Composition analysis and identification of ancient glass objects based on Spearman correlation analysis model

Jinghang Zhou#, Kexin Xu#, Taoyu Xiang#, Yajing Yu*

School of Economics and Management, Hunan University of Technology, Yueyang, China
*Corresponding author: yxu95@njfu.edu.cn

#These authors contributed equally.

Abstract. With the rising fervor of Belt and Road, it has drawn attention and research to the ancient Silk Road commodity trade exchanges, of which ancient glass products are valuable physical evidence of the early commodity trade exchanges. Now the chemical composition of ancient glass is analyzed and identified, which helps to find out the key and know-how of ancient glass products refining technology. In this paper, we establish a chi-square test and Spearman correlation analysis model to quantitatively analyze the relationship between weathering and type, decoration and color to establish a weathering interval estimation model, so as to predict the chemical composition content before weathering. Logistic dichotomous classification model, BP neural network model, BP neural network model optimized by genetic algorithm and random forest classification prediction model were established respectively to predict three high potassium glasses and five lead-barium glasses out of eight compounds.

Keywords: Cardinality test, Interval estimation prediction model, Logistic dichotomous model, BP neural network model.

1. Introduction

Since ancient times, the Silk Road has been an important hub for China to connect the world and a solid bridge for economic, cultural and trade exchanges between the East and West. After a century of transmigration, from the ancient Silk Road to the present-day Belt and Road, the Silk Road has once again taken on new life, creating an open, inclusive and inclusive framework for regional economic cooperation, benefiting a wider range of regions and people, including countries along the route, with the results of joint construction [1]. As the Silk Road was a channel for cultural and economic trade exchanges between China and the West in ancient times, during which the ancients conducted a large amount of commodity trade through the Silk Road, ancient glass is a valuable physical evidence of early commodity trade exchanges. Through the Silk Road, foreign glass manufacturing technology and techniques were introduced to China, and ancient glass makers in China absorbed and improved their technology, relying on local materials to make ancient glass products, making ancient glassmaking technology greatly improved. As a result, they were similar in appearance to foreign glass products, but their chemical composition was different[2,3].

It is important to analyze and identify the composition of ancient glass, and to clarify the chemical composition ratio of ancient glassmaking to improve modern glassmaking technology[4]. At the same time, it also helps to examine the ancient trade and economic exchanges and human conditions, and is of great archaeological significance for the study of ancient economic trade and cultural exchanges.

This paper analyzes the relationship between the surface weathering of glass artifacts and their glass type, decoration and color; analyzes the statistical pattern of the chemical composition content of artifact samples with and without weathering based on the type of glass, and predicts their chemical composition content before weathering based on the weathering point detection data. The chemical composition of the unknown category of glass artifacts is analyzed to discern its type, and then its classification results are analyzed.

2. Model hypothesis and notation

2.1. Hypothesis[5]

Hypothesis 1: It is assumed that the topic measurement data are accurate and have fidelity.

Hypothesis 2:Assume that the calculation error is within a reasonable range during the data processing, and the effect of the error is negligible.

Hypothesis 3: It is assumed that in the process of glass weathering, the change of chemical composition is only considered natural weathering factors, and no other artificial or accidental factors are considered.

Hypothesis 4: The collected relevant information is true and valid, and has a certain degree of credibility.

2.2. Notations

Important notations used in this paper are listed in Table 1.

Table 1. Notations

Symbols	Description
χ^2	Cardinality statistic
f_0	Observed value frequency
f_e	Expected value frequency
ρ	Spearman correlation coefficient
M_e	Median
s	Standard deviation
$t_{a/2}$	t-distribution
h_w	Range of predicted values
W_1	Input layer weight matrix
W_2	Output layer weight matrix

3. Model construction and solving

3.1. Analysis of surface weathering of glass artifacts

In order to do the relevant quantitative analysis, the underlying data need to be defined, and then the next quantitative analysis, the results of the data definition are shown in Table 2[6].

Table 2. Definition of data results

Basic information about glass artifacts	Type	Definition
Ornament	A	1
	B	2
	C	3
Type	High Potassium	1
	Lead Barium	2
Color	Blank	0
	Black	1
	blue-green	2
	Green	3
	light blue	4
	Light Green	5
	Dark Blue	6
	dark green	7
	Violet	8
Degree of surface weathering	No weathering	0
	Weathered	1

3.1.1 Descriptive statistical analysis

In order to investigate the relationship between the surface weathering of glass artifacts and their type, decoration and color, descriptive statistical analysis was performed on the underlying data. Finally, it was found that the presence or absence of weathering was mainly related to decoration B and glass type, and not much related to its color.

3.1.2 Establishing a chi-square test model

The χ^2-test can be used to determine the degree of correlation between two categorical variables, where f_0 is used to denote the observed frequency and f_e is used to denote the expected frequency, then the χ^2-statistic formula is.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \tag{1}$$

3.1.3 Building Spearman correlation models

The Spearman correlation, also known as the Pearson correlation between rank variables, can handle any form of variable that satisfies any distribution and has the following "expression".

$$\rho_{x,y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{2}$$

3.1.4 Establishing a pre-weathering prediction model

The average is the result obtained by adding a set of data and dividing it by the number of data, and its formula is.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3}$$

The median is a set of data sorted in the middle of the value of the variable, denoted by the median, which divides all the data into two parts, each part containing 50% of the data.

The standard deviation is the square root of the variance, which can better reflect the dispersion of the data, where the variance is the average of the values of the variables and their mean spread squared, and the formula for calculating the standard deviation is.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{4}$$

In summary, the statistical law that can be obtained by solving the model is that silica is the main influencing factor on the presence or absence of weathering of high potassium glass and lead-barium glass, while tin oxide and sulfur dioxide have no influence on the weathering of both glass types and can be neglected in the subsequent analysis.

3.1.5 Predicting the chemical content before weathering

(1) Descriptive analysis of compounds before and after weathering

Before making predictions, a simple descriptive analysis was performed by plotting histograms of the content of 14 compounds before and after weathering for high potassium glass and lead-barium glass, and the results are shown in Figure 1 below.

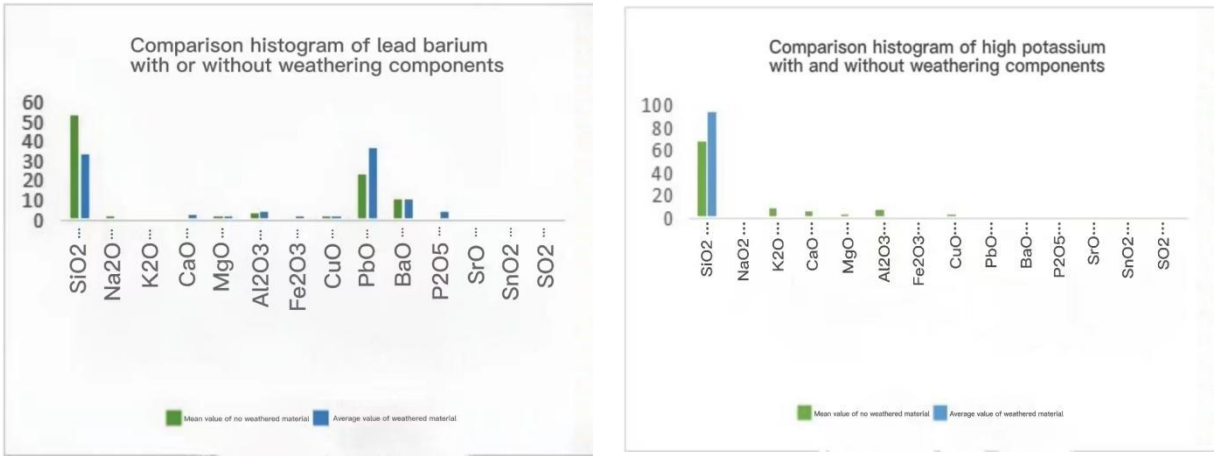

Figure 1. Histogram of compound content before and after weathering of lead-barium and high potassium

The chemical composition of lead barium glass, silica, decreases after weathering, and the content of lead oxide increases. There is no change in the content of barium oxide, so it is clear that barium oxide has no effect on the weathering of lead-barium glass. The silica content of high potassium glass increases after weathering. The mean values of potassium oxide, calcium oxide, and alumina content decreased significantly, indicating that these chemical components have a strong influence on the weathering of high potassium glass, followed by the prediction of the compound content before weathering.

(2) Establishing interval estimation model

Interval estimation is based on confidence coefficients for parameter estimation. By means of a sample of unweathered artifacts, a sample drawn from the overall population, an appropriate confidence interval is constructed according to certain requirements of correctness and accuracy, in order to serve as an estimate of the range in which the true value of the distribution parameter (or a function of the parameter) of the overall population lies. When the overall obeys a normal distribution and the overall variance is unknown, the overall variance is replaced by the sample variance in the case of a small sample, and then a confidence interval for the overall mean at the 1- confidence level is established based on the distribution with the following formula.

$$\bar{x} \pm t_{a/2} \frac{s}{\sqrt{n}} \tag{5}$$

The chemical composition content of unweathered glass then changes over time so that the glass is weathered, so to make predictions about the pre-weathering data is to make interval predictions about the chemical composition content of unweathered glass so that the pre-weathering data can be demonstrated to be within this interval. Therefore, this paper needs to combine the unweathered data of both glass types to make predictions of the pre-weathering chemical composition content. That is, the unweathered data are used as the predicted value to estimate the interval where the pre-weathering content lies. Firstly, this paper needs to classify the glass into two types: high potassium and lead barium, then screen the unweathered and weathered data separately, and then use Excel to estimate the interval of the unweathered data for the next 5 periods, and predict the interval where the compound content and the pre-weathering compound content are located for the next 5 periods, i.e., within this interval is the predicted interval of the pre-weathering compound content, and take the average value of the weathered data and compare it with the The weathering data are averaged and compared with the predicted values.

Take SiO2 as an example for illustration: the confidence interval predicted based on the unweathered data of SiO2 five periods back is [0.61 0.95], and the SiO2 content of high potassium glass is rising after weathering, so weathering occurs after 0.696, which means that the content of SiO2 before weathering should be between [0.696 0.95], and the predicted value of 0.78 is exactly in the interval before weathering. The predicted value of 0.78 is within the interval before weathering,

so it can be used as the predicted value of the chemical composition content of high potassium glass before weathering, and the data after that in turn.

The same prediction and analysis was then performed for the unweathered lead-barium glass.

Take SiO2 as an example for illustration: the confidence interval predicted based on the unweathered data of SiO2 five periods back is [-0.61 1.43], and the SiO2 content of lead-barium glass is decreasing after weathering, so weathering will occur before 0.718, indicating that the content of SiO2 before weathering should be between [-0.61 0.718], and the predicted value of 0.40 is exactly in the interval before weathering. interval, so it can be used as the predicted value of chemical composition content of lead-barium glass before weathering, and the data in the following order.

3.2. Judgment of the type to which the cultural relic belongs

3.2.1 Building a Logistic Dichotomous Model

Logistic regression is a generalized linear regression analysis model, by exploring the relationship between the 14 compounds and the type of glass, to determine the probability that the type of glass is high potassium or lead barium, so it is necessary to set a threshold value, when the calculated probability is greater than this threshold value, the computer determines that the sample belongs to high potassium, when the calculated probability is less than the threshold value, the computer determines that the sample belongs to lead barium, this time it is necessary to use the Sigmoid function, whose calculation formula is as follows[7].

$$h_w = \frac{1}{1+e^{-z}}$$

$$\begin{gathered} z = \omega^T x \\ x = [1, x_1, x_2, \ldots, x_n] \\ z = w_0 + w_1 x_1 + \ldots + w_n x_n \end{gathered} \tag{6}$$

The data were solved using MATLAB, using 89% of the data of known types as training values and 11% of the data of unknown types to predict the type to which they belong. The classification results are calculated as shown in Table 3.

Table 3. Logistic dichotomous classification results table

Artifact Number	Prediction Type	Probability of occurrence	Comparison of thresholds	Output results
A1	High Potassium	0.607	>0.5	1
A2	Lead Barium	0.219	<0.5	0
A3	Lead Barium	0.116	<0.5	0
A4	Lead Barium	0.029	<0.5	0
A5	Lead barium	0.173	<0.5	0
A6	High Potassium	0.830	>0.5	1
A7	high potassium	0.709	>0.5	1
A8	high potassium	0.545	>0.5	1

3.2.2 Building a BP neural network model

$x_1, x_2, \ldots, x_n$ is the input variable, y is the output variable, inside the positive direction is the input and output, inside the circle is the neuron, after the neuron are to pass the activation function, if the activation function is sigmod function, the mapping relationship is[8]

$$f(x) = \frac{1}{1+e^{-x}} \tag{7}$$

MATLAB was used to solve the data, using 89% of the data of known types as training values and 11% of the data of unknown types to predict the types to which they belong. The number of neurons

in the hidden layer is set to 8, the content value of 14 compounds is input, and the type to which the glass belongs is output. Matlab calculates the threshold between the output layer and the hidden layer as -0.0372, and if the predicted value is greater than the threshold, the type is 1 (high potassium), and if the predicted value is less than the threshold, the type is 0 (lead barium), and the classification results are calculated as shown in Table 4.

Table 4. BP neural network classification results table

Artifact Number	Prediction Type	Probability of occurrence	Comparison of thresholds	Output results
A1	High Potassium	1.696	>-0.0372	1
A2	Lead Barium	-0.045	<-0.0372	0
A3	Lead Barium	-0.901	<-0.0372	0
A4	Lead Barium	-0.826	<-0.0372	0
A5	Lead barium	-0.670	<-0.0372	0
A6	High Potassium	1.666	>-0.0372	1
A7	high potassium	1.433	>-0.0372	1
A8	high potassium	1.023	>-0.0372	1

The results are consistent with those of logistic binary classification regression. A graph of its loss function and prediction results can also be obtained for the BP neural network model, as shown in Figure 2 below.

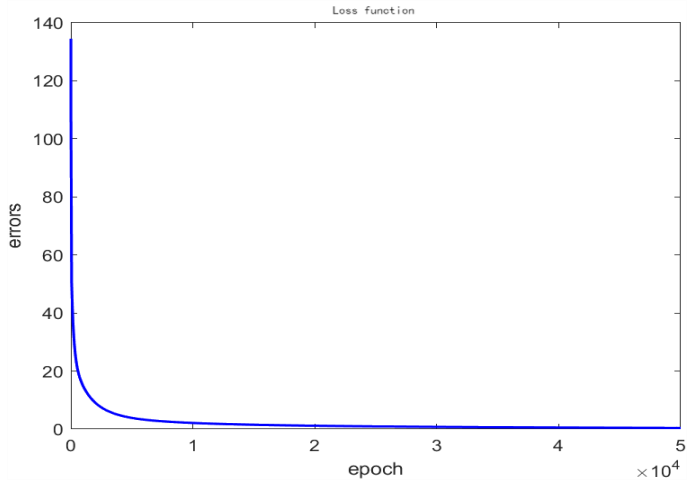

Figure 2. Loss function diagram

3.2.3 Building a BP neural network model optimized by genetic algorithm

Genetic algorithm is a model of biological evolutionary process that simulates the mechanism of natural selection and genetics of Darwin's theory of biological evolution. The algorithm converts the problem solving process into a process similar to the crossover and mutation of chromosomal genes in the biological evolution process by means of mathematical large and using computer simulation[9].

$$\begin{aligned} A_1 &= \text{tansig}(W_1 * \text{X} + \text{b}_1) \\ \hat{y} &= W_2 * A_1 + \text{b}_2 \\ fitness &= (\hat{y} - y)^2 \end{aligned} \tag{8}$$

MATLAB was used to solve the data, using 89% of the data of known types as training values and 11% of the data of unknown types to predict the type to which they belong. The number of neurons in the hidden layer is set to 8, the content values of 14 compounds are input, and the type to which the glass belongs is output. Matlab calculates the threshold value as 0, and if the predicted value is greater than the threshold value, the type is 1 (high potassium), and if the predicted value is less than the threshold value, the type is 0 (lead barium), and the classification results are calculated as shown in Table 5.

Table 5. Table of classification results of BP neural network based on genetic algorithm optimization

T	Prediction Type	Probability of occurrence	Comparison of thresholds	Output results
A1	High Potassium	1.017	>0	1
A2	Lead Barium	-0.999	<0	0
A3	Lead Barium	-1.000	<0	0
A4	Lead Barium	-0.989	<0	0
A5	Lead barium	-1.045	<0	0
A6	High Potassium	1.014	>0	1
A7	high potassium	0.975	>0	1
A8	lead barium	-0.967	<0	0

When the trained data is iterated 30 times, its adaptation degree reaches the optimum of 108, the adaptation, degree is larger, indicating that the probability of individual optimum inheritance to the next generation is also larger, and the error between the predicted and actual values is smaller.

3.2.4 Building a random forest classification prediction model

Random forest refers to a classifier that uses multiple trees to train and predict samples. A random forest is a classifier that contains multiple decision trees and whose output is determined by the plurality of the classes output by the individual trees. In essence the regression tree is a constant bifurcation of the original feature space, and multiple spaces are obtained when different feature bifurcations are satisfied, and finally the mean value of the samples in the space to which it belongs is used as the predicted value for that space[10].

Using MATLAB to solve the data, 89% of the data of known types are used as training values, and 11% of the data of unknown types are predicted to be the types they belong to. The random forest classifier was created and the glass types of 8 artifacts were obtained through simulation tests. Matlab calculated the threshold value of 0.5, and if the predicted value was greater than the threshold value, the type was 1 (high potassium), and if the predicted value was less than the threshold value, the type was 0 (lead barium), and the predicted results are shown in Table 6.

Table 6. Table of prediction results based on random forest classification

Artifact Number	Prediction Type	Probability of occurrence	Comparison of thresholds	Output results
A1	High Potassium	0.781	>0.5	1
A2	Lead Barium	0.224	<0.5	0
A3	Lead Barium	0.286	<0.5	0
A4	Lead Barium	0.198	<0.5	0
A5	Lead barium	0.219	<0.5	0
A6	High Potassium	0.911	>0.5	1
A7	high potassium	0.895	>0.5	1
A8	lead barium	0.034	<0.5	0

3.2.5 Errorability analysis

The accuracy of the model needs to be further tested based on the analysis, therefore the accuracy of the results of the four models are summarized separately and the results are shown in Table 7 below.

Table 7. Accuracy of the results of the four model runs

Serial number	Classification Models	Accuracy of running results
1	Logistic binary classification model	0.781
2	BP neural network model	0.795
3	Genetic algorithm optimized BP neural network model	0.822
4	Random forest classification prediction model	0.804

4. Conclusion

This paper analyzes the relationship between surface weathering of glass artifacts and their types, ornamentation and color, and further explores the statistical pattern of chemical composition content with and without weathering on the surface according to the glass types, and predicts the chemical composition content before weathering based on the weathering point data. On this basis the data in this paper were analyzed to discriminate the types to which they belonged, and sensitivity analysis was performed on the results of the classification.

After grouping the data, descriptive statistical analysis was performed to visually and clearly show the relationship between glass weathering and ornamentation, type and color. To avoid the influence of subjective evaluation on the real situation, this paper continued to use chi-square test and Spearman correlation analysis to scientifically and reasonably determine the correlation relationship between glass weathering and ornamentation, type and color, followed by using interval estimation to predict the unweathered after Then, we used interval estimation to predict the estimated values and confidence intervals for the unweathered period, and compared them with the weathered data to accurately predict the chemical composition content of the glass before weathering, which made the whole model establishment more reasonable and accurate.

Logistic classification prediction model, BP neural network model, BP neural network model optimized by genetic algorithm and random forest model were established to use the data as training set and prediction set, so as to predict the type to which the artifacts belonged, and further compare the accuracy of the operation results of the four models, so as to obtain the optimal classification results and make the thinking deep and comprehensive.

References

[1] Gan Fu Xi. The ancient Silk Road and ancient Chinese glass[J]. Journal of Nature, 2006, 28(5):9.

[2] Wang Jie, Li Mo, Ma Qinglin, et al. Weathering study of a Warring States period octagonal lead-barium glass vessel[J]. Glass and enamel, 2014, 42(2):8.

[3] Sichin Bilig, Li Qinghui, Gan Fuxi. Analysis of ancient Chinese potassium glass components by laser exfoliation-inductively coupled plasma-atomic emission spectrometry/mass spectrometry[J]. Analytical Chemistry, 2013, 41(9):6.

[4] Chen Shuyu, Hou Zhili. An analysis of ancient silk road and ancient Chinese glass[J]. China Ethnic Expo, 2019(5):2.

[5] Jiang Qiyuan. Mathematical models (2nd ed.) [M]. Higher Education Press, 1987.

[6] Zhuo, Jinwu. Applications of MATLAB in mathematical modeling [M]. Beijing University of Aeronautics and Astronautics Press, 2011.

[7] Rong Zirong, Ma Anqing, Wang Zhikai, Zhou Kai.Analysis on Driving Forces of Liaohe Estuary Wetland Landscape Pattern Change Based on Logistic [J]. Environmental Science and Technology, 2012,35 (06): 193-198.

[8] Hou Beiping, Lu Pei. BP Neural Network Modeling and System Simulation Based on MATLAB [J]. Automation and Instrumentation, 2001, (01): 36-38.

[9] Min Xilin, Liu Guohua. Application of Artificial Neural Network and Genetic Algorithm in Modeling and Optimization [J].Computer Application Research, 2002, (01): 79-80.

[10] Li Xinhai. Application of random forest model in classification and regression analysis [J]. Journal of Applied Entomology, 2013, 50(4): 1190-1197.