Component analysis of ancient glass products based on hierarchical analysis clustering algorithm

Yusi Feng*, Hongkai Chen, Xin Zheng

Taiyuan University of Technology, Taiyuan, China

*Corresponding author: tyut_fsy@163.com

Abstract: First, the attachment is pre-processed, and after the abnormal data are removed side by side, the bar chart is drawn to preliminarily analyze the relationship that lead-barium glass is easier to weathering than high potassium glass. Then the chi-square test is carried out to find that whether the weathering of the glass cultural relics surface is related to the glass type of the cultural relics, but there is no obvious relationship with the decoration and color of the cultural relics. Secondly, the statistical model of one-way ANOVA was established, and the difference analysis of various chemical components was conducted before and after the two types of glass weathering. The chemical components that passed the difference test had significant statistical rules. Finally, multiple linear regression equations are constructed to predict the chemical composition content before weathering. The 14 chemical components of glass were regarded as 14 indicators, and the principal component analysis method was used to calculate the principal component contribution rate and the cumulative contribution rate, and then the two principal components were determined. Then the principal component analysis was used to cluster the indicators, and the hierarchical clustering algorithm was used to generate lineage maps. According to the elbow principle, the number of subcategories of high potassium glass is 3, and the number of subcategories of lead-barium glass is 4, so as to divide the chemical composition of glass. The cluster group scatter plot is then plotted and the subclass results are highly reasonable. Then a stepwise regression model is established to analyze the classification sensitivity and obtain several indicators with high sensitivity, which can be used for more targeted protection and restoration of unearthed cultural relics.

Keywords: Chi-square Test,The Multiple Linear Regression Equation,Hierarchical Clustering Algorithm.

1. Introduction

We were asked to select suitable chemical components for each glass category to divide it into the subclass, and to analyze the plausibility and sensitivity of the classification results[1-3]. The idea of this paper is as follows: First, the 14 chemical components of glass are regarded as 14 indicators, and the main component analysis method is used to accurately get the main indicators to distinguish between the two types of glass[4]. Then the principal component analysis is used to cluster the indicators, and the lineage graph generated by the hierarchical clustering algorithm is used to determine the optimal number of clusters K according to the elbow principle, so as to subclass the chemical composition of the glass[5]. Then we plot the cluster grouping scatter plot to analyze the rationality of the classification results, observe the change of the cluster scatter map by changing the chemical composition index of the cultural relic samples, and analyze the sensitivity of the results[6-7].

2. Model building and solution

The analysis of the data in form 1 shows that the color information of cultural relics No.19,40,48 and 58 is missing. In order to more accurately analyze the data of surface weathering degree and glass type, decoration and color, the information of these five groups of cultural relics was removed and then the subsequent correlation analysis was conducted[8].

In order to make a preliminary qualitative analysis of the correlation between each information, the proportion of weathered glass relics of different types, patterns and colors is calculated by using the processed data and makes the column chart as follows Figure 1.

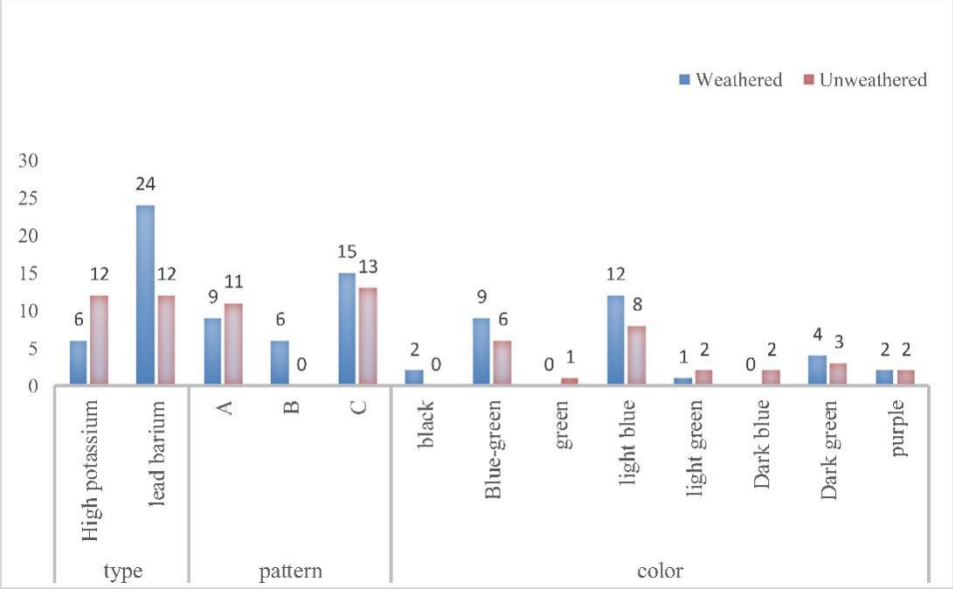

Figure 1. Comparison of weathering degree between lead barium glass and high potassium glass

According to Figure 1, lead-barium glass is easier to be weathered than high potassium glass, class B decoration is up to 100% weathered, much higher than the other two types, while Class C decoration is more weathered than Class A decoration. For glass color, green and dark blue glass relics are not weathered, light green cultural relics are less weathered, while all black glass relics are weathered, and cultural relics of other colors are about 50% weathered. However, only the number and proportion of weathering and unweathering of different types, patterns and colors can only reflect the difference in the value of each information, and a statistically significant relationship can not be obtained[9-10]. Therefore, further correlation analysis is needed based on the preliminary association analysis.

Since the data for surface differentiation, glass type, ornamentation, and color in Form 1 were all categorical data, the chi-square test was used to calculate the correlation between feature quantities and classification outcome quantities. According to the meaning of the question, it can be analyzed: the surface weathering situation is the characteristic amount, the glass type, decoration and color are the classification results, and the corresponding number of cultural relics is summarized in Tables 1,3 and 5. The pre-processed data were imported into Spass, the number of cultural relics corresponding to each classification result quantity was selected as the frequency variable, and the chi-square test results of the corresponding data of glass type, pattern and color were obtained as shown in Tables 2,4 and 6.

For surface weathering and glass type, the chi-square value is 5.400, compared with the chi-square test threshold table, 5% significance level (α =0.05), when the degree of freedom is 1, the threshold threshold value is 3.841. The null hypothesis of significant differences between variables was rejected because 5.400> 3.84103.2. That is, there is a significant correlation between the surface weathering of the glass artifacts and the glass types.

For surface weathering and ornamentation, the chi-square value is 5.747, compared to the chi-square test critical value table, 5% significance level (α =0.05), when the degree of freedom is 2, the chi-square value is 5.991. Since 5.747 <5.991, the null hypothesis that there was no significant difference between the variables was accepted. That is to say, is there a significant correlation between the surface weathering and ornamentation of glass relics.

For surface weathering and color, the chi-square value is 6.287, compared to the chi-square test threshold table, 5% significance level (α =0.05), when the degree of freedom is 7, the chi-square value is 14.067. Since 6.287 <14.067, the null hypothesis that there was no significant difference between

the variables was accepted. In other words, is there a significant correlation between the surface weathering of glass relics and the color.

In a word, whether the surface of glass cultural relics is weathered is related to the type of glass of cultural relics, but there is no obvious relationship with the decoration and color of cultural relics. The reasons for the influence of cultural relic glass type on surface weathering may be related to the chemical composition of different types of cultural relics and some specific chemical elements, so further analysis of the chemical composition content of cultural relic samples is needed.

The accumulation of the proportion of known components and the data between 85% and 105% are regarded as effective data, and the proportion of chemical composition of each sampling point is accumulated as shown in Figure 2. From the figure, the cultural relic sampling points 15 and 17 do not meet the requirements and should be abandoned. Then, the sum of each component content should be 100%, but due to the detection means and other reasons, the sum is not 100%, so each component is normalized, so that its component content sum is 100%.

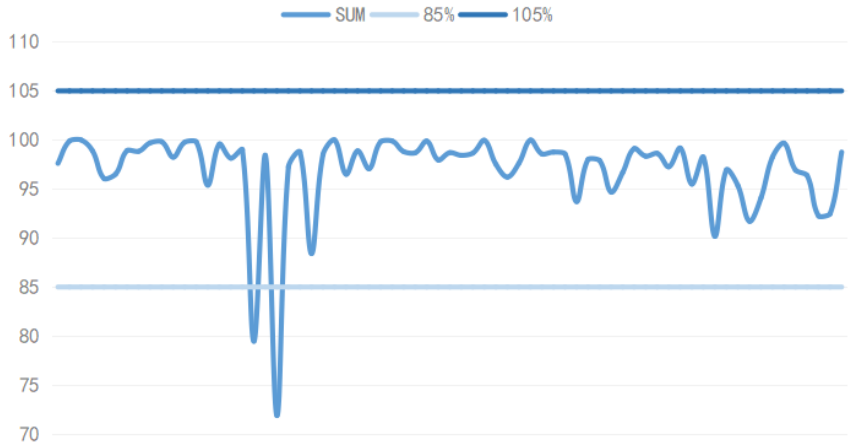

Figure 2. Accumulation of chemical composition proportion at each sampling point.

According to the type of glass, the data are divided into two categories: high-potassium glass and lead-barium glass. The average content of chemical composition before and after weathering of each type of glass is shown in Figure 3 and Figure 4. It is worth noting that although some cultural relics are weathered, the cultural relics sampling sites are unweathered points, and such points should be classified as unweathered points.

One-way analysis of variance method is only consider a factor A concerned about the influence of indicators, assumed in the process of the experiment, our task is to infer from the test results, factor A index has significant influence, namely when A take different levels index has significant difference, is equivalent to test whether the mean of several overall equal.

Let A take r levels, respectively recorded as A_1, A_2,..., A_r. The A in this question is whether the surface of the glass cultural relics is weathered. The null hypothesis is that the r levels of H_0: A will have no significant effect on the chemical composition content. According to the analysis, the factors affecting the chemical composition content of cultural relics include: the first type is an uncontrollable random factor, namely the precision of the detection method; the second type is a controllable factor, namely the difference between samples. The difference in detection results within the same sample is called "within-group difference", which is mainly caused by random error; the differences between different levels are called "difference between groups", caused by both sample difference and random error.

First, the error analysis of the test results, any one i(i=1、2、3.....m)the sample number j (j=1、2、3.....n) the secondary test results can be expressed as the formula:

$$x_{ij} = u_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (1)$$

A statistical model of the one-factor ANOVA can thus be written:

$$\begin{cases} x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \sum_{i=1}^{r} \alpha_i = 0 \\ \varepsilon_{ij} \sim N(0, \ \sigma^2), \ i = 1,...,r, j = 1,2,...n \end{cases} \quad (2)$$

One-way ANOVA was used to analyze the differences between the two chemical components of glass before and after weathering. Taking silica (SiO_2) in the sample point of high potassium glass cultural relics as an example, assuming that there is no significant difference in SiO_2 content before and after weathering of high potassium glass products, the ANOVA table is obtained with a p value of 2.5400e-06 <0.05, which indicates that there is a significant difference in SiO_2 content of high potassium glass products in the case of significant level α =0.05 (95% confidence). Moreover, according from Table 7, the content of SiO_2 after weathering is significantly higher than that before weathering, so it can be concluded that the content of SiO_2 in high potassium glass products is significantly higher after weathering than that before weathering.

The p value of silica (SiO_2), potassium oxide (K_2O), calcium oxide (CaO), magnesium oxide (MgO), alumina (Al_2O_3), iron oxide (Fe_2O_3) is <0.05, It shows that the above chemical components have significant differences before and after weathering, The average content of silica (SiO_2) is increased after weathering, Variance decreases; Lower mean values of potassium oxide (K_2O), calcium oxide (CaO), and alumina (Al_2O_3), The variance has all decreased; Lower average reduction of magnesium oxide (MgO), The variance is basically unchanged; The mean value of iron oxide (Fe_2O_3) content was decreased.

The title requires the analysis of the classification basis of high-potassium glass and lead-barium glass, and its essence is to analyze which kind of (or which kinds) of chemical composition has significant differences in the two glasses, so that it can be used as the basis for distinguishing between different kinds of glass. In order to facilitate the analysis of silica (SiO_2), sodium oxide (Na_2O) and other 14 chemical components in different cultural relics samples content differences.

In order to more accurately get the main indicators to distinguish between the two glass types, the method of principal component analysis is used to retain some of the most important features of the high-latitude data (with multiple indicators) by dimension reduction. In this way, we can not only find the main indicators to distinguish between the different variables, but also remove the noise and the unimportant indicators[2], so as to improve the speed of data processing.

The chemical composition of each sample after removing the invalid data was imported into the Matlab, with a total of 67 samples and 14 chemical composition indicators, resulting in a 6714, sample matrix x. First, the correlation coefficient matrix R of the matrix x is found by using the corrcoef() function in Matlab, and it is visually displayed as shown in Figure 3.

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
x1	1.000	0.092	0.317	-0.132	0.013	0.232	0.086	-0.227	-0.842	-0.617	-0.571	-0.661	0.109	-0.338
x2	0.092	1.000	0.040	-0.101	-0.037	0.115	-0.195	-0.060	-0.122	0.016	-0.325	-0.014	-0.083	-0.108
x3	0.317	0.040	1.000	0.632	0.231	0.369	0.363	0.090	-0.545	-0.382	-0.189	-0.365	0.158	-0.079
x4	-0.132	-0.101	0.632	1.000	0.322	0.335	0.476	0.084	-0.118	-0.244	0.199	-0.096	-0.074	0.029
x5	0.013	-0.037	0.231	0.322	1.000	0.528	0.415	-0.208	-0.091	-0.355	0.251	0.042	0.240	-0.223
x6	0.232	0.115	0.369	0.335	0.528	1.000	0.424	-0.174	-0.391	-0.338	-0.053	-0.192	0.161	-0.193
x7	0.086	-0.195	0.363	0.476	0.415	0.424	1.000	-0.025	-0.227	-0.301	0.115	-0.165	-0.014	-0.151
x8	-0.227	-0.060	0.090	0.084	-0.208	-0.174	-0.025	1.000	-0.093	0.549	0.083	0.131	-0.190	0.221
x9	-0.842	-0.122	-0.545	-0.118	-0.091	-0.391	-0.227	-0.093	1.000	0.328	0.420	0.643	-0.119	0.049
x10	-0.617	0.016	-0.382	-0.244	-0.355	-0.338	-0.301	0.549	0.328	1.000	0.140	0.405	-0.091	0.595
x11	-0.571	-0.325	-0.189	0.199	0.251	-0.053	0.115	0.083	0.420	0.140	1.000	0.385	-0.059	0.196
x12	-0.661	-0.014	-0.365	-0.096	0.042	-0.192	-0.165	0.131	0.643	0.405	0.385	1.000	-0.052	0.216
x13	0.109	-0.083	0.158	-0.074	0.240	0.161	-0.014	-0.190	-0.119	-0.091	-0.059	-0.052	1.000	-0.052
x14	-0.338	-0.108	-0.079	0.029	-0.223	-0.193	-0.151	0.221	0.049	0.595	0.196	0.216	-0.052	1.000

Figure 3. Visualization results

According to the known information in the question, all the cultural relics samples can be divided into high-potassium glass and lead-barium glass. In order to further refine the classification, it is

necessary to classify the two categories according to the chemical composition of cultural relics, which is conducive to the more targeted protection and restoration of [6] for different cultural relics.

Data from high potassium glass and lead barium glass from Form 2 were imported into Spass separately, the systematic cluster was selected, and the square Euclidean distance was used as the distance between the two data points, and the cluster lineage map was generated. However, the optimal number of clusters is still unable to determine the cluster K based solely on the lineage map, so the appropriate number of clusters needs to be determined before further classification using the lineage map.

The line diagram of the aggregation coefficient is drawn from the hierarchical clustering as shown in Figure 4 and Figure 5.

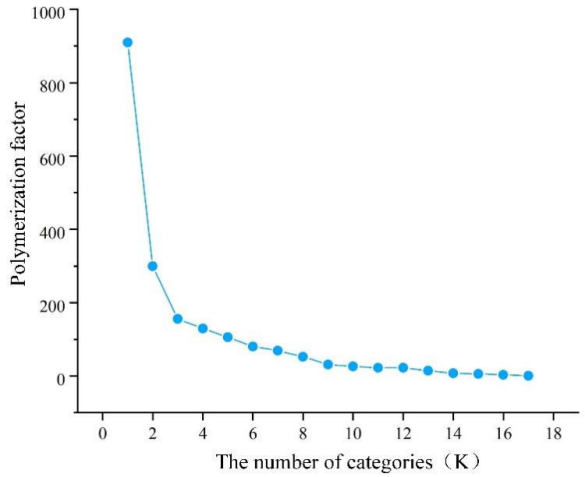

Figure 4. High potassium glass polymerization coefficient is broken line

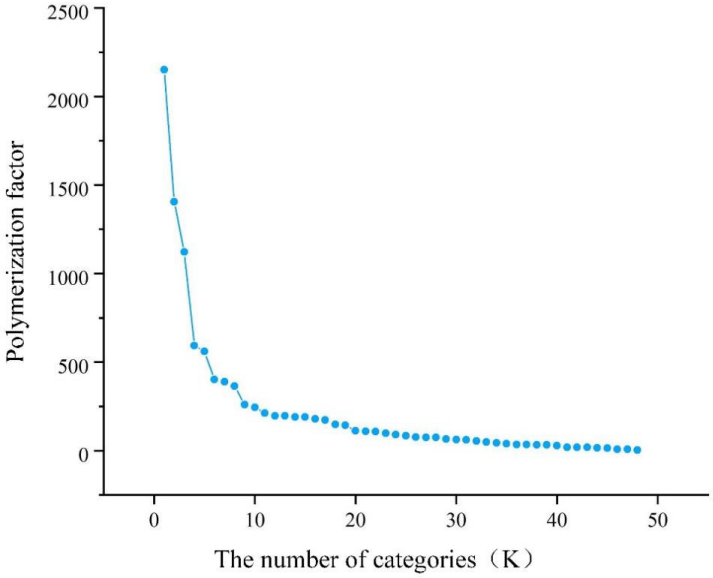

Figure 5. Lead-barium glass polymerization coefficient broken line

It is obviously found that when the K value is taken from 1-3, the aggregation coefficient changes the most, and the change speed decreases significantly after exceeding 3. Therefore, the elbow can be considered as K=3, so the number of subclasses of high potassium glass is set as 3. Similarly, it is obvious that the elbow of the polymerization coefficient fold of lead barium glass is K=4, and the number of subclasses of lead barium glass is 4.

Stepwise regression is an independent variable selection method for linear regression models, and the basic idea is to introduce the variable step by step with the condition that its partial regression squared sum experience is considered significant. At the same time, after each new variable was

introduced, the old variables selected in the regression model were tested one by one, and the insignificant variables were removed to ensure that each variable in the obtained subset of the independent variables was significant. At this time, all the variables in the regression model were significant [5] to the dependent variable.

The regression coefficients obtained after a stepwise linear regression were obtained, in which the independent variable x_9 was negatively correlated with the dependent variable, and all the remaining independent variables were all positively correlated with the dependent variable. Because $\alpha_1 > \alpha_{14} > |\alpha_9| > \alpha_2$, x1 has the strongest sensitivity among x_9, x_1, x_{14}, x_2 and x_2. That is, the sensitivity of silica (SiO_2) has the most influence on the classification situation.

3. Conclusion

Chi-square test and one-way ANOVA were used to make the obtained correlation results more reliable. Multivariate regression model can fully cover the influence of all factors on chemical composition and consider more comprehensively. For the multiple linear regression model, the prediction results may differ greatly from the actual results because of ignoring the interaction effects and the non-linear causality. Although Spearman's correlation coefficient model has a wide range of applications, the accuracy decreases compared with Pearson's correlation coefficient model, which does not well represent the correlation between the two variables. Through the composition analysis, identification and prediction of glass products, the important influencing factors of ancient glass weathering can be found, which can be used for more targeted protection and restoration of unearthed cultural relics. The obtained prediction model can be used as a basis for the investigation of ancient glass age.

References

[1] Liu Dong. Prediction and analysis of water consumption in Chifeng city based on multiple linear regression model [J]. Journal of Chifeng College (Natural Science Edition), 2022,38(07):11-16.DOI:10.13398/j.cnki.issn1673-260x. 2022.07. 026.

[2] Wang Di. —— takes the city of Shanghai as an example [J]. Investment and Entrepreneurship, 2022,33 (14): 58-60.

[3] Li Meng, Bao Lei, Hu Yi, Cheng Song, Hu Xiaobo, Gao Ying. Implementation of a random number online detection method based on the chi-square test [J]. Microelectronics, 2022,52(03):388-392.DOI:10.13911/j.cnki.1004-3365.210329.

[4] Conditions for the application of the Chi-square test [J]. Journal of Clinical Hepatobiliary Diseases, 2022,38 (06): 1292.

[5] Fang Xiangzhong. The Chi-square distribution was compared with the chi-square test [J]. China Statistics, 2022 (05): 29-31.

[6] season Jiang Shuai, Pei Songwen. Intelligent hierarchical clustering algorithm study for heterogeneous gene data [J / OL]. Small microcomputer system: 1-7 [2022-09-21]. http://kns.cnki.net/kcms/detail/21.1106.TP.20210706.1114.018.html

[7] Kai Yu. Quantum Hierarchical Clustering and its Related Subalgorithms Research [D]. Fujian Normal University, 2021.DOI:10.27019/d.cnki.gfjsu. 2021.001183.

[8] Tian Qingyun. Hierarchical clustering algorithm study based on density peaks [D]. Henan University of Economics and Law, 2021.DOI:10.27113/d.cnki.ghncc. 2021.000265.

[9] Kemp, V. , et al. "LA-ICP-MS analysis of Late Bronze Age blue glass beads from Gurob, Egypt." Archaeometry Pt.1(2020):62.

[10] Smith, G. L. . "Sensitivity Analysis of Kinetic Rate-Law Parameters Used to Simulate Long-Term Weathering of ILAW Glass Erratum." (2016).