Analysis of the composition of ancient glass objects based on cluster analysis and random forest methods

Zhiying Chen [1, *, #], Dongchen Shang [2, #]

[1] College of Science, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, 212100

[2] Science of Mathematics College, Baotou Teachers' College, Baotou, Inner Mongolia, 014030

* Corresponding author: 949670318@qq.com

[#]These authors contributed equally.

Abstract. In ancient times fluxes were often added to the production of glass to lower the melting point of pure quartz sand. During the smelting process, the addition of different products resulted in significant changes in the internal composition of the glass products. This paper examines a sample of ancient glassworks, based on the data from several sources. The data are divided into 'weathered' and 'unweathered', and cluster analysis reveals significant differences between the categories, and random forest analysis is used to determine the non-linear relationship between the variables, resulting in a chemical composition correlation of the relationships were more variable for potassium chloride and lead chloride and smoother for the rest. This paper innovatively adopts a combination of K-means cluster analysis and random forest to evaluate the composition and identification problems of ancient glass objects. The model in this paper can also be extended to other studies related to the degree of weathering in ancient wooden shipwrecks, metals, ceramics, and other aspects.

Keywords: Cluster analysis, Random Forest, Glass classification study.

1. Introduction

Ancient glass is highly susceptible to weathering from the environment in which it is buried, and weathering can cause changes in the proportions of its chemical composition, thus affecting the correct judgement of its category [1-2]. The history of glass production in China can be traced back to as early as the Western Zhou Dynasty or even earlier. Later, with the opening of the Silk Road and the increasingly frequent cultural and economic exchanges between China and the West, foreign glass (such as bead-shaped glass made in West Asia and Egypt as ornaments) was introduced into China, and its production techniques were learned by local producers, resulting in the two being similar in appearance, but due to the regional characteristics of locally sourced materials, the chemical composition the chemical composition of the two beads differs due to the regional nature of the materials used [3]. Compared to ceramics, silk, and tea, which were transmitted via the Silk Road and became famous overseas, ancient glass and its production techniques are less well known in China, but as valuable physical evidence of the early trade between China and the West, it is still important to study its composition and production techniques [4-5]. By studying the chemical composition and decoration of ancient glass, it is possible to distinguish between indigenous and exotic varieties of glass and to speculate on when they were produced and their specific origin, which helps us to understand the state of economic, cultural, and technological development in ancient times. In ancient times, fluxes were often added to the production of glass to lower the melting point of pure quartz sand. In this paper, this was modelled and analyzed by constructing models such as cluster analysis and random forest respectively, and the results are described below [6].

2. The fundamental of Classification Model

2.1. The fundamental of cluster analysis methods

Cluster analysis, also known as cluster analysis and point cluster analysis, is a form of unsupervised learning and is known, along with regression analysis and multivariate analysis, as the three major methods of multivariate analysis. Cluster analysis is an exploratory analysis method. Unlike discriminant analysis, cluster analysis does not know in advance the criteria for classification, or even how many categories it should be divided into, but will automatically classify samples based on their data characteristics. It can be established as a method of grouping samples according to the degree of similarity of the data without a given division into categories. Its entry reference is a set of unlabeled samples, divided into groups based on the distance or similarity of the sample data, and the division is based on the principle of minimizing the distance within the group and maximizing the spacing outside the group [7].

The basic idea of cluster analysis: there are varying degrees of similarity (affinity - measured by distance between samples) between the samples or indicators (variables) we are studying. So, based on several observations of a batch of samples, some statistics are specifically identified that measure the degree of similarity between samples or indicators, and these statistics are used as a basis for classifying the type. Some samples (or indicators) that are more like each other are aggregated into one category, and some other samples (or indicators) that are more like each other are aggregated into another category, until all the samples (or indicators) are aggregated, satisfying the basic idea of "small differences within classes and large differences between classes". This is the basic idea of classification. It is also the idea that "things come together in classes" [8].

2.2. The fundamental of random forest models

Random Forest, or RF for short, is a combinatorial classification algorithm that belongs to integrated learning. A forest is built in a random way, with several decision trees, each of which is unrelated to the other. After obtaining the forest, when a new input sample enters, each decision tree in the forest makes a separate judgement as to which class the sample should belong to, and then predicts which class the sample should belong to by seeing which class is selected the most. In the process of building each decision tree, there are two things to note about sampling and complete splitting. The first is the two random sampling processes, where the random forest must sample the input data in rows and columns. This is followed by building a decision tree using a full split on the sampled data, so that either one of the leaf nodes of the tree cannot be split further, or all the samples in it point to the same classification. When the number of layers is low, the random forest uses autonomous sampling (bootstrap) to draw multiple samples from the original dataset in a replayed manner. The samples are first trained with a weak classifier-decision tree, and then these decision trees are combined to produce a final classification or prediction by voting. The final classification or prediction result is obtained by voting [9].

There are many classification trees in a random forest. The research process needs to classify an input sample, and we need to input the input sample into all the trees to classify and make decisions. Because each tree in the random forest is an individual, almost all of the unrelated trees will make predictions that contain all of the predictions, and these predictions will cancel each other out. The final prediction results will lead to the existence of only a small number of good trees whose prediction results will exist and thus make a good prediction. This is the idea of random forest bagging, where the classification results of several weak classifiers are voted on to form a strong classifier [10].

3. Results

3.1. The establishment of simulation model

The chemical composition of the unknown category of glass artefacts in the sample was analyzed to identify the type to which they belonged. They are divided into "weathered" and "unweathered" and then the empty spaces are filled with zeros. Using the conclusions drawn from the cluster analysis, it can be seen that for potassium oxide (K2O) and calcium oxide (CaO) there are significant differences between the categories classified by the cluster analysis

3.2. Analysis of experimental results

The K-means clustering algorithm is an iterative solution-based cluster analysis algorithm in which K objects are randomly selected as the initial cluster centers, and then the distance between each object and each seed cluster center is calculated and each object is assigned to the cluster center closest to it. The clustering centers and the objects assigned to them represent a cluster. For each sample assigned, the cluster centers are recalculated based on the existing objects in the cluster. This process is repeated until a termination condition is met. The termination conditions can be that no (or a minimum number of) objects are reassigned to different clusters, no (or a minimum number of) cluster centers are changing, and the error sum of squares is locally minimal. The computational process is to determine a value of k, i.e., the set of k sets we wish to obtain by clustering the dataset. Then k data points are randomly selected from the dataset as the center of mass. Then, for each point in the data set, its distance from each center of mass is calculated, and it is divided into the set to which the center of mass belongs if it is close to the center of mass.

Then, after all the data are grouped into sets, there are k sets in total. Then recalculate the center of mass of each set. Finally, if the distance between the newly calculated center of mass and the original center of mass is less than a certain set threshold value. If the distance between the new center of mass and the original center of mass changes significantly, the above steps need to be iterated.

The results of the cluster analysis are shown in Tables 1 and 2. Cluster analysis can be performed for potassium oxide and calcium oxide and the preliminary results are shown in Table 1.

After comparing the types, it can be found that so of the sample types A6, A7 type is high potassium and A4 is lead barium. The sensitivity of the classification results was analyzed. Adding alumina, the clustering analysis was performed again and the results were obtained as in Table 2.

Table 1. Results of the preliminary cluster analysis

Clustering categories	Central value_potassium oxide	Central value_Calcium oxide
1	0.744	1.436
2	0.68	6.635

Table 2. Cluster analysis after the addition of alumina

Category	Central value_potassium oxide	Central value_Calcium oxide	Central value_aluminium oxide
1	0.98	2.46	2.9075
2	0.387	3.537	9.017

The process of processing the data reveals a nonlinear relationship between variables, so a random forest analysis model was used to model and analyze the association relationship between chemical components using high potassium and lead-barium as types. By comparing the variability of the correlation relationships of chemical components between different categories. A table of variability was generated for the above random forest analysis to obtain the results.

The results were analyzed and after comparison with the raw data, type A7 was high potassium and A4 was lead barium, indicating better sensitivity and more accurate model classification. For different categories of glass artefact samples, the correlations between their chemical compositions were analyzed. Firstly, the data analysis is processed by removing invalid data and filling the empty space to zero, using random forest analysis. Random forest is an algorithm that integrates multiple trees

through the idea of integration learning, and its basic unit is the decision tree, while its essence belongs to a large branch of machine learning - integration learning methods. In this paper, the types are high potassium and lead-barium, depending on the choice, are used to analyses the correlation between the chemical components. The relationship between the input variables and the output results is highly non-linear, and in order to obtain the specific functional form between the two, a random seen model can be used instead of traditional mathematical functions. Also, the random forest model output results can be used as to achieve more accurate optimization. The results of the calculations are shown in Figure 1.

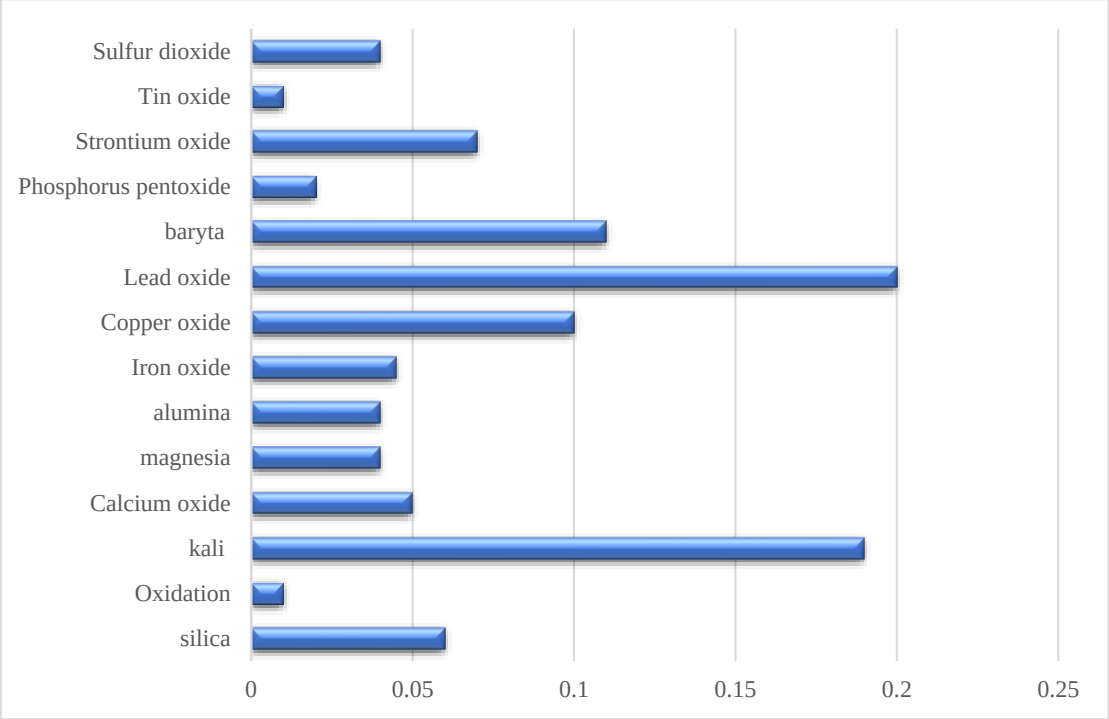

Figure 1. The difference of chemical composition association between different categories

In summary, when the glass type is lead-barium, it is more likely to be influenced by silica, lead oxide and phosphorus pentoxide; when the glass type is high potassium, it is more likely to be influenced by silica, iron oxide, copper oxide and aluminum oxide. When the phosphorus pentoxide content is high, the glass type is more likely to be lead-barium; when the copper oxide content is high, the glass type is more likely to be high potassium, i.e., the glass type is closely related to the environment, and the variability is mainly reflected in the different glass types due to the different content of the above elements.

4. Conclusions

This paper provides a somewhat accurate and clever description of a study of the classification and identification of ancient glass objects, using cluster analysis and random forest models for the classification and identification of components, respectively. K-means cluster analysis was used to give a more intuitive picture of the classification process and a concise form of the conclusions. As the results of the calculations in this paper show, types A6, A7 are high potassium and A4 is lead-barium. The paper then uses random forest analysis, which uses an integrated algorithm that is inherently more accurate than most individual algorithms and has high accuracy, and therefore performs well on the test set. There is no overfitting in this paper using this model and the data obtained are all appropriate values. The analysis of this problem can also be applied to other complex composition product problems, where the parameters can be adjusted to solve the problem with reference to this model. In addition, the model has implications for the planning and selection of industrial production.

References

[1] Liang J, Chen JH, Zhang XUEQIN, Zhou YUE, Lin JIAJUN. Anomaly detection based on unique thermal coding and convolutional neural network[J]. Journal of Tsinghua University (Natural Science Edition), 2019, 59 (07).

[2] WANG Quan, CHENG Xiaofang, FU Teran, LU Shaosong. Definition of the colour gamut of continuous radioluminescence in the visible wavelength band [J]. Science Bulletin, 2002, (13): 972 - 977.

[3] Zhou, Nana, Rao, Zhijian. Analysis of factors influencing grain yield in Yunnan Province based on grey correlation analysis [J]. Agriculture and Technology, 2022, 42 (15): 164 - 167.

[4] Zhang Tie, Chen Jun, Xue Chunzhu, Mu Cunfu. Value analysis of a random forest algorithm-based prediction model for acute kidney injury in postoperative cardiothoracic patients [J]. Journal of Cardiology, 2023, (01): 67 - 71.

[5] He Wenliang, Fu Lianlian, Liao Jingping. Research on pig price forecasting and regulation mechanism based on random forest model [J]. Price Monthly: 1 - 10.

[6] Li Y R. Application of random forests in agriculture [J]. Southern Agricultural Machinery, 2022, 53 (22): 63 - 65+87.

[7] Xia Shuyuan, Dong Yongfeng, Wang Liqin. Research on XGBoost blast block prediction based on feature engineering [J]. Blasting: 1 - 9.

[8] Zhou Yisong, Zhao Chuanping, Huang Yaoming, Zhu Li, Cheng Ming. Research on the prediction of compressive strength of concrete based on machine learning technology [J]. Journal of Anyang Engineering College, 2022, 21 (06): 91 - 95.

[9] XU Qingyu, YU Jing, ZHU Dawei, ZHENG Xiaolong, MENG Lingqi, ZHU Zhiwei, SHAO Yafang. Study on the evaluation of nutritional quality of different rice varieties based on principal component analysis and cluster analysis [J]. China Rice,2022, 28 (06): 1 - 8.

[10] Liu Shuna. Research on glassware of nomadic peoples in ancient China's north [D]. Inner Mongolia Normal University, 2022.