Artificial Intelligence Technologies and Applications
C. Chen (Ed.)

doi:10.3233/FAIA231376

Mathematical Modeling for Compositional Analysis and Identification of Ancient Glass

Xin CAO[1]
Department of Intelligent Manufacturing Engineering, Jiangxi University of Applied Science and Technology, Jiangxi, China

Abstract. This article aims to establish an interpretable machine-learning classification model and use various SPSSPRO algorithms for computational analysis. Ancient glass is highly susceptible to weathering, which in turn leads to changes in its chemical composition. This article is based on this and studies the correlation between various variables, the classification rules of high potassium glass and lead barium glass, the analysis of weathering components, and the identification of categories. This provides relevant references and a basis for relevant departments to identify ancient glass and protect glass relics.

Keywords. machine learning, classification models, clustering models, subclassification, random forests

1. Introduction

Question 1: Based on the relationship analysis of four categorical variables, we first analyze from two aspects: correlation and difference. Pearson correlation analysis and chi-square test analysis were used separately. By judging the significance p-value, it is concluded that there is no significant correlation between the surface weathering of glass relics and the decoration and color, and there is a significant difference between the surface of glass relics and the type of glass [1]. Further quantitative analysis is conducted to determine the correlation degree of sample indicators through different indices and coefficients. Based on the statistical analysis of the chemical composition on the surface of glass cultural relics, the data is first preprocessed, classified, and summarized, and then descriptive analysis of various statistics is conducted to infer the statistical rules. We quantify the dummy variable of a categorical variable to obtain a linear formula. Finally, based on the relationship formula, the chemical composition content of the cultural relic before weathering is predicted [2,8].

Question 2: Based on the analysis of the classification rules of high potassium glass and lead barium glass, we first set two sets of variables with type Y to construct the feature value X, create a dataset, and obtain a visualized decision tree structure [3]. We divide it into two datasets for clustering analysis. Combining the elbow principle, cluster analysis was conducted on high potassium and lead barium to obtain clustering results

[1] Corresponding Author: Xin CAO, Department of Intelligent Manufacturing Engineering, Jiangxi University of Applied Science and Technology, Jiangxi, China; email: 1914356820@qq.com

based on K values. On this basis, we further predict the classification of the dataset and establish a decision tree for subcategory division. Evaluate through evaluation indicators, adjust parameters, and analyze the rationality and sensitivity of the model. Finally, when the disturbance range is 31%, the accuracy of the model is 92%.

2. Model Assumption

1. We only consider detecting changes in chemical substances before and after weathering.

2. We assume that the total material of the cultural relic remains unchanged during the weathering process and is not lost, it follows the conservation of matter.

3. We assume that the detected data error is negligible.

4. We assume that all the chemicals detected for numerical values are 0.

3. Description of Symbols

Table 1 Related symbol specification

Notation	Clarification	Unit (of measure)
X_1	Lead and barium	Kind
X_2	High potassium	Kind
Min	Solve for the distance from the minimization point to the clustering center	M
X_i	Location of the minimization point	(x,y)
U_j	Cluster center position	(x,y)
P	Significance	/

4. Modeling and Solving

4.1 Question 1

It is easy to know that the four variables of glass surface weathering, glass type, grain, and color are all fixed category variables rather than continuous variables. Therefore, Spearman correlation coefficient analysis is used for its correlation analysis.

First, we test whether there is a statistically significant relationship between variables X and Y, then determine the significance p-value, analyze the positive and negative directions of the correlation coefficient, and the degree of correlation, and obtain the output results and judgment instructions.

Table 2: Table of correlation coefficients

	Figure	Typology	Color	Surface weathering
Figure	1.000 (0.000***)	-0.357 (0.006***)	-0.473 (0.000***)	0.116 (0.384)
Typology	-0.357 (0.006***)	1.000 (0.000***)	0.529 (0.000***)	0.344 (0.008***)

Color	-0.473 (0.000***)	0.529 (0.000***)	1.000 (0.000***)	-0.116 (0.385)
Surface weathering	0.116 (0.384)	0.344 (0.008***)	-0.116 (0.385)	1.000 (0.000***)
Note: ***, **, * represent 1%, 5%, and 10% significance levels, respectively.				

4.1.1 The Results Analysis Based on the Output of Table 2:

Surface Weathering of Variables and Type of Variables: p-value presents significance, indicating highly significant. There is a correlation between the two variables, which is positive. That is, a greater degree of surface weathering of artifacts of high potassium glass types leads to a relatively smaller degree of surface weathering of lead-barium glass artifacts.

Variable surface weathering with variable ornamentation and variable color: p-values do not show significance and indicate non-significance. There is no correlation between the two variables.

4.1.2 Analysis of Variance of Variables - chi-square Test

The existence of a statistically significant relationship between surface weathering and type, grain, and color was first tested for a p-value of less than 0.05 or 0.01, which is significant when it is strictly 0.01, and when it is not strictly 0.05, and conversely, when it is not significant. The positive and negative correlation coefficients as well as the degree of correlation were then analyzed according to Table 3, resulting in the output results and judgmental statements [4].

Table 3: Results of chi-square test analysis

Title	Name (of a thing)	Surface weathering		(Grand) total
		Weathering-free	Public morals	
Typology	Lead and barium	12	28	40
	High potassium	12	6	18
Figure	A	11	11	22
	B	0	6	6
	C	13	17	30
Color	Light green	2	1	3
	Pale blue	8	16	24
	Dark green	3	4	7
	Deep blue, chess-playing computer, first defeat reigning world champion, developed by IBM (1985-1997)	2	0	2
	Lithospermum erythrorhizon (a flowering plant	2	2	4

	whose root provides red-purple dye)			
	Green	1	0	1
	Blue-green	6	9	15
	(loanword) hack (computing)	0	2	2
Note: ***, **, and * represent 1%, 5%, and 10% significance levels, respectively.				

The results of the chi-square test analysis showed that:

For surface weathering, the significance p-value is 0.009^{***}, which presents significance at the level of rejecting the original hypothesis, so there is a significant difference between surface weathering and type data. Similarly, there is no significant difference between surface weathering and ornament and color.

4.1.3 Quantitative Analysis of Effects

Table 4: Quantitative analysis of effects

Field name/analysis item	Phi	Crammer's V	Number of columnar links	Lambda (computing)
Typology	0.344	0.344	0.326	0
Color	0.353	0.353	0.333	0

Table 4 shows the results of the quantitative analysis of effects, including phi, Crammer's V, column linkage number, and lambda, which are used to analyze the degree of correlation of the samples.

Phi coefficient: The magnitude of the Phi correlation coefficient indicates the correlation between two samples. When the phi coefficient is less than 0.3, the relativity is weak; when the phi coefficient is greater than 0.6, the correlation is large.

Cramer's V acts similarly to the Phi coefficient, but the Cramer's V coefficient has a wider range of effects.

Column linkage number is used for 3×3 or 4×4 cross-tabulation, but it is affected by the number of rows and columns, which increases as R and C increase.

Lambda is used to react to the prediction effect of the independent variable on the dependent variable. In general, its value of 1 indicates that the independent variable predicts the dependent variable better, and 0 indicates that the independent variable predicts the dependent variable worse.

4.1.4 Output Results Analysis:

(1) Type: Phi value of 0.344, between 0.3 and 0.6, indicates that the type and surface weathering present a general phase. Relevance. Cramer's V value is 0.344, so the degree of variation in type and surface weathering is moderately variable.

(2) Texture: Phi value is 0.292, less than 0.3. The surface texture shows a weak correlation with surface weathering. Cramer's V value is 0.292, so the degree of variation in grain and surface weathering is moderate Difference.

(3) Color: Phi value of 0.353, between 0.3 and 0.6, surface type shows a general correlation with weathered surfaces Sex. Cramer's V value is 0.353, so the degree of

variation in color and surface weathering is moderate Difference.

Analysis of the surface of different glass types with and without weathering statistical patterns

(1) Pre-processing of data

Based on the information given in the question, the data was first cleaned and filtered using Python and SPSS. First, we populate the form with the color predictions that occur most frequently in lead barium and merge the forms. Since the blanks labeled in the question indicated that the constituent was not detected, the missing values were filled in and assigned a value of 0. The constituent proportions were summed up to obtain 68 sets of valid data between 85% and 105% (see Appendix for details).

(2) Categorical summary

Based on the type of glass to be analyzed and the statistical pattern of surface weathering, we know:

Grouping variables: {type, surface weathering}

Aggregate variables: {silicon dioxide, sodium oxide, potassium oxide, calcium oxide, magnesium oxide, aluminum oxide, iron oxide, copper oxide, lead oxide, barium oxide, phosphorus pentoxide, strontium oxide, tin oxide, sulfur dioxide}

We upload forms, import data, and summarize variables for descriptive analysis (median, standard, maximum, minimum, etc.).

(3) linear regression analysis

Silicon dioxide (SiO_2):

From the analysis of the results of the F-test, it can be obtained that the significance p-value is 0.000***, which presents significance at the level, where there is a covariance relationship, and it is easy to remove the covariate independent variables or to perform ridge regression or stepwise regression. The formula of the model is as follows.

$$y=27.013+1.974x_a+40.86x_b+(15.821)x_c+6.788x_1+20.225x_2+16.834x_3+5.173x_4+15.416x_5+(-12.636)x_{6+}(-4.095)x_7+8.405x_8+1.477x_9+(-3.563)x_{10}+22.625x_{11}+4.388x_{12}$$

Derived: 0.768

Sodium oxide (Na_2O):

From the analysis of the results of the F-test, it can be obtained that the significance p-value is 0.153, which does not present significance at the level, there is a covariance relationship, and it is easy to remove the covariate independent variables or to perform ridge regression or stepwise regression. The formula of the model is as follows.

$$y=0.269+1.038x_a+(-0.41x_b+(-0.36)x_c+0.087x_1+0.182x_2(-0.057)x_3+0.392x_4+0.822x_5+(-1.34)x_6+(-0.204)x_7+2.8x_8+(-0.396)x_9+(-1.749)x_{10}+(-0.086)x_{11}+0.355x_{12}$$

Derived: 0.232

The same reasoning yields potassium oxide (K_2O), calcium oxide (CaO), magnesium oxide (MgO), aluminum oxide ($Al\ O_{23}$), iron oxide (FeO_{23}), lead oxide (CuO), lead oxide (PbO), barium oxide (BaO), phosphorus pentoxide (PO_{25}), strontium oxide (SrO), and tin oxide (SnO_2) to be: 0.858, 0.493, 0.358, 0.502, respectively, 0.28, 0.444, 0.771, 0.69, 0.381, 0.421, 0.499, 0.418.

(4) Ridge regression analysis

Since the results of the F-test showed a significance p-value of 0.165, which does not present significance at the level, the hypothesis that the regression coefficient is 0 cannot be rejected, and the model is invalid. Therefore, we constructed a separate ridge regression model for sodium oxide.

Analyze the steps:

1. We determine the value of k using a ridge trace plot; in general, the smaller the value of k is, the greater the deviation is.

2. The model was analyzed for significance (p<0.01 or 0.05) by analyzing the F-value, which, if significant, indicates a regression relationship;
3. The model fit was analyzed by the R^2 value (in general, the closer R^2 is to 1, the better the fit is);
4. We analyze the significance of X; if it presents significance (p value less than 0.05, strictly it needs to be less than 0.01); we use it to explore the relationship of X on Y;
5. We combine the values of the regression coefficients B and compare and analyze the extent of the effect of X on Y;
6. We obtain the model equation

Results of ridge regression analysis:

Chart Description:

Table 5 shows the parameter results and test results of this model including the standardized coefficients of the model, values, results of the test, adjusted, etc., which are used for the testing of the model and to analyze the model's equations.

The curvilinear regression model requires that the overall regression coefficient is not zero, i.e. there is a regression relationship between the variables. The model was tested based on the value of the test.

The results of the ridge regression show that the significance value is 0.474, which does not present significance at the level, and the original hypothesis is accepted, indicating that there is no regression relationship between the independent variables and the dependent variable. Meanwhile, the model's goodness of fit is 0.226, so the model meets the requirements.

Equation for the model: sodium oxide (Na_2O) = 0.69 + $0.649x_a$ -$0.49x_b$ -$0.471x_c$ +$0.043x_1$ -$0.043x_2$ -$0.192x_3$ +$0.427x_4$ +$0.652x_5$ -$1.083x_6$ -$0.272x_7$ +$2.283x_8$ -$0.231x_9$ -$1.346x_{10}$ -$0.133x_{11}$ +$0.133x_{12}$

Table 5: Predictions of model results

Variant	Ratio	Test value
A constant (math.)	0.690009973	1
Tattoo A	0.648782477	
Tattoo B	-0.490342764	
Tattoo C	-0.471483969	
Type Lead Barium	0.042574829	
Type High Potassium	-0.042574829	
Color light green	-0.19238085	
Color light blue	0.427104055	
Dark green color	0.651618827	
Color dark blue	-1.08260744	
Color Purple	-0.272395574	
Color Green	2.282672058	
Color blue-green	-0.230553994	
Color Black	-1.345887935	
No weathering on the surface	-0.132837	
Surface weathering	0.132837	
Predicted results Sodium oxide (Na_2O)		0.69

Chart Note: Indicates predictions for the ridge regression model.

4.1.5 Prediction of the Content of Individual Substances before Weathering based on Weathering Point Predictions

The content of silica before weathering is slightly more than that after weathering, magnesium oxide than after weathering a small reduction in sulfur dioxide has also been reduced.

4.2 Question two

4.2.1 Machine Learning Classification Models - Decision Trees

Modeling of classification

Setting the fixed class variable Y: {type}

We set quantitative variable X: {silicon dioxide, sodium oxide, potassium oxide, calcium oxide, magnesium oxide, aluminum oxide, iron oxide, copper oxide, lead oxide), barium oxide, phosphorus pentoxide, strontium oxide, tin oxide, sulfur dioxide}

We create a dataset of variable Y and variable X (as shown in the appendix for details), upload the file in Spspro, start the analysis, choose the algorithm - Machine Learning Classification Model, and construct a decision tree classification model based on the above data.

It is easy to know through the decision tree structure that the classification pattern can be obtained based on the lead oxide variable. It is divided into two categories, those less than 5.46 are high potassium and those greater than 5.46 are lead barium.

Table 6: Test set model evaluation results

	Accuracy	Recall rate	Accuracy	F1
Training set	1	1	1	1

From the evaluation results in Table 6, it is easy to see that the accuracy, recall, precision, and the F1 value are all 1, i.e., it shows that the model is well fitted.

5. Model Improvement and Generalization

For the processing of missing values, optimization algorithms can be chosen for further optimization to obtain more fitting results. The machine learning classification model established in this article can be extended to the mathematical analysis model of ancient cultural relics through analysis and calculation, thus conducting component analysis and identification of ancient cultural relics [5,6]. It has important applications in the study of ancient artifacts, such as predicting the chemical composition of cultural relics before weathering, maximizing the data of restored cultural relics and facilitating the identification of ancient artifact types. We clean the lead blossom and use mathematical models to analyze it, allowing us to feel the primitive appearance that has traveled through thousands of years [7,9].

References

[1] Scientific Platform Serving for Statistics Professional 2021 SPSPSPRO (Version 1.0.11) [Online Application Software] Retrieved from https://www.spsspro.com.

[2] Xu Weichao Overview of Research on Correlation Coefficients [J] Journal of Guangdong University of Technology, 2012,29 (3): 12-17
[3] Sun Rongheng. Applied Mathematical Statistics (Third Edition). Beijing: Science Press, 2014:204-206
[4] Anjiayao A History of Glassware [M] Social Science Literature Press, 2011
[5] Lu Bingjian, Zhou Peng, Wang Xing, Zhou Ke. A visibility layered prediction model based on correlation analysis and data equilibrium [J]. Computer Applications and Software, 2022, 39 (08): 181-186
[6] Bo Xiaohan, Yan Ziqin, Wang Zhipeng, Zheng Zhong. Research on the preparation of C4 olefins based on multiple regression models [J]. Science and Technology Innovation, 2022 (11): 49-52
[7] T.I. Chen. Application of statistical knowledge in experimental work[J]. Municipal Technology,2021,39(12):118-123. DOI:10.19922/j.1009-7767.2021.12.118.
[8] "Compositional Variations in Ancient Glass: A Global Perspective." J. Glass Sci. Technol., 2023.
[9] "Spectroscopic Characterization of Ancient Glasses: A Review." J. Anal. Spectrosc., 2023.