [2], Liguo Liu [2], Hao Xu [3]

[1] School of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin, China

[2] School of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin, China

[3] College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China

* Corresponding author: lvlei2239729638@163.com

Abstract. As we all know, ancient glass buried in the soil will be weathered, and a large number of chemical elements in it will exchange with the chemical elements of the soil. As a result, the proportion of chemical components inside the glass changes greatly, affecting the analysis and identification of ancient glass products. This paper selects the fresh data from weathering samples, establish a clustering model and use SPSS to cluster analysis, and classify the types of each cluster. If there is a significant difference, it shall be divided into subcategories; The K-Means algorithm is used to analyze again, and the observation results are compared with the results without disturbance.

Keywords: Glass component identification, Weighted average prediction, Systematic clustering, K-Means.

1. Introduction

In the early days of the ancient cultural exchange between China and the West was the Silk Road, and glass was used as valuable physical evidence to prove the trade [1]. Glass was introduced to China in the early days as bead-shaped jewelry, and after China learned the techniques of making it locally, so many glass products are very similar in appearance, but they do not have the same internal chemical composition [2].

The main material of glass is quartz sand, but the melting point of quartz sand is very high, so in the processing to add flux to reduce its temperature, in ancient times there are many fluxes, in addition to them in the addition of limestone as a stabilizer, when the added fluxes are different their chemical composition is also different [3].

In ancient times, glass is currently found buried in the soil, and its chemical composition is easily changed by the influence of the buried soil, so that its internal chemical elements are exchanged with the chemical elements in the soil, resulting in a change in the internal chemical composition and the ratio of each chemical component, thus affecting the analysis and identification of the composition of glass products [4,5].

Classify and sub-classify high potassium glass and lead-barium glass, and summarize their laws. Give the specific division method and the results of the division, and build a machine model and evaluate the model.

Analyze the cultural relics according to their chemical composition and identify the type to which they belong to derive the classification results, and perform sensitivity analysis on the results.

2. Model Assumptions and Notation

2.1. Assumptions[6]

1. The known conditions and parameters have fidelity.
2. Only the main conditions are considered, and no consideration is given to other cases.
3. Assume that the soil pH environmental conditions are stable, ignore the weather, human factors and other uncertainties bar the impact on the glass products.

2.2. Notations

Important notations used in this paper are listed in Table 1.

Table 1. Notations

Symbols	Symbol Description
$x_i(i=1,2,\cdots,14)$	The *i*-th indicator pre-weathering chemical variable
$x_i^{'}(i=1,2,\cdots,14)$	*i*-th indicator post-weathering chemical variable
$w_i(i=1,2,\cdots,14)$	Weight of the *i*-th indicator before weathering
$w_i^{'}(i=1,2,\cdots,14)$	Weight of the *i*-th indicator after weathering
$\alpha_i(i=1,2,\cdots,14)$	Weighted percentage of the *i*-th indicator before and after weathering
k	Number of K-Means clusters
σ	Standard deviation

3. Model Construction and Solving

3.1. SPSS glass products system clustering model

The results of the screening for high potassium glass and lead-barium glass for weathering, decoration, category and the total range of chemical composition are shown in Table 2[7].

Table 2. Screening results

Glass Type	Weathering or not	Color	Ornamentation	Total content range
High Potassium	Weathered	Blue Green	B	99.81%~100%
	No weathering	Blue green, light blue, dark blue	A, C	97.25%~100%
Lead Barium	Weathering	Blue green, light blue, light green, dark green, purple, black	A, C	90.17%~99.89%
	No weathering	Dark blue, light blue, dark green, light green, purple, green	A, C	88.41%~99.89%

Step 1: Perform a cluster analysis for the glass, looking at each sample as a class

Step 2: These variables are clustered into one class each, calculate the distance matrix of class distance.

Step 3: The variables with similar distances are clustered into one class based on the calculated inter-class distances, and the other variables remain in their own class.

Step 4: Further aggregation of classes with similar distances, and so on, until the data are completely grouped into one category, this process results in a tree diagram of different glass types in the system clustering as shown in Figure 1 below.

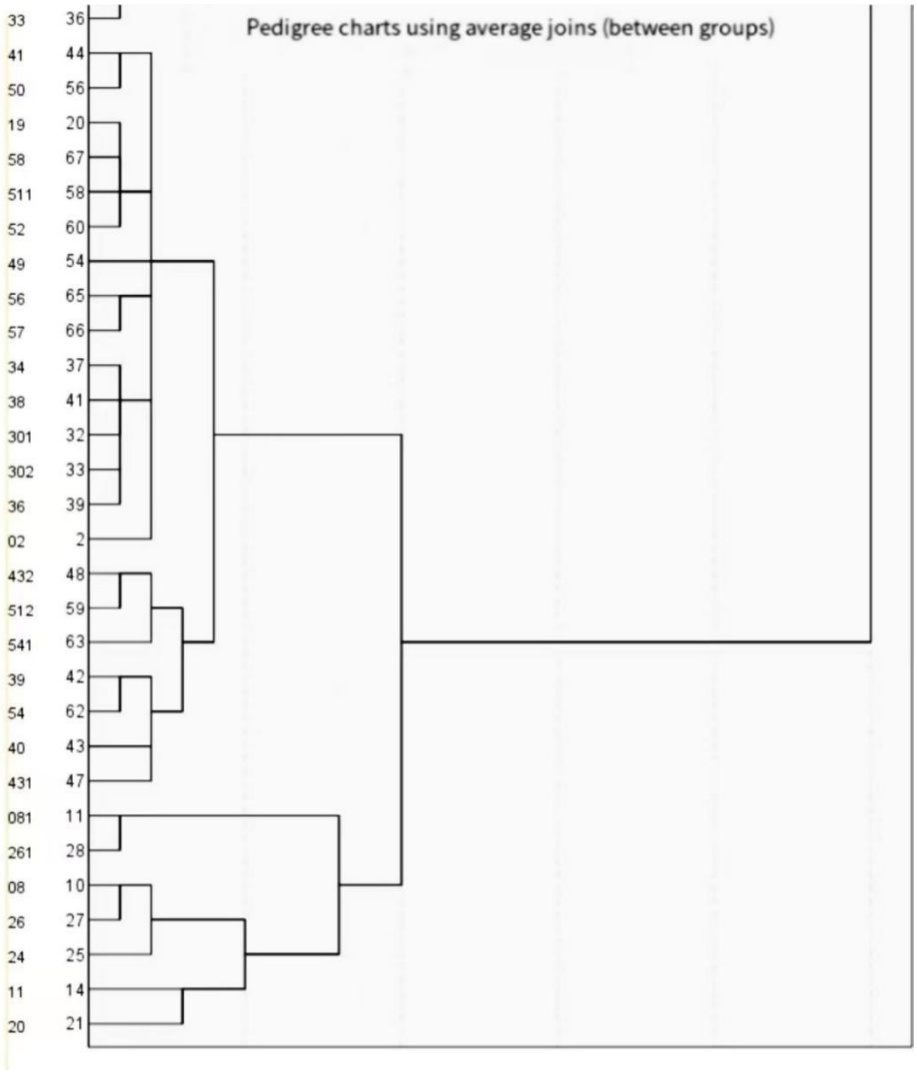

Figure 1. Cluster analysis dendrogram

Step 5: Finally, a clustering analysis was conducted for the type of glass, without dividing the specific glass types, and only based on whether they were weathered or not. The results of the systematic clustering were compared with the actual values, so as to judge the reasonableness of the classification criteria division.

Based on the above classification results, the number of clusters is set to 2. And calculate the number of class artifacts. Calculating the number of correct categories is judged by examining the data of high potassium glass and lead-barium glass with the given title attached to the card. The correct rate of weathering in high potassium glass is 30/32=93.75%, and the correct rate of unweathered glass is 35/35=100%.

In the classification analysis of the subclasses of high-potassium glass, the above classification results were used to compare the various chemical compositions of high-potassium glass and lead-barium glass, and the dispersion was used to reflect the individual chemical compositions, using the standard deviation as the basis for the calculation [8-10].

Overall standard deviation.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \tag{1}$$

Sample standard deviation

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

Standard Error

$$\sigma_n = \frac{\sigma}{\sqrt{n}} \tag{3}$$

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

Using box plot of SPSS calculation reflects its reasonableness and sensitivity, this paper sorts the data from small to large. When the greater the difference then the greater the difference between them; if their difference is smaller, then it is difficult to distinguish the difference of the data. Calculate the quartile and median quartile distance IQR=Q3-Q1. Draw two lines at Q3+1.5IQR and Q1-1.5IQR with the same median to reflect the outlier cut-off point, mark the mild outliers with a "O" and the extreme ones with "*" marks the extreme outliers. Data points of the same value are marked side by side on the same data line position, and data points of different values are marked above and below different data line positions.

The quartile Q1 is the value ranked at 25%, and the upper quartile Q3 is the value ranked at 75%. The reasonable interval is specified as $[Q1-1.5*IQR, Q3+1.5*IQR]$, then the value outside the interval is an outlier.

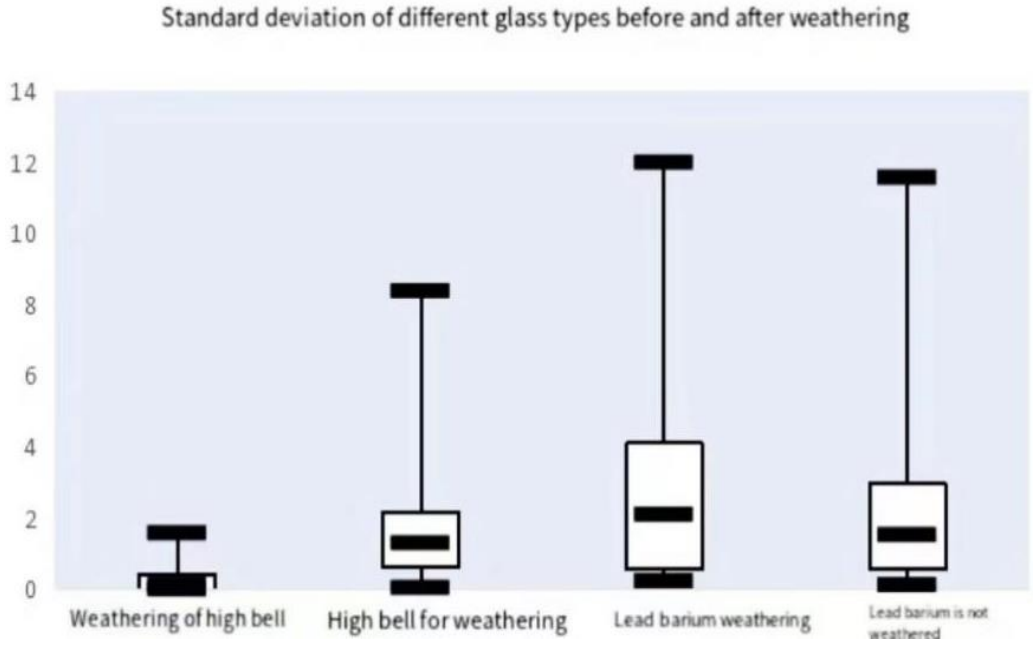

Figure 2. Standard deviation data for box plot analysis

According to the standard deviation analysis of the above Figure 2, it can be concluded that each classification has data that do not match. This paper will find out these data and delete them, and select the data that are reasonable for the second class classification, for the high potassium weathering also use K-Means algorithm for analysis.

The data were clustered and analyzed for differences, the frequencies of each category were analyzed, and after the frequencies were derived for classification, the distance between each sample and the center point was analyzed using the cluster center coordinates, and finally a comprehensive analysis (results of the mean ± standard deviation, F-test results, and significant P-value) was performed to determine whether the P-value of each analyzed item was significant (P<0.05)(Table 3), and the differences were analyzed according to the mean ± standard deviation Analysis was performed to analyze the differences that existed between the two groups of data.

Table 3. Cluster analysis with analysis of variance table

Chemical composition	Clustering categories (mean ± standard deviation)		F	P
	Category 2(n=4)	Category 1(n=2)		
Sodium Oxide(Na2O)	0.0±0.0	0.0±0.0		NaN
Potassium Oxide(K2O)	0.63±0.457	0.37±0.523	0.4	0.561
Calcium Oxide(CaO)	0.655±0.354	1.3±0.509	3.495	0.135
Magnesium Oxide(MgO)	0.0±0.0	0.59±0.071	371.307	0.000***
Aluminum Oxide(Al2O3)	1.392±0.481	3.005±0.7	11.709	0.027**
Sulfur Dioxide(SO2)	0.0±0.0	0.0±0.0		NaN
Tin Oxide (SnO2)	0.0±0.0	0.0±0.0		NaN
Strontium Oxide(SrO)	0.0±0.0	0.0±0.0		NaN
Phosphorus pentoxide(P2O5)	0.277±0.264	0.285±0.106	0.001	0.972
Iron Oxide(Fe2O3)	0.26±0.065	0.275±0.106	0.05	0.834
Barium Oxide(BaO)	0.0±0.0	0.0±0.0		NaN
Lead Oxide(PbO)	0.0±0.0	0.0±0.0		NaN
Copper Oxide(CuO)	1.82±1.013	1.045±0.7	0.898	0.397
Note: ***, **, * represent 1%, 5%, 10% significance level respectively				

The Magnesium Oxide and Alumina indicators were chosen as appropriate variables. The same degree of weathering was calculated for the remaining types.

The final classification corresponding to each type of glass and the coordinates of the clustering centroids in the compounds are derived as shown in the following Table 4.

After weathering in the high potassium category, Magnesium Oxide and Alumina showed significant differences, so the two were selected for differentiation and subcategorization, and similarly the remaining one for comparative analysis, and finally the data were perturbed and sensitivity analysis was performed.

Use Matlab to program the result and substitute its answer into algorithm K-Means for calculation the Euclidean distance from the coordinates of the original clustering centroids and compared with the true values to calculate the accuracy of the model.

Table 4. Table of coordinates of clustering centroids in compounds

	Category	Sodium Oxide (Na2O)	Potassium Oxide (K2O)	Calcium Oxide (CaO)	Magnesium Oxide (MgO)	Aluminum Oxide (Al2O3)	Iron Oxide (Fe2O3)
High potassium weathering	1	0	0.37	1.3	0.59	3.005	0
	2	0	0.63	0.655	0	1.3925	0
High potassium type unweathered	1	1.19142857	11.7585714	0.92714286	1.76857143	2.95428571	0.17428571
	2	0	5.932	1.292	2.16	1.75	0
Lead-barium weathering2	1	0.241	0.127	2.7315	0.7325	2.8275	-2.22E-16
	2	0.13333333	0.155	2.575	0.375	3.445	5.92
Lead-barium weathering3	1	0.30285714	0.12142857	2.95714286	0.785	3.64857143	0
	2	0.1725	0.17125	2.445	0.73875	2.72625	0
	3	0	0.1	2.28	0	1.0825	8.88
Lead-barium unweathered2	1	1.21	0.24375	1.86875	0.53875	3.145	0
	2	1.93466667	0.20533333	1.028	0.69466667	5.15533333	0.244
Lead-barium unweathered3	1	1.38285714	0.27857143	2.06857143	0.61571429	3.36714286	2.78E-17
	2	2.07285714	0.16928571	1.10142857	0.74428571	5.13428571	0.26142857
	3	0	0.355	0.235	0	3.52	0
High potassium weathering	1	0	0.285	0.275	0	0	1.045
	2	0	0.2775	0.26	0	0	1.82
High potassium unweathered	1	2.78E-17	0.02857143	0.44857143	7.40857143	0.40857143	0.71857143
	2	0.472	0.06	0.36	2.426	0.864	2.36
Lead-barium weathering2	1	0.4145	0.7085	5.3205	7.776	48.224	1.326
	2	0.43166667	0.17166667	5.13333333	25.245	26.94666667	5.44166667
Lead and barium weathering3	1	0.31285714	0.83428571	5.91357143	9.95785714	39.46357143	1.52214286
	2	0.56625	0.44	4.2625	4.75375	56.63625	1.2675
	3	0.4925	0	5.08	32.3875	30.145	6.93
Lead-barium unweathered2	1	0.43125	0.80125	1.09125	10.8725	31.5225	1.57
	2	0.18133333	0.702	1.02666667	8.004	17.05133333	1.358
Lead-barium unweathered3	1	0.36285714	0.91571429	1.22714286	8.67857143	31.86285714	0.58571429
	2	0.19428571	0.64428571	0.68928571	6.89357143	17.605	1.11357143
	3	0.455	0.755	2.945	24.89	19.22	6.62

Table 5. Model Accuracy Table

Category	1%	2%	5%	10%	20%	30%
High potassium weathering 2	100%	100%	100%	100%	100%	100%
High potassium category unweathered2	100%	100%	100%	100%	100%	100%
Lead-barium weathering2	100%	100%	100%	100%	100%	100%
Lead-barium unweathered2	100%	100%	100%	100%	100%	91%

It can be seen from Table 5 above that the high potassium glass increases its random disturbance after weathering. When the disturbance value is less than 30%, the accuracy of the model is still 100%, which indicates that the clustering effect is good; For the fresh lead barium glass, the accuracy of the model is 91% only when the disturbance effect is 30%; The accuracy of other cases is 100%. Therefore, it reflects the accuracy and stability of our model to a certain extent.

3.2. Subclass classification based on significant differences

Using the above model for classification, the classification results are shown in Table 6.

Table 6. Unknown category of glass artifacts inferred table

Artifact Number	Surface weathering	Inferred classification	**Inferred subclasses**
A1	No weathering	High Potassium	2
A2	Weathering	Lead barium	2
A3	No weathering	lead barium	2
A4	No weathering	lead barium	2
A5	Weathering	Lead barium	1
A6	Weathering	High Potassium	1
A7	Weathering	high potassium	1
A8	No weathering	lead barium	1

On this basis, the data were perturbed and the variable values were scaled randomly within the range (-130%, +130%) and the results were observed in comparison with the results without perturbation for sensitivity analysis and the results are shown in Table 7 below.

Table 7. Glass artifacts inferred stability table

Artifact Number	1% Perturbation	2% Perturbation	5% Perturbation	10% Perturbation	20% Perturbation	30% Perturbation
A1	Accurate	Accurate	Accurate	Accurate	Accurate	Accurate
A2	Accurate	Accurate	Accurate	Accurate	Accurate	Accurate
A3	Accurate	Accurate	Accurate	Accurate	Accurate	Accurate
A4	Accurate	Accurate	Accurate	Accurate	Accurate	Accurate
A5	Accurate	Accurate	Accurate	Accurate	Accurate	Accurate
A6	Accurate	Accurate	Accurate	Accurate	Accurate	Accurate
A7	Accurate	Accurate	Accurate	Accurate	Accurate	Accurate
A8	Accurate	Accurate	Accurate	Accurate	Accurate	Accurate

The final accuracy of the model is 100% after the perturbation, which illustrates the good robustness as well as the generalizability of the model.

4. Conclusion

This paper processed the data, screened out the weathering samples of unweathered data, established a clustering model, and used SPSS for cluster analysis to divide the types of each of the clusters, when the number of clusters is 2 the two clusters belong to high potassium and lead barium, respectively. After weathering in the high potassium category, magnesium Oxide and alumina showed significant differences, so the two were selected for differentiation and subclassification, and the four types of distribution of whether high potassium and lead barium weathered were analyzed using the K-Means method, and the differences in the results of each clustering when the number of clusters was 2 and 3 were discussed, and subclassification was performed according to the P value. Finally, a sensitivity analysis was performed to perturb the data eventually yielding that when the perturbation range was 30%, the accuracy of the model was 91% the rest of the cases were 100% correct, thus reflecting to some extent the accuracy and stability of our model.

For identifying the chemical composition of glass products of unknown composition grouped into types belonging to, using the clustering model, using SPSS for cluster analysis, after analysis for each type of division, if there is a significant difference then subclass division, using the algorithm of K-Means again for analysis, consider whether the resulting results are different according to the P value for subclass division, based on this perturbation of the data, the results are compared with the results without perturbation, and the accuracy of the model is 100% after perturbation, which shows the robustness and generalizability of the model.

References

[1] Jiang P, Lu Hao-Xiang, Liu Zhen-B. Near-infrared spectroscopic drug identification by random forest combined with CatBoost [J]. Spectroscopy and Spectral Analysis, 2022, 42 (07): 2148 - 2155.

[2] Sun Yifan, Liu Bing, Yu Xuchu, et al. A high-resolution feature network classification method for image-level hyperspectral images[J/OL]. Journal of Surveying and Mapping: 1 - 16 [2022-09-16].

[3] Yao Chun, Deng Junliang, Yang Zhiqi, et al. Comparison of the value of MRI texture analysis and LI-RADS classification for differential diagnosis of small hepatocellular carcinoma with atypical hyperplastic nodules in liver cirrhosis [J]. Radiology Practice, 2022, 37 (08):995 - 999.

[4] Li AY. Research on the prediction of particulate matter concentration in Urumqi based on RF-Kmeans-LIBSVM[J]. Environmental Protection Science, 2022, 48 (04): 118 - 124.

[5] Gu LJ, Si Shoukui, Sun Huijing, Dong Chao. Fuzzy clustering analysis applied to artillery precision strike effectiveness assessment [J]. Military Automation, 2015, 34 (12): 1 - 3.

[6] Qin Risheng, Xiang Hua, He Xin, et al. Clustering analysis of actual daily load curve based on improved PSO-Kmeans algorithm [J]. Electrical Engineering Technology, 2022 (11): 1 - 6.

[7] Zhu Yifei. Research on the application of Kmeans-SMOTE algorithm of clustering fusion in personal credit risk assessment [D]. Shanghai Normal University ,2022.

[8] Chen Chen-Peng, Zhao Xin, Bi Gui-Hong, et al. Short-term wind speed prediction based on Kmeans-VMD-LSTM [J]. Electrical Machines and Control Applications, 2021, 48 (12): 85 - 93.

[9] Cui Rumai, Si Shoukui. Mathematical Model of Data Visualization Processing [J]. Journal of Naval Academy of Aeronautical Engineering, 2010, 25 (01): 105 - 108+112.

[10] Liu Lihong, Chen Yafeng. Prediction method of bumper mold opening shrinkage based on weighted average algorithm and Middle Plane grid [J]. Plastic Science and Technology, 2021, 49 (07): 133 - 136.