Component Analysis Model of Glass Relics Based on Logistic Regression and Fisher Linear Discrimination

Jiayan Xu *, Chenfei Shi #, Chenyuan He #

School of Science, Chang'an University, Xi'an, Shaanxi, 710018

* Corresponding Author Email: vancyxi@163.com

#These authors contributed equally.

Abstract. In recent years, with the attention paid by various countries to the exploitation and protection of cultural relics and the continuous development of industrialization, people have gradually realized the necessity of analyzing the weathering of ancient glass relics. In this paper, the physical characteristics, chemical composition and content of sample cultural relics are analyzed and their types are identified by using multiple algorithms such as Logistic regression model, Fisher linear discrimination, Fisher's exact test, etc. The study of this issue can provide guiding suggestions for the analysis and classification of ancient cultural materials, and help archaeologists carry out a series of work such as cultural relics protection. The results show that the "type" factor has the greatest relationship with weathering, followed by "color" and "decoration". The regression equation between the sample and weathering chemical composition is established, and it is concluded that "type" and "SnO_2" have a large negative correlation with weathering, "MgO" and "SO_2" have a large positive correlation, and other variables have a weak correlation. Finally, Fisher linear discriminant method is used to identify the category of unknown glass relics. By estimating the training group and testing the test group, it is found that the algorithm can correctly classify 100.0% of the original grouped cases. We can identify the unknown cultural relics and find that A1, A5, A6 and A7 cultural relics belong to high potassium glass, and the rest belong to lead barium glass. In order to test the accuracy of the model in identifying the category of glass relics, Fisher's exact test algorithm was used. It was found that the classification results were more sensitive to silica in the fluctuation range of - 50% to - 25%, and less sensitive to other chemical components, it is stable.

Keywords: Logistic regression model, Fisher linear discrimination, Analysis of cultural relics.

1. Introduction

The ancient refining technology made the glass easy to lose crystal water with the passage of time, and the glass that lost crystal water will produce glass alkalization. That is glass weathering [1]. It has major practical significance to study the problem of glass weathering. Firstly, it can effectively improve the storage method of glass relics. Secondly, it can improve the contemporary glass making process and effectively extend the life of contemporary glass products. Therefore, studying the relationship between glass weathering and various factors can foster the development of the protection of cultural relics and the industrialization process of glass products in China [2]. Hao Yunhong, Wu Regen [3] and others used the Taguchi method to design the erosion test of tempered glass, conduct weathering and erosion tests on tempered glass, and analyze and compare the erosion laws of tempered glass under various influence factors before and after weathering. Wang Chengyu and Tao Yingzong [4] described the factors affecting the weathering resistance of float glass, crystalline glass and vessel glass, and used SEM, WDS, EPMA and XPS to detect the morphology and composition of weathering products. Qi Ji, Gulistan [5] and others measured the amount of MgO and Na_2O precipitated from glass by atomic absorption spectrophotometer, and measured the surface composition of glass after weathering by fluorescence X-ray spectrometer.

This paper will study the weathering factors ‘influence of the surface of glass relics, including the type of glass, the patterns and colors of the surface of cultural relics. Analyze the relationship between the surface weathering of glass relics and these factors. Secondly, considering the types of glass, the relationship between surface weathering and chemical composition and glass type is analyzed. At the same time, we will identify the categories of some glass relics and analyze the sensitivity of the classification results. Then we obtain the classified data after fluctuation by periodically fluctuating the content of one chemical component and stabilizing the content of other chemical components. The sensitivity of classification results is obtained by comparing the classification before and after fluctuation.

2. The basic funamental of model

2.1. Logistic regression model

Logistic regression is a multivariate statistical method, which is applicable to the case where the dependent variable is binary or multi-classified variable. The independent variable can be applied under a wide range of conditions, which can be classified variable or continuous variable. The two-class logistic regression model does not have assumptions about the distribution of variables, nor does it need to assume that there is a multivariate normal distribution between them. Finally, the results are provided in the form of event occurrence probability. The parameters of the fitted logistic regression model are estimated by the maximum likelihood estimation method instead of the usual minimum multiplication method.

To explore the relationship between the surface weathering of glass relics and the influencing factors, it is necessary to use the logistic regression model to explore the influence index of different variables on the surface weathering of cultural relics. There are three factors affecting the surface weathering of glass relics. The independent variable X_1 is glass type, X_2 is decoration, and X_3 is color; The dependent variable Y is a binary variable. $Y = 0$ is "unweathered" and $Y = 0$ is "weathered". Define the probability P of "weathering" and the probability $1 - P$ of "weathering" on the surface of glass relics as:

$$P = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \tag{1}$$

$$1 - P = \frac{1}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \tag{2}$$

β_0 is the intercept term independent of the independent variable X_i, β_1、β_2、 ……、β_p is called partial regression coefficient. They reflect the contribution of independent variables X_1、X_2、X_3 to dependent variables respectively. The ratio of P to $1 - P$ is called advantage, and the logarithm of advantage can be obtained from the above formula:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3}$$

It is called linear logistic regression [6]. β_i indicates the natural logarithm of the odds ratio generated by the influence factor X_i changing one measurement unit by one when other independent variables are controlled at a fixed level. An important indicator to measure the contribution of independent variables is the odds ratio. In the case that other influencing factors are the same, the odds ratio's natural logarithm of two different values that c_1 and c_0 of the risk factor Xj is:

$$\ln(OR_j) = ln\left[\frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}\right] = \beta_j(C_1 - C_0) \tag{4}$$

The advantage ratio of this factor is:

$$\ln(OR_j) = e^{[\beta_j(C_1-C_0)]} \tag{5}$$

P_0 and P_1 represent the probability of "weathering" when X_j is c_0 and c_1. OR_j is the adjusted odds ratio, indicating the role of risk factor Xj after deducting the influence of other independent variables.

2.2. Fisher linear discrimination

Fisher linear discriminant analysis is a common classification model, which is mostly used to solve the problem of binary classification of high-dimensional data. Its main idea is to project high-dimensional data into low-dimensional space, so as to solve the problem of steep increase in computation due to high dimension. Figure 1 shows the basic principle of Fisher linear discriminant analysis.

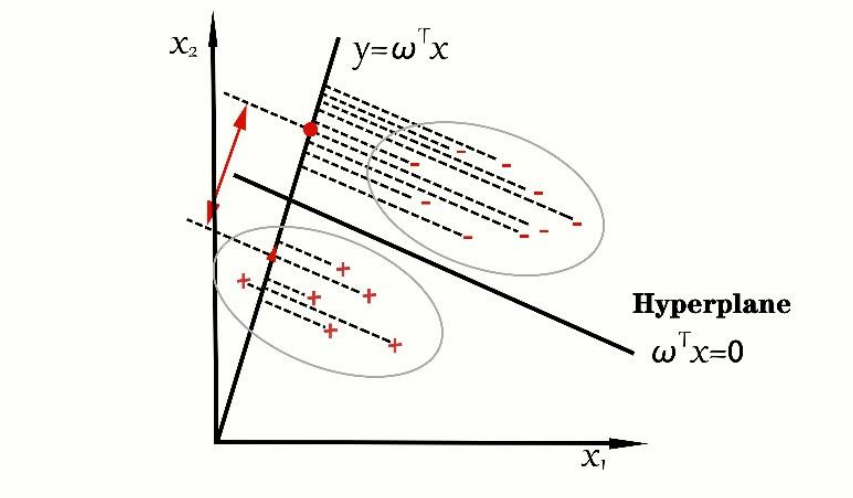

Figure 1. Basic principle of Fisher linear discriminant analysis

Given a data matrix $A = [\alpha_1, \alpha_2, \dots, \alpha_n]^T \in R^{n*n}$ where α_i is a m-dimensional vector, the data matrix A is divided into p matrices of class $A^T = [A_1^T, A_2^T, \dots, A_P^T]$, where $A_i \in R^{n*n}$, and $\sum_{i=1}^{p} n_i = n_0$. Calculating the optimal linear change $F \in R^{m*1}$ is the goal of traditional linear discriminant analysis, which can keep the class structure of the initial space in the low-dimensional space. Therefore, the optimal linear transformation (F) maps the vector α_i in each A in the m-dimensional space to the vector in the q-dimensional spaceβ_j,$F_i\alpha_i \in R^m \to \beta_j = F^T\alpha_i \in R(q < m)$. The two scattering matrices are defined as the following formula:

$$S_b = \frac{1}{n}\sum_{i=1}^{p}\sum_{\alpha\in A_i}(b_i - b)\,(b_i - b)^T = \frac{1}{n}\sum_{i=1}^{p} n_i(b_i - b)(b_i - b)^T \tag{6}$$

$$S_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i - \mathrm{\alpha})(\alpha_i - \alpha)^T \tag{7}$$

Where $b_i = \frac{1}{n_i}\sum_{\alpha\in A_i}\alpha$ is the average of i^{th} and $b = \frac{1}{n}\sum_{\alpha\in A}\alpha$ is the average of all data sets. In the low-dimensional space, the scattering matrix becomes:

$$S_b^Q = F^T S_b F,\ S_t^Q = F^T S_t F \tag{8}$$

The calculation result of scattering matrix is:

$$H_b = \frac{1}{\sqrt{n}}\left(\sqrt{n_1}(b_1 - b)\cdots\sqrt{n_p}\left(c_p - c\right)\right) \tag{9}$$

$$H_t = \frac{1}{\sqrt{n}}(A^{\top} - be^{\top})\,, e = [1, \dots, 1]^{\top} \in R^n \tag{10}$$

Then the scattering matrix becomes:

$$S_b = H_b H_b^T\ , S_b = H_b H_b^T\ ,\ S_t \tag{11}$$

The best transformation F can be obtained by using discriminant analysis for optimization. The calculation formula is:

$$\arg\ max = \left\{\text{trace}\left(\left(S_t^Q\right)^{-1} S_b^Q\right)\right\} \tag{12}$$

Based on the above discriminant function, the weight factor is introduced here to transform the fractional discriminant function into the differential discriminant function, and the model is established to maximize the difference. Using the classical discriminant method, we can get: $max\frac{F^T S_b F}{F^T S_t F}$,Introduce weight factor $\rho, \rho \in (0,1)$,we can get:

$$max = F^\top S_b F - (1-\rho)F^\top S_t F \tag{13}$$

After simplification, we can get the model:

$$max = F^\top[\rho(S_b + S_t) - S_t]F \tag{14}$$

It can be seen that the eigenvector corresponding to the maximum eigenvalue of $\rho(S_b + S_t) - S_t$ is F, and F is also the solution of the formula. Finally, the best value of ρ can be determined by the correct rate of back substitution, and the best discriminant function can be obtained [7].

3. Results

3.1. The establishment of the model

3.1.1 Logistic regression model

Maximum likelihood estimation is a common estimation method and can also be used to solve the parameters of the Logistic regression model. Its main idea is to construct the following sample likelihood function L.

$$L = \prod_{i=1}^{n} {P_i}^{Y_i} * (1-P_i)^{1-Y_i} \tag{15}$$

In the equation, P_i represents the probability of occurrence of "surface weathering" of the ith observation object. If the actual result is "surface weathering", then $Y_i = 1$; otherwise, $Y_i = 0$. Then, the maximum likelihood estimation principle is used to solve the parameters when the maximum value of the likelihood function L is obtained. Through observation and analysis, we use the logarithmic form of the function to simplify the calculation, and the formula is as follows:

$$lnL = \sum_{i=1}^{n}[Y_i lnP_i + (1-Y_i)\ln(1-P_i)] \tag{16}$$

In the case of large sample size, the confidence interval of β_i can be estimated with the following formula using the normal approximation method:

$$e^{(b_j \pm u_{\frac{a}{2}} S_b)} \tag{17}$$

To compare the contribution of each influencing factor in Y, the regression coefficient for each variable can be compared to obtain the contribution of "glass type, decoration, color"(physical characteristics) and various chemical components to surface weathering [8].

3.1.2 Fisher linear discrimination

In this paper, the known chemical composition is used as the prediction variable input, and the high potassium glass is set as 0, and the lead barium glass is set as 1, as the dependent variable output. The data are divided into a test group and a training group, in which the training group data are used to estimate the model, 80% of the total data set is randomly selected, and the remaining data are the test group for detection [9]. Then use SPSS software to analyze the chemical composition data of the sample surface, and obtain the Fisher linear discriminant analysis model. The model flow chart is shown in Figure 2.

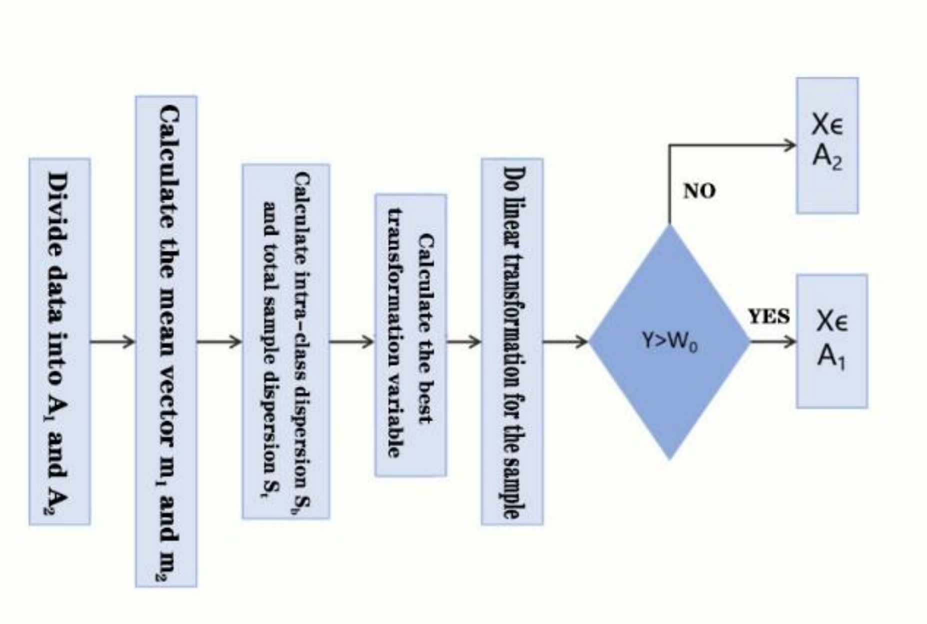

Figure 2. Flow chart of Fisher linear discriminant analysis model

Through the cleaning of experimental data, we selected 50 groups of data for simulation, of which 40 were training groups and 10 were test groups to test the accuracy of the model. Our training results are shown in Table.1.

Table 1. Fisher linear discriminant model training

Type			Forecast member information		total
			0	1	
Original	Count	0	13	0	13
		1	0	28	28
		Ungrouped cases	9	9	18
	%	0	100	0	100
		1	0	100	100
		Ungrouped cases	50	50	100

From Table.1, we can see that this model correctly classifies 100.0% of the original grouped cases. We use ten pre-selected groups of data to test the model, and the results are shown in Table.2.

Tabel 2. Fisher linear discriminant analysis model training

	Type	The probability of being 0	The probability of being 1
03 Part 2	0	1	0
06 Part 1	0	0.99998	0.00002
……		……	
55	1	0	1

It can be seen that the model can classify samples with high probability and high accuracy.

3.1.3 Sensitivity test model

Fisher exact test is a significant test method used in contingency table analysis. It can accurately calculate the significance of the deviation from the zero hypothesis, which greatly avoids the problem that other statistical tests can only be more accurate when the sample size is sufficiently large. Fisher exact test is often used in small sample problems, and often can get reliable results [10]. With the development of computer network technology, Fisher exact test can also be applied to large sample problems, with a wide range of applications. Build a $2 * 2$ contingency table as shown in Table.3.

Table 3. $2 * 2$ contingency table

	1	2	Sum
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Sum	n_{+1}	n_{+2}	n

In fact, for the Table.3, when the zero hypothesis that the edge and $(n_{1+}, n_{2+}. n_{+1}, n_{+2})$ are fixed and the contingency table is independent is true, the random variable n_{11} is subject to hypergeometric analysis, and the probability of random variable n_{11} is obtained by using mathematical statistics knowledge:

$$P = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}} \tag{18}$$

Where$m_- \leqslant n_{11} \leqslant m_+, m_+ = min(n_{1+}, n_{+1}), m_- = max(0, n_{1+}, +n_{+1} - n)$. The P value here represents the sum of the probabilities of the experimental results and the alternative hypothesis results. When the P value is less than the significance level, the independence assumption is rejected, and on the contrary, the independence assumption is accepted. In this paper, Fisher exact test is used to study the sensitivity of glass classification.

3.2. Analysis of experimental results

3.2.1 Logistic regression model

Through the calculation of three factors, the influence of different factors on the prediction results is explored. The results showed that the lack of "pattern" data had little impact on the prediction results, while the lack of "type" data had the greatest impact on the prediction results, followed by "color". We calculate the regression coefficient of variables and get the following Table.4.

Table 4. Variables in the equation

	B	Standard error	Wald	Free degree	Significance	Exp(B)
pattern	-0.0890	0.3410	0.0680	1.0000	0.7940	0.9150
type	1.6550	0.6720	6.0610	1.0000	0.0140	5.2310
color	-0.2230	0.2890	0.5970	1.0000	0.4400	0.8000
constant	-0.2380	1.0500	0.0510	1.0000	0.8210	0.7890

Substituting the correlation coefficient in the table into the equation, we can get:

$$P = \frac{e^{(-0.238+1.655X_1-0.089X_2-0.238X_3)}}{1+e^{(-0.238+1.655X_1-0.089X_2-0.238X_3)}} \tag{19}$$

We use the above equation to predict the known cultural relics samples, and the results are as Table.5.

Table 5. Sample weathering prediction

y_hat	Forecast result	Actual result
0.30171	0	0
0.70714	1	1
0.34042	0	0
0.34042	0	0
0.34042	0	0
0.32076	0	1

According to the comparison between the predicted results of the model and the actual results, our model has a high prediction success rate. It can be seen that the "type" factor has the greatest relationship with weathering, followed by "color" and "decoration". In terms of chemical composition, analyze the data to obtain the confidence level of the secondary interaction term of chemical composition. The results are shown in Table.6.

Table 6. Heteroscedasticity test

Surface2	Coef	P>|t|
$PbO \cdot BaO$	0.000367	0.688
$PbO \cdot P_2O_5$	0.002129	0.44
$PbO \cdot SiO_2$	-0.00027	0.535
$BaO \cdot P_2O_5$	0.00604	0.05
$BaO \cdot SiO_2$	-0.00073	0.497
$PbO \cdot CuO$	0.004762	0.228
$CuO \cdot SiO_2$	0.003817	0.277
$CuO \cdot P_2O_5$	0.006688	0.705

We test the heteroscedasticity and find that the P value is greater than 0.05, which means that we accept the original hypothesis at 95% confidence level, that is, we believe that there is heteroscedasticity in the disturbance term.

Due to the existence of the heteroscedasticity of the disturbance term, the constructed statistics are invalid. Here we eliminate the interaction items with heteroscedasticity. Use Stata software to program and calculate, introduce interaction terms according to the calculation results, and then use Logistic regression model to solve, we get the correlation coefficient of variables, as shown in the Table.7.

Table 7. Correlation coefficient of equation

	B	Standard error	Ward	Freedom	Significance	Exp(B)
Type	-135.185	52513.931	0.000	1.000	0.998	0.000
SiO2	2.076	7109.749	0.000	1.000	1.000	7.971
Na2O	-0.661	15622.465	0.000	1.000	1.000	0.516
…			…			
Constant	-141.700	715312.172	0.000	1.000	1.000	0.000

$$P = \frac{e^{(-141.7-135.185\text{Type}+2.076\text{SiO}_2+\cdots+13.24\text{SO2}-0.018\text{BaOSiO}_2)}}{1+e^{(-141.7-135.185\text{Type}+2.076\text{SiO}_2+\cdots+13.24\text{SO}_2-0.018\text{BaOSiO}_2)}} \tag{20}$$

Through equation analysis, it is considered that the type and SnO_2 have great negative correlation with weathering, MgO and SO_2 have great positive correlation, and other variables have weak correlation.

3.2.2 Fisher linear discrimination

In the solution of the Fisher linear discriminant analysis model, we believe that the model has reached a good classification level. Now the unknown data is substituted into the model for classification. The results are shown in Table.8.

Table 8. Classification results

Classification		The probability of being 0	The probability of being 1
A1	0	1	0
A2	1	0	1
A3	1	0	1
A4	1	0	1
A5	0	1	0
A6	0	1	0
A7	0	1	0
A8	1	0	1

By observing the data in Table.8, we have successfully classified the given samples. Among them, A1, A5, A6 and A7 are "high potassium glass", and the other samples are "lead barium glass".

4. Conclusions

4.1. Advantages of the model

Logistic regression model has very good interpretability. From the weight of features, we can see the impact of different features on the final result. At the same time, it is appropriate for binary classification problems. It is not necessary to scale the input features, but only needs to store the characteristic values of each dimension, which is very efficient. This paper eliminates the influence of heteroscedasticity on the results, so that all parameter estimates and hypothesis tests can be carried out normally.

Fisher linear discrimination can always find a projection direction for the generally linearly separable samples, so that the reduced samples are still linearly separable, and the separability is better, that is, the distance between samples of different categories should be as far as possible, and the same category should be as concentrated as possible [9]. In this paper, Fisher linear discriminant method is utilized to study the classification identification. The logic is rigorous and the calculation results are accurate.

Fisher's exact test can accurately calculate the significance of the deviation from the zero hypothesis, which greatly avoids the problem that other statistical tests can only be more accurate when the sample size is adequately large. This paper uses Fisher's exact test to study the sensitivity of glass classification, which can more accurately test the performance of the constructed model.

4.2. Deficiency and improvement of the model

In order to simplify the variables of the model, the physical factors and chemical factors that have less influence on the weathering of glass products are eliminated in this paper, which affects the accuracy of the solution results of the model. In addition, the selection of article samples is too single, and the data volume needs to be improved. In the subsequent experimental research, the data volume of training samples can be increased to enhance the accuracy of the model.

References

[1] http://www.chinaglassnet.com/info_main/2022920/43797.html

[2] Gu Liangliang, Zhou Jing. Several Issues on the Study of Ancient Chinese Glass Art [J]. Art and Design (Theory), 2021,2 (04): 112-114. DOI: 10.16824/j.cnki.issn10082832.2021.04.029.

[3] Hao Yunhong, Wu Rigen, Zhao Chengguang, Guo Xin, Yaruhan. Study on the impact of weathering on the erosion resistance of tempered glass based on the Taguchi method [J]. Silicate Bulletin, 2020,39 (09): 2980-2986. DOI: 10.16552/j.cnki.issn1001-1625.2020.09.038.

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

[5] Qi Ji, Gulistan, Wang Chengyu, Ning Guiling. Study on alkali evolution on the surface of silicate glass [J]. Glass and Enamel, 2006 (03): 9-13.

[6] Gao Bo, Chen Xiang, Shan Zaihe, Han Xiyue, Shan Yanlong, Yin Sainan, Yu Bo. Study on the probability of forest fire occurrence in the Greater Khingan Mountains region based on the logistic regression model [J]. China Science and Technology of Work Safety, 2022,18 (11): 163-168.

[7] Wang Jiaqing, Zhang Xin, Cao Tong, Wang Nengcai, Zhang Haiying. Prediction of breast microcalcification properties based on Fisher linear discriminant analysis [J]. China Medical Equipment, 2022,19 (02): 5-9.

[8] Xiong Yun. Research on weathering influence factors based on LOGISTIC and CART models [D]. Harbin Institute of Technology, 2010.

[9] Ding Xueli, Qi Changsheng, Fang Li. Classification and identification of traditional Chinese medicine based on Fisher discriminant analysis [J]. Journal of Chifeng University (Natural Science Edition), 2021,37 (11): 19-22. DOI: 10.13398/j.cnki.issn1673-260x.2021.11.006.

[10] Hua Yanfei, Sun Jianan, Yang Wuyue, Wang Pengchao, Chen Qiu. Exploring the influencing factors of gas accidents in coal mining enterprises using statistical modeling and testing [J]. Enterprise Reform and Management, 2016 (05): 185-187. DOI: 10.13768/j.cnki.cn11-3793/f.2016.0889.