基于统计分析对玻璃制品成分的研究

摘要：古代玻璃极易收埋藏环境的影响而风化，在风化过程中，内部元素与环境进行大量交换，导致其成分比例发生变化，从而影响其类别的正确判断。本文根据不同风化情况建立了相应的模型进行求解。

针对问题一，首先建立了多因素方差分析模型并结合斯皮尔曼相关系数在 SPSS 中得出玻璃文物的表面风化与其玻璃类型和纹饰具有显著性关系，与颜色显著性关系不大，并且均具有相关性。然后建立了独立样本t检验模型，得到风化前后高钾玻璃的二氧化硅等含量较多铅钡含量变化不大的统计规律。最后建立加权均值模型，通过风化前后映射关系和 Matlab 软件预测出风化前化学成分含量。

针对问题二，首先建立随机森林模型，并向其中投入题目数据进而得到模型的求解，然后根据模型的表现得出高钾玻璃氧化铅、氧化钡的含量极少，二氧化硅含量一般高于 55%，铅钡玻璃则相反。然后建立 K-means 聚类模型对玻璃进行亚类划分，根据得到的类别和权重的不同划分出不同的亚类。接着控制部分化学成分含量的增加或减少得出敏感性低，最后在使用主成分分析法和 TOPSIS 方法结合对数据重新进行一次分类，对比两次的结果印证了分类结果的合理性。

针对问题三，在求解问题二的基础上，将题目中表单三的数据投入随机森林模型中鉴别出了所归属的基本玻璃类型，然后根据问题二亚类的划分再对表单三的玻璃进行细分。最后同样通过上下更改化学成分含量的比率来分析分类结果的敏感性。

针对问题四，为了分析各成分之间的关联性，可以选择建立相关性分析模型，利用模型计算出斯皮尔曼系数，然后构造相关系数矩阵，代入求解，最后通过结果分析出不同类型的各化学成分之间的关联性与差异性。

关键词：多因素方差分析 独立样本t检验 随机森林 聚类分析 TOPSIS

一、问题重述

1.1 问题背景

早期的玻璃通过丝绸之路由西南亚传入我国，与外来的玻璃制品有比较相似的外观，但其化学成分却大不相同。玻璃的主要原料是石英砂，主要成分是二氧化硅（SiO2）。炼制过程通过添加不同的助燃剂会有不同的化学成分。例如，铅钡玻璃在烧制过程中加入铅矿石作为助熔剂，其氧化铅（PbO）、氧化钡（BaO）的含量较高。高钾玻璃在烧制过程中加入草木灰作为助熔剂，古代玻璃易受埋藏的环境的影响而风化，风化过程中，内外元素发生交换，导致其成分比例发生变化。

1.2 问题的提出

题目中已知：成分比例累加和介于 85%~105%之间的视为有效数据。

问题一：分析玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系；根据玻璃类型，分析文物有无风化化学成分含量的规律；根据风化点检测数据，预测风化前化学成分含量。

问题二：依照数据分析高钾、铅钡玻璃的分类规律；通过选择不同的化学成分对不同类型的玻璃进行亚类划分，并给出划分方法和结果；给出分类结果合理性和敏感性的依据。

问题三：分析表单 3 未知玻璃文物的化学成分，鉴别其所属类型；并对分类结果的敏感性进行分析。

问题四：对不同类型玻璃，分析化学成分的关联关系；比较不同类别间化学成分关联关系的差异性。

二、 问题分析

2.1 问题一的分析

为了分析玻璃的表面风化分别与玻璃类型、纹饰和颜色的关系。可以使用多因素方差分析，首先对数据进行检验是否服从正态分布，然后对其显著性 p 值是否小于 0.05 进行差异性判别并且判断是否需要进行事后多重比较，从而得到了成分之间的差异性关系。接下来可以选择使用斯皮尔曼相关系数[4]来进行相关性的判断。关于分析文物样品表面有无风化化学成分含量的统计规律，可以使用

独立样本 t 检验并结合正态分布检验得到统计规律[2]。为了预测风化前的化学成分含量，可以考虑使用加权均值法找出风化前后的成分占比的函数关系，预测其风化前成分含量，并进行结果比对。

2.2 问题二的分析

为了不同类型的玻璃的分类规律，可以通过数据建立随机森立分类模型，代入题目中的数据得到一个随机森林模型，进而计算不同类型玻璃化学成分的特征重要性，将测试数据导入模型得出分类规律与评估结果。可以使用聚类分析 K-means 对玻璃进行亚类分析，然后根据 K-means 的结果进行划分亚类并且根据所占权重比的不同从而得到具体的划分方法和结果。为了验证分类结果的合理性，可以选择引入主成分分析法（PCA）和优劣解距离法（TOPSIS）进行再次分类，并与 K-means 分类结果进行对比从而验证合理性。最后可以通过在一区间内上下浮动改变化学成分的含量，并将浮动变化后的结果与原记录进行对比，从而得出敏感性的分析。

2.3 问题三的分析

在问题 2 求解结果的基础上，将附件表单 3 中的数据代入问题 2 建立的随机森林模型[5]鉴别出归属高钾或铅钡玻璃类型，然后根据化学成分的不同并结合问题 2 中 K-means 的划分的方法鉴别所归属的亚类。进行同问题 2 的敏感性分析操作，得到分类结果敏感性的分析。

2.4 问题四的分析

为了得到化学成分的关联性，建立相关性分析的模型，将表单 2 中的数据依据类型分成两份导入模型，最终得到两份表格数据，再依据数据构造两个相关性系数矩阵，计算并分析，得到最终的关联性结果。

三、模型假设

假设 1：题目中的数据足够完全正确，且不存在缺失。

假设 2：玻璃文物只受到了环境风化的影响，没有其他的干扰因素。

假设 3：玻璃文物的化学成分含量均匀分布。

假设 4：玻璃文物的分类只与其占有较大比重的化学成分含量有关。

假设 5：模型进行分析预测的时候未出现极端特殊情况。

四、符号说明

为使文章前后一致，特地对文中符号做如下说明

符号	符号说明
S_A	效应平方和
$Z-Score_{ij}$	零均值法对数据进行处理
S_T	多因素方差法中的误差表示
s_p^2	独立样本 t 检验中的样本方差
$Cov(X,Y)$	样本协方差
r_{ij}	构建的协方差矩阵
$\sigma_X(sigmaX)$	X 的标准差

五、模型的建立与求解

5.1 问题一的建模与求解

5.1.1 分析玻璃文物表面风化与其玻璃类型和纹饰的关系

1）数据的检验

表 5.1 正态检验结果

变量名	样本量	中位数	平均值	标准差	偏度	峰度
表面风化	58	2	1.586	0.497	-0.359	-1.939

上表为表面风化统计和正态性检验的结果，用于检验数据的正态性。分析得表面风化样本 N ＜ 5000，采用 S-W 检验，显著性 P 值为 0.000***，水平呈现显著性（P<0.05），数据满足正态分布[1]。

2）模型的建立

$$\begin{cases} X_{ij}=\mu_i+\varepsilon_{ij} \\ \varepsilon_{ij}\sim N(0,\sigma^2)，\ 各\varepsilon_{ij}独立 \\ j=1,2,\cdots,n_i,i=1,2,\cdots,\mathrm{r}. \end{cases}$$

X_{ij} 就是第 i 个水平的第 j 个观测值，μ_i 表示第 i 个水平的理论均值，后面的 ε_{ij} 表示随机误差，假设数据服从正态分布。等式一可以作为某个观测值用某水平下的均值加一个误差表示[5]。

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$$

总偏差平方和，如果这个值越大，就表示 X_{ij} 之间的差异越大

$$\bar{X}=\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}}{n},\quad n=n_1+n_2+,\cdots,+n_n$$

n 表示总的观测值个数

$$S_{\mathrm{E}}=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{\mathrm{i}})^2$$

误差平方和公式，其中 $\bar{X}_{\mathrm{i}}=\frac{\mathrm{X}_{i1}+\mathrm{X}_{i2}+\cdots+\mathrm{X}_{in}}{n_i}$

上面两者相减，得到效应平方和

$$S_A=\sum_{i=1}^{r}n_i(\bar{X}_{\mathrm{i}}-\bar{X})^2$$

$\bar{X}_{\mathrm{i}}$ 可看作是每个水平的理论平均值的估计，故每个理论平均值越大，$\bar{X}_{\mathrm{i}}$ 的差异也会越大，S_A 就是衡量不同水平之间的差异程度。

3）模型的求解

3.1）差异性

将上述模型代入至 SPSS 中得到如下表格

表 5.2 方差分析结果

变量名	变量值	样本量	平均值	标准差	F	P
表面风化	铅钡	40	1.7	0.464	8.804	0.004***
	高钾	18	1.333	0.485		
	合计	58	1.586	0.497		

上表展示了方差分析的结果，说明变量表面风化在类型之间存在显著性差异，需要进行事后多重比较；

表 5.3 事后多重比较结果

	(I)名称	(J)名称	(I)平均值	(J)平均值	差值(I-J)	P
表面风化	铅钡	高钾	1.7	1.333	0.367	0.008***

上表是事后多重比较的结果，对变量之间具体差异进行分析。

使用 LSD 方法的事后多重比较的结果显示：对于变量表面风化，均值大小排序为：铅钡>高钾。其中铅钡与高钾存在显著性差异。

3.2）相关性

将数据代入 SPSS 中进行斯皮尔曼相关性分析得到如下结果

表 5.4 相关性分析结果

	表面风化	类型	纹饰	颜色
表面风化	1.000(0.000***)	0.346(0.016**)	0.106(0.475)	-0.061(0.679)
类型	0.346(0.016**)	1.000(0.000***)	-0.386(0.007***)	0.587(0.000***)

纹饰	0.106(0.475)	-0.386(0.007***)	1.000(0.000***)	-0.646(0.000***)
颜色	-0.061(0.679)	0.587(0.000***)	-0.646(0.000***)	1.000(0.000***)

4）结果描述

由上述过程可知，玻璃文物的表面风化与其玻璃类型有显著性关系并且铅钡玻璃的风化程度高于高钾玻璃。

玻璃文物的表面风化与其纹饰有显著性关系并且纹饰 B 的风化程度高于纹饰 C，纹饰 C 的风化程度高于纹饰 A。

玻璃文物的表面风化与其颜色显著性关系不大，其均值影响大小为：黑>紫>浅蓝>蓝绿>深绿>浅绿。

上表展示了模型检验的参数结果表，包括了相关系数、显著性 P 值和差异性，所以可以发现玻璃文物的表面风化与其玻璃类型、纹饰和颜色具有相关性。

5.1.2 分析玻璃表面有无风化化学成分含量的统计规律

1） 数据的检验

对表单二中的数据进行正态分布检验，将其代入至 SPSS 中得到如下图表

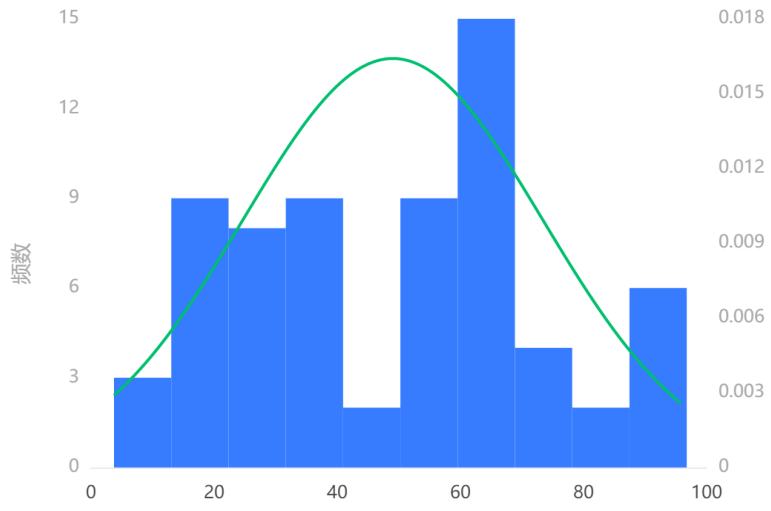

图 5.1 正态分布检验

观察可以发现数据基本满足正态分布，所以可以使用独立样本 t 检验[6]

2）模型的建立

建立独立样本 t 检验的方程

$$t=\frac{(M_1-M_2)-(\mu_1-\mu_2)}{S(M_1-M_2)}$$

其中(M_1-M_2)为样本均值差异，$(\mu_1-\mu_2)$为总体均值差异，$S(M_1-M_2)$为估计标准误

由于样本量不相等所以标准误为

$$S(M_1-M_2)=\sqrt{\frac{s_p^{\ 2}}{n_1^{\ 2}}+\frac{s_p^{\ 2}}{n_2^{\ 2}}}$$

其中$s_p^{\ 2}$为样本方差$\frac{SS_1+SS_2}{df_1+df_2}$

计算 Cohen’sd 系数为

$$d=\frac{M_1-M_2}{\sqrt{s_p^{\ 2}}}$$

测量变异的解释比例为

$$\mathrm{r}^2=\frac{t^2}{t^2+df}$$

置信区间为

$$\mu_1-\mu_2=(M_1-M_2)\pm \mathrm{ts}_{M_1-M_2}$$

3）模型的求解

将上述模型代入附录 13 代码中得到 Cohen’sd 系数为 d=0.2，测量变异解释比例r^2=0.01 均满足小效应，所以可以使用该模型，接着代入 SPSS 可视化处理后得到结果如下

对于高钾玻璃结合附录 6 分析结果得到如下统计规律

表 5.5 高钾玻璃统计规律

风化	二氧化硅含量较多，氧化铜含量基本不变，其余化合物含量较低

未风化	氧化铜含量略微增加，二氧化硅含量较少，其余化合物含量较高

对于铅钡玻璃结合附录 6 分析结果得到如下统计规律

表 5.6 铅钡玻璃统计规律

风化	氧化钠，氧化钙，氧化铝，氧化镁，氧化铜含量较高，其余含量较低
未风化	氧化钠，氧化钙，氧化铝，氧化镁，氧化铜含量较低，氧化钡和氧化锶含量基本没有变化，其余含量较高

5.1.3 根据风化点数据预测风化前化学成分含量

1）模型的建立

建立计算加权平均值的计算模型， A_j ， B_j 为加权平均值

x_{ij} 代表第 j 个化学成分的第 i 次检测

$$A_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}, (i = 1...n_1, j = 1...m)$$

y_{ij} 代表第 j 个化学成分的第 i 次检测

$$B_j = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{ij}, (i = 1...n_2, j = 1...m)$$

C_j 为变化率

$$C_j = \frac{(A_j - B_j)}{B_j} \times 100\% (j = 1...m)$$

然后将 C_j 代入

$$A_j = C_j B_j + B_j$$

得到风化前化学成分的含量

2） 模型的求解

将题目表格二中的数据带入到附录 1.3.1 和 1.3.2 代码中以及带入 SPSS 分析得得到如下结果（缺失表格部分在附录 3）

对于高钾玻璃可得

表 5.7 差值比率

元素	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)
风化前	67.98416667	2.78	10.17909091	6.399	1.295
风华后	93.96333333	0	0.815	0.87	0.59
差值	-25.97916667	2.78	9.364090909	5.529	0.705
比率 C	-0.276481961	0	11.4896821	6.355172414	1.194915254

表 5.8 高钾玻璃风化前的化学成分含量预测结果

编号	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)
22	66.8168909	0	9.242364752	12.20958621	1.404745763
27	67.08459257	0	0	6.913862069	1.185254237
07	67.01947595	0	0	7.870034483	0
09	68.74868406	0	7.368912437	4.560206897	0
10	96.77	0	0.92	0.21	0
12	94.29	0	1.01	0.72	0

表 5.9 铅钡玻璃比值计算

元素	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)
风化后	33.61472222	3.117272727	0.32125	2.483529412	1.096956522
风化前	53.44384615	3.343333333	0.42	1.455454545	0.914285714
差值	19.82912393	0.226060606	0.09875	-1.028074866	-0.182670807
比率 C	0.589894029	0.072518713	0.307392996	-0.413957194	-0.166525112

表 5.10 铅钡玻璃风化前的化学成分含量预测结果

文物编号	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)
02	57.68135538	0	1.372762646	1.371340166	0.983500368

08	32.02046575	0	0	0.867343353	0
08	7.329411475	0	0	1.869476551	0
11	53.40454044	0	0.274552529	2.057010249	0.591767171
19	47.12445903	0	0	1.717105422	0.491750184
23	85.52039983	8.494348206	0	0.293021403	0.591767171
25	80.46453682	2.477518227	0	0.369206968	0
26	31.46400284	0	0	0.843901641	0
26	5.914405789	0	0.522957198	1.763988846	0
28	108.2399855	0	0.339922179	0.78529736	0.833474888
29	100.640292	0.986717216	0.392217899	1.746407562	1.241877583

5.2 问题二的建模与求解

5.2. 分析高钾玻璃和铅钡玻璃的分类规律

1）数据的检验

由问题一可知，表单数据满足基本的正态分布。

2）模型的建立

假设类别的集合$\{c_1,c_2,\ldots,c_n\}$，为了方便讨论，这里将$(h_i^1(x),h_i^2(x),....,h_i^n(x))$在样本$x$上的预测输出表示为一个 n 维向量$(h_i^1(x),h_i^2(x),....,h_i^n(x))$，其中$h_i^j(x)$表示$h_i$在类别$c_j$上的输出根据绝对多数投票法得

$$H(x)=\begin{cases} c_j,\sum_{i=1}^{T}h_i^j(x)>0.5\sum_{k=1}^{n}\sum_{i=1}^{T}h_i^k(x) \\ reject,\text{其他} \end{cases}$$

只要当某个标记得票过半数，该预测为该类别，否则拒绝预测

随机森林的泛化误差

$$\text{泛化误差}\le\frac{\bar{\rho}(1-s^2)}{s^2}$$

$\bar{\rho}$为树之间的平均相关系数，s 是度量树型分类器的“强度的量”

3） 模型的求解

代入附录 13 代码和 SPSS 中，得到如下结果（部分表格在附录）

表 5.11 预测结果

预测结果 Y	类型	预测结果概率_铅钡	预测结果概率_高钾	二氧化硅 (SiO2)	氧化钠 (Na2O)	氧化钾 (K2O)	氧化钙 (CaO)	氧化镁 (MgO)
铅钡	铅钡	0.97	0.03	54.61	0	0.3	2.08	1.2
铅钡	铅钡	0.99	0.01	50.61	2.31	0	0.63	0
铅钡	铅钡	1	0	29.64	0	0	2.93	0.59
高钾	高钾	0.11	0.89	87.05	0	5.19	2.01	0
高钾	高钾	0.03	0.97	69.33	0	9.99	6.32	0.87

由表格得，高钾玻璃的氧化铅、氧化钡的含量极少甚至没有，二氧化硅的含量一般高于 55%；铅钡玻璃的氧化铅、氧化钡的含量较高，二氧化硅的含量一般都低于 55%。

5.2.2 对每个类别进行合适的亚类划分

1） 模型的建立

利用 K-means 聚类分析建立模型，先根据准则函数求出每一个样本到类中心距离的平方之和

对所有 k 个模式类有

$$J=\sum_{j=1}^{k}\sum_{i=1}^{N_j}\| X_i - Z_j \|^2, X_i \in S_j$$

S_j :第j个聚类集，聚类中心为Z_j;

N_j :第j个聚类集S_j中所包含的样本个数；

为了使准则函数 J 极小，所以

$$\frac{\delta J_j}{\delta Z_j}=0$$

即

$$\frac{\delta}{\delta}\sum_{i=1}^{N_j}\| X_i - Z_j \|^2 = \frac{\delta}{\delta Z_j}\sum_{i=1}^{N_j}(X_i - Z_j)^T(X_i - Z_j) = 0$$

可以解得

$$Z_j = \frac{1}{N_j}\sum_{j=1}^{N_j} X_i, X_i \in S_j$$

2）模型的求解

将上述模型代入 SPSS 以及附录 13 代码中得到（表格缺失部分见附录）

表 5.12 高钾玻璃的聚类分析结果

	聚类种类	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)
0	1	69.33	0	9.99	6.32	0.87
1	2	87.05	0	5.19	2.01	0
2	1	61.71	0	12.37	5.87	1.11
3	1	65.88	0	9.67	7.12	1.56
4	1	61.58	0	10.95	7.35	1.77
5	1	67.65	0	7.37	0	1.98

表 5.13 铅钡玻璃的聚类分析结果

	聚类种类	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)
0	1	36.28	0	1.05	2.34	1.18
1	1	20.14	0	0	1.48	0
2	1	4.61	0	0	3.19	0
3	1	33.59	0	0.21	3.51	0.71
4	1	29.64	0	0	2.93	0.59

5	2	37.36	0	0.71	0	0

3）结果分析

对于高钾玻璃的亚类划分：

第一类二氧化硅占比含量基本都小于 70%，部分文物含有少量的氧化纳，且氧化钙占比含量均在 6%以上，因此可将此类高钾玻璃命名为钙钾玻璃。

第二类二氧化硅占比含量基本都大于 80%，且均不含有氧化纳，且氧化钙占比含量均在 6%以下，因此可将此类高钾玻璃命名为硅钾玻璃。

对于铅钡玻璃的亚类划分：

第一类二氧化硅占比含量基本都小于 40%，氧化铅含量基本大于 30%，因此可将此类铅钡玻璃命名为硅铅玻璃。

第二类二氧化硅占比含量基本都大于 50%，氧化铅含量基本小于 20%，因此可将此类铅钡玻璃命名为硅铅钢玻璃。

5.2.3 合理性和敏感性分析

1）模型的建立

合理性：

针对 K-means 的分类结果的合理性检验，可以选择使用主成分分析法（PCA）和优劣解距离法（TOPSIS）计算化学成分的权重比与 K-means 的结果进行对比，从而验证合理性

使用主成分分析法进行数据的降维处理：

进行数据的 m*n 的矩阵构建

$$X=\begin{pmatrix} x_{11}\cdots x_{1n} \\ \ddots \\ x_{m1}\cdots x_{mn} \end{pmatrix}=(X_1,X_2,X_3\cdots X_n)$$

利用零均值法对数据处理，得到均值为 0，标准差为 1 的服从标准正态分布的数据

$$Z-Score_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}$$

其中 $\bar{x}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij}$ 为第 j 个指标的样本均值 $S_j=\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2}$ 为第 j 个指标的标准差

对上述经过处理后的 X 矩阵构建其协方差矩阵

$$r_{ij}=\mathrm{cov}(x_{ki},x_{kj})=\frac{\sum_{k=1}^{m}x_{ki}x_{kj}}{m-1}$$

$$R=\begin{pmatrix}\mathrm{cov}(X_1,X_1)\,\mathrm{cov}(X_1,X_1)\cdots\mathrm{cov}(X_1,X_1)\\ \ddots\\ \mathrm{cov}(X_n,X_1)\,\mathrm{cov}(X_1,X_1)\cdots\mathrm{cov}(X_1,X_1)\end{pmatrix}$$

求解上述协方差矩阵的特征方程 $R-\lambda E=0$ 得到的对应特征向量依次为

$$\beta_1=\begin{bmatrix}\beta_{11}\\ \beta_{21}\\ \vdots\\ \beta_{n1}\end{bmatrix},\beta_2=\begin{bmatrix}\beta_{12}\\ \beta_{22}\\ \vdots\\ \beta_{n2}\end{bmatrix},\cdots\beta_n=\begin{bmatrix}\beta_{1n}\\ \beta_{2n}\\ \vdots\\ \beta_{nn}\end{bmatrix}$$

最后再利用贡献率阈值确定出主成分个数

$$\alpha\le\frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{n}\lambda_i}$$

出主成分个数后，对前 p 个特征值对应的特征向量进行计算

$$Z_i=\beta_i^TX=\beta_{1i}X_1+\beta_{2i}X_2+\cdots+\beta_{ni}{}^TX_n,i=1,2,\cdots p$$

使用 TOPSIS 方法计算理想解：

设多属性决策问题的决策矩阵为规范化决策矩阵为 $A=(a_{ij})_{m\times n}$ 规范化决策矩阵为 $B=(b_{ij})_{m\times n}$，其中

$$b_{ij} = a_{ij} / \sqrt{\sum_{i=1}^{m} a_{ij}^{\ 2}}, i = 1,2,\cdots,n$$

构建加权规范阵$C = (c_{ij})_{m\times n}$并利用 2.3.1 中主成分分析的结果来决定权重向量$\omega = [w_1, w_2, \cdots, w_n]^T$则

$$c_{ij} = w_{ij} \cdot b_{ij}, i = 1,2,\cdots,m; j = 1,2,\cdots n$$

确定正理想解C^*和负理想解C^0设正理想解C^*的第 j 个属性值为c_j^*，负理想解C^0的第 j 个属性值为c_j^0，则

正理想解：

$$c_j^* = \begin{cases} \max_i c_{ij,} j\text{为效益型属性} \\ \min c_{ij,} j\text{为成本型属性} \end{cases} j = 1,2,\cdots,n$$

负理想解：

$$c_j^0 = \begin{cases} \max_i c_{ij,} j\text{为效益型属性} \\ \min c_{ij,} j\text{为成本型属性} \end{cases} j = 1,2,\cdots,n$$

计算方案到正理想解与负理想解的距离

到正理想解的距离为

$$s_i^* = \sqrt{\sum_{j=1}^{n} (c_{ij} - c_j^*)^2}, i = 1,2,\cdots,m$$

到负理想解的距离为

$$s_i^0 = \sqrt{\sum_{j=1}^{n} (c_{ij} - c_j^0)^2}, i = 1,2,\cdots,m$$

计算综合评价指数为：

$$f_i^* = s_i^0 / (s_i^0 + s_i^*), i = 1,2,\cdots,m$$

敏感性：

选取两种玻璃类型中占比重较大的二氧化硅该化学成分，对该成分的比重进行上调 10%和下调 10%的操作。再将操作之后的数据，代入到聚类分析出来的结果中，对比原结果，观察两个结果的差异大小。

2）模型的求解

将上述模型代入到附录 13 代码和 SPSS 中得到如下分析结果

表 5.14 总方差解释

总方差解释			
成分	特征根		
	特征根	方差百分比	累积
二氧化硅	4.055	28.97%	28.97%
氧化钠	2.413	17.23%	46.20%
氧化钾	1.739	12.42%	58.62%
氧化钙	1.124	8.03%	66.65%
氧化镁	1.11	7.93%	74.58%
氧化铝	0.784	5.60%	80.19%

由图表可知这六种成分的总累计百分比超过了百分之八十，所以可以选择这六种成分作为主成分使用。

再将主成分分析结果代入到上述 TOPSIS 模型中，并根据附录 13 代码和 SPSS 求解得到（缺失部分请参看附录）

表 5.15 熵权法结果

熵权法			
项	信息熵值 e	信息效用值 d	权重
二氧化硅 (SiO2)	0.962	0.038	0.032
氧化钠 (Na2O)	0.634	0.366	0.317
氧化钾 (K2O)	0.653	0.347	0.3
氧化钙 (CaO)	0.903	0.097	0.084

由表格可知占据主要分类权重的化学成分以及分类结果与 K-means 分类结果具备一致，证明了 K-means 分类的合理性

敏感性：

表 5.16 高钾 上调 10%

	聚类种类	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)
0	1	62.397	0	9.99	6.32	0.87
1	2	78.345	0	5.19	2.01	0
2	1	55.539	0	12.37	5.87	1.11
3	1	59.292	0	9.67	7.12	1.56
4	1	55.422	0	10.95	7.35	1.77
5	1	60.885	0	7.37	0	1.98

表 5.17 高钾 下调 10%

	聚类种类	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)
0	1	76.263	0	9.99	6.32	0.87
1	2	95.755	0	5.19	2.01	0
2	1	67.881	0	12.37	5.87	1.11
3	1	72.468	0	9.67	7.12	1.56
4	1	67.738	0	10.95	7.35	1.77
5	1	74.415	0	7.37	0	1.98

铅钡的上调 10%和下调 10%结果见附录 4。

经过每个类型的上次操作调整出来的数据，再代入到聚类分析得到的分类结果中[7]，发现与第二小问的结果差别不大，说明该模型的敏感度低，模型十分稳定，不会因为其中一个数据的急剧改变而造成结果的变化。

5.3 问题 3 的建模与求解

1)模型的建立

由于问题二已经建立了随机森林和 K-means 模型，所以可以直接使用问题二的模型。

2）结果分析

将表单 3 中的数据代入问题二中的随机森林模型，得到如下预测结果（缺失化学成分见附录）

表 5.18 随机森林预测结果

预测结果_Y	预测结果概率_铅钡	预测结果概率_高钾	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)
高钾	0.08	0.92	78.45	0	0	6.08	1.86	7.23
铅钡	0.82	0.18	37.75	0	0	7.63	0	2.33
铅钡	0.86	0.14	31.95	0	1.36	7.19	0.81	2.93
铅钡	0.92	0.08	35.47	0	0.79	2.89	1.05	7.07
铅钡	0.9	0.1	64.29	1.2	0.37	1.64	2.34	12.75
高钾	0.12	0.88	93.17	0	1.35	0.64	0.21	1.52
高钾	0.16	0.84	90.83	0	0.98	1.12	0	5.06
铅钡	1	0	51.12	0	0.23	0.89	0	2.12

3）亚类划分

根据问题二中的亚类划分方式可以对数据进行再次细分，得到如下结果

表 5.19 亚类划分结果

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
高钾/铅钡	高钾	铅钡	铅钡	铅钡	铅钡	高钾	高钾	铅钡
亚类	硅钾玻璃	硅铅玻璃	硅铅玻璃	硅铅玻璃	硅钾刚玻璃	硅钾玻璃	硅钾玻璃	硅钾刚玻璃

4）敏感性分析

选取某一特征元素，将其特征元素的化学成分含量在原有数据基础上上下变动 5%和 10%，对变动后的结果与原数据进行比对，给出其敏感性。下调整百分之十（缺失部分参见附录 9）

表 5.20 敏感性分析预测

预测结果_Y	预测结果概率_铅钡	预测结果概率_高钾	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)
高钾	0.11	0.89	70.6	0	0	6.08	1.86
铅钡	0.82	0.18	33.98	0	0	7.63	0
铅钡	0.86	0.14	28.76	0	1.36	7.19	0.81
铅钡	0.92	0.08	31.92	0	0.79	2.89	1.05
铅钡	0.91	0.09	57.86	1.2	0.37	1.64	2.34
高钾	0.14	0.86	83.85	0	1.35	0.64	0.21
高钾	0.18	0.82	81.75	0	0.98	1.12	0
铅钡	1	0	46	0	0.23	0.89	0

再经过上下浮动百分之 5 之后得到如下结果

表 5.21 预测结果对比

0.1	0.05	原数据	-0.05	-0.1
高钾	高钾	高钾	高钾	高钾
铅钡	铅钡	铅钡	铅钡	铅钡
铅钡	铅钡	铅钡	铅钡	铅钡
铅钡	铅钡	铅钡	铅钡	铅钡
铅钡	铅钡	铅钡	铅钡	铅钡
高钾	高钾	高钾	高钾	高钾
高钾	高钾	高钾	高钾	高钾
铅钡	铅钡	铅钡	铅钡	铅钡

5）结果分析

由图表可以知道在局部范围内改变数据时候模型的结果并无太大变化，所以模型有较高的稳定性，敏感性较低。

5.4 问题 4 的建模与求解

1）模型的建立

计算协方差（用于引出相关系数的定义）

若有两组数据 X:$\{X_1, X_2, \cdots, X_n\}$ 和 Y:$\{Y_1, Y_2, \cdots, Y_n\}$ 是总体数据

那么总体均值[8]

$$E(X)=\frac{\sum_{i=1}^{n} X_i}{n}, \quad E(Y)=\frac{\sum_{i=1}^{n} Y_i}{n}$$

总体协方差

$$Cov(X,Y)=\frac{\sum_{i=1}^{n}(X_i-E(X))(Y_i-E(Y))}{n}$$

$E(X)$ 是 X 组数据的均值[9]； $E(Y)$ 是 Y 组数据的均值

由此引出总体 Pearson 相关系数：

$$\rho_{XY}=\frac{Cov(X,Y)}{\sigma_X \sigma_Y}=\frac{\sum_{i=1}^{n}\frac{(X_i-E(X))}{\sigma_X}\frac{(Y_i-E(Y))}{\sigma_Y}}{n}$$

$\sigma_X(sigmaX)\frac{x-\mu}{\sigma}$ 是 X 的标准差

$$\sigma_X=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\mathrm{E}(X))^2}{n}}, \quad \sigma_Y=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\mathrm{E}(Y))^2}{n}}$$

可以证明，$|\rho_{XY}|\le 1$，且当 $Y=aX+b$ 时，$\rho_{XY}=\begin{cases}1, a>0\\ -1, a<0\end{cases}$

样本 Pearson 相关系数

两组样本数据 X:$\{X_1, X_2, \cdots, X_n\}$ 和 Y:$\{Y_1, Y_2, \cdots, Y_n\}$

样本均值

$$\bar{X}=\frac{\sum_{i=1}^{n}X_i}{n},\quad \bar{Y}=\frac{\sum_{i=1}^{n}Y_i}{n}$$

样本协方差

$$Cov(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

样本 Pearson 相关系数

$$\gamma_{XY}=\frac{Cov(X,Y)}{S_X S_Y}$$

其中：$S_X(sigmaX)$是 X 的样本标准差

$$S_X=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}},$$

同理

$$S_Y=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}$$

3）模型的求解

将上述公式代入附录 4.1 的代码中得到下面的结果（缺失部分参见附录）

表 5.22 高钾玻璃的化学成分之间的关联关系

	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
氧化纳	0.362									
氧化钾										
氧化钙										
氧化镁				0.417		0.45	0.293			

表 5.23 铅钡玻璃的化学成分之间的关联关系

	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)	氧化铅 (PbO)	氧化钡 (BaO)
氧化钠 (Na2O)	0.362							
氧化镁 (MgO)				0.417				
氧化铝 (Al2O3)	0.401		0.306		0.45			
氧化铁 (Fe2O3)				0.387	0.293			

将上面两个表格中的数据构造出两个对称的相关系数矩阵，并代入到附录 13 中的代码，得到结果

4）结果描述

玻璃制品中的大部分化学成分都与其他一两个化学成分有明显的关联性。氧化钡与二氧化硫之间的关联性较强，氧化钙与其他化学成分有明显关联性的数量较多。其中在高钾玻璃中，氧化铜与氧化钡的关联性很强，并且氧化铝与氧化锡存在明显的关联性；在铅钡玻璃中，氧化镁与氧化铁存在明显的关联性。其他化学成分在两种玻璃类型中的差别不大，并且关联性相差不大。

六．模型优缺点分析

6.1 问题一的分析

优点： 1、采用柱状图进行可视化分析，对数据的有更加直观的了解。

2、多因素方差分析不受比较组别的限制，具有良好的相互独立性。

3、独立样本 t 检验可以比较不同定类变量之间的差异性。

缺点： 1、数据的正态分布拟合不是很好，数据大于 30，利用大样本补全

2、独立样本 t 检验忽略数据的细微差别，不够精确。

6.2 问题二的分析

优点：1、随机森林非常灵活，并且它的准确性较高。

2、与单个决策树相比，随机森林在较大范围的数据项上的效果非常好。

3、可以克服过度拟合的问题。

缺点： 1、算法整体比较复杂，耗时严重且占用空间大。

2、进行聚类分析时可能聚类成链状，导致数据很乱。

6.3 问题三的分析

优点： 1、运用了问题二中构造的模型，让预测结果与前文有更好的关联性。

2、在该问题中对玻璃分类还进行了亚类分类，使得结果更加精准。

3、在分析敏感性中，运用了占比排名第三但是数据充足的二氧化硅成分，避免了运用其他变量却由于空值较多影响结果。

缺点： 1、运用问题二构造的模型，若问题二的步骤错误，则问题三的预测结果基本也错误。

2、在对敏感性分析时，只改变了二氧化硅含量的比重数值，可能存在一定的预测误差值。

6.4 问题四的分析

优点： 1、计算斯皮尔曼，并运用了相关系数矩阵，得到的相关度可信度较高。

2、分别对两种玻璃类别进行了相关系数分析，准确性较高。

缺点： 1、得出的表格数据结合自己的主观进行判断，可能主观性较强。

2、数据含量成分较多，影响结果的准确性。

七．模型的推广与改进

本文中所使用的独立样本 t 检验和聚类分析等算法要求数据满足正态分布，但是题目中的数据的正态拟合数据并不是特别完美，存在数据优化模型进步的可能。此外本文所使用的随机森林模型的训练样本太少，如果题目中可以给到足够的数据样本量，本文的随机模型会更加的完善与精准。

参考文献

[1] 宗序平，姚玉兰. 利用Q-Q图与P-P图快速检验数据的统计分布[J]. 统计与决策，2010(20):2.

[2] 倪安顺编著，Excel 统计与数量方法应用，北京：清华大学出版社，1998。

[3] 徐维超. 相关系数研究综述[J]. 广东工业大学学报,2012,29(3):12-17。

[4] 姜启源.《数学模型（第二版）》.高等教育出版社，1993。

[5] 司守奎.《数学建模算法与程序》. 海军航空工程学院，2007。

[6] 丁丽娟，数值计算方法，北京：北京理工大学出版社，1997。

[7] 高惠璇，应用多元统计分析，北京：北京大学出版社，2006。

[8] 杨虎，刘琼荪，钟波编著，数理统计，北京：高等教育出版社，2004。

[9] 王惠文著，偏最小二乘回归方法及其应用，北京：国防工业出版社，2000。

附 录

附录 1：1.3.1 代码

```
#include <iostream>
#include <algorithm>
#include <cstring>
#include <vector>
using namespace std;
const int N = 100;
double a[N];  // 数据数组
double b[N];
int main(void)
{
    /*
        求出风化前后的均值
    */
    memset(a,0,sizeof a);    //清空数组
    while(1) // 控制死循环
    {
```

```
        cout << "请输入未风化数据样本数量： " ;
        int n;          //输入样本数量
        cin >> n;
        cout << "请输入数据样本编号：" ;
        int flag;
        cin >> flag;
        double sum = 0;
        double x ; // 记录中间累加值
        for(int i = 0; i <  n ; i ++)
        {
            cout << "请输入样本的值：" ;
            cin >> x;
            sum += x; // 求出数据多次采样的 Sum 和
        }
        sum /= n; // 求出数据多次采样的均值
        a[flag] = sum;
        cout << "当前数据均值为 :  " << sum << endl;
        cout <<"如果以得到所需要数据，并且想退出请输入 1，输入别的数据则继续进行";
        int mm;
        cin >> mm;
        if(mm == 1) break;
    }
    /*
    ---得到了风化前的所有指标
    */
    memset(b,0,sizeof b);    //清空数组
    while(1) // 控制死循环
    {
```

```
        cout << "请输入风化数据样本数量： " ;
        int n;          //输入样本数量
        cin >> n;
        cout << "请输入数据样本编号：" ;
        int flag;
        cin >> flag;
        cout << "请输入样本的值：" ;
        double sum = 0;
        double x ; // 记录中间累加值
        for(int i = 0; i <  n ; i ++){
        cin >> x;
        sum += x; // 求出数据多次采样的 Sum 和
    }
        sum /= n; // 求出数据多次采样的均值
        b[flag] = sum;
        cout << "当前数据均值为 :  " << sum << endl;
        cout <<"如果以得到所需要数据，并且想退出请输入 1，输入别的数据则继续进行";
        int mm;
        cin >> mm;
        if(mm == 1) break;
    }
    /*
    ---得到了风化后的所有指标
    */

    /*
    --- 计算 C 的值
    */
```

```
    double c[100];
    memset(c,0,sizeof c);
    cout <<"输入样本数量" << endl;
    int n;
    cin >> n;
    for(int i = 0 ; i < n ; i ++)
    {
        c[i] = (a[i] - b[i]) / b[i]; // 得到每一个成分的 C[i]
    }
    for(int i = 0 ; i < n; i ++)
    {
        if(!c[i]) cout <<"编号为 ：" << i <<" 比率为 ：" << i * 100 <<"%"
<< endl;
    }
    return 0;
}
```

附录 2：1.3.2 代码

```
#include <iostream>
using namespace std;
/*
--- 利用  A = C*B + B 得出风化前的数据
*/
int main(void) {
    while(1) {
        double a b c;
        cout <<"请输入元素名称 :" << endl;
        string flag;
        cin >> flag;
        cout <<"请输入 B，C :"<<endl;
        cin b >> c;
```

```
            a = c * b + b;
            cout    <<flag<<"元素的风化前的含量为:" << a << endl;
        }
        return 0;
}
```

附录 3：1.3.2 分析结果请参考支撑材料

附录 4：2.3 高钾敏感性请参考支撑材料

附录 5：TOPSIS 分析结果请参考支撑材料

附录 6：独立样本 t 检验结果请参考支撑材料

附录 7：高钾聚类分析结果请参考支撑材料

附录 8：铅钡聚类分析结果请参考支撑材料

附录 9：问题三敏感性分析结果请参考支撑材料

附录 10：问题四关联分析结果请参考支撑材料

附录 11：问题一预测风化前成分含量分析结果请参考支撑材料

附录 12：主成分分析结果请参考支撑材料

附录 13：随机森林代码、相关系数矩阵、主成分分析代码请参考支撑材料