Component Analysis and Identification of Ancient Glass Products Based on Statistical Methods

ABSTRACT

This paper analyzes its role in the composition analysis and identification of ancient glass products by flexible use of statistical methods, and emphasizes four statistical methods: systematic clustering algorithm, K-means algorithm, logistic regression model and grey correlation analysis. Taking the C project of CUMCM in 2022 as an example, this paper systematically introduces these four common data classification and statistical methods to classify and analyze the given data. In this paper, suitable chemical components of high potassium and lead barium glass were selected for subdivision, and the specific division methods and results w ere given. The chemical composition of glass relics of unknown category was analyzed to identify their type. The grey correlation matrix of surface weathering of high-potassium cultural relics was obtained, and the correlation degree of chemical components was analyzed. This greatly promotes the composition analysis and identification of chemical components in ancient relics.

Keywords:System cluster analysis algorithm, K-means clustering analysis, Logistic regression model, Grey correlation method

1. INTRODUCTION

With the development of society, statistics has been widely used in market research, financial analysis, risk assessment and other fields, which has important significance and status. There are many statistical models, and the thinking of statistical models is also applied in all aspects of social life. Based on the C project of CUMCM in 2022, this paper will focus on the application of systematic clustering algorithm, K-means algorithm, logistic regression model and grey correlation analysis in data analysis and identification.

Both the system clustering algorithm and the K-means clustering algorithm are based on the proximity of distance as the standard. K-means clustering algorithm mainly adopts the iterative calculation method, which is fast and efficient with high accuracy, but it needs to specify the number of cluster categories by itself. The system clustering algorithm can clearly show the tree relationship between classes, and can estimate the most effective number of clusters by Elbow Method. Moreover, system clustering can produce a series of clustering results for different class numbers, while K-means clustering algorithm can only produce clustering results for specified class numbers. Therefore, we use the results of system clustering as a reference for K-means clustering algorithm to determine the

number of classes. If the two classification results are the same, it proves the accuracy of our results.The logic regression model is clear and concise, and the training speed is very fast, and the adjustment of the output results can be easily realized by adjusting the threshold value. Grey correlation analysis is widely applicable to correlation comparison and evaluation of various index variables, and its application scenarios are wide and flexible.

Then, these statistical methods are improved in the analysis and identification of chemical components of ancient glass products, and provide help for their development and optimization.

2. SYSTEMATIC CLUSTER ANALYSIS AND K-MEANS CLUSTER ANALYSIS

2.1 Subdivision of High Potassium Glass

2.1.1 Data Processing and Analysis

The data of high potassium glass were preprocessed and the appropriate chemical components were selected to classify them into sub-classes. It can be seen from the sorted data table that the contents of strontium oxide, tin oxide, sulfur dioxide, lead oxide and barium oxide at different sampling points are very different, so they can be ignored in the cluster analysis. Systematic clustering algorithm is applied to nine chemical components of silicon dioxide, sodium oxide, potassium oxide, calcium oxide, magnesium oxide, alumina, iron oxide, copper oxide and phosphorus pentoxide.

Then, a clustering model is established, and the data are analyzed respectively through the systematic clustering analysis algorithm and K-means algorithm. Finally, the division results are summarized, and the similarities of the same class and the differences between different classes are found to form the basis and method of division[1].

2.1.2 Results of Systematic Cluster Analysis

In this paper, the square Euclidean distance is used as a measurement method to calculate the distance between the above 16 samples and the 9 index factors, and the closest two classes are combined into a new type. The new type will keep all its properties and calculate the distance between the new class and each class, repeating the process until all objects have been clustered together[2].

In the line diagram of polymerization coefficient shown in Figure 1, the horizontal coordinate is the number of categories of clustering(K), and the vertical coordinate is the polymerization coefficient. According to the Elbow Method documented in existing articles, we generally choose the end value of the steepest section of the line graph before it reaches a smooth level as the number of clusters K[3]. If the number before this value is chosen as K, the error sum of squares is too large. If the number after this value is selected as K, the sum of squares of error does not decrease much and has no obvious research significance. In Figure 1, it can be seen that when the value of K is from 1 to 3, the degree of distortion changes the most. When the number of categories is 3, the downward trend of the curve is significantly slowed down, so the number of categories can be set to 3. The 16 samples are analyzed in the following three categories.

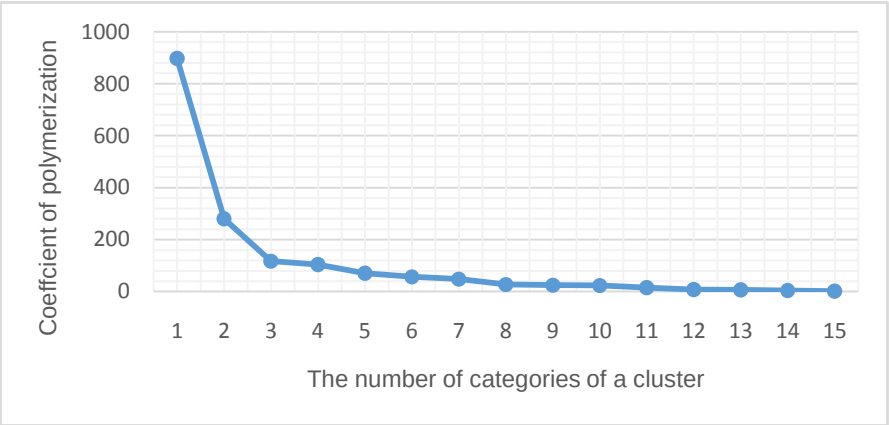

Figure 1. The curve of polymerization coefficient with K value of high potassium type

As shown in Figure 2, when the number of categories is known, high-potassium glass is divided into three categories according to the results of cluster analysis, and the results of systematic cluster analysis of high-potassium glass are obtained, among which there are three cultural relics of the first category, seven cultural relics of the second category, and six cultural relics of the third category. The first category contains 03, 21, 18, the second category contains 01, 04, 05, 06, 13, 16, 14, and the third category contains 07, 09, 10, 12, 22, 27.

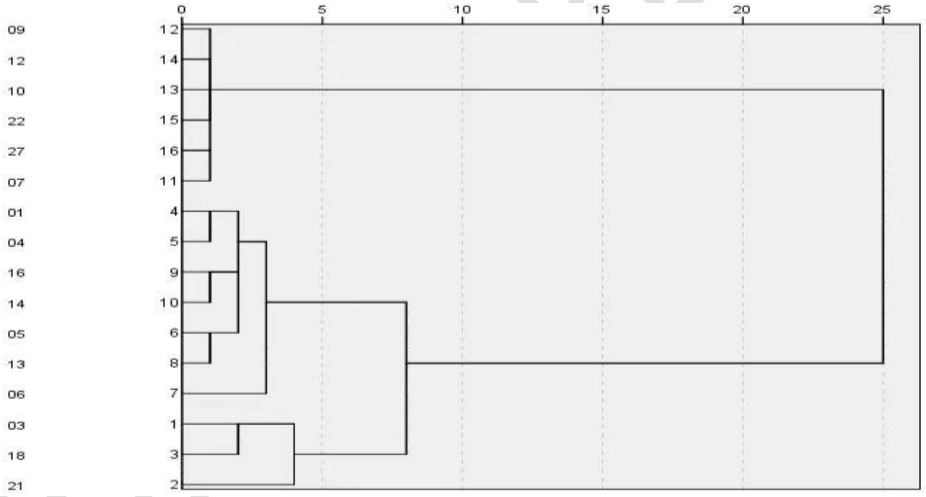

Figure 2.Systematic cluster analysis diagram of high potassium glass

2.1.3 The Result Analysis of K-means Algorithm

By SPSS, K-means algorithm was used to cluster the classification of high potassium glass. With the result of system cluster analysis, the number of categories is set to 3. Here, the K value of k-means algorithm is still set to 3[4]. The final output result is as follows:

Table 1. Number of cases in each cluster

Number of cases in each cluster		
Cluster	1	3.000
	2	7.000
	3	6.000

	effective	16.000
	missing	.000

In addition, it can be found that the K-means clustering analysis algorithm, taking K value as 3, can obtain the same results as the systematic clustering analysis of high potassium glass, as shown in Table 2.Through the combination of these two methods, the accuracy of the classification results is proved.

Table 2. Classification result

Category	Quantity	Cultural relic number
First kind	3	03,21,18
Second kind	7	01,04,05,06,13,16,14
Third kind	6	07,09,10,12,22,27

Therefore, according to the analysis and comparison data, it can be concluded that the sub-category of high-potassium glass is divided into three categories according to the different contents of silica, potassium oxide and calcium oxide in 16 high-potassium glass relics[5]. When the proportion of silica is greater than 90% and the proportion of potassium oxide is very low, it is divided into a class. Then the remaining data, that is, the data of silicon dioxide accounting for less than 90%, is analyzed, and the calcium oxide accounting for less than 5% is divided into one category, and the rest is divided into another category, a total of three categories.

2.2 Subdivision of Lead Barium Glass

2.2.1 Data Processing

The lead barium glass data were pretreated and the appropriate chemical components were selected for subdivision. It can be seen from the sorted data table that the contents of strontium oxide, tin oxide, sulfur dioxide and potassium oxide at different sampling points differ little, so they can be ignored in cluster analysis. The systematic clustering algorithm is applied to the ten chemical components of silicon dioxide, sodium oxide, calcium oxide, magnesium oxide, alumina, iron oxide, copper oxide, lead oxide, barium oxide and phosphorus pentoxide[6].

2.2.2 Results of Systematic Cluster Analysis

The polymerization coefficient output by SPSS is converted into the line diagram of the polymerization coefficient as shown in Figure 3. When the value of K ranges from 1 to 5, it can be seen that the distortion degree of the curve changes the most, and when the value exceeds 5, the distortion degree changes significantly decrease. Therefore, the number of categories can be set to 5, and 45 samples are analyzed in the following 5 categories.

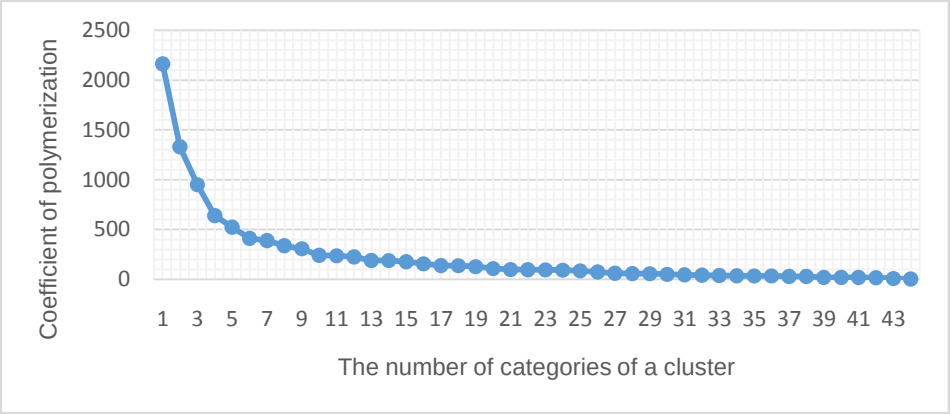

Figure 3. The curve of polymerization coefficient of lead barium type with K value

As shown in Figure 4, when the number of categories is known, lead barium glass is divided into five categories according to the results of cluster analysis, and the results of systematic cluster analysis of lead barium glass are obtained[7]. Among them, there are 5 cultural relics of the first category, 2 cultural relics of the second category, 16 cultural relics of the third category, 3 cultural relics of the fourth category and 19 cultural relics of the fifth category. The first category contains 08 (two sampling points), 24, 26 (two sampling points), the second category contains 11, 20, the third category contains 02, 19, 30, 34, 36, 38, 41, 43 (two sampling points), 50, 51, 52, 54(sampling point 1), 56, 57, 58, and the fourth category contains 39, 40 ,54(sampling point 2). The fifth category includes 23, 25, 28, 29, 31, 32, 33, 35, 37, 42, 44, 45, 46, 47, 48, 49 (two sampling points), 53, 55.

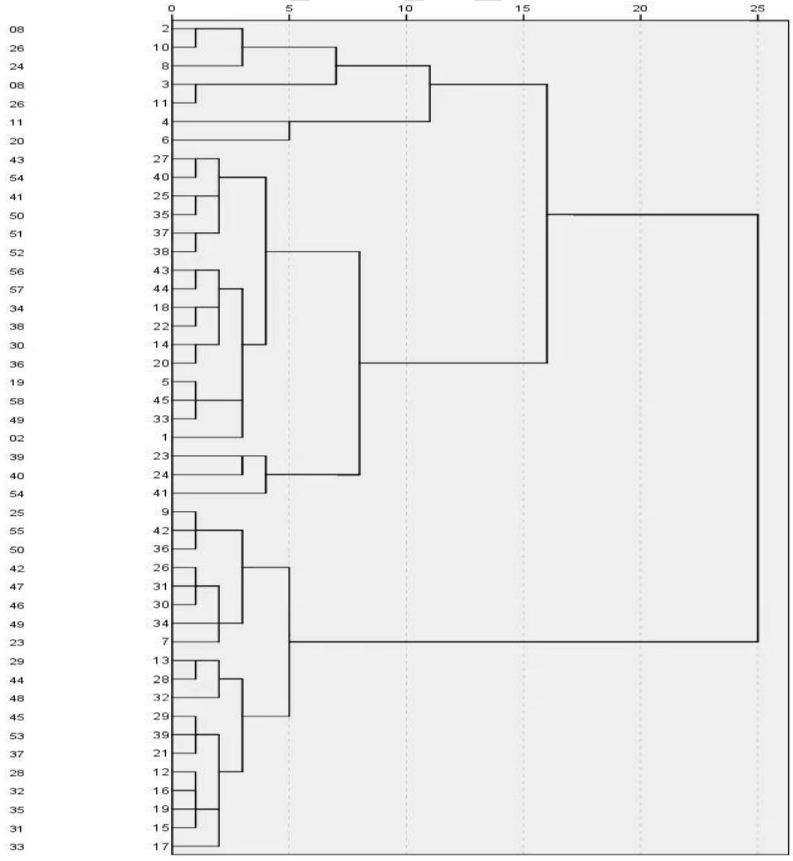

Figure 4. Systematic cluster analysis diagram of lead barium glass

2.2.3 The Result Analysis of K-means Algorithm

By SPSS, K-Means algorithm was used to classify lead-barium glass. In the systematic cluster analysis, the number of categories is set to 5, and the K value is still entered here as 5, and the same result is obtained as the systematic cluster analysis of lead barium glass,as shown in Table 3.

Table 3. Classification result of K-means algorithm when K=5

Category	Quantity	Cultural relic number
First kind	5	08 (two sampling points), 24, 26 (two sampling points)
Second kind	2	11, 20
Third kind	16	02, 19, 30, 34, 36, 38, 41, 43 (two sampling points), 50, 51, 52, 54(sampling point 1), 56, 57, 58
Fourth kind	3	39, 40 ,54(sampling point 2)
Fifth kind	19	23, 25, 28, 29, 31, 32, 33, 35, 37, 42, 44, 45, 46, 47, 48, 49 (two sampling points), 53, 55

Therefore, according to the analysis and comparison data, it can be concluded that the classification method can be divided into five categories according to the different proportions of silicon dioxide, lead oxide and barium oxide in 45 high-potassium glass relics[8]. According to the difference of silica content for the first classification, silica content greater than 50% is classified as a class, less than 50% is a class. Those with a content of less than 50% are further divided according to the content of barium oxide. Those with a content of more than 55% are classified as one category, those with a content of more than 35% and less than 55% are classified as one category, and those with a content of less than 35% are classified as one category. The content of less than 30% is one category, a total of five categories.

3. LOGISTIC REGRESSION MODEL

Firstly, the data were processed. Considering that the contents of sodium oxide, strontium oxide, tin oxide and sulfur dioxide were almost the same among each sample, which had little impact on sample classification, it was not adopted in this paper. The remaining chemical components of silicon dioxide, potassium oxide, calcium oxide, magnesium oxide, alumina, iron oxide, copper oxide, lead oxide, barium oxide and phosphorus pentoxide are set as $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$ in order.

The species is set as the dependent variable y, where y is a qualitative variable, so corresponding dummy variables need to be created[9]. In this paper, the high potassium class is recorded as 1, and the lead barium class is recorded as 0. Consider y as the probability of event occurrence, when $y \geq 0.5$ indicates occurrence, when $y<0.5$ indicates that this parameter does not occur. The independent variable $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$ was used to predict y[10].

Let the connection function be $F(x,\alpha)$, and $F(x,\alpha)$ is the function defined on [0,1], where $F(x,\alpha)$ is the Sigmoid function,then $F(x,\alpha)=S(x_i^T\alpha)$, and then for this nonlinear model, the maximum likelihood estimation method (MLE) is used to estimate, and $\hat{\alpha}$ is obtained[11].

Therefore,$\hat{y_i}=P(y_i=1|x)=S(x_i^T\hat{\alpha})$.If it is $\hat{y_i} \geq 0.5$, it is considered that the predicted y=1, that is, high potassium glass; Otherwise, it is considered that the predicted y=0 is lead barium glass.

Finally, through SPSS, the types of glass relics belonging to unknown categories can be obtained from the final data list, as shown in Table 4 below.

Table 4. Statistical table of the type of glass relics belonging to unknown category

species	Cultural relic number
High potassium type	A1、A6、A7、A8
Lead barium type	A2、A3、A4、A5

4. THE GREY CORRELATION METHOD

4.1 Establishment of Model

At present, taking the weathered glass relics of high potassium type as an example, grey correlation method is used to solve the correlation problem between chemical components[12].Among the high-potassium glass relics, there are 6 weathered glass relics. The chemical composition of glass relics is analyzed by gray correlation analysis, and the dimensionless matrix A of weathered glass relics is obtained[13].

$$A=\begin{bmatrix} 92.63 & 0 & 0 & 1.07 & 0 & 1.98 & 0.17 & 3.24 & 0 & 0 & 0.61 & 0 & 0 & 0 \\ 95.02 & 0 & 0.59 & 0.62 & 0 & 1.32 & 0.32 & 1.55 & 0 & 0 & 0.35 & 0 & 0 & 0 \\ 96.77 & 0 & 0.92 & 0..21 & 0 & 0.81 & 0.26 & 0.84 & 0 & 0 & 0 & 0 & 0 & 0 \\ 94.29 & 0 & 1.01 & 0.72 & 0 & 1.46 & 0.29 & 1.65 & 0 & 0 & 0.15 & 0 & 0 & 0 \\ 92.35 & 0 & 0.74 & 1.66 & 0.64 & 3.5 & 0.35 & 0.55 & 0 & 0 & 0.21 & 0 & 0 & 0 \\ 92.72 & 0 & 0 & 0.94 & 0.54 & 2.51 & 0.2 & 1.54 & 0 & 0 & 0.36 & 0 & 0 & 0 \end{bmatrix}$$

Since most of the data in the seven columns 2, 5, 9, 10, 12, 13 and 14 are 0, that is, the above seven groups of chemical indicators are not detected, so they are deleted, and the other seven groups of values are preprocessed to obtain the following matrix:

$$Z=\begin{bmatrix} 0.99 & 0 & 1.23 & 1.03 & 0.64 & 2.07 & 2.18 \\ 1.01 & 1.09 & 0.71 & 0.68 & 1.21 & 0.99 & 1.25 \\ 1.03 & 1.69 & 0.24 & 0.42 & 0.98 & 0.54 & 0 \\ 1.00 & 1.86 & 0.83 & 0.76 & 1.09 & 1.06 & 0.54 \\ 0.98 & 1.36 & 1.91 & 1.81 & 1.32 & 0.35 & 0.75 \\ 0.99 & 0 & 1.08 & 1.30 & 0.75 & 0.99 & 1.29 \end{bmatrix}$$

Through calculation:

$$\alpha = \min_{1\le s\le m}\min_{1\le t\le n} | a_{1t} - a_{st} | = 0.00064$$

$$\beta = \max_{1\le s\le m}\max_{1\le t\le n} | a_{1t} - a_{st} | = 1.19276$$

The calculated values of ξ_{ij} are as follows:

Table 5. The calculated value of ξ_{ij}

	K_2O	CaO	Al_2O_3	Fe_2O_3	CuO	P_2O_5
07	0.3773	0.7104	0.9380	0.6347	0.3543	0.3337
09	0.8897	0.6671	0.6463	0.7532	0.9706	0.7149
10	0.4739	0.4311	0.4948	0.9254	0.5486	0.3671
12	0.4112	0.7731	0.7079	0.8687	0.9192	0.5610
22	0.6120	0.3924	0.4184	0.6390	0.4866	0.7200
27	0.3771	0.8652	0.6560	0.7207	1.0000	0.6668

The value of grey correlation degree $\omega_{ij} = \frac{1}{n}\sum_{i=1}^{n}\xi_{ij}$ is shown in the following Table 6:

Table 6. Grey correlation degree ω_{ij}

ω_{i3}	ω_{i4}	ω_{i6}	ω_{i7}	ω_{i8}	$\omega_{i,11}$
0.5236	0.6399	0.6436	0.7569	0.7132	0.5606

As a result:$\omega_{i7} > \omega_{i8} > \omega_{i6} > \omega_{i4} > \omega_{i,11} > \omega_{i3}$

Therefore, it shows that among the weathered glass relics of high potassium type, the relationship between silica and iron oxide is the most close, greater than copper oxide, greater than alumina, greater than calcium oxide, greater than phosphorus pentoxide, greater than potassium oxide.

Matlab is now used to evaluate the grey correlation degree of $\omega_{i7}, \omega_{i8}, \omega_{i6}, \omega_{i4}, \omega_{i,11}, \omega_{i1}, \omega_{i3}$, and the obtained data are shown in the following Table 7:

Table 7. Grey correlation matrix of weathered cultural relics with high potassium

	ω_{i1}	ω_{i3}	ω_{i4}	ω_{i6}	ω_{i7}	ω_{i8}	$\omega_{i,11}$
ω_{i1}	1	0.5236	0.6399	0.6436	0.7569	0.7132	0.5606
ω_{i3}	0.6761	1	0.5751	0.5823	0.7395	0.5830	0.5439
ω_{i4}	0.7164	0.4922	1	0.8913	0.6416	0.6803	0.6438
ω_{i6}	0.6969	0.4765	0.8694	1	0.6218	0.6318	0.6363
ω_{i7}	0.8366	0.6789	0.6462	0.6519	1	0.6870	0.6058
ω_{i8}	0.7873	0.5515	0.7091	0.6877	0.7085	1	0.7560
$\omega_{i,11}$	0.6959	0.5312	0.6990	0.7078	0.6576	0.7745	1

4.2 Result analysis

It can be seen from the data in the table that the substance most closely related to the content of chemical composition silica is iron oxide, followed by copper oxide, and the correlation degree of the two indicators is more than 0.7000, which has a strong correlation[14]. Both alumina and calcium oxide are more than 0.6000, which indicates that the content of alumina and calcium oxide is related to silicon dioxide. Since the values of potassium oxide and copper oxide are between 0.5-0.6, it can be preliminarily determined that the relationship between potassium oxide and copper oxide and silicon dioxide content is not significant[15].

The substance most closely related to the content of potassium oxide is iron oxide, which has a strong correlation, and the other five indicators have a smaller correlation with potassium oxide than iron oxide.

The substance associated with calcium oxide is alumina, and the correlation between the two is strong, while the potassium oxide content is less than 0.5, indicating that the correlation between the two is small.

The substance with the strongest association with alumina is calcium oxide, although the correlation between the two is the strongest, but the correlation distribution of alumina is more dispersed than that of calcium oxide.

The strongest correlation with iron oxide is silica, reaching 0.8366, which is higher than the gray correlation of iron oxide to silicon dioxide of 0.7569.

The substance with the highest correlation degree with copper oxide is silicon dioxide, with an index of 0.7873, slightly lower than the gray correlation coefficient of 0.8366 of silicon dioxide compared with iron oxide. In the calculation of the gray correlation coefficient with silicon dioxide as the reference series, the obtained results are still highly logical and authentic, and conform to the cognitive law.

The gray correlation coefficient of phosphorus pentoxide as reference series shows that copper oxide has the greatest correlation with it. In the gray coefficient of phosphorus pentoxide as reference series, the two values are close to each other, only lower than the content of silicon dioxide.

5.CONCLUSION

This paper focuses on the application of four statistical methods in data analysis and identification, namely, systematic clustering algorithm, K-means algorithm, logistic regression model and grey correlation analysis.

Taking the C project of CUMCM in 2022 as an example, by constructing the above four models, we obtained the subclass classification results, and finally divided high-potassium glass into three categories and lead-barium glass into five categories. By comparing the result and the actual situation, it is found that the two agree with each other, which indicates that the result is reasonable.Unknown categories of glass artifacts are also classified. In this paper, A1, A6, A7 and A8 are classified as high potassium glass, and A2, A3, A4 and A5 are classified as lead barium glass. And the correlation analysis of chemical constituents in the group was also carried out through the correlation matrix data.

Thus, the application value and practical significance of statistics in the chemical composition and identification of ancient cultural relics are realized. Statistics plays an important role in all fields of society, facilitating people to analyze, process and apply all kinds of data, and promoting social progress and development.

REFERENCES

[1] Deng X,Zheng K,Xiong Y. Cluster Analysis Based on Indicator System on the Development of Digital Economy in Guangdong[J]. IAENG International Journal of Applied Mathematics,2021,51.0(3.0).

[2] Yue Z,Claire B,Maxime V, et al. Performance comparisons between clustering models for reconstructing NGS results from technical replicates[J]. Frontiers in Genetics,2023,14.

[3] Singh A,Koju R. Healthcare Vulnerability Mapping Using K-means ++ Algorithm and Entropy Method: A Case Study of Ratnanagar Municipality[J]. International Journal of Intelligent Systems and Applications(IJISA),2023,15(2).

[4] Ping L,Hilal B A. Automatic Knowledge Integration Method of English Translation Corpus Based on Kmeans Algorithm[J]. Applied Mathematics and Nonlinear Sciences,2023,8(1).

[5] Tao Y,Yang J,Chang H. Enhanced iterative projection for subclass discriminant analysis under EM-alike framework[J]. Pattern Recognition,2014,47(3).

[6] Pang Z. Composition identification of ancient glass products based on cluster analysis[J]. Academic Journal of Computing & Information Science,2022,5(13).

[7] Ma K,Hu C C,Feng Q S, et al. Application and Research of Systematic Cluster in IED Switch Online Condition Monitoring[J]. Applied Mechanics and Materials,2014,3546(672-674).

[8] Yuheng G,Wei Z,Weihao L. Application of Support Vector Machine Algorithm Incorporating Slime Mould Algorithm Strategy in Ancient Glass Classification[J]. Applied Sciences,2023,13(6).

[9] F. A L,Golam M B K,K. C N, et al. K-L Estimator: Dealing with Multicollinearity in the Logistic Regression Model[J]. Mathematics,2023,11(2).

[10] Sofia R,E Y S K,John W, et al. Insight into Faculty Job Satisfaction at a Southern California Dental School Using a Logistic Regression Model.[J]. Journal of dental education,2022.

[11] József D,Tamás J. Generalizing the sigmoid function using continuous-valued logic[J]. Fuzzy Sets and Systems,2022,449.
[12] Xing W,Aihong K,Said E, et al. Performance Evaluation of Reinforced Asphalt Using Six Organic and Inorganic Fibers[J]. Journal of Materials in Civil Engineering,2023,35(6).
[13] Bing D,Danli L,Lei Z, et al. Rock Mass Classification Method Based on Entropy Weight–TOPSIS–Grey Correlation Analysis[J]. Sustainability,2022,14(17).
[14] Liangshan L. Measurement of the Correlation Degree between Rural Family Fertility Willingness and the Development of China's Labor Original Equipment Manufacturing Industry[J]. Computational Intelligence and Neuroscience,2022,2022.
[15] Hu Y,Ying Q,Yang L, et al. Research on the Positive Correlation between the Degree of Self-Organization and the Effectiveness of Social Work Intervention in Elderly Care Institutions[J]. Open Journal of Social Sciences,2019,07(11).