古代玻璃制品的成分分析与鉴别

摘要

我国古代玻璃工艺的起源很早，近年来随着“丝绸之路”上的文化与工艺品的广泛传播，跟随“丝绸之路”传来的玻璃制品受到人们越来越多的关注。古代玻璃极易受埋藏环境的影响而风化。本文采用显著性检验，斯皮尔曼系数，Kruskal-Wallis 检验等统计学方法，建立了 DBSCAN 聚类模型以及 k-means 聚类模型，以用于研究玻璃制品文物风化元素含量以及种类划分等问题。

针对问题一，题目要对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，使用了列联表和卡方检验来计算是否相关；用图表的形式来展现化学成分的变化趋势以分析样品表面有无风化化学成分含量的统计规律，统计规律的分析使用了平均数、方差、五数概括等方法。利用附录表单中的数据，得到不同玻璃类型分化前后的化学成分含量比例，进而得到风化前的成分含量。

针对问题二，通过数据可视化所得到的图表，从玻璃类型、风化状态、颜色等方面可以通过直观数据来判断大致的分类情况，同时利用了 DBSCAN 聚类模型来进行数字化模型分类，实测在 Esp=46.5，MinPts=60 时分类效果最优。通过显著性检验来寻找与颜色有显著性的化学成分，得到在高钾文物中，氧化纳、氧化钙、氧化锡的含量对文物颜色有着显著作用；在铅钡文物中，氧化镁、氧化铝、氧化铜、氧化锡、二氧化硫对文物颜色有着显著作用。以这些化学成分为变量使用 k-means 聚类算法分类，聚类得到的中心点用来验证聚类效果，同时也用来预测的待测数据的亚类情况，进行亚类划分，对于两种分类各自亚类划分的正确性分别达到了 95%和 93%。

针对问题三，直接使用第二问所建立的模型，使用 DBSCAN 模型来对其进行高钾-铅钡大类分类，再使用 k-means 聚类得到的中心点来预测其亚类分类，多次循环改变数据值（改变范围最高 30%）其输出均无变动，证明模型稳定性较强。

针对问题四，由此想到利用斯皮尔曼系数来判断相关性，得出氧化铝和氧化锡含量在两种玻璃类型的文物中具有相关性的结论，利用 Kruskal-Wallis 检验来判断差异性，得出二氧化硅含量、氧化钾含量、氧化铁含量、氧化铅含量、氧化钡含量、氧化锶含量在两种不同玻璃类型中有显著差异的结论。

关键词：DBSCAN 模型，显著性检验，k-means 聚类，斯皮尔曼系数，Kruskal-Wallis 检验

一、 问题重述

1.1 研究背景

玻璃的主要原料是石英砂，主要化学成分是二氧化硅（SiO2）。煅烧过程中添加的助熔剂不同，其主要化学成分也不同。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）的含量较高，通常被认为是我国自己发明的玻璃品种，楚文化的玻璃就是以铅钡玻璃为主。钾玻璃是以含钾量高的物质如草木灰作为助熔剂烧制而成的，主要流行于我国岭南以及东南亚和印度等区域。现有一批我国古代玻璃制品的相关数据，考古工作者依据这些文物样品的化学成分和其他 检测手段已将其分为高钾玻璃和铅钡玻璃两种类型。

1.2 问题提出

基于上述研究背景本文将要解决以下几个问题：

问题一：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。

问题二：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，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。

问题四：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。

二、问题分析

针对问题一，可以分为三个小问：首先题目要对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，进而想到相关关系、差异关系以及关系是否显著，结合是否风化、玻璃类型、纹饰、颜色这些数据都是定类数据，故想到用列联表和卡方检验来看是否相关；接着题目问根据玻璃类型来分析样品表面有无风化化学成分含量的统计规律，根据附录表单的数据可以看出高钾类型由未风化到风化，二氧化硅呈现一个由少变多的趋势，而铅钡类型的玻璃文物刚好相反，故联想到用图表的形式来展现化学成分的变化趋势，统计规律的分析使用了平均数、方差、五数概括等方法，较为完善的反应了各种化学元素的统计规律。最后题目要求根据风化点的检测数据，预测风化前的化学成分含量，利用附录表单中的数据，得到不同玻璃类型分化前后的化学成分含量比例，进而得到风化前的成分含量。

针对问题二，通过数据可视化所得到的图表，从玻璃类型、风化状态、颜色等方面可以通过直观数据来判断大致的分类情况，同时，我们也利用了 DBSCAN 聚类模型进行多次分类模拟来进行数字化模型分类。在进行亚类划分时，由于是从化学元素来考虑，参考第一问所作的相关性分析，我们认为利用颜色来划分亚类最为妥当，故首先通过显著性检验来寻找与颜色有显著性的化学成分，分别得

到高钾和铅钡两大类中的对颜色有显著作用的化学成分，再以这些化学成分为变量使用 k-means 聚类算法分类，聚类得到的中心点用来验证聚类效果，同时也用来预测的待测数据的亚类情况。

针对问题三，我们可以直接使用第二问所建立的模型，首先使用 DBSCAN 模型来对其进行高钾-铅钡大类分类，再使用 k-means 聚类得到的中心点来预测其亚类分类，多次循环改变数据值观察其敏感度情况。

针对问题四，题目要求分析不同类型的玻璃文物的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。此问题和问题一的第二个小问相似，不同之处是题目要求分析关联关系以及差异性，由此想到利用斯皮尔曼系数来判断相关性，利用 Kruskal-Wallis 检验来判断差异性。

三、模型假设

针对本文提出的有关玻璃制品成分鉴别与分析的问题，我们做了如下模型假设：

1. 假设在考虑风化前后玻璃类型以及化学成分含量的变化情况只受风化因素的影响；

2. 假设同种风化程度下，同种属性的玻璃受风化因素的影响是基本相同的；

3. 在假设检验时，若无特殊说明，显著性水平α均取 0.05；

4. 假设提供的数据集能广泛的代表出土文物玻璃数据的普遍情况；

四、模型的建立与求解

4.1 数据的预处理

图 1 为所作的数据初处理的流程图

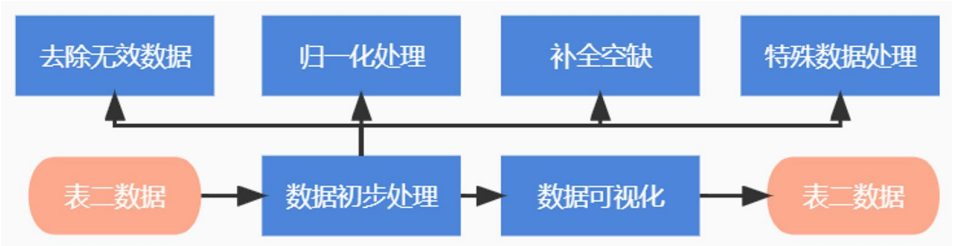

图 1

去除无效数据是根据题目中所说的各个化学成分的含量之和在 85%~105%之间为正常数据这一条件，可以将附录表单 2 中的文物采集点 15 以及 17 作为异常值舍去。

将异常值舍去之后，再将剩下的数据进行归一化处理，以使得各个化学成分含量之和均为 100%。

然后是补全空缺内容，这一步是由于后续使用 python 以及 c++处理数据时对

空值的数据处理方便。

特殊数据处理的过程处理的是一些标记数据，比如数据含有一些同一文物上风化类型相同的点，会取其平均值代表该文物的整体概况。以及还有一些数据为同一文物的风化点以及严重风化点，该数据在我们的假设中研究风化的影响时可以看作时未风化点和风化点来看（此时重点是风化的过程对化学成分的影响），但是由于其原本就已经收到一定的风化，因此在亚类判别（对其本身所含化学成分进行分类）时会将严重和非严重的两个都看作为风化点。

数据可视化是为了对数据概况有一个初步的认知，方便后续对题目的分析与认知。

4.2 问题一的分析与求解

4.2.1 文物各属性关系分析

要对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，首先要确定各个数据的数据类型，由分析可知，文物风化状态、玻璃类型、纹饰、颜色，都是定类数据，因此选用列联表独立检验的方式来判断相关性。

由题目所给数据可得如下列联表

	风化	无风化	总计
铅钡	28	12	40
高钾	6	12	18
总计	34	24	58

表 1

	风化	无风化	总计
A	11	11	22
B	6	0	6
C	17	13	30
总计	34	24	58

表 2

	风化	无风化	总计
浅蓝	12	8	20
蓝绿	9	6	15
深绿	4	3	7
紫	2	2	4
无颜色数据	4	0	4
浅绿	1	2	3
黑	2	0	2
深蓝	0	2	2
绿	0	1	1
总计	34	24	58

表 3

使用列联表独立性检验，设出原假设为 H_0：数据之间是独立的，则备选假设为 H_1：数据之间不具有独立性，即数据之间有影响。通过该检验的计算公式

计算可得表一二三的检验统计量分别为

$$\chi_1^2 = 6.8804、\chi_2^2 = 4.9565、\chi_3^2 = 9.4324$$

其中由于三个列联表的行数不相同，则它们的拒绝域也不相同，这里假设显著性水平α=0.05，则表一二三的拒绝域分别为

$$W_1 = \chi^2 \geqslant \chi_{0.95}^2(1)，$$

$$W_2 = \chi^2 \geqslant \chi_{0.95}^2(2)，$$

$$W_3 = \chi^2 \geqslant \chi_{0.95}^2(8)。$$

通过查阅卡方分布的分位数表，可得

$$\chi_{0.95}^2(1) = 3.8415，\ \chi_{0.95}^2(2) = 5.9915，\ \chi_{0.95}^2(8) = 15.5073$$

因此对于列联表一可拒绝原假设，即可以认为文物是否风化和玻璃类型之间有相关关系；对于列联表二和三，由于检验统计量均未落入拒绝域内，故应接受原假设，可以认为文物是否风化和纹饰之间相互独立、和颜色之间相互独立，即是数据之间的相关性很小。

接着，利用程序将经过数据预处理之和的附录表单 2 中的数据按照高钾-风化、高钾-无风化、铅钡-风化、铅钡-无风化的类别进行作图，结果如图 2 所示：

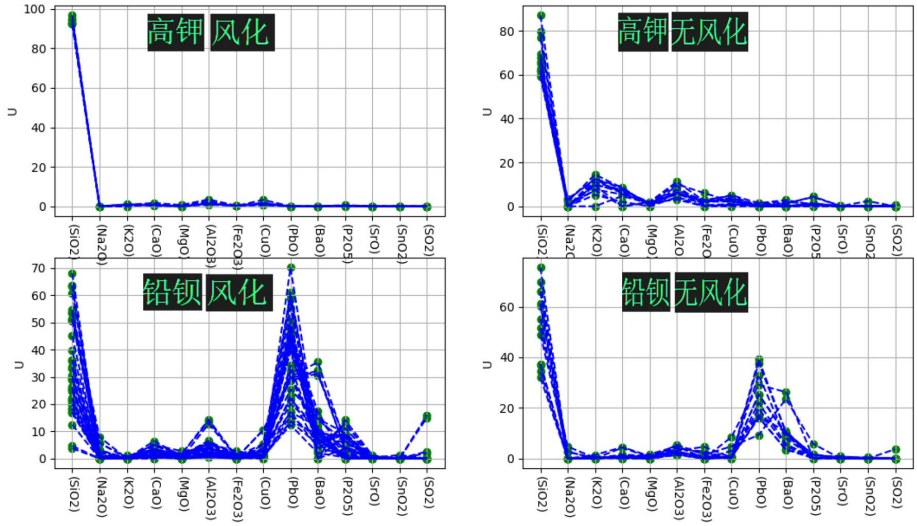

图 2

由图 2 可知，高钾类型的文物在风化之前氧化钾和氧化铝的含量明显多于其风化之后的含量，由此推测玻璃类型为高钾的文物在进行风化过程中，氧化钾和氧化铝的含量会随着风化的进程逐渐减少；对于铅钡类型的文物，可以推测其氧化铅和氧化铝的含量会随着风化的进程逐渐减少。

此外，对于二氧化硅来说，这两种玻璃在从无风化到风化的过程中，二氧化硅的含量都有所的下降，但下降的程度有大有小，由此推测二氧化硅的含量和文物风化程度有关，风化程度越大，则二氧化硅的含量就越低。

4.2.2 通过类型以及风化分析统计规律

五数概括法是可以表示数据离散程度的方法，五数概括指的是通过最小数值，

第一个四分位数、中位数、第三个四分位数和最大数据值，其数据见图 3。

	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_3)	氧化铁(Fe_2O_3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)	氧化锶(SrO)	氧化锡(SnO_2)	二氧化硫(SO_2)
高钾-无风化														
最小值	60.1284	0	0	0	0	3.13625	0	0	0	0	0	0	0	0
1/4处值	62.409	0	7.45047	2.01	0.528402	4.06	0.426786	0.78	0	0	0.66	0	0	0
中位数	66.2331	0	10.0666	5.93649	1.12257	6.27985	2.14449	2.26941	0.111777	0	0.973992	0	0	0
3/4处值	77.8319	2.13393	12.5101	8.31313	1.7503	9.32323	2.71474	3.96476	1.01502	1.39507	1.39846	0.101133	0	0.374766
最大值	87.05	3.41414	14.7546	8.86489	2.00162	11.2717	6.11089	5.14765	1.63636	2.89239	4.55281	0.121408	2.42674	0.486996
铅钡-风化														
最小值	3.7241	0	0	0	0	0.456528	0	0	12.3248	0	0	0	0	0
1/4处值	20.1763	0	0	0.80604	0	1.63884	0	0.696507	23.4324	5.41717	0.195755	0.225341	0	0
中位数	30.1466	0	0	2.159	0.735751	2.82	0.294207	0.891049	39.8441	7.75618	3.13564	0.341969	0	0
3/4处值	52.3008	1.29677	0.30036	3.53382	1.18488	5.50719	1.03862	3.09845	48.7967	12.2902	7.96073	0.501547	0	0
最大值	68.9837	8.20725	1.05116	6.62046	2.91324	14.3572	2.86911	10.5891	71.2286	35.489	14.579	1.15559	1.32096	15.9676
高钾-风化														
最小值	92.35	0	0	0.2104	0	0.811542	0.170512	0.55	0	0	0	0	0	0
1/4处值	92.9087	0	0	0.621429	0	1.32304	0.202409	0.841599	0	0	0.150648	0	0	0
中位数	93.8367	0	0.59136	0.723109	0	1.46631	0.260495	1.55357	0	0	0.21	0	0	0
3/4处值	96.9542	0	1.01436	1.66	0.64	3.5	0.35	3.24975	0	0	0.611836	0	0	0
最大值	96.9542	0	1.01436	1.66	0.64	3.5	0.35	3.24975	0	0	0.611836	0	0	0
铅钡-无风化														
最小值	32.3018	0	0	0	0	1.49673	0	0	10.5192	3.47455	0	0	0	0
1/4处值	37.4164	0	0	0.394969	0	1.60801	0	0.111494	16.2599	4.94628	0.102817	0	0	0
中位数	55.9769	0	0.15006	0.854179	0.516717	3.14621	0	0.538947	22.9186	10.1997	0.202778	0.304785	0	0
3/4处值	68.5064	2.7049	0.29817	1.62552	1.0004	4.85653	2.16437	3.0106	34.1565	11.145	1.46029	0.486322	0	0
最大值	75.5402	4.79128	1.43951	4.58397	1.6932	6.16446	4.66321	8.55583	40.0408	26.6373	6.50379	0.920307	0.445795	3.66073

图 3

通过五数概括表格可以做出其箱线图如图 4 所示

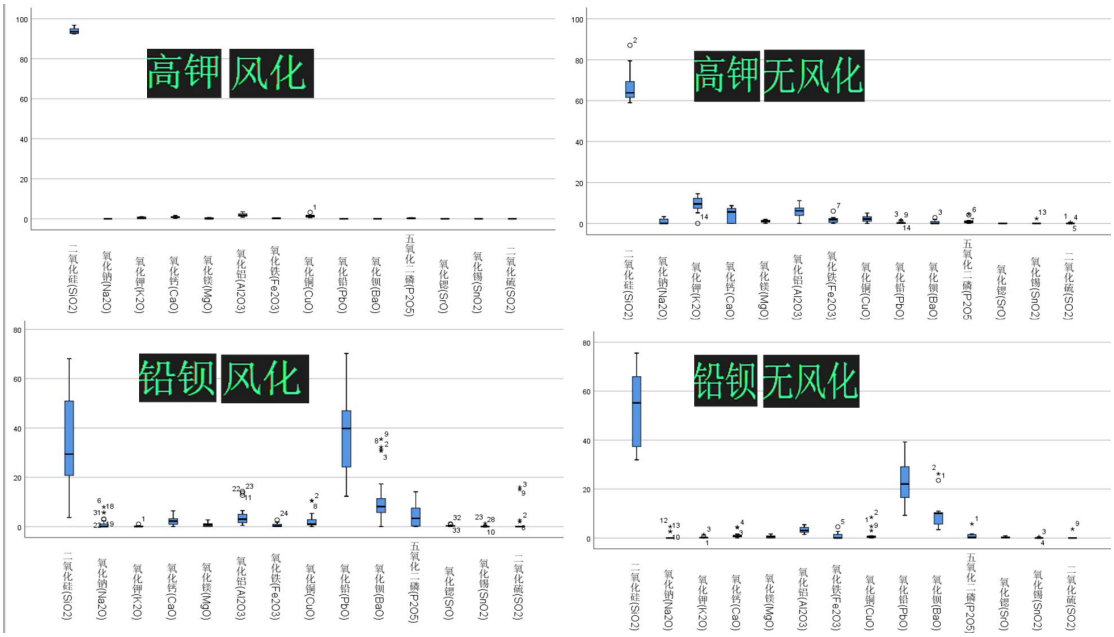

图 4

由此可以大致看出一些显著的分布规律，比如说二氧化硅(SiO2)在所有分类中都占据主要部分，高钾类型的二氧化硅含量显著高于铅钡类型，铅钡分类的主要特征时氧化铅含量较多等等。

	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_3)	氧化铁 (Fe_2O_3)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷(P_2O_5)	氧化锶 (SrO)	氧化锡 (SnO_2)	二氧化硫 (SO_2)
高钾-无风化														
平均值	69.2315	0.70519	9.51503	5.44041	1.10254	6.73855	1.96878	2.50033	0.416285	0.605703	1.42601	0.0423781	0.202228	0.105109
方差	69.4087	1.56126	14.6751	9.17465	0.439228	5.79845	2.61847	2.62167	0.325258	0.905928	1.92364	0.0022067	0.449858	0.0337232
标准差	8.33119	1.2495	3.83081	3.02897	0.662743	2.408	1.61817	1.61916	0.570314	0.951802	1.38695	0.0469759	0.670714	0.183639
铅钡-风化														
平均值	34.518	0.975706	0.14507	2.43106	0.724796	3.92971	0.573456	2.04353	38.1319	10.7877	4.31036	0.376974	0.0568623	0.994862
方差	292.182	3.78132	0.0457206	2.80744	0.454651	11.3991	0.498662	6.1141	244.284	78.2474	18.1655	0.0625342	0.0534981	12.8871
标准差	17.0933	1.94456	0.213824	1.67554	0.674278	3.37625	0.70616	2.47267	15.6296	8.84575	4.2621	0.250068	0.231297	3.58986
高钾-风化														
平均值	94.331	0	0.544579	0.873246	0.197751	1.93785	0.265901	1.56843	0	0	0.281271	0	0	0
方差	2.33858	0	0.166078	0.198397	0.078939	0.779067	0.0039803	0.733417	0	0	0.0370439	0	0	0
标准差	1.52924	0	0.407526	0.445418	0.280961	0.882647	0.0630899	0.856398	0	0	0.192468	0	0	0
铅钡-无风化														
平均值	54.6896	0.792921	0.269157	1.25334	0.499592	3.29506	0.961685	1.6191	24.1208	10.8813	0.968287	0.301739	0.0657052	0.281595
方差	190.595	2.30812	0.158595	2.03032	0.28217	2.07151	2.00688	6.15553	78.2435	51.6629	2.87951	0.0942417	0.0237983	0.951547
标准差	13.8056	1.51925	0.39824	1.42489	0.531197	1.43927	1.41665	2.48103	8.84554	7.18769	1.69691	0.306988	0.154267	0.975473

图 5

平均值，方差，标准差的数据如图 5 所示。

4.2.3 预测化学成分含量

在上一小问中得到了风化前后的五数概括数据，由于五数能够很好的反应的变量分布的特性，并且由于风化的物理化学因素，经历相同情况下的风化并不会让变量变为逆序，因此有理由认为，变量风化前的化学成分可以使用风化前后相邻五数所组成区间的比例拟合。

设数据的五数，即最小值、第一 4 分位数、中位数、第二 4 分位数、最大值，分别为 min、Q1、mid、Q3、maxx，则可令风化时的数据五数为 $\min_1$、$Q1_1$、mid_1、$Q3_1$、$maxx_1$，风化前的数据五数为 $\min_2$、$Q1_2$、mid_2、$Q3_2$、$maxx_2$，则五数将数据分为四组，假设风化时的化学成分含量为 X1，风化前的化学成分含量为 X2，若风化时的数据落在 $\min_1$~$Q1_1$之间，则有近似关系：$\frac{Q1_1-min_1}{Q1_2-min_2}=\frac{X_1}{X_2}$，同理，若 X1 落在其他三组中，也有类似的近似关系。

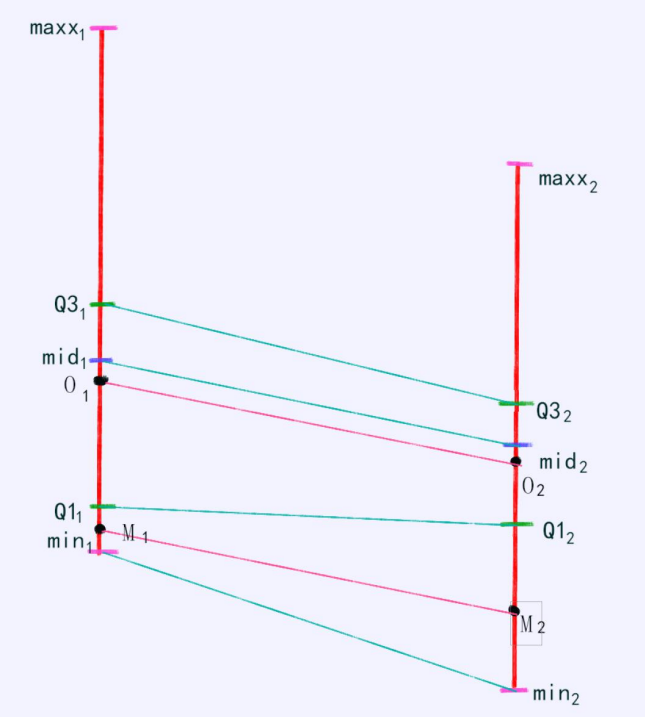

图 6

如图 6 所示，图中 O1 点和 M1 点为风化前的点，按照其所属区间等比例投到风化后的区间中，O2 和 M2 即为其所对应的风化后的点，由此也可以从风化后推出风化前的值。

使用程序进行计算，得到的结果见附件（第一问第三小问预测数据表.xls）。

4.3 问题二的分析与求解

4.3.1 对于铅钡玻璃、高钾玻璃的分类规律的分析

4.3.1.1 初步文字概况分析

高钾和铅钡的分类，由问题一中求得的统计规律可以直接分别看风化和无风化的化学成分含量以及分布规律情况。

可以直接观察到的是，部分成分（氧化铅(PbO),、氧化钡(BaO)、二氧化硅(SiO2)等）在四个分类之间的差别明显

在无风化的时候，铅钡玻璃中氧化钡(BaO)的化学成分含量在 10%附近震荡，最高可以超过 20%，而反观高钾玻璃中氧化钡(BaO)化学成分趋向平稳在 5%附近。同样对于铅钡玻璃，氧化铅(PbO)的含量集中在 20%到 40%，但是对于高钾玻璃中的氧化铅(PbO),成分在 10%以下，差距明显。

在风化的情况下，铅钡玻璃中的氧化铅(PbO),氧化钡(BaO)成分再次明显增高，并且远大于高钾玻璃。氧化铅(PbO)数值最高达到了 70%左右并且氧化钡(BaO)达到了 10%~20%，高钾玻璃中的这几种成分平稳趋向 0 值。而对于二氧化硅(SiO2)，高钾玻璃中某些文物的该成分集中逼近 100%，而铅钡玻璃种的二氧化硅(SiO2)含量较为离散，较为均匀的分布在 0 到 70%的区间内。

而部分成分（氧化锡(SnO2)、氧化钾(K2O)、氧化锶(SrO)、氧化铁(Fe2O3)、

氧化镁(MgO)）差距不大

这些化学成分，根据图表可以明显看出，无论对于高钾玻璃还是铅钡玻璃，抑或是在风化前后，这几种成分均不占主要成分并且变动范围不明显。

高钾和铅钡的分类除了体现在化学成分组成的区别上，其所包含颜色也有一定的不同：

对于高钾玻璃类型，风化过后颜色只有蓝绿色一种；而在其无风化状态时，颜色有蓝绿、浅蓝、深蓝三种。

对于铅钡玻璃类型，风化后的颜色有黑、蓝绿、浅蓝、浅绿、深绿、紫；其无风化状态时有深蓝、浅蓝、深绿、浅绿、紫、绿。

高钾玻璃类型的颜色主要以蓝为主并且较为单一，铅钡玻璃类型的颜色较杂，种类较多。

4.3.1.2 聚类算法分析

聚类算法是一种无监督的算法，是一种广泛使用的数据分析技术，不同的聚类算法对于不同的问题的表现是各有优势，也就是说没有使用于所有数据的单一最佳方法。

我们使用了多种聚类算法(K-means, Mean-Shift, DBSCAN)来进行分类任务并且对结果进行评判，其中 DBSCAN 在本数据上表现的结果最优。

DBSCAN 是一种基于密度来判断的无噪声应用空间聚类算法，该算法的主要优点是能找到任意形状以及带有噪声的簇（也就是异常值），其主要思想在于，若是一个点接近于该簇内的许多点，呢么就可以认为该点属于该簇。

DBSCAN 主要有两个关键参数，ESP 和 MinPts；ESP 为指定邻域之间的距离，若是两点之间距离小于 ESP 则可以认定两点是可相互到达的；MinPts 为集群的最小数据点数，即若 $P\_num > minPts$，则认为该样本是核心点。

该算法确定好 ESP 以及 MinPts 后，随机选择一个起点，查看当前邻域中点数是否满足 MinPts 的数目，若满足则确立该点为中心点，不满足则该点标记为噪声，一旦开始形成簇（假设形成的簇为簇 A），初始点邻域内的所有点都将成为簇 A 的一部分。如果这些新点也是核心点，则它们所代表的簇里的所有点也将添加到集群 A。然后一直随机直到所有点都存在于集群中。

DBSCAN 所使用的样本距离为欧式距离，欧式距离与曼哈顿距离是两种很常见的衡量数据样本距离的方法，假设有样本 A(a1,a2,…,an)和样本 B(b1,b2,…,bn)，那么 A 与 B 的欧式距离为：

$$d(A,B) = \sqrt{(a1-b1)^2 + (a2-b2)^2 + \cdots + (an-bn)^2}$$

曼哈顿距离：

$$d(A,B) = |a1-b1| + |a2-b2| + \cdots + |an-bn|$$

聚类算法进行完后，需要通过评判来决定该聚类方法是否适用，对于聚类结果的评判，同类之间的数据距离越近，聚类就越准确，反之，数据之间的距离越远越好

这里我们使用了 Calinski-Harabaz 评价模型，对于有 K 个簇的聚类，

Calinski-Harabaz 的分数 S 被定义为组间离散与组内离散的比率，该分值越大说明聚类效果越好

$$S(K) = \frac{T_r(B_K)}{T_r(W_K)} * \frac{N-K}{K-1}$$

其中 B_K 是组间离散矩阵，W_K 是组内离散矩阵

$$W_K = \sum_{q=1}^{k} \sum_{x \in C_q} * (x - c_q)(x - c_q)^T$$

$$B_K = \sum_{q} n_q (c_q - c)(c_q - c)^T$$

变量	含义
K	簇的数目
N	数据量大小
$\boldsymbol{C_q}$	簇 q 点集
c_q	簇 q 的聚类中心点
$\boldsymbol{N_q}$	簇 q 中点的量
c	所有数据集中心点

表 4

由于各种化学成分数据是一个高维变量，普通展示方式较难清晰展示出分类结果，我们使用了 t-SNE 降维工具，t-SNE 是一种可视化高维数据的工具，它将高维数据点之间的相似度成分转化成为了联合概率的成分，并且最小化低维嵌入和高维数据的联合概率之间的 Kullback-Leibler 散度，由于初始化所用的随机种子为当前时间，因此每次显示的图像分布会有些许差别。

运行模型修改参数多次尝试，得到的 Calinski-Harabaz 值最大为 67.2314，其对应的参数为 Esp=46.5 MinPts=60 ,该模型对数据表二中的数据运行出的最终结果如图 7 所示

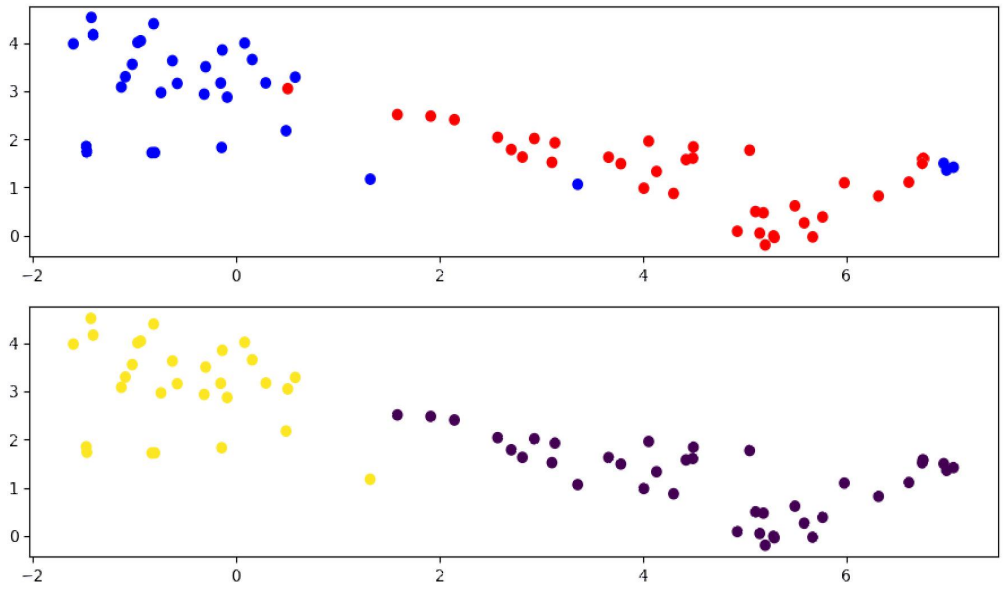

图 7

上半部分图为基于玻璃类型（高钾/铅钡）所划分的标准结果，下半部分图为程序预测分类结果，由图像可以看出，其分类结果较好，分类正确率达到了 93%，比较精准的达到了分类要求，模型的合理性较好。

4.3.2 亚类划分

首先明确，我们认为通过化学成分进行亚类划分应该以颜色为划分亚类的基准，因为不同的颜色也可以代表其主成成分中含有不同的成分，而纹饰在第一问的结果中表现为和组成成分为相互独立的，因此从化学成分的角度考虑的化选择了颜色作为亚类基准，划分亚类的流程图如图 8 所示

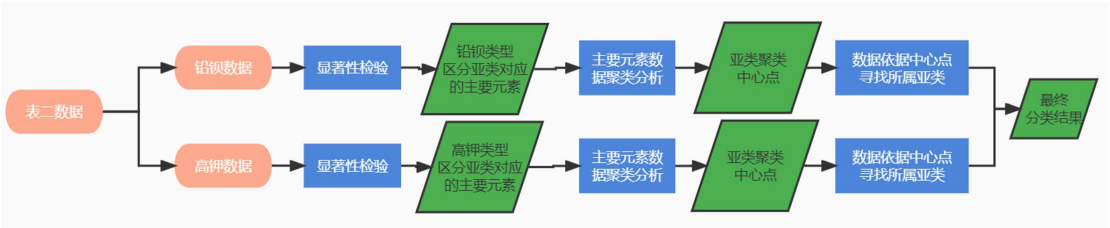

图 8

首先将表二中数据根据主要成分（高钾文物数据，铅钡文物数据）区别分为两组，对其各个化学成分做进行显著性检验；这样做的目的在于，我们所做的亚类分析是基于其颜色变量的，因此需要知道对颜色变量有显著作用的化学成分，由于文物的本身属性原因（通常在制作时颜色只会涂在文物表面），导致其所代表的成分含量应该较少，若是不分类可能其显著性会被文物主要组成成分所覆盖；并且高钾和铅钡的颜色组成也有着有较大的差异（高钾只有四种，铅钡有八种），因此我们想排除玻璃的主要成分对颜色的分类产生额外的影响。

得到两种类型亚类分类的主要成分后我们按照着几种变量分别进行了聚类分析，得到聚类的中心点，然后就可以通过计算得到各个数据到中心点的距离，从而可以划分其归属于哪个亚类

具体操作时，我们首先将数据表里的颜色从黑、绿、蓝绿、浅蓝、浅绿、深蓝、深绿到紫以此赋值 1 至 8，并将无颜色属性的数据赋值为 0，这样，就运用颜色将已经按照玻璃类型来分类的数据分为了十八组。在利用数据之前，由于数据中有定类数据,故采用非参数检验中的 Kruskal-Wallis 显著性检验,利用 SPSS 软件分别对高钾文物、铅钡文物的数据进行 Kruskal-Wallis 显著性检验，结果如下：

检验统计[a,b]

	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al2O3)	氧化铁(Fe2O3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P2O5)	氧化锶(SrO)	氧化锡(SnO2)	二氧化硫(SO2)
克鲁斯卡尔-沃利斯 H (K)	4.308	16.958	6.986	8.755	2.093	2.316	3.118	5.150	3.910	1.361	2.662	1.992	17.000	.962
自由度	3	3	3	3	3	3	3	3	3	3	3	3	3	3
渐近显著性	.230	.001	.072	.033	.553	.510	.374	.161	.271	.715	.447	.574	.001	.810

a. 克鲁斯卡尔-沃利斯检验

图 9

检验统计[a,b]

	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al2O3)	氧化铁(Fe2O3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P2O5)	氧化锶(SrO)	氧化锡(SnO2)	二氧化硫(SO2)
克鲁斯卡尔-沃利斯 H (K)	6.434	11.459	7.716	13.323	17.535	22.836	15.699	18.392	2.742	15.287	11.483	10.702	20.167	24.745
自由度	8	8	8	8	8	8	8	8	8	8	8	8	8	8
渐近显著性	.599	.177	.462	.101	.025	.004	.047	.018	.949	.054	.176	.219	.010	.002

a. 克鲁斯卡尔-沃利斯检验

图 10

高钾文物的数据见图 9，铅钡文物的数据见图 10。

由表可知，在高钾文物中，氧化纳、氧化钙、氧化锡的含量对文物颜色有着显著作用；在铅钡文物中，氧化镁、氧化铝、氧化铜、氧化锡、二氧化硫对文物颜色有着显著作用。

由此可以想到通过氧化纳、氧化钙、氧化锡的含量将高钾文物分为四类（蓝绿，浅蓝，深蓝，深绿），通过氧化镁、氧化铝、氧化铜、氧化锡、二氧化硫将铅钡文物分成八类（浅绿，浅蓝，深绿，深蓝，紫，绿，蓝绿，黑），这次分类利用的是 K-means 聚类的方法，K-means 算法解决的问题是，在事先不知道如何分类的情况下，让程序根据距离的远近，把 N 个对象最优的划分为 k 个类。仍利用 SPSS 进行 K-means 聚类，得到的结果如图 11 所示：

最终聚类中心

	聚类			
	1	2	3	4
氧化钠(Na2O)	.0000000000	2.820758364	.0000000000	.0000000000
氧化钙(CaO)	.0000000000	8.527211909	.9061848491	6.282218010
氧化锡(SnO2)	2.426735219	.0000000000	.0000000000	.0000000000

图 11

聚类成员

个案号	聚类	距离
1	4	.193
2	3	1.104
3	4	.346
4	4	1.130
5	4	1.334
6	3	.906
7	4	.809
8	3	.167
9	3	.285
10	3	.696
11	3	.183
12	2	.350
13	2	.631
14	2	.698
15	1	.000
16	4	1.501
17	3	.754
18	3	.045

图 12

图 12 为高钾文物中最终预测的结果，由图中数据可以看到给出了聚类中心的情况，现在已经知道了聚类中心各主成分的含量，再根据聚类成员到聚类点的距离，可大致判断每一类的颜色类别；根据程序计算，图中的聚类中心 1 代表了深蓝色；聚类中心 2 代表浅蓝色；聚类中心 3 代表深绿色；聚类中心 4 代表浅绿色。同理，也可得到铅钡文物的聚类情况，由于数据较多，其余的数据在附件中给出。

4.3.3 合理性和敏感性分析

在进行亚类划分之前先进行了非参数检验，并挑选了对颜色具有显著特征的化学成分，再对这些数据进行亚分类，且在风化过程中，随着不同的化学成分的增多或者减少，对文物的颜色也可能有所影响，故在选取化学成分之前要对数据进行显著性检验，否则若数据不显著，则说明颜色与化学成分无关，不能按照此方法来进行分类。

同时，我们在进行亚类划分之前将其大类数据进行分类，有效的防止了区别大类之间的含量较高的成分对亚类分类计算时进行妨碍。

而且我们通过验算检验了其分类结果，在高钾大类中的亚类分类中只有一个数据与原数据不符，正确率为 95%，在铅钡大类中的亚类分类中有三个数据与原数据不符，正确率为 93.7%，两个大类的亚类分析正确率都比较高。结果如图 13 所示。

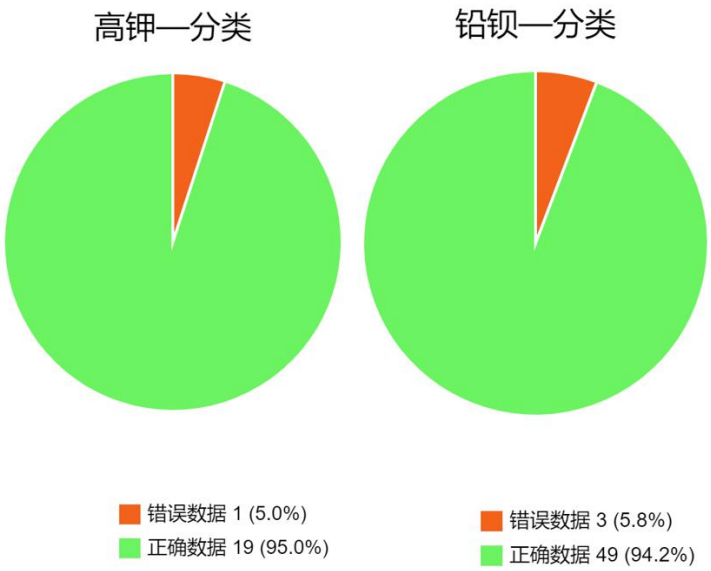

图 13

因此，我们有理由认为所做模型合理。

敏感度分析是指在给定的一组假设情况下自变量的不同取值如何影响特定的因变量。因此我们使用程序随机对数据进行了改动后重新输入模型，在 30%的浮动范围区间内随机并且对随机数据进行标准化，最终所有的改动后数据与未改动时的结果一致，由此可以看出模型的稳定性较强，以及我们对化学成分做的显著性检验是正确的。

4.4 问题三的分析与求解

问题三的分析流程如图 14 所示

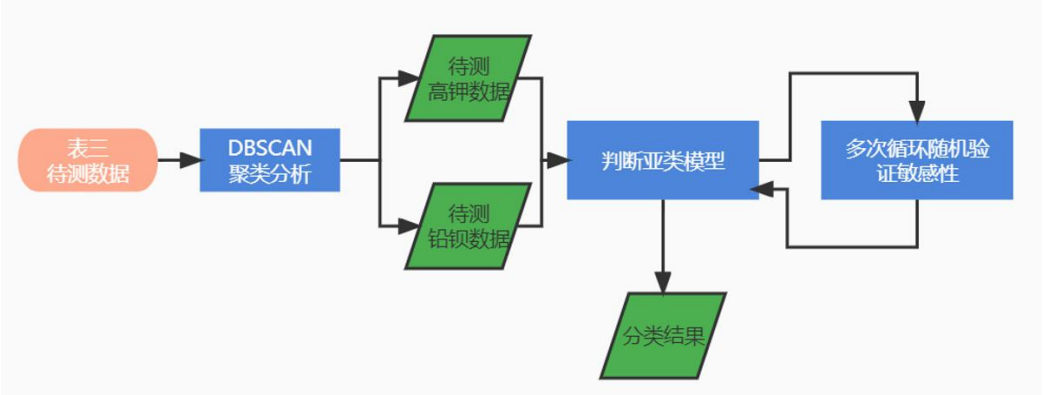

图 14

首先是运用问题二所作的 DBSCAN 聚类进行大类分类，处理得到高钾和铅钡的两类数据后通过问题二中所作的亚类分析模型进行亚类判断，由于题目要求对分类结果的敏感性进行分析，因此进行了多次循环随机改变输入变量进行敏感性验证。

4.4.1 大类分类结果

高钾-铅钡的分类结果如图 15 所示

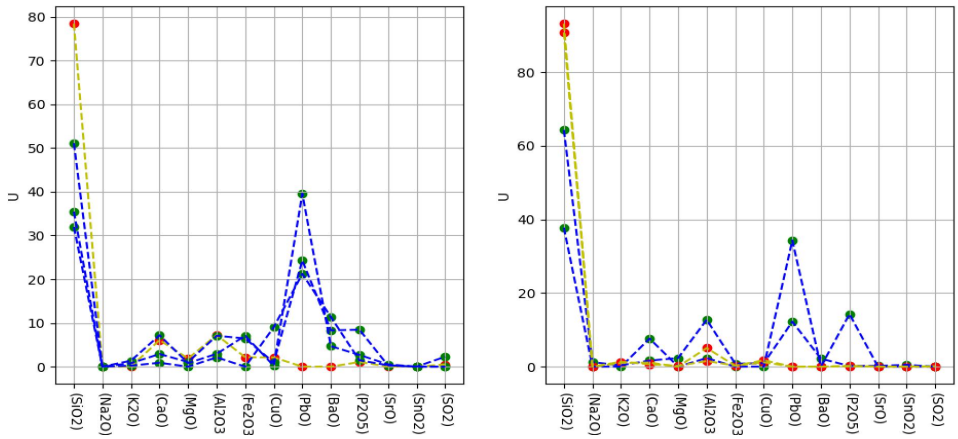

图 15

图中左图为无风化的数据，右图为风化的数据，其中蓝线绿点的折线表示其预测结果为铅钡类，黄线红点的折现表示其预测结果为高钾类。

4.4.2 亚类分类

文物编号	表面风化	预测主要类型	预测亚类类型
A1	无风化	高钾	深绿色
A2	风化	铅钡	深绿色
A3	无风化	铅钡	紫
A4	无风化	铅钡	浅蓝
A5	风化	铅钡	浅蓝
A6	风化	高钾	浅绿色
A7	风化	高钾	浅绿色
A8	无风化	铅钡	深绿色

表 5

经过模型预测，附件三中的数据所对应的的亚类类型如表 5 所示。

4.4.3 敏感性分析

循环使用程序对数据进行了 30%范围内的随机调整，观察重复输入模型的结果见图 16

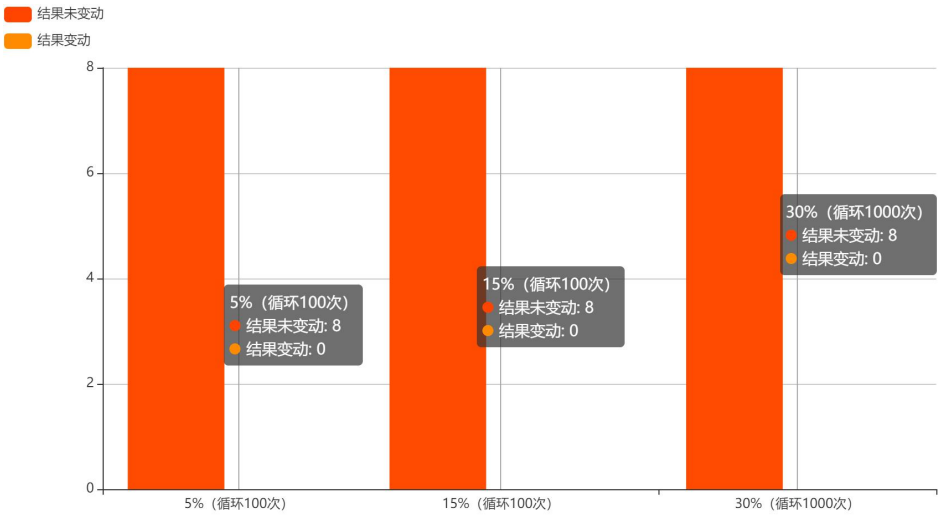

图 16

分别以多种范围随机并且多次循环后均无结果变动，表明模型十分稳定。

4.5 问题四的分析与求解

首先根据玻璃进行文物类型的划分，即高钾文物和铅钡文物，再将经过预处理的数据表 2 中的数据按照高钾和铅钡来分成两组，比较这两组数据的相关性和差异性。

利用 SPSS 软件进行斯皮尔曼相关系数的检验，得到十四组相关系数以及 sig 值，由于篇幅限制，现只展示通过检验的相关系数，结果如图 17、18：

相关性

			高钾氧化铝(Al2O3)	铅钡氧化铝(Al2O3)
斯皮尔曼 Rho	高钾氧化铝(Al2O3)	相关系数	1.000	.544*
		Sig.（双尾）	.	.020
		N	18	18
	铅钡氧化铝(Al2O3)	相关系数	.544*	1.000
		Sig.（双尾）	.020	.
		N	18	49

*. 在 0.05 级别（双尾），相关性显著。

图 17

相关性

			高钾氧化铝(Al2O3)	铅钡氧化铝(Al2O3)
斯皮尔曼 Rho	高钾氧化铝(Al2O3)	相关系数	1.000	.544*
		Sig.（双尾）	.	.020
		N	18	18
	铅钡氧化铝(Al2O3)	相关系数	.544*	1.000
		Sig.（双尾）	.020	.
		N	18	49

*. 在 0.05 级别（双尾），相关性显著。

图 18

即氧化铝和氧化锡的含量在两种不同玻璃类型中呈现一定的正相关关系，这说明了这两种化学成分在不同玻璃类型中的含量差别不大。

接下来，将高钾类型的化学成分含量数据定为第一类，将铅钡类型的化学成分含量数据定为第二类，仍然利用 SPSS 软件进行 Kruskal-Wallis 检验，得到下表：

检验统计[a,b]

	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al2O3)	氧化铁(Fe2O3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P2O5)	氧化锶(SrO)	氧化锡(SnO2)	二氧化硫(SO2)
克鲁斯卡尔-沃利斯 H(K)	30.276	.908	22.307	2.335	.549	3.381	5.302	3.078	39.083	33.692	1.626	22.890	.264	.278
自由度	1	1	1	1	1	1	1	1	1	1	1	1	1	1
渐近显著性	.000	.341	.000	.126	.459	.066	.021	.079	.000	.000	.202	.000	.607	.598

a. 克鲁斯卡尔-沃利斯检验

图 19

由图 19 可知，只有二氧化硅含量、氧化钾含量、氧化铁含量、氧化铅含量、氧化钡含量、氧化锶含量在两种不同玻璃类型中有显著差异，可以利用这些化学成分含量的不同来判断文物属于何种玻璃类型。对于其他既不相关，也没有显著差异的化学成分，可能是由于这些化学成分本身就含量极少，如氧化锡、二氧化硫，也有可能是因为在复杂风化过程的不确定因素造成的化学成分的减少或增多。

五、模型评价

5.1 模型总体评价

本文首先分析了各种数据的类型，接着根据不同的数据类型利用列联表来或者斯皮尔曼系数分析各类数据之间的相关性，利用五数概况来预测风化前的化学成分，并用显著性检验来检验结果的合理性。分类时，运用 DBSCAN 聚类和 K-means 聚类进行分类，并利用题中所给的已知的分类情况，即玻璃类型和风化状态，来检验分类的准确性。

5.2 模型优缺点

5.2.1 模型优点

（1） 利用斯皮尔曼系数来判断相关性，可以不必关注数据的线性性质，而重点

关注数据的相关性。
（2） 由于所给数据部分不符合正态分布，利用 Kruskal-Wallis 检验这种非参数检验，可以避免由于数据的非正态性而导致的结果不符合实际情况。
（3） 在分亚类时，运用 K-means 聚类方法，这种方法简单易操作，可以推广至更复杂的情况。

5.2.1 模型缺点

（1） 在运用五数概况预测风化前的化学成分时，由于各化学成分含量的箱线图有明显的不稳定现象，可能会导致最终的预测结果精度较低。
（2） 由于文物采样点的随机性，无法得知各个文物采样点的风化程度，这种情况也会降低预测结果的精度以及影响后续的分类标准。

六、参考文献

[1]崔剑锋,吴小红,谭远辉,王永彪.湖南沅水流域战国时期楚墓出土古代玻璃器的成分分析[J].硅酸盐学报,2009,37(11):1909-1913+1918.

[2]温睿,赵志强,马健,王建新.新疆哈密巴里坤西沟遗址 1 号墓出土玻璃珠的科学分析[J].文物,2016(05):92-97.DOI:10.13619/j.cnki.cn11-1532/k.2016.05.010.

[3]胡志中,李佩,蒋璐蔓,王通洋,杜谷,杨波.古代玻璃材料 LA-ICP-MS 组分分析及产源研究[J].岩矿测试,2020,39(04):505-514.DOI:10.15898/j.cnki.11-2131/td.201909210134.

[4]赵华玲. 淄博地区古代玻璃历史发展的研究[D].苏州大学,2008.

[5]茆诗松,程依明,濮晓龙. 概率论与数理统计[m].3 版 . 北京 : 高等教育出版社,2019.