基于成分数据分析的玻璃制品分析与分类

摘　要

玻璃是人类最早批发明的人造材料之一，古代玻璃除少数天然玻璃外，其余均为人造玻璃。虽外观相似，但化学成分不同。研究玻璃制品的化学成分及鉴别对古代玻璃文物的保护具有重要意义。本文从成分数据分析视角出发，对玻璃的化学成分进行分析，并对不同类型的玻璃进行分类及预测。

首先对附件表单中数据进行预处理。表单 1 数据按照众数以及热卡填充进行缺失值插补。由于玻璃制品的化学成分属于成分数据，但因检测手段等原因可能导致其成分比例的累加和非 100%，另外很多成分检测不到或为 0 值。假设这种情况是由于仪器精度首先或四舍五入得到的。因此，选用乘法替换法对表单 2 和 3 中空白处及 0 值进行插补，最后得到的玻璃化学成分数据为成分数据。

对于问题 1，首先基于 spearman 相关系数以及卡方检验分析玻璃文物表面风化与类型、纹饰、颜色的关系，结果均表明玻璃文物表面风化与玻璃类型有关，与纹饰、颜色无关。其次，根据成分数据的 Aitchison 几何结构，在单形空间分别计算每种玻璃风化与无风化的化学成分均值，结果表明高钾玻璃风化后二氧化硅占比增加较大，氧化钾显著减少，氧化钠、氧化铝、氧化铜、氧化钡占比均有所上升；铅钡玻璃风化后二氧化硅占比降低，氧化铅、氧化钾升高，五氧化二磷降低。最后，以玻璃的化学成分为响应变量，纹饰、类型、颜色与表面风化为自变量，构建 Dirichlet 回归模型，进而根据该模型预测风化点风化前的化学成分含量，具体结果见支撑材料。

对于问题 2，首先分析高钾玻璃与铅钡玻璃的分类规律，基于问题 1 结果，选择表面风化、化学成分为分类特征，以玻璃类型为类别。由于化学成分为成分数据，因此对化学成分做对称对数比率(clr)变换。采用决策树建立分类模型，以氧化铅为特征进行两类玻璃分类，若氧化铅值大于 1.5 则为铅钡玻璃，反之为高钾玻璃。同时建立偏最小二乘判别分析(PLS-DA)，通过不同特征的投影重要性，得到对分类有影响的特征分别为氧化铅、氧化钾、氧化钡和氧化锶。两种方法得到结果一致。其次，采用 kmeans 聚类分析分别对两种类型的玻璃进行聚类，确定最优聚类个数都为 3，然后基于 PLS-DA 对每种玻璃选择合适的化学成分，结果为高钾玻璃以氧化钾、二氧化硅、氧化钙、氧化锡和氧化钡为分类特征分为三类，铅钡玻璃以五氧化二磷、氧化铜、氧化钠和氧化铁为分类特征分为三类。

对于问题 3，依据问题 2 构建模型对未知类别的玻璃进行划分。方法一是基于决策树模型，选择氧化铅成分值对未知玻璃进行划分。方法二是对表单 3 数据做 clr 变换，选择氧化铅、氧化钾、氧化钡和氧化锶成分为自变量，玻璃类型为因变量，建立偏最小二乘回归(PLS)，通过交叉验证法选择主成分个数为 3，通过预测值确定玻璃类型。两种方法得到的结果一致，A1、A6、A7 为高钾玻璃，A2、A3、A4、A5、A8 为铅钡玻璃。进一步结合问题 2 中两类玻璃的亚类划分特征，基于 PLS 得到 A1、A6、A7 都是高钾玻璃的同一亚类，A2 和 A4 是铅钡玻璃的同一亚类，A3 和 A8 是铅钡玻璃的同一亚类，A5 是铅钡玻璃的另一亚类。

对于问题 4，首先分别对每种类型玻璃的化学成分计算相关系数，观察和分析相关性热力图。然后对两类玻璃化学成分之间的相关系数进行配对 Wilcoxon 检验，结果为两类玻璃化学成分之间的相关关系无显著差异。

关键词：成分数据；决策树；偏最小二乘判别分析；聚类分析；相关分析

一、问题重述

玻璃的发展历史悠久，在中国经历了从舶来品到自主生产的过程，虽外观相似，但化学成分不同[1,2]。玻璃主要化学成分为二氧化硅，因其添加的助熔剂不同，两者成分比例不同，其可分为铅钡玻璃（助熔剂主要为铅矿石）与高钾玻璃（助熔剂主要为草木灰等含钾量较高的物质）。

图 1：玻璃制作过程

同时，因古代玻璃极易受埋藏环境的影响产生风化，导致其内外部元素进行交换，成分比例发生变化，影响其类别判断。现有已完成分类的一批铅钡玻璃与高钾玻璃制品的相关数据，其中附件表单 1 给出了文物的基本信息，表单 2 给出了已分类玻璃文物的化学成分比例，表单 3 给出了未分类玻璃文物的化学成分比例。

本文基以上信息建立数学模型解决以下问题：

问题 1：（1）分析玻璃文物表面风化情况与玻璃类型、纹饰、颜色是否有关；（2）以玻璃类型为基础，分析铅钡玻璃表面有风化与无风化的化学成分的统计规律与高钾玻璃表面有风化与无风化的化学成分的统计规律；（3）根据附件表单 2 数据中风化点的检测数据，预测风化点未风化前的化学成分含量。

问题 2：（1）根据表单 1 中玻璃的基本信息与表单 2 中玻璃检测点所含化学成分分析玻璃如何进行分类；（2）已完成分类的两种玻璃依据合适的化学成分再分别进行更进一步的分类；（3）分析分类的合理性与敏感性。

问题 3：（1）分析附件表单 3 中未知玻璃类型的化学成分，依据问题 2 中得出的分类规律对其类型进行鉴别；（2）分析分类结果的敏感性。

问题 4：（1）分析两种类型的玻璃文物样品化学成分之间的关联关系；（2）比较两种类型的化学成分之间关联关系的差异性。

二、问题分析

在回答问题之前首先对附件表单中数据进行预处理。表单 1 数据按照众数以及热卡填充进行缺失值插补。表单 2 和 3 中空白处及 0 值，选用乘法替换法进行插补，然后将化学成分数据转化为成分数据[3,4]。

2.1 问题 1

首先基于 spearman 相关系数以及卡方检验分析玻璃文物表面风化与类型、纹饰、颜色的关系。其次，根据成分数据的 Aitchison 几何结构，在单形空间分别计算每种玻璃风化与无风化的化学成分均值，再通过环形图进行比较分析。最后，以文物的化学成分为

响应变量，纹饰、类型、颜色与表面风化为自变量，构建 Dirichlet 回归模型，进而根据该模型预测风化点风化前的化学成分含量。

2.2 问题 2

为研究高钾玻璃与铅钡玻璃的分类规律，基于问题 1 的结果，选择玻璃基本信息、化学成分为分类特征，以玻璃类型为类别。由于化学成分为成分数据，因此对化学成分做对称对数比率(clr)变换。分别采用决策树和偏最小二乘判别分析进行分类，选择对于分类重要的特征。对每种每种类型的玻璃的亚类划分，采用 kmeans 聚类分析分别对两种类型的玻璃进行聚类分析，然后基于偏最小二乘判别分析对每种玻璃选择合适的化学成分。

2.3 问题 3

依据问题 2 构建的模型对未知类别的玻璃文物进行类别划分。方法一是基于决策树模型，选择筛选的特征对未知玻璃进行划分。方法二是对表单 3 数据做 clr 变换，选择偏最小二乘判别分析筛选的特征为自变量，玻璃类型为因变量，建立偏最小二乘回归，通过交叉验证法确定主成分个数，通过预测值确定玻璃类型。

2.4 问题 4

对于每种玻璃类型，我们首先采用 Pearson 相关系数分析化学成分之间的相关关系，并将此相关性以热力图的形式进行数据可视化展示，进而通过配对 Wilcoxon 检验确定两类玻璃化学成分之间的相关关系有无显著差异。

三、模型假设

1. 假设化学成分指采样点处的化学成分。
2. 假设未检测到的化学成分的原因为仪器精度受限，故暂时将化学成分缺失值记为 0，后续进行插补处理。
3. 假设检测到的化学成分 0 值是由于四舍五入得到的，后续进行插补处理。

四、符号说明

表 1：符号说明

符号	含义
X	数据集
c	成分数据的常数和约束
e	探测范围向量
e_j	成分数据集的第 j 个部分对应的探测范围
δ_{ij}	一个小于 e_j 的数
D	数据集分为 D 个部分

五、数据预处理

5.1 附件表单 1 数据预处理

本题附件表单 1 给出了玻璃文物编号、纹饰、类型、颜色、表面风化的基本信息，由于数据量较大且存在一定的缺失值，故对所给数据进行预处理弥补缺失值，防止因数据缺失对后续建模产生不利影响。

附件表单 1 中缺失值均为玻璃文物颜色，其中具有缺失值两个文物的其他信息均为“纹饰 A、表面风化、铅钡”，其他两个具有缺失值的文物的其他信息均为“纹饰 C、表面风化、铅钡”，按此分为两个类别分别进行填补。

在所给文物中，符合“纹饰 C、表面风化、铅钡”条件并已知颜色的共 15 件文物，颜色分布如下：

表 2：颜色分布（1）

颜色	蓝绿	浅蓝	浅绿	深绿	紫
对应文物编号	56，57	11，25，43，51，52，54	41	34，36，38，39	08，26

因数据为定性数据，故采用热卡填充弥补缺失值。对 15 件文物及待填充的 40，58 号文物的化学成分进行比较，用成分相似的文物颜色进行填充。（一件文物在两个部位进行采样时取均值代表，在一个部位和严重分化点采样时按 1：2 的权重计算后进行代表，在一个部位和未分化点采样时将部位的化学成分作为代表）经比较可知：40 号与 39 号成分相似，58 号与 11 号成分相似，故将 40 号、58 号文物颜色分别填充为深绿，浅蓝。

符合“纹饰 A、表面风化、铅钡”条件并已知颜色的共 9 件文物，颜色分布如下：

表 3：颜色分布（2）

颜色	黑	蓝绿	浅蓝
对应文物编号	49，50	23	02，28，29，42，44，53

由附件表单 2 可知，9 件文物中有 6 件文物在分析化学成分时采样点为未风化点，所得到的成分含量参考性较低，故对此采用众数进行填补，即 19、48 号文物颜色填充为浅蓝。

5.2 附件表单 2、3 数据预处理

5.2.1 缺失值插补

因成分比例累加和介于 85%—105%之间的数据视为有效数据，故将各成分比例累加，得到无效数据为“文物 15 号、文物 17 号”，将其从表单中去除。表单 2 和 3 中出现的空白为未检测到该成分。假设因仪器精度受限未检测到成分，所以不能单纯将缺失值插补为“0”，可看作近似零值。另外，表单 2 和 3 中有 0 值，假设是由于四舍五入得到，属于近似零值。采用如下乘法替换方法对近似零值进行插补，具体过程为：

考虑成分数据集

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

假定成分数据中有近似零值，近似零值是由于小于已知的探测范围观测不到而产生的，且不同成分数据相同部分对应的探测范围是相同的。记探测范围向量为 $e=(e_1,e_2,\cdots,e_D)^T$，其中 e_j 为成分数据集 X 的第 j 个部分对应的探测范围。后提出乘法简单替换法，x_{ij} 替换后的数据为

$$xr_{ij}=\begin{cases} \delta_{ij}, & x_{ij}=0 \\ x_{ij}\left(1-\dfrac{\sum_{k|x_{ik}=0}\delta_{ik}}{c}\right), & x_{ij}>0 \end{cases}$$

其中 δ_{ij} 为一个小于 e_j 的数，c 为成分数据的常数和约束，即

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

通过实验发现，当成分数据集中近似零值比例不高时，δ_{ij} 等于探测范围的 65%时可以最小化协方差阵的扭曲，即 $\delta_{ij}=0.65e_j$。记附件表单 2、3 中空白处所在列的最小值为临界值，并将其乘以 0.65，得到插补值。（在此列出部分插补值，具体数据见支撑材料）

表 4：表单 2 成分数据插补缺失值（部分）

文物采样点	SiO_2	Na_2O	K_2O	……	SnO_2	SO_2
01	69.3300	0.5275	9.9900		0.1517	0.3900
03 部位 1	87.0500	0.5287	5.1900		0.1520	0.0727
03 部位 2	61.7100	0.5239	12.3700		0.1506	0.0720
……						
57	25.4200	0.5264	0.0724		0.1513	0.0724
58	30.3900	0.5239	0.3400		0.1506	0.0720

表 5：表单 3 成分数据插补缺失值（部分）

文物编号	表面风化	SiO_2	Na_2O	K_2O	……	SnO_2	SO_2
A1	无风化	78.4500	0.5277	0.0726		0.1517	0.5100
A2	风化	37.7500	0.5298	0.0728		0.1523	0.0728
A3	无风化	31.9500	0.5239	1.3600		0.1506	0.0720
A4	无风化	35.4700	0.5240	0.7900		0.1507	0.0721
A5	风化	64.2900	1.2000	0.3700		0.4900	0.0716
A6	风化	93.1700	0.5278	1.3500		0.1517	0.0726
A7	风化	90.8300	0.5281	0.9800		0.1518	0.1100
A8	无风化	51.1200	0.5248	0.2300		0.1509	2.2600

5.2.2 成分数据处理

因成分数据相加需为 100，可利用相对信息对数据进一步处理。相对信息指的是成分数据仅有的信息反映在成分间的比率中，每个成分的绝对数据是无关的，如果成分数据的每个成分乘以相同的正常数，则成分间的比率是不变的，因此成分数据可以看成是等价类，这个类里面的成分数据含有相同的信息，都可以通过适合的尺度因子表示为相同的比例向量。这可进行闭合运算：

$$C(x)=C(x_1,x_2,\cdots,x_D)^T=\left(\frac{k\cdot x_1}{\sum_{i=1}^{D}x_i},\frac{k\cdot x_2}{\sum_{i=1}^{D}x_i},\cdots,\frac{k\cdot x_D}{\sum_{i=1}^{D}x_i}\right)^T$$

闭合运算就是对初始向量乘以合适的尺度因子，使得闭合后的成分和为常数 k（这里为 100），对于任意的两个向量 $x,y\in\mathbb{R}_+^D$，如果 $C(x)=C(y)$，则 x 和 y 是成分等价的[3]。故将每个玻璃文物化学成分乘以相应计算出的因子得到最终数据。表单 2、3 数据处理结果见支撑材料。

六、模型建立与求解

6.1 问题 1

6.1.1 玻璃表面风化与纹饰、玻璃类型、颜色的相关性分析

鉴于以上分析，我们首先进行斯皮尔曼相关系数分析。在分析之前对定性数据进行虚拟变量处理，见表 4。然后利用 SPSS 软件分别对纹饰、玻璃类型、颜色与文物表面风化进行 spearman 相关性检验，得到结果见表 5 至表 7。结果表明纹饰与颜色对于玻璃表面是否风化无相关关系，玻璃类型对于玻璃表面是否风化存在相关关系。

表 6：虚拟变量处理

纹饰		颜色	
A	1	浅蓝	1
B	2	深蓝	2
C	3	蓝绿	3
类型		浅绿	4
铅钡	1	深绿	5
高钾	2	绿	6
表面风化		紫	7
无风化	0	黑	8
风化	1		

表 7：纹饰与表面风化相关性

	纹饰	表面风化
纹饰	1.000	0.080
表面风化	0.080	1.000

表 8：玻璃类型与表面风化相关性

	玻璃类型	表面风化
玻璃类型	1.000	-0.301
表面风化	-0.301	1.000

表 9：颜色与表面风化相关性

	颜色	表面风化
颜色	1.000	0.088
表面风化	0.088	1.000

其次，为保证结果的可信度，我们同时利用了卡方检验对数据进行了分析，在进行分析之前对数据进行了频数统计，见图 2 至图 4：

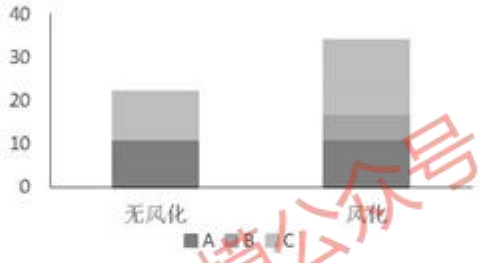

图 2：纹饰

由图 2 可知，在无风化与风化两种情况下，纹饰类型 A、B、C 变化差异较小，故可初步推断表面风化与纹饰类型无关。

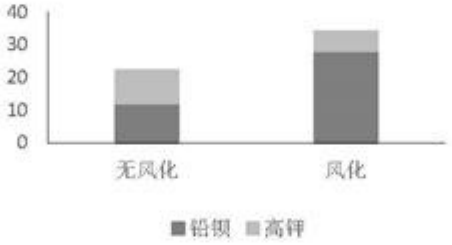

图 3：玻璃类型

由图 3 可知，风化与无风化相比，铅钡玻璃类型占比增加，高钾玻璃类型占比减少，且增加与减少幅度较大，故可初步推断表面风化与玻璃类型有关。

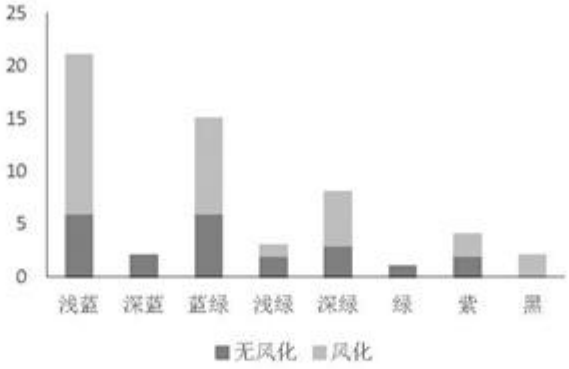

图 4：颜色

由图 4 可知，风化与无风化相比，各种颜色变化较小，故可初步推断表面风化与颜色无关。

为进一步验证推断是否合理，利用 SPSS 软件进行了卡方检验，结果见表 10。

表 10：卡方检验结果

	系数
纹饰	0.085
玻璃类型	0.024
颜色	0.325

因只有玻璃类型的系数小于 0.05，故玻璃类型与文物表面风化有相关关系，纹饰与颜色的系数均大于 0.05，故纹饰、颜色与文物表面风化无相关关系。

6.1.2 不同类型玻璃表面有风化与无风化化学成分含量的统计规律分析

此问运用附件表单 2 处理后的成分数据进行分析。分别计算出属于铅钡玻璃的无风化与风化检测点数据均值、属于高钾玻璃的无风化与风化检测点数据均值。具体计算方法如下：

因成分数据为几何结构，故可进行扰动运算来类似实数空间上的加法运算。对于任意成分数据 $x=(x_1,x_2,\cdots,x_D)^T$， $y=(y_1,y_2,\cdots,y_D)^T\in S^D$,x 与 y 的扰动运算定义为：

$$x\oplus y=C(x_1y_1,x_2y_2,\cdots,x_Dy_D)^T\in S^D$$

故可进一步计算均值。其可运用 R 语言进行成分数据均值计算，得到结果如下表：

表 11：高钾玻璃、铅钡玻璃风化与无风化均值

	SiO_2	Na_2O	K_2O	……	SnO_2	SO_2
高钾未风化	74.9529	0.8818	7.2326		0.2110	0.1234
高钾风化	93.6615	0.5268	0.3579		0.1515	0.0724
铅钡未风化	59.5046	1.3090	0.1851		0.1871	0.0953
铅钡风化	27.0317	0.7432	0.1502		0.2105	0.1745

为更直观地观察化学成分变化，作出下图：

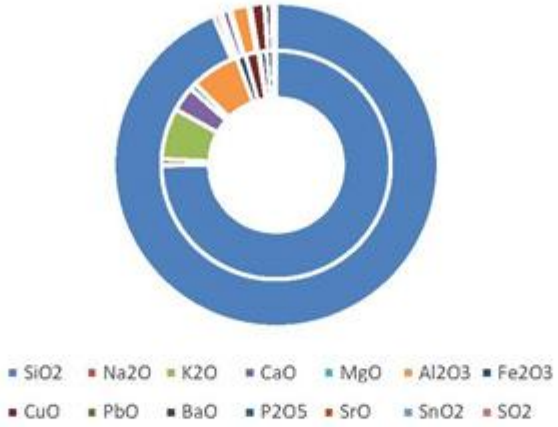

图 5：高钾玻璃风化前后化学成分对比，外圈为风化，内圈为无风化

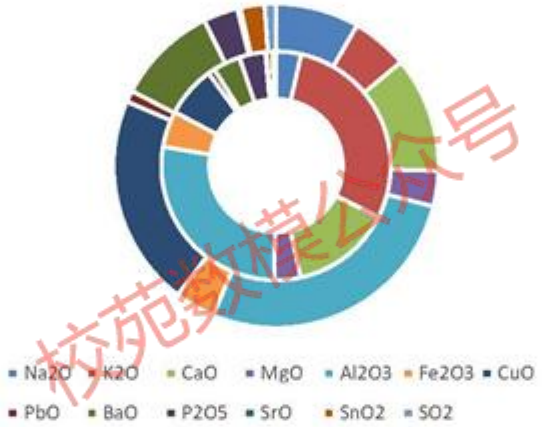

图 6：高钾玻璃风化前后化学对比（未含 SiO2），外圈为风化，内圈为无风化

对于高钾玻璃，因二氧化硅占比较大，图 5 不能很好地体现其他化学成分的变化，故将二氧化硅去掉，单独制作图 6 更好地体现其他化学成分占比变化。结合图 5 与图 6 进行分析，风化后二氧化硅占比增加较大，氧化钾显著减少，氧化钠、氧化铝、氧化铜、氧化钡占比均有所上升。

图 7：铅钡玻璃分化前后元素对比，外圈为风化，内圈为无风化

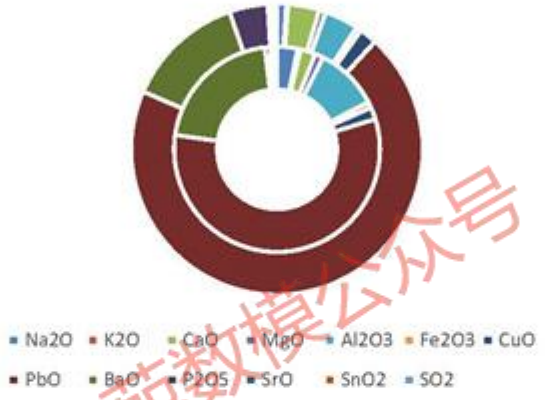

图 8：铅钡玻璃风化前后元素对比（未含 SiO2），外圈为风化，内圈为无风化

对于铅钡玻璃，由图 7 可知，无风化与风化相比，二氧化硅占比降低，氧化铅的占比升高。同样由于二氧化硅占比较大影响其他化学成分变化的分析，将二氧化硅去除得到图 8，可分析得出氧化钾，五氧化二磷变化程度较大，变化趋势分别为升高、降低。

6.1.3 风化前化学成分预测

首先对附件表单 1 中的数据进行取虚拟变量处理，见上表 4。将附件表单 1 中的变量作为自变量，附件表单 2 中的化学变量作为因变量，做 Dirichlet 回归。得到回归模型后，将自变量中的风化（1）替换为未风化（0），得到预测部分结果如下表（因预测结果数据量较大，故此处只写出了部分预测结果，所有预测结果见支撑材料）。

表 12：预测风化前的化学成分含量（部分）

文物编号	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

6.2 问题 2

6.2.1 玻璃的分类规律分析

在进行建模之前首先要对成分数据进行 clr 变换，对于任意成分数据 $x=(x_1,x_2,\cdots,x_D)^T\in S^D$，clr 变换将 $x\in S^D$ 变换为 ξ^D 上的系数，clr 系数为

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

记 clr 变换后数据为 $clr(x)=\xi=(\xi_1,\xi_2,\cdots,\xi_D)^T$，则 clr 逆变换为

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

用转化后的成分数据建立分类模型，为使分类结果直观明了，使用决策树模型对其进行特征的选择与分类，进一步探究主要分类特征从而推断两种玻璃的分类规律。在此模型中，本题将附件表单 1 中的表面风化与附件表单 2 中各化学成分作为分类特征、将玻璃的类型作为类别构建决策树模型，且将 67 个样本划分为训练集与测试集两类，其中 47 个样本为训练集剩余则为测试集。最后采用 R 语言进行软件实现，结果如图 9。再用测试集数据进行预测，正确率为 100%。

图 9：决策树结果

从该图中可知，最终以氧化铅（PbO）为特征进行两类玻璃分类，若氧化铅值大于 1.5 则为铅钡玻璃，反之为高钾玻璃。

下面采用偏最小二乘判别分析进行高钾玻璃和铅钡玻璃的分类。由于 clr 变换后数据求和为 0，以玻璃类型为分类变量，风化、二氧化硅(SiO_2)、氧化钠(Na_2O)、氧化钾(K_2O)、氧化钙(CaO)、氧化镁(MgO)、氧化铝(Al_2O_3)、氧化铁(Fe_2O_3)、氧化铜(CuO)、氧化铅(PbO)、氧化钡(BaO)、五氧化二磷(P_2O_5)、氧化锶(SrO)、氧化锡(SnO_2)、二氧化硫(SO_2)为特征变量，建立偏最小二乘判别分析。由图 10 可知，高钾玻璃和铅钡玻璃有明显的分离趋势。由图 11 可知，Q2 左侧交于 Y 轴负半轴，说明模型构建成功。

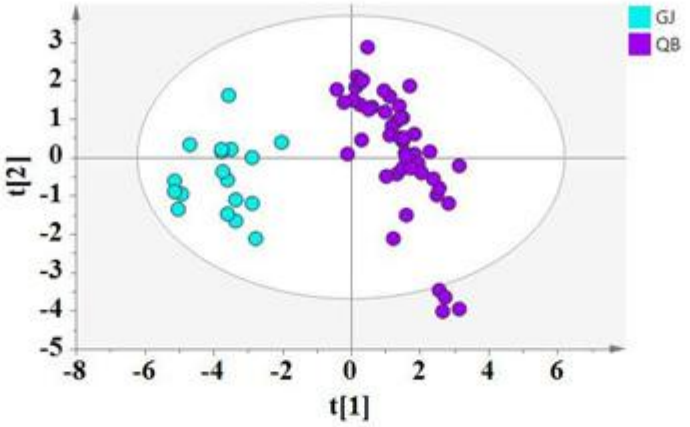

图 10：偏最小二乘判别分析结果

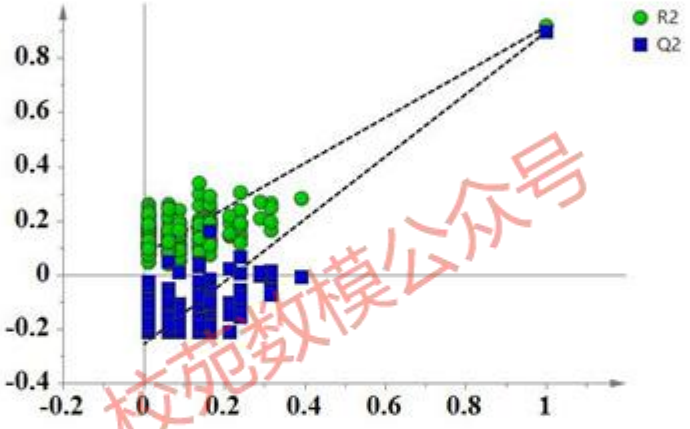

图 11：偏最小二乘判别分析模型验证

表 13：偏最小二乘判别分析的不同特征的 VIP 值

变量	VIP	变量	VIP
风化	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		

计算不同变量的投影重要性（VIP），结果见表 10。筛选 VIP 值大于 1 的特征，对于分类有影响的特征分别为 PbO、K_2O、BaO 和 SrO。上面图表结果是在 simca 软件操作完成。

6.2.2 亚类划分方法结果及敏感性分析

将两类玻璃分别进行更进一步亚类的划分，本次划分并不知亚类划分的结果故该批数据并未有明确的类型，因此该类问题为无监督学习问题没有已知的因变量，且需进行分类所以选择聚类分析模型。首先需要进行选择所分类别的个数，我们采用 R 语言中的 fviz_nbclust 函数进行判断，如图根据其间断点可知两种玻璃的亚分类最终均选取 3 类：

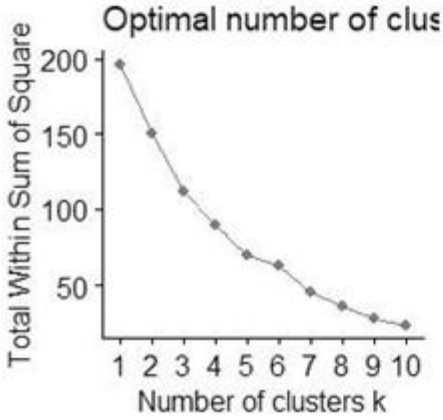

图 12：高钾分类个数确定

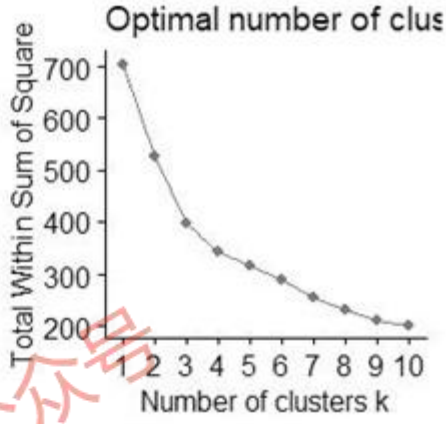

图 13：铅钡分类个数确定

其次，在确定聚类个数的情况下用 K-means 进行聚类,结果如图：

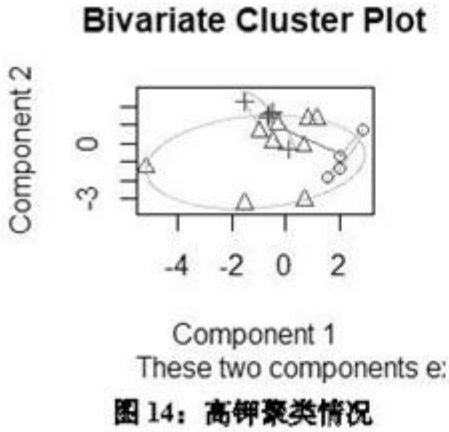

图 14：高钾聚类情况

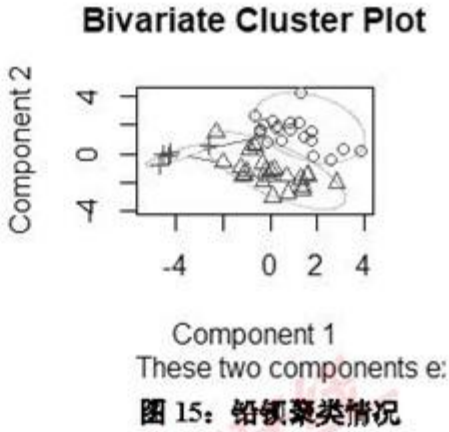

图 15：铅钡聚类情况

分析运行结果图可知铅钡玻璃可被明显的分为三类，故其亚分类包含三类；高钾玻璃三类划分并未有非常明显的类别，仅可粗略将其分为三类。

表 14：亚分类

玻璃类型	亚分类	文物编号
高钾玻璃	1	06 部位 1，18
	2	21，07，09，10，12，22，27
	3	01，03 部位 1，03 部位 2，04，05，06 部位 2，13，14，16
铅钡玻璃	1	20，37，50 未风化点，08，08 严重风化，11，19，26，26 严重风化，39，40，43 部位 2，50，51 部位 1，52，54，54 严重风化，56，58
	2	28 未风化点，29 未风化点，30 部位 1，30 部位 2，31，32，35，49 未风化点，02，41，48，49，51 部位 2
	3	23 未风化点，24，25 未风化点，33，42 未风化点 1，42 未风化点 2，44 未风化点，45，46，47，53 未风化点，55，34，36，38，43 部位 1，57

在确定分类个数时，不同分类个数的选择会使得聚类出现不同的结果，故应对此模型的敏感性进行分析。因高钾玻璃并未有明显的分类效果，故尝试多个分类个数进行聚类从而选择出较优的聚类情况。

下面对每类玻璃的亚类划分选择合适的化学成分。

对于高钾玻璃，亚类划分为 3 类。由于其中一类只有两个文物玻璃，因此删除这一类文物，以其余类为分类变量，化学成分为特征，构建偏最小二乘判别分析，由图 16 可知，高钾玻璃的亚类有明显的分离趋势。由图 17 可知，Q2 左侧交于 Y 轴负半轴，说明模型构建成功。

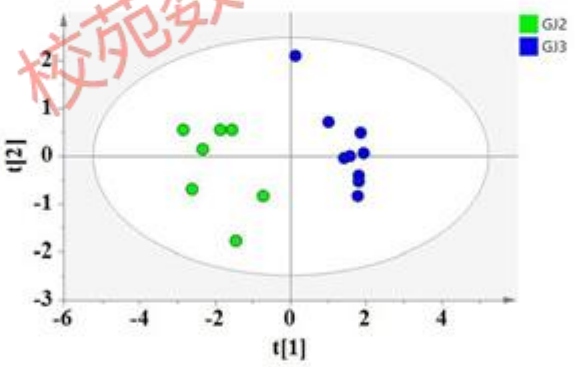

图 16：高钾玻璃亚类偏最小二乘判别分析结果

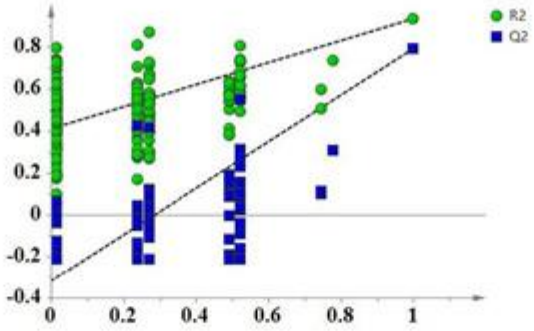

图 17：高钾玻璃亚类偏最小二乘判别分析模型验证

表 15：高钾玻璃亚类偏最小二乘判别分析的不同特征的 VIP 值

变量	VIP	变量	VIP
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

计算不同变量的投影重要性（VIP），结果见表 15。筛选 VIP 值大于 1 的特征，对于高钾玻璃亚类分类有影响的特征分别为 K_2O、SiO_2、CaO、SnO_2 和 BaO。上面图表结果是在 simca 软件操作完成。

对于铅钡玻璃，亚类划分为 3 类，以这 3 类为分类变量，化学成分为特征，构建偏最小二乘判别分析，由图 18 可知，铅钡玻璃的三类有明显的分离趋势。由图 19 可知，Q2 左侧交于 Y 轴负半轴，说明模型构建成功。

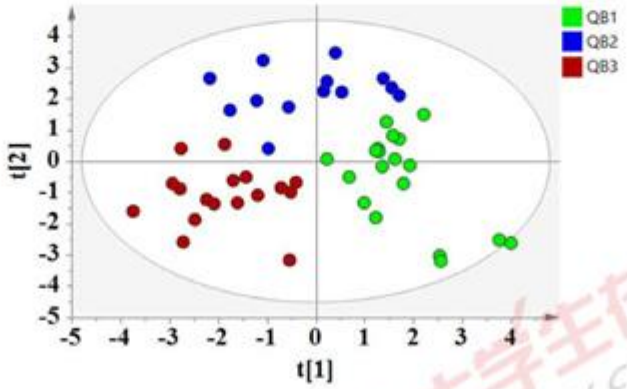

图 18：铅钡玻璃亚类偏最小二乘判别分析结果

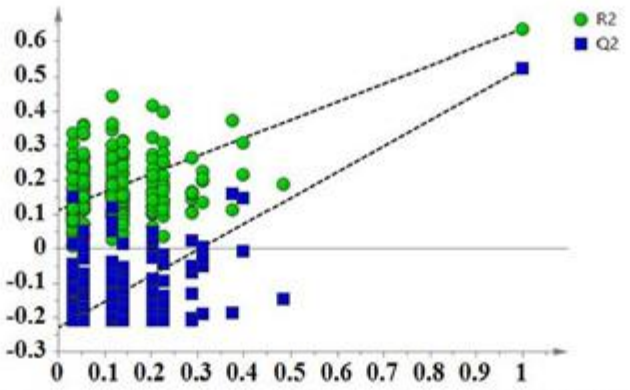

图 19：铅钡玻璃亚类偏最小二乘判别分析模型验证

表 16：铅钡玻璃亚类偏最小二乘判别分析的不同特征的 VIP 值

变量	VIP	变量	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

计算不同变量的投影重要性（VIP），结果见表 16。筛选 VIP 值大于 1 的特征，对于铅钡玻璃亚类分类有影响的特征分别为 P_2O_5、CuO、Na_2O 和 Fe_2O_3。上面图表结果是在 simca 软件操作完成。

6.3 问题 3

方法一：对于表单 3 中未知玻璃类型的类别划分，基于问题 2 决策树的结果，以 PbO 为特征进行两类玻璃分类，若氧化铅值大于 1.5 则为铅钡玻璃，反之为高钾玻璃。结果为 A1、A6、A7 为高钾玻璃，A2、A3、A4、A5、A8 为铅钡玻璃。

方法二：选取问题 2 对于高钾玻璃和铅钡玻璃筛选的特征 PbO、K_2O、BaO 和 SrO，建立这四个特征与玻璃类型的偏最小二乘回归分析，其中因变量玻璃类型中高钾玻璃取值为 1，铅钡玻璃取值为 0。

基于 CV 交叉验证方法计算 RMSEP，使用所有主成分进行回归得到的结果如下图：

```
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
```

图 20：所有主成分下偏最小二乘回归结果

从回归结果可以看出，主成分个数为 3 时，模型在经 CV 交叉验证后得到的 RMSEP 综合最小，同时 3 个主成分对各变量的累计贡献率已经达到 93%，因此将偏最小二乘回归的主成分个数设定为 3。

主成分个数确定后计算得到偏最小二乘回归系数如下图 17，PbO、K_2O、BaO 和 SrO 回归系数分别为 0.0134、-0.1582、-0.0288、-0.0206。

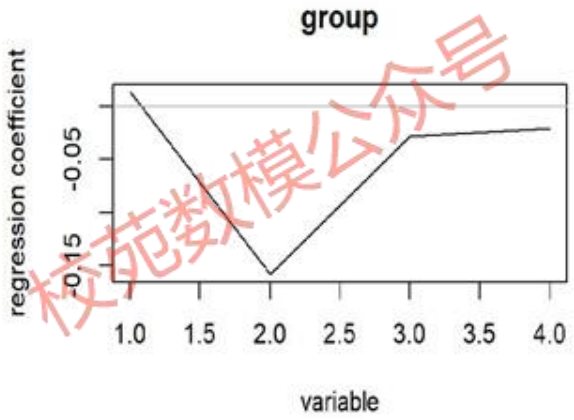

图 21：偏最小二乘回归系数

将表单 3 中 PbO、K_2O、BaO 和 SrO 化学成分数据带入偏最小二乘回归模型，预测得到不同文物的预测值，如果预测值接近 1，则为高钾玻璃，预测值接近 0，则为铅钡玻璃。预测结果见下表 17。

表 17：表单 3 未知玻璃文物的类型预测

文物编号	预测值	预测玻璃类型
A1	1.0408	高钾
A2	0.0177	铅钡
A3	0.0980	铅钡
A4	0.1898	铅钡
A5	0.2558	铅钡
A6	0.9615	高钾
A7	0.9631	高钾
A8	0.1122	铅钡

通过上面分析，可以看出两种方法预测结果一致。

为了更进一步分析高钾玻璃与铅钡玻璃的亚类划分，基于问题 2 的亚类划分结果。对于高钾玻璃，以亚类 3 类为因变量，化学成分 K_2O、SiO_2、CaO、SnO_2 和 BaO 为自变量，建立偏最小二乘回归，通过交叉验证确定主成分个数为 2，对表单 3 中 A1，A6，A7 文物进行预测，结果是 A1，A6，A7 都是高钾玻璃的同一亚类。

对于铅钡玻璃，以亚类 3 类为因变量，化学成分 P_2O_5、CuO、Na_2O 和 Fe_2O_3 为自变量，建立偏最小二乘回归，通过交叉验证确定主成分个数为 2，对表单 3 中 A2，A3，A4，A5，A8 文物进行预测，结果是 A2 和 A4 是铅钡玻璃的同一亚类，A3 和 A8 是铅钡玻璃的同一亚类，A5 是铅钡玻璃的另一亚类。

6.4 问题 4

6.4.1 相关性热力图分析

将玻璃按照高钾与铅钡玻璃进行分类讨论，因为题目中要求分析不同类别文物化学成分之间的关联关系，且因变量较多，热力图更可直观地体现两两化学成分之间的相关关系，根据颜色深浅比较相关关系的大小，故可分别制作高钾玻璃与铅钡玻璃的化学元素热力图，如下图所示：

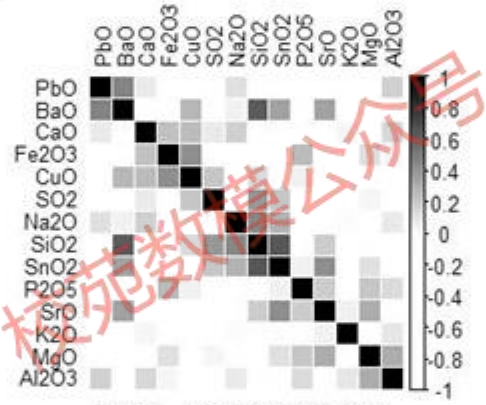

图 22：高钾玻璃热力图

由高钾玻璃热力图可知，按照热力图上方从左到右元素顺序，每个元素与左右相邻元素相关性较强，SiO_2 与 BaO 关联性较强，其余两两成分之间关联性相对较弱。

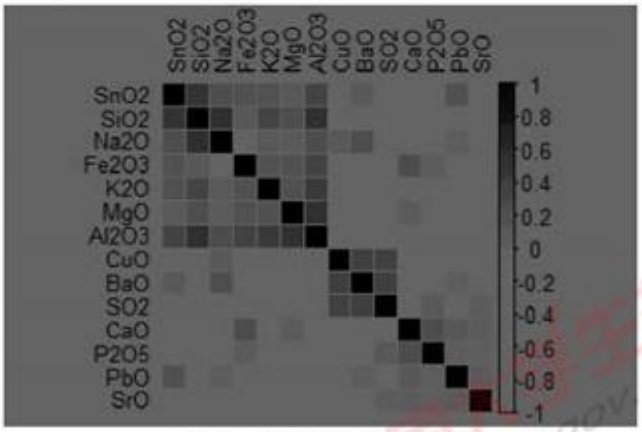

图 23：铅钡玻璃热力图

由铅钡玻璃热力图可知，SnO_2、SiO_2、Na_2O、Fe_2O_3、K_2O、MgO、Al_2O_3这七个成分两两之间均存在相关性（Fe_2O_3和Na_2O之间除外）。CuO、BaO、SO_2两两之间均存在较为强烈的相关性。

6.4.2 差异性分析

为了比较不同类别之间的化学成分关联关系的差异性，由于热力图中相关系数是对称矩阵。因此仅选取高钾玻璃相关系数上三角矩阵与铅钡玻璃相关系数上三角矩阵。由于每个化学成分之间的相关关系是配对的，因此选取非参数配对样本 wilcoxon 检验，得到结果 p 值为 0.905。由于 p 值大于 α（0.05），因此不拒绝原假设，即两种玻璃类型的化学成分之间的关联关系没有显著差异。

七、模型检验及可靠性分析

7.1 针对问题一的检验

在进行问题 1 之前，首先对表单中数据进行预处理。对于玻璃的化学成分，近似零值插补后转换为成分数据，化学成分比例和为 100%。数据预处理合理。对于问题 1，考虑到成分数据的特殊结构，在成分数据单形空间上计算均值，选择适用于成分数据的 Dirichlet 回归模型，分析方法相比传统分析方法更加合理。

7.2 针对问题二的检验

对于问题 2，选择两种方法分析高钾玻璃与铅钡玻璃的分类规律，两种方法结果一致，进一步验证了分类模型合理性。对每类玻璃亚类划分时，基于 kmeans 聚类分析确定了最优聚类个数，分析结果真实可靠。

7.3 针对问题三的检验

偏最小二乘回归方法基于交叉验证确定了最优主成分个数。基于两种方法对未知玻璃文物进行类别预测，结果一致。因此预测结果合理。

7.4 针对问题四的检验

采用 Pearson 相关系数分析化学成分之间的关联关系，基于 wilcoxon 检验比较两种类型玻璃化学成分之间关联关系的差异性。结果真实可靠。

八、模型评价与展望

8.1 模型的优点

本文优点是基于成分数据分析对玻璃化学成分进行分析，并对不同玻璃类型进行分类，具体体现在如下方面：

（1） 在数据预处理阶段，采用乘法替换法对空白处及 0 值进行插补。

（2） 结合成分数据的几何结构，计算不同文物化学成分的均值。

（3） 采用 Dirichlet 回归模型对风化前的化学成分进行预测。

（4） 对于不同玻璃类型的分类，考虑到成分数据求和为 100%的约束，在模型构建前，首先对化学成分进行 clr 变换，使得成分数据变换为欧式空间上的普通数据。

（5） 采用不同分类模型对玻璃类型进行分类，不同模型结果一致。

（6） 对于每种方法，都对参数进行敏感性分析，选择最优参数。

8.2 模型的缺点

对于问题 4，不同玻璃类型的化学成分之间的关联关系，采用 Pearson 相关系数。但 Pearson 相关系数只能度量变量之间的线性相关关系，在使用之前未进行线性相关关系检验。

8.3 模型的展望

对于问题 2，分析不同玻璃之间的分类规律时，由于两种类型玻璃样本量不是很接近，因此后续可以考虑不平衡样本分类模型，通过对训练集样本重采样或方法修正来进去分类。对于问题 4，考虑其他相关系数方法来度量化学成分之间的关联关系，例如灰色关联分析，最大距离相关系数、互信息等。

九、参考文献

[1] 赵志强. 新疆巴里坤石人子沟遗址群出土玻璃珠的成分体系与制作工艺研究[D]. 西北大学, 2016.

[2] 安家瑶. 玻璃器史话[M]. 北京: 社会科学文献出版社, 2011: 7-11.

[3] Pawlowsky-Glahn V, Buccianti A. Compositional Data Analysis: Theory and Applica - tions [M]. Wiley, 2011.

[4] Pawlowsky-Glahn V, Egozcue J J, Tolosana-Delgado R. Modeling and Analysis of Compositional Data[M]. Wiley, 2015.

附　录

支撑材料：

1.数据：

- 数据1：附件表单1数据插补缺失值.xlsx
- 数据2：附件表单2数据处理.xlsx
- 数据3：附件表单3数据处理.xlsx
- 数据4：风化点风化前化学成分预测数据.xlsx
- 数据5：决策树.xlsx
- 数据6：聚类.xlsx
- 数据7：热力图.xlsx
- 数据8：myydata.xlsx

2.辅助资料：

- Dirichlet回归结果.docx
- 高钾玻璃、铅钡玻璃风化与无风化均值.docx

3.图：

- 图1：玻璃制作过程.docx
- 图2：纹饰.docx
- 图3：玻璃类型.docx
- 图4：颜色.docx
- 图5：高钾玻璃风化前后化学成分对比，外圈为风化，内圈为无风化.docx
- 图6：高钾玻璃风化前后化学对比（未含SiO2），外圈为风化，内圈为无风化.docx
- 图7：铅钡玻璃分化前后元素对比，外圈为风化，内圈为无风化.docx
- 图8：铅钡玻璃风化前后元素对比（未含SiO2），外圈为风化，内圈为无风化.docx
- 图9：决策树结果.docx
- 图10：偏最小二乘判别分析结果.docx
- 图11：偏最小二乘判别分析模型验证.docx
- 图12：高钾分类个数确定 .docx
- 图13：铅钡分类个数确定.docx
- 图14：高钾聚类情况 .docx
- 图15：铅钡聚类情况.docx

图16：高钾玻璃亚类偏最小二乘判别分析结果.docx

图17：高钾玻璃亚类偏最小二乘判别分析模型验证.docx

图18：铅钡玻璃亚类偏最小二乘判别分析结果.docx

图19：铅钡玻璃亚类偏最小二乘判别分析模型验证.docx

图20：所有主成分下偏最小二乘回归结果.docx

图21：偏最小二乘回归系数.docx

图22：高钾玻璃热力图.docx

图23：铅钡玻璃热力图.docx

4.表：

表1：符号说明.docx

表2：颜色分布（1）.docx

表3：颜色分布（2）.docx

表4：表单2成分数据插补缺失值（部分）.docx

表5：表单3成分数据插补缺失值（部分）.docx

表6：虚拟变量处理.docx

表7：纹饰与表面风化相关性.docx

表8：玻璃类型与表面风化相关性.docx

表9：颜色与表面风化相关性.docx

表10：卡方检验结果.docx

表11：高钾玻璃、铅钡玻璃风化与无风化均值.docx

表12：预测风化前的化学成分含量（部分）.docx

表13：偏最小二乘判别分析的不同特征的VIP值.docx

表14：亚分类.docx

表15：高钾玻璃亚类偏最小二乘判别分析的不同特征的VIP值.docx

表16：铅钡玻璃亚类偏最小二乘判别分析的不同特征的VIP值.docx

表17：表单3未知玻璃文物的类型预测.docx

5.代码：

1附录一：数据处理及问题1代码.R

2附录二：决策树代码.R

3附录三：聚类分析代码.R

4附录四：问题3代码.R

5附录五：热力图代码.R

6附录六：问题4检验代码.R

附录一 数据处理及问题 1 代码

```
y <- read.table("C:\\Users\\86182\\Desktop\\data.txt",header = T,skipNul = T)
#成分数据缺失值填补
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.csv")
y3 <- read.table("C:\\Users\\86182\\Desktop\\data.txt",header = T,skipNul = T)
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/data.csv")
#成分数据求均值及 clr 变换
gaojia <- y_bianhan[1:18,]
mean(acomp(gaojia))
qianbei <- y_bianhan[19:67,]
mean(acomp(qianbei))
x_gaojia0 <- mean(acomp(x_gaojia0))
x_gaojia1 <- mean(acomp(x_gaojia1))
x_qianbei0 <- mean(acomp(x_qianbei0))
x_qianbei1 <- mean(acomp(x_qianbei1))
c1 <- clr(y_bianhan)
c2 <- clr(y3_bianhan)
write.csv(c2,file = "C:/Users/86182/Desktop/data.csv")
#dirichlet 回归
y_chuli <- read.table("C:\\Users\\86182\\Desktop\\data.txt",header = T,skipNul = T)
x <- read.table("C:\\Users\\86182\\Desktop\\data.txt",header = T,skipNul = T)
fenghuadata <- cbind(y_chuli,x)
fenghuadata$y <- DR_data(y_chuli[,1:14])
res1 <- DirichReg(y ~ emblazonry + class + color + weathering, fenghuadata)
summary(res1)
#预测
```

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

附录二 决策树代码（R 语言）

```
library(rpart)
library(tibble)
library(bitops)
library(rattle)
library(rpart.plot)
library(RColorBrewer)
mydata<-read.table("clipboard",header=T)
sub<-sample(1:67,47)
train<-mydata[sub,]
test<-mydata[-sub,]
model <- rpart(类型~.,data = train)
fancyRpartPlot(model)
x<-subset(test,select=-类型)
pred<-predict(model,x,type="class")
k<-test[,"类型"]
table(pred,k)
```

附录三 聚类分析代码（R 语言）

```
library(NbClust)
library(factoextra)
library(ggplot2)
mydata<-read.table("clipboard",header=T)
head(mydata)
dim(mydata)
fviz_nbclust(mydata, kmeans, method = "wss")
kmeansO<-kmeans(mydata[,-14],centers=4)
print(kmeansO)
fit.km <- kmeans(mydata, 4, nstart=25)
fit.km$size
fit.km$centers
aggregate(mydata[-1], by=list(cluster=fit.km$cluster), mean)
library(cluster)
set.seed(1234)
fit.pam <- pam(mydata[-1], k=3, stand=TRUE)
fit.pam$medoids
clusplot(fit.pam, main="Bivariate Cluster Plot")
```

附录四 问题 3 代码

```
##表单 3 类型划分
mydata=read.table("clipboard", header = T, sep = '\t')
mydata.pls <- plsr(group ~K2O+PbO+BaO+SrO,data=mydata, 4, validation = "CV")
summary(mydata.pls)
mydata.pls <- plsr(group ~K2O+PbO+BaO+SrO,data=mydata, 3, validation = "CV")
as.matrix(coef(mydata.pls))
as.matrix(mydata)[,-1]%*%as.matrix(coef(mydata.pls))
yuce=as.matrix(read.table("clipboard", header = F, sep = '\t'))
predict(mydata.pls,yuce)

###表单 3 高钾玻璃亚类划分
mydata=read.table("clipboard", header = T, sep = '\t')
mydata.pls <- plsr(subgroup ~SiO2+K2O+CaO+BaO+SnO2
,data=mydata, 5, validation = "CV")
summary(mydata.pls)
mydata.pls <- plsr(subgroup ~SiO2+K2O+CaO+BaO+SnO2,data=mydata, 2, validation =
"CV")
yuce=as.matrix(read.table("clipboard", header = F, sep = '\t'))
predict(mydata.pls,yuce)

###表单 3 铅钡玻璃亚类划分
mydata=read.table("clipboard", header = T, sep = '\t')
mydata.pls <- plsr(subgroup ~Na2O+Fe2O3+CuO+P2O5,data=mydata, 4, validation = "CV")
summary(mydata.pls)
mydata.pls <- plsr(subgroup ~Na2O+Fe2O3+CuO+P2O5,data=mydata, 2, validation = "CV")
yuce=as.matrix(read.table("clipboard", header = F, sep = '\t'))
predict(mydata.pls,yuce)
```

附录五 热力图代码（R 语言）

```
mydata<-read.table("clipboard",header=T)
library(corrplot)
cor<-cor(mydata)
corrplot(cor,method="square")
col2 <- colorRampPalette(c("#FFFFFF","white" ,"#000000"),alpha = TRUE)
corrplot(cor, order = "hclust",method = "square",
         tl.col="black",tl.cex = 0.8)
corrplot(cor, order = "hclust",col = col2(100),method = "color",
         tl.col="black",tl.cex = 0.8,cl.pos = "r",cl.ratio = 0.2,
         insig = "blank",addgrid.col="white")
```

附录六 问题 4 检验代码

```
gaojia=as.matrix(read.table("clipboard", header = F, sep = '\t'))
p1=cor(gaojia)
p1[!upper.tri(p1,diag=FALSE)]=0
x=as.matrix(as.vector(p1))
x[which(x!=0)]
qianbei=as.matrix(read.table("clipboard", header = F, sep = '\t'))
p2=cor(qianbei)
p2[!upper.tri(p2,diag=FALSE)]=0
y=as.matrix(as.vector(p2))
y[which(y!=0)]
wilcox.test(x,y,paired = T)
```