模式识别-作业 2

计科 210X 甘晴void 202108010XXX

题目：

查阅文献资料，构建一个面向应用的分类系统。

要求：

1．至少 3 页 A4 纸，文字部分手写；

2．有系统背景，系统原理和系统设计部分；

3．有系统实现和系统评价更好。

目录

作业综述

我国古代玻璃技术外观上吸取了西亚和埃及地区的珠形饰品，但化学成分就地取材，有所差别，玻璃炼制过程需要添加助熔剂，添加的助熔剂不同，其主要化学成分也不同。其中，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，钾玻璃是以含钾量高的物质如草木灰作为助熔剂烧制而成的，古代玻璃易受埋藏环境的影响而风化，风化过程中，内外部元素大量交换使其成分比例发生变化，从而影响其类别的正确判断，因此需要建立相关的数学模型研究风化过程中古代玻璃制品的特征以及化学成分变化规律，以研究鉴别其种类。

本系统背景与训练数据来源于 CUMCM2022 ProblemsC 赛题。

解题思路来源于我参与的数学建模小组提交的论文。由于时间仓促，暂未将各部分代码汇总形成完整系统代码，各部分解决的详细代码附在文章后面。

古代玻璃制品成分分析分类器

（一）系统背景

丝绸之路是古代中西方文化交流的通道，其中玻璃是早期贸易往来的宝贵物证。早期的玻璃在西亚和埃及地区常被制作成珠形饰品传入我国，我国古代玻璃吸收其技术后在本土就地取材制作，因此与外来的玻璃制品外观相似，但化学成分却不相同。

玻璃的主要原料是石英砂，主要化学成分是二氧化硅（SiO2）。由于纯石英砂的熔点较高，为了降低熔化温度，在炼制时需要添加助熔剂。古代常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等，并添加石灰石作为稳定剂，石灰石煅烧以后转化为氧化钙（CaO）。添加的助熔剂不同，其主要化学成分也不同。例如，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，通常被认为是我国自己发明的玻璃品种，楚文化的玻璃就是以铅钡玻璃为主。钾玻璃是以含钾量高的物质如草木灰作为助熔剂烧制而成的，主要流行于我国岭南以及东南亚和印度等区域。

古代玻璃极易受埋藏环境的影响而风化。在风化过程中，内部元素与环境元素进行大量交换，导致其成分比例发生变化，从而影响对其类别的正确判断。如图 1 的文物标记为表面无风化，表面能明显看出文物的颜色、纹饰，但不排除局部有较浅的风化；图 2 的文物标记为表面风化，表面大面积灰黄色区域为风化层，是明显风化区域，紫色部分是一般风化表面。在部分风化的文物中，其表面也有未风化的区域。

图 1　未风化的蜻蜓眼玻璃珠样品

图 2　风化的玻璃棋子样品

现有一批我国古代玻璃制品的相关数据，考古工作者依据这些文物样品的化学成分和其他检测手段已将其分为高钾玻璃和铅钡玻璃两种类型。附件表单 1 给出了这些文物的分类信息，附件表单 2 给出了相应的主要成分所占比例（空白处表示未检测到该成分）。这些数据的特点是成分性，即各成分比例的累加和应为 100%，但因检测手段等原因可能导致其成分比例的累加和非 100%的情况。本题中将成分比例累加和介于 85%~105%之间的数据视为有效数据。

依据附件中的相关数据进行分析建模，解决亚类划分问题。目标是构建一个分类器，可以依据玻璃文物的表面风化、其玻璃类型、纹饰、颜色、化学成分含量等特征，对其进行分类，判断其玻璃产生类别。简单来说，是属于高钾玻璃还是铅钡玻璃。

正确判断该玻璃文物的分类对于文物保护有重大意义。

古代玻璃制品易受环境的影响而风化，风化过程中的元素变化使玻璃化学成分比例发生变化，从而影响其鉴别过程。本系统通过已知的一批我国古代玻璃制

品的相关数据，建立数学模型，分析玻璃制品风化前后化学成分的变化规律，玻璃类型分类规律并运用规律对玻璃进行相关鉴别。

（二）系统需求分析

根据设计系统的背景，结合题目提出以下目标：

（1）对这些玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系进行分析；结合玻璃的类型，分析文物样品表面有无风化化学成分含量的统计规律，并根据风化点检测数据，预测其风化前的化学成分含量。

（2）依据附件数据分析高钾玻璃、铅钡玻璃的分类规律；对于每个类别选择合适的化学成分对其进行亚类划分，给出具体的划分方法及划分结果，并对分类结果的合理性和敏感性进行分析。

（3）对附件表单 3 中未知类别玻璃文物的化学成分进行分析，鉴别其所属类型，并对分类结果的敏感性进行分析。

（4）针对不同类别的玻璃文物样品，分析其化学成分之间的关联关系，并比较不同类别之间的化学成分关联关系的差异性。

这 4 个目标分别对数据进行预处理，亚类划分，所属类型划分，差异性分析，并有分类的敏感性分析。在达成这 4 个目标的过程中构建系统。

下面分析这些目标：

2.1 问题一分析

问题一的第一小问，首先我们对样本数据进行了剔除和预处理，为了更加清晰的判断其关系，本文分别建立了玻璃文物的表面风化与玻璃类型、玻璃文物的表面风化与纹饰和颜色、玻璃文物的表面风化与颜色的线性函数关系，根据函数关系中相关参数进行分析。

问题一的第二小问，将玻璃类型与有无风化的情况进行组合，组成四种形式，再分别做出这四种组合形式玻璃的化学成分统计图，通过比较得出统计规律。针对预测问题，本文采用四分位法对两种玻璃类型风化前后各种化学成分进行标准化处理，将风化前后标准值相除得到各种化学成分变化比率，结合此比率预测其风化浅的化学成分含量。需要注意的是，在标准化过程中，标准值为零的特殊情况，直接将无风化的标准值作为这些数据的预测值。

2.2 问题二分析

问题二第一小问，本文首先对化学成分比例进行归一化处理，分别算出风化前后玻璃类型与各化学元素成分的相关系数，并结合小提琴图以筛选出与玻璃类型相关性较强的化学成分。其次，分别在无风化与风化两种条件下对玻璃类型与筛选出化学成分进行多元线性回归，得到有无风化条件下的两个方程，根据函数值判断玻璃类型，函数值大于 0.5 的即为铅钡玻璃，反之为高钾玻璃。

问题二第二小问主要采用聚类分析法，首先对数据进行剔除和预处理。在组数划分上，经过多次试错后确定最佳的分类组数--高钾玻璃为 3 组、铅钡玻璃为 5 组。将 14 种化学成分定为指标，采用 Q 型聚类法，即将常用的统计量用距离来表达。对于距离的表示，我们采用欧式距离。关于聚类方法的选择，本文在尝试了最短距离法和重心距离法后选择内平方距离法并得到了最终的分组结果。针

对分类结果的合理性、敏感性分析，本文选择轮廓系数作为聚类性能的评估指标，计算聚类分析后聚类总的轮廓系数值衡量其合理性及敏感性。

2.3 问题三分析

在问题二的基础上，考虑到是否风化对于化学成分含量比例的影响，我们将表单 3 中的未知文物分成“表面风化”与“表面无风化”两类。对于每一类分别采取随机森林算法，使用表单 2 的对应类别数据作为训练集与测试集，进行模型训练。完成模型训练后，使用表单 3 中未知文物的化学成分数据进行预测。

2.4 问题四分析

首先，我们分别计算高钾玻璃、铅钡玻璃各自化学成分之间的 Pearson(皮尔逊)相关系数，并形成相关系数矩阵，考虑到使用皮尔逊相关系数来刻画不同化学成分之间的关系，要求数据具有符合正态分布的特性，我们对每种化学成分 W_i所对应的各样本数据进行正态性检验。我们使用 matlab 绘制 qq 图，发现数据的正态性的偏差性较大，结合 Kolmogorov-Smirnov 算法与 Jarque-Bera 算法，发现大部分的数据其实不符合正态性分布，使用 Pearson(皮尔逊)相关系数不能来很好地刻画它们的相关性。因此我们采取对于正态性要求较低的斯皮尔曼（spearman）相关系数来刻画这种相关性，在计算出化学成分的相关系数矩阵之后，我们使用 python 的 heatmap 库可以绘制出相关系数热点图，方便较为直接地观察化学成分之间的关联关系。

（三）系统假设与符号说明

（1）系统假设如下

①假设除附件表单二中化学成分外，古代玻璃制品不含其他化学成分；

②假设问题二第一小问被排除的化学成分对分类的影响忽略不计。

（2）系统符号说明如下

表 1 列出了本文需要的符号，文中出现的其它符号将在出现时进行解释。

表 1 本文符号说明

符号	含义
Y	有无风化
T	玻璃类型
D1	B 纹饰的虚拟变量
D2	C 纹饰的虚拟变量
R	颜色的十进制对应 RGB 值中的 R 值
G	颜色的十进制对应 RGB 值中的 G 值
B	颜色的十进制对应 RGB 值中的 B 值
Wi(i=1,2.....，14）	分别对应表单 2 中从左到右化学成分含量

（四）系统原理与设计

4.1 问题一模型的建立与求解

4.1.1 玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系的分析

在建立模型前对数据进行预处理，首先，排除表单一中颜色变量未知的文物数据。其次，针对表面风化与其玻璃类型、纹饰和颜色等变量，风化变量采用 0（未风化）、1（风化）变量形式，文物纹饰及类型以虚拟变量方式引入（以纹饰 A，类型高钾作为参考项），颜色变量则采用十进制 RGB 值，引入 R、G、B 三个变量，然后利用计量软件 Eviews 分别对玻璃文物的表面风化与其玻璃类型、纹饰和颜色变量进行 OLS 线性回归，得出：

①玻璃文物风化与玻璃类型的函数表达式为：Y=0.33333+0.33333T

②玻璃文物风化与纹饰函数表达式为：Y=0.45+0.55D1+0.085714D2

③玻璃文物与颜色表达式为：Y=0.59225+0.00068R-0.00001G-0.00060B

对函数方程①分析可知，铅钡玻璃相较于高钾玻璃更易风化；对函数方程②分析可知，三种纹饰中，A 纹饰的风化可能性最小，B 纹饰的风化可能性中等，C 纹饰的风化可能性最大；对函数方程③分析可知，颜色 R 值越大，越易风化，G 值 B 值越大，越不易风化。

4.1.2 结合玻璃的类型，分析文物样品表面有无风化化学成分含量的统计规律

首先，将玻璃类型与文物表面有无风化进行排列组合，得到“风化高钾”、“无风化高钾”、“风化铅钡”、“无风化铅钡”四种玻璃形式。其次，根据题目中要求的数据有效范围，剔除了成分比例累加和为 79.47%的 15 号文物及成分比例累加和为 71.89%的 17 号文物。然后，对矩阵进行行标准化处理，对列化学成分取平均值，运用 MATLAB 对以上处理过的数据进行可视化分析，行成四种玻璃文物化学成分统计图，如下：

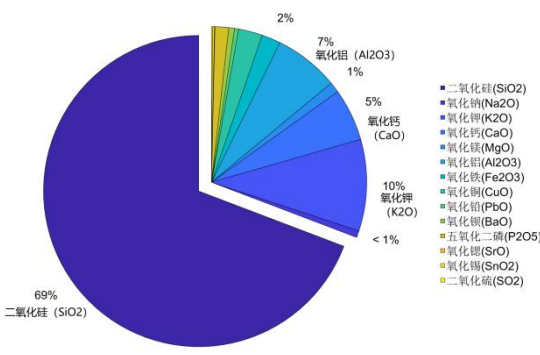

图 1 无风化高钾玻璃制品化学成分饼状图

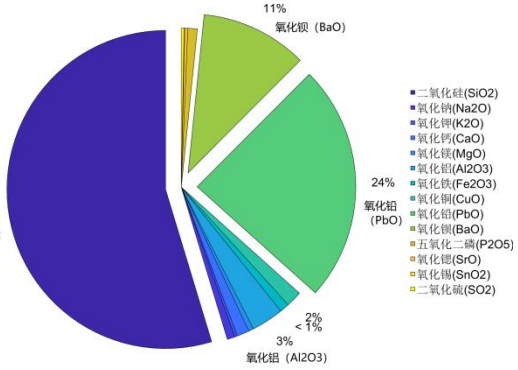

图 2 无风化铅钡玻璃制品化学成分饼状图

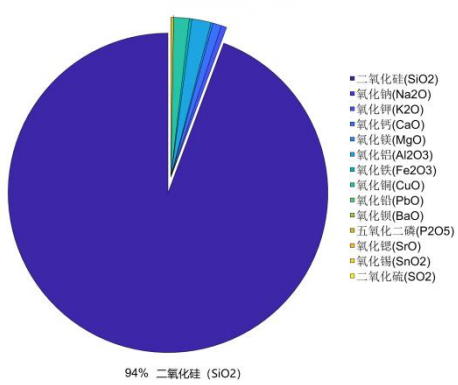

图 3 风化高钾玻璃制品化学成分饼状图

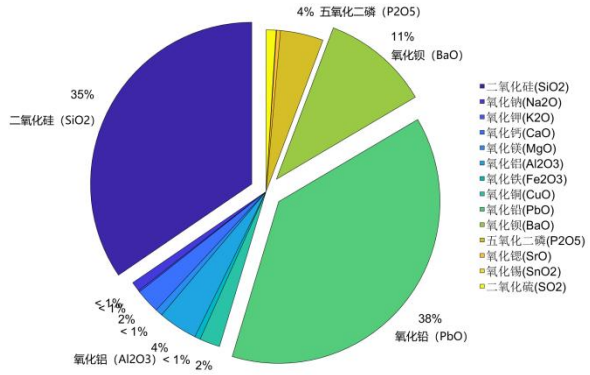

图 4 风化铅钡玻璃制品化学成分饼状图

对比图 1、3 饼状图可知，在玻璃类型是高钾的前提下，风化玻璃二氧化硅含量明显提高（69.23%~94.33%），氧化钾（9.52%~0.54%）、氧化钙（5.44%~0.87%）、氧化铝（6.74%~1.94%）明显下降，同时结合化学成分比例值（见附录）发现其他化学成分含量均降低。

对比 2、4 饼状图可知，在玻璃类型是铅钡的前提下，风化玻璃二氧化硅含量明显降低（54.69%~34.52%），氧化铅（24.12%~38.13%）、五氧化二磷（0.97%~4.31%）含量明显提高，氧化钡含量基本不变，其余大多数化学成分含量基本不变或有所提高。

总结看来，高钾玻璃风化后二氧化硅成分含量最高，其他化学成分含量均较少，且风化前后二氧化硅均含量第一；铅钡玻璃风化后二氧化硅含量减少，氧化钡含量基本不变，氧化铅含量增加，且风化前二氧化硅为含量第一的成分，风化后氧化铅成分含量第一。

4.1.3 对风化前的化学成分含量的预测

首先，采用四分位法对两种玻璃类型风化前后各种化学成分进行标准化处理，将风化前后标准值（具体标准值见附录）相除得到各种元素变化比率，结合此比率预测其风化前的化学成分含量。需要注意的是，在标准化过程中，标准值为零的特殊情况，直接将无风化的标准值作为这些数据的预测值。下表列出几个文物风化采样点的预测值，因篇幅原因，其他预测值在附录中展示。

表 2 文物采样点 02、07、08、09、10 的化学成分预测值

文物采样点	02	07	08	08 严重风化点	09	10
类型	铅钡	高钾	铅钡	铅钡	高钾	高钾
二氧化硅(SiO2)	57.43	67.28	31.90	7.42	68.97	70.21
氧化钠(Na2O)	0.00	0.00	0.00	0.00	0.00	0.00
氧化钾(K2O)	1.64	0.00	0.00	0.00	11.24	17.51
氧化钙(CaO)	0.92	7.49	0.58	1.27	4.34	1.47
氧化镁(MgO)	0.89	0.00	0.00	0.00	0.00	0.00
氧化铝(Al2O3)	5.68	6.99	1.33	1.12	4.66	2.86
氧化铁(Fe2O3)	2.70	1.09	0.00	0.00	2.06	1.67

氧化铜(Cu0)	0.12	5.63	4.68	1.43	2.69	1.46
氧化铅(Pb0)	29.04	0.00	17.57	20.20	0.00	0.00
氧化钡(Ba0)	0.00	0.00	31.83	31.71	0.00	0.00
五氧化二磷(P205)	0.55	2.33	0.56	1.19	1.34	0.00
氧化锶(Sr0)	0.13	0.00	0.26	0.37	0.00	0.00
氧化锡(Sn02)	0.00	0.00	0.00	0.00	0.00	0.00
二氧化硫(S02)	0.00	0.00	2.58	15.30	0.00	0.00

4.2 问题二模型的建立与求解

4.2.1 问题二模型建立基础

（1）斯皮尔曼（Spearman）系数

斯皮尔曼（Spearman）系数，又称秩相关系数，根据随机变量的等级而不是其原始值衡量相关性的一种方法。对原始变量的分布不作要求，适用范围更广些。不服从正态分布的变量、分类或等级变量之间的关联性可采用 Spearman 秩相关系数，：W_i和W_j为两组数据，其斯皮尔曼（等级）相关系数为：

$$r_s = 1 - \frac{6\sum_{k=1}^{n} d_i^{\,2}}{n(n^2-1)}$$

di 为W_i和W_j之间的等级差。可以证明：rs 位于-1 和 1 之间。其中等级差为将它所在的一列按照从小到大排序后，这个数所在的位置。这样做的目的是让它符合单调性。

斯皮尔曼（spearman）相关系数的假设检验，计算显著性水平：

①小样本(n<30) :直接查临界值表

样本相关系数r_s必须大于等于表格中的临界值（临界表见附录），才能得出显著的结论。

②大样本情况（n > 30）: P 值法，依据：$r_s\sqrt{n-1} \sim N(0,1)$

a.确立原假设H_0：$r_s = 0$，即两变量之间不存在线性关联和备择假设H_1：$r_s \neq 0$，即两变量之间存在线性关联。

b.在原假设成立的条件下，根据需要检测的量构造一个分统计量检验值$r_s\sqrt{n-1}$

c.求出对应的 p 值。并与 0.05 相比（注意：双侧检验时，p 值需要乘 2，假如 p>0.05，则无法拒绝原假设；假如 p≤0.05，则拒绝原假设。）

（2）欧氏距离（Euclidean distance）

$d(x,y) = /sqrt(x-y)'(x-y) = /sqrt/sum_j(x_j - y_j)^2$,

其中x_j, y_j为向量的横纵坐标

（3）Ward 方法（内平方法）

Ward 法（Ward’s method），又称方差平方和增量法（Incremental sum of squares），由合并前后的族群内方差平方和的差异$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$定义距离。记为族群和合并而得的族群，则合并前后的族群内方差平方和分别为：

$$SSE_A = \sum_{I=1}^{n_A} (y_i - y_A)'(y_j - \overline{y_A}) for y_i \in A$$

$$SSE_B = \sum_{I=1}^{n_B} (y_i - y_B)'(y_j - \overline{y_B}) for y_i \in B$$

$$SSE_{AB} = \sum_{I=1}^{n_{AB}} (y_i - y_{AB})'(y_j - \overline{y_{AB}}) for y_i \in AB$$

其中，$\overline{y_{AB}} = \frac{\sum_{i=1}^{n_A} y_i + \sum_{i=1}^{n_B} y}{n_A + n_B} = \frac{n_A \overline{y_A} + n_B \overline{y_B}}{n_A + n_B}$

由 Ward 法合并的两个族群和，应使得在合并前后的增量$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$最小。

（4）轮廓系数[1]

当文本类别未知时，可以选择轮廓系数作为聚类性能的评估指标。轮廓系数取值范围为[-1,1]，取值越接近 1 则说明聚类性能越好，相反，取值越接近-1 则说明聚类性能越差。

则针对某个样本的轮廓系数 s 为：$s = \frac{b-a}{max(a,b)}$,聚类总的轮廓系数 SC 为:$SC = \frac{1}{N}\sum_{i=1}^{N} SC(d_i)$

(a：某个样本与其所在簇内其他样本的平均距离,b：某个样本与其他簇样本的平均距离)

4.2.2 高钾玻璃、铅钡玻璃的分类规律

首先，使用 excel 对原始数据进行归一化处理，然后分别计算玻璃类型与各化学元素含量的斯皮尔曼（Spearman）系数，得到如下表 3、表 4：

表 3 无风化条件下玻璃类型与各化学元素含量的 sperman 相关系数

化学成分	相关系数	化学成分	相关系数
二氧化硅(SiO2)	-0.499630041	氧化铜(CuO)	-0.355566091
氧化钠(Na2O)	-0.014816522	氧化铅(PbO)	0.869375611
氧化钾(K2O)	-0.754540909	氧化钡(BaO)	0.880363417
氧化钙(CaO)	-0.578463771	五氧化二磷(P2O5)	-0.266674568
氧化镁(MgO)	-0.440181708	氧化锶(SrO)	0.38295537
氧化铝(Al2O3)	-0.688379167	氧化锡(SnO2)	0.088482615
氧化铁(Fe2O3)	-0.38652339	二氧化硫(SO2)	-0.208665676

表 4 风化条件下玻璃类型与各化学元素含量的 sperman 相关系数

化学成分	相关系数	化学成分	相关系数
二氧化硅(SiO2)	-0.606263394	氧化铜(CuO)	-0.106661771
氧化钠(Na2O)	0.239535069	氧化铅(PbO)	0.607124932

化学成分	相关系数	化学成分	相关系数
二氧化硅(Si02)	-0.606263394	氧化铜(Cu0)	-0.106661771
氧化钾(K20)	-0.315412294	氧化钡(Ba0)	0.558463238
氧化钙(Ca0)	0.353667976	五氧化二磷(P205)	0.31507447
氧化镁(Mg0)	0.296252809	氧化锶(Sr0)	0.542539897
氧化铝(A1203)	0.269450398	氧化锡(Sn02)	0.113135648
氧化铁(Fe203)	0.034472513	二氧化硫(S02)	0.132246245

为了筛选出与玻璃类型相关性较强的化学成分，本文在相关系数的基础上，结合绘制的小提琴图进行筛选。我们需要学习一下如何从小提琴图中提取信息，例如，在无风化状态中，氧化钡与氧化铅各自的 0（高钾玻璃）1（铅钡玻璃）状态的中位数看起来是比较分开的，因此对分类很有用。而事实上它们的相关系数绝对值分别达到 0.86，0.88。这也可以说明实际上它们分别与玻璃类型的相关性是比较高的；但是，在无风化状态中，氧化钠（Na2O）的 0（高钾玻璃）1（铅钡玻璃）状态下的中位数看起来不像是分开的，因此它不能提供很好的分类信息。从斯皮尔曼（Spearman）系数来看，Na2O 的相关系数绝对值确实也较低，这两种方法得出的结果比较符合。

结合 sperman 相关系数与小提琴图，我们筛选出最终与玻璃类型相关的化学成分：

（1）无风化条件下

在十四种化学成分中排除：氧化钠(Na2O)、氧化锡(SnO2)和二氧化硫(SO2)，因为这三种化学成分小提琴图中位数过于相近，且相关系数绝对值过低。

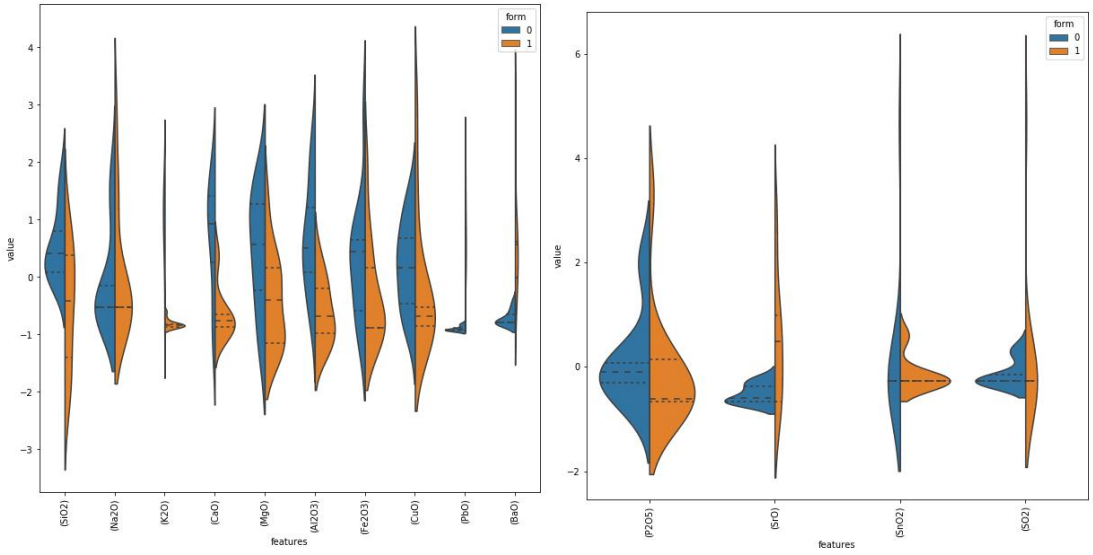

图 4，5 无风化条件下 14 中化学成分小提琴图

最后用 Eviews 将玻璃类型与二氧化硅(SiO2)、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)成分含量进行多元线性回归，得到函数表达式①：

T=5.01-4.47W1-5.97W3-8.40W4-5.82W5-6.54W6+4.37W7-9.01W8-3.46W9-0.

59W10-12.05W11-37.58W12

（2）风化条件下

在十四种化学成分中排除：氧化钠(Na2O)、氧化铁(Fe2O3)、氧化铜(CuO)、五氧化二磷(P2O5)、氧化锡(SnO2)、二氧化硫(SO2)，因为这六种化学成分小提琴图中位数过于相近，且相关系数绝对值过低。

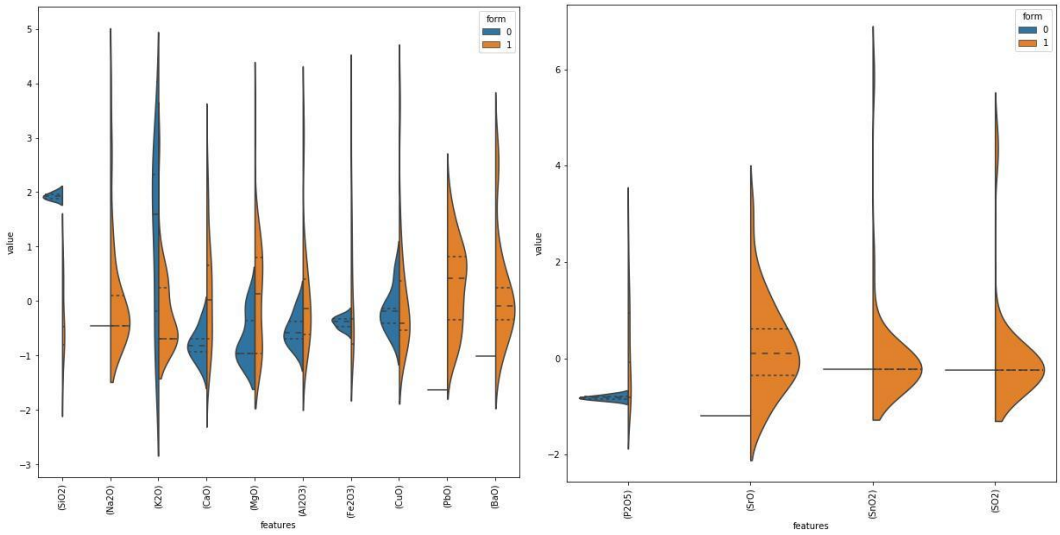

图 5、6 风化条件下 14 种化学成分小提琴图

最后用 Eviews 将玻璃类型与二氧化硅(SiO2)、氧化钾(K2O)、氧化钙(CaO)、氧化镁(MgO)、氧化铝(Al2O3)、氧化铅(PbO)、氧化钡(BaO)、氧化锶(SrO)成分含量进行多元线性回归，得到函数表达式②为：

T=1.06-0.94W1-18.85W3-5.85W4+10.17W5+4.48W6+0.49W9+0.55W10-22.00W12

结合①②得到的函数表达式得到高钾玻璃、铅钡玻璃的分类规律：在无风化条件下，将①函数中对应的化学成分含量比例代入，得到的函数值大于 0.5 则为铅钡玻璃，否则为高钾玻璃；在风化条件下，将②中函数对应化学成分含量比例代入，得到的函数值大于 0.5 则为铅钡玻璃，否则为高钾玻璃。

而查阅了相关参考文献[4,5]对高钾、铅钡玻璃的划分，高钾玻璃 Al2O3, CaO 和 Na2O 的含量皆很低 (<3%)，而 K2O 含量很高(>10%)；铅钡玻璃二氧化硅约 45%~50%，氧化铅约 15%~20%，氧化钡约 10%，显然符合得出的结论。

4.2.3 对每个类别选择合适的化学成分进行亚类划分

首先对数据进行归一化和标准化处理，剔除了异常值。由于不确定数据可以划分的组数，多次试错后确定最佳的分类组数高钾玻璃为 3 组、铅钡玻璃为 5 组。将 14 个化学元素定为指标，采用 Q 型聚类法，即将常用的统计量用距离来表达。对于距离的表示，我们采用欧式距离。关于聚类方法的选择，我们先尝试了最短距离法和重心距离法（结果见附录），因为数据非单调簇树，所以重心距离法和最短距离法连接可能不合适，因而放弃，最终选择内平方距离法，得出较好结论，聚类结果如下：

（1）高钾玻璃：

表 5 高钾玻璃最终分组

分组	文物编号
一	21
二	01、03 部位 2、04、05、06 部位 1、06 部位 2、13、14、16
三	03 部位 1、07、09、10、12、18、22、27

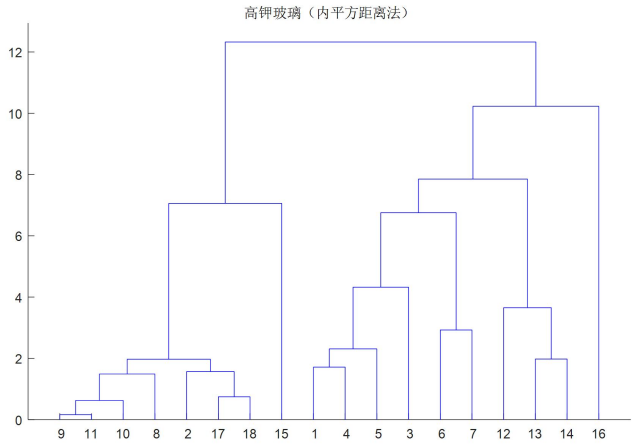

图 7 高钾玻璃（内平方距离法）的聚类图

（2）铅钡玻璃:

表 6 高钾玻璃最终分组

分组	文物编号
一	20、28 未风化点、29 未风化点、31、32、33、35、37、44 未风化点、45、46、49 未风化点、53 未风化点
二	23 未风化点、25 未风化点、34、36、38、42 未风化点 1、42 未风化点 2、47、55、56、57
三	02、30 部位 1、48
四	11、19、30 部位 2、39、40、41、43 部位 1、43 部位 2、49、50、50 未风化点、51 部位 1、51 部位 2、52、54、54 严重风化点、58
五	08、08 严重风化点、24、26、26 严重风化点

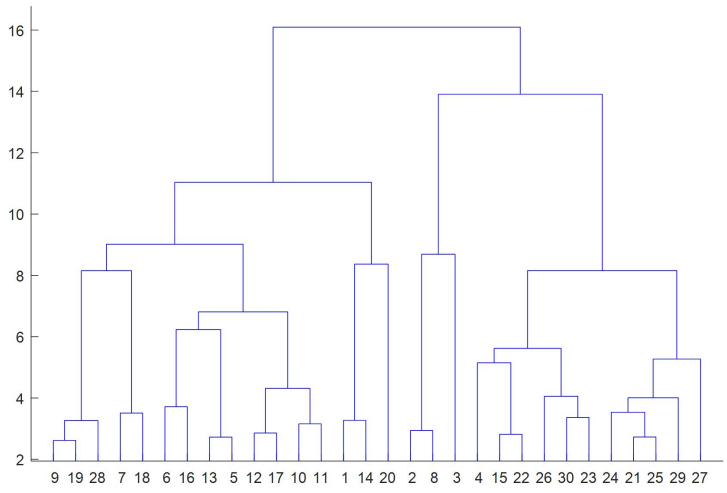

图 8 铅钡玻璃（仅前 30 个）（内平方距离法）的聚类结果

4.2.4 合理性分析

根据轮廓系数的相关知识，我们分别算出了以上高钾玻璃和铅钡玻璃的轮廓值，如下图：

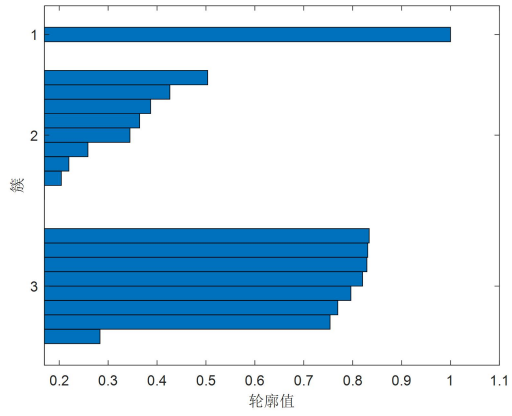

图 9 高钾玻璃轮廓值

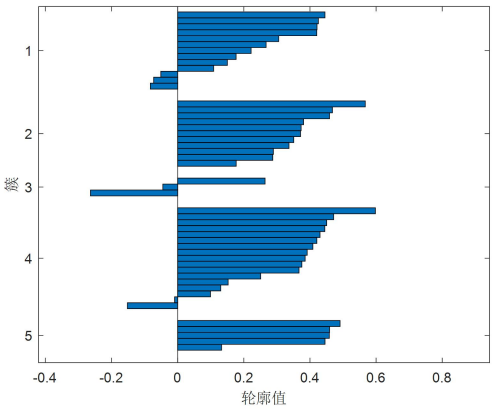

图 10 铅钡玻璃轮廓值

轮廓系数取值范围为[-1,1]，取值越接近 1 则说明聚类性能越好，相反，取值越接近-1 则说明聚类性能越差。

根据图 9 可看出高钾玻璃的轮廓值均大于零，说明聚类性能良好，聚类较为合理。根据图 10，可看出铅钡玻璃的轮廓值大多数大于 0（更偏向于 1），只有 7 个样本小于 0 且均为小于-0.4，说明聚类性能良好，聚类较为合理。

4.3 问题三模型的建立与求解

4.3.1 问题三模型建立基础

（1）随机森林算法

随机森林算法属于集成算法中的 bagging 类别，其本质就是集成多个弱分类器达成较强的分类效果。对于本模型而言，其意义在于集成不同化学成分对于最

终结果的影响，通过集成投票的方式选择更为可能的结果，从而完成对于玻璃文物类别的预测。

4.3.2 未知类别玻璃文物类型鉴别及敏感性分析

考虑到是否风化对于化学成分含量比例的影响，我们将表单 3 中的未知文物分成“表面风化”与“表面无风化”两类。对于每一类分别采取随机森林算法，使用表单 2 的对应类别数据作为训练集与测试集，进行模型训练。完成模型训练后，使用表单 3 中未知文物的化学成分数据进行预测。建立玻璃文物类别随机森林算法模型：

（1）数据集说明

总数据集：已知文物的类别 T（高钾玻璃或铅钡玻璃）与其对应的化学成分W_i的含量a_{ij}

训练集 D：总数据集的随机 70%

测试集 E：总数据集的剩余 30%

（2）决策树的构建

①特征选择

在问题 2 中，我们已经探究了高钾玻璃、铅钡玻璃的分类规律。因此在此处的特征选择，我们将沿用之前所确定的相关特征。

即对于“表面无风化”类别，我们采用二氧化硅(SiO2)、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)上述 11 种，作为特征；

对于“表面风化”类别，我们采用二氧化硅(SiO2)、氧化钾(K2O)、氧化钙(CaO)、氧化镁(MgO)、氧化铝(Al2O3)、氧化铅(PbO)、氧化钡(BaO)、氧化锶(SrO)。上述 6 种，作为特征。

②分类决策树T_0生成

根据训练集，从根节点开始，使用递归方式遍历各节点，生成二叉决策树。遍历中需要对每个节点$Node_i$进行如下操作：

a.计算现有特征对该数据集的基尼指数。对于每一个特征 A 的任意可能值 a，根据样本点对 A==a 是否成立将 D 分为D_1和D_2两部分，并计算 A=a 时的基尼指数 Gini(D)。

$$\mathrm{Gini(D)} = 1 - \sum_{i=1}^{n} \left(\frac{|C_i|}{|D|} \right)^2$$

其中C_i为 D 中第 i 类的样本子集，n 为类总数

b.在任意可能的 A 与其可能的切分点 a 中，选择基尼指数最小的特征A_k以及其对应的切分点a_k作为最优特征与最优切分点，并依此从该节点生成两个子节点$Node_m$和$Node_n$，将训练集依照特征分配到其子节点中去。

③剪枝处理

以损失函数（loss function）作为剪枝的标准，用测试集对已生成的树进行剪枝并选择最优子树。首先从产生的决策树底端剪枝，以根节点为先决条件，形成一个子树序列；之后通过交叉验证法在独立的测试集上对子序列进行测试，按照测试的结果选择最优子树。由决策树T_0生成最优决策树T_α的算法如下：

a.i=0，$T=T_0$

b.令$\alpha = +\infty$

c.自下而上地对各内部节点 node 计算 $C(T_{node})$, $|T_{node}|$ 以及 $g(node)=\frac{C(node)-C(T_{node})}{|T_{node}|-1}$, $\alpha=\min\left(\alpha, g(node)\right)$,上述式子中，$T_{node}$表示以 node 为根节点的子树，$C(T_{node})$表示对训练集的预测误差，$|T_{node}|$表示$T_{node}$根节点的个数。

d.自上而下地访问内部节点 node，若$\exists g(t)=\alpha$，则可以进行剪枝操作，并对于节点 node 以投票计数的方法确定其种类，可以获取树 T

e.令 i++，$\alpha_i=\alpha$，$T_i=T$

f.若 T 存在子树，则返回步骤（4）

g.采用交叉验证法在字数序列$T_0, T_1, T_2, T_3 \ldots\ldots T_n$中选取最有子树$T_\alpha$

（4）由决策树形成随机森林

对于每一棵分类决策树，最后都会产生一个分类结果。这里采取投票的方式对这些结果进行计数，获取票数较高的结果即为最终的结果。

上述过程由 python 编程实现。“表面风化”类别代码详见“T3s1_differentiation.py”，“表面无风化”类别代码详见“T3s1_non_differentiation.py”。

结合以上过程，得出未知文物的具体类型如下表：

表 7 未知文物分类表

文物编号	类型
A1	高钾
A2	铅钡
A3	铅钡
A4	铅钡
A5	铅钡
A6	高钾
A7	高钾
A8	铅钡

敏感性分析：

采用改变参数值，观察评价因变量随参数的变化的方法：

①对模型的参数--即化学成分，对其按照一定的数量（参数数据值的 1%）变动

②评价参数的变化对因变量预测的影响--发现几乎无影响（改变量为 0）

③观察得知变动幅度很小，灵敏度差、稳定度高

4.4 问题四模型的建立与求解

4.4.1 问题四模型基础

（1）皮尔逊相关系数

求皮尔逊相关系数矩阵 B 的方法（矩阵种某一位b_{ij}有 1<=i,j<=14），第 i 个化学成分为W_i（1<=i<=14），第 i 组数据的第 j 个化学成分归一化后的占比为a_{ij}（假定样本容量为 n，有 1<=i<=n,1<=j<=14），则$b_{ij}=\rho(W_i,W_j)=\frac{cov(W_i,W_j)}{\sigma_{W_i}\sigma_{W_j}}=\frac{E\left[(W_i-\mu_{W_i})(W_j-\mu_{W_j})\right]}{\sigma_{W_i}\sigma_{W_j}}$，估算样本协方差和标准差，可得到样本相关系数（这里是样

本皮尔逊系数）r：

$$r=\frac{\sum_{k=1}^{n}\left(a_{ki}-\frac{\sum_{t=1}^{n}a_{ti}}{n}\right)\left(a_{kj}-\frac{\sum_{t=1}^{n}a_{tj}}{n}\right)}{\sqrt{\sum_{k=1}^{n}\left(a_{ki}-\frac{\sum_{t=1}^{n}a_{ti}}{n}\right)^{2}}\sqrt{\sum_{k=1}^{n}\left(a_{kj}-\frac{\sum_{t=1}^{n}a_{tj}}{n}\right)^{2}}}$$

r 亦可由(a_{ki},a_{kj})样本点的标准分数均值估计，得到与上式等价的表达式：

$$r=\frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{a_{ki}-\frac{\sum_{t=1}^{n}a_{ti}}{n}}{\sigma_{W_i}}\right)\left(\frac{a_{kj}-\frac{\sum_{t=1}^{n}a_{tj}}{n}}{\sigma_{W_j}}\right)$$

其中$\frac{a_{ki}-\frac{\sum_{t=1}^{n}a_{ti}}{n}}{\sigma_{W_i}}$为对$W_i$的标准分数，$\frac{\sum_{t=1}^{n}a_{ti}}{n}$为$W_i$的各样本平均值，$\sigma_{W_i}$为$W_i$的样本标准差。

4.4.2 不同类别玻璃文物样化学成分间的关系及化学成分关系的差异性

我们首先对数据进行预处理，包括剔除无效数据（15、17）、对所有数据进行行归一化。分别计算高钾玻璃、铅钡玻璃各自化学成分之间的 Pearson(皮尔逊)相关系数（相关基础公式见问题二模型基础），并形成相关系数矩阵。详细的计算的过程由 python 程序（T4s1_pearson.py）完成。在计算出化学成分的相关系数矩阵之后，我们使用 python 的 heatmap 库可以绘制出相关系数热点图，方便较为直接地观察化学成分之间的关联关系。（实际意义：对于b_{ij}的每一个值 $\rho = b_{ij}$。当ρ=0 时，X 和 Y 不存在线性关系，即不相关；当 0<ρ<1 时，X 和 Y 正相关；当-1<ρ<0 时，X 和 Y 负相关；当|ρ|=1 时，X 和 Y 存在严格的线性关系，二者线性相关。）体现在热点图上即为，颜色越深，越倾向于负相关；颜色越浅，越倾向于正相关。

但是考虑到如果使用皮尔逊相关系数来刻画不同化学成分之间的关系，要求数据具有符合正态分布的特性。因此我们对每种化学成分W_i所对应的各样本数据进行正态性检验。我们使用 matlab 绘制 qq 图（Quantile-Quantile Plots）（程序见附录 T4s1qq.m）发现数据的正态性的偏差性较大，需要进一步使用定量算法刻画正态性。

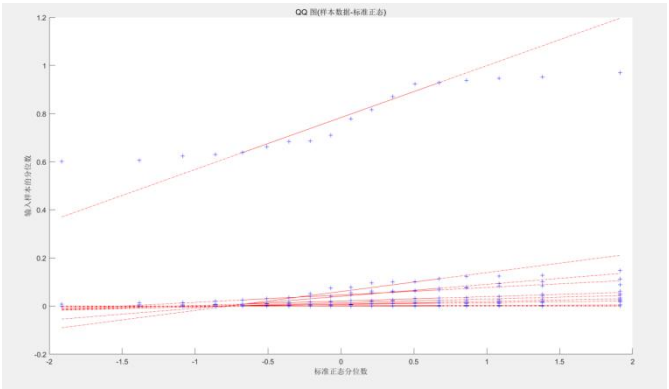

图 11 各化学成分的 qq 图（Quantile-Quantile Plots）

此外使用 Kolmogorov-Smirnov 算法（程序见附录

T4s1_Kolmogorov_Smirnov_check.m ） 与 Jarque-Bera 算 法 （ 程 序 见 附 录 T4s1_Jarque_Bera_check.m）检验数据集的正态分布特性。检验的结果如下：

表 8 算法检验结果

不符合正态性的数据组	高钾玻璃	铅钡玻璃
Jarque-Bera 法	2，7，9，10，11，12，13，14	2，3，4，5，6，7，8，10，11，13，14
Kolmogorov-Smirnov 法	2，9，10，12，13，14	2，3，5，7，8，10，11，13，14

我们可以发现，大部分的数据其实不符合正态性分布，所以使用 Pearson(皮尔逊)相关系数不能来很好地刻画它们的相关性。（因为符合的较差，Pearson(皮尔逊)相关系数热力图不再此处赘述，改为在附录中给出）因此我们采取对于正态性要求较低的斯皮尔曼（spearman）相关系数来刻画这种相关性，详细的计算的过程由 python 程序（附录 T4s1_spearman.py）完成。在计算出化学成分的相关系数矩阵之后，我们使用 python 的 heatmap 库可以绘制出相关系数热点图，方便较为直接地观察化学成分之间的关联关系。

由于高钾玻璃有 18 个样本，铅钡玻璃有 49 个样本，故对于高钾玻璃的相关性显著性检验采取直接查阅临界值表的方法，对于铅钡玻璃的显著性检验采取计算 p 值的方法。对于高钾玻璃，需要直接在相关系数矩阵种寻找r_s绝对值高于 0.4 的元素；而对于铅钡玻璃，则需要在 p 值矩阵种寻找 p 值小于 0.05 的元素。我们使用 p 值计算算法利用 matlab 编写程序（详见附录 T4s1_p.m）计算 p 值矩阵。下面给出根据上述规则筛选后的具体有效结果（显著性较低的值已经剔除，显示为空白）：

（1）高钾玻璃：

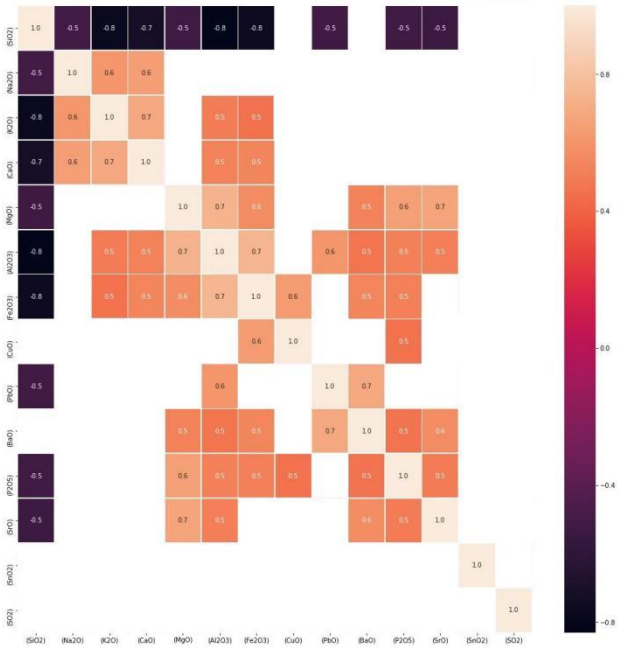

图 12 高钾玻璃相关系数热点图

针对高钾玻璃，二氧化硅(SiO2)与其余化学成分均呈负相关，且相关系数值

大多大于 0.5，因此二氧化硅(SiO2)与其余化学成分均的负相关性较强，除二氧化硅(SiO2)外其余各化学成分之间均为正相关，且相关系数均≥0.5，因此其余各化学成分间正相关性较为明显。（其中二氧化硅(SiO2)与氧化钾(K2O)氧化铁(Fe2O3)的负相关性最强，均为-0.8；氧化镁(MgO)与氧化铝(Al2O3)正相关性最强，为 0.7）。

（2）铅钡玻璃：

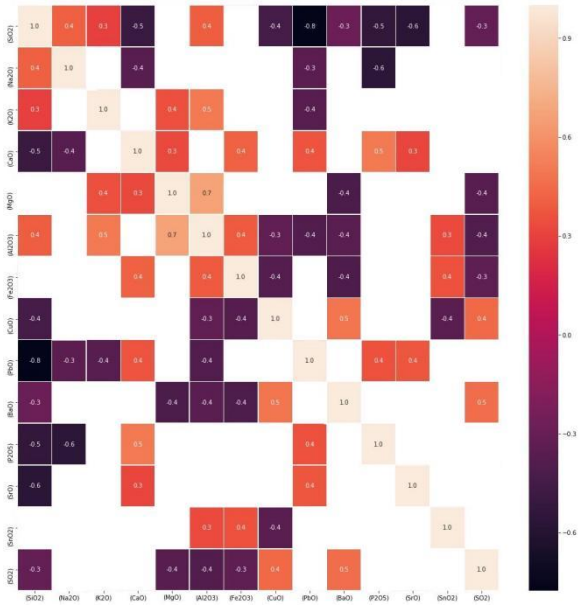

图 13 铅钡玻璃相关系数热点图

针对铅钡玻璃，各化学成分之间正负相关关系均存在，但二氧化硅(SiO2)与其余大多数化学成分有负相关关系。（其中二氧化硅(SiO2)与氧化铅(PbO)的负相关性最强，为-0.8；氧化镁(MgO)与氧化铝(Al2O3)正相关性最强，为 0.7）

结合分析上图，不同玻璃类别之间的化学成分关联关系的差异性在于：高钾玻璃相对于铅钡玻璃各化学成分间相关性较强，且除二氧化硅(SiO2)与其余化学成分均呈负相关外，其余各化学成分之间均为正相关，各化学成分相关关系较为简单；而铅钡玻璃各化学成分之间正负相关关系不一，各化学成分关系较为复杂。

（五）系统检验

针对问题二第一小问析高钾玻璃、铅钡玻璃的分类规律建立的数学模型，进行了模型检验。在化学成分选取上，在排除了关联性较小的分类影响因素之后，我们将剩下的因素作为重要影响因素，使用配对网格图的方式绘制二元的核密度图[3]以及散点图：

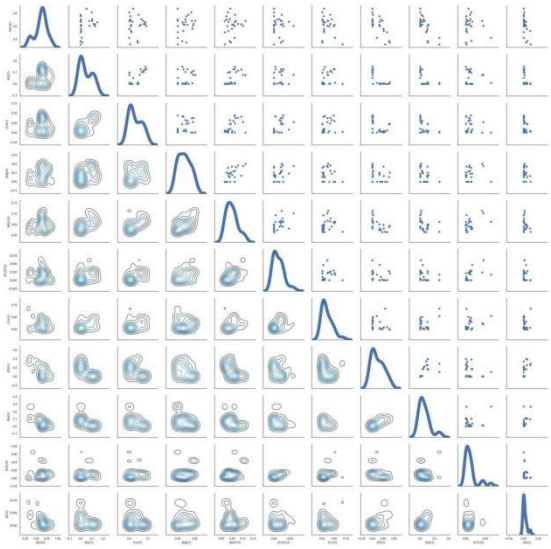

图 14 无风化二元核密度统计散点图

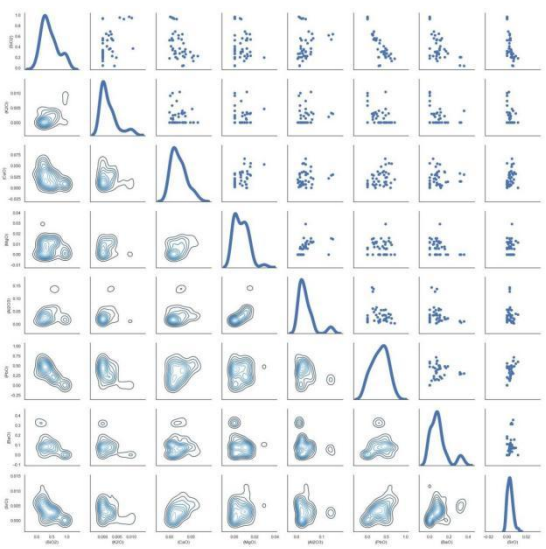

图 15 风化二元核密度统计散点图

通过以上两图，我们不难发现，剩下的影响因素彼此之间关联度较小，在散点图以及核密度统计图上未出现聚集的线性形状。可以认为他们彼此之间可以相互独立，因此建立的模型较为合理。

（六）系统评价

6.1 系统优点

（1）本文分析分析高钾玻璃、铅钡玻璃的分类规律时，在根得出函数关系的基础上，结合了已有文献进行分析，使分析的统计规律更具有现实性。

（2）在考虑玻璃类型与各种化学成分相关关系时，除了运用 Spearman 相关系数外，本文还结合了小提琴图进行相关性分析，保证结果的准确性。

（3）在用欧氏距离测度统计量时，本文在分析比较了内平方法、重心距离法和最短距离法后选择内平方法，全面考虑问题，保证合理性和准确性。

6.2 系统缺点

（1）问题三模型中运用到的随机森林算法，当样本数据过大时，程序运行时间较长。

（七）系统代码

由于之前代码主要解决的是部分问题，尚未形成完整的代码，故暂不整合完整系统代码，而给出各部分实现的代码

```
# T 2 S 1 (表示第 2 题第 1 小问，下同)
import pandas as pd
```

```
import xlrd2
# 打开 excel 表格
data_excel = xlrd2.open_workbook('T2s1.xlsx')
# 获取所有 sheet 名称
names = data_excel.sheet_names()
# 获取 book 中的 sheet 工作表,返回一个 xlrd.sheet.Sheet()对象
table = data_excel.sheets()[1]  # 通过索引顺序获取 sheet

# excel 工作表的行列操作
n_rows = table.nrows  # 获取该 sheet 中的有效行数
n_cols = table.ncols  # 获取该 sheet 中的有效列数
#cols_list = table.col(colx=0)  # 返回某列中所有的单元格对象组成的列表
y = table.col_values(0, start_rowx=0, end_rowx=None)
#print(y)
chemistrytype = ['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)']

for i in range(1,15):
    # 返回某列中所有单元格的数据组成的列表
    now = table.col_values(i, start_rowx=0, end_rowx=None)
    for j in range(0,len(now)):
        if now[j]=='':
            now[j]=0

    data = pd.DataFrame({
        '类型':y,
        '化学成分':now})
    #print(now)
    ans=data.corr(method='spearman')

    print(chemistrytype[i-1])
    print(ans.iloc[0][1])
    #print(type(ans))
```

```
# T 2 S 1 violin
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import time
#%matplotlib inline

data = pd.read_excel(r"C:\Users\y\Desktop\math-py\T2s1-violin.xlsx")
data.head()
y = data.diagnosis
x = data.drop(['diagnosis'],axis = 1 )
```

```
ax = sns.countplot(y,label="Count")            # M = 212, B = 357
B, M = y.value_counts()
print('高钾的数量: ',B)
print('铅贝的数量 : ',M)
x.describe()



data_dia = y
data = x
data_n_2 = (data - data.mean()) / (data.std())
data = pd.concat([y,data_n_2.iloc[:,0:10]],axis=1)
# 由于特征太多，我们每十个展示一次，前十个小提琴图
data = pd.melt(data,id_vars="diagnosis",
                                var_name="features",
                                value_name='value')
plt.figure(figsize=(10,10))
sns.violinplot(x="features", y="value", hue="diagnosis", data=data,split=True,
inner="quart")
plt.xticks(rotation=90)

data = pd.concat([y,data_n_2.iloc[:,10:15]],axis=1)
#20-最后
data = pd.melt(data,id_vars="diagnosis",
                                var_name="features",
                                value_name='value')
plt.figure(figsize=(10,10))
sns.violinplot(x="features", y="value", hue="diagnosis", data=data,split=True,
inner="quart")
plt.xticks(rotation=90)
```

```
# T 3 S 1 differentitation
#风化
import numpy as np
import pandas as pd

data = pd.read_excel(r"C:\Users\hxc\Desktop\t2s1-violin.xlsx",sheet_name=' 风
化')
data.head()
y = data.form
x = data.drop(['form'],axis = 1 )

drop_list1 = ['(Fe2O3)','(CuO)','(Na2O)','(SnO2)','(SO2)','(P2O5)']
x_1 = x.drop(drop_list1,axis = 1 )

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.metrics import f1_score,confusion_matrix
from sklearn.metrics import accuracy_score

# split data train 70 % and test 30 %
x_train, x_test, y_train, y_test = train_test_split(x_1, y, test_size=0.3, random_state=42)

#random forest classifier with n_estimators=10 (default)
clf_rf = RandomForestClassifier(random_state=43)
clr_rf = clf_rf.fit(x_train,y_train)

ac = accuracy_score(y_test,clf_rf.predict(x_test))
print('Accuracy is: ',ac)

cm = confusion_matrix(y_test,clf_rf.predict(x_test))
sns.heatmap(cm,annot=True,fmt="d")

datatest = pd.read_excel(r"C:\Users\hxc\Desktop\t3.xlsx")
datatest1=datatest.iloc[[2,5,6,7],[2,4,5,6,7,10,11,13]]
ddd=clf_rf.predict(datatest1)
```

```
# T 3 S 1 non_differentitation
#未风化
import numpy as np
import pandas as pd

data = pd.read_excel(r"C:\Users\hxc\Desktop\t2s1-violin.xlsx")
data.head()
y = data.form
x = data.drop(['form'],axis = 1 )

drop_list1 = ['(Na2O)','(SnO2)','(SO2)']
x_1 = x.drop(drop_list1,axis = 1 )

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import f1_score,confusion_matrix
from sklearn.metrics import accuracy_score

# split data train 70 % and test 30 %
x_train, x_test, y_train, y_test = train_test_split(x_1, y, test_size=0.3, random_state=42)

#random forest classifier with n_estimators=10 (default)
clf_rf = RandomForestClassifier(random_state=43)
clr_rf = clf_rf.fit(x_train,y_train)

ac = accuracy_score(y_test,clf_rf.predict(x_test))
print('Accuracy is: ',ac)
cm = confusion_matrix(y_test,clf_rf.predict(x_test))
sns.heatmap(cm,annot=True,fmt="d")
```

```
datatest = pd.read_excel(r"C:\Users\hxc\Desktop\t3.xlsx")
datatest1=datatest.iloc[[0,2,3,7],[2,4,5,6,7,8,9,10,11,13,14]]
ddd=clf_rf.predict(datatest1)
```

```
# T 3 S 1 sensitive
import numpy as np
import pandas as pd

data = pd.read_excel(r"C:\Users\hxc\Desktop\t2s1-violin.xlsx",sheet_name='风化')
data.head()
y = data.form
x = data.drop(['form'],axis = 1 )

drop_list1 = ['(Fe2O3)','(CuO)','(Na2O)','(SnO2)','(SO2)','(P2O5)']
x_1 = x.drop(drop_list1,axis = 1 )

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import f1_score,confusion_matrix
from sklearn.metrics import accuracy_score

# split data train 70 % and test 30 %
x_train, x_test, y_train, y_test = train_test_split(x_1, y, test_size=0.3, random_state=42)

#random forest classifier with n_estimators=10 (default)
clf_rf = RandomForestClassifier(random_state=43)
clr_rf = clf_rf.fit(x_train,y_train)

ac = accuracy_score(y_test,clf_rf.predict(x_test))
print('Accuracy is: ',ac)


endi=0.001
temp=0.0001
cr_num=clf_rf.predict(x_1)
tt=x_1
jieguo = [[0]*20]*8
nn=0
for i in range(8):
    n=0
    for j in np.arange(0,endi,temp):
        tt1=tt
        for l in range(len(cr_num)):
            tt1.iloc[l,i]=tt1.iloc[l,i]*(1+j)
        print(tt)
        tt1pr=clf_rf.predict(tt1)
        for l in range(len(cr_num)):
            if tt1pr[l]==cr_num[l]:
```

```
                    nn+=1
    mgx=nn/(endi/temp*8*len(cr_num))
```

```
    # T 4 S 1_pearson
    import numpy as np
    import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    import time
    #%matplotlib inline

    data = pd.read_excel(r"C:\Users\y\Desktop\高钾铅钡分开来（归一化）.xlsx",sheet_name='Sheet2')
    x=data
    f,ax = plt.subplots(figsize=(18, 18))
    sns.heatmap(x.corr(), annot=True, linewidths=.5, fmt= '.1f',ax=ax)
```

```
    # T 4 S 1 _spearman
    import numpy as np
    import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    import time
    #%matplotlib inline

    data = pd.read_excel(r"C:\Users\hxc\Desktop\高钾铅钡分开来（归一化）.xlsx",sheet_name='Sheet1')
    x=data
    f,ax = plt.subplots(figsize=(18, 18))
    sns.heatmap(x.corr(method='spearman'), annot=True, linewidths=.5, fmt= '.1f',ax=ax)
```

参考文献

[1]Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics,1987,20.

[2]刘学聪,李欣海.常用的生物统计方法及其R语言实现[J].应用昆虫学报,2021,58(01):220-232.

[3]李存华,孙志挥,陈耿,胡云.核密度估计及其在聚类算法构造中的应用[J].计算机研究与发展,2004(10):1712-1719.

[4]干福熹,承焕生,李青会.中国古代玻璃的起源——中国最早的古代玻璃研究[J].中国科学(E辑:技术科学),2007(03):382-391.

[5]张治国,马清林,海因兹·贝克,梅建军.中国古代人造硅酸铜钡颜料研究进展[J].中国文物科学研究,2011(04):45-49.