

基于二分类模型的玻璃制品的成分分析与鉴别

摘要

对古代玻璃制品的成分分析与鉴别是考古工作中重要的部分。本文首先建立二元 logistic 回归模型用于分析玻璃成分规律和预测玻璃风化前的成分，然后建立基于单因素回归分析的系统聚类模型用于鉴别玻璃样品的类型，最后使用皮尔逊相关系数分析玻璃样品成分的关联性，用斯皮尔曼相关系数分析关联性的差异性，还给出了模型的优化方法。

针对问题一，首先考虑到需要分析的是分类数据之间的关系，本文采用卡方检验分别进行分析，得出仅玻璃类型对表面风化有影响，铅钡玻璃更易风化。为了研究风化与化学成分含量的关系，本文采用二元 logistic 回归模型，先使用单因素分析筛选与是否风化显著相关的化学成分，再分析高钾玻璃和铅钡玻璃在风化前后化学成分的变化规律，最后通过回判分析和 H-L 拟合度对回归方程进行了检验。并且建立基于 Adaboost 改进的风化判别模型，在 logistic 模型的比较中表现出高精度的优点。接着使用得到的 logistic 回归方程预测风化前的化学成分含量。

针对问题二，使用类似问题一的方法建立二元 logistic 回归模型，通过单因素分析和经验分析筛选出显著性较强的八个化学成分，在风化和未风化两种情况下分别比较其含量，得到玻璃类型与化学成分的规律，如高钾玻璃中氧化钾含量高。然后根据筛选出的八个化学成分对风化前后的两种玻璃类型分别进行建立系统聚类模型，通过观察聚合系数和聚类个数的图像确定最优聚合个数，从而得到玻璃亚型的 11 个分类。最后根据玻璃的制造原理和 ANOVA 方差分析进行合理性检验，定义“步进距离”定量分析模型中的灵敏度，得出该聚类模型对二氧化硅最敏感。

针对问题三，本文使用问题二建立的二元 logistic 回归模型对样品的类型进行鉴别，得出样品 A1,A6,A7 是高钾玻璃，A2,A3,A4,A5,A8 是铅钡玻璃。然后根据指标的 Odds Ratio 值分析模型的灵敏度，得到该模型对氧化锆的含量最敏感。最后，还用 Fisher 判别法进行鉴别结果的检验，得到的结果与 logistic 回归模型一致。

针对问题四，本文首先使用皮尔逊相关系数分析数据呈正态分布的化学成分的相关性。在比较不同类别的玻璃文物样品的化学成分关联的差异性时，为了获得化学成分含量两两之间的相关性，使用斯皮尔曼相关系数得到两种玻璃类型的化学成分相关系数矩阵，作差分析得出有显著差异的化学成分关系：对于未风化的玻璃，氧化钾与氧化镁的关系在高钾玻璃中成负相关，在铅钡玻璃中呈正相关；对于风化的玻璃，二氧化硅与氧化铝的关系在高钾玻璃中成负相关，在铅钡玻璃中呈正相关。

关键词：二元 logistic 回归模型；Adaboost；单因素分析；系统聚类模型；相关系数

一 问题背景与重述

在西方文化沟通的桥梁——丝绸之路中，玻璃是双方贸易往来的宝贵物证。我国的古代玻璃在吸收外来玻璃技术后，使用本土材料制作而成，所以与外来的玻璃制品相比，外观相似而化学成分不同。

古代玻璃在风化的过程中，由于受到环境元素的影响，成分比例发生变化，从而其玻璃类型将会难以判断。标记为表面无风化或者风化的文物，都存在局部风化区和未风化区。

现有我国古代玻璃制品的相关数据，考古工作者已根据其化学成分和其他的检测手段将其分为两种类型：高钾玻璃和铅钡玻璃。

本文需要根据附件的数据分析建模，解决如下四个问题：

- 分析玻璃文物的表面风化情况与玻璃的类型、纹饰、颜色之间的关系；结合玻璃类型分析化学成分含量对文物样品表面有无风化情况的规律；根据风化点的化学成分检测数据对玻璃风化前的化学成分含量进行预测。
- 依据附件中的数据，分析高钾、铅钡两种玻璃类型的分类规律；对于每种玻璃的类别，挑选合适的化学成分指标进行亚类划分，给出划分方法和结果并分析结果的合理性和敏感性。
- 分析附件 3 未知类别的玻璃文物的化学成分，对其所属类型进行鉴别并分析分类结果的敏感性。
- 分析不同类别的玻璃样品的化学成分的关联性；比较不同类别的化学成分关联的差异性。

二 模型假设

为了建立模型，我们提出如下的假设。

1. 假设题中所给数据真实可靠，化学成分测量的误差对结果的影响可以忽略。
2. 将严重风化点划入风化类别中，不做额外分类。
3. 假设考古工作者的分类手段是具有权威性的，分类结果是准确的。
4. 假设检测设备没有出现故障，未检测到的成分确实不存在于玻璃制品中。
5. 忽略风化环境对风化结果的影响。
6. 假设玻璃的颜色是由额外的加工工艺决定的，与玻璃类型无关。

三 符号说明

所使用的符号及说明如表3.1所示。

表 3.1: 符号说明

符号	说明
F	风化情况 (0 表示未风化, 1 表示风化)
T	玻璃类型 (0 表示铅钡玻璃, 1 表示高钾玻璃)
P	风化概率
L	似然函数
x_i	化学成分含量
β_i	各项回归系数
D	距离矩阵
d_{ij}	聚类空间中两点间的欧式距离
r	皮尔逊相关系数
ρ	斯皮尔曼相关系数

四 模型的建立与求解

4.1 数据预处理

4.1.1 缺失值的处理

对于表单 1 颜色缺失的样品, 由于样品量过少, 颜色无法预测, 故剔除表单 1 中颜色存在缺失值的样品, 即 19,40,48,58 号样品。

对于表单 2 中的缺失值, 化学成分缺失的原因是未检测到该成分, 故其成分比例均视作为 0%。

4.1.2 异常值的处理

对于表单 2 中的异常值, 根据题中所给的有效数据的定义, 15 和 17 号样品的各成分比例的累加和分别为 79.47% 和 71.89%, 均小于 85%, 不属于有效数据, 因此在后文的分析中均剔除 15 和 17 号样品。

4.1.3 特殊情形的处理

对于从风化样品中未风化部分取样这种特殊情况, 考虑到同一样品不同部位之间的化学成分存在差异, 本文的处理方式是视作未风化的新样品。

对于从风化样品中的两个风化部分取样的情况, 考虑到两个取样点化学成分含量差异不大, 本文以平均数作为该样品的化学成分含量。

4.2 问题一

表单 1 的数据给出了 58 种文物的纹饰、类型、颜色和表面风化四项数据，首先要分析前三项数据与表面风化的相关性。由于四项数据都是分类数据，本文采用卡方检验分别分析前三项数据与表面风化的关系。

该问还要求分析化学成分含量对文物样品表面有无风化的统计学规律，这实际上是对分类变量和数值变量的关系分析。由于是否风化是分类变量且是二值变量，本文采用二元 logistic 回归模型分析其规律。最后，使用得到的 logistic 回归方程预测风化前的化学成分含量。

4.2.1 基于卡方检验的风化影响因素的分析

Step1: 通过交叉图分析，初步判断是否存在相关关系。

为了分别研究玻璃类型、纹饰和颜色三个因素对玻璃文物的表面风化的差异关系，首先分别作出玻璃类型、纹饰和颜色与表面风化的交叉图。

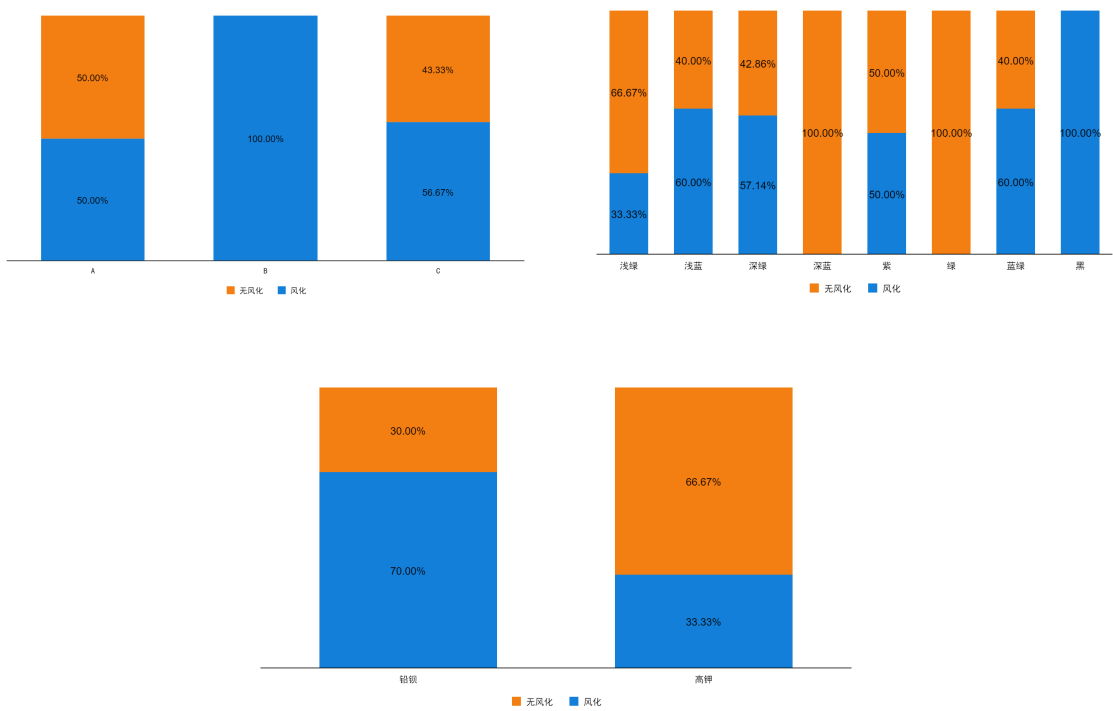


图 4.1: 玻璃类型、纹饰和颜色与表面风化的交叉图

由图4.1可以看出：

- 1. 类型与表面风化有较大的关系，铅钡玻璃风化的概率大，高钾玻璃风化的概率小。
- 2. 纹饰 A 和 C 的风化都有一半的几率风化，也就是 A 和 C 对是否风化没有影响，纹饰 B 虽然都风化了，但考虑到个数只有六个，具有一定的偶然性，无法推断出纹饰与表面风化具有相关性。

3. 颜色方面，除了少数几种颜色如深蓝、绿色由于样本量过小的缘故导致的大幅度比例差，各种颜色风化概率基本一致。

Step2: 卡方分析

表 4.2: 类型、纹饰和颜色与是否发生风化的卡方分析

		是否风化		总计	χ^2	p
		未风化	风化			
纹饰	A	11(45.83)	11(32.35)	22(37.93)	4.957	0.084
	B	0(0.00)	6(17.65)	6(10.34)		
	C	13(54.17)	17(50.00)	30(51.72)		
类型	铅钡	12(50.00)	28(82.35)	40(68.97)	6.88	0.009
	高钾	12(50.00)	6(17.65)	18(31.03)		
颜色	浅绿	2(8.33)	1(3.33)	3(5.56)	6.287	0.507
	浅蓝	8(33.33)	12(40.00)	20(37.04)		
	深绿	3(12.50)	4(13.33)	7(12.96)		
	深蓝	2(8.33)	0(0.00)	2(3.70)		
	紫	2(8.33)	2(6.67)	4(7.41)		
	绿	1(4.17)	0(0.00)	1(1.85)		
	蓝绿	6(25.00)	9(30.00)	15(27.78)		
	黑	0(0.00)	2(6.67)	2(3.70)		

从表4.2可以看出玻璃类型、纹饰和颜色的 p 值分别为 0.009、0.084 和 0.507，只有玻璃类型的 p 值小于 0.05，而颜色和纹饰的 p 值大于 0.05，因此在 0.05 的显著性水平下，只有玻璃类型对表面有无风化表现出显著性差异。

Step3: 得出结论

根据得到的三个 p 值，在玻璃类型、纹饰和颜色三个因素中，只有玻璃类型对于表面风化会表现出显著性差异，其中铅钡玻璃更容易风化。

由于不同类型的玻璃化学成分含量不同，而不同的化学元素在风化过程中的迁移能力也有所不同，所以玻璃类型与玻璃受风化的程度存在一定的关系；玻璃的纹饰和颜色仅改变玻璃的浅表形态与颜料种类，在风化环境中对与玻璃表面受风化程度的影响较小，这与卡方分析的结果相符合^[2]。

4.2.2 基于二元 logistic 回归模型的风化判别模型

结合玻璃的类型，对同种玻璃类型的风化和未风化两种玻璃的化学成分含量进行分析，研究化学成分含量对有无风化的影响，即何种化学成分含量与样品表面有无风化有显著相关性以及其如何对有无风化产生显著性差异。

为了证明对于同种类型的玻璃，表面有无风化的两种样品的化学成分含量确实有显著性差异，首先对数据进行初步的整理与分析。

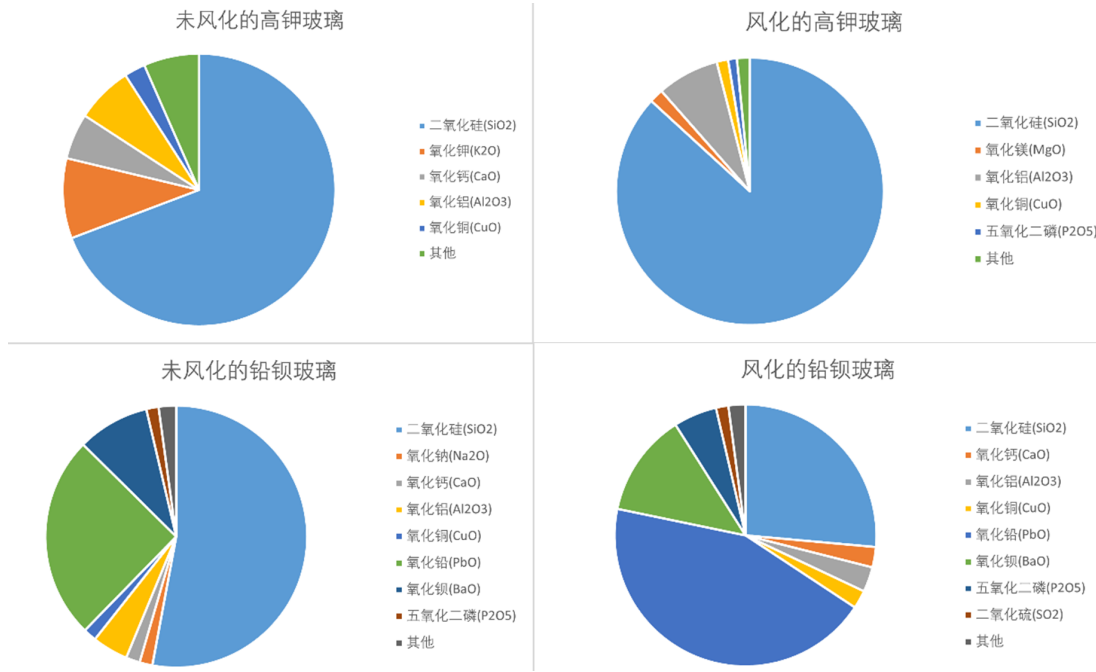


图 4.2: 四种类型玻璃化学成分含量饼图

由图4.2可以看出：

1. 无论风化与否，高钾玻璃中二氧化硅 (SiO₂) 含量均高于铅钡玻璃，且其高钾玻璃含量很高。
2. 无论风化与否，铅钡玻璃中氧化铅 (PbO) 和氧化钡 (BaO) 含量都很高，而在高钾玻璃中含量几乎为 0。
3. 对于未风化的高钾玻璃，其氧化钾 (K₂O) 含量远高于未风化的铅钡玻璃，但风化后在两者中含量基本一致。

下面建立二元 logistic 回归模型分析高钾玻璃与铅钡玻璃的分类规律。

(1) 模型的建立

玻璃的风化情况 (F) 有两种情形：未分化 (0) 和分化 (1)。

$$F = \begin{cases} 0 & , \text{未风化} \\ 1 & , \text{风化} \end{cases} \quad (4.1)$$

影响 D 取值的 14 个化学成分含量自变量为 x_1, x_2, \dots, x_{14} 。记 $P(Y = 1|x_1, x_2, \dots, x_{14})$ 表示在 14 个自变量作用下风化的概率，二元 logistic 模型可表示为：

$$P = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{14} x_{14} + \mu)]} \quad (4.2)$$

其中 β_0 为常数项, $\beta_1, \beta_2, \dots, \beta_{14}$ 为回归系数。若用 z 表示 14 个化学成分含量的线性组合:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{14} x_{14} \quad (4.3)$$

作对数变换的得 logistic 回归模型的如下线性形式:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{14} x_{14} \quad (4.4)$$

$\ln\left(\frac{P}{1-P}\right)$ 为风化与未风化的发生概率之比的自然对数, 称为 P 的 logistic 变换, 用 logistic P 表示。虽然在 0-1 之间, logistic 却没有数值界限。

在根据实际数据估计模型参数时, 通常采用似然估计建立一个样本似然函数:

$$L = \prod_{i=1}^n P_i^{Y_i} (1 - P_i)^{1-Y_i} \quad (4.5)$$

其中 P_i 表示第 i 个样品风化的概率, 如果实际出现了风化, 取 $Y_i = 1$, 否则 $Y_i = 0$ 。根据最大似然原理, 似然函数 L 应达到最大值, 取 L 的对数形式:

$$\ln L = \sum_{i=1}^n [Y_i \ln P_i + (1 - Y_i) \ln (1 - P_i)] \quad (4.6)$$

即为目标函数。然后使用 Newton-Raphson 迭代方法使似然函数达到极限, 此时参数的取值即为 $\beta_1, \beta_2, \dots, \beta_{14}$ 的最大似然估计值。

(2) 模型的求解

理论上在样本足够多的情况下, 应把 14 个自变量都放到似然方程中, 以考虑所有的因素。但自变量之间可能存在一定的关联性, 所以样本量与变量之比最好为 10:1-5:1^[4]。考虑到本题表中给出的样品数仅有 58 个, 本文利用逐步回归的思想, 先进行单因素分析, 筛选出具有显著性的变量放入方程进行多因素分析。

首先对高钾玻璃样品进行分析。单因素分析的结果如表4.3所示:

表 4.3: 高钾玻璃样品进行单因素分析的 p 值

化学成分	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铜
p	0.184	0.999	0.173	0.087	0.075	0.052	0.299
化学成分	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫	氧化铁
p	1.000	0.999	0.069	0.983	1.000	0.999	0.197

高钾样品共 18 个, 为满足例数与变量 10:1-5:1 的较佳比例, 取置信区间为 0.925, 则 $p \leq 0.075$, 有意义的化学成分为氧化镁、氧化铝和五氧化二磷。

通过 SPSS 软件进行二元 logistic 回归分析, 得到结果如表4.4所示:

表 4.4: 高钾玻璃的二元 logistic 回归分析结果

项	氧化镁	氧化铝	五氧化二磷	常量
回归系数	1.983	-1.293	-4.838	5.143

可以得到高钾玻璃的二元 logistic 回归方程为:

$$\ln \left(\frac{P}{1-P} \right) = 1.983x_5 - 1.293x_6 - 4.838x_{10} + 5.143 \quad (4.7)$$

采用同样的方法对铅钡玻璃样品进行分析, 由于铅钡玻璃有 49 个样品, 取置信区间为 0.965, 则 $p \leq 0.035$ 的有意义的化学成分为氧化钠、二氧化硅、氧化铅、五氧化二磷四个。进行二元 logistic 回归分析结果如下:

表 4.5: 铅钡玻璃的二元 logistic 回归分析结果

项	二氧化硅	氧化钠	氧化铅	五氧化二磷	常量
回归系数	0.107	-0.083	0.045	0.107	1.245

得到高钾玻璃的二元 logistic 回归方程为:

$$\ln \left(\frac{P}{1-P} \right) = -0.085x_2 - 0.083x_8 + 0.045x_{10} + 0.107x_1 + 1.245 \quad (4.8)$$

(3) 模型的检验

1. 回判分析是常用的检验分类结果准确性的方法。通过将样本数据代回到回归方程, 若回代准确率较高, 说明分类模型较合理。

2. H-L 拟合度检验可以用于判断拟合模型的优度。本文 H-L 检验的原假设为: 模型拟合值和观测值的吻合, 如 p 值 >0.05 , 则说明通过 H-L 检验。

模型检验的结果如下表所示:

表 4.6: logistic 模型的回判准确率和 H-L 检验的 p 值

类型	回判准确率	H-L 检验 p 值
高钾玻璃	88.89%	0.808
铅钡玻璃	87.76%	0.22

可以看出两次 logistic 回归模型的回判准确率都较高, 且 H-L 检验的 p 值都 >0.05 , 说明模型通过 H-L 检验, 拟合较优。

4.2.3 基于 Adaboost 改进的风化判别模型

Boosting 是一种集成学习技术, 能够将弱学习器增强为预测精度高的强学习器。而 adaboost 算法不要求预知学习算法的先验知识, 是一种较优的二分类 Boosting 算法, 也

可以用于解决本文的问题。因此可以考虑采用 **adaboost** 算法作为 **logistic** 的比较，验证 **logistic** 结果的正确性。

Adaboost 的步骤分为输入和输出。在分类和预测问题中，要输入的有训练样本集，迭代次数 T 和弱分类器 *Weaklearn*，训练的步骤如下：

Step1: 初始化样本权值 β_i

Step2: 对于 $t = 1, 2, \dots, T$ ，每次迭代产生分布 $p^t = \frac{x_i}{\sum_{i=1}^m x_i^t}$ ，基于 p^t 调用 *Weaklearn*

得到假设 $h_t : X \rightarrow [0, 1]$

Step3: 计算假设的错误率

$$\epsilon_t = \sum_{i=1}^m p_i^t |h_t(x_i) - y_i| \quad (4.9)$$

Step4: 更新权重

$$\beta_i^{t+1} = \frac{\beta_i^t \epsilon_t^{1-|h_t(s_{i1}-s_{i2})|}}{1 - \epsilon_t} \quad (4.10)$$

$$\text{输出的 } H_f(s_i) = \begin{cases} 1, & \sum_{t=1}^T \left(\log \frac{1}{\beta_t} \right) h_t(s_i) \geq \frac{1}{2} \sum_{t=1}^T \left(\log \frac{1}{\beta_t} \right) \\ 0, & \sum_{t=1}^T \left(\log \frac{1}{\beta_t} \right) h_t(s_i) < \frac{1}{2} \sum_{t=1}^T \left(\log \frac{1}{\beta_t} \right) \end{cases}$$

在问题三中，对应的结果为： $\begin{cases} \text{风化, } H_f(s_i) = 1 \\ \text{未风化, } H_f(s_i) = 0 \end{cases}$

分别对高钾玻璃与铅钡玻璃的风化情况进行判断，结果如下：

表 4.7: **Adaboost** 改进的风化判别模型的回判准确率

类型	准确率
高钾玻璃	100%
铅钡玻璃	97.9%

当迭代次数足够时，**Adaboost** 分类器准确率基本都能达到 100%，比 **logistic** 回归模型更加精确。

4.2.4 基于二元 **logistic** 回归的样本成分预测模型

先对已有的未风化的样品数据进行分析，计算它们的 **logistic** 目标值，发现集中在 0.01-0.04 之间，故本文将风化前样品的 **logistic** 目标值取为 0.01-0.04 之间的随机值。又考虑到每一样品的风化环境相同，则可得到：

$$\ln\left(\frac{P}{1-P}\right)' \in (0.01, 0.04)$$

$$\Delta x_i = \left[\ln\left(\frac{P}{1-P}\right)' - \ln\left(\frac{P}{1-P}\right) \right] \frac{\beta_i}{\sum_{i=1}^m \beta_i} \quad (4.11)$$

那么某一化学成分的预测值就可以用风化后的值与 Δx_i 之和来表示：

$$x'_i = x_i + \Delta x_i \quad (4.12)$$

使用 matlab 编程（代码见附录2.1）对风化的高钾玻璃的化学成分进行预测，得到的结果如表4.8所示：

表 4.8: 高钾玻璃的化学成分预测

编号	氧化镁	氧化铝	五氧化二磷
7	1.54	7.79	1.43
9	1.72	7.83	1.26
10	1.76	7.47	0.94
12	1.70	7.85	1.05
22	2.21	9.43	1.04
27	2.08	8.33	1.18

使用同样的方法对风化的铅钡玻璃进行化学成分预测，得到的结果见附录1.1。

4.3 问题二

问题二首先仍然是要求找出两种玻璃类型的分类规律，考虑到玻璃类型与玻璃的化学成分存在一定的相关性，依然可以使用二元 logistic 回归分析类型与化学成分的关系，以寻找玻璃分类的规律。

接下来要求选择合适的化学成分进行亚类划分，考虑到化学成分之间可能存在一定的相关性，二元 logistic 回归的单因素分析可以得到与玻璃类型典型相关的化学成分，接着本文采用系统聚类的方法对玻璃样品进行分类，通过观察聚合系数和聚类个数的关系图得到最优聚合个数，从而得到不同的玻璃亚型分类。

在对问题进行检验时，本文首先观察到文中挑选的典型相关的化学成分中存在氧化铅、氧化钡、氧化钾等在化学分类原理中与玻璃类型明显相关的化学成分，然后使用方差分析检验聚类模型的合理性。

最后，通过计算增加一定的化学成分含量对该样品点到聚类中心的距离的影响程度，定量表示该化学成分的灵敏度。距离的改变越大，该化学成分的灵敏度就越高。

4.3.1 基于 logistic 单因素分析模型的玻璃类型分类规律

在问题一的基础上，对二元 logistic 回归模型进行修改。
因变量 (T) 取值为：

$$T = \begin{cases} 1 & , \text{高钾玻璃} \\ 0 & , \text{铅钡玻璃} \end{cases} \tag{4.13}$$

影响 T 取值的 14 个化学成分含量自变量为仍 x_1, x_2, \dots, x_{14} 。logistic P 仍可表示为：

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_{14}x_{14} \tag{4.14}$$

对两种玻璃样品进行单因素分析结果如下：

表 4.9: 玻璃样品进行单因素分析的 p 值							
化学成分	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铜
p	0.002	0.343	0.039	0.008	0.435	0.111	0.654
化学成分	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫	氧化铁
p	0.990	0.001	0.038	0.001	0.452	0.509	0.042

取置信区间为 0.95，p 值 ≤ 0.05 的化学成分：二氧化硅、氧化钾、氧化钡、氧化锶、氧化钙、氧化铁、五氧化二磷是与玻璃类型具有显著相关性的七种成分。

需要注意的是，氧化铅的 p 值出现了异常增高的情况。观察氧化铅的数据可知，所有铅钡玻璃的氧化铅数据都高于高钾玻璃的氧化铅数值。在这种完美分离的情况下，不存在最大似然距离，导致了 0.99 的极高的 p 值。由此，本文对单因素分析结果进行人工干预，将氧化铅同样列入具有显著相关性的化学成分。

比较两种玻璃之间这八种化学成分的含量关系，得到的详细数据比较见附录1.2。结论是在未风化的情况下，高钾玻璃中的氧化钾远高于铅钡玻璃，氧化铅和氧化钡远低于铅钡玻璃。而在风化的情况下，除了以上三种化学成分的规律外，还可以得到高钾玻璃中的二氧化硅远高于铅钡玻璃，五氧化二磷和氧化锶远低于铅钡玻璃。

4.3.2 基于系统聚类模型的玻璃亚类划分

(1) 系统聚类法

考虑到风化对化学成分含量的影响，本文分别对未风化高钾玻璃、风化高钾玻璃、未风化铅钡玻璃和风华铅钡玻璃进行内部样品的聚类。各组样品量分别为 10,6,24,21，指标为单因素分析得到的 8 种化学成分含量。

系统聚类的步骤如下：

Step1: 数据预处理

考虑到每种化学成分含量的量纲相同，但数值变化的范围不同，如果直接进行聚类，容易出现“大数吃小数”的情况，使得聚类结果不准确。因此，本文首先对数据进行标准化预处理。每种化学成分的标准方差为

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(x_{ij} - \frac{1}{n} \sum_{i=1}^n x_{ij} \right)^2} \quad (4.15)$$

则标准化后的数据为：

$$x'_{ij} = \frac{x_{ij} - \frac{1}{n} \sum_{i=1}^n x_{ij}}{s_j} \quad (4.16)$$

其中 n 为需要聚类的样品个数， x_{ij} 为第 i 个样品的第 j 个化学成分含量。

Step2: 每个样品独自视为一类，计算各样品间的距离矩阵。

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots & d_{2n} \\ d_{31} & d_{32} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & 0 \end{bmatrix} \quad (4.17)$$

其中 d_{ij} 为第 i 个样品与第 j 个样品之间的欧氏距离。

Step3: 将距离最小的两个样本合并成新的类，计算此时新类与其余类之间的距离矩阵。

Step4: 重复直至全部样品被并为一类。

聚类结束后，本文使用肘部法则估计最优的聚合数量。定义所有类的总畸变程度（聚合系数） J 为：

$$J = \sum_{k=1}^K \sum_{i \in C_k} |x_i - u_k| \quad (4.18)$$

其中 C_k 表示第 k 类， x_i 表示某一点在空间中的位置， u_k 表示该类的重心位置。聚合系数越小，聚类效果越好。作出聚合系数与聚合种类的关系图，考虑到图像转折点之后的聚类效果提升程度较小，可以得到转折点处的聚类个数就是较优的聚类个数。

(2) 系统聚类模型的求解

首先将每种玻璃根据是否分化进行分类，得到高钾风化玻璃、高钾未风化玻璃、铅钡风化玻璃、铅钡未风化玻璃四种类型。再对每一种类型进行关于八种具有显著相关性的化学成分的系統聚类。

对四种类型的玻璃分别使用 SPSS 对进行系統聚类分析，聚类结果（以高钾未风化玻璃为例）如下：

表 4.10: 高钾未风化玻璃的聚合系数与聚合种类

聚类个数	1	2	3	4	5	6	7	8	9
聚合系数	261.08	106.67	78.07	52.05	45.41	31.52	15.13	12.89	8.44

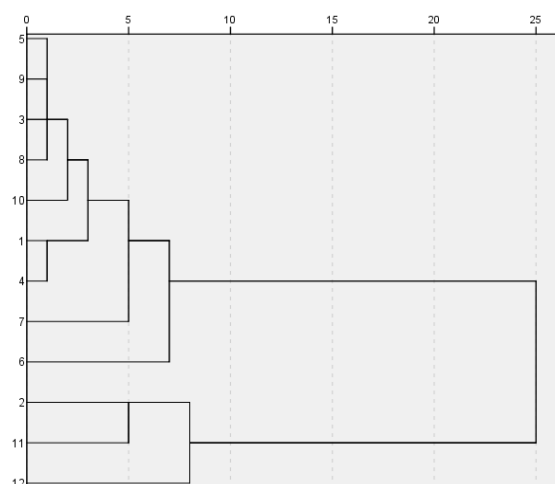


图 4.3: 高钾未风化玻璃的聚类谱系图

将得到的四类玻璃的聚合系数与聚合种类数量以折线图的方式展现。

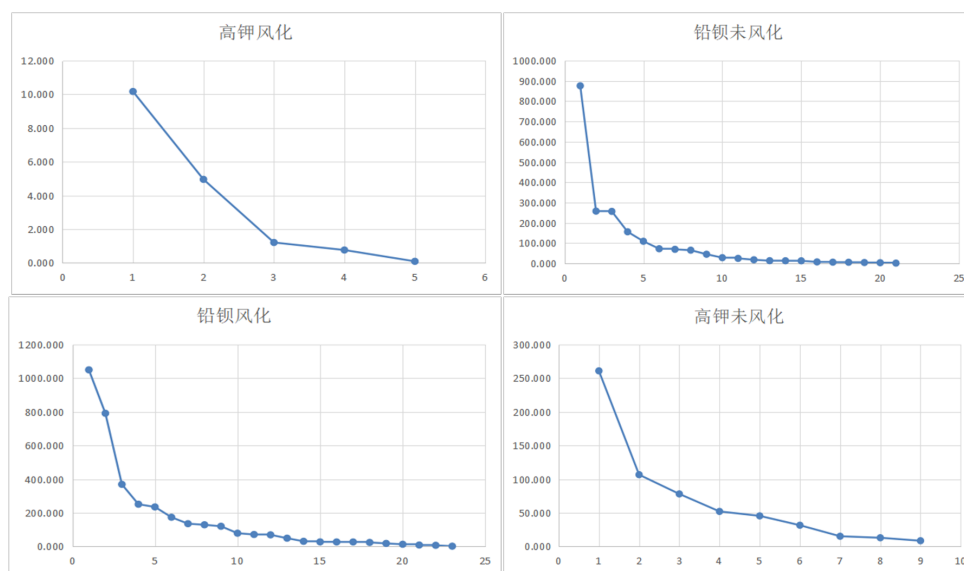


图 4.4: 四种玻璃聚合系数与聚合种类关系图

由图可以看出高钾未风化玻璃的图像在聚合种类为 2 处发生转折，根据肘部法则，聚合最佳结果是分为两类。同理得到高钾风化聚为 3 类、铅钡风化玻璃聚为 4 类、铅钡未风化玻璃聚为 2 类。再根据聚类谱系图（见附录 1.3）可以得出玻璃的亚型分类：

图 4.5: 玻璃样品亚型的聚类结果

玻璃类型		所属样品编号	亚类编号
高钾	风化	7, 22, 27	1
		9, 12	2
		10	3
	未风化	1, 4, 5, 6, 13, 14, 16	4
		3, 18, 21	5
铅钡	风化	34, 38, 36, 2, 19, 58, 49, 56, 57, 41, 50, 52, 51, 11	6
		39, 40, 54	7
		8, 26, 43	8
		48	9
	未风化	20, 24, 30	10
		47, 42, 23, 46, 25, 55, 49, 50, 37, 45, 53, 29, 44, 28, 32, 35, 31, 33, 20, 24, 30	11

4.3.3 基于方差分析和化学原理的亚型分类的合理性分析

1. 根据玻璃的原料，铅钡玻璃的制作过程中加入了铅矿石，氧化铅、氧化钡的含量较高；高钾玻璃的助熔剂是含钾量高的草木灰，氧化钾的含量较高。而本文的聚类模型使用的八种化学成分中包含了氧化铅、氧化钡和氧化钾，与玻璃制作的化学原理相符合。

2. 对聚类结果进行 ANOVA 方差检验。以高钾未风化的 4、5 两个亚类为例，对他们的化学成分进行方差分析的结果如下：

表 4.11: 4、5 亚类的方差分析结果

化学成分	二氧化硅	氧化钾	氧化钙	氧化铁	氧化钡	氧化铅	氧化锶	五氧化二磷
p	0.000	0.059	0.026	0.342	0.06	0.445	0.784	0.832

其中二氧化硅和氧化钙、氧化钡的 p 值小于 0.05，在 4、5 亚类中存在显著性差异，可见区分 4、5 亚类的分类依据主要二氧化硅和氧化钙、氧化钡的含量差异，因此聚类结果具有一定的合理性和可解释性。

同样对铅钡未风化的 10、11 两种亚类进行方差分析，可以看出二氧化硅、氧化钾、氧化钡在 10、11 亚类中存在显著性差异，是 10、11 亚类划分的主要依据。

表 4.12: 10、11 亚类的方差分析结果

化学成分	二氧化硅	氧化钾	氧化钙	氧化铁	氧化钡	氧化铅	氧化锶	五氧化二磷
p	0.001	0.034	0.244	0.443	0.02	0.442	0.323	0.088

4.3.4 玻璃分类的敏感性分析

本文通过化学成分的取值变化对玻璃分类结果的影响表示某种化学成分在模型中的敏感性。每个化学成分改变的值为该亚类中该化学成分的均值的 1%:

$$\Delta x_j = \frac{\sum_{i=1}^m x_{ij}}{m} \times 1\% \quad (4.19)$$

定义“步进距离” Δd 表示化学成分含量改变 Δx_i 时该样品点到聚类中心的距离变化量:

$$\Delta d = \sqrt{\sum (x_i - x_j - \Delta x_i)^2} - \sqrt{\sum (x_i - x_j)^2} \quad (4.20)$$

以第 58 号样品为例, 将其 8 项化学成分分别增加 $k\Delta x_{ij} (k = 1, 2, 3, 4, 5)$ 后得到新的亚类划分结果, 计算亚类之间的距离, 即亚类之间的差异性, 得到的结果如下表所示:

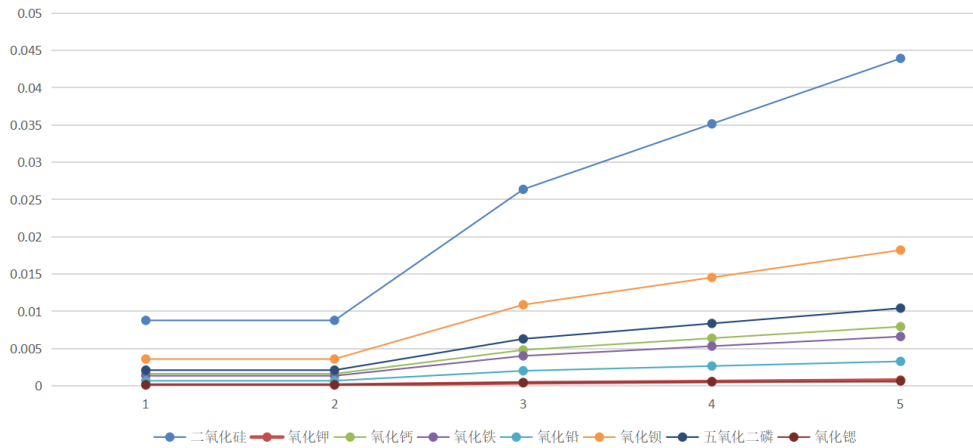


图 4.6: 化学成分变化对分类结果的影响

在逐步改变化学成分的过程中, 仅二氧化硅含量的改变对样品点与聚类中心之间的距离造成了较大的影响, 其余 7 种化学成分的改变对距离的影响程度都较小, 可见铅钡未风化玻璃的聚类模型仅对二氧化硅的敏感性较大, 对其余化学成分的依赖性不强。

4.4 问题三

本问要求鉴别不同的玻璃样品的类别。在问题一的 logistic 模型的基础上, 使用 logistic 回归分析鉴别样品的类型。

4.4.1 基于二元 logistic 回归模型的玻璃分类模型

根据需要分类样品的化学成分 x_i ，计算玻璃类型 Y_i 的两点分布概率：

$$\begin{cases} P(Y = 1|x) = F(x, \beta) \\ P(Y = 0|x) = 1 - F(x, \beta) \end{cases} \quad (4.21)$$

由于 $E(y | x) = 1 \times P(Y = 1 | x) + 0 \times P(Y = 0 | x) = P(Y = 1 | x)$ 将 \hat{Y} 理解为 $Y = 1$ 发生的概率。

$$\hat{Y}_i = P(Y_i = 1 | x) = S(x_i \beta) = \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \quad (4.22)$$

如 $\hat{Y} \geq 0.5$ ，认为 $Y = 1$ 发生，即样品属于高钾玻璃。

使用 matlab 编程得到的回归方程为：

$$\ln\left(\frac{P}{1-P}\right) = 1.842x_1 - 2.575x_3 + 0.1.793x_4 + 1.009x_8 + 1.714x_9 + 1.561x_{10} + 10.694x_{11} + 1.532x_{14} - 154.156 \quad (4.23)$$

结果如下表：

表 4.13: 高钾玻璃的化学成分预测

样本	风化类型	所属类别
A1	无风化	高钾
A2	风化	铅钡
A3	无风化	铅钡
A4	无风化	铅钡
A5	风化	铅钡
A6	风化	高钾
A7	风化	高钾
A8	无风化	铅钡

4.4.2 分类模型的灵敏度分析

SPSS 分析得到的 Odds Ratio 值如下表所示。Odds Ratio 是改变一单位的化学成分带来的 odds 值的变化率，Odds Ratio 值越大，该化学成分在模型中越敏感。

表 4.14: 不同化学成分变化的 Odds Ratio 值

化学成分	Odds Ratio 值
二氧化硅 (SiO ₂)	6.312
氧化钾 (K ₂ O)	0.076
氧化钙 (CaO)	6.009
氧化铅 (PbO)	2.744
氧化钡 (BaO)	5.553
五氧化二磷 (P ₂ O ₅)	4.766
氧化锶 (SrO)	44089.593
氧化铁 (Fe ₂ O ₃)	4.628

由表可以看出氧化锶 (SrO) 对分类结果的影响最大，而其余成分影响不大。

4.4.3 基于 Fisher 判别的玻璃分类结果检验

Fisher 判别的思想是从 k 个总体中抽取 p 个化学成分的样品观测数据，借助方差分析构造线性判别函数：

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_8 x_8 = c'X \quad (4.24)$$

此处系数 β 的确定原则是使得风化和未风化的玻璃样品之间的区别最大，内部的离差最小。对于新的玻璃样品，将其化学成分含量代入判别函数，求出的函数值与判别临界值进行比较即可确定样品的归类。

使用 SPSS 软件进行 Fisher 判别后，预测结果（见附录1.7）和使用 logistic 回归的结果完全符合，可见二元 logistic 回归的分类结果具有较高的可信度。

4.5 问题四

本问要求针对玻璃文物样品的不同类别，分析化学成分的关联性。在统计学中，Pearson 相关系数、Spearman 秩相关系数和 kendall 等级三种相关系数都可以表示两种变量之间的相关程度。考虑到部分变量之间的关联性较低，先使用需要正态分布的皮尔逊相关系数分析相关性。

在比较不同类别的玻璃文物样品的化学成分关联的差异性时，为了获得化学成分含量两两之间的相关性，使用斯皮尔曼相关系数得到两种玻璃类型的化学成分相关系数矩阵，做差分析其差异。

4.5.1 基于皮尔逊相关系数的化学成分关系模型

Step1: 正态性检验

皮尔逊相关系数用于处理的数据需要满足连续和服从正态分布的前提。本文需要对四种玻璃的化学成分含量数据进行正态分布检验。Shapiro-Wilk 检验适用于对小于 50 个样品的正态性检验，检验统计量为：

$$W = \frac{\left[\sum_{i=1}^n (a_i - \bar{a}) (X_{(i)} - \bar{X}) \right]^2}{\sum_{i=1}^n (a_i - \bar{a}) \sum_{i=1}^2 (X_{(i)} - \bar{X})^2} \quad (4.25)$$

以未风化高钾玻璃为例，正态性检验结果如下表所示：

表 4.15: 未风化高钾玻璃化学成分含量 S-W 检验结果

化学成分	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铜
p	0.384	0.000	0.073	0.234	0.545	0.715	0.805
化学成分	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫	氧化铁
p	0.003	0.000	0.002	0.025	0.452	0.000	0.000

二氧化硅、氧化钾、氧化钙、氧化镁、氧化铝、氧化铜、氧化锡没有呈现出显著性 ($p > 0.05$)，说明符合正态分布。

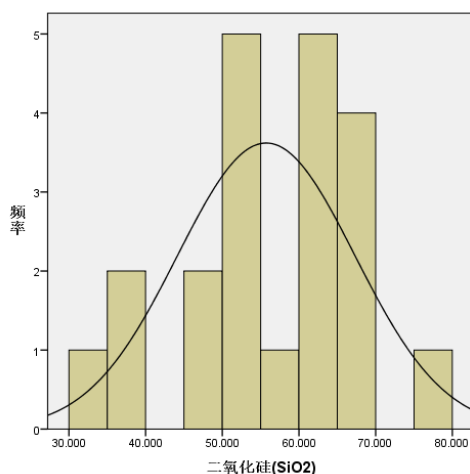


图 4.7: 二氧化硅的正态分布拟合

Step2: 作出散点图判断是否呈现线性关系

以二氧化硅与氧化钾为例，做出成分含量散点图。可见二者满足一定的线性关系。

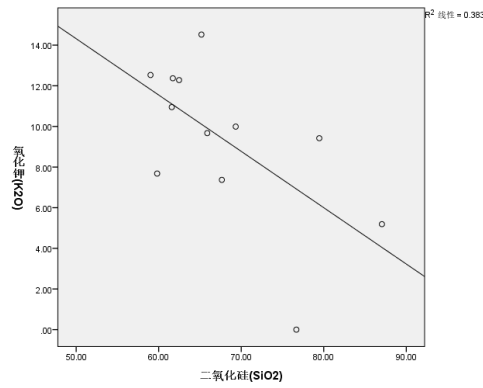


图 4.8: 二氧化硅和氧化钾的成分含量散点图

Step3: 计算各化学成分间的皮尔逊相关系数

皮尔逊相关系数用来衡量变量间的线性关系，具体计算公式如下：

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X}) (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (4.26)$$

式中 r 代表相关系数， n 为样本个数， X_i 与 Y_i 分别表示第 i 个样本的两组属性值。当 $r = 1$ 时，称 X, Y 完全相关，此时 X, Y 之间具有线性函数关系； $r > 0.8$ 时称为高度相关，当 $r < 0.3$ 时称为低度相关，其它时候为中度相关。

得到的皮尔逊相关系数如下：

	二氧化硅	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁	氧化铜
二氧化硅	1.000	-0.619	-0.678	-0.187	-0.558	-0.571	-0.473
氧化钾	-0.619	1.000	0.526	-0.206	-0.001	-0.157	0.053
氧化钙	-0.678	0.526	1.000	-0.352	0.081	0.185	0.348
氧化镁	-0.187	-0.206	-0.352	1.000	0.437	0.433	0.004
氧化铝	-0.558	-0.001	0.081	0.437	1.000	0.551	0.009
氧化铁	-0.571	-0.157	0.185	0.433	0.551	1.000	0.499
氧化铜	-0.473	0.053	0.348	0.004	0.009	0.499	1.000

图 4.9: 二氧化硅的正态分布拟合

采用同样的方法，得到各种玻璃符合正态分布的化学成分为：

表 4.16: 各类玻璃符合正态分布的化学成分

玻璃类型	符合正态分布的化学成分
高钾风化	二氧化硅、氧化钾、氧化钙、氧化铝、氧化铁、氧化铜、五氧化二磷
铅钡风化	二氧化硅、氧化钙、氧化铅、五氧化二磷、氧化锶
铅钡未风化	二氧化硅、氧化铅

分别得到的皮尔逊相关系数见附录1.4。

4.5.2 基于斯皮尔曼相关系数的成分关系差异性分析

斯皮尔曼相关系数是经过排行的两个随机变量的皮尔逊相关系数，表示为：

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (4.27)$$

其中 n 是样本的数量， d_i 表示 x_i 与 y_i 之间的等级差。

Step1: 计算斯皮尔曼相关系数

通过 matlab 计算风化的两种玻璃和未风化的两种玻璃的斯皮尔曼相关系数矩阵见附录1.5。

Step2: 求相关系数的差异性

将风化的两种玻璃和未风化的两种玻璃的斯皮尔曼相关系数矩阵作差，得到相关系数的差值，也即相关程度的差异性，结果见附录1.6。

Step3: 取出具有显著差异的相关系数

将大于均值的相关系数视作具有显著差异的化学成分关系。相关系数的均值为：

表 4.17: 不同类别玻璃之间化学成分斯皮尔曼相关系数之差的均值

两种玻璃类型	相关系数均值
铅钡风化与高钾风化	0.4536
铅钡未风化与高钾未分化	0.4378

Step4: 得出结论

具有显著差异的化学成分关系详见附录1.6。

对于高钾未风化与铅钡未风化的玻璃，氧化钾和氧化镁的相关关系的差异性最大，斯皮尔曼相关系数分别是-0.55 和 0.39，前者是强负相关，后者则是强正相关。

对于高钾风化与铅钡风化的玻璃，二氧化硅和氧化铝的相关关系的差异性最大，斯皮尔曼相关系数分别是-0.94 和 0.38，前者是强负相关，后者则是较强的正相关。

五 模型的优缺点分析

5.1 模型的优点

- 使用二元 logistic 回归模型分析是否风化和玻璃类型的规律，训练的速度较快，占用的内存资源小，算法成熟、预测准确，且模型的可解释性强，特征的权重可以直观展示不同特征对结果的影响权重。
- 在划分亚类时本文建立系统聚类模型，该模型算法简单，层次聚类的方法适用于任意形状的聚类并且对样本的输入顺序不敏感；在确定聚类个数时使用观察聚类个数与聚类系数的关系图象转折点的方法得到最优聚类个数，较好的解决了聚类终止条件具有不精确性的问题。

5.2 模型的缺点

- 由于 logistic 回归模型类似线性模型，形式较为简单，难以极为准确的拟合数据的真实分布情况。
- 由于系统聚类每一步都需计算类间的距离，在变量较多或样本量较大的情况下，运算速度可能较慢。

5.3 模型的优化

5.3.1 Firth 惩罚的最大似然法

针对第二问的单因素分析中，氧化铅的数据完全分离导致 p 值异常的情况，可以采用 Firth 惩罚的最大似然法解决。Firth 提出在回归方程中加入惩罚项将渐进偏移去除，通过求解惩罚得分函数，得到移除偏差后参数的估计值。最后通过轮廓似然估计参数的可信区间，从而可以达到单因素分析的效果。

参考文献

- [1] 李莉. 统计学原理与应用 [M]. 南京大学出版社, 2019: 354.
- [2] 陈骏. 最近 2.5Ma 以来黄土高原风尘化学组成的变化与亚洲内陆的化学风化 [J]. 中国科学 (D 辑: 地球科学), 2001(02): 136-145.
- [3] 王蕾, 刘婷. 用交叉列表评价法解决企业技术经济效益评价问题 [J]. 技术经济, 2006(10): 36-38.
- [4] 任萱煜, 朱林慧. 基于多元 logistic 模型的互联网医疗使用意愿影响因素分析 [J]. 经济研究导刊, 2022(20): 46-49.

- [5] 覃玉冰, 邓春林, 杨柳. 基于皮尔逊相关系数的网络舆情评估指标体系构建研究 [J]. 情报探索, 2018(10):15-19.
- [6] 李洪成, 张茂军, 马广斌. *SPSS* 数据分析实用教程 [M]. 人民邮电出版社, 2017:338.
- [7] 曹莹, 苗启广. *AdaBoost* 算法研究进展与展望 [J]. 自动化学报, 2013, 39(06):745-758.

附录

A 问题结果

1.1 风化铅钡玻璃的化学成分预测结果

采样点	二氧化硅	氧化钠	氧化铅	五氧化二磷
02	64.48	1.88	27.56	0.17
08	51.94	2.12	20.73	0.18
08 严重风化点	47.21	2.84	21.80	0.46
11	58.19	1.64	19.24	1.00
19	63.84	2.28	34.27	1.02
26	49.19	1.96	22.18	0.16
26 严重风化点	43.92	2.68	19.87	0.18
34	58.58	1.52	33.56	0.18
36	60.57	3.62	15.78	0.17
38	62.33	3.34	22.28	0.15
39	60.45	2.28	19.97	0.15
40	59.31	2.84	19.63	0.18
41	56.86	2.56	34.52	1.06
43 部位 1	52.61	2.68	31.13	0.15
43 部位 2	61.30	2.64	34.85	1.03
48	59.93	1.24	14.06	0.18
49	60.59	2.12	26.23	1.00
50	55.78	2.52	34.55	0.04
51 部位 1	61.21	2.44	31.09	1.01
51 部位 2	60.95	2.64	29.33	1.00
52	56.94	3.30	19.86	0.51
54	63.08	2.72	21.13	0.18
54 严重风化点	63.31	3.08	25.44	1.00
56	57.35	1.88	34.20	0.15
57	57.82	2.16	23.09	0.15
58	60.99	2.04	31.70	1.03

1.2 玻璃表面有无风化化学成分含量的数据比较

未风化情况

二氧化硅：高钾 (67.77) > 铅钡 (55.72571429)

氧化钾：高钾 (9.567) > 铅钡 (0.194047619)

氧化钙：高钾 (5.735) > 铅钡 (1.21952381)

氧化铁：高钾 (1.789) > 铅钡 (0.690952381)

氧化铅：高钾 (0.384) < 铅钡 (21.35571429)

氧化钡：高钾 (0.458) < 铅钡 (8.859285714)

五氧化二磷：高钾 (1.181) > 铅钡 (1.113571429)

氧化锶：高钾 (0.034) < 铅钡 (0.265714286)

风化情况

二氧化硅：高钾 (93.96333333) > 铅钡 (26.45761905)

氧化钾：高钾 (0.543333333) > 铅钡 (0.157619048)

氧化钙：高钾 (0.87) < 铅钡 (2.596904762)

氧化铁：高钾 (1.789) < 铅钡 (0.690952381)

氧化铅：高钾 (0.435) < 铅钡 (43.38166667)

氧化钡：高钾 (0) < 铅钡 (12.54309524)

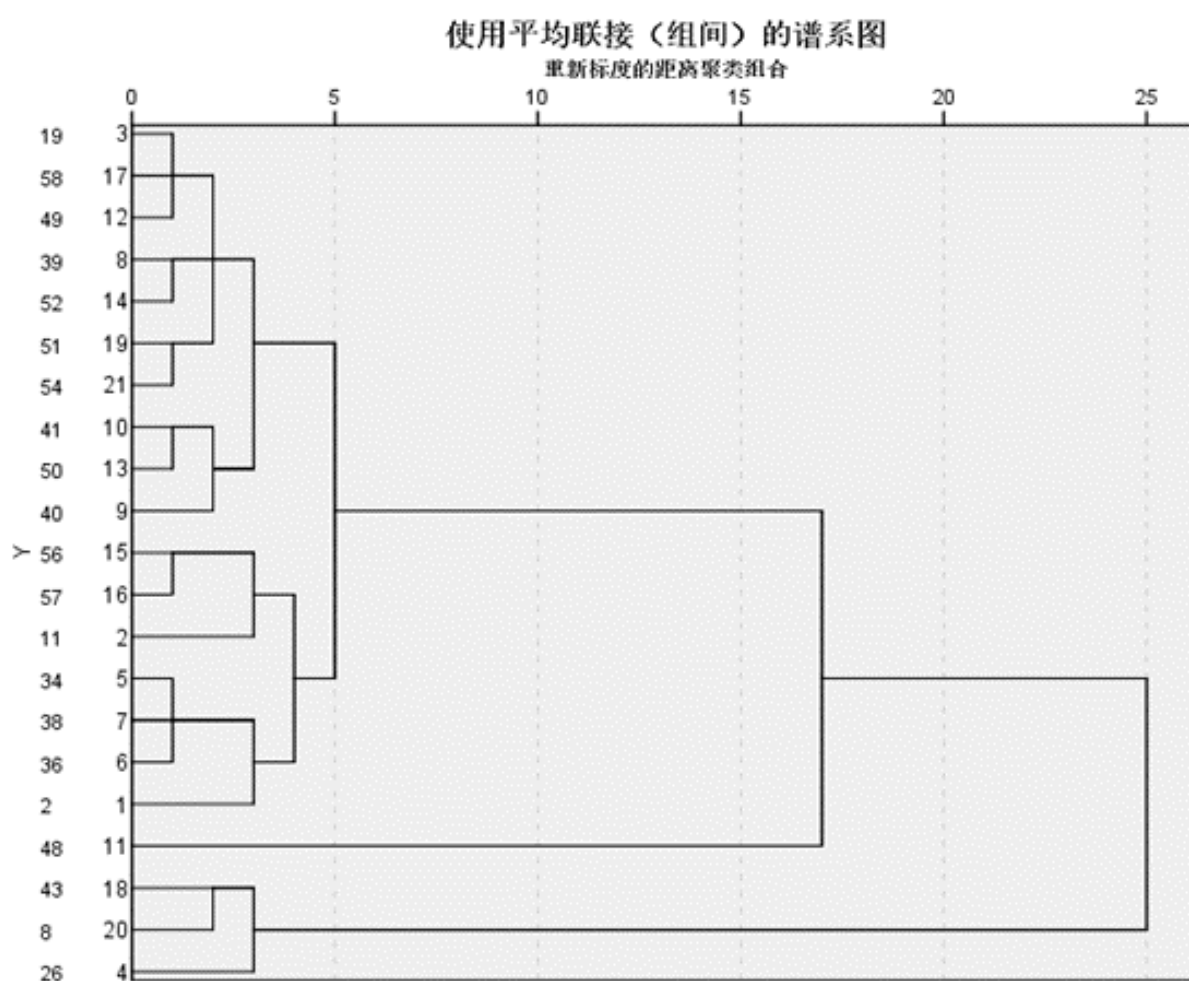
五氧化二磷：高钾 (0.28) < 铅钡 (4.975238095)

氧化锶：高钾 (0) < 铅钡 (0.391904762)

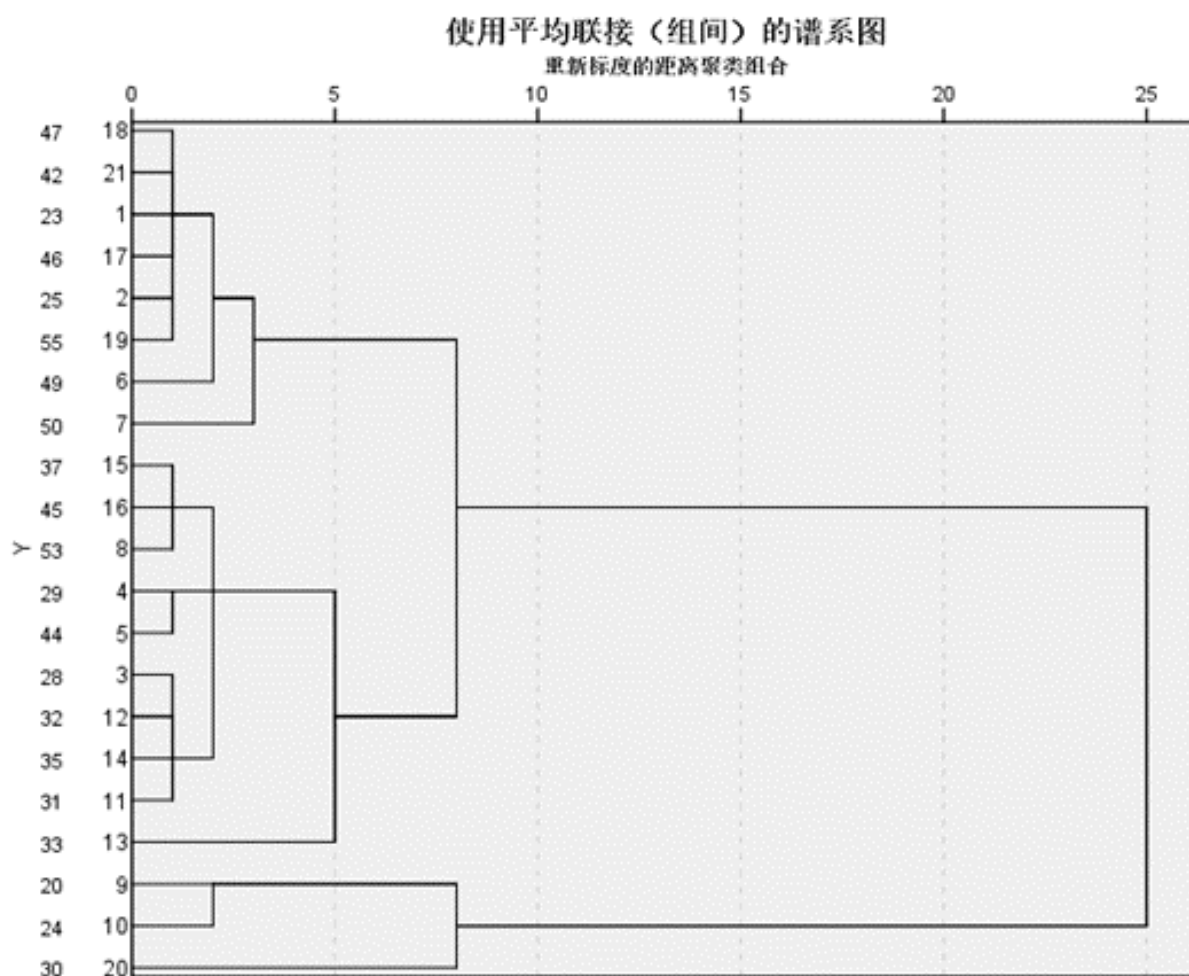
注：高亮部分表示有显著差异。

1.3 四种玻璃的聚类谱系图

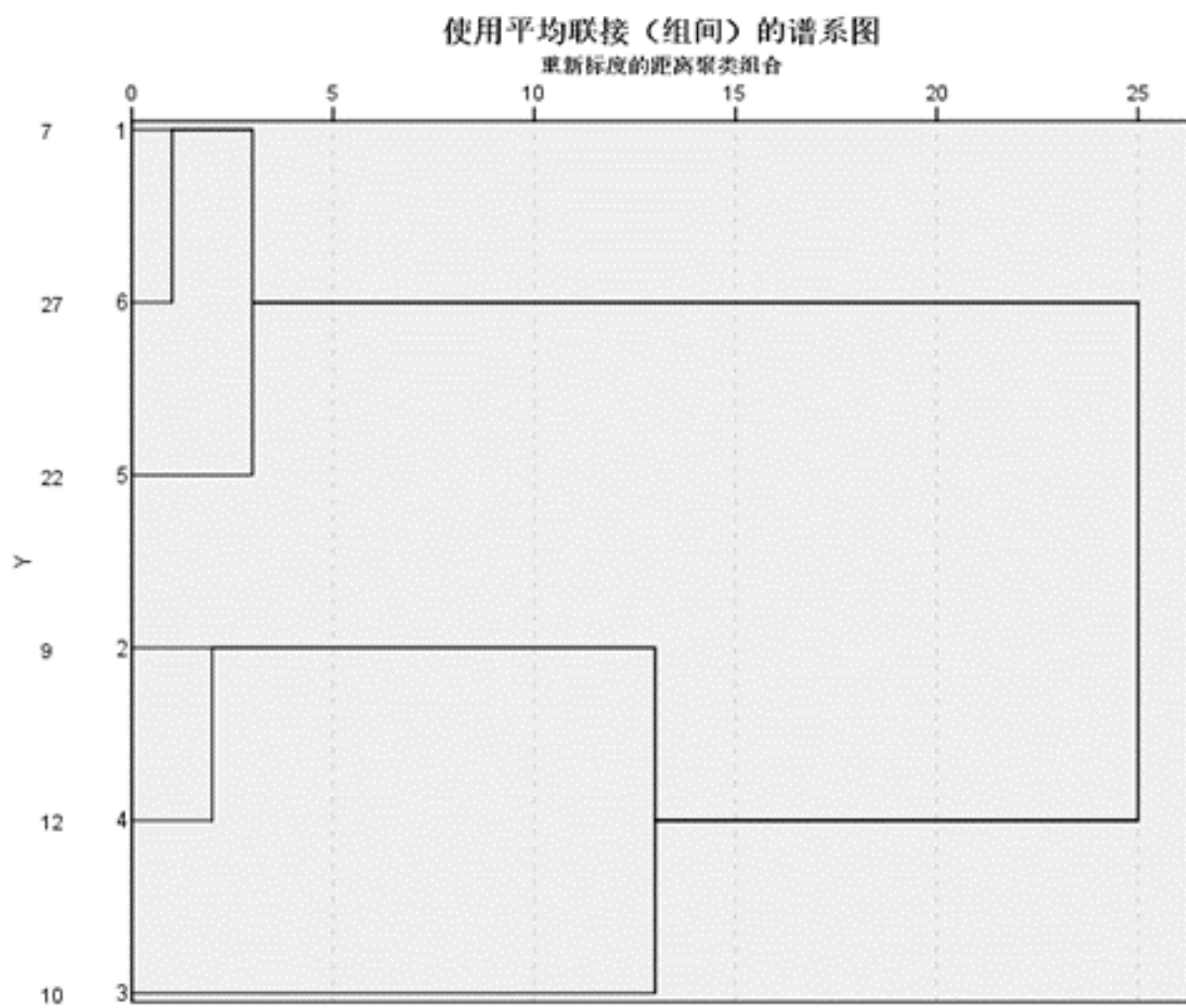
1.3.1 铅钡未风化



1.3.2 铅钡风化



1.3.3 高钾风化



1.4 各类玻璃的化学成分的皮尔逊相关系数

	二氧化硅(SiO ₂)	氧化钾(K ₂ O)	氧化钙(CaO)	氧化铝(Al ₂ O ₃)	氧化铁(Fe ₂ O ₃)	氧化铜(CuO)	五氧化二磷(P ₂ O ₅)
二氧化硅(SiO ₂)	1.000						
氧化钾(K ₂ O)	0.607	1.000					
氧化钙(CaO)	-0.897	-0.296	1.000				
氧化铝(Al ₂ O ₃)	-0.684	-0.036	0.895	1.000			
氧化铁(Fe ₂ O ₃)	0.204	0.750	0.173	0.451	1.000		
氧化铜(CuO)	-0.319	-0.639	0.000	-0.438	-0.735	1.000	
五氧化二磷(P ₂ O ₅)	-0.480	-0.786	0.225	-0.161	-0.540	0.870	1.000

	二氧化硅(SiO ₂)	氧化钙(CaO)	氧化铅(PbO)	五氧化二磷(P ₂ O ₅)	氧化锶(SrO)
二氧化硅(SiO ₂)	1				
氧化钙(CaO)	-0.29022	1			
氧化铅(PbO)	-0.31081	-0.23718	1		
五氧化二磷(P ₂ O ₅)	-0.34193	0.725488	-0.20306	1	
氧化锶(SrO)	-0.53595	0.147965	0.355522	0.304889	1

二氧化硅(SiO ₂)		氧化铅(PbO)
氧化硅(SiO ₂)	1	
氧化铅(PbO)	-0.54904	1

1.5 各类玻璃的化学成分的斯皮尔曼相关系数

1.5.1 高钾未风化

表 1.18: Add caption

	二氧化硅 (SiO ₂)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	五氧化二磷 (P ₂ O ₅)
二氧化硅 (SiO ₂)	1.000	-0.636	-0.733	0.115	-0.564	-0.515	-0.248	0.091
氧化钾 (K ₂ O)	-0.636	1.000	0.867	-0.552	-0.030	-0.224	-0.042	-0.442
氧化钙 (CaO)	-0.733	0.867	1.000	-0.564	0.200	0.091	0.236	-0.503
氧化镁 (MgO)	0.115	-0.552	-0.564	1.000	0.455	0.345	-0.188	0.430
氧化铝 (Al ₂ O ₃)	-0.564	-0.030	0.200	0.455	1.000	0.503	-0.127	-0.139
氧化铁 (Fe ₂ O ₃)	-0.515	-0.224	0.091	0.345	0.503	1.000	0.673	0.479
氧化铜 (CuO)	-0.248	-0.042	0.236	-0.188	-0.127	0.673	1.000	0.285
五氧化二磷 (P ₂ O ₅)	0.091	-0.442	-0.503	0.430	-0.139	0.479	0.285	1.000

1.5.2 高钾风化

表 1.19: Add caption

	二氧化硅 (SiO ₂)	氧 化 钾 (K ₂ O)	氧 化 钙 (CaO)	氧 化 镁 (MgO)	氧 化 铝 (Al ₂ O ₃)	氧 化 铁 (Fe ₂ O ₃)	氧 化 铜 (CuO)	五氧化二 磷 (P ₂ O ₅)
二氧化硅 (SiO ₂)	1.000	0.406	-1.000	-0.676	-0.943	-0.029	0.029	-0.543
氧 化 钾 (K ₂ O)	0.406	1.000	-0.406	-0.223	-0.406	0.522	-0.232	-0.928
氧 化 钙 (CaO)	-1.000	-0.406	1.000	0.676	0.943	0.029	-0.029	0.543
氧 化 镁 (MgO)	-0.676	-0.223	0.676	1.000	0.845	0.338	-0.676	0.135
氧 化 铝 (Al ₂ O ₃)	-0.943	-0.406	0.943	0.845	1.000	0.086	-0.200	0.486
氧 化 铁 (Fe ₂ O ₃)	-0.029	0.522	0.029	0.338	0.086	1.000	-0.543	-0.486
氧 化 铜 (CuO)	0.029	-0.232	-0.029	-0.676	-0.200	-0.543	1.000	0.486
五 氧 化 二 磷 (P ₂ O ₅)	-0.543	-0.928	0.543	0.135	0.486	-0.486	0.486	1.000

1.5.3 铅钡未风化

表 1.20: Add caption

	二氧化硅 (SiO ₂)	氧 化 钾 (K ₂ O)	氧 化 钙 (CaO)	氧 化 镁 (MgO)	氧 化 铝 (Al ₂ O ₃)	氧 化 铁 (Fe ₂ O ₃)	氧 化 铜 (CuO)	五氧化二 磷 (P ₂ O ₅)
二氧化硅 (SiO ₂)	1.000	-0.040	-0.028	0.280	0.105	0.051	-0.559	-0.163
氧 化 钾 (K ₂ O)	-0.040	1.000	0.122	0.387	0.614	0.290	-0.131	0.229
氧 化 钙 (CaO)	-0.028	0.122	1.000	0.171	0.335	0.247	-0.377	0.277
氧 化 镁 (MgO)	0.280	0.387	0.171	1.000	0.522	-0.065	-0.165	-0.044
氧 化 铝 (Al ₂ O ₃)	0.105	0.614	0.335	0.522	1.000	0.219	-0.160	0.081
氧 化 铁 (Fe ₂ O ₃)	0.051	0.290	0.247	-0.065	0.219	1.000	-0.425	0.460
氧 化 铜 (CuO)	-0.559	-0.131	-0.377	-0.165	-0.160	-0.425	1.000	-0.056
五 氧 化 二 磷 (P ₂ O ₅)	-0.163	0.229	0.277	-0.044	0.081	0.460	-0.056	1.000

1.5.4 铅钡风化

表 1.21: Add caption

	二氧化硅 (SiO ₂)	氧 化 钾 (K ₂ O)	氧 化 钙 (CaO)	氧 化 镁 (MgO)	氧 化 铝 (Al ₂ O ₃)	氧 化 铁 (Fe ₂ O ₃)	氧 化 铜 (CuO)	五氧化二 磷 (P ₂ O ₅)
二氧化硅 (SiO ₂)	1.000	0.317	-0.327	0.114	0.375	0.299	-0.285	-0.251
氧 化 钾 (K ₂ O)	0.317	1.000	0.154	0.341	0.324	0.229	-0.099	0.082
氧 化 钙 (CaO)	-0.327	0.154	1.000	0.715	0.470	0.503	0.195	0.764
氧 化 镁 (MgO)	0.114	0.341	0.715	1.000	0.828	0.660	-0.256	0.636
氧 化 铝 (Al ₂ O ₃)	0.375	0.324	0.470	0.828	1.000	0.632	-0.127	0.453
氧 化 铁 (Fe ₂ O ₃)	0.299	0.229	0.503	0.660	0.632	1.000	-0.344	0.263
氧 化 铜 (CuO)	-0.285	-0.099	0.195	-0.256	-0.127	-0.344	1.000	0.314
五 氧 化 二 磷 (P ₂ O ₅)	-0.251	0.082	0.764	0.636	0.453	0.263	0.314	1.000

1.6 具有显著性的化学成分关系

1.6.1 高钾未分化与铅钡未风化

表 1.22: Add caption

	二氧化硅 (SiO ₂)	氧 化 钾 (K ₂ O)	氧 化 钙 (CaO)	氧 化 镁 (MgO)	氧 化 铝 (Al ₂ O ₃)	氧 化 铁 (Fe ₂ O ₃)	氧 化 铜 (CuO)	五 氧 化 二 磷 (P ₂ O ₅)
二氧化硅 (SiO ₂)	0.000	0.596	0.705	0.165	0.669	0.566	0.311	0.254
氧 化 钾 (K ₂ O)	0.596	0.000	0.745	0.938	0.645	0.515	0.089	0.671
氧 化 钙 (CaO)	0.705	0.745	0.000	0.734	0.135	0.156	0.613	0.780
氧 化 镁 (MgO)	0.165	0.938	0.734	0.000	0.068	0.410	0.023	0.474
氧 化 铝 (Al ₂ O ₃)	0.669	0.645	0.135	0.068	0.000	0.284	0.032	0.221
氧 化 铁 (Fe ₂ O ₃)	0.566	0.515	0.156	0.410	0.284	0.000	1.098	0.019
氧 化 铜 (CuO)	0.311	0.089	0.613	0.023	0.032	1.098	0.000	0.341
五 氧 化 二 磷 (P ₂ O ₅)	0.254	0.671	0.780	0.474	0.221	0.019	0.341	0.000

1.6.2 高钾风化与铅钡风化

表 1.23: Add caption

	二氧化硅 (SiO ₂)	氧 化 钾 (K ₂ O)	氧 化 钙 (CaO)	氧 化 镁 (MgO)	氧 化 铝 (Al ₂ O ₃)	氧 化 铁 (Fe ₂ O ₃)	氧 化 铜 (CuO)	五 氧 化 二 磷 (P ₂ O ₅)
二氧化硅 (SiO ₂)	0.000	0.088	0.673	0.790	1.318	0.328	0.313	0.292
氧 化 钾 (K ₂ O)	0.088	0.000	0.560	0.564	0.730	0.292	0.133	1.010
氧 化 钙 (CaO)	0.673	0.560	0.000	0.039	0.473	0.474	0.224	0.221
氧 化 镁 (MgO)	0.790	0.564	0.039	0.000	0.017	0.322	0.420	0.500
氧 化 铝 (Al ₂ O ₃)	1.318	0.730	0.473	0.017	0.000	0.547	0.073	0.032
氧 化 铁 (Fe ₂ O ₃)	0.328	0.292	0.474	0.322	0.547	0.000	0.199	0.749
氧 化 铜 (CuO)	0.313	0.133	0.224	0.420	0.073	0.199	0.000	0.171
五 氧 化 二 磷 (P ₂ O ₅)	0.292	1.010	0.221	0.500	0.032	0.749	0.171	0.000

1.7 Fisher 判别法得到的分类结果

表 1.24: Add caption

文物编号	表面风化	预测	判别函数得分	属于第 1 组的后验概率	属于第 1 组的后验概率
A1	无风化	1	2.93749	0.00001	0.99999
A2	风化	0	1.74513	1	0
A3	无风化	0	-1.19182	1	0
A4	无风化	0	-1.52054	1	0
A5	风化	0	-1.84583	1	0
A6	风化	1	-16.72055	0	1
A7	风化	1	-13.70229	0	1
A8	无风化	0	0.00173	0.99995	0.00005

B 代码

2.1 问题一：logistic

```
1 syms x
2 list=[];
3 for i=1:6
4 p=0.01;
5 eqn = log(p/(1-p))==1.983*(1.883*x)-1.293*(7.117*x)-4.838*(x)
    +5.143
6 S=solve(eqn,p);
7 S=vpa(S);
8 list = [list S];
9 end
10
11 syms x
12 list=[];
13 for i=1:6
14 p=0.01+(0.04-0.01)*rand(1,6);
15 eqn = log(p(i)/(1-p(i)))==1.983*(1.883*x)-1.293*(7.117*x)-4.838*(x)
    )+5.143;
```

```

16 S=solve (eqn , x) ;
17 S=vpa (S) ;
18 S=double (S) ;
19 S=roundn (S, -3) ;
20 list = [list S] ;
21 end
22 t=1;
23 q=1;
24 while t<7
25 yct1 (q,1)=yct1 (q,1)+1.883*(list (q));
26 yct1 (q,2)=yct1 (q,2)+7.117*(list (q));
27 yct1 (q,3)=yct1 (q,3)+list (q);
28 q=q+1;
29 t=t+1;
30 end
31 yct1
32
33 syms x
34 list=[];
35 for i=1:26
36 p=0.02+(0.05-0.02)*rand (1,26);
37 eqn = log (p(i)/(1-p(i))) == -0.085*(84.0515*x+yct2 (i,1))
      -0.083*(4.2165*x+yct2 (i,2))+0.045*(-x+yct2 (i,3))
      +0.107*(-13.8144*x+yct2 (i,4))+1.245;
38 S=solve (eqn , x) ;
39 S=vpa (S) ;
40 S=double (S) ;
41 S=roundn (S, -2) ;
42 list = [list S] ;
43 end
44 list
45 t=1;
46 q=1;
47 while t<27
48 yct2 (q,1)=yct2 (q,1)+60*list (q) ;
49 yct2 (q,2)=yct2 (q,2)+4*(list (q)) ;
50 yct2 (q,3)=yct2 (q,3)-15*list (q) ;
51 if yct2 (q,3)>=35
52 yct2 (q,3)=15+(35-15)*rand (1) ;
53 end
54 yct2 (q,4)=yct2 (q,4)-10*list (q) ;

```

```

55 if yct2(q,4) <=0
56 yct2(q,4) = 0.15 + (0.05 - 0.02) * rand(1);
57 elseif yct2(q,4) >=2
58 yct2(q,4) = 1 + (0.05 - 0.02) * rand(1);
59 end
60 q = q + 1;
61 t = t + 1;
62 end
63 yct2
64
65 syms p
66 eqn = log(p/(1-p)) == -0.085*(20.14) - 0.083*(0) + 0.045*(28.68)
        + 0.107*(3.59) + 1.245;
67 S = solve(eqn, p);
68 S = vpa(S)

```

2.2 问题一：Adabost

```

1  clc
2  clear
3  W = [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1];
4  p = [];
5
6  Y = [1 1 1 1 1 1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1];
7  Y_1 = Y;
8
9  X = xlsread('分类汇总', sheet1);
10 error = [];
11
12 X_process = zeros(21, 14);
13 cut = [0.5 1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5 9.5 10.5 11.5 12.5
        13.5];
14 p_1 = [];
15 f = [];
16 a = [];
17 m = 15;
18 for i = 1:m
19 sumW = sum(W);
20 W = W ./ sumW;
21 for j = 1:length(cut)
22 for k = 1:length(X)

```

```

23  if X(k) < cut(j)
24  p(j,k) = 1;
25  else
26  p(j,k) = -1;
27  end
28  end
29  end
30  for j = 1:size(p,1)
31  if isempty(find(X_process==j))==0
32  f(j) = 999;
33  else
34  f(j) = sum(W.*((Y - p(j,:)).*Y) / 2);
35  if f(j) > 0.5
36  f(j) = 1 - f(j);
37  for k=1:length(X)
38  if X(k) < cut(j)
39  p(j,k) = -1;
40  else
41  p(j,k) = 1;
42  end
43  end
44  end
45  end
46  end
47
48  miE = min(f);
49  miI = find(f == miE);
50  if (size(miI,2) > 1)
51  miI = miI(1);
52  end
53  miE;
54  miI;
55  a(i) = log((1 - miE)/miE)/2;
56  Y_1 = p(miI,:);
57  Z = sum(W.*exp(-a(i) * p(miI,:).*Y));
58  W = W/Z .* exp(-a(i) * p(miI,:).*Y);
59  X_process(i) = miI;
60  fprintf('第%d迭代，划分点为 %f，对于的弱分类函数为：\n',i,cut(miI)
    );
61
62  p(i,:) =a(i)*Y_1;

```

```

63 sign(sum(p));
64 error(i)=sum(abs(sign(sum(p))-Y))/2;
65 if sign(sum(p))-Y==0
66 break;
67 end
68
69 end
70 fprintf('迭代结束，误差率为0：\n');

```


























2.3 问题四

```

1 z = ones(8,8);
2 for i=1:8
3 for q = (i+1):8
4 x = S8(:,i);
5 y = S8(:,q);
6 t1 = corr(x,y,'type','Spearman');
7 z(q,i)=t1;
8 end
9 end
10
11 list4=[];
12 for i=1:64
13 if list2(i)~=0
14 list4 = [list4,list2(i)];
15 end
16 end
17 q1 = prctile(list4,25);
18 q2 = prctile(list4,75);
19 iqr = q1-q2;
20 iqr1 = abs(q1-q2)
21
22 list3=[];
23 for i=1:64
24 if list2(i)~=0
25 list3 = [list3,list1(i)];
26 end
27 end
28 pjz=mean(list3,2)

```

2.4 文件列表

 问题1	2022/9/18 18:27	文件夹	
 问题2	2022/9/18 18:27	文件夹	
 问题3	2022/9/17 20:49	文件夹	
 问题4	2022/9/18 9:47	文件夹	
 1	2022/9/17 23:35	PNG 图片文件	50 KB
 2	2022/9/17 23:36	PNG 图片文件	42 KB
 3	2022/9/17 23:36	PNG 图片文件	8 KB
 chart	2022/9/18 16:55	PNG 图片文件	77 KB
 q21 (1)	2022/9/17 13:41	PNG 图片文件	77 KB
 q21 (2)	2022/9/17 13:41	PNG 图片文件	73 KB
 q21 (3)	2022/9/17 13:49	PNG 图片文件	96 KB
 q21 (4)	2022/9/17 13:52	PNG 图片文件	89 KB
 高钾风化	2022/9/18 10:01	PNG 图片文件	23 KB
 类型和表面风化的交叉图	2022/9/18 16:55	PNG 图片文件	58 KB
 铅钡风化	2022/9/18 10:00	PNG 图片文件	29 KB
 铅钡未风化	2022/9/18 10:00	PNG 图片文件	32 KB
 图片1	2022/9/17 21:13	PNG 图片文件	239 KB
 图片2	2022/9/18 14:36	PNG 图片文件	10 KB
 图片3	2022/9/17 22:44	PNG 图片文件	17 KB
 图片4	2022/9/18 10:03	PNG 图片文件	107 KB
 图片5	2022/9/18 10:32	PNG 图片文件	388 KB
 微信图片_20220917224527	2022/9/17 22:45	PNG 图片文件	121 KB
 微信图片_20220917224704	2022/9/17 22:47	PNG 图片文件	20 KB
 微信图片_20220918112206	2022/9/18 11:40	PNG 图片文件	53 KB
 纹饰和表面风化的交叉图	2022/9/18 16:54	PNG 图片文件	64 KB