Compositional analysis of glass products based on XGBOOST

Zhiyang Li [1, *], Yu Chen [2], Wenxuan Pan [1], Wentao Kuang [3], Manfang Hu [2]

[1] School of Electronic Information and Automation, University of Science and Technology of Tianjin, Tianjin, China, 300222

[2] School of Chemical Engineering and Materials, University of Science and Technology of Tianjin, Tianjin, China, 300457

[3] School of Mechanical Engineering, University of Science and Technology of Tianjin, Tianjin Municipality, 300222

* Corresponding author: a17730209318@163.com

Abstract. This paper presents a model for composition analysis and identification of glass products. Based on the XGBOOST algorithm, the XGBOOST model was constructed to find the type of glass, ornamentation, and color in the weathering and non-weathering points. The main component of glass products is SiO2, which imparts the desired functional properties by mixing different oxides. Ancient glass products are very susceptible to weathering caused by the buried environment, resulting in changes in the proportions of the various components inside. Since glass products of different materials have different properties, the classification of glass products is worth studying.

Keywords: Glass products, Classify, Differential, XGBOOST model.

1. Introduction

There are many kinds of ancient Chinese glass, each with its characteristics, which not only has artistic and practical value, but its historical and scientific value is more worthy of attention. Ancient glass products are witnesses of history and carriers of civilization [1]. The study of ancient glass is of great significance for excavating the value of cultural relics and inheriting Chinese civilization [2]. This paper will study the chemical composition of these glass cultural relics, analyze the differences in the composition and content of different types of glass, and establish a model to solve the problem of identification and identification of cultural relics [3].

2. Materials and methods

2.1. Data

The data in this article is derived from Mathematical Modeling Question C (http://www.mcm.edu.cn/) in 2022. Form 1 is missing 4 sets of color data, which are removed in order not to affect the model analysis. Form 2 gives the corresponding proportion of the main components, with some missing values filled with 0. Considering the valid data required by the title, the two groups of cultural relics sampling points No. 15 and No. 17 were finally screened as invalid data and rejected.

2.2. Introduction to the method

For datasets containing n m-dimensions, the XGBOOST model [4] can be represented as:

$$\hat{\mathbf{y}}_i = \sum^{k} f_k(x_i),\ f_k \in F(i = 1, 2...n) \tag{1}$$

$$F = \{ f(x) = w_{q(x)} \}(q : R^m \longrightarrow \{1, 2, ..., T\}, w \in R^T) \tag{2}$$

Where (6) is the set of decision tree structures [5], and when constructing the model, it is necessary to find the optimal parameters according to the principle of objective function minimization to

establish the optimal model. The model's objective function can be divided into the error function term L and the model complexity function term Q. The objective function can be written as:

$$\hat{L} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{3}$$

$$\Omega = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

$$Obj = L + \Omega \tag{5}$$

When using the training data to optimize the training of the model [6], you need to keep the original model unchanged and add a new function f to the model to reduce the objective function as much as possible, the specific process is:

$$\hat{y}_i^{(0)} = \mathbf{0} \tag{6}$$

$$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i) \tag{7}$$

$$\hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(x_i) \tag{8}$$

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{9}$$

Where $y_i^{(t)}$ is the predicted value of the tth model and ft (xi) is a new function added to tth. The target function is represented as:

$$Obj^{(t)} = \sum \left(y_i - (y^{(t-1)^2} + f_i(\hat{x}_i)) \right)^2 + \Omega \tag{10}$$

In this algorithm, to quickly find the parameters that minimize the objective function, the approximate objective function is obtained by second-order Taylor expansion:

$$Oby^{(t)} \approx \sum_{i=1}^{n} (y_i - y^{(t-1)})^2 + 2(y_i - y_i^{(t-1)}) \hat{f}_t(x_t) - \hat{h}_t f_t^{\,2}(x_i)) + \Omega \tag{11}$$

When the constant term is removed, the objective function is expressed as:

$$Obj^{(t)} \approx \sum_{i=1}^{n} [g_i w_{q(x_i)} + \frac{\mathbf{1}}{2} h_i w_{q(x_i)}^2] + \frac{\mathbf{1}}{2} \sum_{j=1}^{T} w_j^2 \tag{12}$$

$$= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_i + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{13}$$

If q is known, the objective function can be used to find the optimal wj and obtain the optimal objective function value. Its essence can be classified as the minimum solution problem of quadratic functions. Solution:

$$w_j^* = -\frac{\sum_{i=I_j} \boldsymbol{g}_i}{\sum_{i=I_j} h_i + \lambda} \tag{14}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i=I_j} \mathbf{g}_i)^2}{\sum_{i=I_{j,}} h_i + \lambda} + \gamma T \tag{15}$$

Among them, Obj can be used as the scoring function [7] of the evaluation model, and the smaller the value, the better the model effect. By recursively calling the above tree establishment method, a large number of regression tree structures can be obtained, and the optimal XGBOOST model can be established by using Obj to search for the optimal tree structure and put it into the existing model.

3. Model building and solving

3.1. Problem analysis

The topic requires analyzing the correlation between the chemical composition of different types of glass artifacts and comparing the differences.

Based on the above analysis, we can see that the chemical composition is not only related to the type of glass, the content is different, but also affects the surface weathering. The four situations before and after the differentiation of the two types of glass cultural relics were classified and discussed, and after preliminary scatter plotting between the index variables, it was found that there was a linear relationship between the variables, which met the premise requirements of Pearson's correlation coefficient calculation. The Pearson correlation [8] coefficient is used to illustrate the correlation between different chemical components, and the difference is analyzed by the solution of the model.

3.2. The composition of the model

(1) Assume that two sets of data sum, the specific formula is as follows:

Sample mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{16}$$

Sample covariance:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{17}$$

$$\sigma_x = \sqrt{\frac{(X_i - \bar{X})}{n}} \tag{18}$$

$$\sigma_y = \sqrt{\frac{(Y_i - \bar{Y})}{n}} \tag{19}$$

Overall correlation coefficient:

$$\rho_x = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{20}$$

Sample correlation coefficient:

$$\gamma_{xy} = \frac{\sum_{k=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{21}$$

By calculating the size of the correlation coefficient between the indicators, the correlation between the correlation coefficients is judged according to the correlation coefficient interval table, as follows in Table 1:

Table 1. Correlation Judgment

correlation	negative	correct
No correlation	-0.09~0.00	0.00~0.09
Weak correlation	-0.3~-0.1	0.1~0.3
in relevance	-0.5~-0.3	0.3~0.5
Strong correlation	-1.0~-0.5	0.5~1.0

(2) Hypothesis testing

The null hypothesis is proposed: correlation coefficient, alternative hypothesis. By constructing a distribution statistic, this article selects a distribution statistic as follows:

$$t = \gamma_{xy}\sqrt{\frac{n-2}{1-\gamma_{xy}}} \tag{22}$$

By proving, it is obtained that obeys the distribution of degrees of freedom. The P value between the indicators can be calculated to reject the null hypothesis at a 99% confidence level, and if it rejects the null hypothesis at a 95% confidence level [9], it cannot reject the null hypothesis anyway.

3.3. Solvingthe model

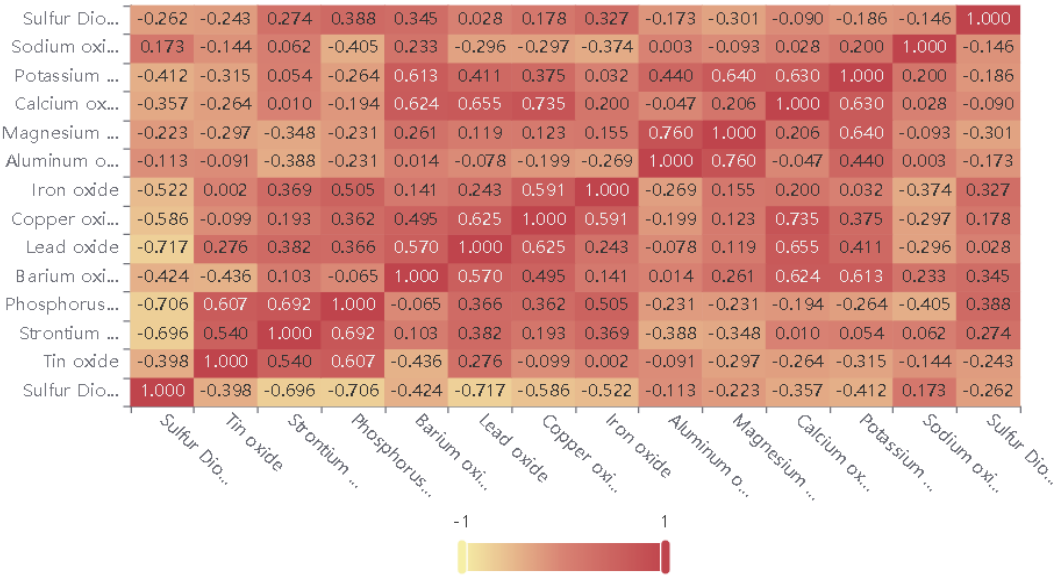

Figure 1. Heat map of the correlation coefficient between chemical components before weathering of high-potassium glass

According to the results, at a 99% confidence level, there was a strong positive correlation between strontium oxide, phosphorus pentoxide, and magnesium oxide, barium oxide and lead oxide, phosphorus pentoxide, and iron oxide had a strong positive correlation, and strontium oxide had a strong positive correlation between strontium oxide and lead oxide, calcium oxide, and sodium oxide.

Similarly, we use Pearson correlation analysis to obtain heat maps of correlation coefficients between chemical components under different conditions after weathering of high potassium glass, before weathering of lead barium glass, and after weathering of lead barium glass in Figure 1.

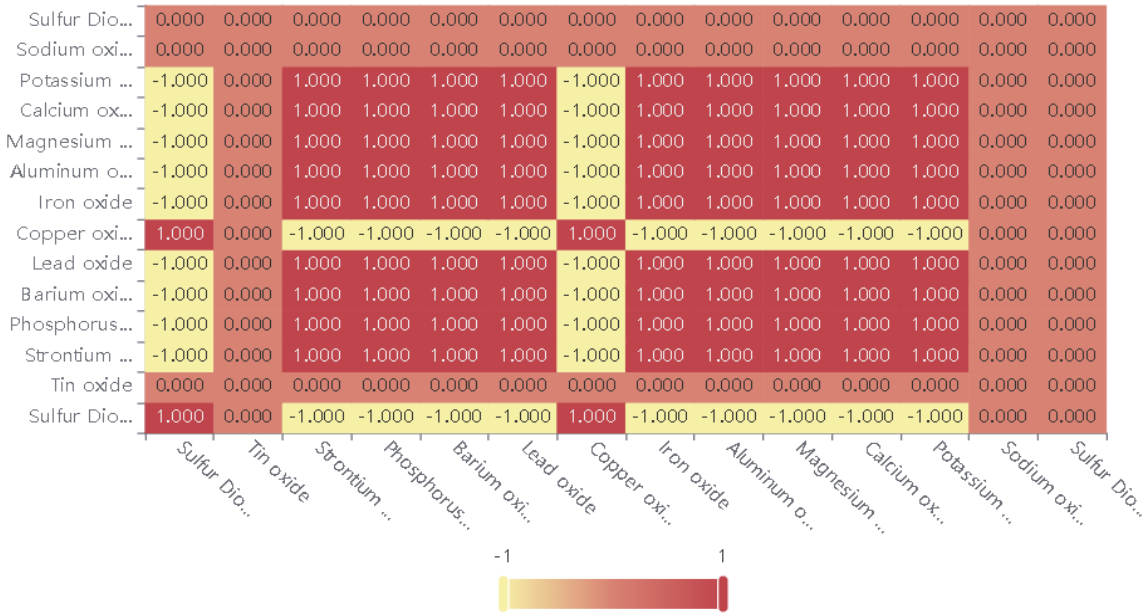

Figure 2. Correlation coefficient heat map between chemical components after weathering of high-potassium glass

It can be seen from the heat map in Figure 2 of the correlation coefficient between the chemical components of high-potassium glass weathering that the chemical components such as strontium oxide, phosphorus pentoxide, barium oxide, and copper oxide have a strong negative correlation with silica, while potassium oxide, calcium oxide, magnesium oxide, alumina, and iron oxide have a strong

positive correlation, and copper oxide, lead oxide, barium oxide, phosphorus pentoxide also have a strong positive correlation with potassium oxide, calcium oxide, magnesium oxide, alumina, respectively.

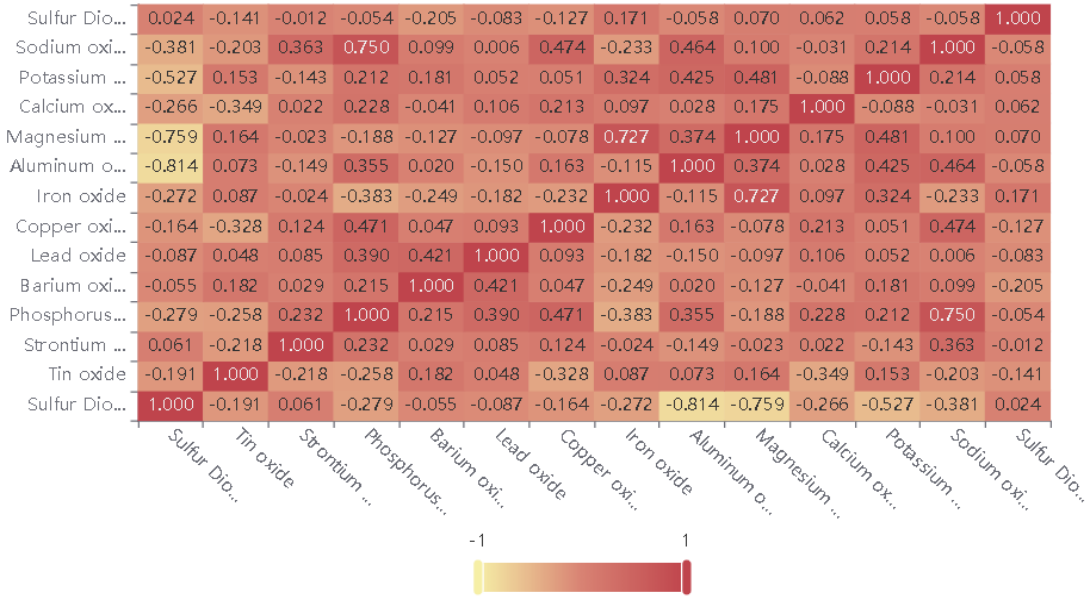

Figure 3. Heat map of correlation coefficients between chemical components before weathering of lead-barium glass

From the heat map inFigure 3 [10] of the correlation coefficient before weathering of lead-barium glass, it can be seen that there is a strong positive correlation between tin oxide and calcium oxide, barium oxide, and copper oxide, and the correlation between other components is not obvious. A heat map of the correlation coefficient of chemical composition after weathering of lead-barium glass is shown in Figure 4.

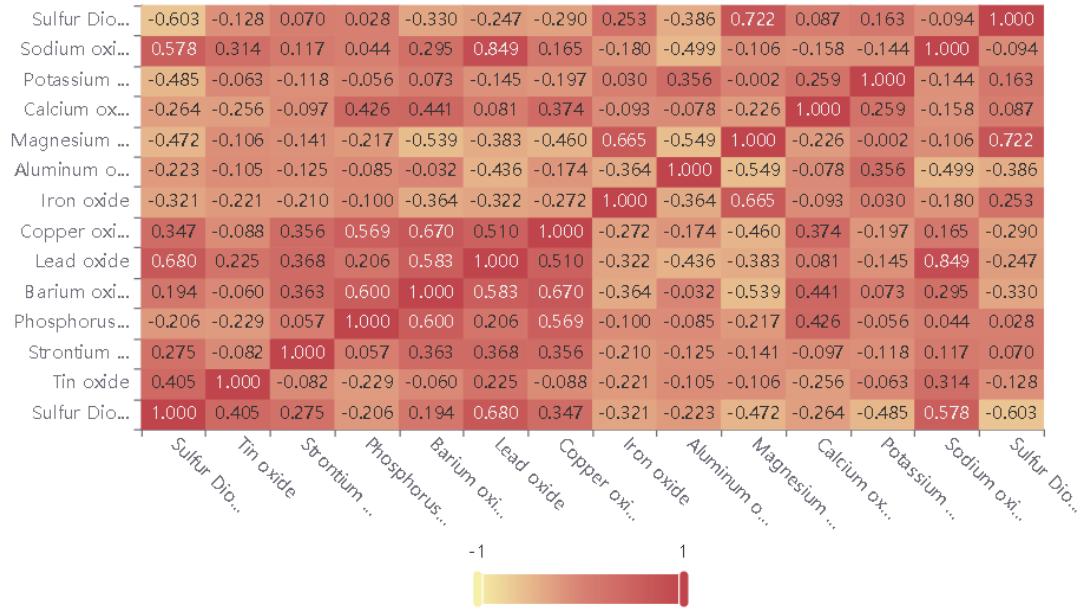

Figure 4. Heat map of correlation coefficients between chemical components after weathering of lead-barium glass

From the heat map of the correlation coefficient after weathering of lead barium glass, it can be seen that tin oxide has a strong positive correlation with alumina, sulfur dioxide has a strong positive correlation with barium oxide, iron oxide, and magnesium oxide, copper oxide, and barium oxide,

and there is a certain positive correlation between alumina and magnesium oxide, calcium oxide and magnesium oxide, calcium oxide, and iron oxide, but the correlation is weak.

4. Conclusion

In this paper, a model for the analysis and identification of glass product composition is proposed. Based on the XGBOOST model, the glass is classified according to weathering point data, decoration type and color. The performance of the model is good, which can well complete the task of glass composition analysis and identification, and has strong research significance for analyzing the correlation and difference between the chemical composition of different types of glass cultural relics.

References

[1] Qing Li, Chuang Wang, Jianwei Zhai. Research on Composition Analysis and Identification Based on Ancient Glass Products [J]. Academic Journal of Materials & Chemistry, 2022, 3 (2).

[2] Zhenghu Pang. Composition identification of ancient glass products based on cluster analysis [J]. Academic Journal of Computing & Information Science, 2022, 5 (13).

[3] Hui Xu. A study on the composition analysis and identification of ancient glass products based on SVM model [J]. Academic Journal of Computing & Information Science, 2022, 5 (13).

[4] Shang L, Li J, Jia X L, et al. Search for Pair-Produced vectorlike lepton singlet at the ILC by the XGBoost method [J]. Nuclear Physics B, 2023: 116071.

[5] Gorgan-Mohammadi F, Rajaee T, Zounemat-Kermani M. Decision tree models in predicting water quality parameters of dissolved oxygen and phosphorus in lake water [J]. Sustainable Water Resources Management, 2023, 9 (1): 1.

[6] Wang Y, Shi C, Wang J, et al. Efficacy of Yun-type pelvic floor optimal training therapy and PFMT on middle aged women with mild to moderate overactive bladder: a randomized controlled trial [J]. Annals of Translational Medicine, 2022, 10 (14).

[7] Rzęsikowska K, Kalinowska-Tłuścik J, Krawczuk A. Hierarchical analysis of the target-based scoring function modification for the example of selected class A GPCRs [J]. Physical Chemistry Chemical Physics, 2023.F.

[8] Sciuti L F, Mercante L A, Correa D S, et al. Random laser in dye-doped electrospun nanofibers: study of laser mode dynamics via temporal mapping of emission spectra using Pearson's correlation [J]. Journal of Luminescence, 2020, 224: 117281.

[9] Jia R M, Yang Z L, Zhou J, et al. Evaluating the confidence level of Traditional Chinese Medicine in nursing undergraduates at Chinese medical university [J]. Frontiers of Nursing, 9 (3): 263 - 268.

[10] Wark P A B, Hew M, Xu Y, et al. regional variation in prevalence of difficult-to-treat asthma and oral corticosteroid uses for patients in Australia: heat map analysis [J]. Journal of Asthma, 2022: 1 - 10.