Glass Classification Accuracy Based on Spearman Coefficient and K-Means ++ Algorithm

Zhenzhen Wang[1], Zecheng Meng[2,*], Zeyang Jin[3]

[1]School of Management, Hebei University, Baoding, China, 071000

[2]Institute of Electrical Information, Qing Gong College, North China University of Science and Technology, Tangshan, China, 063020

[3]The Department of Business Administration, Hebei Vocational University of Industry and Technology, Shijiazhuang, China, 050091

* Corresponding Author Email: meng4568522024@163.com

Abstract. The study of the physical and chemical properties of artifactual glass can contribute greatly to the attribution of artifacts and to the study of their history. In this paper, we study the interplay of physical and chemical properties of glass and classify them. In order to make the classification more accurate, the correlation coefficients of compounds were calculated using Spearman's coefficient, and the content of lead oxide was determined as the basis for the classification. After that, this paper used K-means++ to cluster the data and selected calcium oxide and alumina as the sub-classification basis for high-potassium glass, and lead oxide and barium oxide as the sub-classification basis for lead-barium glass. Finally, the contents of some artifacts calcium oxide, alumina, lead oxide and barium oxide were changed with a perturbation of 0.2%, and the clustering points did not move significantly, and the model was stable.

Keywords: Glass, Chemical Composition, Classification, K-means++, Spearman Coefficient.

1. Introduction

Silk Road is a comprehensive trade road in ancient China, and it is a channel for political and cultural exchanges between China and Eurasian countries, and a friendship branch road for East-West exchanges[1-3]. As the main raw material of glass, quartz sand has a high melting point, so it is necessary to add fluxes when making glass, and the common fluxes used in ancient China are grass wood ash, natural bubble soda, saltpeter, etc[4]. Fluxes can have an impact on the chemical composition of glass, such as lead barium glass flux is lead ore, which is usually considered to be a variety of glass in China[5]; potassium glass flux is grass wood ash, which is mainly popular in China's Lingnan, Southeast Asia and India and other regions. Ancient glass is susceptible to weathering by the burial environment, causing changes in the proportion of its chemical composition. Therefore, predicting and analyzing the chemical composition content of glass has an important influence on its classification[6-8].

A batch of ancient glass data is available, and archaeologists have classified them into two categories: high potassium glass and lead-barium glass, and have obtained information on the classification of the glass and the proportion of the corresponding main chemical composition, of which the cumulative proportion of chemical composition between 85% and 105% is valid data. Based on the above background and information, we need to establish a mathematical model to analyze the classification pattern of high potassium glass and lead-barium glass and to classify them into subclasses, and to derive the classification methods and results, and to analyze the reasonableness and sensitivity of the classification results.

2. Problem Analysis

This study can be divided into two steps, the first one is to analyze the classification law of high potassium glass and lead-barium glass i.e., the classification is related to

What factors are related to the classification; the second is to select the appropriate chemical composition for each category for subclassification, and give the classification method and classification results. The technical roadmap is shown in Fig. 1.

For the first problem, firstly, the necessity of silica and lead oxide for classification is obtained by analyzing the data. After that, the K-means++ algorithm is used to verify the classification prediction of the sample data using the proportion of lead oxide content, and the correct rate is 100%, which proves that lead oxide is the main factor affecting high potassium glass and lead-barium glass.

For the second problem, the influence of each chemical element on glass subclassification was first considered. The role of alumina in glass is to improve the stability of glass, and the role of calcium oxide is to enhance the hardness of glass, but the high content of calcium oxide reduces the thermal stability of glass, so alumina and calcium oxide, which have a greater influence, were considered, and the data were clustered using K-means++ to obtain the classification law and perform sensitivity analysis to test the stability of the model.

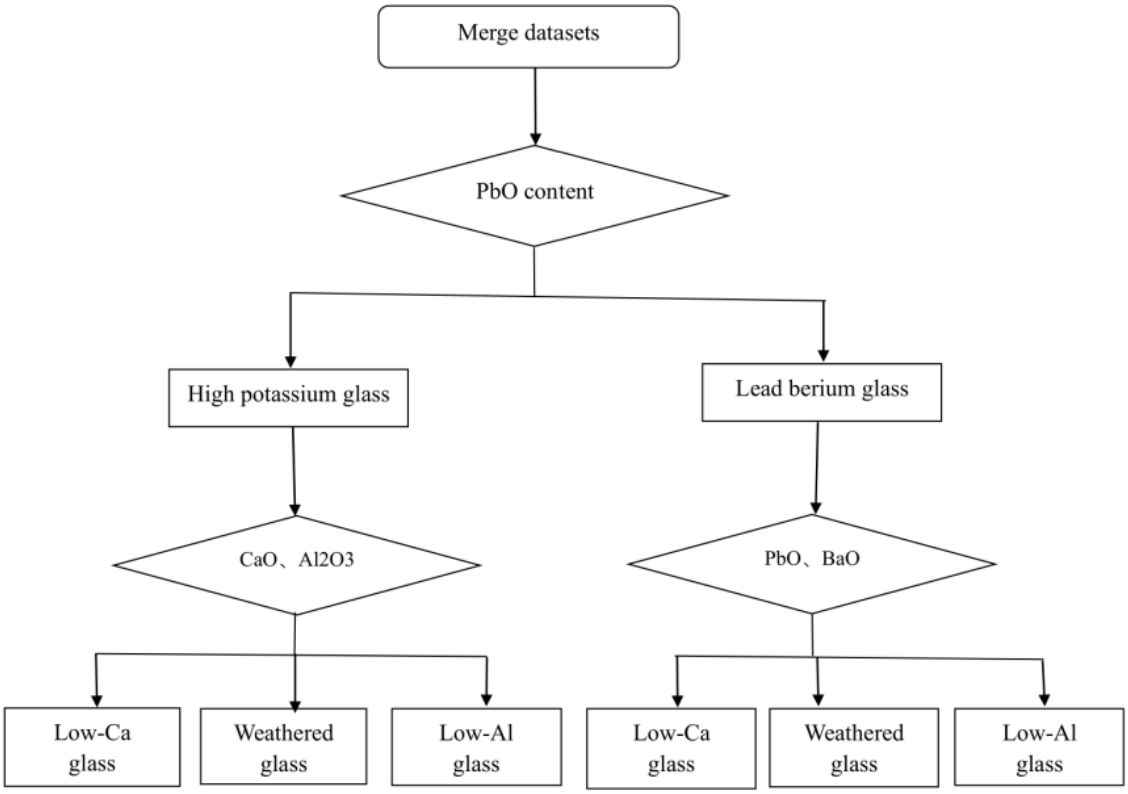


Figure 1. Technology Roadmap

3. Classification modeling of glass types

In this study, the data were quantified and tested for normality, followed by an analysis of the correlation between the properties and types of chemicals using the Spearman coefficient, from which the classification could be obtained; for subclassification, we clustered the data using K-means++, and then obtained classification laws and performed model sensitivity analysis.

3.1. Normality test and data preprocessing

The data for this study comes from market research. We quantified the data for high potassium glass and lead-barium glass, so that high potassium glass was 1 and lead-barium glass was 0. The normality test was performed for the fourteen compounds, as follows in Table 1. Our sample size is less than 5000, so the S-W test is performed, and if the p-value is greater than 0.05 means that the normal distribution is met. Combined with the above table, for all the above indicators do not conform to the normal distribution.

Table 1. Normality test

Variable Name	Num	Median	Mean	Standard deviation	Bias	Peakiness	S-W test (p-value)
SiO2	69	51.54	49.378	24.046	0.142	-0.794	0.963(0.039**)
Na2O	69	0	0.84	1.662	2.288	5.381	0.584(0.000***)
K2O	69	0.21	1.984	3.907	1.956	2.43	0.562(0.000***)
CaO	69	1.6	2.459	2.33	1.059	0.262	0.874(0.000***)
P2O5	69	0.94	2.613	3.516	1.521	1.525	0.754(0.000***)
SrO	69	0.22	0.254	0.267	1.082	0.856	0.866(0.000***)
SnO2	69	0	0.076	0.333	5.735	35.763	0.244(0.000***)
SO2	69	0	0.586	2.66	5.367	28.76	0.229(0.000***)
MgO	69	0.66	0.69	0.642	0.588	-0.07	0.888(0.000***)
Al2O3	69	3.33	3.969	3.062	1.621	2.839	0.850(0.000***)
Fe2O3	69	0.33	0.855	1.162	2.1	5.728	0.748(0.000***)
CuO	69	1.09	1.932	2.213	2.225	5.722	0.746(0.000***)
PbO	69	22.05	23.76	19.657	0.303	-0.965	0.921(0.000***)
BaO	69	6.22	7.554	8.403	1.68	2.82	0.800(0.000***)

Note: ***, **, and * represent 1%, 5%, and 10% significance levels, respectively.

3.2. Model building

3.2.1 Analysis and classification of indicators

According to the data preprocessing, we learned that the data did not conform to normal distribution, so we used the Spearman coefficient for the correlation between the nature and type of chemicals to analyze.

(1) Visual analysis of the data

The analysis of the data revealed the difference of lead oxide content in lead-barium glass and high-potassium glass. From the above, it can be seen that lead barium glass contains a higher amount of lead oxide than high potassium glass when it is unweathered, and the amount of lead oxide increases further after weathering. High potassium glass contains almost no lead oxide.

(2) Spearman's coefficient

Then we perform the solution of Spearman's correlation coefficient, which is defined as the Pearson's correlation coefficient between rank variables(Figure 2). For a sample of sample size n, n raw data were converted into rank data, and the correlation coefficient ρ

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

First quantify the type as 1 for high potassium glass and 2 for lead-barium glass. The corresponding Pearson correlation plot thus obtained is 2.

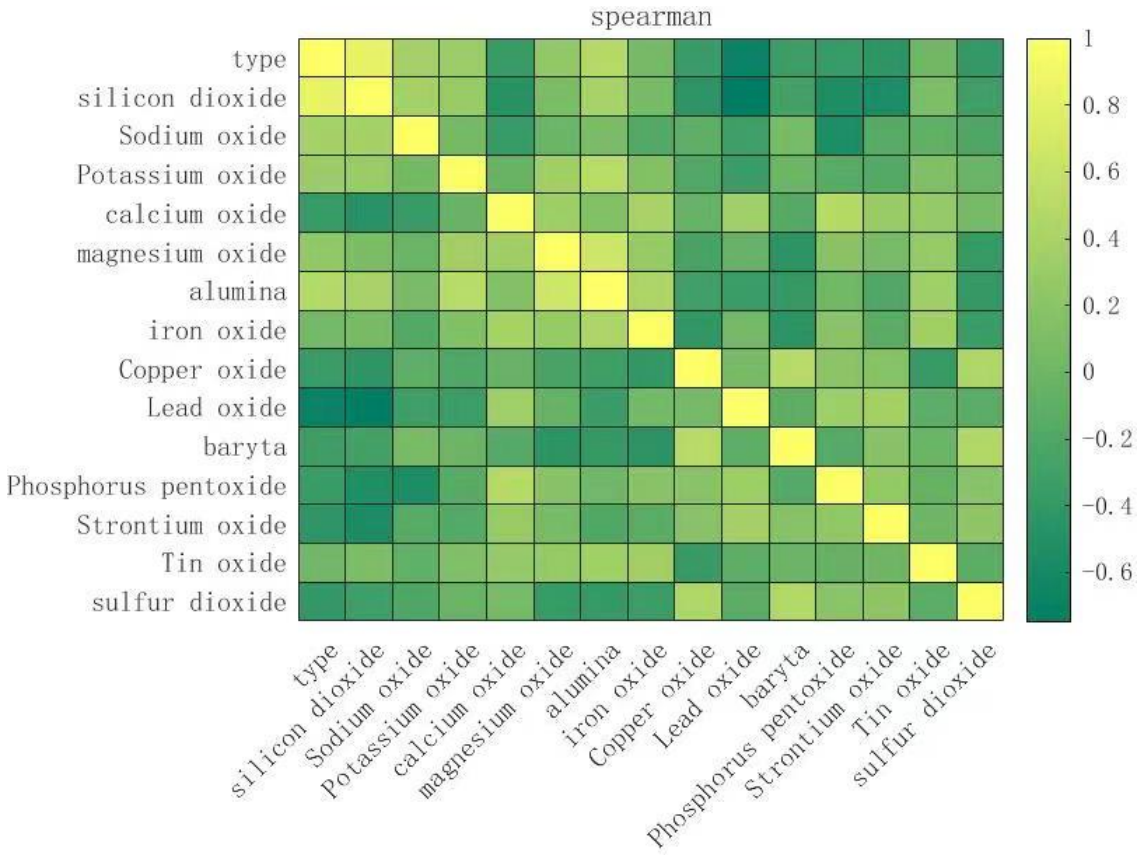


Figure 2. Pearson correlation diagram

According to the Spearman correlation coefficient, we can also obtain that lead oxide has the greatest effect on the type, followed by potassium oxide, and after that by silica.

(3) K-means algorithm

Then we use the K-means++ algorithm to classify the data according to the content of lead oxide, and the K-means++ is calculated as follows

a) Randomly select a point from the input set of data points as the first clustering center μ1

b) For each point xi in the data set, calculate its distance from the nearest cluster center among the cluster centers that have been selected: $D(x_i) = \arg\min \square x_i - \mu_r \square_2^2 \; r = 1, 2, \ldots k_{\text{selected?}}$.

c) Select a new data point as the new cluster center, the principle of selection is: the point with larger D(x) has a higher probability of being selected as the cluster center

d) Repeat b and c until k clustering centers of mass are selected

Using K-means++, combined with the information found above, so that K = 2, the classification points obtained.

Lead oxide content below 10% is recognized as high potassium glass; lead oxide content above 10% is recognized as lead barium glass.

3.2.2 Sub-classification of high potassium glass and lead-barium glass

(1) Sub-classification of high potassium glass

For high-potassium glasses, based on the concentration of CaO and Al2O3 in potassium glasses, previous studies further classified ancient potassium glasses found throughout Asia into three subclasses: medium calcium-aluminum (m-K-Ca-Al), low calcium (m-K-Al), and low aluminum (m-K-Ca), with m indicating that the K2O in the sample is from potassium-rich minerals [9, 10]. These three subclasses of potassium glasses all use K2O as a flux, but the content of CaO and Al2O3 differs somewhat [11]. Although the trace elements Sr and Rb also have some influence on the three glass subclasses, the trace elements Sr and Rb were too little or even undetected in the samples due to the lack of detection technology, so we do not consider the influence of Sr and Rb on the classification of high potassium glasses.

For high potassium glasses, the compound content changes before and after weathering, so the clustering results include the distribution before and after weathering.

Since the data range is relatively small, the data can be clustered using K-means++. By clustering the data with K-means++, the division points of medium chalcogenide potassium glass, low chalcogenide potassium glass and low aluminum low chalcogenide potassium glass can be obtained, and then the classification laws can be obtained. Using K-means++, combined with the information checked above, so that K=3, the classification points obtained are as follows:

The clustering centroids are as follows Table 2:

Table 2. Clustering centroids

Meeting Point	Calcium oxide	Aluminum oxide
Cluster I (low calcium)	0	11.15
Cluster II (weathering)	6.8867	6.7978
Cluster III (low aluminum and low calcium)	0.9038	2.3363

The results after clustering are shown in Figure 3. As can be seen from the figure below, there is one point of clustering, which does not exclude the error caused by the relatively small content of compounds and the imprecision of the measurement equipment.

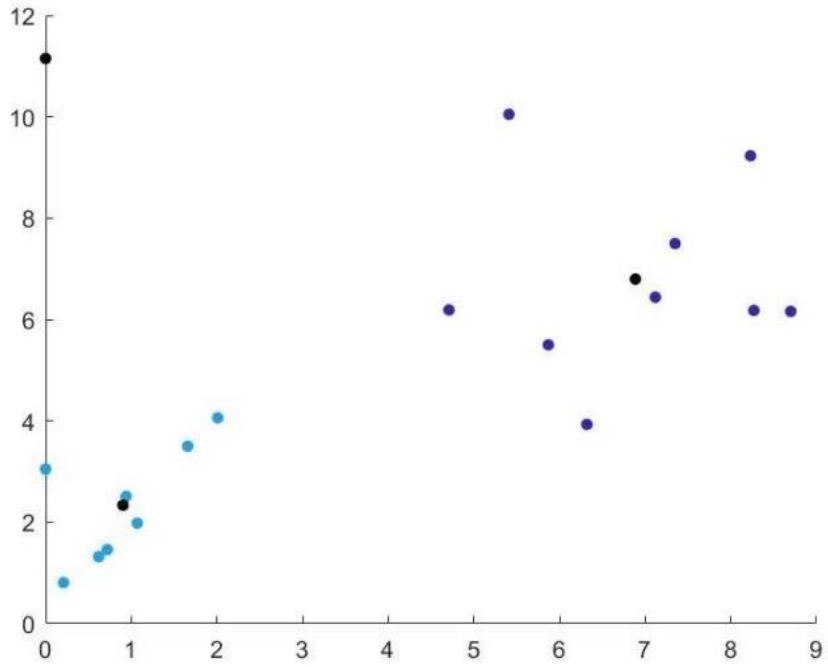


Figure 3. Clustering results

High potassium glass sensitivity analysis. The content of calcium oxide of some artifacts was changed randomly (the amount of change was around 0.1) and the change of clustering points was observed as shown in Table 3.

Table 3. Changes in clustering points

Clustering type	Calcium oxide (partially changed)	Aluminum oxide (unchanged)
Clustering I	6.875555555555556	6.797777777777777
Cluster II	0.92875	2.3362499999999997
Clustering III	0	11.15

It can be obtained that the change of clustering points is relatively small, so the model classification meets the basic requirements.

The content of some artifacts alumina was changed randomly (the change amount was around 0.2), and the change of clustering points was observed as shown in Table 4.

Table 4. Changes in clustering points

Clustering type	Central value Aluminum oxide (partially changed)	Central value Calcium oxide (unchanged)
Clustering I	6.82	6.886666666666667
Cluster II	2.34875	0.90375
Clustering III	11.15	0

The clustering points can be obtained with the same small change, so it meets the classification requirements.

The role of alumina in glass is to improve the stability of glass, the role of glass species calcium oxide is to enhance the hardness of glass For lead barium glass, lead barium glass combustion aid, in the classification of high potassium glass, this measure of glass stability. For the weathered class of glass, calcium and aluminum content is relatively balanced more stable, for low aluminum and low calcium as well as low aluminum and high calcium type glass, low aluminum and low calcium glass stability is the worst, low calcium and high aluminum glass stability is moderate.

(2) Lead and barium glass

Lead-barium glass is a unique glass in ancient China, which mainly appeared in the Chu period. And ancient artisans may control the acidity of glass by adjusting the ratio of lead-barium-silica content in glass to get the expected glass color [12], which is based on the principle that copper ions only appear blue under acidic conditions. Silica is the main acidic compound in lead-barium glass, and lead oxide and barium oxide are the main basic compounds, so in studying this part, we use the sum of silica and lead oxide and barium oxide as the main classification basis using K-means++ to classify the clustering categories obtained by K-means++, for which P is the indicator to test whether the clustering is valid, and when $P< 0.05$, it means that the classification is valid.

For the following table, we get that the P of both categories divided is less than 0.05, then the classification is valid. Lead oxide and barium oxide can not only be used as fluxes for glass, but also can increase the refractive index density, which is a measure of the transparency of glass.

The clustering categories are as follows, and the clustering points are divided as follows in Table 5 and Figure 4.

Table 5. Clustering categories

	Clustering categories (mean ± standard deviation)		P
	Type II(n=28)	Type I(n=21)	
Unnamed: 21	55.825±8.551	27.859±7.145	0.000***
Q7- SiO2	24.915±9.204	57.49±9.133	0.000***

Note: ***, **, and * represent 1%, 5%, and 10% significance levels, respectively.

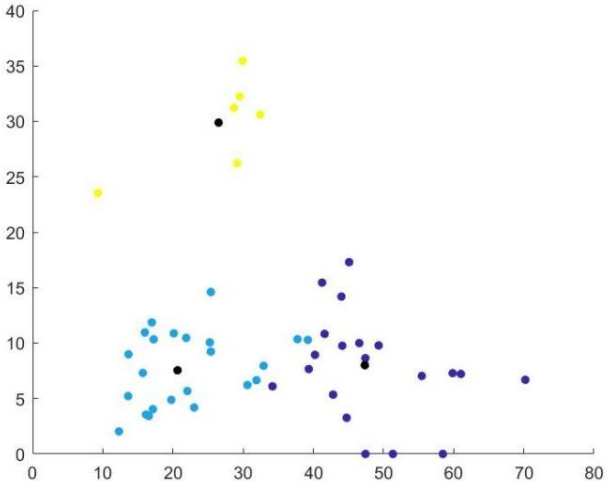


Figure 4. Clustering result

Note: Black dots are cluster centers, blue dots are low transparency clusters, purple dots are high lead transparency clusters, and yellow dots are high barium transparency clusters.

Sensitivity analysis of lead barium glass. The content of barium oxide of lead barium glass was changed randomly within 0.3, and K-means++ cluster analysis was performed, and the clustering points were obtained as shown in Table 6 below. It can be seen that the change of clustering points is small and the model is more stable.

Table 6. Clustering categories

Clustering type	Central Value - Lead Oxide	Central value - Barium oxide (change)
1	47.338181818181816	8.011818181818182
2	20.134090909090908	8.263181818181817
3	29.944000000000003	31.136000000000003

The content of lead oxide of lead barium glass was changed randomly within 0.3, and K-means++ cluster analysis was performed, and the clustering points were obtained as shown in Table 7.

Table 7. Clustering categories

Clustering type	Central Value - Lead Oxide	Central value - Barium oxide (change)
1	47.36181818181818	8.007272727272728
2	20.138636363636365	8.276818181818182
3	29.904	31.156000000000002

It can be seen that the clustering points change less and the model is more stable.

4. Conclusion

The study of the physical and chemical properties of artifactual glass can contribute greatly to the attribution of artifacts and to the study of their history. In this paper, we study the interplay of physical and chemical properties of glass and classify them. In order to make the classification more accurate, the correlation coefficients of compounds were calculated using Spearman's coefficient, and the content of lead oxide was determined as the basis for the classification. After that, this paper used K-means++ to cluster the data and selected calcium oxide and alumina as the sub-classification basis for high-potassium glass, and lead oxide and barium oxide as the sub-classification basis for lead-barium glass. Finally, the contents of some artifacts calcium oxide, alumina, lead oxide and barium oxide were changed with a perturbation of 0.2%, and the clustering points did not move significantly, and the model was stable.

Reference

[1] Wu Chi,Fu Xingyao,Shi Jinlong,Liu Quanwei. Research on Chemical Composition and Weathering of Ancient Glass Based on Grey Correlation[J]. Journal of Physics: Conference Series,2023,2508(1).

[2] Jiang Yicheng,Wang Yu,Chen Shuai,Liu Zhixin,Yang Haojun,Jiao Yuwen,Tang Liming. Screening of Biomarkers in Liver Tissue after Bariatric Surgery Based on WGCNA and SVM-RFE Algorithms[J]. Disease Markers,2023,2023.

[3] Zhu Manqin,Hu Shengbo,Yang Kai,Yan Tingting,Ye Lv,Li Hongjuan,Jin Yang. GMSK Demodulation Combining 1D-CNN and Bi-LSTM Network Over Strong Solar Wind Turbulence Channel[J]. Radio Science,2023,58(1).

[4] Zaheer Shahzad,Anjum Nadeem,Hussain Saddam,Algarni Abeer D.,Iqbal Jawaid,Bourouis Sami,Ullah Syed Sajid. A Multi Parameter Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model[J]. Mathematics,2023,11(3).

[5] Han Dongyang,Liu Pan,Xie Kang,Li He,Xia Qian,Cheng Qian,Wang Yibo,Yang Zhikai,Zhang Yanjun,Xia Jun. An attention-based LSTM model for long-term runoff forecasting and factor recognition[J]. Environmental Research Letters,2023,18(2).

[6] Vilarinho R,Weber M C,Guennou M,Miranda A C,Dias C,Tavares P,Kreisel J,Almeida A,Moreira J Agostinho. Author Correction: Magnetostructural coupling in RFeO3 (R = Nd, Tb, Eu and Gd).[J]. Scientific reports,2023,13(1).

[7] Zou Ying. Molecular-Composition Analysis of Glass Chemical Composition Based on Time-Series and Clustering Methods[J]. Molecules,2023,28(2).

[8] Singh Sanjeet,Bansal Pooja,Hosen Mosharrof,Bansal Sanjeev K.. Forecasting annual natural gas consumption in USA: Application of machine learning techniques- ANN and SVM[J]. Resources Policy,2023,80.

[9] Hu Junying,Qian Xiaofei,Tan Changchun,Liu Xinbao. Point and interval prediction of aircraft engine maintenance cost by bootstrapped SVR and improved RFE[J]. The Journal of Supercomputing,2022,79(7).

[10] ThenObłuska Joanna,Dussubieux Laure. Beads from a mediaeval pilgrim centre: Chemical composition and provenance of glass from Banganarti, Nubia, Sudan[J]. Archaeometry,2022,65(2).

[11] Zlámalová Cílová Zuzana,Blažková Gabriela. Chemical composition of early modern and modern window glass from archaeological excavations in Prague-Hradčany, Czech Republic[J]. Archaeometry,2022,64(6).

[12] Tkaczewska Ewelina. Effect of Chemical Composition and Network of Fly Ash Glass on the Hydration Process and Properties of Portland-Fly Ash Cement[J]. Journal of Materials Engineering and Performance,2021,30(12).