基于多模型的古代玻璃制品分类与预测

摘要

随着古代玻璃考古工作的日益发展，玻璃制品的鉴别与成分分析也愈加重要。总结玻璃制品的判别规律和化学成分的关联性，可以帮助考古工作者更高效地分析文物的特性。本文借助 python 编程，采用层次聚类、K 近邻算法和相关性分析等模型对玻璃制品进行分类与预测。

首先对附件数据进行预处理，根据题目要求，剔除化学成分比列累加之和不在 85%~105% 之间的数据，并对化学成分含量中空白的部分补 0。

针对问题一，我们借助 spss 软件来绘制堆积直方图，得出三个因素与风化情况的相关性大小排序为：颜色＞纹饰＞类型，并利用卡方检验验证。通过绘制箱线图，可以较为直观地得出风化前后文物样品化学成分的变化规律。在定性分析的基础上，我们借助 python 进一步计算出具体变化率。在预测风化文物未被风化前的化学成分含量时，我们通过计算风化前后化学成分含量的均值变化率来求解。

针对问题二，通过统计分析，并结合基尼系数，筛选出鉴别玻璃类型最重要的几种化学成分（按重要程度依次为氧化铅、氧化钡、氧化锶），并进一步定量确定分类阈值。在选择化学成分进行亚类划分时，需要筛选同父类玻璃中含量波动较大的化学成分，并排除风化对数据波动的影响。分别计算高钾玻璃、铅钡玻璃化学成分含量的标准差 σ 和风化前后的均值差。通过参考这两个指标，我们选择氧化铜、五氧化二磷用于区分高钾玻璃亚类；氧化钠、氧化钙、氧化铝、氧化铜、五氧化二磷用于区分铅钡玻璃亚类，并进行层次聚类分析，从而完成亚类划分。在进行合理性分析时，我们引入信息熵的相关理论进行佐证。进行灵敏性分析时，采用数据假设，对数据进行 2

针对问题三，首先基于问题 2 的分类规律，将待分类样品初步分为高钾玻璃和铅钡玻璃，接着根据有无风化将两类玻璃进一步分类为：高钾无风化玻璃（A1），高钾风化玻璃（A6、A7），铅钡无风化玻璃（A3、A4、A8），铅钡风化玻璃（A2、A5）。根据同样的分类将表单 2 中的有效数据划分为四个数据集，并建立 k 近邻模型进行纹饰的分类。对于敏感度分析，采用数据假设，对数据进行微小扰动来获得测试数据，代入原模型检验得灵敏度较好。

针对问题四，我们先按玻璃类型将数据分为两类，再各自对各化学成分进行相关性分析。采用 Spearman 相关系数作为分析系数，可以得到各个化学成分之间的相关系数，并结合各成分的化学性质进行后续分析。对于关联关系，我们选择将数据转化为热力图，从而让结果更直观地反应，便于分析。在差异性上，我们通过纵向对比相关系数来体现。

关键字： 基本统计量　层次聚类　K 近邻算法　相关性分析

一、 问题重述

1.1 问题背景

玻璃最早出现在美索不达米亚地区，而考古学表明，中国最早的玻璃很有可能是舶来品。公元前 138 年，汉武帝派遣张骞西行，与西域各国达成合作贸易关系。从此开辟了陆上丝绸之路。中国的丝绸织品运往中亚、西亚等地区，同时也从西域各国运回了各种物品，玻璃就是其中之一。从 20 世纪 30 年代开始，研究古代玻璃的考古工作日益发展，通过对玻璃的考古，可以得知中国玻璃的发展历程，甚至得以一窥当时的经济发展。因此，古玻璃考古工作越来越受到关。

现有一批从某地出土的古玻璃制品，通过对这些文物样品进行化学成分分析，以及通过 XRF、XRD、SEM-EDX 等分析测试仪器 [1]，将这些文物分为铅钡玻璃、高钾玻璃这两种类型；并在每件文物表面上随机选择 1-2 个采样点，分析采样点的化学成分，得出各个主要成分所占的比例。通过这些有效的科学手段得出的数据，可以帮助考古工作者高效地分析文物的特性，总结出玻璃制品的风化规律，以及各个化学成分的关联性等。

1.2 需要解决的问题

现经过对化学成分的分析以及其他检测手段，给出了附件信息。表单 1 给出了这些玻璃文物样品的分类信息，通过玻璃类型、表面有无风化、颜色和纹饰进行分类。表单 2 给出了各个文物表面上（1-2 个）检测点的化学成分分析数据，即相应的主要成分所占比例，空白处表示未检测到该成分，但由于检测手段精确度等问题，各成分比例累加可能非 100%，因此在实际工作中，将 85% ~100% 之间的数据视为有效数据。根据以上信息，解决以下问题：

（1）分析玻璃文物的纹饰、颜色和玻璃类型与表面风化的关系；根据玻璃类型对采样点数据进行归纳，分析文物样品表面风化化学成分含量的统计规律；并预测风化点在风化前的化学成分含量。

（2）参考附件数据，分析将文物分为高钾玻璃、铅钡玻璃的分类规律；选择合适的化学成分，依此对各个类别的文物进行亚类划分，详细解释划分方法、展示划分结果，并对分类结果的合理性和敏感性进行分析。

（3）表单 3 中的玻璃文物为未知类型，对其化学成分进行分析，鉴别其所属类型，并分析以上分类结果的敏感性。

（4）针对不同类别的玻璃文物样品，对其各主要化学成分之间的关联关系进行分析，通过分析比较，总结出不同类别之间的化学成分关联关系的差异性。

二、 问题分析

对附件数据进行预处理，根据题目要求，剔除表单 2 中的无效数据。由于题目说明表单 2 中空白处表示未检测到，因而我们考虑对空白处补 0。

针对问题一： 题目要求探究玻璃类型、纹饰、颜色与表面风化的关系。为了让数据可视化，我们选择绘制堆积直方图，分别统计各个类型、纹饰、颜色的玻璃文物在风化与未风化两种情况下的数量。为进一步说明各个因素与风化的相关性，我们选择卡方检验，计算各个因素与风化情况的 χ^2 值，综合分析出各个分类与玻璃风化情况的关系。

在分析风化前后文物样品化学成分的变化规律时，我们需要根据玻璃类型、是否风化两个指标，对数据进行归纳分类。为了更真实地展示数据分布情况、更直观地识别数据异常值、更充分地利用数据，我们采用绘制箱线图的方法，对文物样品的化学成分含量进行统计分析，从而归纳得出统计规律。

要想预测风化点风化前的化学成分含量，我们必须要知道风化前后化学成分含量的具体变化情况。可以通过求取风化前后化学成分含量的均值变化率，来进行初步的预测。

针对问题二： 题目要探究高钾玻璃、铅钡玻璃的分类规律。只按玻璃类型进行初步归类，并观察两种玻璃的化学成分分布的特点。通过计算基尼系数，筛选出在分类过程中最重要的化学成分，以此作为分类规律。

对高钾、铅钡玻璃进行亚类划分，应从二者的化学成分含量入手，选择能够体现出同父类玻璃之间差异性的化学成分。此处可分别计算各个化学成分在同父类玻璃占比的标准差，建立关于标准差的筛选指标，再考虑是否受风化影响，得出所需化学成分后，将这些化学成分用来做聚类分析，从而得出亚类划分的结果。合理性分析需要对模型整体的普适性，严谨性做出分析，包括划分标准、算法，以及最后分类的结果；灵敏性则需要考虑数据波动的情况下，该模型分类结果受影响程度。

针对问题三： 基于问题 2 的分类规律，我们可以初步区分玻璃种类。经数据初步统计知纹饰 B 所有玻璃均风化，说明是否风化和纹饰种类有关系，在后续对样品预测时将样品分为风化和无风化两大类进行预测。由于风化文物上的未风化点只是局部特征，难以说明风化文物的整体情况，我们将风化文物上的未风化点的数据剔除。数据统计分析发现存在少量异常值，故选择对异常值不敏感的分类模型 KNN 来预测纹饰。由于单种颜色的样本数量较少，数据提取到的信息可靠度无法满足要求，故不区分颜色。对于敏感度分析，我们可以进行数据假设，对原始数据进行扰动来获得测试数据，

代入原分类模型进行检验。

针对问题四： 对于不同类型玻璃文物化学成分之间的关联关系、差异性分析，都需要相关系数作为重要指标，故应先计算相关系数。首先按照玻璃类型把数据归为两类，然后各自对各个化学成分含量进行相关性分析，此处采用 Spearman 相关系数作为分析系数，可以得到各个化学成分之间的相关系数。

对于关联关系，分别绘制 Spearman 相关系数热力图，直观看出关联关系。综合热力图的直观性与数据的客观性，评价化学成分之间的关联关系。对于差异性，也需要相关系数的对比来得出结论，将高钾玻璃和铅钡玻璃的各化学成分的相关系数进行纵向对比，从而体现出差异性。

三、 模型的假设

- 采集的样本数据有一定代表性，能反应该类别的特征
- 文物表面出土后不会受到二次破坏
- 化学成分异常仅跟检测手段有关，不代表文物的性质

四、 符号说明

符号	含义
X_{ij}	第 i 个化学成分在 j 号采样点处的占比
ΣX	j 号采样点所有化学成分占比之和
σ	标准差
Δ_i	风化前后各化学成分均值差
D_i	各个样本集
ρ_s	spearman 相关系数
χ^2	观察值与理论值之间的偏离程度
$\widehat{J_j^2}(T)$	特征 j 在单棵树中的特征重要性

五、 模型的建立与解决

5.1 数据预处理

选取表单 2 中各个主要成分所占比例累加之和介于 85%~105% 的作为有效数据，将 15 和 17 号文物的数据剔除。

对于表单 2 中的空缺数据，根据题目所给信息，空缺数据表示未检测到该成分，同时我们对各个采样点的所有化学成分占比进行求和，得出 ΣX 与 100% 极为接近，因此可以忽略空缺数据。为方便后续分析，将空缺数据补全为 0。

观察表单 1 所给信息，可以发现 19、50、48、58 号文物的颜色信息缺失，因此仅在讨论其颜色与表面风化的关系时，将这四个数据剔除，保证模型合理性。

5.2 问题一：模型的建立与解决

首先分析玻璃类型、纹饰、颜色与风化的关系。我们统计各个种类玻璃风化与未风化的数量，通过 SPSS 绘制堆积直方图，可以直观看出各个种类玻璃风化的占比，从而初步直观地判断各个因素与风化的关系，图示如下：

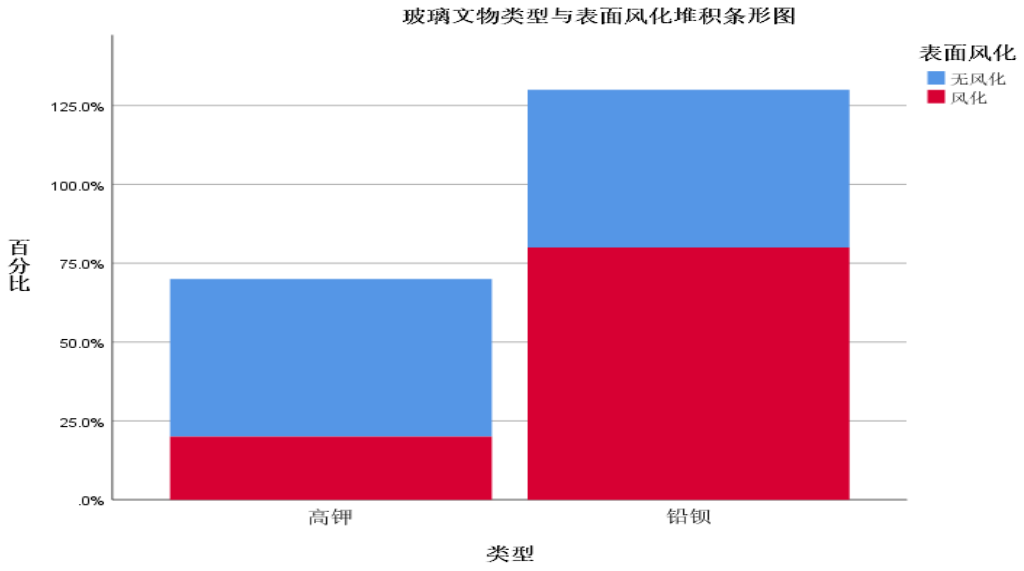

图 1　玻璃类型与表面风化堆积条形图

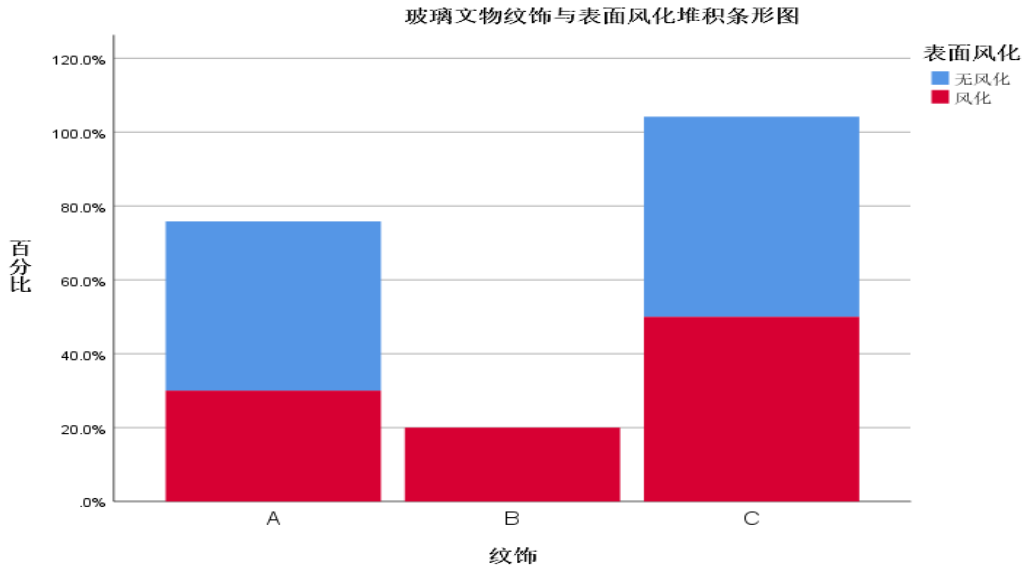

图 2　玻璃纹饰与表面风化堆积条形图

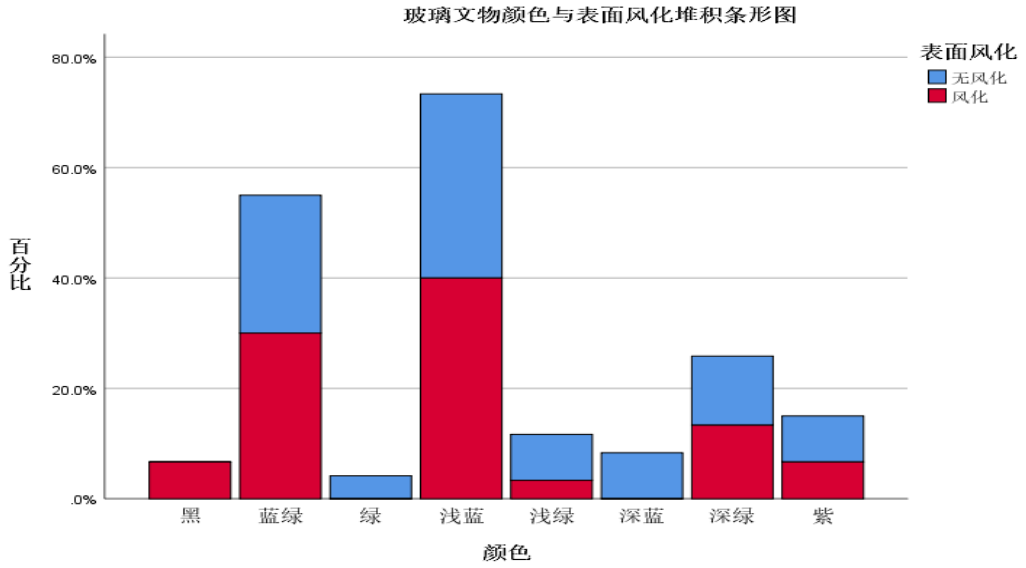

图 3　玻璃颜色与表面风化堆积条形图

从类型来看，高钾玻璃风化的占比较小，铅钡玻璃的风化占比较高；从纹饰来看，A、C 纹饰玻璃文物风化与未风化占比相当，B 纹饰全部风化；从颜色来看黑色玻璃全部风化，绿色、深蓝色玻璃没有出现风化情况。从整体来看，颜色与风化情况的相关性最强，不同类型对风化情况也有影响，但效果不明显。

为进一步定量评估三个因素与风化情况的关系，现采用卡方检验，计算各个因素与风化情况的 χ^2 值，公式为

$$\chi^2 = \Sigma \frac{(B-A)^2}{A} \tag{1}$$

其中 A={表面风化情况}，B={纹饰，类型，颜色}, χ^2 可用于表明两组定类变量的关联程度，χ^2 越大，二者偏差程度越大，即相关性越强。统计结果如表 1 所示。从表 1 结果可看出，三个因素与风化情况的相关性大小排序为：颜色＞纹饰＞类型，与堆积直方图所呈现的结果相符合，进一步验证了该评价模型的合理性。

表 1　玻璃文物三个因素与风化情况的卡方检验结果

	类型	纹饰	颜色
χ^2	5.4	5.747	6.287
校正 χ^2	4.134	5.747	6.287

针对“分析风化前后，文物样品化学成分变化规律”的问题，我们利用 Excel 软件将数据分为“高钾玻璃 -风化文物”、“高钾玻璃 -无风化文物”、“铅钡玻璃 -风化文物”、“铅钡玻璃 -无风化文物”四大类。对于风化文物上未风化的采样点（采样点为 23、25、28、29、42(1)、42（2）、44、49、50、53），我们考量它的化学成分含量特征，将其归类到“铅钡玻璃 -风化文物”组中。借助 python 编程，采用绘制箱线图的方法，对文物样品的化学成分含量进行分析。箱线图详见附录 F 图 1~4：

由附录 F 图 12、图 33 可以看出，对于铅钡玻璃，风化后与风化前相比，二氧化硅、氧化钠含量总体减小，氧化钙、氧化铅、五氧化二磷、二氧化硫含量总体增加。由附录 F 图 10、图 11 可以看出，对于高钾玻璃，风化后与风化前相比，氧化钾、氧化钙、氧化铝含量总体减小，二氧化硅含量总体增加。

从箱线图我们可以看到有异常值的出现，因而我们选取 25% 分位数、75% 分位数作为化学成分的区间边界。借助 python 编程对变化情况进一步定量评估，总结如下：

表 2　铅钡玻璃类文物风化前后化学成分含量的变化情况

化学成分	风化前含量范围	风化后含量范围	均值变化率
二氧化硅	37.36%-65.61%	18.79%-30.20%	-54.42%
氧化钠	31.94%-63.48%	1.12%-1.59%	-63.70%
氧化钙	0.64%-2.09%	1.44%-3.51%	84.60%
氧化铅	16.36%-27.67%	35.47%-48.84%	96.12%
五氧化二磷	0.17%-1.46%	2.35%-8.77%	302.78%
二氧化硫	0-3.66%	2.43%-15.26%	142.62%

表 3　高钾玻璃类文物风化前后化学成分含量的变化情况

化学成分	风化前含量范围	风化后含量范围	均值变化率
二氧化硅	61.68%-71.17%	92.65%-94.84%	38.32%
氧化钾	8.55%-12.33%	0.70%-0.94%	-92.9%
氧化钙	5.53%-8.01%	0.65%-1.04%	-86.4%
氧化铝	5.14%-7.93%	1.36-2.38%	-70.85%

其中均值变化率的计算公式为：

$$\Delta X = \frac{\bar{X}_t - \bar{X}}{\bar{X}} \tag{2}$$

其中，$\bar{X}_t$ 表示各样品风化后 A 的平均含量，$\bar{X}$ 表示各样品风化前 A 的平均含量。

通过查阅文献 [2]，我们得知：对于钾玻璃，风化后表层玻璃中的 K_2O 会大量流失，而 SiO_2 显著提高，这与我们统计得出的规律也相符。

针对“预测风化前化学成分含量”的问题，我们延续第二小问的思路，将所有文物样本分为四大类，注意风化文物上未风化采样点的划分。分别计算每种化学成分含量的均值变化率，并以该变化率预测风化前的化学成分含量。易得预测数据的计算公式为:

$$X' = (1 + \Delta X) * X_t \tag{3}$$

以 7 号文物为例，预测数据结果如下：

表 4　7 号文物风化采样点的化学成分含量及风化前的预测数据

	二氧化硅 (SiO_2)	氧化钙 (CaO)	氧化铝 (Al_2O_3)	氧化铁 (Fe_2O_3)	氧化铜 (CuO)	五氧化二磷 (P_2O_5)
风化后（%）	92.63	1.07	1.98	0.17	3.24	0.61
预测风化前（%）	67.03	7.87	6.78	1.49	5.55	2.77

借助 excel 软件，我们很方便就求得其余风化文物的预测结果，详见附录 F。

5.3 问题二：模型的建立与解决

5.3.1 分析分类规律

高钾玻璃的氧化铅含量极少（不超过 2%），而铅钡玻璃的氧化铅含量较高（严重风化时也超过 9%）。高钾玻璃的氧化铅含量范围的上界远小于铅钡玻璃的氧化铅含量范围的下界，因此可以较为严谨地得出结论：氧化铅含量是这两种玻璃最重要的分类依据。其余 13 种化学成分并无上述特性，借助决策树分类模型对其重要性进行量化。

图 4 为决策树结构。通过该图可以直观地看出如何对类型进行精准分类、回归，基尼指数 gini 作为该模型的重要参数之一，可以用来表示该化学成分样本集的稳定性，gini 值越大，代表该成分越不稳定，它在分类决策中起到的作用就更大，计算公式如下:

$$Gini\_index(D,a) = \Sigma_{m=1}^{m} \frac{|D^m|}{|D|} Gini(D^m) \tag{4}$$

$$Gini(D) = \Sigma_{k=1}^{|y|} \Sigma_{k \neq k'} p_k p_{k'} \tag{5}$$

在该决策树中。氧化钡的 gini 值最大，说明氧化钡的决策重要度最高。再从特征重要性的角度分析，特征 j 在单棵树中的特征重要性计算公式如下:

$$\widehat{J_j^2}(T) = \Sigma_{t=1}^{L-1} \widehat{I_t^2} \mathbb{I}(v_t = j) \tag{6}$$

其中，L 为树的叶子节点数量，v_t 是和节点 t 相关联的特征，$\widehat{I_t^2}$ 是节点 t 分裂之后平方损失的值。从公式可以看出来，$\widehat{J_j^2}(T)$ 越大，说明该节点降低损失的能力就越大，在决策中的重要性就更大。计算结果如图 5 所示，氧化钡的特征重要性高达 75.1%，进一步证明了氧化钡的决策重要性之高；氧化锶的特征重要性为 17.7%，与氧化钡的重要性相比较小；为次要的分类依据，相比之下，五氧化二磷的特征重要性可以忽略不计，与其他 10 种化学成分一起划分为不影响这两种玻璃分类规律的因素。

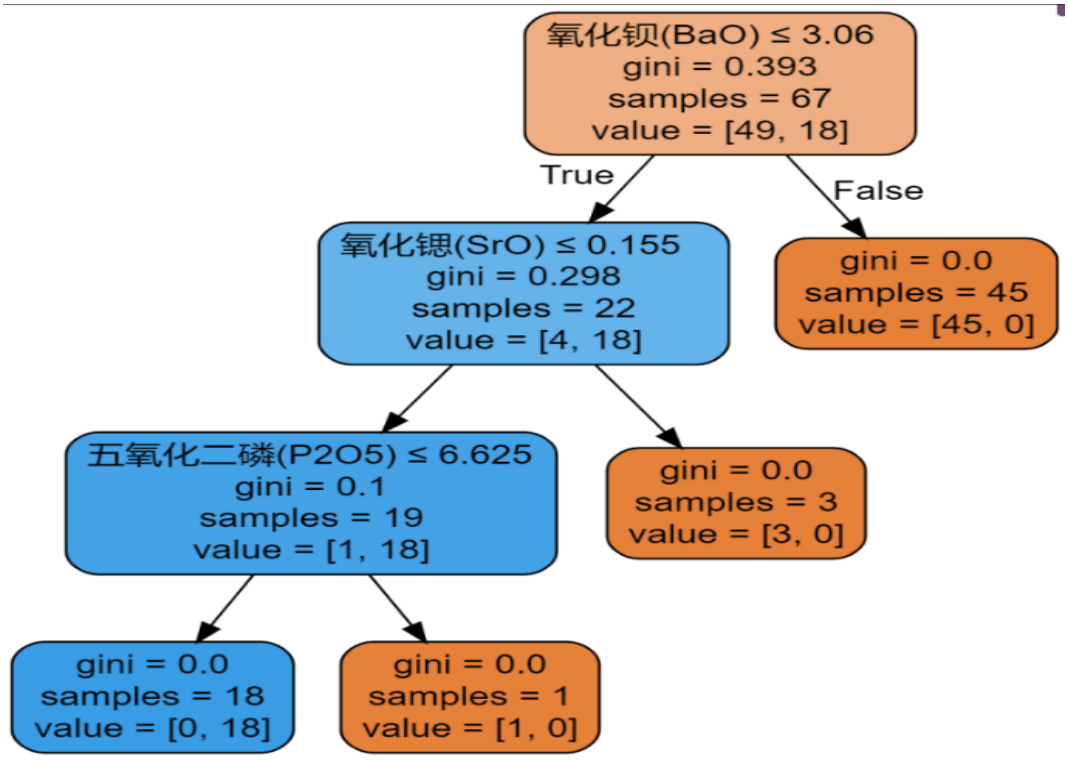

图 4　决策树结构

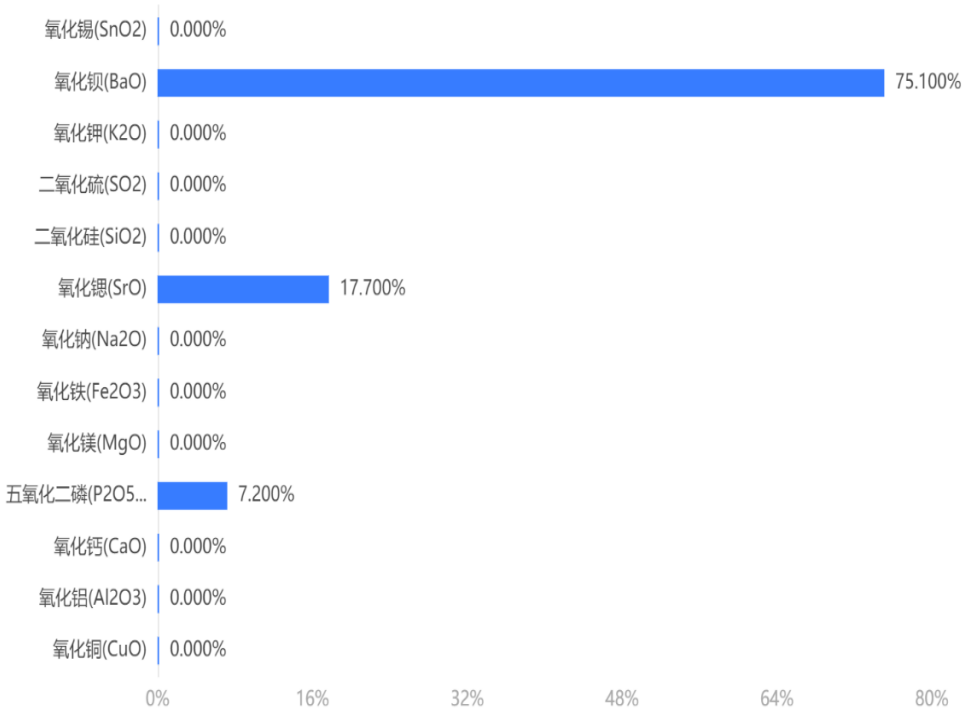

图 5　其余 13 种化学成分对分类规律的重要性 -条形图

在该分类规律下，氧化钡的特征重要性达到 75.1%，这也是一个重要的分类依据；氧化锶的特征重要性为 17.7%，与氧化钡的重要性相比较小，为次要的分类依据；相比

之下，五氧化二磷的特征重要性可以忽略不计，与其他 10 种化学成分一起划分为不影响这两种玻璃分类规律的因素。

因而我们选取氧化铅、氧化钡、氧化锶作为分类指标。从题目中我们得知，铅钡玻璃在烧制过程中加入铅矿石作为助熔剂，其氧化铅（PbO）、氧化钡（BaO）的含量较高，适于作为分类指标，这进一步检验我们的结论。

接下来对重要指标氧化铅进一步定量分析。氧化铅在铅钡玻璃中含量较高，在高钾玻璃中含量较低。利用边界值来确定定量指标，具体如下：

$$I = \frac{X_{min} - X_{kmax}}{2} = \frac{9.3\% - 1.62\%}{2} = 3.84\% \tag{7}$$

其中 I 表示氧化铅指标，X_{min} 表示样本铅钡玻璃中氧化铅占比最小值，X_{kmax} 表示样本高钾玻璃中氧化铅占比最大值。

当样品中的氧化铅含量高于 3.84% 时，可将其分类为铅钡玻璃；低于 3.84% 时，可将其分类为高钾玻璃。当氧化铅含量与 3.84% 相当时，再结合氧化钡、氧化锶的含量进行判断。

5.3.2 进行亚类划分

5.3.2.1 划分方法

在考虑所选化学成分时，不仅要充分体现同父类玻璃之间的差异性，还要排除风化的影响。在问题 2 的第一小问中，我们选择了氧化铅、氧化钡作为区分父类玻璃的依据，因此在分析亚类时，应排除这两个化学成分。此处引入标准差 σ 作为指标之一，公式如下:

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n}((S_i - \bar{S})^2}{n}} \tag{8}$$

标准差越大，数据波动性越强，对于分类的决策作用就越强。此处设置标准差 σ ＞ 1 的筛选指标。为排除风化的影响，我们计算化学成分含量在风化前后的均值差，均值差越小，表明该化学成分在风化前后的变化越小。计算结果如下表所示（已排除 σ ＜ 1 的化学成分）

表 5　高钾玻璃各成分含量风化前后数据

	二氧化硅 (SiO_2)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铝 (Al_2O_3)	氧化铁 (Fe_2O_3)	氧化铜 (CuO)	五氧化二磷 (P_2O_5)
标准差	14.47	4.87	3.19	3.08	1.58	1.43	1.30
风化均值 (%)	93.96	0.82	0.87	1.93	0.27	1.56	0.34
未风化均值 (%)	67.98	10.18	6.40	6.62	2.32	2.68	1.53
均值差 Δ(%)	25.98	-9.36	-5.53	-4.69	-2.05	-1.11	-1.19

表 6　铅钡玻璃各成分含量风化前后数据

	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钙 (CaO)	氧化铝 (Al_2O_3)	氧化铜 (CuO)	五氧化二磷 (P_2O_5)	二氧化硫 (SO_2)
标准差	18.65	2.10	1.58	3.00	2.51	3.97	7.02
风化均值 (%)	24.91	1.41	2.80	2.97	2.47	5.72	8.88
未风化均值 (%)	54.66	3.87	1.52	4.46	1.57	1.42	3.66
均值差 (%)	-29.75	-2.47	1.28	-1.49	0.90	4.30	5.22

对于高钾玻璃化学成分的均值差 Δ，选择 $|\Delta| < 2$ 的化学成分；对于铅钡玻璃，考虑到它的化学成分标准差普遍较大，因此提高阈值，选择 $|\Delta| < 6$ 的化学成分。然后得出以下结论：氧化铜、五氧化二磷用于区分高钾玻璃亚类；氧化钠、氧化钙、氧化铝、氧化铜、五氧化二磷用于区分铅钡玻璃亚类。将以上主要化学成分作为特征，来进行层次聚类分析。

5.3.2.2 层次聚类分析

过上述分析，可以给定样本集 D_1={氧化铜，五氧化二磷}，D_2={氧化钠，氧化钙，氧化铝，氧化铜，五氧化二磷}，用层次聚类算法对聚类所得簇进行划分，程序输出的聚类谱系图 6、图 7 如下:

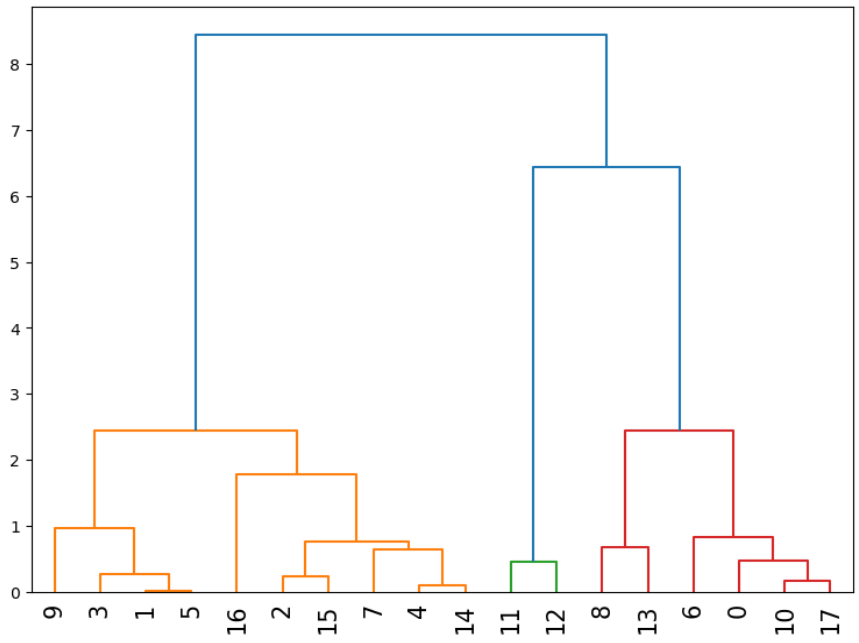

图 6　高钾玻璃聚类谱系图

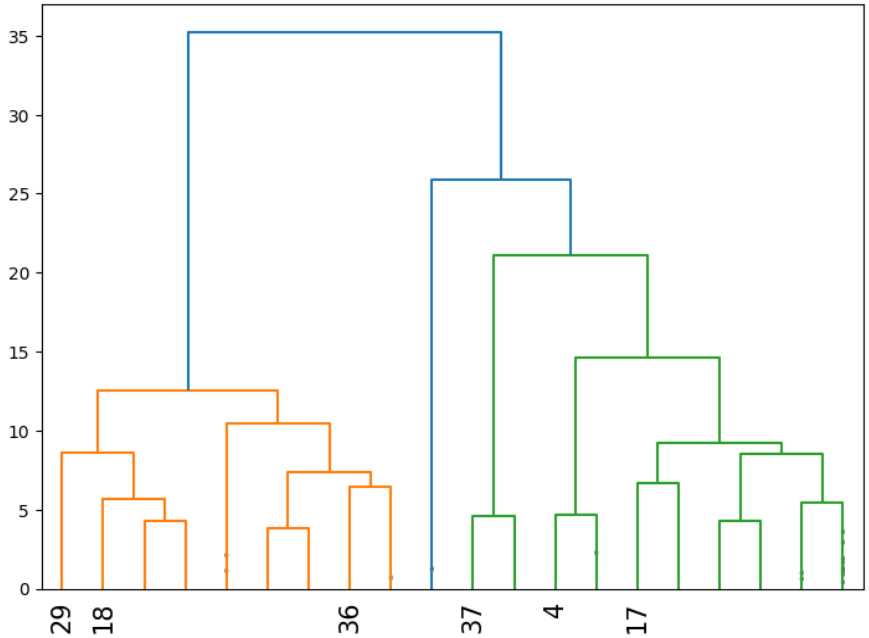

图 7　铅钡玻璃聚类谱系图

将所得簇划分为 3 类，分别得到以下簇 C_1={C_{1A},C_{1B},C_{1C}}，C_2={C_{2A},C_{2B},C_{2C}}。参考国外学者 J. W. Lankton 和 L. Dussubieu 对硅酸盐亚类的划分原则 [4]，我们按照关键化学成分含量高低进行亚类划分及命名，命名规则如表 7：（注：NCACP 代表化学元素钠、钙、铝、铜、磷）

表 7　亚类命名

	全称
高钾 A	高铜磷高钾玻璃
高钾 B	中铜磷高钾玻璃
高钾 C	低铜磷高钾玻璃
铅钡 A	高 NCACP 铅钡玻璃
铅钡 B	中 NCACP 铅钡玻璃
铅钡 C	低 NCACP 铅钡玻璃

具体的分类结果见表 8。

表 8　分类结果

采样点	分类结果	采样点	分类结果
7	高钾 A	42 未风化点 1	铅钡 C
9	高钾 B	42 未风化点 2	铅钡 C
10	高钾 B	43 部位 1	铅钡 C
12	高钾 B	43 部位 2	铅钡 B
22	高钾 B	44 未风化点	铅钡 A
27	高钾 B	49	铅钡 B
1	高钾 A	49 未风化点	铅钡 B
03 部位 1	高钾 B	50	铅钡 B
03 部位 2	高钾 A	50 未风化点	铅钡 B
4	高钾 B	51 部位 1	铅钡 B
5	高钾 A	51 部位 2	铅钡 B
06 部位 1	高钾 C	52	铅钡 B
06 部位 2	高钾 C	53 未风化点	铅钡 C
13	高钾 A	54	铅钡 B
14	高钾 B	54 严重风化点	铅钡 B
16	高钾 B	56	铅钡 C
18	高钾 B	57	铅钡 C
21	高钾 A	19	铅钡 B
		40	铅钡 C
2	铅钡 B	48	铅钡 A
8	铅钡 C	58	铅钡 B
08 严重风化点	铅钡 B	20	铅钡 B
11	铅钡 B	24	铅钡 C
23 未风化点	铅钡 C	30 部位 1	铅钡 C
25 未风化点	铅钡 C	30 部位 2	铅钡 C
26	铅钡 C	31	铅钡 C
26 严重风化点	铅钡 B	32	铅钡 C
28 未风化点	铅钡 C	33	铅钡 C
29 未风化点	铅钡 A	35	铅钡 C
34	铅钡 C	37	铅钡 C
36	铅钡 C	45	铅钡 C
38	铅钡 C	46	铅钡 C
39	铅钡 C	47	铅钡 C
41	铅钡 B	55	铅钡 C

5.3.2.3 合理性分析

整个模型建立到求解的过程中，划分标准、划分算法是两个最重要的分析目标。为证明标准差的作用，此处引入信息熵的相关理论进行佐证。信息熵来源于香农的信息论，信息熵 H(x) 是用来度量信息量多少的重要指标，同时也可以用来衡量信息的混乱程度。对于连续变量 x 的信息熵计算公式 [3] 如下:

$$H(x) = -\int p(x) ln p(x) dx \tag{9}$$

其中 p(x) 表示概率密度函数。信息熵越大，意味着信息混乱程度越高，同时说明信息量越大，对于因变量的影响也就更大。而标准差刚好能用于衡量变量混乱程度，故标准差越大，变量的信息熵越大，对于因变量的影响也就更大。因此，该模型具有合理性。

5.3.2.4 灵敏性分析

灵敏性分析是指研究与分析一个系统（或模型）的状态或输出变化对系统参数或周围条件变化的敏感程度。对本模型的分析，需要考虑每一个有效因子的敏感程度，分别考量各个因子的数据波动对于分类结果的影响。此处决定采用数据假设的方法进行分析。选择所有高钾玻璃、所有铅钡玻璃作为两个样本集 D_3 、D_4 ，分别从中选出一些样本，对各自主要化学成分的数据进行 2%~5% 的扰动，分别得到新的样本集 D_3' 、D_4' 作为他们的测试样本。将 D_3' 、D_4' 作为测试数据输入到原分类模型中。此处展示高钾玻璃的测试数据，如表 9 所示。

表 9　扰动数据后的预测结果

编号		氧化铜	五氧化二磷	模型输出结果
10	实际	3.27	0.94	A
	测试	3.5	0.9	A
17	实际	3.28	1.1	A
	测试	3.2	1.18	A
7	实际	0.78	0.66	B
	测试	0.55	0.68	B
15	实际	1.07	0	B
	测试	0.93	0.01	B
11	实际	2.51	4.18	C
	测试	2.45	4.32	C

测试数据的预测结果与实际数据的分类结果完全符合，因此可以认为，样本数据在受到一定程度的扰动后，仍然适用于该模型，证明该分类模型具有较好的灵敏性。

5.4 问题三：模型的建立与解决

首先利用问题 2 的分类规律将待分类样品分为高钾玻璃和铅钡玻璃，接着根据有无风化将两类玻璃进一步分类为：高钾无风化玻璃（A1），高钾风化玻璃（A6、A7），铅钡无风化玻璃（A3、A4、A8）, 铅钡风化玻璃（A2、A5）, 根据同样的方法将表 2 中有效数据划分为四个数据集。

设特征空间 n 维实数向量空间 R^n,$x_i, x_j \in X, x_i = (x_i^1, x_i^2, ..., x_i^n)^T$,$x_j = (x_j^1, x_j^2, ..., x_j^n)^T$，按照以下步骤建立 k 近邻模型：

（1）根据以下公式计算已知类别数据集中的点与当前点之间的距离:

$$L_p(x_i, x_j) = (\Sigma_{l=1}^{n}|x_i^l - x_j^l|^p)^{1/p}, (p \geqq 1) \tag{10}$$

（2）按照距离递增次序排序；

（3）选取与当前点距离最小的 k 个点；

（4）确定前 k 个点所在类别的出现频率；返回前 k 个点所出现频率最高的类别作为当前点的预测分类 y:

$$y = arg \max_{c_j} \Sigma_{x_i \in N_k(x)} I(y_i = c_j), i = 1, 2, ..., N; j = 1, 2, K \tag{11}$$

式中，I 为指示函数，即为 $y_i = c_j$ 时 I 为 1，否则 I 为 0.

设置不同的 k 值结合模型的加权平均评价指标取到最优的 k 解，分别将四组数据集的化学成分数据求取距离代入模型，得到预测结果为：

{A1: 高钾纹饰 A}{A2: 铅钡纹饰 A}{A3: 铅钡纹饰 A}{A4: 铅钡纹饰 A}{A5: 铅钡纹饰 A}{A6: 高钾纹饰 B}{A7: 高钾纹饰 B}{A8: 铅钡纹饰 A}

对于敏感度分析，我们对 A1 进行 2%-5% 的扰动，分别得到新的样本集 A12、A13 作为测试样本。将 A12、A13 作为测试数据输入到原分类模型中。测试数据及结果如表 10 所示：

表 10　扰动数据后的预测结果

样本	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO	BaO	P_2O_5	SrO	SuO_2	SO_2	检测结果
A1	78.45	0	0	6.08	1.86	7.23	2.15	2.11	0	0	1.06	0.03	0	0.51	A
A12	76.6	0	0	7	2	6	2	2	0	0	0.8	0.03	0	0.6	A
A13	79.7	0	0	5.8	1.69	7.12	2.03	2.10	0	0	1.20	0.02	0	0.49	A

测试数据的预测结果与实际数据的分类结果完全符合，可以认为，样本数据在受到一定程度的扰动后，仍然适用于该模型，证明该分类模型具有较好的敏感度。

5.5 问题四：模型的建立与解决

5.5.1 计算 spearman 相关系数

针对高钾玻璃和铅钡玻璃，进行相关性分析。

计算 spearman 相关系数公式为：

$$\rho_s = \frac{\Sigma_i(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i(x_i - \bar{x})^2 \Sigma_i(y_i - \bar{y})^2}} \tag{12}$$

计算结果制作成表格形式写入 excel 表，并绘制成热力图如下：

	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
二氧化硅	1	-0.4749	-0.7808	-0.7517	-0.5455	-0.8308	-0.7827	-0.4295	-0.4425	-0.3457	-0.5297	-0.5236	0.07012	-0.2843
氧化钠	-0.4749	1	0.60646	0.64048	-0.2233	0.31293	0.21773	-0.1128	0.26559	-0.2355	-0.2209	-0.1097	-0.1079	-0.198
氧化钾	-0.7808	0.60646	1	0.69012	0.25596	0.50569	0.45732	0.179	0.24113	-0.0099	0.15623	0.40402	0.07027	0.30721
氧化钙	-0.7517	0.64048	0.69012	1	0.08729	0.52142	0.53667	0.34401	0.25602	-0.073	0.02273	-0.0517	-0.3742	0.38937
氧化镁	-0.5455	-0.2233	0.25596	0.08729	1	0.73475	0.56949	0.20403	0.29531	0.52102	0.64311	0.66608	0.21427	0.42876
氧化铝	-0.8308	0.31293	0.50569	0.52142	0.73475	1	0.74755	0.24574	0.60216	0.47326	0.51833	0.51135	-0.1636	0.24621
氧化铁	-0.7827	0.21773	0.45732	0.53667	0.56949	0.74755	1	0.64256	0.32296	0.53872	0.51653	0.40892	-0.3742	0.31309
氧化铜	-0.4295	-0.1128	0.179	0.34401	0.20403	0.24574	0.64256	1	0.04698	0.48272	0.46591	0.14082	-0.3976	0.35441
氧化铅	-0.4425	0.26559	0.24113	0.25602	0.29531	0.60216	0.32296	0.04698	1	0.68174	0.12096	0.36771	-0.1861	-0.3415
氧化钡	-0.3457	-0.2355	-0.0099	-0.073	0.52102	0.47326	0.53872	0.48272	0.68174	1	0.47067	0.56455	-0.1284	-0.2355
五氧化二磷	-0.5297	-0.2209	0.15623	0.02273	0.64311	0.51833	0.51653	0.46591	0.12096	0.47067	1	0.49993	0.30402	0.24634
氧化锶	-0.5236	-0.1097	0.40402	-0.0517	0.66608	0.51135	0.40892	0.14082	0.36771	0.56455	0.49993	1	0.30625	-0.0227
氧化锡	0.07012	-0.1079	0.07027	-0.3742	0.21427	-0.1636	-0.3742	-0.3976	-0.1861	-0.1284	0.30402	0.30625	1	-0.1079
二氧化硫	-0.2843	-0.198	0.30721	0.38937	0.42876	0.24621	0.31309	0.35441	-0.3415	-0.2355	0.24634	-0.0227	-0.1079	1

图 8　高钾玻璃化学成分之间 spearman 相关系数热力图

	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
二氧化硅	1	0.40481	0.30816	-0.4898	0.12306	0.42465	0.04381	-0.4468	-0.7573	-0.2796	-0.5378	-0.5524	0.09828	-0.3134
氧化钠	0.40481	1	0.06673	-0.3669	0.02341	0.14513	-0.2044	-0.1024	-0.3569	0.08652	-0.5697	-0.1489	-0.0927	-0.2089
氧化钾	0.30816	0.06673	1	-0.0348	0.36193	0.51242	0.14923	-0.1979	-0.3359	-0.0005	-0.1398	-0.1669	0.14001	-0.0224
氧化钙	-0.4898	-0.3669	-0.0348	1	0.32505	0.12255	0.41127	-0.0371	0.351	-0.1569	0.50189	0.30829	0.28397	0.07821
氧化镁	0.12306	0.02341	0.36193	0.32505	1	0.67356	0.25758	-0.2656	-0.0759	-0.435	0.1527	0.07448	0.25914	-0.3678
氧化铝	0.42465	0.14513	0.51242	0.12255	0.67356	1	0.38704	-0.3249	-0.3886	-0.369	-0.0193	-0.1966	0.33339	-0.3849
氧化铁	0.04381	-0.2044	0.14923	0.41127	0.25758	0.38704	1	-0.3896	0.08115	-0.4427	0.2248	-0.1161	0.36336	-0.3368
氧化铜	-0.4468	-0.1024	-0.1979	-0.0371	-0.2656	-0.3249	-0.3896	1	0.09356	0.49468	0.22054	0.1822	-0.356	0.44854
氧化铅	-0.7573	-0.3569	-0.3359	0.351	-0.0759	-0.3886	0.08115	0.09356	1	-0.0996	0.36135	0.36215	-0.092	-0.1079
氧化钡	-0.2796	0.08652	-0.0005	-0.1569	-0.435	-0.369	-0.4427	0.49468	-0.0996	1	-0.1689	0.18168	-0.0201	0.46657
五氧化二磷	-0.5378	-0.5697	-0.1398	0.50189	0.1527	-0.0193	0.2248	0.22054	0.36135	-0.1689	1	0.25478	-0.0328	0.2
氧化锶	-0.5524	-0.1489	-0.1669	0.30829	0.07448	-0.1966	-0.1161	0.1822	0.36215	0.18168	0.25478	1	0.01393	0.22677
氧化锡	0.09828	-0.0927	0.14001	0.28397	0.25914	0.33339	0.36336	-0.356	-0.092	-0.0201	-0.0328	0.01393	1	-0.1132
二氧化硫	-0.3134	-0.2089	-0.0224	0.07821	-0.3678	-0.3849	-0.3368	0.44854	-0.1079	0.46657	0.2	0.22677	-0.1132	1

图 9　铅钡玻璃化学成分之间 spearman 相关系数热力图

5.5.2 根据热力图分析化学成分之间的关联关系并比较不同关联关系的差异性

高钾玻璃化学成分之间的关联关系:

（1）二氧化硅和氧化钾、氧化钙、氧化铝、氧化铁、氧化锡有较好的负相关性。根据第一问得出的规律，高钾玻璃风化程度越大，二氧化硅含量越少，相反，氧化钙、氧化铝、氧化铁、氧化锡的含量越来越大，因此，具有较好的负相关性。

（2）氧化钠、氧化钾、氧化钙三者有较好的正相关性。根据氧化钠、氧化钾、氧化钙的化学性质，三者都会与水发生化学反应，生产氢氧化物，因此三者具有较好的正相关性。

（3）氧化镁、氧化铝、氧化铁三者有较好的正相关性。氧化铝会与氧化钠潮解后生成的氢氧化钠发生化学反应，而氧化镁与氧化铁都会潮解，因此，三者都会因水而减少，故具有较好的正相关性。

铅钡玻璃化学成分之间的关联关系：

（1）二氧化硅和氧化铅有较好的负相关性。二氧化硅会与氧化钙发生缓慢的化学反应 $SiO_2 + CaO = CaSiO_3$，而氧化铅本身化学性质稳定，不会与空气中、玻璃中的其他成分发生反应，因此二者具有较好的负相关性。

（2）与高钾玻璃相同，铅钡玻璃中，氧化镁和氧化铝的相关性也是较好的正相关性。

不同关联关系的差异性：

（1）铅钡玻璃中氧化铝和氧化铁的正相关性一般。两种玻璃的氧化铝与氧化铁相关性存在差异。通过对这些化学成分的性质的查询，发现 BaO 易与水作用生成 $Ba(OH)_2$，进而继续吸收空气中的水与二氧化碳。因此，虽然铅钡玻璃的氧化铝与氧化铁仍有一定的正相关性，但是受到氧化钡的影响，其正相关性是弱于高钾玻璃的。

（2）铅钡玻璃中二氧化硅和氧化铁、氧化锡无相关。在高钾玻璃中，由于氧化钾会与水反应生产氢氧化钾，同时氧化锡能溶于氢氧化钾，因此随着风化程度加深，二氧化硅与氢氧化钾呈现较好的正相关性；而在铅钡玻璃中，几乎不含氧化钾，因此，不存在上述关联关系。

（3）铅钡玻璃中二氧化硅和氧化纳、氧化钾、氧化铝的正相关性中等，高钾玻璃中二氧化硅与氧化钾氧化铝呈负相关性。

六、 模型的评价与推广

6.1 模型的优点

1. 本文采用绘制堆积直方图、箱线图、热力图、决策树结构图等方法，使数据可视化，能较为直观地反映数据的关系和规律。

2. 本模型对初始数据的接受性强，预测准确且具有很好的普适性。
3. 问题三中采用 k 近邻算法来区分文物纹饰，只需保存训练样本和标记，就可以用来预测新的样品，具有较高的推广价值。

6.2 模型的缺点

1. 在计算风化前后化学成分的均值变化率时，只考虑了不同玻璃类型会有不同的均值变化率，未考虑纹饰、颜色对均值变化率的影响，模型的兼容性还具有提升空间。
2. 由于特定种类的样本数量较少（例如铅钡绿色只有 1 个样本），在问题三区分未知类别文物时未对颜色进行区分。

6.3 模型的改进

1. 进一步挖掘数据中纹饰、颜色与风化前后化学成分含量的关系，更准确地预测不同风化样品未风化前化学成分的含量。
2. 查阅资料增加初始样本量，同时使用更先进的算法和统计知识处理数据。

6.4 模型的推广

本文建立的分类和鉴别模型适用范围广泛，对研究一个因变量与多个自变量的关系有一定参考价值。通过改进，该模型能更好运用于考古工作中，具有很强的现实意义。

参考文献

[1] 王婕. 一件战国时期八棱柱状铅钡玻璃器的风化研究 [J]. 玻璃与搪瓷,2014,42(2)

[2] 史美光. 何欧里. 周福征. 一批中国汉墓出土钾玻璃的研究 [J]. 硅酸盐学报,1986,14(3)

[3] 赵秀菊. 风险的两种度量方法——信息熵与方差 [J]. 襄樊学院学报,2010,2:12-15

[4] 赵凤燕, 陈斌, 柴怡, 董俊卿, 李青会. 西安出土若干玻璃器的 pXRF 分析及相关问题探讨 [J]. 考古与文物,2015(04):111-119.

附录 A 支撑材料声明

对附件.excel 数据预处理后得到的新表：预处理数据.xlsx

程序运行使用的预处理表格：高钾.xlsx

程序运行使用的预处理表格：铅钡.xlsx

问题一代码：cum.py

问题二代码：cum2.py

问题三代码：cum3.py

问题四代码：cum4.py

附录 B 数据预处理及第一问 –python 源程序

```
import matplotlib.pyplot as plt
import pandas as pd

pd.set_option("display.max_columns", 15) # 可显示列1000

'''将字符串数据转化为数字'''
def data_preprocessing(df):
    wenshi = []
    leixing = []
    result = []
    color = []
    # 列数据为字符串转为为数字标识,
    for i in df['纹饰']:
        if i == "A":
            wenshi.append(1)
        elif i == "B":
            wenshi.append(2)
        elif i == "C":
            wenshi.append(3)
    df['纹饰'] = wenshi
    for i in df['类型']:
        if i == "高钾":
            leixing.append(1)
```

```
        elif i == "铅钡":
            leixing.append(2)
    df['类型'] = leixing
    for i in df['表面风化']:
        if i == "无风化":
            result.append(1)
        elif i == "风化":
            result.append(2)
    df['表面风化'] = result
    for i in df['颜色']:
        if i == "黑":
            color.append(1)
        elif i == "蓝绿":
            color.append(2)
        elif i == "绿":
            color.append(3)
        elif i == "浅蓝":
            color.append(4)
        elif i == "浅绿":
            color.append(5)
        elif i == "深蓝":
            color.append(6)
        elif i == "深绿":
            color.append(7)
        elif i == "紫":
            color.append(8)
        elif i == 0:
            color.append(9)
    df['颜色'] = color
    # print(df)
    return df

'''将预处理的数据按需求组成并输出为格式便于统计分析等excel'''
if __name__ == '__main__':
    # 数据预处理
    df = pd.read_excel(r'高钾./.xlsx', engine='openpyxl')
    df = df.iloc[:, 6:]
```

```
df = df.fillna(0)
df.to_excel('高钾./1.xlsx') # 路径和文件名

df = pd.read_excel(r'铅钡./.xlsx', engine='openpyxl')
df = df.iloc[:, 6:]
df = df.fillna(0)
df.to_excel('铅钡./1.xlsx') # 路径和文件名

df1 = pd.read_excel(r'高钾./.xlsx', engine='openpyxl')
df2 = pd.read_excel(r'铅钡./.xlsx', engine='openpyxl')
df = pd.concat([df1, df2], ignore_index=True)
df = df.iloc[:, 2:]
# 列数据为字符串转为为数字标识,
leixing = []
for i in df['类型']:
    if i == "高钾":
        leixing.append(1)
    elif i == "铅钡":
        leixing.append(2)
df['类型'] = leixing
df = df.iloc[:, [0, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]]
df = df.fillna(0)
# print(df)
df.to_excel('表./2.xlsx') # 路径和文件名

# 对表数据的处理3
df = pd.read_excel(r'附件./.xlsx', sheet_name=2, engine='openpyxl')
df = df.iloc[:, 1:]
df.columns = ['表面风化', '二氧化硅', '氧化纳', '氧化钾', '氧化钙', '氧化
    镁', '氧化铝', '氧化铁', '氧化铜', '氧化铅', '氧化钡', '五氧化二磷', '氧化
    锶', '氧化锡',
              '二氧化硫']
df = df.fillna(0)
result = []
for i in df['表面风化']:
    if i == "无风化":
        result.append(1)
    elif i == "风化":
```

```
            result.append(2)
    df['表面风化'] = result
    print(df)
    df.to_excel('表./3.xlsx') # 路径和文件名
    '''数据预处理生成表2_2.xlsx'''
    df1 = pd.read_excel(r'高钾./.xlsx', engine='openpyxl')
    df2 = pd.read_excel(r'铅钡./.xlsx', engine='openpyxl')
    df = pd.concat([df1, df2], ignore_index=True)
    df = df.iloc[:, :]
    df = df.fillna(0)
    # pd.set_option('display.max_rows',None)
    print(df)
    df = data_preprocessing(df)
    df.to_excel('表./2_2.xlsx') # 路径和文件名

if __name__ == '__main__':
    df1 = pd.read_excel(r'附件./.xlsx', sheet_name=0, engine='openpyxl')
    # 删除空缺数据所在行
    df = df1.dropna()
    pd.set_option('mode.chained_assignment', None)
    # print(df)
    df = data_preprocessing(df)
    # print(df)
    df.to_excel('./data1.xls')

if __name__ == '__main__':
    plt.rcParams['font.sans-serif'] = ['SimHei'] # 显示中文
    df = pd.read_excel(r'附件./.xlsx', sheet_name=1, engine='openpyxl')
    # print(df)
    df['Row_sum'] = df.iloc[:,1:].apply(lambda x: x.sum(), axis=1) # 按行求和,
        添加为新列
    # print
    # 筛选出成分比例累加和介于之间的数据，即有效数据 85%~105%
    df = df[((df['Row_sum'] >= 85) & (df['Row_sum'] <= 105))]
    # print(df)
    # plt.figure()
    # p = df.boxplot()
    # plt.show()
```

```
    # print(df.describe())
    df['文物编号'] = [x[:2] for x in df['文物采样点']]
    print(df)
    print(df1.iloc[:,[0,-1]])

'''绘制箱线图，调用
describ()得到各成分的均值，分位数，分位数等。25%75%
'''
if __name__ == '__main__':
    plt.rcParams['font.sans-serif'] = ['SimHei'] # 显示中文

    df = pd.read_excel(r'高钾风化./.xlsx', engine='openpyxl')
    df = df.iloc[:, 6:]
    print(df.describe())
    plt.figure()
    p = df.boxplot()
    plt.title('高钾玻璃风化文物化学成分分析-')
    # plt.savefig高钾风化("./.png")
    plt.show()

    df = pd.read_excel(r'高钾无风化./.xlsx', engine='openpyxl')
    df = df.iloc[:, 6:]
    print(df.describe())
    plt.figure()
    p = df.boxplot()
    plt.title('高钾玻璃无风化文物化学成分分析-')
    # plt.savefig高钾无风化("./.png")
    plt.show()

    df = pd.read_excel(r'铅钡风化./.xlsx', engine='openpyxl')
    df = df.iloc[:, 6:]
    print(df.describe())
    plt.figure()
    p = df.boxplot()
    plt.title('铅钡玻璃风化文物化学成分分析-')
    # savefig铅钡风化("./.png")
    plt.show()
```

```
    df = pd.read_excel(r'铅钡无风化./.xlsx', engine='openpyxl')
    df = df.iloc[:, 6:]
    print(df.describe())
    plt.figure()
    p = df.boxplot()
    plt.title('铅钡玻璃无风化文物化学成分分析-')
    # savefig铅钡无风化("./.png")
    plt.show()

df1 = pd.read_excel(r'高钾风化./.xlsx', engine='openpyxl')
df2 = pd.read_excel(r'高钾无风化./.xlsx', engine='openpyxl')
df = pd.concat([df1,df2], ignore_index=True)
df = df.iloc[:, 6:]
print(df.describe())
# plt.figure()
# p = df.boxplot()
# plt.title高钾玻璃文物化学成分分析('')
# plt.show()

df3 = pd.read_excel(r'铅钡风化./.xlsx', engine='openpyxl')
df4 = pd.read_excel(r'铅钡无风化./.xlsx', engine='openpyxl')
df = pd.concat([df3,df4], ignore_index=True)
df = df.iloc[:, 6:]
print(df.describe())
# plt.figure()
# p = df.boxplot()
# plt.title铅钡玻璃文物化学成分分析('')
# plt.show()
```

附录 C 第二问 –python 源程序

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
'''使用层次聚类，对玻璃进行亚分类与评价。
```

```
'''
if __name__ == '__main__':
    pd.set_option("display.max_columns", 15) # 可显示列1000
    df = pd.read_excel(r'高钾./1.xlsx', engine='openpyxl')

    from scipy.cluster.hierarchy import linkage

    df = df.iloc[:, [8, 11]]
    # print(df)

    Z = linkage(df, method='ward', metric='euclidean')
    # print(Z.shape)
    # print(Z[: 5])

    from scipy.cluster.hierarchy import dendrogram

    plt.figure(figsize=(10, 8))
    dendrogram(Z, truncate_mode='lastp', p=20, show_leaf_counts=False,
        leaf_rotation=90, leaf_font_size=15,
              show_contracted=True)
    plt.show()

    from scipy.cluster.hierarchy import fcluster
    # 根据聚类数目返回聚类结果
    k = 3
    labels_2 = fcluster(Z, t=k, criterion='maxclust')
    print(labels_2[: 100])

    #层次聚类法得到分三个亚类后用等价的算法来进行灵敏性分析kmeans
    from sklearn.cluster import KMeans
    kmeans = KMeans(n_clusters=3) # 创建一个K均值聚类对象-
    kmeans.fit(df.values) # 拟合算法
    # print训练得分: ('', kmeans.score(df))
    # print(kmeans.fit_predict(df))
    data = np.array([[3.5, 1.17], [0.55, 0.68], [3.2, 1.18], [2.45, 4.32],
        [0.93, 0.05]]) #通过随机选取主要化学成分的部分数据微小波动进行验
```

```
        证
    df = pd.DataFrame(data)
    cluster_assignment = kmeans.predict(df) # 获取聚类分配
    print(cluster_assignment)

    #使用层次聚类算法对铅钡玻璃进行亚分类
    df = pd.read_excel(r'铅钡./1.xlsx', engine='openpyxl')
    df = df.iloc[:, [2, 4, 6, 8, 11]]
    Z = linkage(df, method='ward', metric='euclidean')
    plt.figure(figsize=(10, 8))
    dendrogram(Z, truncate_mode='lastp', p=20, show_leaf_counts=False,
        leaf_rotation=90, leaf_font_size=15,
              show_contracted=True)
    plt.show()
    k = 3
    labels_2 = fcluster(Z, t=k, criterion='maxclust')
    print(labels_2[: 100])
```

附录 D 第三问 –python 源程序

```
import warnings
import pandas as pd
from sklearn import neighbors
warnings.filterwarnings("ignore")

if __name__ == '__main__':
    pd.set_option("display.max_columns", 15) # 可显示列1000
    #导入训练集
    training_set = pd.read_excel(r'表./2_2.xlsx', engine='openpyxl')
    #导入测试集
    test_set = pd.read_excel(r'表./3.xlsx', engine='openpyxl')
    #调库生成K近邻算法模型-
    # for k in [4,6,8]:
    k = 6
    knn = neighbors.KNeighborsClassifier(n_neighbors=k, weights='uniform',
        algorithm='auto', p=2)
```

```
trainX, trainY = training_set.iloc[:, 7:], training_set.iloc[:, 3:4]
knn.fit(trainX, trainY.values.ravel()) # 通过训练集对模型进行拟合
score = knn.score(trainX, trainY)
print(score)
# print(test_set)
testX = test_set.iloc[:, 2:]
# print(testX)
predict_result = knn.predict(testX)
print(predict_result)

# for k in [1,2,3,4,5,6,7]:
k = 1
knn = neighbors.KNeighborsClassifier(n_neighbors=k, weights='uniform',
    algorithm='auto', p=2)
trainX, trainY = training_set.iloc[:, 7:], training_set.iloc[:, 2:3]
knn.fit(trainX, trainY) # 通过训练集对模型进行拟合
score = knn.score(trainX, trainY)
print(score)
testX = test_set.iloc[:, 2:]
predict_result = knn.predict(testX)
print(predict_result)



index1 = training_set[training_set["表面风化"] == 1].index.tolist()
t1 = training_set.iloc[index1,:]
index1_1 = t1[t1["类型"] == 1].index.tolist()
index1_2 = t1[t1["类型"] == 2].index.tolist()
t1_1 = training_set.iloc[index1_1,:] #无风化高钾
t1_2= training_set.iloc[index1_2,:] #无风化铅钡
y1_1 = test_set.iloc[0:1,2:].reset_index(drop=True) #无风化高钾待预测数据的
    读取
y1_2 = test_set.iloc[[2,3,7],2:].reset_index(drop=True) #无风化铅钡待预测数
    据的读取
# print(y1_1)
# print(y1_2)
index2 = training_set[training_set["表面风化"] == 2].index.tolist()
t2 = training_set.iloc[index2,:]
index2_1 = t2[t2["类型"] == 1].index.tolist()
```

```
index2_2 = t2[t2["类型"] == 2].index.tolist()
t2_1 = training_set.iloc[index2_1,:] #风化高钾
t2_2= training_set.iloc[index2_2,:] #风化铅钡
y2_1 = test_set.iloc[5:7,2:].reset_index(drop=True) #风化高钾待预测数据的读取
y2_2 = test_set.iloc[[1,4],2:].reset_index(drop=True) #风化铅钡待预测数据的读取
# print(y2_1)
# print(y2_2)

# for k in [1,2,3]: 通过循环可得到最优#值k
k = 1
knn = neighbors.KNeighborsClassifier(n_neighbors=k, weights='uniform', algorithm='auto', p=2)
trainX, trainY = t1_1.iloc[:, 7:], t1_1.iloc[:, 2:3]
knn.fit(trainX, trainY) # 通过训练集对模型进行拟合
score = knn.score(trainX, trainY)
print(score)
testX = y1_1.iloc[:, :]
predict_result = knn.predict(testX)
print(predict_result)

k = 1
knn = neighbors.KNeighborsClassifier(n_neighbors=k, weights='uniform', algorithm='auto', p=2)
trainX, trainY = t1_2.iloc[:, 7:], t1_2.iloc[:, 2:3]
# print(trainX)
knn.fit(trainX, trainY) # 通过训练集对模型进行拟合
score = knn.score(trainX, trainY)
print(score)
testX = y1_2.iloc[:, :]
predict_result = knn.predict(testX)
print(predict_result)

k = 1
knn = neighbors.KNeighborsClassifier(n_neighbors=k, weights='uniform', algorithm='auto', p=2)
trainX, trainY = t2_1.iloc[:, 7:], t2_1.iloc[:, 2:3]
```

```
    knn.fit(trainX, trainY) # 通过训练集对模型进行拟合
    score = knn.score(trainX, trainY)
    print(score)
    testX = y2_1.iloc[:, :]
    predict_result = knn.predict(testX)
    print(predict_result)

    k = 1
    knn = neighbors.KNeighborsClassifier(n_neighbors=k, weights='uniform',
        algorithm='auto', p=2)
    trainX, trainY = t2_2.iloc[:, 7:], t2_2.iloc[:, 2:3]
    # print(trainX)
    knn.fit(trainX, trainY) # 通过训练集对模型进行拟合
    score = knn.score(trainX, trainY)
    print(score)
    testX = y2_2.iloc[:, :]
    predict_result = knn.predict(testX)
    print(predict_result)

    #灵敏性分析
    k = 1
    knn = neighbors.KNeighborsClassifier(n_neighbors=k, weights='uniform',
        algorithm='auto', p=2)
    trainX, trainY = t1_1.iloc[:, 7:], t1_1.iloc[:, 2:3]
    knn.fit(trainX, trainY) # 通过训练集对模型进行拟合
    score = knn.score(trainX, trainY)
    print(score)
    testX =
        [[76.6,0,0,7,2,6,2,2,0,0,0.8,0.03,0,0.6],[79.7,0,0,5.8,1.69,7.12,2.03,2.10,0,0,1.20,0
    predict_result = knn.predict(testX)
    print(predict_result)
```

附录 E 第三问 –python 源程序

```
import pandas as pd
```

```
pd.set_option("display.max_columns", 15) # 可显示列1000
data = pd.read_excel(r'高钾./.xlsx', engine='openpyxl')
df = data.iloc[:, 6:]
df.columns = ['二氧化硅','氧化纳','氧化钾','氧化钙','氧化镁','氧化铝','氧化
    铁','氧化铜','氧化铅','氧化钡','五氧化二磷','氧化锶','氧化锡','二氧化硫']
df = df.fillna(0)
spearman = df.corr('spearman')
print(spearman)
spearman.to_excel('./spearman1.xlsx') # 路径和文件名


data = pd.read_excel(r'铅钡./.xlsx', engine='openpyxl')
df = data.iloc[:, 6:]
df.columns = ['二氧化硅','氧化纳','氧化钾','氧化钙','氧化镁','氧化铝','氧化
    铁','氧化铜','氧化铅','氧化钡','五氧化二磷','氧化锶','氧化锡','二氧化硫']
df = df.fillna(0)
spearman = df.corr('spearman')
print(spearman)
spearman.to_excel('./spearman2.xlsx') # 路径和文件名
```

附录 F 两种玻璃在有无风化情况下化学成分含量统计箱线图

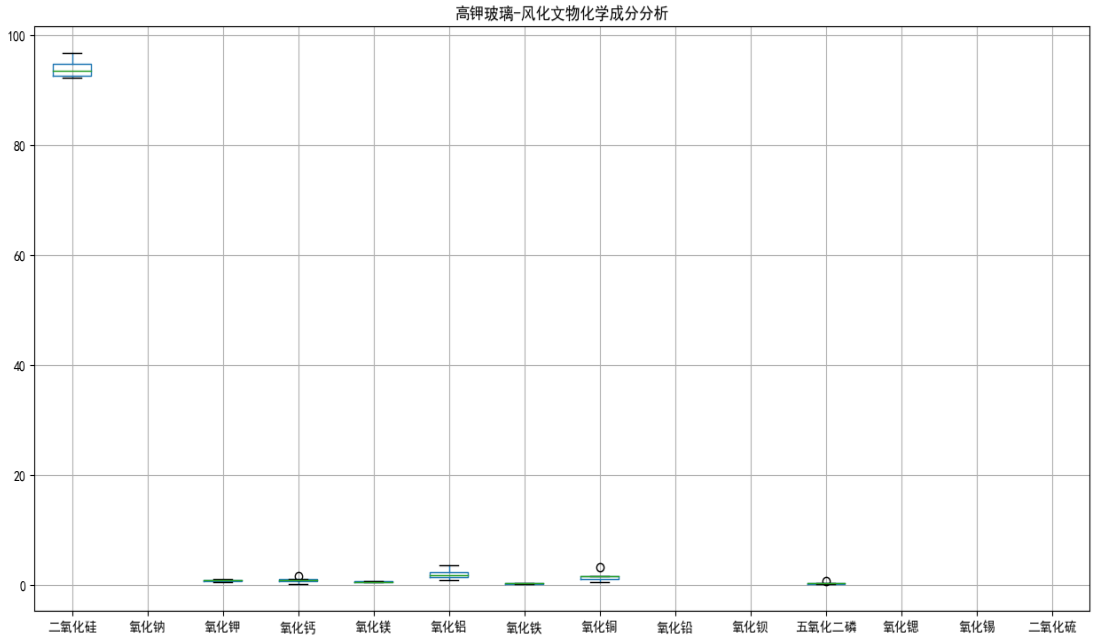

图 10　高钾玻璃 -风化文物化学成分含量统计结果

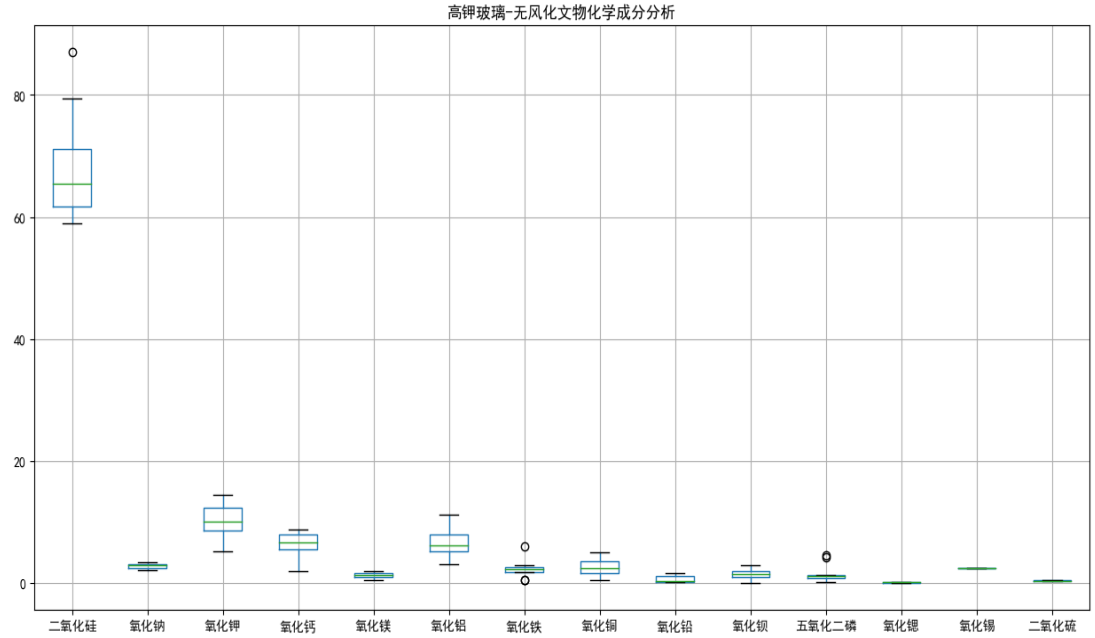

图 11　高钾玻璃 -无风化文物化学成分含量统计结果

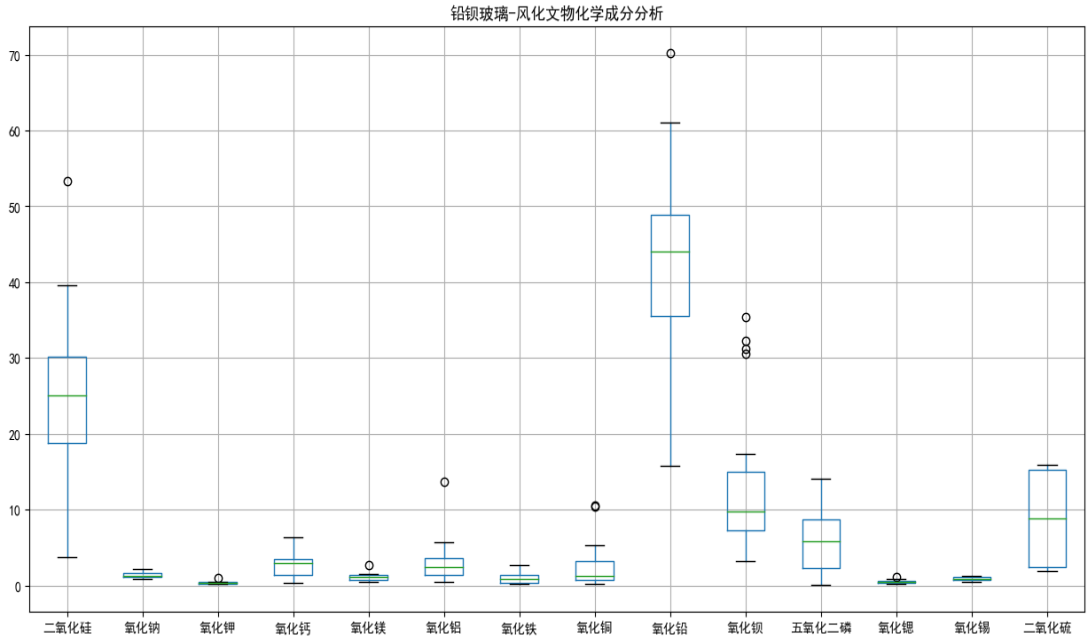

图 12　铅钡玻璃 -风化文物化学成分含量统计结果

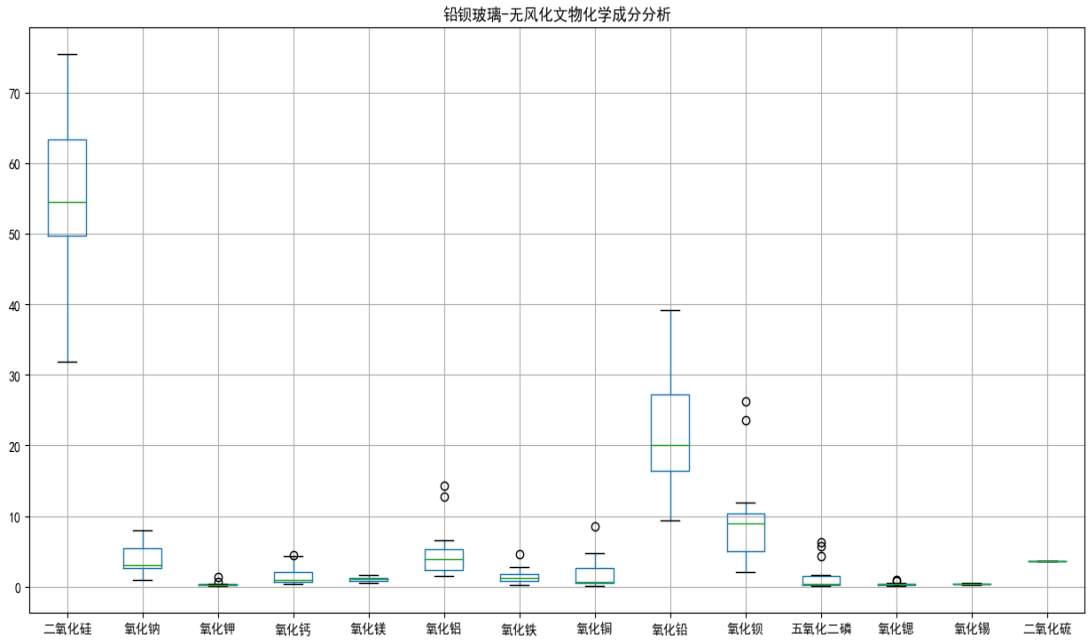

图 13　铅钡玻璃 -无风化文物化学成分含量统计结果

附录 G 各个风化点风化前的化学成分含量预测值

次级采样点	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al2O3)	氧化铁(Fe2O3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P2O5)	氧化锶(SrO)	氧化锡(SnO2)	二氧化硫(SO2)
B 高钾 铸造 风化前	07	67.035	7.87	6.78	1.49	5.55					2.77			
B 高钾 铸造 风化前	09	68.76	7.375	4.56	4.32	2.81	2.65				1.59			
B 高钾 铸造 风化前	10	70.02	11.5	1.54	2.77	2.28	1.44				0.00			
B 高钾 铸造 风化前	12	68.23	12.625	5.29	5.00	2.54	2.83				0.68			
B 高钾 铸造 风化前	22	66.02	9.25	12.21	1.40	11.99	3.07	0.04			0.95			
B 高钾 铸造 风化前	27	67.09	6.91	11.18	8.40	1.75	2.64				1.14			

				文物采样点	二氧化硅 (SiO2)	氧化钠 (Na2O)	氧化钾 (K2O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al2O3)	氧化铁 (Fe2O3)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P2O5)	氧化锶 (SrO)	氧化锡 (SnO2)	二氧化硫 (SO2)
A	铅钡	浅蓝	风化前	02	79.56	0.00	2.89	6.45	3.26	15.79	5.12	0.72	13.66	0.00	9.83	0.52	0.00	0.00
C	铅钡	紫	风化前	08	44.17	0.00	0.00	4.77	0.00	3.69	0.00	28.68	79.83	86.34	9.89	1.19	0.00	7.17
C	铅钡	紫	风化前	08 严重风化点	1.20	0.00	0.00	8.79	0.00	3.58	0.00	8.65	89.39	84.35	2.83	1.47	0.00	41.45
C	铅钡	浅蓝	风化前	11	73.66	0.00	0.58	9.67	1.96	7.41	0.00	13.58	69.94	4.25	25.84	1.19	0.00	0.00
C	铅钡	紫	风化前	26	43.40	0.00	0.00	3.97	0.00	1.93	0.00	29.12	81.35	88.84	8.62	1.24	0.00	5.40
C	铅钡	紫	风化前	26 严重风化点	8.16	0.00	1.12	8.29	0.00	3.26	0.00	9.92	82.42	97.66	16.64	1.78	0.00	43.94
C	铅钡	深绿	风化前	34	78.46	0.00	0.69	2.15	0.00	4.46	1.29	4.16	128.24	27.55	0.94	0.67	0.00	0.00
C	铅钡	深绿	风化前	36	86.78	6.12	0.39	1.19	0.00	4.48	0.88	1.87	114.63	29.83	0.19	0.67	0.00	0.00
C	铅钡	深绿	风化前	38	72.21	3.82	0.00	1.87	0.00	7.80	0.80	2.11	135.84	26.97	1.32	1.13	0.00	0.00
C	铅钡	深绿	风化前	39	57.57	0.00	0.00	3.58	0.00	1.38	0.00	2.42	168.13	19.89	3.20	1.68	0.00	0.00
C	铅钡	浅绿	风化前	41	4.48	0.00	1.21	13.66	7.53	9.17	4.93	0.52	121.54	26.89	2.56	1.29	0.00	0.00
C	铅钡	浅蓝	风化前	43 部位 1		27.21	0.00	0.00	14.44	2.45	6.20	2.94	14.74	164.88	2.83	0.00	1.76	0.00
0.00																		
C	铅钡	浅蓝	风化前	43 部位 2		47.59	0.00	0.00	17.64	2.62	9.39	3.83	4.16	123.28	8.99	35.34	1.29	0.00
0.00																		
A	铅钡	黑	风化前	49	63.14	0.00	0.00	12.62	4.50	14.83	7.55	1.93	94.16	16.84	3.58	1.27	0.00	0.00
A	铅钡	黑	风化前	50	39.43	0.00	0.00	8.79	1.29	5.15	1.00	3.11	121.21	39.12	17.47	1.82	0.00	0.00
C	铅钡	浅蓝	风化前	51 部位 1		53.97	0.00	0.00	9.86	3.28	14.46	3.28	3.77	11.85	24.63	22.31	1.74	1.29
0.00																		
C	铅钡	浅蓝	风化前	51 部位 2		46.82	0.00	0.00	14.13	3.99	6.91	1.16	2.66	141.43	0.00	24.15	0.00	0.00
0.00																		
C	铅钡	浅蓝	风化前	52	56.45	3.37	0.00	6.25	1.52	3.20	0.63	1.93	13.63	23.82	15.73	1.21	0.00	0.00
C	铅钡	浅蓝	风化前	54	48.86	0.00	0.88	8.79	3.53	11.43	0.00	2.29	152.78	19.39	11.68	2.42	0.00	0.00
C	铅钡	浅蓝	风化前	54 严重风化点	37.52	0.00	0.00	0.00	3.58	1.56	0.00	3.69	161.47	0.00	38.93	3.85	0.00	0.00
C	铅钡	蓝绿	风化前	56	63.93	0.00	0.00	3.33	0.00	5.96	0.00	2.18	113.64	42.56	7.00	0.00	0.00	0.00
C	铅钡	蓝绿	风化前	57	55.75	0.00	0.00	3.69	0.00	6.56	0.00	3.20	124.24	47.66	0.00	0.00	0.00	0.00
A	铅钡		风化前	19	65.00	0.00	0.00	8.72	1.63	9.83	3.66	9.67	117.96	14.74	24.33	0.52	0.00	0.00
C	铅钡		风化前	40	36.64	0.00	0.00	5.15	0.00	1.24	0.52	0.00	193.42	18.43	4.88	1.87	0.00	0.00
A	铅钡		风化前	48	116.95	2.24	0.88	7.77	4.24	37.63	2.84	0.00	43.28	2.14	3.33	0.69	3.69	0.00
C	铅钡		风化前	58	66.64	0.00	0.94	9.61	2.18	9.70	2.37	8.62	18.42	21.12	24.77	0.66	0.00	0.00