

Based on the CART Decision Tree Model of Prediction and Classification of Ancient Glass-Related Properties

Yanyan Han¹, Ke Xu¹, Jiayin Qin²

¹ Department of Logistics Management, Hefei University of Technology, China

² Department of Mechanical Engineering, Hefei University of Technology, China

* Corresponding Author Email: hanyanyan505@163.com

Abstract. Weathering occurs when glass interacts with the atmosphere. Ancient glass is susceptible to the influence of the burial environment. In the process of glass weathering, its composition ratio will be changed due to a large number of internal elements and environmental elements exchange. In this paper, Fisher's exact test and frequency analysis are used to test the glass pattern, type, color, and surface weathering respectively, and the conclusion is that there are significant differences between glass type and surface weathering, while there are no significant differences between pattern, color, and surface weathering. Then, the CART decision tree algorithm was used to classify the glass into weathered and unweathered, and a binary tree was established with 14 chemical components as characteristics. The average chemical composition of all the sampling points in the sibling nodes of the node where the weathering point is located was used as the predictive value of the chemical composition before the weathering point. Finally, the correlation analysis of the chemical constituents of high potassium glass and lead barium glass was carried out by hierarchical clustering method. In a high-potassium glass, silica is one category, and other chemical components are another category. In a lead-barium glass, silica and lead oxide are one category, and other chemical components are another category. Therefore, it can be considered that the chemical composition relationship of lead-barium glass is more complex.

Keywords: CART decision Tree, Hierarchical clustering, Fisher Coefficient, prediction model.

1. Introduction

The erosion caused by the interaction of glass and the atmosphere is called weathering [1]. Ancient glass is susceptible to weathering under the influence of the burial environment, and weathering is an important reason why ancient cultural relics cannot be completely preserved today. In the process of glass weathering, the proportion of its components will change due to a large amount of exchange between internal elements [2] and environmental elements, thus affecting the correct judgment of its category, the identification of ancient cultural relics, and the implementation of protection measures. Due to the influence of environmental factors and other factors [3], the degree of weathering of glass relics also varies. With the strengthening of people's patriotic feelings and the development trend of paying more and more attention to ancient cultural relics, the phenomenon of glass weathering has gradually been paid more and more attention by researchers. The degree of weathering of cultural relics is generally there are four kinds: no weathering, local shallow weathering, general weathering, and obvious weathering [4]. The researchers need to consider many factors of weathering reason carries on the analysis of historical relics and protection.

2. Prediction of composition of glass products before weathering based on CART decision tree classification model

2.1. Data preprocessing

After filtering out the vacant values in the data, it was found that the corresponding glass types were all lead barium, with weathering on the surface and the decoration falling into A or C. Therefore, the data of the decoration were classified and analyzed respectively. The data of pattern A, type lead barium, and surface weathering are screened out for frequency statistics. Light blue is 6, blue-green

is 1 and black is 2. Therefore, the color with the largest frequency is added to the vacancy value, namely light blue. For the vacancy value with a pattern of C, the filling color is light blue. As it is proposed in the title that the sum of component proportions and the data between 85% and 105% are regarded as valid data [5], that is, the sum of component proportions and the data of cultural relics sampling points outside this range are regarded as invalid data. According to the analysis, the total components of No. 15 and No. 17 are less than 85%, so these two sets of data will not be considered in the following calculation. At the same time, due to the detection means and other reasons, the proportion of components may be accumulated and not 100%, so the missing value should be changed to 0 without any other consideration and treatment. The results are shown in Table 1 and Table 2 below

Table.1. Decorative design A data filter color frequency

Light blue	Blue-green	Black
6	1	2

Table.2. Decorative design A data filter color frequency

Light blue	Dark green	Blue-green	Purple
6	4	2	2

2.2. Variable relationship analysis based on Fisher's exact test and frequency analysis

By checking the pattern, type, color, and surface weathering, it is found that there are significant differences between the type and surface weathering. However, there is no significant difference between ornamentation data and color data, and surface weathering. The relationship between surface weathering and type is shown in Table 3 and Table 4 below.

Table.3. Whether weathered * Type cross table

			Type		Total
			High potassium	Lead barium	
Weathered or not	Weathered	Count	6	28	34
		Count of expectation	10.6	23.4	34.0
		Percentage of weathered or not	17.6%	82.4%	100.0%
	Not weathered	Count	12	12	24
		Count of expectation	7.4	16.6	24.0
		Percentage of weathered or not	50.0%	50.0%	100.0%
Total		Count of expectation	18	40	58
		Percentage of weathered or not	18.0	40.0	58.0
		Percentage of weathered or not	31.0%	69.0%	100.0%

Table.4. Surface weathered * Type Chi-square test

	value	Degree of freedom	Progressive significance (bilateral)	Accuracy significance (bilateral)	Accuracy significance (bilateral)
Pearson chi-square	6.880 ^a	1	0.009		
Correction of continuity	5.452	1	0.020		
Likelihood ratio (L)	6.889	1	0.009		
Fisher's exact test				0.011	0.010
Linear correlation	6.762	1	0.009		
The effective number of cases	58				
a. The expected count of 0 cells (0.0%) is less than 5. The minimum expected count is 7.45.					

The relationship between the three variables can be obtained from the above two analyses, but it is not specific enough, so the frequency recording analysis of the four variables is carried out. For example, for decoration A, the degree of influence on surface weathering [6] is not biased, but more balanced; As for ornamentation B, both are weathering, so it can be concluded that ornamentation B is easy to weather; For ornamentation C, it can be concluded that there is a bias towards weathering but the degree is not deep. The results are shown in Table 5.

Table.5. Frequency record analysis

	Name	Whether the surface is weathered weathering		Total
		Not weathered	Weathered	
Grain	A	11	11	22
	B	0	6	6
	C	13	17	30
Total		24	34	58
Type	High potassium	12	6	18
	Lead barium	12	28	40
Total		24	34	58
Color	Blue-green	6	9	15
	Light blue	8	16	24
	Purple	2	2	4
	Dark green	3	4	7
	Dark blue	2	0	2
	Light green	2	1	3
	Black	0	2	2
	Green	1	0	1
Total		24	34	58

2.3. Prediction model of each chemical composition before weathering of weathering point based on decision tree algorithm

(1) Regression model

Through access to relevant data, this paper collected the sample data before and after weathering. The data source is the results obtained after the composition test of the unweathered and weathered areas of the glass bead samples by X-ray fluorescence spectrometer [1]. There are altogether 26 pieces of data, 13 pieces of data before weathering, and 13 pieces of data corresponding to weathering. However, the feature dimension of the data is large, with 15 chemical composition characteristics. To fit the deep learning neural network model, the number of existing samples is far less than the required number.

Therefore, principal component analysis was performed on the sample data before and after weathering. The SPSS analysis results are shown in Table 6:

Table.6. KMO test and Bartlett test

Feature	Weight
KMO value	0.185
Approximate chi-square	448.596
df	105.000
P	0.000***

The result of the KMO test shows that the value of KMO is 0.185. At the same time, the Bartlett sphericity test results show that the significance P value is 0.000***, showing significance at the level, there is a correlation between variables, the principal component analysis is effective, and the degree is not suitable. Therefore, feature dimension reduction cannot be carried out, so the number of samples required for model fitting cannot be reduced, and regression analysis is abandoned.

(2) Decision tree model training set preparation

Given the data sets of U2, the sample points are $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$. Where x is the 14-dimensional feature vector, representing 14 different chemical components), and y is the category output, indicating whether the surface is aerated.

(3) CART decision tree model

Establish the CART decision tree algorithm using the Gini coefficient to measure the decision-making process of the decision tree. For a given sample D , suppose there are K categories, and the number of the K th category is C_k , then the Gini coefficient expression of sample D is:

$$Gini(D) = 1 - \sum_{k=1}^K \left(\frac{|C_k|}{|D|} \right)^2 \quad (1)$$

Since the weathering model is only a binary classification, if the probability of weathering output is p , the Gini coefficient can be expressed as:

$$Gini(p) = 2p(1 - p) \quad (2)$$

In particular, for sample D , if D is divided into D_1 and D_2 parts according to some value A of feature a , then under the condition of feature A , the Gini coefficient of D is expressed as:

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (3)$$

The continuous chemical component values are discretized, m samples of continuous feature A has m , from smallest to largest order $a_1, a_2, \dots, a_m$, then the CART algorithm takes the average of two adjacent sample values to obtain a total of $m-1$ partition points, in which the i -th partition point T_i is expressed as $T_i = \frac{a_i + a_{i+1}}{2}$. For the $m-1$ point, the Gini coefficient when it is taken as the binary classification point is calculated respectively. The point with the lowest Gini coefficient is selected as the binary discrete classification point of the continuous feature. The algorithm in this paper starts from the root node and uses the training set to recursively build the CART tree.

For the data set of the current node is D , if the number of samples is less than the threshold or there is no feature, the decision subtree will be returned, and the current node will stop recursion.

Calculate the Gini coefficient of the sample set D. If the Gini coefficient is less than the threshold, the decision tree subtree is returned and the current node stops recursion.

Calculate the Gini coefficient of each eigenvalue of each feature of the current node to the data set D.

From the calculated Gini coefficient of each feature value to the data set D, select Feature A with the smallest Gini coefficient and the corresponding feature value a. According to the optimal feature and the optimal eigenvalue, the data set is divided into two parts D1 and D2. At the same time, the left and right nodes of the current node are established. The data set D of the node is D1, and the data set D of the right node is D2.5. Call the left and right child nodes recursively in steps 1-4 to generate the decision tree. When predicting the generated decision tree, suppose sample A in the test set falls to a leaf node, and there are multiple training samples in the node. The category prediction for A is the category with the highest probability in this leaf node. Finally, the following decision tree classification model is obtained, as shown in Figure 1.

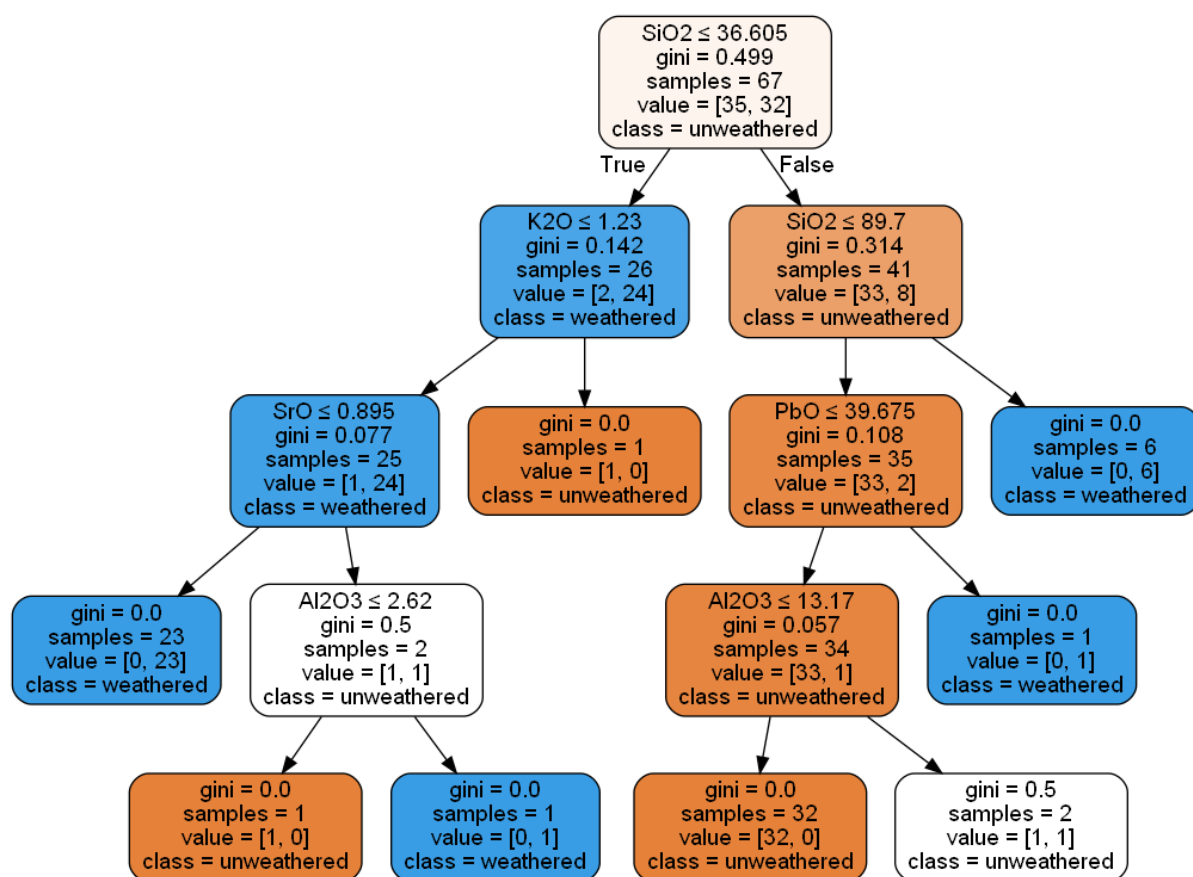


Figure 1. Decision tree classification model

(4) Chemical composition prediction model of weathering point before weathering

First, make the decision tree a full binary tree, and then number each node from top to bottom, from left to right, from 1 to 2 times the depth of the decision tree -1. Then, the leaf node whose classification result is weathered [7] is labeled as 1, and the leaf node whose fruit is not weathered is labeled as 0. The weathered sample points are input, the 14 chemical components are used as input features and the final output is the sample numbers assigned to the leaf nodes. The mean chemical components of all samples in their sibling nodes were calculated, and an array of length 14 was obtained. The number of the obtained classification result is set as x. If x is even, the number of the sibling node is x+1; otherwise, it is X-1. The sample mean of the corresponding number is calculated. Table 7 shows the chemical composition prediction before weathering.

Table.7. Chemical composition prediction of weathering point before weathering

	(SiO ₂)	(Na ₂ O)	(K ₂ O)	(CaO)	(MgO)	(Al ₂ O ₃)
02	24.525	0	0	0.235	0.555	2.62
07	59.9886	1.4302	3.3157	2.645	0.8068	5.464
08	24.525	0	0	0.235	0.555	2.62
08 Severe weathering point	24.525	0	0	0.235	0.555	2.62
09	59.9886	1.430	3.316	2.645	0.8069	5.464
10	59.9886	1.430	3.316	2.645	0.8069	5.464
11	24.525	0	0	0.235	0.555	2.62
12	59.9886	1.430	3.316	2.645	0.8069	5.464
19	24.525	0	0	0.235	0.555	2.62
22	59.9886	1.430	3.316	2.645	0.8069	5.464
	(Fe ₂ O ₃)	(CuO)	(PbO)	(BaO)	(P ₂ O ₅)	(SrO)
02	0	4.9	43.8	13.115	7.135	1.015
07	1.124	1.559	14.339	5.5954	1.19971	0.168
08 Severe weathering point	0	4.9	43.8	13.115	7.135	1.015
09	1.124	1.559	14.339	5.5954	1.19971	0.168
10	1.124	1.559	14.339	5.5954	1.19971	0.168
11	0	4.9	43.8	13.115	7.135	1.015
12	1.124	1.559	14.339	5.5954	1.19971	0.168
19	0	4.9	43.8	13.115	7.135	1.015
22	1.124	1.559	14.339	5.5954	1.19971	0.168

3. Qualitative analysis of glass samples based on CART decision tree

3.1. Establishment of CART decision tree model

According to the above methods, high-potassium glass and lead-barium glass are classified, and unweathered, weathered and severely weathered are taken as attributes for decision tree classification. The following decision tree classification results are obtained:

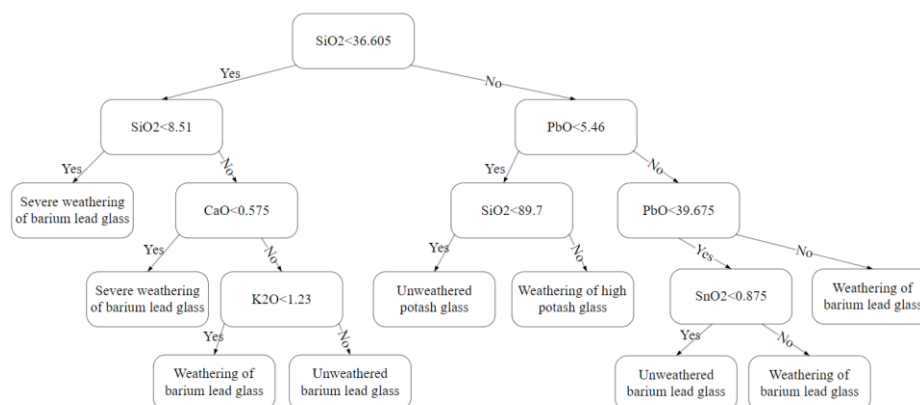


Figure 2. Decision tree classification results

As can be seen from Figure 2, there are five nodes: silica, calcium oxide, potassium oxide, lead oxide, and tin oxide. According to the judgment of each node and then choosing a different path to continue to subdivide, it will be found that the classification of high potassium glass and lead barium glass depends on their chemical composition and content. For example, if the proportion of silica [8] is less than 36.605%, further judge whether it is less than 8.51%; if it is less than, it is seriously weathered lead-barium glass; if it is greater, judge whether the content of calcium oxide is less than 0.575%; if it is less than, it is seriously weathered lead-barium; otherwise, judge whether the

proportion of potassium oxide is less than 1.23%; if it is, it is lead-barium weathering; Otherwise for lead barium no weathering.

3.2. Identify the type of unknown glass artifact

The attributes of unweathered, weathered, and severely weathered are not considered, but only high-potassium glass and lead-barium glass are considered. The classification results of the decision tree are shown in Figure 3 below:

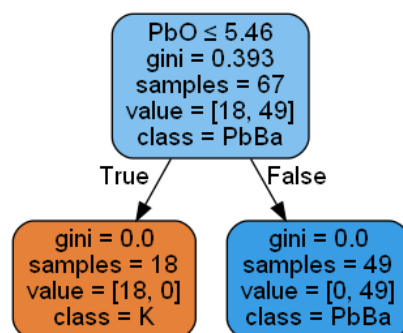


Figure 3. Decision tree classification results

The decision tree classification of cultural relics is carried out, and the results are shown in Table 8:

Table.8. Identify the type of unknown glass artifact

Heritage Number	Whether the surface is weathered	(PbO)	Type
A1	Not weathered		High potassium glass
A2	Weathered	34.3	Lead barium glass
A3	Not weathered	39.58	Lead barium glass
A4	Not weathered	24.28	Lead barium glass
A5	Weathered	12.23	Lead barium glass
A6	Weathered		High potassium glass
A7	Weathered		High potassium glass
A8	Not weathered	21.24	Lead barium glass

Because this article the classification number is only related to the lead oxide, so the sensitivity of the decision tree classification is lead oxide content change on the result of classification. The decision tree model relies on the content of lead oxide as the classification standard of glass type, and the classification accuracy is up to 100%. There is a large difference in the content of lead oxide in high-potassium glass and lead-barium glass, and the proportion of the decrease of lead oxide content in lead-barium glass [9] after weathering is small. Therefore, this classification standard can effectively avoid the judgment of weathering on the type of glass. If the classification point of lead oxide is artificially set as 10%, the overall classification accuracy will be increased and the generalization is good. Taking the content of lead oxide as the classification point is the most reasonable and stable.

4. Chemical composition analysis of different classes of glass artifacts based on hierarchical clustering

4.1. Spearman coefficient analysis

In this paper, Spearman order correlation coefficients of different chemical compositions of high potassium glass and lead barium glass were calculated respectively. The correlation between alumina and ferric oxide is the highest in high-potassium glass, while the correlation between alumina and magnesium oxide is the highest in lead-barium glasses. In addition, the correlation of chemical compositions in different kinds of glass is quite different. For example, alumina has a medium or

above correlation with many substances in high-potassium glass, but the correlation degree of alumina with other substances in lead-barium glass is small. Table 9 and Table 10 list Spearman order correlation coefficients among some variables.

Table.9. Spearman order correlation coefficient of chemical composition in high potassium glass

MgO	Al ₂ O ₃	0.673557
K ₂ O	Al ₂ O ₃	0.512423
CaO	P ₂ O ₅	0.501893
CuO	BaO	0.494679
SiO ₂	Al ₂ O ₃	0.424654
CaO	Fe ₂ O ₃	0.411274
SiO ₂	Na ₂ O	0.404812
Al ₂ O ₃	Fe ₂ O ₃	0.387043

Table.10. Spearman order correlation coefficient of chemical composition of lead barium glass

MgO	Al ₂ O ₃	0.673557
K ₂ O	Al ₂ O ₃	0.512423
CaO	P ₂ O ₅	0.501893
CuO	BaO	0.494679
SiO ₂	Al ₂ O ₃	0.424654
CaO	Fe ₂ O ₃	0.411274
SiO ₂	Na ₂ O	0.404812
Al ₂ O ₃	Fe ₂ O ₃	0.387043
BaO	SO ₂	0.370916
Fe ₂ O ₃	SnO ₂	0.363359
PbO	SrO	0.362154

4.2. Hierarchical cluster analysis

In this paper, the chemical composition of high-potassium glass and lead-barium glass were analyzed by hierarchical clustering. The analysis results are shown in Figure 4 and Figure 5 below:



Figure 4. High potassium cluster analysis diagram

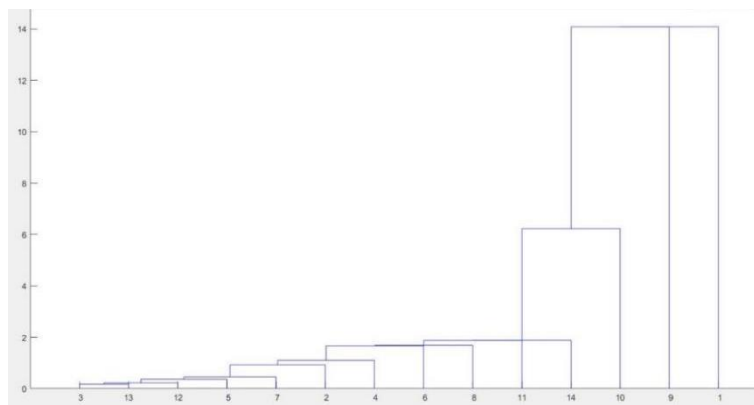


Figure 5. Lead barium cluster analysis diagram

1 represents silicon dioxide, 2 sodium oxide, 3 potassium oxide, 4 calcium oxide, 5 magnesium oxide, 6 alumina, 7 iron oxide, 8 copper oxide, 9 lead oxide, 10 barium oxide, 11 phosphorus pentoxide, 12 strontium oxide, 13 tin oxide, and 14 sulfur dioxides. Through hierarchical clustering, those with small distance differences can be regarded as the same class. Therefore, it can be seen from the two figures that the clustering result of high potassium is silica in one category and other categories. The clustering result of lead barium is a class of silica and lead oxide.

Through the analysis and comparison of the clustering results of high-potassium glass and lead-barium glass, we can find that high-potassium glass has fewer associated chemical groupings than lead-barium glass, only silica is a single category [10], and other chemical elements have little difference. While the silica and lead oxide of lead-barium glass belong to the same class, barium oxide belongs to one class alone, and there are more related chemical groups. It can be concluded that the chemical composition of lead-barium glass is more complex than that of high-potassium glass.

5. Conclusion

First of all, based on Fisher's exact test and frequency analysis, it is concluded that glass relics with pattern type B are easier to weather. Secondly, the use of the CART decision tree algorithm to determine whether glass relics weathering is mainly based on the chemical composition content of silica, potassium oxide, lead oxide, alumina, and strontium oxide. Distinguish the key characteristics of high potassium glass and lead barium glass. The difference is the lead oxide content. When the lead oxide content is less than 5.46%, the glass is high potassium glass. Finally, the correlation between the contents of various chemical components in glass was analyzed by calculating Spearman coefficient, and the chemical composition contents of high-potassium glass and lead-barium glass were respectively clustered by the hierarchical clustering method. It was found that the chemical composition relationship of lead-barium glass was more complex than that of high-potassium glass.

The model can be used in weathering analysis or corrosion analysis of other cultural relics after proper adjustment. The model can also provide a new method for the identification of a large class of articles by identifying them more carefully.

References

- [1] Hu Zhi-zhong, LI Pei, JIANG Lu-man, WANG Tong-Yang, DU Gu, Yang Bo. Ancient glass material LA - icp-ms component analysis and source study [J]. Journal of testing, 2020, 33 (4) 6:505-514. The DOI: 10.15898 / j.carol carroll nki. 11-2131 / td., 201909210134.
- [2] HU Xiaozhong. The subordinate position of ancient Chinese glass and its influence -- The ancient Chinese ceramics and glass Chemistry from the perspective of ecological niche Theory [J]. Journal of guangxi university for nationalities (natural science edition), 2009 (04) : 54-58, DOI: 10.16177 / j.carol carroll nki GXMZZK. 2009.04.017.

- [3] Zhao Zhiqiang, Wen Rui. Study on the composition system and making technology of glass beads unearthed from Shirenzigou Site Group in Barikun, Xinjiang. M.s., Northwestern University, 2016.
- [4] Wang Chengyu, Tao Ying. Weathering of silicate glass [J]. Journal of the Chinese Ceramic Society, 2003(01):78-85.
- [5] Wen Rui, ZHAO Zhiqiang, Ma Jian, Wang Jianxin. Composition analysis of glass beads unearthed from Shirenzigou Site in Balikun, Xinjiang [J]. Spectroscopy and Spectral Analysis, 2016, 36(09):2961-2965.
- [6] Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man-Machine Studies. 27(3): 221–234. CiteSeerX10.1.1.18.4267. doi:10.1016/S0020-7373(87)80053-6.
- [7] Wang Zifan. Characteristics of Glass in Early Ancient China [J]. Tiangong, 2020(03):10-11.
- [8] R. Sibson (1973). "SLINK: an optimally efficient algorithm for the single-link cluster method" (PDF). The Computer Journal. British Computer Society. 16(1): 30–34. doi:10.1093/comjnl/16.1.30.
- [9] Feng age. China unearthed ancient glass bead database construction and application [D]. Northwest university, 2021. The DOI: 10.27405 /, dc nki. Gxbdu. 2021.001676.
- [10] Hotelling, Harold (1953). "New Light on the Correlation Coefficient and its Transforms". Journal of the Royal Statistical Society, Series B (Methodological). 15(2): 193–225. doi:10.1111/j.2517-6161.1953.tb00135.x. ISSN0035-9246.