Chemical compositions prediction of glass artifacts based on logistic regression and linear weighting

Kewen Zuo *, Yi Yang, Xingyu Li

College of Electrical and New Energy, China Three Gorges University, Yichang, China

* Corresponding Author Email: KevinZuo0110@163.com

Abstract. Chemical composition analysis is of great significance for studying the category, origin and production age of the ancient glass. In order to consider the influence of weathering on chemical composition, this paper proposes a prediction method of glass chemical composition based on logistic regression and linear weighting. First, the chi-square test shows that whether the glass surface is differentiated is related to the type of glass. Secondly, the chemical composition of glass relics is divided into two categories by binary logistic regression: significantly affected by weathering and not significantly affected by weathering. Then, for the chemical components not significantly affected by weathering, combined with descriptive statistical results, the mean value was used as the chemical composition content before weathering. For the chemical components significantly affected by weathering, the linear weighting method was used to predict the content of chemical components before weathering. The final results show that this method can better reflect the influence of weathering factors on the chemical composition of glass, which is of great significance for the prediction of the glass composition.

Keywords: Glass artifacts, chemical composition, logistic regression, linear weighting.

1. Introduction

The Silk Road was a channel for cultural exchange between China and the West in ancient times [1], in which glass was valuable physical evidence of early trade exchanges, and although similar in appearance, the chemical composition was different. The main raw material of glass is quartz sand, whose main chemical composition is silicon dioxide (SiO_2). Due to the high melting point of pure quartz sand, to reduce the melting point, fluxes need to be added during refining [2]. The main chemical composition differs depending on the fluxes added to it. As ancient glass is susceptible to weathering due to the environment of storage, the external elements will exchange a lot with the internal elements of the glass [3], and thus the composition percentage of the glass changes, which eventually affects the judgment of the researcher on the glass category. Therefore, it is necessary to make a reasonable prediction of the chemical composition of glass artifacts under the influence of weathering factors.

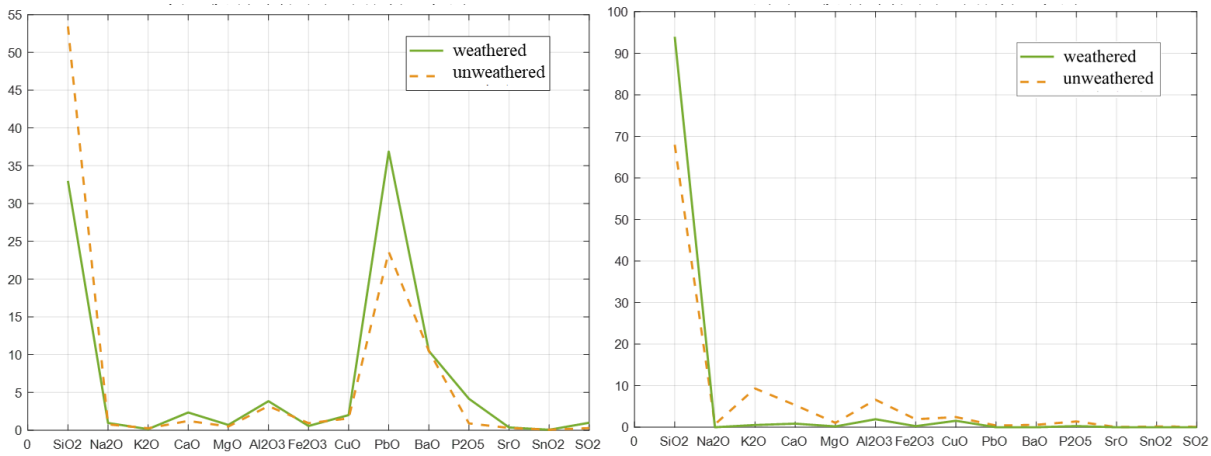
In order to reflect the influence of weathering on the chemical composition of glass, this paper proposes a prediction method of glass chemical composition based on logistic regression and linear weighting. Firstly, the relationship between whether the surface of glass relics is weathered and type, decoration and color is analyzed. Then, the chemical compositions of glass relics are divided into two categories: significantly affected by weathering and not significantly affected by weathering by binary logical stepwise regression, and the chemical composition was predicted respectively. For the chemical components that are not significantly affected by weathering, the mean value is used as the chemical composition content before weathering. For the chemical components significantly affected by weathering, the linear weighting method is used to predict the content of chemical components before weathering.

2. Materials and Methods

2.1. Data Description and Preprocessing

The data in this paper come from the 2022 China Student Mathematical Modeling Competition (http://www.mcm.edu.cn/). The dataset gives the categories of different glass artifacts, their weathering, and the percentage of different chemical compositions at the sampling points. The cumulative sum of each chemical composition of the same glass artifact should be between 85% and 105%. By analyzing the dataset, it is found that artifacts 19, 40, 48, and 58 in Form 1 are missing color information, and the data from Annex Form 1 are processed using hot card filling, and the colors of artifacts 19, 40, 48, and 58 are obtained as light blue. Studying Form 2 revealed that the data of group 15 and group 17 are invalid, so they are excluded. Missing values in the data set indicate that the chemical compositions are not detected, so they are filled with 0.

According to the type of glass, the percentage of chemical composition data of the processed glass are divided into two categories of glass artifacts, lead barium glass and high potassium glass. The percentages of the fourteen chemical components in the glass are shown in Figure 1.



(a) Lead barium glass (b) High potassium glass

Figure 1. Relationship between chemical composition and weathering of different types of glass

For lead barium type glass, the difference in the content of chemical composition SiO_2 and PbO is very obvious in weathered and unweathered cases, and there is a slight difference in the proportion of P_2O_5 composition with or without weathering. The proportions of the chemical components Na_2O, Fe_2O_3, and SnO_2 in the two cases are almost indistinguishable. Therefore, according to the above analysis, we can make an inference that Na_2O, Fe_2O_3, and SnO_2 are not the main chemical components that affect whether the lead-barium type glass is weathered or not.

For high potassium type glass, it can be seen that the contents of SiO_2, K_2O, MgO, and Al_2O_3 are larger in both cases, while the contents of SrO, SnO_2, and SO_2 are lower.

2.2. Chi-square Test

Since the four heritage characteristics indicators of glass type, decoration, color, and whether the glass surface is weathered are all fixed-class data, correlation analysis of fixed-class data cannot be performed by the commonly used quantitative data correlation analysis method. The chi-square test is a widely used hypothesis testing method, which is applicable to random samples and fixed-class data. The chi-square test is the degree of deviation between the actual observed value of the sample and the theoretical inferred value, and the degree of deviation between the actual observed value and the theoretical inferred value determines the size of the chi-square value. Therefore, the chi-square test is chosen for the analysis. The chi-square value is calculated as shown in Equation (1).

$$\chi^2 = \sum_{i=1}^{n} \frac{(A_i - T_i)^2}{T_i} \tag{1}$$

where, A_i is the actual frequency, T_i is the theoretical frequency, and n is the number of classification terms.

2.3. Binary Logistic Stepwise Regression

In 2.1, according to Figure 1, it is known that not all chemical compositions are closely related to weathering, so when predicting the chemical compositions content of artifacts before weathering, it is necessary to classify chemical compositions into two categories according to the degree of influence by weathering: significantly affected by weathering and insignificantly affected by weathering. Logistic regression is a model that can be used to establish the relationship between categorical independent variables (X) and dependent variables (Y) [4]. The dependent variable in this study is whether weathering, so binary logistic stepwise regression is used to screen the chemical compositions that are differently affected by weathering.

Assuming that the conditional probability that an artifact is weathered is $P_i = P\left(Y_i = 1 \middle| X\right)$, the binary logistic regression model assumes that the expression for this probability is as in Equation (2).

$$P_i = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_{14} X_{14})}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_{14} X_{14})}} \tag{2}$$

where P_i represents the probability that the cultural relics are weathered cultural relics, X represents each chemical component, β_0 is a constant, and $\beta_1 \sim \beta_{14}$ represents the correlation between each chemical component and weathering

Since Logistic functions are covariant nonlinear functions [5], they are able to be converted into linear functional form.

$$\frac{P_i}{1-P_i} = e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_{14} X_{14})} \tag{3}$$

This ratio is called the likelihood ratio [6]. A linear function can be obtained by performing a natural logarithm transformation on the likelihood ratio. It is calculated by Equation (4).

$$\log\frac{P_i}{1-P_i} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{14} X_{14} \tag{4}$$

Through the binary Logistic stepwise regression method, the chemical components significantly affected by weathering can be screened out. Based on this, the chemical compositions can be divided into two categories : significantly affected by weathering and not significantly affected by weathering.

2.4. Prediction Method of Original Chemical Compositions

For all chemical compositions of weathered glass artifacts, they are divided into those significantly affected by weathering, and those not significantly affected by weathering [7]. Different methods should be used for different compositions to predict the content of the compositions before weathering.

1) Chemical compositions not significantly affected by weathering

From Figure 1, it can be clearly seen that the differences between most of the chemical components before weathering and after weathering are not significant. Then for the same category of weathered glass artifacts in the chemical composition affected by weathering is small, the average value of the corresponding chemical composition in the category of unweathered glass artifacts is used as the content of chemical composition before weathering.

2) Chemical compositions significantly affected by weathering

Among all chemical compositions, there are also chemical compositions that are strongly affected by weathering. These chemical components have large differences in values before weathering and after weathering in the same category of artifacts, reflecting that the chemical components are

significantly affected by weathering. For these chemical components, a linear weighting method is adopted to predict their content before weathering. Based on the above analysis, the following steps are taken for the prediction of the pre-weathering chemical composition content.

Step1 Data Segmentation

The glasses are classified into four categories: lead-barium weathered glass, lead-barium unweathered glass, high potassium weathered glass, and high potassium unweathered glass. In each glass category, the chemical components that are less affected by weathering are excluded and those that are more affected by weathering are retained according to the results of the logistic stepwise regression model.

Step2 Calculate the distance vector

Taking lead-barium weathered glass and lead-barium unweathered glass as examples, the chemical component of lead-barium glass that is more affected by weathering is assumed to be a. For a lead-barium weathered glass artifact, the Euclidean distance of the artifact for each unweathered glass artifact is calculated using Equation (5) [8].

$$d_{ij} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{in} - x_{jn}\right)^2} \tag{5}$$

where i represents unweathered glass artifacts, $i = 12$ in the case of lead-barium glass artifacts and $i = 12$ in the case of high-potassium glass artifacts. j represents weathered glass artifacts, $j = 28$ in the case of lead-barium glass artifacts and $j = 6$ in the case of high-potassium glass artifacts. The final distance vector $D_i = \left[d_{i,1}, d_{i,2}, \ldots, d_{i,j} \right]$ is obtained.

Step3 Linear weighted prediction

As the Euclidean distances of weathered artifacts to different unweathered artifacts are different, the small Euclidean distance between them indicates that the composition contents of the two are approximately close to each other. The closer the weathered artifacts are to the unweathered glass artifacts, the greater their contribution to the chemical composition of the weathered glass artifacts, so the linear weighting method is used to reflect the variability between different glass artifacts.

For the j-th weathered glass artifact, the similarity of its chemical composition to the i-th unweathered glass artifact is defined as the inverse of the Euclidean distance between the two. Equation (6) is the formula for similarity calculation.

$$\alpha_{ij} = \frac{1}{d_{ij}} \tag{6}$$

Then, the similarity is normalized, and the normalized value is defined as the weight of the weathered glass artifacts. In this paper, we use Z-score normalization process and Equation (7) as the conversion function.

$$w_i = \frac{\alpha_{ij} - \overline{\alpha_{ij}}}{\sigma_{ij}} \tag{7}$$

Where, $\overline{\alpha_{ij}}$ is the mean value of similarity and σ_{ij} b is the standard deviation of similarity. For chemical component X_k, which is more affected by weathering, its original content is calculated using equation (8).

$$X_k = w_1 \cdot X_{1,k} + w_2 \cdot X_{2,k} + \ldots + w_i \cdot X_{i,k} \tag{8}$$

Where, $X_{i,k}$ a represents the k-th chemical composition of the i-th unweathered glass artifact that is highly affected by weathering. $k = 4$ in lead-barium glass artifacts and $k = 6$ in high-potassium glass artifacts.

In summary, in order to predict the chemical composition content of weathered glass artifacts before weathering, the data are firstly preprocessed and the characteristics that correlate most strongly with whether the glass surface is weathered are analyzed using chi-square tests. Secondly, the glass artifacts are classified by lead-barium weathered glass, lead-barium unweathered glass, high potassium weathered glass, and high potassium unweathered glass, and for different categories of artifacts, the chemical compositions are classified into two categories, significantly affected by weathering and insignificantly affected by weathering, using binary logistic stepwise regression. For the chemical components that are not significantly affected by weathering, the mean value is used as the chemical component content before weathering. For the chemical components that are significantly affected by weathering, the linear weighting method is used to predict the chemical component content before weathering. The process of chemical composition prediction is shown in Figure 2.

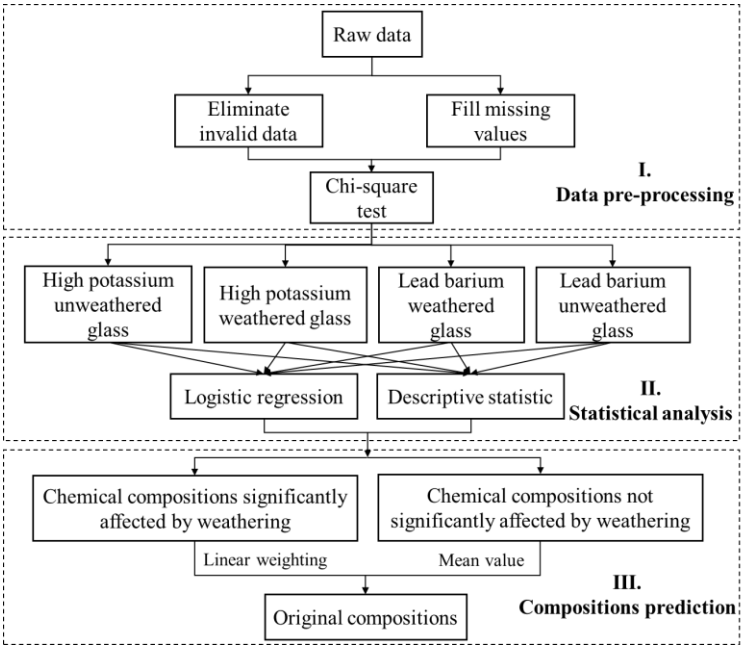


Figure 2. Flow chart of chemical compositions prediction

3. Results

Chi-square analysis is performed using SPSS to calculate the significance between whether the glass surface is weathered and color, glass type, and decoration. The results are shown in Table 1.

Table 1. Results of chi-square test

Influencing Factors	χ^2	P
Color	6.287	0.507
Glass Type	6.88	0.009
Decoration	4.957	0.084

As can be seen from Table 1, weathering does not show significant differences for ornamentation and color ($p>0.05$), while weathering presents significant differences for glass type ($p<0.05$). The results indicate that weathering on the glass surface is related to glass type and not to ornamentation and color.

Binary logistic stepwise regressions are performed for lead-barium glass artifacts and high-potassium glass artifacts, respectively, to obtain chemical compositions that are more affected by weathering for subsequent compositional content prediction work. The results of the binary logistic stepwise regression for the lead-barium glass artifacts are shown in Table 2.

Table 2. Results of binary logistic stepwise regression analysis of lead barium artifacts

Item	Regression coefficient	*p*
SiO_2	-0.892*	0.018
CaO	-0.292*	0.030
PbO	0.156**	0.005
P_2O_5	0.579*	0.021
Constant term	-7.193**	0.005
Hosmer-Lemeshow test	$\chi^2 = 10.893$	0.208

From Table 2, it can be seen that SiO_2, CaO, PbO, and P_2O_5 in lead barium glass are significantly affected by weathering since the value of *p* for each is less than 0.05. The Hosmer-Lemeshow goodness-of-fit test is used to analyze the goodness-of-model-fit situation [9]. The original hypothesis of the model test was that the model fit values and the observed values are in agreement [10]. *p* is greater than 0.05 (p=0.208>0.05) and thus the original hypothesis is accepted, which means that this model passes the HL test and the model fit is good.

The binary logistic stepwise regression results for the high potassium glass artifacts are shown in Table 3.

Table 3. Results of binary logistic stepwise regression analysis of high potassium artifacts

Item	Regression coefficient	*p*
SiO_2	-0.759**	0.000
K_2O	-0.261**	0.000
CaO	-0.215**	0.003
MgO	0.156**	0.008
Al_2O_3	0.289**	0.000
Fe_2O_3	0.156*	0.028
Constant term	-9.251**	0.005
Hosmer-Lemeshow test	$\chi^2 = 9.834$	0.208

As can be seen from Table 3: Since the p-values of each item are less than 0.05, the six components of SiO_2, K_2O, CaO, MgO, Al_2O_3, and Fe_2O_3 in the high potassium glass are significantly affected by weathering. This indicates that the present model passes the HL test and the model fit is good.

Based on the chemical content of different types of glass after weathering, the percentages of each chemical content of high potassium glass and lead-barium glass before weathering are predicted using a linear weighting treatment. Due to space limitation, only some of the predicted results for chemistry are shown in Table 4.

Table 4. Partial chemical composition content of glass before weathering

Glass Type	Sampling points	SiO_2	CaO	K_2O	SrO	SnO_2	SO_2
High Potassium Glass	07	69.33	6.32	9.99	0.04	0.20	0.10
	09	87.05	2.01	5.19	0.04	0.20	0.10
	10	61.71	5.87	12.37	0.04	0.20	0.10
	12	65.88	7.12	9.67	0.04	0.20	0.10
	22	61.58	7.35	10.95	0.04	0.20	0.10
	27	67.65	0.00	7.37	0.04	0.20	0.10
Lead barium glass	02	36.28	2.34	0.14	0.37	0.06	0.98
	11	33.59	3.51	0.14	0.37	0.06	0.98
	19	29.64	2.93	0.14	0.37	0.06	0.98
	26	19.79	1.44	0.14	0.37	0.06	0.98
	57	25.42	1.31	0.14	0.37	0.06	0.98
	58	30.39	3.49	0.14	0.37	0.06	0.98

To better visualize the changes in the composition of different types of glass before and after weathering, the chemical composition percentage diagrams of different types of glass are made. As shown in Figure 3 and Figure 4.

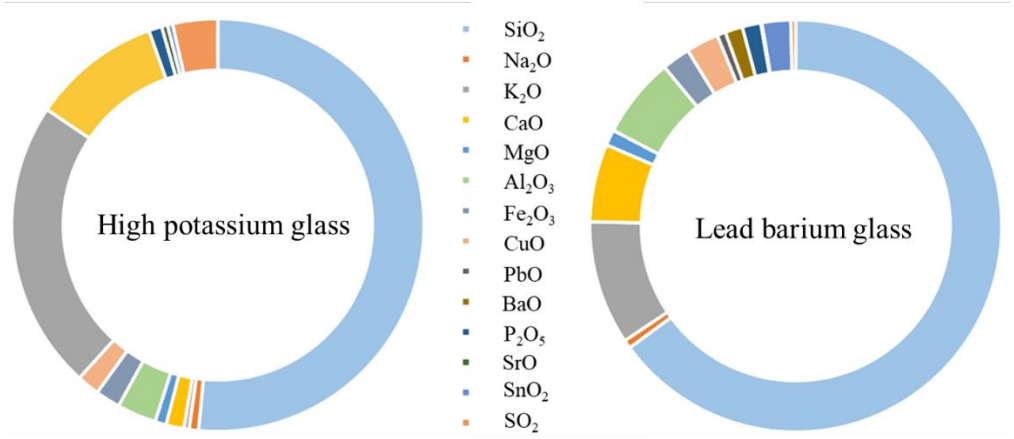


Figure 3. The proportion of chemical composition of different types of unweathered glass

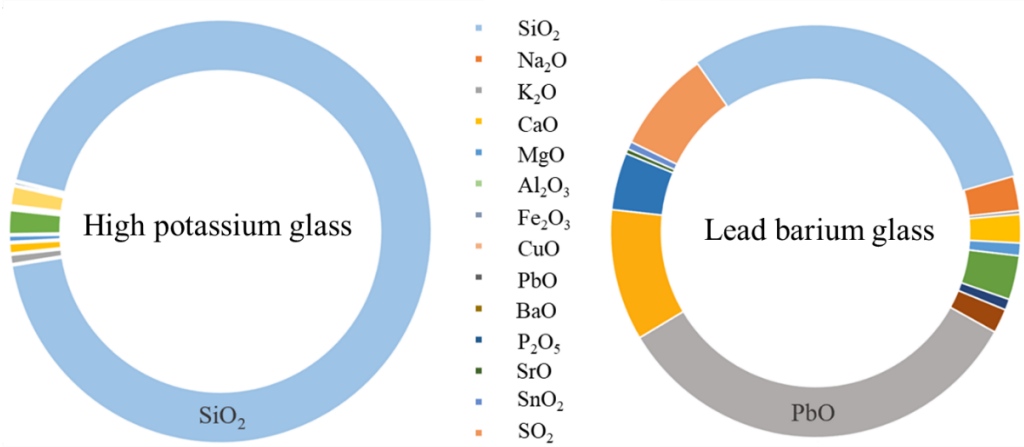


Figure 4. The proportion of chemical composition of different types of weathered glass

It can be seen from Fig.3 and Fig.4 that for high potassium glass relics, the proportion of silica after weathering increases significantly, and potassium oxide decreases significantly. For lead barium glass, the proportion of silica after weathering is significantly reduced, and the proportion of lead oxide is significantly increased.

4. Conclusions

In this paper, a method for predicting the chemical composition of glass based on logistic regression and linear weighting is proposed. Firstly, whether the glass surface is weathered or not is related to the glass type by chi-square test analysis. Then, the binary logistic regression algorithm is used to classify the chemical composition of glass artifacts into two categories, which are significantly affected by weathering and insignificantly affected by weathering, reflecting the influence of weathering factors on the chemical composition of glass. For the chemical components with large differences in content before and after weathering, a linear weighting method is used to predict their chemical composition content before weathering. This method makes better use of the similarities and differences in chemical composition between unweathered glass and weathered glass, and provides a new idea for predicting the chemical composition of glass before weathering.

References

[1] Zeng Weisheng. Ancient Chinese and Western glassware and their cultural differences from the glass unearthed in the Han Dynasty [J]. Journal of Gemology and Gemology, 2022,24 (06): 84-92.

[2] Xu Changwei, Chen Yong, Meng Qihan, Ma Shifang, Zeng Wei.Effects of flux on sintering and properties of CaO-MgO-Al _ 2O _ 3-SiO _ 2 system glass-ceramics [J]. Material Guide, 2015,29 (S2): 443-445 + 488.

[3] Wang Chengyu, Tao Ying. Weathering of Silicate Glass [J]. Silicate Journal, 2003, (01): 78-85.

[4] Wu and Xu. Research on automobile modeling design method based on binary logistic regression analysis [J]. Design art research, 2022,12 (06): 1-5 + 29.

[5] Li Qiang, Ren Zhiyuan. Spatial statistics and simulation of land use change based on Logistic regression analysis [J]. Forum on Statistics and Information, 2012,27 (03): 98-103.

[6] Sun Qian, Jia Ru, Zheng Jingli, Lei Peiyu, Hui Xiaofen.Analysis of influencing factors of health behavior of employees in public places in Guanzhong area of Shaanxi Province [J]. Journal of Yan 'an University (Medical Science Edition), 2022,20 (04): 42-46.

[7] Wang, Zhao, Li, Guo, Xiao and Liu.Prediction and sub-classification method of original composition of weathered silicate glass based on machine learning [J]. Journal of Silicate: 1-11.

[8] Fan Yanyan, Han Shiyu, Jiang Xu, Lv Yuying. Resilience evaluation of safety system in plateau railway operation period based on combination weighting-Euclidean distance [J]. Journal of Railway Science and Engineering: 1-11.

[9] Wang Hongqing, Chen Bo, Zhao Zuping, Song Jing.Study on the influencing factors of college students ' academic problems based on binary Logit regression model [J]. Technology and innovation, 2022, (01): 40-43.

[10] Pan Junjie, Zhou Jie.Empirical study on the influencing factors of consumers ' purchase intention of swimwear based on Logistic model [J]. Journal of Xi 'an Polytechnic University, 2020,34 (04): 19-25.