Glass Classification and Composition Study based on systematic clustering

Jiadong Xie*, Yixun Li, Bochen Huang, Yiyi Chen, Yimeng He

Beijing Institute of Technology, Zhuhai Campus, Zhuhai, China, 519088

*Corresponding author: jiadongxie23@gmail.com

Abstract. To obtain the influence of the production of raw materials and buried environment changes on their composition, this study analyzes a batch of ancient glass products in China. In the history of the Silk Road, glass was an important trade item. The main raw material of ancient glass is quartz sand, with a high melting point, so it is necessary to add solvents and stabilizers for refining. Different solvents contain different main components. The results show that the burial environment change of ancient glass products can affect their color, ornamentation, and composition ratio. The addition of different cosolvents to the production process can lead to different compositions of ancient glass products. The analysis of ancient glass products plays an important role in providing a better understanding of the Silk Road trade and technology exchange.

Keywords: Hierarchical Clustering, Chi-squared Test, Component Analysis.

1. Introduction

In the history of the Silk Road, glass was the most important item of the early merchants' trade[1]. The bead ornaments imported from West Asia and Egypt got the reference and technical absorption and obtained the ancient glass with similar appearance but different compositions. Because the main raw material of glass is quartz sand, the melting point is very high, so in ancient times, the ancient people need to add such as plant ash, calcium oxide, and other auxiliary solvents and limestone as a stabilizer for refining[2]. Of course, the added cosolvent is different, its main ingredients are also different. Because of the unearthed ancient glass burial environment has changed greatly, in these changes, including the color, decoration and composition, and proportion of the glass changes. Therefore, the relevant data from a batch of ancient Chinese glass products are analyzed now.

2. Establishment and solution of the glass classification and composition research model

2.1. Model building

System clustering is also known as hierarchical clustering[3]. Compared with K-means clustering and KNN clustering methods with more frequency, this method avoids the problem of subjectively determining the number of clusters k during the initial clustering. The specific flow chart is shown in Figure 1:

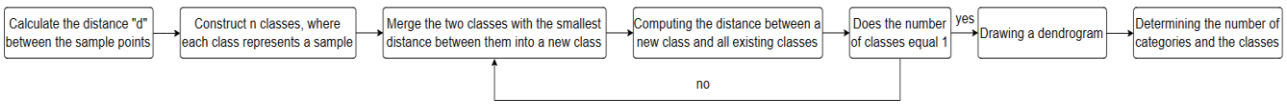


Figure 1. Flow-chart of the system clustering model

Where, for the distance between samples, using the Euclidean distance[4] is calculated by the following formula:

$$d^2(\vec{x}_i, \vec{x}_j) = \sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2 \quad (1)$$

On the type x_{ik} is the value of the k-th indicator for the i-th glass artifact sample, $d^2(\vec{x}_i, \vec{x}_j)$ is the distance between the i-th glass artifact sample and the j-th glass artifact sample. For the distance between classes, the inter-group distance method is calculated using the following formula:

$$D(G_p, G_q) = \frac{d_1+d_2+\cdots+d_n}{n} \tag{2}$$

Where $d_i(i = 1 \ldots n)$ is the distance between the samples in the G_p and G_q categories and n is the total number of samples between the two categories.

2.1.1 Chi-square test

The chi-square test is a method of hypothesis testing[5]. If the test statistic is obtained, the greater the value of the test statistic, the greater the degree of deviation between variables. First, propose the null hypothesis H_0: there is no significant difference between the two variables and the alternative hypothesis H_1: there is a significant difference between the two variables.

Next, construct the test statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(A-T)^2}{T} \tag{3}$$

Where, A is the observed value of the variable, and T is the theoretical value of the variable. After finding the statistic, comparisons were made at the 95% confidence level. If the statistic is greater than the corresponding F value obtained by checking the table, the original hypothesis is rejected, indicating that there is a correlation between the two variables; if the statistic is less than the corresponding F value obtained by checking the table, then the original hypothesis cannot be rejected, indicating that there is no correlation between the two variables.

2.1.2 Variance Analysis

Analysis of variance, also known as "variation number analysis" or "F test"[6], this method is mainly used for the significance test of the difference between two or more samples. First of all, the original hypothesis H_0 is proposed: there is no significant difference in the proportion of chemical components before and after the differentiation of glass relics and the alternative hypothesis H_1: there is a significant difference in the proportion of chemical components before and after the differentiation of glass relics.

Next, construct the F test statistic:

$$F = \frac{MSR}{MSE} = \frac{SSR/df_B}{SSE/df_E} = \frac{SSR/(k-1)}{SSE/(n-k)} \tag{4}$$

In the formula, SSR represents the sum of squares between groups and SST represents the overall variance. SSE Is the sum of squares of the residual differences within the group and $SSR = SST - SSE$.

Among

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}\right)^2$$
$$SSE = \sum_{j=1}^{k} SSE_j = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x_j}\right)^2 \tag{5}$$

x_{ij} is the proportion of components corresponding to each sample under each index、$\bar{x_J}$ represents the mean of the proportion of individual indicators in individual samples、$\bar{x}$ represents the mean of the chemical composition content ratio of all samples、k is the number of indicators、n is the number of samples. df_R is SSR the degree of freedom, df_E is SSE the degree of freedom. Then judge the value of the statistic and the corresponding F distribution table, and conclude.

2.2. Model solution

After determining the distance formula between samples and the distance formula between classes, the clustering results such as the pedigree diagram Figure 2:

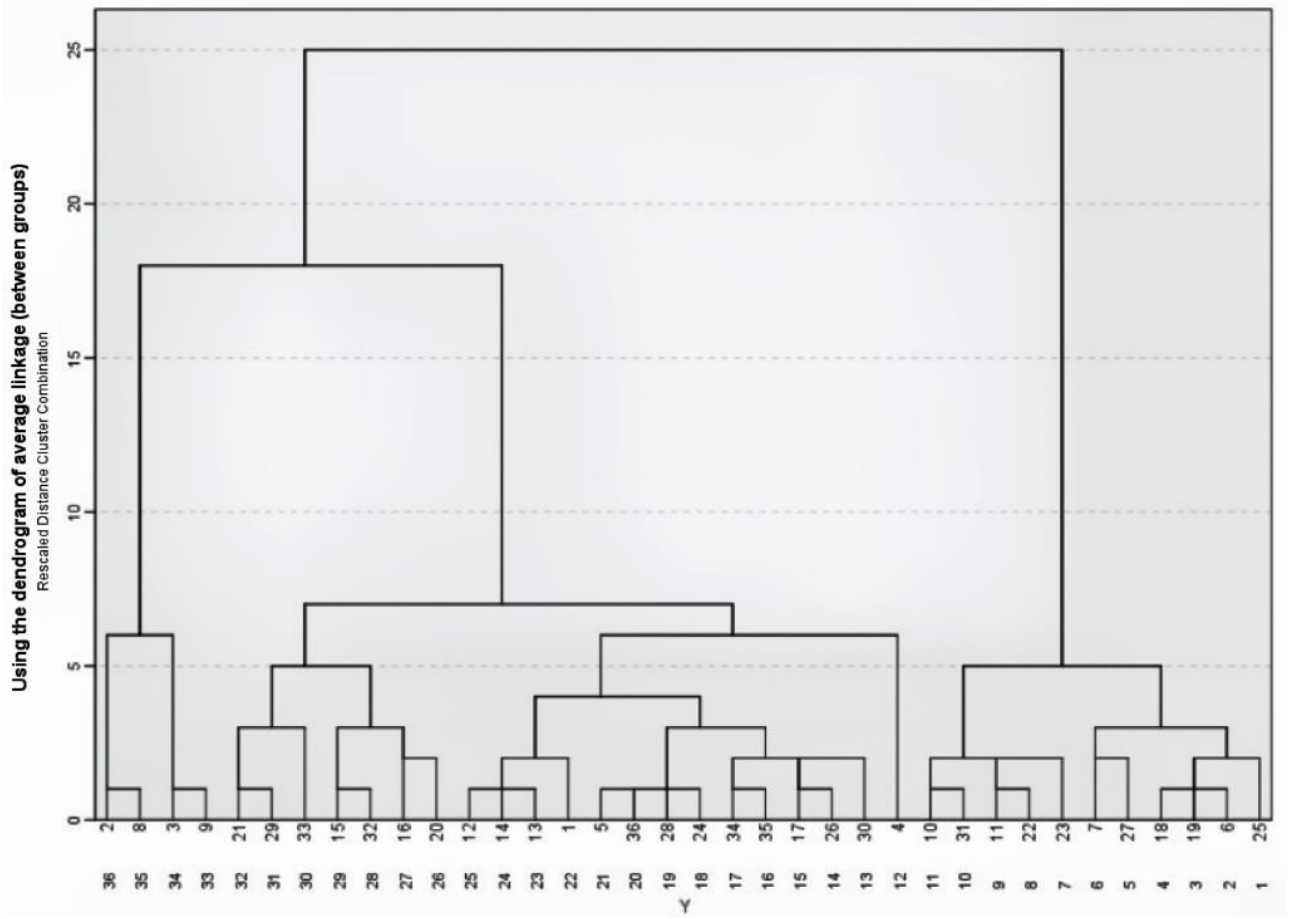


Figure 2. Flow-chart of the system clustering model

According to the number of color categories in the data, we determined cluster number K as 8. Combined with the pedigree diagram and the idea of the voting device minority, the missing color data of glass relics numbered 19,40,48, and 58 were classified as dark green, light blue, light blue, and dark green respectively, and the results were filled in Table 1. After filling in the missing color data, a chi-square test is then conducted between weathering and color, glass type, and ornamentation. The result of the chi-square test is shown in Table 1:

Table 1. Table of the Chi-square test for each factor

Variable	P-Value
type	0.0009***
sculpture	0.084*
dyestuff	0.423

Available from the above table, the P -values of the chi-square test for type, pattern, color, and surface weathering are 0.009,0.084,0.423, respectively. It can be concluded that at the 95% confidence level, only the test P value of type and surface weathering is less than 0.05, and the rest are greater than 0.05. That is, only the chi-square test of type and surface weathering refuses the null hypothesis, and the rest cannot reject the null hypothesis. This indicates that there were significant differences in the surface weathering status of different glass relic types, and the remaining variables were not significantly different from the surface weathering status.

To explore the statistical law of the chemical composition content on the surface of cultural relics samples, variance analysis was first conducted to test whether the chemical composition content of the surface of cultural relics samples was significantly different, to ensure the feasibility and necessity of subsequent analysis and research. After inspection, the F statistics corresponding to the chemical content of the two types of glass are shown in Table 2 below:

Table 2. Table of ANOVA results

Chemical composition	F distribution value (high K)	F distribution value(lead and barium)
SiO_2	50.33	86.17
Na_2O	1.7	9.981
K_2O	29.06	1.172
CaO	11.98	10.31
MgO	9.066	0.002
Al_2O_3	19.3	3.108
Fe_2O_3	5.813	0.309
CuO	1.464	1.438
PbO	2.842	49.48
BaO	2.16	1.395
P_2O_5	3.531	19.89
SrO	4.312	4.232
SnO_2	0.485	0.127
SO_2	1.747	1.836

Finding the F distribution table at the significance level of α=0.01, the F value of high potassium glass is 3.05 and the F value corresponding to lead barium glass is 2.82. Based on the results of the two tests, the conclusion is drawn: if the statistic is greater than the corresponding F value, the original hypothesis is rejected, indicating that the chemical composition content ratio is significant before and after weathering; otherwise, the null hypothesis cannot be rejected, indicating that there is no significant difference between the chemical composition content ratio before and after weathering.

As can be seen from Table 2, except for sodium oxide, barium oxide, lead oxide, copper oxide, tin oxide, sodium oxide, and lead oxide, there is no significant difference in the proportion of components before and after weathering.

Therefore, according to the analysis of variance, only a small part of the chemical composition of the two types of glass is not significant before and after weathering at the significance level of α=0.01. It can be deduced that the significance level is more general and required than the significance level of α=0.05, the insignificant chemical composition can be almost ignored. Therefore, it is feasible and necessary to study the statistical law of chemical composition content before and after weathering.

3. Descriptive statistics as well as trend, and distribution analysis

To analyze the statistical law of the weathering chemical content on the surface of cultural relics samples, and more effectively display the chemical content law between different glasses[7], glass cultural relics are divided into four types: high potassium unweathered glass, high potassium weathered glass, lead barium unweathered glass and lead barium weathered glass. Descriptive statistics and trend and distribution analysis were conducted for the four categories.

The proportion of sodium oxide, tin oxide, and sulfur dioxide is almost unchanged before and after the weathering of glass relics, while the proportion of lead oxide in most glass relics before weathering is very small, but the proportion of lead oxide in most glass relics has increased significantly after weathering. On the contrary, the proportion of silica decreased significantly after weathering of glass relics.

Descriptive statistics were performed for the four categories of glasses. Descriptive statistics[8] refers to the description of the overall data through tables, appropriate figures, and various numerical values, which has achieved the purpose of summarizing the characteristics of the data. Here, the mean, standard deviation, minimum value, quartile, quartile, and maximum value are selected to make descriptive statistics and analyze the data characteristics. Descriptive statistics for HK glass before and after weathering are shown in Table 3.

Table 3. Statistical table described before and after weathering of high potassium glass

Chemical composition	Mean value	Standard deviation	Minimum	Lower quartile	The dichotomous	Upper quartile	Maximal value
SiO_2	67.98(93.36)	8.76(1.73)	59.01(92.35)	61.68(92.65)	65.53(93.51)	71.17(94.84)	87.05(96.77)
Na_2O	0.70(0.00)	1.29(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.53(0.00)	3.38(0.00)
K_2O	9.33(0.54)	3.92(0.45)	0.00(0.00)	7.60(0.15)	9.83(0.67)	12.30(0.88)	14.52(1.01)
CaO	5.33(0.87)	3.09(0.49)	0.00(0.21)	4.04(0.65)	6.10(0.83)	7.57(1.04)	8.70(1.66)
MgO	1.08(0.20)	0.68(0.31)	0.00(0.00)	0.63(0.00)	1.17(0.00)	1.60(0.41)	1.98(0.64)
Al_2O_3	6.62(1.93)	2.49(0.96)	3.05(0.81)	5.14(1.36)	6.19(1.72)	7.93(2.38)	11.15(3.50)
Fe_2O_3	1.93(0.27)	1.67(0.07)	0.00(0.17)	0.48(0.22)	2.11(0.28)	2.45(0.31)	6.04(0.35)
CuO	2.45(1.56)	1.66(0.93)	0.00(0.55)	1.00(1.02)	2.35(1.55)	3.43(3.24)	5.09(3.24)
PbO	0.41(0.00)	0.59(0.00)	0.00(0.00)	0.00(0.00)	0.16(0.00)	0.51(0.00)	1.62(0.00)
BaO	0.60(0.00)	0.98(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.07(0.00)	2.86(0.00)
P_2O_5	1.40(0.28)	1.43(0.21)	0.00(0.00)	0.69(0.17)	1.02(0.28)	1.29(0.61)	4.50(0.61)
SrO	0.04(0.00)	0.05(0)	0.00(0.00)	0.00(0.00)	0.02(0.00)	0.08(0.00)	0.12(0.00)
SnO_2	0.20(0.00)	0.68(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	2.36(0.00)
SO_2	0.10(0.00)	0.19(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.09(0.00)	0.47(0.00)

The brackets in the table above are the descriptive statistics of chemical content after weathering. It can be seen from the descriptive statistics of the variance before and after weathering of high potassium glass that the standard deviation before weathering is greater than the standard deviation after weathering, indicating that the fluctuation degree of the chemical composition of high potassium glass is small after weathering; weathering is a change process of large exchange of internal elements and environmental elements[9]. To reflect the difference in chemical composition before and after weathering, we introduce the difference coefficient R[10]. If the differential coefficient is a positive value, that is, the overall distribution moves to the right, that is, the overall data is increasing and vice versa. The calculation formula is as follows:

$$R_i = Q_{ki} - H_{ki} \tag{6}$$

In the above formula,$R_i, i = 1, \ldots, 14$is the difference coefficient before and after weathering of 14 chemical components, Q_{ki} is the upper quartile of the i-th chemical content of high potassium glass after weathering (75%), and H_{ki} is the upper quartile of the i-th chemical content after weathering (75%). The silica chemical content difference coefficient is positive and the value is larger, namely high potassium glass weathering silica chemical content overall data have obvious growth, potassium oxide, calcium oxide, aluminum oxide content difference coefficient are negative, namely the three chemical content after weathering appeared obvious decline trend, other chemical composition has no obvious change.

Table 4. Statistics described before and after weathering of lead and barium glass

Chemical composition	Mean value	Standard deviation	Minimum	Lower quartile	The dichotomous	Upper quartile	Maximal value
SiO_2	53.44(33.61)	14.59(17.32)	31.94(3.72)	37.36(21.05)	55.21(29.40)	65.91(50.77)	75.51(68.08)
Na_2O	0.77(0.95)	1.54(1.92)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(1.00)	4.66(7.92)
K_2O	0.26(0.14)	0.40(0.21)	0.00(0.00)	0.00(0.00)	0.15(0.00)	0.25(0.27)	1.41(1.05)
CaO	1.23(2.35)	1.46(1.61)	0.00(0.00)	0.46(1.03)	0.84(2.21)	1.13(3.19)	4.49(6.40)
MgO	0.49(0.70)	0.55(0.66)	0.00(0.00)	0.00(0.00)	0.51(0.71)	0.89(1.71)	1.67(2.73)
Al_2O_3	3.19(3.84)	1.39(3.41)	1.44(0.45)	2.35(1.62)	3.06(3.01)	4.35(4.84)	5.45(14.84)
Fe_2O_3	0.93(0.56)	1.45(0.69)	0.00(0.00)	0.00(0.00)	0.00(0.31)	1.51(0.90)	4.59(2.74)
CuO	1.56(2.00)	2.49(2.49)	0.00(0.00)	0.16(0.70)	0.53(1.00)	0.86(2.97)	8.46(10.57)
PbO	23.59(36.87)	9.09(15.16)	9.30(12.31)	16.55(24.8)	22.05(39.8)	29.14(46.77)	39.22(70.21)
BaO	10.50(10.49)	6.95(8.87)	3.42(0.00)	5.68(5.91)	10.06(8.15)	10.35(11.13)	26.23(35.45)
P_2O_5	0.90(4.16)	1.57(4.15)	0.00(0.00)	0.13(0.30)	0.20(3.35)	1.41(7.49)	5.75(14.13)
SrO	0.30(0.37)	0.31(0.25)	0.00(0.00)	0.00(0.22)	0.30(0.34)	0.43(0.47)	0.91(1.12)
SnO_2	0.06(0.06)	0.16(0.23)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.44(1.31)
SO_2	0.28(0.99)	1.02(3.61)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	3.66(15.95)

According to the statistical table values before and after the weathering of lead barium glass, as shown in Table 4, it can be found from the mean value that the silica content of glass relics decreased significantly, while the content of lead oxide and phosphorus pentoxide on the surface of glass relics increased significantly; the greater the standard deviation, the greater the fluctuation of the data. From the table, the standard deviation of silica, alumina, lead oxide, and phosphorus pentoxide all increased significantly after weathering, indicating that the weathering process increased the volatility of these chemical composition content data. From the perspective of difference coefficient, the difference coefficient of difference in chemical content of lead oxide and phosphorus pentoxide is large, that is, the overall data of weathered chemical content of lead oxide in lead barium glass has increased significantly.

4. Conclusions

The agglomerative clustering model, which does not require a priori specification of the number of clusters, is suitable for large datasets and exploratory data analysis. It can achieve different clustering results by using different distance measures and linkage methods. However, it is computationally expensive, sensitive to noise and outliers, and may not handle non-Euclidean distance measures well, making it difficult to determine an appropriate number of clusters. In this study, the agglomerative clustering model was used to analyze data related to ancient Chinese glass artifacts. The results showed a significant increase in the overall chemical content of silica after weathering of high-potassium glass, with a clear decrease in the content of potassium oxide, calcium oxide, and aluminum oxide. No significant changes were observed in other chemical components. In contrast, the silica content decreased significantly after weathering of lead-barium glass, while the content of lead oxide and diphosphorus pentoxide on the surface of the glass artifacts increased significantly.

References

[1] Zhang, X. Glassmaking in China: Origins and Evolution[M]. Routledge. 2020.

[2] Li, C., Chen, S., Li, J., & Li, H. A study of the chemical compositions and possible fluxes of ancient Chinese glass samples[J]. Journal of Cultural Heritage, 2017:24, 1-7.

[3] Luo, Y., Zhang, H., Zhou, Y., & Sun, T. Analysis of the classification of ancient glasses based on hierarchical clustering algorithm[J]. IEEE Access,2020:8, 27614-27622.

[4] Zaki, M. J., & Meira Jr, W. Data mining and analysis: fundamental concepts and algorithms[M]. Cambridge University Press. 2014.

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

[6] Zhang, J., Yu, Z., & Cao, X. Variance analysis of the chemical composition of ancient glass in the Western Regions of China[N]. Journal of Cultural Heritage, 2018:32, 97-103.

[7] Liu, S., Wang, S., Zhao, Y., & Feng, J. Chemical compositions and corrosion of ancient Chinese glasses excavated from the Imperial Palace, Beijing[J]. Journal of Non-Crystalline Solids, 2019:521, 119507.

[8] Schiavon, N., Kockelmann, W., & Ciliberto, E. An investigation on the effects of weathering on ancient glasses[J]. Journal of Cultural Heritage, 2019:36, 98-105.

[9] Trochim, W. M. K. Descriptive statistics[M]. Research Methods Knowledge Base. 2021.

[10] Duan, J., Chen, Y., & Wang, Y. Optimal design of differential evolution algorithm based on coefficient of variation for optimization problems[J]. IEEE Access, 2022:10, 3257-3265.