Fisher's linear discriminant analysis model for inscribing the effect of chemical composition on cultural heritage

Zihan Wang *

Shandong University of Finance and Economic, Jinan, China

* Corresponding Author Email: 15106330077@163.com

Abstract. In this study, the relationship between weathering and glass type, decoration and colour was first analysed using Goodman and Kruskal's tau-y coefficients, and a chi-squared test was performed, which showed that weathering was most influenced by glass type, and was influenced by glass decoration but very little by glass colour. A Fisher linear discriminant analysis model was then established, and the eight chemical components with the greatest correlation to artefact type were calculated using SPSS software under the conditions of the two glass types respectively, and optimised using a grey correlation algorithm with specified thresholds for screening components, resulting in four final elements affecting high potassium glass as potassium oxide, magnesium oxide, iron oxide and copper oxide, and affecting lead and barium types as calcium oxide, aluminium oxide, lead oxide and barium oxide, aluminium oxide, lead oxide, barium oxide and strontium oxide.

Keywords: Cardinality test, Fisher's linear discriminant analysis model, Grey correlation algorithm, Threshold.

1. Introduction

As a historical witness to the trade between East and West, glass products are a valuable crystallisation of the Silk Road and an important historical basis for the study of ancient history and culture.

The ancient glass making process in China was based on the essence of the technology spread from the West in ancient times, and was improved locally to suit the local conditions, especially in the refining of lead ore, a common local material, as a flux to make our unique lead and barium glass. Local materials such as grass ash, which has a high potassium content, would have been chosen as a flux to make potassium glass.

Artefacts buried in the soil for thousands of years are susceptible to the effects of weathering. The chemical elements on the surface of the artefacts react strongly with the elements in the air and soil due to factors such as temperature or humidity, which can easily lead to oxidation and other chemical reactions that can affect the archaeologist's ability to accurately determine their category.

2. Materials and Methods

2.1. Data

In order to ensure the accuracy and completeness of the data, and in terms of the subsequent research for the collation and extraction of the data, the data in this study refer to the first question data in question C of the 2022 National Student Mathematical Modelling Competition (mcm.edu.cn). And the data were sorted and cleaned in a basic way. For example, it can be seen from the form that there are different heritage detection at the sampling point, different parts of the same heritage detection, the location of the detection point is complex, in order to maintain the independence and equality of the data and facilitate data processing, we will use each detection point as a data processing unit.

2.2. Introduction to the method

In this study, a correlation model was developed to measure the magnitude of the correlation between the data using Goodman's and Kruskal's tau-y coefficients, and a chi-square test was performed to investigate the relationship between colour, decoration and glass type on the weathering of the objects.

A table showing the percentage composition of the fourteen chemical components for different objects and for different parts of the same object is given in Annex Table 2. The table shows that the proportions of the fourteen chemical components vary according to the type of glass used. In this regard, we can first establish a Fisher linear discriminant analysis model, using SPSS software to calculate the correlation coefficients of the chemical composition on the type of artefact in each of the two glass types, prioritising the top eight influential chemical compositions as the more important factors, and taking the mean values of the remaining six less influential factors as secondary influences. In order to optimise the regression model, we can then use the grey correlation algorithm to refine the most important factors twice.

3. Model building and solving

3.1. Correlation analysis

The data in Form 1 shows that the surface weathering of glass artefacts and their glass type, decoration and colour are all definite class variables, so we can use the tau-y coefficient to investigate the correlation between whether the glass is weathered and its glass type, decoration and colour, and carry out a chi-square test [1][2].

(1) Principle of correlation analysis

The original hypothesis is that the correlation coefficient between glass decoration, glass colour and surface weathering is zero, and the alternative hypothesis is that the correlation coefficient between glass decoration, glass colour and surface weathering is not zero.

The tau-y coefficients of Goodman and Kruskal were used to measure the magnitude of the correlation between weathering and glass type, decoration and colour (at 90% confidence level) and the tau-y coefficients and significance levels of glass decoration, glass type and glass colour were derived using SPSS software.

Next, in order to demonstrate that the same conclusion exists for inferences to be drawn from random sample data to the overall population, a chi-square test was conducted on the data (at 95% confidence level).

(2) Solving for correlation coefficients and significance levels

We used the SPSS software to measure the magnitude of the correlation between the data (at 90% confidence level) for the 58 glass artefacts in Table 1, using Goodman's (Goodman) and Kruskal's (Kruskal) tau-y coefficients, with the following results.

Table 1. Correlation of glass properties with surface weathering

Nature of artefact	tau-y coefficient	tau-y coefficient significance level
Glass decoration	0.085	0.088
Glass Type	0.119	0.009
Glass colour	0.163	0.320

As can be seen from Table 1, the tau-y coefficient significance level of glass colour is greater than 0.1, and the original hypothesis is accepted, i.e. the correlation between glass colour and surface weathering is not strong; the tau-y coefficient significance level of glass decoration is close to 0.1, and the tau-y coefficient significance of glass type is significantly less than 0.1, then both glass decoration and glass type reject the original hypothesis, i.e. there is a strong correlation between glass type and surface weathering The correlation coefficient was 0.119, and the correlation between glass decoration and surface weathering was weak.

Using SPSS software, we conducted a chi-square test (90% confidence level) on the data and found that the p-values for both glass decoration and glass colour were greater than 0.1 and did not pass the significance test; the p-values for glass type were greater than 0.1 and passed the significance test.

The chi-square test further indicates that the surface weathering of glass artefacts has a strong correlation with glass type, a correlation with glass decoration but a weak correlation, and almost no correlation with glass colour.

3.2. Optimisation of Fisher's linear discriminant analysis model and grey correlation algorithm

In order to analyse the statistical pattern of the content of chemical components on the surface of artefact samples with or without weathering, we divided the glass into two categories: high-potassium glass and lead-barium glass.The content of each chemical component on the surface of the artefacts was analysed by Fisher's linear discriminant analysis [3-5] to obtain the main chemical components associated with weathering, and then the main chemical components were optimised according to a grey correlation algorithm to obtain the most important chemical components.

(1) Fisher's linear discriminant analysis model and the establishment of the grey correlation algorithm

① Fisher's linear discriminant analysis method is to find a linear combination that makes the difference between classes as large as possible and the variation within classes as small as possible, and to classify the data by finding a suitable hyperplane and placing different classes of data on either side of the hyperplane. A linear combination is defined as follows.

$$Z = C_1X_1 + C_2X_2 + \cdots + C_mX_m \tag{1}$$

where C_1 、 $C_2 \ldots C_m$ denotes the discriminant coefficient of each component and, X_1,$X_2 \ldots X_m$denotes the content of each chemical component. Z is the scoring criterion in the linear discriminant analysis method and satisfies the following conditions.

$$\begin{cases} Z_0 > Z_1, \text{Awarded as category 0} \\ Z_0 < Z_1, \text{ Awarded as category 1} \\ Z_0 = Z_1, \text{Awarded in either category} \end{cases} \tag{2}$$

The different glass artefacts are classified by comparing the magnitude of Z values. To find the relationship between each chemical component and whether it is weathered or not, we identify the main chemical component associated with whether it is weathered or not by specifying a threshold, i.e. when the value of the component is greater than the specified threshold, the chemical component is retained as the main chemical component.

We also verify the reasonableness of Fisher's discriminant by calculating the success rate of the predicted data.

Since the Fisher model, as a traditional regression model, has non-negligible shortcomings in the algorithm, such as the need for a large amount of data as the basis for analysis and the discrepancy between quantitative results and qualitative data, the above shortcomings lead to the traditional method being difficult to meet the computational needs of this paper, so we apply the above results to the secondary analysis using the grey correlation algorithm to compensate for the small amount of data and the qualitative data in the Fisher model.

② The basic idea of grey correlation analysis [6-8] is to determine whether a series of curves are closely related based on their similarity in geometry. We first made a line chart of the component series from known data to roughly show the primary and secondary relationships of the chemical components.

In order to obtain an accurate hierarchical relationship between the components, we need to first process the existing data, we set the sum of the components as the reference series and the rest of the chemical components as the comparison series. In order to de-scale, reduce the range of variables and simplify the calculation, the mean value of each chemical component of the reference and comparison series was calculated and each ratio of each chemical component was divided by the mean value, and then the grey correlation data was calculated according to the grey correlation formula, which is as follows.

$$\gamma(x_0)=\frac{1}{n}\sum_{k=1}^{n}(x_0(k)-x_i(k)) \tag{3}$$

where $\gamma(x_0(k)-x_i(k))=\frac{a+\rho b}{|x_0(k)-x_i(k)|+\rho b}$ is the discrimination coefficient, usually ρ taken as ρ= 0.5, i is the i-th chemical component of the sequence, k is the k-th element of the chemical component, a is the bipolar minimum difference, b is the bipolar maximum difference, and $a=\min_i\min_k|x_0(k)-x_i(k)|$ $b=\max_i\max_k|x_0(k)-x_i(k)|$, $|x_0(k)-x_i(k)|$ is the correlation coefficient. Finally, the grey correlation coefficients are compared and ranked to arrive at the most significant chemical components affecting the weathering of the artefacts.

(2) Solving the model

① Solving the Fisher linear discriminant analysis model

We transformed the weathered or unweathered column in Form 2 into a dummy variable, the 0-1 variable, where 0 represents no weathering and 1 represents weathering. The glass was divided into two categories, high potassium glass and lead-barium glass, and the 0-1 variables were subjected to Fisher linear discriminant analysis using SPSS with the 14 chemical components of the two types of glass, respectively, and the results are shown in Table 2.

Table 2. High potassium glass and lead-barium glass superflats

Chemical composition	High potassium glass	Lead barium glass
SiO_2	2.249	0.081
Na_2O	5.586	0.498
K_2O	2.303	-1.660
CaO	0.108	0.053
MgO	23.953	0.619
Al_2O_3	-1.734	0.503
Fe_2O_3	4.758	-0.109
CuO	5.465	0.208
PbO	6.905	0.207
BaO	-10.517	0.143
P_2O_5	-1.988	0.245
SrO	30.329	-3.069
SnO_2	-11.345	-0.643
SO_2	-40.589	0.188

The hyperplane [9-10] data for high potassium glasses are as follows:

ω_1 =(2.249,5.586,2.303,0.108,23.953,-1.734,4.758,5.465,6.905,-10.517,-1.988,30.329,-11.345,-40.589)

The hyperplane data for lead-barium glass are as follows:

ω_2=(0.081,0.498,-1.660,0.053,0.619,0.503,-0.109,0.208,0.207,0.143,0.245,-3.069,-0.643,0.188).

The discriminant coefficients for high potassium glass and lead-barium glass were calculated. In the high potassium glass group, the threshold value of 1800 was specified based on the analysis of the corresponding numbers of the two weathering groups, and when the composition values were greater than 1800, we retained them as the main chemical composition, whereby the following chemical compositions were obtained: sodium oxide, potassium oxide, magnesium oxide, iron oxide, copper oxide, lead oxide, strontium oxide and sulphur dioxide.

In the lead-barium glass group, according to the analysis of the weathering correspondence numbers of the two groups, the threshold value is specified as 30, when the value of the composition is greater than 1800, we retain it as the main chemical composition, according to which the chemical composition obtained is as follows: sodium oxide, calcium oxide, magnesium oxide, aluminium oxide, lead oxide, barium oxide, phosphorus pentoxide, strontium oxide and tin oxide.

② Solving for grey correlation

We first organised the data by glass type to facilitate a categorical study of the constituent components in relation to glass type.

Based on the annexed data and the processing needs of this question, we collated the following data for the 18 groups of artefact components under the high potassium glass type and created the following diagram accordingly.

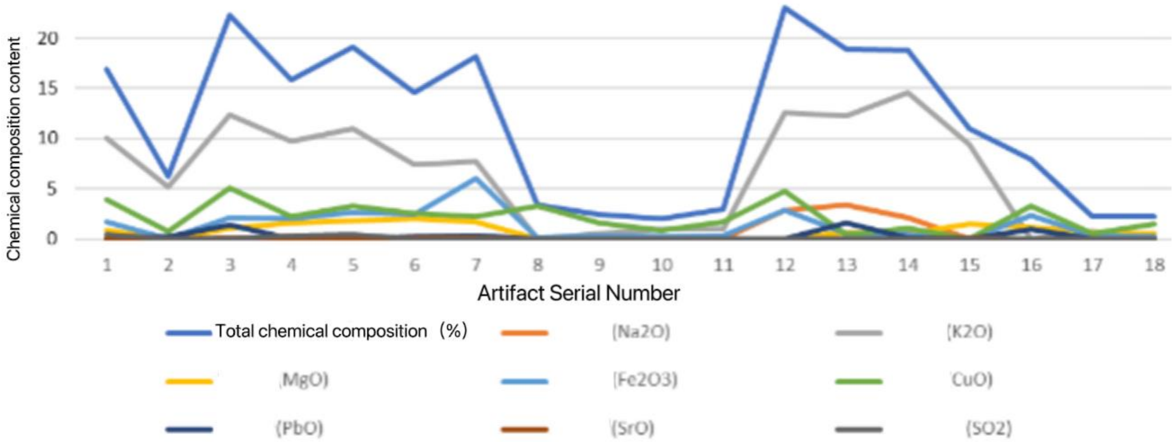

Figure 1. Sequence folding diagram

From Figure 1, we can roughly see that closer to the top blue line representing the sum of the components are the grey line representing potassium oxide, the green line representing copper oxide, the yellow line representing magnesium oxide and the blue line representing iron oxide. We can then roughly conclude that potassium oxide, magnesium oxide, iron oxide and copper oxide correlate well with the sum of the components.

To demonstrate the accuracy of the conclusions on the correlations obtained in the above figure, we need to further analyse and process the data, First we need to determine the analytical sequence.

a. Reference sequence: i.e. for the chemical composition sum column, noted as $x_0, x_0 = (x_0(1), x_0(2) \dots x_0(18))^T$. $x_i(k)$.is the kth element of the i-th column in the data table.

b. Comparison sequence: i.e. the eight chemical composition columns: sodium oxide, potassium oxide, magnesium oxide, iron oxide, copper oxide, lead oxide, strontium oxide, and sulphur dioxide, respectively, noted as $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$.

where

$$\begin{cases} x_1 = \left(x_1(1), x_1(2) \dots x_1(18)\right)^T \\ x_2 = (x_2(1), x_2(2) \dots x_2(18))^T \\ \dots\dots \end{cases} \tag{4}$$

In order to de-scale, reduce the range of variables and simplify the calculations, we pre-process the variables, i.e. we calculate the average value for each chemical component of the reference and comparison series and divide each ratio of each chemical component by the average value, based on the results of the calculations we can calculate the grey correlation between each component of the reference series and the reference series. After substituting the data into the grey correlation formula and ranking the results in order of magnitude we can obtain potassium oxide, magnesium oxide, iron oxide and copper oxide as the most dominant elements.

Based on the lead-barium glass data and repeating the above steps we can obtain grey correlations of 0.7286, 0.9328, 0.8196, 0.8683, 0.8537, 0.7545, 0.7709, 0.7189 for each element of lead-barium type glass, then the most dominant elements in lead-barium glass are calcium oxide, aluminium oxide, lead oxide, barium oxide and strontium oxide.

4. Conclusion

This study began with a chi-square test, which illustrated that surface weathering of glass artefacts has a strong correlation with glass type, a correlation with glass decoration but a weak correlation, and almost no correlation with glass colour. This is followed by solving the Fisher linear discriminant analysis model to obtain the main influencing chemical constituents, and then a grey correlation to further conclude that in high potassium glass potassium oxide, magnesium oxide, iron oxide and copper oxide are the most dominant influencing elements.In lead-barium glass, the most important factors are calcium oxide, aluminium oxide, lead oxide, barium oxide and strontium oxide.

References

[1] Wichitchan Supawadee, Yao Weixin, Yu Chun. A new class of multivariate goodness of fit tests for multivariate normal mixtures[J]. Communications in Statistics - Simulation and Computation,2022,51(11).

[2] Klass Michael J. An [formula omitted] asymptotically nearly minimax goodness of fit test[J]. Stochastic Processes and their Applications,2022,150.

[3] Sablik Maciej, Stapor Katarzyna. EXTENDED CAUCHY-SCHWARZ INEQUALITY AND ITS APPLICATION FOR THE TWO-CLASS FISHER DISCRIMINANT ANALYSIS[J]. STUDIA SCIENTIARUM MATHEMATICARUM HUNGARICA,2020,57(4).

[4] Aimin Miao, Fei Tao, Peng Li, Wenping Ren, Qiwei Guo. An improved Fisher discriminant analysis algorithm based on Procrustes analysis for adaptive fault recognition[J]. Measurement and Control,2019,52(7-8).

[5] Mingyang Zou, Caihong Sun, Shuang Liang, Yi Sun, Dexin Li, Ling Li, Lili Fan, Lijie Wu, Wei Xia. Fisher discriminant analysis for classification of autism Spectrum disorders based on folate-related metabolism markers[J]. The Journal of Nutritional Biochemistry,2018,64.

[6] Hu Yi Chung, Chiang Shu hen, Chiu Yu Jing. Applying Grey Relational Analysis to Detect Change Points in Time Series[J]. Journal of Mathematics,2022,2022.

[7] Shariff Jamilah Mohd, Mohamad Daud, Hasan Noorfahida. Sharia-compliant stock preference using similarity-based Fuzzy Grey Relational Analysis[J]. Journal of Physics: Conference Series,2021,1988(1).

[8] Determination of Influencing Factors for Slope Stability Using Grey Relational Analysis (GRA) Technique[J]. Systematic Reviews in Pharmacy,2020,11(04).

[9] Maatouk Hassan, Bay Xavier, Rullière Didier. A note on simulating hyperplane-truncated multivariate normal distributions[J]. Statistics and Probability Letters,2022,191.

[10] Rosen Julian. How to Tee a Hyperplane[J]. The American Mathematical Monthly,2022,129(8).