Analysis and Identification of the Composition of Ancient Glass Objects Based on Logistic Regression Analysis

Zhihao Chen *, Jingmin Ji

Westa college, Southwest University, Chongqing, China

* Corresponding Author Email: 282660764@qq.com

Abstract. Ancient glass objects were weathered by environmental influences, and the analysis of their main components and identification of their categories is a prerequisite for all subsequent research work. This paper uses Spearman's correlation coefficient and chi-square test of variance to analyses the correlation between weathering and type, color and decoration on the surface of cultural relics; secondly, descriptive statistical tests are conducted separately for two different types of glass objects to obtain the relationship between the changes in the content of each chemical component of different types of glass objects before and after weathering; finally, by analyzing the above statistical laws, a weighted average prediction model is established to The paper analyzed the classification patterns of two types of glass, namely high potassium and lead-barium, by establishing a logistic regression model; the Spearman coefficient and analysis of variance were used to investigate the correlation between surface weathering and each component. A principal component analysis of 14 different chemical components was carried out to achieve the effect of dimensionality reduction and to interpret the meaning of each principal component; a logistic regression model was established to regress the data after the principal component analysis, which showed that the classification criteria were more reasonable and the sensitivity of the model was high.

Keywords: Spearman's correlation coefficient, principal component analysis, logistic regression analysis.

1. Introduction

The Silk Road was a historical route of trade and communication across Asia and Europe, and a gateway for cultural exchange between China and the West in ancient times, when glass was introduced to China from the West Asian and Egyptian regions via the Silk Road [1-2].

Glass is mainly made from silica and due to its high melting point, co-solvents are often added to lower the melting temperature during refining. For example, lead barium glass often uses lead ore as a co-solvent, while high potash glass often uses grass ash as a co-solvent [3-4]. Throughout history, most of the glass that has been handed down from ancient times to the present day has suffered from weathering to varying degrees [5]. During the weathering process, elements from the external environment are exchanged with elements within the glass, causing changes in the proportions of their composition, which provides a challenge in identifying the type of glass. It is therefore particularly important to develop effective mathematical models to accurately analyses and identify the composition of ancient glass objects [6].

2. The basic fundamental of data processing analysis

The Spearman correlation coefficient [1, 8] was chosen to analyses the correlation between the surface weathering of glass artefacts and their type, color and ornamentation.

Suppose X_i and Y_i are two sets of data with Spearman correlation coefficients of

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

Where d_i is the grade difference between X_i and Y_i and n is the number of samples.

This paper performs a hypothesis test on Spearman's correlation coefficient, which is a hypothesis test in the case of a large sample due to the data size of the sample of 58. The statistic is constructed according to the method of equation (2) and the value of p is calculated [7].

$$r_s\sqrt{n-1} \sim N(0,1) \tag{2}$$

When $p < 0.05$, a significant difference is considered to exist, i.e., a correlation exists.

Table 1. Spearman's correlation coefficient table

	Ornamentation	Type	Color	Surface weathering
Ornamentation	1.000 (0.000***)	-0.357 (0.006***)	-0.433 (0.001***)	0.116 (0.384)
Type	-0.357 (0.006***)	1.000 (0.000***)	0.557 (0.000***)	0.344 (0.008***)
Color	-0.433 (0.001***)	0.557 (0.000***)	1.000 (0.000***)	-0.019 (0.885)
Surface weathering	0.116 (0.384)	0.344 (0.008***)	-0.019 (0.885)	1.000 (0.00***)

Note: ***, ** and * represent 1%, 5% and 10% significance levels respectively

As can be seen from Table 1, after hypothesis testing, when the significance level is 5%, the significance P value for type and surface weathering is 0.008, rejecting the original hypothesis, and the Spearman coefficient is 0.344, indicating a moderate positive correlation between glass type and surface weathering; while the color and decoration of the glass artefacts do not pass the hypothesis test, indicating that the color and decoration of the glass artefacts and surface weathering do not significant correlation with surface weathering [8].

Analysis of Variance (ANOVA) is a method used to test whether there is a difference between the experimental and control groups and whether the difference is significant. In this paper, ANOVA is used to test the variability between the weathering of the glass surface and its type, color and decoration. As the data examined were all qualitative variables, chi-square analysis was chosen [9].

Table 2. Results of the chi-square analysis

Title	Name	Surface weathering		χ^2	Correction	P
		No weathering	Weathering			
Color	Blue green	6	9	9.432	9.432	0.307
	Pale blue	8	12			
	Purple	2	2			
	Dark green	3	4			
	Deep Blue	2	0			
	None	0	4			
	Light green	2	1			
	Black	0	2			
	Green	1	0			
Type	High Potassium	12	6	6.88	5.452	0.009***
	Lead Barium	12	28			
Ornamentation	C	13	17	4.957	4.957	0.084*
	A	11	11			
	B	0	6			

Note: ***, ** and * represent 1%, 5% and 10% significance levels respectively

Based on the results in Table 1, with a significance level of 1%, the following conclusions can be drawn.

For surface weathering and color, the significance P value of 0.307 does not present significance at the level, so there is no significant difference between the surface weathering and color data.

For surface weathering and type, the significance p-value is 0.009***, showing significance at the level, so there is a significant difference between the surface weathering and type data.

For surface weathering and decoration, the significance p-value is 0.084*, which does not show significance at the level, so there is no significant difference between the surface weathering and decoration data.

A weighted average prediction model was developed. For a certain type of glass artefact, it is divided into two categories: weathered and unweathered, and the individual chemical components of its different groups of test samples are weighted and averaged, and the difference between the mean value of each chemical component of the unweathered test samples and that of the weathered test samples is obtained as the average consumption value of weathering for each chemical component, so the content of chemical components before weathering is equal to the content of each chemical component after weathering plus the average consumption value of each The average depletion value of the chemical components [10]. The results of the chi-square analysis are shown in Table 2.

For a given type of glass artefact, with a total of s chemical compositions and m sample data, the mean value of the i chemical composition before weathering is

$$\bar{K}_{1i} = \frac{1}{m}\sum_{k=1}^{m} x_{ik} \quad i = 1, 2, , ..., s \tag{3}$$

The mean value of the chemical composition of the first i after weathering is

$$\bar{K}_{2i} = \frac{1}{m}\sum_{k=1}^{m} y_{ik} \quad i = 1, 2, , ..., s \tag{4}$$

Therefore, it can be concluded that the average consumption value of the i chemical component is

$$S_i = \bar{K}_{1i} - \bar{K}_{2i} \quad i = 1, 2, , ..., s \tag{5}$$

For a given sample after weathering, the predicted value of the pre-weathering chemical composition of the i

$$P_i = z_i + S_i \quad i = 1, 2, , ..., s \tag{6}$$

Bringing in the data in Annex V.xlsx, the corresponding predictions can be derived using Python.

Three severely weathered points of the test samples and randomly added weathering test points of high potassium type glass artefacts were selected to predict their chemical composition content prior to weathering. In the process of data analysis, the large number of independent variables will increase the difficulty and complexity of the problem analysis and also reduce the accuracy of the model, so this paper performs principal component analysis [3] to reduce the dimensionality of the data.

First assume that the original data has n samples and p indicators, which can form a sample matrix of size $n \times p$ x:

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix} \tag{7}$$

Calculate the correlation coefficient matrix for the sample matrix x R:

$$R = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}} \tag{8}$$

Where the mean value is $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.

Subsequently, the eigenvalues and eigenvectors of R were calculated and the eigenvalues were used to calculate the contribution margin $contribution$ and the cumulative contribution margin $sumcontribution$:

$$contribution = \frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k} \quad (i = 1, 2, ..., p) \tag{9}$$

$$sumcontribution = \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k} \quad (i = 1, 2, ..., p) \tag{10}$$

In practice, the first, second, ..., and m $(m \le p)$ principal components corresponding to the eigenvalues whose cumulative contribution exceeds 80% are generally taken, so that the i principal component can be expressed as

$$F_i = a_{1i}X_1 + a_{2i}X_2 + ... + a_{pi}X_p \quad (i = 1, 2, ..., m) \tag{11}$$

Using the data processed by MATLAB, the operation of principal component analysis is performed, when the cumulative contribution is greater than 80%, the number of indicators corresponds to 6, so the number of principal components is taken to be 6, i.e. $m = 6$, and the processed principal component coefficient data is saved followed by the original data brought into the principal component expression, and the calculated results obtained are saved in, as samples for the subsequent logistic regression analysis and training of neural networks in this paper data.

Logistic regression models [2] are generalized linear regression models in which the dependent variable is a qualitative variable, i.e., a categorical variable, and often the predicted outcome follows a 0-1 distribution. The dependent variable is considered to be the probability of the event occurring, and is considered to have a predictive value of 1 when it is greater than 0.5 and 0 when it is less than 0.5.

This paper examines the classification patterns of two types of glass artefacts, high potassium and lead barium, where the dependent variable is a categorical variable and the usual multiple linear regression model is not applicable in this case, so a logistic regression model is chosen to analyses the classification patterns of glass artefacts.

The connection function was selected as the $Sigmoid$ function, which is shown in the following form.

$$F(x, \beta) = S(x_i^{'}\beta) = \frac{\exp(x_i^{'}\beta)}{1 + \exp(x_i^{'}\beta)} \tag{12}$$

Where the independent variable $x_i = (x_{1i}, x_{2i}, ..., x_{ki})$ corresponds to the principal component factor of F_i, and the regression coefficient $\beta = (\beta_0, \beta_1, ..., \beta_k)$, so the logistic regression model developed is

$$\hat{y}_i = \frac{\exp(x_i^{'}\hat{\beta})}{1+\exp(x_i^{'}\hat{\beta})} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+...+\hat{\beta}_k x_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+...+\hat{\beta}_k x_{ki}}} \tag{13}$$

If $\hat{y}_i \geq 0.5$, the prediction is considered to be $y=0$; otherwise, the prediction is considered to be $y=1$.

3. Results

3.1. The establishment of simulation model

This paper analyses the pattern of change in the chemical composition of the surface of the artefacts, refines the information on the sampling points and examines the statistical patterns of the chemical composition of the sampling points before and after weathering.

3.2. Analysis of experimental results

The interpretation of the significance for the six principal components derived is shown in Table 3 below.

Table3. Interpretation of principal components

Principal Components	Significance
F_1	This principal component has relatively high positive loadings on barium oxide, lead oxide and strontium oxide, mainly reflecting the effect of heavy metal oxides on the classification of glass artefacts
F_2	This principal component has relatively high positive loadings on calcium oxide, magnesium oxide and iron oxide, mainly reflecting the effect of the more reactive metal oxides on the classification of glass artefacts
F_3	This principal component has relatively high positive loadings on copper oxide and Sulphur dioxide, mainly reflecting the effect of basic and acidic oxides on the classification of glass artefacts
F_4	This principal component has a high negative loading on oxidation na, mainly reflecting the weighting factor of the active metal oxides in all chemical components
F_5	This principal component has a high negative loading on tin oxide, mainly reflecting the weight factor of inactive non-metallic oxides in all chemical compositions
F_6	This principal component has high positive loadings on potassium oxide and calcium oxide, mainly reflecting the effect of active metal alkaline oxides on the classification of glass artefacts

Regression analysis of the data calculated from the principal component analysis was performed using Python to obtain the logistic regression equation.

$$\hat{y}_i = \frac{e^{-1.23+0.18x_{1i}+0.01x_{2i}-0.47x_{3i}-0.29x_{4i}+0.08x_{5i}-0.32x_{6i}}}{1+e^{-1.23+0.18x_{1i}+0.01x_{2i}-0.47x_{3i}-0.29x_{4i}+0.08x_{5i}-0.32x_{6i}}} \tag{14}$$

The principal components F_3 and F_6 have large negative regression coefficients, indicating that high potassium glass artefacts have relatively large values for F_3 and F_6, i.e. larger values for the principal components of F_3 and F_6 have a higher probability of being high potassium glass artefacts and, conversely, a higher probability of being lead-barium glass artefacts.

After comparative testing, it was found that the model tested 73% correct and could better describe the classification pattern of both high potassium and lead-barium glass artefacts, but the accuracy was lower if the model was used as a predictive analysis.

4. Conclusion

The classification model based on glass artefacts can be extended to the predictive analysis of non-glass artefact categories by determining the content of certain chemical components on their surfaces or inside them; in addition, it can be extended to the classification model of artefact categories, not limited to a certain type (e.g. glass, ceramics, etc.), by entering the values of the main chemical composition parameters to predict which broad category the artefact belongs to (e.g. bronze artefacts).

This paper uses Spearman's correlation coefficient to correlate surface weathering with its type, decoration and color, and then uses the chi-square test for variance to analyses the differences. The average content of each chemical constituent was then used to predict the content of each chemical constituent at the weathering site before weathering. Then, a principal component analysis was carried out on the 14 chemical components to calculate the cumulative contribution of each component, and the components with a cumulative contribution greater than 85% were extracted.

References

[1] Chen Meigu, Lin Xing'e, Li Xinguo, Liu Saki Di, Gao Hongmao, Ming Jianhong, Dai Minjie, Zhou Zhaoxi. A comprehensive evaluation of durian quality based on principal component analysis and cluster analysis [J]. Food Industry Science and Technology: 1-19.

[2] Du Manru. Research on quantitative stock selection based on kernel principal component analysis and fuzzy C-mean clustering algorithm [D]. Nanjing University of Technology, 2020.

[3] Gao Yun, Yu Zhifang. Quality evaluation of celery based on principal component analysis [J]. Food Industry Science and Technology, 2020, 41(03): 308-314+320.

[4] Liang Zhuo. Research on complex two-dimensional irregular part ranking algorithm based on principal component analysis [D]. Yanshan University, 2019.

[5] Hu Jianping, Li Ling, Xie Qi, Zhang Daochang. A new image segmentation method for aerial photography of glass insulators [J]. Journal of Northeastern Electric Power University, 2018, 38(02): 87-92.

[6] WANG Jianchao, CHEN Zhifeng, LIU Xinming, CHEN Ting, GUO Linrong. Analysis and evaluation of nutritional composition of different varieties of Yu Ganzi fruit [J]. Journal of fruit trees, 2018, 35(01): 108-117.

[7] Wang R. Principal component analysis of primitive porcelain from Hulushan site and related issues [D]. Xiamen University, 2017.

[8] Zhao Qinghua. Research on near-infrared identification technology of plastics for end-of-life vehicles based on principal component analysis [D]. Shanghai Jiaotong University, 2014.

[9] Ben Youhong. An empirical study on the construction of a financial early warning model for listed companies in China based on principal component analysis [D]. Suzhou University, 2005.

[10] Yasujiro Yamada, Hisashi Kawano, Mamoru Murata. A low dilution ratio glass fused bead method for X-ray fluorescence analysis of major and trace constituents in rocks [J]. Journal of Analytical Testing Instruments, 1996, (02): 22-29.