关于古代玻璃制品的成分分析与鉴别分类

摘要

本文对一批我国古代玻璃制品的相关数据进行分析，探究不同玻璃文物的化学成分与类型划分。主要基于数据统计，多种相关性分析手段与 k-means 聚类等方法进行分析。

针对问题一，第一小问探究玻璃风化与玻璃类型、纹饰和颜色的关系，首先对表单一中数据进行预处理，从主观上初步分析它们的相关关系，然后进一步采用卡方检验的方法客观分析玻璃风化分与类型、纹饰和颜色三者的关系，得出类型与玻璃风化显著相关。第二小问分析化学成分的统计规律，首先将所有文物划分为高钾类和铅钡类，进行数据可视化，得到这两类文物的成分含量特点，再结合平均值、标准差等统计数据，分析化学成分含量规律。第三小问预测风化前的成分含量，通过各属性的信息增益来判断其对风化的影响程度。根据文物的纹饰和类型，为风化文物选取几个可能风化过程相似的未风化文物，通过计算这些文物各成分向量之间的距离来选取最相似未风化文物进行预测。

针对问题二，对于不同文物的分类规律，基于未风化玻璃的数据集，通过主成分分析法，根据不同原成分变量关于贡献度较大的主成分的载荷因子的大小，分析部分成分变量对于文物分类情况的影响程度大小，确定主要影响因素。对于每个文物类别进行亚类划分问题，首先找出能够作为亚类划分依据的化学成分。通过对表单二进行统计学分析，求出方差最大的几个化学成分，并对它们是否能作为划分依据进行分析，并最终选取合适的化学成分。接着采用 k-means 聚类算法基于所选取的化学成分来对样本进行聚类分析，通过穷举试验，选择合适 k 值后得到分类结果。最后对分类结果的合理性进行敏感性分析。

针对问题三，建立了基于 k-means 算法的分类模型。为了实现类型划分，首先需要对待测数据进行预处理，采用问题一的方法，得到所有文物风化前的成分数据。基于问题二得出的分类规律，提取 PbO，BaO，K2O 作为分类依据。基于这三个属性，对原未风化数据集进行 k-means 聚类，聚类数为 2，得到 2 个聚类中心，将 2 个聚类中心分别作为该类的代表。根据待测数据距两聚类中心的距离大小，判断其所属类别，即待测数据属于聚类中心较近的类。最后通过对变量数据施加扰动，判断结果变化的方法对模型进行敏感性分析。

针对问题四，基于不同类型与风化情况对文物数据集进行分类，对于每一类数据，采用皮尔逊相关系数得到各个化学成分间的相关性，并选出相关性较强的几组成分进行了简要的说明。

关键词：卡方检验；主成分分析；k-means；皮尔逊相关系数；聚类

一、 问题重述

1.1 问题背景

古代丝绸之路是中西方文化交流的重要通道，而玻璃是中西方早期贸易往来的宝贵物证。我国早期的玻璃是通过吸收了西方制造工艺并结合本土材料制成，因此化学成分具有一定差异性。古代玻璃文物的化学成分分析对于文物的鉴别与分类十分重要，但是由于古代玻璃极易受埋藏环境的影响而风化，风化后的玻璃其风化部分的各种化学成分比例将会发生变化，而这会影响对玻璃类别的判断。因此解决风化和未风化玻璃制品的分析和分类问题对于考古工作十分迫切和重要。

1.2 问题重述

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。对这些文物有一些基本的统计数据。现在需要根据所给出的数据进行分析和建模，解决以下问题：

1) 一是需要通过对表单一中给出的玻璃文物的基本信息分析玻璃有无表面风化与其玻璃类型、纹饰和颜色的关系；二是结合玻璃类型，分析文物样品表面风化与未风化化学成分含量的统计规律，并且根据风化点的检测数据，预测该玻璃制品风化前的化学成分含量。

2) 根据附件表单二的数据分析高钾玻璃、铅钡玻璃的分类规律；并且对于每个类别的玻璃依据合适成分进行亚类的再分类，需要给出具体的划分方法和划分结果，并对分类结果进行合理性和敏感性分析。

3) 在前面分析基础上对附件表单三中未知类别玻璃文物的化学成分数据进行分析，并依据前面标准进行分类，再对分类结果进行分析。

4) 根据表单一和二中的数据，针对不同类别的玻璃文物样品，分析玻璃样品化学成分之间的关联关系，并且比较不同类别玻璃之间的化学成分关联关系的差异。

二、 问题分析

2.1 问题一的分析

对于本问，首先要根据表 1，分析玻璃表面是否分化与其玻璃类型、纹饰和颜色间的关系。分别对“是否风化”与另外三个属性间的数据进行统计、分类与可视化，得到初步的分析结果。为了进一步探究其相关性，也考虑到这些数据自身属性为定类数据，本问中也分别进行了卡方检验。接着根据表 2 分析不同玻璃化学成分的统计规律。首先需要对数据进行预处理，而由于“氧化锡”“二氧化硫”两个属性中缺失值较多，故不采用，单独进行分析。根据不同的玻璃种类与风化情况，本问将数据分组，分别进行了数据可视化呈现、平均值等统计数据的计算与展示，并据此分析规律；对于风化前各检测点的成分预测，首先要选取属性相同的文物，根据各属性的信息增益来判断对风化的影响程度，选取影响程度最大的两种属性，找出几组与需要预测的文物中这两种属性相同且无风化的文物，计算这些文物的成分数据之间的距离，距离越近，表示两个文物越相似，选取最相似的无风化文物成分含量来代表预测值。

2.2 问题二的分析

问题二首先要求探究不同类型玻璃的分类规律，联系实际需要根据未风化的玻璃数据进行分析，可以通过分析不同化学成分对于分类情况的影响程度反映出其规律。本文选择通过主成分分析的方法，得到贡献度不同的主成分，而各原化学成分变量可以通过不同的主成分反映其对数据集的影响贡献。占比最大的几个主成分对于某些原变量的解释程度越高，说明这些原变量对于玻璃的分类影响越大。在无风化的的高钾和铅钡两种类型的玻璃中分别对除二氧化硅外的化学成分进行统计学分析，求出平均值，中位数以及方差等数据，选取含量差异性较大的化学成分作为 k-means 分类依据，在判断差异性时主要关注方差这一参数。然后选择合适的聚类个数结合分析所得的化学成分依据进行聚类，再对聚类结果进行分析。通过改变作为 k-means 的化学成分含量的大小来分析对分类结果的影响来进行敏感度分析，以此作为亚类划分是否合理的检验。

2.3 问题三的分析

问题三要求对未知类型的玻璃进行分类。根据问题二的分析可得，玻璃类型的划分与未风化时的化学成分有关，因此，需要对数据进行预处理，根据问题一的预测思路，将风化的待测玻璃的成分含量转化为未风化时的含量。分类问题很多时候考虑机器学习的分类模型，但由于此次数据集较小，使用机器学习模型容易出现过拟合现象，因此我们可通过聚类实现类型预测。根据问题二中的分析，玻璃类型与 PbO，BaO，K2O 的成分含量关系密切，为了方便分析，本问只根据这三个成分进行分类。基于这三个属性的数据，可以对原未风化数据集进行聚类，分为两类，由聚类中心的三维坐标代表该类的特征，用待测点到聚类中心的距离代表其与该类的相似性，将依据特征相似性的分类问题转化为根据距离远近的分类问题。最后的判别标准是待测点距哪个聚类中心更近，其即属于哪一类。最后进行模型的敏感性检验，通过对待测数据的值产生扰动，重新进行结果判别，通过分析结果的变化情况检验模型的敏感性。

2.4 问题四的分析

问题四要求对各化学成分间的关联关系与关联关系的差异进行分析。由第一问对不同类型及风化情况的玻璃的统计性分析可得玻璃类型及风化情况对不同的变量各有不同的影响，因此需要将数据集分为“未风化的铅钡类”“风化的铅钡类”“未风化的高钾类”“风化的高钾类”四类。由于各化学成分是连续型数据，为了分析它们的相关性，可以采用皮尔逊相关系数反映各变量两两间的关系。皮尔逊系数在 1 和-1 之间，其绝对值越大，说明两变量间的线性相关性越强。

三、 模型假设及符号说明

现给出本文的假设如下：

1. 假设一风化文物的未风化点可以列为未风化情况.
2. 假设玻璃文物未风化处的化学成分与风化处风化前的化学成分相同.
3. 假设文物样品的高钾铅钡与铅钡玻璃的分类全都正确.
4. 假设不含有除附件表中化学成分以外的其他化合物

现给出本文的符号说明如下表 1 所示：

表 1：符号说明

符号	说明	单位
S	文物特征	
M	文物样本数据集	
k	文物样本数据集的子集个数	个
D	成分数据维数	
S^D	D 维成分数据空间	
x	成分数据	
ω	成分数据向量	
ρ	皮尔逊相关系数	
n	亚类中的文物数量	个

四、 模型的建立与求解

4.1 问题一的模型建立与求解

4.1.1 文物的表面风化与其属性分析

表单一所给的样本数据共有 58 个，其中有四个样本的颜色数据没有给出。通过对所给的数据分析发现，纹饰有 A、B、C 三种，玻璃类型有高钾和铅钡两种，颜色有八种，其中黑色的样本有 2 个，绿色仅有 1 个，紫色有 4 个，这三种颜色的数据较少。由排列组合原理可以推算出不同纹饰、类型和颜色玻璃制品的组合在数学层面最多为 48 种，但是由于样本数据不够多和玻璃制造工艺的影响，表中数据的类型只有 18 种。通过这 18 中不同纹饰玻璃类型和颜色有无风化的数据相关性分析可以得到玻璃文物表面风化与玻璃类型、纹饰和颜色的关系。

本文对玻璃文物表面是否风化分别与玻璃类型、纹饰和颜色进行一对一的关系分析，主要通过 Excel 图表的形式来直观判断二者关系。并结合卡方检验分析方法对相关性进行客观证明[5]。从数据表一中的数据统计得出风化的玻璃样本有 34 件，无风化的玻璃样本 24 件。首先从玻璃类型分析，如图 1 所示：

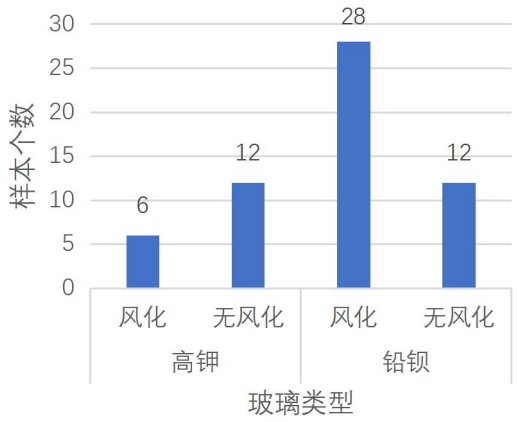


图 1：不同玻璃类型风化情况

从图中可以观察到高钾类型的玻璃制品风化数占比 1/3，而铅钡玻璃制品风化数占比 4/7，明显高于高钾类型，由此可以得出铅钡类型的玻璃制品比高钾类型的玻璃制品风化的概率更大。卡方检验分析[8]结果如表 2 所示：

表 2：类型的卡方检验结果

题目	名称	类型		总计	X^2	P
		高钾	铅钡			
表面风化	无风化	12	12	24		
	风化	6	28	30	6.880	0.009***
合计		18	40	54		

（注：***、**、*分别代表 1%、5%、10%的显著性水平）

卡方检验结果表示基于类型和表面风化，显著性 P 值为 0.009***，水平上呈现显著性，因此对于类型和表面风化数据存在显著性差异。

从纹饰样式分析，如图 2 所示：

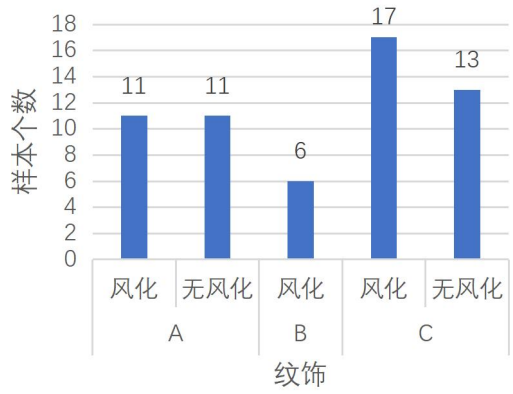


图 2：不同玻璃纹饰风化情况

从图中可以观察到纹饰为 A 类和 C 类的玻璃制品均有有风化和无风化的玻璃制品但是数量相当，而 B 类纹饰的玻璃制品虽然均为风化，但是样本数只有 6 个，数据量小，偶然性大，综合判断为纹饰对玻璃制品是否风化的影响较小。卡方检验结果如表 3 所示：

表 3：纹饰的卡方检验结果

题目	名称	纹饰			总计	X^2	P
		C	B	A			
表面风化	无风化	13	11	0	24		
	风化	17	11	6	34	4.957	0.084*
合计		30	22	6	58		

（注：***、**、*分别代表 1%、5%、10%的显著性水平）

结果显示，基于纹饰和表面风化，显著性 P 值为 0.084*，水平上不呈现显著性，接受原假设，因此对于纹饰和表面风化数据不存在显著性差异。

从颜色种类进行分析，如图 3 所示：

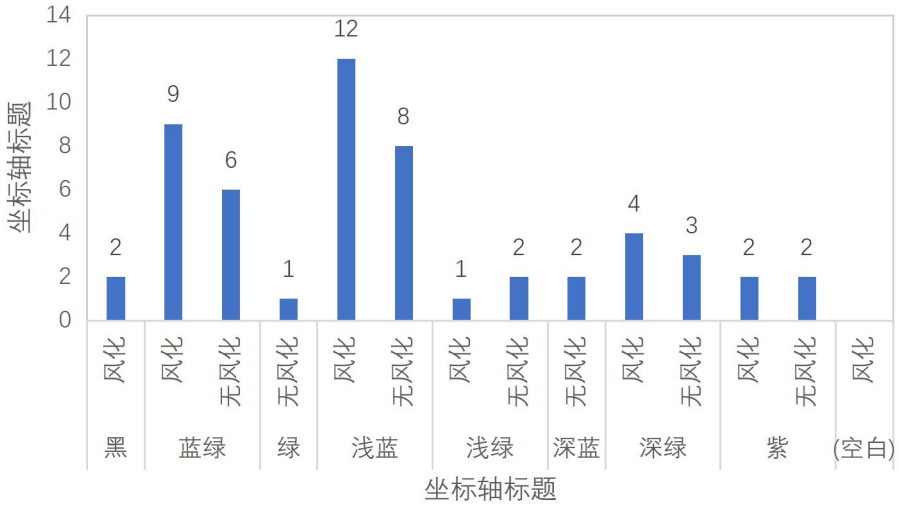


图 3:不同颜色玻璃风化情况

由于颜色种类较多，一些颜色种类的样本数很少，偶然性可能会较大。其中绿色和深蓝色均为无风化，黑色为风化，但是这三个颜色样本数均在 2 个以内，因此不能直接判断这几种颜色与有无风化的关系。卡方检验分析结果如表 4 所示：

表 4：类型的卡方检验结果

题目	名称	颜色								总计	X^2	P
		蓝绿	浅蓝	紫	深绿	深蓝	浅绿	绿	黑			
表面风化	无风化	6	8	2	3	2	2	1	0	24		
	风化	9	12	2	4	0	1	0	2	30	6.287	0.507
合计		15	20	4	7	2	3	1	2	54		

（注：***、**、*分别代表 1%、5%、10%的显著性水平）

卡方检验分析的结果显示，基于颜色和表面风化，显著性 P 值为 0.507，水平上不呈现显著性，接受原假设，因此对于颜色和表面风化数据不存在显著性差异。与 Excel 图表分析结果一致。

4.1.2 文物化学成分含量的统计规律分析

在探究文物样品表面有无风化化学成分含量的统计规律之前，首先对表单二中各文物的成分进行加和，发现文物 15 和文物 17 得到的结果有较大的偏差，为无效数据，故剔除掉。

在表单二中某些成分可能因为技术原因而没有被检测到，但实际上这些成分可能存在也可能不存在，假设将空白部分看做是 0，通过对各成分的加和结果可知，除了文物 15 和文物 17 外，其余的均为有效数据，数据值具有参考价值，因此可将空白部分的成分含量看做是 0，补全表单。另外还可以看出，氧化锡、氧化硫这些成分含量几乎为 0，且只存在于极少数文物中，将其列为特定情况，因此本文将这些数据抛开，后面进行单独分析。

根据玻璃的类别进行划分，可以得到高钾类和铅钡类两种类型，再根据类型对风化和未风化的文物进行表格排序，其中采样点为风化文物的未风化点情况归于未风化区，得到两组完善后的数据表格，将其进行数据可视化分析，得到如图 4 所示结果。图 4（a）中呈现了高钾类文物各采样点的化学成分含量情况，从图

中发现，二氧化硅作为玻璃的主要成分，无论在风化前还是风化后，成分含量都较高，而其余含量都在 20%以下。为了更清晰地展现其余含量的变化情况，对图 4（a）进行放大，如图 4（b）所示：

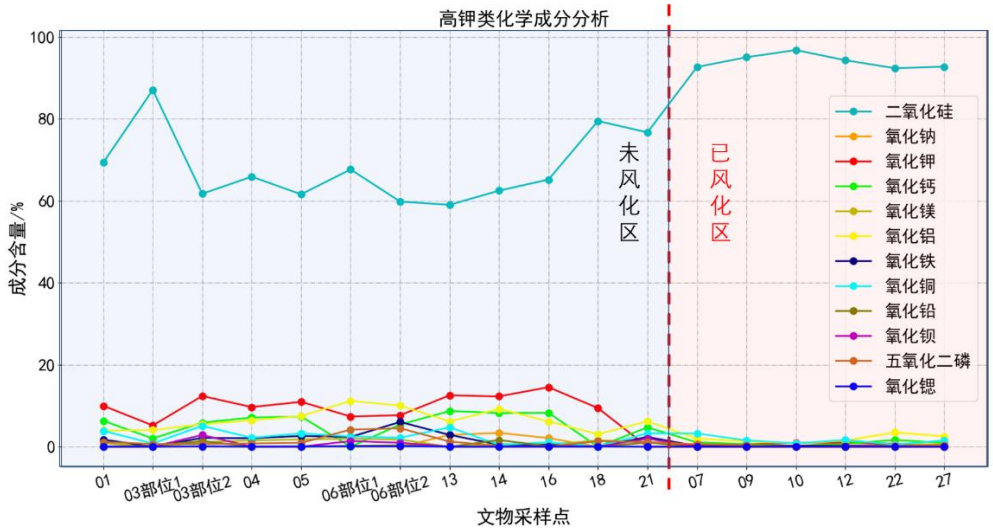


图 4(a):高钾类化学成分含量

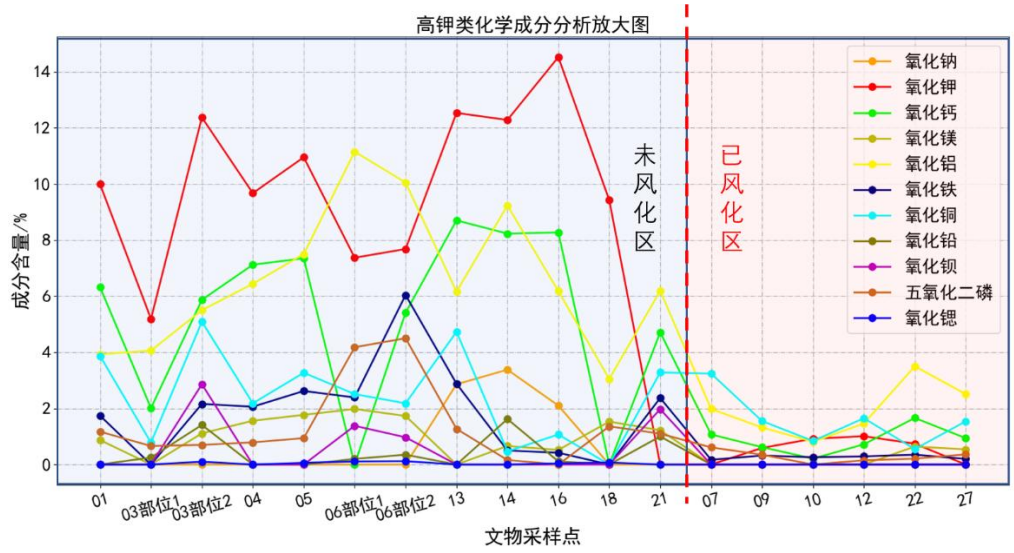


图 4(b):高钾类化学成分含量放大图

（注：因篇幅所限，此处仅展示高钾类文物化学成分含量图，铅钡类文物化学成分图详见附录）

从图 4（a）、（b）中得到高钾类文物的统计规律如下：不论该文物是否风化，二氧化硅作为其主要成分之一，总是在成分中占据绝大部分，但已风化文物相对于未风化文物，二氧化硅的含量明显升高；在其余成分含量中发现，文物未风化时，文物中含有多种氧化物成分，其中氧化钾的含量最多，文物风化后，本存在于文物中的其他微量成分含量都出现了明显的下降，且含量不大于 5%。通过查阅文献，玻璃风化主要原因是水中 H^+与玻璃中碱性离子交换，形成的碱性金属氢氧化物进一步腐蚀玻璃，在表面生成风化产物[3],其中绝大部分是钠、钾、钙的碳酸盐。所以其他微量成分含量出现明显的下降是由于生成了风化产物，而二氧化硅不会参与反应，造成了二氧化硅含量百分比上升，以上分析符合实际情况。

通过分析铅钡类成分含量变化图可得到如下结论：未风化文物中，二氧化硅为主要成分，而在已风化文物中，二氧化硅的含量下降，氧化铅的含量明显上升，超过二氧化硅，这与高钾类文物氧化前后二氧化硅的变化刚好相反；除了氧化铅的含量明显上升以外，一小部分已风化文物的氧化钡含量也大幅上升。其余氧化物在已风化文物和未风化文物中均变化不大。

根据类型和有无分化情况进行分类讨论，对各成分含量进行分析，得到平均数、中位数等统计数据如表 5 所示：

表 5:未风化高钾类数据统计情况

化学成分	平均数	最大值	最小值	中位数	标准差
SiO_2	9.33	14.52	0	9.83	3.92
Na_2O	5.33	8.70	0	6.10	3.09
K_2O	1.08	1.98	0	1.17	3.68
CaO	6.62	11.15	3.05	6.19	2.49
MgO	1.93	6.04	0	2.11	1.66
Al_2O_3	2.45	5.09	0	2.34	1.68
Fe_2O_3	0.41	1.62	0	0.15	0.59
CuO	0.60	2.86	0	0	0.98
PbO	1.40	4.50	0	1.02	1.43
BaO	0.04	0.12	0	0.02	0.05
P_2O_5	0.20	2.36	0	0	0.68
SrO	0.10	0.47	0	0	0.19

（注：因篇幅所限，此处仅以未风化高钾类文物化学成分含量数据统计情况为例，其余三种类型化学成分数据统计情况详见附录）

由表 5 可知，在未风化高钾类文物中，各个成分都有一定含量，二氧化硅成分维持在较高水平但同时波动幅度也是最大的，与其相对，氧化钡成分稳定在极低的水平，除了二氧化硅以外，氧化钠、氧化钙成分相对较高。

通过分析其余三种类型的数据统计情况得知，每一种类型的二氧化硅都处于较高水平，在铅钡类文物中，含量多的成分标准差也变大，除二氧化硅以外，氧化铅、氧化钡占据大部分比重，并且风化后的铅钡类文物的氧化铅、氧化钡含量超过二氧化硅，但同时这两种成分的波动幅度也变大；在风化的高钾类文物中，二氧化硅含量最高，相较于未风化高钾类有所增加，但所有化学成分含量的均稳定在一个范围内，方差较小。

对于氧化锡、氧化硫这两种成分，在高钾类文物中，只存在于无风化文物中，在铅钡类文物中均有分布，总体来看分布在风化文物中较多，和高钾类文物相反。

4.1.3 风化前化学成分含量的预测

表单二中给出的数据均是不同文物的成分含量值，本文不能得到同一文物风化前后的数据关联性，因此只能在相似文物风化前后的数据间寻找关系，来预测某一风化后文物风化前的化学成分含量。

在定义相似文物时，考虑到如果选择纹饰、类型、颜色均相同的文物得到的样本数据较少，因此本文在这三个属性中选择两个对风化影响较大的两个属性作为判定是否为相似文物的标准。为选出这两种属性，本文以信息增益来判断属性

对风化的影响程度。

将表单一中的文物样本作为一个数据集 M，在这些数据集中有两个类，即：已风化（N_1）和未风化(N_2)。纹饰、类型、颜色作为三个特征(S)，每个特征将数据集分为 k 个不同的子集，记为 M_k，例如纹饰特征将数据集分为三个子集，即：A、B、C。这些子集中，有的是已风化的（M_{1k}），有的是未风化的（M_{2k}），于是得到信息增益的算法如下：

$$H(M) = -\sum_{i=1}^{2}\frac{|N_i|}{|M|}\log_2\frac{|N_i|}{|M|} \tag{1}$$

$$H(M|S) = -\sum_{i=1}^{k}\frac{|M_i|}{|M|}\sum_{i=1}^{2}\frac{|M_{ik}|}{|M_k|}\log_2\frac{|M_{ik}|}{|M_k|} \tag{2}$$

$$g(M,S) = H(M) - H(M|S) \tag{3}$$

将表单一中不同属性的数据带入公式（1）（2）（3），可得到如表 6 所示的结果：

表 6:各属性的信息增益

属性	信息增益
颜色	0.07278022578373267
类型	0.11895096242495139
纹饰	0.1067732519232758

由上表可知，类型的信息增益最大，其次是纹饰，因此下文将选择这两种属性对文物相似度进行分析。

以文物 2 为例，该文物是纹饰为 A、类型为铅钡、颜色为浅蓝的风化样品，在表单一中查找同样纹饰为 A、类型为铅钡且未风化的样本，得到文物 20、30、45、46、47 均满足条件。

这里引入成分数据的概念[4],当 D 维向量$x = [x_1, x_2, ..., x_D]^T$满足:

$$S^D = \left\{x = [x_1, x_2 ... x_D]^T; x_i > 0; i = 1,2,3, ..., D, \sum_{i=1}^{D} x_i = c\right\}$$

则称向量 x 为成分数据，S^D为 D 维成分数据空间，c 为任意常数。

在本文中，$x_1, x_2, ..., x_D$分别按顺序表示表单二中各风化成分，因此文物 2、20、30、45、46、47 的成分数据为：

$$\omega_{02} = [36.28,0,1.05,2.34,1.18,5.73,1.86,0.26,47.43,0,3.57,0.19,0,0]$$

$$\omega_{20} = [37.36,0,0.71,0,0,5.45,1.51,4.78,9.3,23.55,5.75,0,0,0]$$

$$\omega_{30(1)} = [34.34,0,1.41,4.49,0.98,4.35,2.12,0,39.22,10.29,0,0.35,0.4,0]$$

$$\omega_{30(2)} = [36.93,0,0,1,4.24,0.51,3.86,2.74,0,37.74,10.35,1.41,0.48,0.44,0]$$

$$\omega_{45} = [61.28,2.66,0.11,0.84,0.74,5,0,0.53,15.99,10.96,0,0.23,0,0]$$

$$\omega_{46} = [55.21,0,0.25,0,1.67,4.79,0,0.77,25.25,10.06,0.20.43,0,0]$$
$$\omega_{47} = [51.54,4.66,0.29,0.87,0.61,3.06,0,0.65,25.4,9.23,0.1,0.85,0,0]$$

要预测风化前的含量，则要寻找这些不同风化类型相似文物之间的成分规律，首先计算文物 2 与其相似文物的成分数据间的距离，距离越短，说明相似度越高。通过 excel 进行计算，各相似文物间成分数据的距离如表 7 所示：

表 7:成分数据间距离

文物	距离
20	45.18925979
30（1）	14.02904131
30（2）	14.69301535
45	41.96200543
46	31.20696076
47	29.17837384

由表可知，编号为 2 的文物和编号为 30 的文物相似度更高，且在成分含量上，可用 30 号文物部位 1 的成分含量情况对 2 号文物风化前的含量进行预测。

4.2 问题二的模型建立与求解

4.2.1 玻璃分类规律模型的建立与求解

考虑到在实际情况中，一般需要通过玻璃未风化的成分数据才能直接判断其类别，因此在分析玻璃分类规律的过程中，本文选用的均是玻璃未风化检测点的数据记录。

分析玻璃分类的规律，即要分析玻璃各化学成分的差异对玻璃类型的影响，本问采用主成分分析的方法寻找起主要作用的成分。通过 SPSS 实现主成分分析。选取累计贡献率在 70%以上的前几个主成分，这里选取前 5 个主成分，各主成分的方差贡献率与累计贡献率如表 8 所示，各变量的因子载荷矩阵如表 9 所示。

表 8：5 个主成分的方差贡献率与累计贡献率

	方差贡献率	累计贡献率
主成分 1	35.8923	30.8645
主成分 2	15.7834	44.9068
主成分 3	12.3101	56.0119
主成分 4	9.6057	65.6310
主成分 5	7.5579	73.1267

表 9：因子载荷矩阵

成分矩阵	主成分 1	主成分 2	主成分 3	主成分 4	主成分 5
PbO	-.856	-.021	.242	.121	-.136
BaO	-.847	.351	-.041	-.051	.165
K2O	.800	.150	-.335	-.042	.311
CaO	.680	.336	-.345	.116	-.019
SrO	-.655	.252	.068	.228	.300
SiO2	.653	-.590	-.031	-.219	-.103
Al2O3	.548	.227	.133	.449	-.181
MgO	.480	.020	.408	.379	.317
CuO	-.032	.679	-.447	-.211	.328
Fe2O3	.495	.552	.291	.091	-.098
P2O5	.158	.520	.571	-.221	-.131
Na2O	-.307	-.218	-.543	.525	.052

成分矩阵又叫“因子载荷矩阵”。因子载荷矩阵是各个原始变量的主成分表达式的系数，表达提取的主成分对原始变量的影响程度。简单说，通过因子载荷矩阵可以得到原始指标变量的线性组合。该变量对应该主成分的因子载荷越大，该主成分对于该变量的解释性就越强。通过观察因子载荷系数值，可以分析出每个因子与变量的对应关系情况，也针对不合理的变量进行删除。

而判断该变量存在的是否合理的具体操作是需要观察“旋转成分矩阵”表，表中各变量对于各主成分的载荷如果都小于 0.4，可以考虑删除或修改这个变量。该数据集的“旋转成分矩阵”表中所有因子都至少有一个载荷值大于 0.4，所以认为这 14 个变量都与主成分有关，不做删除处理。

通过观察因子载荷矩阵可以得到，PbO，BaO，K2O 在主成分 1 上的载荷最大，CuO，SiO2，Fe2O3 在主成分 2 上的载荷因子最大，Na2O，P2O5，CuO 在主成分 3 上的载荷因子最大。

说明主成分 1 主要包含 PbO，BaO，K2O 等的成分信息，主成分 2 主要包含 CuO，SiO2，Fe2O3 等的成分信息，主成分 3 主要包含 Na2O，P2O5，CuO 等的成分信息[7]。那么这些与前几个主成分关系密切的变量可以相对较好地反映两类样品之间的差别。

综上， PbO，BaO，K2O，CuO，Fe2O3，Na2O，P2O5（由于 SiO2 是玻璃的主成分，这里不将其列出）的差异均会导致玻璃类型的不同。最后我们取主成分 1 对应的变量：PbO，BaO，K2O 的差异作为划分玻璃类型的主要依据。

其实结合问题一中对于不同类型的玻璃未风化时的数据统计规律，我们可以发现在未风化的高钾玻璃中 K2O 的平均值远大于铅钡玻璃中的，PbO，BaO 的平均值远小于铅钡玻璃中的且这些数据值分布较均匀。通过主成分分析法所得的分类规律正与此相吻合，印证了模型结果的合理性。

4.2.2 不同类型文物亚类划分模型的建立与求解

通过问题一的分析，本文已将表单二修改，删除无效数据并补全空白数据，将风化文物的无风化点采样结果作为无风化类，在此基础上，对铅钡无分化的 23 个数据进行描述性统计得到总体描述结果如表 10 所示：

表 10：无风化铅钡类统计数据

化学成分	平均数	最大值	最小值	中位数	标准差
PbO	22.08	39.22	9.30	20.12	8.22
BaO	9.00	26.23	2.03	8.99	5.83
CuO	1.43	8.46	0	0.65	1.97
…	…	…	…	…	…

（注：因篇幅所限，此处仅展示无风化铅钡类标准差较大的几组数据，其余数据详见附录）

上表展示了描述性统计的结果，包括平均数、中位数、标准差等统计量，用于研究定量数据的整体情况。从上表可以看出氧化铜、氧化铅和氧化钡的方差较大，表明在未风化的铅钡类型玻璃样本中这三种化学成分含量具有较大差异性，可作为亚类划分的分类依据。

以此三个数据进行 k-means 聚类分析，可得到如表 11 所示的结果：

表 11：铅钡类聚类结果

	聚类类别（平均值±标准差）			F	P
	类别 1(n=14)	类别 2(n=7)	类别 3(n=2)		
氧化铜(CuO)	6.894±3.473	8.679±1.746	24.89±1.895	31.772	0.000***
氧化铅(PbO)	17.605±3.333	31.863±5.429	19.22±14.029	18.657	0.000***
氧化钡(BaO)	7.265±3.197	0.586±0.428	6.62±2.602	24.005	0.000***

（注：***、**、*分别代表 1%、5%、10%的显著性水平）

同理，无风化高钾类统计数据见附录，以氧化钙、氧化钾、氧化铝为亚类划分的分类依据，可得到聚类结果如表 12 所示：

表 12：高钾类聚类结果

	聚类类别（平均值±标准差）			F	P
	类别 1(n=7)	类别 2(n=3)	类别 3(n=2)		
氧化钙(CaO)	7.409±1.059	2.24±2.363	2.705±3.825	10.053	0.005***
氧化钾(K2O)	11.759±1.684	4.87±4.718	7.525±0.219	7.854	0.011***
氧化铝(Al2O)3	6.42±1.645	4.433±1.603	10.6±0.778	9.481	0.006***

（注：***、**、*分别代表 1%、5%、10%的显著性水平）

上表展示了定量字段差异性分析的结果，包括均值±标准差的结果、F 检验结果、显著性 P 值。三个化学成分含量的显著性 P 值都小于 0.05，说明它们在聚类分析划分的类别之间存在显著性差异，类别划分合理。

对于这两类数据具体归类情况，将在支撑材料中展示。

4.3 问题三的模型建立与求解

4.3.1 数据的准备

由于对玻璃类型的划分是基于无风化的化学成分，所以需要将原附件表三中

存在的 4 条表面风化玻璃的记录通过问题一建立的模型的思想，预测其无风化时的成分再用于分类。

4.3.2 模型的建立与求解

考虑到该未风化的数据集的数据量过于小，不适用于一般的机器学习分类模型，容易出现过拟合的问题。因此，本问采用 k-means 聚类算法构造“类型判断模型”[6]，通过比较待分类点到聚类中心的距离进行类型划分。参考问题二中的玻璃类型划分规律：PbO，BaO，K2O 的成分差异对类型影响较大，故本问对类型的划分可以只考虑这 3 种化学成分。

“类型判断模型”建立与求解的具体步骤及结果如下：

Step1.选取已知未风化数据集中的 PbO，BaO，K2O 成分，根据这三个变量对数据集进行聚类划分，聚类数为 2。聚类效果良好，只有 2 个铅钡类被划分为高钾类，高钾类全部划分正确。得到 2 个聚类中心如下表 13：

表 13：聚类中心

	K2O	PbO	BaO
高钾类	9.33	0.41	0.60
铅钡类	0.22	22.08	9.00

Step2.预处理表面风化的 8 条待测数据，预测所需的 3 种成分未风化前的含量，结果如表 14 所示，

表 14：待测玻璃风化前的个别成分含量

	K2O	PbO	BaO
A2	0	31.90	6.65
A5	0.3	12.31	2.03
A6	0	1.00	1.97
A7	0	1.00	1.97

Step3.计算 8 条数据到 2 个聚类中心的距离。

Step4.比较这 2 个距离，取聚类中心较近的一类作为该待测数据的所属类别。以上可以通过 matlab 编程实现，得到的划分结果如表 15：

表 15：待测玻璃的分类结果

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

4.3.3 模型的敏感性分析

通过改变自变量的值，不断重复模型的判断过程，根据判断结果的变化分析模型的敏感性。

本问选取一条数据 A3，分别按一定倍数不断增加其 K2O，PbO，BaO 的值，对数据产生扰动，用接受扰动的数据重新分类，得到的分类结果如下图 5 所示，（用 1 表示分类为铅钡类，-1 表示分类为高钾类）

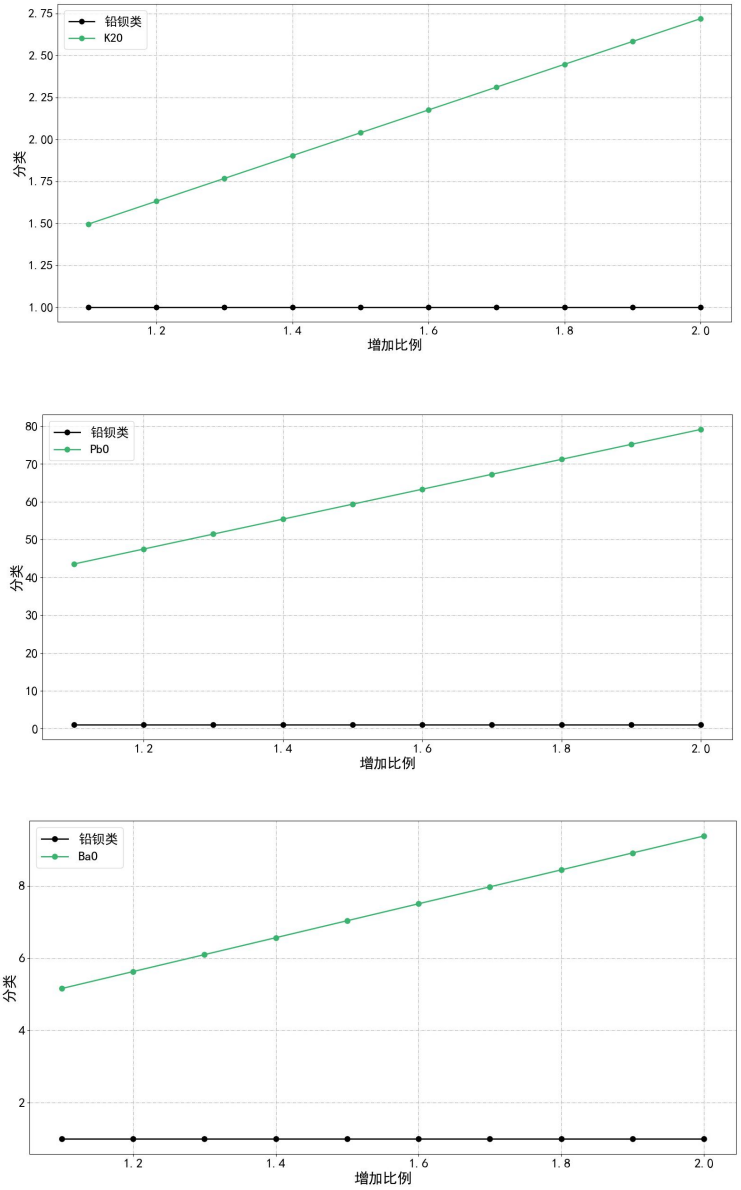


图 5:敏感性分析

由图 5 可以判断，不管随着哪种成分的变化，该玻璃均被分为铅钡类，分类情况不受一定程度上数据变化的影响，说明模型的稳定性很好。

4.4 问题四的模型建立与求解

对于连续型变量，本文采用皮尔逊相关性分析，在皮尔逊相关性分析中，通过皮尔逊相关系数的大小来表征相关度强弱，该系数值介于-1 和 1 之间，越接近 1 说明越趋向于正相关，越接近-1 说明越趋向于负相关。计算公式如下：

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x \sigma_y}$$

ρ的数值大小对于属性相关性的大致划分如下[2]：

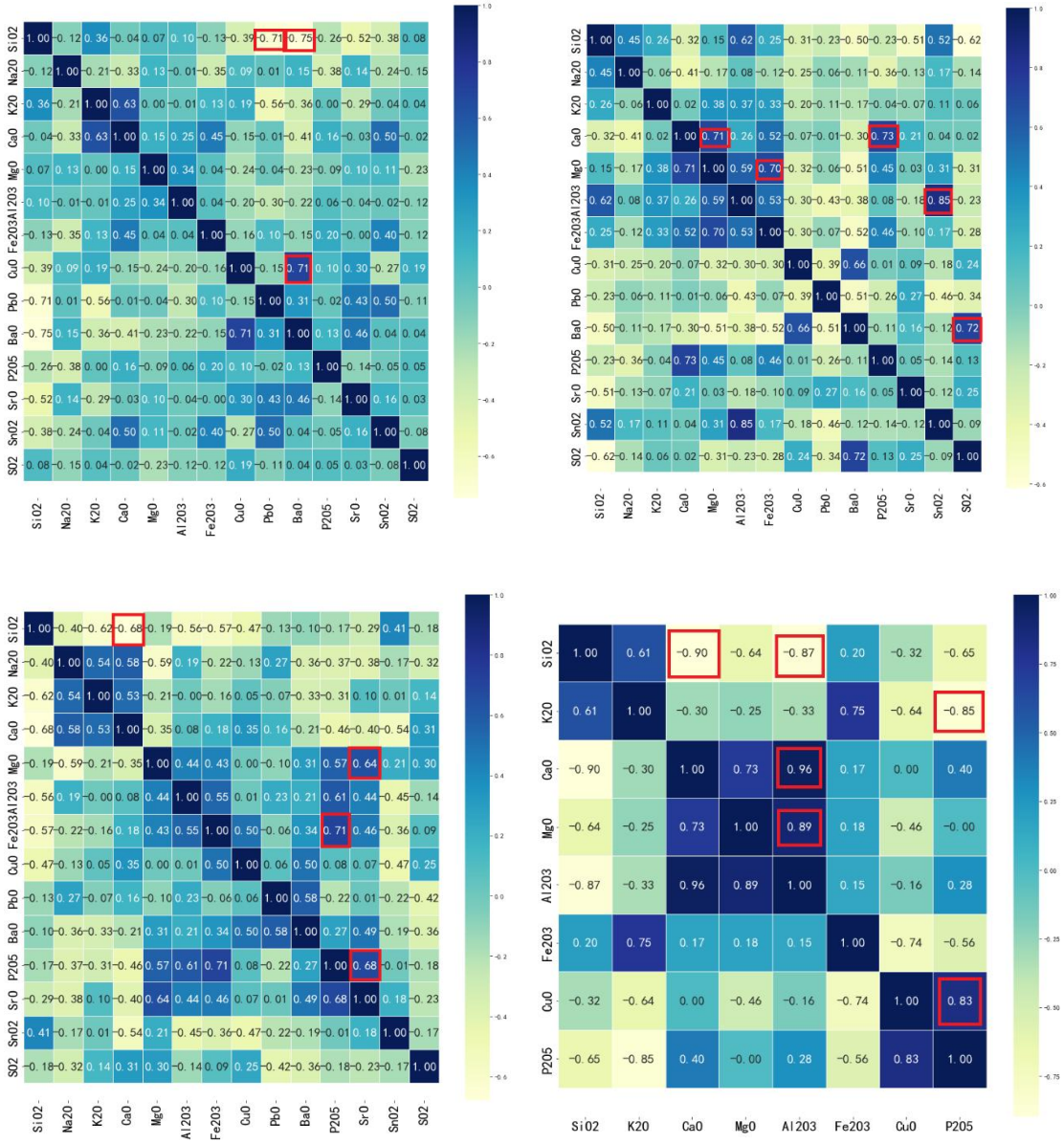
当 ρ >0 时，表示两变量正相关，ρ <0 时，两变量为负相关；

当$|\rho|$ =1 时，表示两变量为完全相关，当$|\rho|$=0 时，表示两变量间无相关关系；

当 0< $|\rho|$ <1 时，表示两变量存在一定程度的相关。且 $|\rho|$越接近 1，两变量间线性关系越密切;| $\rho|$越接近于 0，表示两变量的线性相关越弱；

一般可按三级划分：$|\rho|$<0.4 为低度相关；0.4<$|\rho|$<0.7 为显著性相关；0.7<$|\rho|$<1 为高度线性相关。

分别对“未风化的铅钡”“风化的铅钡”“未风化的高钾”“风化的高钾”进行分析，得到皮尔逊相关系数矩阵结果如图 6(a)、(b)、(c)、(d)所示：



（a）未风化的铅钡类型　　（b）风化的铅钡类型

（c）未风化的高钾类　　（d）风化的高钾类

图 6：各类皮尔逊相关系数

对于未风化的铅钡类玻璃，其中 SiO2 与 PbO、BaO，CuO 与 BaO 的相关系

数的绝对值均达到 0.7 以上（如图中红框所标示），为高度线性相关，且 SiO2 与 PbO、BaO 是负相关关系，CuO 与 BaO 是正相关关系。此外，如 K2O 和 CaO 的相关系数的绝对值也达到 0.63，是显著相关的。

对于风化的铅钡类玻璃，其中 Al2O3 与 SnO2，CaO 与 P2O5，MgO，MgO 与 Fe2O3 的相关系数的绝对值均达到 0.7 以上（如图中红框所标示），且都为正相关，为高度线性相关。

对于未风化的高钾玻璃，其中 Fe2O3 和 P2O5 的相关系数较大，呈正相关关系；SiO2 与 CaO 的相关系数的绝对值接近 0.7，呈现较强的负相关关系，SrO 与 P2O5,MgO 呈现较强的正相关关系。

对于风化的高钾类玻璃，由于其中几个属性值全部为 0，因此将其删除，分析剩余 6 个化学成分间的关系。可以发现，该组数据各属性呈现出很高的相关程度。Al2O3 与 MgO，CaO，SiO2 与 CaO，Al2O3，P2O5 与 K2O,CuO 的相关系数的绝对值都达到 0.8 以上，呈现很强的线性相关性，其中 SiO2 与 CaO，Al2O3 是负相关关系，其余均为正相关。

五、 模型的评价

5.1 模型的优点

5.1.1 K-means 模型（K 均值聚类算法）

K-means 算法属于无监督的聚类算法，主要优点是原理比较简单，实现相对于相同类型的算法容易，而且在实现过程中的收敛速度很快。算法的可理解性也好，适于开始接触聚类分析的同学使用，而且能达到的聚类效果较优。使用时主要参数较少，主要考虑的是簇数 k。在本次问题模型中，将化学成分含量差异较大的几种属性作为聚类的依据，从结果上看效果也是十分明显，直观地展示了化学成分含量范围与类之间的关系。

5.1.2 相关性的检验

本文用到几个相关性检验的方法，首先是卡方检验，它针对的是分类变量，比较定类变量与定类变量之间的差异性分析，操作简单，可得到卡方交叉热力图、效应量化分析等结果，清晰得看出交叉列联表的值、Phi、列联系数等，来分析样本的相关程度。其次是皮尔逊相关系数，在评分类型问题中优点尤为显著，它可以避免评分等级膨胀。

5.1.3 主成分分析

主成分分析可以消除各评价指标的影响，对最终结果形成新的影响因素，产生彼此独立的主成分，使得数据集更容易使用，且结果容易理解。当各评价指标较多时，还可以选择保留一部分，选取具有代表性的几个指标进行分析，并且在使用过程中没有参数限制。在本文中，通过主成分分析，选取对文物风化影响程度较大的几个成分，实现了评价指标的降维，提取较强特征。

5.2 模型的缺点

5.2.1 K-means 模型（K 均值聚类算法）

K-means 模型的缺点首先就是 K 值的选取不好把握，而这常常需要通过合并迭代解决，这也增大了模型调用的复杂度；除此以外，还对数据集有一定要求，比如不是凸的数据集就比较难收敛，而且如果隐含类别的数据不平衡，聚类后的效果不佳。

5.2.2 相关性的检验

卡方检验模型要求样本最好是有大量数据，因此在使用范围上有一定限制。除此以外，在区间划分时，选定区间数目也不宜过多。在皮尔逊相关系数中，协方差可以很好的解释两个变量之间相关的方向,但在可信度方面，相对较差；使用皮尔逊相关系数进行检验还有很多约束条件，比如要求两变量之间要有线性关系，因此就造成了一定的局限。

5.2.3 主成分分析

通过主成分分析得到的成分没有具体的名字代表，也没有指向性，难以理解它到底指的是什么，可能会在使用过程中造成逻辑混乱，除此之外，它要求参与变换的矩阵必须是方阵因此，在特征值的分解上也造成了一些局限性。它的合理性也有一定的前提，那就是分析出的前几个主成分需要在贡献率上有较高的水平，否则，分析就会失去它的意义。

六、 模型的改进与推广

6.1 模型的改进

问题一中根据检测风化点检测数据来预测风化前的化学成分含量模型中，对于一些数据的选择和处理可以采取更为可靠的算法如机器学习等。

对于问题二中我们采用了主成分分析法来分析两种玻璃的分类规律，在亚分类过程中遇到样品数据较少，聚类分析效果不是很好，可以通过收集更多样本数据来改进。

对于问题三中的敏感性分析，没有得出分类结果与划分依据的关系表达式，没能很好的从数学的角度来分析。

6.2 模型的推广

本文主要解决的是非数值型变量间的相关性问题，在转化为数值型数据后进行数据分析的方法上适用于其它相似的模型问题上。使用的主成分分析方法和卡方检验可以很好地对数据各个变量相关关系进行分析和验证。在根据化学成分进行亚分类的模型中，不仅可应用于古代玻璃文物的鉴别与分类，也可应用于根据化学成分进行分类的研究中。

在对分类依据的选择方法上也是具有普适性的，先对各个变量的数据进行统计分析，选择具有显著性差异的数据作为分类划分依据，在利用聚类的方法进行分类，这个分类模型逻辑清晰，理解上也较为简单，可以作为其他领域的推广使用。

参考文献

[1]巫锡勇,罗健,魏有仪.岩石风化与岩石化学成分的变化研究[J].地质与勘探,2004(04):85-88.

[2]https://baike.baidu.com/item/Pearson%E7%9B%B8%E5%85%B3%E7%B3%BB%E6%95%B0/6243913?fr=aladdin.Pearson 相关系数.2022.9.16

[3]王承遇,陶瑛,陈敏,黄明.钠钙铝镁硅酸盐玻璃和碱铅硅酸盐玻璃的风化[J].硅酸盐通报,1989(06):1-9.DOI:10.16552/j.cnki.issn1001-1625.1989.06.001.

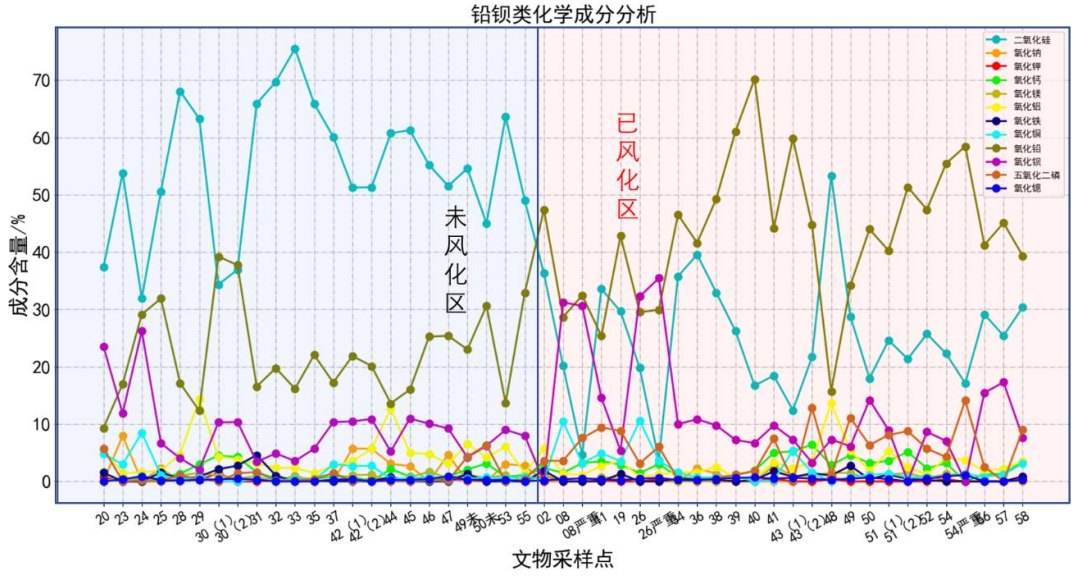
[4]陈倩倩. 基于成分数据的灰色预测模型及其应用研究[D].江南大学,2020.DOI:10.27169/d.cnki.gwqgu.2020.000890.

[5] 茆诗松, 王静龙, 濮晓龙, 等. 高等数理统计 (第二版)[M]. 北京:高等教育出版社, 2006.
[6] Saroj,Kavita.Review:study on simple k mean and modified K mean clustering technique[J].International Journal of Computer Science Engineering and Technology,2016,6(7)：279-281.
[7]张国文,邱萍,倪永年.主成分分析法用于食品样品分类研究[J].食品科技,2003,(12):72-75.
[8]Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

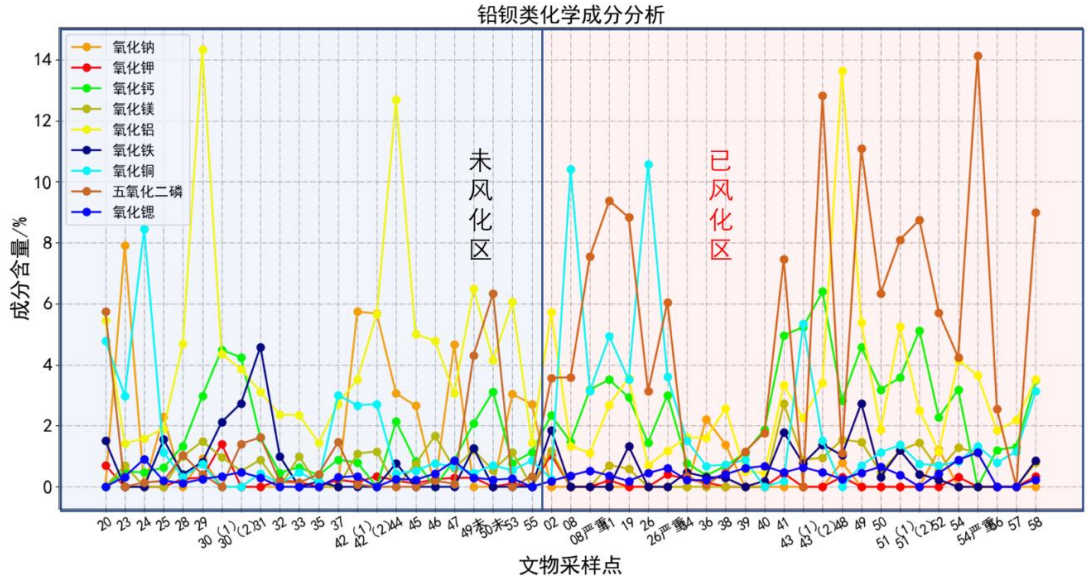
附录

支撑材料含：高钾无风化.xlsx、高钾无风化_数据集标注.csv、卡方检验交叉热力图_类型.png、卡方检验交叉热力图_纹饰.png、卡方检验交叉热力图_颜色.png、铅钡无风化（更新）、铅钡无风化_集聚类标注.csv、预测 2 号文物.xlsx、主成分分析高钾类.xlsx、主成分分析铅钡类.xlsx、mainPb.py、Pb Ba.py、q2 pb.py、relation1.py、工作簿 1.py、工作簿 1.xlsx、表 3.xlsx、信息熵.ipynb、统计规律.ipynb、皮尔逊系数计算.ipynb、表 1.xlsx、q2.m

- 问题一中铅钡类文物化学成分图



铅钡类化学成分含量图



铅钡类化学成分含量放大图

- 问题一中三种类型化学成分数据统计情况

未风化铅钡类数据统计情况

化学成分	平均数	最大值	最小值	中位数	标准差

SiO_2	54.66	75.51	31.94	54.61	11.83
Na_2O	1.68	7.92	0	0	1.37
K_2O	0.22	1.41	0	0.15	0.31
CaO	1.32	4.49	0	0.84	1.28
MgO	0.64	1.67	0	0.71	0.55
Al_2O_3	4.46	14.34	1.42	3.86	3.26
Fe_2O_3	0.74	4.59	0	0	1.15
CuO	1.43	8.46	0	0.65	1.97
PbO	22.08	39.22	9.30	20.12	8.22
BaO	9.00	26.23	2.03	8.99	5.83
P_2O_5	1.05	6.34	0	0.19	1.85
SrO	0.27	0.91	0	0.26	0.24

风化铅钡类数据统计情况

化学成分	平均数	最大值	最小值	中位数	标准差
SiO_2	25.22	53.33	3.72	25.43	10.70
Na_2O	0.22	2.22	0	0	0.57
K_2O	0.14	1.05	0	0	0.24
CaO	2.80	6.40	0.37	2.93	1.60
MgO	0.63	2.73	0	0.55	0.71
Al_2O_3	2.94	13.65	0.45	2.25	2.68
Fe_2O_3	0.61	2.74	0	0.32	0.74
CuO	2.31	10.57	0	1.13	2.87
PbO	42.71	70.21	15.71	44.00	12.08
BaO	12.28	35.45	0	8.94	9.88
P_2O_5	4.92	12.83	0	4.24	3.87
SrO	0.39	0.88	0	0.41	0.23

风化高钾类数据统计情况

化学成分	平均数	最大值	最小值	中位数	标准差
SiO_2	93.96	96.77	92.35	93.50	1.73
Na_2O	0	0	0	0	0
K_2O	0.54	1.01	0	0.67	0.54
CaO	0.87	1.66	0.21	0.83	0.87
MgO	0.20	0.64	0	0	0.20
Al_2O_3	1.93	3.60	0.81	1.72	1.93

Fe_2O_3	0.27	0.35	0.17	0.28	0.27
CuO	1.56	3.24	0.55	1.54	1.56
PbO	0	0	0	0	0
BaO	0	0	0	0	0
P_2O_5	0.28	0.61	0	0.28	0.28
SrO	0	0	0	0	0

- 问题二中未风化铅钡类和高钾类统计数据

未风化铅钡类数据统计情况

化学成分	平均数	最大值	最小值	中位数	标准差
PbO	22.08	39.22	9.30	20.12	8.22
BaO	9.00	26.23	2.03	8.99	5.83
CuO	1.43	8.46	0	0.65	1.97
SiO_2	54.66	75.51	31.94	54.61	11.83
Al_2O_3	4.46	14.34	1.42	3.86	1.26
P_2O_5	1.05	6.34	0	0.19	1.85
Na_2O	1.68	7.92	0	0	1.37
CaO	1.32	4.49	0	0.84	1.28
Fe_2O_3	0.74	4.59	0	0	1.15
MgO	0.64	1.67	0	0.71	0.55
K_2O	0.22	1.41	0	0.15	0.31
SrO	0.27	0.91	0	0.26	0.24

未风化高钾类数据统计情况

化学成分	平均数	最大值	最小值	中位数	标准差
SiO_2	9.33	14.52	0	9.83	3.92
Na_2O	5.33	8.70	0	6.10	3.09
K_2O	1.08	1.98	0	1.17	3.68
CaO	6.62	11.15	3.05	6.19	2.49
MgO	1.93	6.04	0	2.11	1.66
Al_2O_3	2.45	5.09	0	2.34	1.68
Fe_2O_3	0.41	1.62	0	0.15	0.59
CuO	0.60	2.86	0	0	0.98
PbO	1.40	4.50	0	1.02	1.43
BaO	0.04	0.12	0	0.02	0.05
P_2O_5	0.20	2.36	0	0	0.68
SrO	0.10	0.47	0	0	0.19

- 问题一中折线图代码（改变 plt.plot 语句可更改数据读取情况，其余折

线图代码类似）

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
plt.rcParams[&apos;font.sans-serif&apos;]=[&apos;SimHei&apos;]  # 用来正常显示中文标签
plt.rcParams[&apos;axes.unicode_minus&apos;]=False
df = pd.read_excel("C:/Users/86181/Desktop/数模/国赛/铅钡.xlsx")
plt.plot(df[" 文 物 采 样 点 "],df[" 二 氧 化 硅 (SiO2)"],label=&apos; 二 氧 化 硅 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;c&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 钠 (Na2O)"],label=&apos; 氧 化 钠 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;orange&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 钾 (K2O)"],label=&apos; 氧 化 钾 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;r&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 钙 (CaO)"],label=&apos; 氧 化 钙 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;lime&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 镁 (MgO)"],label=&apos; 氧 化 镁 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;y&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 铝 (Al2O3)"],label=&apos; 氧 化 铝 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;yellow&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 铁 (Fe2O3)"],label=&apos; 氧 化 铁 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;navy&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 铜 (CuO)"],label=&apos; 氧 化 铜 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;cyan&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 铅 (PbO)"],label=&apos; 氧 化 铅 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;olive&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 钡 (BaO)"],label=&apos; 氧 化 钡 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;m&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 五 氧 化 二 磷 (P2O5)"],label=&apos; 五 氧 化 二 磷 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;chocolate&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 锶 (SrO)"],label=&apos; 氧 化 锶 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;b&apos;,markersize=8)
plt.xlabel("文物采样点",fontsize=22)
plt.yticks(fontsize=20)
plt.xticks(fontsize=15,rotation=30)
#横坐标为物品编号
plt.ylabel(&apos;成分含量/%&apos;,fontsize=22)
#纵坐标为各类指标
plt.title("铅钡类化学成分分析",fontsize=22)
plt.legend(fontsize=10)
#显示虚线网格
plt.grid(linestyle=&apos;-.&apos;)
#显示图像
plt.show()
```

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
plt.rcParams[&apos;font.sans-serif&apos;]=[&apos;SimHei&apos;]  # 用来正常显示中文标签
plt.rcParams[&apos;axes.unicode_minus&apos;]=False
df = pd.read_excel("C:/Users/86181/Desktop/数模/国赛/高钾.xlsx")
plt.plot(df[" 文 物 采 样 点 "],df[" 二 氧 化 硅 (SiO2)"],label=&apos; 二 氧 化 硅 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;c&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 钠 (Na2O)"],label=&apos; 氧 化 钠 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;orange&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 钾 (K2O)"],label=&apos; 氧 化 钾 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;r&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 钙 (CaO)"],label=&apos; 氧 化 钙 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;lime&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 镁 (MgO)"],label=&apos; 氧 化 镁 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;y&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 铝 (Al2O3)"],label=&apos; 氧 化 铝 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;yellow&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 铁 (Fe2O3)"],label=&apos; 氧 化 铁 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;navy&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 铜 (CuO)"],label=&apos; 氧 化 铜 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;cyan&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 铅 (PbO)"],label=&apos; 氧 化 铅 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;olive&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 钡 (BaO)"],label=&apos; 氧 化 钡 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;m&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 五 氧 化 二 磷 (P2O5)"],label=&apos; 五 氧 化 二 磷 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;chocolate&apos;,markersize=8)
plt.plot(df[" 文 物 采 样 点 "],df[" 氧 化 锶 (SrO)"],label=&apos; 氧 化 锶 &apos;,linewidth=2,marker=&apos;o&apos;,color=&apos;b&apos;,markersize=8)
plt.xlabel("文物采样点",fontsize=22)
plt.yticks(fontsize=20)
plt.xticks(fontsize=15,rotation=30)
#横坐标为物品编号
plt.ylabel(&apos;成分含量/%&apos;,fontsize=22)
#纵坐标为各类指标
plt.title("高钾类化学成分分析",fontsize=22)
plt.legend(fontsize=10)
#显示虚线网格
plt.grid(linestyle=&apos;-.&apos;)
#显示图像
plt.show()
```

- 问题一信息增益

```
def info_entropy(attr):
    prob = pd.value_counts(attr) / len(attr)    # 对于一个特征不同类所占的比例类
    return sum( np.log2( prob )* prob * (-1) )   # 经验熵

# 信息增益    （返回值越大，attr1 与 attr2 相关性越强）
def info_gain(ssh1):
    ent1= ssh1.groupby('表面').apply(lambda x: info_entropy(x['颜色']))
    prob = pd.value_counts(ssh1['表面']) / len(ssh1['表面'])
    ent2= sum( ent1 * prob )                    # 经验条件熵
    return info_entropy(ssh1['颜色']) - ent2       #   信息增益

print(info_gain(ssh1),'颜色','表面')

def info_entropy(attr):
    prob = pd.value_counts(attr) / len(attr)    # 对于一个特征不同类所占的比例类
    return sum( np.log2( prob )* prob * (-1) )   # 经验熵

# 信息增益    （返回值越大，attr1 与 attr2 相关性越强）
def info_gain(ssh1):
    ent1= ssh1.groupby('表面').apply(lambda x: info_entropy(x['类型']))
    prob = pd.value_counts(ssh1['表面']) / len(ssh1['表面'])
    ent2= sum( ent1 * prob )                    # 经验条件熵
    return info_entropy(ssh1['类型']) - ent2       #   信息增益

print(info_gain(ssh1),'类型','表面')
def info_entropy(attr):
    prob = pd.value_counts(attr) / len(attr)    # 对于一个特征不同类所占的比例类
    return sum( np.log2( prob )* prob * (-1) )   # 经验熵

# 信息增益    （返回值越大，attr1 与 attr2 相关性越强）
def info_gain(ssh1):
    ent1= ssh1.groupby('表面').apply(lambda x: info_entropy(x['纹饰']))
    prob = pd.value_counts(ssh1['表面']) / len(ssh1['表面'])
    ent2= sum( ent1 * prob )                    # 经验条件熵
    return info_entropy(ssh1['纹饰']) - ent2       #   信息增益

print(info_gain(ssh1),'纹饰','表面')
```

- 问题二统计规律

```
#高锰风化
b1=header_row                                                    =
['SiO2','Na2O','K2O','CaO','MgO','Al2O3','Fe2O3','CuO','PbO','BaO','P2O5','SrO']
b1 = pd.read_excel('c:\\Users\\huawei\\Desktop\\ 风 化 的 高 钾 .xlsx',header=None,
```

```
names=header_row)
    pd.set_option('display.float_format',lambda X:'%.2f% X)
    b1.describe()
    #高锰未风化
    b2=header_row                                                    =
['SiO2','Na2O','K2O','CaO','MgO','Al2O3','Fe2O3','CuO','PbO','BaO','P2O5','SrO']
    b2 = pd.read_excel('c:\\Users\\huawei\\Desktop\\未风化的高钾类.xlsx',header=None,
names=header_row)
    pd.set_option('display.float_format',lambda X:'%.2f% X)
    b2.describe()
    #铅钡风化
    b3=header_row                                                    =
['SiO2','Na2O','K2O','CaO','MgO','Al2O3','Fe2O3','CuO','PbO','BaO','P2O5','SrO']
    b3 = pd.read_excel('c:\\Users\\huawei\\Desktop\\风化的铅钡.xlsx',header=None,
names=header_row)
    pd.set_option('display.float_format',lambda X:'%.2f% X)
    b3.describe()
    #铅钡未风化
    b4=header_row                                                    =
['SiO2','Na2O','K2O','CaO','MgO','Al2O3','Fe2O3','CuO','PbO','BaO','P2O5','SrO']
    b4 = pd.read_excel('c:\\Users\\huawei\\Desktop\\未风化的铅钡类.xlsx',header=None,
names=header_row)
    pd.set_option('display.float_format',lambda X:'%.2f% X)
    b4.describe()
```

- 问题二主成分分析

```
DATASET ACTIVATE 数据集 1.
FACTOR
  /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
  /MISSING LISTWISE
  /ANALYSIS 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
  /PRINT INITIAL CORRELATION KMO EXTRACTION ROTATION FSCORE
  /PLOT ROTATION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION.
```

- 问题三聚类

```
QUICK CLUSTER 氧化铅 PbO 氧化钡 BaO 氧化钾 K2O
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(2) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER
  /PRINT INITIAL ANOVA.
```

- 问题三分类代码

```
% clear;
% clc;
% close;
% %1 是铅钡，-1 是高钾
%  x = [1.36 39.58 4.69];
%  result1=zeros(30,3);
%  y=repmat(x,30,1);
%  m=0;
%  for k=1:3
%      for i=0.1:0.1:1
%          m=m+1;
%          y(m,k)=x(1,k)*(1+i);
%          [result1(m,k)]=judge(y)
%      end
%  end
x=xlsread(&apos;C:\Users\huawei\Desktop\表 3.xlsx&apos;,&apos;E2:G25&apos;);
judge(x);
function [j]=judge(x)
[m,n]=size(x);
q1=[9.33 0.41 0.60];
q2=[0.22 22.08 9.00];
for i =1:m
    d1=norm(x(i,:)-q1);
    d2=norm(x(i,:)-q2);
    if d1>d2
       j=1;
       %disp(&apos;铅钡类&apos;);
    else
       j=-1;
       %disp(&apos;高钾类&apos;);
    end
end
end
```