Glass classification and identification based on Lasso-logistic model and K-means clustering

Hesheng Chen [1,#,*], Wenya Fan [1,#], Lanxin Zhang [2,#]

[1] College of Safety Science and Engineering, Civil Aviation University of China, Tianjin, China, 300000

[2.]College of Electronic Information and Automation, Civil Aviation University of China, Tianjin, 300000

* Corresponding Author Email: cauchschen@163.com

[#]These authors contributed equally.

Abstract. Ancient glass is highly susceptible to weathering by the burial environment, with a large number of internal elements exchanging with environmental elements, resulting in changes in the proportions of their composition. In order to analyse the classification patterns of High potassium-rich glass and Lead and barium enriched glass, and to classify each type of glass into appropriate subcategories, Decision tree was first chosen to broadly classify the decision factors for the glass types, and then lasso-logistic was chosen for the exact solution. K-means clustering was used in combination with Factor Analysis Method to achieve subcategories for both types of glass. Finally, using the above model, the chemical composition of the unknown glass is used to identify the type.

Keywords: Decision tree, Lasso-logistic, Cluster analysis, Fisher discriminant analysis.

1. Introduction

The Silk Road was a gateway for cultural exchange between China and the West in ancient times, and glass was a valuable physical evidence of this early trade. The main chemical component of glass is silica (SiO2) and is highly susceptible to weathering by the environment in which it is buried. During the weathering process, a large number of internal elements are exchanged with environmental elements, resulting in a change in the proportion of its composition [1], which affects the correct determination of its category. To address this issue, the classification rules for High potassium-rich glass and Lead and barium enriched glass were analysed, appropriate subclasses were assigned to each type of glass, the division methods and the results were described, and the types of unknown glass were identified. It is therefore of great theoretical value and importance to carry out research into the composition of glass. [2]

2. Glass classification model building and solving

This paper will complete a summary of the classification laws for High potassium-rich glass and Lead and barium enriched glass, as the relevant data and the type to which they belong are already available. So it is reasonable to use a supervised learning approach to classification. As the criteria for determining the type of glass will not change whether it is weathered or not, the paper will discuss before and after weathering together.

2.1. The decision tree

Using SPSSAU to classify the decision tree on the pre-processed data(CLR)[3][4][5], 80% was selected as the training set and 20% as the test set, and the samples were subjected to decision tree classification[6], and the results obtained were 42.21% of the feature weights accounted for by lead oxide (PbO), 19.89% by barium oxide (BaO); 15.29% by potassium oxide (K_2O); 13.43% by sulphur dioxide (SO_2) The weight of the above four features together accounted for 90.82%, and the confusion matrix of the test set is shown in Figure 1.

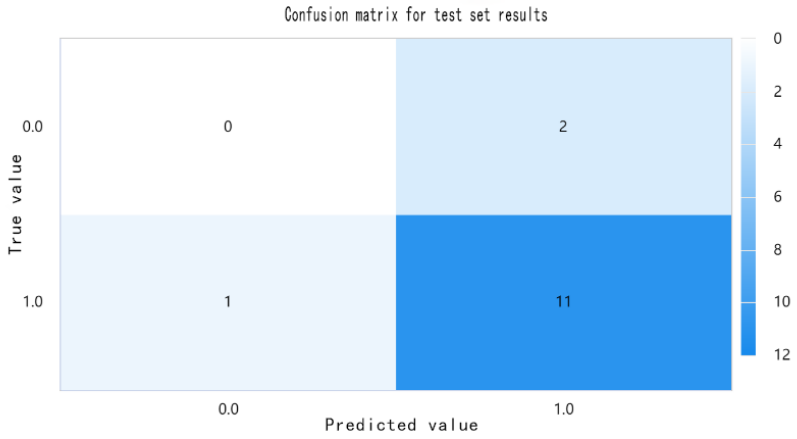


Figure 1. Confusion matrix

The accuracy, recall and F1-score (a combined precision and recall evaluation metric) for the training and test sets were calculated for model evaluation and the results are shown in Table 1.

Table 1. Model evaluation

	Accuracy	Recall	F1-score	Number of samples
Training set mean	1.00	1.00	1.00	53
Test set average	0.73	0.79	0.75	14

Analysis of the above table shows that the decision tree model obtained an accuracy (combined) of 72.53%, a recall (combined) of 78.57% and an F1-score (combined) of 0.75 on the test set. the overall effectiveness of the model is poor, but it can be roughly determined that the main decisive chemical components for glass type are lead oxide (PbO), barium oxide (BaO) and potassium oxide (K_2O).

2.2. Lasso-logistic

(1) Introduction to Lasso regression optimization

Lasso regression, by generating a penalty function is a regression model in the variable coefficients for compression, to prevent over-fitting, to solve the problem of severe covariance, in the Newgran literature, some experts proposed that for too many variables and the number of variables is small model fitting, first consider using the Lasso penalty function. lasso regression introduced a penalty term in the model, in the log-likelihood function L1 parametric is added to the model, and the L1 regularized logistic regression model is obtained, with the following equation [7].

$$\beta = arg\min_{\beta}\left\{\sum_{i=1}^{n}\left[-y_i(x_i\beta^T) + \ln\left(1 + exp(x_i\beta^T)\right)\right] + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (1)$$

(2) Lasso regression optimization solving process [8]

In this paper, we use Stata 17 for logistic regression based on Lasso optimization, meanwhile, logistic regression combined with the method based on K-folding cross-validation (splitting all samples into K groups of equal size, taking one of them at a time as the test set and the others as the training set, and finally evaluating the model based on the test results), using the data multiple times to further prevent overfitting, and the results after filtering variables are shown in Table 2.

Table 2. Variable screening table

Variate	Active
K_2O	-0.3565986
Fe_2O_3	-0.273008
PbO	0.4920091
BaO	0.5348074
SnO_2	0.151927
SO_2	-0.0318351
_cons	1.199952

Analysis of the above table shows that for the decision variables of glass type, the main ones are potassium oxide(K_2O), iron oxide(Fe_2O_3), lead oxide (PbO), barium oxide (BaO), tin oxide (SnO_2), and sulfur dioxide (SO_2).

Meanwhile, the cross-check diagram was drawn to determine the parameter values of the penalty function, as in Figure 2.

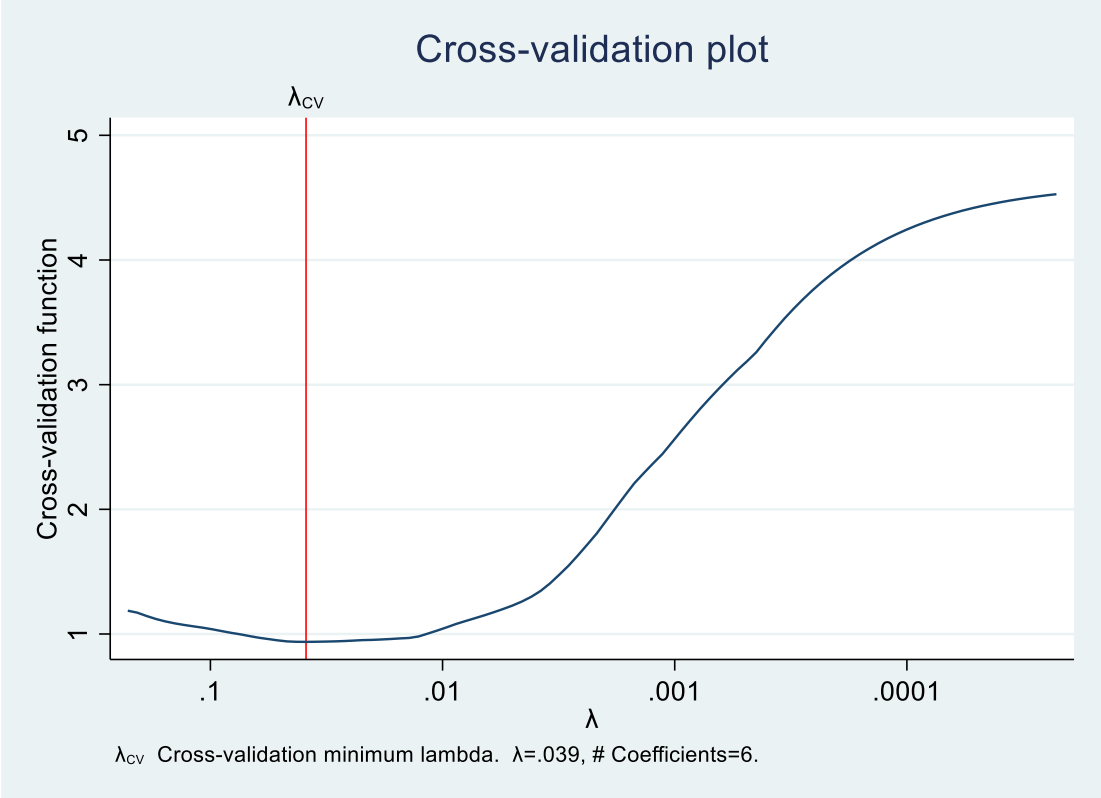


Figure 2. Cross-check diagram

The optimal values of the regulation parameters are given in the above figure. From the above figure, it can be seen that near the optimal value λ = 0.039, the vertical coordinate is small and the function is relatively flat, which means that it is stable for the predictive ability of the model when it is varied near the optimal value. Meanwhile, using the Coefpath function, the coefficient path of Lasso regression is plotted on a logarithmic scale, and λ=0.039 is chosen as the baseline, as shown in Figure 3.

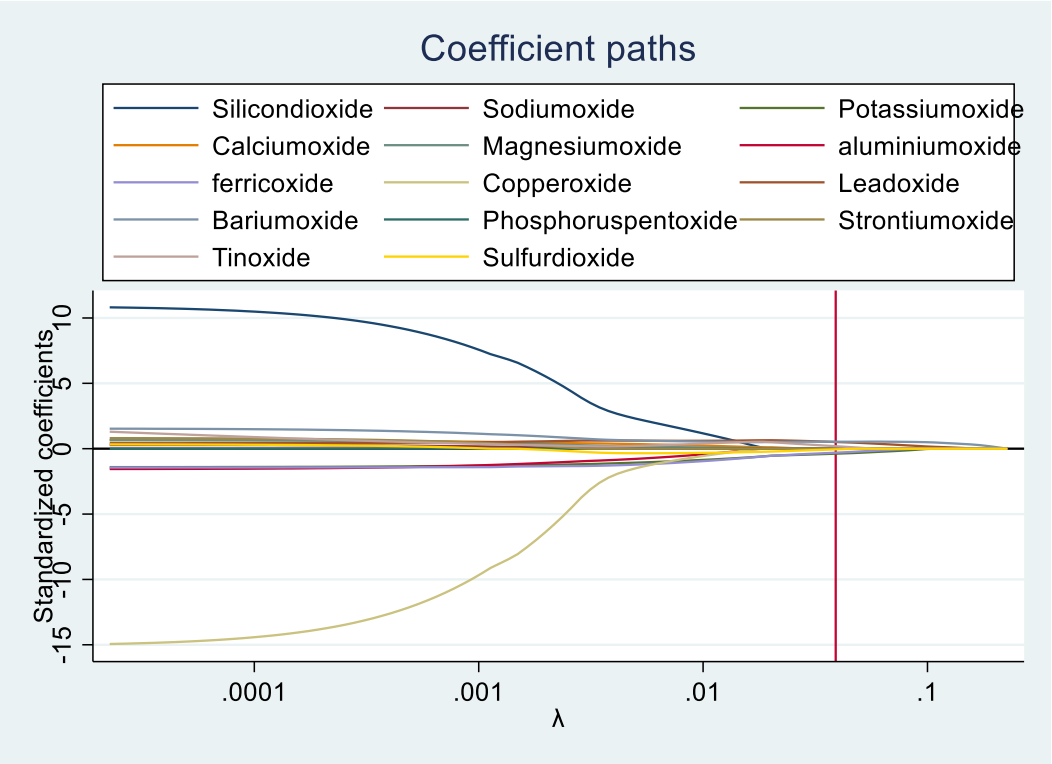


Figure 3. Coefficient path diagram

The number of points where the baseline line intersects is the number of variables incorporated in the final model, and the corresponding vertical coordinate is the regression coefficient of that variable.

(3) Cross-testing of the model

The binary outcome model was cross-validated using Stata's Kfoldclass command [9], and ROC plots were drawn using the full data and test data, as shown in Figure 4.

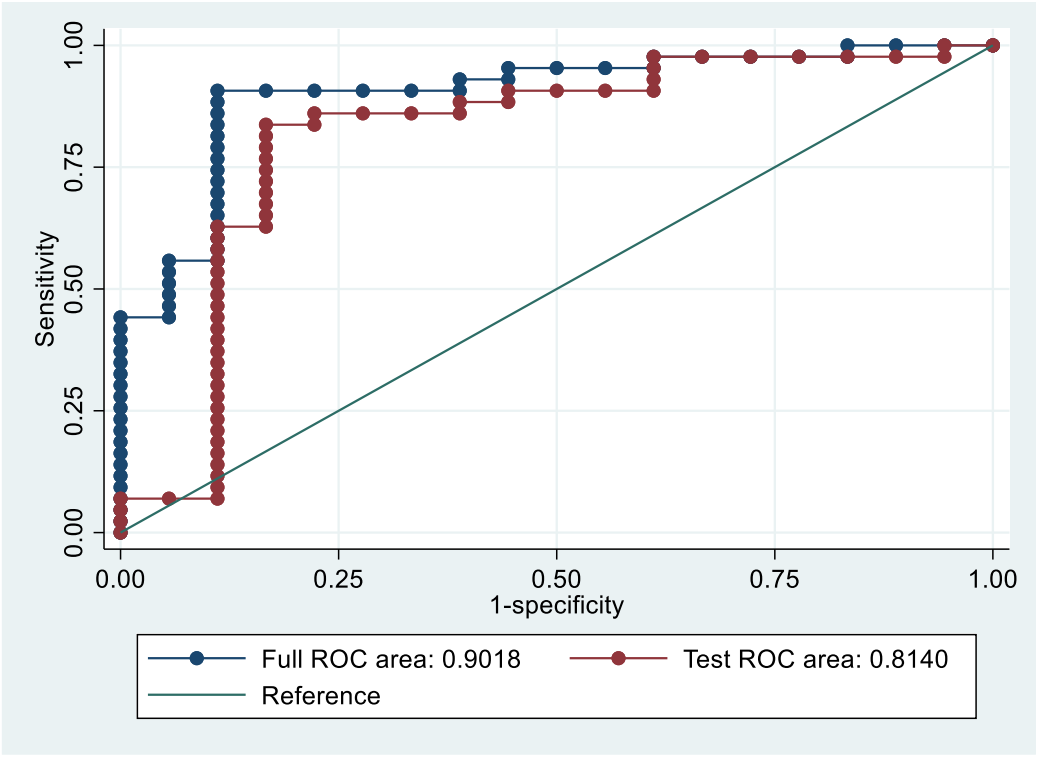


Figure 4. ROC Plot

According to the calculated AUC value (i.e., the area under the curve from 0 to 1), the AUC of the test data set is 0.8140, which is close to the AUC value of 0.9018 of the full data set, both of which are close to 1. The larger AUC value of the model indicates better prediction.

The evaluation results of the model's accuracy, recall and precision are shown in Figure 5.

Classified + if predicted Pr(D) >= .5
True D defined as != 0

		Full	Test
Sensitivity	Pr(+| D)	**91.84%**	**89.80%**
Specificity	Pr(-|~D)	**61.11%**	**66.67%**
Positive predictive value	Pr(D| +)	**86.54%**	**88.00%**
Negative predictive value	Pr(~D| -)	**73.33%**	**70.59%**
False + rate for true ~D	Pr(+|~D)	**38.89%**	**33.33%**
False - rate for true D	Pr(-| D)	**8.16%**	**10.20%**
False + rate for classified +	Pr(~D| +)	**13.46%**	**12.00%**
False - rate for classified -	Pr(D| -)	**26.67%**	**29.41%**
Correctly classified		**83.58%**	**83.58%**
ROC area		**0.9018**	**0.8036**
p-value for Full vs Test ROC areas			**0.0061**

Figure 5. Evaluation results of accuracy, recall and precision rate

(4) Determination of classification law

The results of Lasso-logistic and decision tree analysis were integrated, and the following variables were considered for the classification of High potassium-rich glass and Lead and barium enriched glass: potassium oxide(K_2O), iron oxide (Fe_2O_3), lead oxide (PbO), barium oxide (BaO), tin oxide (SnO_2), and sulfur dioxide (SO_2).

2.3. Classification of chemical composition subclasses based on factor analysis clustering method

(1) Factor analysis

Factor analysis was conducted on the elemental variables in the 49 measured lead-barium relic sampling points [10], and 14 chemical components in the pre-processed experimental data were downscaled and linearly combined according to factor analysis, resulting in four factors containing 73.116% of the information of 14 chemical components, which are representative and can be judged as a comprehensive index for the classification of lead-barium relic sampling points with practical significance.

After the main factor is determined, the factor score coefficient matrix is obtained by using the regression method, and the four factor score functions are established.

$$F1 = 0.103X_1 + 0.197X_2 + 0.143X_3 + 0.091X_4 - 0.011X_5 + 0.135X_6 - 0.138X_7 + 0.007X_8 + 0.199X_9 + 0.247X_{10} - 0.118X_{11} + 0.236X_{12} + 0.073X_{13} + 0.145X_{14} \quad (2)$$

$$F2 = 0.076X_1 - 0.371X_2 - 0.162X_3 + 0.190X_4 - 0.051X_5 + 0.071X_6 - 0.424X_7 + 0.140X_8 - 0.027X_9 - 0.084X_{10} + 0.519X_{11} - 0.107X_{12} + 0.009X_{13} - 0.012X_{14} \quad (3)$$

$$F3 = 0.133X_1 + 0.206X_2 + 0.150X_3 - 0.080X_4 + 0.404X_5 + 0.018X_6 + 0.135X_7 + 0.029X_8 - 0.096X_9 - 0.188X_{10} - 0.113X_{11} - 0.122X_{12} + 0.022X_{13} - 0.561X_{14} \quad (4)$$

$$F4 = -0.051X_1 - 0.177X_2 + 0.007X_3 - 0.023X_4 + 0.109X_5 + 0.028X_6 + 0.020X_7 - 0.505X_8 - 0.006X_9 + 0.058X_{10} - 0.062X_{11} + 0.169X_{12} + 0.598X_{13} + 0.107X_{14} \quad (5)$$

(2) k-means clustering

In this paper, using R software, the four principal factor scores were substituted for the original data as new variables for cluster analysis of lead barium glass, and the Sum of Squared Error (SSE) within-group squared error and -inflection point plots were cited to determine the optimal number of clusters. As the number of clusters becomes larger and the number of individual categories becomes smaller, the value of WSS definitely decreases as the number of clusters increases, so the change of the slope is of interest, and when the WWS decreases slowly, we therefore believe that the effect cannot be enhanced despite increasing the number of clusters, and the value of k corresponding to the "elbow point" is the optimal number of clusters. For lead barium glass, the inflection point diagram is shown in Figure 6, and the comprehensive analysis shows that the number of clusters k should be taken as 4.

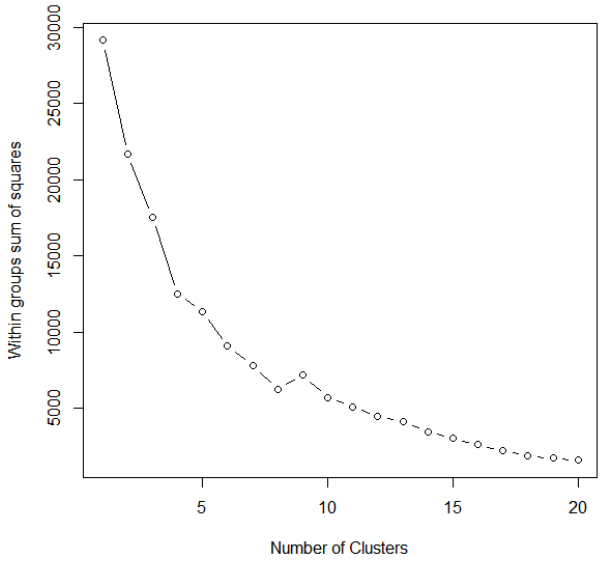


Figure 6. Inflection point diagram

For Lead and barium enriched glass sub-classification, the clustering result can be viewed as Figure 7.

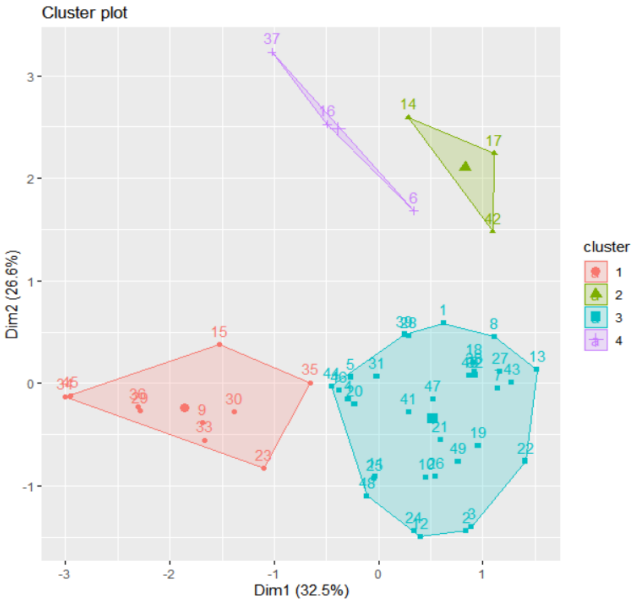


Figure 7. Visualization of clustering results

Due to the limitation of space in this paper, the detailed description of the subclassification of High potassium-rich glass is not given in the main text, which is verified to be exactly the same as the subclassification process of lead-barium glass. The results of subclassification for High potassium-rich glass and Lead and barium enriched glass are given here directly in Table 3.

Table 3. Subclassification results for High potassium-rich glass and Lead and barium enriched glass

Glass Type	Subcategories	Artifact Number
High potassium-rich glass	1	20
	2	25,30,38,43,43,46,47,49,54
	3	30,31,32,49,52
	4	02,08,08,11,19,23,24,26,26,28,29,33,34,35,36,37,39,40,41,42,42,44,45,48, 50,50,51,51,53,54,55,56,57,58
Lead and barium enriched glass	1	1,4,5
	2	3,6,6,21,27
	3	3,7,9,10,12,13,14,16,18,22

2.4. Fisher's discriminant analysis of clustering results[11]

The accuracy of the above clustering results based on factor analysis was verified by Fisher's discriminant, and three discriminant functions were established based on the clustering results, and the summary of the discriminant functions is shown in Table 4. As can be seen from the table, the variance contributions of the eigenvalues of discriminant functions 1, 2 and 3 were 73.900%, 25.100% and 1%, and the typical correlations were 0.930, 0.827 and 0.28%, respectively. This indicates that the three discriminant functions can be used as the basis for discriminating the accuracy of the clustering results of the subclasses of Lead and barium enriched glass.

Table 4. Summary table of discriminant functions

Eigenvalue				
Function	Eigenvalue	Percentage of variance	Cumulative percentage	Typical correlation
1	6.379a	73.9	73.9	930
2	2.169a	25.1	99.0	.827
3	.087a	1.0	100.0	.282

a. The first 3 canonical discriminant functions were used in the analysis.

From the clustering results, it can be seen that when the number of clusters is 4, the distribution of the group mass centers of each group in the Fisher discriminant function plot is shown in Figure 8. From Figure 3, it can be seen that when k=4, the group cores are evenly dispersed, and the samples between different groups have strong differentiation.

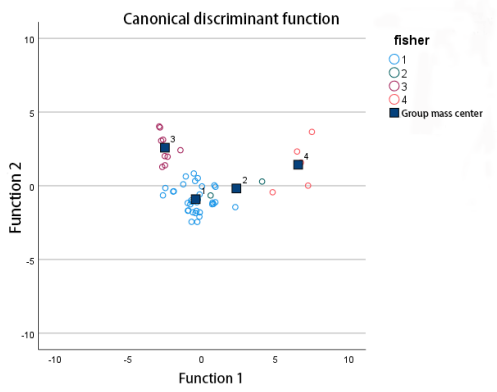


Figure 8. Distribution of Fisher discriminant function

The Fisher discriminant function model was used to verify the original classification and cross-classification of lead-barium glass. The results showed that when 49 Lead and barium enriched glass

samples were classified into 4 groups, the original classification was correct at 95.9% and the cross-validation was correct at 89.8%. This verified that the cluster analysis based on factor analysis could correctly classify the lead-barium glass into subclasses. Similarly, for the subclass classification of High potassium-rich glass the clustering result test also passed.

3. The establishment and solution of the glass identification model

As can be seen above, the solution of the model is first established in the data pre-processing, so the data is first centralized log-ratio conversion.

3.1. Solving logistic regression

From the conclusion of question two, it can be seen that the type of glass is mainly related to potassium oxide (K_2O), iron oxide (Fe_2O_3), lead oxide (PbO), barium oxide(BaO), tin oxide(SnO_2)and sulphur dioxide(SO_2), and has been verified to have a high degree of accuracy, so again using Stata to binary logistic regression of the variables after the Lasso regression screening, the correlation coefficients obtained are shown in Table 5.

Table 5. Table of logistic regression correlation coefficients

Type of artefact	Correlation coefficient value
K_2O	-0.050
Fe_2O_3	-0.484
PbO	0.078
BaO	0.036
SnO_2	0
SO_2	-0.053
cons_	0.298

So, $\theta(x) = -0.050w(K_2O) - 0.484w(Fe_2O_3) + 0.078w(PbO) + 0.036w(BaO) - 0.053w(SO_2) + 0.298$.

So for the determination of the glass type, the logistic regression obtained is : $y = \frac{1}{1+e^{-\theta(X)}}$

3.2. Category classification

After data pre-processing, the relevant chemical composition data of the type samples to be identified were substituted into the logistic regression function equation obtained above, and the function values obtained were as follows[12], for the function values obtained from the logistic regression, greater than 0.5 for lead and barium enriched glass and less than 0.5 for high potassium-rich glass, so the category identification results are shown in Table 6.

Table 6. Glass type identification results

Artifact Number	Logistic regression function value	Predicted glass type
A1	0.057	high potassium-rich glass
A2	0.733	lead and barium enriched glass
A3	0.794	lead and barium enriched glass
A4	0.790	lead and barium enriched glass
A5	0.796	Lead Barium
A6	0.062	lead and barium enriched glass
A7	0.014	lead and barium enriched glass
A8	0.729	Lead Barium

4. Conclusion

This paper analyzes and obtains the classification rules of high potassium-rich glass and lead and barium enriched glass, and applies them to the identification scheme of glass types.

For the classification law of glass, the surface weathering and chemical composition of glass were selected as features for central logarithmic ratio conversion (CLR). A decision tree was used to build a classification model through SPSS, and the main chemical components that determined the glass type were found to be lead oxide, barium oxide and potassium oxide through cross-validation. Secondly, as ordinary logistic regression showed significant over-fitting, Lasso-logistic was used to further accurately solve for the chemical components determining glass type by Stata, while k-fold cross-validation was chosen to further prevent over-fitting and to accurately solve for them. The results obtained from the logistic regression were cross-validated with the decision tree to obtain that for the classification of high potassium and lead and barium enriched glasses, the paper considers the following variables as the main reference: potassium oxide (K_2O), iron oxide (Fe_2O_3), lead oxide (PbO), barium oxide(BaO), tin oxide(SnO_2) and sulphur dioxide (SO_2). Secondly, the subcategories of high potassium-rich glass and lead and barium enriched glass were classified separately.Secondly, the high potassium-rich glass and lead and barium enriched glass are divided into sub-classes respectively. Factor analysis is used to extract the main factors, determine the optimal number of clusters, and K-means clustering is used to determine the sub-class division results, and R language is used to realize the visualization of clustering results. Finally, Fisher discrimination is used to verify the accuracy and rationality of clustering results.

For the identification of glass types, the model established above was used to identify unknown categories of glass. Firstly, the correlation coefficients of each decision factor of the logistic regression above were further solved to obtain the logistic regression equation, and secondly, the chemical composition data were centrally log-ratio transformed and the most likely types of glass were obtained through the logistic regression equation: high potassium-rich glass (A1, A6, A8), lead and barium enriched glass (A2 , A3, A4, A5, A7).

References

[1] YAN Jingping, DENG Lu, HU Lili. Research progress on the structure and properties of phosphate glasses [J]. Journal of Silicates,2022,50(04):1006-1021.

[2] Zhang Ran. Preparation, structure and luminescence properties of molybdenum-borosilicate glass and glass-ceramics[D]. Guilin University of Electronic Science and Technology, 2022.

[3] Chen Hui. Research on regression modeling method for multi-component data in monomorphic space[D]. Wuhan University of Technology, 2021.

[4] Fu Yalong. Research on component data processing methods[D]. Chang'an University,2019.

[5] Pawlowsky-Glahn V, Egozcue J J, Tolosana-Delgado R. Modeling and Analysis of compositional Data [M]. Wiley,2015.

[6] Sun Xiaoming, Yu Wujing, Ren Ruobing, Xiong Wang, Wang Yalan. Research on disambiguation of patent inventors' names based on decision tree algorithm[J/OL]. Science and Management:1-20 [2022-11-28].

[7] WANG Yaqi,SU Yiwei,LIU Migration. Development of lasso-logistic regression analysis prediction model for prognosis of paraquat poisoning[J]. Occupational Health and Emergency Rescue,2022,40(03):259-264.

[8] Qin Yao, Han Hongjuan, Chen Durong, Wang Haoji, Ge Xiaoyan, Bai Wenlin, Cui Jing, Yu Hongmei. A predictive model for reversal of mild cognitive impairment based on LASSO logistic regression model[J]. China Health Statistics,2022,39(05):653-658.

[9] Liu F. Stock value analysis based on logistic regression method--an example of the digital currency sector[J]. Investment and Cooperation,2022(12):32-34.

[10] Chen Z,Jiang H. XRF analysis of disposable paper cups based on factor analysis and systematic clustering [J/OL]. Packaging Engineering:1-10 [2022-11-28].

[11] Wang Xiao,Xu Qiuyue,Huang Benlin,Zhang Hui,Duan Yong.Clinical value of logistic regression and Fisher's discriminant analysis model in the differential diagnosis of coronary heart disease[J]. Journal of Clinical Laboratory,2022,40(09)

[12] Baofu, Ma Wen, Gao Yudou, Yang Tianchun, Li Night. Forecasting power outage sensitivity based on logistic regression model[J]. Microcomputer Applications,2022,38(07):67-68+72.