

Glass Classification and Identification Based on Random Forest Model and Systematic Clustering

Ruixing Zheng^{*,†}, Lezhi Wu[†], Qiyang Liu[†], Heming Chen[†], Xiangrui Qiu[†]

School of Finance, University of International Business and Economics, Chaoyang, Beijing, 100020

* Corresponding Author Email: z1843471706@163.com

[†] These authors contributed equally.

Abstract. A large exchange of material elements occurs during glass weathering, which will affect the identification and judgment of ancient glass. To this end, the data is first used to give the basis for dividing lead barium and high potassium glass, and the difference between the chemical components of each type of glass can be compared to obtain a conclusion. Since the final number of subclasses is unknown, this paper uses systematic cluster analysis to obtain the number of subclasses and classification results at the same time. In order to identify the glass types of unknown cultural relics, this paper uses the existing data to establish a random forest model and construct a method and standard for dividing glass types.

Keywords: Random forest, systematic clustering, glass differentiation.

1. Introduction

Glass is an important item in the early Sino-foreign trade exchanges, China's ancient glass combined with foreign technology and local materials, developed a unique chemical composition. The main raw material of glass is quartz sand, and the main chemical composition is SiO₂ [1-2]. Considering the high melting point of SiO₂, fluxes and stabilizers need to be added during the refining process. Most of the stabilizers are limestone (the main component is CaCO₃). Flux is roughly divided into two categories, one is lead ore (the main component is lead oxide (PbO), barium oxide (BaO)), forming lead barium glass often found in Chu culture areas of China [3-4]. The other type uses high potassium substances such as grass ash (the main component is potassium carbonate (K₂CO₃)) to form potassium glass that often appears in Lingnan and Southeast Asia in China [5-6].

Modern mining and research have found that ancient glass is affected by the buried environment and weathering to varying degrees. A large exchange of material elements occurs during the weathering process, which will affect the identification and judgment of ancient glass [7-8]. Therefore, it is necessary to use the existing ingredient test data for analysis and inference. In general, the sum of the ingredients should be 100% [9-10].

2. Establishment and solution of glass classification model

2.1. Classification law: Random forests are used to assess the importance of features

According to the known data, the chemical composition data of high-potassium glass and lead-barium glass are compared one by one. A simple observation reveals that the lead oxide (PbO) content is the key basis for the division. In this composition data, the average value of high-potassium glass is 0.274, which is much smaller than that of lead-barium glass of 33.35. And the maximum lead oxide content of high potassium glass is 1.62 is also less than the minimum value of lead barium glass 9.3. In fact, this can be explained by the use of lead ore as a flux in lead barium glass.

However, the solution process of question 3 inspired further reflection on this question. One of the advantages of the random forest model used in question 3 is that it can obtain the ranking of importance of different classification features.

We express variable importance measures as VIM and Gini coefficients as GI, and the current classification is based on k features, $X(1), X(2), \dots, X(k), \dots, X(14)$, M decision trees, N categories. Now we want to calculate the Gini index score $VIM_k^{(Gini)}$ for each feature $X(k)$, which is the average change in node division impurity of all decision trees of the k th feature.

The Gini index of the m th tree, node α is calculated as

$$GI_{\alpha}^{(m)} = 1 - \sum_{n=1}^{|N|} (p_{\alpha n}^{(m)})^2 \quad (1)$$

Where N represents N categories, $P_{\alpha n}$ represents the proportion of category n in the node α .

The importance of feature $X(k)$ α the m th tree node, i.e. the change in the Gini index before and after the node α branching is

$$VIM_{k\alpha}^{(Gini)(m)} = GI_{\alpha}^{(m)} - GI_{\beta}^{(m)} - GI_{\gamma}^{(m)} \quad (2)$$

Where $GI_{\beta}^{(m)}$ and $GI_{\gamma}^{(m)}$ represent the Gini indices of the two new nodes after branching, respectively.

If the node in which feature $X(k)$ appears in the decision tree m is the set α , then the importance of $X(k)$ in the m tree is

$$VIM_k^{(Gini)(m)} = \sum_{\alpha \in A} VIM_{k\alpha}^{(Gini)(m)} \quad (3)$$

Now there are M trees in the RF forest, then

$$VIM_k^{(Gini)} = \sum_{m=1}^M VIM_k^{(Gini)(m)} \quad (4)$$

Finally, all the importance scores obtained can be normalized:

$$VIM_k^{(Gini)} = \frac{VIM_k^{(Gini)}}{\sum_{k=1}^K VIM_k^{(Gini)}} \quad (5)$$

2.2. Subclass division: the establishment of a systematic clustering model

After a simple analysis of the data, it is found that the chemical composition of high potassium and lead barium glass before and after weathering has changed greatly, so firstly, the high potassium and lead barium glass are divided into two categories according to whether they are weathered or not, and four glass types are obtained: weathered high potassium glass, non-weathered high potassium glass, lead barium glass, and non-weathered lead barium glass.

On this basis, each category is further subclassified, and a general cluster analysis method can be used. Considering that the final number of subclasses obtained is uncertain, a systematic clustering model, that is, a hierarchical clustering model, should be adopted. The core idea of system clustering is to calculate the "distance" between samples, which is calculated using the Euclidean distance method, that is, the square sum of the differences between the eigen quantities. The closest sample is first clustered into a class, regarded as a type of sample, repeats the above calculation process, and finally the sample can be gradually aggregated into a large category.

The systematic clustering model can restore the specific situation of each cluster, calculate the degree of distortion in various classification situations, and finally use the elbow rule to determine the number of subclasses. Calculate the position of the center of gravity of the sample data in each category, and then calculate the position of each sample in the category to the center of gravity, and sum to obtain the degree of distortion. Finally, the degree of total distortion, also known as the aggregation coefficient R , is obtained.

The specific process is as follows:

Step 1: Calculate the distance between individual samples, using the European distance method:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{14} (x_i(k) - x_j(k))^2} \quad (6)$$

Where x_i is the sampling point in row i , and $x_i(k)$ is the chemical composition of item k x_i at the sampling point.

Step 2: Get the two samples closest and combine them into one class.

Step 3: Calculate the distance between the new class and other samples (classes), and when calculating the distance, take the minimum value of the distance between each sample in the new class and the calculated object.

Step 4: Repeat steps 2 and 3 until all samples are combined into one class to plot the systematic clustering analysis.

Step 5: Solve the aggregation coefficient R corresponding to the number of different subclasses, and select the flat turning point as the result based on the elbow rule.

The flowchart is shown as follows figure 1:

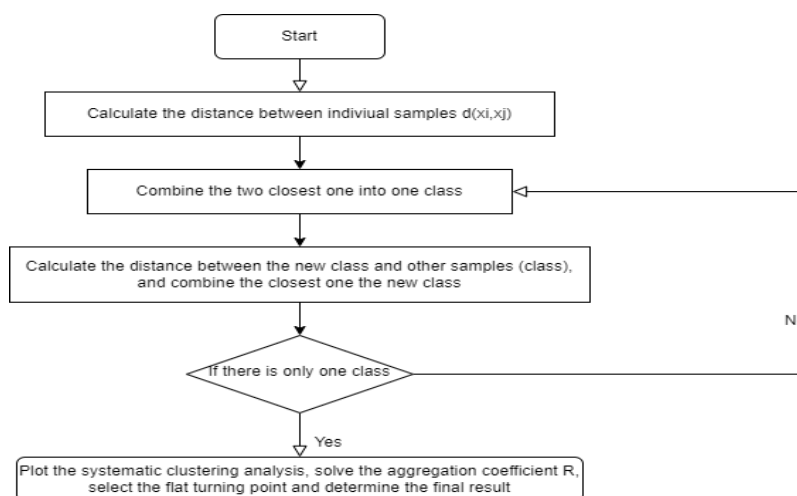


Figure 1. Flowchart of systematic clustering model

2.3. Classification law: the solution of the random forest model

Use Python code to implement the solution process of the model, taking into account randomness and perform it multiple times:

The sequential arrangement of the data is the arrangement of the corresponding component indicators. It can be analyzed that lead oxide (PbO), barium oxide (BaO), and silicon dioxide (SiO₂) are the top three important classification indicators of high potassium and lead barium glass, respectively.

2.4. Subclass Division: Solving the clustering model of the system

In order to reasonably calculate the distance between each sample, the blank value in the attached form 2 is uniformly treated as 0.

(1) Weathered high-potassium glass: system clustering results and aggregation coefficient

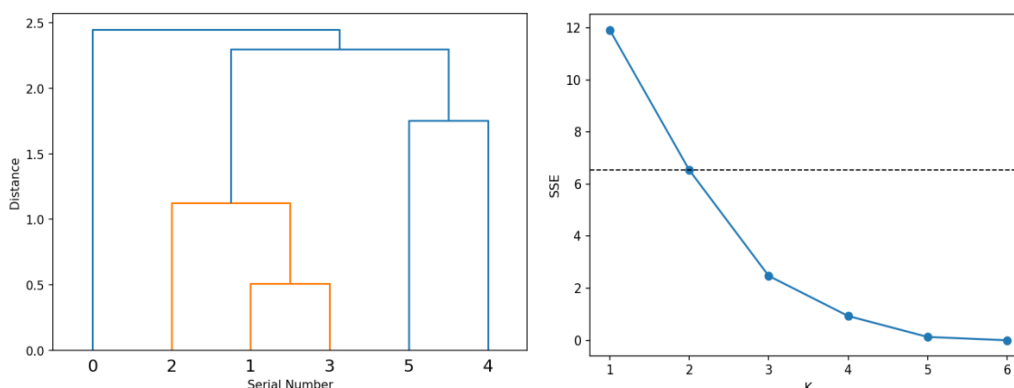


Figure 2. System clustering results and aggregation coefficient

According to the line chart of aggregation coefficients, when $K=3$, the change trend of aggregation coefficients slows down significantly.

When K takes 1 to 3, the aggregation coefficient changes significantly, and when $K \geq 3$, the value of aggregation coefficient is significantly smaller than the value when K takes 1 to 2. Therefore, the number of classifications is set at 4. The result is shown in the figure 3.

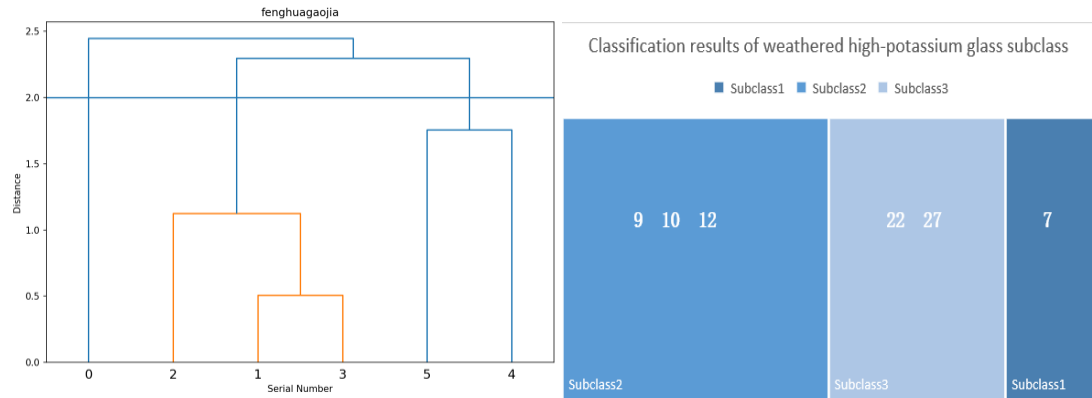


Figure 3. The result when K takes 1 to 3, the aggregation coefficient changes significantly
(2) Non-weathering high-potassium glass: system clustering results and aggregation coefficient plot. The result is shown in the figure 4.

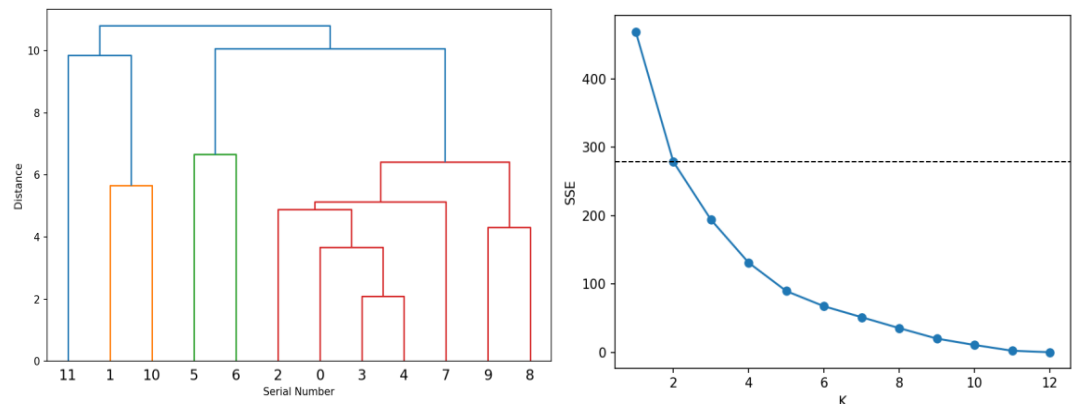


Figure 4. System clustering results and aggregation coefficient plot

Chart analysis:

According to the aggregation coefficient line chart, when $K=4$, the change trend of aggregation coefficient slows down significantly.

When K takes values from 1 to 4, the aggregation coefficient changes significantly, and when $K \geq 4$, the value of aggregation coefficients is significantly smaller than the value when K takes 1 to 3. Therefore, the number of classifications is set at 4. The result is shown in the figure 5.

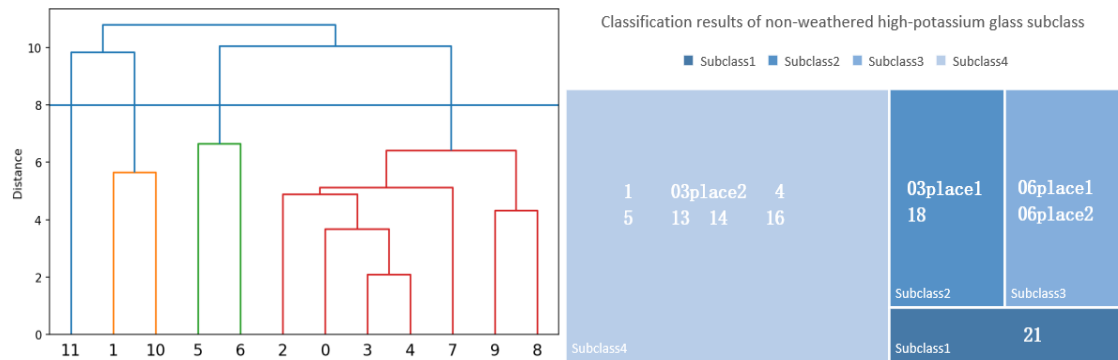


Figure 5. The result when K takes values from 1 to 4, the aggregation coefficient changes significantly

(3) Weathered lead barium glass: system clustering results and aggregation coefficient diagram. The result is shown in the figure 6.

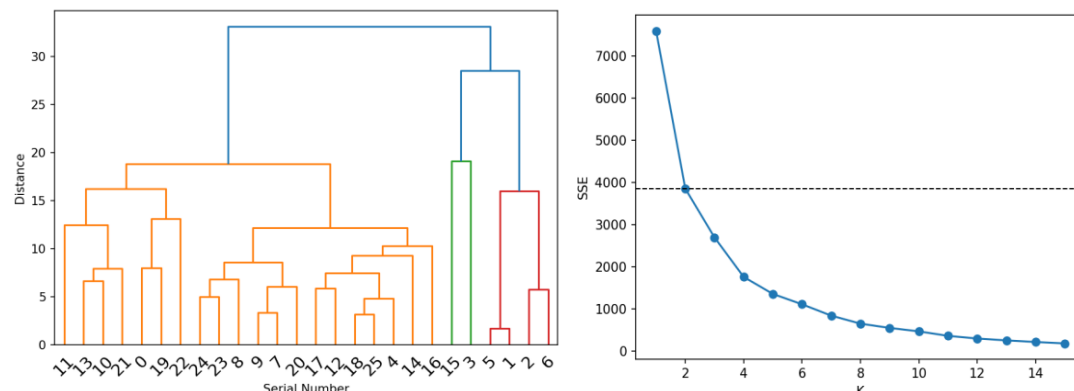


Figure 6. system clustering results and aggregation coefficient diagram

Chart analysis:

According to the aggregation coefficient line chart, when $K=4$, the change trend of aggregation coefficient slows down significantly.

When K takes values from 1 to 4, the aggregation coefficient changes significantly, and when $K \geq 4$, the value of aggregation coefficients is significantly smaller than the value when K takes 1 to 3. Therefore, the number of classifications is set at 4. The result is shown in the figure 7.

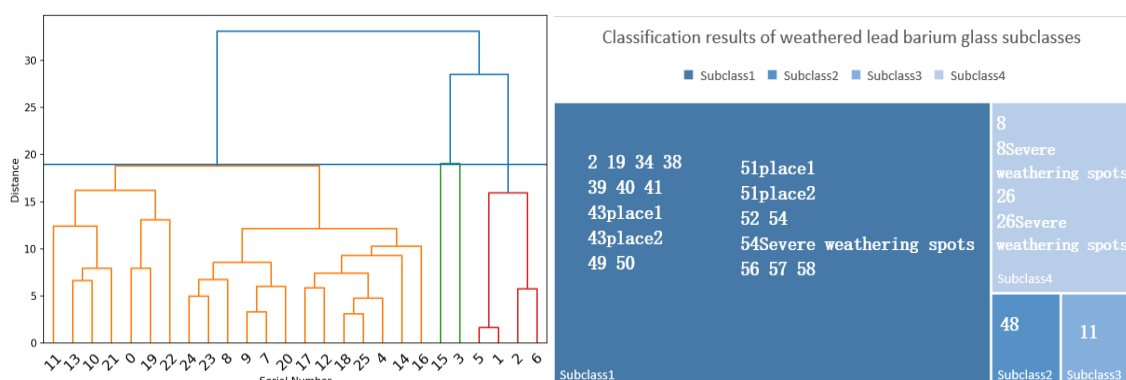


Figure 7. The result when $K \geq 4$

(4) Non-weathered barium lead-barium glass: systematic clustering results and aggregation coefficient figure 8.

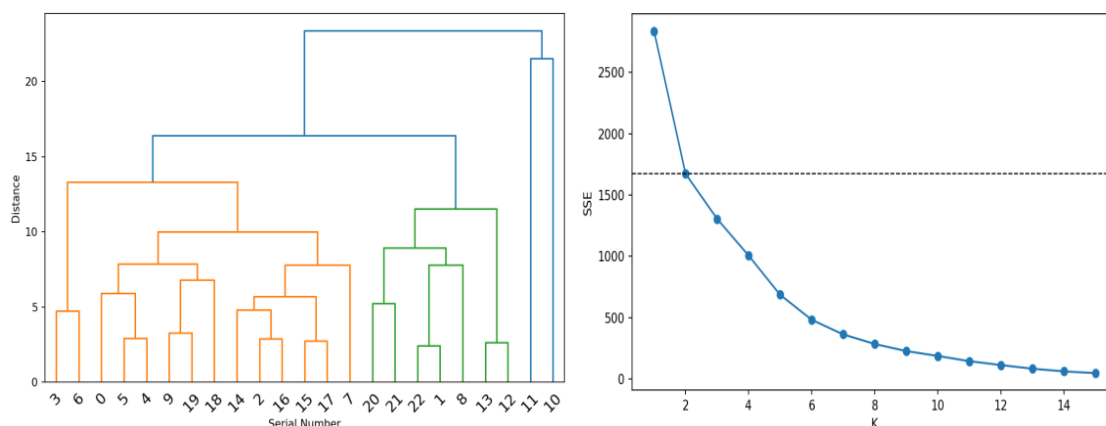


Figure 8. Systematic clustering results and aggregation coefficient

Chart analysis:

According to the aggregation coefficient line chart, when $K=5$, the change trend of aggregation coefficient slows down significantly.

When K takes values from 1 to 5, the aggregation coefficient changes significantly, and when $K \geq 5$, the value of aggregation coefficients is significantly smaller than the value when K takes 1 to 4. Therefore, the number of classifications is set at 5. The result is shown in the figure 9.

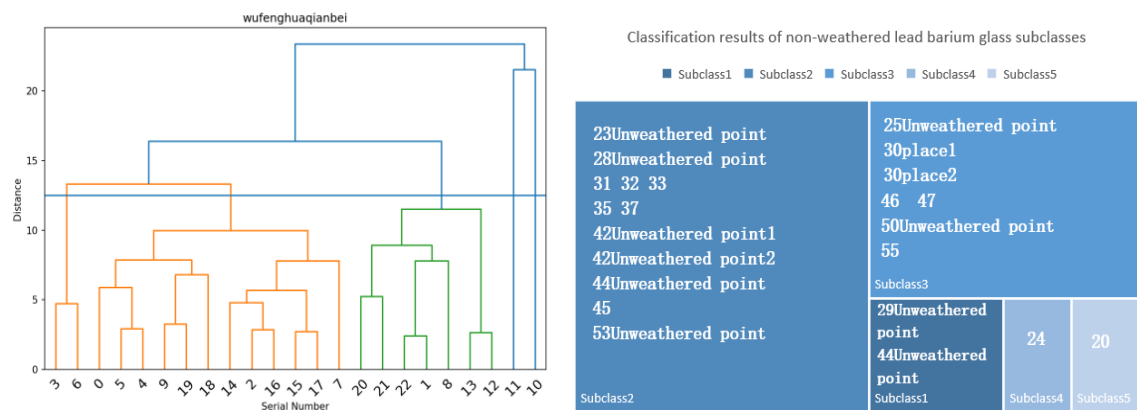


Figure 9. The result when the number of classifications is set at 5

The final list of results is as follows:

Table 1. Classification results of weathered high-potassium glass subclasses

Subclass 1	7
Subclass 2	09, 10, 12
Subclass 3	22, 27

Classification results of non-weathered high-potassium glass subclass	
Subclass 1	21
Subclass 2	03 place 1, 18
Subclass 3	06place1, 06place2
Subclass 4	01, 03place2, 04, 05, 13, 14, 16

Classification results of weathered lead barium glass subclass	
Subclass 1	02, 19, 34, 38, 39, 40, 41, 43place1, 43place2, 49, 50, 51place1, 51place2, 52, 54, 54Severe weathering spots, 56, 57, 58
Subclass 2	48
Subclass 3	11
Subclass 4	08, 08Severe weathering spots, 26, 26Severe weathering spots

Classification results of non-weathered lead barium glass subclass	
Subclass 1	29Unweathered point, 44Unweathered point
Subclass 2	23Unweathered point, 28Unweathered point31, 32, 33, 35, 37, 42Unweathered point1, 42Unweathered point2, 44Unweathered point, 45, 53Unweathered point
Subclass 3	25Unweathered point, 30place1, 30place2, 46, 47, 50Unweathered point, 55
Subclass 4	24
Subclass 5	20

2.5. Subclass Division: Sensitivity analysis of systematic clustering models

The result is shown in the figure 10. According to the elbow principle, combined with the above four aggregation coefficient graphs, the most reasonable classification K value is obtained at the gentle turning point by continuously adjusting the K value, and the sensitivity test and rationality test are completed.

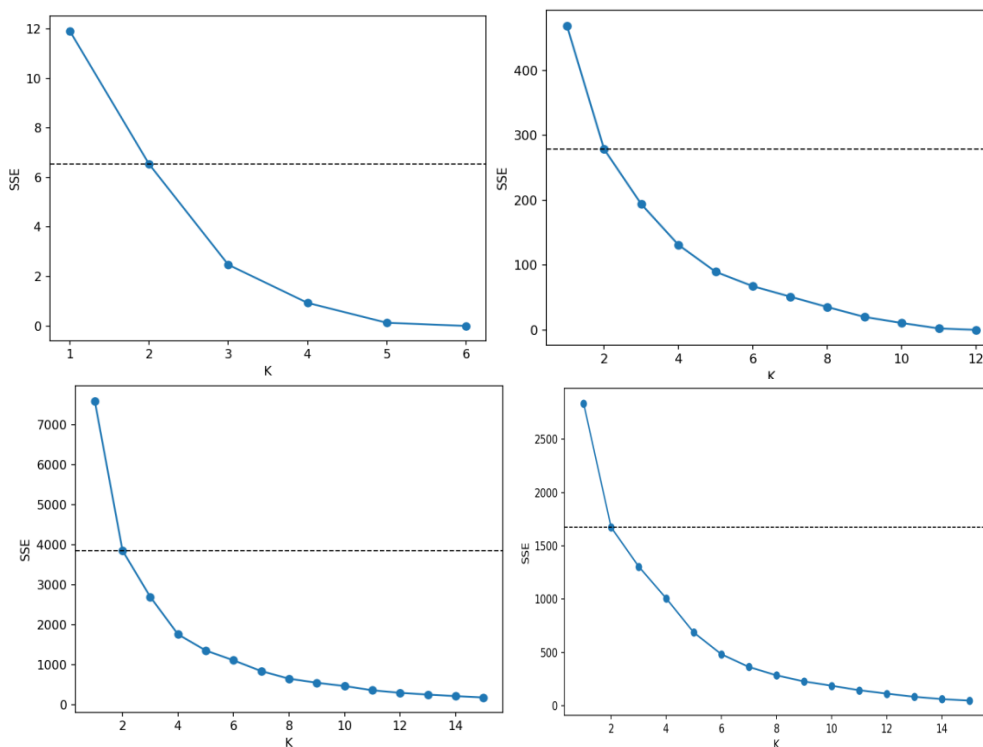


Figure 10. The result when the sensitivity test and rationality test are completed

3. Establishment and solution of glass identification model

3.1. Establishment of random forest models

Since the weathering of unknown cultural relics is also determined, this paper uses weathering and unweathering as the basis for classification, and constructs two random forests respectively. Based on the analysis of problem 2, in order to improve the validity and accuracy of machine learning, the following operations are performed in the data processing and input process:

(1) Assign values to discrete variables: record high potassium glass as 1 and lead barium glass as 0.

(2) Random data input: For each decision tree, 70% of the data volume is selected as the input data, and each data extraction is put back, and the data selection of different decision trees does not interfere with each other.

3.2. Solving the random forest model

The samples A1, A3, A4, and A8 were all non-weathered samples, and their type values were solved by being brought into a weathered random forest. According to the high potassium set to 1 and lead barium set to 0 when the model is established, it can be determined according to the trend of the data results between 0 and 1. Since the process is completely random, there will be subtle differences in each output result, but it does not affect the final type of determination. The three judgment results are listed here to verify the universality of the conclusions, and it can be obtained that A1 is high-potassium glass, and A3, A4, and A8 are lead-barium glass.

The samples A2, A5, A6, and A7 were all weathered samples, and their type values were solved by being taken into a weathered random forest. According to the high potassium set to 1 and lead barium set to 0 when the model is established, it can be determined according to the trend of the data results between 0 and 1. Since the process is completely random, there will be subtle differences in each output result, but it does not affect the final type of determination. The three judgment results are listed here to verify the universality of the conclusions, and it can be obtained that A6 and A7 are high-potassium glass, and A2 and A5 are lead-barium glass.

3.3. Categorical sensitivity analysis of the model

The remaining 30% of the data volume in the model building process is used for testing and sensitivity analysis of the model. With code to achieve an accuracy of 1, it fits the real situation very well.

The final results are summarized in the table 2 below:

Table 2. The final results

Artifact number	Type (inference)
A1	High potassium
A2	Lead barium
A3	Lead barium
A4	Lead barium
A5	Lead barium
A6	High potassium
A7	High potassium
A8	Lead barium

4. Conclusion

In this paper, on a simple observation basis, the random forest model is first used to sort the feature importance, so as to find the classification law, and obtain the PbO, BaO, and SiO₂ contents as the top three important features that distinguish high potassium from lead barium glass. Secondly, in the problem of subclass analysis, this paper adopts a systematic clustering model and uses a cyclic algorithm to reproduce the whole process of clustering, and a total of 16 subclasses are obtained on the basis of the original four categories (weathered high-potassium glass, weathered high-potassium glass, weathered lead barium glass, and non-weathered lead barium glass). Finally, the trend change of the aggregation coefficient elbow chart was used for sensitivity analysis, and it was found that all kinds of fitting were optimal.

In order to identify the glass type, this paper first uses the known data, whether weathering is the classification criterion, and 70% of the data volume as the machine learning basis, constructs two random forest models, and then brings the 8 unknown cultural relics samples in Form 3 into different random forest models according to whether they are weathered or not, and completes the estimation of their glass types. Since the process is completely random, this paper has determined A1, A6 and A7 as high-potassium glass, and A2, A3, A4, A5 and A8 as lead-barium glass. Finally, the sensitivity test was carried out using the unentered 30% data, and the fit degree was found to be 100%, indicating that the accuracy of the prediction was excellent.

References

- [1] Wang Dong, Wen Rui, Zhu Yingpei, Hu Xingjun, Li Wenying. Study on glass beads with metal foil layers excavated from Yingpan Cemetery, Yuli County, Xinjiang [J]. Archaeology and Cultural Relics, 2022(04): 117-122.
- [2] Do Khai Ly, Su Miao, Yang Yuanyuan. Propagation and communion: The influence of Chinese culture on decorative arts along the Silk Road [J]. Journal of Donghua University (Social Science Edition), 2022, 22(02): 37-46. DOI:10.19883/j.1009-9034.2021.0319.
- [3] Cui Ning, Wang stack. Chinese and foreign exchanges on the grassland silk road in Tongliao during the Liao Dynasty [J]. Journal of Inner Mongolia University of Nationalities (Social Science Edition), 2022,48(01):23-28. doi:10.14045/j.cnki.nmsx.2022.01.001.
- [4] Jiang Bo. Ports, shipwrecks and trade goods: archaeological findings and research on the Maritime Silk Road [J]. Studies in the history of maritime transport,2021(04):8-22. DOI: 10.16674/j.cnki.cn35-1066/u.2021.04.002.

- [5] An Jiayao. The Silk Road and glassware [J]. Cultural Heritage World, 2021(12): 87-92.
- [6] Xu Siwen, Gu Zhou, Yang Yimin. The distribution of Indo-Pacific beads and the Maritime Silk Road before the 2nd century A.D. [J]. Chinese Journal of the History of Science and Technology, 2021, 42(02): 281-292+321.
- [7] Wang S. Y. Maritime silk road relics of the Han dynasty in Lingnan from the exhibition of "The Silk Road Begins" [J]. Collection, 2021(02): 162-171.
- [8] 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.
- [9] 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.gjnu.2020.000412.
- [10] Mao Mingxia. Ancient economic and social development of Nanliujiang and the Maritime Silk Road [D]. Guangxi Normal University, 2020. doi: 10.27036/d.cnki.ggxsu.2020.000888.
- [11] Qi Haoyue. Exploration of foreign cultural elements of the Liao Dynasty in the context of the Silk Road [D]. Inner Mongolia Normal University, 2020. doi: 10.27230/d.cnki.gnmsu.2020.000051.