基于多元统计分析的玻璃文物特征规律的挖掘

摘要

古代玻璃是东西方经济文化交流的重要见证，其成分、类别上的差异反映东西方不同的工艺特色与地质特征。本文基于主成分分析、聚类分析等多元统计分析方法，从定性和定量两种视角，对不同类别的古代玻璃开展多层次的关联性分析，试图挖掘各个特征及其关系的统计规律，为古代玻璃样本的类型判断、成分分析提供统计依据。

针对问题 1：对于第（1）小问，需要分析玻璃文物表面风化与三个定性变量之间关系。本文基于 Apriori 关联规则挖掘和数据侧写提供的先验关联性，采取定性的**卡方检验**与定量的对数线性模型，发现其中类别对表面风化的影响较为显著；对于第（2）、（3）小问，需要分析文物采样点化学成分的统计规律，本文通过定性的方差分析与差异性分析，分别挖掘各化学成分与类别、表面风化之间关系，通过**多因素方差分析和多重比较思想**下的聚类分析，量化统计规律并据此预测风化点风化前的化学成分含量。可知其中二氧化硅、氧化钾等 7 种化学成分与类别和表面风化显著关联。

针对问题 2：对于第（1）小问，分析玻璃不同类别的分类规律是一个**分类预测问题**，在可视化等定性分析的基础上，本文基于 **Fisher 判别分析**发掘并量化分类规律；对于第（2）小问，对每个类别的玻璃划分亚类是一个**聚类问题**，本文基于现有研究成果，结合主成分分析，构建铅钡玻璃不同年代隶属度指标进行聚类，最终划分为 3 个亚类；选取高钾玻璃中氧化铝、氧化钙的等主要成分进行聚类，最终划分为 5 个亚类。对两组聚类结果的合理性和敏感性分析表明，各个亚类的解释能力突出，聚类性能对样本数据 10%以内的随机扰动表现稳健。

针对问题 3：首先需要对表单 3 中未知类别的玻璃文物化学成分进行判别，本文沿用了问题 2 的 Fisher 判别分析进行判断，结果表明除了样本 A6 和 A7 外均属于铅钡类别。

其次，将 Fisher 判别分析的预测结果与逻辑回归、朴素贝叶斯分类的预测结果进行 **Kappa 一致性检验**，结果表明一致性显著，说明 Fisher 判别模型较为稳健。

针对问题 4：对于第（1）小问，对于不同类别的样品，首先通过皮尔逊相关系数进行定性判断，其次在因子分析的基础上进行因子之间定量的**典型相关分析**；对于第（2）小问，本文将不同类别之间化学成分关联关系的差异归结于结构性差异和规律性差异。

关键词: 玻璃文物 聚类模型 判别模型 方差分析 卡方检验

问题重述

1.1 问题背景

早期玻璃通过丝绸之路由西方地区传入我国，我国古代玻璃吸收其技术后采用本土材料制作玻璃，因此中西方玻璃制品的化学成分不尽相同。古代玻璃埋藏时易受环境影响而风化，造成其内部成分比例的改变，影响对其类别的判断。

1.2 基本信息

已知一批我国古代玻璃文物的数据，附件中表单 1 给出了编号、纹饰、类型、颜色和表面风化情况等信息，表单 2 给出了相应主要成分所占比例。

1.3 问题提出

我们需要解决的问题如下：

1. **问题 1 第(1)问**：根据表单 1，对玻璃文物表面风化情况和其玻璃类型、纹饰与颜色进行定类数据的相关性分析；
   问题 1 第(2)问：根据表单 2 并结合玻璃类型，分析文物样品表面有无风化化学成分含量的统计规律，
   问题 1 第(3)问：并预测风化点数据风化前的值。
2. **问题 2 第(1)问**：分析两种类型玻璃的分类规律；
   问题 2 第(2)问：在每一类中选取适当的化学成分作为分类指标对其划分亚类，给出划分方法和结果，并分析结果的合理性和敏感性。
3. **问题 3**：关于表单 3 中未知分类的玻璃文物，通过对其化学成分的合理分析鉴别其所属玻璃类型，并分析结果的敏感性。
4. **问题 4**：关于不同类别的玻璃文物样本，进行其化学成分的关联性分析与不同类别之间的化学成分关联关系的差异性分析。

2 问题分析

2.1 问题 1 的分析

2.1.1 第(1)问分析

文物表面风化情况、纹饰、类型、颜色都是定类数据，彼此间构成列联表，显然不适用连续型数据分析方法。在列联分析中，卡方检验可用于分类资料相关分析和比较分类数据的差异性；线性对数模型可将列联表中的单元格频数的对数表示为各个变量及其之间交互效应的线性模型，用于纯粹定类变量间关系的评价。在这两种方法之前，可以使用关联规则挖掘算法(Apriori)先验地找到类型数据间的关联关系，再加以检验。

2.1.2 第(2)问分析

要求根据玻璃文物的采样点的数据，结合其类型、是否风化，对于化学成分分布的统计规律进行分析，本质上是一个定性数据与定量数据的关联性分析问题。（在第二小问中我们仅研究定性规律，定量规律的测定我们会在第三问中说明）；研究内容是玻璃文物采样点的化学成分与类别、风化条件的关系，研究对象是玻璃文物采样点。本文首先通过单因素方差分析，研究不同类别下化学成分的分布的差异性；在此基础上进行差异性分析，即进行不同类别分组下风化前后化学成分分布的差异性分析。之所以没有采用双因素方差分析，是为了先将不同类型作为分组条件，再研究风化与否对于化学成分分布的差异性与关联性。

2.1.3 第(3)问分析

本文利用多因素方差分析和多重比较的思想。问题 1 第三小问的中心问题在于估计风化对于化学成分的偏效应，偏效应是指在其他因素不变的情况下（比如类别、纹饰、颜色）未风化样本被风化导致的化学成分的水平变化。鉴于不同的类别、纹饰、颜色分组下，偏效应有一定的差异，所以我们针对上述三个变量进行聚类分析，聚类而得来的每一小类有着较为集中的类别、纹饰、颜色，所以我们针对每一小类的偏效应进行估计，预测数据较为有效。 在对此类问题的定性分析中，如果定性数据划分的子样本较少、子样本之间趋于均衡，可以采用参数方法下的双样本 t 检验或者非参数方法下的双总体 Wilcoxon 符号秩检验的方法做差异性分析；如果定性数据维度较高，可以采用多因素方差分析，研究不同变量分组下因变量（所研究的定量数据）的水平差异。在对该问题的定量分析中，可以采取回归分析、多重比较等方法量化子样本之间的差异。 该问题的研究内容是玻璃文物采样点的化学成分与类别、风化条件的关系，研究对象是玻璃文物采样点，所以玻璃文物采样点的化学成分以表单 2 为唯一依据，玻璃文物采样点的类别以该采样点所属的玻璃文物的类别为唯一依据。在未特殊注明下，玻璃文物采样点是否风化的判断以所属文物的是否风化为唯一依据。但对于已风化文物的未风化点，基于文献结果和模型假设，本文认为文物采样点是否风化的判断最终以其化学成分作为唯一依据，所以更正为“未风化”。

2.2 问题 2 的分析

2.2.1 第(1)问分析

需要分析铅钡玻璃和属于典型的分类预测问题；关于变量的选取，可以选取问题 1 中与类别关联性比较大的硅、钾、铅、钡等连续型指标，以及“是否风化”这一定性变量，采用 Fisher 判别的方法对样本的类型进行归类。

2.2.2 第(2)问分析

文首先进行信息去噪，将风化后的样本各项指标反推回风化前样本的各项指标，更容易挖掘数据内在的集聚性与规律性。对于铅钡类的数据，首先通过主成分分析构建聚类指标体系，再根据玻璃文物对各个年代的隶属程度划分为不同年代的玻璃文物；对于高钾类的数据，根据草木灰、钙、铝、显色剂、钠钾比五项指标进行聚类。最后是灵敏度分析，通过为聚类样本增加噪音并观察聚类结果和聚类性能的变化是否显著，判断得出，两个聚类模型都比较稳健

2.3 问题 3 的分析

对于有连续型指标的分类数据的预测，可以使用问题 2 第(1)问中的判别分析法对文物预测归类，为了确保分类模型的强稳健性，同时进行 Logit 回归和朴素贝叶斯回归，将三种方法得到结果放在一起进行 Kendall 一致性检验，进行灵敏度分析时对样本数据随机加 1—10%的噪声，再进行检验，看是否影响模型的稳健性。

2.4 问题 4 的分析

要求针对不同类型的玻璃文物样本，分析其化学成分之间的关联关系。 关联分析首要在于关联关系的存在性。对此，本节通过求解两种类型的样本所有化学成分两两之间的皮尔逊相关系数；在相关性显著的基础上，找出两两之间具有较强相关性的化学成分。 其次，为了简化变量之间的相关关系，本节利用主成分分析将两种类型玻璃样本的化学成分进行

分划，以便后续分组变量之间的研究。 然后基于降维后的数据，开展定量的关联分析。借助典型相关分析，对两组之间作关联性分析，探索并量化不同组别之间的关联关系。 最后针对每组数据的关联性分析结果，做描述性的差异性分析。

3 模型假设

3.1 假设 1

表单中所给数据对文物采样点的观测结果无误。

3.2 假设 2

考虑到“严重风化”样本个体体量太小，将其视同“风化”；因缺少同一个体不同风化情况部位的对比，认为风化个体的未风化点数据反映了该文物未风化前的理化性质。

3.3 假设 3

为方便玻璃亚类划分，此批玻璃文物的来源地和出产年代可能有较大不同

4 符号说明

符号	含义
X_i	玻璃文物的类型
Y_j	玻璃文物的装饰
Z_k	玻璃文物表面风化
f_{ijk}	类型、装饰和表面风化构成的三维列联表的频数。
$MAIN\_i$	主成分分析后的第 i 种主成分
$color$	颜色属性
$type$	类型属性
$pattern$	纹饰属性
d	是否风化属性
ω_i	变异系数法算的对应成分的权重
KC	草木灰指标
$pattern\_color$	显色剂指标
$ratio$	第一主族和第二主族的比例

5 模型建立与求解

5.1 .模型准备

查阅古代玻璃成分和分析的相关文献，得到一些信息如下：

1） 关于助熔剂：按照不同的助熔剂，古代玻璃一般可分为（1）K2O-CaO-SiO2 系统（其中 K2O/Na2O> 1）;（2）BaO-PbO-SiO2 系统和 K2O-SiO2 系统;（3）PbO-SiO2 系统;（4）K2O-PbO-SiO2 系统;（5）K2O-CaO-SiO2 系统[1]。

2） 关于风化：不同程度的风化腐蚀会造成助熔剂 Na2O、K2O、PbO 和 BaO 不同程度的流失，随着风化残损程度加深，流失越严重，不少样品甚至形成磷酸钙的风化产物[2]。风化

造成的风化条痕中，风化颗粒中 SiO2、Al203 的含量明显增高，CaO、MgO 则明显减少，Na2O 却基本消失[3]。

3） 关于玻璃颜色：蓝(绿) 色着色剂以 CuO 和 Fe2O3 为主，其中 CuO 含量在 2% 以上，系人为引入，不是杂质，黑色和红色着色剂以 Fe2O3 为主，Mn3+着色为红色，与其他离子如 Fe2+或 Fe3+组合成紫色。

4） 关于污染物：P2O5 含量可能和样品的污染有关[4]。

5） 关于铅钡玻璃：含铅量高的玻璃，由于其 SiO2 含量相应较低，故其化学稳定性较差，在长期的埋藏过程中，易受到各种因素如 SO2、H2S、CO2 等酸腐蚀，生成结构疏松的风化层[5]。

6） 关于高钾玻璃：在钾玻璃中，其风化表面的 Si，Al，Ca 及 Fe 元素含量相对于新鲜面是富集的，而 K 元素流失很明显; Na 元素在风化比较严重的玻璃中含量显著增加，在风化较轻的玻璃中稍有减少，而 Mn，Ti 及其它微量元素变化不十分显著，表明风化对于钾玻璃中的微量元素的影响远小于主量及次量元素[6]

根据上述资料数据，我们把玻璃中化学成分划分为以下六种：

1、 玻璃中的主要成分：二氧化硅(SiO2)

2、 玻璃中的次要成分：氧化铝(Al2O3)、氧化镁(MgO)

3、 玻璃制造的助熔剂：氧化钠(Na2O)、氧化钾(K2O)、氧化铅(PbO)、氧化钡(BaO)、氧化钙(CaO)。

4、 玻璃中的颜色物质：氧化铁(Fe2O3)、氧化铜(CuO)

5、 玻璃中的杂质：氧化锶(SrO)、氧化锡(SnO2)

6、 风化导致的污染物：五氧化二磷(P2O5)、二氧化硫(SO2)

5.2 数据预处理

为方便后续模型的建立，精准挖掘附件数据的统计规律与分布特征，首先需要对原始数据进行清洗和预处理，去除噪音和无效数据，发掘数据规律，避免在后续统计分析中由于数据不准导致的错误。

Step1：数据完整性判断

数据的完整与否，首要在于是否有缺失值：附件数据中缺失值较为集中的区域在于表单 2 中各个样本点的化学成分，根据题意，样本点中某化学成分存在缺失，意味着该成分未被检测到，所以缺失值的零值替换有一定的合理性；不难发现，氧化锡、二氧化硫缺失值分布达到了 91.0%和 89.5%，一般而言，如果某个特征缺失值比例如果超过 75%~80%，那么该指标的观测数据随机性较大、极易得到缪误结论，所以在后续的统计分析中可以忽略氧化锡、二氧化硫这两个特征。表单 1 中有 4 组玻璃文物的颜色观测是缺失的，为了保证模型的合理性，并与现实情况相符合，可以将该类缺失值替换为单独一类的观测，称为“颜色无法识别”。

数据的完整性还在于，数据之间的观测是否是正交的；不难发现，表单 1 中缺失了纹饰为“B”并且类别为“铅钡”的样本数据。对于这一样本缺失，可以认为是由纹饰、类别两个定性变量之间的相关性导致的。

Step2：数据有效性判断

首先，根据题意，成分比例累加和在 85%以下或在 105%以上视为无效数据，出现了数据的谬误观测，所以表单 2 中文物采样点为 15、17 的两个样本可以删去；表单 3 中待预测类别的数据均是正常的；

其次，数据的有效性还在于数据的集中分布，不存在离群值。玻璃文物采样点的若干化学成分的箱线如图 5.2.1 和图 5.2.2 所示，氧化钾、氧化钡的分布呈现正偏态，离群数据较多；为保证信息的完整性，可以对偏态数据取对数处理，使得数据具有分布的正态性。

图 5.2.1 氧化钾含量比例箱线图

图 5.2.2　氧化钡含量比例箱线图

Step3：样本均衡判断

由于样本数据存在较多定类变量，比如类别、是否风化这类指标，所以有必要考虑样本不均衡问题；样本不均衡问题是指，各个类别之间的样本数据体量差异过大，导致数据挖掘中出现潜在的过拟合问题或者欠拟合问题。如图 5.2.3 玻璃类型数量比例图所示，类别为“铅钡”的玻璃文物样本量明显多于类型为“高钾”的样本量，对此可以采用下采样和过采样相结合的方法，可以对体量较多的数据随机选取一定的样本，对体量较小的数据随机取样并加入原数据，作为数据的重复观测。最终均衡后的样本中两种类别体量均为 30 组数据。

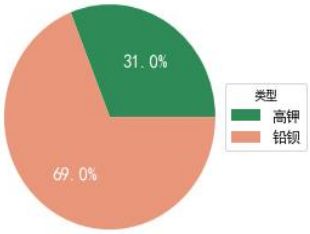


图 5.2.3 玻璃类型数量比例图

Step4：标准化处理

由于数据的数量级不同，离散程度不同，所以有必要对数据进行 z-score 标准化处理。处理后的数据大致服从于标准正态分布，例如“二氧化硅”这一指标的分布情况如图 5.2.4 和图 5.2.5 所示：

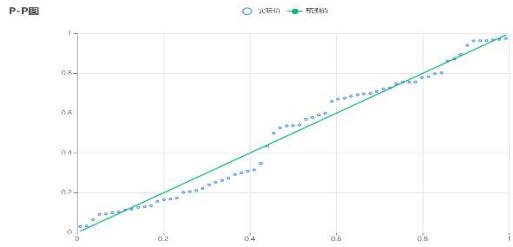


图 5.2.4 二氧化硅 P-P 图

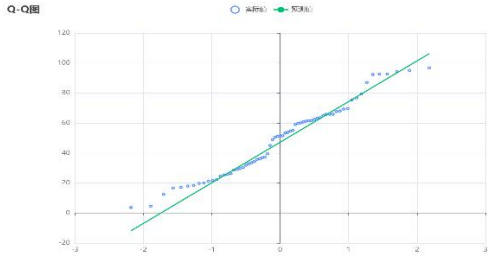


图 5.2.5 二氧化硅 Q-Q 图

5.3 问题 1 第（1）问的建模与求解

问题 1 的第一问要求研究文物表面风化和纹饰、类型和颜色之间的关系，为了严谨性，也防止由于表单 1 中文物数据体量较小，单一的数据分析方法可能无法发掘数据内在规律或被偶然现象所误导，故我们采用数据可视化与描述统计、先验分析、定性检验和定量分析相结合的方法，发掘数据特点并进行检验。具体流程如图 5.3.1 问题 1 第(1)问求解流程示意图所示

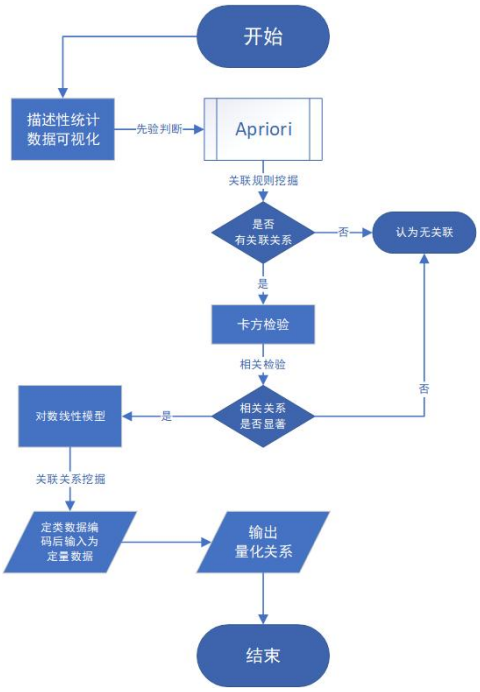


图 5.3.1 问题 1 第(1)问求解流程示意图

5.3.1 定性分析：数据可视化与描述性统计

由图 5.3.2 可以明显看出两种玻璃的特征纹饰，铅钡玻璃只有 A、C 两种纹饰类型，C 纹饰数量较高，可能与类型相关程度较高；高钾玻璃有 A、B、C 三种纹饰类型，每种类型数量表现为均衡分布，可以知道 B 是高钾玻璃的特征纹饰，结合题目，可以解释为古代中西方制作材料或文化及审美差异所致。

由图 5.3.3 可以明显看出两种玻璃的颜色特征，高钾玻璃主要为“蓝绿”、“浅蓝”，铅钡玻璃以“浅蓝”、“深绿”为主。玻璃颜色共有八种类型，其中有四个数据点的颜色无法识别，均为铅钡玻璃。高钾玻璃的颜色情况只有四种，分别为“蓝绿”、“深绿”、“浅蓝”和“深蓝”，且集中表现为“蓝绿”色；铅钡玻璃各个颜色均有，其中“浅蓝色”最多，“深绿色”次之。可以初步推断，高钾玻璃和铅钡玻璃的化学成分及含量不同，导致其表现出一定的特有的颜色。

由图 5.3.4 可以比较直观地看出两种玻璃各自的风化比例，高钾玻璃中风化个体与未风化个体的比例为 6：12，铅钡玻璃中风化个体与未风化个体比例为 28：12，铅钡玻璃风化比例明显高于高钾玻璃。

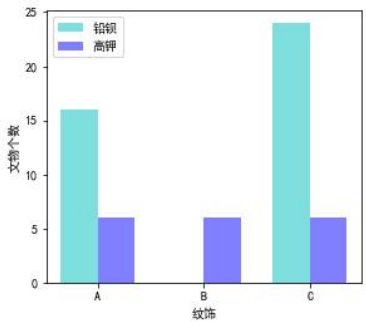


图 5.3.2 不同玻璃类型

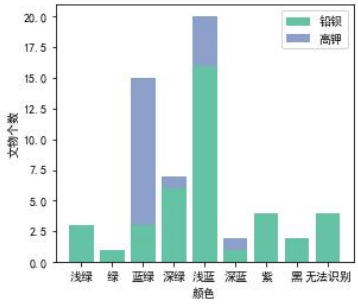


图 5.3.3 不同玻璃类型颜色情况图

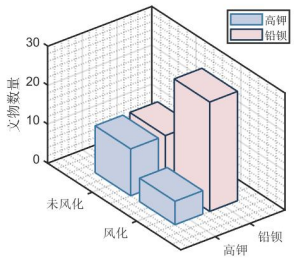


图 5.3.4 不同玻璃类型风化情况图

5.3.2 定性分析：基于 Apriori 算法的关联规则挖掘

Apriori 关联规则挖掘属于机器学习范畴下数据关联性分析的一种，适用于发掘“某一事件发生而引起的另外一件事件发生”的统计规律。可以结合案例“啤酒和尿布的故事”来说明算法的原理。

Step1：基于统计数据发现，67%的顾客在购买啤酒的同时也会购买尿布；在这一观测下，“购买啤酒”称为规则的“**前项**”，购买尿布称为规则的“**后项**”。所有前项和后项构成一个**项集**，每一前项和后项之间构成项集中的一个**规律**，记作：

$$\{Diaper\} \Rightarrow \{Beer\}$$

Step2：介绍项集的**支持度**这一概念，表示所有的事务（在本例下，等同于所有的交易）下出现该项集的概率，可以通过前项后项概率的加和来计算：

$$Support\left(\{Diaper, beer\}\right) = P\left(Diaper \cup Beer\right) \tag{1}$$

其中 Support 表示该项集的支持度，表示该类规律的频繁程度；

Step3：对各个项集的支持度进行排序，若某项集的支持度高于频繁程度，称之为**频繁项集**。

Step4：最后对频繁项集做显著性检验；置信度为前项发生、则后项也发生的概率，记作：

$$Coffidence\left(Diaper \Rightarrow Beer\right) = P\left(Beer \mid Diaper\right) \tag{2}$$

结果和分析在问题 1 第(1)问的情境下，研究对象为是否风化与颜色、纹饰、类别之间的关系，这四个变量均为定性变量，并存在着一定的关联性，如：

（1）纹饰为 B 类型玻璃文物均属于“高钾”类别，“铅钡”类别玻璃文物没有观察到 B 纹饰；

（2）已被风化的玻璃文物不会出现深蓝色和绿色，且颜色无法识别的样本均被风化的；

由于定类变量较多，每个定类变量的取值也比较多，所以人工发掘类似的规律较为困难；所以本文选择将 Apriori 关联规则挖掘算法作为正式研究之前的先验判断和样本依据，精准研究问题、节省人力物力。

通常设置阈值为 0.5，得到频繁项集及其显著性检验如表 5.3.1 所示（部分结果）：

表 5.3.1 关联规则挖掘的频繁项集表

前项	后项	先验支持度
无风化	纹饰 A	0.527777778
纹饰 A	无风化	0.388888889

可见，纹饰 A 风化规律在一定程度上都是有章可循的。这些规律性的结果将会作为接下来卡方检验（定性数据的相关性分析）的起点和依据。在下面的章节将会逐步引入纹饰、类型、颜色三个变量，与是否风化这一变量做相关性分析和检验。

5.3.3 模型建立：基于卡方检验和对数线性模型的关联性定量分析模型

我们根据 Aprioi 算法得到的先验结果，得知一些变量之间存在关联关系或联合关联关系，首先需要对全体定类数据进行列联分析，定性地判断其关联关系。

Ⅰ.卡方检验：

在列联分析中，卡方检验可用于分类资料相关分析和比较分类数据的差异性，下面简单介绍一下卡方检验模型。

一个简单实现卡方检验的方法是 SPSS 软件。SPSS 中默认使用的统计量为皮尔逊卡方，其统计量计算公式为

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i} \tag{3}$$

其中 O_i 代表观测频数， E_i 代表期望频数。

卡方检验的原假设为H_0：$O_i = E_i$，因此若观测值对于期望值出现了较大程度的偏离，则倾向于有较大的卡方，说明数据在不同组别中有较大的差异性。进行卡方检验时，大部分检验结果只需要看皮尔逊卡方输出值即可，只有一类数据情况较为特殊，即 2×2 的列联表，该表的自由度为 1，根据卡方分布曲线图 5.3.5[7]，自由度小于 2 的曲线与其他曲线分布形状有很大差异。

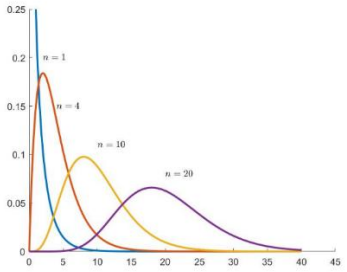


图 5.3.5 卡方分布曲线图

SPSS 输出 2×2 列联表的卡方检验结果时，一共输出皮尔逊卡方，连续性修正和费希尔精确检验三个数值。假设 n 为样本总数，E 为单元格内的期望频数，选择三种检验结果的规则如下：

当 n≥40，且 E>5 时，看皮尔逊卡方值；

当 n≥40，且 1<E<5 时，看连续性修正值；

当 n<40，或 E<1 时，选择费希尔精确检验值

其中费希尔精确检验使用的是超几何分布理论，与卡方分布无关；连续性修正的公式如公式所示

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(\left|O_i - E_i\right| - 0.5\right)^2}{E_i} \tag{4}$$

Ⅱ.对数线性模型：

对数线性模型将列联表中的单元格频数的对数表示为各个变量及其之间交互效应的线性模型，可用于纯粹定类变量间关系的评价。

基于卡方检验判别结果，类型、装饰与表面风化之间存在较大的相关性，故筛去“颜色”数据，对“表面风化情况”、“类型”、“纹饰”构成的三维列联表数据资料代入对数线性模型，进一步探讨三者之间的具体量化关系。

假定不同的行代表第一个变量 X 的不同水平，不同的列代表第二个变量 Y 的不同水平，不同的层代表第三个变量 Z 的不同水平,用 f_{ijk} 表示三维列联表第 i 行，第 j 列，第 k 层所唯一对应的样本类型的频数。

则当三维列联表三个变量相互独立时，对数线性模型可表示为：

$$\ln\left(f_{ijk}\right) = u + X_i + Y_{\mathrm{j}} + Z_k + \varepsilon_{ijk} \tag{5}$$

行变量的第 i 个水平对 $\ln\left(f_{ijk}\right)$ 的影响用 X_i 来表示，列变量的第 j 个水平对 $\ln\left(f_{ijk}\right)$ 的影响用 Y_{j} 来表示，层变量的第 k 个水平对 $\ln\left(f_{ijk}\right)$ 的影响用 Z_{k} 来表示，这三者的影响均为主效应；$\varepsilon_{\mathrm{ijk}}$ 代表随机误差；u 为截距项，表示总平均水平，不属于单独的效应，当对总效应作差时会将其减掉。

当三维列联表三个变量互相都不独立时，则表示为饱和对数线性模型：

$$\ln\left(f_{ijk}\right)=u+X_i+Y_j+Z_k+\left(XY\right)_{ij}+\left(XZ\right)_{ik}+\left(YZ\right)_{jk}+\left(XYZ\right)_{ijk}+\varepsilon_{ijk} \tag{6}$$

这里比主效应模型多出三个两两交互项$\left(XY\right)_{ij}$，$\left(XZ\right)_{ik}$，$\left(YZ\right)_{jk}$，和一个三变量共同交互项$\left(XYZ\right)_{ijk}$。其中，交互项表示不同变量对应的不同层次对$\ln\left(f_{ijk}\right)$的共同影响作用，称为交互效应。交叉项的各个参数的大小也是相对的，也需要约束条件来得到其“估计”。其中，需要注意的是，无论模型中假定了多少种效应，并不见得全都有意义，有些可能是多余的。本来没有交互影响的作用，将其写入模型也可以，但在分析过程中，通过独立性检验的方法一般可以知道哪些是显著的，哪些不是，然后决定取舍。

5.3.4 模型求解：基于卡方检验和对数线性模型的关联性定量分析模型

Ⅰ.卡方检验

Step1:首先借助 SPSS 中描述统计功能探究了三类关系：①纹饰*表面风化②类型*表面风化③颜色*表面风化，得到输出结果中有两类显著，如所示

Step2:分析第一步中结果，引入“类型”作为分类基础，探究纹饰*表面风化*类型，得到的结果如所示。

Ⅱ.对数线性模型

Step1: 设类型为X，装饰为Y，表面风化为Z,对类型、纹饰、表面风化及其之间的交互项进行独立性检验：

Step2:根据独立性检验得到的小表格，分析三者之间的相关情况，选取最适合的对数线性模型。

Step3:估计参数，代入所选取的模型中，计算预测值$\ln\left(f\right)$，再取指数运算求得频率值f。

5.3.5 结果分析

卡方检验共得到三个显著结果，其余均不显著，认为无关联关系。

结果一：

表 5.3.2 **纹饰*表面风化检验表**

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	4.957a	2	0.084
似然比	7.12	2	0.028

分析：皮尔逊卡方值显著，说明纹饰和表面风化之间存在关联关系。

结果二：

表 5.3.3 **类型*表面风化检验表**

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	6.880a	1	0.009
连续性修正 b	5.452	1	0.02
似然比	6.889	1	0.009

分析：皮尔逊卡方显著，说明类型和表面风化之间存在关联关系

结果三：

表 5.3.4 **纹饰*表面风化*类型检验表**

类型		值	自由度	渐进显著性（双侧）
高钾	皮尔逊卡方	18.000b	2	0
	似然比	22.915	2	0
铅钡	皮尔逊卡方	.020c	1	0.888
	连续性修正 d	0	1	1

似然比	0.02	1	0.888

高钾类型下的皮尔逊卡方值显著，铅钡不显著，说明在高钾类型下，纹饰和表面风化之间存在一定的关联关系。

对数线性模型

结果一:

通过独立性检验得到，在给定 Z 的情况下，X 与 Y 之间独立，即表明在任何表面风化下，类型与纹饰是不相关的。根据独立性检验的小标格可得，对于类型 X，纹饰 Y，表面风化 Z 三种变量，最适合的对数线性模型应为:

$$\ln\left(f_{ijk}\right)=u+X_i+Y_j+Z_k+\left(XZ\right)_{ik}+\left(YZ\right)_{jk}+\varepsilon_{ijk} \tag{7}$$

结果二:

表 5.3.5 对数线性模型变量系数表

参数	估算	标准错误	Z	显著性	95% 置信区间	
					下限	上限
常量	1.576	0.371	4.242	0	0.848	2.303
[类型 = 0]	1.517	0.371	4.085	0	0.789	2.245
[类型 = 1]	0	.	.	.	.	.
[表面风化 = 0]	0.345	0.525	0.657	0.511	-0.685	1.374
[表面风化 = 1]	0	.	.	.	.	.
[纹饰 = 1]	-0.621	0.455	-1.365	0.172	-1.512	0.271
[纹饰 = 2]	-1.138	0.455	-2.502	0.012	-2.029	-0.246
[纹饰 = 3]	0	.	.	.	.	.
[类型 = 0] * [表面风化 = 0]	-1.517	0.525	-2.889	0.004	-2.547	-0.488
[类型 = 0] * [表面风化 = 1]	0	.	.	.	.	.
[类型 = 1] * [表面风化 = 0]	0	.	.	.	.	.
[类型 = 1] * [表面风化 = 1]	0	.	.	.	.	.
[表面风化 = 0] * [纹饰 = 1]	0.414	0.643	0.643	0.52	-0.847	1.675
[表面风化 = 0] * [纹饰 = 2]	-0.207	0.643	-0.322	0.748	-1.468	1.054
[表面风化 = 0] * [纹饰 = 3]	0	.	.	.	.	.
[表面风化 = 1] * [纹饰 = 1]	0	.	.	.	.	.
[表面风化 = 1] * [纹饰 = 2]	0	.	.	.	.	.
[表面风化 = 1] * [纹饰 = 3]	0	.	.	.	.	.

设计: 常量 + 类型 + 表面风化 + 纹饰 + 类型 * 表面风化 + 表面风化 * 纹饰

以类型=1，纹饰=1，表面风化=0 为例计算频数结果：

$$\ln\left(f_{110}\right)=1.576+0-0.621+0+0.414=1.369$$

$$\exp\left(1.369\right)\approx 3.931417312$$

5.4 问题 1 第(2)问：基于差异性分析的统计规律挖掘

针对问题 1 的第二小问，要求根据玻璃文物的采样点的数据，结合其类型、是否风化，对于化学成分分布的统计规律进行分析，本质上是一个定性数据与定量数据的关联性分析问题。本节将通过定性分析的方法(定量分析见第(3)问)挖掘统计规律

值得注明的是，我们采用单因素方差分析——差异性分析而不采用双因素方差分析，目的在于观察不同类别分组下，风化条件对化学成分的影响以及差异性。

5.4.1 定性分析：数据侧写与描述性统计

结合类型和表面风化，对表单 2 中化学成分的数据进行分析得到关系图如图 5.4.1 所示

根据对风化前后化学成分含量数据的比较，区分了流失显著的化学成分与流失不显著或者含量升高的化学成分，前者包括玻璃的主要成分二氧化硅(SiO2)，助熔剂氧化钾(K2O)、氧化铅(PbO)、氧化钡(BaO)和氧化钙(CaO)，次要成分氧化锶(SrO)，以及与颜色相关的成分氧化铜(CuO)，后者主要为玻璃的次要成分氧化镁(MgO)和氧化铝(Al2O3)，以及一些杂质污染物诸如氧化锡(SnO2)、二氧化硫(SO2)和五氧化二磷(P2O5)等。

其中有三种助熔剂的含量在不同玻璃类型中的含量有显著差异，分别为氧化钾(K2O)、氧化铅(PbO)、氧化钡(BaO)，刚好对应两种玻璃类型的重要成分，而结合第一问的关联性分析得到，类别与纹饰之间存在相关关系；玻璃中的两种颜色成分，氧化铜(CuO)随风化有流失，氧化铁(Fe2O3)随风化流失不明显。

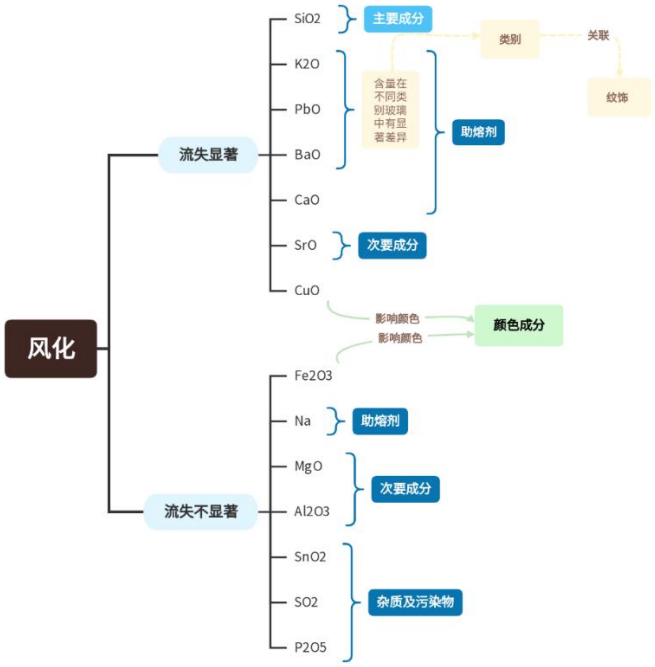


图 5.4.1 化学成分变化与风化的关系图

结合可以看出二氧化硅在铅钡玻璃中随风化流失显著，氧化铅、氧化钡和氧化钙在高钾玻璃中随风化流失显著。表中化学成分的数据为百分比含量。

表 5.4.1 不同玻璃类别和风化情况下部分化学物质含量比例

		二氧化硅	氧化铅	氧化钙	氧化钡
铅钡	严重风化	8.48	40.27666667	2.066666667	22.02333333
	无风化	54.37578947	22.47631579	1.171052632	9.601052632
	风化	31.34518519	40.23074074	2.666666667	9.834814815
高钾	无风化	67.98416667	0.411666667	5.3325	0.598333333
	风化	93.96333333	0	0.87	0

5.4.2 定性分析：基于方差分析的类别差异性判断

I.模型的建立和求解

根据上一小节的描述性统计分析和可视化等方法，不难推断得出，玻璃文物采样点的若干化学成分在不同类别的玻璃文物之下存在着较为显著的差异，比如高钾类采样点的二氧化硅和氧化钾的含量相比于铅钡类较高，铅钡类采样点的氧化铅和氧化钡含量相比于高钾类较高。从直观上理解，类别的不同反映产地、工艺的不同，高钾类来源于岭南和海外，制作工序中会加入更多的草木灰等钾含量较高的元素；铅钡类更倾向于本土工艺，制作过程中会使用铅、钡等本土玻璃制品常见元素。

在量化化学成分与所属类别的关系之前，首先对这一潜在关系进行检验；在定量分析之前首先进行定性分析，更易精准的把握问题所在。

Step1：确定观测变量和状态变量

本题观测变量是玻璃文物采样点的化学成分，是连续型的变量，统一记为 X ；本题控制变量是玻璃文物采样点的类型，编码为 0-1 变量。确定本题方法是单因素方差分析，内容是控制变量的不同水平是否会对观测变量产生影响，以及影响是否显著。

Step2: 建立数学模型

控制变量有两个水平，即为 j = 0,1；每个水平有 30 个左右的样本（已均衡后的数据），记为每个控制变量水平的 n 个观测。在两个水平下第 i 个样本的观测 X_{ij} ，可以记作：

$$X_{ij} = \mu_i + e_{ij}, i = 1,2,...,n, j = 0,1 \tag{8}$$

其中 μ_i 为第 i 个样本的期望值， e 为抽样误差。

在统计分析中，为便于估计，模型通常采用下面的式子：

$$X_{ij} = \mu + a_j + e_{ij}, i = 1,2,...,n, j = 0,1 \tag{9}$$

其中 a 表示控制变量水平对观测变量水平的影响，是下一步需要检验的目标，记作：

$$a_j = \mu_j - \bar{\mu} \tag{10}$$

Step3: 构造检验统计量并进行假设检验

方差分析的检验对象在于 a ，即控制变量的某一或某些取值是否对观测变量水平产生显著影响，所以原假设和备择假设为：

$$H_0 : a_1 = a_2 = 0 \tag{11}$$

$$H_1 : a_1, a_2 \text{中至少一个不为} 0 \tag{12}$$

随后构造检验统计量：需要用到观测变量的总离差平方和 SST、组间离差平方和 SSA、组内离差平方和 SSE，定义分别为：

$$SST = \sum_{j=1}^{2}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}\right)^2 \tag{13}$$

$$SSA = \sum_{j=1}^{2} n_j \left(X_{ij} - \bar{X}\right)^2 \tag{14}$$

$$SSE = \sum_{j=1}^{2}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2 \tag{15}$$

根据数理统计的分布知识，构造出的检验统计量 F 服从以下分布：

$$F=\frac{SSA/(2-1)}{SSE/(n-2)}\sim F(1,n-2) \tag{16}$$

随后根据 F 统计值和显著性 p 值判断；如果结果显著则拒绝原假设，说明控制变量各个取值的不同会对观测变量产生显著影响。

II.模型检验和结果分析

12 个化学成分（筛选后的）分别对玻璃文物的类型做单因素方差分析，组内离差平方和及其自由度、组间离差平方和及其自由度，以及显著性 p 值如下表所示（篇幅有限，这里仅展示若干较为典型的化学成分）：

表 5.4.2 **典型化学成分单因素方差分析结果**

SiO2	自由度	总离差平方和	组间离差平方和	F	PR(>F)
模型	1	18778.04892	18778.04892	60.28411	7.66E-11
残差	65	20247.01484	311.492536		

Na2O	自由度	总离差平方和	组间离差平方和	F	PR(>F)
模型平方和	1	2.561999695	2.561999695	0.936175	0.33685
残差平方和	65	177.8834122			

K2O	自由度	总离差平方和	组间离差平方和	F	PR(>F)
模型平方和	1	510.6443174	510.6443174	68.78128992	8.79387E-12
残差平方和	65	482.5713602	7.424174772		

PbO	自由度	总离差平方和	组间离差平方和	F	PR(>F)
模型平方和	1	14400.8007	14400.8007	87.24702676	1.24984E-13
残差平方和	65	10728.75581	165.0577817		

BaO	自由度	总离差平方和	组间离差平方和	F	PR(>F)
模型平方和	1	1340.623693	1340.623693	26.06035775	3.10491E-06
残差平方和	65	3343.79677	51.44302722		

将 12 个指标分别对类型做单因素方差分析**部分结果**如上，得到的结论是：玻璃采样点中钠、镁、铝、铜、磷 5 种元素成分的含量与类型关系不显著，硅、钾、钙、铁、铅、钡、锶 7 种元素的成分含量在两类样本中差异显著。

对于该结果的合理化解释如下：

1)硅元素是古玻璃的主要成分；不同类别的产地不同，当地的地质条件和原料也不同，所以硅含量存在差异是可以解释的；

2)高钾类型倾向使用含钾元素的原料，铅钡类型倾向使用含铅钡的原料，所以必然存在差异；

3)铁、铜是玻璃显色的主要成分，根据题目背景，两类玻璃存在颜色上的差异；但是铜元素不显著的原因可能是含量比较低、流失较为剧烈导致的，因为二价铜离子比三价铁离子氧化性更强；

4)铝、镁较为活泼，且氧化物的溶解性也相对较高，许多文献中将铝、镁作为研究玻璃风化和玻璃不同类型的次要因素；

5)五氧化二磷，包括二氧化硫、氧化锡，这些元素含量较低，并且不同采样点中方差比较小，在许多专业研究中被认作是污染物；

6)钙、锶同属第二主族元素，可以一起研究：这两种元素是某种助熔剂的主要成分，所以存在着不同工艺背景下用料的差异；

7)钠元素也是助熔剂，但是表单 2 中数据大多和 0 较为接近，所以在统计学意义上，和类别是关系不大的；如果存在统计规律和科学规律之间的不一致性，可以认为是由小样本的不稳健性导致。

5.4.3 定性分析：基于 Wilcoxon 符号秩检验的风化差异性判断

Ⅰ.模型的建立与求解

在上一小节的基础上，我们将玻璃采样点的 12 种化学成分分为两类：第一类与玻璃采样点所属的类别相关，另一类无关。所以在研究化学成分与“是否风化”之间的关系时，需要控制“所属类别”这一因素不变。具体做法是：

（1）如果研究的化学成分与所述类型有关，那么在研究该化学成分与“是否风化”之间的关系时，需要将样本数据分为高钾样本和铅钡样本，分别对每一类样本进行不同风化类型下化学成分的差异性分析（之所以采用差异性分析，是因为样本数据减少，方差分析拟合效果不佳）。

（2）如果所研究化学成分与所属类型无关，则对全样本进行不同风化情况下化学成分的差异性分析。

下面我们以氧化钙这一化学成分，在高钾和铅钡不同类别中，与风化情况关系的分析为例：在每种类别下，将“未风化”和“风化”形成配对样本（由于经过样本均衡，所以数据可以配对）。

Step1：对高钾类别样本的氧化钙配对数据做正态性检验，结果如表 5.4.3 所示

表 5.4.3 高钾类别样本正态性检验数据

变量名	样本量	平均值	标准差	偏度	峰度	S-W 检验
无风化	6	4.778	3.038	-1.021	-0.822	
风化	6	0.87	0.488	0.504	0.988	
无风化配对风化	6	3.908	2.97	-1.157	-0.365	0.806(0.067*)

由于样本量过小（<<5000），所以采用 S-W 检验，得到的 p 值表现为不显著，所以数据呈现明显的非正态性；所以针对非正态性的数据，差异性分析应采用配对样本的 Wilcoxon 符号秩检验。

Step2：进行差异性检验：

符号秩检验属于一种非参数符号检验，将两个总体进行作差，取其中正值和负值，并求正值、负值的秩（即为表 5.4.4 中的“配对 1”“配对 2”），最终检验统计量依据的是正值的秩和与负值的秩和。

表 5.4.4 总体变量的符号秩检验

配对变量	平均值±标准差			t	df	p	Cohen's d
	配对 1	配对 2	配对差值				
无风化配对风化	4.778±3.038	0.87±0.488	3.908±2.551	3.224	5	0.023**	1.316

Ⅱ.模型检验与结果分析

结合上一小节的结果，12 种成分与类别、是否风化的关系图如表 5.4.5 所示：

表 5.4.5 化学成分与类别、表面风化之间的联系

分布与类型无关的化学成分	分布与类型相关的化学成分	
	铅钡类别	高钾类别
氧化钠 NaO	二氧化硅 SiO2	二氧化硅 SiO3
氧化镁 MgO	氧化钾 K2O	氧化钾 K3O
氧化铝 Al2O3	氧化钙 CaO	氧化钙 CaO
氧化铜 CuO	氧化铁 Fe2O3	氧化铁 Fe2O4
五氧化二磷 P2O5	氧化铅 PbO	氧化铅 PbO
	氧化锶 SrO	氧化锶 SrO
	氧化钡 BaO	氧化钡 BaO

图表解释：左侧的化学成分表示分布与类型无关，右侧的化学成分表示分布与类型相关；左侧的方框染成蓝色，表示这些化学成分与是否风化关系并不显著；右侧标为蓝色的方框，表示该化学成分在该类别下，与“是否风化“关系不显著；反之，如果被标为红色，则表示该化学成分在该类别下，与“是否风化”有着较高的关联性。

结果的合理解释：

（1）钾、铅、钡的氧化物是典型的助熔剂，根据文献，由于玻璃文物埋藏条件差，会导致助熔剂或多或少的流失；所以这类化学成分的流失应当是显著的。

（2）氧化钠虽然是助溶剂但是与类型、风化均不显著，前面已经说明，因为氧化钠的含量都在一个比较低的水平，总方差和离差平方和均不显著，所以我们的结果表现为不显著，与先验事实不符，是一个样本问题；氧化钠的化学性质过于活泼，可以与空气中的水、二氧化碳、硫化氢等成分进行还原反应，极不稳定，所以氧化钠的低水平也是可以预见的。

（3）同理，高钾类玻璃中氧化钙含量较少，铅钡类玻璃中氧化钾含量较少，所以也表现为与流失的关系不显著。氧化钾的化学性质比氧化钠更为活泼。

（4）二氧化硅的化学性质及其稳定性，所以和风化应当是显著相关的。

（5）通过查阅文献，参考了某一北宋玻璃文物的数据[1]，同属于铅钡体系：发现风化后的颗粒，二氧化硅、氧化铝的含量明显减少，钙、镁、锶等第二主族的氧化物也有所减少，所以可以作为解释的依据。

5.5 问题 1 第(3)问：基于回归与聚类的定量规律与数据预测模型

针对问题 1 第(3)问，要求根据上一问中发掘的的文物表面有无风化的统计规律，预测风化（或严重风化）样本点风化前的化学成分分布。

该问题是问题 1 第(2)问的延续：我们在本文中，将会通过两个相互对照的模型：回归与聚类进行规律挖掘和预测。考虑到模型的精度、准度和稳健性，最终结果我们选定基于 K-modes 聚类分析的预测模型。

首先明确**本节的研究对象为风华前后存在显著差异的化学成分比例**：也就是说，如果某化学成分在风化前后的变化并不显著，则其风化前的该化学成分的预测数据与风化后的该化学成分保持一致。所以我们仅需要研究**硅、钾、钙、铁、铅、锶、钡**随着风华前后的统计规律的变化。

其次明确**本节的研究方法是定量分析**而非定性分析，只有通过定量分析，才可以预测风化前相应化学成分的分布。

5.5.1 模型建立：基于回归模型的定量规律和预测模型

在此问中，我们想要研究风化对玻璃化学成分含量的影响，但表面风化并不是定量数据，无法直接进行回归。于是我们将表面风化编码成 0-1 变量，在回归模型中可设置为虚拟变量(Dummy Variable)，考虑到纹饰、类型和颜色与化学成分含量也有关系，且风化对于不同类型玻璃影响不同，故同样将纹饰、颜色设置成虚拟变量，对应编号为(0,1,2)和(0,1,⋯,9)。

将研究的化学成分记作 X，将 X 对所有虚拟变量（颜色、类别、纹饰）的水平做回归，记作：

$$X = F\left(color, type, pattern, d\right) + E\left(X \mid color, type, pattern, d\right) + e \tag{17}$$

其中，X 表示化学成分的含量占比，color 表示与颜色相关的所有虚拟变量，type 表示与类别相关的所有虚拟变量，pattern 表示与纹饰相关的所有虚拟变量，d 表示与“是否风化”相关的虚拟变量（d=1 表示已风化，d=0 表示未风化）；F 是模型设定形式，可以采用线性或者多项式映射，拟合的是在一定的类别、纹饰、颜色、风化条件下化学成分 X 的**条件期望**。

定义：记“是否风化”对于化学成分 X 的影响的**偏效应** β_{X_d} 为:

$$\beta_{X_d} = E\left(X \mid color, type, pattern, d=1\right) - E(X \mid color, type, pattern, d=0) \tag{18}$$

其现实含义在于，在其他变量不变的情况下（颜色、类别、纹饰、是否风化）玻璃文物风化对于化学成分 X 的水平影响或平均影响。

那么，如果预测某一风化样本未风化前的数据，则可以用下面的式子：

$$\begin{aligned}
&\widehat{X} = X(color, type, pattern, d=1) - \beta_{\mathrm{d}} \\
&= X(color, type, pattern, d=1) - [E(X \mid color, type, pattern, d=1) - E(X \mid color, type, pattern, d=0)] \\
&= E(X \mid color, type, pattern, d=0) + [X - E(X \mid color, type, pattern, d=1)] \\
&= E(X \mid color, type, pattern, d=0) + dev(X)
\end{aligned}$$

$dev\left(X\right)$表示风化样本的离差，反映的是样本数据的自身特性。综上，**X 在风化条件下的观测值减去风化的偏效应 β_{X_d}，就可以预测风化前的化学成分观测值**。本题的思想就在于此。

回归分析的步骤：

Step1: 区分玻璃类别，用所有虚拟变量对化学成分逐一进行回归(Stata 软件实现)。

Step2: 若出现**风化情况变量**因多重共线性而被遗漏的情况，则逐步删去其他多余虚拟变量，直到结果中同时出现风化和未风化变量的系数。

Step3: 检验统计显著性以及观察调整 R 方的大小，对于检验显著的化学成分，则计算无风化和风化变量回归系数的差值；对于检验不显著，调整 R 方较低的化学成分，认为风化对其含量改变极小，即上述回归系数差值为 0.

Step4: 给风化数据的每一化学成分加上对应回归系数的差值，得到风化前的预测数据。

结果：预测后的数据效果不佳；本文认为，回归模型的缺陷有三点：

1)回归模型的精度。回归模型的核心在于如何精准的估计“是否风化”这一变量的**偏效应**。

但是由于该偏效应是基于所有样本数据估计的得来的、在其他条件不变的情况下某化学成分在风化前后的水平差异，所以忽视了不同颜色、类型、纹饰分组下这个“偏效应”的差异，导致偏效应估计有偏差和误差。

```
. reg 氧化铁Fe2O3 纹饰_A 纹饰_B 纹饰_C 颜色_浅蓝 颜色_深绿 颜色_深蓝 颜色_蓝绿 表面风化_无风化 表面风化_风化,noc
note: 纹饰_A omitted because of collinearity
```

Source	SS	df	MS	Number of obs	=	67
				F(8, 59)	=	6.91
Model	67.736897	8	8.46711213	Prob > F	=	0.0000
Residual	72.301003	59	1.22544073	R-squared	=	0.4837
				Adj R-squared	=	0.4137
Total	140.0379	67	2.09011791	Root MSE	=	1.107

氧化铁Fe2O3	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
纹饰_A	0	(omitted)				
纹饰_B	-1.228881	.6376709	-1.93	0.059	-2.504858	.0470952
纹饰_C	-.1722093	.3386881	-0.51	0.613	-.8499226	.5055041
颜色_浅蓝	-.2760116	.3564499	-0.77	0.442	-.9892662	.437243
颜色_深绿	-.5821557	.5068395	-1.15	0.255	-1.596339	.4320278
颜色_深蓝	.4199885	.7347653	0.57	0.570	-1.050273	1.89025
颜色_蓝绿	.6787513	.4532045	1.50	0.140	-.2281088	1.585611
表面风化_无风化	1.200011	.3624911	3.31	0.002	.4746685	1.925354
表面风化_风化	.8151301	.3884645	2.10	0.040	.0378143	1.592446

2)预测结果的有效性；如果按照回归分析的方法，预测出来的各个化学成分中，比如预测高钾类型玻璃样本点风化前 6 个指标的数据，有 4 个指标（二氧化硅、氧化钾、氧化铅、氧化钡）的预测数据，全部大于 100 或者小于 0，即无效数据；总体无效数据比例在 74.8%左右。所以回归预测是不准确的，预测数据是无效的。

3)回归模型的稳健性：回归模型将“严重风化”的样本点归结于“风化”，忽视了一个事实，即风化程度的不同会较为显著的反映在化学成分上，所以模型也是不稳健的。

5.5.2 改进模型的建立：基于 K-modes 聚类的定量规律和预测模型

由于回归分析中忽视了颜色、类别、纹饰等不同分组之间“是否风化”对于化学成分**偏效应**的差异性，所以我们试图对不同颜色、类别、纹饰的分组数据一一研究。

鉴于类别过多，每一小类的数据量过小，难以挖掘规律，所以本节首先对颜色、纹饰、类别三个定性变量进行降维，并发现三个变量存在集聚性：

Step1: 确定聚类变量：纹饰、类别、颜色；

首先对这三个定性变量进行数据编码，编码的对应结果参见表 5.5.1；图 5.5.1 展示了风化与无风化类型中纹饰、类别、颜色的分布规律。可见，样本数据有较强的集聚性分布特征，且每一类中风化数据和未风化数据的差异性较为显著：

表 5.5.1 类别变量编码对应表

纹饰编码		类别编码		颜色编码	
A	1	铅钡	0	浅绿	0
B	2	高钾	1	绿	1
C	3			蓝绿	2
				深绿	3
				浅蓝	4
				深蓝	5
				紫	6
				无法识别	7
				黑	8

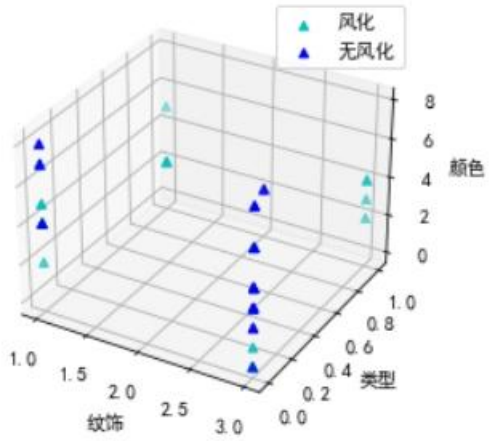


图 5.5.1 基于风化情况的特征指标三维散点图

Step2：K-modes 原理分析与具体做法；

思路分析：首先根据定性变量，聚类出多个类型分组，每一类型分组下，颜色、类别、纹饰有较为集中的取值。

类型的集合记为：

$$Mode = \{Mode1, Mode2, Mode3, ..., Modek\}$$

在第 i 种类型下，颜色、类型、纹饰的分类较为集中，记作：

$$Mode(i) = \{color = color(i), type = type(i), pattern = pattern(i)\}$$

具体方法：K-modes 聚类是 K-均值聚类的一个变种，相似之处在于，均为基于样本中心点的聚类，适用于簇状数据的聚类，使用之前需要指定聚类数目以及聚类中心点的初始值；不同之处在于，K-modes 通过离散距离衡量样本之间的相似性。如，对于 n 维整数数据点 X1，X2，他们之间的离散距离为：

$$d_{ij} = \max_i \left| X_i^1 - X_i^2 \right| = \lim_{n\to\infty} \left(\sum_{i=1}^{n} \left(X_i^1 - X_i^2 \right)^n \right)^{1/n} \tag{19}$$

K-modes 算法的伪代码如下所示：

```
%伪代码来源：https://wenku.baidu.com/view/17711298c47da26925c52cc58bd63186bceb928f.html
Begin <K-modes>
     input k:= 聚类数目
     input 样本集 D = {纹饰，类型，颜色}
     while true
     num = 0
          for i=0 to k
               Ci:= 空集
          end
          for j = 1 to 3
               for i = 0 to k
                    ui = int(ui) %取整
                    Dij = max||xj - ui||2;
                    %离散距离
               end
               min = dj1
               for i = 2 to k
                    if dij<min
                         min :=dij
                         temp:=1
                    end
               end
               lambda = temp
               C(lambda) = C(lambda)+int(xj) %取整
          end
          for i = 1 to k
               ui:=
               基于前一次产生的簇更新均值向量
               if ui ! == ui
                    ui := ui
               else
                    num+=1
               end
          end
          if num == k
               break
          end
     end
     output Ci
end
```

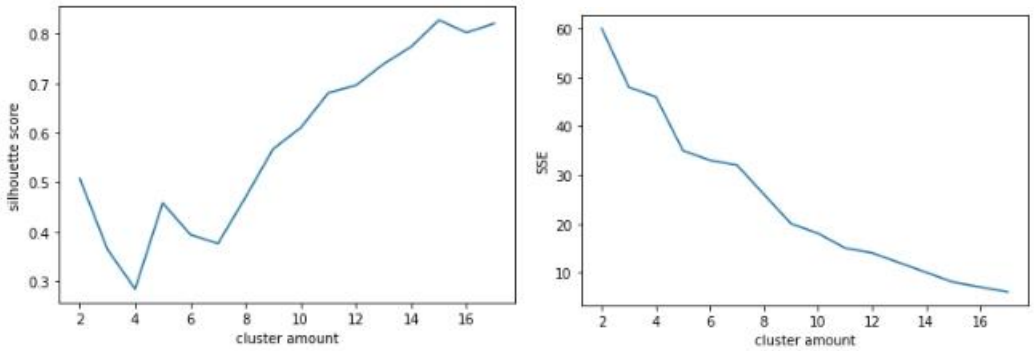
Step3：选择合适的聚类数目：

K-modes 算法最重要的参数为聚类的数目 k，本节聚类数目的选取需要依据两个原则：

（1）类别之间的样本相似度较低

（2）类别之内的样本相似度较高

下面是不同聚类数目下，聚类模型的性能，采用轮廓系数和组内离差平方和测度：



轮廓系数和组内离差平方和均为衡量模型性能的极大型指标，所以最终聚类数目 k=5，因为在 k=5 时各项指标达到了局部最优。

Step4：预测

到目前为止，我们已经将所有样本分为了四类；每一种小类型中，颜色、类别（高钾 or 铅钡）、纹饰具有较高的一致性。所以我们可以对四种小类型中风化样本的未风化数据分别进行预测。

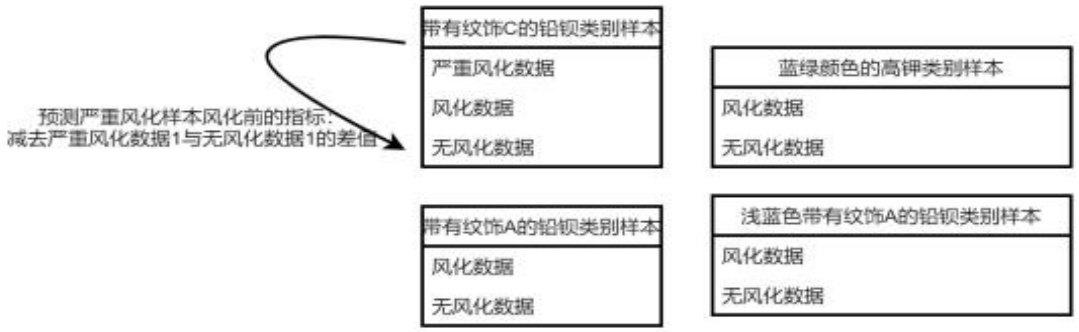
比如，在第 i（i=1,2,3,4）种小类型中，估计偏效应 β_{X_d}：

$$\beta_{X_i} = E\left(X \mid Mode(i), d=1\right) - E\left(X \mid Mode(i), d=0\right) \tag{20}$$

进一步的，我们的预测为：

$$\widehat{X} = X(Mode(i), d=1) - \beta_{\hat{X}}$$
$$= E(X \mid Mode(i), d=0) + [X - E(X \mid Mode(i), d=1)]$$

直观理解就是减去每个小类别中风化样本和未风化样本的**均值差**，可以用下面的概念图表示：



5.5.3 模型求解结果：对风化样本无风化点的预测

最终对风化样本的预测如图 5.5.2 所示：

文物采样点	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)	氧化钡(PbO)	氧化锡(BaO)	二氧化硫(P2O5)	氧化锶(SrO)	氧化锑(Sb2O3)	氧化硫(SO2)
02	47.61625	0	0.812083	2.415417	0.867917	7.767917	1.6775	0	34.2025	2.141333	3.290833	0.31125	0.02875	0
07	66.96067	0.792	6.720667	4.129	1.051333	5.979	1.843	4.293333	2.019	1.904	1.87	0.079	0.236	0.122
08	54.08267	0.730222	2.669556	1.129222	0	1.983778	0.621556	9.925778	1.641333	26.54667	0	0.164889	0	2.693778
严重风化	55.421	0.998	2.630667	3.440333	0	2.056	1.061	2.489667	10.7563	15.46667	0	0	0	5.069333
09	69.35067	0.792	7.310667	3.679	1.051333	5.319	1.993	2.603333	2.019	1.904	1.61	0.079	0.236	0.122
10	71.10067	0.792	7.640667	3.269	1.051333	4.809	1.933	1.893333	2.019	1.904	1.26	0.079	0.236	0.122
11	67.53267	0.730222	2.879556	3.159222	0.365444	3.333778	0.621556	4.445778	0	9.926667	5.591667	0.164889	0	0.113778
12	68.62067	0.792	7.730667	3.779	1.051333	5.459	1.963	2.703333	2.019	1.904	1.41	0.079	0.236	0.122
15	39.93	0	0.3475	3.0325	0.38	2.17	1.505	2.4625	41.29	4.8725	5.005	0.14	0	0
22	66.68067	0.792	7.460667	4.719	1.691333	7.499	2.023	1.603333	2.019	1.904	1.47	0.079	0.236	0.122
26	53.73267	0.730222	2.669556	1.089222	0	1.383778	0.621556	10.08578	2.491333	27.56667	0	0.244889	0	2.073778
严重风化	54.531	0.998	3.030667	3.260333	0	2.126	1.061	2.949667	8.226333	20.29667	0	0.039333	0	5.989333
27	67.05067	0.792	6.720667	3.999	1.591333	6.509	1.873	2.593333	2.019	1.904	1.62	0.079	0.236	0.122
34	69.72267	0.730222	2.919556	0.429222	0	2.263778	1.091556	1.025778	19.51133	5.316667	0	0.014889	0	0.113778
36	73.51267	2.950222	2.809556	0.019222	0	2.243778	0.941556	0.195778	14.57133	6.146667	0	0.014889	0	0.113778
38	66.87267	2.110222	2.669556	0.129222	0	3.213778	0.911556	0.245778	22.27133	5.106667	0	0.204889	0	0.113778
39	60.19267	0.730222	2.669556	0.759222	0	1.143778	0.621556	0.395778	33.99133	2.536667	0	0.404889	0	0.113778
40	50.65267	0.730222	2.669556	1.519222	0	1.093778	0.811556	0	43.17133	2.006667	0	0.474889	0	0.113778
41	52.40267	0.730222	3.109556	4.609222	2.385444	3.973778	2.411556	0	17.08133	5.076667	3.671667	0.264889	0	0.113778
2未风化点	62.61625	3.725833	0	0.865417	0.777917	5.567917	0	1.882917	8.6525	12.61333	0	0.47125	0.02875	0
2未风化点	62.68625	3.665833	0.112083	0.075417	0.847917	7.697917	0	1.932917	6.8925	13.02333	0	0.12125	0.02875	0
43等位1	46.35267	0.730222	2.669556	4.889222	0.545444	2.893778	1.381556	4.865778	32.81133	2.606667	0	0.434889	0	0.113778
43等位2	55.64267	0.730222	2.669556	6.049222	0.605444	4.053778	2.011556	1.025778	17.71133	0	9.041667	0.264889	0	0.113778
48	61.62	0.6	0.6675	2.9225	1.33	12.25	1.205	0	14.18	6.8325	0	0.2	1.1925	0
49	39.08	0	0.3475	4.6825	1.26	3.98	2.915	0	32.65	5.6225	7.275	0.41	0	0
50	28.27	0	0.3475	3.2925	0.26	0.47	0.505	0.0825	42.47	13.7225	2.515	0.61	0	0
51等位1	58.55267	0.730222	2.669556	3.229222	0.845444	5.893778	1.811556	0.885778	13.20133	4.256667	4.311667	0.184889	0.443889	0.113778
51等位2	55.29267	0.730222	2.669556	4.779222	1.105444	3.153778	1.041556	0.265778	24.30133	0	4.961667	0	0	0.113778
52	59.68267	1.950222	2.669556	1.919222	0.205444	1.803778	0.851556	0.215778	20.38133	3.956667	1.921667	0.234889	0	0.113778
54	56.22267	0.730222	2.989556	2.839222	0.935444	4.793778	0.621556	0.345778	28.42133	2.356667	0.451667	0.674889	0	0.113778
严重风化	67.921	0.998	2.630667	0.250333	0.981	4.596	1.061	0.689667	36.7663	0	5.461667	0.539333	0	0
56	63.09267	0.730222	2.669556	0.859222	0	2.493778	0.621556	0.305778	14.21133	10.76667	0	0	0	0.113778
57	59.36267	0.730222	2.669556	0.959222	0	2.823778	0.621556	0.675778	18.06133	12.61667	0	0	0	0.113778
58	64.33267	0.730222	3.009556	3.139222	0.445444	4.163778	1.481556	2.645778	12.31133	2.976667	5.201667	0.014889	0	0.113778

图 5.5.2 风化样本风化前化学成分百分含量预测数据

其中第一列表示文物采样点，对应行为风化之前化学成分的百分含量。

5.6 问题 2 第（1）问的建模与求解

本问的求解思路如流程示意所示图 5.6.1 所示

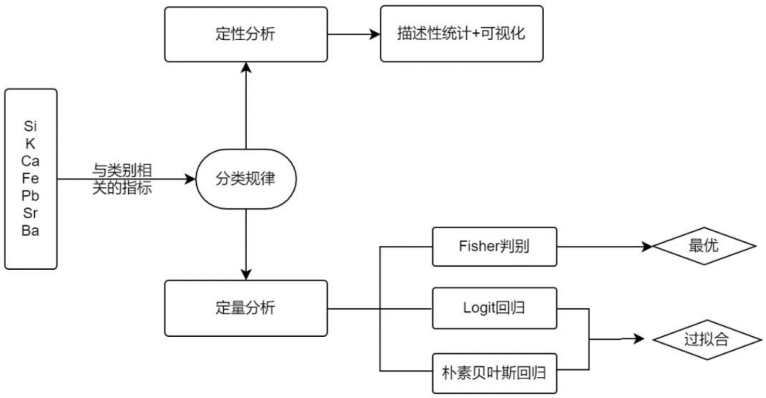


图 5.6.1 问题 2 第(1)问求解思路图

5.6.1 定性分析：描述性统计和数据可视化

为研究不同玻璃类型的分类规律，画出在不同类别中含量百分比差异较大的元素，如下所示

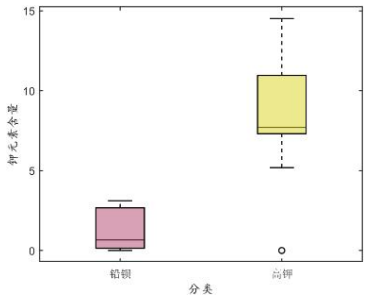


图 5.6.2 钾元素

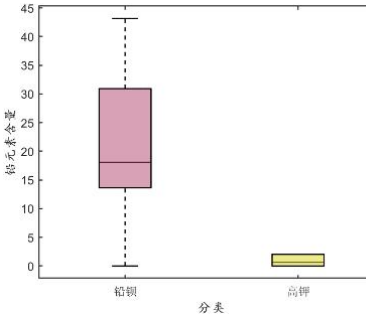


图 5.6.3 铅元素

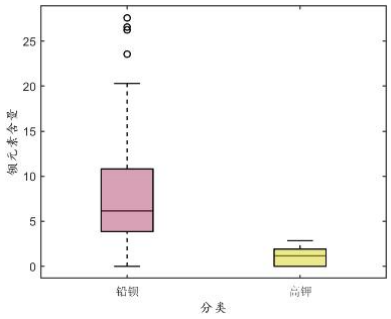


图 5.6.4 钡元素

可以看出铅钡玻璃中钾元素含量几乎为零值，高钾玻璃也表现出铅钡含量极低的特性，可知玻璃的主要成分(除二氧化硅外)是其分类的重要依据。这其实相当容易理解，各类玻璃内不同个体之间主要成分的差异并不会过大。因此要对类别进行亚类划分，则要寻找某一类型玻璃组内差异性较大的元素。这里借助 SPSS 软件绘制了两张小提琴图，反映了铝元素和铜元素在不同类别中的含量水平情况，分别如图 5.6.5、图 5.6.6 所示

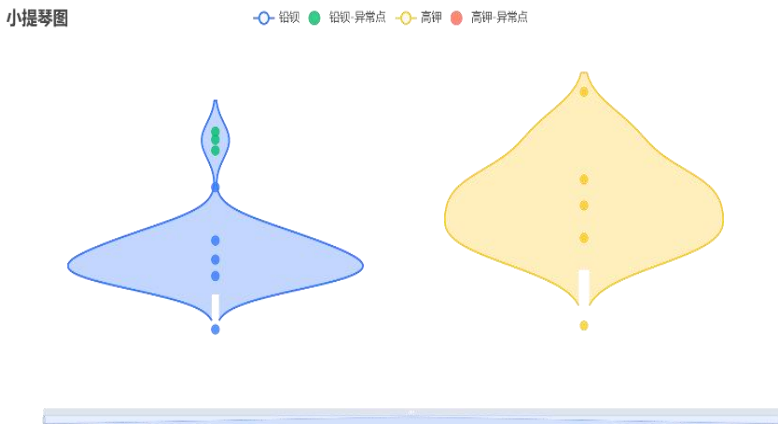


图 5.6.5 铝元素

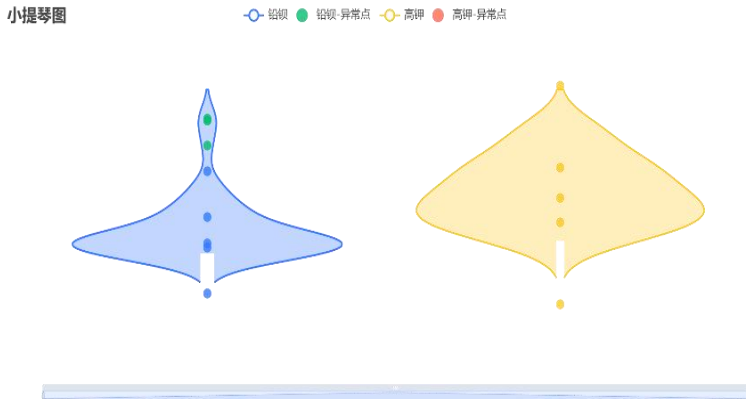


图 5.6.6 铜元素

如图中所示，氧化铝、氧化铜在高钾中的分布更“散”，可能存在较大的组内差异。因此可能考虑作为判别亚类的依据。然而由于数据体量过小，存在异常数据可能性大，不能简单根据描述性统计划定依据，还要依靠聚类的结果来找到其分类。

5.6.2 定量分析：Fisher 判别分析

判别分析是一种类别分辨的方法,其适用于有限确定的分类条件下，根据研究对象的各种特征值判别其类型归属。基本原理为按照一定的判别准则，建立一至多个判别函数，用所研究对象的特征值资料确定判别函数中的待定系数，并计算判别指标。根据计算得出的判别函数则可以判别样本应该归属于何类总体。

常用的判别准则有 1)最大似然法 2)距离判别 3)Fisher 判别 4)Bayes 判别，其中最大似然法判别适用于定性资料的两类或多类判别，其基本原理为用独立事件的概率乘法定理得到判别对象归属某类的概率。

1) 判别指标 $X_i, i=1,2,...,n$，n 个指标之间互相独立；

2) 判别类型 $Y_k, k=1,2,...,m$，m 个类型互斥；

3) 假定样本已知属于第 k 类时，指标 X_i 的取值为 S_l 概率为 $P\left(X_i\left(S_l\right)|Y_k\right)$；

4) 当某个判别样品的各类指标 $X_i, i=1,2,...,n$ 的取值分别对应为 $S_i, i=1,2,...,n$ 时，其似然函数值为：

$$P_k = \prod_{i=1}^{n} P\left(X_i\left(S_{li}\right) | Y_k\right) \tag{21}$$

5) 计算样品属于每一类的概率大小并比较，判定样品为概率最大的类。

Step1:确定样品各判别指标之间的独立性

结果：由第一问对数线性模型得到的独立性检验小表格可知，纹饰、颜色、类型指标之间并非互相独立，因此不能用于最大似然判别。

而若是使用不同化学成分含量作为指标，以类别作为判别类型，则适用于 Fisher 判别分析。

对于 A、B 两类数据判别，不妨假定 A 类有 α 个样本，B 类有 β 个样本，有 $X_i, i=1,2,...,n$ 个观测指标，则目标函数为

$$\max \lambda = \frac{\left|\overline{Z}_A - \overline{Z}_B\right|}{S_A^{\ 2} + S_B^{\ 2}} \tag{22}$$

s.t.

$$Z = C_1X_1 + C_2X_2 + ... + C_nX_n \tag{23}$$

即找到一个线性组合，使得综合指标 Z 在 A 类均数和 B 类均数的差异尽可能大，而两类的类内综合指标变异 $\left(S_A^{\ 2} + S_B^{\ 2}\right)$ 尽可能小，即组间差异尽可能大，组内差异尽可能小。

此综合指标公式便称为 Fisher 判别函数，$C_i, i=1,2,...,n$ 即为判别系数

建立好判别函数后，逐例计算出综合指标 Z_i，求得 A 类和 B 类的均数以及总均数，按照公式判别界值：

$$Z_c = \frac{\overline{Z}_A + \overline{Z}_B}{2} \tag{24}$$

若 A 类均值＞B 类均值，则判别规则如下：

$$\left\{\begin{array}{l} Z_i > Z_c, \text{判为A类} \\ Z_i < Z_c, \text{判为B类} \\ Z_i = Z_c, \text{判为任意一类} \end{array}\right\} \tag{25}$$

Step2:根据问题一第(2)问选出与类别有关的化学元素，对表面风化编码为 0-1 变量，共同作为观测指标 X

Step3:利用 SPSS 软件进行判别分析并进行交叉验证

结果：对原始数据的预测如表 5.6.1 所示

表 5.6.1 **对样本数据分类的** Fisher **判别结果**

		类型	0	1	总计
原始	计数	0	49	0	49

		1	1	17	18
	百分比	0	100	0	100
		1	5.6	94.4	100

其中类型 0 为铅钡玻璃，共有 49 个个体，全部分类正确，类型 1 为高钾玻璃，共有 18 个个体，正确分类 17 个，错误分类 1 个，正确率达到 98.5%。

5.7 问题 2 第（2）问：基于聚类分析的亚类划分模型

5.7.1 模型建立：基于主成分分析的玻璃文物指标体系构建

Ⅰ.铅钡类玻璃文物化学成分的指标体系

根据干福熹《中国古代玻璃的起源和发展》[1]一文，铅钡类古玻璃是我国本土的玻璃类型，最古老的玻璃样本可以追溯到战国及以前，在成分上有着鲜明的本土特色。不同于西方钠钙硅酸盐古玻璃（Na2O-CaO-SiO2）成分较为单一，中国铅钡类古玻璃富含铅、钡、钠、钾多种元素，并且化学成分随着年代的演进而发生变化，可以大致划分为五个历史时期：

（1）春秋战国时期：春秋战国时期处于青铜时代，金属冶炼工艺相对而言不成熟，冶铁技术还未引入，所以在玻璃制作工艺中铁、铜等金属元素引入较少，烧制玻璃中所需的助熔剂以天然的草木灰为主（主要成分是碳酸钾），还会包括石灰石（主要成分是碳酸钙），碳酸盐在一定条件下容易分解成为氧化物；所以该类型玻璃的钾-钠比值较高，钙的含量较高，称该时期的古玻璃属于钾-钙-硅系统。

（2）从战国至东汉时期：在这一时期，冶铁技术日臻成熟，多种金属诸如铅、钡元素的冶炼技术也更加完善，**钡-铅-硅系统**在这一时期出现；

与此同时，由于分离、提纯等制作工艺更加完善，传统的的钾-钙-硅系统的古玻璃烧制技术也有所发展，成分更为纯净的**钾-硅系统**成为该时期中国古玻璃的另一种代表

（3）从东汉至唐代时期：由于制作工艺中分离、提纯的技术更加完善，在原先的钡-铅-硅系统的基础上，**铅-硅系统**成为主流；

（4）从唐代到元代，中西交流最为频繁，本土的铅-硅系统和外来的高钾系统相结合，**钾-铅-硅系统**问世。**钠-铅-硅系统**占比相对较少。

（5）从元至清，技术出现了一定程度上的倒退，主流玻璃是春秋战国时期流行的**钾-钙-硅系统**。

综上，根据历史年代对于铅钡类玻璃文物的划分，大体有钾-钙-硅系统、铅-钡-硅系统，钾-铅-硅系统三大类。

本文对于铅钡类玻璃文物亚类的划分，以主要化学成分的占比（或者是主要化学成分组合的占比）为依据。在对于指标的具体构建方面，本文采用主成分分析，得到相互正交、相关性较小、解释能力较好的主成分代替单一化学成分进行分析。

Step 1:绘制碎石图

碎石图是提取的成分个数与方差解释比例之间的关系图，可用于引入主成分个数的判断；一般而言，如果所引入主成分的边际方差解释能力较小，碎石图在该点变得平缓，那么该主成分不应引入，否则会增加模型复杂度。按照以上原则，最终引入 7 个主成分。

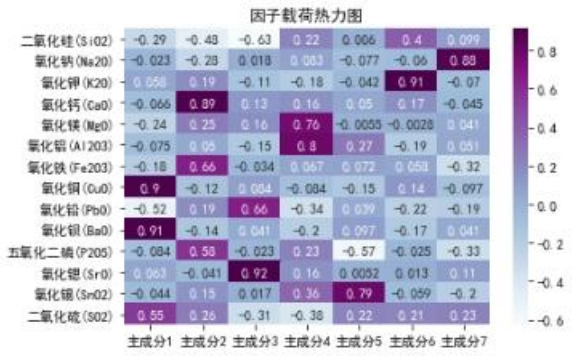


图 5.7.1 因子载荷矩阵

Step2：得到因子载荷矩阵如图 5.7.1 所示

因子载荷矩阵反映提取出来的 7 个主成分与原先 12 个变量之间的相关性，数值恒在 [-1,1] 之间。数值的绝对值越大，反映的主成分与变量之间的相关性就越显著，主成分对于变量信息的保留成分就越多。

Step3：主成分的关联分析。根据因子载荷热力图可以得出：

主成分 1 与氧化钡高度正相关，与氧化铅负相关，与硅相关性较大；所以主成分 1 越高，样本越有可能属于铅-钡-硅体系，并且是其中“钡含量高、铅含量低”的类型；

主成分 3 与氧化铅高度正相关，并且与二氧化硅存在负相关；所以主成分 3,越高，样本越有可能属于铅-钡-硅体系中“钡含量低、铅含量高”的类型；并且主成分 3 越高，样本越有可能不属于“铅-硅”等体系；

主成分 2 与钙相关性较高，所以主成分 2 越高，样本越有可能属于钾-钙-硅体系；

主成分 7 与钠含量高度正相关，可看做是用于解释钠含量的专用指标；

主成分 6 与钾含量高度正相关，可看做是用于解释钾含量的专用指标；并且可以用于反映样本是否属于“钾-硅体系”

主成分 4 与镁、铝相关，可看做是反映次要成分的指标，该指标可用作聚类模型的灵敏度分析；

主成分 5 只和氧化锡有关，数据不是很有效，所以该指标也用作灵敏度分析。

标准化后的第 i 个主成分（i=1,2,…7）记作 $MAIN\_i$

降维得到的主成分表(见附录),根据主成分及其对于不同历史时期（不同系统）的解释能力，本节为进行聚类分析，构建 2 层、3 准则、6 指标的中国铅钡类古玻璃指标体系如图 5.7.2 所示。

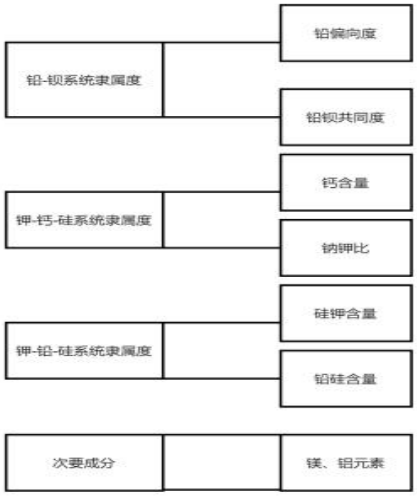


图 5.7.2 中国铅钡类古玻璃指标体系

下面对于各个指标的概念和衡量手段，我们将逐个说明：

准则层 1：铅-钡系统隶属度

极大型指标；衡量样本属于铅钡系统的可能性与置信程度；铅-钡系统隶属度越高，样本就更有可能属于铅钡系统。衡量指标有以下两个：

指标 1：铅偏向度

极大型指标；衡量样本中铅的绝对含量和相对含量；可以用标准化后的主成分衡量；主成分 1 越高，铅钡系统中铅-钡比例越大，铅的绝对含量越多、相对含量越高，进而铅-钡系统隶属度也就越高。计算方法为：

$$Pb = MAIN\_1 \tag{26}$$

指标 2：铅钡共同度

极大型指标；衡量铅钡两种相关性较强的元素在样本中的含量比例之和。可以直接用二者的含量比例之和来判断，用 X 表示标准化的某化学成分占比，计算方法为：

$$Pb\_Ba = X_{P_b} + X_{B_a} \tag{27}$$

最终铅偏向度和铅钡共同度按照各自的变异系数加权，得到铅-钡隶属度。

准则层 2：钾-钙-硅系统隶属度

极大型指标；衡量样本属于钾-钙-硅系统的可能性与置信程度；钾-钙-硅系统隶属度越高，样本属于钾-钙-硅系统的可能性就越大

钾-钙-硅系统的突出特征在于较高的钙含量；这类样本的集中特征在于，钠-钾比通常大于 1；准则层 2 中需要同时引入这两个特征，保证模型的稳健性，正确衡量样本属于钾-钙-硅系统的可能性与置信程度。

指标 1：钙含量

极大型指标；衡量样本中的钙含量。钾-钙-硅系统的突出特征在于较高的钙含量；钙含量越高，样本就越有可能属于钾-钙-硅系统。钙的含量可以采用标准化的主成分 2 来代替；主成分 2 越高，钙含量就越高。

$$Ca = MAIN\_2 \tag{28}$$

指标 2：钠钾比

极大型指标；钾-钙-硅系统的集中特征是钠钾比含量高于 1；钠元素含量可用主成分 7 表示，钾元素含量可用主成分 6 表示；鉴于数据为 0 的情况，可以用标准化后的钠（主成分 7）和钾（主成分 6）的差异代替比值：差异和比值同向变动。

$$Na-K-ratio = MAIN\_7 - MAIN\_6 \tag{29}$$

两个指标以变异系数加权，得到钾-钙-硅系统隶属度的最终指标。

准则层 3：钾-铅-硅系统隶属度

极大型指标；可以分为钾-硅系统隶属度和铅-硅系统隶属度两部分；分别用主成分 6 和主成分 3 来指代，并做根据其变异系数加权。主成分 6 和主成分 3 均为极大型指标：任意一个变量越高，样本越有可能属于钾-铅-硅系统。

准则层 4：次要成分；

极大型指标；衡量样本中镁、钙的比例之和；该指标可以用标准化后的主成分 4 来指代，主成分 4 越高，次要成分越多。

II.高钾类玻璃文物化学成分的指标体系构建

本节引用斯琴毕力格等《激光剥蚀-电感耦合等离子体-原子发射光谱 /质谱法分析中国古代钾玻璃组分》一文作为构建指标体系的最终依据。

Step1：文献综述

上述文献结合对 14 件古代钾玻璃的 LA-ICP-AES 分析（一种红外光谱分析方法），划分了多个亚类；

（1）首先，MgO 的含量有一定的差异；13 个样本点的 MgO 的含量低于 0.5%，说明这 13 个样本点可能使用了草木灰（主要成分为碳酸钾）作为原料之一。

（2）其次，钾玻璃亚类的划分依据在于氧化铝和氧化钙，进而划分为三个亚类，比如第一个亚类氧化铝含量较多、氧化钙含量较少，记作 m-Al-K 亚类。

（3）随后，颜色之间会有一定的差异；影响显色的成分主要是氧化铁和氧化铜

（4）最后，铷元素和锶元素的比例；在 m-Al-K 亚类中，倾向于低铷、高锶；而在其他亚类中，倾向于高铷、低锶；铷和锶的比值对于分类的可靠性与稳健性有较大贡献；

Step2：选择指标

根据从文献综述得出的四个先验结论，本节构造五个准则层，并选择合适的指标如图 5.7.3 所示：

图 5.7.3 高钾玻璃分类指标

特征层 1：草木灰

极大型指标；草木灰影响的是高价类别样本的总体分类，即在制作工艺中是否用到了草木灰。通过以下两个指标衡量：

其一是氧化镁，是一个极小型指标。氧化镁越少，可能用到的草木灰就越多；

其二是氧化钾；是一个极大型指标；因为草木灰的主要成分是钾，所以氧化钾越多，可能用到的草木灰就越多。

草木灰这一特征，可以根据氧化镁和氧化钾的变异系数线性加权和来求得，记作：

$$KC = \omega_{21} * (-X_{Ca}) + \omega_{22} * (X_K) \tag{30}$$

其中，ω_{21}，ω_{22} 分别代表两个指标的变异系数权重。

特征 2、3：关键指标-铝和关键指标-钙

均为极大型指标；用标准化的铝、钙含量比例衡量。

$$AL = X_{Al} \tag{31}$$

$$CA = X_{Ca} \tag{32}$$

特征 4：显色剂

显色剂主要和氧化铁、氧化铜的含量相关；所以取二者标准化后的含量比例进行变异系数加权即可，记作 $pattern\_color$.

特征 5：第一主族和第二主族比值

极大型指标；文献表明铷和锶的比值对于分类的可靠性与稳健性有较大贡献；本文采用所有第一主族元素与第二主族元素的含量比值来近似衡量这一特征；一共有八个指

标：钠镁比值、钠钙比值、钠钡比值、钠锶比值、钾镁比值、钾钙比值、钾钡比值、钾锶比值。

比如，钠镁比值的衡量方法是标准化的氧化纳含量比例和标准化的氧化镁含量比例的差值，均为极大型指标。

最终这一特征记作 *ratio*，采用八种指标的平均值。

III模型建立：K-means 聚类

模型求解和结果分析：基于聚类分析的亚类划分

I. 铅钡类玻璃文物的亚类划分结果，如表 5.7.1

表 5.7.1 铅钡类玻璃文物亚类划分结果

文物采样点	铅-钡系统隶属度	钾-钙-硅系统隶属度	钾-铅-硅系统隶属度	次要成分
08 严重风化点	-0.823142394	-0.560901486	1.326646872	0.20399
26 严重风化点	-0.889522136	-0.530768253	1.542296287	0.549374
54 严重风化点	0.751235984	0.21545577	0.214090284	-1.06854
08	-1.258954296	1.056901921	0.413059423	1.296818
11	-1.155775817	0.624780924	0.649042792	0.047491
26	-1.149226007	1.212215191	0.291849137	1.245166
34	-0.502348143	0.014484844	0.41850227	-1.33767
36	-0.831801902	-0.769579559	0.636758335	-1.30422
38	-0.333809734	-0.487574518	0.441514353	-1.17232
39	0.273230486	0.005921536	0.221680256	-1.61355
40	0.875898802	-0.001532056	0.14982442	-1.67263
41	0.406140831	-0.243997676	1.214249436	0.372718
43 部位 1	0.513957244	0.187386622	0.719621884	-0.24144
43 部位 2	0.153054949	0.27971869	0.826302925	-0.52225
51 部位 1	-0.116873571	-0.019693938	0.899158659	0.767615
51 部位 2	0.278881966	-0.139744299	0.717750232	-0.77927
52	-0.285654635	-0.37071373	0.530776154	-1.12416
54	0.37577414	-0.300588887	0.827350177	-0.35458
56	-0.584456022	0.017213254	0.243773765	-0.88325
57	-0.305467382	0.11490571	0.136400038	-0.75692
58	-0.545193419	0.312812775	0.703847268	-0.56709
24	0.603950522	1.860978937	-1.56277433	1.287749
25 未风化点	0.350839286	-0.181353108	-0.779260262	-0.77986
31	0.033233659	0.331749068	-0.167771766	-0.19365
32	-0.401672559	0.185929437	-0.627240844	-1.04144
33	-0.586412293	-0.038319473	-0.403754638	-0.63534
35	-0.329558311	0.251071307	-0.86700645	-1.1211
37	-0.654138444	0.201168268	-0.206059711	0.103241
55	0.302666239	-0.406979948	-0.875527643	-0.9047
23 未风化点	-0.589783492	-1.843940483	-0.084765728	0.242285
19	1.208220528	0.912362165	-0.968101744	-0.68309
48	0.265774144	-0.727528159	0.7822209	3.536848
49	1.383155603	0.651680431	-0.45534366	0.157515

根据上一小节建立好的铅钡类玻璃文物指标体系，构建 K-means 聚类模型，步骤如

下：

Step1：确定聚类所需指标；如上表所示，包括样本对三个系统的隶属程度和一个次要成分综合指标

Step2：根据聚类模型的性能，选定聚类数目 K

下面是引入四个变量时，不同的聚类数目选定下，模型的性能判断；三个性能指标都是正向性指标（极大型指标），可见聚类数目 K=3 时各项指标达到了一个局部最优；考虑到模型的复杂程度，最终选择 K=3 的聚类数目作为最优模型。

Step3: 聚类结果展示与分析

聚类结果如下：

（1）第一类如图 5.7.4 所示，有较高含量的铅钡、钾还有钠；还有镁铝含量较低，可以判断是古代的工艺（战国至东汉），由于技术不先进导致的；不能判断是铅钡硅系统，还是钾钙硅（或者是钠钙硅系统）；所以第一类判定为**从战国到东汉时期(400B.C .～200A.D.)Ba -Pb-Si 系统和 K -Si 系统**；

文物采样点	铅-钡系统隶属度	钾-钙-硅系统隶属度	钾-铅-硅系统隶属度	次要成分	cluster
24	0.603951	1.860979	-1.562774	1.287749	0
19	1.208221	0.912362	-0.968102	-0.683091	0
49	1.383156	0.651680	-0.455344	0.157515	0
50	1.543239	0.726376	-1.055662	-0.371862	0
30部位1	1.627067	0.088582	0.129900	0.900869	0
30部位2	1.587859	0.328133	-0.394756	0.958445	0
49未风化点	0.367647	0.345544	-0.517249	0.289666	0
50未风化点	0.711870	0.658448	-1.011891	-0.229901	0
02	0.949560	0.236721	-0.415999	0.064784	0
20	-0.216318	1.739722	-1.428702	1.012469	0
46	0.357988	0.247861	-0.723747	0.500000	0

图 5.7.4 铅钡玻璃第一类划分结果

（2）第二类结果如图 5.7.5 所示，铅-钡系统隶属度较低，钾-钙-硅系统隶属度不显著，钾-铅-硅系统隶属度较高，可以初步判断是东汉-唐代或者唐代-元代的样本；但同时次要成分镁铝含量较高，表示技术条件较为先进；再加上唐代到元代中西方交流密切，本土玻璃吸收外来的高钾玻璃的成分较多，钾含量偏高也较为正常；所以确定为**从唐代到元代时期(600～1200A.D.)的 K -Pb-Si 系统**；

文物采样点	铅-钡系统隶属度	钾-钙-硅系统隶属度	钾-铅-硅系统隶属度	次要成分	cluster
08严重风化点	-0.823142	-0.560901	1.326647	0.203990	1
26严重风化点	-0.889522	-0.530768	1.542296	0.549374	1
08	-1.258954	1.056902	0.413059	1.296818	1
11	-1.155776	0.624781	0.649043	0.047491	1
26	-1.149226	1.212215	0.291849	1.245166	1
41	0.406141	-0.243998	1.214249	0.372718	1
51部位1	-0.116874	-0.019694	0.899159	0.767615	1
37	-0.654138	0.201168	-0.206060	0.103241	1
23未风化点	-0.589783	-1.843940	-0.084766	0.242285	1
48	0.265774	-0.727528	0.782221	3.536848	1
42未风化点1	-0.732297	-0.847272	-0.187897	0.932151	1
42未风化点2	-0.862604	-0.845850	-0.334776	1.059196	1
28未风化点	-0.259413	-0.031112	-0.248993	0.262161	1
29未风化点	-0.280557	-0.710923	0.156471	1.654298	1
44未风化点	-0.502879	-1.083217	-0.077478	0.919163	1
45	-0.354196	-0.660391	-0.396349	0.292577	1
47	0.104298	-1.099675	-0.192792	0.140862	1
53未风化点	-0.473487	-0.906468	-0.209189	0.561395	1

图 5.7.5 铅钡玻璃第二类划分结果

（3）第三类结果如图 5.7.6 所示，铅-钡系统隶属度不显著，钾-钙-硅系统隶属度较低，钾-铅-硅系统隶属度较高，但不同的是镁、铝等次要成分含量较低，所以技术条件稍显落后、年代略微久远。本文的判断是**从东汉到唐代时期(200～700A.D.)：Pb-Si 系统；**；

文物采样点	铅-钡系统隶属度	钾-钙-硅系统隶属度	钾-铅-硅系统隶属度	次要成分	cluster
54严重风化点	0.751236	0.215456	0.214090	-1.068544	2
34	-0.502348	0.014485	0.418502	-1.337673	2
36	-0.831802	-0.769580	0.636758	-1.304221	2
38	-0.333810	-0.487575	0.441514	-1.172323	2
39	0.273230	0.005922	0.221680	-1.613546	2
40	0.875899	-0.001532	0.149824	-1.672632	2
43部位1	0.513957	0.187387	0.719622	-0.241441	2
43部位2	0.153055	0.279719	0.826303	-0.522254	2
51部位2	0.278882	-0.139744	0.717750	-0.779273	2
52	-0.285655	-0.370714	0.530776	-1.124158	2
54	0.375774	-0.300589	0.827350	-0.354582	2
56	-0.584456	0.017213	0.243774	-0.883247	2
57	-0.305467	0.114906	0.136400	-0.756918	2
58	-0.545193	0.312813	0.703847	-0.567085	2
25未风化点	0.350839	-0.181353	-0.779260	-0.779856	2
31	0.033234	0.331749	-0.167772	-0.193655	2
32	-0.401673	0.185929	-0.627241	-1.041444	2
33	-0.586412	-0.038319	-0.403755	-0.635344	2
35	-0.329558	0.251071	-0.867006	-1.121100	2
55	0.302666	-0.406980	-0.875528	-0.904696	2

图 5.7.6 铅钡玻璃第三类划分结果

Step4: 聚类模型的稳健性分析和灵敏度分析

（1）稳健性分析

在这一小节中，本文需要研究的对象是引入聚类模型的变量的稳健性；我们可以

尝试将次要成分剔除构建一个对照模型，和对照模型和原模型的聚类效果做对比分析；对照模型的部分性能如图 5.7.7，图 5.7.8 所示：

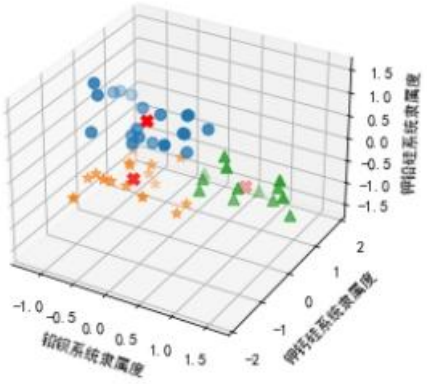


图 5.7.7

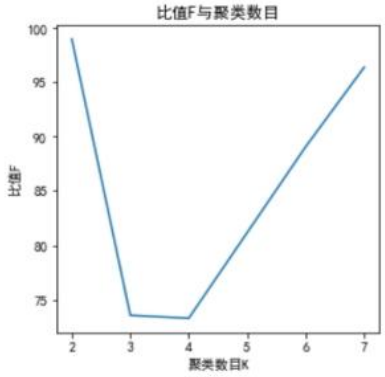


图 5.7.8

左图是对照模型（只引入了三个隶属度，未引入次要成分）的聚类效果，可见黄、蓝两个亚类别之间的差异性不显著；右图的比值 F 是聚类模型的量化指标，而且是正向指标（指标越大，聚类模型性能越优异），可见，模型性能相对于原模型有了很大程度上的下降。

（2）敏感性分析

由于聚类分析是一种无监督学习算法，所以对于较为敏感；衡量一个聚类模型好坏与否的唯一标准在于，样本数据的微小变化是否会显著影响聚类结果。

对数据依施加 1%~10%的噪音，通过观察加噪音前后数据聚类性能变化是否显著，判断聚类结果的敏感性。如图 5.7.9 灵敏度分析所示：

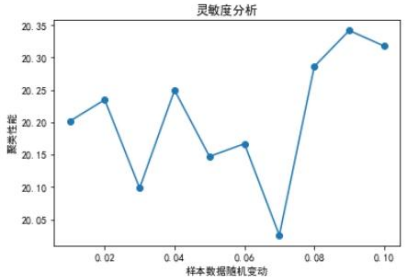


图 5.7.9 灵敏度分析

结果：这里选取的聚类模型性能指标是 CH 指数，可见对模型施加 10%以内的扰动，模型性能的变尽在 0.20 上下，所以完全有理由认为模型是稳健的。

II.高钾类玻璃文物的亚类划分

根据上一节做出的高钾类文物的指标体系（如下表所示），开展 K-means 聚类；

文物采样点	KC	AL	CA	pattern_color	ratio
13	-0.402154	-0.109456	1.479341	1.129878	0.800357
16	0.350262	-0.099925	1.313484	-1.066493	0.905399
07	-0.362727	-0.195719	-0.283757	0.590660	-0.247654
09	-0.273813	-0.510267	-0.457328	0.055866	-0.088649
10	-0.224082	-0.753327	-0.615470	-0.214671	0.025191
12	-0.210518	-0.443545	-0.418756	0.079636	-0.041587
22	0.305739	0.528695	-0.056186	-0.282485	-0.252714
27	0.107197	0.056873	-0.333899	0.007635	-0.246506
01	-0.027835	-1.172248	0.561342	0.404293	-0.121474
03部位1	-1.508302	-1.110292	-1.101084	-1.324431	-0.129971
03部位2	0.539689	-0.424005	0.387771	0.987299	-0.032131
04	0.524399	0.023988	0.869913	-0.067126	-0.319839
05	0.900045	0.529172	0.958627	0.522631	-0.181550
06部位1	0.543282	2.268719	-1.876369	0.171242	0.273123
06部位2	0.372442	1.744472	0.210343	1.416596	-0.479385
18	0.460616	-1.591646	-1.876369	-1.597068	0.753020
21	-1.228762	-0.095159	-0.059657	0.432928	-1.589938
14	0.134523	1.353669	1.298055	-1.246390	0.974309

Step1：明确聚类指标：

聚类指标有五个，分别是草木灰（KC）、关键指标-AL（AL）、关键指标-Ca（CA），显色剂（pattern_color）、第一主族元素和第二主族元素的比例（ratio）；

Step2: 根据模型性能指标选定聚类数目；

根据不同聚类数目下聚类模型的三个评价指标（F 值、轮廓系数、CH 指数），权衡模型复杂性、模型性能，最终选择聚类数目为 6 的聚类模型，因为在聚类数目=6 的时候模型性能达到一个局部最优

Step3：聚类的结果分析与解释

本节将所有样本聚成 6 类，所有结果如下表：labels 是定性变量，表示样本的所述类型，取值为 0,1,2,3,4,5,；labels 相同的两组样本所属类型相同：

文物采样点	KC	AL	CA	pattern_color	ratio	labels
13	-0.402154	-0.109456	1.479341	1.129878	0.800357	2
16	0.350262	-0.099925	1.313484	-1.066493	0.905399	1
07	-0.362727	-0.195719	-0.283757	0.590660	-0.247654	0
09	-0.273813	-0.510267	-0.457328	0.055866	-0.088649	0
10	-0.224082	-0.753327	-0.615470	-0.214671	0.025191	0
12	-0.210518	-0.443545	-0.418756	0.079636	-0.041587	0
22	0.305739	0.528695	-0.056186	-0.282485	-0.252714	0
27	0.107197	0.056873	-0.333899	0.007635	-0.246506	0
01	-0.027835	-1.172248	0.561342	0.404293	-0.121474	0
03部位1	-1.508302	-1.110292	-1.101084	-1.324431	-0.129971	4
03部位2	0.539689	-0.424005	0.387771	0.987299	-0.032131	2
04	0.524399	0.023988	0.869913	-0.067126	-0.319839	2
05	0.900045	0.529172	0.958627	0.522631	-0.181550	2
06部位1	0.543282	2.268719	-1.876369	0.171242	0.273123	3
06部位2	0.372442	1.744472	0.210343	1.416596	-0.479385	5
18	0.460616	-1.591646	-1.876369	-1.597068	0.753020	4
21	-1.228762	-0.095159	-0.059657	0.432928	-1.589938	0
14	0.134523	1.353669	1.298055	-1.246390	0.974309	1

Step4:聚类结果的合理性分析

聚类结果的合理性分析在于，划分出来的类别之间的差异性是否显著，划分的各个类别内部的一致性是否明显。聚类的最终结果如图 5.7.10 所示：

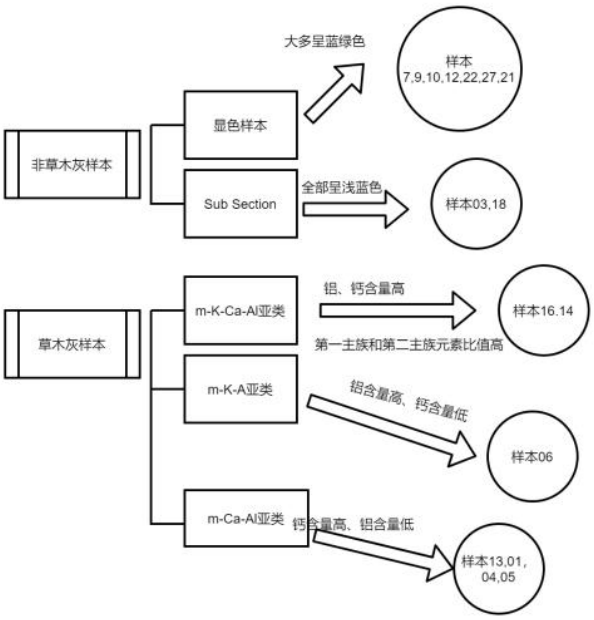


图 5.7.10 高钾玻璃亚类划分结果

划分出来的类别基本与参考文献中的分类相符，找到了其中的 m-K-Ca-Al 亚类，第一主族和第二主族的元素比值偏高，一致性良好，可以断定模型合理、有效。

Step5:聚类模型的敏感性分析；

聚类模型本质上是基于样本的无监督学习模型，并未基于任何先验的理论。样本数据的小范围波动是否显著影响聚类结果，是检验聚类模型稳健性与敏感性的关键所在。

小范围改变数据分布情况，具体做法是给待聚类数据加入 1%到 8%的随机噪音，观察聚类的性能是否发生显著变化：

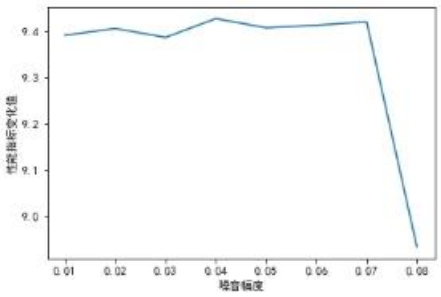


图 5.7.11 模型性能指标受扰动冲击的变化情况

图 5.7.11 是模型性能指标（用轮廓系数衡量）随着不同体量的扰动冲击的变化情况；可见，模型在 7%的范围内是较为稳健的，但是如果扰动项超过 7%，聚类性能会有 5%的波动，因此可以认为模型在一定情境下是稳健的。

部分聚类结果：

文物采样点	铅-钡系统隶属度	钾-钙-硅系统隶属度	钾-铅-硅系统隶属度	次要成分	cluster
24	0.603950522	1.860978937	-1.56277433	1.287749199	0
19	1.208220528	0.912362165	-0.968101744	-0.683091067	0

49	1.383155603	0.651680431	-0.45534366	0.157514501	0
50	1.54323854	0.726375746	-1.055661769	-0.371861987	0
30 部位 1	1.627066861	0.088581744	0.129900182	0.90086891	0
30 部位 2	1.587858921	0.328133067	-0.394756436	0.958444903	0
49 未风化点	0.367646701	0.345543669	-0.517248927	0.289666123	0
50 未风化点	0.711870312	0.658447738	-1.011891407	-0.22990131	0
02	0.949559916	0.236721137	-0.41599907	0.064784126	0
20	-0.2163178	1.739722172	-1.428701543	1.01246852	0
46	0.35798803	0.247861233	-0.723747097	0.500000448	0

5.8 问题 3 的求解与结果分析

Step1:根据问题 1 第(2)问对不同玻璃类型化学成分的分析结果选取了与玻璃类型有显著关联的化学物质作为指标，分别采用问题 2 第(1)问中做出的判别分析以及 Logit 回归和朴素贝叶斯回归对表单 3 中未知类型玻璃文物进行了归类判别。这里 Logit 回归与朴素贝叶斯回归具体方法不列出，仅列出结果

结果：得到的分类预测结果如表 5.8.1 所示

表 5.8.1 表单 3 文物分类的预测结果

文物编号	朴素贝叶斯	逻辑回归	Fisher 判别	文物编号	朴素贝叶斯	逻辑回归	Fisher 判别
A1	高钾	高钾	铅钡	A5	铅钡	铅钡	铅钡
A2	铅钡	铅钡	铅钡	A6	高钾	高钾	高钾
A3	铅钡	铅钡	铅钡	A7	高钾	高钾	高钾
A4	铅钡	铅钡	铅钡	A8	铅钡	铅钡	铅钡

结果解读：朴素贝叶斯和逻辑回归预测的结果完全一致，Fisher 判别对文物 A1 的预测结果为铅钡，另两者为高钾，其他文物相同。

Step2:合理性分析：

为说明预测结果的合理性，对三种方法得到的结果进行一致性检验，利用 SPSSpro 软件进行 Kendall 一致性检验。

结果：检验结果如表 5.8.2 所示：

表 5.8.2 一致性检验分析结果

Kendall's W 分析结果					
名称	秩平均值	中位数	Kendall's W 系数	X^2	P
A1	5.833	1	0.889	18.667	0.009***
A4	3.167	0			
A2	3.167	0			
A3	3.167	0			
A5	3.167	0			
A6	7.167	1			
A7	7.167	1			
A8	3.167	0			

结果解读：Kendall 系数一致性检验的结果显示，总体数据的显著性 P 值为 0.009***，水平上呈现显著性，拒绝原假设，因此数据呈现一致性，同时模型的 Kendall 协调系数W值为 0.889，因此相关性的程度为几乎完全一致性。

通过一致性检验表明判别分析模型有足够强的稳健性。

另外，朴素贝叶斯和逻辑回归的预测结果都是 100%正确，可能存在过拟合；Fisher 判

别分析的预测结果正确率为 98.5%，合理性更强。故采用 Fisher 判别分析的预测结果作为表单 3 文物的分类预测结果。

5.9 问题 4 模型的建立与求解

对于第四问，要求针对不同类型的玻璃文物样本，分析其化学成分之间的关联关系。

关联分析首要在于关联关系的存在性。对此，本节通过求解两种类型的样本所有化学成分两两之间的皮尔逊相关系数；在相关性显著的基础上，找出两两之间具有较强相关性的化学成分。

其次，为了简化变量之间的相关关系，本节利用主成分分析将两种类型玻璃样本的化学成分进行分划，以便后续分组变量之间的研究。

然后基于降维后的数据，开展定量的关联分析。借助典型相关分析，对两组之间作关联性分析，探索并量化不同组别之间的关联关系。

最后针对每组数据的关联性分析结果，做描述性的差异性分析。

5.9.1 相关性分析

Step1:对铅钡类型样本的化学成分做相关性分析。

在数据预处理部分容易得知，大部分化学成分的分布均为正态的；所以，对铅钡类型的化学成分作相关性分析和相关性检验，可以采用皮尔逊相关系数：相关系数表及其显著性检验的部分结果如下所示：

		SiO2	Na2O	K2O	CaO	MgO	Al2O3	Fe2O3
SiO2	皮尔逊相关性	1	.362*	0.087	-.488**	0.088	.401**	0.082
	Sig.（双尾）		0.011	0.553	0.000	0.549	0.004	0.574
	个案数	49	49	49	49	49	49	49
Na2O	皮尔逊相关性	.362*	1	-0.059	-.373**	0.026	0.103	-0.244
	Sig.（双尾）	0.011		0.689	0.008	0.861	0.483	0.091
	个案数	49	49	49	49	49	49	49
K2O	皮尔逊相关性	0.087	-0.059	1	0.096	0.269	.306*	0.248
	Sig.（双尾）	0.553	0.689		0.510	0.061	0.033	0.086
	个案数	49	49	49	49	49	49	49
CaO	皮尔逊相关性	-.488**	-.373**	0.096	1	.417**	0.135	.387**
	Sig.（双尾）	0.000	0.008	0.510		0.003	0.354	0.006
	个案数	49	49	49	49	49	49	49

Step2:绘制铅钡类型样本各项化学指标两两之间的皮尔逊相关系数矩阵热力图，进行可视化分析如下所示：

	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁	氧化铜	氧化铅	氧化钡	氧化二磷	氧化锶	氧化锡	二氧化硫
SiO2	1	0.361626	0.086932	-0.48825	0.087697	0.401109	0.082391	-0.35442	-0.73761	-0.43619	-0.56537	-0.5032	0.085382	-0.38578
Na2O	0.361626	1	-0.05865	-0.3734	0.025719	0.102628	-0.24444	-0.05807	-0.34948	-0.05645	-0.39564	-0.09276	-0.08254	-0.12976
K2O	0.086932	-0.05865	1	0.096395	0.269204	0.305767	0.247599	-0.14639	-0.1308	-0.04106	-0.10555	-0.12115	0.194767	0.010246
CaO	-0.48825	-0.3734	0.096395	1	0.417369	0.135241	0.387489	-0.06817	0.356158	-0.12413	0.535134	0.170992	0.209672	0.103686
MgO	0.087697	0.025719	0.269204	0.417369	1	0.449918	0.293465	-0.30896	-0.03614	-0.45265	0.272303	0.086694	0.245807	-0.26644
Al2O3	0.401109	0.102628	0.305767	0.135241	0.449918	1	0.229807	-0.28163	-0.44126	-0.3394	-0.0743	-0.15526	0.468594	-0.20605
Fe2O3	0.082391	-0.24444	0.247599	0.387489	0.293465	0.229807	1	-0.24698	-0.04761	-0.29492	0.152629	-0.09086	0.225548	-0.17985
CuO	-0.35442	-0.05807	-0.14639	-0.06817	-0.30896	-0.28163	-0.24698	1	-0.08485	0.718749	0.096344	0.195984	-0.18359	0.240639
PbO	-0.73761	-0.34948	-0.1308	0.356158	-0.03614	-0.44126	-0.04761	-0.08485	1	-0.14016	0.331325	0.433521	-0.1264	-0.06515
BaO	-0.43619	-0.05645	-0.04106	-0.12413	-0.45265	-0.3394	-0.29492	0.718749	-0.14016	1	-0.02047	0.16354	-0.07655	0.634378
P2O5	-0.56537	-0.39564	-0.10555	0.535134	0.272303	-0.0743	0.152629	0.096344	0.331325	-0.02047	1	0.279552	-0.08198	0.172677
SrO	-0.5032	-0.09276	-0.12115	0.170992	0.086694	-0.15526	-0.09086	0.195984	0.433521	0.16354	0.279552	1	-0.03789	0.181192
SnO2	0.085382	-0.08254	0.194767	0.209672	0.245807	0.468594	0.225548	-0.18359	-0.1264	-0.07655	-0.08198	-0.03789	1	-0.07101
SO2	-0.38578	-0.12976	0.010246	0.103686	-0.26644	-0.20605	-0.17985	0.240639	-0.06515	0.634378	0.172677	0.181192	-0.07101	1
	SiO2	Na2O	K2O	CaO	MgO	Al2O3	Fe2O3	CuO	PbO	BaO	P2O5	SrO	SnO2	SO2

在上述皮尔逊相关系数的矩阵热力图中，绿色和红色的深度分别代表了两种化学成分之间的正负相关性大小，颜色越深，相关性系数越大。通过观察热力图，选取皮尔逊相关性系数在 60%以上，并且通过显著检验的几组化学成分。

Step3：最终的结果如下：在铅钡类型的玻璃样本中，

（1）二氧化硅的含量比例和氧化铅的含量比例呈显著的负相关；

（2）氧化铜的含量比例和氧化钡的含量比例呈显著的正相关；

（3）氧化钡的含量比例和二氧化硫的含量比例呈显著的负相关。

Step4：铅钡类型样本下相关性检验结果的合理化分析：

（1）二氧化硅和氧化铅呈现显著的负相关；这是由于铅钡类型玻璃样本的主要助熔剂采用的是氧化铅，而二氧化硅是所有玻璃的主要成分，二者呈现一个显著的互斥关系：如果作为主要成分的氧化铅大量流失，势必会引起二氧化硅含量比例相对上升；

（2）氧化铜的比例和氧化钡的比例呈现正相关：从化学上理解，二者的最外层电子数相同，进而最高价离子价数相同，氧化物、硫化物、碳酸盐等性质相近；因为流失意味着金属离子与外界物质反应，所以化学性质相近的元素一同流失也是可以理解的；

（3）最后一对关系，氧化钡容易和二氧化硫形成稳定的沉淀，所以空气中的二氧化硫往往会和氧化钡结合形成稳定的化合物。所以钡含量高时，吸收的二氧化硫会增多，二氧化硫含量减少。

Step5：对高钾类型的化学成分作相关性检验，得出的结论如下，高钾类型中（以下所述化学成分，指代的都是其含量比例）

（1）二氧化硅与氧化钾，氧化钙，氧化铝都呈负相关；

（2）氧化钾和氧化钙呈正相关，氧化镁和氧化锶呈正相关；，

（3）氧化铝和五氧化二磷呈正相关；

（4）氧化铁和五氧化二磷呈正相关；

（5）氧化铅和氧化钡呈正相关；

（6）五氧化二磷和氧化锶呈正相关。

Step6:上述高钾类样本中化学成分关联性的合理化分析：

对于结论 1，可以认为钾、钙、铝是高钾类的主要成分，他们的比例下降均会导致二氧化硅的上升，因为二氧化硅是最稳定的；对于结论 2，钾、钙同为活泼金属，镁、锶同为二价金属，性质相近；由于磷元素分布不显著，所以可以忽略；对于结论 5，铅、钡同为铅钡类的助溶剂，往往同时出现，所以出现正相关也是可以解释的；

5.9.2 因子分析

1、模型建立

主成分分析法是建立在相关系数矩阵或相关的协方差矩阵的基础上的，通过因子分析，得到因子载荷矩阵中主成分与各原始指标的相关系数的绝对值大小，称为因子载荷。再根据因子载荷判断对原始指标的取舍，从而实现对指标的降维。其因子载荷的大小也能反应变量之间的相关性。在同一主成分下，因子载荷越大，证明该因子对主成分的解释性越强。

对同一主成分解释较强的因子，其相关性也更强。综上可将关联性较强的因子聚在一起讨论。

2、模型求解

Step1:对高钾类型的所有化学成分作主成分分析，得到碎石图如图 5.9.1 所示

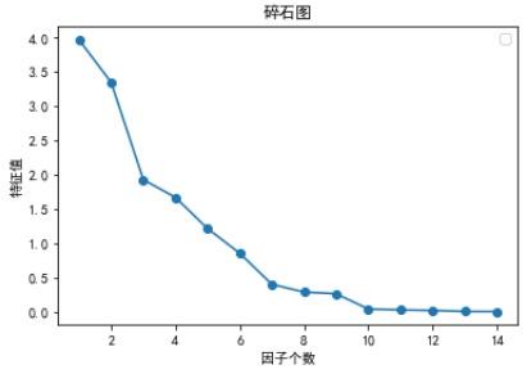


图 5.9.1 碎石图

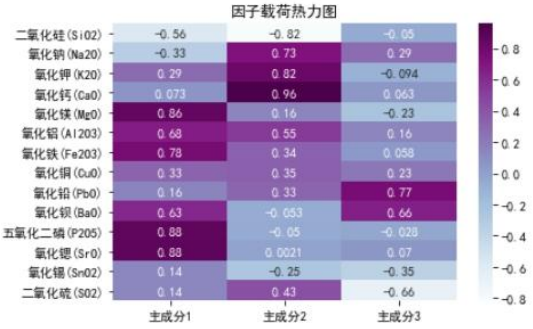


图 5.9.2 因子载荷热力图

根据碎石图可知，折线在主成分为 3 以后趋向平缓，并在之前急剧下降，说明 14 种化学成分指标取 3 个主成分较为合适。得到因子载荷图如图 5.9.2 因子载荷热力图。

Step2：观察因子载荷热力图，主成分 1 中因子载荷较高的为氧化镁，三氧化二铝，三氧化二铁，五氧化二磷，氧化锶；主成分二中因子载荷较高的为二氧化硅，氧化纳，氧化钾，氧化钙，氧化铜；主成分三中因子载荷较高的为氧化铅，氧化钡，二氧化硫。根据因子载荷将化学成分指标划分。划分结果无重复项，无缺失项，说明划分较好。

构建的高钾化学成分指标体系如表 5.9.1 所示

表 5.9.1 高钾化学成分指标体系

A	B	C
MgO	SiO2	PbO
Al2O3	Na2O	BaO
Fe2O3	K2O	SnO2
P2O5	CaO	SO2
SrO	CuO	

Step3：对铅钡类型样本的所有化学成分作主成分分析，最终提取 4 个主成分，构建的铅钡化学成分指标体系划分如表 5.9.2 所示。

表 5.9.2 铅钡化学成分指标体系

A	B	C	D
CuO	SiO2	K2O	Fe2O3
BaO	CaO	MgO	Na2O
SO2	PbO	Al2O3	
	P2O5	SnO2	
	SrO		

5.9.3 典型相关分析

1、模型建立

典型相关分析研究两组[x1,x2…，xp],[y1,y2…yq]定量变量间整体的线性相关关系，是将每一组变量作为一个整体进行研究，并不是分析每一组变量内部的各个交量，所研究的两组变量可以是一组为自变量，另一组为因变量的情况，也可以是两组变量处于同一地位。

本节开展的典型相关分析借助主成分分析的思想和结果。对每组变量分别寻找线性组合。使生成的新变量在能代表原始变量的大部分信息的同时，与另一组变量生成的新变量的相关性最大。除此之外，还要求同一组变量生成的新变量之间线性无关，生成的新变量成为典型变量。

2、模型求解

(1)对高钾类型变量组进行两两之间的典型相关分析（以组 A 和组 B 为例），

设组 A 中的变量构成集合 Y，命名为被解释变量组；组 B 中的变量构成集合 X，称为解释变量组。

Step1：首先提取典型变量对，计算成对典型变量之间的典型相关系数，并对其相关性进行显著性检验。

表 5.9.3 **典型相关系数表**

典型变量对	典型相关系数	特征值	Wilks	模型自由度	误差自由度	F	P
第 1 对	0.994	0.989	0	25	31.221	12.593	0.000***
第 2 对	0.969	0.94	0.012	16	28.133	5.745	0.000***
第 3 对	0.789	0.623	0.197	9	24.488	2.58	0.031**
第 4 对	0.597	0.357	0.523	4	22	2.107	0.114
第 5 对	0.433	0.187	0.813	1	12	2.769	0.122

由表 5.9.3 可知，前 3 对典型变量通过显著性检验，认为前 3 对典型变量之间的相关性显著，第 1 对典型变量的相关系数为 0.994。第 2 对典型变量的相关系数为 0.969。第 3 对典型变量的相关系数为 0.789，相关性都比较强。由此可以取前三对典型变量进行分析。

Step2: 其次分别计算集合 X 和集合 Y 的典型变量的系数，可得到典型变量的组成公式。

	典型变量 v1	典型变量 v2	典型变量 v3		典型变量 u1	典型变量 u2	典型变量 u3
氧化镁	-0.618	0.337	-1.422	二氧化硅	0.145	-0.032	-0.042
氧化铝	-0.116	-0.196	0.456	氧化钠	0.251	-0.156	1.101
氧化铁	-0.211	-0.658	-0.605	氧化钾	0.162	0.118	-0.003
五氧化二磷	-0.014	0.831	0.246	氧化钙	0.102	-0.456	-0.319
氧化锶	0.24	11.296	6.092	氧化铜	0.207	-0.041	-0.066

由此可得出具体的典型相关模型为：

第一组

v1=-0.618y1*-0.116y2*-0.211y3*-0.014y4*+0.24y5*

u1=0.145x1*+0.251x2*+0.162x3*+0.102x4*+0.207x5*

第二组

v2=0.337y1*-0.196y2*-0.658y3*+0.831y4*+11.296y5*

u2=-0.032x1*-0.156x2*+0.118x3*-0.456x4*-0.041x5*

第三组

v3=-1.422y1*+0.456y2*-0.605y3*+0.246y4*+6.092y5*

u3=-0.042x1*+1.101x2*-0.003x3*-0.319x4*-0.066x5*

第一组典型相关方程可知，A 主成分的主要因素是 y1 (典型系数为-0.618)，说明 A 中影响 B 的主要因素是氧化镁；A 主成分的第一典型变量 v1，与 y1 呈高度相关，说明在 A 主成分中, 氧化镁占有主要地位。根据第三组典型相关方程，B 主成分的主要因素是 x2(典型系数为 1.101)，说明 B 中影响 A 的主要因素是氧化钠；B 主成分的第三典型变量 u3，与 x2 呈高度相关，说明在 B 主成分中, 氧化钠占有主要地位。

Step3:根据典型载荷系数进行典型结果分析，载荷系数得到六个个典型变量所有变量的相关性系数。

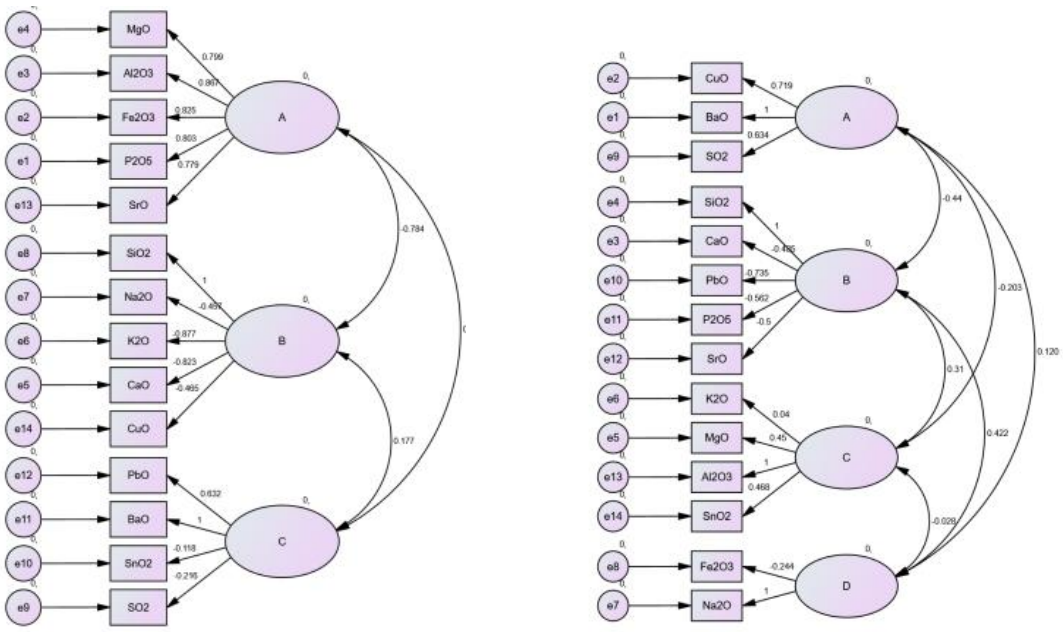
结构分析

表 5.9.4 **典型载荷系数**

	典型变量 v1	典型变量 v2	典型变量 v3	典型变量 u1	典型变量 u2	典型变量 u3
氧化镁	-0.891	0.251	-0.202	-0.886	0.243	-0.159
氧化铝	-0.899	-0.132	0.415	-0.894	-0.128	0.328
氧化铁	-0.847	-0.197	-0.184	-0.842	-0.191	-0.145
五氧化二磷	-0.791	0.391	0.078	-0.787	0.379	0.061
氧化锶	-0.724	0.487	0.055	-0.72	0.472	0.043
二氧化硅	0.792	0.328	-0.185	0.797	0.338	-0.235
氧化钠	-0.007	-0.531	0.646	-0.007	-0.547	0.819
氧化钾	-0.491	-0.219	0.285	-0.494	-0.226	0.362
氧化钙	-0.432	-0.762	0.111	-0.435	-0.786	0.141
氧化铜	-0.313	-0.315	-0.229	-0.314	-0.325	-0.291

典型载荷系数绝对值越大说明该项与典型变量之间的相关关系越强。由表 5.3.1 关联规则挖掘的频繁项集表可得，y1,y2,y3,y4,y5 与 A 主成分的第一典型变量 v1 均呈高度相关，同时又与 B 主成分的第一典型变量 u1 均呈高度相关，说明典型变量 v1 对该集合变量的代表性较好，又能反应主成分 A 与主成分 B 之间的线性相关性。由于第一对典型变量之间的高度相关，导致二氧化硅与 A 主成分的第一典型变量 v1 高度相关；而二氧化硅与 B 主成分的第一典型变量 u1 也高度相关。这种一致性从数量上体现了 A 主成分对 B 主成分的本质作用，说明典型相关分析结果具有较高的可行度。

Step4：将典型相关分析的结果用结构方程的方式可视化；基于典型相关分析的结果，能找出各个主成分（或者因子）之间的整体差异，并按照结构方程的方式可视化如下图所示（左侧是高钾类型下各个化学成分及其组合之间的关系，右侧是铅钡类型下各个化学成分及其组合之间的关系）



以高钾类型下的数据为例说明：由上表可知，A 与 B 之间具有较强的相关性，其相关性系数为-0.784，为负相关，并通过了显著性检验。主成分 A 与主成分 B 之间的相互影响作用较大，其中主成分 B 与因子 A 的相关性最大的为二氧化硅，主成分 A 中所有化学成分

与因子 A 的相关性相似，且都较大。说明，主成分 B 中起主要作用的为二氧化硅，主成分 A 对其中所有因子的代表性较好。此结果与典型相关分析所得出的结果相吻合。

5.9.4 规律之间的差异性分析

两种类别之间的差异性分析请见上面的结构方程图

1)**结构性差异**：根据结构方程模型，两类样本化学成分之间的关联性存在结构性差异；比如，在高钾类型下，氧化铜和二氧化硫隶属于同一因子，是有着较高的关联性的；

还有比如：在主成分分析中，我们根据因子载荷图将高钾类的化学成分划分为 3 组，而将铅钡类的化学成分划分为 4 组。其中，在高钾类中，我们把二氧化硅和氧化钾放在了一组，但并不包含氧化铅；在铅钡类中，我们把二氧化硅和氧化铅放在一组，但并不包含氧化钾。这样的分组差异可以在一定程度上帮助去区分高钾和铅钡的不同类型。

2)**规律性差异**：在相关性分析中，高钾类玻璃化学成分之间两两相关的为 10 对，而铅钡类玻璃化学成分之间两两相关的为 3 对。这说明高钾类中化学成分之间的比率更有规律。

我们可以试图解释该规律的合理性：其中，高钾类中与二氧化硅相关的为氧化钾，氧化钙，氧化铝；铅钡类中与二氧化硅相关的为氧化铅。二氧化硅是玻璃的主要成分，氧化钾，氧化钙，氧化铅为玻璃的助熔剂，不同类型助熔剂与二氧化硅之间相关性的差异也反映了玻璃类型的不同。

6 模型评价及推广

6.1 模型的优点

（1）坚持定性分析与定量分析相结合

例如，定性分析作为定量分析的先导，能够节省物力、精准把握问题。问题 1 中的规律性挖掘是基于方差分析和差异性分析的，问题 4 中典型相关分析是基于相关性检验的。

（2）使用一系列创新算法，效果良好

例如，问题 1 的 Apriori 算法，以数据挖掘方法中的规律关联分析进行定性变量之间的规律性寻找；问题 1 的 K-modes 算法，是基于簇状分布的定性数据的聚类，轮廓系数良好，方便研究不同类型分组下的偏效应。

（3）聚类和分类模型的稳健性较好

例如，聚类性能对聚类样本增加 10%以内的随机扰动表现稳健，便于模型推广；

（4）聚类结果的可解释能力强

例如，对于高钾数据的聚类结果符合 m-Ka-Ca 等亚类的划分。

6.2 模型的缺点

（1）未形成一个较为完整的化学成分指标体系

（2）对于小样本数据，部分规律解释不充分或者没有被充分挖掘

7 参考文献

[1]干福熹.中国古代玻璃的起源和发展[J].自然杂志,2006(04):187-193+184.

[2]董俊卿,李青会,干福熹,胡永庆,程永建,蒋宏杰.一批河南出土东周至宋代玻璃器的无损分析[J].中国材料进展,2012,31(11):9-15.

[3]吴宗道,周福征,史美光.几个古玻璃的显微形貌、成分及其风化的初步研究[J].电子显微学报,1986(04):65-71.

[4]李青会,黄教珍,李飞,干福熹.中国出土的一批战国古玻璃样品化学成分的检测[J].文物保护与考古科学,2006(02):8-13.DOI:10.16334/j.cnki.cn31-1652/k.2006.02.002.

[5]崔天兴,张继华,韩国河.登封南洼遗址出土的宋代玻璃饰品[J].文物保护与考古科

学,2020,32(02):51-56.DOI:10.16334/j.cnki.cn31-1652/k.2020.02.006.

[6]斯琴毕力格,李青会,干福熹.激光剥蚀-电感耦合等离子体-原子发射光谱/质谱法分析中国古代钾玻璃组分[J].分析化学,2013,41(09):1328-1333.

[7]科学网—卡方分布的密度曲线 - 王福昌的博文 (sciencenet.cn)

[8]多元统计分析/何晓群编著. —5 版. —北京：中国人民大学出版社，2019.6

附录

AP_eventual_1.py：Apriori 算法 python 程序

```
#!/usr/bin/env python
# coding: utf-8

# 程序来源：
#
file:///C:/Users//Desktop/%E5%86%B2%E5%88%BA%E9%98%B6%E6%AE%B5/%E6%95%B0%E6%8D%AE%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99%E6%8C%96%E6%8E%98/(89%E6%9D%A1%E6%B6%88%E6%81%AF)%20%E6%95%B0%E6%8D%AE%E5%85%B3%E8%81%94%E5%88%86%E6%9E%90_%E9%A5%AD%E9%A5%AD%E9%A5%AD%E9%A5%AD%E9%A5%AD%E7%82%92%E8%9B%8B%E7%9A%84%E5%8D%9A%E5%AE%A2-CSDN%E5%8D%9A%E5%AE%A2_%E6%95%B0%E6%8D%AE%E5%85%B3%E8%81%94%E5%88%86%E6%9E%90.html

import numpy as np
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
import warnings
warnings.filterwarnings('ignore')

data_raw = pd.read_excel("附件.xlsx",sheet_name = 1)
data_raw.replace(np.nan, 0, inplace = True)
data_raw["总和"] = data_raw.sum(axis = 1)
data = data_raw[(data_raw["总和"]< 105) & (data_raw['总和']> 85)]

data_pattern = pd.read_excel("附件.xlsx", sheet_name = 0)
data_pattern.set_index("文物编号",inplace = True)
data_pattern.dropna(inplace = True)
data_pattern

data_pattern = pd.get_dummies(data_pattern)
data_pattern
frequent_itemsets = apriori(data_pattern,min_support=0.25,use_colnames=True)

#frequent_itemsets = apriori(data_pattern,min_support=0.25,use_colnames=True)
association_rules(frequent_itemsets,metric='lift')

data0 = pd.read_excel("附件.xlsx", sheet_name = 0)
data0.dropna(inplace = True)
data0
```

```
data_new0 = data0[data0["类型"] == "高钾"]
data_new1 = data0[data0["类型"] == "铅钡"].iloc[np.random.randint(1,36,18),:]

data_new = pd.concat([data_new0, data_new1],axis = 0, join = 'inner')
data_new.set_index("文物编号",inplace = True)
data_pattern_new = pd.get_dummies(data_new)
data_pattern_new

frequent_itemsets = apriori(data_pattern_new,min_support=0.25,use_colnames=True)
association_rules(frequent_itemsets,metric='lift')
```

cleaning1.py: 数据清洗 python 程序

```
#!/usr/bin/env python
# coding: utf-8


import numpy as np
import pandas as pd

data0= pd.read_excel("附件.xlsx", sheet_name = 0)
data0.replace(np.nan, "无法识别", inplace = True)
data0.set_index("文物编号", inplace = True)
data0["颜色"].unique()

#在 k-modes 分类中，定类变量和定序变量没有本质区别，对数据标签化处理即可
"""
data0["纹饰"] = data0["纹饰"].map({"C":3, "B":2, "A":1})
data0["类型"] = data0["类型"].map({"铅钡":0, "高钾"：1})
data0["颜色"] = data0["颜色"].map({"浅绿":0, "绿"：1, "蓝绿": 2, "深绿":3,"浅蓝":4, "深蓝
":5, "紫":6, "无法识别":7,"黑":8 })
"""

data1 = pd.read_excel("附件.xlsx", sheet_name = 1)
data1.set_index("文物采样点", inplace = True)
data1.replace(np.nan, 0, inplace = True)
data1["总和"] = data1.sum(axis = 1)
data1 = data1[(data1["总和"]< 105) & (data1['总和']> 85)]
data1


[m,n] = data1.shape
```

```
data1[["纹饰"]] = np.zeros(m)
data1[["类型"]] = np.zeros(m)
data1[["颜色"]] = np.zeros(m)
data1[["表面风化"]] = np.zeros(m)


data1.iloc[:,-4:]
for i in  range(m):
    if data1.index[i][:1] == "0":
        data1.iloc[i,-4] = data0.loc[int(data1.index[i][1:2]),"纹饰"]
        data1.iloc[i,-3] = data0.loc[int(data1.index[i][1:2]),"类型"]
        data1.iloc[i,-2] = data0.loc[int(data1.index[i][1:2]),"颜色"]
        data1.iloc[i,-1] = data0.loc[int(data1.index[i][1:2]),"表面风化"]
    else:
        data1.iloc[i,-4] = data0.loc[int(data1.index[i][:2]),"纹饰"]
        data1.iloc[i,-3] = data0.loc[int(data1.index[i][:2]),"类型"]
        data1.iloc[i,-2] = data0.loc[int(data1.index[i][:2]),"颜色"]
        data1.iloc[i,-1] = data0.loc[int(data1.index[i][:2]),"表面风化"]
#data1.to_excel("洗好的数据.xlsx")
data1.isna().sum()
data1.index[0][-4:]

for i in range(m):
    if data1.index[i][-4:] == "未风化点":
        data1.iloc[i,-1] = "无风化"
data1.to_excel("洗好的数据 1.xlsx")
```

crosstab.py : 列联分析：python 程序

```
#!/usr/bin/env python
# coding: utf-8


import numpy as np
import pandas as pd

data0= pd.read_excel("附件.xlsx", sheet_name = 0)
data0.replace(np.nan, "无法识别", inplace = True)
data0.set_index("文物编号", inplace = True)
data0["颜色"].unique()

#在 k-modes 分类中，定类变量和定序变量没有本质区别，对数据标签化处理即可
"""
```

```
data0["纹饰"] = data0["纹饰"].map({"C":3, "B":2, "A":1})
data0["类型"] = data0["类型"].map({"铅钡":0, "高钾"：1})
data0["颜色"] = data0["颜色"].map({"浅绿":0, "绿"：1, "蓝绿": 2, "深绿":3,"浅蓝":4, "深蓝":5, "紫":6, "无法识别":7,"黑":8 })
"""

data1 = pd.read_excel("附件.xlsx", sheet_name = 1)
data1.set_index("文物采样点", inplace = True)
data1.replace(np.nan, 0, inplace = True)
data1["总和"] = data1.sum(axis = 1)
data1 = data1[(data1["总和"]< 105) & (data1['总和']> 85)]
data1



[m,n] = data1.shape
data1[["纹饰"]] = np.zeros(m)
data1[["类型"]] = np.zeros(m)
data1[["颜色"]] = np.zeros(m)
data1[["表面风化"]] = np.zeros(m)



data1.iloc[:,-4:]
for i in  range(m):
    if data1.index[i][:1] == "0":
        data1.iloc[i,-4] = data0.loc[int(data1.index[i][1:2]),"纹饰"]
        data1.iloc[i,-3] = data0.loc[int(data1.index[i][1:2]),"类型"]
        data1.iloc[i,-2] = data0.loc[int(data1.index[i][1:2]),"颜色"]
        data1.iloc[i,-1] = data0.loc[int(data1.index[i][1:2]),"表面风化"]
    else:
        data1.iloc[i,-4] = data0.loc[int(data1.index[i][:2]),"纹饰"]
        data1.iloc[i,-3] = data0.loc[int(data1.index[i][:2]),"类型"]
        data1.iloc[i,-2] = data0.loc[int(data1.index[i][:2]),"颜色"]
        data1.iloc[i,-1] = data0.loc[int(data1.index[i][:2]),"表面风化"]
#data1.to_excel("洗好的数据.xlsx")
data1.isna().sum()
data1.index[0][-4:]

for i in range(m):
    if data1.index[i][-4:] == "未风化点":
        data1.iloc[i,-1] = "无风化"
data1.to_excel("洗好的数据 1.xlsx")
```

Cluster.py 聚类分析 python 程序

```
#!/usr/bin/env python
# coding: utf-8

# In[2]:


import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
get_ipython().run_line_magic('matplotlib', 'inline')
import seaborn as sns

from sklearn.decomposition import PCA
from scipy.stats import bartlett
from factor_analyzer import FactorAnalyzer
from scipy.stats import zscore

plt.rcParams['font.sans-serif']=['SimHei'] #解决中文显示乱码问题
plt.rcParams['axes.unicode_minus']=False
#本章需导入的模块
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import warnings
warnings.filterwarnings(action = 'ignore')
get_ipython().run_line_magic('matplotlib', 'inline')
plt.rcParams['font.sans-serif']=['SimHei'] #解决中文显示乱码问题
plt.rcParams['axes.unicode_minus']=False
from sklearn.datasets import make_blobs
from sklearn.feature_selection import  f_classif
from sklearn import decomposition
from sklearn.cluster import KMeans,AgglomerativeClustering
from sklearn.metrics import silhouette_score,calinski_harabasz_score
import scipy.cluster.hierarchy as sch
from itertools import cycle
from matplotlib.patches import Ellipse
from sklearn.mixture import GaussianMixture
```

In[3]:

#index_1 #pb_ba_si_1
    #pbba

#index_2 #kca_si
    #na_k

#index_3 #k_si
    #pb_si
#index_4

data = pd.read_excel("after_PCA.xlsx")

In[5]:

data.set_index("文物采样点", inplace = True)

In[6]:

pb_ba_si_1 = ((data["主成分 1"] -data["主成分 1"].mean())/data["主成分 1"].std()).values

In[7]:

data0 = pd.read_excel("data_pro.xlsx")
data0 = data0[data0["类别"] == 0]

In[8]:

pbba = (data0["氧化铅(PbO)"] + data0["氧化钡(BaO)"])

```
# In[9]:

pbba = (pbba - pbba.mean())/pbba.std()

# In[10]:

pbba = pbba.values

# In[11]:

kca_si = ((data["主成分 2"] -data["主成分 2"].mean())/data["主成分 2"].std()).values

# In[12]:

std2_0 = data["主成分 7"] -data["主成分 6"]

# In[13]:

na_k = (std2_0 - std2_0.mean())/std2_0.std()

# In[14]:

na_k = na_k.values

# In[15]:

len(na_k)

# In[16]:
```

k_si = ((data["主成分 6"] -data["主成分 6"].mean())/data["主成分 6"].std()).values

In[17]:

pb_si = ((data["主成分 3"] -data["主成分 3"].mean())/data["主成分 3"].std()).values

In[18]:

index_4 = ((data["主成分 4"] -data["主成分 4"].mean())/data["主成分 4"].std()).values

In[44]:

```
import numpy as np
def entropy(data):
    #data 是从 DataFrame 中提取出来的 ndarray，横向为不同样本观测，纵向为不同特征指标
    data = (data-data.min())/(data.max() - data.min())
    [m,n] = data.shape
    k = 1/np.log(m)
    yij = data.sum(axis = 0)
    pij = li/yij
    test = pij*np.log(pij)
    test =np.nan_to_num(test)
    ej = -k*(test.sum(axis = 0))
    wi = (1-ej)/np.sum(1-ej)
    return wi
```

In[19]:

var_1 = pb_ba_si_1.std()

In[20]:

```
var_2 = pbba.std()


# In[21]:


index_1 = (var_1/(var_1+var_2) ) * pb_ba_si_1 + (var_2/(var_1+var_2) ) *pbba


# In[22]:


varr_1 = kca_si.std()
varr_2 = na_k.std()
index_2 = (varr_1/(varr_1+varr_2))*kca_si + (varr_2/(varr_1+varr_2))*na_k


# In[23]:


index_2


# In[24]:


#k_si
#pb_si
va_1 = k_si.std()
va_2 = pb_si.std()
index_3 = (va_1/(va_1+va_2))*k_si + (va_2/(va_1+va_2))*pb_si


# In[25]:


index_3


# In[26]:


zhibiao = pd.DataFrame([index_1, index_2, index_3]).T
zhibiao.columns = ["index_1", "index_2", "index_3"]
```

```
zhibiao.index = data.index

# In[27]:

zhibiao = zhibiao.values

# In[28]:

KM= KMeans(n_clusters=4, n_jobs = 4, max_iter = 500)
KM.fit(zhibiao)

# In[ ]:

zhibiao = pd.DataFrame(zhibiao)

# In[29]:

zhibiao["cluster"] = KM.labels_

# In[76]:

zhibiao[KM.labels_ == 0]

# In[97]:

zhibiao = zhibiao.iloc[:,:-1].values

# In[98]:

zhibiao
```

```
# In[ ]:


KM= KMeans(n_clusters=4, max_iter = 500)
KM.fit(zhibiao)
fig, ax = plt.subplots()
fig = plt.figure(figsize=(4.5))
ax=plt.subplot(224, projection='3d')
labels=np.unique(KM.labels_)
markers='o*^+'
for i,label in enumerate(labels):

ax.scatter(zhibiao[KM.labels_==label,0],zhibiao[KM.labels_==label,1],zhibiao[KM.labels_==label,2],
                label="cluster %d"%label,marker=markers[i],s=50)
    ax.scatter(KM.cluster_centers_[:,0],KM.cluster_centers_[:,1],KM.cluster_centers_[:,2],
           marker='X',s=60,c='r',label="小类中心")
    #ax.legend(loc="best",framealpha=0.5)
    ax.set_xlabel("X1")
    ax.set_ylabel("X2")
    ax.set_zlabel("X3")
    #ax.set_title("%d 个样本观测点的聚类结果"%N)
    plt.show()


    # In[ ]:


    第一个聚类的小类，是 Na-K-Si 体系和 Pb-Ca-Si 体系的综合，


    # In[109]:


    #第一小类，x1 和 x2 都比较高，对应的是蓝色的线
    zhibiao[KM.labels_==0,:]


    # 第二个聚类的小类，代表着典型的 Pb-Ba 体系，其他条件无法识别

    # In[112]:
```

#第二小类，x2 和 x3 的含量都显著的低，x1 分布不显著；由于是比值变量，其他成分显著减少，一定会导致第一类的铅钡含量上升；
zhibiao[KM.labels_==1,:]

In[111]:

#第三类，x1 比较低，x2 偏高，x3 较高水平分布
zhibiao[KM.labels_==2,:]

In[45]:

```
KM= KMeans(n_clusters=3,max_iter = 500)
KM.fit(zhibiao)
fig = plt.figure(figsize=(10,10),facecolor= 'white',edgecolor='black')#设置画布大小和颜色
ax = fig.add_subplot(224, projection = '3d')#相当于在原画板上加一个子模板
ax.patch.set_facecolor("white") # 设置 ax1 区域背景颜色
ax.patch.set_alpha(1.0) # 设置 ax1 区域背景颜色透明度

labels=np.unique(KM.labels_)
markers='o*^+'
for i,label in enumerate(labels):

ax.scatter(zhibiao[KM.labels_==label,0],zhibiao[KM.labels_==label,1],zhibiao[KM.labels_==label,2],
                label="cluster %d"%label,marker=markers[i],s=50)
ax.scatter(KM.cluster_centers_[:,0],KM.cluster_centers_[:,1],KM.cluster_centers_[:,2],
          marker='X',s=60,c='r',label="小类中心")
#ax.legend(loc="best",framealpha=0.5)
ax.set_xlabel("铅钡系统隶属度")
ax.set_ylabel("钾钙硅系统隶属度")
ax.set_zlabel("钾铅硅系统隶属度")
#ax.set_title("%d 个样本观测点的聚类结果"%N)
plt.show()
```

In[118]:

```
    KM= KMeans(n_clusters=3,max_iter = 500)
    KM.fit(zhibiao)
    fig = plt.figure(figsize=(10,10),facecolor= 'white',edgecolor='black')#设置画布大小和颜
色
    ax = fig.add_subplot(224, projection = '3d')#相当于在原画板上加一个子模板
    ax.patch.set_facecolor("white") # 设置 ax1 区域背景颜色
    ax.patch.set_alpha(1.0) # 设置 ax1 区域背景颜色透明度


    labels=np.unique(KM.labels_)
    markers='o*^+'
    for i,label in enumerate(labels):

ax.scatter(zhibiao[KM.labels_==label,0],zhibiao[KM.labels_==label,1],zhibiao[KM.labels_==la
bel,2],
               label="cluster %d"%label,marker=markers[i],s=50)
    ax.scatter(KM.cluster_centers_[:,0],KM.cluster_centers_[:,1],KM.cluster_centers_[:,2],
          marker='X',s=60,c='r',label="小类中心")
    #ax.legend(loc="best",framealpha=0.5)
    ax.set_xlabel("主成分 1")
    ax.set_ylabel("主成分 2")
    ax.set_zlabel("主成分 3")
    #ax.set_title("%d 个样本观测点的聚类结果"%N)
    plt.show()


    # In[116]:


    plt.figure(figsize=(15,15))
    K=[2,3,4,5,6,7]
    markers='o*^+X<>'
    Fvalue=[]
    silhouettescore=[]
    chscore=[]
    i=0
    for k in K:
        KM= KMeans(n_clusters=k, init='k-means++',random_state=1,max_iter = 500)
        KM.fit(zhibiao)
        tmp=f_classif(zhibiao, KM.labels_)
        Fvalue.append(sum(tmp[0]))
        score=calinski_harabasz_score(zhibiao,KM.labels_)
        chscore.append(score)
```

```
    score=silhouette_score(zhibiao,KM.labels_)
    silhouettescore.append(score)
    labels=np.unique(KM.labels_)
    plt.subplot(3,3,i+1)
    i+=1

    plt.xlabel("X1")
    plt.ylabel("X2")
    plt.title("聚类结果(K=%d)"%k)

plt.subplot(3,3,i+1)
plt.plot(K,Fvalue)
Fvalue = 100-Fvalue
plt.xlabel("聚类数目 K")
plt.ylabel("比值 F")
plt.title("比值 F 与聚类数目")
plt.subplot(3,3,i+2)
plt.plot(K,chscore)
chscore = 35 - chscore
plt.xlabel("聚类数目 K")
plt.ylabel("CH 指数")
plt.title("CH 指数与聚类数目")
plt.subplot(3,3,i+3)
plt.plot(K,silhouettescore)
plt.xlabel("聚类数目 K")
plt.ylabel("轮宽")
plt.title("轮宽与聚类数目")
plt.subplots_adjust(hspace=0.3)
plt.subplots_adjust(wspace=0.3)


# In[30]:


zhibiao_2 = pd.DataFrame([index_1, index_2, index_3, index_4]).T
zhibiao_2.columns = ["指标 1","指标 2","指标 3","指标 4"]
zhibiao_2.index = data.index


# In[32]:


zhibiao_2.to_excel("zhibiao_2.xlsx")
```

In[ ]:

In[125]:

```
plt.figure(figsize=(15,15))
K=[2,3,4,5,6,7]
markers='o*^+X<>'
Fvalue=[]
silhouettescore=[]
chscore=[]
i=0
for k in K:
    KM= KMeans(n_clusters=k, init='k-means++',random_state=1,max_iter = 500)
    KM.fit(zhibiao_2)
    tmp=f_classif(zhibiao_2, KM.labels_)
    Fvalue.append(sum(tmp[0]))
    score=calinski_harabasz_score(zhibiao_2,KM.labels_)
    chscore.append(score)
    score=silhouette_score(zhibiao_2,KM.labels_)
    silhouettescore.append(score)
    labels=np.unique(KM.labels_)
    plt.subplot(3,3,i+1)
    i+=1

    plt.xlabel("X1")
    plt.ylabel("X2")
    plt.title("聚类结果(K=%d)"%k)

plt.subplot(3,3,i+1)
plt.plot(K,Fvalue)
Fvalue = Fvalue
plt.xlabel("聚类数目 K")
plt.ylabel("比值 F")
plt.title("比值 F 与聚类数目")
plt.subplot(3,3,i+2)
plt.plot(K,chscore)
chscore =  chscore
plt.xlabel("聚类数目 K")
```

```
plt.ylabel("CH 指数")
plt.title("CH 指数与聚类数目")
plt.subplot(3,3,i+3)
plt.plot(K,silhouettescore)
plt.xlabel("聚类数目 K")
plt.ylabel("轮宽")
plt.title("轮宽与聚类数目")
plt.subplots_adjust(hspace=0.3)
plt.subplots_adjust(wspace=0.3)


# In[33]:


KM_best = KMeans(n_clusters=3, init='k-means++',random_state=1,max_iter = 500)
KM_best.fit(zhibiao_2)
KM_best.labels_


# In[142]:


KM_best = KMeans(n_clusters=3, init='k-means++',random_state=1,max_iter = 500)
KM_best.fit(zhibiao_2)
KM_best.labels_


# In[34]:


zhibiao_2["cluster"] = KM_best.labels_


# In[37]:


zhibiao_2.columns = ["铅-钡系统隶属度","钾-钙-硅系统隶属度","钾-铅-硅系统隶属度","次要成分","cluster"]


# In[ ]:


铅-钡系统隶属度 钾-钙-硅系统隶属度  钾-铅-硅系统隶属度  次要成分
```

In[43]:

zhibiao_2[zhibiao_2["cluster"] == 0].to_excel("类型 1.xlsx")

第一类：有较高含量的铅钡、钾还有钠；还有镁铝含量较低，可以判断是古代的工艺（战国至东汉），由于技术不先进导致的；不能判断是铅钡硅系统，还是钾钙硅（或者是钠钙硅系统）；
第二类，有较高含量的钾钠、钾铅、镁铝，是近代的工艺，由于唐代中西方交流较多，所以确定是钾、铅、硅系统
第三类，只有较高的铅钡，镁铝含量比较低（战国前），所以判断是铅钡硅系统

In[147]:

type_1 = zhibiao_2[zhibiao_2["cluster"] == 0]

In[153]:

type_1["main_1"] = (data.loc[type_1.index]["主成分 1"] - data.loc[type_1.index]["主成分 1"].mean())/data.loc[type_1.index]["主成分 1"].std()

In[154]:

type_1["main_3"] = (data.loc[type_1.index]["主成分 3"] - data.loc[type_1.index]["主成分 3"].mean())/data.loc[type_1.index]["主成分 3"].std()

In[161]:

type_1["nak"] = na_k.loc[type_1.index]

In[163]:

```
type_1.iloc[:,-3:]


# In[170]:


KM_1 = KMeans(n_clusters=5, init='k-means++',random_state=1,max_iter = 500)
KM_1.fit(type_1.iloc[:,-3:])
KM_1.labels_


# In[171]:


type_1["labels"] = KM_1.labels_


# In[172]:


type_1


# In[165]:


def dif_clu(data):
    plt.figure(figsize=(15,15))
    K=[2,3,4,5,6,7]
    markers='o*^+X<>'
    Fvalue=[]
    silhouettescore=[]
    chscore=[]
    i=0
    for k in K:
        KM= KMeans(n_clusters=k, init='k-means++',random_state=1,max_iter = 500)
        KM.fit(data)
        tmp=f_classif(data, KM.labels_)
        Fvalue.append(sum(tmp[0]))
        score=calinski_harabasz_score(data,KM.labels_)
        chscore.append(score)
        score=silhouette_score(data,KM.labels_)
        silhouettescore.append(score)
        labels=np.unique(KM.labels_)
```

```
        plt.subplot(3,3,i+1)
        i+=1

        plt.xlabel("X1")
        plt.ylabel("X2")
        plt.title("聚类结果(K=%d)"%k)

    plt.subplot(3,3,i+1)
    plt.plot(K,Fvalue)
    Fvalue = Fvalue
    plt.xlabel("聚类数目 K")
    plt.ylabel("比值 F")
    plt.title("比值 F 与聚类数目")
    plt.subplot(3,3,i+2)
    plt.plot(K,chscore)
    chscore =  chscore
    plt.xlabel("聚类数目 K")
    plt.ylabel("CH 指数")
    plt.title("CH 指数与聚类数目")
    plt.subplot(3,3,i+3)
    plt.plot(K,silhouettescore)
    plt.xlabel("聚类数目 K")
    plt.ylabel("轮宽")
    plt.title("轮宽与聚类数目")
    plt.subplots_adjust(hspace=0.3)
    plt.subplots_adjust(wspace=0.3)
    return ax


# In[173]:


type_2 = zhibiao_2[zhibiao_2["cluster"] == 1]


# In[175]:


type_2["nak"] = na_k.loc[type_2.index]


# In[176]:
```

type_2

In[178]:

type_3 = zhibiao_2[zhibiao_2["cluster"] == 2]

In[179]:

type_3

In[180]:

type_3["main_1"] = (data.loc[type_3.index]["主成分 1"] - data.loc[type_3.index]["主成分 1"].mean())/data.loc[type_3.index]["主成分 1"].std()
type_3["main_3"] = (data.loc[type_3.index]["主成分 3"] - data.loc[type_3.index]["主成分 3"].mean())/data.loc[type_3.index]["主成分 3"].std()

In[182]:

type_3

In[323]:

zhibiao_2_min = zhibiao_2.iloc[:,:-1]
zhibiao_2_min.shape

In[324]:

e = np.random.rand(15,4)

In[325]:

```
e = 0.1*e
```

```
# In[326]:
```

```
e = 1+e
```

```
# In[327]:
```

```
RAND = np.random.randint(1,49,15)
```

```
# In[328]:
```

```
len(RAND)
```

```
# In[329]:
```

```
e[1,2]
```

```
# In[330]:
```

```
zhibiao_2_min.iloc[RAND[1] 2]
```

```
# In[331]:
```

```
for i in range(15):
    for j in range(4):
        zhibiao_2_min.iloc[RAND[i], j] = zhibiao_2_min.iloc[RAND[i], j]*e[i,j]
```

```
# In[332]:
```

```
test = KMeans(n_clusters=3, init='k-means++',random_state=1,max_iter = 500)
test.fit(zhibiao_2_min)
score=calinski_harabasz_score(zhibiao_2_min,KM.labels_)


# In[333]:


sc_list.append(score)


# In[334]:


sc_list


# In[343]:


list(range(1,11))


# In[49]:


a = [0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07,0.08,0.09,0.1]


# In[48]:


sc_list = [20.20156008549269,
 20.23469855401065,
 20.09830205943167,
 20.249424738744327,
 20.14711956213794,
 20.166743468085617,
 20.02516856919142,
 20.286120241125076,
 20.3417389531109,
 20.317798943835637]
```

In[50]:

```
fig,ax = plt.subplots()
ax.plot(a,sc_list, marker = "o")
ax.set_xlabel("样本数据随机变动")
ax.set_ylabel("聚类性能")
ax.set_title("灵敏度分析")
```

In[345]:

```
fig,ax = plt.subplots()
ax.plot(a,sc_list, marker = "o")
ax.set_xlabel("样本数据德宾变动")
ax.set_ylabel("聚类性能")
ax.set_title("灵敏度分析")
```

In[46]:

```
fig,ax = plt.subplots()
ax.plot(a,sc_list, marker = "o")
ax.set_xlabel("样本数据德宾变动")
ax.set_ylabel("聚类性能")
ax.set_title("灵敏度分析")
```

In[226]:

```
score1 = score
```

In[235]:

```
score2 = score
```

In[249]:

score3 = score

In[261]:

score4 = score

In[273]:

score5 = score

In[274]:

sc_list = [score1, score2, score3, score4, score5]

In[ ]:

sc_list