Glass Classification Method Based on K-Means++ Algorithm

Yuhao Chen *, Kun Ji

School of Automation, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210000

2109737597@qq.com

* Corresponding Author Email: 2109737597@qq.com

Abstract. Ancient glass is highly susceptible to weathering by the influence of the buried environment, resulting in changes in the proportion of its components. In order to select the appropriate chemical components for each category and divide them into subclasses, and give specific classification methods and classification results, this paper considers first using the K-means++ algorithm for clustering, and calculating the aggregation coefficient of each k value to determine the number of clusters. The three-level nomenclature is used to name each type according to the amount of chemical component content contained in each category, lists its division basis and the names of each type of high potassium glass and lead barium glass, and then conducts rationality analysis, uses the contour coefficient to evaluate the quality of the cluster, and concludes that the number of clusters of high potassium glass and lead barium glass is 3 and 5 respectively. Finally, sensitivity analysis is performed on the aggregation coefficients.

Keywords: K-means++, glass classification, profile coefficients.

1. Introduction

The Silk Road was a channel for cultural exchanges between China and the West in ancient times, of which glass was valuable material evidence of early trade. Early glass in West Asia and Egypt is often made into bead-shaped jewelry into China, China's ancient glass absorbs its technology after local materials production, so and foreign glass products similar in appearance, but the chemical composition is different [1-3].

The main raw material of glass is quartz sand, and the main chemical component is silica (SiO2). Due to the high melting point of pure quartz sand, in order to reduce the melting temperature, flux needs to be added during refining. Commonly used fluxes in ancient times include grass ash, natural bubble alkali, saltpeter and lead ore, etc [4-5], and limestone is added as a stabilizer, and limestone is converted into calcium oxide (CaO) after calcination. Different fluxes are added, and their main chemical composition is also different. For example, lead barium glass in the firing process added lead ore as a flux, its lead oxide (PbO), barium oxide (BaO) content is higher, usually considered to be China's own invented glass varieties [6-7], Chu culture glass is based on lead barium glass. Potassium glass is fired with substances with high potassium content such as grass ash as flux, which is mainly popular in Lingnan, Southeast Asia and India [8-9].

Ancient glass is highly susceptible to weathering by the influence of the buried environment. During weathering, internal elements undergo a large exchange with environmental elements, resulting in a change in the proportion of their components, which affects the correct judgment of their category [10-11].

2. Model building and solving

2.1 K-means++ algorithm

K-means algorithm is one of the main algorithms in clustering algorithms, MacQueen first proposed K-means algorithm in 1967, the basic principle of the algorithm is: predetermine the number of categories and the initial cluster center to cluster the sample into several classes, and obtain the final clustering result through iterative optimization. The K-means algorithm has a simple principle, fast convergence speed, and can efficiently cluster large sample sets. However, the K-means

algorithm is greatly affected by the initialization of cluster centers, and when the random initialization cluster centers are not selected properly, the clustering process is very time-consuming and the clustering results have a large error. In response to this problem, the K-means++ algorithm is proposed. The K-means++ algorithm is basically the same as the principle of the K-means algorithm, which is an optimization of the random initialization of cluster centers by the K-means algorithm. Compared with the K-means algorithm, the basic idea of the K-means++ algorithm is: first, randomly select a point from the data set as the initial centroid; Secondly, other centroids are selected by random seeding technology, and it is believed that the point farthest away from the currently existing centroid has a higher probability of becoming a new centroid. Finally, the data set is iteratively updated with clustering and centroids until convergence conditions are reached. Although K-means++ increases the amount of computation when initializing the cluster center, it significantly improves the computational efficiency and improves the error of clustering results throughout the clustering process. The algorithm flow is shown in Figure 1.

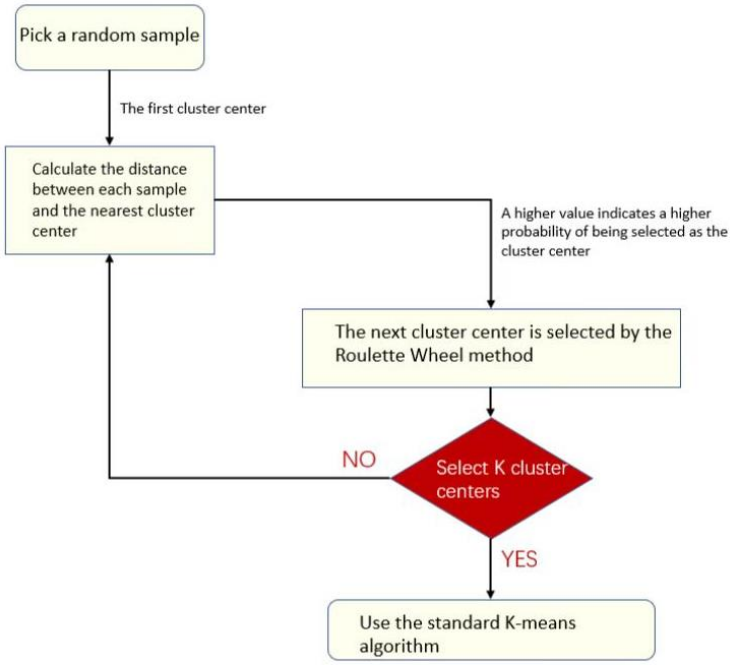


Figure 1. K-means++ algorithm flowchart

2.2 Clustering model based on K-means++ algorithm

In order to divide the subcategories of high potassium glass and lead barium glass, the high potassium glass and lead barium glass were first clustered separately, divided according to the clustering results, and the characteristics of each cluster were summarized according to each chemical composition.

1) Determine the center of gravity k-value

Since there are 19 groups and 40 groups of pretreated data in the high-potassium glass group and lead-barium glass group, this paper is to obtain the optimal results of clustering, according to the elbow law (Elbow Method), that is, the optimal number of clusters is roughly estimated by graphs, and the aggregation coefficient is calculated first.

Evaluation index of clustering effect SSE: SSE (Sum of Square Error, sum of squared errors), the smaller the SSE value, the closer the data points are to their centroid, and the better the clustering effect.

Suppose that a total of n samples are divided into K classes (K≤n-1, that is, there are at least two elements in one class) to represent the kth class (k=1, 2, ..., K), and the position of the center of gravity of the class is denoted as, then the aggregation coefficient of the kth class is:

$$\sum_{i \in C_k} | x_i - u_k |^2 \tag{1}$$

Define the total aggregation coefficient for all classes:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} | x_i - u_k |^2 \quad (2)$$

The data of high potassium glass and lead-barium glass were respectively subjected to K-means++ clustering of K centers of gravity (k=1, 2, ..., K), the number of cluster centers of gravity K and the aggregation coefficient J were recorded, and a line chart of the aggregation coefficients was drawn: the abscissa was the number of clustering barys K, and the ordinate was the aggregation coefficient J, as shown in Figure 2 and 3 below.

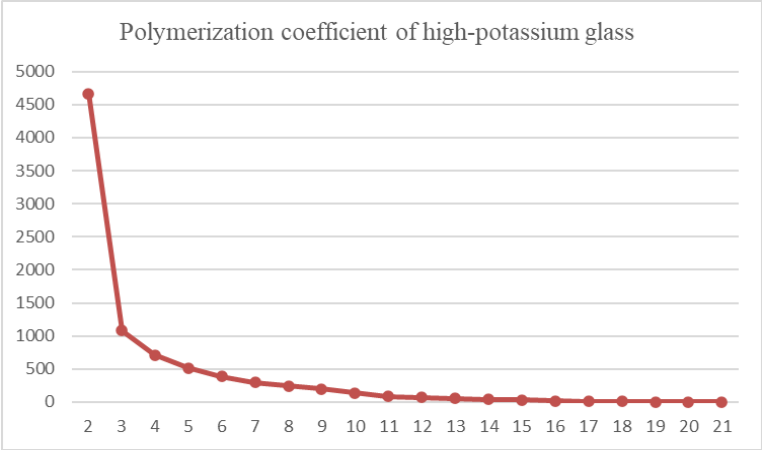


Figure 2. Statistics of polymerization coefficient of high-potassium glass

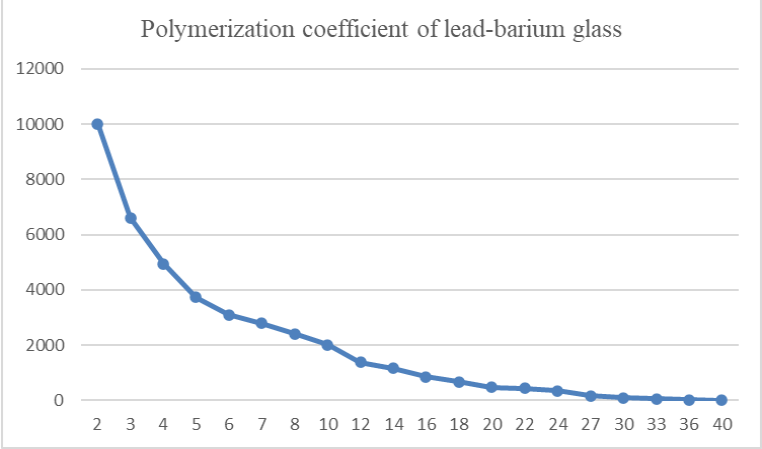


Figure 3. Statistics of polymerization coefficient of lead-barium glass

According to Figure 2, when K changes the most from 1 to 3, the polymerization coefficient changes the most, and the change of the aggregation coefficient slows down significantly after 3, so the "elbow" is set to 3, that is, the center of gravity of the high-potassium glass cluster is set to 3. Similarly, the center of gravity of the lead-barium glass cluster is set at 5.

1) Perform K-means++ cluster analysis

- High potassium glass clustering
  - Generate an initial cluster center

As shown in Table 1, K-means++ cluster modeling first randomly selects a sample as the first cluster center, calculates the shortest distance between each sample and the current existing cluster center (that is, the distance from the nearest cluster center), the larger this value, the greater the probability of being selected as the cluster center; Finally, use the roulette method (selected according to the probability size) to select the next cluster center, and repeat step 2 until K cluster centers are selected.

Table 1. Initial clustering centers

Initial cluster center			
chemical composition	clustering		
	1	2	3
Silica (SiO2)	59.01	96.77	4.61
Sodium oxide (Na2O)	2.86	0.00	0.00
Potassium oxide (K2O)	12.53	0.92	0.00
Calcium oxide (CaO)	8.70	0.21	3.19
Magnesium oxide (MgO)	0.00	0.00	0.00
Alumina (Al2O3)	6.16	0.81	1.11
Iron oxides (Fe2O3)	2.88	0.26	0.00
Copper oxide (CuO)	4.73	0.84	3.14
Lead oxide (PbO)	0.00	0.00	32.45
Barium oxide (BaO)	0.00	0.00	30.62
Phosphorus pentoxide (P2O5)	1.27	0.00	7.56
Strontium oxide (SrO)	0.00	0.00	0.53
Tin oxide (SnO2)	0.00	0.00	0.00
Sulfur dioxide (SO2)	0.00	0.00	15.03

- Iteration recording

Table 2. Iteration history

Iteration history			
iterate	Changes in the cluster center		
	1	2	3
1	6.757	7.601	0.000
2	0.000	0.000	0.000

According to Table 1, convergence is achieved because there is no or only small change in the cluster center. The current iteration is 2. The minimum distance between the initial centers is 41.147.

- Final cluster center

Table 3. Final cluster centers

The final cluster center			
chemical composition	clustering		
	1	2	3
Silica (SiO2)	63.20	89.66	4.61
Sodium oxide (Na2O)	1.24	0.00	0.00
Potassium oxide (K2O)	10.05	1.99	0.00
Calcium oxide (CaO)	5.21	1.33	3.19
Magnesium oxide (MgO)	1.10	0.44	0.00
Alumina (Al2O3)	6.30	2.76	1.11
Iron oxides (Fe2O3)	2.08	0.44	0.00
Copper oxide (CuO)	2.52	1.49	3.14
Lead oxide (PbO)	0.37	0.14	32.45
Barium oxide (BaO)	0.47	0.22	30.62
Phosphorus pentoxide (P2O5)	1.29	0.53	7.56
Strontium oxide (SrO)	0.04	0.01	0.53
Tin oxide (SnO2)	0.00	0.26	0.00
Sulfur dioxide (SO2)	0.11	0.00	15.03

As shown in Table 3, three cluster centers can be obtained according to repeated iteration of the K-means algorithm.

- Lead-barium glass clustering

- Generate an initial cluster center

Table 4. Initial clustering centers

Initial cluster center					
chemical composition	clustering				
	1	2	3	4	5
Silica (SiO2)	16.71	21.70	49.01	3.72	75.51
Sodium oxide (Na2O)	0.00	0.00	2.71	0.00	0.00
Potassium oxide (K2O)	0.00	0.00	0.00	0.40	0.15
Calcium oxide (CaO)	1.87	6.40	1.13	3.01	0.64
Magnesium oxide (MgO)	0.00	0.95	0.00	0.00	1.00
Alumina (Al2O3)	0.45	3.41	1.45	1.18	2.35
Iron oxides (Fe2O3)	0.19	1.39	0.00	0.00	0.00
Copper oxide (CuO)	0.00	1.51	0.86	3.60	0.47
Lead oxide (PbO)	70.21	44.75	32.92	29.92	16.16
Barium oxide (BaO)	6.69	3.26	7.95	35.45	3.55
Phosphorus pentoxide (P2O5)	1.77	12.83	0.35	6.04	0.13
Strontium oxide (SrO)	0.68	0.47	0.00	0.62	0.00
Tin oxide (SnO2)	0.00	0.00	0.00	0.00	0.00
Sulfur dioxide (SO2)	0.00	0.00	0.00	15.95	0.00

As shown in Table 4, K-means++ cluster modeling first randomly selects a sample as the first cluster center, and calculates the shortest distance between each sample and the current existing cluster center (that is, the distance from the nearest cluster center), the larger this value, the greater the probability of being selected as the cluster center; Finally, use the roulette method (selected according to the probability size) to select the next cluster center, and repeat step 2 until K cluster centers are selected.

- Iteration recording

Table 5. Iteration history

Iteration history					
iterate	Changes in the cluster center				
	1	2	3	4	5
1	6.658	8.338	6.632	15.243	10.934
2	6.390	5.280	2.209	5.089	2.184
3	0.000	1.958	3.203	0.000	0.000
4	0.000	0.000	1.270	0.000	1.248
5	0.000	0.000	0.000	0.000	0.000

As shown in Table 5, convergence is achieved because there is no or only small change in the cluster centers. The current iteration is 5. The minimum distance between the initial centers is 29.003.

- Final cluster center

As shown in Table 6, according to the repeated iteration of the K-means algorithm, five cluster centers can be obtained.

Table 6. Final cluster centers

The final cluster center					
chemical composition	clustering				
	1	2	3	4	5
Silica (SiO2)	20.59	31.91	49.60	18.90	64.32
Sodium oxide (Na2O)	0.00	0.33	2.90	0.00	0.95
Potassium oxide (K2O)	0.08	0.29	0.16	0.10	0.17
Calcium oxide (CaO)	1.54	3.08	1.26	1.60	1.35
Magnesium oxide (MgO)	0.60	0.70	0.77	0.00	0.71
Alumina (Al2O3)	2.19	3.71	3.52	1.20	6.22
Iron oxides (Fe2O3)	0.05	1.39	0.28	0.00	0.80
Copper oxide (CuO)	0.76	1.22	1.79	8.26	0.61
Lead oxide (PbO)	61.29	42.11	25.35	29.32	16.38
Barium oxide (BaO)	5.24	7.51	9.21	31.29	6.04
Phosphorus pentoxide (P2O5)	5.33	5.07	2.10	3.23	0.58
Strontium oxide (SrO)	0.82	0.33	0.31	0.59	0.18
Tin oxide (SnO2)	0.00	0.12	0.00	0.00	0.14
Sulfur dioxide (SO2)	0.00	0.00	0.00	5.12	0.33

2.3 Data analysis and subspecies classification

(1) Three-level nomenclature

(2) The tertiary nomenclature is a common grain size classification naming principle for clastic rocks.

(3) A grain content greater than or equal to 50% of the main name of the rock, that is, the basic name; Grain levels between 50%~25% are written before the basic name in the form of the adjective "×× quality"; The granular content of 25%~10% is used as a secondary adjective, and it is written in the form of "containing ××" at the front; Grain levels containing less than 10% are generally not reflected in the name of the rock.

(4) Second, if the particle size of the clastic rock is poorly sorted and contains more grain grades, but the content of none of the grain grades is greater than or equal to 50%, and the content of 50%~25% is more than one, then the particle level with a content of 50%~25% is compounded, and the form of "××-×× rocks" is expressed, and the content is written later. Other grain grades with low content are still treated according to the first principle.

(5) If the particle size sorting of clastic rock is worse, not only is there no grain grade with content greater than 50%, but there is no or only one grain grade with a content of 50%~25%, all the particle size components of this rock should be combined into three major grades: gravel, sand and silt, and then named according to the first two principles.

(6) Nomenclature of high-potassium glass and lead-barium glass based on the three-level nomenclature

(7) According to the clustering results, the chemical content of the first item of the first category (the silica content in the first class) of the final clustering result can be set as E11, and the chemical content of each category can be defined as Eij (i=1, 2, 3) (j=1, 2, ..., 14). Silica was removed and the ratio of chemical content in each class was calculated Eij (i=1, 2, 3) (j=2, 3, ..., 14).

$$W_{ij} = \frac{E_{ij}}{\sum_{j=1}^{14} E_{ij}} (j = 1, 2, \cdots, 14) \tag{3}$$

According to the three-level nomenclature:

(1) There are no chemical components in cluster 1 with a content greater than or equal to 50%; The content is between 50%~25% with potassium oxide; The content of 25%~10% is calcium oxide and alumina, so it is named aluminum-containing calcium potassium glass

(2) There are no chemical components in cluster 2 with a content greater than or equal to 50%; The content is between 50%~25% with alumina; The content of 25%~10% is copper oxide, potassium oxide and calcium oxide, so it is named copper-containing calcium potassium aluminum glass.

(3) No chemical components in cluster 3 with a content greater than or equal to 50%; The content is between 50%~25%, there are lead oxide, barium oxide; The content of 25%~10% has sulfur dioxide, so it is named sulfur-containing lead barium glass.

According to the three-level nomenclature:

(1) The content of cluster 1 is greater than or equal to 50% as lead oxide; No chemical composition between 50%~25%; There is no chemical composition between 25%~10%, so it is named lead glass.

(2) The content of cluster 2 is greater than or equal to 50% is lead oxide; No chemical composition between 50%~25%; The content of barium oxide is 25%~10%, and the sodium content is lower than that of cluster 3, so it is named low-sodium barium-containing lead glass.

(3) The content of cluster 3 is greater than or equal to 50% as lead oxide; No chemical composition between 50%~25%; The content is 25%~10% barium oxide, and the sodium content is higher than that of cluster 2, so it is named high-sodium barium-containing lead glass.

(4) There are no chemical components in cluster 4 with a content greater than or equal to 50%; The content is between 50%~25% with lead oxide and barium oxide; The content of 25%~10% has copper oxide, so it is named copper-containing lead barium glass.

(5) No chemical components in cluster 5 with a content greater than or equal to 50%; The content is between 50%~25% with lead oxide; The content of 25%~10% is alumina and barium oxide, so it is named aluminum-containing barium lead glass.

3. Model plausibility and sensitivity analysis

3.1 Plausibility analysis

(1) Silhouette Coefficient uses a measure of similarity between objects in a dataset to evaluate the quality of clustering, which is an indicator of the density and dispersion of clusters. Profile factors are useful when the actual category information is unknown. The following steps explain and calculate the contour coefficients for individual points. This process consists of the following three steps. Take Euclidean distance as an example, but similarity can be used using a similar approach.

1) For the i-th object, calculate its average distance to all other objects in the cluster, which is denoted as ai.

2) For the i-th object and any cluster that does not contain the object, calculate the average distance of that object to all objects in the given cluster. For all clusters, find out the minimum; The value is denoted as b_i.

3) For the ith object, the contour coefficient is $o_i = \frac{b_i - a_i}{\max(a_i, b_i)}$.

The value of the profile coefficient varies between -1 and 1. We don't want negative values because negative values indicate that the average distance from a point to a point within a cluster a_i greater than the minimum average distance from a point to other cluster b_i s, and we want the profile coefficient to be positive (a_i<b_i), and the closer the a_i is to 0, the better, because when a _i=0, the contour coefficient takes its maximum value of 1. We can simply take the average of the contour coefficients of the midpoints of the cluster and calculate the average contour coefficients of the clusters. By calculating the average profile coefficient for all points, a total measure of cluster goodness can be obtained.

(2) Plausibility analysis based on profile coefficients

According to the above steps, the contour coefficient plot of high potassium glass and lead barium glass can be obtained, as can be seen from the figure 4 and Figure 5 below, the contour coefficient can be longitudinally compared to the model clustering effect under the condition that the model takes different centroids, when k is 3 and 5 respectively, the contour coefficient is closer to 1, so k takes 3 and 5 respectively is more reasonable.

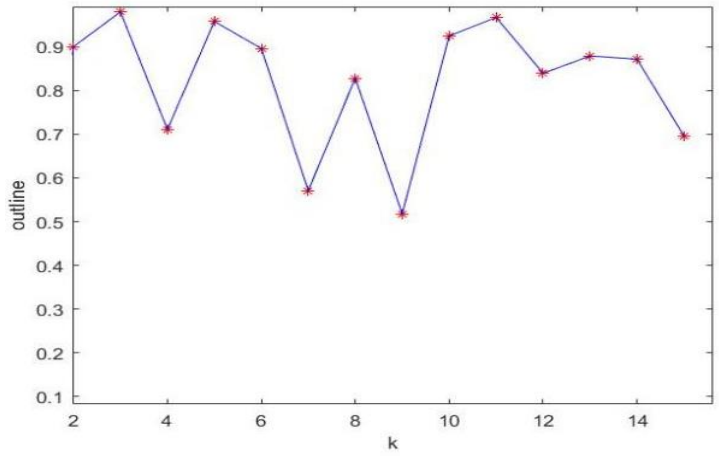


Figure 4. High potassium profile coefficient

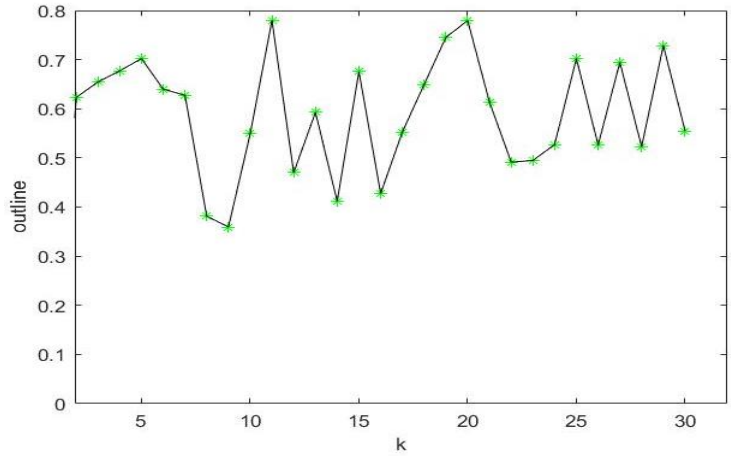


Figure 5. Lead barium profile coefficient

3.2 Sensitivity analysis

When the chemical composition fluctuates, it has an impact on the clustering effect, which is first reflected in the aggregation coefficient, so our team chose the aggregation coefficient as a parameter for sensitivity analysis. The effect on the aggregation coefficient when each chemical component fluctuates by 10% continuously was examined to measure the sensitivity of the clustering model in this paper.

Figure X shows the sensitivity analysis of these 14 chemical components, which shows that sulfur dioxide and strontium oxide have little influence on the polymerization coefficient and low sensitivity. Copper oxide, alumina and silica have a great influence on the polymerization coefficient and have a high degree of sensitivity.

At the same time, the analysis can be obtained, with copper oxide as a typical representative, it shows a general negative correlation with the polymerization coefficient, such as silica, calcium oxide and other components show a rough positive correlation with the polymerization coefficient, in addition, such as sulfur dioxide, strontium oxide, etc., there is no obvious correlation with the polymerization coefficient.

4. Conclusion

In this paper, the K-means++ algorithm is used for clustering, and the aggregation coefficient of each k value is calculated to determine the number of clusters, and the high-potassium glass is divided into 3 categories and the lead-barium glass is divided into 5 categories. The three-level nomenclature is used to name each type according to the amount of chemical component content contained in each category, and its division is listed according to the names of various types of high-potassium glass

and lead-barium glass, among which, high-potassium glass can be divided into aluminum-calcium potassium-containing glass, copper-containing calcium potassium-containing aluminum glass, sulfur-containing lead barium glass; Lead barium glass can be divided into lead glass, low sodium barium-containing lead glass, high sodium barium-containing lead glass, copper-containing barium-containing barium glass and aluminum-containing barium-containing lead glass. Then, the rationality analysis is carried out, and the profile coefficient is used to evaluate the quality of the clustering, and it is reasonable to obtain that the number of clusters of high potassium glass and lead barium glass is 3 and 5 respectively. Finally, the sensitivity analysis of the polymerization coefficient shows that sulfur dioxide and strontium oxide have little influence on the polymerization coefficient and have low sensitivity. Copper oxide, alumina and silica have a great influence on the polymerization coefficient, and the sensitivity is high. Copper oxide and other components, which show a general negative correlation with the polymerization coefficient, such as silica, calcium oxide and other components show a rough positive correlation with the polymerization coefficient, in addition, such as sulfur dioxide, strontium oxide, etc., there is no obvious correlation with the polymerization coefficient.

References

[1] WANG Dong, WEN Rui, ZHU Yingpei, HU Xingjun, LI Wenying. Study on metal foil layer glass beads excavated from Yingpan cemetery in Weili County, Xinjiang [J]. Archaeology and Cultural Relics, 2022(04): 117-122.)

[2] Do Khai Ly, Su Miao, Yang Yuanyuan. Communication and Inclusion: The Influence of Chinese Culture on Decorative Arts along the Silk Road [J]. Journal of Donghua University (Social Sciences), 2022, 22(02): 37-46. DOI:10.19883/j.1009-9034.2021.0319.

[3] Cui Ning, Wang Gong. Sino-foreign exchanges on the grassland Silk Road in Tongliao region of the Liao Dynasty [J]. Journal of Inner Mongolia University for Nationalities (Social Sciences Edition), 2022, 48(01): 23-28. DOI:10.14045/j.cnki.nmsx.2022.01.001.

[4] JIANG Bo. Ports, shipwrecks and trade goods: archaeological discovery and research of the Maritime Silk Road [J]. Studies in the History of Maritime Communication, 2021(04): 8-22. DOI:10.16674/j.cnki.cn35-1066/u.2021.04.002.

[5] An Jiayao. Silk Road and Glassware [J]. Cultural Relics Tiandi, 2021(12): 87-92.)

[6] Xu Siwen, Gu Zhou, Yang Yimin. Distribution of Indo-Pacific pearls before the 2nd century and maritime Silk Road [J]. Chinese Journal of History of Science and Technology, 2021, 42(02): 281-292+321.)

[7] Wang Siyu. Cultural relics of the Lingnan Han Dynasty Maritime Silk Road from the "Silk Road Sailing" exhibition [J]. Collection, 2021(02): 162-171.)

[8] Wang Xiaofu. Journal of Tsinghua University (Philosophy and Social Sciences), 2020, 35(04): 1-14+211. DOI:10.13613/j.cnki.qhdz.002956.

[9] Tan Xin. The transmission and influence of Islamic glass on the Silk Road (8th-16th centuries) [D]. Jinan University, 2020. DOI:10.27167/d.cnki.gjinu.2020.000412.

[10] Mao Mingxia. Nanliujiang Ancient Economic and Social Development and Maritime Silk Road [D]. Guangxi Normal University, 2020. DOI:10.27036/d.cnki.ggxsu.2020.000888.

[11] Qi Haoyue. Analysis of cultural relics of foreign cultural factors in the Liao Dynasty under the background of the Silk Road [D]. Inner Mongolia Normal University, 2020. DOI:10.27230/d.cnki.gnmsu.2020.000051.