Analysis and Classification of Glass Products Based on Component Data Analysis

Bo Wu, Beijing International Studies University 2021221257@stu.bisu.edu.cn
[Official]:https://github.com/sxjs1st

Abstract

Glass is one of the earliest man-made materials invented by humanity. In ancient times, except for a few naturally occurring glasses, most glasses were man-made. Although they appear similar, their chemical compositions differ. The study and identification of the chemical compositions of glass products are of great significance for the preservation of ancient glass artifacts. This paper analyzes the chemical composition of glass from the perspective of component data analysis and classifies and predicts different types of glass.

First, the data in the attached tables were preprocessed. In Table 1, missing values were imputed using the mode and heat capacity filling. Because the chemical composition of glass products is reflected in the component data, the sum of the component proportions might not be 100% due to detection methods and other factors; in addition, many components may be undetectable or recorded as zero. It is assumed that this situation is due to limitations in instrument precision or rounding. Therefore, a multiplicative replacement method was used to impute the blanks and zero values in Tables 2 and 3, and the resulting glass chemical composition data is treated as component data.

For Issue 1, first, the relationship between surface weathering of glass artifacts and their type, decoration, and color was analyzed based on the Spearman correlation coefficient and chi-square test. The results indicated that the surface weathering of glass artifacts is related to the type of glass, but not to the decoration or color. Next, based on the Aitchison geometric structure of the component data, the mean chemical compositions of weathered and unweathered glass were calculated in the simplex space for each type of glass. The results showed that in weathered high-potassium glass, the proportion of silicon dioxide increased considerably while potassium oxide decreased significantly, and the proportions of sodium oxide, aluminum oxide, copper oxide, and barium oxide all increased; in weathered lead-barium glass, the proportion of silicon dioxide decreased, while lead oxide and potassium oxide increased, and phosphorus pentoxide decreased. Finally, using the chemical composition of the glass as the response variable and decoration, type, color, and surface weathering as the explanatory variables, a Dirichlet regression model was constructed to predict the pre-weathering chemical composition content of weathered points. Specific results can be found in the supplementary materials.

For Issue 2, the classification patterns of high-potassium glass and lead-barium glass were first analyzed. Based on the results from Issue 1, surface weathering and chemical composition were chosen as classification features, with glass type as the category. Since the chemical composition is component data, a centered log-ratio (clr) transformation was applied to the chemical composition. A decision tree was used to build a classification model, classifying the two types of glass based on the lead oxide content: if the lead oxide value is greater than 1.5, it is classified as lead-barium glass; otherwise, it is high-potassium glass. A partial least squares discriminant analysis (PLS-DA) was also established, and through the projection importance of different features, it was found that the features affecting classification were lead oxide, potassium oxide, barium oxide, and strontium oxide. Both methods yielded consistent results. Then, k-means clustering analysis was applied to separately cluster the two types of glass, determining that the optimal number of clusters for each was 3. Based on PLS-DA, appropriate chemical components were selected for each type of glass: high-potassium glass was divided into three categories based on potassium oxide, silicon dioxide, calcium oxide, tin oxide, and barium oxide, while lead-barium glass was divided into three categories based on phosphorus pentoxide, copper oxide, sodium oxide, and iron oxide.

For Issue 3, models constructed in Issue 2 were used to classify glasses of unknown categories. Method one is based on the decision tree model, which uses the lead oxide content to classify unknown glasses. Method two applies a clr transformation to the data in Table 3, selects lead oxide, potassium oxide, barium oxide, and strontium oxide as the explanatory variables and glass type as the dependent variable, and builds a partial least squares (PLS) regression model. Through cross-validation, the number of principal components was chosen as 3, and the glass type was determined based on the predicted values. Both methods yielded consistent results: A1, A6, and A7 are high-potassium glass, while A2, A3, A4, AS, and A8 are lead-barium glass. Furthermore, combining the subclassification features of the two types of glass from Issue 2 and based on the PLS results, it was determined that A1, A6, and A7 belong to the same subclass of high-potassium glass; A2 and A4 belong to the same subclass of lead-barium glass; A3 and A8 belong to another subclass of lead-barium glass; and AS belongs to yet another subclass of mis-barium glass.

For Issue 4, first, the correlation coefficients of the chemical components for each type of glass are calculated separately, and a correlation heatmap is generated for observation and analysis. Then, a paired Wilcoxon test is conducted on the correlation coefficients between the two types of glass, with the results indicating that there is no significant difference in the correlation relationships of the chemical components between the two types.

Keywords:**Component Data, Decision Tree, Partial Least Squares Discriminant Analysis, Cluster Analysis, Correlation Analysis**

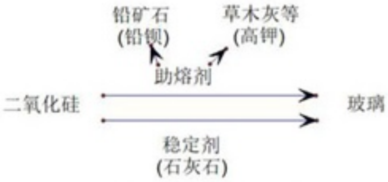


Figure 1: **Glass Making Process**

1 Restatement of the Problem

Glass has a long development history, and in China it has evolved from an imported product to being domestically produced(Zhao, 2016). Although the glass products may look similar, their chemical compositions are different. The primary chemical component of glass is silicon dioxide, but due to the different flux additives used, the proportions of components vary. Thus, glass can be classified into lead-barium glass (where the flux is mainly derived from lead ore) and high-potassium glass (where the flux is mainly derived from substances such as plant ash that are high in potassium **Figure 1**).(201, 2011)

Moreover, ancient glass is highly susceptible to weathering due to the burial environment, which causes an exchange of elements between the interior and exterior, leading to changes in component proportions and affecting classification. A set of classified lead-barium and high-potassium glass artifacts is available, with Table 1 providing the basic information of the artifacts, Table 2 giving the chemical composition proportions of the classified glass artifacts, and Table 3 presenting the chemical composition proportions of the unclassified glass artifacts.

Based on the above information, this paper establishes mathematical models to address the following issues:

Issue 1: (1) Analyze whether the surface weathering of glass artifacts is related to glass type, decoration, and color; (2) Based on glass type, analyze the statistical patterns of chemical compositions between weathered and unweathered surfaces for both lead-barium and high-potassium glass; (3) Using the detection data of weathered points in Table 2, predict the chemical composition content of these points before weathering.

Issue 2: (1) Analyze how to classify the glass based on the basic information in Table 1 and the chemical components detected in Table 2; (2) For the two classified types of glass, further subdivide them based on appropriate chemical components; (3) Analyze the rationality and sensitivity of the classification.

Issue 3: (1) Analyze the chemical compositions of unknown glass types in Table 3 and identify their types according to the classification rules derived in Issue 2; (2) Analyze the sensitivity of the classification results.

Issue 4: (1) Analyze the correlation between the chemical components of the two types of glass artifact samples; (2) Compare the differences in the correlation relationships between the chemical components of the two types.

2 Problem Analysis

Before addressing the issues, the data in the attached tables were preprocessed. For Table 1, missing values were imputed using the mode and heat capacity filling methods.(Pawlowsky-Glahn et al., 2015) For Tables 2 and 3, blanks and zero values were imputed using the multiplicative replacement method, and then the chemical composition data was converted into component data (Pawlowsky-Glahn & Buccianti, 2011).

2.1 Issue 1:

First, the relationship between the surface weathering of glass artifacts and their type, decoration, and color was analyzed using the Spearman correlation coefficient and chi-square test. Next, based on the Aitchison geometric structure of the component data, the mean chemical compositions of weathered and unweathered surfaces for each type of glass were calculated in the simplex space, and circular plots were used for comparative analysis. Finally, a Dirichlet regression model was constructed using the artifact's chemical composition as the response variable and decoration, type, color, and surface weathering as the explanatory variables to predict the pre-weathering chemical composition at weathered points.

2.2 Issue 2:

To study the classification patterns of high-potassium glass and lead-barium glass, and based on the results from Issue 1, the basic information and chemical composition of the glass were chosen as classification features with glass type as the category. Since the chemical composition is component data, a centered log-ratio (clr) transformation was applied. Classification was performed using both a decision tree and partial least squares discriminant analysis (PLS-DA) to select features important for classification. For the subclassification of each glass type, k-means clustering analysis was applied separately to the two types, and then appropriate chemical components for each type were selected based on the PLS-DA results.

2.3 Issue 3:

Based on the models constructed in Issue 2, the unknown glass artifacts were classified. Method one utilized a decision tree model, using the selected features to classify the unknown glass. Method two involved applying a clr transformation to the data in Table 3, then using the features selected via partial least squares discriminant analysis as explanatory variables and glass type as the dependent variable to build a partial least squares (PLS) regression model. The number of principal components was determined through cross-validation, and the glass type was predicted based on the resulting values.

2.4 Issue 4:

For each glass type, the Pearson correlation coefficient was used to analyze the relationships among the chemical components, and the correlations were visualized as a heatmap. Then, a paired Wilcoxon test was performed to determine whether there was a significant difference in the correlation relationships of the chemical components between the two types of glass.

Table 1: **Explanation of Symbols**

Symbol	Meaning
X	Dataset
c	Constant sum constraint of compositional data
e	Exploration range vector
e_j	Exploration range corresponding to the j-th part of the compositional dataset
δ_{ij}	A number smaller than e_j
D	The dataset is divided into D parts

3 Model Assumptions and Notation Explanation **Table 1**

1. It is assumed that the chemical composition refers to the composition at the sampling points.

2. It is assumed that the undetected chemical components are due to limitations in instrument precision; hence, missing values are temporarily recorded as 0 and later imputed.

3. It is assumed that the 0 values for detected chemical components are a result of rounding, and these values will be imputed accordingly.

4 Data Preprocessing

Table 2: **Color Distribution (1)**

Color	**Blue-green**	**Light Blue**	**Light Green**	**Dark Green**	**Purple**
Corresponding Artifact Numbers	56, 57	11, 25, 43, 51, 52, 54	41	34, 36, 38, 39	08, 26

Table 3: **Color Distribution (2)**

Color	**Black**	**Ochre Green**	**Light Blue**
Corresponding Artifact Numbers	49, 50	23	02, 28, 29, 42, 44, 53

4.1 Preprocessing of Attachment Form 1 Data

Attachment Form 1 for this problem provides basic information on glass artifact number, ornamentation, type, color, and surface weathering. Due to the large volume of data and the presence of certain missing values, the given data is preprocessed to compensate for these omissions and to prevent any adverse impacts on subsequent modeling.

In Attachment Form 1, the missing values are all in the glass artifacts' color field. Specifically, two artifacts with missing values have their other attributes recorded as "Ornamentation A, Surface Weathering, Lead-Barium," and the other two artifacts with missing values have their other attributes recorded as "Ornamentation C, Surface Weathering, Lead-Barium." These are divided into two categories for separate imputation.

Among the artifacts provided, there are 15 artifacts meeting the "Ornamentation C, Surface Weathering, Lead-Barium" criteria with known colors, whose color distribution is shown in **Table 2**. Since the data is qualitative, a heat map imputation method is used to fill in the missing values. A comparison of the chemical compositions is made between these 15 artifacts and artifacts No. 40 and No. 58 that require imputation; the colors from artifacts with similar compositions are used for the imputation. (For an artifact sampled from two parts, the average is taken as representative; for an artifact sampled from one part and a severely differentiated point, a weighted average of 1:2 is computed; for an artifact sampled from one part and an undifferentiated point, the chemical composition of the part is used as representative.) After the comparison, it is found that artifact No. 40 has a composition similar to artifact No. 39, and artifact No. 58 is similar to artifact No. 11. Therefore, the colors of artifacts No. 40 and No. 58 are imputed as dark green and light blue, respectively.

For artifacts meeting the "Ornamentation A, Surface Weathering, Lead-Barium" criteria with known colors, there are 9 artifacts in total, whose color distribution is shown in **Table 3**. According to Attachment Form 2, among these 9 artifacts, 6 have chemical composition analysis where the sampling point was an unweathered point, resulting in relatively low reliability of the obtained composition data. Therefore, the mode is used for imputation; specifically, the colors of artifacts No. 19 and No. 48 are imputed as light blue.

4.2 Preprocessing of Data in Attachment Forms 2 and 3

Table 4: **Component Data Imputation Missing Values (Partial)**

Sampling Point	SiO_2	Na_2O	K_2O	…	SnO_2	SO_2
01	69.3300	0.5275	9.9900	…	0.1517	0.3902
03 Unit 1	87.0500	0.5287	1.5900	…	0.1520	0.0727
03 Unit 2	61.7100	0.5239	12.3700	…	0.1506	0.1720
⋮	⋮	⋮	⋮	…	⋮	⋮
57	25.4200	0.5264	0.0724	…	0.1513	0.3902
58	30.9900	0.5300	0.0200	…	0.1506	0.0902

Table 5: **Component Data Imputation Missing Values (Partial)**

Artifact No.	Surface Weathering	SiO_2	Na_2O	K_2O	…	SnO_2	SO_2
A1	No Weathering	78.4500	0.5277	0.0726	…	0.1512	0.0728
A2	Weathered	67.3500	0.5239	0.0130	…	0.1523	0.0785
A3	No Weathering	31.9500	0.5239	3.8600	…	0.1500	0.0706
A4	No Weathering	31.7900	0.5287	1.3600	…	0.1507	0.0750
A5	Weathered	69.2400	0.5300	0.0724	…	0.1515	0.0807
A6	Weathered	28.6900	0.5242	2.0200	…	0.1517	0.0893
A7	Weathered	67.3500	0.5239	0.0130	…	0.1513	0.0750
A8	No Weathering	32.0500	0.5300	1.3600	…	0.1509	0.2600

4.2.1 Missing Value Imputation

Since data with cumulative component proportions between 85% and 105% are considered valid, the cumulative proportions of the components are calculated, revealing that the invalid data belong to "Artifact No. 15" and "Artifact No. 17." These artifacts are subsequently removed from the forms. The blanks in Forms 2 and 3 indicate that the corresponding component was not detected. It is assumed that due to instrument precision limitations, the component was not detected; therefore, the missing values cannot simply be replaced with "0" but are considered approximate zero values. Additionally, the 0 values present in Forms 2 and 3 are assumed to result from rounding and are also treated as approximate zero values. A multiplicative replacement method is then employed to impute these approximate zero values; the specific process is as follows: **Consider the compositional data set Formula 1**

$$X = [x_{ij}]_{n \cdot D} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nD} \end{pmatrix} \tag{1}$$

Assume that the compositional data contains near-zero values, which arise because the true values are below a known detection limit and thus unobservable. The detection limits are assumed to be the same for corresponding parts across different components. Let the detection limit vector be **Formula 2**:

$$\mathbf{e} = (e_1, e_2, \ldots, e_D)^T, \tag{2}$$

where e_j is the detection limit corresponding to the j-th component of the data set X. A multiplicative simple substitution method is proposed. The substituted value x_{ij}^r is given by **Formula 3**:

$$x_{ij}^r = \begin{cases} \delta_{ij} \cdot \left(\dfrac{\sum\limits_{k,l} x_{kl} \cdot \delta_{ik}}{1 - \sum\limits_{j=1}^{D} x_{ij}/c} \right), & x_{ij} = 0 \\ x_{ij}, & x_{ij} > 0 \end{cases} \tag{3}$$

Here, δ_{ij} is a small number less than e_j, and c is a constant sum constraint of the compositional data, i.e. **Formula 4**,

$$\sum_{j=1}^{D} x_{ij} = c \tag{4}$$

Through experimental results, it is found that when the proportion of near-zero values in the compositional dataset is not high, setting δ_{ij} equal to 65% of the detection limit can minimize the distortion of the covariance matrix, i.e. **Formula 5**,

$$\delta_{ij} = 0.65 \cdot e_j \tag{5}$$

For the blank cells in Tables 2 and 3, the minimum value in the respective columns is taken as the threshold, and 0.65 times that value is used as the imputed value. (Some of the imputed values are listed here **Table 4 Table 5**; for detailed data, see supporting material.)

4.2.2 Component Data Processing

Since the component data must sum to 100, the relative information within the data can be further utilized. "Relative information" refers to the fact that the only information present in the component data is reflected in the ratios between components; the absolute values of each component are irrelevant. If every component is multiplied by the same constant, the ratios remain unchanged. Therefore, the component data can be regarded as an equivalence class—each set of component data in this class contains the same information and can be represented as the same proportional vector using an appropriate scale factor. In this way, the closure operation can be realized, the **Formula 6** is:

$$C(\mathbf{x}) = C(x_1, x_2, \ldots, x_D)^T = \left(\frac{k \cdot x_1}{\sum\limits_{i=1}^{D} x_i}, \frac{k \cdot x_2}{\sum\limits_{i=1}^{D} x_i}, \ldots, \frac{k \cdot x_D}{\sum\limits_{i=1}^{D} x_i} \right)^T \tag{6}$$

The closure operation is defined as multiplying the original vector by an appropriate scaling factor such that the sum of the resulting components equals a constant k (here $k = 100$). For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$, if $C(\mathbf{x}) = C(\mathbf{y})$, then $\mathbf{x}$ and $\mathbf{y}$ are compositionally equivalent.

5 Model Establishment and Solution

5.1 Problem 1

Table 6: **Dummy Variable Processing**

Pattern	Code	Type	Code	Color	Code
A	1	Aluminum	1	Light Blue	1
B	2	High Potassium	2	Dark Blue	2
C	3			Cyan Blue	3
				Light Green	4
				Dark Green	5
				Green	6
				Purple	7
				Black	8
Surface Weathering					
None	0				
Weathered	1				

Table 7: **Correlation between Dummy Variables and Surface Weathering**

	Dummy Variable	Surface Weathering
Dummy Variable	1.000	0.080
Surface Weathering	0.080	1.000

Table 8: **Correlation Between Species Type and Surface Roughness**

Species Type	Species Type	Surface Roughness
Species Type	1.000	-0.301
Surface Roughness	-0.301	1.000

5.1.1 Correlation Analysis Between Glass Surface Weathering and Ornamentation, Glass Type, and Color

Based on the above analysis, we first conducted a Spearman correlation analysis. Prior to the analysis, the qualitative data were transformed into dummy variables (see **Table 6**). Using SPSS software, Spearman correlation tests were performed for ornamentation, glass type, and color versus the surface weathering of the artifacts, with the results presented in **Table 7 Table 8** and **Table 9**. The results indicate that neither ornamentation nor color is correlated with whether the glass surface is weathered, whereas glass type does exhibit a correlation with surface weathering.

Secondly, to further ensure the credibility of the results, a chi-square test was also conducted on the data. Prior to the analysis, frequency statistics were compiled (see **Figure 2 Figure 3** and **Figure 4**):

From **Figure 2**, it is evident that under both unweathered and weathered conditions, the variations among ornamentation types A, B, and C are relatively small, leading to the preliminary inference that surface weathering is not related to ornamentation type.

From **Figure 3**,, when comparing weathered to unweathered conditions, the proportion of lead-barium glass increases while that of high-potassium glass decreases, with significant changes in both directions. This suggests that surface weathering is related to glass type.

From **Figure 4**,, the variations in different colors are minor between weathered and unweathered conditions, leading to the preliminary inference that surface weathering is not related to color.

Table 9: **Correlation Between Color and Surface Roughness**

Color	**Color**	**Surface Roughness**
Color	1.000	0.088
Surface Roughness	0.088	1.000

Table 10: **Chi-square Test Results**

	Coefficient
Decoration	0.085
Glass type	0.024
Color	0.325

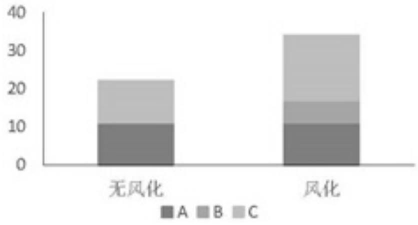


Figure 2: **Decoration**

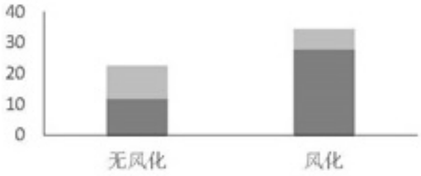


Figure 3: **Glass Type**

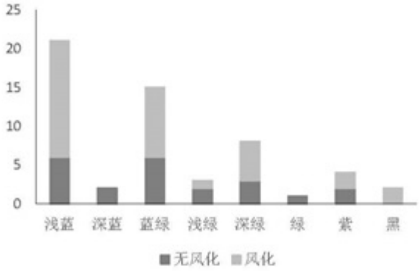


Figure 4: **Color**

To further validate these inferences, a chi-square test was performed using SPSS software, and the results are shown in **Table 10**. Since only the coefficient for glass type is less than 0.05, it can be concluded that glass type is correlated with the surface weathering of the artifacts, while the coefficients for ornamentation and color are both greater than 0.05—indicating that neither ornamentation nor color is correlated with surface weathering.

Table 11: **Average Values of Weathered and Unweathered High-Lead Glass and Lead-Barium Glass**

	SiO_2	Na_2O	K_2O	...	SnO_2	SO_2
High-lead (unweathered)	74.9529	8.1890	7.2236	...	0.2110	0.1234
High-lead (weathered)	93.6615	2.6528	0.3579	...	0.2191	0.0123
Lead-barium (unweathered)	59.5046	1.3090	0.1981	...	0.1602	0.0925
Lead-barium (weathered)	27.0317	0.7581	0.1535	...	0.0973	0.0529

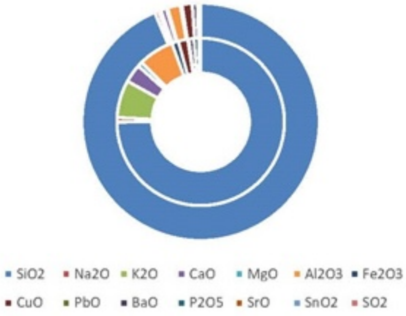


Figure 5: **Comparison of chemical composition of high potassium glass before and after weathering. The outer circle is weathered, and the inner circle is weathered.**

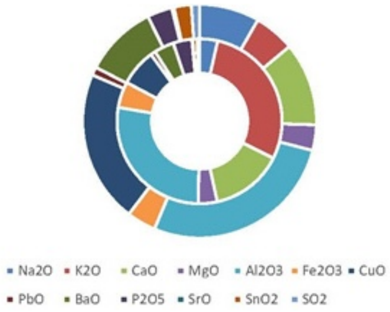


Figure 6: **Comparison of chemical composition of high potassium glass before and after weathering (excluding SiO2)**

5.1.2 Analysis of the Statistical Patterns in the Chemical Composition Content of Weathered and Unweathered Surfaces of Different Types of Glass

This analysis uses the component data from Attachment Form 2 after processing. The mean values of the chemical composition data for the unweathered and weathered sampling points were computed separately for lead-barium glass and high-potassium glass. The specific calculation method is as follows:

Because compositional data has a geometric structure, perturbation operations can be performed to resemble addition operations in real space. For any compositional data $\boldsymbol{x}$ and $\boldsymbol{y}$, the perturbation operation is defined as **Formula 7** and **Formula 8** :

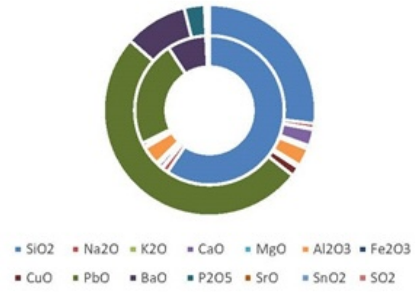


Figure 7: **Comparison of elements before and after lead-barium glass differentiation. The outer circle is weathered, and the inner circle is not weathered.**

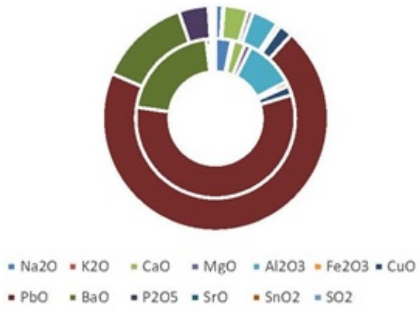


Figure 8: **Comparison of elements in lead-barium glass before and after weathering (excluding SiO2), the outer circle is weathered, the inner circle is non-weathered**

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_D)^T, \quad \boldsymbol{y} = (y_1, y_2, \ldots, y_D)^T \in S^D \tag{7}$$

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}\left(x_1 y_1, x_2 y_2, \ldots, x_D y_D\right)^T \in S^D \tag{8}$$

Thus, the mean values are further computed. Using R language to calculate the means of the component data, the results are shown in **Table 11**.

To visualize the changes in chemical composition more intuitively, the following figures were produced:

For high-potassium glass, because silicon dioxide occupies a large proportion, **Figure 5**, does not effectively reflect the changes in the other chemical components. By removing silicon dioxide and producing **Figure 6**, separately, the changes in the proportions of the other chemical components are better illustrated. Analyzing **Figure 5** and **Figure 6** together, after weathering, the proportion of silicon dioxide increases significantly, potassium oxide decreases noticeably, and the proportions of sodium oxide, aluminum oxide, copper oxide, and barium oxide all increase.

For lead-barium glass, **Figure 7** shows that, when comparing unweathered and weathered conditions, the proportion of silicon dioxide decreases while that of lead oxide increases. Similarly, by removing silicon dioxide to mitigate its dominant influence on the analysis, **Figure 8** reveals that potassium oxide and phosphorus pentoxide exhibit significant changes, with their trends being an increase and a decrease, respectively.

5.1.3 Prediction of Chemical Composition Prior to Weathering

First, the data in Attachment Form 1 were converted into dummy variables (see **Table 4** above). The variables from Attachment Form 1 were then used as independent variables, while the chemical variables from Attachment Form 2 were used as dependent variables to perform Dirichlet regression. After obtaining the regression model, the "weathering" variable in the independent variables (originally coded as 1) was replaced with "unweathered" (coded as 0), and partial prediction results are

Table 12: **Chemical Composition Before Prediction (in percentage)**

Sample Number	SiO_2	Na_2O	K_2O	...	SnO_2	SO_2
02	32.812	2.525	1.359	...	1.016	0.883
07	72.503	1.816	2.830	...	0.944	0.813
08	46.822	2.567	1.677	...	1.234	0.964
⋮	⋮	⋮	⋮	...	⋮	⋮
57	43.538	2.406	1.253	...	1.099	0.972
58	41.464	2.301	1.070	...	1.025	0.963

shown in **Table 12** (only a portion of the prediction results is presented here due to the large amount of data; all prediction results can be found in the supporting materials).**Code A**

5.2 Problem 2

Table 13: **VIP Values of Different Variables Analyzed by Partial Least Squares Discriminant Analysis**

Variable	VIP Value	Variable	VIP Value
Weathering	0.2098	CuO	0.5694
SiO_2	0.8907	PbO	2.4614
Na_2O	0.1203	BaO	1.3472
K_2O	1.8000	P_2O_5	0.1226
CaO	0.5095	SrO	1.1113
MgO	0.3871	SnO_2	0.3548
Al_2O_3	0.6157	SO_2	0.1247
Fe_2O_3	0.7284		

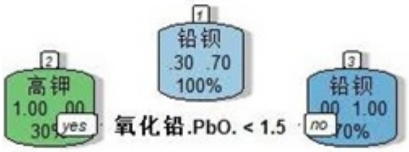


Figure 9: **Decision Tree Results**

5.2.1 Analysis of Glass Classification Rules

Before modeling, compositional data must be transformed using the clr (centered log-ratio) transformation. For any compositional data $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)^T \in S^D$, the clr transformation maps $\boldsymbol{x} \in S^D$ to coefficients on $\boldsymbol{\xi}^D$, and the clr coefficients are defined as **Formula 9**:

$$clr(\boldsymbol{x}) = \left(\log\left(\frac{x_1}{g_m(\boldsymbol{x})}\right), \log\left(\frac{x_2}{g_m(\boldsymbol{x})}\right), \ldots, \log\left(\frac{x_D}{g_m(\boldsymbol{x})}\right) \right)^T \tag{9}$$

Let the clr-transformed data be $clr(\boldsymbol{x}) = \boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_D)^T$, then the inverse clr transformation is given by **Formula 10**:

$$\boldsymbol{x} = clr^{-1}(\boldsymbol{\xi}) = C\left(\exp(\xi_1), \exp(\xi_2), \ldots, \exp(\xi_D)\right)^T \tag{10}$$

A classification model was established using the transformed component data. To present the classification results intuitively and clearly, a decision tree model was employed for feature selection and classification, with the aim of further exploring the primary classification features to deduce the classification rules for the two types of glass. In this model, the weathering data from Attachment Form 1 and the chemical component data from Attachment Form 2 were used as classification

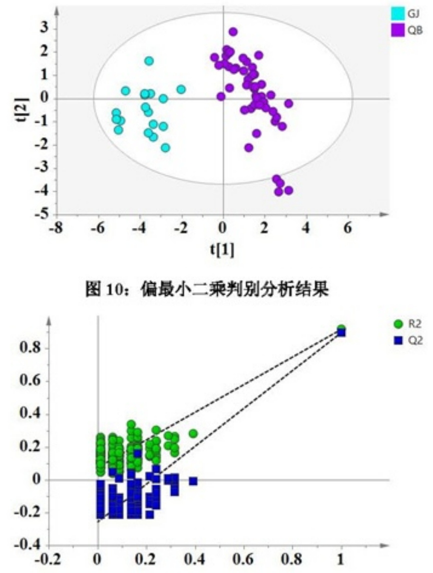


Figure 10: **Partial minimum quadratic discriminant analysis model validation**

features, while the type of glass served as the target category for constructing the decision tree model. The 67 samples were divided into a training set and a test set, with 47 samples allocated to the training set and the remaining samples to the test set. Finally, the implementation was carried out using R, and the results are shown in **Figure 9**,.

Subsequently, predictions were made using the test set data, achieving an accuracy rate of 100%.

From this figure, it can be observed that lead oxide (PbO) is ultimately used as the key feature for classifying the two types of glass: if the lead oxide value is greater than 1.5, the glass is classified as lead–barium glass; otherwise, it is classified as high-potassium glass.

Next, partial least squares discriminant analysis (PLS-DA) was used to classify high-potassium glass and lead–barium glass. Since the clr-transformed data sum to zero, glass type was used as the classification variable, and weathering, silicon dioxide (SiO_2), sodium oxide (Na_2O), potassium oxide (K_2O), calcium oxide (CaO), magnesium oxide (MgO), aluminum oxide (Al_2O_3), iron oxide (Fe_2O_3), copper oxide (CuO), lead oxide (PbO), barium oxide (BaO), diphosphorus pentoxide (P_2O_5), strontium oxide (SrO), tin oxide (SnO_2), and sulfur dioxide (SO_2) were used as feature variables to build the PLS-DA model. **Figure 10**, shows a clear separation trend between high-potassium glass and lead–barium glass.

As shown in **Figure 10**,, Q² crosses the negative half-axis of the Y-axis, indicating that the model has been successfully constructed.

Projection importance (VIP) for the different variables was calculated, as shown in **Table 13**. By screening features with VIP values greater than 1, the variables affecting the classification were determined to be PbO, K_2O, BaO, and SrO. (The above table results were obtained using the SIMCA software.)**Code B**

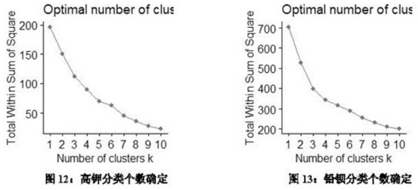


Figure 11: **The number of high potassium, lead and barium classifications is determined**

Table 14: **Subclassification of Glass Types**

Glass Type	Subclass	Artifact ID(s)
Borosilicate Glass	1	06 section 1, 18
	2	21, 07, 09, 10, 12, 22, 27
	3	01, 03 section 1, 03 section 2, 04, 05, 06 section 2, 13, 14, 16, 20, 37, 50 unweathered, 08 heavily weathered, 11, 19, 26 heavily weathered, 39, 40, 43 section 2, 50.1, 51 section 2, 54, 54 heavily weathered, 56, 58
Lead Glass	1	20, 37, 50 unweathered, 08 heavily weathered, 11, 19, 26 heavily weathered, 39, 40, 43 section 2, 50.1, 51 section 2, 54, 54 heavily weathered, 56, 58
	2	28 unweathered, 29 unweathered, 30 section 1, 30 section 2, 31, 32, 35, 49 unweathered, 02, 41, 48, 49, 51 section 2
	3	23 unweathered, 24, 25 unweathered, 33, 34, 38, 42 unweathered, 43 section 2, 44 unweathered, 45, 46, 47, 53 unweathered, 55, 34, 36, 38, 43 section 1, 57

Table 15: **VIP Values of Different Features in High-Temperature Silicate System**

Variable	VIP Value	Variable	VIP Value
SiO_2	1.5267	CuO	0.7509
Na_2O	0.3815	PbO	0.2068
K_2O	2.3016	BaO	1.2158
CaO	1.3078	P_2O_5	0.0160
MgO	0.0216	SrO	0.4147
Al_2O_3	0.4654	SnO_2	1.2763
Fe_2O_3	0.5235	SO_2	0.3721

Table 16: **VIP values of different variables in the analysis of ceramic frits for glass ceramics using partial least squares (PLS) regression.**

Variable	VIP	Variable	VIP
SiO_2	0.9550	CuO	1.3711
Na_2O	1.3046	PbO	0.2813
K_2O	0.5825	BaO	0.8605
CaO	0.6696	P_2O_5	2.0512
MgO	0.7716	SrO	0.4435
Al_2O_3	0.8339	SnO_2	0.4663
Fe_2O_3	1.2119	SO_2	0.7193

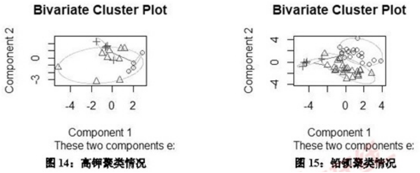


Figure 12: **High potassium, lead and barium clustering**

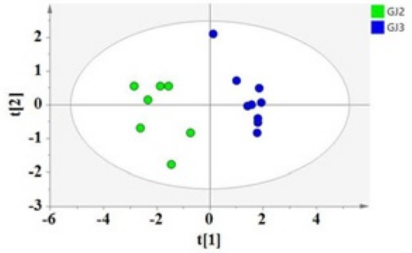


Figure 13: **High potassium glass subclass partial minimum quadratic discriminant analysis results**

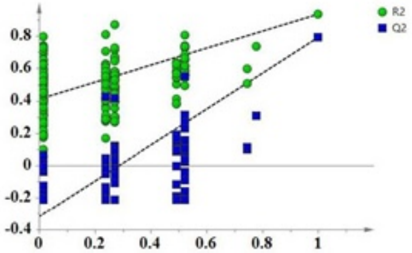


Figure 14: **Validation of the partial minimum quadratic discriminant analysis model for high potassium glass subclasses**

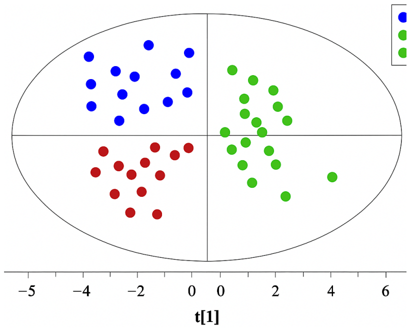


Figure 15: **PLS discriminant analysis results for lead-barium glass subclasses**

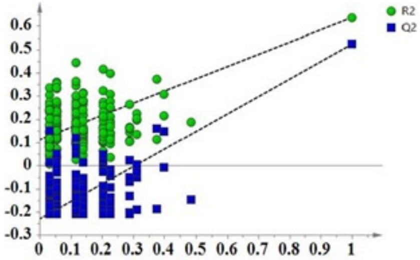


Figure 16: **Validation of the partial least quadratic discriminant analysis model for lead-barium glass subclasses**

5.2.2 Results of Subclassification Methods and Sensitivity Analysis

Further subclassification was conducted for the two types of glass. Since the subclassification results are not predetermined and the batch of samples does not have a definitive type, this problem represents an unsupervised learning scenario with no known dependent variable; hence, clustering analysis was chosen for classification. First, it was necessary to determine the number of clusters. The fviznbclust function in R was used for this purpose, and based on the inflection points, it was determined that the subclassification of both types of glass should ultimately be divided into 3 clusters **Figure 11**.

With the number of clusters determined, K-means clustering was applied, and the results are shown in the following figures **Figure 12**:

Analysis of the clustering results indicates that lead–barium glass can be clearly divided into three clusters; thus, its subclassification consists of three classes. For high-potassium glass, however, the clustering results were not as distinct, and the glass could only be roughly divided into three clusters **Table 14**.

The choice of the number of clusters can influence the clustering outcome, so a sensitivity analysis of the model is necessary. Since high-potassium glass did not show a very clear clustering effect, multiple cluster numbers were tried to determine an optimal clustering scenario.

Next, appropriate chemical components were selected for the subclassification of each type of glass.

For high-potassium glass, subclassification into 3 clusters was performed. Since one of the clusters contained only two samples, these samples were removed. Using the remaining clusters as the classification variable and the chemical components as features, a partial least squares discriminant analysis was built. As shown in **Figure 13**,, the subclasses of high-potassium glass exhibit a clear separation trend. **Figure 14**, shows that Q² crosses the negative half-axis of the Y-axis, indicating that the model was successfully constructed.

Projection importance (VIP) for different variables was calculated, as shown in **Table 15**. By screening features with VIP values greater than 1, the variables affecting the subclassification of high-potassium glass were identified as K_2O, SiO_2, CaO, SnO_2, and BaO. (The above table results were obtained using simca software.)

For lead–barium glass, subclassification into 3 clusters was performed. Taking these three clusters as the classification variable and the chemical components as features, a partial least squares discriminant analysis was constructed. **Figure 15**, shows a clear separation trend among the three clusters of lead–barium glass. Similarly, **Figure 16** shows that Q² crosses the negative half-axis of the Y-axis, indicating successful model construction.

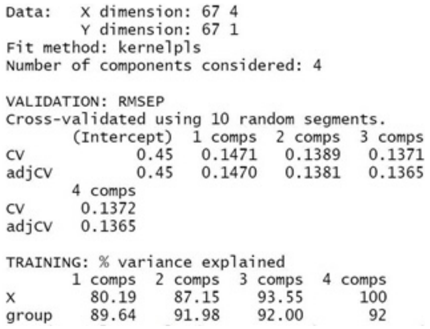
Projection importance (VIP) for the different variables was calculated, as shown in **Table 16**. By screening features with VIP values greater than 1, the variables affecting the subclassification of lead–barium glass were identified as P_2O_5, CuO, Na_2O, and Fe_2O_3. (The above table results were obtained using simca software.)**Code C**

Table 17: **Type Prediction for Unknown Contaminants**

Specimen ID	Predicted Value	Predicted Contaminant Type
A1	1.0408	High Silica
A2	0.0177	Low Silica
A3	0.0980	Low Silica
A4	0.1898	Low Silica
A5	0.2558	Low Silica
A6	0.9615	High Silica
A7	0.9631	High Silica
A8	0.1122	Low Silica

5.3 Problem 3

Method 1: For the classification of unknown glass types in Form 3, based on the decision tree results from Problem 2, a two-class classification is carried out using PbO as the feature. If the PbO



```
Data:   X dimension: 67 4
        Y dimension: 67 1
Fit method: kernelpls
Number of components considered: 4

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps
CV            0.45   0.1471   0.1389   0.1371
adjCV         0.45   0.1470   0.1381   0.1365
       4 comps
CV      0.1372
adjCV   0.1365

TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps
X        80.19    87.15    93.55      100
group    89.64    91.98    92.00       92
```

Figure 17: **The results of partial least quadratic regression under all principal components**

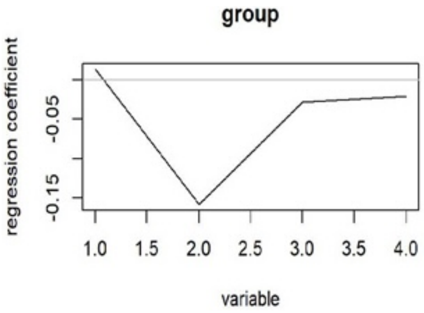


Figure 18: **Partial minimum quadratic regression coefficient**

(lead oxide) value is greater than 1.5, the sample is classified as lead–barium glass; otherwise, it is classified as high–potassium glass. The result is that A1, A6, and A7 are high–potassium glass, while A2, A3, A4, A5, and A8 are lead–barium glass.

Method 2: Using the features PbO, K_2O, BaO, and SrO (selected in Problem 2 for screening high–potassium and lead–barium glasses), a partial least squares regression (PLSR) is established between these four features and the glass type. In this regression, the dependent variable is assigned a value of 1 for high–potassium glass and 0 for lead–barium glass. Based on CV cross–validation, the RMSBP is calculated, and the regression result using all principal components is shown in **Figure 17**,.

From the regression results, it can be seen that when the number of principal components is 3, the model yields a relatively low overall RMSEP after CV cross–validation. Moreover, the cumulative contribution rate of these three principal components reaches 93% for the variables. Therefore, the number of principal components for the PLSR is set to 3.

After determining the number of principal components, the PLSR coefficients are calculated (see **Figure 18**). The regression coefficients for PbO, K_2O, BaO, and SrO are 0.0134, –0.1582, –0.0288, and –0.0206.

By inputting the chemical composition data (PbO, K_2O, BaO, and SrO) from Form 3 into the PLSR model, predicted values for the different artifacts are obtained. If a predicted value is close to 1, it indicates high–potassium glass; if it is close to 0, it indicates lead–barium glass. The prediction results are shown in **Table 17**.

From the above analysis, the prediction results from both methods are consistent. To further analyze the subclassification within high–potassium and lead–barium glasses, the subclassification results from Problem 2 are used:

For **high–potassium glass**, taking the three subclasses as the dependent variable and the chemical compositions K_2O, SiO_2, CaO, SnO_2, and BaO as independent variables, a partial least squares regression is established. Cross–validation determines that 2 principal components are optimal. The prediction on artifacts A1, A6, and A7 in Form 3 shows that A1, A6, and A7 all belong to the same subclass of high–potassium glass.

For **lead–barium glass**, taking the three subclasses as the dependent variable and the chemical compositions PbO, CuO, Na_2O, and FeO as independent variables, a partial least squares regression is established.

Cross–validation determines that 2 principal components are optimal. The prediction on artifacts A2, A3, A4, A5, and A8 in Form 3 shows that A2 and A4 belong to one subclass of lead–barium glass, A3 and A8 belong to another subclass, and A5 belongs to yet another subclass.**Code D**

5.4 Problem 4

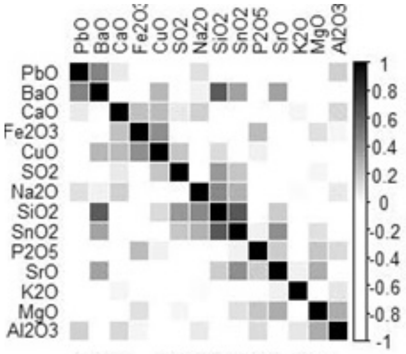


Figure 19: **High Potassium Glass Thermal Diagram**

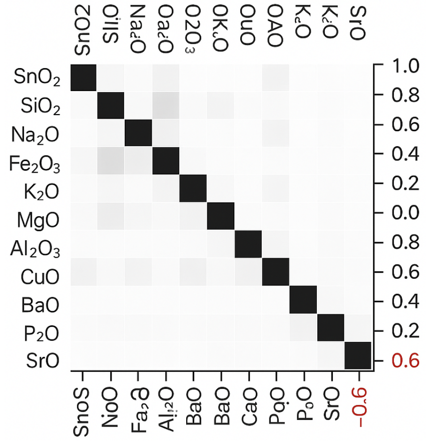


Figure 20: **Heat map of lead-barium glass**

5.4.1 Correlation Heatmap Analysis

The glass samples are analyzed separately as high–potassium and lead–barium glasses. Since the task requires analyzing the relationships among the chemical components of artifacts from different categories—and given that there are many variables—a heatmap can intuitively display the pairwise correlations between chemical components. The magnitude of the correlations can be inferred from the depth of the colors. Therefore, separate chemical element heatmaps for high–potassium and lead–barium glasses are produced, as shown in the figures below.

From the high–potassium glass heatmap **Figure 19**, it can be observed that, following the element order (from left to right) at the top of the heatmap, each element shows strong correlations with its adjacent elements. In particular, SiO_2 and BaO are strongly correlated, while the other pairs of components exhibit relatively weaker correlations.

From the lead–barium glass heatmap **Figure 20**, it can be seen that among the seven components—SnO_2, SiO_2, Na_2O, FeO, K_2O, MgO, and BaO—almost every pair is correlated (with the exception of the pair FeO and Na_2O). In addition, the components CuO, BaO, and SrO exhibit relatively strong pairwise correlations.

5.4.2 Difference Analysis

To compare the differences in the correlations of chemical components between the different glass types, only the upper triangular parts of the symmetric correlation matrices for high–potassium and lead–barium glasses are selected. Because the correlations between each pair of chemical components are paired, a non–parametric paired sample Wilcoxon test is performed. The resulting p–value is 0.905. Since the p–value is greater than 0.05, the null hypothesis is not rejected. This indicates that there is no significant difference in the correlations among the chemical components between the two types of glass.**Code** E

6 Model Testing and Reliability Analysis

6.1 Testing for Problem 1

Before addressing Problem 1, the data in the form were preprocessed. For the chemical composition of glass, near-zero values were imputed, and the data were converted into compositional data with proportions summing to 100%. This preprocessing is deemed reasonable. Considering the unique structure of compositional data, the mean was computed on the simplex of compositional data, and a Dirichlet regression model suitable for compositional data was chosen. This method is more appropriate than traditional analysis methods.

6.2 Testing for Problem 2

For Problem 2, two methods were employed to analyze the classification patterns between high-potassium glass and lead-barium glass. Both methods yielded consistent results, which further validate the rationality of the classification model. When subdividing each glass type into subcategories, the optimal number of clusters was determined based on kmcans clustering analysis, and the results are genuine and reliable.

6.3 Testing for Problem 3

The partial least squares regression method was used, with cross-validation determining the optimal number of principal components. Two methods were then applied to predict the categories of unknown glass artifacts, and the results were consistent, indicating that the predictions are reasonable.

6.4 Testing for Problem 4

Pearson correlation coefficients were used to analyze the associations between chemical components, while the Wilcoxon test was applied to compare the differences in these associations between the two types of glass. The results are genuine and reliable.**Code** F

7 Model Evaluation and Outlook

7.1 Advantages of the Model

The strengths of this study lie in its compositional data-based analysis of the chemical compositions of glass and in the classification of different glass types, as evidenced by the following points:

1. In the data preprocessing stage, the multiplicative replacement method was employed to impute missing values and zeros.

2. Considering the geometric structure of the compositional data, the mean of the chemical compositions of different artifacts was calculated.

3. The Dirichlet regression model was used to predict the chemical composition before weathering.

4. For classifying different glass types, given the constraint that compositional data sum to 100%, a centered log-ratio (clr) transformation was first applied to the chemical compositions to convert the data into a standard Euclidean space.

5. Various classification models were employed to categorize the glass types, and the results were consistent across different models.

6. Sensitivity analysis was performed for each method to select the optimal parameters.

7.2 Disadvantages of the Model

For Problem 4, the Pearson correlation coefficient was used to assess the association between the chemical compositions of different glass types. However, since the Pearson correlation coefficient only measures linear relationships between variables, and no prior test for linearity was conducted, this approach has its limitations.

7.3 Outlook for the Model

For Problem 2, when analyzing the classification patterns among different glass types, the sample sizes for the two glass types were not very balanced. Future research could consider imbalanced sample classification models, potentially improving classification by resampling the training set or adjusting the methodology. For Problem 4, alternative methods for measuring the association between chemical components—such as grey relational analysis, maximum distance correlation coefficient, or mutual information—could be explored.

References

A Brief History of Glassware. Chinese Historical Series: Material History Series. Social Sciences Academic Press, 2011. ISBN 9787509725818. URL https://books.google.com/books?id=YZR-EAAAQBAJ.

Vera Pawlowsky-Glahn and Antonella Buccianti. *Compositional data analysis*. Wiley Online Library, 2011.

Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and analysis of compositional data*. John Wiley & Sons, 2015.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zhiqiang Zhao. A study on the composition system and manufacturing techniques of glass beads unearthed from the shiren zigou site group in barkol, xinjiang. Master's thesis, Northwest University, 2016.

Part I

Appendix

Table of Contents

A Data processing and problem 1 code

```
  # Reading Data
  y <- read.table("C:/Users/86182/Desktop/data.txt", header = TRUE,
skipNul = TRUE)

  # Missing value filling
  library(zCompositions)
  library(compositions)
  library(DirichletReg)

  y_bianhan <- multRepl(y, label = 0, dl = c
(0,0.8,0.11,0.21,0.21,0,0.17,0.11,0.11,0.97,0.07,0.03,0.23,0.11))
  write.csv(y_bianhan, file = "C:/Users/86182/Desktop/data_bianhan.csv")

  y3 <- read.table("C:/Users/86182/Desktop/data.txt", header = TRUE,
skipNul = TRUE)
  y3_bianhan <- multRepl(y3, label = 0, dl = c
(0,0.8,0.11,0.21,0.21,0,0.17,0.11,0.11,0.97,0.07,0.03,0.23,0.11))
  write.csv(y3_bianhan, file = "C:/Users/86182/Desktop/data3_bianhan.csv"
)

  # CLR Transform
  gaojia <- y_bianhan[1:18, ]
  qianbei <- y_bianhan[19:67, ]
  mean(acomp(gaojia))
  mean(acomp(qianbei))

  el <- clr(y_bianhan)
  c2 <- clr(y3_bianhan)
  write.csv(c2, file = "C:/Users/86182/Desktop/c2_clr.csv")

  # Dirichlet regress
  y_chuli <- read.table("C:/Users/86182/Desktop/data.txt", header = TRUE,
 skipNul = TRUE)
  x <- read.table("C:/Users/86182/Desktop/data.txt", header = TRUE,
skipNul = TRUE)
  fenghuadata <- cbind(y_chuli, x)
  fenghuadataSy <- DR_data(y_chuli[, 1:14])

  res1 <- DirichReg(y ~ emblazonry + class + color + weathering, data =
fenghuadata)
  summary(res1)

  # predict
  x_yuce <- read.table("C:/Users/86182/Desktop/data.txt", header = TRUE,
skipNul = TRUE)
  x_pred <- predict(res1, x_yuce)
  write.csv(x_pred, file = "C:/Users/86182/Desktop/prediction.csv")
```

B Decision Tree Code

```
\lstinputlisting[style=Rstyle]
library(rpart)
library(tibble)
library(bitops)
library(rattle)
library(rpart.plot)
library(RColorBrewer)

mydata <- read.table("clipboard", header = TRUE)
sub <- sample(1:67, 47)
```

```
train <- mydata[sub, ]
test <- mydata[-sub, ]

model <- rpart(type ~ ., data = train)
fancyRpartPlot(model)

x <- subset(test, select = -type)
pred <- predict(model, x, type = "class")
k <- test[["type"]]
table(pred, k)
```

C Cluster analysis code

```
library(NbClust)
library(factoextra)
library(ggplot2)
library(cluster)

mydata <- read.table("clipboard", header = TRUE)
head(mydata)
dim(mydata)

fviz_nbclust(mydata, kmeans, method = "wss")

kmeans0 <- kmeans(mydata[, -14], centers = 4)
print(kmeans0)

fit.km <- kmeans(mydata, 4, nstart = 25)
fit.km$size
fit.km$centers

aggregate(mydata[-1], by = list(cluster = fit.km$cluster), mean)

set.seed(1234)
fit.pam <- pam(mydata[-1], k = 3, stand = TRUE)
fit.pam$medoids
clusplot(fit.pam, main = "Bivariate Cluster Plot")
```

D Problem 3 PLS regression code

```
# Form 3 Type Classification
library(pls)

mydata <- read.table("clipboard", header = TRUE, sep = "\t")
mydata.pls <- plsr(group ~ K2O + PbO + BaO + SrO, data = mydata, ncomp = 4, validation = "CV")
summary(mydata.pls)

coef_matrix <- as.matrix(coef(mydata.pls))
data_matrix <- as.matrix(mydata)[, -1]
data_matrix %*% coef_matrix

yuce <- as.matrix(read.table("clipboard", header = FALSE, sep = "\t"))
predict(mydata.pls, yuce)

# High potassium glass subclassification
mydata <- read.table("clipboard", header = TRUE, sep = "\t")
mydata.pls <- plsr(subgroup ~ SiO2 + K2O + CaO + BaO + SnO2, data = mydata, ncomp = 5, validation = "CV")
summary(mydata.pls)

```

```
mydata.pls <- plsr(subgroup ~ SiO2 + K2O + CaO + BaO + SnO2, data = mydata, ncomp = 2, validation = "CV")
yuce <- as.matrix(read.table("clipboard", header = FALSE, sep = "\t"))
predict(mydata.pls, yuce)

# Lead-barium glass subclassification
mydata <- read.table("clipboard", header = TRUE, sep = "\t")
mydata.pls <- plsr(subgroup ~ Na2O + Fe2O3 + CuO + P2O5, data = mydata, ncomp = 4, validation = "CV")
summary(mydata.pls)

mydata.pls <- plsr(subgroup ~ Na2O + Fe2O3 + CuO + P2O5, data = mydata, ncomp = 2, validation = "CV")
yuce <- as.matrix(read.table("clipboard", header = FALSE, sep = "\t"))
predict(mydata.pls, yuce)
```

E HEATMAP CODE

```
mydata <- read.table("clipboard", header = TRUE)
library(corrplot)

cor_matrix <- cor(mydata)
corrplot(cor_matrix, method = "square")

col2 <- colorRampPalette(c("#FFFFFF", "white", "#000000"), alpha = TRUE)

corrplot(cor_matrix, order = "hclust", method = "color",
         col = col2(100), tl.col = "black", tl.cex = 0.8,
         cl.pos = "r", cl.ratio = 0.2, insig = "blank",
         addgrid.col = "white")
```

F QUESTION 4 TEST CODE (WILCOXON TEST)

```
gaojia <- as.matrix(read.table("clipboard", header = FALSE, sep = "\t"))
p1 <- cor(gaojia)
p1[upper.tri(p1, diag = FALSE)] <- 0
x <- as.matrix(as.vector(p1))
x <- x[which(x != 0)]

qianbei <- as.matrix(read.table("clipboard", header = FALSE, sep = "\t"))
p2 <- cor(qianbei)
p2[upper.tri(p2, diag = FALSE)] <- 0
y <- as.matrix(as.vector(p2))
y <- y[which(y != 0)]

wilcox.test(x, y, paired = TRUE)
```