Identification of glass products based on cluster analysis and random forest model

Mengqi Lu[1, *], Xin Xu[1], Di Shao[2], Rui Tang[3]

[1]Maynooth International Engineering College, Fuzhou university, Fuzhou, China

[2]College of Electrical Engineering and Automation, Fuzhou university, Fuzhou, China

[3]College of Advanced Manufacturing, Fuzhou university, Fuzhou, China

*Corresponding author: 1687673531@qq.com

Abstract The main chemical composition of glass is silica (SiO_2). Due to the high melting point of pure quartz sand in the raw material of glass production, flux should be added in order to reduce the melting temperature. The main chemical composition varies with the flux added. In this paper, through the grey correlation analysis can be concluded that the key factors influencing the classification to the content of lead oxide, and barium oxide, and through the random forest, hierarchical clustering, and K - Means method to classify different types of glass and the division of the class, the glass can be divided into high potassium, high potassium not weathering, lead, barium, lead, barium not weathering four categories of two classes, a total of eight classes, Through the test of classification sensitivity and reliability, it shows that the three models are reasonable and can be applied.

Keywords Grey correlation analysis, K-Means, hierarchical clustering, random forest, glass type division

1. Introduction

The main raw material of glass is quartz sand, the main chemical composition is silica (SiO_2). Due to the high melting point of pure quartz sand, in order to reduce the melting temperature, it is necessary to add flux in refining. The main chemical composition varies with the flux added. For example, lead-barium glass is usually regarded as a kind of glass invented in China because of its high content of lead oxide (PbO) and barium oxide (BaO) when lead ore is added as flux in the firing process [1]. Potassium glass is made of high potassium content material such as plant ash as flux. It is mainly popular in Lingnan, Southeast Asia and India. In this article, through analyzing the classification of high potassium and lead barium glass, for each category to choose the appropriate chemical composition to the class division, the division of specific methods and results, and analyze the unknown categories the chemical composition of glass, identify its type, and the classification results of sensitivity analysis.

2. Model Building

2.1. Grey correlation method

2.1.1 Definition

In the process of system development, if the change trend of two factors tends to be the same, that is, the synchronous change degree of the two factors is higher, that is, the correlation degree of the two factors is higher [2-5]. Otherwise, it is lower. Therefore, grey correlation analysis method is usually based on the degree of similarity or dissimilarity between the development trend of factors, such as the comparison of the similarity between the geometric relationship of data series and the geometric shape of curves, which is a method to measure the degree of correlation between factors [6,7].

2.1.2 Main Steps

Step 1: characteristic sequence and parent sequence.

Step 2: Unify the dimensions of the index data. Eliminate the influence caused by the different units of each index and the great difference between the numerical orders of magnitude, and avoid the occurrence of unreasonable phenomena.

Step 3: Calculate the correlation coefficient. The correlation coefficient between each comparison sequence and the corresponding element of the reference sequence is calculated from the following formula: ρ is the resolution coefficient, and the value is within (0,1). The smaller the resolution coefficient, the greater the difference between the correlation coefficients and the stronger the discrimination ability, usually 0.5.

Step 4: Calculate the association order r_{0i}(Eq (1)). The weighted average of the correlation coefficients between each index and the corresponding elements of the reference sequence is calculated to reflect the correlation between each manipulation device object and the reference sequence, which is called the correlation degree.

$$r_{0i} = \frac{1}{m}\sum_{k=1}^{m} W_k \zeta_i(k) \tag{1}$$

Step 5: Analyze the calculation results (Eq (2)). According to the grey weighted correlation degree, the correlation order of each evaluation object is established. The greater the correlation degree, the more important the evaluation object is to the evaluation standard.

$$\begin{gathered} \gamma\big(x_0(k), x_i(k)\big) = \frac{\triangle min + \rho \triangle max}{\triangle_{ik} + \rho \triangle max} \\ \triangle min = \min_i \min_k |x_0(k) - x_i(k)| \\ \triangle max = \max_i \max_k |x_0(k) - x_i(k)| \\ \triangle_{ik} = |x_0(k) - x_i(k)| \end{gathered} \tag{2}$$

2.2. Hierarchical clustering and K-means algorithm

Hierarchical clustering creates a hierarchical nested clustering tree by calculating the similarity between data points of different categories.

K-means algorithm is a simple iterative clustering algorithm, which uses distance as the similarity index and belongs to unsupervised learning, so as to discover K classes in a given data set, and the centroid of each class is obtained according to the mean value of all values in the class, and the centroid of each class is described by the cluster centroid [8]. For a given dataset X (containing n one-dimensional and more than one-dimensional data points) and the number of categories K to be obtained, the Euclidean distance is selected as the similarity index, and the clustering sum of squares of the classes implemented by the clustering objective is minimized, namely, the minimization (Eq. (3)).

$$J = \sum_{k=1}^{k}\sum_{i=1}^{n} \|x_i - u_k\|^2 \tag{3}$$

2.3. Random Forest

Classification generates many decision trees by randomly sampling the sample observations and characteristic variables of the modeling dataset [9]. Each sampling result is a tree, and each tree will generate rules and classification results that conform to its own attributes, and the forest finally integrates the rules and classification results of all decision trees to realize the classification of random forest algorithm [10,11].

3. Results

3.1. Gray correlation analysis method to screen variables

The gray correlation degree results of the 14 evaluation items are shown in Table 1. PbO has the highest evaluation (correlation degree: 1.0), followed by BaO (correlation degree: 0.949). Therefore, the chemical components with correlation degree less than 0.5 were removed, and hierarchical clustering was carried out for the remaining variables.

Table 1. Grey correlation degree

Items	PbO	BaO	SrO	SO_2	Na_2O	P_2O_5	Al_2O_3	MgO	SiO_2	CuO	Fe_2O_3	SnO_2	CaO	K_2O
Correlation degree	1	0.949	0.871	0.651	0.641	0.517	0.497	0.477	0.466	0.462	0.439	0.413	0.41	0.342
Ranking	1	2	3	4	5	6	7	8	9	10	11	12	13	14

3.2. Establishment of subclass division model of glass species based on systematic clustering and K-means clustering

First of all, specific glass types are not divided, but cluster analysis is carried out according to whether the glass is weathered or not. The result of systematic clustering is compared with the actual value, so as to judge the rationality of classification criteria. The dentree of different glass types in systematic clustering is shown in Fig.1.

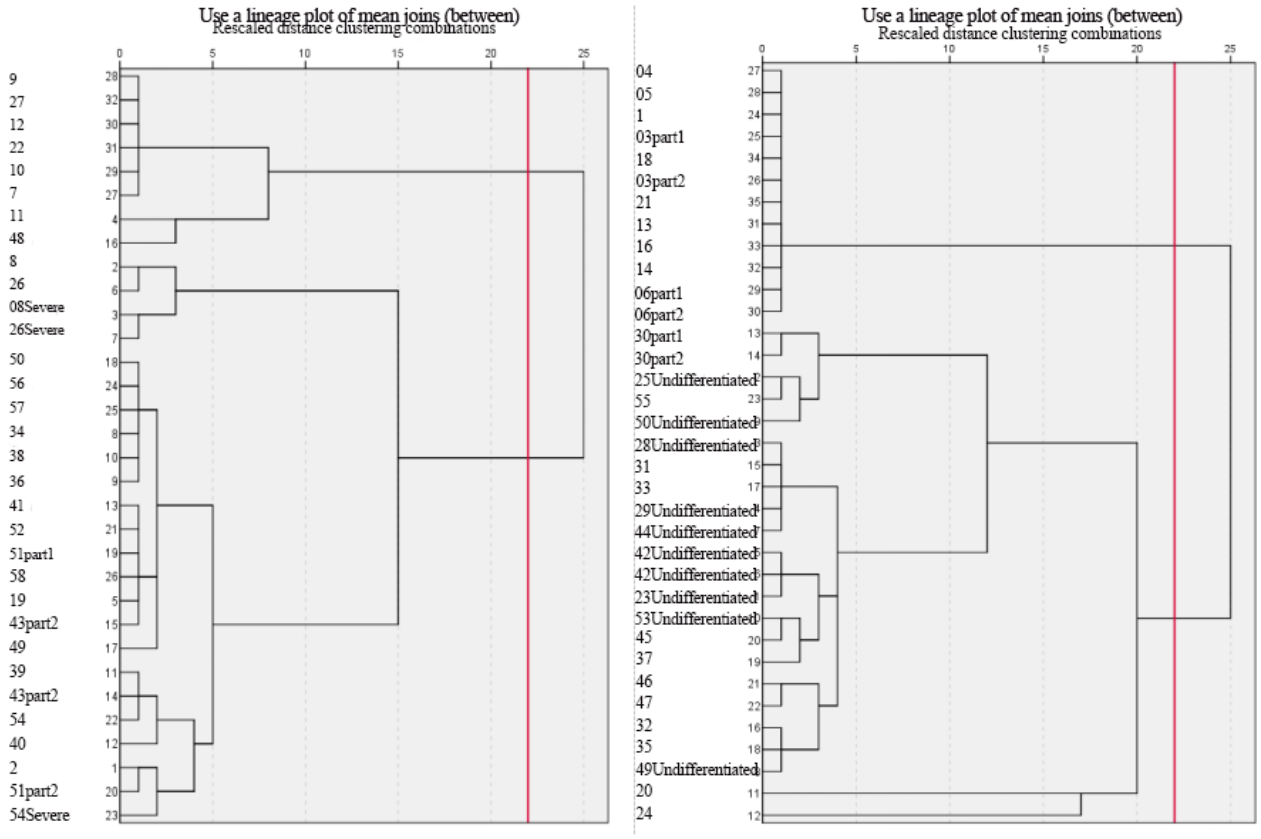

Fig. 1 Pedigree diagram of the clustering results of weathered and unweathered systems of all glass types

According to the classification results shown in Figure 1, the number of clusters is set as 2, and the cultural relics in each category are divided into categories according to their numbers in the pedigree map, and then analyzed with whether they belong to high-potassium glass or lead-barium glass in reality. The inconformity results are marked in the table, and the classification results are shown in Table 2.

Table 2. Comparison table between system clustering results and actual values

	High potassium glass	Lead barium glass
Weathered	9, 27, 12, 22, 10, 7, 11 (abnormal),48 (abnormal)	8, 26, 50, 56, 57, 34, 38, 36, 41, 52, 51, 58, 19, 43, 49, 39, 54, 40, 2, 54
Unweathered	4, 5, 1, 3, 18, 21, 13, 16, 14, 6	30, 25, 55, 50, 28, 31, 33, 29, 44, 42, 23, 53, 45, 37, 46, 47, 32, 35, 49, 20, 24

As can be seen from the above table, the number of abnormal data is very small, indicating that the model is reasonable.

Finally, four categories can be obtained: high potassium weathering, high potassium unweathering, lead-barium weathering and lead-barium unweathering. When subcategories are divided, K-Means algorithm can be used for analysis, and difference analysis (F-test and P-value) can be conducted for various chemical substances.

It can be seen from Table 3 and 4 that Na_2O showed significant differences before weathering of high-potassium glass, so it was selected for subclass division. P_2O_5 showed significant difference after weathering of high-potassium glass, so it was selected for subclass division. Before weathering of lead-barium glass, there were significant differences in PbO, BaO and SO_2, so they were selected for subclassification. PbO and SrO showed significant differences after weathering of lead-barium glass, so they were selected for subclassification. The classification results are shown in Fig. 2

Table 3. Difference analysis of weathered and unweathered components of high potassium glass

	Weathered				Unweathered			
	cluster		significance		cluster		significance	
	mean square	degree of freedom	F	P	mean square	degree of freedom	F	P
Na_2O	.000	1	.	.	17.389	1	209.808	.000
PbO	.000	1	.	.	.109	1	.294	.600
BaO	.000	1	.	.	1.432	1	1.560	.240
P_2O_5	.154	1	9.198	.039	3.429	1	1.787	.211
SrO	.000	1	.	.	.003	1	1.424	.260
SO_2	.000	1	.	.	.041	1	1.226	.294

Table 4. Difference analysis of weathered and unweathered components of lead barium glass

	Weathered				Unweathered			
	cluster		significance		cluster		significance	
	mean square	degree of freedom	F	P	mean square	degree of freedom	F	P
Na_2O	.054	1	.167	.686	2.740	1	.476	.498
PbO	2089.500	1	30.393	.000	1092.598	1	58.509	.000
BaO	1408.458	1	31.279	.000	42.930	1	1.281	.270
P_2O_5	.162	1	.009	.926	.022	1	.006	.939
SrO	.001	1	.019	.893	.326	1	6.997	.015
SO_2	161.753	1	13.839	.001	.311	1	.522	.478

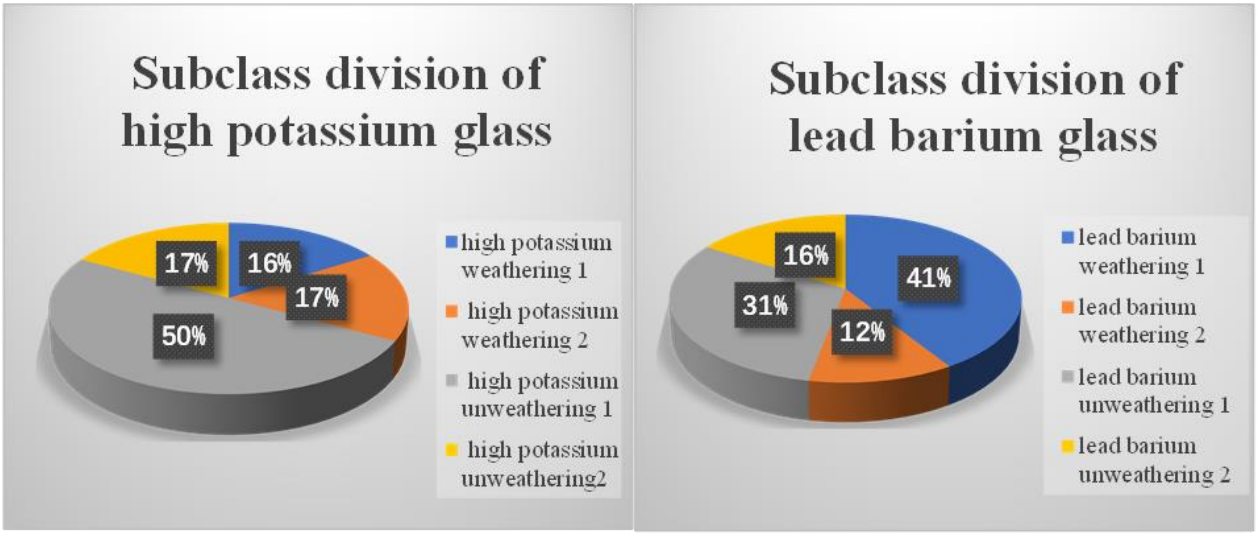

Fig. 2 Different clustering results of subclass division

3.3. Prediction of glass types by random forest

The prediction results of random forest are shown in Table 5, Table 6 and Table 7. It is concluded that A1, A6, A7 belong to high potassium glass, A2, A3, A4, A5, A8 belong to lead-barium glass, and it is found that the simulation evaluation effect is well.

Table 5. Evaluation results of random forest model

	Accuracy rate	Recall rate	Precision ratio	F1
Training set	1	1	1	1
Test set	1	1	1	1

Table 6. Prediction results

No.	predicting results	given type	Predicted result probability lead barium	Predicted outcome probability high potassium	CaO	SiO_2	Na_2O	MgO	K_2O	SO_2	Al_2O_3	SnO_2	BaO	Fe_2O_3	CuO
A1	high potassium	high potassium	0.03	0.97	6.08	78.45	0	1.86	0	0.51	7.23	0	0	2.15	2.11
A2	lead barium	lead barium	0.66	0.34	7.63	37.75	0	0	0	0	2.33	0	0	0	0
A3	lead barium	lead barium	0.9	0.1	7.19	31.95	0	0.81	1.36	0	2.93	0	4.69	7.06	0.21
A4	lead barium	lead barium	0.92	0.08	2.89	35.47	0	1.05	0.79	0	7.07	0	8.31	6.45	0.96
A5	lead barium	lead barium	0.89	0.11	1.64	64.29	1.2	2.34	0.37	0	12.75	0.49	2.16	0.81	0.94
A6	high potassium	high potassium	0	1	0.64	93.17	0	0.21	1.35	0	1.52	0	0	0.27	1.73
A7	high potassium	high potassium	0.01	0.99	1.12	90.83	0	0	0.98	0.11	5.06	0	0	0.24	1.17
A8	lead barium	lead barium	1	0	0.89	51.12	0	0	0.23	2.26	2.12	0	11.34	0	9.01

Table 7. Subclass division result of random forest

Predicting results_Y	Predicted resultd probability_1	Predicted resultd probability_2	Predicted resultd probability_3	Predicted resultd probability_4	Predicted resultd probability_5	Predicted resultd probability_6	Predicted result probability_7	Predicted result probability_8	CaO	Fe_2O_3	SnO_2	Al_2O_3	CuO	MgO
3	0.18	0.04	0.72	0	0	0.01	0.05	0	6.08	2.15	0	7.23	2.11	1.86
7	0.16	0.03	0.12	0.04	0.18	0.09	0.21	0.17	7.63	0	0	2.33	0	0
5	0	0	0.02	0.02	0.39	0.13	0.05	0.39	7.19	7.06	0	2.93	0.21	0.81
5	0.01	0.01	0.05	0.02	0.28	0.15	0.28	0.2	2.89	6.45	0	7.07	0.96	1.05
7	0.02	0.03	0.16	0.03	0.02	0	0.67	0.07	1.64	0.81	0.49	12.705	0.94	2.34
2	0.25	0.69	0.03	0	0.01	0	0.01	0.01	0.64	0.27	0	1.52	1.73	0.21
2	0.15	0.72	0.1	0.02	0	0	0.01	0	1.12	0.24	0	5.06	1.17	0
7	0.04	0	0.01	0	0.02	0.12	0.73	0.08	0.89	0	0	2.12	9.01	0

4. Conclusion

Fluxes are often added in the making of glass, and the addition of different fluxes will lead to certain differences in the composition of glass. In this paper, grey correlation analysis, hierarchical clustering and random forest are used to classify glass. The results show that several models can be used to classify glass types, and the results are accurate and reasonable.

References

[1] WANG Xiao-mo. The Status of Glass in the Chinese History [J]. Art & Design, 2006(10):30-31.

[2] JIA Hua-ping. Application of Intelligent Algorithm in Mathematical Modeling Contest[J]. Computer Systems & Applications,2016(8):149-154.

[3] XU Meifang. Ingenious application of Lingo and Excel in classroom teaching for higher mathematic[J]. Laboratory Science, 2020(2):97-99.

[4] XU Zhao-di; LI Xiao-yi. Integration of Mathematical Modeling and Mathematics Experimental Curriculum[J]. Journal of Shenyang Normal University (Natural Science Edition), 2010(3): 350-352.

[5] Yang Jianfeng; Qiao Peirui; Li Yongmei; Wang Ning. A Review of Machine-learning Classification and Algorithms[J]. Statistics & Decision, 2019(6):36-40.

[6] GUO San-dang; WANG Ling-ling; LIU-Si-feng; FANG Zhi-geng. Grey Cluster Analysis Based on the Biggest Relational Grades. [J] Mathematics in Practice and Theory, 2013(6): 195-210.

[7] Zhou Wenhao; Zeng Bo. A Research Review of Grey Relational Degree Model [J]. Statistics & Decision, 2020(15):29-34.

[8] WANG Qian; WANG Cheng; FENG Zhen-yuan; YE Jin-feng. Review of K-means clustering algorithm [J]. Electronic Design Engineering, 2012(7): 21-24.

[9] A Machine Learning Combination Clustering Algorithm and Its Application[J]. Journal of Shandong Agricultural University (Natural Science Edition), 2018(3): 463-466.

[10] DONG Shi-shi; HUANG Zhe-xue. A Brief Theoretical Overview of Random Forests [J]. Journal of Integration Technology,2013,01:1-7.

[11] WANG Cheng; WANG Kai. An improved random forest algorithm based on decision trees clustering reduction[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2019(3):91-97.