基于玻璃分类体系的构建对古代玻璃制品的成分分析与鉴别

摘　　要

古代玻璃在风化过程中，由于内部元素与环境元素进行大量交换，导致其成分比例发生变化，使得难以进行类别判断。本文通过建立风化前后化学成分含量变化模型、聚类分析模型、针对玻璃亚分类的支持向量机模型构建玻璃分类体系，实现对玻璃的准确分类。

针对问题一：针对表 1 的数据进行**卡方检验**，得到玻璃表面是否风化与玻璃纹饰以及玻璃颜色相关性较弱，与玻璃类型相关性显著这一结论。筛选主要化学成分，并对其进行 K-S **检验**，验证其是否符合正态分布。在正态分布的基础上，运用3σ准则，对成分含量的分布规律进行描述，并基于正态分布的规律，得到风化前后化学成分含量转化公式，建立**风化前后化学成分含量变化模型**。运用该模型预测了所有风化点在风化前的化学成分含量。

针对问题二：通过对比平均值，对高钾玻璃和铅钡玻璃的分类规律进行进一步描述。对未风化的不同类别的玻璃样本，利用系统聚类建立针对玻璃样品亚分类的**聚类分析模型**，实现不同类别玻璃的亚类划分。运用该模型，铅钡玻璃样品可分为四类，分别是高铅低钡系统、低铅低钡系统、低铅高钡系统、高铅高钡系统；高钾玻璃样品可分为四类，分别是无钾系统、钾硅系统、钾钙系统、钾铝系统。对模型进行敏感性检验，用对数据进行不同程度干扰的方法，对主要指标进行干扰，铅钡玻璃亚分类准确率为 91.30%，高钾玻璃亚分类准确率为 100%；对次要指标进行干扰，铅钡玻璃亚分类准确率、高钾玻璃亚分类准确率均为 100%。

针对问题三：利用高钾玻璃和铅钡玻璃的分类规律，将未知类别文物分为高钾、铅钡两类。运用**纠错输出编码**，将多分类问题转化为二分类问题，建立针对玻璃亚分类问题的**支持向量机模型**进一步对文物进行亚类鉴别。高钾玻璃中 A1 为亚类中的类别 1，A6、A7 为亚类中的类别 3；铅钡玻璃中 A2、A3 为亚类中的类别 1，A4、A5、A8 为亚类中的类别 2。

针对问题四：运用**斯皮尔曼相关系数**对其化学成分的关联程度进行判断。在铅钡玻璃中二氧化硅与氧化铅负相关性显著；在高钾玻璃中二氧化硅与氧化钾、氧化钙、氧化铝、氧化铁负相关性显著，相关系数的绝对值均大于 0.7。结合文献资料对相关性和差异性进行分析，得到的结论是：化学成分之间具有相关性主要是因为玻璃制作和风化过程中的化学反应；产生差异性的原因则是助熔剂不同导致的。

关键词：　卡方检验　聚类分析　支持向量机　斯皮尔曼相关系数

一、问题重述

玻璃的主要原料是石英砂，主要化学成分是二氧化硅。为降低其熔点，炼制时需添加助熔剂。由于助熔剂类型不同，使得主要化学成分不同。根据其化学成分常把玻璃分为高钾玻璃和铅钡玻璃。同时古代玻璃极易受埋藏环境的影响而风化从而导致其成分比例发生变化。现有一批我国古代玻璃制品的相关数据，解决以下问题：

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

问题 2：依据附件数据分析高钾玻璃、铅钡玻璃的分类规律；对于每个类别选择合适的化学成分对其进行亚类划分 4，给出具体的划分方法及划分结果，并对分类结果的合理性和敏感性进行分析。

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

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

二、问题分析

问题一的分析：首先对附件中的数据进行预处理，利用题中所给条件对数据进行简单清洗。针对文物表面有无风化和其他玻璃属性关系的分析，由于玻璃类型、纹饰等都属于分类变量，故我们采用卡方检验对其进行分析。对文物表面有无风化化学成分含量以及根据风化点数据推测未风化数据时，我们建立了风化前后化学含量变化模型。先运用化学元素占比对指标进行进一步筛选，再运用 K-S 检验发现其统计规律，并根据其统计规律预测风化点风化前的化学成分含量。

问题二的分析：在问题一中对化学成分含量的统计规律的基础上，进一步研究其平均值等统计量，对高钾玻璃和铅钡玻璃的分类规律进行进一步描述。对每一类玻璃进行亚分类时，首先将化学成分指标分为主要指标和次要指标，建立针对玻璃的聚类分析模型，并对模型进行合理性分析，实现玻璃的亚类划分。对于模型的敏感性分析，我们采用了两种不同分析方式，分别针对主要指标和次要指标进行数值改变探究其分类的变化。

问题三的分析：问题三我们将分为两步进行求解。第一步是将样品进行粗分类，针对表 3 中的数据，利用玻璃类型差异性特征，把玻璃样品分为铅钡和高钾两类。第二步是对两种不同类比的玻璃进行亚类细分。首先由于亚类是仅针对未风化样品的亚分类，所以我们需要利用问题一的风化前后化学含量变化模型，将已风化玻璃化学成分含量数据转化为风化前的化学成分含量。之后建立多分类的支持向量机模型，对样品的亚类进行进一步甄别。

问题四的分析：将样品分为高钾和铅钡两类，分别对其进行相关性显著检验，运用斯皮尔曼相关系数对得出其化学成分的关联程度。结合相关文献分别对同类之间的关联性和不同类别之间化学成分的差异性进行分析。

三、模型假设

1、假设文物样品之检测过程相互独立；

2、假设检测过程中文物样品表面化学成分不发生变化。

四、符号说明

符号	说明
df	卡方检验自由度
K^2	卡方值
$d(x_i, x_j)$	聚类分析中的距离
r_s	斯皮尔曼相关系数

五、模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 数据预处理

首先，我们需要对附件中表 1 表 2 的数据进行数据清洗。根据题中所给条件，我们将各个样品的成分比例进行累加，其中介于 85%-105%的数据为有效数据，根据这一条件可将表 2 中文物编号为 15 和 17 的数据删除。

针对表 1 中部分样品颜色存在缺失，在研究表面风化与颜色关系时，删除缺失数据。

结合表 1 与表 2，当表 2 中检测的样品为未风化样品的已风化点时作为已风化样品处理，若未表明是否风化则与表 1 类型保持一致。

5.1.2 表面风化和其他指标的卡方检验

针对玻璃文物表面是否风化与玻璃类型、纹饰和颜色关系的分析时，我们需要分析其差异性和相关性。表 1 中是否风化、玻璃类型、纹饰、颜色都属于分类变量。对于类别与类别型变量的数据，我们采用针对分类变量的卡方检验对其相关系进行分析。

1. 卡方检验

卡方检验是统计样本实际观测值与理论推测值之间的偏离度的统计方法。其具体操作方法如下：

（1）确定原假设

H_0：两样本相互独立。

（2）统计频数列联表

对于两个分类变量X和Y，它们的值域分别为$\{x_1, x_2\}$和$\{y_1, y_2\}$。

得到样本频数列联表为：

表 1　样本频数列联表示意

	y_1	y_2	总计
x_1	a	b	$a+b$
x_2	c	d	$c+d$
总计	$a+c$	$b+d$	$a+b+c+d$

以表面是否风化与玻璃类型之间的关系为例，构建样本频数列联表如下所示：

表 2　表面是否风化与玻璃类型频数列联表

	风化	未风化	总计
高钾	6	10	16
铅钡	28	12	40
总计	34	22	56

（1）计算卡方值

$$K^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{1}$$

其中$n=a+b+c+d$为样本容量。

（2）计算自由度

$$df=(\text{行数}-1)(\text{列数}-1) \tag{2}$$

（3）得到p值

当p值大于 0.05 时，接受原假设；反之则拒绝原假设。

2. 模型求解

将风化与玻璃类型，风化与纹饰，风化与颜色分别进行卡方检验，得到结果如下：

表 3　表面风化与其他指标卡方检验结果

	玻璃类型	纹饰	颜色
p值	0.0245	0.0845	0.4277

当p值大于 0.05 时接受原假设，即表面是否风化与纹饰以及颜色相关性弱；当p值小于 0.0 时拒绝原假设，即表面是否风化与玻璃类型相关性显著。我们可以得出结论：玻璃的风化程度只与玻璃类型有关也就是与玻璃的化学成分含量有关。

根据表 2 对不同玻璃类型风化数量频数的统计，我们可以发现高钾玻璃未风化量大于风化量。由此我们可以初步推断铅钡玻璃比高钾玻璃更容易风化。研究表明，铅钡玻璃中高含量的铅与周围环境中的二氧化碳、水蒸气等反应形成大量氧化铅，使得玻璃更容易风化；而高钾玻璃中二氧化硅含量高，由于硅氧四面体相互连接程度大，化学稳定性好，同时玻璃成分中的氧化铝形成了铝氧四面体，对硅氧网络起到补充作用，使得高钾玻璃抗风化能力远远大于铅钡玻璃。

5.1.3 建立风化前后化学成分含量变化模型

1. 模型建立

（1）指标的初步筛选

在研究文物样品表面有无风化化学成分含量统计规律以及文物风化前后化学成分含量变化时，由于化学成分多并且部分化学成分在特定种类玻璃中存在明显缺失，给接下来的研究带来巨大的困难。所以我们将化学成分占比（含该元素的样品数除以总样品数）少于 50%的化学成分去除。通过该方法只保留主要化学成分，从而实现指标降维的目的。

经过筛选，高钾玻璃的有效指标为二氧化硅、氧化钾、氧化钙、氧化铝、氧化铁、氧化铜、五氧化二磷七个；铅钡玻璃的有效指标为二氧化硅、氧化钙、氧化铝、氧化铜、氧化铅、氧化钡、五氧化二磷、氧化锶八个。

（2）K-S 检验

在探究化学成分的分布特征时，我们先假设所有指标都服从正态分布，运用直方图和 Q-Q 图对其进行初步判断，以铅钡玻璃的二氧化硅为例，其直方图和 Q-Q 图如下：

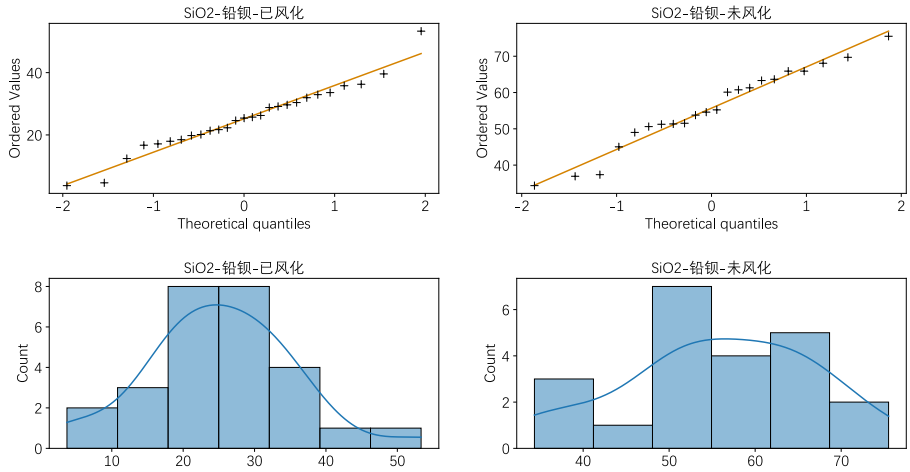

图 1　部分化学成分直方图和 Q-Q 图

接下来我们用 K-S 检验对各化学成分是否服从正态分布进行进一步分析。K-S 检验是检验单个样本是否来自某一特定分布的方法，其检验方式如下：

1、建立原假设：

H_0：构成样本的总体服从正态分布。

2、定义公式如下：$D = \max|F_n(x) - F_0(x)|$

其中，$F_0(x)$表示理论分布的分布函数，本文中为正态分布函数$X \sim N(\mu, \sigma^2)$；$F_n(x)$表示一组随机样本的累计频率函数；D为$F_0(x)$与$F_n(x)$差距的最大值。

3、当$D > D(n, \alpha)$则接受原假设，反之则拒绝原假设。

2. 模型的求解

（1）各成分的分布情况

在对已筛选过的指标进行 K-S 检验后，得到检验结果如下：

表 4　高钾玻璃化学成分 K-S 检验

化学成分	二氧化硅	氧化钾	氧化钙	氧化铝	氧化铁	氧化铜	五氧化二磷
p（已风化）	0.7138	0.8715	0.9766	0.9581	0.9861	0.5735	0.9615
p（未风化）	0.7187	0.7930	0.7880	0.6802	0.6452	0.9697	0.0883

表 5　铅钡玻璃化学成分 K-S 检验

化学成分	二氧化硅	氧化钙	氧化铝	氧化铜	氧化铅	氧化钡	氧化锶	五氧化二磷
p（已风化）	0.9317	0.7889	0.1839	0.0136	0.9540	0.1086	0.7993	0.7909
p（未风化）	0.9220	0.1605	0.3583	0.0250	0.5517	0.4049	0.5962	0.0161

由上表可知，只有铅钡玻璃氧化铜成分和未风化的铅钡玻璃的五氧化二磷成分不服从正态分布，其余均服从正态分布。这可能是因为氧化铜多作为着色剂使用使玻璃呈色，而在地下埋藏过程中玻璃受各种因素的影响使得表面形成腐蚀层，导致五氧化二磷含量发生变化。

根据正态分布这一结论我们可以得到化学成分的分布规律。由3σ原则可以得到服从正态分布的化学成分的范围。

表 6　高钾化学成分含量的分布规律

表面风化	二氧化硅	氧化钾	氧化钙	氧化铝	氧化铁	氧化铜	五氧化二磷
无风化	(41.7,94.3)	(0,21.1)	(0,14.6)	(0,14.1)	(0,7.0)	(0,7.4)	(0,5.7)
风化	(88.8,99.2)	(0,1.9)	(0,2.3)	(0,4.8)	(0.1,0.5)	(0.4.4)	(0,0.9)

表 7　铅钡化学成分含量的分布规律

表面风化	二氧化硅	氧化钙	氧化铜	氧化铅	氧化钡	氧化锶	氧化铝	五氧化二磷
无风化	(22.7,88.7)	(0,5.3)	(0,4.9)	(0,46.5)	(0,21.9)	(0,3)	(0,14.4)	(0,6.7)
风化	(0,56.6)	(0,7.7)	(0,11.5)	(5.9,79.7)	(0,42.9)	(0,1.3)	(0,10.7)	(0,17.78)

（2）风化前的成分推测

由于大部分化学成分在风化前后都服从正态分布规律，我们可以由正态分布的性质对风化点风化前的数据进行预测。方法如下：

$$\frac{x-\mu_{\text{after}}}{\sigma_{\text{after}}}=\frac{y-\mu_{\text{before}}}{\sigma_{\text{before}}} \tag{3}$$

其中μ_{after}为风化后正态分布的均值，σ_{after}为风化后正态分布的均方差，μ_{before}为风化前正态分布的均值，σ_{before}为风化前正态分布的均方差，x为风化后各化学成分的含量，y为待预测的化学成分的含量。

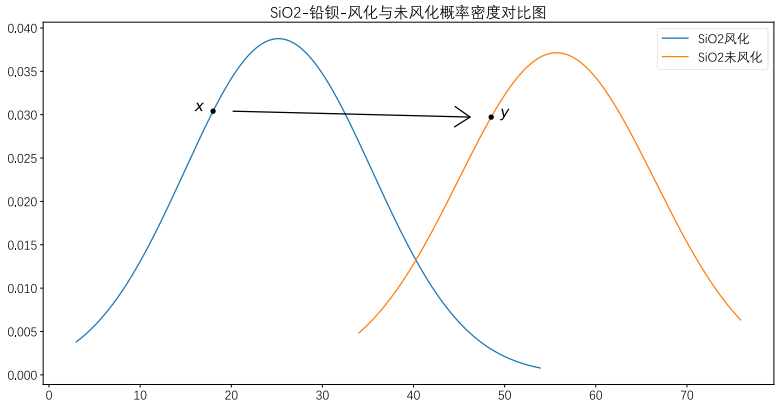

图 2　铅钡玻璃风化前后概率密度对比图

对不符合正态分布的元素，我们采用计算风化前风化后平均值的差值的方式，对该种元素进行预测，公式如下：

$$y=x+\left(\bar{x}_{before}-\bar{x}_{after}\right) \tag{4}$$

对次要元素，我们认为其风化前后含量不变。

在进行指标数值转换后，将转换后的指标化为数据百分比的形式得到预测结果，部分预测结果如下表所示（完整预测结果见附件）。

表 8　风化点的化学成分预测

类型	高钾			铅钡		
文物编号	7	9	10	11	19	24
二氧化硅	69.41	73.76	80.02	66.91	60.06	64.77
氧化钠	0.00	0.00	0.00	0.00	0.00	0.00
氧化钾	5.16	9.80	12.32	0.22	0.00	0.00
氧化钙	7.48	3.77	1.12	2.13	1.59	0.00
氧化镁	0.00	0.00	0.00	0.74	0.59	0.00
氧化铝	7.65	5.07	3.63	4.46	5.38	3.00
氧化铁	0.00	3.27	1.76	0.00	1.32	0.00
氧化铜	6.15	2.44	1.14	2.21	1.53	3.73
氧化铅	0.00	0.00	0.00	10.45	21.67	13.00
氧化钡	0.00	0.00	0.00	9.58	5.06	14.89
五氧化二磷	4.15	1.89	0.00	3.11	2.74	0.00
氧化锶	0.00	0.00	0.00	0.20	0.06	0.61
氧化锡	0.00	0.00	0.00	0.00	0.00	0.00
二氧化硫	0.00	0.00	0.00	0.00	0.00	0.00

5.2 问题二模型的建立与求解

5.2.1 高钾玻璃与铅钡玻璃分类规律的描述

为探究高钾和铅钡玻璃的分类规律，首先我们将玻璃分成已风化和未风化并计算出高钾玻璃和铅钡玻璃各元素的平均值。问题一中我们已将不同类型玻璃的化学成分分为主要化学成分和次要化学成分，对次要化学成分我们不再进行研究；对于主要化学成分，又因为不同类型玻璃主要成分不同，我们先对其共有的化学成分进行研究，再研究特有成分的差异性。

首先，绘制不同类型玻璃的共有元素对比图，如图 3，图 4 所示。

从图中我们可以看出，对于已风化的玻璃，铅钡玻璃二氧化硅含量占比远远小于高钾玻璃二氧化硅含量占比，而五氧化二磷则相反。但根据图 4，两类玻璃二氧化硅含量的均值和五氧化二磷含量的均值相差并不明显。对于未风化高钾玻璃氧化钙的含量远大于未风化铅钡玻璃氧化钙的含量，氧化铝和氧化铜的含量也明显多于未风化铅钡玻璃同种化学成分的含量，在风化后这些成分明显变少，而高钾玻璃减少更剧烈，可能是某种化学成分剧烈增加导致其相对含量降低或是参与了某些化学反应产生了消耗。

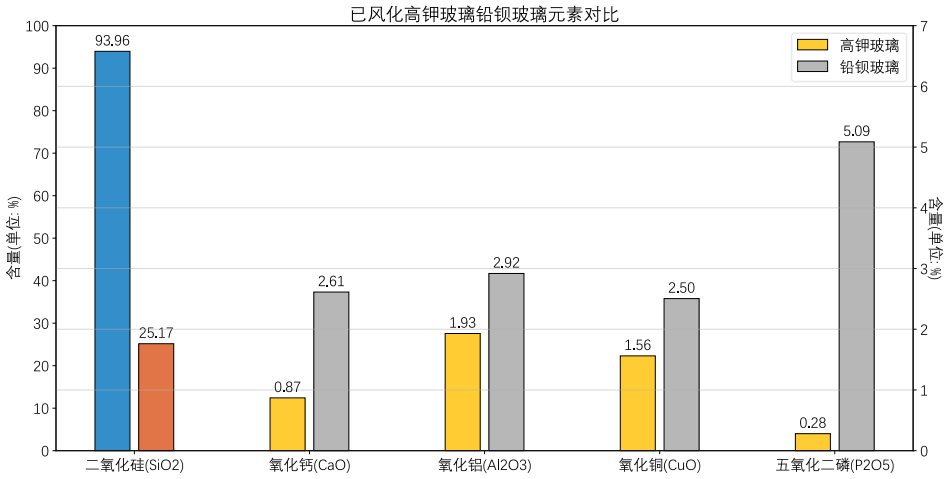

图 3　已风化高钾玻璃铅钡玻璃元素对比

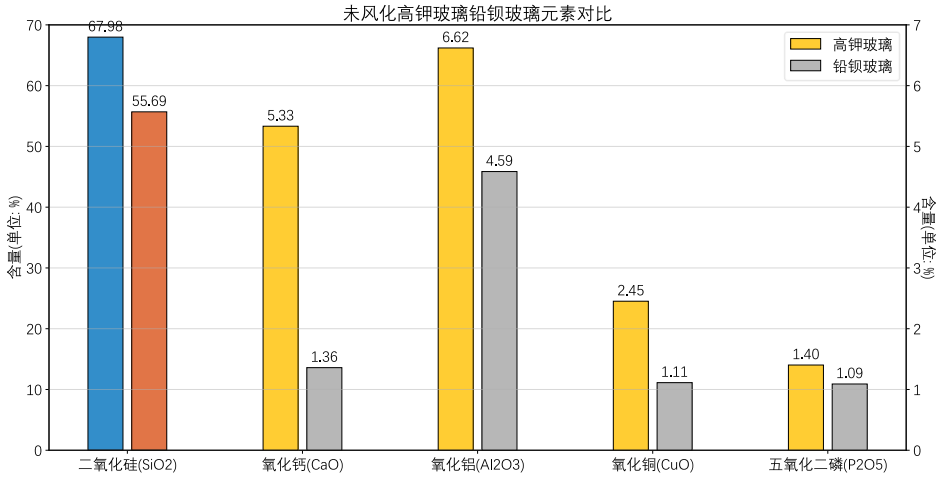

图 4　未风化高钾玻璃铅钡玻璃元素对比

接着研究不同类别玻璃特有成分，我们根据是否风化列出两个表格如下：

表 9　已风化不同类型玻璃特有化学成分含量

类型	铅钡			高钾	
化学成分	氧化铅	氧化钡	氧化锶	氧化钾	氧化铁
平均值	42.79	8.22	0.49	0.82	0.27

表 10　未风化不同类型玻璃特有化学成分含量

类型	铅钡			高钾	
化学成分	氧化铅	氧化钡	氧化锶	氧化钾	氧化铁
平均值	21.76	8.22	1.50	10.18	2.32

其中铅钡玻璃的特有元素有氧化铅、氧化钡、氧化锶；高钾玻璃的特有元素有氧化钾、氧化铁。根据它们含量的不同可以很好的区分玻璃类型。

5.2.2 不同类型玻璃亚分类方法

在对不同类型玻璃进行亚类划分时，我们采用多元统计中广泛应用的聚类方法。以样品中的化学成分的含量为变量，以玻璃样品为事件进行聚类，从而实现亚类的划分。

1. 建立聚类模型

聚类是指按照某个特定的标准（如距离）把一个数据集分割成不同类别的方法。我们将要分类的所有玻璃样品称为总体G，每个玻璃样本数据记为$X_i = (x_{il}, \dots, x_{ip}), i = 1,2, \dots, n$，每个$X_i$称为一个样品。

有以下公式：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}, S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \tag{5}$$

$\bar{X}$称为玻璃样品均值矩阵，S为玻璃样品协方差矩阵，矩阵X称为玻璃样本数据矩阵。

在进行样品间距离的计算时，我们采用欧式距离和平均值距离计算，公式如下：

$$d(x_i, x_j) = \frac{1}{n_r n_s} dist(x_{ri}, x_{sj}) \tag{6}$$

其中n_r为第一个簇样本个数，n_s第二个簇样本个数。

对于未知类别数量的玻璃样品，采用系统聚类法对其进行积累分析，具体步骤如下：

（1）将容量为n的玻璃样本中每个玻璃样品看成一类，计算两两之间的平均距离；

（2）将所有距离小的合并成一个新类；

（3）重新计算新类与其他类的距离；

（4）重复步骤（2）（3）直至最后合并成一类。

2. 聚类数量的判断

由于聚类数量具有强烈的主观性，我们运用肘部法则对分类数量进行确定。

肘部法则是运用图像的方式确定最优聚类数量的方法。我们可以定义畸变程度这一概念，即每个类别距离其该类中心点的距离。

假设将n个样本划分到K个类中，用C_k表示第k个类$(k = 1,2, \dots, K)$,且该类中心位置记为u_k，那么第k个类的畸变程度为：

$$\sum_{i \in C_k} |x_i - u_k|^2 \tag{7}$$

定义总畸变程度：

$$J = \sum_{k=1}^{K}\sum_{i \in C_k} |x_i - u_k|^2 \tag{8}$$

通过肘部法则进行聚类数量的判断可以很好的减少主观因素对分类的干扰。

3. 聚类模型合理性检验

为检验模型合理性，我们将所有数据只进行风化和未风化的划分，将其带入上述聚类模型，将聚类结果与表 1 中所属类别进行对比。

我们发现若将玻璃样本的全部化学成分进行聚类分析，则得到的分类结果与题中所给出的分类相差过大。在对玻璃化学成分进行研究之后，我们发现由于二氧化硅为玻璃的主要元素，在成分含量中的占比巨大，故对分类结果有很大的影响。在剔除二氧化硅这一化学成分后，我们重新对全部类型玻璃样本进行聚类分析，得到聚类分析谱系图。

根据谱系图将聚类结果设定为 2 种，我们可以得出每一编号的所属类别，并与实际所属类别进行比较。发现只有风化的高钾玻璃样品 11 和风化的高钾玻璃样品 48 分类错误。

计算得到风化的高钾玻璃正确率为 93.75%，其余正确率为 100%。模型建立合理，具有较强可信度。

4. 亚分类方法

接着我们运用上述模型对未风化的高钾玻璃和未风化的铅钡玻璃进行亚类划分。首先基于上述模型我们得到聚类分析谱系图如下：

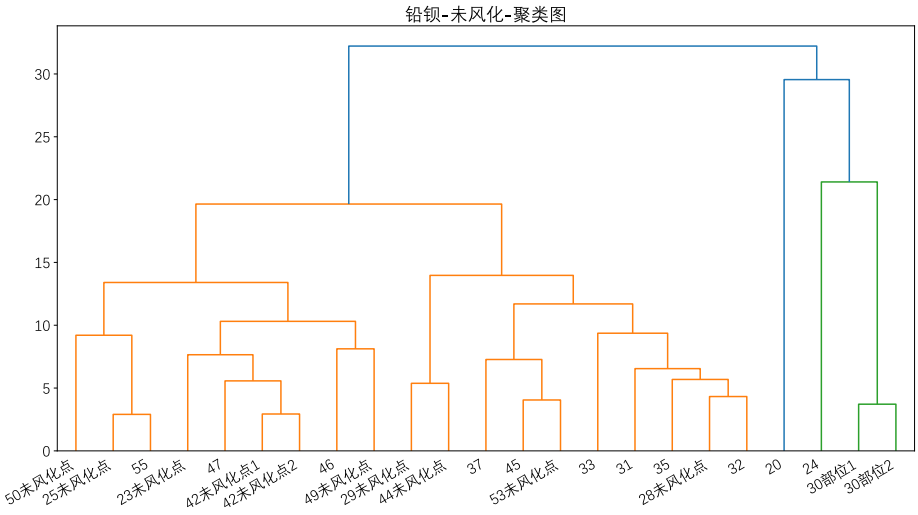

图 5　未风化的铅钡玻璃聚类分析谱系图

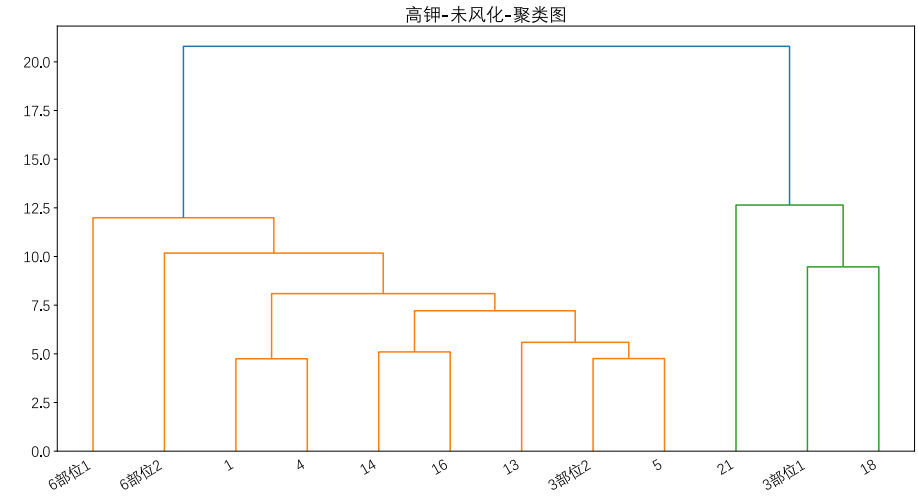

图 6　未风化的高钾玻璃聚类分析谱系图

并根据每个类别距离该类中心的距离得到距离图像，如图 7 和图 8 所示。

根据肘部法则我们可以得到未风化的铅钡玻璃可分为 4 类，未风化的高钾玻璃可分为 4 类。分类结果见表 11。

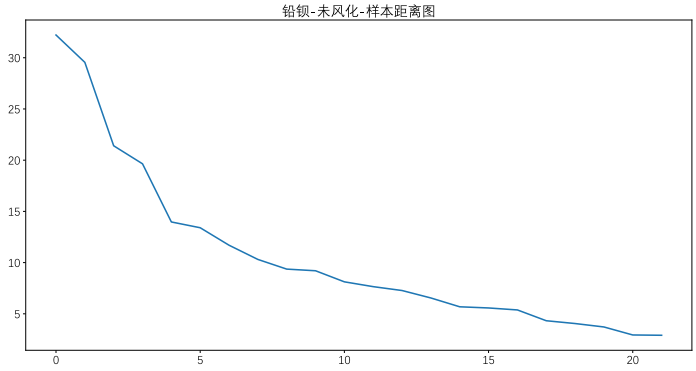

图 7　未风化的铅钡玻璃距离图

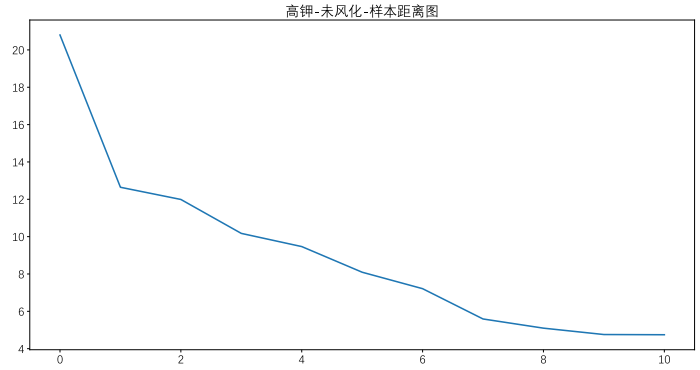

图 8　未风化的高钾玻璃距离图

表 11　未风化铅钡玻璃亚分类

类别	名称	编号
类别 1	高铅低钡系统	25 未风化点、30 部位 1、30 部位 2、46、47、50 未风化点、55
类别 2	低铅低钡系统	23 未风化点、28 未风化点、29 未风化点、31、32、33、35、37、42 未风化点 1、42 未风化点 2、44 未风化点、45、49 未风化点、53 未风化点
类别 3	低铅高钡系统	20
类别 4	高铅高钡系统	24

表 12　未风化高钾玻璃亚分类

类别	名称	编号
类别 1	无钾系统	21
类别 2	钾硅系统	3 部位 1、18
类别 3	钾钙系统	1、3 部位 2、4、5、13、14、16
类别 4	钾铝系统	6 部位 1、6 部位 2

在对玻璃类别进行划分时，由于不同玻璃所含元素种类不同，我们将主要影响玻璃分类的元素称为主要影响元素。其划分依据包括元素占比（含该元素的样品数除以总样品数）、文献查找等。根据查到的资料可知，氧化铁、氧化铜主要作为着色剂使玻璃呈色，不能作为亚分类依据；五氧化二磷含量高可能由于地下埋藏过程中形成了腐蚀层，而未受腐蚀的玻璃不含有该元素不能说明玻璃制作过程中曾用草木灰作为原料，故也不能作为分类依据；氧化

钠、氧化钾、氧化钙这三个元素虽也作为铅钡玻璃的助熔剂出现，但由于其含量低且氧化铅的相对含量远大于这三种元素之和，故我们认为我们所研究的铅钡玻璃不以它们为助熔氧化物，也作为次要影响元素。

基于上述分析，对于铅钡玻璃我们选取二氧化硅、氧化铝、氧化铅、氧化钡作为主要影响元素；对于高钾玻璃主要选取二氧化硅、氧化钾、氧化钙、氧化铝作为主要影响元素。

对于铅钡玻璃，我们将其分为四个亚类。

第一类高铅低钡系统，其主要特点是氧化铅含量大于 25%而氧化钡含量小于 20%，该类亚类玻璃的主要特点是铅含量高的同时二氧化硅含量相对较少，其化学性质不稳定容易被腐蚀。第二类为低铅低钡系统，其主要特点是氧化铅含量小于 25%并且氧化钡含量小于 20%，该亚类玻璃由于助熔剂的含量相对较少使二氧化硅成分占比增加，二氧化硅多使得化学性质相较其他亚类玻璃更加稳定，使玻璃能够长久保存，我们推测这是该类玻璃多于其他亚类玻璃的原因。第三类为低铅高钡系统，其主要特点为助熔剂以氧化钡为主，并且氧化钡的相对含量远大于一二类玻璃，二氧化硅含量少。第四类为高铅高钡系统，该类的主要特点是氧化铅和氧化钡的占比和远大于其他类从而导致二氧化硅含量过低。我们认为三四类样品少的原因就在于二氧化硅含量低使得玻璃制品易被腐蚀，难以长久保存。

对于高钾玻璃，我们将其分为四个亚类。

第一类为无钾系统，我们给该类玻璃命名为高硅无钾玻璃，主要特点是具有较高的二氧化硅含量但无氧化钾。第二类为钾硅系统，我们给该类命名为高硅低钾玻璃，主要特点是具有较高的二氧化硅含量但具有较少量的氧化钾。第三类为钾钙系统，与前两类不同，该类玻璃中含有较高的而氧化钙，证明氧化钾和二氧化钙同时起到助熔作用，该玻璃系统属于钾钙玻璃系统。第四类为钾铝系统，该类的主要特点为氧化钾含量明显小于第三类玻璃系统但含有大量二氧化铝。

5.2.3 分类结果敏感性分析

针对上述分类结果，我们采用两种方式对其进行敏感性检验。

方法一：改变主要化学成分的含量，观察分类变化。

方法二：改变次要化学成分的含量，观察分类变化。

（1）运用方法一进行敏感性分析

运用方法一对分类结果进行敏感性检验，得到分类准确度如下：

表 13　运用方法一的敏感性检验

	铅钡玻璃	高钾玻璃
准确度	91.30%	100%

根据上表我们可知，对主要化学成分进行小幅度扰动后，分类的准确率仍然较高，说明外界干扰对模型的影响程度小，模型稳定性强，同时不同类别有较好的区分度。

（2）运用方法二进行敏感性分析

运用方法二对分类结果进行敏感性检验，得到分类准确度如下：

表 14　运用方法二的敏感性检验

	铅钡玻璃	高钾玻璃
准确度	100%	100%

若仅对次要化学成分进行扰动，我们发现玻璃亚分类不发生变化。说明次要化学成分对分类不产生影响，不作为亚分类的分类依据，可以证明分类的合理性。

5.3 问题三模型的建立与求解

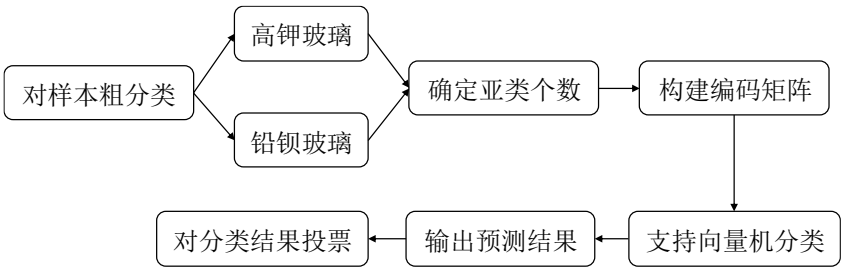

图 9　问题三流程图

5.3.1 未知类别玻璃文物的粗分类

根据问题二中对两种类型玻璃的特征分析，我们发现不同风化状态的不同类型玻璃都有其对应特征。对于已风化玻璃，由于不同玻璃类型二氧化硅含量差别巨大，我们可直接通过二氧化硅推测其所属类别。对于未风化玻璃，虽然二氧化硅含量的差别不明显，但我们可以通过其特有元素比如高钾玻璃中的氧化钾含量、铅钡玻璃中的二氧化铅、氧化钡的含量推断其所属类别。

根据规律，我们判断出玻璃类型如下表：

表 15　未知类别玻璃文物所属类别粗判断

	高钾	铅钡
风化	A6、A7	A2、A5
无风化	A1	A3、A4、A8

5.3.2 建立多分类支持向量机模型

（1）模型准备

①玻璃样品化学成分的筛选

在分类过程中，过多的化学指标会对分类产生干扰，所以在进行亚类鉴别前我们对化学进行筛选。问题二在对玻璃亚类进行鉴别时，我们已经筛选出了对亚类鉴别有重要作用的主要指标，所以在进行问题三的亚分类时我们只需要考虑主要指标即可。对于铅钡玻璃我们只保留二氧化硅、氧化铝、氧化铅、氧化钡四个指标；对于高钾玻璃只保留二氧化硅、氧化钾、氧化钙、氧化铝四个指标。

②将风化数据转化为风化前数据

由于问题二我们只针对未风化玻璃样本进行分类，因此针对表 3 中风化玻璃的数据，我们需根据它的所属类型，利用风化前后化学含量变化模型，将风化数据转化为风化前数据。

转化结果如下：

表 16　风化数据转化为风化前数据

文物编号	二氧化硅	氧化钾	氧化钙	氧化铝
A2	47.92	2.27	27.60	5.00
A5	63.47	6.90	15.92	11.44
A6	74.17	12.10	3.15	4.86
A7	61.79	6.56	7.52	8.38

（2）纠错输出编码 ECOC 法

为解决多分类问题中训练难度大等问题，我们采用纠错输出编码将玻璃亚分类这一多分类问题转化成多个二分类问题。

①确定分类个数K。问题二中将铅钡玻璃分为 4 个亚类，将高钾玻璃分为 4 个亚类，所以在进行纠错输出编码时K取 4。

②采用一对一分类方法，构建编码矩阵。构建的编码矩阵为：

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{bmatrix}$$

取其中一个分类为正，一个分类为负，其余分类为 0，行对应于类别，列对应于某种分类情况（分类器）。总共分类方法有C_k^2种，本题分类方法共 6 种。

（3）支持向量机模型的建立

支持向量机是一种二分类模型。

①定义超平面

与二维平面类似我们可以定义超平面方程：$w^T x + b = 0$。

②计算样本点到平面的距离

$$d = \frac{|w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b|}{\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}} = \frac{|W^T X + b|}{\|W\|} \tag{9}$$

其中$\|W\|$为超平面的范数。

距离超平面最近的点被称为支持向量，两个异类支持向量到超平面的距离之和称为间隔。

③支持向量机模型

找出所有间隔中最大值对应的超平面，即确定w，b使得距离最大。

$$\max_{w,b} \frac{2}{\|W\|} \tag{10}$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1, i = 1,2,\ldots,m$$

5.3.3 分类方法

（1）模型求解

①构建拉格朗日函数

$$L(w,b,\xi,\alpha,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum \alpha_i(\gamma_i(w \cdot x_i + b) - \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i \tag{11}$$

其中，$\alpha_i \geq 0, \mu_i \geq 0$。

②原始问题对偶化

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_i\gamma_i\gamma_i(x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \tag{12}$$

$$\text{s.t.}\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, i = 1,2,\ldots,N$$

③运用 SMO 算法求解

（2）分类结果

运用该模型我们对表 3 的数据进行分类结果如下：

表 17　未分类玻璃样品分类表

文物编号	粗分类	类别	名称
A1	高钾	类别 1	无钾系统
A2	铅钡	类别 1	高铅低钡系统
A3	铅钡	类别 1	高铅低钡系统
A4	铅钡	类别 2	低铅低钡系统
A5	铅钡	类别 2	低铅低钡系统
A6	高钾	类别 3	钾钙系统
A7	高钾	类别 3	钾钙系统
A8	铅钡	类别 2	低铅低钡系统

5.3.4 灵敏度分析

针对表 3 数据亚分类的结果，对表中样品化学成分含量进行不同程度的噪声干扰，我们可以得到分类准确度的变化。

表 18　灵敏度分析

扰动	准确度
1 倍	87.50%
3 倍	75%
5 倍	75%

当噪声干扰增加时，分类的准确度相应降低，但在当干扰小时准确度较高，说明模型稳定性好，当噪声干扰增加时准确度也相应降低，说明不同类别之间区分度明显。

5.4 问题四模型的建立与求解

在研究不同文物样品化学成分相关性时，由于不同化学成分不服从正态分布，我们采用斯皮尔曼相关系数对不同类别文物化学成分相关性进行分析。

斯皮尔曼相关系数是衡量两变量依赖性的指标，已知有X，Y两组数据，其计算公式如下：

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{13}$$

其中d_i为X_i和Y_i之间的等级差。

根据玻璃属性，我们分别对高钾玻璃和铅钡玻璃的化学成分进行分析得到其相关系数图。

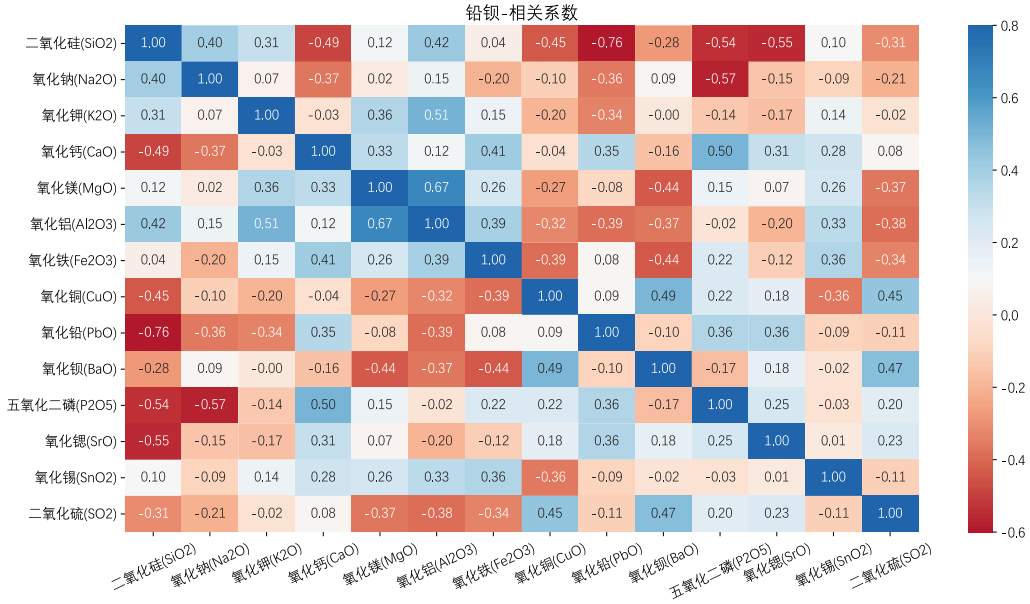

图 10　铅钡玻璃相关系数图

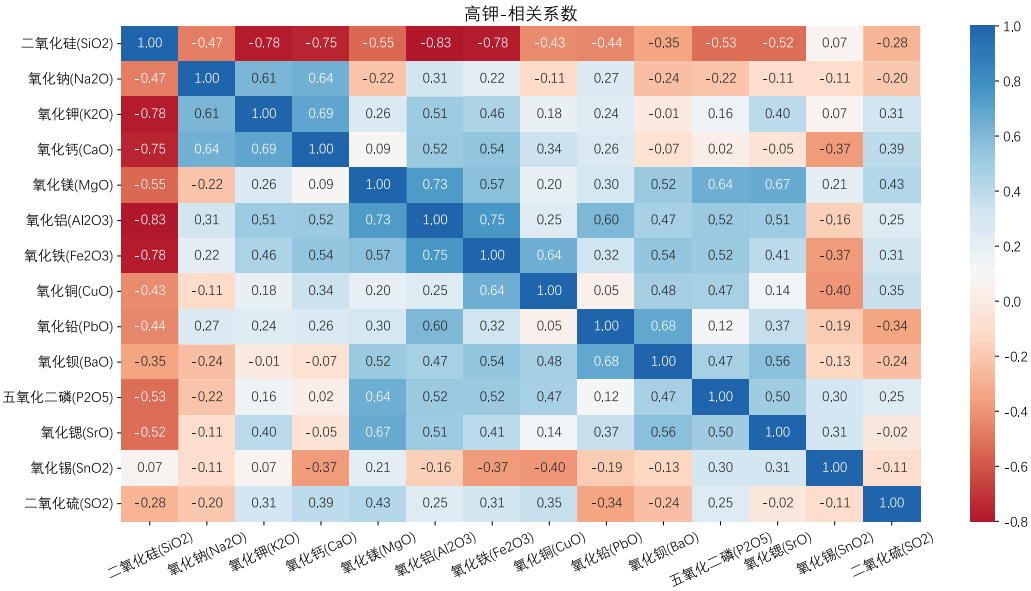

图 11　高钾玻璃相关系数图

根据上图所示，我们可得不同化学成分之间的相关性。

在铅钡玻璃中二氧化硅与氧化铅、五氧化二磷、氧化锶呈负相关；氧化钠与五氧化二磷呈负相关；氧化钾与氧化铝呈正相关；氧化钙与五氧化二磷呈正相关；氧化镁与氧化铝呈正相关。

在高钾玻璃中二氧化硅与氧化钙、氧化镁、氧化铝、氧化铁、五氧化二磷、氧化锶呈负相关；氧化钠与氧化钾、氧化钙呈正相关；氧化钾与氧化钙、氧化铝呈正相关；氧化钙与氧化铝、氧化铁呈正相关；氧化镁与氧化铝、氧化铁、氧化钡、五氧化二磷、氧化锶呈正相关；

氧化铝与氧化铁、氧化铅、五氧化二磷呈正相关、氧化锶呈正相关；氧化铁与氧化铜、氧化钡、五氧化二磷呈正相关；氧化铅和氧化钡呈正相关；氧化钡和氧化锶呈正相关。

在铅钡玻璃中氧化硅与氧化铅负相关性显著；在高钾玻璃中氧化硅与氧化钾、氧化钙、氧化铝、氧化铁负相关性显著。根据研究可知 Al^{3+}和 Fe^{2+}可以代替硅氧四面体中的 Si^{4+}，由此我们推断氧化铝和氧化铁的含量减少会导致氧化硅的含量增加。对于铅钡玻璃，常以氧化硅、氧化铅作为助熔剂，而氧化硅的增速远大于助熔剂的增速，所以氧化硅与助熔剂呈显著负相关。

与高钾玻璃相比铅钡玻璃化学成分相关性较低，我们推测可能是由于助熔剂不同所致。高钾玻璃助熔剂类型多为草木灰，由于草木灰中含有大量矿质元素，当草木灰含量增加时，氧化钠、氧化钾、氧化钙、氧化铝、氧化铁、氧化镁、氧化钡、五氧化二磷和氧化锶的含量同时增加，这使得高钾玻璃的化学成分相关性明显。而铅钡玻璃助熔剂种类繁多，例如天然泡碱、铅矿石等，而其中含有的杂质使得铅钡玻璃化学成分总类增加，但由于杂质含量各异使各成分之间关联性变弱。

六、模型的评价与推广

6.1 模型评价

（1）优点

①针对问题一的分类变量，我们选取了卡方检验。并且在进行 K-S 检验前先利用直方图和 Q-Q 图对其进行判断，减少工作量。

②针对问题三的灵敏度分析，我们通过改变干扰程度使分析更加全面合理。

（2）缺点

并未对已风化玻璃进行类别划分。

6.2 模型推广

本文构造了一套完整的玻璃分类体系，该体系由问题一的建立风化前后化学成分含量变化模型问题二的聚类分析模型问题三的针对玻璃亚分类的支持向量机模型构成。该模型利于推广，如对其他非玻璃文物进行研究时仍可利用该模型对所要研究样品进行分类。

七、参考文献

[1]伏修锋,干福熹.基于多元统计分析方法对一批中国南方和西南地区的古玻璃成分的研究[J].文物保护与考古科学,2006(04):6-13.DOI:10.16334/j.cnki.cn31-1652 /k.2006.04.002.

[2]石军,熊苡.多元统计、聚类分析法在自然资源开发中的应用[J].山东理工大学学报(自然科学版),2003(01):81-83.

[3]李清临,徐承泰,汪大海,姚政权.河南禹县阳翟遗址出土古玻璃的科学分析[J].考古与文物,2011(04):105-110.

[4]王婕,李沫,马清林,张治国,章梅芳,王菊琳.一件战国时期八棱柱状铅钡玻璃器的风化研究[J].玻璃与搪瓷,2014,42(02):6-13.DOI:10.13588/j.cnki.g.e.1000-2871. 2014.02.002.