Subclassification of Chemical Composition of Glass Based on Principal Component Analysis and K-Means Clustering

Jiaxin Xie

School of Mathematics and Statistics, Lanzhou University, Lanzhou, China, 730030

xiejx@lzu.edu.cn

Abstract. Accurately carrying out the subclassification of the chemical composition of glass products is of great significance for the compositional analysis and identification of archaeological artefact samples. Firstly, this study collected the relevant data of glass composition and divided it into two groups: high potassium glass and lead-barium glass. Combined with the controlled variable method, the analysis concluded that the weathering condition of the surface of high-potassium glass artifacts has a great correlation with the decoration, while the weathering condition of the lead-barium glass artifacts has no correlation with the decoration and colour. The artefacts were divided into four categories according to the glass type and weathering condition, resulting in a law indicating the effect of weathering on chemical composition. Thus, the weathering point is predicted inversely. Then, this paper carries out a significance analysis of the various components of the two types of glass, respectively, and obtains that the F-values of silica, potassium oxide, and barium oxide are 38.28, 30.31, and 26.99, respectively, which are the most significant. Secondly, using the comprehensive determination model, the principal component analysis method was firstly adopted to obtain the greatest influence of CaO on the overall variance. Then, K-means clustering method was used, and through many iterations, it was finally concluded that: for high potassium glass, according to whether the CaO content is greater than 4 as the subclass division method; for lead-barium glass, according to whether the CaO content is greater than 3 as the subclass division method, and listed all the samples of the subclass division method, and finally verified the reasonableness of the subclass division method and the results, and verify that the model has good sensitivity.

Keywords: Glass chemical composition, Subclassification, Principal component analysis, K-means clustering, Comprehensive determination model.

1. Introduction

Glass is a precious evidence of early trade between China and the West through the Silk Road, China absorbed foreign technology and then used local materials to make, so although the appearance is similar, the chemical composition of glass products at home and abroad is very different. And because of the ancient glass weathering and burial environment has a great correlation, and by the glass type, decoration, colour and other factors. During the weathering process, its proportional composition will also change. Therefore, the analysis and identification of the composition of cultural relic samples is a crucial part of archaeological work.

This study intends to solve the following 2 problems:

(1) To analyse the relationship between the weathering of the surface of glass artefacts and their glass types, decorations and colours; to analyse the statistical patterns of the weathering of the surface of the artefact samples and the content of each chemical composition in relation to the glass types; and to predict the content of the chemical compositions of artefacts before weathering by backward extrapolation from the monitoring data of the weathering points.

(2) Analyse the classification rules of high-potassium glass and lead-barium glass; for each type, select the appropriate chemical composition for subclassification, and give the corresponding classification scheme and its classification results, and finally analyse the reasonableness and sensitivity of the classification results.

2. Materials and methods

2.1. Data acquisition and pre-processing

Accessed from Open Source Data. (http://www.mcm.edu.cn/index_cn.html)

2.2. Methodology

2.2.1 Advantages and Principles of Principal Component Analysis (PCA)

The advantages of Principal Component Analysis are: data dimensionality reduction, de-correlation, data visualisation, feature extraction, noise filtering.

The goal of PCA is to project the data into a new coordinate system in order to retain the maximum amount of variance and to reduce the dimensionality of the data by selecting principal components while maintaining as much information as possible. These principal components are linear combinations of the original data and allow for easier analysis and understanding of the data.

2.2.2 Advantages and principles of K-means clustering

The advantages of K-means clustering are simplicity and efficiency, scalability, intuition, wide range of applications, and applicability.

The goal of K-mean clustering is to minimise the variance of the data points within a cluster while maximising the differences between different clusters. This is achieved through a continuous iterative allocation and updating process. Although K-mean clustering performs well in many cases, it makes certain assumptions about the shape and size of the clusters and therefore may not perform well in some cases. In practice, it is often necessary to choose an appropriate value of k, as well as to cope with the choice of initial centre of mass and local minima [1].

3. Results

3.1. Subclassification using statistical laws

In this paper, the proposed problem is divided into two steps. In the first step, according to the glass type, the sample space is grouped by preprocessing, and bar charts and bar graphs are made by data visualisation and analysis [2]. Observe the relationship between weathering on the surface of artifacts and glass type, decoration and colour, and perform chi-square test. In the second step, this paper divides the samples into four categories according to glass type and weathering, and calculates the statistics of the four categories and fourteen component indicators. After deriving the law of weathering's influence on the indicators, the weathering point indicators are predicted inversely. Then the hypothesis test is carried out on the predicted indicators and unweathered indicators, if it is valid, the hypothesis is accepted; if not, the sample is reprocessed.

3.1.1 Data pre-processing

Firstly, the given samples were roughly classified into the following two categories based on the type of glass, as shown in Table 1.

Table 1. Glass classification table

Group	Serial number of cultural relic	Classification standard
One	01, 03, 04, 05, 06, 07, 09, 10, 12, 13, 14, 15, 16, 17, 18, 21, 22, 27	High potassium glass
Two	02, 08, 11, 19, 20, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58	Lead barium glass

Next, the first set of samples was subjected to data visualisation by drawing a bar chart of the amount of surface weathering versus glass type, grain, and colour, as shown in Figure 1.

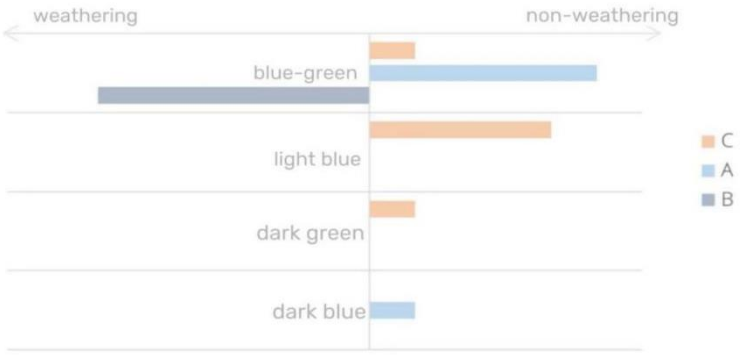

Figure 1. Weathering quantities plotted against each factor

Observation of the statistical graphs reveals that for the high-potassium glass artefacts, there are and only B patterns with surface weathering. While A and C patterns do not show surface weathering regardless of the colour [3]. This leads to Conjecture 1: For high-potassium glass artefacts, the type of pattern has a greater relationship with the surface weathering. The following chi-square analysis is performed for this conjecture. We calculated Table 2.

Table 2. Chi-square analysis results of high potassium glass

	Chi-square	P value	DOF
Pattern	18	0.0001	2
Color	4.5	0.21	3

From the data in the table, grain $P \leq 0.05$ indicates that the degree of surface weathering of high-potassium type of glass has a great relationship with the type of grain, while colour $P > 0.05$ indicates that the degree of surface weathering of high-potassium type of glass has less correlation with the colour, which is in line with Conjecture 1.

The second set of samples was then analysed.It was found that for lead-barium glass artefacts, whether weathered or unweathered, the A and C ornamentation ratios were approximately the same. By calculating Table 3.

Table 3. Scale of weathering and non-weathering

Type	Proportion
Weathering	0.76
Non-weathering	0.71

Observing the statistical images, we obtained that the difference in grain pattern of lead-barium glass artefacts has less influence on the degree of weathering. Therefore, we ignore the texture difference and reclassify the lead-barium glass only by colour, as shown in Figure 2.

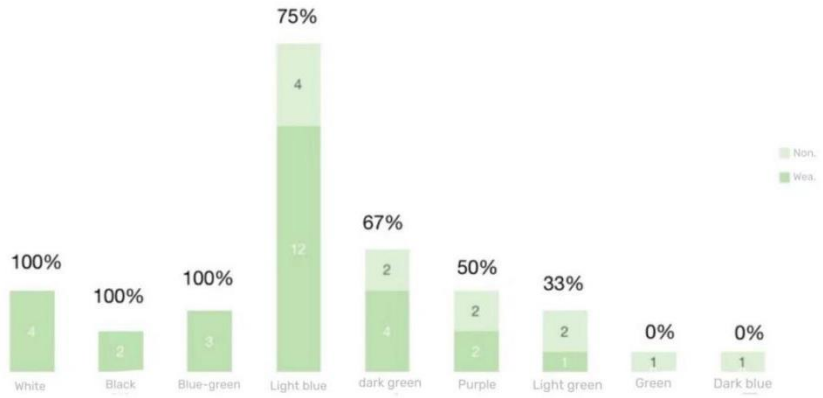

Figure 2. Statistical chart of lead barium glass bar

The numbers above the bars represent the proportion of weathered samples of that colour to the total number of samples of that colour, and a decreasing trend can be clearly seen. This leads to conjecture two: for lead-barium glass artifacts, grain type has no relationship with surface weathering,

and colour has a greater relationship with surface weathering. The following chi-square analysis was carried out for this conjecture, and after python calculation, Table 4 was obtained.

Table 4. Results of chi square analysis of lead barium glass

	Chi-square	P value	DOF
Pattern	0	1	1
Color	9	0.25	7

Since P≥0.05, the hypothesis is not valid. Therefore, for lead-barium glass artifacts, there is no great correlation between pattern type, colour and surface weathering [4].

In summary, we get the following conclusions: for the high potassium glass artifacts, the pattern has a great correlation with the surface weathering, while the colour has little correlation with the surface weathering; for the lead-barium artifacts, neither the pattern nor the colour has a great correlation with the surface weathering.

3.1.2 Modelling and solving

Based on the glass type and degree of weathering, the data given in this paper are subdivided into the following four categories [5], as shown in Table 5.

Table 5. Classification table according to glass type and degree of weathering

Group	Serial number of cultural relic	classification standard
One	01, 03.1, 03.2, 04, 05, 06.1, 06.2, 13, 14, 15, 16, 17, 18, 21	High potassium glass without weathering
Two	07, 09, 10, 12, 22, 27	High potassium glass with weathering
Three	20, 23.0, 24, 25.0, 28.0, 29.0, 30.1, 30.2, 31, 32, 33, 35, 37, 42.1, 42.2, 44.0, 45, 46, 47, 49.0, 50.0, 53.0, 55	Lead barium glass without weathering
Four	02, 08, 08.9, 11, 19, 26, 26.9, 34, 36, 38, 39, 40, 41, 43.1, 43.2, 48, 49, 50, 51.1, 51.2, 52, 54, 54.9, 56, 57, 58	Lead barium glass with weathering

Where X.0 denotes the unweathered point, X.1 and X.2 denote the partial one and two, and X.9 denotes the severe weathering point. The statistical values of the 14 indicators before and after weathering of the two types of glass are derived according to the question [6, 7].

From this, the chemical composition content of the weathering points before weathering was predicted, and the results are shown in Tables 6 and 7.

Table 6. Prediction of chemical composition of high potassium glass

High potassium glass	SiO_2		K_2O	
	Actual value	Predicted value	Actual value	Predicted value
09	95.02	67.87	0.59	9.83
10	96.77	69.12	0.92	15.33
12	94.29	67.35	1.01	16.83
22	92.35	65.96	0.74	12.33

Table 7. Prediction of chemical composition of lead barium glass

lead barium glass	SiO_2		K_2O	
	Actual value	Predicted value	Actual value	Predicted value
02	36.28	78.87	47.43	24.20
43.2	21.70	47.17	44.75	22.83
51.2	21.35	46.41	51.34	26.19
52	25.74	55.96	47.42	24.19
54	22.28	48.43	55.46	28.30
57	25.42	55.26	45.10	23.01

3.2. K-means clustering and integrated evaluation models

In order to find the classification law of high potassium glass and lead-barium glass, this paper carried out the significance analysis of 14 indexes of the two types of glass. Arrive at potassium chloride and lead oxide have the smallest P-value among the two types of glass, i.e., the probability of the difference between the two types of data due to statistical error is almost 0. Thus, these three types of indicators are obtained as the basis for the classification of glass artefacts. The medians of the three types of indicators are listed for comparison, and the results are shown in Table 8.

Table 8. Table of the median values of three categories of indicators

	SiO_2	K_2O	PbO
Lead barium glass	35.78	0	31.9
High potassium glass	83.255	3.1	0

From the table, it can be seen that: for high-potassium glass, its silica content is 83.255 per cent up and down, potassium oxide content is 3.1 per cent up and down, and lead oxide content is almost 0 per cent; while for lead-barium glass, its silica content is 35.78 per cent up and down, potassium oxide content is almost 0 per cent up and down, and lead oxide content is 31.9 per cent up and down.

Step1: Integrated decision-making model

Method1: Principal Component Analysis

Define the variable correlation matrix as,

$$\begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix} \tag{1}$$

$$\rho_{ij} = \frac{cov(X_i^*,\ X_j^*)}{\sqrt{D(X_i^*)D(X_j^*)}} \tag{2}$$

The covariance matrix eigenvalues obtained through python are shown in Table 9.

Table 9. Eigenvalues of covariance matrix

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11-14}
238.15	10.45	8.02	4.23	2.06	0.87	0.6	0.3	0.16	0.065	0

From the knowledge of principal component analysis model, it is known that the percentage of each component is.

$$F_k = \sum_{i=1}^{n} a_{ik}(X_i - \overline{X_\iota}) \tag{3}$$

The variance contribution of each component is

$$P_k = \frac{\lambda_k}{\sum_{i=1}^{n} \lambda_i} \tag{4}$$

Bringing in the data table yields that when n = 3,

$$\sum_{k=1}^{3} P_k = 0.968 \tag{5}$$

It can be seen that the contribution of variance from the first two principal components is 96.8 per cent, which is sufficiently high to be credible, so the data were reduced to three dimensions, and the results are shown in Table 10, as follows.

Table 10. Contribution Table

	F1	F2	F3		F1	F2	F3		F1	F2	F3
SiO_2	-0.281	-0.131	0.936	CuO	0.094	-0.181	0.052	Al_2O_3	-0.018	-0.263	-0.011
Na_2O	0.467	0.291	0.316	PbO	0.115	0.009	0.003	Fe_2O_3	0.103	0.028	0.063
K_2O	-0.440	0.703	0.018	BaO	-0.097	-0.127	-0.062	SnO_2	0.012	0.001	0.002
CaO	0.604	0.411	0.093	P_2O_5	-0.037	-0.030	-0.007	SO_2	-0.006	0.006	0.001
MgO	-0.315	0.289	-0.051	SrO	-0.051	0.177	-0.050				

In the above table, the data in the right three columns indicate the coefficients. As can be seen from the table, among the two components with the largest contribution, the term with the largest coefficient is CaO. The same operation was performed on the samples of lead-barium glass artefacts, and consistent results were obtained. Therefore, it is concluded that CaO is a suitable indicator for differentiating subclasses for high potassium and lead-barium glass artefacts.

Method2: Modified K-means cluster analysis

In the same way as model I firstly, the samples were divided into two groups according to the glass type, and K-means clustering analysis was carried out on each group of artefacts for each index, respectively. After several rounds of iteration, two clustering points are obtained for each [8-10].

The clustering diagram of CaO was drawn and the two clustering points were marked with different colours, and the results are shown in Figure 3.

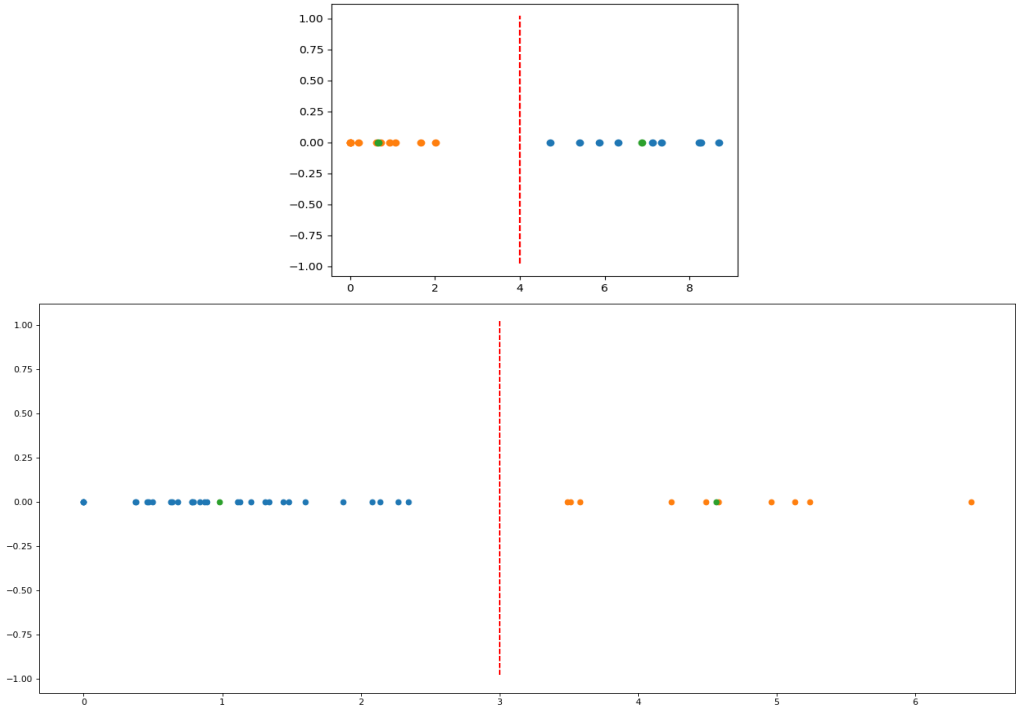

Figure 3. Cluster diagram of CaO content of high-potassium glasses and Lead-barium glass

The above figure proves that CaO has more obvious clustering characteristics in both high-potassium and lead-barium glass artefacts, and is more suitable than other indicators to be used as a subclassification indicator. The division methods and results are listed in Tables 11 and 12.

Table 11. Classification of lead barium glass

Lead barium glass	Sampling number	Classification
Subclass I	54.9, 20, 42.02, 46, 36, 35, 32, 24, 23.0, 25.0, 38, 38, 34, 53.0, 42.01, 45, 47, 37, 39, 55, 56, 57, 28.0, 26, 08, 31, 40, 49.0, 44.0, 52, 02, 48, 19, 29.0	CaO values closer to clustering point 1
Subclass II	26.9, 50.0, 08.9, 50, 54, 58, 11, 51.1, 30.2, 30.1, 49, 41, 51.2, 43.1, 43.2	CaO values closer to clustering point 2

Table 12. Classification of high-potassium glass

high-potassium glass	Sampling number	Classification
Subclass I	06.1, 15, 17, 18, 10, 09, 12, 27, 07, 22, 03.1	CaO values closer to clustering point 1
Subclass II	21, 06.2, 03.2, 01, 04, 05, 14, 16, 13	CaO values closer to clustering point 2

Step2: Reasonableness and Sensitivity Analysis

In this paper, the data visualisation operation is carried out to observe intuitively whether the classification situation of the model is credible and reasonable. According to the above classification criteria, the Form 2 cultural relics samples were classified into two categories of high potassium and lead-barium glass according to the glass category, and then, through Excel data visualisation, the scatter plots of 14 indicators for each of the two categories of cultural relics were drawn respectively (the horizontal axis is the cultural relics sampling points, and the vertical axis is the indicator content), and the results are shown in Figure 4.

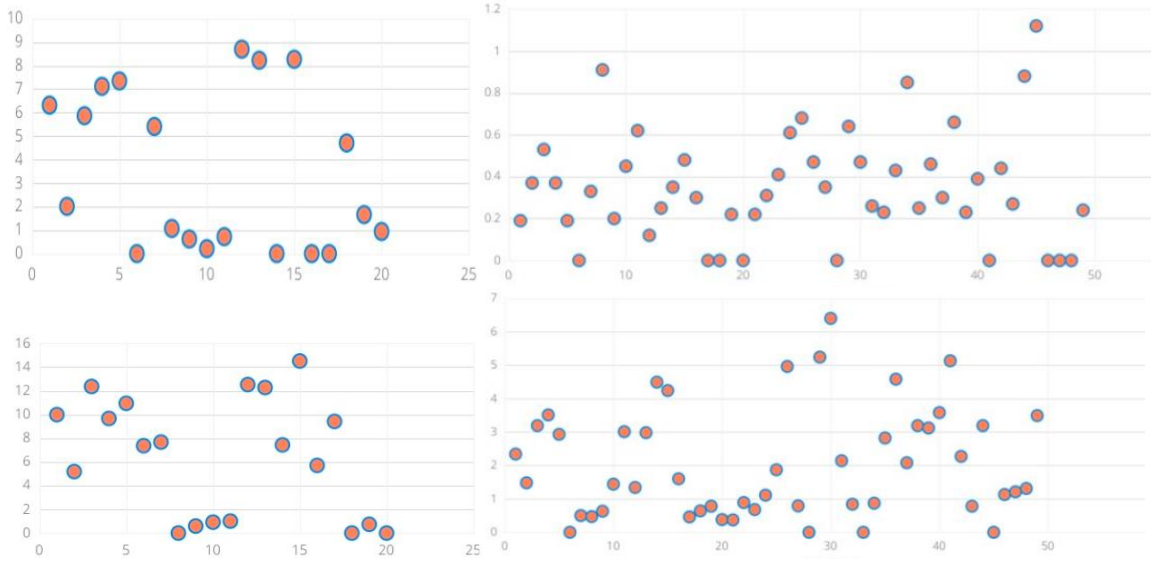

Figure 4. Scatterplot of the content of an indicator in an artefact

Due to space limitations, only four images are picked here, and it is not difficult to find that the scatter plots of CaO indicators have more obvious clustering tendency in both types of artefacts. Thus, the model is proved to be reasonable.

From the comprehensive judgement model, we know that CaO is an important indicator for the two types of glass to distinguish subclasses, so this paper carries out sensitivity analysis on the model through this indicator. By changing its value 7%-10%, observe the change of the clustering point and the actual clustering, the results are shown in Table 13 and Figure 5.

Table 13. Lead and barium glass heritage sensitivity table

CaO rate of change	new meeting point1	rate of change	new meeting point2	rate of change
7	1.029	5.00%	4.92	7.89%
8	1.039	6.02%	4.98	9.21%
9	1.062	8.37%	5.011	9.89%
10	1.075	9.69%	5.061	10.99%

From the above charts we can conclude that the variation of the independent variable within 7% to 10% caused a variation of 5.00% to 9.69% and 7.89% to 10.99% for the two clustering points of high potassium glass, and 6.01% to 11.23% and 5.93% to 10.10% for the two clustering points of lead-barium glass, which is relatively stable, indicating that the model has good sensitivity.

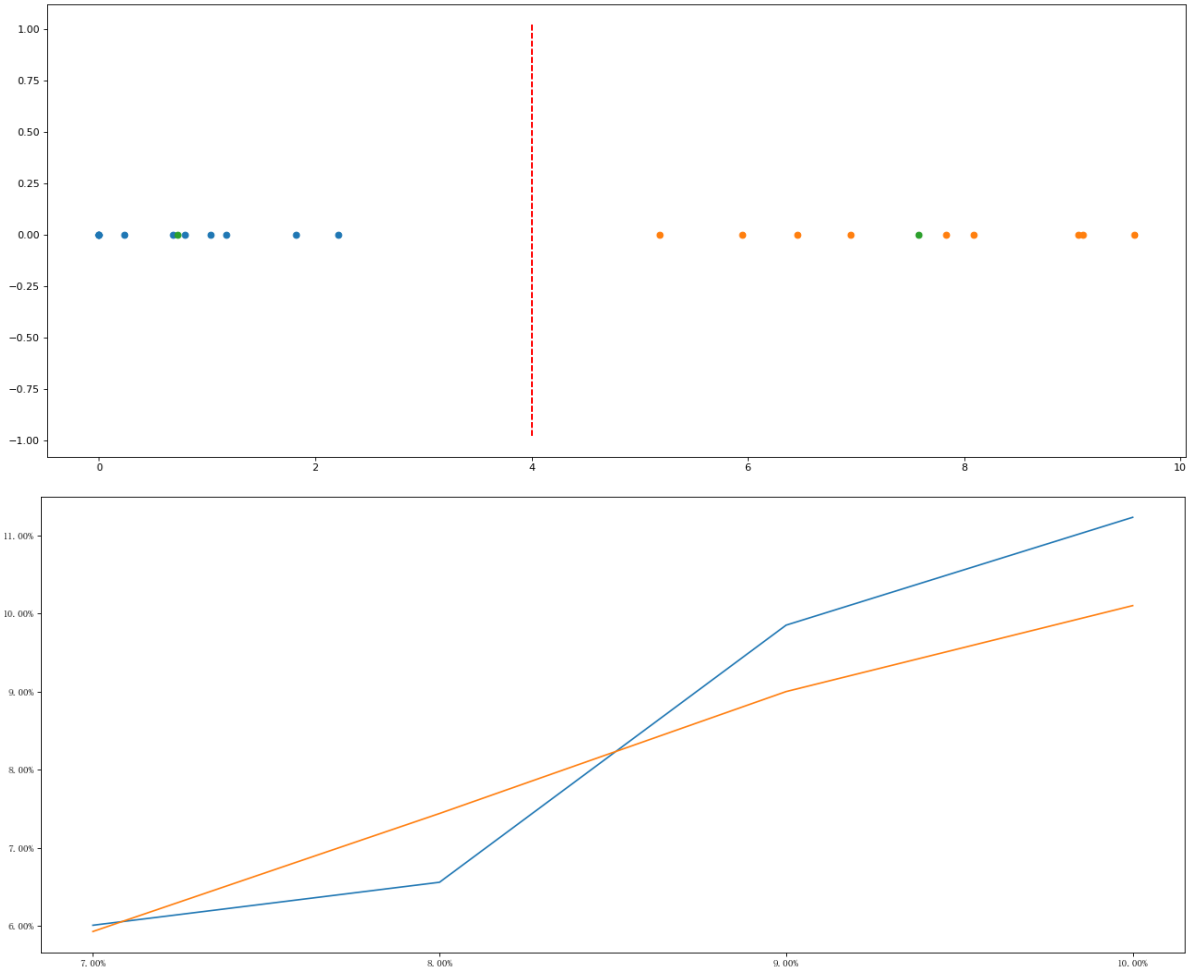

Figure 5. Sensitivity change map for high-potassium glass with new clustering points

4. Conclusion

In this study, the first use of statistical law analysis of high potassium glass artifacts surface weathering conditions and decorations have great relevance, lead-barium glass artifacts weathering conditions and decorations, colour has no relevance.

Secondly, the two types of glass components were analysed for their significance, and silica, potassium oxide and barium oxide were found to be the most significant by F value. Then, using the comprehensive decision model, the principal component analysis was used first, and the CaO had the greatest influence on the overall variance. Then through the K-means clustering method, after many iterations, the subclassification method of high potassium glass and lead-barium glass according to the CaO content was obtained, and the subclassification method of all samples was listed, and finally the subclassification method and the results were verified to be reasonable. Finally, this paper carries out the sensitivity test of the comprehensive determination model, using the change of CaO content of high-potassium glass to get two clustering point changes 5.00%~9.69%, 7.89%~10.99%, and then verified that the model has good sensitivity.

References

[1] SHEN Xiujuan, XUE Shuo. A weighted K-means clustering algorithm and its application [J]. Journal of Qujing Normal College, 2022, 41 (03): 1-7.

[2] WEN Limei, LIANG Guohao, WEI Tongbian, ZHANG Liang, WEI Tongming. Research on data visualization [J]. Information Technology and Informatisation, 2022 (05): 164-167.

[3] Huang Huihong. Detection of mudstone lithology and weathering degree based on deep learning [D]. University of Electronic Science and Technology, 2022.

[4] WEI Manman, LI Shihu, ZHOU Qin. Teaching design of the basic idea of hypothesis testing and related concepts [J]. Mathematics Learning and Research, 2022 (08): 5-7.

[5] LIU Guangyuan, CAO Changyang, ZHANG Ning. Parameter estimation of broken glass composition and its implementation [J]. Glass and Enamel, 2015, 43 (02): 24-26.

[6] Zhao X. The use of interactive data visualisation technology in cultural digital heritage display using edge computing [J]. International Journal of Grid and Utility Computing, 2022, 13 (2-3).

[7] REN Ni, WU Qiong, LI Huoluan. Analysis and research on data visualisation techniques [J]. Electronic Technology and Software Engineering, 2022 (16): 180-183.

[8] XING Tao, YONG Yi, HOU Jiang et al. Comprehensive evaluation of water quality based on principal component analysis and hierarchical cluster analysis [J]. Sichuan Environment, 2022, 41 (04): 131-139.

[9] CHEN Zhuang, JIA Chenghe, JIANG Hong. Convenient sticker micro confocal laser Raman spectroscopy classification based on principal component analysis and K-means [J]. Journal of Chinese People's Public Security University (Natural Science Edition), 2022, 28 (03): 9-14.

[10] ZHANG Zhongwei, YANG Hailong, FU Jun et al. Comprehensive evaluation of maize varieties based on principal component analysis and cluster analysis [J]. Agricultural Science and Techn ology Newsletter, 2022 (06): 30-35.