摘要

玻璃是早期贸易往来的宝贵物证，而风化会影响对其类别的正确判断。为了帮助考古工作者完成文物样品的分类和来源鉴定，需要处理不同的样本信息，权衡相关性。本文针对古代玻璃制品的成分分析与鉴别问题，以卡方检验、敏感性分析和 K-means 算法为理论基础建立了数学模型。

针对问题一，根据数据统计分析可知，玻璃类型、纹饰、颜色对表面风化的影响从数量上有直观表现。因此，对于文物是否风化的数据统计结果比较可以有效衡量是否相关。本文对统计后数据进行卡方检验，根据计算结果来衡量玻璃类型、纹饰、颜色对表面风化的影响。依据题目要求总含量在 85%-105% 为合理数据，15 号和 17 号样本数据属于异常值，将其剔除。在此基础上对剩余数据进行归一化处理，与表单一合并使用，定义影响程度计算公式和范围，不考虑缺失较多的化学成分，通过筛选计算用数据给出分析结果，同时借此预测同类型玻璃风化前各成分含量。

针对问题二，分析上述已处理的表单，可初步得出分类规律：归一化后，铅钡玻璃中铅钡总含量一定大于 14%，高钾玻璃中钾含量普遍高于 5%。对于个别异常数据，考虑样本风化与否和硅含量的关系。因此建立聚类模型，选择 K-means 算法进行聚类，将风化前后的数据带入聚类，比较发现结果与所求完全一致，则认定分类时同时考虑了硅、钾、铅、钡的含量以及风化对不同的样本造成影响的关系，完成分类。将铅钡玻璃的数据再次导入 K-means 算法时，多次尝试聚类，得到不同类型的亚分类方法。高钾玻璃总量较少，亚分类选择综合考虑统计学方法与玻璃风化对成分的影响，给出分类标准：归一化后未风化的样本铝含量是否大于 8%，风化的样本铝含量是否大于 2.3%。

针对问题三，采用与问题二相同的处理方式，将未知样本数据进行归一化处理，与已知样本数据一同通过 K-means 算法进行聚类，分析结果发现 A1、A5、A6、A7 分为高钾，其他为铅钡，结合问题二对此分类方法比较，其所属类型与化学成分分布十分吻合。

针对问题四，结合问题一中对文物样品表面有无风化化学成分含量统计规律的分析，发现风化的作用使得不同的化学成分展现出相同或相近的变化趋势。最后分析具体数据得：高钾中钾、钙、铝、铁、磷有较强关联性，铅钡类别中有较强关联性的分组有：硅和铝，钙、铅和磷，铜、钡和锶。

关键词: 卡方检验；K-means 聚类；归一化处理

1 问题重述

1.1 问题背景

我国古代玻璃与外来玻璃制品外观相似，但化学成分不同。玻璃的主要原料是石英砂，主要化学成分是二氧化硅（SiO_2）。同时为了降低玻璃熔点，加入草木灰、天然泡碱、硝石和铅矿石等助熔剂，并添加石灰石作为稳定剂，石灰石煅烧后转化成为氧化钙 (CaO)。不同的助熔剂的化学成分不同。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) 含量较高；钾玻璃以含钾量高的物质如草木灰为助熔剂烧制而成。

古代玻璃容易受埋藏环境影响而风化。在风化过程中，其化学成分比例、颜色、纹饰都有可能发生变化，且同一文物的不同区域风化程度可能不同。

1.2 问题提出

现有一批我国古代玻璃制品的相关数据，已被分为高钾玻璃和铅钡玻璃两种类型。附件一给出了玻璃文物的基本信息，包括每个文物对应的纹饰、玻璃类型、颜色以及风化程度。附件二给出了附件一中对应编号文物表面某部位的随机采样结果，主要为各项化学成分所占比例（各成分比例的累加和应为 100% 但因检测手段等原因可能导致其成分比例的累加和非 100% 的情况。本题中将成分比例累加和介于 $85\% - 105\%$ 之间的数据视为有效数据）。附件三给出了未知类别文物的风化程度以及化学成分。

现依据附件中的数据进行建模解决以下问题：

问题一：依据附件一中的数据对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。依据附件二，结合附件一中已知的玻璃的类型，分析有无风化的玻璃的化学成分差异，并预测风化前的化学成分含量。

问题二：依据附件二数据分析高钾玻璃、铅钡玻璃的分类规律；对于高钾玻璃和铅钡玻璃分别选择合适的化学成分对其进行亚类划分，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。

问题三：对附件三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。

问题四：结合附件二和附件三，针对不同类别的玻璃文物样品，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。

2 问题假设

1. 分类中空白视为一种颜色。

2. 不考虑检测手段对数据准确性的影响。

3. 将空白数据统一视为成分含量为 0: 由于空白处表示未检测到该成分，且不考虑检测手段对数据准确性的影响，故空白处成分占比可视为 0.

4. 个别数据过少的化学成分视为取样特有成分，不考虑风化影响. 我们的数据分析主要对于成功测量出的数据进行。

3 符号说明

表 1: 符号说明

符号	含义
i	玻璃文物编号
j	已进行的归一化次数
k	化学成分
$\omega_i^j(k)$	第 i 个玻璃文物 j 次归一化后 k 所占比例
ω_i^j	第 i 个玻璃文物 j 次归一化后的化学成分
S_i^j	第 i 个文物 j 次归一化后各成分比例累加和
x	玻璃类型/花纹/颜色属性
$N_1(x)$	风化 x 玻璃数量
$N_2(x)$	无风化 x 玻璃数量
$\overline{\omega_1}(k)$	高钾风化玻璃 k 成分含量的平均值
$\overline{\omega_2}(k)$	高钾无风化玻璃 k 成分含量的平均值
$\overline{\omega_3}(k)$	铅钡风化玻璃 k 成分含量的平均值
$\overline{\omega_4}(k)$	铅钡无风化玻璃 k 成分含量的平均值
$p_{12}(k)$	高钾玻璃 k 成分含量风化与无风化相比的变化率
$p_{34}(k)$	铅钡玻璃 k 成分含量风化与无风化相比的变化率
$\omega_i^{'}(k)$	风化玻璃风化前 k 成分含量的预测结果

4 问题分析

4.1 问题一

该问主要探究有无风化玻璃的不同属性之间的差异和内部化学成分之间的统计规律。首先对附件一中的数据进行处理，针对玻璃类型、纹饰和颜色三个不同属性，统计有无风化玻璃的数量，对数据做可视化处理，得出初步结论，然后进行 χ^2 检验，进一步探究风化程度与三种属性的相关性。

对于第二问，先将各个编号玻璃的各化学成分比例累加，删去累加和非介于 85%-105% 之间的玻璃所在行，剩下编号的数据视为有效数据。后续处理均在有效数据上进行。为了使不同成分比例累加和的玻璃成分在数值上有一定比较性，提高准确性，需进行两次归一化处理。第一次，将各个编号的化学成分对各成分累加和进行归一化处理，即计算不同成分在累加和中的占比；第二次，为

探究各种氧化物的成分分布规律，计算除去二氧化硅后各成分占比。之后将附件一附件二整合，整理出各编号文物对应的化学成分占比、风化属性、玻璃类型、花纹以及颜色，以便后续分析。

为了探究高钾和铅钡玻璃表面风化与各化学成分的关系，先分开计算两种类型风化和无风化中各编号玻璃的不同化学成分的平均值，随后将风化与无风化的数据进行比较。为了使结论更加精确，进行定量计算，得出风化与否的各化学成分含量平均值变化率，根据变化率大小判断风化对各成分的影响，得出定性结果。

对于第三问，在第二问的基础上，可将各平均值变化率视为风化前后变化率，计算各编号风化玻璃的未风化数据，即为预测结果。

4.2 问题二

对于第一问，初步计算两种类型玻璃的 $\omega_i^1(K_2O)$ 以及 $\omega_i^1(PbO)+\omega_i^1(BaO)$，尝试将有较明显区分度的化学成分及其含量作为分类标准。对于该种化学成分占比数据缺漏的玻璃，探究其余化学成分与有无风化的关系，推测考古工作者将其分类至该类型的依据。

对于后一问，需分别对高钾和铅钡玻璃进行亚分类。首先根据史实资料查阅常见助熔剂和玻璃成分，找出经常与钾或钡同时使用于玻璃中的化学成分，并结合首次归一化后的数据进行统计，做出初步推断。为在每一项数据确定情况下，将不同化学成分的玻璃正确地进行更细致的划分，从而使其与该亚类中的其他玻璃相似，但与其他亚类中的玻璃不同，我们使用聚类算法。此处建立 K-means 聚类模型，将附件表单 2 中风化的数据和未风化的数据分开执行 K-means 聚群分类，并将风化前和风化后的玻璃分别归类为高钾和铅钡，检验结果与考古工作者分类的异同，从而得出该模型的合理性和敏感性。

4.3 问题三

首先将未知类别的玻璃对总成分进行归一化处理，随后再次使用问题二的 K-means 模型，将未知样本的数据加入到问题二的数据中进行分类，得到高钾和铅钡玻璃，再通过类似操作进行亚分类。

4.4 问题四

结合问题一中对文物样品表面有无风化化学成分含量统计规律的分析，进一步探究风化作用对不同类型玻璃的不同化学成分的影响，从而得出化学成分之间的关联关系及其差异性。

5 模型的建立与求解

5.1 问题一

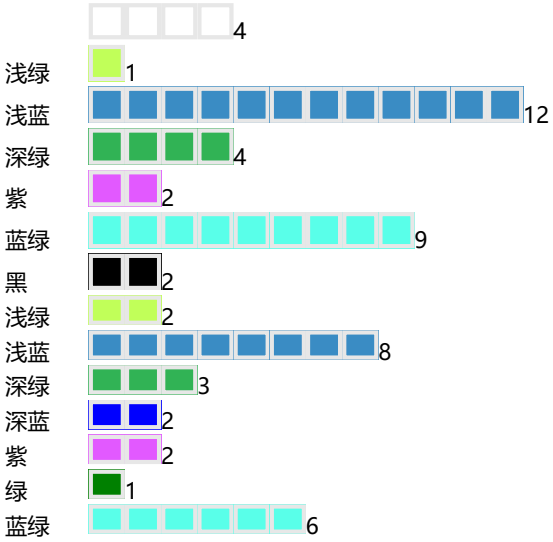
5.1.1 分析玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系

针对附件一中玻璃类型、花纹和颜色三种属性，分别统计有无风化的玻璃数量，利用数据可视化的方法得到如图5.1.1结果。

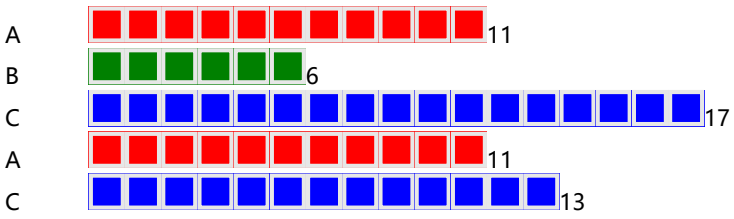
从图5.1.1中可以初步看出：（1）铅钡玻璃中风化玻璃数量远高于无风化，高钾玻璃中风化玻璃数量较低于无风化，故猜测文物表面风化与玻璃类型相关联，铅钡玻璃较容易风化；（2）A 花纹玻璃风化数量与无风化相同，B 花纹玻璃全部风化，C 花纹玻璃风化数高于无风化数，故猜测 B 花纹极易使玻璃风化，C 花纹玻璃较易风化，A 花纹与玻璃风化基本无关；（3）深绿、浅绿玻璃虽风化数小于未风化数，但样本数低，难以做出推断；浅蓝玻璃和蓝绿玻璃风化数略高于无风化数，故猜

图 1: 统计图谱

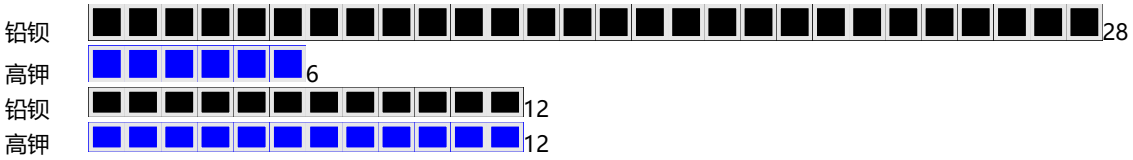
风化与否和颜色的关系统计图



风化与否和花纹的关系统计图



风化与否和玻璃类型的关系统计图



测此二种玻璃较易风化；紫色玻璃风化数与无风化数相同，黑色玻璃全部风化，深绿玻璃全无风化，但同样由于样本数较少，难以做出推断。

为验证猜想并进一步探究，现对三种属性的有无风化玻璃数进行 χ^2 检验。卡方检验通过统计实际情况与理论情况的偏差程度来推断两个变量之间的关联度。卡方值越大，偏差程度越大，相关性越大，独立性越小；卡方值越小，偏差程度越小，相关性越小，独立性越大。计算公式为

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

其中 χ^2 为卡方值，O_i 为实际频数，E_i 为预期频数 [1], 现用此方法检验有无风化与玻璃类型、花纹、颜色的关联度：

(1) 玻璃类型：

$$\chi_1^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(N_i(j) - E_i(j))^2}{E_i(j)} \tag{2}$$

其中 $E_i(j)$ 为预期频数。

(2) 花纹：

$$\chi_2^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(N_i(j) - E_i(j))^2}{E_i(j)} \tag{3}$$

(3) 颜色：

$$\chi_3^2 = \sum_{i=1}^{2} \sum_{j=1}^{9} \frac{(N_i(j) - E_i(j))^2}{E_i(j)} \tag{4}$$

计算结果如表2：

表 2: 卡方检验结果

	高钾	铅钡
有风化	6	28
无风化	12	12
χ^2	6.88039	
自由度	1	

A	B	C
11	6	17
11	0	13
4.95654		
2		

黑	蓝绿	浅蓝	浅绿	深绿	紫	(空白)	深蓝	绿
2	9	12	1	4	2	4	0	0
0	6	8	2	3	2	0	2	1
9.43245								
8								

故可得以下结论：

1. 有 99% 的把握，认为文物表面风化与玻璃类型有关；

2. 有 90% 的把握，认为文物表面风化与花纹有关；

3. 有不超过 75% 的把握，认为文物表面风化与颜色相关联，不存在显著性差异。

5.1.2 不同类型玻璃文物的表面风化与各化学成分之间的关系

首先对附件二中原始数据进行数据处理。计算出各行成分比例累加和 $S_i^0 (i = 1, 2, 3, ..., 14)$，删去非介于 85%-105% 之间的数据，即编号 15 和 17。随后分别算出 ω_i^0 对 S_i^0 占比，得到第一次归一化的结果。通过类似的方式算出除去 SiO_2 外其余成分的占比，得到二次归一化的结果（见附件二（归一化））。

将附件一中有关玻璃类型、花纹、颜色的数据拼接至附件二（归一化）中，注意表一中某些编号的风化玻璃在表二取样时取的是未风化部位。拼合结果见附件“拼合数据”。

为了探究高钾和铅钡玻璃表面风化与各化学成分的关系，先分开计算两种类型风化和无风化中各编号玻璃的不同化学成分的平均值 $\overline{\omega_1}(k)$,$\overline{\omega_2}(k)$,$\overline{\omega_3}(k)$,$\overline{\omega_4}(k)$, 其中

$$\overline{\omega_1}(k) = \sum_i \frac{\omega_i^1(k)}{6} \tag{5}$$

$$\overline{\omega_2}(k) = \sum_i \frac{\omega_i^1(k)}{12} \tag{6}$$

$$\overline{\omega_3}(k) = \sum_i \frac{\omega_i^1(k)}{28} \tag{7}$$

$$\overline{\omega_4}(k) = \sum_i \frac{\omega_i^1(k)}{12} \tag{8}$$

随后计算各化学成分风化与无风化相比平均值的变化率。

$$p_{12}(k) = \frac{\overline{\omega_2}(k) - \overline{\omega_1}(k)}{\overline{\omega_1}(k)} \tag{9}$$

$$p_{34}(k) = \frac{\overline{\omega_4}(k) - \overline{\omega_3}(k)}{\overline{\omega_3}(k)} \tag{10}$$

根据各种化学成分的变化率大小，制定风化影响的评判标准后，可得出定性结果，即统计规律，如图2所示:

表面风化	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾(K_2O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al_2O)	氧化铁 (Fe_2O)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷(P_2O)	氧化锶 (SrO)	氧化锡 (SnO)	二氧化硫 (SO_2)
风化影响统计														
高钾	上升			下降		下降	下降	下降			略有下降			
无风化	69.23		9.52	5.44		6.74	1.97	2.5			1.43			
有风化	94.33		0.54	0.87		1.94	0.27	1.57			0.28			
	36.256		-94.327731	-84.00735		-71.217	-86.294	-37.2			-80.42			
铅钡	下降			明显上升	几乎不变	下降		上升	上升	上升	明显上升	上升		
无风化	55.83			1.34	0.65	4.55		1.48	22.58	9.29	1.1	0.28		
有风化	25.75			2.8	0.68	3.06		2.33	44.89	12.16	5.48	0.43		
	-53.878			108.95522	4.615385	-32.747		57.4324	98.8043	30.8934	398.182	53.5714		
		风化影响	评判标准	p≤-100	-100<p≤-	-10<p<10	10≤p<10	p≥100						
		p（比率）		明显下降	下降	几乎不变	上升	明显上升						

图 2: 有无风化化学成分含量变化

5.1.3 根据数据预测风化前的化学成分含量

第二问计算得出的各平均值变化率 $p_{12}(k), p_{34}(k)$ 可近似视为风化前后变化率，故将无风化数据进行计算即可预测风化前结果。

$$\omega_i^{'}(k) = (1 + p_{mn}(k))\overline{\omega_m}(k) \tag{11}$$

其中 $m = 1, 3, n = 2, 4$. 计算结果如表5.1.3所示。

5.2 问题二

5.2.1 分析高钾玻璃、铅钡玻璃的分类规律

根据数据分析可初步得出如下分类规律：

表 2: 预测结果

编号	类型	$\omega_i'(SiO_2)$	$\omega_i'(Na_2O)$	$\omega_i'(K_2O)$	$\omega_i'(CaO)$	$\omega_i'(MgO)$	$\omega_i'(Al_2O_3)$	$\omega_i'(Fe_2O_3)$	$\omega_i'(CuO)$	$\omega_i'(PbO)$	$\omega_i'(BaO)$	$\omega_i'(P_2O_5)$	$\omega_i'(SrO)$	$\omega_i'(SnO2)$	$\omega_i'(SO_2)$
2	铅钡	67.12	0	0.9	0.96	0.96	7.27	1.59	0.14	20.36	0	0.61	0.11	0	0
8	铅钡	46.06	0	0	0.75	0	2.1	0	6.97	15.22	25.17	0.76	0.25	0	2.72
08严重	铅钡	13.93	0	0	2.13	0	2.3	0	2.78	22.74	32.59	2.11	0.48	0	20.94
11	铅钡	67.07	0	0.19	1.55	0.63	3.68	0	2.88	11.76	10.28	1.73	0.22	0	0
19	铅钡	62.62	0	0	1.37	0.55	5.17	1.3	2.17	20.99	3.98	1.73	0.12	0	0
26	铅钡	10.95	0	0.54	1.96	0	2.38	0	3.1	20.44	36.77	1.65	0.55	0	21.66
26严重	铅钡	45.78	0	0	0.74	0	1.11	0	7.16	15.85	26.29	0.67	0.31	0	2.09
34	铅钡	68.47	0	0.22	0.33	0	2.13	0.41	0.85	20.67	6.74	0.06	0.13	0	0
36	铅钡	71.01	1.84	0.12	0.15	0	1.97	0.26	0.36	17.32	6.85	0.01	0.12	0	0
38	铅钡	64.72	1.25	0	0.29	0	3.46	0.26	0.42	22.48	6.78	0.09	0.24	0	0
39	铅钡	59.54	0	0	0.56	0	0.78	0	0.58	32.11	5.77	0.24	0.42	0	0
40	铅钡	45.74	0	0	1.13	0	0.84	0.24	0	44.59	6.45	0.45	0.56	0	0
41	铅钡	47.78	0	0.53	2.83	3.12	5.91	2.14	0.14	26.49	8.9	1.79	0.37	0	0
43/1	铅钡	36.43	0	0	3.4	1.15	4.53	1.03	4.6	40.76	7.54	0	0.56	0	0
43/2	铅钡	54.5	0	0	3.55	1.05	5.87	1.61	1.11	26.08	2.89	2.98	0.35	0	0
48	铅钡	74.08	0.51	0.21	0.86	0.94	13	0.66	0	5.06	3.58	0.14	0.1	0.84	0
49/1	铅钡	61.45	0	0	2.16	1.38	7.87	2.7	0.44	16.92	4.59	2.19	0.29	0	0
50/1	铅钡	49.05	0	0	1.92	0.57	3.5	0.42	0.9	27.85	13.65	1.6	0.54	0	0
51/1	铅钡	55.87	0	0	1.79	1.19	8.17	1.25	0.91	21.2	7.15	1.7	0.27	0.49	0
51/2	铅钡	56.22	0	0	2.98	1.68	4.53	0.51	0.58	31.36	0	2.13	0	0	0
52	铅钡	60.06	1.31	0	1.17	0.57	1.86	0.25	0.48	25.67	7.1	1.23	0.31	0	0
54/1	铅钡	52.07	0	0.34	1.65	1.32	6.65	0	0.57	30.07	5.8	0.92	0.62	0	0
54/2严重	铅钡	47.92	0	0	0	1.37	7.01	0	1.1	37.99	0	3.66	0.94	0	0
56	铅钡	63.14	0	0	0.58	0	2.75	0	0.5	20.73	11.79	0.51	0	0	0
57	铅钡	57.64	0	0	0.66	0	3.39	0	0.77	23.72	13.82	0	0	0	0
58	铅钡	63.15	0	0.33	1.6	0.72	5.02	0.82	1.91	18.97	5.61	1.73	0.15	0	0
7	高钾	74.65	0	0	7.35	0	7.55	1.36	5.67	0	0	3.42	0	0	0
9	高钾	73.26	0	10.93	4.07	0	4.82	2.45	2.59	0	0	1.88	0	0	0
10	高钾	75.07	0	17.14	1.39	0	2.97	2.01	1.41	0	0	0	0	0	0
12	高钾	67.78	0	17.44	4.41	0	4.97	2.07	2.57	0	0	0.75	0	0	0
22	高钾	62.46	0	12.02	9.57	0.59	11.21	2.35	0.81	0	0	0.99	0	0	0
27	高钾	76.51	0	0	6.61	0.61	9.81	1.64	2.76	0	0	2.07	0	0	0

1. 先考察铅加钡含量：对于任意编号为 i 的铅钡玻璃，$\omega_i^1(PbO)+\omega_i^1(BaO)$ 占比最低值为 14.35%，显著高于高钾玻璃的 $\omega_i^1(PbO)+\omega_i^1(BaO)$ 最高值 4.32%，故可将铅钡含量高于 14% 的玻璃归类为铅钡玻璃。其余铅钡含量低于 14% 的数据中，将 $\omega_i^1(K_2O)$ 降序排列会发现从 5.19% 到 1.01% 有较明显的突变，故将 $\omega_i^1(K_2O)$ 高于 5% 的玻璃分为高钾玻璃。其余铅钡和钾含量都较低的玻璃继续进行更细致的划分。

2. 对于剩下铅钡和钾含量都较低的玻璃，可根据“高钾和铅钡玻璃风化前后硅含量的变化相反”判断。观察到高钾玻璃风化后 $\omega_i^1(SiO_2)$ 明显升高，铅钡玻璃风化后 $\omega_i^1(SiO_2)$ 明显降低，故可根据玻璃是否风化以及硅占比将其分类至高钾玻璃或铅钡玻璃，即将风化且硅含量高的分为高钾玻璃，将风化且硅含量低的分为铅钡玻璃，将无风化且硅含量低的分为高钾玻璃，将风化且硅含量高的分为铅钡玻璃。

10	96.95	高钾	风化
09	95.24	高钾	风化
12	94.7	高钾	风化
27	93.84	高钾	风化
07	92.91	高钾	风化
22	92.35	高钾	风化

(a) 高钾风化玻璃二氧化硅含量

03部位1	87.05	高钾	无风化
18	81.71	高钾	无风化
21	77.83	高钾	无风化
01	71.03	高钾	无风化
04	68.58	高钾	无风化
06部位1	68.39	高钾	无风化
16	66.23	高钾	无风化
05	63.81	高钾	无风化
14	63.1	高钾	无风化
03部位2	62.41	高钾	无风化
06部位2	60.51	高钾	无风化
13	60.13	高钾	无风化

(b) 高钾无风化玻璃二氧化硅含量

编号	SiO2	玻璃类型	风化属性
48	53.78	铅钡	风化
36	40.53	铅钡	风化
34	36.69	铅钡	风化
02	36.32	铅钡	风化
11	35.21	铅钡	风化
38	33.41	铅钡	风化
56	31.6	铅钡	风化
58	30.77	铅钡	风化
49	30.15	铅钡	风化
19	30.01	铅钡	风化
57	27.49	铅钡	风化
52	27.36	铅钡	风化
39	26.58	铅钡	风化
51部位1	25.82	铅钡	风化
51部位2	23.28	铅钡	风化
43部位2	22.45	铅钡	风化
54	22.35	铅钡	风化
08	20.18	铅钡	风化
50	19.94	铅钡	风化
26	19.83	铅钡	风化
41	19.7	铅钡	风化
54严重风化点	17.65	铅钡	风化
40	16.95	铅钡	风化
43部位1	13.11	铅钡	风化
08严重风化点	4.69	铅钡	风化
26严重风化点	3.72	铅钡	风化

(c) 铅钡风化玻璃二氧化硅含量

编号	SiO2	玻璃类型	风化属性
33	75.54	铅钡	无风化
32	70.66	铅钡	无风化
28未风化点	68.98	铅钡	无风化
35	68.51	铅钡	无风化
31	66.96	铅钡	无风化
53未风化点	64.79	铅钡	无风化
29未风化点	63.38	铅钡	无风化
45	62.31	铅钡	无风化
44未风化点	61.28	铅钡	无风化
37	60.13	铅钡	无风化
46	55.98	铅钡	无风化
23未风化点	55.74	铅钡	无风化
49未风化点	55.59	铅钡	无风化
47	52.99	铅钡	无风化
42未风化点2	52.43	铅钡	无风化
42未风化点1	52.3	铅钡	无风化
25未风化点	52.14	铅钡	无风化
55	50.85	铅钡	无风化
50未风化点	46.44	铅钡	无风化
20	42.26	铅钡	无风化
30部位2	37.42	铅钡	无风化
30部位1	35.06	铅钡	无风化
24	32.3	铅钡	无风化

(d) 铅钡无风化玻璃二氧化硅含量

现用 K-means 聚群算法进行检验。K-means 聚群算法是一种数据挖掘中常用的算法 [2]，即它能够帮助我们从数据中提取其中隐藏的相关性 [3]。K-means 的一般运作方式为：对于给定的数据集 $S \subseteq \mathbb{R}^n$ 和 $k \in \mathbb{N}$，先根据某种方式选取 $\mathbb{R}^n$ 中的 k 个点 $P_1, P_2, \ldots, P_k$ 并称之为划分的基准点，并且将 S 中的每个点映射到离它最近的基准点；将相同的点划分为一类即得到集合 S 的一个分划。随后对此分划进行迭代：每次迭代过程中，先取分划中的每个子集的平均值（若某子集为空集，则取

这个集合在上一次迭代过程中对应的点），随后以取出的这 k 个新的点为划分的基准点再进行划分并得到分划，此为一次迭代；当某一次迭代所得到的分划和上一次一致时，我们认为分划已经收敛，K-means 算法已经得出聚群分类的结果。

K-means 算法中，两个重要的参数是：初始划分基准点的选取，和 k 值的选取。在本文所描述的工作中，我们使用 mathematica 内建函数 $FindCluster[data, n]$ 并指定参数 $Method \to "KMeans"$，以调用 mathematica 的 K-means 算法实现以处理数据，并采用了 mathematica 默认的划分基准为选取策略；对于后者，我们先人为观察数据，并且预估分为的聚群的数目，并在运行结束后评估结果的合理性。

mathematica 的运行笔记本见附件 mma.nb，我们将附件表单 2 中风化的数据和未风化的数据分开执行 K-means 聚群分类，并将风化前和风化后的玻璃分别归类为高钾和铅钡，得到的分类结果和考古工作者的分类结果完全一致。

5.2.2 对于不同类型玻璃进行亚分类

古代玻璃常见成分为高铅硅酸盐、钾铅硅酸盐、钾钙硅酸盐 [4], 不同助熔剂成分如下表所示：

表 3: 常见助熔剂成分表

草木灰	天然泡碱	硝石	铅矿石
钠、磷、钙、硅、镁、硫、铁、铜	钠	钾	铅（方铅矿）与铜矿、闪锌矿伴生

现结合上述资料进行分析。

1. 高钾玻璃：

   对于已有数据处理整合发现，高钾玻璃总样本数为 18，其中风化的样本数为 6，未风化的样本数为 12，样本数较少，故采取直观的数据统计处理。归一化处理后的数据经排序分类发现：无风化样本中，根据铝含量是否大于等于 8% 作为分类依据，可将高钾样本分为钾铝玻璃和钾玻璃两个亚类。同时比对风化样本，发现风化后的玻璃氧化铝含量明显下降，故 22 号和 27 号样本铝含量反常高于其余风化样本平均水平的原因可解释为，二者均为高铝玻璃，但由于风化，铝含量降低。剩余风化样本可分类至低铝玻璃即钾玻璃。

   综上所述，高钾玻璃的分类规律为：无风化样本中，氧化铝含量大于等于 8% 分类为钾铝玻璃，其余分类为钾玻璃；风化样本中，氧化铝含量大于等于 2.3% 分类为钾铝玻璃，其余分类为钾玻璃.

2. 铅钡玻璃：

   类似上一问，我们对铅钡玻璃进行 K-means 聚群分析。我们先对于任意两个不同成分，以其中一个为横坐标，另一个为纵坐标画出各个铅钡玻璃样品的散点图（共计 196 张散点图，见附件 mma.nb），并通过散点图观察发现可以使用二氧化硅含量和氧化铝、氧化铜的含量之和进行聚群分类得到比较好的区分效果。在对这些数据进行运行后得到如图2的聚群分类。

   图2 中左侧蓝色数据点二氧化硅含量较低，右侧绿色数据点二氧化硅含量较高，故命名为低硅铅钡玻璃、高硅铅钡玻璃；红色数据点氧化铜和氧化铝的行之和相对其它数据较高，且二氧化硅含量不低，下方橙色数据点二氧化硅和氧化铝、氧化铜含量均适中，因此分别命名为铝铜铅钡玻璃、中硅铅钡玻璃。具体的划分方式如下：

   - 二氧化硅含量低于 44%：低硅铅钡玻璃；

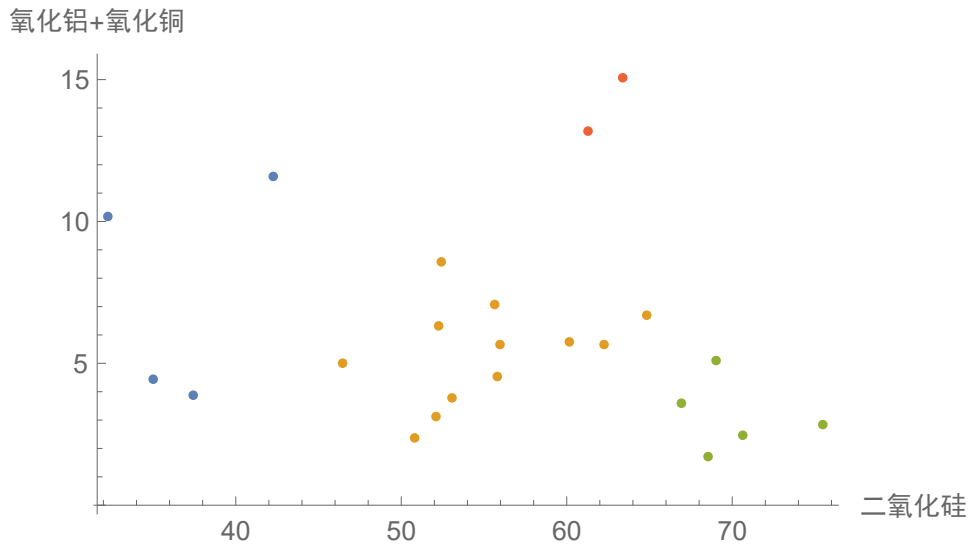


图 3: 铅钡玻璃聚群分类

- 二氧化硅含量高于 66%，且氧化铝和氧化铜含量之和低于 12%：高硅铅钡玻璃；
- 二氧化硅含量介于 44% 和 66% 之间，且氧化铝和氧化铜含量直和低于 12%：中硅铅钡玻璃；
- 二氧化硅含量高于 44%，且氧化铝和氧化铜含量直和高于 12%：铝铜铅钡玻璃。

5.2.3 合理性和敏感度分析

合理性分析：高钾类别中，将无风化的样本以铝含量 8% 作为划分依据，可将无风化的高钾样本分为两类。而应用问题一解法，风化后 2.3% 的铝含量预测风化前含量恰为划分标准的 8%，则对于风化后的高钾样本，可被分入对应的亚类；铅钡类型中，使用 K-means 算法对所有铅钡中无风化聚类得到，可得结果亚类聚类相同，则有分类合理。

敏感性分析：讨论作为分类依据的化学成分有微小变化时对分类结果造成的影响由于归一化后的数据不连续，相比于分类依据的确切数值，样本化学成分的微小变化对分类结果几乎不会造成影响。

5.3 问题三

5.3.1 成分分析鉴别类型

第二题中已经得出了较为合理的分类方法，现应用于本题对未知类别的玻璃进行分类。首先对附件三中的数据进行归一化处理，得到各成分占总累加和的占比，结果见附件三（归一化）。

我们将归一化后的数据与问题二中的数据合并，并且再次运行 K-means 聚类算法，得到的结果中，A1 被归类为未风化的高钾玻璃，A3、A4、A8 被归类为未风化的铅钡玻璃；A2 被归类为风化的铅钡玻璃，A5、A6、A7 被归类为风化的铅钡玻璃。

可以注意到，A2、A3、A4、A8 的氧化铅与氧化钡含量合计高于 14%，且二氧化硅含量小于 60%，被归类为铅钡玻璃，与我们经过观察总结的规律相符合；同时 A1 玻璃二氧化硅含量高于 60%，被归类为高钾玻璃；A5、A6、A7 玻璃已风化且二氧化硅含量高于 60%，故被归类为高钾玻璃。

根据高钾和铅钡玻璃各自的亚分类标准，将高钾和铅钡玻璃进行更细致的分类。高钾玻璃中，基于其铝含量，A1、A5、A6、A7 均被分类为钾铝玻璃；而铅钡玻璃中，A3、A4 根据二氧化硅含量

被归类为低硅铅钡玻璃；A8 根据其二氧化硅含量和氧化铝、氧化铜含量被归类为中硅铅钡玻璃；根据问题 1 中的预测模型，预测得 A2 未风化前的二氧化硅含量为 75.03%，被归类为高硅铅钡玻璃。

5.3.2 分类结果的敏感性分析

敏感性分析：由于题目中未知样本数据划分时存在估计和预测，所以对数据微小变化对分类结果的影响对于初始未知数据划分时采用归一化后一同聚类分析给定初步的结果支撑，对于参与判定的钾硅铅钡含量，其数值上为准确数据，则聚类结果不受数据处理的影响。而对于风化后的数据处理，进一步验证分类时，可能对风化前数据预测发生波动，由于数据较为离散，分析认定数据的微小变化对模型结果的影响不大。

5.4 问题四

1. 对于高钾玻璃样本：分析问题一中计算所得数据，发现：高钾玻璃样本在发生风化后，硅含量上升，铜含量下降，钾、钙、铝、铁、磷的下降幅度十分相近，即受风化影响近乎相同，其具有较强的关联性。

2. 对于铅钡玻璃样本：同理分析已有数据得：铅钡玻璃样本在发生风化后，硅和铝的含量一同下降，钙、铅和磷均有明显上升，而铜、钡和锶同时有中等幅度的上升。以风化影响作衡量，分析得铅钡类型中，此三组化学成分有较强的关联性。

差异性：比较高钾类型和铅钡类型化学成分关联关系，铅钡中硅铝之间的关联性在高钾类型中截然相反，而高钾中铜未与其他元素有关联性，但在铅钡中，铜存在与锶和钡的关联。不同类别差异的来源在于，不同类别的文物样本化学组成不完全相同，有交集的成分的有效数据较少。

6 模型的评价

本文所用的模型有以下有点：

- 本文使用的卡方检验适用范围广，原理较为简单；
- 本文所使用的卡方检验和 K-means 模型适用范围广，且得出的结果和考古工作者所得出的结果相同，准确性较高。

本文所用的模型有以下局限性：

- 本文使用的模型的自动化不充足，包含大量需要手动计算或人工检验的部分；
- 本文的模型依赖于 K-means 聚群算法 [5]，由于 K-means 算法随着数据增多，其耗时显得较长，对模型处理大量数据的能力产生了影响；
- 本文的模型采用卡方法检验相关性，其计算量相对较大，同样影响了此模型处理大量数据的能力。

参考文献

[1] 李濛, 包蕾, 胡毅, 成嵩, 胡晓波, and 高鹰, "基于卡方检验的随机数在线检测方法的实现 [j]," 微电子学, pp. 388–392, 2022. [Online]. Available: https://doi.org/10.13911/j.cnki.1004-3365.210329 5.1.1

[2] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63–67. 5.2.1

[3] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003, biometrics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320302000602 5.2.1

[4] 干福熹, "中国古代玻璃的起源和发展 [j]," 自然杂志, pp. 187–193+184, 2006(04). 5.2.2

[5] 刘文佳 and 张骏, "一种改进的 k-means 聚类算法 [j]," 现代商贸工业, vol. 39(19), pp. 196–198, 2018. [Online]. Available: https://doi.org/10.19311/j.cnki.1672-3198.2018.19.086 6

A 支撑材料列表

-> 我的支撑材料.rar

|- mma.nb * (mathematica 11.1.0 笔记本)

|- 附件二（归一化）.xls

|- 附件三（归一化）.xlsx

|- 拼合数据.xlsx

|- system.rar

    |- README.md (程序使用说明)

    |- .gitignore

    |- index.js

    |- package.json

    |- package-lock.json

    |- router.js

    |- server.js

    |- src/

        |- index.html

        |- index.js

        |- mma.js

        |- mma.nb

        |- mma.pdf

        |- mma1.pdf

        |- mma2.pdf

        |- mma3.pdf

        |- preload.js

        |- wcccall.js

    |- data/

        |- 表单 1.csv

        |- 表单 2.csv

        |- 表单 2 归一化.csv

        |- 表单 2 归一化 mathematicaform.csv

        |- 表单 3.csv

        |- 除二氧化硅归一化.csv

        |- 除二氧化硅归一化 mmaform.csv

        |- 附件二（归一化）.xls

        |- 数据 1.csv

|- 无风化.csv
|- data.csv
|- qbfh.csv
|- sep.csv
|- tolatex.csv
|- tolatex.tex
|- tolatex2.aux
|- tolatex2.csv
|- tolatex2.log
|- tolatex2.pdf
|- tolatex2.tex