Analysis and identification of ancient glass components based on qualitative-quantitative data mining

Jiaqi Xu *, Yifeng Xu, Yuan Wan

School of Civil Engineering, Tongji University of China, Shanghai, China, 200092

* Corresponding Author Email: 2050523@tongji.edu.cn

Abstract. First of all, the data is visualized by grouping the data in the form and drawing a scatter diagram. By observing the distribution of the data in the scatter plot, the classification indexes of high-potassium glass and lead-barium glass were found. Secondly, the Spearman coefficient is used to test these classification indicators. The results show that the constructed classification system is more accurate, the Spearman coefficient is small, and the fitting effect is excellent. Secondly, through the combination of qualitative and quantitative methods, the cultural relics sub-category classification index was double screened, and the chemical composition of the sub-category classification index was determined by combining the law analyzed from the scatter plot and the index weight solved by the entropy weight method. Then, the high-potassium glass was divided into low-calcium and low-copper group and high-calcium and high-copper group by K-means method. The glass was divided into high copper group, low sodium and high calcium group and high sodium and low calcium group. Finally, all the unknown types of cultural relics in the table are divided into subclasses, and the sensitivity and rationality of the classification system of major classes and subclasses are tested one by one by fitting the known data and calculating the contour coefficient. It is concluded that the classification system has a good fit to the data.

Keywords: Data mining; Distribution fitting; Correlation coefficient; BP neural network; K - means clustering

1. Introduction

The Silk Road was an important cultural exchange channel between China and the West in ancient times, and it has important historical significance and cultural symbolism in the long history of Chinese civilization [1-3]. Among them, ancient glass products, as an important symbol of early trade between Chinese and foreign cultures, are of great value [4]. Early glass often flowed into China through the Silk Road from West Asia and Egypt in the form of jewelry, and the Chinese glass manufacturing technology was improved after learning the production methods of Western glass products, resulting in the appearance of Chinese ancient glass products and foreign products are very close, but the chemical composition is different [5].

In order to analyze and study the differences between Chinese and foreign ancient glass products, it is necessary to analyze and classify the composition of excavated glass products, because most ancient glass products may have weathered to different degrees when excavated [6-8]. Therefore, it is of great research value to identify the types of products whose types are not clear by means of chemical composition analysis based on known data to find out the rules of composition analysis [9-10].

This study will address the following questions. Analyze the relevant basic information given in the form to find out the correlation between glass type, decoration and color and whether the product is weathered; Secondly, the regularity of chemical composition of weathered and unweathered ancient glass products is summarized from the perspective of statistical law. Finally, according to the detection data of weathering points in the form, the chemical composition content of these points before weathering is predicted.

2. Model establishment and solution of problem

2.1. Look for the classification rules of the two glass types

Step 1: Scatter plot analysis

We will according to the information in the form 3, 14 kinds of chemicals to study separately, and observation points can be divided into weathering and unweathered group, according to the two kinds of classification methods, we put all the data sample is divided into 28 groups, with the number of point as the abscissa, chemical component content as the ordinate, draw corresponding scatter plot of data one by one ", We selected two representative pictures, namely, the distribution of silica content in two kinds of glass before and after weathering, and the "chemical dispersion dot plot" in "Problem 1" of the supporting material can be seen in the rest scatter plots. (In the Figure 1, blue dots represent high-potassium glass and red dots represent lead-barium glass)

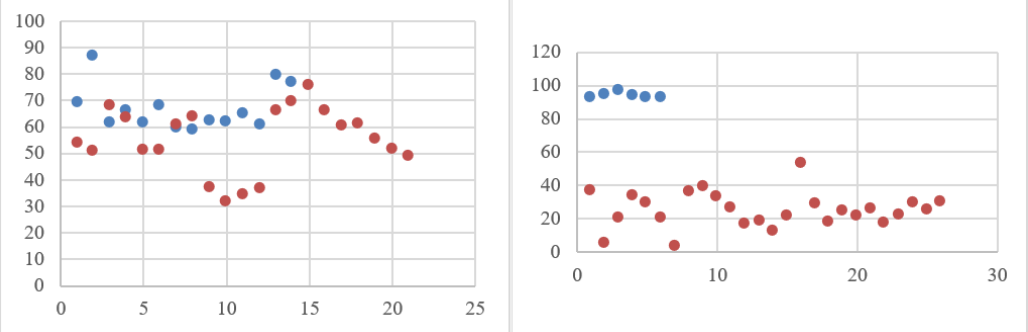

Figure 1. Scatter plot of silica content distribution before (left panel) and after (right panel) weathering

By observing the graphic features of the scatter plot we drew, we found out the content change rules of different chemical components on the surface of two types of glass before and after weathering and summarized them as follows:

(1) In high-potassium glass, the contents of sodium oxide, lead oxide, barium oxide, strontium oxide, tin oxide and sulfur dioxide all decreased to 0 after weathering, while in the other three groups, the contents of these chemical components of cultural relics were not all 0;

(2) For the high-potassium glass and lead-barium glass before weathering, the content of potassium oxide and lead oxide between the two kinds of glass are very different;

(3) The content of silica, the main component of weathered high-potassium glass, can reach more than 90%, while the content of other chemical components are greatly reduced or even disappeared in the process of weathering.

Step 2: Selection of major classification indicators

In the selection of classification indicators, entropy weight method was used to calculate the weight of chemical components in the unweathered Pb and barium cultural relics and the high potassium cultural relics, and the components with greater differences were selected as the indicators to identify the types of unweathered cultural relics. We also adopt the same index selection scheme for the classification of weathered cultural relics. The weights of the data of each group are recorded in the "chemical composition weights" in "Problem 2" of the support material. The specific indicators we selected are as follows: potassium oxide, lead oxide and barium oxide were selected for the cultural relics before weathering; Silica, magnesium oxide, lead oxide and barium oxide were selected as classification indexes for the differentiated cultural relics.

Step 3: Spielman's method was used to test major classification indexes

After determining the categories of indicators, we need to test the reliability of the indicators. Here, Spearman method and Euclidean distance method are used again to test the rationality of the indicators we select. Since the specific principles and procedures have been analyzed in detail before, the details are not repeated here.

Finally, according to the results we obtained, the correlation coefficients between the indicators we selected and the types are all less than 0.05, which is suitable for category judgment. The specific correlation coefficients can be seen in the "Spearman method to test the results of major classification indicators" in the "Question II" of the support material.

After determining the rationality of the index, we used Euclidean method to calculate the weighted distance between the sample data and the center points of different categories by combining the weight of each chemical content solved by entropy weight method in the above content. The final clustering centers of the major classification indexes were shown in Table 1. Then the type corresponding to the center point with the smallest distance from the sample data is taken as the type of the data sample.

Table 1. The cluster center of the major classification index

The final cluster center		
	Clustering	
	1	2
K_2O	8.94	.26
PbO	.38	23.59
BaO	.51	10.50

2.2. Division of subclasses

When classifying subclasses, we first make it clear that, since the chemical composition content of cultural relics will change in the process of weathering, we only consider the value of unweathered cultural relics, and do not consider the weathered cultural relics.

Step 1: Qualitative screening of subcategory indicators

We observed the scatter plots drawn in the above steps and tried to preliminarily determine the sub-classification indexes through the scatter distribution rules in the figures. Therefore, we took out 14 unweathered scatter plots and summarized the data classification rules of high-potassium glass and lead-barium glass respectively.

For high-potassium glass, the data points of sodium oxide are obviously divided into two parts. There are 3 data points with zero content and 5 data points with more than 2% and less than 4% content, which can be used as one of the criteria for subclass classification. Similarly, the data points of calcium oxide and copper oxide are also more obvious block phenomenon. So we chose copper oxide, calcium oxide and sodium oxide.

The same analysis of lead and barium glass can be performed.

Step 2: Screening of quantitative subclass indicators based on weight

In order to screen the sub-category indicators at the quantitative level, we used the entropy weight method to screen the weights of the two categories of glass in unweathered cultural relics again. The weight distribution of specific chemical components is shown in the following table.

Table 2. Chemical composition weight distribution in unweathered cultural relics

Chemical composition	**CuO**	**Al_2O_3**	**BaO**	**P_20_5**	**SrO**	**SnO**	**$S0_2$**
High potassium glass	0.021339	0.118147	0.06757	0.04265	0.054041	0.027642	0.081062
Lead barium glass	0.043483	0.097381	0.01073	0.03613	0.020775	0.013838	0.033579
Chemical composition	**CuO**	**Al_2O_3**	**BaO**	**P_20_5**	**SrO**	**SnO**	**$S0_2$**
High potassium glass	0.068732	0.016611	0.03304	0.073168	0.047203	0.147204	0.201593
Lead barium glass	0.023888	0.078574	0.12346	0.045207	0.084443	0.244941	0.14357

In this table, we can see that for high-potassium glass: the five chemical components with the largest weight are sulfur dioxide, tin oxide, sodium oxide, iron oxide and phosphorus pentoxide; For lead-barium glass: the five chemical components with the largest weight are respectively tin oxide, sulfur dioxide, barium oxide, sodium oxide and strontium oxide.

Step 3: Combined with qualitative and quantitative subcategory indicators for secondary screening

After the qualitative and quantitative subcategory index analysis, we selected a large number of indicators. In order to reduce the number of categorical indicators, we also need to combine the qualitative analysis results with the quantitative analysis results for secondary screening.

Before we start filtering, let's stipulate a few rules to simplify the following operations:

1, extremely trace elements do not participate in the classification consideration. That is, when the overall content of a chemical component is less than or equal to 0.1%, we choose to reject the component as a classification index.

2. Categorization of indexes with too few cases of non-zero chemical composition content is rejected. For example, there is only a non-zero data value of 2.36% for tin oxide in high-potassium glass, so the data amount is too small and the data value is too small, which lacks reliability.

3. Conduct crony analysis on relevant implicated data. Spearman correlation analysis was used to establish the significance value of each chemical component, and the correlation between the two indexes <0.05 was analyzed.

4. Index classification was conducted only for the data before weathering, and the data of chemical composition content after weathering were distorted.

After determining the rules, we began to conduct secondary screening for sub-classification indicators, and the specific screening process was as follows:

First of all, we selected all the indexes determined by the entropy weight method, and added the corresponding indexes selected by the qualitative analysis part. For high potassium glass, we selected sodium oxide, calcium oxide, copper oxide, tin oxide, barium oxide, sulfur dioxide and strontium oxide to participate in the following comparative screening process; For lead-barium glass, we selected potassium oxide, magnesium oxide, iron oxide, sodium oxide, barium oxide, phosphorus pentoxide, strontium oxide, copper oxide seven indicators to participate in the following comparative screening process.

Then, we delete three indexes of tin oxide, sulfur dioxide and strontium oxide in high potassium glass index according to rule one and rule two. Strontium oxide in lead barium glass was deleted.

Secondly, according to the analysis of barium oxide content in the data table, the classification significance of high potassium glass is not obvious, so it is also rejected. As for lead-barium glass, the content classification significance of potassium oxide, magnesium oxide and barium oxide is not particularly obvious, so it is rejected. It is worth mentioning that there is only one outlier in the data of phosphorus pentoxide index that affects its weight, so it is also rejected.

Finally, we with high potassium residual sodium oxide, calcium oxide in the glass and copper oxide three component index correlation analysis of nepotism inspection, found that in addition to sodium oxide and cupric oxide correlation coefficient is 0.01, the other no correlation, considering the 0.01 is not strong correlation, therefore, we believe that these three indicators are more freedom indicators, It is an independent index for cluster analysis. For lead-barium glass, the correlation of iron oxide, sodium oxide and copper oxide is not significant, so it is considered that the three indexes are relatively free and independent indexes for cluster analysis.

In summary, the subcategories of sodium oxide, calcium oxide and copper oxide were selected for high-potassium glass. Iron oxide, sodium oxide and copper oxide were selected as the subindexes of lead and barium glass.

Step 4: K-means subclass classification

After selecting appropriate subcategory indicators, we need to further carry out specific subclassification of the two categories of glass products. Therefore, we choose to use K-means method for clustering division.

The basic principle of K-means is based on Euclidean distance method. It can achieve the goal of cluster analysis by continuously calculating the distance between sample data and the center point -- dividing it into clusters corresponding to the nearest center point -- recalculating the coordinates of the center point calculating the distance between sample data and the center point again. The algorithm complexity of K-means is simple and suitable for the working environment where a large amount of data is processed in a short time, so it is more consistent with our needs.

When conducting K-means cluster analysis, we also studied high-potassium glass and lead-barium glass separately.

First, high potassium glass:

The value of K is 2, and the corresponding data center points of various types of cultural relics in the initial state are calculated. Then, the distance between each high-potassium glass and these center points is solved to represent the fitting degree of samples with different categories and subclasses. The specific calculation formula is as follows:

Bringing the data into the functional calculation leads to a set of sub-classification cases, as described in "Classification Results for high-potassium Glass" in the supporting material "Question 1".

Therefore, we finally get two kinds of high-potassium glass subclasses, and according to the characteristics of chemical composition and content respectively named: low calcium and low copper group (red data points) and high calcium and high copper group (blue data points), and respectively denoted.

Two, lead and barium glass:

Lead barium glass when dealing with specific ideas and practices are consistent with high potassium glass, go, finally, we concluded that lead the class there are three kinds of barium glass, and respectively named: high copper set of data points (red), low sodium and high calcium group of data points (in blue) and higher sodium (green data points) and the low calcium group as.

The location of each data point in the cluster analysis is shown in the Figure 2.

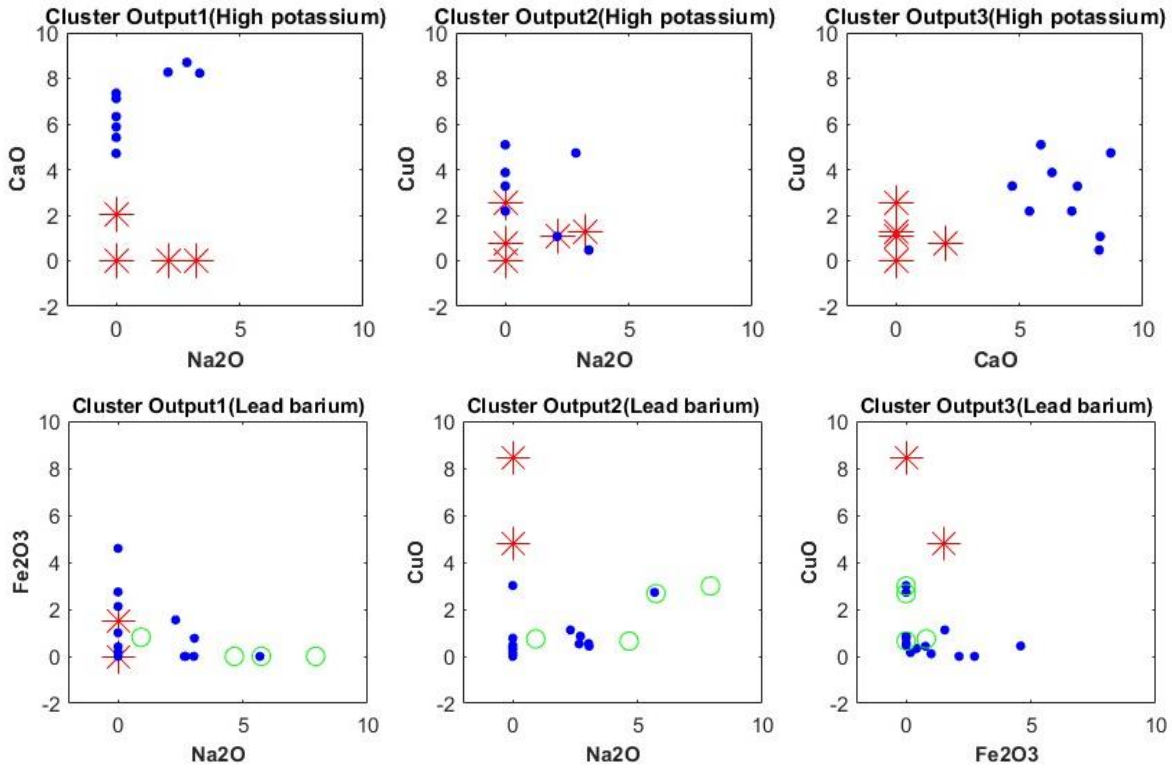

Figure. 2 Scatter plot of cluster analysis results

2.3. Analysis of reasonableness and sensitivity of classification model of major class and subclass

After the establishment of the classification system of major classes and subclasses, we further analyzed the sensitivity and rationality of the model.

First of all, we analyze the reasonableness of the system. And this kind of method to form groups of original data in the 2 kinds of forecast, after comparing actual situation and forecasting results we found that the vast majority of predict the type of the sample data group were in line with the real situation, only the division of 48 test point will appear error, therefore, we can roughly determine the categories of classification model is relatively accurate.

Then we analyze the rationality and sensitivity of the subclass model. In the fourth step of the subclass model establishment process, when the K-means method was used for subclass division, we respectively selected K=2 and K=3 as the number of subclasses of high-potassium glass and lead-barium glass. However, the change of K value will change the accuracy of the model. We try to cluster the high-potassium glass data into different number of subclasses, and compare the advantages and disadvantages of clustering results with the value of contour coefficient. The specific data table is shown in "Problem Diclustering Analysis and Calculation Results of contour Coefficient.txt" in the supporting materials. It can be seen from the table that, for high-potassium glass, the contour coefficient at K=2 is about 0.58, which is the best one in the simulated data group. Similarly, we have done the same operation for the lead and barium glass group. Finally, the contour coefficient when K=3 is about 0.49, which is the second best one in the simulated data set. The contour coefficients of both are relatively large, and both are at a high level when the value of K fluctuates. Therefore, we believe that the selected value of K is reasonable to a certain extent. We can see that after changing the value of K, the fluctuation range of the contour coefficient is not large, so we believe that the established model also has good sensitivity.

3. Conclusions

3.1. Advantages of the model

1. Deep data mining was carried out to analyze the relationship between basic attributes and weathering respectively qualitatively and quantitatively, and the correlation coefficient test method was used to support the conclusion

2. In the analysis of the statistical law of chemical composition before and after weathering, the corresponding treatment of each component index was carried out to ensure the reliability and perfection of the model.

3. The model shows a good fit in terms of clustering and classification, which indicates that the clustering index obtained through qualitative-quantitative analysis has strong feasibility.

4. In Question 1, Spearman correlation coefficient was used to test the basic attribute model; in question 2, K-means clustering parameters were adjusted to analyze the rationality and sensitivity; in question 3, BP neural network was used to test the classification results; in question 4, grey correlation degree matrix was used to test the model rules. All of them show good fitting and strong reliability.

3.2. Disadvantages of the model

1. If there is a large difference in chemical composition content due to different sampling points in the classification of the model, there is no good classification effect.

2. Some subclassification indexes are specialized in the association rules, which may affect the classification results.

3. The model does not take into account the general differences in chemical composition between the unweathered spots and the unweathered materials.

References

[1] Wu Jun, Yin Li, Zhang Maolin, Wu Junming, Li Qijiang. Comparison of Artificial Neural Network and multivariate statistical discriminant Analysis in fault-source dating of ancient ceramics [J]. Chinese Journal of Ceramics,2014,35(04):429-435.

[2] Li Daoquan, Yang Qianqian, Lu Xiaov. Intrusion Classification and Detection model of SDN Network based on Decision Tree [J]. Computer Engineering and Design, 222,43(08):2146-2152.

[3] Shi Shoukui, Sun Zhaoliang. Mathematical Modeling Algorithm and Its Application [M]. 2nd Ed. National Defense Industry Press,2019.

[4] Jiang Qiyuan, Xie Jinxing Ye Jun. Mathematical Models [M]. 5th Ed. Higher Education Press,2018.

[5] LUO Hongjie. Ancient Chinese ceramics and multivariate statistical analysi[M]. Beijing：China Light Industry Press,1997.

[6] Zhou Ting, Zhang Junying, Luo Cheng. Implementation of K-means Clustering Algorithm based on Hadoop [J]. Computer Technology and Development, 2013(7): 18-21.

[7] Zhao Qin. An efficient Canopy-Kmeans algorithm based on Hadoop Platform [J]. Electronic Science and Technology, 2014, 27(2): 29 -- 31.

[8] Shan Yibin. [J]. Research on Financial and Economic Issues.1999(03): 68.

[9] Yang Zunqing. [J]. Journal of Beijing Technology and Business University (Social Science Edition)1985(02): 102 -- 106.

[10] Zhang Liming. The Model and Application of Artificial Neural Network Shanghai: Fudan University Press, 1994.