# Composition Analysis and Identification of Glass Products Based on Hierarchical Clustering

Cunnan Jia *, Zhuangwen Gong

College of Information and Intelligence Science and Technology, Hunan Agricultural University, Changsha, China

* Corresponding Author Email: 3110993808@qq.com

**Abstract.** Ancient glass is an important carrier of the development of Chinese civilization. Affected by the buried environment of cultural relics, ancient glass is very easy to weather. A large number of internal elements of glass exchange with environmental elements, resulting in changes in its composition and appearance, thus affecting the judgment of its correct classification. In this paper, a comprehensive evaluation model was established to analyze the chemical composition of glass. First of all, the chi-square test is used to determine the factors related to surface weathering. Then the chemical composition of different glass was analyzed by mathematical statistics, and the number of clustering categories was determined by the method of system clustering, combined with the determination of square Euclidean distance and elbow rule. Finally, the final classification result is obtained by judging the size of the grey correlation degree of the type classification law and the sub-classification division method obtained in the model. The results show that the comprehensive model can be used to analyze and identify the composition of glass products.

**Keywords:** Chi-square test, entropy-weighted mean combination forecasting, hierarchical clustering, grey correlation.

## 1. Introduction

Glassware was an important commodity in the trade between China and the West in ancient times. But ancient glass is very easy to weather. A large number of internal elements [1] of glass exchange with environmental elements, resulting in changes in its composition and appearance, and thus affecting the judgment of its correct classification, so it is urgent to propose a comprehensive evaluation model to analyze the chemical composition of glass products.

First of all, the chi-square test was used to test the four variables of glass classification. According to the chi-square value and P value, it was judged that the type had a significant relationship with the surface weathering, while the color and texture had no significant relationship with the surface weathering. The classification rules of the two types of glasses were obtained by the mathematical statistics of the chemical composition of different glasses, combined with the visual processing of decorative patterns, colors and other characteristics. Then through the method of hierarchical clustering, combined with the determination of square Euclidean distance and the elbow rule, the number of clustering categories is determined. Divide into two categories according to whether there is weathering or not, according to the type classification law and the grey correlation degree of the sub-classification division method obtained by each category of data in the second model, judge and obtain the final classification results, and then establish a grey correlation analysis model to calculate the grey correlation degree between each category of chemical components. Therefore, the model can be used to analyze and judge the chemical composition of ancient glass [2].

## 2. Combination forecasting model

### 2.1. Surface weathering in relation to glass type, ornamentation, and colour

We can roughly understand the relationship between each glass type, ornamentation and colour and weathering through the accumulation scale diagram. When the colour is black or the ornamentation is B, the glass is weathered. As shown in Figure 1.

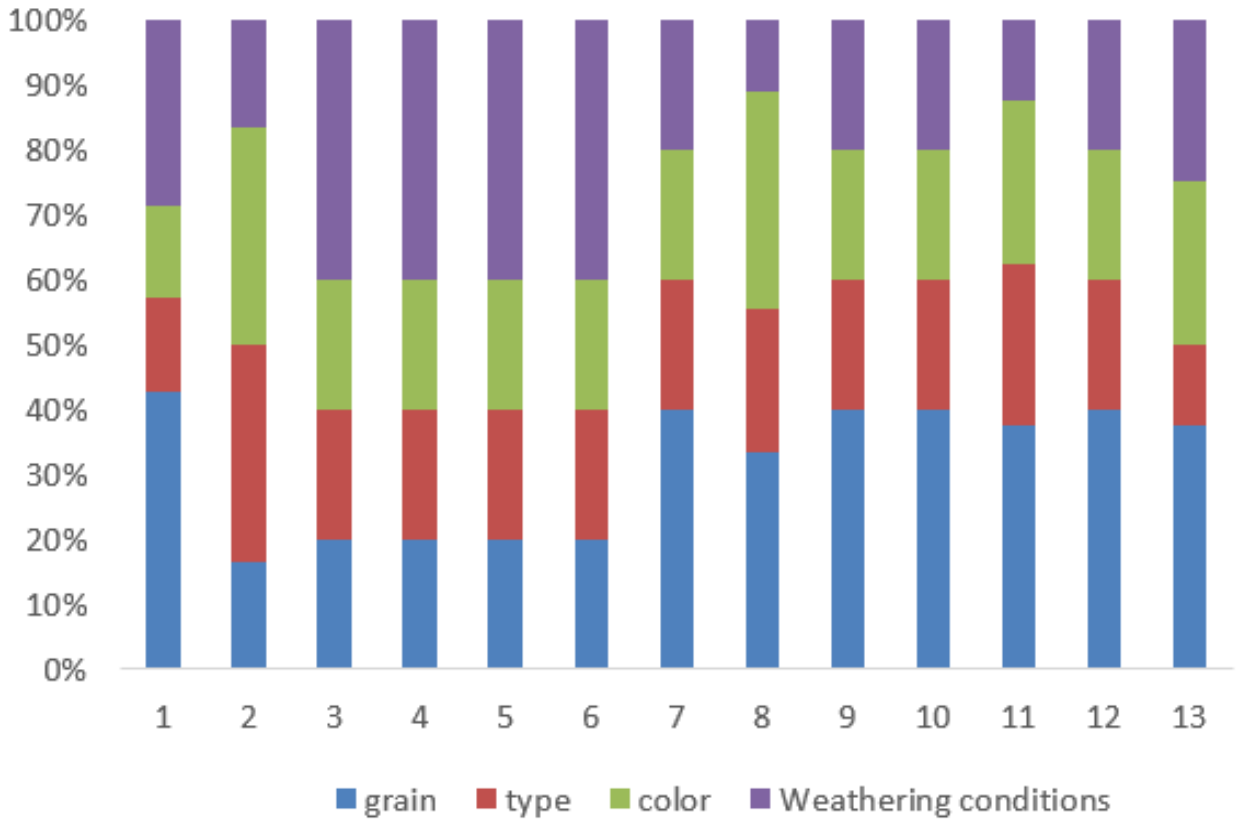


**Figure 1.** Weathering Scale in Glass Type, Texture and Color.

To quantify the differences between multiple groups of categorical variables, the chi-square test was used for analysi [3]. The Chi-square test is a widely used hypothesis testing method, which belongs to the category of non-parametric test. The main purpose is to compare two or more sample rates and the correlation analysis of two categorical variables. The degree of deviation between the two is judged by the chi-square value, that is, the smaller the chi-square value is, the smaller the observed value (actual value) and the expected value (theoretical value) are. The closer they are, the more consistent the chi-square distribution is between the two variables. The premise of a chi-square distribution is that variables are independent of each other, so the smaller the chi-square value is, the more independent the two variables are. In this paper, with the help of categorical variable analysis in the data, the chi-square test steps are as follows:

(1) Calculate the proportion of the chemical composition of each column in the composition of each cultural relic as formula (1).

$$q_{ij} = \frac{\delta_{ij}}{\sum_{i=1}^{n} \delta_{ij}}, i = 1,2,\cdots,58, j = 1,2,\cdots 14 \tag{1}$$

Where $\delta_{ij}$ the content of the chemical composition of the cultural relic is numbered i in the j column, and $\sum_{i=1}^{n} \delta_{ij}$ is the sum of the content of the chemical composition of the cultural relic numbered i.

(2) Calculate that entropy value $\sigma_j$ determined by the content of the chemical components in each row as formula (2).

$$\sigma_j = -a\sum_{i=1}^{n} q_{ij} \ln q_{ij}, i = 1,2,\cdots,58, j = 1,2,\cdots 14 \tag{2}$$

(3) Calculate the coefficient of variation $\lambda_j$ of the predicted weight of the chemical component in the jth column as formula (3).

$$\lambda_j = 1 - \sigma_j, j = 1,2,\cdots,14 \tag{3}$$

Since $\sigma_j \in [0,1]$, we can make it clear that there is a negative correlation between the degree of variation of the chemical composition in column j and its entropy.

(4) Calculating a combined weighting coefficient of each column of indexes as formula (4).

$$w_j = \frac{1}{m-1}\left[1 - \frac{\lambda_j}{\sum_{i=1}^{n} w_i}\right] \tag{4}$$

Through the above steps, we can obtain the combined weighting coefficient of each chemical component when the content of some chemical elements is not detected, to facilitate the subsequent calculation.

## 2.2. Combination Forecasting Model Based on Component Distribution Matrix

Having obtained the combined weighting coefficients for each column of data, we first establish a matrix of chemical element distributions as formula (5).

$$R_{ij} = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1\rho} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{\tau 1} & \delta_{\tau 1} & \cdots & \delta_{\tau\rho} \end{bmatrix} \cdot \tau = 58, \rho = 14 \tag{5}$$

We multiply the chemical element composition distribution matrix with the combined weighting coefficient to obtain the weighted composition distribution matrix as formula (6).

$$R_{ij}^{'} = w_j \cdot \begin{bmatrix} \delta_{11} & \delta_{1\rho} \\ \delta_{\tau 1} & \delta_{\tau\rho} \end{bmatrix} \cdot \tau = 58, \rho = 14 \tag{6}$$

Each data in $R_{ij}^{'}$ is the chemical element composition after the weighted assignment, and we can obtain its weighted harmonic mean. Here we introduce the harmonic mean calculation formula (7).

$$\overline{x}_t^{'} = \frac{\sum_{i=1}^{n} w_j^{'}}{\sum_{i=1}^{n} \frac{w_j^{'}}{\delta_{ij}}}, i = 1,2,\cdots,58, j = 1,2,\cdots,14 \tag{7}$$

Where $\overline{x_t}^{'}$ is the weighted average of the columns of chemical composition that have been weighted and combined after weathering

Obtain the weighted value of the composition distribution, as shown in the formula (8).

$$\begin{cases} \xi_{ij} = \overline{x}_t^{'} \cdot w_j \cdot \begin{bmatrix} \delta_{11} & \delta_{1\rho} \\ \delta_{\tau 1} & \delta_{\tau\rho} \end{bmatrix} \\ \xi_{ij}^{'} = \overline{x}_t^{'} \cdot w_j^{'} \cdot \begin{bmatrix} \delta_{11}^{'} & \delta_{1\rho}^{'} \\ \delta_{\tau 1}^{'} & \delta_{\tau\rho}^{'} \end{bmatrix} \end{cases} \tag{8}$$

## 2.3. Predicted results

We screened and weighted the data [4], and through the above process and programming with Matlab, we predicted the chemical composition content of high-potassium and lead-barium glass before weathering, as shown in Table 1 below.

**Table 1.** Predicted results.

| Type | Cultural relic number | $\delta_1$ | …… | $\delta_2$ | …… | $\delta_{13}$ | $\delta_{13}$ |
|---|---|---|---|---|---|---|---|
| High potassium | 7 | 75.67639851 | | 5.587284085 | | 0 | 0 |
| High potassium | 27 | 72.47684669 | | 2.575203209 | | 0 | 0 |
| High potassium | 9 | 73.8967498 | …… | 2.497491032 | …… | 0 | 0 |
| High potassium | 22 | 59.01382236 | | 0.773504076 | | 0 | 0 |
| High potassium | 12 | 67.5628886 | | 2.552126451 | | 0 | 0 |
| …… | | | | | | | |
| Lead and barium | 39 | 60.88844235 | | 0.550766587 | | 0 | 0 |
| Lead and barium | 48 | 69.79115372 | | 0 | | 0.580698724 | 0 |
| Lead and barium | 38 | 59.66077176 | …… | 0.358655396 | …… | 0 | 0 |
| Lead and barium | 34 | 70.00846143 | | 0.790962647 | | 0 | 0 |
| Lead and barium | 36 | 61.43020589 | | 0.285697951 | | 0 | 0 |

## 3. Glass classification model based on clustering

Glass classification model [5] based on clustering following figure shows the statistics of chemical content and color decoration of high-potassium glass and lead-barium glass:

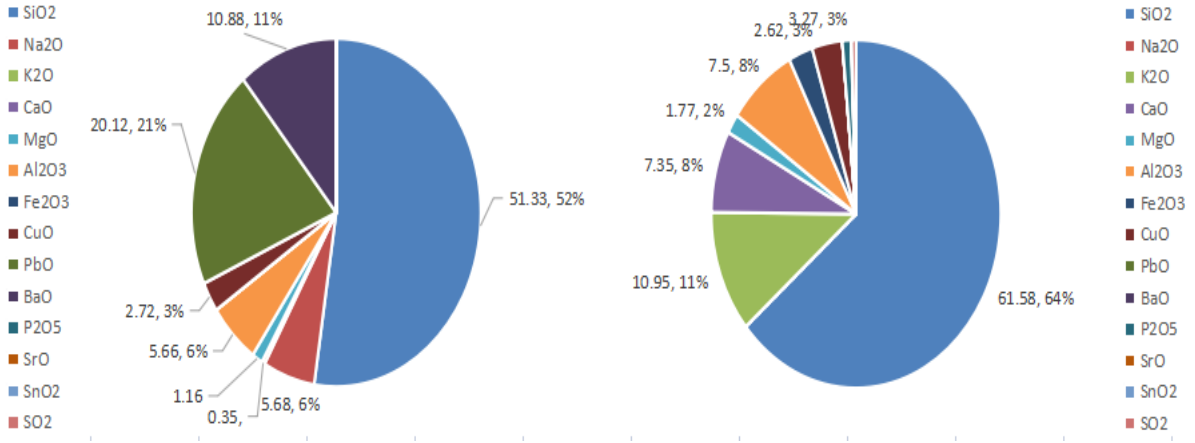


**Figure 2.** Visualization of the chemical content of lead-barium glass (left) and high-potassium glass.

From the above figure 2, we can see that silicon dioxide (SiO2) is the main component in the two types of glass. Next, we analyze the colour and texture of the two types of glass, and the results are shown in the following table:

For the two kinds of glass that have been divided: high-potassium glass and lead-barium glass, because the sample data of each kind is not large, this paper can use the method of system clustering to divide the two kinds of glass twice. It can be seen that 22 iterations have been carried out. Draw the hierarchical clustering pedigree diagram and the centralized planning coefficient as Table 2.

**Table 2.** Statistical table of principal components of color and ornamentation.

| Glass type | Weathering or not | Ornamentation | Color | Coverage |
|---|---|---|---|---|
| High potassium | Weathering | B | Blue and green | 99.79%-100% |
| | No weathering | A, C | Blue-green, light blue, dark blue | 96.25%-100% |
| Lead and barium | Weathering | A, C | Black, blue-green, light blue, light green, dark green, purple | 91.17%-99.86% |
| | No weathering | A, C | Dark blue, light blue, dark green, light green, purple, green | 88.31%-99.98% |

The following conclusions [6] can be drawn from the above diagram: the glass with potassium oxide content of more than 1% or decorative type B shall be classified as high-potassium glass; the glass with barium oxide and lead oxide content of more than 1% or dark colors such as black and purple shall be classified as lead-barium glass. As shown in Figure 3.

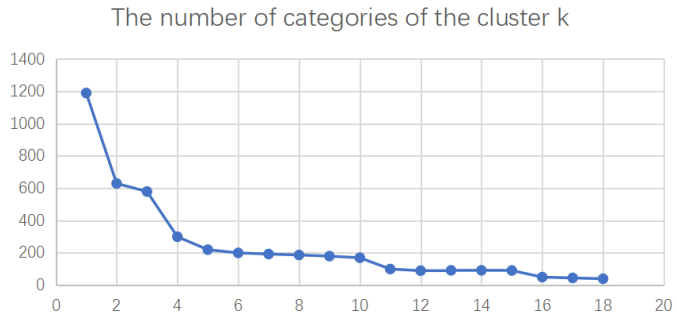


**Figure 3.** Scatter diagram of centralized planning coefficient.

According to the above figure, we can use the elbow rule to determine the number of categories. According to the broken line chart of the aggregation coefficient, when the number of categories is 2, the downward trend of the broken line slows down, so the number of categories can be set to 2.

According to the above steps, the unweathered lead-barium glass, the unweathered high-potassium glass and the weathered high-potassium glass were subjected to hierarchical cluster analysis, and the classification results were visualized as Figure 4.

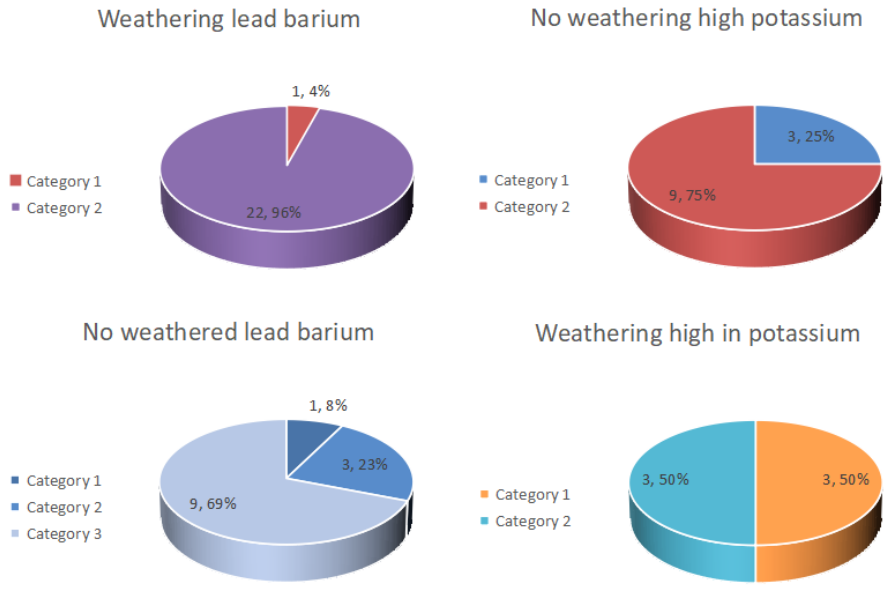


**Figure 4.** Cluster Visualization Results Error! No sequence specified.

Because in the process of classification [7], we select representative chemical components for the main analysis and discrimination, there are deviations, so we make a specific analysis of the rationality and sensitivity of the results.

First, we perform data perturbation on the main chemical components, and control the amplitude change of the variable based on (-30% -30%) itself to obtain the results after perturbation, as shown in the following Table 3.

**Table 3.** Classification accuracy after data disturbance.

| Category | -30% | -15% | 5% | 10% | 20% | 30% |
|---|---|---|---|---|---|---|
| High potassium weathering | 97% | 100% | 100% | 100% | 100% | 100% |
| High potassium unweathered | 100% | 100% | 100% | 100% | 100% | 90% |
| Lead-barium weathering | 100% | 100% | 100% | 100% | 100% | 100% |
| Lead-barium unweathered | 95% | 100% | 100% | 100% | 100% | 92% |

It can be seen from the above table that when the data is disturbed at the amplitude of -20% -20%, the accuracy of the model results reaches 100%. When the threshold value is 30% of the positive and negative disturbance amplitude, the accuracy of the unweathered results of lead and barium is 92% and 95% respectively. When the disturbance amplitude is -30%, the result accuracy of high-potassium weathering is 97%. With a perturbation amplitude of 30%, the accuracy of the unweathered results of the high-valence class is 90%. The average classification accuracy in this range is 98.91%, which shows that the model has high rationality and stability.

## 4. Glass-type discrimination model based on grey correlation analysis

As the change in chemical composition of the weathered glass will have a greater impact on the discrimination results, we first divide the cultural relics in Table 3 according to the weathering of the cultural relics, and the division is as Table 4.

**Table 4.** Classification of unknown data with or without weathering.

| | Unweathered | Weathering |
|---|---|---|
| Cultural relic number | A1,A3,A4,A8 | A2,A5,A6,A7 |

The mean value of chemical components of each class and subclass in the second model is obtained, and it is used as the discrimination data for dividing the unknown data into this class or subclass. The average value of major categories is shown in the Figure 5.

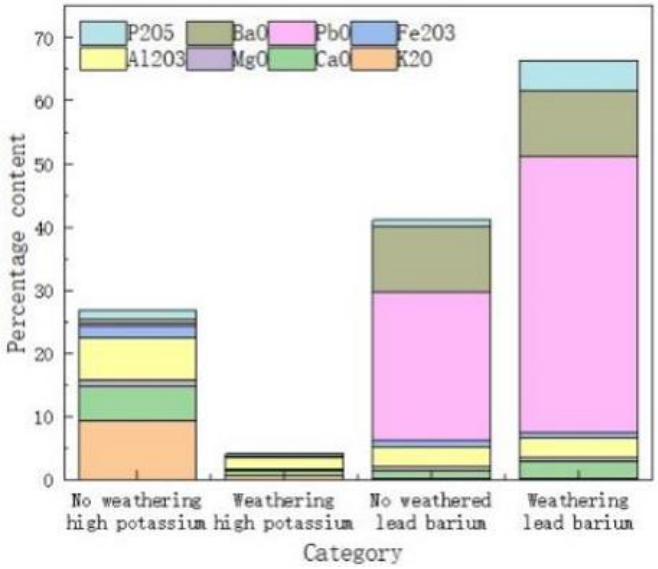


**Figure 5.** Average Data Value of Major Chemical Composition.

The grey correlation degree between A1 and unweathered high-potassium and unweathered lead barium obtained by MATLAB is shown in Figure 6.

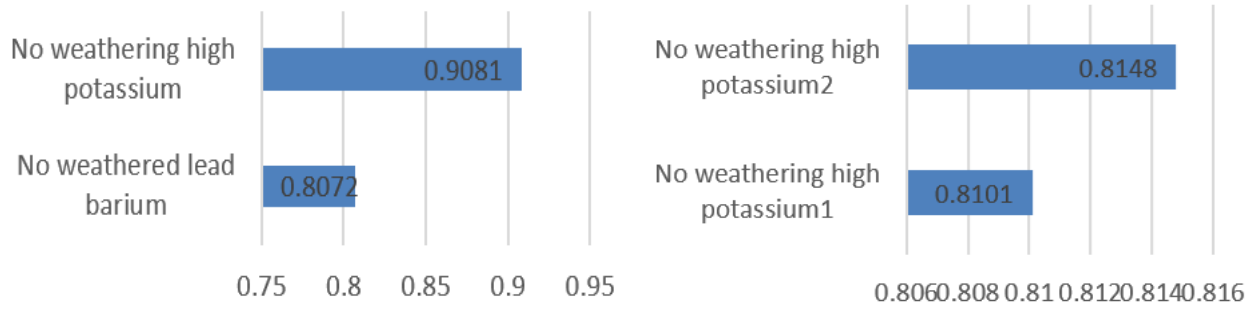


**Figure 6.** Grey Correlation Degree of A1 with Major Categories and Subcategories.

It can be seen that the unknown cultural relic A1 should belong to the unweathered high-potassium category; at this time, further considering the classification of subcategories, the chemical composition content of A1 is still set as the parent sequence, and the unweathered high-potassium subcategories 1 and 2 are set as the sub-sequences respectively, and the grey correlation degree is calculated by repeating the above steps, as shown in the right of Figure 6. It can be seen that A1 belongs to unweathered high potassium subclass 2[8].

We do the perturbation to each kind of discrimination main chemical composition content, controls the data in-10%-10% scope fluctuation, obtains in this foundation discrimination result, when the data in-10%-10% scope perturbation, obtains the model result the rate of accuracy to achieve 100%, the proof this model has the good robustness and the stability. According to the classification and weathering, the model can be divided into four categories for correlation analysis. The classification results are as follows: the unweathered high-potassium glass is 4, 5, 1, 3, 13, 14, 16, 6, 8, 21, and the weathered glass is 7, 9, 10, 12, 22, 27;Lead-barium glass unweathered is 30, 46, 47, 25, 55, 50, 29, 44, 32, 35, 28, 33, 31, 49, 42, 23, 53, 37, 20, 24, Weathering is 2, 8, 11, 26, 34, 38, 52, 56, 57, 36, 51, 58, 19, 41, 50, 43, 49, 39, 54, 40, 48.

Because of the different research objects, we can use the correlation coefficient method to solve the problem. Through MATLAB to solve the correlation coefficient [9] between the various components, we get the following conclusions.

The correlation coefficient between each component in the four types of potassium-weathered glass, high-potassium glass without weathering glass, lead-barium weathered glass and lead-barium glass without weathering glass, wherein for the correlation coefficient R, $R > 0.7$ represents the high correlation between the two, and the result is that the difference of different types of chemical components depends on the difference of the content of chemical components with high correlation in different types of glass. The effect of potassium oxide is more significant in the high potassium type glass, while the effect of lead oxide and barium oxide is more significant in the lead-barium type glass [10].

## 5. Conclusion

In this paper, the chi-square test was used to analyze the influencing factors of glass weathering, and then the mathematical statistics of the chemical composition of different glasses were carried out, combined with the visual processing of decoration, colour and other characteristics, the classification rules of two types of glasses were obtained, and then the system clustering was carried out, and finally, the chemical composition of ancient glass was analyzed and judged by the grey correlation analysis model. The entropy weight method is used for combination weighting and the component distribution weighting value is used for prediction, which is suitable for the combination prediction model with more zeros, and the classification algorithm in supervised learning can be used to solve the classification problem with a large amount of data, which can achieve better fitting and classification results for the problem with a large amount of data, and improve the accuracy of cross-validation.

## References

[1] Dai Huajuan. Research on combination forecasting model and its application [D]. Central South University, 2007.

[2] Zheng Yingxin. Application of clustering analysis based on elbow rule in data mining to route optimization design of primary and secondary school students [J]. Electronic World, 2017 (9): 1.

[3] Fang Xiangzhong. Chi-square distribution and chi-square test. China Statistics, 2022 (05): 29-31.

[4] Yang Chaoran, Chang Guangping. Bootstrap method for normality test based on L _ 2 Wasserstein distance [J]. Applied Probability and Statistics, 2022, 38 (02): 179-194.

[5] Dai Huajuan. Research on combination forecasting model and its application [D]. Central South University, 2007.

[6] Cao Siming, Wu Yi, Cao Kai, Chen Meng. Evaluation of integrated energy system operation service based on entropy weight and analytic hierarchy process [J]. Chinese Science and Technology Bulletin, 2021, 37 (12): 56-60. DOI: 10. 13774/J. CN ki. Kjtb. 2021.12.011.

[7] Hu Leifang. Five common hierarchical cluster analysis methods and their comparison [J]. Zhejiang Statistics, 2007 (04): 11-13.

[8] Zheng Yingxin. Application of clustering analysis based on elbow rule in data mining to route optimization design of primary and secondary school students [J]. Electronic World, 2017 (9): 1.

[9] Cao Mingxia. Study on the Model of Grey Relational Analysis and Its Application. Nanjing University of Aeronautics and Astronautics, 2007.

[10] Kong Yinghui, Che, yuan Jinsha, Jing Jing, Liu Yunfeng. Power quality disturbance recognition method based on wavelet decomposition and decision tree algorithm in data mining. Power Grid Technology, 2007 (23): 78-82.