全国大学生数学建模竞赛论文模板

摘要

摘要

对于问题一，

对于问题二，

对于问题三，

对于问题四，

最后，

关键字： 关键词　关键词　关键词　关键词　关键词

一、 问题重述

1.1 问题背景

丝绸之路作为古代中西方文化交流的核心通道，玻璃是早期贸易往来的重要物证。早期西亚和埃及的玻璃多以珠形饰品传入我国，我国古代吸收其技术后，利用本土原料制作玻璃，虽外观与外来品相似，但因助熔剂差异（如铅矿石、草木灰等），化学成分截然不同，形成了铅钡玻璃（我国自创，以楚文化为代表）、高钾玻璃（流行于岭南及东南亚、印度等区域）等本土特色品种。古代玻璃因埋藏环境易风化，风化过程中元素交换导致成分比例改变，影响类别判断，而部分风化文物表面仍保留未风化区域，为成分研究提供了特殊样本，对于研究古代中国社会和玻璃工艺具有很高的价值。

1.2 问题要求

问题 1 分析玻璃文物的表面风化状态与其类型（高钾玻璃 / 铅钡玻璃）、纹饰、颜色之间的关联；结合玻璃类型，总结文物表面有无风化时化学成分含量的统计规律；并基于风化点的检测数据，预测其风化前的化学成分含量。

问题 2 依据附件数据，提炼高钾玻璃与铅钡玻璃的分类规律；针对这两类玻璃，分别选取合适的化学成分进行亚类划分，明确具体的划分方法及结果，并分析该分类结果的合理性与敏感性。

问题 3 对附件表单 3 中未知类别的玻璃文物，通过分析其化学成分鉴别其所属类型（高钾玻璃或铅钡玻璃），并对该分类结果的敏感性进行分析。

问题 4 针对高钾玻璃和铅钡玻璃这两类不同的文物样品，分别分析其内部化学成分之间的关联关系，并比较两类玻璃在化学成分关联关系上的差异性。

二、 问题分析

2.1 问题一分析

对于问题一，

2.2 问题二分析

对于问题二，

2.3 问题三分析

对于问题三，

2.4 问题四分析

对于问题四，

三、 模型假设

为简化问题，本文做出以下假设：

- 假设 1
- 假设 2
- 假设 3

四、 符号说明

符号	说明	单位
m	质量	kg
V	体积	m^3

五、 问题一的模型的建立和求解

5.1 玻璃类型、颜色、纹饰与风化的关系

首先我们对表单 2 中各文物采样点的化学成分进行累加，其中样本编号为 15、17 的文物化学成分总和分别为 79.47%、71.89%，不满足题目对成分比例累加和介于 85% 105% 之间的要求，因此我们将其剔除。

为了分析表面风化与玻璃类型、纹饰、颜色之间的关系，我们分别统计（表面风化，玻璃类型）（表面风化，纹饰）（表面风化，颜色）这三个二元组的列联表数据，并进行了可视化。

表 **1**　表面风化与颜色的列联表

表面风化	颜色							
	浅绿	浅蓝	深绿	深蓝	紫	绿	蓝绿	黑
无风化	2	6	3	2	2	1	6	0
风化	1	12	4	0	2	0	9	2

表 **2**　表面风化与纹饰、玻璃类型的列联表

(a) 表面风化与纹饰的列联表

表面风化	纹饰		
	A	B	C
无风化	11	0	11
风化	11	6	17

(b) 表面风化与玻璃类型的列联表

表面风化	类型	
	铅钡	高钾
无风化	12	10
风化	28	6

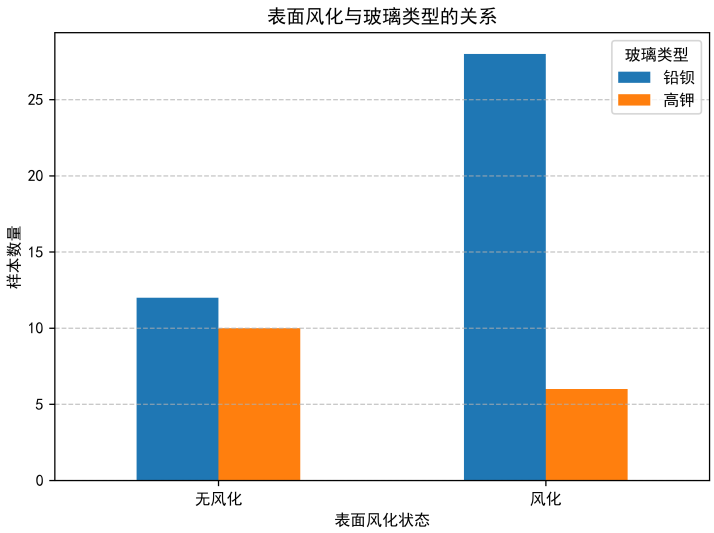


图 **1**　表面风化与玻璃类型

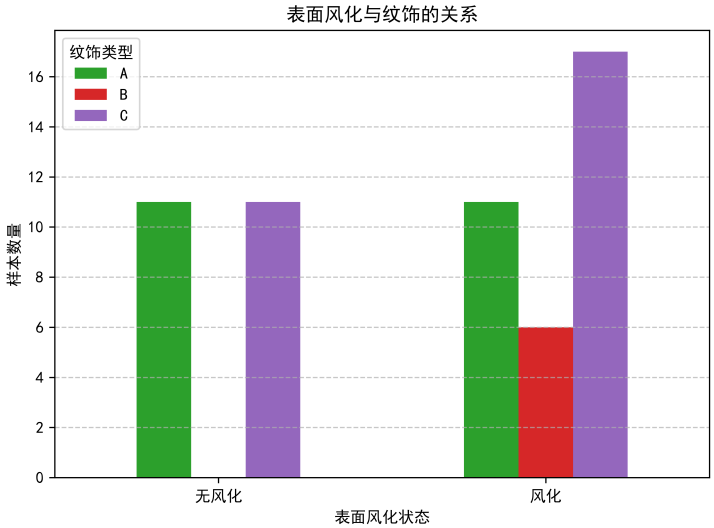


图 **2**　表面风化与纹饰

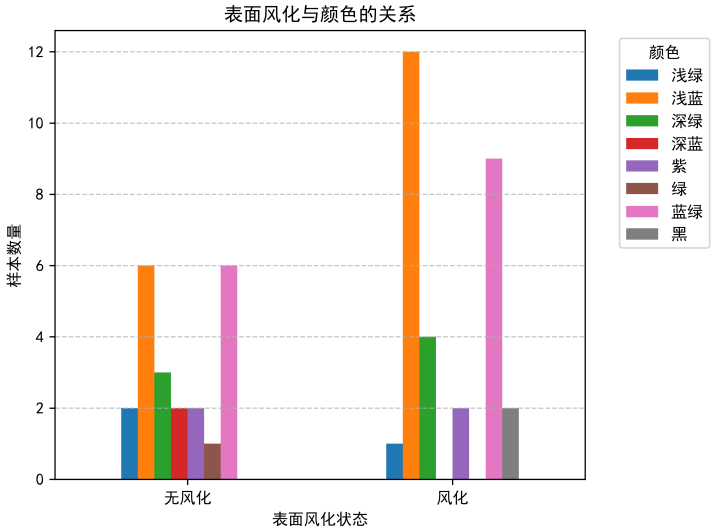


图 **3**　表面风化与颜色

为了量化表面风化与玻璃类型、纹饰、颜色之间的关系，我们引入了卡方检验。卡方检验用于检验两个分类变量是否独立，通过比较观测值与期望值的差异，用 χ^2 统计量判断关联是否显著，适用于计数数据。

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{1}$$

其中：χ^2：卡方统计量；O：实际观测频数；E：理论期望频数；$\sum$：对所有单元格求和。

分别带入（表面风化，玻璃类型）（表面风化，纹饰）（表面风化，颜色）的列联表数据可以求出 χ^2 值和 p 值，我们这里取 $p < 0.005$。从下面的表格中我们可以看出，是否风化与玻璃类型之间存在显著关系，而风化与纹饰、颜色之间则不存在显著关系。

表 **3**　卡方检验结果

关系	χ^2	df	p 值	是否显著
风化 × 颜色	7.0114	7	$p \approx 0.426$	否
风化 × 纹饰	4.9412	2	$p \approx 0.085$	否
风化 × 类型	5.0610	1	$p \approx 0.024$	是

5.2 玻璃是否风化化学成分含量的统计规律

以文物采样点为单位，以玻璃类型、是否风化为分组依据，将数据分为四个组别：

1. 无风化铅钡玻璃
2. 无风化高钾玻璃
3. 风化铅钡玻璃
4. 风化高钾玻璃

我们对预处理后的数据进行统计，计算出了每种组别的化学成分含量的均值、极差、方差、有效样本数。同时，为了更直观的看出化学成分的变化，我们将同一化学成分的风化前后含量做成了柱状图，以下进行部分展示。

表 **4**　铅钡无风化样本化学成分统计数据

化学成分	均值	极差	方差	有效样本数
二氧化硅 (SiO_2)	53.4438	43.5700	212.7885	13
氧化钠 (Na_2O)	3.3433	2.0000	1.3008	3
氧化钾 (K_2O)	0.4356	1.4600	0.2193	9
氧化钙 (CaO)	1.3909	4.1100	2.2167	11
氧化镁 (MgO)	1.4812	4.9400	2.6958	8
氧化铝 (Al_2O_3)	2.8915	3.5600	1.6320	13
氧化铁 (Fe_2O_3)	2.1240	4.4200	2.8855	5
氧化铜 (CuO)	1.8400	8.3500	6.8727	11
氧化铅 (PbO)	23.5938	29.9200	82.7080	13
氧化钡 (BaO)	24.4662	23.2300	62.8734	13
五氧化二磷 (P_2O_5)	1.0682	5.6500	2.7670	11
氧化锶 (SrO)	0.4825	0.6800	0.0664	8
氧化锡 (SnO_2)	0.4000	0.0000	nan	1
二氧化硫 (SO_2)	2.0500	3.2200	5.1842	2

表 **5**　铅钡风化样本化学成分统计数据

化学成分	均值	极差	方差	有效样本数
二氧化硅 (SiO_2)	33.6147	64.3600	296.5795	36
氧化钠 (Na_2O)	3.1173	7.1200	5.4794	11
氧化钾 (K_2O)	0.3937	1.3000	0.1208	16
氧化钙 (CaO)	2.4835	6.0300	2.4015	34
氧化镁 (MgO)	1.0970	2.2600	0.2306	23
氧化铝 (Al_2O_3)	3.8383	13.8900	11.6460	36
氧化铁 (Fe_2O_3)	0.9529	2.5500	0.4407	21
氧化铜 (CuO)	2.1135	10.3800	6.3094	34
氧化铅 (PbO)	36.8719	57.9000	229.9385	36
氧化钡 (BaO)	34.5803	64.8600	297.0158	35
五氧化二磷 (P_2O_5)	4.9863	14.0600	16.5052	30
氧化锶 (SrO)	0.4122	1.0000	0.0484	32
氧化锡 (SnO_2)	0.7700	1.0800	0.5832	2
二氧化硫 (SO_2)	7.1980	15.4800	57.9916	5

表 **6**　高钾无风化样本化学成分统计数据

化学成分	均值	极差	方差	有效样本数
二氧化硅 (SiO_2)	67.9842	28.0400	76.6518	12
氧化钠 (Na_2O)	2.7800	1.2800	0.4144	3
氧化钾 (K_2O)	9.7233	9.8100	9.2269	12
氧化钙 (CaO)	6.0500	7.4800	6.6337	10
氧化镁 (MgO)	1.3033	1.4600	0.2784	9
氧化铝 (Al_2O_3)	6.6200	8.1000	6.2076	12
氧化铁 (Fe_2O_3)	2.3180	5.6200	2.4002	10
氧化铜 (CuO)	2.6755	4.6200	2.3750	11
氧化铅 (PbO)	0.7057	1.5100	0.3939	7
氧化钡 (BaO)	1.4360	2.8600	1.1488	5
五氧化二磷 (P_2O_5)	1.5300	4.3400	2.0473	11
氧化锶 (SrO)	0.0833	0.0800	0.0010	6
氧化锡 (SnO_2)	2.3600	0.0000	nan	1
二氧化硫 (SO_2)	0.4067	0.1100	0.0032	3

表 7　高钾风化样本化学成分统计数据

化学成分	均值	极差	方差	有效样本数
二氧化硅 (SiO_2)	93.9633	4.4200	3.0054	6
氧化钾 (K_2O)	0.7040	0.7500	0.0879	5
氧化钙 (CaO)	0.8700	1.4500	0.2379	6
氧化镁 (MgO)	0.5900	0.1000	0.0050	2
氧化铝 (Al_2O_3)	1.9300	2.6900	0.9302	6
氧化铁 (Fe_2O_3)	0.2650	0.1800	0.0048	6
氧化铜 (CuO)	1.5617	2.6900	0.8739	6
五氧化二磷 (P_2O_5)	0.3360	0.4600	0.0316	5

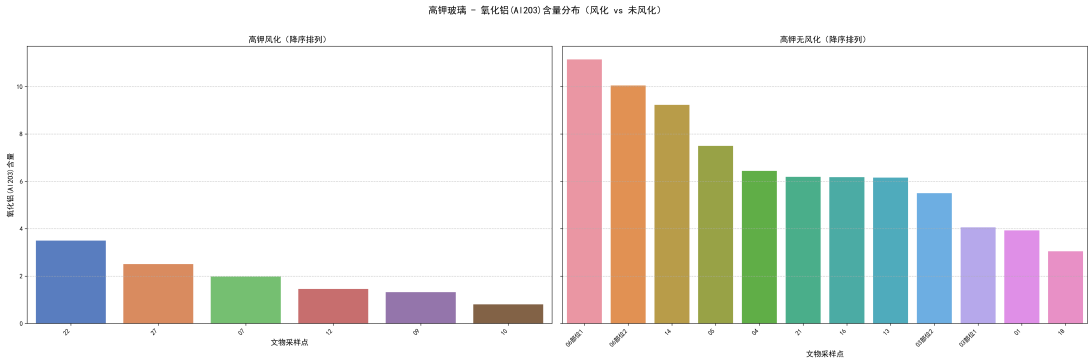


图 4　高钾玻璃 - 氧化铝 (Al_2O_3) 风化前后含量分布

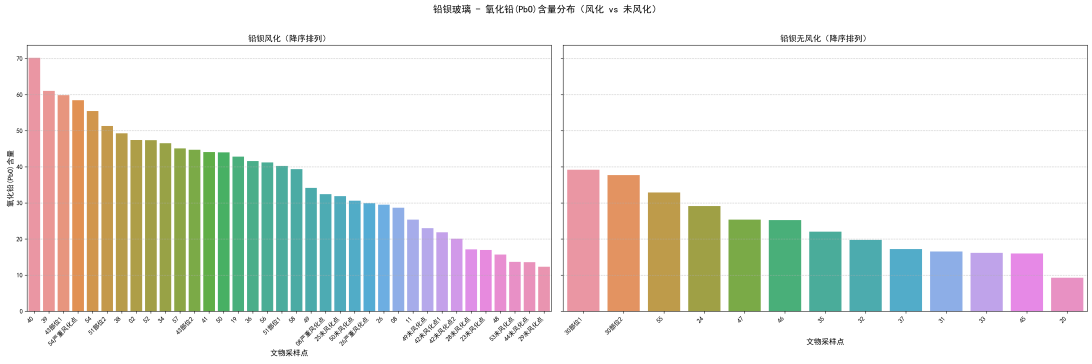
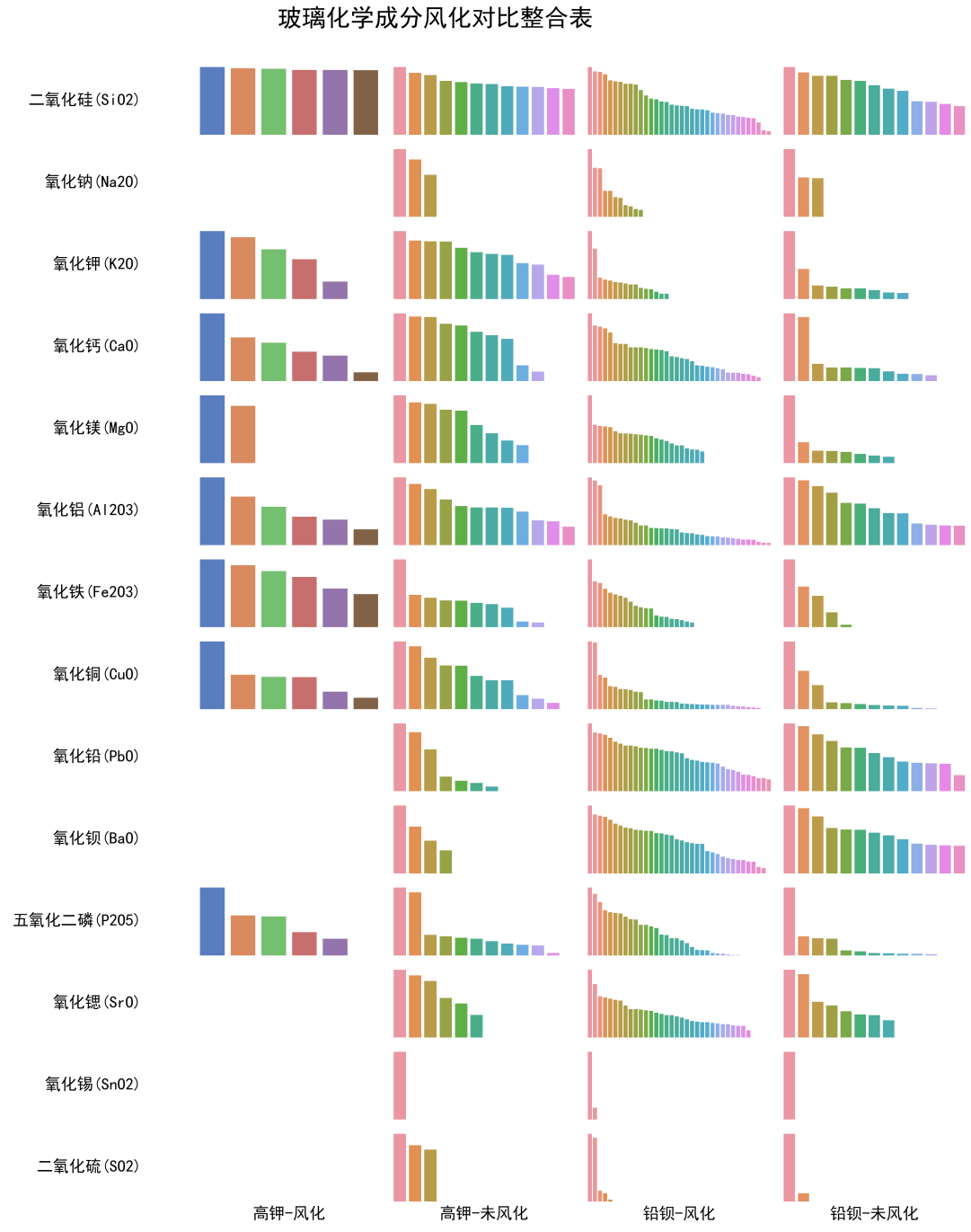


图 5　铅钡玻璃 - 氧化铅 (PbO) 风化前后含量分布



注：整合表展示了高钾和铅钡两种玻璃类型在风化前后的化学成分分布对比，按含量降序排列。

图 **6**　高钾-铅钡玻璃风化前后各化学含量对比

通过这些图表，可以直观地观察到：

- 铅钡玻璃在风化后，SiO_2 含量略微增加，而 SrO、Fe_2O_3、CuO、Na_2O、MgO、CaO、K_2O、Al_2O_3、PbO、BaO、P_2O_5 和 SO_2 含量均有不同程度的下降；
- 高钾玻璃在风化后，SiO_2 含量显著降低，而 SrO、Fe_2O_3、CuO、Na_2O、MgO、CaO、K_2O、Al_2O_3、PbO、BaO、P_2O_5 和 SO_2 含量明显增加，SnO_2 含量基本不变；
- 风化过程对不同类型玻璃的化学成分影响存在明显差异。

5.3 风化前的化学成分含量的预测

5.3.1 数据分析

通过阅读数据，我们发现，除了编号为 49、50 的样本具有成对的 (风化前，风化后) 的化学成分数据，其他的样本都是基于风化前后两群体的横截面数据，可以抽象为 (风化前，NULL) 或 (NULL，风化后) 的形式。同时，考虑到需要预测的是化学成分，是总和为 1 的成分数据，各成分之间彼此依赖，不适宜使用一般的回归模型。

因此，我们基于成分数据分析 CoDA 模型，参考文献 [**?**]，在模型中添加先验正则项，最终建立风化前化学含量预测模型。

5.3.2 模型记号与数据映射

- 模型记号
  - $\mathbf{a}_i$ （原始成分百分比）
  - $\mathbf{p}_i$ （闭合后，和为 1 的比例）
  - $\mathbf{z}_i$ （clr 后的向量）
  - $\bar{\mathbf{z}}_{\mathrm{w}}$, $\bar{\mathbf{z}}_{\mathrm{u}}$
  - $\boldsymbol{\delta}_0$ （按玻璃类型用文献导向值初始化）
  - $\mathbf{w}$ （按成分设置权重）
  - λ （正则强度）
  - n （样本数）
- 令成分列为：

$$\mathcal{P} = \{\mathrm{SiO_2}, \mathrm{Na_2O}, \mathrm{K_2O}, \ldots, \mathrm{SO_2}, \mathrm{unknown}\}$$

  共 $D = 15$ 个成分
- 将每行按百分比除以 100 进行闭合，得到组成比例矩阵 $\mathbf{p} \in \mathbb{R}^{n \times D}$。
  - 样本 i 的成分向量为 $\mathbf{a}_i$，闭合后

$$\mathbf{p}_i = \mathcal{C}(\mathbf{a}_i) = \frac{\mathbf{a}_i}{\sum_{j=1}^{D} a_{ij}}, \quad \sum_{j=1}^{D} p_{ij} = 1.$$

- 为避免 $\log(0)$，在闭合前/或闭合后对零值做微小替换（伪计数）ε。
- CLR 变换（把单纯形映到实向量空间）：对每行 i

$$\mathbf{z}_i = \mathrm{clr}(\mathbf{p}_i) = \left(\ln \frac{p_{i1}}{g_i}, \ldots, \ln \frac{p_{iD}}{g_i} \right), \quad g_i = \left(\prod_{j=1}^{D} p_{ij} \right)^{1/D},$$

5.3.3 风化模型

我们在 CLR 空间做建模，主要假设（能使问题可解且可解释）：

线性位移假设 对同类玻璃，风化后与风化前在 CLR 空间上满足近似的平移关系（均值差）：

$$\mathbf{z}_{\mathrm{w}} \approx \mathbf{z}_{\mathrm{pre}} + \boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

其中 $\boldsymbol{\delta}$ 是同类玻璃的平均风化位移向量（在 CLR 空间），$\boldsymbol{\varepsilon}$ 是噪声。

这样就可以得到基于群体差的回推策略而不需要成对样本。

- 设 $\bar{\mathbf{z}}_{\mathrm{w}}$、$\bar{\mathbf{z}}_{\mathrm{u}}$ 分别为风化与未风化样本在 CLR 空间的均值，观测到的平均位移为：

$$\hat{\boldsymbol{\delta}}_{\mathrm{obs}} = \bar{\mathbf{z}}_{\mathrm{w}} - \bar{\mathbf{z}}_{\mathrm{u}}.$$

- 对任一风化样点 i，其风化前的 CLR 估计为：

$$\hat{\mathbf{z}}_{\mathrm{pre},i}^{(\mathrm{pure})} = \mathbf{z}_{\mathrm{w},i} - \hat{\boldsymbol{\delta}}_{\mathrm{obs}}.$$

- 然后逆 CLR 得到比例并乘以 100 得到百分比：

$$\hat{\mathbf{p}}_{\mathrm{pre},i} = \mathrm{clr}^{-1}(\hat{\mathbf{z}}_{\mathrm{pre},i}) = \frac{\exp(\hat{\mathbf{z}}_{\mathrm{pre},i})}{\sum_{k=1}^{D} \exp(\hat{z}_{\mathrm{pre},i,k})}.$$

5.3.4 将文献先验融合进 CoDA 模型

为了把“机理知识”引入（例如高钾玻璃倾向于 K_2O 严重流失、SiO_2 相对富集；铅钡体系可能出现 Pb/Ba 比例变化与硫酸盐富集等），我们在 CLR 空间对 $\boldsymbol{\delta}$ 加带方向性的岭惩罚：

- 设观测到的 $\boldsymbol{d} = \hat{\boldsymbol{\delta}}_{\mathrm{obs}}$。我们引入先验中心向量 $\boldsymbol{\delta}_0$（来源于机理/文献方向性），以及对每个成分的权重向量 $\mathbf{w} = (w_1, \ldots, w_D)$（表示你对该分量先验的信心和强度），并设正则强度为 $\lambda \geq 0$。最小化目标：

$$\min_{\boldsymbol{\delta}} \|\boldsymbol{\delta} - \boldsymbol{d}\|_2^2 + \lambda \|\mathbf{W}(\boldsymbol{\delta} - \boldsymbol{\delta}_0)\|_2^2,$$

其中 $\mathbf{W} = \mathrm{diag}(\mathbf{w})$。这是二次型问题，有闭式解：

$$\hat{\boldsymbol{\delta}}_{\mathrm{reg}} = (\mathbf{I} + \lambda \mathbf{W}^2)^{-1}(\boldsymbol{d} + \lambda \mathbf{W}^2 \boldsymbol{\delta}_0).$$

当 $\lambda = 0$ 或者 $\mathbf{w} = 0$ 时，退化为纯数据驱动（$\hat{\boldsymbol{\delta}} = \boldsymbol{d}$）。当 λ 较大且某些 w_j 很大时，$\hat{\boldsymbol{\delta}}$ 会更靠近 $\boldsymbol{\delta}_0$（先验主导）。

5.3.5 先验权重 λ 和噪声 ϵ 的调优

基于已知文物编号 49、50 风化前后的化学成分变化，我们将其作为测试集，通过网格搜索来寻找先验权重 λ 和噪声 ϵ 的最优值。我们使用以下公式检验模型性能：

$$\mathrm{MAE}_{\text{prior\_percent}} = \frac{\mathrm{MAE}_{\text{model}}}{\mathrm{MAE}_{\text{prior}}} \times 100\%$$

$$\text{其中} \quad \mathrm{MAE}_{\text{model}} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \quad \mathrm{MAE}_{\text{prior}} = \frac{1}{n}\sum_{i=1}^{n} |y_i - p_i|$$

经过网格搜索，最终得到化学成分预测模型如下：

$$\widehat{\mathbf{a}}_{i,\text{pre}}^{\text{prior}}(\%) = 100 \cdot \frac{\exp\left(\mathbf{z}_i - \widehat{\boldsymbol{\delta}}_g^{\text{prior}}\right)}{\mathbf{1}^\top \exp\left(\mathbf{z}_i - \widehat{\boldsymbol{\delta}}_g^{\text{prior}}\right)}, \tag{2}$$

$$\text{其中} \quad \widehat{\boldsymbol{\delta}}_g^{\text{prior}} = \left(\mathbf{I} + \lambda \mathbf{W}_g^2\right)^{-1} \left[\mathbf{d}_g + \lambda \mathbf{W}_g^2 \, \boldsymbol{\delta}_{0,g}\right], \tag{3}$$

$$\text{s.t.} \quad \begin{cases} \lambda \geq 0, \quad \mathbf{w}_g \in \mathbb{R}_{\geq 0}^D, \\ \boldsymbol{\delta}_{0,g} \in \mathcal{H}, \quad \mathbf{d}_g \in \mathcal{H}, \quad \mathbf{z}_i \in \mathcal{H}, \\ \mathcal{H} = \left\{ \mathbf{z} \in \mathbb{R}^D \;\middle|\; \sum_{j=1}^D z_j = 0 \right\}, \\ \mathbf{p} \in \mathcal{S}^D, \quad \mathcal{S}^D = \left\{ \mathbf{p} \in \mathbb{R}_{>0}^D \;\middle|\; \sum_{j=1}^D p_j = 1 \right\}, \\ \mathbf{W}_g = \mathrm{diag}(w_{g1}, \ldots, w_{gD}), \quad w_{gj} \geq 0, \; \forall j, \\ i \in S_g^{(w)} \;\Rightarrow\; g \text{ 是样本 } i \text{ 的玻璃类型标签.} \end{cases} \tag{4}$$

参考链接: `legest.ufpr.br`, `econ-papers.upt.edu`

这句话引用了文献 [**?**]。

这句话引用了文献[?]。

5.4 模型求解

Step1：

Step2：

Step3：

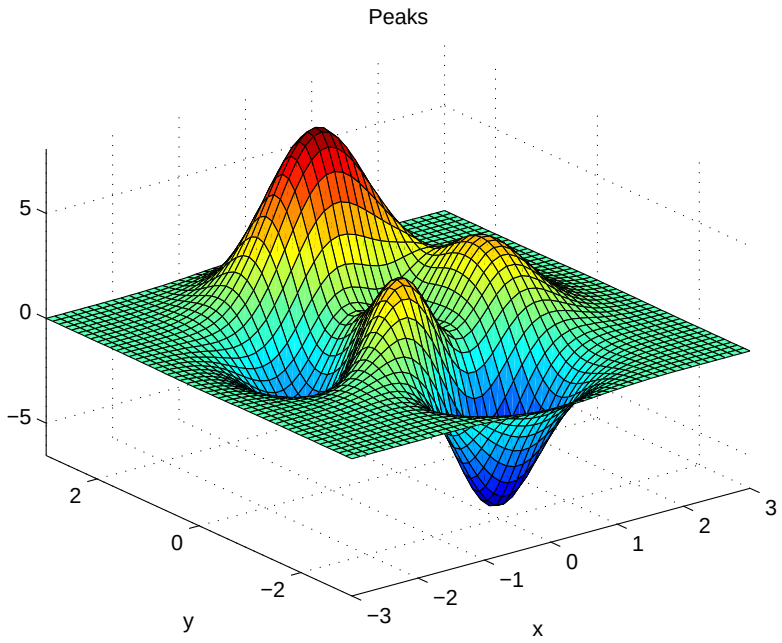


图 7　单图

5.5 求解结果

六、 问题二的模型的建立和求解

6.1 基于决策树的高钾、铅钡玻璃的分类

6.1.1 模型构建与求解

根据题目中已知的玻璃类型（高钾 / 铅钡），基于附件中各样本的化学成分，我们分别在“风化样本”和“未风化样本”中构建决策树分类模型，用于提取分类规律。

数据与符号　共有 N 个有效样本（文物采样点），每个样本的化学成分向量为

$$\mathbf{x}_j = \big(x_{j,1}, x_{j,2}, \ldots, x_{j,12}\big)^\top, \quad j = 1, \ldots, N,$$

对应 12 种主要成分：

$$(\mathrm{SiO_2},\ \mathrm{Na_2O},\ \mathrm{K_2O},\ \mathrm{CaO},\ \mathrm{MgO},\ \mathrm{Al_2O_3},\ \mathrm{Fe_2O_3},\ \mathrm{CuO},\ \mathrm{PbO},\ \mathrm{BaO},\ \mathrm{P_2O_5},\ \mathrm{SrO}).$$

类别标签 $y_j \in \mathcal{Y} = \{\text{高钾}, \text{铅钡}\}$，表面风化指示 $w_j \in \{0, 1\}$（1 表示“风化”，0 表示“无风化”）。

数据有效性筛选与归一化　原始比例（含缺测记为 0）在样本 j 的总和记为

$$s_j = \sum_{i=1}^{12} \tilde{x}_{j,i},$$

仅保留满足

$$85 < s_j < 105$$

的样本（单位：%）。对每一维成分做 min–max 归一化：设有效集合 $\mathcal{I}$ 上

$$m_i = \min_{j \in \mathcal{I}} \tilde{x}_{j,i}, \qquad M_i = \max_{j \in \mathcal{I}} \tilde{x}_{j,i},$$

则归一化后

$$x_{j,i} = \frac{\tilde{x}_{j,i} - m_i}{M_i - m_i}, \quad i = 1, \ldots, 12.$$

分组与数据划分 按风化与否将数据分为两组

$$\mathcal{D}^{(w)} = \{(\mathbf{x}_j, y_j) : w_j = w\}, \quad w \in \{0, 1\}.$$

对每组分别做训练/测试划分（比例 7 : 3，固定随机种子）：

$$\mathcal{D}^{(w)} = \mathcal{D}^{(w)}_{\text{train}} \cup \mathcal{D}^{(w)}_{\text{test}}, \quad \mathcal{D}^{(w)}_{\text{train}} \cap \mathcal{D}^{(w)}_{\text{test}} = \varnothing.$$

CART 决策树（Gini 准则，最大深度 = 3） 设当前节点样本集合为 S，其 Gini 不纯度定义为

$$\text{Gini}(S) = 1 - \sum_{c \in \mathcal{Y}} p(c \mid S)^2, \qquad p(c \mid S) = \frac{1}{|S|} \sum_{(\mathbf{x}, y) \in S} \mathbf{1}\{y = c\}.$$

在候选特征 $i \in \{1, \ldots, 12\}$ 与阈值 $t \in [0, 1]$ 中搜索最优二分

$$S_L(i, t) = \{(\mathbf{x}, y) \in S : x_i \le t\}, \qquad S_R(i, t) = S \setminus S_L(i, t),$$

以最小化加权 Gini：

$$(i^\star, t^\star) = \arg\min_{i,t} \Phi(i, t; S) := \frac{|S_L(i,t)|}{|S|} \text{Gini}\big(S_L(i,t)\big) + \frac{|S_R(i,t)|}{|S|} \text{Gini}\big(S_R(i,t)\big).$$

递归地在子节点 S_L, S_R 上重复该过程，直到满足停止条件之一：

(a) 节点纯度$\text{Gini}(S) = 0$; (b) 无法进一步降低Φ; (c) 当前深度$d = 3$ (最大深度).

叶节点预测与经验概率 落入叶节点 L 的样本以经验概率

$$\hat{p}(c \mid L) = \frac{1}{|L|} \sum_{(\mathbf{x}, y) \in L} \mathbf{1}\{y = c\}$$

进行投票预测：

$$\hat{f}(\mathbf{x}) = \arg\max_{c \in \mathcal{Y}} \hat{p}(c \mid L(\mathbf{x})).$$

测试集评估指标（准确率） 对组 w 的测试集准确率为

$$\mathrm{Acc}^{(w)} = \frac{1}{|\mathcal{D}_{\text{test}}^{(w)}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{test}}^{(w)}} \mathbf{1}\{\hat{f}^{(w)}(\mathbf{x}) = y\}.$$

在风化组（$w = 1$）中，训练得到的决策树如下：

$$\hat{f}^{(1)}(\mathbf{x}) = \begin{cases} \text{铅钡}, & x_{\text{PbO}} \le 0.088, \\ \text{高钾}, & x_{\text{PbO}} > 0.088. \end{cases}$$

该树的根节点以 PbO 为唯一分裂特征，阈值为 0.088。若 PbO 含量较低，则分类为铅钡玻璃；反之，则为高钾玻璃。对应的叶子节点纯度均为 100%（即 $gini = 0.0$），说明该划分在训练样本上完全分离了两类。

在未风化组（$w = 0$）中，训练得到的决策树为：

$$\hat{f}^{(0)}(\mathbf{x}) = \begin{cases} \text{铅钡}, & x_{\text{PbO}} \le 0.124, \\ \text{高钾}, & x_{\text{PbO}} > 0.124. \end{cases}$$

该树同样以 PbO 为唯一分裂特征，阈值略高，为 0.124。此时两类样本在根节点被划分为两个子集，左子树全部为铅钡（$gini = 0.0$），右子树全部为高钾（$gini = 0.0$）。

最终的决策模型表达式 结合风化组与未风化组的情况，可以统一表示为：

$$\hat{f}^{(w)}(\mathbf{x}) = \begin{cases} \text{铅钡}, & x_{\text{PbO}} \le \tau_w, \\ \text{高钾}, & x_{\text{PbO}} > \tau_w, \end{cases}$$

其中风化组的阈值为 $\tau_1 = 0.088$，未风化组的阈值为 $\tau_0 = 0.124$。该结果表明：

无论是否风化，氧化铅（PbO）含量是区分高钾与铅钡玻璃的唯一关键特征，且风化状态会轻微

6.1.2 特征重要性和敏感性分析

通过 Gini 指标的贡献度计算，得到的特征重要性结果如下：

$$I_{\text{PbO}} = 1.0, \qquad I_{\text{其他成分}} = 0.$$

即 PbO 在两组数据中都是唯一的分裂特征，其它 11 种化学成分的重要性均为 0。这进一步验证了 PbO 对玻璃类型分类的决定性作用。

为了检验模型对阈值变化的稳健性，我们在风化组样本上测试了不同 PbO 阈值下的分类准确率，结果如图 **??** 所示。

从图中可以看出，在 $\theta = 0.088$ 与 $\theta = 0.124$ 附近，模型准确率达到 100%，表明该模型对阈值的选择相对稳定；若偏离该区间，则准确率迅速下降。

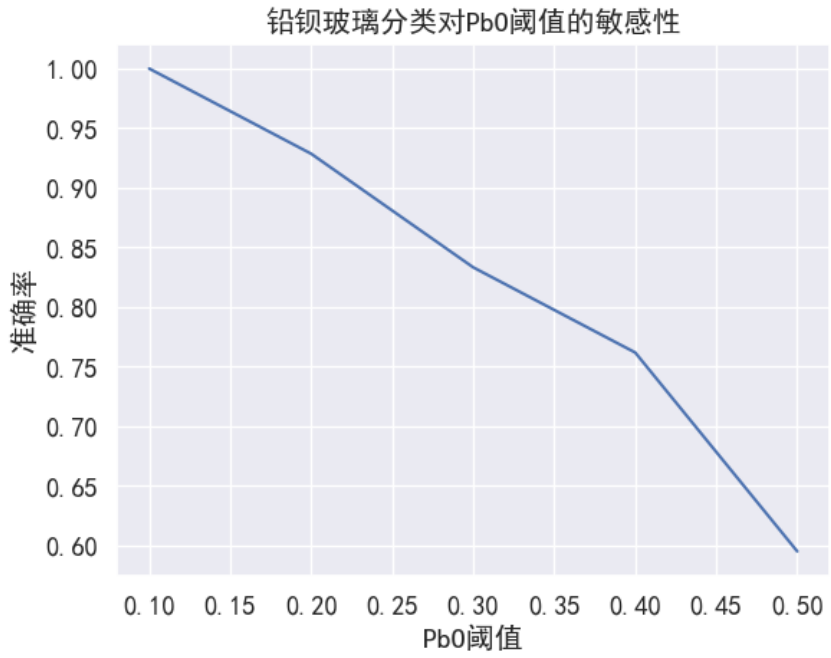


图 **8**　铅钡玻璃分类对 **PbO** 阈值的敏感性。横轴为阈值 θ，纵轴为分类准确率。

6.1.3 结论性说明

综上，决策树训练结果显示：1. PbO 是唯一关键判别特征；2. 风化组与未风化组的最佳阈值分别为 0.088 与 0.124；3. 该模型能够以简单的单一阈值规则实现对高钾玻璃与铅钡玻璃的完全区分。

6.2 基于聚类的亚类划分与显著性检验

6.2.1 问题与方法概述

在已知主类（高钾 / 铅钡）以及样本的风化状态（风化/无风化）的条件下，我们对每一类内部进一步进行亚类划分，目的是挖掘同一主类内部可能存在的不同配方或来源。具体流程为：对每一类（类型 × 风化状态）独立进行标准化后采用 K-means 聚类；并通过单因素方差分析（ANOVA）与 Tukey HSD 事后检验识别不同亚类间显著区分的化学成分。

记号与数据分组　延续前文符号，令所有有效样本归一化后的化学成分向量为 $\mathbf{x}_j \in \mathbb{R}^{12}$（含 12 种主要成分），类型标签 $y_j \in \{\text{高钾}, \text{铅钡}\}$，风化指示 $w_j \in \{0, 1\}$。对类型 $t \in \{\text{高钾}, \text{铅钡}\}$ 与风化状态 w，定义子数据集

$$\mathcal{G}^{(t,w)} = \{\, \mathbf{x}_j : y_j = t,\ w_j = w \,\}.$$

6.2.2 K-means 聚类模型与选择

K-means 的目标是将 $\mathcal{G}$ 中的样本分成 K 个簇，使簇内平方和（SSE，inertia）最小。设簇心为 $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$，簇划分为 $\{C_1, \ldots, C_K\}$，则优化目标为

$$J(\{C_k\}, \{\boldsymbol{\mu}_k\}) = \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in C_k} \left\| \mathbf{x}_j - \boldsymbol{\mu}_k \right\|^2, \tag{5}$$

其中 $\|\cdot\|$ 为欧氏范数。K-means 通过交替步骤（分配最近簇心与更新簇心）逼近最优解。

模型选择：为确定每个子组的簇数 K，采用肘部法（观察 SSE 随 K 的拐点）并结合实际可解释性选择最终 K。在实现中，对每个子组计算了 $K = 1, \ldots, 5$ 的 SSE 并绘制拐点图（见图 **??** 的 4 个子图）。最终选定的簇数如下：

$$K(\text{高钾 \_ 无风化}) = 3, \quad K(\text{高钾 \_ 风化}) = 2,$$
$$K(\text{铅钡 \_ 无风化}) = 3, \quad K(\text{铅钡 \_ 风化}) = 3.$$

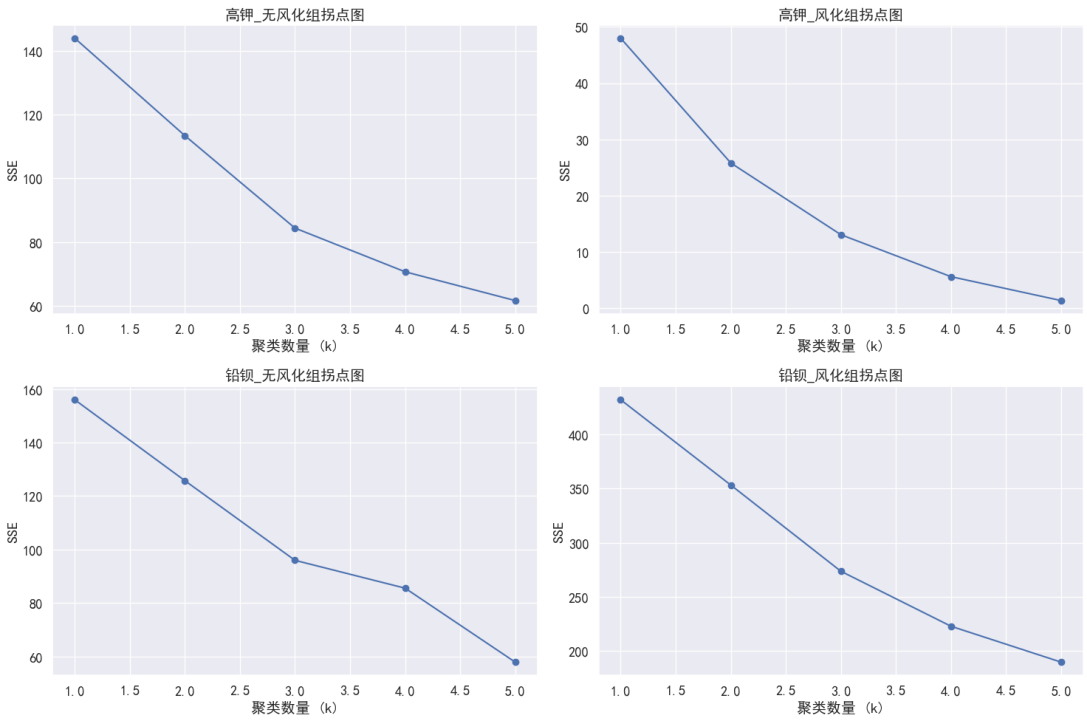


图 **9**　四个子组的肘部图（每个子组 **SSE** 随 K 的变化）。左上：高钾 _ 无风化；右上：高钾 _ 风化；左下：铅钡 _ 无风化；右下：铅钡 _ 风化。

6.2.3 聚类实施与亚类标注

对每个子组 $\mathcal{G}^{(t,w)}$，在选定的 K 下执行 K-means 并将聚类标签记为 $z_j \in \{0, \ldots, K-1\}$。由此得到一个分段映射

$$\kappa^{(t,w)} : \mathcal{G}^{(t,w)} \to \{0, \ldots, K-1\}, \qquad \kappa^{(t,w)}(\mathbf{x}_j) = z_j.$$

聚类后，我们对每个亚类的中心 $\boldsymbol{\mu}_k$ 及其样本集合 C_k 做统计描述并用于后续显著性检验与解释。

在此以“铅钡 _ 风化组”为例，部分样本的亚类划分结果如表 **??** 所示：

表 **8**　铅钡 _ 风化组部分样本的亚类划分结果

文物采样点	类型	表面风化	亚类
02	铅钡	风化	0
08	铅钡	风化	2
08 严重风化点	铅钡	风化	2
19	铅钡	风化	1
23 未风化点	铅钡	风化	0
26	铅钡	风化	2
41	铅钡	风化	1
43 部位 1	铅钡	风化	1
44 未风化点	铅钡	风化	0
49	铅钡	风化	1

从表中可以看出，同一主类（如铅钡 _ 风化）内部确实被划分为若干亚类（$0, 1, 2$），这为后续的显著性成分分析提供了基础。

6.2.4 亚类间成分显著性检验

为识别哪些化学成分在亚类间具有显著差异，我们对每个子组按其亚类标签分别对 12 种成分做单因素 ANOVA 检验。设某成分在不同亚类 $k = 1, \ldots, K$ 上的样本集为 $\{x_j^{(k)}\}$，ANOVA 的 F 统计量定义为

$$F = \frac{\text{组间均方}}{\text{组内均方}} = \frac{\dfrac{1}{K-1}\displaystyle\sum_{k=1}^{K} n_k (\bar{x}_{\cdot}^{(k)} - \bar{x})^2}{\dfrac{1}{N-K}\displaystyle\sum_{k=1}^{K}\sum_{j \in C_k} (x_j^{(k)} - \bar{x}_{\cdot}^{(k)})^2}, \tag{6}$$

其中 $n_k = |C_k|$，$\bar{x}_{\cdot}^{(k)}$ 为第 k 簇的样本均值，$\bar{x}$ 为总体均值，$N = \sum_k n_k$。当 F 对应的 p 值小于显著性水平 α（此处取 $\alpha = 0.05$）时，拒绝“各簇均值相等”的原假设。

若 ANOVA 显著，再对该成分做 Tukey HSD（Honestly Significant Difference）多重比较以辨别哪一对簇之间存在显著差异。Tukey HSD 的检验量可表示为

$$q_{ab} = \frac{|\bar{x}^{(a)} - \bar{x}^{(b)}|}{\sqrt{\dfrac{\text{MSE}}{n'}}},$$

其中 MSE 为均方误差估计，n' 为做比较时的等效样本数（在样本量不等时使用调整量），并与学生化范围分布（Studentized range）比较以得 p 值。

6.2.5 检验结果（将代码运行输出代入）

根据程序运行结果，对每个子组筛选得到的在亚类间差异显著的成分（按 p 值升序列出，取前若干项）如下：

表 **9**　各子组亚类间显著性成分检验结果

子组类型	显著成分	p 值
高钾 _ 无风化	氧化钠 (Na_2O)	9.14×10^{-7}
	二氧化硅 (SiO_2)	4.23×10^{-4}
	氧化钾 (K_2O)	1.43×10^{-2}
	氧化钙 (CaO)	3.00×10^{-2}
	氧化镁 (MgO)	4.00×10^{-2}
高钾 _ 风化	二氧化硅 (SiO_2)	1.99×10^{-2}
	氧化铝 (Al_2O_3)	3.95×10^{-2}
铅钡 _ 无风化	五氧化二磷 (P_2O_5)	1.30×10^{-5}
	氧化铁 (Fe_2O_3)	1.29×10^{-4}
	氧化钙 (CaO)	7.59×10^{-4}
铅钡 _ 风化	氧化钡 (BaO)	8.34×10^{-10}
	二氧化硅 (SiO_2)	3.14×10^{-8}
	氧化铜 (CuO)	1.35×10^{-6}
	五氧化二磷 (P_2O_5)	2.55×10^{-5}
	氧化铅 (PbO)	4.44×10^{-5}
	氧化锶 (SrO)	4.87×10^{-4}

这些结果表明：在高钾類內部，Na_2O 與 SiO_2 等成分對亞類區分貢獻最大；在铅钡類中，BaO、SiO_2、CuO 等成分在風化样本的亚类区分中尤为显著，且在铅钡 _ 风化组中 PbO 自身也对亚类区分显著（$p = 4.44 \times 10^{-5}$），提示风化与铅物质迁移/富集有关。

6.2.6 模型总结性表述

综上，亚类划分流程可以表示为两步映射：

$$\mathbf{x}_j \xrightarrow{\text{标准化}} \tilde{\mathbf{x}}_j \xrightarrow{\kappa^{(t,w)}} z_j \in \{0, \ldots, K-1\},$$

并对每一对 (t, w) 给出 $K = K^{(t,w)}$（上文已列出）。随后对每个成分做 ANOVA 与 Tukey HSD 检验以识别在亚类间具有统计显著性的化学成分集合 $\mathcal{S}^{(t,w)}$。最终得到的数学模型既给出了“样本 → 亚类”的划分规则（由 K-means 中心与最近原则隐含），也给出了“亚类的化学表征”（簇心 $\boldsymbol{\mu}_k$ 与显著成分 $\mathcal{S}^{(t,w)}$）。

6.2.7 对结果的解释与讨论

1. 亚类划分揭示了同一主类内可能存在不同的配方或原料来源，例如高钾 _ 无风化组中 Na_2O 与 SiO_2 的显著差异可能反映了助熔剂或熔炼温度的不同工艺路线。
2. 铅钡 _ 风化组中 BaO、PbO、CuO 等成分的显著性提示风化过程可能导致某些元素的迁移或富集，从而在亚类间产生差异。
3. 方法学上，K-means 依赖于欧氏距离与簇数选择，存在对簇形态（球状簇）与初始值敏感的问题。为增强稳健性，可采用多次随机初始化、Bootstrap 验证或结合层次聚类等方法进一步验证亚类稳定性。

6.2.8 小结

本节通过 K-means 聚类结合 ANOVA/Tukey HSD 的统计检验，在主类内部识别出若干意义明确的亚类，并给出每一亚类的化学成分差异性证据（显著性 p 值）。这些亚类与显著化学成分将用于后续问题（如未知样本的鉴别、成分来源推断）提供更细粒度的判据与解释。

七、 问题三模型应用与未知类别玻璃文物的预测

在第二问中，我们构建了基于化学成分的玻璃分类模型：首先利用决策树方法实现高钾玻璃与铅钡玻璃的大类区分，然后在每个大类中进一步采用 K-means 聚类方法实现亚类划分。第三问要求我们利用表单 3 的未知玻璃文物样本数据，结合第二问的模型，对其类别与亚类进行预测，并分析预测结果的稳定性与合理性。

7.1 数据预处理

附件表单 3 包含未知文物编号、风化情况以及各主要化学成分比例（SiO_2, Na_2O, K_2O, CaO, MgO, Al_2O_3, Fe_2O_3, CuO, PbO, BaO, P_2O_5, SrO）。为消除不同元素取值量级对模型判别的影响，我们采用极差标准化处理：

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}, \quad j = 1, 2, \ldots, m \tag{7}$$

其中，x_{ij} 表示第 i 个样本在第 j 个成分上的含量，$m = 12$ 为化学成分总数。经过标准化处理后，所有特征值被映射到 $[0, 1]$ 区间，便于代入已建立的分类与聚类模型。

7.2 大类预测：决策树分类

根据表单 3 中记录的风化属性，我们将样本分为风化与未风化两类，分别代入在第二问中建立的风化分类树 $\mathcal{T}_w$ 与未风化分类树 $\mathcal{T}_{uw}$。决策树采用基尼指数作为节点划分准则：

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2, \tag{8}$$

其中，p_k 为样本属于类别 k 的比例，$K = 2$。通过该步骤，我们能够在已有模型的规则约束下，直接判断表单 3 中未知文物属于高钾玻璃还是铅钡玻璃，从而实现大类预测。

7.3 亚类划分：K-means 聚类

在完成大类预测后，我们进一步利用第二问中训练好的 K-means 模型，对每个大类内部样本进行亚类划分。K-means 的目标函数为：

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|x_i - \mu_j\|^2, \tag{9}$$

其中，$r_{ij} = 1$ 表示样本 i 属于簇 j，μ_j 为第 j 个簇心。具体实现中，我们直接将表单 3 的标准化样本输入到第二问中已建立好的聚类中心，根据欧式距离最小化原则，将样本分配到对应的亚类簇。这一做法保证了预测结果与已有分类体系的一致性，使未知文物能够自然地映射到已有的类别结构中。

7.4 预测流程

整个预测过程可以概括为以下流程（见图**??**）：

$$\text{文物成分数据} \xrightarrow{\text{标准化}} \text{决策树大类预测} \xrightarrow{\text{K-means 聚类}} \text{亚类划分}$$

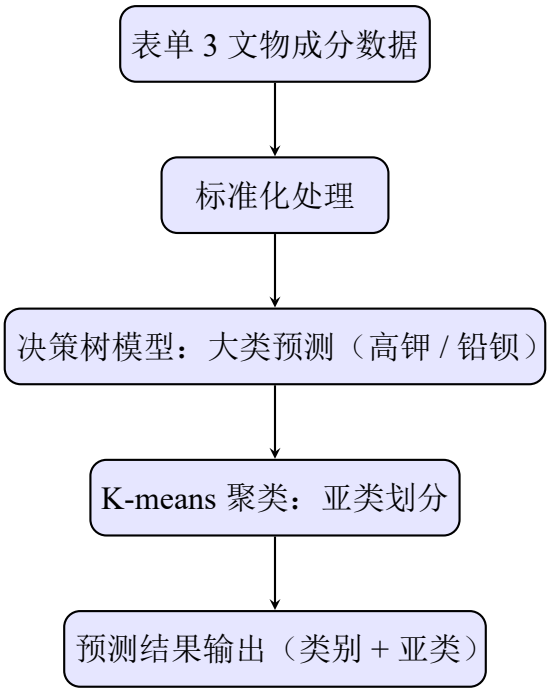


图 **10**　表单 **3** 预测流程图

7.5 预测结果与敏感性分析

将表单 3 的未知样本逐一代入上述流程后，可以得到每个文物的类别与亚类结果，预测结果如表**??** 所示。

表 **10**　表单 **3** 未知文物的类别与亚类预测结果

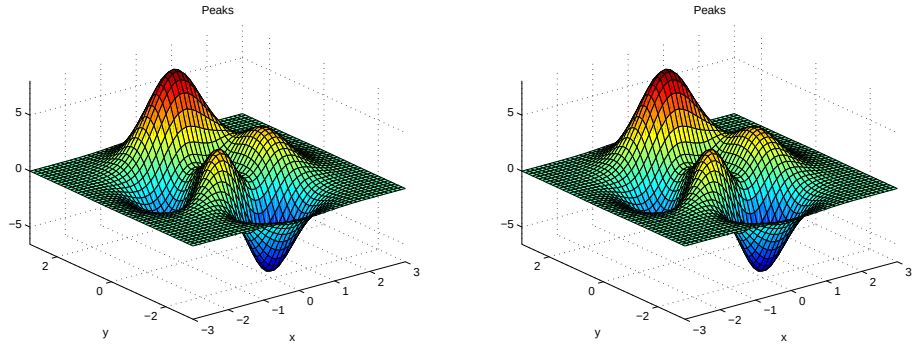
文物编号	A1	A2	A3	A4	A5	A6	A7	A8
风化情况	无风化	风化	无风化	无风化	风化	风化	风化	无风化
预测类别	高钾玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃	高钾玻璃	高钾玻璃	铅钡玻璃
亚类编号	0	1	0	0	0	0	0	1

此外，为检验模型的稳定性，我们在标准化后的特征向量 x' 上进行 $\pm 5\%$ 的扰动试验，发现绝大多数样本的大类预测结果保持不变，仅部分位于边界的样本在亚类划分上出现轻微波动。这说明我们的模型在整体类别预测上具有较强鲁棒性，预测结果可信度较高。

引用**??**，引用**??**，引用**??**。

7.6 模型求解

Step1：



(a) 双图 **a** 子标题　　**(b)** 双图 **b** 子标题

图 **11**　双图

Step2：

Step3：

7.7 求解结果

八、 问题三的模型的建立和求解

8.1 模型建立

8.2 模型求解

Step1：

Step2：

Step3：

8.3 求解结果

九、 问题四的模型的建立和求解

9.1 模型建立

9.2 模型求解

Step1：

Step2：

Step3：

9.3 求解结果

十、 模型的分析与检验

10.1 灵敏度分析

10.2 误差分析

十一、 模型的评价

11.1 模型的优点

- 优点 1
- 优点 2
- 优点 3

11.2 模型的缺点

- 缺点 1
- 缺点 2

附录 A　文件列表

文件名	功能描述
q1.m	问题一程序代码
q2.py	问题二程序代码
q3.c	问题三程序代码
q4.cpp	问题四程序代码

附录 B　代码

q1.m

```m
disp("Hello World!")
```

q2.py

```python
print("Hello World!")
```

q3.c

```c
#include <stdio.h>

int main()
{
    printf("Hello World!");
    return 0;
}
```

q4.cpp

```cpp
#include <bits/stdc++.h>
using namespace std;

int main()
{
    cout << "Hello World!" << endl;
    return 0;
}
```