# Classification and prediction of the chemical composition of glass based on grey and principal component Logistic regression

Hanyi Zhang [#, *], Yuwen Luo [#], Yang Chen [#]

Chang'an Dublin International College of Transportation, Chang'an University, Xi'an, China, 710000

* Corresponding author: zhanghanyi@chd.edu.cn

[#]These authors contributed equally.

**Abstract.** As a momentous witness of the ancient Silk Road trade, glass's variety identification and measurement or even prediction of its chemical constituent content are significantly important for people to have a systematic understanding of it. However, glass weathering tends to occur because of the ambient conditions of the place where it is buried, which may bring some difficulties to the classification and content prediction. Based on the data provided in problem C of the National Undergraduate Mathematical Modeling Contest in 2022, the GM (1,1) grey prediction model has first constructed so that the content ranges of various pivotal components are achieved. Then several glass samples whose category is unknown could be classified based on a principal component logistic regression model of the chemical composition's content.

**Keywords:** GM (1,1) Grey Prediction, Principal Component Logistic Regression Model, Classification of Ancient Glass, Glass Weathering.

## 1. Introduction

Due to the different glass firing techniques mastered by the trading parties, the composition of the glass unearthed in various places is also different. The use of modern techniques for classifying and predicting the content of glass chemical compositions is an important research direction in the field of glass science. Dimitrov and Komatsu used relationships between refractive index-based oxide ion polarizability, optical basicity, metallization criterion, Yamashita-Kurosawa's interaction parameter, ion charge, and binding energies in XPS spectra to classify oxide Glass [1]. In addition, some mathematical classification methods such as the logistic regression (LR) model have also been applied in numerous regions such as the classification of gene expression data [2]. As for the prediction of constituent content, Wang et al. Took advantage of the GM (1,1) gray model to predict critical odorants' content variation during the aging procedure [3]. This paper will introduce a grey prediction for the content prediction in ancient oxide glass as well as a modified logistic regression model to specify the category of provided glass samples.

## 2. Materials and Methodology

### 2.1. Data

It is worth to be mentioned that the research would focus on two types of glass, one is lead-barium glass and other glass with the high potassium content. The statistics are provided in three independent Excel sheets. The first sheet indicates the 58 glass samples' ornamentation, category (lead-barium type or high potassium content type), color, and whether it has experienced weathering. The second sheet provides the category and the percent of 14 oxide components measured from 67 detection points of those samples. The last sheet offers similar figures compared with the second sheet but the category is to be determined.

### 2.2. Introduction of methodology

#### 2.2.1. The introduction of the gray prediction method

The gray prediction model GM (1,1) is a core element of gray systems theory, which abstracts and quantifies system information from known partial data, then finds the optimal model and makes predictions about future data. The gray forecasting method uses the time series as the base for forecasting, and the higher order differential terms of the gray model can reflect the dynamic change process. In addition, its advantages contain requiring a small sample size of data, not requiring typical distribution laws, non-linearity, and high forecasting accuracy [4-6].

#### 2.2.2. The introduction of the LR model based on principal component analysis (PCA)

The figures for 14 oxides are considered original indicators and the 67 measurement points are samples, which could form a 67×14 matrix. To start with, to explain the variance we derive 14 principal components (PCs) and the contribution rate for each of them, and each PC is a linear combination of the original indicators. Then q (q>14) PCs are selected to explain the variance of original data to a large extent. The new q-dimension PC span could not only reserve the primitive information but achieve the dimensionality reduction of the research space [7]. Then the new statistics of these eight unknown samples used for logistic regression could be obtained by these PCs.

Besides the merits of PCA itself, the potential issues of the logistic regression model could be ameliorated with the help of PCA. One is the number of explicative variables should not be too large, which could be solved with the PCs with fewer numbers, and the other is the multicollinearity among the predictors, which could be avoided because those linear combinations are linearly independent [8-10].

## 3. The establishment and solution of the model

### 3.1. The establishment and solution of the GM (1,1) gray prediction model

#### 3.1.1. The establishment of GM (1,1) gray prediction model

Take high potassium content glass as an example, extract chemical components (such as $Al_2O_3$, $SiO_2$, etc.) that are strongly associated with weathering, and make data statistics on their contents. Weathering under natural conditions takes a long time, and the weathering process has a strong correlation with time. It is determined that the period (a certain year) is taken as the unit, and the data of these components before and after weathering are tested and analyzed, to obtain the statistical rule of these components. The number series forecasting principle of the GM (1,1) model is as follows.

**Step1**: Original sequence of numbers:

$$X^{(0)} = \{X^{(0)}(1), X^{(0)}(2), X^{(0)}(3), \cdots, X^{(0)}(n)\} \tag{1}$$

**Step2**: After a cumulative generation of the original sequence, the generated sequence can be obtained:

$$X^{(1)}(k) = \sum_{i=1}^{k} X^{(0)}(i), k = 1, 2, \cdots, n \tag{2}$$

Similarly, the randomness of the generated sequence is weakened, and its regularity is enhanced after multiple accumulations.

**Step3**: Constructing the data Matrix $B$ and data vectors $Y_n$. The least square method is used to calculate the parameter vector to be estimated $\gamma$:

$$B=\begin{bmatrix} -\frac{1}{2}\left[X^{(1)}(1)+X^{(1)}(2)\right] & 1 \\ -\frac{1}{2}\left[X^{(1)}(2)+X^{(1)}(3)\right] & 1 \\ \cdots & \cdots \\ -\frac{1}{2}\left[X^{(1)}(n-1)+X^{(1)}(n)\right] & 1 \end{bmatrix},\ Y_n=\begin{bmatrix} X^{(0)}(2) \\ X^{(0)}(3) \\ \cdots \\ X^{(0)}(n) \end{bmatrix} \tag{3}$$

$$\gamma=\begin{bmatrix}\alpha \\ b\end{bmatrix}=\left(B^TB\right)^{-1}B^TY_n \tag{4}$$

$\alpha$ and $b$ is the parameter vector to be estimated element of $\gamma$, among $\alpha$is the development gray number, and $b$ is the endogenous control gray number.

**Step4**: To establish the differential equation of the GM (1,1) model:

$$\alpha X^{(1)}+\frac{dX^{(1)}}{dt}=b \tag{5}$$

**Step5**: Solve the differential equation and build the prediction model:

$$\chi^{(1)}(k+1)=\left[X^{(0)}(1)-\frac{b}{a}\right]e^{-ak}+\frac{b}{a},k=0,1,2,\ldots,n \tag{6}$$

**Step6**: The accuracy and test of the GM (1,1) model, and the construction of residual sequence:

$$\varepsilon^{(0)}(k)=\left[X^{(0)}(k)-\chi^{(0)}(k)\right],k=0,1,2,\ldots,n \tag{7}$$

**Step7**: Calculate the mean relative error:

$$\bar{\varepsilon}=\frac{1}{n-1}\sum_{k=2}^{n}\left(\frac{X^{(0)}(1)-\chi^{(0)}(1)}{X^{(0)}(1)}\times 100\%\right) \tag{8}$$

Observation: The smaller $\bar{\varepsilon}$, the better. General requirements, $\bar{\varepsilon}<20\%$, preferably if $\bar{\varepsilon}<10\%$.

Due to the long weathering cycle of glass products, there are many factors affected, such as oxygen content in the air, moisture in the surrounding environment, and other factors that cannot be completely determined, so glass weathering is a gray system. The original data of weathered glass is not regular and relatively discrete. Through the accumulation of the original data to generate a series, and then the mathematical modeling of the generated series, the GM (1,1) model is established to obtain the data of chemical composition before weathering.

### 3.1.2. The results of GM (1,1) gray prediction model

In the process of glass weathering, the precipitation of alkali occurs, and more silicates are formed on its surface, so the content of $SiO_2$ increases over time. Firstly, a time series diagram is drawn according to the weathering composition of $SiO_2$. In Figure 1, the highest content of $SiO_2$ in period 1 indicates the highest degree of weathering, while the lowest content of $SiO_2$ in period 6 indicates the lowest degree of weathering.

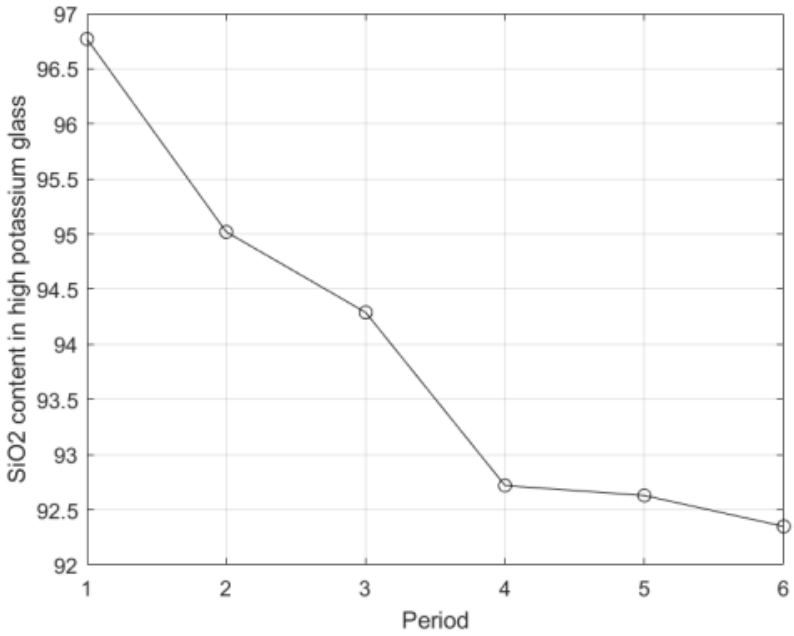


**Figure 1.** Time series plot of SiO2 content in high potassium glasses

The cumulative data were tested for quasi-exponential regularity, and the percentage of data with a smooth ratio less than 0.5, excluding the most preceding data, reached 100%, so this group passed the test. The six groups of $SiO_2$ content changes were divided into four training groups and two experimental groups; the traditional GM (1,1) model, the new information GM (1,1) model and the metabolism GM (1,1) model were used for the prediction of the original data respectively. The results are shown in Table.1. Finally, the traditional model was chosen for the prediction because it had the smallest sum of squared errors.

**Table 1.** Data results predicted by three GM (1,1) models

| Traditional GM (1,1) model | New information GM (1,1) model | Metabolism GM (1,1) model |
|---|---|---|
| 3.7862 | 3.7899 | 4.6576 |

The average relative residual of the model is 0.0031976, and the average grade ratio deviation is 0.0062621. The results show that the model has a good fitting effect on the original weathering data. The data from 40 periods under this development law are predicted and are illustrated in Figure 2.

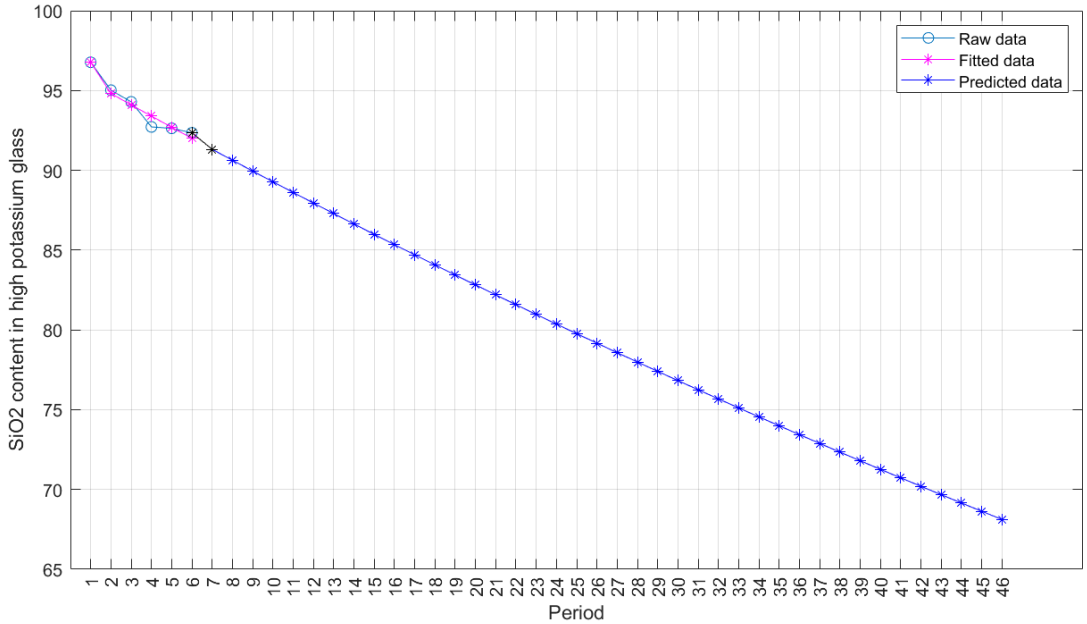


**Figure 2**. Predicted SiO2 content of high potassium glass

### 3.1.3. The verification of the prediction model

To further validate the reasonableness of the model, pre-weathering data were collected and collated from the sampling points of the high potassium content glass. As shown in Figure 3, the predicted contents of the 12 groups were all lower than the post-weathering $SiO_2$ content.

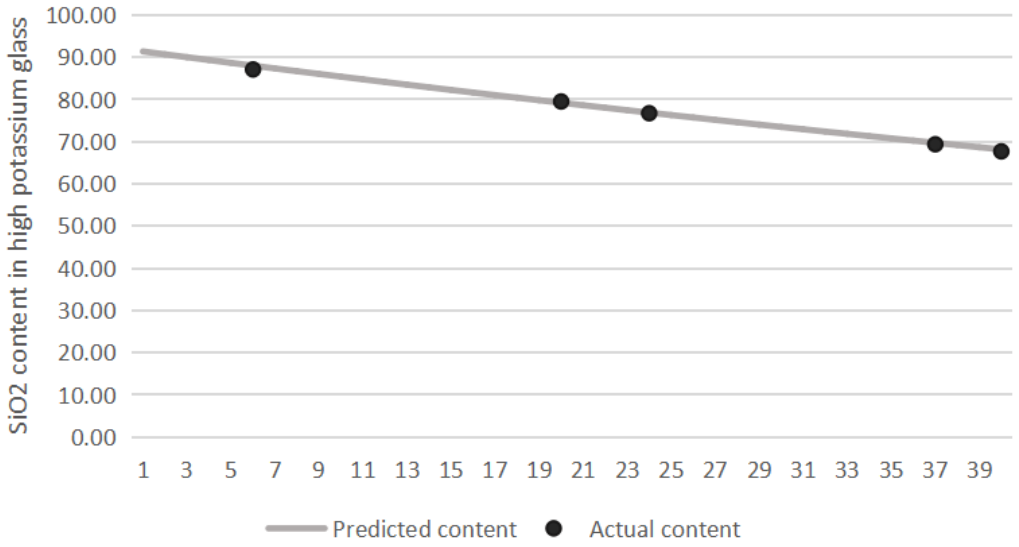


**Figure 3**. Comparison of predicted results with actual content

The GM (1,1) model can be used to predict the content with a high degree of accuracy, so we can obtain the threshold values of $SiO_2$, $Al_2O_3$, and CaO in high potassium content glass and lead-barium glass before weathering.

### 3.2. The establishment and solution of the PCA model

The PCA model is first established in this section so that some PCs could be extracted from the original figures. Then the LR model is constructed with the PCs playing the role of explicative predictors. The final classification result is derived by comparing the calculated probability with 0.5.

### 3.2.1. The establishment of the PCA model

With the provided 14 indicators and 67 samples, we could establish the PCA model, the specific steps are illustrated in the following section:

**Step1**: A 67×14 matrix $x$ could be formed with the data in the second sheet and $X$ is its standardization form. Then the covariance matrix $R$ could also be obtained:

$$x=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1\,14} \\ x_{21} & x_{22} & \cdots & x_{2\,14} \\ \vdots & \vdots & \ddots & \vdots \\ x_{67\,1} & x_{67\,2} & \cdots & x_{67\,14} \end{bmatrix}=(x_1,x_2,\cdots,x_{14}) \tag{9}$$

**Step2**: Find the desired linear combination of original variables, the conditions to be satisfied are labeled "s.t.":

$$\begin{cases} P_1=v_{11}X_1+v_{21}X_2+\cdots+v_{14\,1}X_{14}=v_1^{\,T}X \\ P_2=v_{12}X_1+v_{22}X_2+\cdots+v_{14\,2}X_{14}=v_2^{\,T}X \\ \cdots \\ P_{14}=v_{1\,14}X_1+v_{2\,14}X_2+\cdots+v_{14\,14}X_{14}=v_{14}^{\,T}X \end{cases} \tag{10}$$

$$P=(P_1,P_2,\cdots,P_{14})^T,\ V=(v_1,v_2,\cdots,v_{14}) \tag{11}$$

$$s.t.\begin{cases} u_{1i}^2+u_{2i}^2+\cdots+u_{14i}^2=1,\ i=1,2,\cdots,14 \\ Cov(P_i,P_j)=0,\ i\neq j \text{ and } i,j=1,2,\cdots,14 \\ D(P_1)\geq D(P_2)\geq\cdots\geq D(P_{14}) \end{cases} \tag{12}$$

**Step3**: Calculate the eigenvector $Main\ i$ (corresponding to the PCs) and eigenvalue $\lambda$ (corresponding to each PC's contribution rate) of the covariance matrix:

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1\,14} \\ r_{21} & r_{22} & \cdots & r_{2\,14} \\ \vdots & \vdots & \ddots & \vdots \\ r_{14\,1} & r_{14\,2} & \cdots & r_{14\,14} \end{bmatrix} \xrightarrow{U} URU^T=\begin{bmatrix} \lambda_1 & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \lambda_{14} \end{bmatrix} \tag{13}$$

$$\sum_{i=1}^{p}D(X_i)=r_1^2+r_2^2+\cdots+r_{14}^2=\lambda_1+\lambda_2+\cdots+\lambda_{14}=\sum_{i=1}^{p}D(P_i) \tag{14}$$

**Step4**: Calculate the accumulative contribution rate $\lambda_{acumu}$ of PCs, select them when $\lambda_{acumu}>80\%$:

$$\lambda_{acumu}=\sum_{i=1}^{k}\lambda_i \tag{15}$$

### 3.2.2. The results of PCA model

The PCs and the corresponding contribution rates are illustrated in Table 2.

**Table 2.** The results of PCA (eigenvectors and eigenvalues)

| | **a1** | **a2** | **a3** | **a4** | **a5** | **a6** | **a7** | **a8** | **a9** | **a10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$: **$SiO_2$** | -0.39 | -0.33 | 0.00 | 0.15 | -0.06 | -0.15 | 0.07 | 0.15 | 0.21 | -0.21 |
| $x_2$: **$Na_2O$** | -0.05 | -0.21 | -0.02 | -0.77 | 0.30 | 0.11 | 0.06 | 0.11 | -0.37 | -0.27 |
| $x_3$: **$K_2O$** | -0.33 | 0.16 | 0.34 | -0.08 | -0.02 | 0.43 | -0.20 | 0.07 | 0.16 | -0.07 |
| $x_4$: **Cao** | -0.17 | 0.43 | 0.28 | -0.01 | 0.21 | 0.45 | -0.02 | 0.00 | 0.00 | 0.15 |
| $x_5$: **MgO** | -0.20 | 0.40 | -0.23 | -0.23 | -0.14 | -0.27 | 0.00 | 0.26 | 0.09 | -0.25 |
| $x_6$: **$Al_2O_3$** | -0.30 | 0.24 | -0.03 | -0.36 | -0.08 | -0.33 | 0.22 | -0.09 | 0.17 | 0.65 |
| $x_7$: **$Fe_2O_3$** | -0.23 | 0.37 | 0.09 | 0.14 | 0.12 | -0.30 | 0.05 | -0.62 | -0.28 | -0.42 |
| $x_8$: **CuO** | 0.15 | 0.02 | 0.56 | -0.04 | 0.03 | -0.37 | -0.54 | 0.18 | 0.01 | 0.01 |
| $x_9$: **PbO** | 0.38 | 0.18 | -0.32 | 0.01 | 0.17 | 0.23 | -0.12 | -0.23 | -0.07 | 0.17 |
| $x_{10}$: **BaO** | 0.38 | -0.04 | 0.33 | -0.21 | -0.19 | -0.20 | 0.01 | -0.19 | -0.16 | 0.14 |
| $x_{11}$: **$P_2O_5$** | 0.20 | 0.43 | -0.08 | 0.20 | -0.03 | -0.09 | 0.12 | 0.60 | -0.35 | -0.06 |
| $x_{12}$: **SrO** | 0.34 | 0.22 | -0.15 | -0.26 | 0.01 | -0.02 | -0.14 | -0.08 | 0.68 | -0.32 |
| $x_{13}$: **SnO** | -0.10 | 0.03 | -0.18 | -0.15 | -0.80 | 0.21 | -0.34 | -0.12 | -0.23 | -0.04 |
| $x_{14}$: **$SO_2$** | 0.23 | 0.05 | 0.40 | -0.05 | -0.34 | 0.15 | 0.66 | -0.01 | 0.11 | -0.20 |
| **Contribution Rate** | 0.29 | 0.17 | 0.12 | 0.08 | 0.08 | 0.06 | 0.05 | 0.04 | 0.03 | 0.03 |
| **Accumulative Contribution Rate** | 0.29 | 0.46 | 0.59 | 0.67 | 0.75 | 0.80 | 0.85 | 0.89 | 0.92 | 0.95 |
| **Attention:** Because of the page limit and the low contribution rate of the other PCs, only 10 of 14 PCs are illustrated. | | | | | | | | | | |

As shown in Table.1., the accumulative contribution rate from a1 to a8 has reached 89%, so we only extract them among all 14 PCs. The linear combination expression is shown below:

$$\begin{aligned} Main1 &= -0.39x_1 - 0.05x_2 - 0.33x_3 + \ldots + 0.23x_{14} \\ Main2 &= -0.33x_1 - 0.21x_2 + 0.16x_3 + \ldots + 0.05x_{14} \\ &\vdots \\ Main8 &= 0.15x_1 + 0.11x_2 + 0.07x_3 + \ldots - 0.01x_{14} \end{aligned} \tag{16}$$

### 3.2.3. The establishment of LR model

It is proposed that LR model works well when used to predict a binary response variable which could be explained with numerous predictors [9]. Then this event is one type of Bernoulli distribution, so the expectation value could be obtained by:

$$E(y|x) = 1 \times P(y=1|x) + 0 \times P(y=0|x) = P(y=1|x) \tag{17}$$

It could be found that the expression is exactly the probability of the event "1". With the purpose of connecting the explicative predictors and the response variables, a copula function is constructed, and the following conditions are to be satisfied:

$$P(y=1|x) = F(x,\beta) = \hat{y} \tag{18}$$

Here $\hat{y}$ is used to represent the probability of the event "1".

The sigmoid function is selected as the copula function. Hence, a regression function whose range of value is between 0 and 1.

$$F(x,\beta) = S(x_i'\beta) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \tag{19}$$

It is obvious that this is not a linear model, so the maximum likelihood estimation (MLE) method is used to estimate the parameter β. The log-likelihood function of the sample is constructed and solves the problem:

$$\begin{cases} P(y=1|x) = S(x_i'\beta) \\ P(y=0|x) = 1 - S(x_i'\beta) \end{cases} \Rightarrow \ln L(\beta|y,x) = \sum_{i=1}^{n} y_i \ln\left[S(x_i'\beta)\right] + \sum_{i=1}^{n}(1-y_i)\left[1-S(x_i'\beta)\right] \tag{20}$$

As for the derived $\hat{y}$, 0.5 is taken as the critical value. For $\hat{y} > 0.5$, y is considered as 1 (corresponding to lead-barium glass), or y is considered as 0 (corresponding to glass with high potassium content).

### 3.2.4. The result of the LR model based on PCA

Here the 8 PCs obtained in 3.2.2 are used as explicative variables x1, x2, ..., x8 and the binary LR package could help to derive the LR function model:

$$\hat{y} = \frac{e^{20.439+22.437x_1+1.733x_2-26.625x_3-4.266x_4-3.353x_5-11.414x_6+5.194x_7-12.961x_8}}{1+e^{20.439+22.437x_1+1.733x_2-26.625x_3-4.266x_4-3.353x_5-11.414x_6+5.194x_7-12.961x_8}} \tag{21}$$

The new figures for the given samples under 8 selected PCs could be derived by the method of matrix multiplication and the result is shown in Table.4.

**Table 3.** The PC figures for unknown samples

| PC Sample | Main1 | Main2 | Main3 | Main4 | Main5 | Main6 | Main7 | Main8 |
|---|---|---|---|---|---|---|---|---|
| A1 | -1.37 | 0.11 | -0.04 | 0.28 | 0.04 | -0.85 | 0.76 | 0.44 |
| A2 | 1.15 | 0.90 | -1.00 | 1.48 | 1.02 | 1.26 | 0.32 | 1.14 |
| A3 | 0.69 | 2.22 | -0.13 | -0.04 | 1.07 | 1.26 | -0.72 | -1.80 |
| A4 | 0.68 | 1.69 | -0.17 | -0.05 | 0.27 | -0.77 | 0.04 | -0.87 |
| A5 | -1.43 | -0.15 | -1.68 | -3.48 | -1.63 | -0.59 | -0.32 | 0.15 |
| A6 | -1.38 | -1.95 | 0.39 | 1.10 | 0.09 | 0.44 | -0.38 | 0.65 |
| A7 | -1.38 | -1.84 | 0.18 | 0.85 | 0.04 | 0.09 | 0.14 | 0.40 |
| A8 | 3.03 | -0.97 | 2.46 | -0.14 | -0.90 | -0.84 | 0.16 | -0.11 |

The derived probability and the final results of classification are shown in Table.5.

**Table 4.** The calculation and classification result

| Glass Sample | probability | y | Classification result |
|---|---|---|---|
| A1 | 0.0732211 | 0 | High potassium content glass |
| A2 | 1-7E-17 | 1 | Lead-barium glass |
| A3 | 1-2.8E-20 | 1 | Lead-barium glass |
| A4 | 1-5.7E-28 | 1 | Lead-barium glass |
| A5 | 1-3.7E-25 | 1 | Lead-barium glass |
| A6 | 5.2E-20 | 0 | High potassium content glass |
| A7 | 8.6E-13 | 0 | High potassium content glass |
| A8 | 1-1.1E-16 | 1 | Lead-barium glass |

## 4. Conclusions

Owing to the smaller number of samples, the fuzzy influencing factors and incompleteness of information, GM (1,1) gray prediction is introduced to predict the composition content of glass before weathering, following with a comparison of the prediction result and the actual result. When it comes to the binary classification of unknown samples, LR model based on PCA is applied to derive the probability of specifying one sample's category and it is found that the obtained probability is extremely closed to either 0 or 1, which indicates that this model has a high possibility of success.

## References

[1] Dimitrov V, Komatsu T. Classification of Oxide Glasses: A polarizability approach [J]. Journal of Solid-State Chemistry, 2005, 178: 831 – 846.

[2] Liu Z, Chen D, Bensmail H. Gene expression data classification with kernel principal component analysis [J]. Journal of Biomedicine and Biotechnology, 2005, 2005: 155 – 159.

[3] Wang Z, Wei J, Wang Y, Zhu T, Huang M, Wu J, et al. A new method to predict the content changes of aroma compounds during the aging process of Niulanshan Baijiu using the GM (1, 1) gray model [J]. Flavour and Fragrance Journal, 2021, 37: 5 – 19.

[4] Ju-Long D. Control Problems of gray Systems [J]. Systems & Control Letters, 1982, 1: 288 – 294.

[5] Tien T-L. A research on the gray Prediction Model GM (1, N) [J]. Applied Mathematics and Computation, 2012, 218: 4903 – 4916.

[6] Chen C-I, Huang S-J. The necessary and sufficient condition for GM (1,1) gray prediction model. Applied Mathematics and Computation, 2013, 219: 6152 – 6162.

[7] Kwak N, Kim C, Kim H. Dimensionality reduction based on ICA for regression problems [J]. Neurocomputing, 2008, 71: 2596 – 2603.

[8] Musa AB. A comparison of ℓ1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression [J]. International Journal of Machine Learning and Cybernetics, 2014, 5: 861 - 873.

[9] Aguilera AM, Escabias M, Valderrama MJ. Using principal components for estimating logistic regression with high-dimensional multicollinear data [J]. Computational Statistics & Data Analysis, 2006, 50 (8): 1905 - 1924.

[10] Hosmer DW, Lemeshow S. Applied logistic regression. Wiley series in probability and statistics, 2nd edn [M]. Wiley, New York, 2013.