基于分类模型对古代玻璃制品的分析

摘　要

本文是通过建立卡方分析模型、有序逻辑回归模型、Fisher 线性判别分析模型、系统聚类模型、灰色关联分析模型和熵权—TOPSIS 模型来分析与玻璃有关的几个指标间和相关化学成分间的关系，并对未知玻璃类型进行预测。

对于问题一，本题首先对原始数据进行预处理，预测出缺失的颜色数据并剔除无效数据。然后分析各个指标对是否风化的影响和化学成分对风化影响的统计规律，最后预测出风化部分风化前的化学成分。针对问题的解决，这里会用到**卡方检验**根据检验结果的显著性得出不同指标对风化的判断的相关程度得出相对关系，再建立**有序逻辑回归模型**得出两个类别下玻璃的回归方程，从而得到统计规律，最后求出风化前后的均值之差和风化前的标准差得到随机数区间，从中取得随机数与风化后的数据相加所得结果即为预测值。最后的结果表明，纹饰和颜色与是否风化**相关度较低**，而类型与是否风化的**相关度较高一些**，得到的统计规律为两个回归方程与预测的结果均列在了问题一的模型求解当中。

对于问题二，本题通过 **Fisher 线性判别分析**得出了分类规律，并基于**系统聚类法**建立模型得到了聚类谱系图，对两个类型分别进行了亚类的划分。再通过分析划分亚类结果发现结果的划分中类内区别较小，类间有较大差别，肯定了分类的合理性。最后小范围内改变原始数据后再带入公式重新进行亚类划分，得到的聚类谱系图与原图差异较小，则说明本结论的**灵敏性较低**。

对于问题三，本题依然运用到了问题二中的 **Fisher 线性判别分析**来给未鉴别的玻璃制品进行划分类别，得出分类结果表后，根据测试组的预测结果可以看出其预测**正确率达到 100%**，可以说明分类结果准确性高。最后依旧是小范围内改变原始数据后再带入公式重新进行判别分析，发现分类结果相同，只是预测的概率略有不同。可以认为该结论**存在一定的灵敏性但是灵敏性不高**。

对于问题四，首先建立**灰色关联分析模型**将同一类别下不同化学成分进行纵向分析，得到各自之间的相关关系，然后通过**熵权—TOPSIS 法**将两个类别下同一种成分进行横向对比得到相关关系的差异性。灰色关联分析得到的关联系数表已列在问题四的模型求解中，其根据相关系数大小进行比较反映出了不同化学成分之间的相关关系。对于不同类型中同一成分的差异性比较则是按十四个成分的权重来考虑，因此我们总结了**“5+3+6”** 的形式，将其按权重从高到底来分成三部分概述他们之间相关关系的差异性。

关键词：卡方分析　有序逻辑回归　Fisher 线性判别分析　系统聚类　灰色关联分析　熵权—TOPSIS 法

一、 问题重述

1.1 问题背景

我国古代玻璃是吸收外来技术后在本地取材制作，化学成分与外来玻璃制品不相同。玻璃的主要原料为石英砂，其主要的化学成分是二氧化硅（SiO2）。因为纯石英砂的熔点较高，在炼制时需要添加助熔剂来降低熔化温度。在古代常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等，并且需要添加石灰石作为稳定剂，石灰石煅烧以后会转化为氧化钙（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通常加入铅矿石作为助熔剂，其中氧化铅（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。古代玻璃极易受其埋藏环境的影响而风化。在风化过程中，玻璃内部元素与环境中元素进行大量交换，导致其成分比例发生改变， 从而影响人们对其类别的正确判断。

现有一批我国古代玻璃制品的相关数据，考古工作者根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出现非 100%的情况。 本题将成分比例累加和介于 85%~105%之间的数据视为有效数据。

1.2 具体问题重述

问题 1： 对玻璃制品的表面风化与其玻璃类型、颜色和纹饰的关系进行分析，并结合玻璃的类型，分析文物样品表面是否风化的化学成分含量的统计规律， 再根据风化点分析数据，预测其风化前的化学成分含量。

问题 2：依据附件数据分析铅钡玻璃、高钾玻璃的分类规律；对于每个类别选择出合适的化学成分对其进行亚类划分，给出具体的划分方案及划分结果， 并对结果的合理性和敏感性进行分析。

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

问题 4：对于不同类别的玻璃文物样品，分析其化学成分之间所具有的关联关系，并比较不同类别之间的化学成分关联关系的差异性。

二、 问题分析

2.1 问题一的分析

该问题要求分析玻璃制品风化与其它指标间的关系，得出是否风化的统计关系，并根据风化点来预测风化前的化学含量。分析之前对附件的数据进行预处理，这里采用了采用有序逻辑回归的方法以颜色作为因变量，纹饰、类型和表面风化三个指标作为自变量建立回归方程来填充表单 1 中缺失的颜色；其次根据题目要求，删去成分比例累加和在 85%~105%之外的数据，从而得到有效数据以防止因为数据的原因对后续问题的分析产生影响。对于该问题，首先在分析风化与其他指标的关系时，对预处理后的数据进行卡方检验，根据检验结果的显著性得出不同指标对风化的判断的相关程度得出相对关系；之后根据玻璃的类型分为高钾和铅钡，基于有序逻辑回归以是否风化为因变量，表单 2 中十四个化学成分为自变量分别得出两个类别下

玻璃的回归方程，从而得到统计规律；最后分别求出两个类别中所有在风化点上面的所有化学物质的平均值x_1，对未风化点操作相同取x_2。令$u = x_1 - x_2$，并求得这个风化前的标准差m，然后在区间$[u-m, u+m]$中取随机数n,把每一个化学成分风化后的值与随机数相加得到的值就是预测出来的风化前的值。

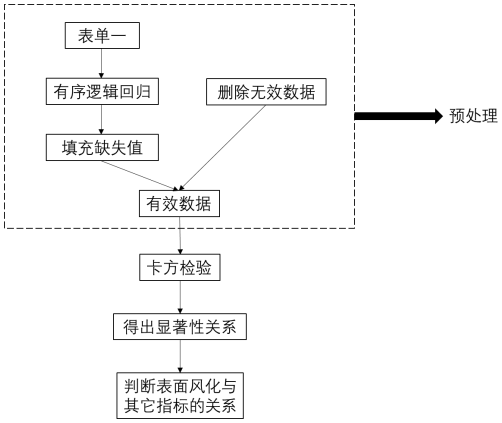

图 1 问题一第一部分思维导图

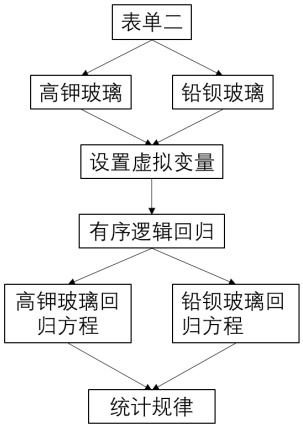

图 2 问题一第二部分思维导图

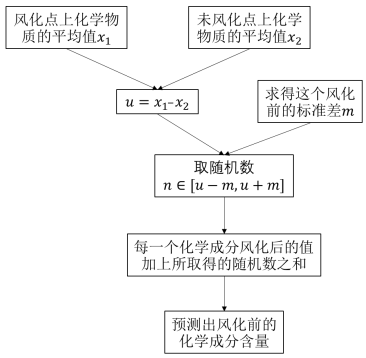

图 3 问题一第三部分思维导图

2.2 问题二的分析

该问题要求根据附件的数据分析铅钡玻璃、高钾玻璃的分类规律，并选择合适成分进行亚类的划分，最后再验证其合理性和灵敏度。所以首先针对铅钡和高钾的数据，由于原始数据之间难以区分，可以运用 Fisher 线性判别分析把数据投影至一维坐标。投影后使得同类样例的投影点尽可能接近和密集，异类投影点尽可能远离，这样可以使不同组别间的区别更加明显，便于得出分类规律；然后对于亚类的划分则需要用到系统聚类法将高钾和铅钡两个大类分别进行划分，划分依据则是分别计算玻璃数据中两类数据点间的距离，将最为接近的两类数据点进行组合划分，并反复迭代这一过程，直到将所有数据点合成一类，并生成聚类谱系图，从而划分出了所需要的亚类。最后再根据划分结果分析其合理性并进行灵敏度测试。

2.3 问题三的分析

该问题需要我们分析未鉴别的玻璃制品中的化学成分来鉴别其所属类型。这里也用到了问题二中的 Fisher 线性判别分析，然后再通过在合理范围内更改输入的数据数据进行灵敏度测试。

2.4 问题四的分析

本题需要我们先将同一类别下不同化学成分进行纵向分析，得到各自之间的相关关系，然后再将两个类别下同一种成分进行横向对比得到相关关系的差异性。首

先对于纵向分析，我们会用到灰色关联分析模型分别将不同化学成分提出来分析其与其它化学成分之间的相关关系。这里考虑到化学成分较多，所以将数据精简化取了最为主要的六个成分(二氧化硅、氧化钙、氧化铅、氧化钡、氧化纳、氧化钾)来进行分析，依次选取化学成分作为母序列，其余为子序列，算出关联系数后确定之间的相关关系；其次对于不同类别间的横向分析，需要先用熵权法算出 14 个化学成分的权重，再根据权重利用 TOPSIS 法得到不同成分的评分，假设评分越高越偏向于高钾，评分越低越偏向于铅钡。最后根据评分情况来解释重要性的大小关系，从而比较出差异。

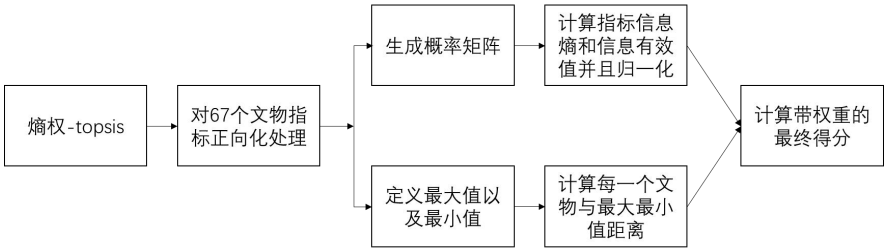

图 4 问题四思维导图

三、 模型假设

1. 假设所测出的实验数据允许存在一定范围内的误差。
2. 假设不考虑除了题中的十四个化学成分之外的化学成分影响。
3. 假设玻璃在出土后风化区域没有受到破坏，没有非正常情况的化学成分损失。
4. 假设成分累加和不为 100%的按系统误差处理。

四、 符号说明

表 1 符号说明表

符号	说明	单位
Df	卡方自由度	\
β	回归系数	\
s_1	类间距离	\
ω	法向量	\
s_2	类内距离	\
S	全局散度距离	\
J_ω	目标函数	\
d	平方欧氏距离	\
J	聚合系数	\
W_j	熵权	\
ω_i	权重	\
ρ	分辨系数	\

五、 模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 数据预处理

为了保证后面问题分析的合理性和准确性，现要将表单一中的数据预先处理，总共分为两步来进行：

1. 基于有序逻辑回归的方法，以颜色作为因变量，纹饰、类型和表面风化三个指标作为自变量建立回归方程。

1) 设置虚拟变量

在有序逻辑回归中，因变量可能有多个元素，为了能将其进行逻辑回归，通常都是将其拆分为多个二元逻辑回归。所以这里设置虚拟变量时先假设某个元素为 1，再将其余元素先设为 0，待回归完成后在剩余元素中重新设置虚拟变量直至每个元素可以被辨别出来。

2) 进行逻辑回归分析

以纹饰、类型和表面风化三个指标作为自变量利用 SPSS 软件进行分析得到回归结果如下表：

表 2 回归结果参数表

	估算	标准错误	瓦尔德	自由度	显著性	95% 置信区间	
						下限	上限
[纹饰=A]	-1.9699	0.5870	11.2620	1.0000	0.0008	-3.1203	-0.8194
[纹饰=B]	-0.9985	1.1935	0.6999	1.0000	0.4028	-3.3377	1.3407
[纹饰=C]	0[a]			0.0000			
[类型=高钾]	-2.2412	0.8091	7.6733	1.0000	0.0056	-3.8269	-0.6554
[类型=铅钡]	0[a]			0.0000			
[表面风化=风化]	-1.5174	0.6617	5.2581	1.0000	0.0218	-2.8143	-0.2204
[表面风化=无风化]	0[a]			0.0000			

表 3 预测结果统计表

原始数据(1-20)	预测数据	原始数据(21-40)	预测数据	原始数据(41-58)	预测数据
蓝绿	浅蓝	蓝绿	蓝绿	浅绿	浅蓝
浅蓝	蓝绿	蓝绿	蓝绿	浅蓝	蓝绿
蓝绿	蓝绿	蓝绿	蓝绿	浅蓝	浅蓝
蓝绿	蓝绿	紫	深绿	浅蓝	蓝绿
蓝绿	蓝绿	浅蓝	浅蓝	浅蓝	浅蓝
蓝绿	蓝绿	紫	浅蓝	浅蓝	浅蓝
蓝绿	蓝绿	蓝绿	蓝绿	浅蓝	浅蓝
紫	浅蓝	浅蓝	蓝绿		蓝绿
蓝绿	蓝绿	浅蓝	蓝绿	黑	蓝绿
蓝绿	蓝绿	深蓝	浅蓝	黑	蓝绿

浅蓝	浅蓝	紫	深绿	浅蓝	浅蓝
蓝绿	蓝绿	浅绿	深绿	浅蓝	浅蓝
浅蓝	浅蓝	深绿	深绿	浅蓝	蓝绿
深绿	浅蓝	深绿	浅蓝	浅蓝	浅蓝
浅蓝	浅蓝	浅绿	深绿	绿	深绿
浅蓝	浅蓝	深绿	浅蓝	蓝绿	浅蓝
浅蓝	浅蓝	深绿	深绿	蓝绿	浅蓝
深蓝	蓝绿	深绿	浅蓝		浅蓝
	蓝绿	深绿	浅蓝		
浅蓝	浅蓝		浅蓝		

3) 查验结果正确率及显著性

根据表三结果可知 19 号预测结果为蓝绿、40 号预测结果为浅蓝、48 号结果预测为蓝绿、58 号预测结果为浅蓝。通过预测结果与原始数据进行分析发现预测的成功率为 51.85%，该结果相对来说较为可以接受。在表二的结果中每一个自变量中都选择了一个元素作为参考来得出其它元素的系数。我们可以看出大部分元素的显著性都小于 0.05，满足 95%的置信区间，由此我们可以判定这些自变量与因变量的相关性较大，所预测出的结果也相对准确。

2. 得到补全过的表单一后，接下来要根据题目中所给出的条件来剔除表单一中成分比例累加和不在 85%—105%区间内的数据，以此来得到有效数据进行更深一步的分析。通过分析发现 15 号和 17 号的成分比例累加和均小于 85%，因此我们选择剔除这两组数据后来得到有效数据。

5.1.2 模型的建立

1. 基于卡方检验判断指标间的关系

本部分要解决的是判断玻璃制品的表面风化与其玻璃类型、颜色和纹饰的关系。而卡方检验就是通过分析指标之间实际观测值和期望值之间的偏离程度得到一个卡方值来判断各指标间的相关性。卡方值越小，对应来说的偏差值就越小，实际值就更趋于符合期望值。所以就可以通过计算玻璃制品表面风化与其它各指标之间的卡方值来判断它们之间的相关性，从而得到它们之间的关系。

1) 对于本题我们可以先列出原假设(玻璃制品的表面风化与其玻璃类型、颜色和纹饰之间相互独立)和备择假设(玻璃制品的表面风化与其玻璃类型、颜色和纹饰之间具有相关性)

2) 然后将预处理之后的表单一中各指标的数量统计起来，按列联表的方式放入表格中。然后根据期望值的计算公式得出期望值并放入表格中：

$$Expected = \frac{RowTotals * ColumnTotals}{Totals} \tag{1}$$

即期望值等于行求和乘以列求和除以总和

3) 最后根据所得出的实际观测值表和期望值表带入卡方的计算公式中得到计算结果：

$$\chi^2 = \frac{(Observed - Expected)^2}{Expected} \tag{2}$$

即卡方(χ^2)等于观测值减期望值的差的平方除以期望值

4) 得到卡方值后根据卡方的自由度计算公式得出自由度，然后对照卡方界

值表判断是否拒绝原假设来得出各指标间的关系。

$$Df = (a-1)*(b-1) \tag{3}$$

其中Df为卡方自由度，a 和 b 分别为检验条件的分类数，本文中就是各指标与表面风化的分类数。

2. 利用有序逻辑回归分析化学成分含量的统计规律

本部分要解决的是结合玻璃的类型，分析文物样品表面是否风化的化学成分含量的统计规律。逻辑回归就是通过分析多个自变量的值将因变量进行分类从而判断出因变量的类别。本题分为高钾玻璃和铅钡玻璃两个类别分别来进行对是否风化的判断来得到一个回归方程，即为统计规律。

1) 首先要进行数据预处理，设置虚拟变量。我们关注到在风化的类型中还存在严重风化的情况，我们将此单独讨论。对于高钾玻璃来说，设置未风化为 0，风化为 1；对于铅钡玻璃来说，设置未风化为 0，风化为 1，严重风化为二。所以对于高钾玻璃类别来说就是逻辑回归的二分类，而铅钡玻璃就涉及到了有序逻辑回归，可以分成两个二分类的逻辑回归来解决。

2) 接下来要建立有序逻辑回归模型，所利用的回归公式为：

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \mu_i \tag{4}$$

也可以写成向量乘积形式：

$$y_i = \boldsymbol{x_i'\beta} + \mu_i (i = 1, 2, \ldots, n) \tag{5}$$

其中y_i为被解释变量，x_i为解释变量，β为回归系数。因为预测结果可能会出现$\hat{y}_i$<0 或$\hat{y}_i$>1 的情况，所以需要引入函数$F(x,\beta)$将解释变量x与被解释变量y_i连接起来，通常$F(x,\beta)$取$Sigmoid$函数：

$$F(x,\beta) = S(\boldsymbol{x_i'\beta}) = \frac{\exp(\boldsymbol{x_i'\beta})}{1+\exp(\boldsymbol{x_i'\beta})} \tag{6}$$

该函数的值域为[0,1]，于是在给定x的情况下，我们考虑y的两点分布：

$$P(y=1|x) = F(x,\beta) \tag{7}$$

$$P(y=0|x) = 1 - F(x,\beta) \tag{8}$$

因为

$$E(y|x) = 1*P(y=1|x) + 0*P(y=0|x) = P(y=1|x) \tag{9}$$

所以我们可以将$\hat{y}$理解为‘$y=1$’发生的概率

$$P(y_i=1|x) = S\left(x_i'\hat{\beta}\right) = \frac{\exp\left(x_i'\hat{\beta}\right)}{1+\exp\left(x_i'\hat{\beta}\right)} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\ldots+\hat{\beta}_k x_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\ldots+\hat{\beta}_k x_{ki}}} \tag{10}$$

如果$\hat{y}_i \geq 0.5$，则认为其预测的$y=1$；否则则认为其预测的$y=0$

3) 在通过 SPSS 软件分别对两个类型玻璃的数据进行逻辑回归后，得到了逻辑回归系数表和预测的成功率表格。根据成功率可以得知该预测的准确性，然后从逻辑回归系数表中可以得到回归系数并将其带入回归公式从而得到回归方程，这就是统计规律。

3. 基于随机取差值法来对风化前的化学成分含量进行预测

1) **首先**求出两个类别中所有在风化点上面的所有化学物质的平均值x_1，对未风化点操作相同取x_2。令$u = x_1 - x_2$，并求得这个风化前的标准差m。
2) 其次在区间$[u-m, u+m]$中取随机数n,把每一个化学成分风化后的值与随机数相加得到的值就是预测出来的风化前化学成分含量的值。

5.1.3 模型的求解

1. 卡方检验求解过程

1) 首先设立
原假设：玻璃制品的表面风化与其玻璃类型、颜色和纹饰之间相互独立。
备择假设：玻璃制品的表面风化与其玻璃类型、颜色和纹饰之间具有相关性。
2) 根据表单一的数据得到下列实际观测值表：

表 4 颜色的实际观测值表

	风化	无风化	总计
蓝绿	9	6	15
浅蓝	16	8	24
紫	2	2	4
深绿	4	3	7
深蓝	0	2	2
黑	2	0	2
绿	0	1	1
浅绿	1	2	3
总计	34	24	58

表 5 纹饰的实际观测值表

	风化	无风化	总计
A	11	11	22
B	6	0	6
C	17	13	30
总计	34	24	58

表 6 类型的实际观测值表

	风化	无风化	总计
高钾	6	12	18
铅钡	28	12	40
总计	34	24	58

根据三个表的数据和公式一可得出三个指标的期望值表：

表 7 颜色的期望值表

	风化	无风化	总计
蓝绿	8.79	6.21	15.00
浅蓝	14.07	9.93	24.00
紫	2.34	1.66	4.00
深绿	4.10	2.90	7.00
深蓝	1.17	0.83	2.00
黑	1.17	0.83	2.00

绿	0.59	0.41	1.00
浅绿	1.76	1.24	3.00
总计	34.00	24.00	58.00

表 8 纹饰的期望值表

	风化	无风化	总计
A	12.90	9.10	22.00
B	3.52	2.48	6.00
C	17.59	12.41	30.00
总计	34.00	24.00	58.00

表 9 类型的期望值表

	风化	无风化	总计
高钾	10.55	7.45	18.00
铅钡	23.45	16.55	40.00
总计	34.00	24.00	58.00

3) 根据实际观测值表和期望值表的数据以及公式二、三我们可以算出每个指标表面风化的卡方值和自由度如下表所示：

表 10 不同指标表面风化的卡方值和自由度表

	卡方值	自由度	渐进显著性（双侧）
纹饰	4.957a	2.000	0.084
类型	6.880a	1.000	0.009
颜色	7.234[a]	7.000	0.405

4) 对于纹饰来说，其显著性水平为 0.05、自由度为 2 的情况下的临界值为 5.991，而实际的卡方值为 4.957 小于其临界值，所以我们认为其无法拒绝原假设，即纹饰与表面风化之间相关程度较小；对于类型来说，其显著性水平为 0.05、自由度为 1 的情况下的临界值为 3.841，而实际的卡方值为 6.880 大于其临界值，所以我们认为其可以拒绝原假设，即类型与表面风化之间有较强的相关性；对于颜色来说，其显著性水平为 0.05、自由度为 7 的情况下的临界值为 14.067，而实际的卡方值为 7.234 小于其临界值，所以我们认为其无法拒绝原假设，即颜色与表面风化之间相关性较小。

2. 有序逻辑回归模型求解过程

1) 对于高钾玻璃，先进行虚拟变量的设置，将未风化为 0，风化为 1；而对于铅钡玻璃则先设未风化为 0，没有未风化的为 1。得出结果后，将风化的设为 1，严重风化的设为 2 再进行逻辑回归。并得出最终结果

2) 运用 SPSS 软件将表单二中数据导入得到逻辑回归系数表如下：

表 11 铅钡玻璃的逻辑回归系数表

	Beta	显著性		Beta	显著性
二氧化硅 SiO2	1.071	0.051	氧化铅 PbO	0.731	0.138
氧化钠 Na2O	0.597	0.402	氧化钡 BaO	1.746	0.041
氧化钾 K2O	7.109	0.013	五氧化二磷 P2O5	2.434	0.022
氧化钙 CaO	-1.35	0.098	氧化锶 SrO	11.763	0.007

氧化镁 MgO	-1.146	0.582	氧化锡 SnO2	11.399	0.065
氧化铝 Al2O3	-0.792	0.285	二氧化硫 SO2	0.603	0.274
氧化铁 Fe2O3	1.824	0.072	常量	97.439	0.066
氧化铜 CuO	-1.809	0.017			

表 12 高钾玻璃的逻辑回归系数表

	Beta	显著性		Beta	显著性
(常量)		0.356	氧化铜(CuO)	-0.884	0.151
二氧化硅(SiO2)	-3.528	0.353	氧化铅(PbO)	-0.385	0.133
氧化钠(Na2O)	-0.659	0.192	氧化钡(BaO)	0.960	0.097
氧化钾(K2O)	-1.325	0.401	五氧化二磷(P2O5)	0.276	0.539
氧化钙(CaO)	-0.039	0.964	氧化锶(SrO)	-0.144	0.642
氧化镁(MgO)	-1.849	0.049	氧化锡(SnO2)	0.684	0.078
氧化铝(Al2O3)	0.579	0.509	二氧化硫(SO2)	0.692	0.151
氧化铁(Fe2O3)	-0.808	0.233			

从系数表中可以得出两个回归方程中不同项的回归系数，从而得到了回归方程，即为统计规律。在此为方便公式的表达，将回归系数进行转至，即

$$\beta_{铅钡} = (97.439,1.071,0.597,7.109,-1.35,-1.146,-0.792,1.824,-1.809,0.731,1.746,2.434,11.763,11.399,0.603)$$

$$x_{铅钡} = x_{高钾} = (SiO2, Na2O, K2O, CaO, MgO, Al2O3, Fe2O3, CuO, PbO, BaO, P2O5, SrO, SnO2, SO2)$$

$$\beta_{高钾} = (0,-3.528,-0.659,-1.325,-0.039,-1.849,0.579,-0.808,-0.884,-0.385,0.960,0.276,-0.144,0.684,0.692)$$

铅钡回归方程：

$$\hat{y}_i = \frac{e^{\beta_{铅钡} x_{铅钡}}}{1 + e^{\beta_{铅钡} x_{铅钡}}} \tag{1}$$

高钾回归方程：

$$\hat{y}_i = \frac{e^{\beta_{高钾} x_{高钾}}}{1 + e^{\beta_{高钾} x_{高钾}}} \tag{2}$$

3. 随机取差值对化学成分含量进行预测求解过程

1) 求出两个类别中所有在风化点上面的所有化学物质的平均值x_1和未风化点上面的所有化学物质的平均值x_2如下图：

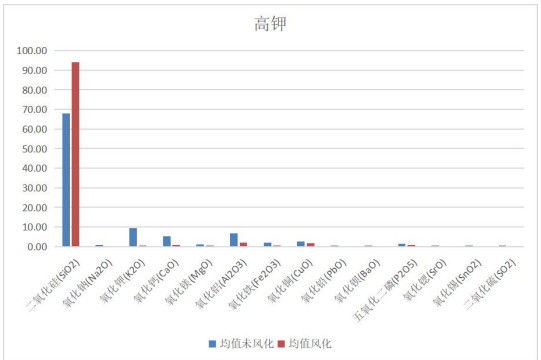

图 5 高钾化学物质风化前后平均值图

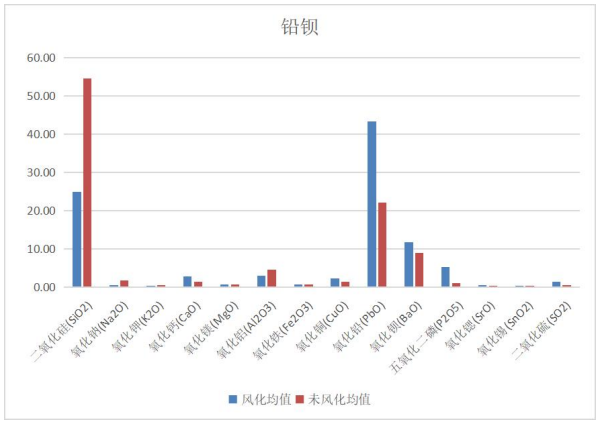

图 6 铅钡化学物质风化前后平均值图

2) 分别求得平均之后，可以得到$u = x_1 - x_2$的值，再算出每个化学物质风化前的标准差m，就可以得到不同的随机数n的取值区间$[u-m, u+m]$如下表所示：

表 13 均值、标准差和取值区间表

	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)
均值未风化	67.984	0.695	9.331	5.333	1.079	6.620	1.932
均值风化	93.963	0.000	0.543	0.870	0.197	1.930	0.265
均值差	25.979	-0.695	-8.788	-4.463	-0.883	-4.690	-1.667
方差未风化	70.264	1.518	14.088	8.766	0.419	5.690	2.546
方差风化	2.505	0.000	0.165	0.198	0.078	0.775	0.004
标准差未风化	8.382	1.232	3.753	2.961	0.647	2.385	1.596
标准差风化	1.583	0.000	0.406	0.445	0.280	0.880	0.063
区间下限	17.600	-1.930	-12.540	-7.420	-1.530	-7.080	-3.270
区间上限	34.360	0.530	-5.040	-1.500	-0.230	-2.300	-0.070
	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)	氧化锶(SrO)	氧化锡(SnO_2)	二氧化硫(SO_2)
均值未风化	2.453	0.412	0.598	1.403	0.042	0.197	0.102
均值风化	1.562	0.000	0.000	0.280	0.000	0.000	0.000
均值差	-0.891	-0.412	-0.598	-1.123	-0.042	-0.197	-0.102
方差未风化	2.526	0.318	0.884	1.885	0.002	0.425	0.032
方差风化	0.728	0.000	0.000	0.037	0.000	0.000	0.000
标准差未风化	1.589	0.564	0.940	1.373	0.046	0.652	0.178
标准差风化	0.853	0.000	0.000	0.192	0.000	0.000	0.000
区间下限	-2.480	-0.970	-1.540	-2.490	-0.090	-0.850	-0.280
区间上限	0.700	0.150	0.340	0.250	0.010	0.450	0.080

3) 最后将化学物质风化后的值与所取得随机数相加得到预测出来的风化前化学成分含量的值，具体结果如下表所示：

表 14 高钾预测结果图(部分)

文物采样点	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)
01	69.33	0	9.99	6.32	0.87	3.93	1.74
03 部位 1	87.05	0	5.19	2.01	0	4.06	0
03 部位 2	61.71	0	12.37	5.87	1.11	5.5	2.16
04	65.88	0	9.67	7.12	1.56	6.44	2.06

05	61.58	0	10.95	7.35	1.77	7.5	2.62
随机数	-18	0	12	4	1	3	1

文物采样点	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)	氧化锶(SrO)	氧化锡(SnO_2)	二氧化硫(SO_2)
01	3.87	0	0	1.17	0	0	0.39
03 部位 1	0.78	0.25	0	0.66	0	0	0
03 部位 2	5.09	1.41	2.86	0.7	0.1	0	0
04	2.18	0	0	0.79	0	0	0.36
05	3.27	0	0	0.94	0.06	0	0.47
随机数	1	0	1	1	0	0	0

表 15 铅钡预测结果图(部分)

文物采样点	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)
02	73.28	2.00	1.05	1.34	1.18	7.73	2.86
08	57.14	2.00	0.00	0.48	0.00	3.34	1.00
08 严重风化点	41.61	2.00	0.00	2.19	0.00	3.11	1.00
11	70.59	2.00	0.21	2.51	0.71	4.69	1.00
19	66.64	2.00	0.00	1.93	0.59	5.57	2.33
随机数	37	2	0	−1	0	2	1

文物采样点	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)	氧化锶(SrO)	氧化锡(SnO_2)	二氧化硫(SO_2)
02	1.26	33.43	0.00	0.57	0.19	0.00	0.00
08	11.41	14.68	27.23	0.59	0.37	0.00	1.58
08 严重风化点	4.14	18.45	26.62	4.56	0.53	0.00	14.03
11	5.93	11.39	10.61	6.38	0.37	0.00	0.00
19	4.51	28.82	1.35	5.83	0.19	0.00	0.00
随机数	1	−14	−4	−3	0	0	−1

上表是预测结果的部分表格，我们取了十组数据作为展示，详细预测结果放在了支撑材料的“第一题预测”Excel 表格中。这里通过取平均值的方法更容易得到较为准确的风化后的数值，再从区间中取随机数相加则可以更好的预测出风化前的化学成分含量。

5.2 问题二模型的建立与求解

5.2.1 模型的建立

1. 运用 Fisher 线性判别分析得出分类规律

本部分要解决的就是根据化学成分含量信息来对类型这个指标进行分类的结果，从中选取区分度大的信息来得到分类规律。这里可以用到 Fisher 线性判别分析，这个方法可以运用投影的方法，把原始玻璃数据的点向一条直线上投影从而进行判断。在原来的坐标系下，可能很难把样品相互分开，而投影后就可能区别明显，方便进行分类规律的寻找。具体步骤如下：

1) 首先将数据中所有点分隔在超平面$\omega^r x = 0$ 两侧，即将每个点投影到ω这个法向量上，要保证类内小，类间大(有内聚，松耦合)。

2) 其次，我们定义类间距离:

$$s_1 = \omega^T(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^T\omega \tag{1}$$

$$\bar{x}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} x_i \tag{2}$$

$$\bar{x}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} x_i \tag{3}$$

类内距离:

$$s_2 = \omega^T(s_{c_1} + s_{c_2})\omega \tag{4}$$

$$s_{c_1} = \frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \bar{x}_1)(x_i - \bar{x}_1)^T \tag{5}$$

$$s_{c_2} = \frac{1}{N_2}\sum_{i=1}^{N_2}(x_i - \bar{x}_2)(x_i - \bar{x}_2)^T \tag{6}$$

式中N是样本的个数，x是样本各指标的值，定义全局散度距离 $S = s_1 + s_2$

现在有了类内距离和类间距离，我们可以定义目标函数:

$$J_\omega = \frac{\text{类间距}}{\text{类内距}} = \frac{\omega^T(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^T\omega}{\omega^T(s_{c_1} + s_{c_2})\omega} \tag{7}$$

要使目标函数最大，应对其求偏导，使得:

$$\frac{\partial J_\omega}{\partial \omega} = 0 \tag{8}$$

3) 最后根据得到的分类函数系数表判断出对于高钾、铅钡分类过程中影响分类结果较大的化学成分，并把系数带入判别函数中比较得出的函数值，哪个函数值较大就将该样品归于哪一类。这就是对于高钾和铅钡两个类别进行分类的规律。

2. 基于系统聚类法对亚类进行划分

本部分要求从高钾、铅钡中分别选择出合适的化学成分对其进行亚类划分，这里可以用系统聚类法通过计算原始数据中两类数据点间的距离，将最为接近的两类数据点进行组合，并反复迭代这一过程，直到将所有数据点合成一类，并生成聚类谱系图。接下来从聚类谱系图中就可以得到所需要划分的亚类，在进行聚类分析时，我们选取了七个化学成分（氧化锶、氧化镁、五氧化二磷、氧化钡、氧化钙、氧化钾、氧化铅），一部分选自得出分类规律时得到的五个化学成分，另外两个考虑到玻璃的主要成分进行选择，分析具体过程如下：

1) 将每个数据点看成一类，计算出两两之间的距离，距离计算方式如下：

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{9}$$

其中d是平方欧氏距离，x是数据点的横坐标，y是数据点的纵坐标。

2) 将所有距离数据进行分析，将距离最小的两类合并为一类后重新计算新的数据间的距离直至全部数据合并为一个大类。

3) 最后需要判断聚类数量，可以用到肘部法则。首先引入各个类畸变程度之和的概念，即各个类的畸变程度等于该类重心与其内部成员位置距离的平方和，假设一共将 n 个样本划分到 K 个类中(K≤n－1,即至少有一类中有两个元素)，用 C,表示第 k 个类(k=1,2,…,K)，且该类重心的位置记为u_k，那么第 k 个类的畸变程度为:

$$\sum_{i\in c_k} |x_i - u_k|^2 \tag{10}$$

定义所有类的总畸变程度(聚合系数):

$$J = \sum_{k=1}^{K} \sum_{i\in c_k} |x_i - u_k|^2 \tag{11}$$

根据此公式可画出聚合系数折线图，根据折线变化趋势就可得到聚类的数量。

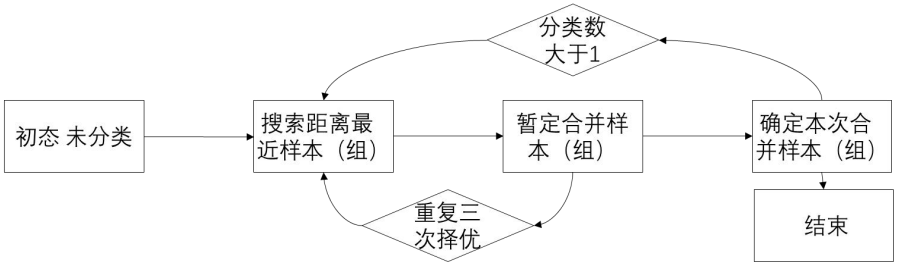

图 7 算法流程图

5. 2. 2 模型的求解

1. Fisher 线性判别分析求解过程

1) 由所给数据可以得出ω法向量的典则判别函数系数表：

表 16 典则判别函数系数表

化学成分	系数	化学成分	系数
二氧化硅(SiO2)	0.043	氧化铅(PbO)	0.118
氧化钠(Na2O)	0.32	氧化钡(BaO)	0.216
氧化钾(K2O)	-0.188	五氧化二磷(P2O5)	0.095
氧化钙(CaO)	0.023	氧化锶(SrO)	-0.608
氧化镁(MgO)	0.204	氧化锡(SnO2)	0.015
氧化铝(Al2O3)	0.209	二氧化硫(SO2)	-0.102
氧化铁(Fe2O3)	0.098	(常量)	-7.377
氧化铜(CuO)	-0.182		

这是所有数据所需投影到的一维直线，在这上面数据之间的区别将会更明显，便于判断出分类规律

2) 根据对目标函数的求解我们可以得到分类函数系数：

表 17 分类函数系数表

	铅钡	高钾		铅钡	高钾
二氧化硅(SiO2)	32.691	32.462	氧化铅(PbO)	33.769	33.141
氧化钠(Na2O)	37.930	36.224	氧化钡(BaO)	42.620	41.471
氧化钾(K2O)	35.124	36.128	五氧化二磷(P2O5)	42.962	42.455
氧化钙(CaO)	38.980	38.859	氧化锶(SrO)	-78.006	-74.769
氧化镁(MgO)	49.418	48.332	氧化锡(SnO2)	24.316	24.235
氧化铝(Al2O3)	28.135	27.023	二氧化硫(SO2)	13.745	14.288

氧化铁(Fe2O3)	32.963	32.441	(常量)	-1639.053	-1606.316
氧化铜(CuO)	15.343	16.315			

对于不同化学成分的重要性判断我们可以比较他们系数中绝对值的大小，在此取 5 个影响较大的化学成分，高钾(氧化锶、氧化镁、五氧化二磷、氧化钡、氧化钙)和铅钡(氧化锶、氧化镁、五氧化二磷、氧化钡、氧化钙)。之后将数据点带入判别函数中比较得出的函数值，哪个函数值较大就将该样品归于哪一类。

2. 系统聚类法求解过程

1) 将每个数据点看为一类，算出其距离后再将距离最小的合为新的一类，再通过不断迭代得到了聚类谱系图如下：

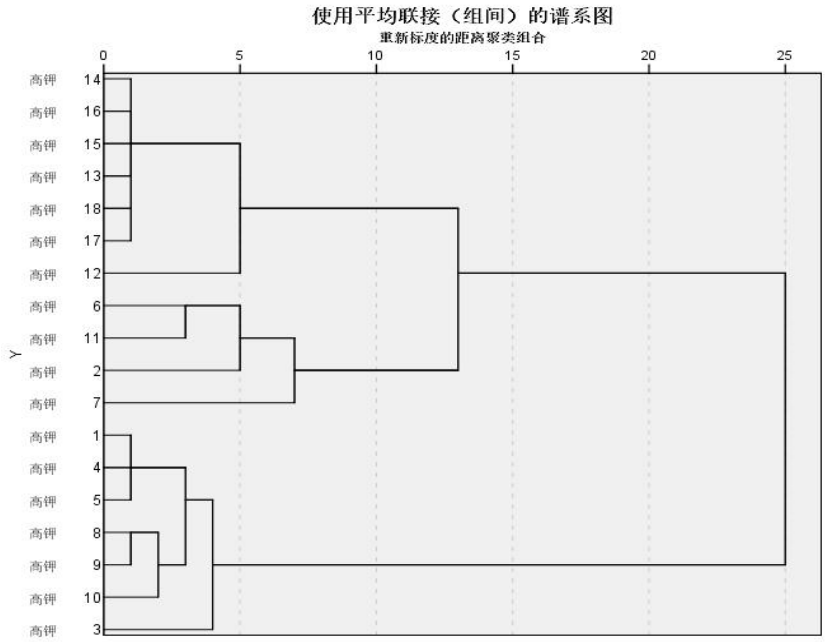

图 8 高钾聚类谱系图

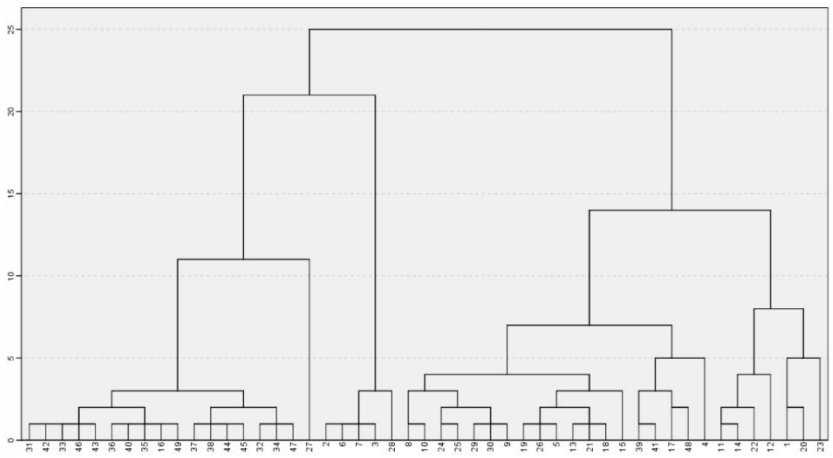

图 9 铅钡聚类谱系图

解决问题所需的亚类就是根据谱系图来划分的。

2) 接下来需要判断聚类数量来确定合适的亚类划分，这里根据肘形法则得到的聚合系数折线图如下：

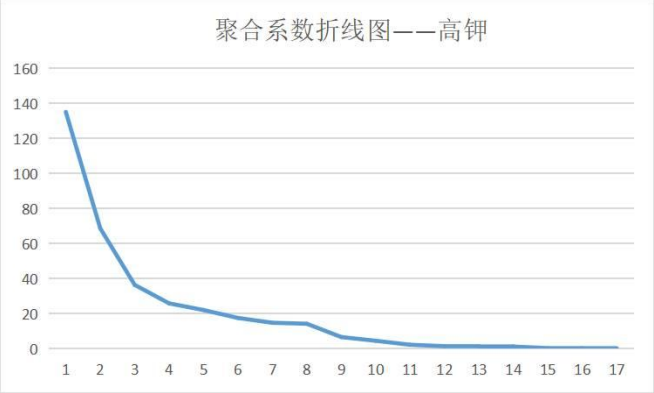

图 10 高钾的聚合系数折线图

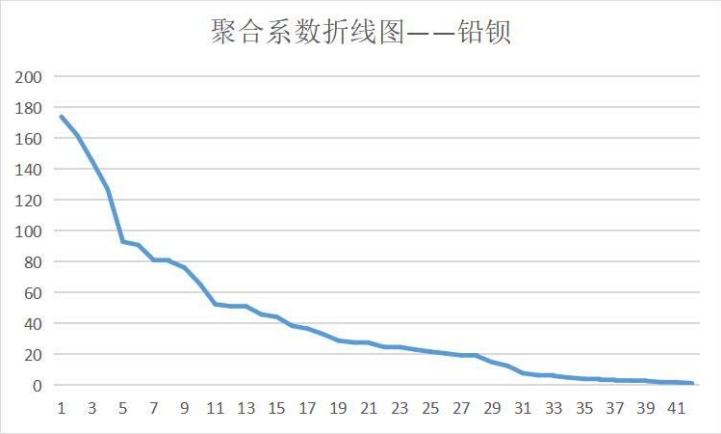

图 11 铅钡的聚合系数折线图

根据聚合系数折线图可知，

- 对于高钾来说，当类别数为 3 时，折线的下降趋势趋缓，故可将类别数设定为 3；对于铅钡来说，当类别数为 4 时，折线的下降趋势趋缓，故可将类别数设定为 4。
- 对于高钾来说，从图中可以看出 K 值从 1 到 3 时，畸变程度变化最大。超过 3 以后，畸变程度变化显著降低。因此肘部就是 K=3，故可将类别数设定为 3；对于铅钡来说，从图中可以看出 K 值从 1 到 4 时，畸变程度变化最大。超过 4 以后，畸变程度变化显著降低。因此肘部就是 K=4，故可将类别数设定为 4。

3. **合理性分析**(分类结果放在了“亚类划分结果”Excel 表格中)

1) 对于高钾的亚分类来说，分为了三类。在纹饰方面，三组各有侧重总体来说是按 A、B、C 进行分类；在风化程度方面，第一二组均为未风化，第三组基本为风化。可以看出该分类结果的类内区别较小，类间有较大差别，因此分为三个类别是相对合理的。

2) 对于铅钡的亚分类来说，分为了四类。在纹饰方面，除了第一组中 A 类占大部分，其余三组均含 C 类较多；在风化程度方面，第一组风化程度类内差异较小，风化个体与未风化个体各占一半，而其余三组基本为风化。在颜色方面，第一、四组以浅蓝为主导，第二组颜色倾向于深色系，第三组全为紫色，可以看出该分类结果的类内区别较小，类间有较大差别，因此分为四个类别是相对合理的。

4. **灵敏性分析**

这里在原始数据的±5%内设立一个区间，在区间内取随机数后组成新的数据带入到系统聚类的公式中重新迭代得到新的聚类谱系图如下：

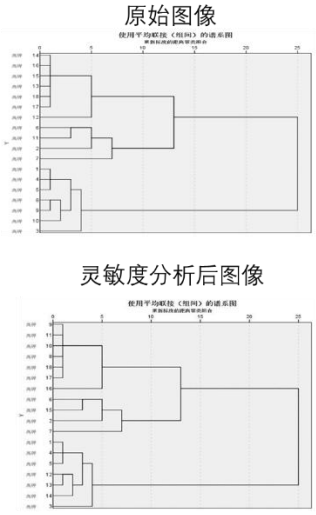

图 12 原始图像与灵敏度分析后图像的对比

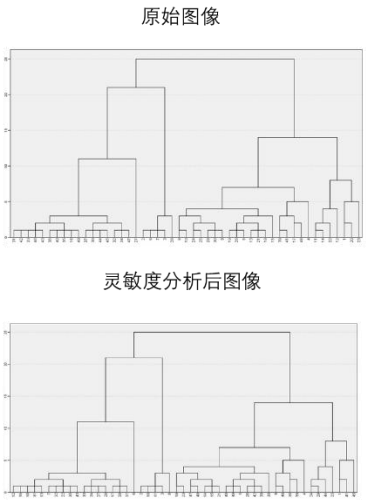

图 13 原始图像与灵敏度分析后图像的对比

从图中可以看出生成的新图像与原始图像差别较小，由此说明其灵敏性较低。

5.3 问题三模型的建立与求解

5.3.1 模型的建立

- **运用 Fisher 线性判别分析鉴别未知玻璃制品所属类型**

这里运用到的 Fisher 线性判别分析与问题二中的较为相似，本题以表单二中已知数据作为测试组(约占总数据的 89.7%)，表单三中未知数据作为预测组(约占总数据的 10.3%)这样确定出来的模型较为合适于本题的解答。具体的分析过程与问题二的相同，这里不再赘述。

5.3.2 模型的求解

1. Fisher 线性判别分析求解过程

1) 由所给数据可以得出ω法向量的典则判别函数系数表：

表 18 典则判别函数系数表

化学成分	系数	化学成分	系数
二氧化硅(SiO2)	0.057	氧化铅(PbO)	0.152
氧化钠(Na2O)	0.416	氧化钡(BaO)	0.211
氧化钾(K2O)	-0.287	五氧化二磷(P2O5)	0.157
氧化钙(CaO)	0.062	氧化锶(SrO)	-1.22
氧化镁(MgO)	0.211	氧化锡(SnO2)	0.114
氧化铝(Al2O3)	0.271	二氧化硫(SO2)	-0.043
氧化铁(Fe2O3)	-0.052	(常量)	-8.338
氧化铜(CuO)	-0.102		

这些系数便是一维直线的各项自变量前的系数，数据就可以通过投影呈现在一维空间中加以区分。

2) 根据对目标函数的求解我们可以得到分类函数系数：

表 19 分类函数系数表

	铅钡	高钾		铅钡	高钾
二氧化硅(SiO2)	32.89	32.55	氧化铅(PbO)	34.26	33.36
氧化钠(Na2O)	39.29	36.84	氧化钡(BaO)	42.79	41.55
氧化钾(K2O)	33.87	35.56	五氧化二磷(P2O5)	43.72	42.80
氧化钙(CaO)	39.43	39.06	氧化锶(SrO)	-85.22	-78.04
氧化镁(MgO)	49.70	48.46	氧化锡(SnO2)	25.39	24.72
氧化铝(Al2O3)	29.02	27.43	二氧化硫(SO2)	14.27	14.53
氧化铁(Fe2O3)	31.45	31.76	(常量)	-1649.54	-1608.47
氧化铜(CuO)	16.02	16.62			

将数据点带入判别函数中比较得出的函数值，哪个函数值较大就将该样品归于哪一类。由此可以对预测组进行归类处理

3) 最后得到分类结果如下表(0 为铅钡，1 为高钾)：

表 20 分类结果表

0 概率	1 概率	预测结果
0.53005	0.46995	0
1	0	0
1	0	0
1	0	0
0.99997	0.00003	0
0	1	1
0	1	1
0.99668	0.00332	0

4) 由测试组的结果得知，预测正确率为 100%(具体预测结果放在“第三题预测结果”Excel 表中),由此表明预测结果较为可靠，可以认为已经将未检测的玻璃制品成功地鉴别其所属类型。

2. 灵敏性分析

针对于所得出的结果，还需要经过灵敏性分析来检测其灵敏程度，这里在原始数据的±5%内设立一个区间，在区间内取随机数后组成新的数据再进行 Fisher 线性判别分析得到结果如下：

表 21 改变数据后的结果

0 概率	1 概率	预测结果
0.97366	0.02634	0
0.99989	0.00011	0
0.99967	0.00033	0
0.99995	0.00005	0
0.99953	0.00047	0
0.00053	0.99947	1
0.01672	0.98328	1
0.99799	0.00201	0

改变原始数据后，预测结果没有发生改变，但对于 0 概率和 1 概率的数值来说产生了变动，由此说明本结果具有灵敏性但灵敏性不高。

5.4 问题四模型的建立与求解

5.4.1 模型的建立

1. 基于灰色关联分析模型对化学成分进行关联关系分析

对于该模型的建立，首先筛选出六个较为重要的成分(二氧化硅、氧化钙、氧化铅、氧化钡、氧化纳、氧化钾)来分析,选择这六个的原因包括题目中所含有的四个以及与高钾、铅钡玻璃主要成分相关的氧化纳、氧化钾两个成分。这里要将两个类型中的六个成分都进行权重的计算，而后就是建立模型的过程：

1) 首先要计算出子序列中多个指标与母序列的关联系数，首先将母序列表现为：

$x_0 = (x_0(1), x_0(2), ..., x_0(n))^T$,

子序列为：

$x_1 = (x_1(1), x_1(2), ..., x_1(n))^T$.....$x_m = (x_m(1), x_m(2), ..., x_m(n))^T$

之后计算出：

$$a = \min_i \min_k |x_0(k) - x_i(k)| \tag{1}$$

$$b = \max_i \max_k |x_0(k) - x_i(k)| \tag{2}$$

其中a为两级最小差，b为两级最大差。

再定义：

$$y(x_0(k), x_i(k)) = \frac{a + \rho b}{|x_0(k) - x_i(k)| + \rho b} \tag{3}$$

其中ρ为分辨系数，通常取值 0.5，$i = 1,2, ..., m$，$k = 1,2, ..., n$。

最后再定义：

$$y(x_0, x_i) = \frac{1}{n} \sum_{k=1}^{n} y(x_0(k), x_i(k)) \tag{4}$$

此为x_0和x_i的灰色关联度，由此可以得出子序列与母序列的关联度。

2) 根据关联度得出结论。

2. 熵权法求权重

在横向比较当中，首先会用到熵权法来求出十四个成分的权重，熵权法求权重的原理是判断一个指标的变异程度，若指标的变异程度越小，其所反映的信息量也会越少，其对应的权值也应该越低。具体过程如下：

1) 将要评价的对象和评价指标构成正向化矩阵，然后判断输入的矩阵中是否存在负数，如果有则要重新标准化到非负区间。

原正向化矩阵为：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

若其中不存在负数，则通过该公式：

$$z_{ij}=x_{ij}/\sqrt{\sum_{i=1}^{n}x_{ij}^2} \tag{5}$$

对原矩阵进行标准化处理。若存在负数，则需要用到另一种标准化公式来得到z̃矩阵。

$$\tilde{z}_{ij}=\frac{x_{ij}-\min\{x_{1j},x_{2j},\ldots,x_{nj}\}}{\max\{x_{1j},x_{2j},\ldots,x_{nj}\}-\min\{x_{1j},x_{2j},\ldots,x_{nj}\}} \tag{6}$$

2) 得到标准化矩阵后，需要计算第 j 项指标下第 i 个样本所占的比重，并将其看作相对熵计算中用到的概率。此时可以计算概率矩阵 P，其中元素P_{ij}的计算公式如下：

$$p_{ij}=\frac{\tilde{z}_{ij}}{\sum_{i=1}^{n}\tilde{z}_{ij}} \tag{7}$$

3) 最后需要计算每个指标的信息熵得到每个指标的熵权，信息熵计算公式为：

$$e_j=-\frac{1}{\ln\ n}\sum_{i=1}^{n}p_{ij}\ln(p_{ij})\ \ (j=1,2,\ldots,m) \tag{8}$$

之后需要引入信息效用值这个概念，即信息效用值越大，所得到的信息越多，其计算公式为：

$$d_j=1-e_j \tag{9}$$

然后将其归一化处理后就可得到每个指标的熵权：

$$W_j=d_j/\sum_{j=1}^{m}d_j\ \ (j=1,2,\ldots,m) \tag{10}$$

3. TOPSIS 法算得分

在得到每个指标的权重后，结合权重为不同化学成分进行评分，这里假设评分高偏向于高钾，评分低则偏向于铅钡。这样就可以通过得分来为评价不同类型下同一化学成分关联关系的差异性。评分公式为：

$$S_j=s_{j1}\omega_1+s_{j2}\omega_2+\ldots+s_{j14}\omega_{14} \tag{11}$$

其中S_j为不同文物检测点的得分，s_{ij}为十四个评估指标，ω_i为评估指标的权重。

5.4.2 模型的求解

1. 灰色关联分析模型的求解

根据 MATLAB 得出的关联系数表为：

表 22 铅钡中六个母序列情况下的关联系数表

	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铅 (PbO)	氧化钡 (BaO)
二氧化硅 (SiO_2)	0.7740	0.8158	0.8105	0.8363	0.8468
	二氧化硅 (SiO_2)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铅 (PbO)	氧化钡 (BaO)
氧化钠 (Na_2O)	0.7992	0.7963	0.7611	0.7693	0.7796
	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钙 (CaO)	氧化铅 (PbO)	氧化钡 (BaO)
氧化钾 (K_2O)	0.8370	0.7963	0.8179	0.8011	0.8157
	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化铅 (PbO)	氧化钡 (BaO)
氧化钙 (CaO)	0.8296	0.7566	0.8142	0.8785	0.8313
	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化钡 (BaO)
氧化铅 (PbO)	0.8557	0.7650	0.7978	0.8814	0.8758
	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铅 (PbO)
氧化钡 (BaO)	0.8498	0.7585	0.7966	0.8166	0.8614

表 23 高钾中六个母序列情况下的关联系数表

	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铅 (PbO)	氧化钡 (BaO)
二氧化硅 (SiO_2)	0.7308	0.8242	0.8146	0.7737	0.7351
	二氧化硅 (SiO_2)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铅 (PbO)	氧化钡 (BaO)
氧化钠 (Na_2O)	0.7198	0.7853	0.7897	0.8252	0.7792
	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钙 (CaO)	氧化铅 (PbO)	氧化钡 (BaO)
氧化钾 (K_2O)	0.7549	0.7421	0.8862	0.7557	0.7077
	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化铅 (PbO)	氧化钡 (BaO)
氧化钙 (CaO)	0.7534	0.7537	0.8905	0.7556	0.7154
	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化钡 (BaO)
氧化铅 (PbO)	0.7302	0.8096	0.7746	0.7686	0.8648
	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铅 (PbO)
氧化钡 (BaO)	0.7242	0.7792	0.7557	0.7554	0.8775

由表可知不同化学成分之间的关系，例如在铅钡类型下，当二氧化硅为母序列时，氧化钡和氧化铅与其关联系数较大，可以认为相比于其它成分，这两个成分对二氧化硅的影响较大；氧化纳与其关联系数较小，可以认为相比于其它成分，这个成分对二氧化硅的影响较小；而氧化钾和氧化钙与二氧化硅的相关系数处于中间值，这两个成分对二氧化硅的影响趋于适中。其余的关系也可以由此表得出，因此就可以得出不同类型中化学成分之间所具有的关联关系。

2. 熵权法求解

权重表如下所示：

表 24 评价指标权重表

计算数值	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)
权重（w_j）	0.0215	0.0051	0.1992	0.056	0.0758	0.0434	0.1112
计算数值	**氧化铜(CuO)**	**氧化铅(PbO)**	**氧化钡(BaO)**	**五氧化二磷(P_2O_5)**	**氧化锶(SrO)**	**氧化锡(SnO_2)**	**二氧化硫(SO_2)**
权重（w_j）	0.0705	0.0141	0.0087	0.0086	0.0082	0.3739	0.0038

3. TOPSIS 求解

结合公式，得出不同检测点的得分如下：

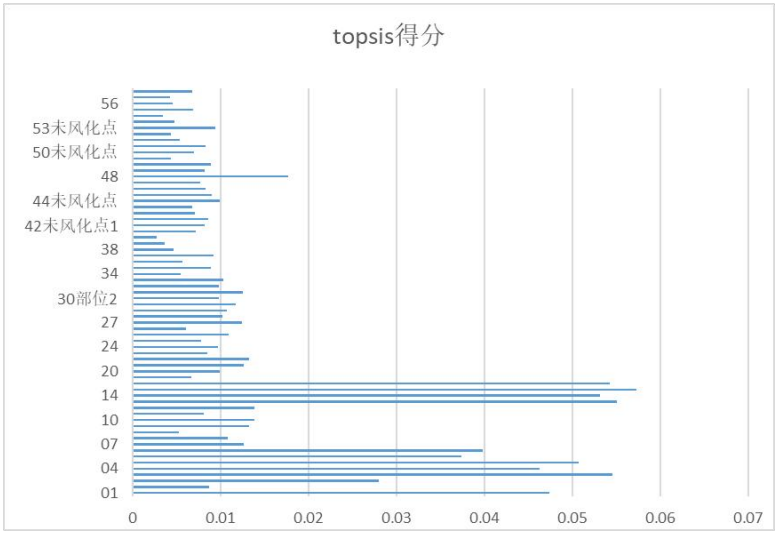

图 14 TOPSIS 得分表

4. 结论

各文物检测点的得分情况已放在了“topsis 得分”Excel 表中，由得分表可以看出有大部分偏向于低得分，即偏向于铅钡。接下来分析十四个化学成分中，每个成分对不同类型的差异性。根据权重分析结果可将十四个成分分为“5+3+6”的形式，“5”为五个较高权重的成分(氧化锡、氧化钾、氧化铁、氧化镁、氧化铜)，其权重较高，对高钾的影响要大于对铅钡的影响；“6”为六个较低权重的成分(二氧化硫、氧化钠、氧化锶、氧化钡、五氧化二磷、氧化铅)，其权重较低，对铅钡的影响要大于对高钾的影响。“3”为剩下三个权重处于中间的成分，其对两个类别均有影响但偏向不大。由此可以对比出不同成分的偏向程度，即为化学成分关联关系的差异性。最后分析一下产生偏向可能的原因：

1) 对于铅钡玻璃来说，其原材料会涉及到铅矿石，而铅矿石的成分中是含

有硫元素的，所以二氧化硫对铅钡玻璃偏向更大。[1]

2) 对于高钾玻璃来说，在采用草木灰作为助溶剂时，其中的聚天冬氨酸会与铁离子和铜离子形成螯合物，因此可在玻璃制品中找到氧化铁和氧化铜的成分，所以氧化铁和氧化铜会更偏向于高钾玻璃。[2]
3) 由于铅钡玻璃和高钾玻璃主要成分可以确定，所以氧化钾、氧化铅和氧化钡的偏向则易于解释。

六、 模型的评价

6.1 模型的优点

1) 数据预处理时对颜色的补充扩大了样本数，在后续回归分析时原始数据更多，避免了对样本的浪费且提高了回归精确度。
2) 在问题一中在铅钡类型中将严重风化单独提出来作为一个考虑因素使模型的建立更加合理，结果也更加精确。
3) 在问题四中，在运用 TOPSIS 法分析时结合熵权法为评分加上了权重这个指标，使我们在评分时又多了一个评分依据。
4) 用高钾和铅钡代替了传统评价模型的好和坏，即将灰色关联分析和 TOPSIS 这样的传统评价类模型用于解决第四题的分析类问题。

6.2 模型的缺点

1) 数据预处理时，给颜色的分类结果不是很准确，预测的成功率仅为 51.85%
2) 熵权法中的评价指标不是很客观，因为模型的参数由人为自行规定。而且对原始数据的标准化暂无统一的标准，不同人用的标准可能不同，导致结果差异较大。

6.3 模型的改进

1) 有序逻辑回归时可以采用逐步回归的方法消除多重共线性的影响。
2) 熵权法可以结合层次分析法，使对权重的计算更为合理化。

6.4 模型的推广

1) 有序逻辑回归可推广于医疗领域，根据病人的身体情况来划分病人的病情严重程度。
2) 灰色关联分析可应用于经济领域，探究各种经济主体对 GDP 的影响情况。

七、 参考文献

[1] 丁真贞,张一.铅矿石中可溶性氧化态和硫化态铅的物相分析[J].化学工程,2020,48(12):68-72.

[2] 梁星星.生物质电厂废弃物草木灰成分分析及资源化利用[D].北京化工大学,2020.

附录

附录 1
介绍：支撑材料的文件列表
Excel 数据结果：

1. topsis 得分.xlsx 为第四题中各文物检测点的得分
2. 第三题预测结果.xlsx 为第三题的检测组预测出的结果
3. 第一题预测.xlsx 为第一题预测出不同类型下的化学成分风化前含量
4. 合理性分析结果.xlsx 为第二题的亚类划分结果

matlab 代码：

1. topsis 代码.zip 为第四题 topsis 算法的求解过程
2. 灰色关联.zip 为第四题灰色关联分析算法的求解过程

spss 命令行：为文中用 SPSS 求解算法的命令行

SPSS 数据结果：为文中用 SPSS 求解算法的最终结果

附录 2

介绍：使用 Matlab 求第四题高钾类别下氧化钡的关联系数(仅以此为例，求关联系数的全部代码详见支撑材料中“灰色关联.zip”)

```
clear;clc
load BAO.mat
Mean = mean(BAO);
BAO= BAO./ repmat(Mean,size(BAO,1),1);
disp('预处理后为：'); disp(BAO)
X = BAO(:,2:end);
Y = BAO(:,1);
ABSX0_Xi = abs(X - repmat(Y,1,size(X,2)));
a = min(min(ABSX0_Xi));
b = max(max(ABSX0_Xi));
FBXS = 0.5;
G = (a+FBXS*b) ./ (ABSX0_Xi  + FBXS*b);
disp('各个指标的灰色关联度分别为：')
disp(mean(G))
```

附录 3

介绍：使用 Matlab 计算熵权法中的权重和 TOPSIS 评分

```
function [Z] = JX(x)
   Z = max(x) - x;
     end
function [Z] = JD(x,best)
   Q= max(abs(x-best));
   Z = 1 - abs(x-best) / Q;
     end
function [lnp] =  mh(p)
n = length(p);
lnp = zeros(n,1);
```

```
    for i = 1:n
        if p(i) == 0
            lnp(i) = 0;
        else
            lnp(i) = log(p(i));
        end
    end
      end
function [Z] = ZJ(x,c,d)
    r_x = size(x,1);
    Q = max([c-min(x),max(x)-d]);
    Z = zeros(r_x,1);
    for i = 1: r_x
        if x(i) < c
            Z(i) = 1-(c-x(i))/Q;
        elseif x(i) > d
            Z(i) = 1-(x(i)-d)/Q;
        else
            Z(i) = 1;
        end
    end
      end
function [T] = SQ(Z)
    [n,m] = size(Z);
    D = zeros(1,m);
    for i = 1:m
        x = Z(:,i);
        p = x / sum(x);%%避免 p 为 0,再次定义一个函数 mh
        e = -sum(p .* mh(p)) / log(n);
        D(i) = 1- e;
    end
    T = D ./ sum(D);
      end
function [Z] = ZXHHS(x,LX,i) %%定义这个函数进行正向化%LZ 指的是指标的类型 %Z 表示正向化
后的向量
    if LX == 1  %LX=1 意味着指标类型为极小型
        Z = JX(x);
        disp(['第' num2str(i) '列极小型处理完成'] )
    elseif LX == 2  %LX=2 意味着指标类型为中间型
        best = input('请输入最佳的那一个值：  ');
```

```
        Z = JD(x,best);
        disp(['第' num2str(i) '列中间型处理完成'] )
    elseif LX == 3  %LX=3 意味着指标类型为区间型
        c= input('请输入区间的下界：  ');
        d = input('请输入区间的上界：  ');
        Z = ZJ(x,c,d);
        disp(['第' num2str(i) '列区间型处理完成'] )
    else
        disp("无此类型指标")
    end
      end
clear;clc
load XT4.mat
[a,b] = size(XT3);
ZXH= input('请以行向量的形式输入需要正向化处理的指标所在的列: ') ;
disp('请输入这些列的指标类型（1 为极小型/2 为中间型/3 为区间型）  ');
LX = input('以行向量的形式：  ') ;

for i = 1 : size(ZXH,2)
    XT3(:,ZXH(i)) = ZXHHS(XT3(:,ZXH(i)),LX(i),ZXH(i));
end
disp('正向化后的矩阵为：')
disp(XT3)

Z = XT3 ./ repmat(sum(XT3.*XT3) .^ 0.5, a, 1);
disp('标准化矩阵 Z = ')
disp(Z)

if sum(sum(Z<0)) >0
   disp('原来标准化得到的 Z 矩阵中存在负数，所以需要对 X 重新标准化')
   for i = 1:n
      for j = 1:m
         Z(i,j) = [X(i,j) - min(X(:,j))] / [max(X(:,j)) - min(X(:,j))];
      end
   end
   disp('X 重新进行标准化得到的标准化矩阵 Z 为:  ')
   disp(Z)
end
P = SQ(Z);
disp('熵权法确定的权重为：')
```

```
    disp(P)
clear;clc
load XT1.mat
[a,b] = size(XT2);
ZXH= input('请以行向量的形式输入需要正向化处理的指标所在的列: ') ;
disp('请输入这些列的指标类型（1 为极小型/2 为中间型/3 为区间型） ');
LX = input('以行向量的形式：  ') ;
for i = 1 : size(ZXH,2)
   XT2(:,ZXH(i)) = ZXHHS(XT2(:,ZXH(i)),LX(i),ZXH(i));
 end
 disp('正向化后的矩阵为：')
 disp(XT2)
Y=[0.0181,0.0052,0.2019,0.0567,0.0768,0.0340,0.1127,0.0715,0.0143,0.0088,0.0087,0.0084,0.3
790,0.0038];%根据熵权法得到的各指标权重
C=XT2 .* Y;
DN = sum([(C - repmat(min(C),a,1)) .^ 2 ] .* repmat(Y,a,1) ,2) .^ 0.5;
DP = sum([(C - repmat(max(C),a,1)) .^ 2 ] .* repmat(Y,a,1) ,2) .^ 0.5;
S = DN ./ (DP+DN);
disp('最后的得分为：')
stand_S = S / sum(S)
    [sorted_S,index] = sort(stand_S ,'descend')
```

附录 3

介绍：使用 SPSS 预测预处理中的缺失值

```
    DATASET ACTIVATE 数据集 1.
    PLUM 颜色 BY 纹饰 类型 表面风化
      /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
      /LINK=LOGIT
      /PRINT=FIT PARAMETER SUMMARY TPARALLEL
 /SAVE=ESTPROB PREDCAT PCPROB ACPROB.
```

附录 3

介绍：使用 SPSS 计算第一题的卡方值

```
    DATASET ACTIVATE 数据集 1.
    CROSSTABS
      /TABLES=表面风化 BY 纹饰 类型 颜色
      /FORMAT=AVALUE TABLES
      /STATISTICS=CHISQ
      /CELLS=COUNT
      /COUNT ROUND CELL.
```

附录 3

介绍：使用 SPSS 将第一题中是否风化转变为虚拟变量

```
DATASET ACTIVATE 数据集 3.
SPSSINC CREATE DUMMIES VARIABLE=V21
ROOTNAME1=A
/OPTIONS ORDER=A USEVALUELABELS=YES USEML=YES
OMITFIRST=NO
```

附录 3

介绍：使用 SPSS 计算铅钡回归系数

```
DATASET ACTIVATE 数据集 4.
PLUM V21 WITH 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
 /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
 /LINK=LOGIT
 /PRINT=FIT PARAMETER SUMMARY TPARALLEL.
```

附录 3

介绍：使用 SPSS 计算高钾回归系数

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT A_2
 /METHOD=ENTER 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.
```

附录 3

介绍：使用 SPSS 将第二题中指标转变为虚拟变量

```
DATASET ACTIVATE 数据集 1.
SPSSINC CREATE DUMMIES VARIABLE=V19
ROOTNAME1=B
/OPTIONS ORDER=A USEVALUELABELS=YES USEML=YES OMITFIRST=NO.
```

附录 3

介绍：使用 SPSS 计算第二题 Fisher 线性判别分析结果

```
DISCRIMINANT
 /GROUPS=B_2(0 1)
 /VARIABLES=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
 /ANALYSIS ALL
 /PRIORS EQUAL
 /STATISTICS=COEFF RAW TABLE
 /CLASSIFY=NONMISSING POOLED
```

附录 3

介绍：使用 SPSS 计算第二题聚类分析结果

```
DATASET ACTIVATE 数据集 4.
CLUSTER  氧化镁 MgO 氧化锶 SrO 五氧化二磷 P2O5 氧化钾 K2O 氧化钙 CaO 氧化铅 PbO 氧化钡 BaO
 /METHOD BAVERAGE
 /MEASURE=SEUCLID
 /ID=V19
 /PRINT SCHEDULE
 /PLOT DENDROGRAM VICICLE.

DATASET ACTIVATE 数据集 5.
CLUSTER  氧化钾 K2O 氧化钙 CaO 氧化镁 MgO 氧化铅 PbO 氧化钡 BaO 五氧化二磷 P2O5 氧化锶 SrO
 /METHOD BAVERAGE
 /MEASURE=SEUCLID
 /ID=V19
 /PRINT SCHEDULE
/PLOT DENDROGRAM VICICLE.
```

附录 3

介绍：使用 SPSS 将第三题中指标转变为虚拟变量

```
DATASET ACTIVATE 数据集 9.
SPSSINC CREATE DUMMIES VARIABLE=V21
ROOTNAME1=D
/OPTIONS ORDER=A USEVALUELABELS=YES USEML=YES OMITFIRST=NO.

SPSSINC CREATE DUMMIES VARIABLE=V19
ROOTNAME1=C
/OPTIONS ORDER=A USEVALUELABELS=YES USEML=YES OMITFIRST=NO.
```

附录 3

介绍：使用 SPSS 计算第二题 Fisher 线性判别分析结果

```
DISCRIMINANT
  /GROUPS=C_3(0 1)
  /VARIABLES=D_1 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
  /ANALYSIS ALL
  /SAVE=CLASS PROBS
  /PRIORS EQUAL
  /STATISTICS=COEFF RAW TABLE
  /CLASSIFY=NONMISSING POOLED.
```