

基于 K-Means 聚类与 XGBoost 的玻璃成分分析与鉴别

摘要

玻璃，我国丝绸之路早期贸易往来的宝贵物证。随着我国各地考古事业的推进，越来越多的玻璃文物被发现。值得注意的是，目前在预测玻璃文物风化前化学成分含量，玻璃文物的亚类划分，未知类型玻璃文物的归类及探究玻璃文物化学成分之间关联关系等问题的处理上，仍然存在一定的问题。本文将以铅钡和高钾两种玻璃文物为例，依据收集到的玻璃文物的基本信息，已分类及未分类玻璃文物的化学成分比例建立相关模型来解决上述问题。

对于问题一，主要需要对数据进行预处理、以合适的方式探求与玻璃文物表面风化有关的因素、统计分析文物样品表面有无风化的化学成分统计规律及预测文物风化前的化学成分含量。由于数据存在多维、稀疏的特征，我们选择将空缺行的数据删除以减少对后续问题求解的影响。对于子问题一，我们发现纹饰、颜色及玻璃类型为定类变量，我们选择使用卡方检验进行求解；对于子问题二，我们统计计算了风化前后各化学成分的均值，筛选出变化率较大的化学成分，并分类绘制箱线图，对其风化前后化学成分含量的变化进行可视化分析，具体见图 3，图 4，图 5 及图 6。对于子问题三，我们引入风化前后变化比的概念，对数据进行计算处理后，得出计算结果，具体结果见图 7 和图 8。

对于问题二，主要需要依据文物化学成分给出文物类别划分依据并验证该分类依据的合理性及敏感性。由于数据存在大量空缺，我们将空缺值视为该化学成分低于仪器可检测到的最低值，并将所有空缺值以 0.04 替代，减少对后续建模处理的影响。对于子问题一，我们首先筛选出高钾和铅钡玻璃风化前后的各化学成分比例，并绘制柱状体进行可视化处理以便分析，具体处理结果见图 9 和图 10。对于子问题二，我们分析认为该问题属于数据的分类处理问题，故我们采用 K-means 算法解决相关问题。为选定合理 k 值，我们使用肘部法则及可视化进行分析，具体处理结果见表 4 和表 5。

对于问题三，主要需要对未知文物类型的玻璃文物进行分类并验证分类的合理性和敏感性。为合理解决该问题，我们建立了 XGBoost 模型，将数据集划分为训练集和测试集，利用 XGBoost 算法得出了未知文物的分类结果，具体分类结果见表 6。为验证该分类结果的敏感性，我们调整了部分化学成分的含量，并再次对未知类型的玻璃文物进行分类处理，具体结果见表 7。

对于问题四，主要需要分析不同类别的玻璃文物中化学成分的关联关系及化学成分间的差异性。为合理探求各化学成分间的关系，我们选择采用斯皮尔曼相关性分析并绘制热力图进行分析，具体结果见图 13 和图 14。

最后我们对所建立的模型进行合理性分析，同时对模型进行进一步推广。

关键词：可视化；XGBoost；K-means；斯皮尔曼相关性分析

一、问题的提出

1.1 问题背景

玻璃，其在西亚和埃及地区往往被制作成珠形饰品，早期通过丝绸之路传入我国。为使玻璃在我国可以顺利生产，我国吸收其技术并将本土化改造，使之与外来玻璃制品外观相似而化学成分不同。

玻璃的主要成分为二氧化硅，炼制时往往需要助熔剂，而助熔剂的不同往往会导致**玻璃的成分比例不同**。我们据此将玻璃分类，例如：铅钡玻璃（**铅矿石**为助熔剂），钾玻璃（**草木灰**等含钾量较高的物质为助熔剂）。示意图如图 1 所示。

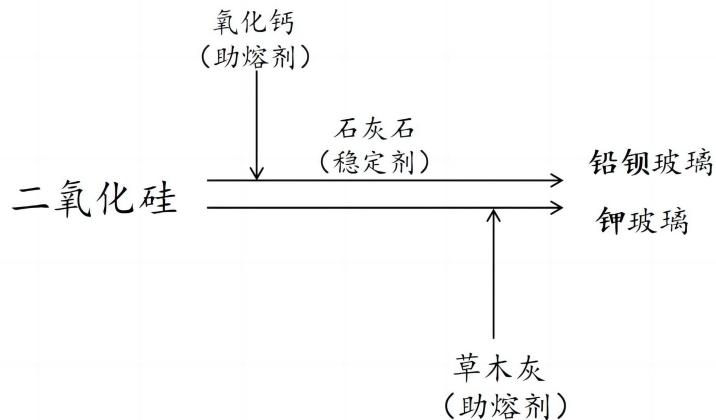


图 1 玻璃制造示意图

古代玻璃极易受埋藏环境的影响而风化，在此过程中**玻璃的内部和外部元素会发生大量交换**，其成分比例会发生一定的变化，对玻璃类型的判别产生一定的干扰，故我们需要利用已有数据建立通用，可靠的模型以判断玻璃所属类别。

1.2 问题要求

- **问题一：**（1）分析**玻璃类型、纹饰和颜色**与表面风化是否有关；（2）根据**玻璃类型**，分析文物表面有无风化的统计规律；（3）根据**风化点监测数据**预测出玻璃在风化前的各化学成分含量。
- **问题二：**（1）根据**表单 1、2 中的数据**分析并得出玻璃的分类规律；（2）**依据分类规律**，使用合适的方法，在各类别选择恰当的化学成分进行亚类划分并展示结果；（3）分析上述结果的**合理性和敏感性**。
- **问题三：**（1）分析**表单 3**中未知类别玻璃的化学成分并鉴别其所属类别；（2）分析结果的**敏感性**。
- **问题四：**（1）分析**不同类别的玻璃文物样品**化学成分之间的关联关系；（2）比较不同类别之间化学成分关联关系的差异性。

二、问题的分析

2.1 问题的整体分析

该问题是一个关于玻璃文物风化前后化学成分比例分析，建立判定玻璃文物类别模型的问题。

从分析目的看，本题需要在风化导致的化学成分比例偏移的情况下鉴别玻璃文物的类别。其中任务包括了：1、分析风化对不同类型的玻璃文物化学成分含量的影响，并能通过风化后成分预测风化前成分；2、选择合适的化学成分对玻璃文物合理分类，确定分类标准；3、选择位置类别的玻璃文物进行化学成分分析，检验分类模型的稳定性和包容性。

从数据来源、特征看，本题的数据来源于考古工作者。数据包括：玻璃文物的基本信息；已分类玻璃文物的化学成分比例；未分类玻璃文物的化学成分比例。这些数据具有多维，稀疏的特性且其成分比例累加和应为 100%。基于本题中数据的特征，应对数据进行一定的预处理。

从模型的选择看，本题的数据量较小，且需预测玻璃中的成分含量并对玻璃进行分类，因此本文并未采用过于复杂的模型，而是建立了风化后化学成分变化比和 k-means 聚类两类模型。

从编程软件的选择看，本题为数据分析类，需要将相关数据进行预处理、分析及可视化，并依据各设问建立不同类别的模型，因此我们选择基于 Python 的 Jupyter Notebook 对问题进行求解，其交互式的编程方式较为轻量化，相较于其它软件更为方便高效。

2.2 问题一的分析

问题一的核心目的有以下几点：其一，对附件中的数据进预处理，剔除不符合要求的数据；其二，分析并得出文物样品表面的有无分化化学成分含量统计规律；其三，依据统计规律推算出玻璃文物风化前化学成分含量。对于已给定的数据集，数据在完整性方面存在着一定的缺陷，故不可由原始数据进行直接分析，须对数据进行预处理。我们选择直接剔除不符合要求的数据。此外我们发现玻璃类型、纹饰和颜色均为定类数据，故采用卡方检验对表面风化与以上三者的关系进行研究。为了直观展现文物样品表面有无风化化学成分含量的统计规律，我们选择绘制两种玻璃有无风化时各化学成分占比的均值和各化学成分占比的箱线图。针对预测玻璃文物风化前化学成分含量这一问题，我们分析认为将附件中表单 1 与 2 的数据相互关联，建立风化前后化学成分变化比例模型以合理预测。

2.3 问题二的分析

问题二的核心目的在于根据文物的化学成分提出合理的文物分类的的依据并分析其合理性及敏感性。对于子问题一，我们选择对高钾和铅钡类型的玻璃数据绘制可视化图像，通过图像分析高钾、铅钡玻璃的分类规律；对于子问题二，我们分析采用 K-means 聚类分析，通过肘部法则选择合适的 k 值，对高钾、铅钡玻璃进行亚分类。为了

验证该模型是否足够稳定即验证风化文物原始成分的预测误差是否会造成明显的影响，我们微调了样本数据以测试其敏感性。

2.4 问题三的分析

问题三的核心目的在于利用未分类的玻璃文物以进一步验证分类模型的稳定性。我们分析选用 **XGboost 模型**以预测未知玻璃文物对应类型，通过轻微改变成分占比来测试分类模型的敏感性。

2.5 问题四的分析

对于该问题我们认为其核心目的在于以合理的方式提取数据中的有效特征。我们分析认为若需知晓两种类型的玻璃制品化学成分之间的关联关系，我们需要进行斯皮尔曼相关性分析并画出热力图，同时结合两个玻璃类型之间化学成分的横向对比，最终得出其化学成分内部关联关系的差异性。

三、模型的假设

- **假设一：**假设成分比例数据中的空缺值是由于成分含量过低且低于仪器所能检测到的含量下限造成并默认记为 0.4。
- **假设二：**假设各未分类玻璃文物的化学成分比例数据分别来自于各文物的某一采样点。
- **假设三：**假设所给化学成分数据的检测手段有效，结果准确无误。
- **假设四：**假设忽略颜色、纹饰、玻璃种类三者之间的内部影响。

四、符号说明

符号	符号说明
χ^2	卡方
i	自由度
y	因变量实际值
\hat{y}	因变量预测值
$L^{(t)}$	目标函数
r_s	斯皮尔曼相关系数

注：这里并未列出其余变量，这是由于它们在不同小节处有不同的含义，同时该表中也未列出专有定义的变量，这些变量在使用时会在相应位置进行详细说明。

五、模型的建立与求解

对于本题，本文模型的建立与求解部分主要分为数据的准备，模型的建立、求解、结果分析。¹

- **数据的准备：**由题设可知，当成分比例累加和介于 85%~105% 之间的数据视为有效数据，分析附件二后我们发现第 15 号及第 17 号采样点玻璃文物的化学成分比例累加和均小于 85%，显然为无效数据，应将其剔除。对于数据集中的空缺数据，我们认为是由于相关化学成分含量过低且低于仪器所能检测下限造成的，我们将所有的空缺数据以仪器检测含量下限 0.04% 填充。
- **模型的建立、求解、结果分析：**对于给定的数据集，我们依据其特点建立合适的模型，研究并量化分析玻璃文物风化与其相关特征的关系；有无风化的化学成分比例的统计规律及玻璃文物的分类及亚类划分依据。同时还需要对位置类别的玻璃文物进行类别划分，并解释该划分方式的合理性和稳定性。最后依据不同类别的玻璃文物分析化学成分之间的关联关系及差异性。

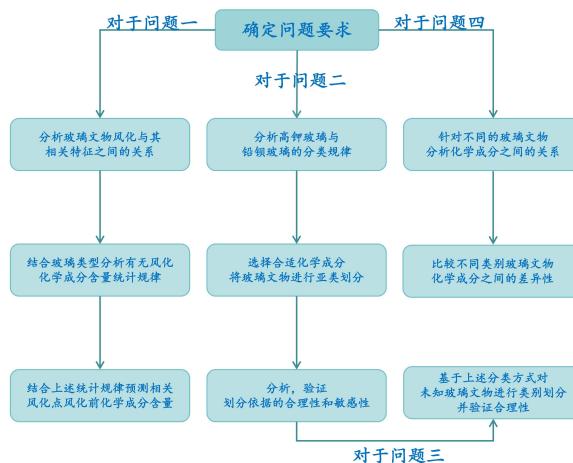


图 2 问题解决流程

5.1 问题一模型的建立与求解

对于问题一，我们将其分为三个子问题进行求解。对于子问题一，分析表单数据，我们发现纹饰、玻璃类型、颜色、表面风化程度均为定类变量，因此我们选用卡方检验对其求解。即统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度决定卡方值的大小。在进行卡方检验时需提出两个互为对立的假设：原假设和备择假设。原假设是我们想要推翻的假设，即：**玻璃文物表面风化与玻璃类型，纹饰和颜色无关**，反之则为**备择假设**。我们通常先假设随机变量相互之间独立。然后通过观察样本数据来判断是否可以拒绝原假设，从而推断两个变量之间是否存在相

¹本文所有可视化图示均为矢量图，若读者在阅读时发现图示字体过小，可适当放大 PDF 页面，详细查看图示数据等。此外，本文所有图示、表格均已交叉引用，读者阅读 PDF 时可点击对应图表，进行跳转。

关系。其中皮尔逊卡方统计量，计算公式如下

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i} \quad (1)$$

两者差异越大，卡方值越大，若 $P < a$ ，则拒绝原假设，若 $P \geq a$ ，则不拒绝原假设。（本题 $a=0.05$ ）我们首先分别统计了纹饰，玻璃类型和颜色类别下风化和未风化的玻璃文物数量。其中纹饰类别下玻璃文物得到风化情况及相应计算结果如表 1 所示；颜色类别下玻璃文物的风化情况及相应计算结果如表 2 所示；不同玻璃类型对应玻璃文物的风化情况及相应计算结果如表 3 所示，

表 1 纹饰类别下玻璃文物的风化情况

纹饰	风化	未风化
A	11	11
B	6	0
C	17	13
卡方计算值	4.9565	
P 计算值	0.0839	

表 2 颜色类别下玻璃文物的风化情况及相关计算结果

颜色	风化	未风化
黑	2	0
蓝绿	9	6
绿	0	7
浅蓝	12	8
浅绿	1	2
深蓝	0	2
深绿	4	3
紫	2	2
卡方计算值	6.2871	
P 计算值	0.5066	

表 3 不同玻璃类型对应的玻璃文物风化情况及相应计算结果

玻璃类型	风化	未风化
铅钡	28	12
高钾	6	12
卡方计算值	5.4518	
P 计算值	0.0195	

由于计算过程中涉及到自由度的计算,在此处补充自由度的计算方法: **自由度** = [行数-1]*[列数-1]。

从卡方检验的结果可知,对于玻璃类型这一类别, $P=0.0195$, P 小于 a , 有理由拒绝原假设,即认为是否风化与玻璃类型有关联性。而对于纹饰和颜色这两个变量, P 分别为 0.0839 和 0.5066, P 均大于 a , 故没有理由拒绝原假设,即认为是否风化与纹饰和颜色均没有关联性。

对于子问题二,我们首先将预处理过的数据按照玻璃类型分类计算风化前后各化学成分的平均值,由于数据量较大,具体结果详见附录图示部分中图 15,图 16,图 17和图 18。通过比较两组数据可知在所有化学成分中二氧化硅、氧化钾、氧化钙、氧化铝、氧化铁、五氧化二磷这五个化学成分的含量风化前后变化较大。为了使结果更直观,我们分别将组成两类玻璃的化学成分与各自对应的风化情况绘制为箱线图。

其中高钾玻璃如图 3, 图 4所示,

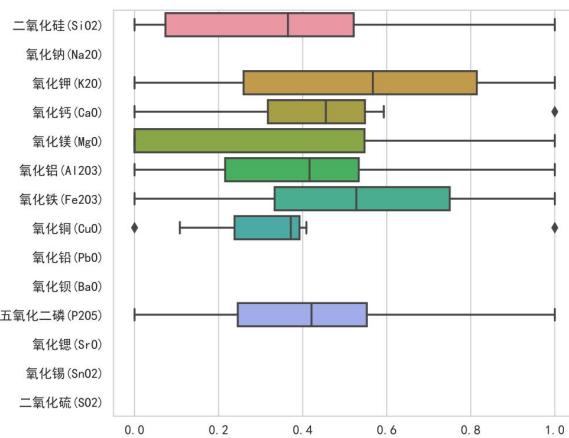


图 3 高钾风化箱线图.jpg

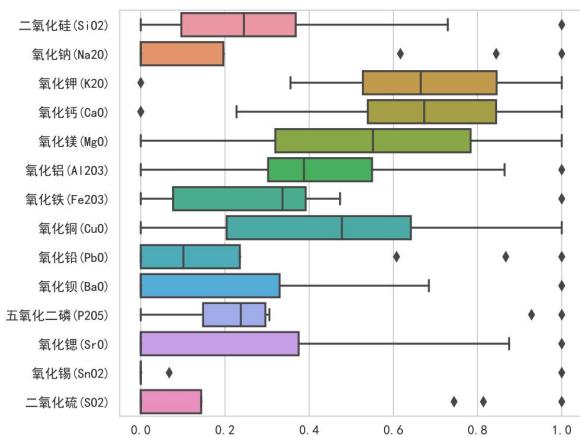


图 4 高钾未风化箱线图.jpg

铅钡玻璃如图 5及图 6所示,

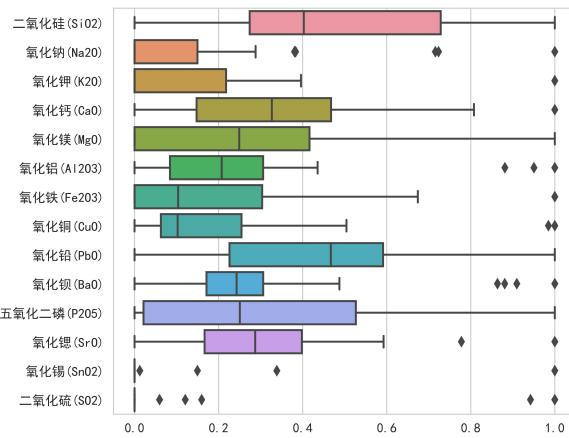


图 5 铅钡风化箱线图

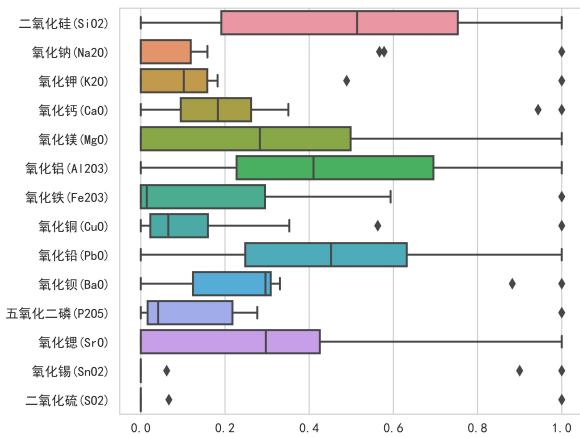


图 6 铅钡未风化箱线图

通过观察比较上图,我们发现对于高钾玻璃,在风化后二氧化硅、氧化铁和五氧化二磷的含量明显上升,而氧化铜、氧化镁和氧化钙的含量明显下降。而铅钡玻璃,在风

化后五氧化二磷、氧化铜和氧化钙的含量小幅上升，而氧化硅和氧化铝的含量小幅降低。此外，我们发现高钾类型的玻璃在风化后其主要化学成分呈下降趋势，铅钡类型的玻璃在风化后其主要化学成分呈上升趋势。

对于子问题三，我们通过附件可知每块玻璃文物由二氧化硅等共 14 个化学成分组成，我们设风化前某类型玻璃的某化学成分为 a_i ，其中 ($i=1, 2 \dots 14$)，风化后其含量为 b_i 。则风化前某类型玻璃的某化学成分的占比我们设为 A_i ，同理风化后其含量占比记为 B_i 。为合理预测风化前化学成分含量，我们引入风化前后变化比 M_i 的概念。风化前后变化比由分化前后的占比计算得出，具体公式如下

$$M_i = \frac{A_i - B_i}{B_i} \quad (2)$$

其中 A_i 和 B_i 的计算公式分别为 $\frac{a_i}{\sum_{i=1}^{14} a_i}$, $\frac{b_i}{\sum_{i=1}^{14} b_i}$ 。

经计算，铅钡玻璃（图 7）和高钾玻璃（图 8）的预测结果如下，

	二氧化硅 (SiO ₂)	氯化钠(Na ₂ O)	氯化钾(K ₂ O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氯化铜(CuO)	氯化钡(BaO)	五氧化二磷 (P ₂ O ₅)	氯化锶(SrO)	氧化锶(SrO)	二氧化硫 (SO ₂)	
02	58.371594	0.03990186	1.25801543	0.20176981	0.20176981	0.159771754	0.02124746	0.02124746	0.02124746	0.02124746	0.02124746	0.02124746	0.02124746	
08	31.7719343	0.03990186	0.06371631	0.72164309	0.025341688	1.16661465	0.060180288	0.87393661	18.3944154	0.0391871	0.0427446	0.0427446	0.0427446	
09	69.6188321	0.734176282	10.01831517	3.853082062	0.196939104	4.584974825	2.370254133	2.465337784	0.433754723	0.632910588	1.734027823	0.062447178	0.23628662	0.133333164
10	70.90101434	0.734176282	15.62177959	1.305076182	0.196939104	2.813507279	1.9258831483	1.337679832	0.433754723	0.632910588	0.198174608	0.062447178	0.23628662	0.133333164
12	69.0397894	0.734176282	14.1499971	4.474546911	0.196939104	5.071260034	2.140842808	2.627585383	0.433754723	0.632910588	0.743154781	0.062447178	0.23628662	0.133333164
22	67.66258834	0.734176282	12.31635445	3.151025658	1.659301631	0.509405949	0.540501516	1.544704379	0.433754723	0.632910588	0.104041664	0.062447178	0.23628662	0.133333164
27	67.9336783	0.734176282	0.679207808	5.841769578	2.658677899	8.7183991	1.481408833	2.452413024	0.433754723	0.632910588	1.783571475	0.062447178	0.23628662	0.133333164
B	0.943312251	0.000401566	0.005584862	0.008734063	0.002242079	0.019375561	0.002660303	0.01567781	0.00401566	0.00401566	0.00287789	0.000401566	0.000401566	0.000401566
A	0.691141836	0.007730508	0.04949317	0.05429173	0.011038818	0.019705519	0.067300036	0.009739215	0.019494886	0.24671978	0.07111567	0.006283403	0.001004528	0.006358388
C	-0.267324436	17.354407006	15.9801952	5.214648487	3.92347759	2.473465777	6.407044166	0.59247599	9.843866076	14.8227647	3.954365208	0.561179451	4.907165489	2.33332909

图 7 铅钡玻璃预测结果

文物采样点	二氧化硅(SiO ₂)	氯化钠(Na ₂ O)	氯化钾(K ₂ O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氯化铜(CuO)	氯化钡(BaO)	五氧化二磷 (P ₂ O ₅)	氯化锶(SrO)	氧化锶(SrO)	二氧化硫 (SO ₂)	
07	67.8677375	0.734176282	0.679207808	6.649673882	0.196939104	6.877462238	1.259197508	5.159622207	0.433754723	0.632910588	3.022162777	0.062447178	0.23628662	0.133333164
09	69.6188321	0.734176282	10.01831517	3.853082062	0.196939104	4.584974825	2.370254133	2.465337784	0.433754723	0.632910588	1.734027823	0.062447178	0.23628662	0.133333164
10	70.90101434	0.734176282	15.62177959	1.305076182	0.196939104	2.813507279	1.9258831483	1.337679832	0.433754723	0.632910588	0.198174608	0.062447178	0.23628662	0.133333164
12	69.0397894	0.734176282	14.1499971	4.474546911	0.196939104	5.071260034	2.140842808	2.627585383	0.433754723	0.632910588	0.743154781	0.062447178	0.23628662	0.133333164
22	67.66258834	0.734176282	12.31635445	3.151025658	1.659301631	0.509405949	0.540501516	1.544704379	0.433754723	0.632910588	0.104041664	0.062447178	0.23628662	0.133333164
27	67.9336783	0.734176282	0.679207808	5.841769578	2.658677899	8.7183991	1.481408833	2.452413024	0.433754723	0.632910588	1.783571475	0.062447178	0.23628662	0.133333164
B	0.943312251	0.000401566	0.005584862	0.008734063	0.002242079	0.019375561	0.002660303	0.01567781	0.00401566	0.00401566	0.00287789	0.000401566	0.000401566	0.000401566
A	0.691141836	0.007730508	0.04949317	0.05429173	0.011038818	0.019705519	0.067300036	0.009739215	0.019494886	0.24671978	0.07111567	0.006283403	0.001004528	0.006358388
C	-0.87753892	17.354407006	15.9801952	5.214648487	3.92347759	2.473465777	6.407044166	0.59247599	9.843866076	14.8227647	3.954365208	0.561179451	4.907165489	2.33332909

图 8 高钾玻璃预测结果

5.2 问题二模型的建立与求解

对于问题二，我们将其分为两个子问题进行求解。

对于子问题一，为分析出高钾和铅钡玻璃的分类规律，我们首先筛选出高钾和铅钡玻璃风化前后各化学成分的占比，并对筛选出的数据进行可视化处理。如图 9 和图 10 所示。

观察柱状图，我们发现高钾玻璃中氧化铅与氧化钡的含量远低于铅钡玻璃而其氧化钾的含量远超铅钡玻璃。

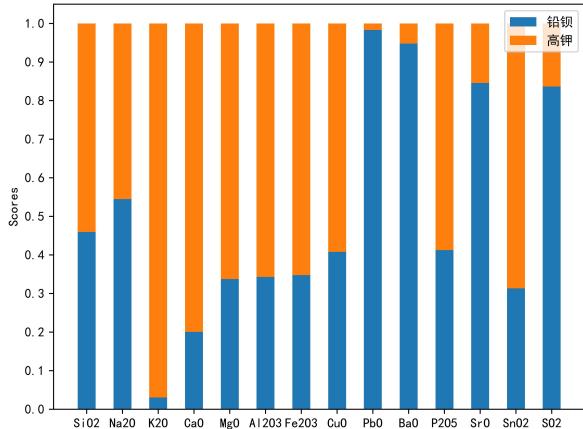


图 9 高钾玻璃风化前后各化学成分的占比

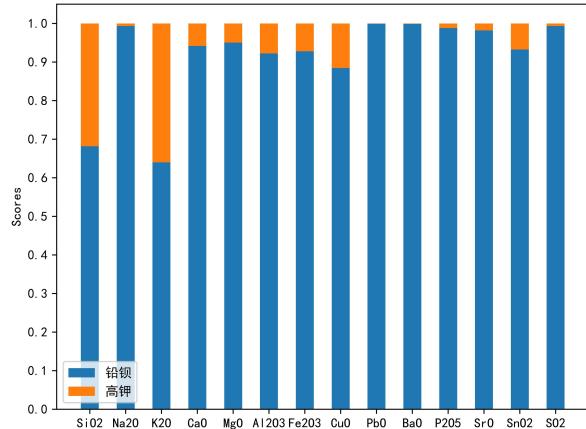


图 10 铅钡玻璃风化前后各化学成分的占比

由以上分析可知，高钾玻璃相较于铅钡玻璃而言含有更多的氧化钾，但其氧化铅和氧化钡的含量远没有铅钡玻璃高。此结果与题目所说铅钡玻璃在烧制过程中加入铅矿石作为助熔剂，高钾玻璃的助熔剂是含钾较多的草木灰一致。

对于子问题二，题目要求对各类别选择合适的化学成分对玻璃文物进行亚类划分，为数据的分类处理的问题，而**聚类分析**的方法在处理这类问题时起到重要的作用。需要注意的是，**聚类算法**属于无监督算法，其可通过样本的相似度将样本分为若干类，使得同一类内部的样本相似度尽可能高，不同类别之间的样本相似度尽可能低。本题我们采用 **k-means 算法**进行亚类划分，为了简化模型，我们剔除了含量过高和过低的数据和一些很难作为聚类依据的均匀数据。

使用 **k-means 算法**，其目的是将目标数据点分成类簇，找到每个簇的中心并使其度量最小化。过程为将数据集聚集成 k 个类簇，从数据集中随机选择 N 个数据点作为数据中心，分别计算出每个点到每个数据中心的距离，并将每个点划分到离其最近数据中心的类簇，在数据中心聚集了一些点后，重复上述过程，选出新的数据中心，比较第一次和第二次得到的数据中心。若两个数据中心之间的距离小于某一临界值，则此聚类达到了期望，算法终止；若距离相差很大，则继续执行算法，直到算法终止。

K-means 算法首先需要从给定的数据对象中随机指定初始聚类数 k 和相应的初始聚类中心 C 。然后计算从初始聚类中心到其余数据对象的距离。本文选择欧氏距离进行计算。从聚类中心到空间中其他数据对象的欧氏距离公式为

$$d(x, C_i) = \sqrt{\sum_{j=1}^m (x_j - C_{ij})^2} \quad (3)$$

其中 x 为数据对象， C_i 是第 i 个距离中心， m 为数据对象的维度， x_j 为数据 x 的第 j 个维度与聚类中心 C_i 的属性值。

根据欧氏距离，测量相似度，并将与聚类中心相似度最高的目标数据分配到 C_i 类别。同类别之后，对 k 个聚类中的数据对象进行平均计算，形成新一轮的聚类中心，从

而降低数据集的误差平方和 (Sum of Square Error, SSE), 其计算公式如下

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} |d(x, C_i)|^2 \quad (4)$$

SSE 被用来衡量聚类结果的效果。当其不再变化或收敛时，即停止迭代，得到最终结果。

为了更好地对玻璃类型进行亚类划分聚类，我们需要确定合适的 k 值，因此我们使用肘部法则，绘制肘部法则可视化，确定最适 k 值，如图 12 和图 11 所示。

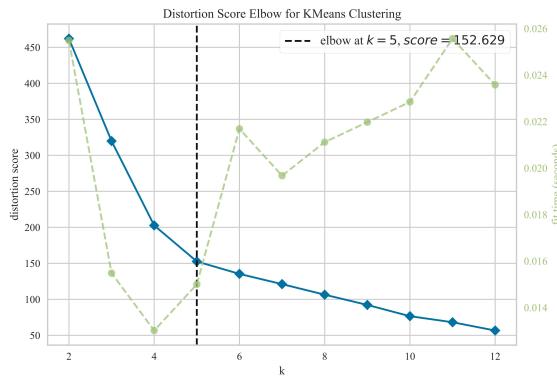


图 11 针对高钾玻璃的肘部法则可视化

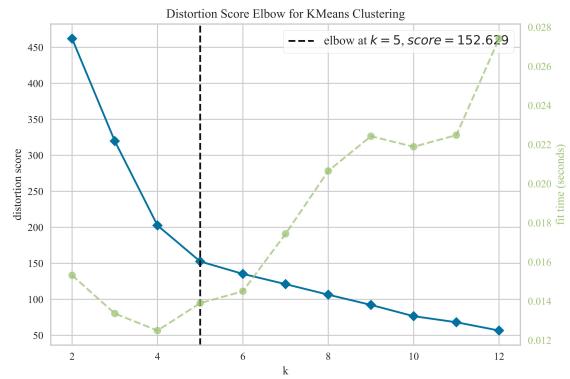


图 12 针对铅钡玻璃的肘部法则可视化

根据上图，我们可以选择出最适 k 值，即 $k=5$ 。据此我们对高钾玻璃和铅钡玻璃进行亚类划分，如表 4 和表 5 所示。

表 4 高钾玻璃亚类划分结果

Class	文物编号
0	1、7、8、9、10、14、16、17
1	11、12、13
2	5
3	6
4	0、2、3、4、15

表 5 铅钡玻璃亚类划分结果

Class	文物编号
0	3、4、13、14、15、25、28、29、35、38、39、40、43、48
1	12、30、34
2	1、2、7、8、9、10、16、17、18、19、20、21、22、23、24、37、41、45、46、47
3	6、26、27、33
4	0、5、11、31、32、36、42、44

5.3 问题三模型的建立与求解

问题三要求我们分析未知类别玻璃文物的化学成分以确定其所属类别，首先我们分析表单三中的数据是否为有效数据，发现此表单数据均为有效数据。经过综合考虑，我们建立了**极端梯度提升 (eXtreme Gradient Boosting, XGBoost)** 模型，对表单三中的未知文物进行合理的分类。

XGBoost 算法是一种基于树模型的优化模型，该算法通过多次迭代，生成一个新的树模型用于优化前一棵树模型，随着迭代次数的增多，该模型的预测精度也会相应提高^[1]。

记通过数据处理后的数据集特征为 $R(x_{ij})_{m \times n}$ ，表示其包含 m 天的游戏情况， n 个特征，在训练中形成的 CART 数的集合记为 $F = \{f(x) = w_{q(x)}, q : \mathbf{R}^n \rightarrow T, w \in \mathbf{R}^T\}$ ， q_{zhongq} 为树模型的叶节点决策规划， T 为某一树模型叶节点数量， w 为叶节点对应得分^[2]。对于预测的 y 值，即 \hat{y} ，计算公式为

$$\hat{y} = \varphi(x_i) = \sum_{k=1}^K f_k(x_i) \quad (5)$$

XGBoost 算法在每一次迭代过程中会保存前面所学习的模型，会将这些模型加入到新一轮迭代过程中，因此我们记第 i 个模型的预测结果为

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (6)$$

XGBoost 算法的目标函数计算公式如下

$$L^{(t)} = \sum_{i=1}^n l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 + const \quad (7)$$

上述公式中， l 为模型误差损失，描述在该模型下预测值与实际值之间的出差异损失， Ω 为模型叶节点的正则项惩罚系数， γ 和 λ 为模型的超参数^[2]。

通常情况下，我们难以用枚举法得到在模型中所训练出来的树结构，因此这里采用贪婪算法，从单子叶结点开始，通过迭代方法，将其加入到树结构中，从而得到最优解，其计算公式^[3]如下

$$\zeta_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (8)$$

其中 $I_j = \{i | q(x_i) = j\}$ 为叶节点 j 上的样本集合^[2]，且有

$$g_i = \partial_{\hat{y}(t-1)} l\left(y_i, \hat{y}_i^{(t-1)}\right) \quad (9)$$

$$h_i = \partial_{\hat{y}(t-1)}^2 l\left(y_i, \hat{y}_i^{(t-1)}\right) \quad (10)$$

依据上述分析，我们首先以表单三中的文物为分类目标，再将表单二的数据集按比例 7: 3 划分为训练集与测试集并通过 **XGBoost 算法**对训练集和测试集进行训练和评估。结果如表 6 所示。

表 6 未知玻璃文物分类结果

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
类别	高钾	铅钡	铅钡	铅钡	铅钡	高钾	高钾	铅钡

为检验该模型的敏感性，我们改变了部分文物的部分化学成分比例。调整后的分类结果如表 7 所示。

表 7 调整后未知玻璃文物分类结果

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
类别	高钾	高钾	铅钡	铅钡	铅钡	高钾	高钾	高钾

由上述调整后的结果可见此模型的敏感性较高，可靠性较高。

5.4 问题四模型的建立与求解

问题四需要我们针对不同类别的玻璃文物样品，分析化学成分之间的关联关系及不同类别之间的化学成分关联关系的差异性。关联关系是指对两个或两个以上存在相关性的因素进行分析，通过计算相关系数并比较，从而判断它们之间的关联程度，得到它们的联系与变化规律。这里我们采用斯皮尔曼相关性分析。

斯皮尔曼相关性分析可以用来评估两个变量相关性的强烈程度，以及它们之间的方向性关系。其原理为考察两个变量的均值和标准偏差，将各自的大小值依次排列，然后计算排列后的值之间的差值 (Δ)，最后依据斯皮尔曼相关系数计算公式算得斯皮尔曼相关系数 (r_s)，该系数可以判断两个变量正相关，负相关或无关。

斯皮尔曼相关系数计算公式如下

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (11)$$

其中 d_i 为两组数据之间的等级差。

通过上述公式的计算我们得到了相关结果，为了更加直观的展现我们的结果，我们绘制了如图 13 和图 14 所示的热力图。

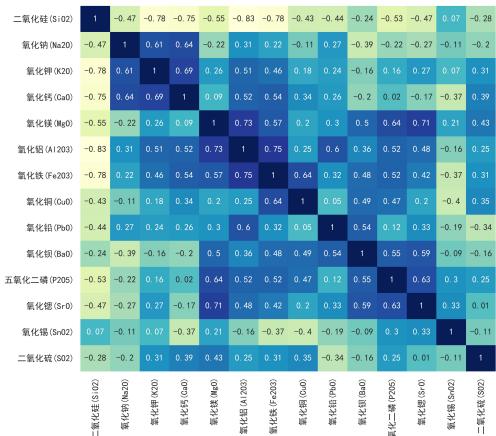


图 13 高钾玻璃化学成分热力图

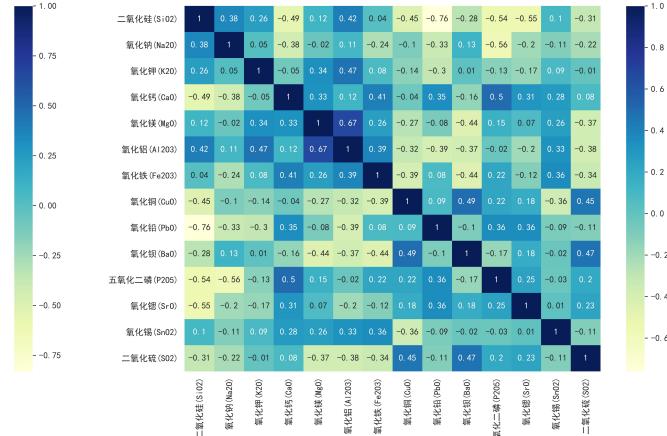


图 14 铅钡玻璃化学成分热力图

我们发现**高钾玻璃化学成分热力图**的色阶图的颜色均较深，我们猜测高钾玻璃各化学成分之间的相关性都较高。其中除主对角线外颜色均较浅，所以高钾玻璃各化学成分之间的关联性都较低。从图中分析可知，高钾玻璃中，两种化学成分呈明显正相关的有氧化镁与氧化铝，呈明显负相关的有二氧化硅与氧化铅。其他成分的相关性较低，可认为其基本没有联系。

对于**铅钡玻璃化学成分热力图**，我们发现除主对角线外颜色均较浅，所以铅钡玻璃各化学成分之间的关联性都较低。从图中分析可知，铅钡玻璃中化学成分呈明显正相关的有氧化镁与氧化铝，呈明显负相关的有二氧化硅与氧化铅。其他成分的相关性较低，可认为其基本没有联系。

对于不同类别之间的化学成分关联关系的差异性，我们通过对高钾玻璃和铅钡玻璃的热力图，颜色差异较大处对应的两个化学成分斯皮尔曼相关系数数值上相差较大，即两种玻璃关于这两种化学成分的相关性差异较大。由于涉及化学成分较多，此处选取一组具有明显差异性的化学成分进行分析作为示例：两种玻璃中二氧化硅和氧化铝的相关性虽然接近，但在铅钡玻璃中两者为正相关性，而在高钾玻璃中两者为负相关性。

六、模型的评价与推广

6.1 模型的评价

- **模型的优点:**

1. 对数据进行预处理时，我们以 0.04 而非 0 填充空缺值，数据的误差更小且更真实。
2. 利用 K-means 聚类，结合肘部法则，选择最优聚类 k 值，聚类结果更精确。
3. 可视化分析问题，结果直观易理解。
4. 对玻璃进行聚类之后与题目的文字叙述进行对比，验证了结果的合理性。

- **模型的缺点:**

1. 在研究定类变量的关系时，由于数据较少，卡方检验的误差增大，当存在很大误差时，可使用 Fisher 精确检验代替卡方检验。
2. 数据量偏少，可能会影响模型的准确程度，易出现“过拟合”现象
3. 对于分类的结果未采用其他模型进行验证。

6.2 模型的推广

XGboost 聚类预测模型可较好地对事物进行分类并预测未知的类别，因此其有利于对已知未分类事物进行分类，并分析预测出新事物的类别。我们认为该模型可应用于物种鉴定等领域。总而言之，本文所建立模型有着泛化能力较强、精度较高的优秀品质，可为医疗，生物，考古等多行业提供有利帮助。

参考文献

- [1] 陈振宇, 刘金波, 李晨, 季晓慧, 李大鹏, 黄运豪, 狄方春, 高兴宇, 徐立中. 基于 LSTM 与 XGBoost 组合模型的超短期电力负荷预测 [J]. 电网技术, 2020, 44(02):614-620. DOI:10.13335/j.1000-3673.pst.2019.1566.
- [2] 杨贵军, 徐雪, 赵富强. 基于 XGBoost 算法的用户评分预测模型及应用 [J]. 数据分析与知识发现, 2019, 3(01):118-126.
- [3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785-794. <https://doi.org/10.1145/2939672.2939785>.
- [4] 肖杨, 李亚, 王海瑞, 常梦容. 基于皮尔逊相关系数的滚动轴承混合域特征选择方法 [J]. 化工自动化及仪表, 2022, 49(03):308-315. DOI:10.20030/j.cnki.1000-3932.202203009.
- [5] 王殿武, 赵云斌, 尚丽英, 王凤刚, 张震. 皮尔逊相关系数算法在 B 油田优选化学防砂措施井的应用 [J]. 精细与专用化学品, 2022, 30(07):26-28. DOI:10.19482/j.cn11-3237.2022.07.07.
- [6] 汪海燕, 黎建辉, 杨风雷. 支持向量机理论及算法研究综述 [J]. 计算机应用研究, 2014, 31(05):1281-1286.
- [7] 百度百科【卡方检验】[EO/BL].<https://baike.baidu.com/item/%E5%8D%A1%E6%96%B9%E6%A3%80%E9%AA%8C/2591853>.
- [8] 百家号【如何简单理解卡方检验?】[EO/BL].<https://baijiahao.baidu.com/s?id=1771127457087657402&wfr=spider&for=pc>.
- [9] 百度百科【斯皮尔曼相关性分析】[EO/BL].https://wenku.baidu.com/view/f4e06f5975c66137ee06eff9aef8941ea76e4b38.html?_wkts_=1690549232319&bdQuery=%E6%96%AF%E7%9A%AE%E5%B0%94%E6%9B%BC%E7%9B%B8%E5%85%B3%E6%80%A7%E5%88%86%E6%9E%90%E5%8E%9F%E7%90%86

附录

[A] 图示

文物编号	纹饰	类型	颜色	表面风化	文物采样点	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P ₂ O ₅)	氧化锶 (SrO)	氧化锡 (SnO ₂)	二氧化硫 (SO ₂)
20	A	铅钡	浅蓝	无风化	20	37.36	0.04	0.71	0.04	5.45	1.51	4.78	9.3	23.55	5.75	0.04	0.04	0.04	
24	C	铅钡	紫	无风化	24	31.94	0.04	0.04	0.47	0.04	1.59	0.04	8.46	29.14	26.23	0.14	0.91	0.04	
30	A	铅钡	深蓝	无风化	30部位1	34.34	0.04	1.41	4.49	0.98	4.35	2.12	0.04	39.22	10.29	0.04	0.35	0.4	
30	A	铅钡	深蓝	无风化	30部位2	36.93	0.04	0.04	4.24	0.51	3.86	2.74	0.04	37.74	10.35	1.41	0.48	0.44	
31	C	铅钡	紫	无风化	31	65.91	0.04	0.04	1.6	0.89	2.11	4.59	0.44	16.55	3.42	1.62	0.2	0.04	
32	C	铅钡	浅绿	无风化	32	69.71	0.04	0.21	0.46	0.04	2.36	1	0.11	19.76	4.88	0.17	0.04	0.04	
33	C	铅钡	深绿	无风化	33	75.51	0.04	0.15	0.64	1	2.35	0.04	0.47	16.16	3.55	0.13	0.04	0.04	
35	C	铅钡	浅绿	无风化	35	65.91	0.04	0.04	0.38	0.04	1.44	0.17	0.16	22.05	5.68	0.42	0.04	0.04	
37	C	铅钡	深绿	无风化	37	60.12	0.04	0.23	0.89	0.04	2.72	0.04	3.01	17.24	10.34	1.46	0.21	0.04	
45	A	铅钡	深蓝	无风化	45	61.28	2.66	0.11	0.84	0.74	5	0.04	0.53	15.99	10.96	0.04	0.23	0.04	
46	A	铅钡	浅蓝	无风化	46	55.21	0.04	0.25	0.04	1.67	4.79	0.04	0.77	25.25	10.06	0.2	0.43	0.04	
47	A	铅钡	浅蓝	无风化	47	51.54	4.66	0.29	0.87	0.61	3.06	0.04	0.65	25.4	9.23	0.1	0.85	0.04	
55	C	铅钡	绿	无风化	55	49.01	2.71	0.04	1.13	0.04	1.45	0.04	0.86	32.92	2.95	0.35	0.04	0.04	
均值					53.44384615	0.802307692	0.2738461514	1.237692308	0.510769231	3.194615385	0.954615385	1.563076923	23.59354615	10.49923077	0.91	0.312307692	0.098461538	0.623076923	

图 15 铅钡玻璃未风化相关数据

文物编号	纹饰	类型	颜色	表面风化	文物采样点	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P ₂ O ₅)	氧化锶 (SrO)	氧化锡 (SnO ₂)	二氧化硫 (SO ₂)	
02	A	铅钡	浅蓝	风化	02	36.28	0.04	1.05	2.34	1.18	5.73	1.86	0.26	47.43	0.04	3.57	0.19	0.04	0.04	
05	C	铅钡	紫	风化	08	20.14	0.04	0.04	1.48	0.04	1.34	0.04	10.41	28.68	31.23	3.59	0.37	0.04	2.58	
08	C	铅钡	紫	风化	08严重风化	4.61	0.04	0.04	3.19	0.04	1.11	0.04	3.14	32.45	30.62	7.56	0.53	0.04	15.03	
11	C	铅钡	浅蓝	风化	11	33.59	0.04	0.21	3.51	0.71	2.69	0.04	4.93	25.39	14.61	9.38	0.37	0.04	0.04	
19	A	铅钡	蓝绿	风化	19	29.64	0.04	0.04	2.93	0.59	3.57	1.33	3.51	42.82	5.35	8.83	0.19	0.04	0.04	
23	A	铅钡	蓝绿	风化	23未风化	53.79	7.92	0.04	0.5	0.71	1.42	0.04	2.99	16.98	11.86	0.04	0.33	0.04	0.04	
25	C	铅钡	浅蓝	风化	25未风化	50.61	2.31	0.04	0.63	0.04	1.9	1.55	1.12	31.9	6.65	0.19	0.2	0.04	0.04	
26	C	铅钡	紫	风化	26	19.79	0.04	0.04	1.44	0.04	0.7	0.04	10.57	29.53	32.25	3.13	0.45	0.04	1.96	
26	C	铅钡	紫	风化	26严重风化	3.72	0.04	0.4	3.01	0.04	1.18	0.04	3.6	29.92	35.45	6.04	0.62	0.04	15.95	
28	A	铅钡	浅蓝	风化	28未风化	68.08	0.04	0.26	1.34	1	4.7	0.41	0.33	17.14	4.04	1.04	0.12	0.23	0.04	
29	A	铅钡	浅蓝	风化	29未风化	63.3	0.92	0.3	2.98	1.49	14.34	0.81	0.74	12.31	2.03	0.41	0.25	0.04	0.04	
34	C	铅钡	深绿	风化	34	35.78	0.04	0.25	0.78	0.04	1.62	0.47	1.51	46.55	10	0.34	0.22	0.04	0.04	
36	C	铅钡	深绿	风化	36	39.57	2.22	0.14	0.37	0.04	1.6	0.32	0.68	41.61	10.83	0.07	0.22	0.04	0.04	
38	C	铅钡	深绿	风化	38	32.93	1.38	0.04	0.68	0.04	2.57	0.29	0.73	49.31	9.79	0.48	0.41	0.04	0.04	
39	C	铅钡	深绿	风化	39	26.25	0.04	0.04	1.11	0.04	0.5	0.04	0.88	61.03	7.22	1.16	0.61	0.04	0.04	
40	C	铅钡	风化	40		16.71	0.04	0.04	1.87	0.04	0.45	0.19	0.04	70.21	6.69	1.77	0.68	0.04	0.04	
41	C	铅钡	风化	41		18.46	0.04	0.44	4.96	2.73	3.33	1.79	0.19	44.12	9.76	7.46	0.47	0.04	0.04	
42	A	铅钡	浅蓝	风化	42未风化	51.26	5.74	0.15	0.79	1.09	3.53	0.04	2.67	21.88	10.47	0.08	0.35	0.04	0.04	
42	A	铅钡	浅蓝	风化	42未风化	51.33	5.68	0.35	0.04	1.16	5.66	0.24	2.72	20.12	10.88	0.04	0.4	0.04	0.04	
43	C	铅钡	浅蓝	风化	43部位1	12.41	0.04	0.04	5.24	0.89	2.25	0.76	5.35	59.85	7.29	0.04	0.64	0.04	0.04	
43	C	铅钡	浅蓝	风化	43部位2	21.7	0.04	0.04	6.4	0.95	3.41	1.39	1.51	44.75	3.26	12.83	0.47	0.04	0.04	
44	A	铅钡	浅蓝	风化	44未风化	60.74	3.06	0.2	2.14	0.04	12.69	0.77	0.43	13.61	5.22	0.04	0.26	0.04	0.04	
45	A	铅钡	风化	48		53.33	0.8	0.32	2.82	1.54	13.65	1.03	0.04	15.71	7.31	1.1	0.25	1.31	0.04	
49	A	铅钡	黑	风化	49	28.79	0.04	0.04	4.58	1.47	5.38	2.74	0.7	34.18	6.1	11.1	0.46	0.04	0.04	
49	A	铅钡	黑	风化	49未风化	54.61	0.04	0.3	2.08	1.2	6.5	1.27	0.45	23.02	4.19	4.32	0.3	0.04	0.04	
50	A	铅钡	黑	风化	50	17.98	0.04	0.04	3.19	0.47	1.87	0.33	1.13	44	14.2	6.34	0.66	0.04	0.04	
50	A	铅钡	黑	风化	50未风化	45.02	0.04	0.04	3.12	0.54	4.16	0.04	0.7	30.61	6.22	6.34	0.23	0.04	0.04	
51	C	铅钡	浅蓝	风化	51部位1	24.61	0.04	0.04	3.58	1.19	5.25	1.19	1.37	40.24	8.94	8.1	0.39	0.47	0.04	
51	C	铅钡	浅蓝	风化	51部位2	21.35	0.04	0.04	5.13	1.45	2.51	0.42	0.75	51.34	0.04	8.75	0.04	0.04	0.04	
52	C	铅钡	浅蓝	风化	52	25.74	1.22	0.04	2.27	0.55	1.16	0.23	0.7	47.42	8.64	5.71	0.44	0.04	0.04	
53	A	铅钡	浅蓝	风化	53未风化	63.66	3.04	0.11	0.78	1.14	6.06	0.04	0.54	13.66	8.99	0.04	0.27	0.04	0.04	
54	C	铅钡	浅蓝	风化	54	22.28	0.04	0.32	3.19	1.28	4.15	0.04	0.83	55.46	7.04	4.24	0.88	0.04	0.04	
54	C	高钾	浅蓝	风化	54严重风化	17.11	0.04	0.04	1.11	3.65	0.04	1.34	58.46	0.04	14.13	1.12	0.04	0.04	0.04	
56	C	铅钡	蓝绿	风化	56	29.15	1.846667	0.04	1.21	1.73	0.04	1.85	0.04	0.79	41.25	15.45	2.54	0.04	0.04	0.04
57	C	铅钡	蓝绿	风化	57	25.42	0.04	0.04	1.31	0.04	2.18	0.04	1.16	45.1	17.3	0.04	0.04	0.04	0.04	
58	C	铅钡	风化	58		30.39	0.04	0.34	3.49	0.79	3.52	0.86	3.13	39.35	7.66	8.99	0.24	0.04	0.04	
均值					67.98417	0.725	9.334167	5.339167	1.085833	6.62	1.938333	2.455833	0.428333	0.625	1.4025	0.061667	0.233333	0.131667		

图 16 铅钡玻璃风化相关数据

<table border="

文物编号	纹饰	类型	颜色	表面风化文物采样点	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P ₂ O ₅)	氧化锶 (SrO)	氧化锡 (SnO ₂)	二氧化碳 (SO ₂)
07	B	高钾	蓝绿	风化 07	92.63	0.04	0.04	1.07	0.04	1.98	0.17	3.24	0.04	0.04	0.61	0.04	0.04	0.04
09	B	高钾	蓝绿	风化 09	95.02	0.04	0.59	0.62	0.04	1.32	0.32	1.55	0.04	0.04	0.35	0.04	0.04	0.04
10	B	高钾	蓝绿	风化 10	96.77	0.04	0.92	0.21	0.04	0.81	0.26	0.84	0.04	0.04	0.04	0.04	0.04	0.04
12	B	高钾	蓝绿	风化 12	94.29	0.04	1.01	0.72	0.04	1.46	0.29	1.65	0.04	0.04	0.15	0.04	0.04	0.04
22	B	高钾	蓝绿	风化 22	92.35	0.04	0.74	1.66	0.64	3.5	0.35	0.55	0.04	0.04	0.21	0.04	0.04	0.04
27	B	高钾	蓝绿	风化 27	92.72	0.04	0.04	0.94	0.54	2.51	0.2	1.54	0.04	0.04	0.36	0.04	0.04	0.04
				均值	93.96333	0.04	0.556667	0.87	0.223333	1.93	0.265	1.561667	0.04	0.04	0.286667	0.04	0.04	0.04

图 18 高钾玻璃风化相关数据

[B] 支撑文件列表

支撑文件列表如下（按文件夹进行分类）：

文件夹名	描述
Code	解决问题所有源程序，包括 ipynb 及其对应的 py 文件
Figures	论文中所有矢量图示，均为 pdf 文件
Data	解决问题所用数据，均为 xlsx 文件
Result	程序输出结果，均为 html 文件

[C] 使用的软件、环境

C.1: 为解决该问题，我们所使用的主要软件有：

- TeX Live 2022
- Visual Studio Code 1.77.3
- WPS Office 2023 春季更新 (14036)
- Python 3.10.4
- Pycharm 2023.1.1 (Professional Edition)

C.2: Python 环境下所用使用到的库及其版本如下：

库	版本	库	版本
copy	内置库	matplotlib	3.5.2
jupyter	1.0.0	nltk	3.7
jupyter-client	7.3.1	numpy	1.22.4+mkl
jupyter-console	6.4.3	openpyxl	3.0.10
jupyter-contrib-core	0.4.0	pandas	1.4.2
jupyter-contrib-nbextensions	0.5.1	scikit-learn	0.22.2 psot1
jupyter-highlight-selected-word	0.2.0	seaborn	0.11.2
jupyterlab-pygments	0.2.2	sklearn	0
jupyterlab-widgets	1.1.0	xgboost	1.6.1
jupyter-latex-envs	1.4.6	yellowbrick	1.4
jupyter-nbextensions-configurator	0.5.0		

[D] 问题解决源程序

D.1 DataPreProcessing

```
1 #!/usr/bin/env python
2 # coding: utf-8
3
4 # In[1]:
5
6
7 import pandas as pd
8 data=pd.read_excel("Data_Wordle.xlsx",sheet_name="Sheet1")
9 data
10
11
12 # In[2]:
13
14
15 data['Date']=pd.to_datetime(data['Date'])
16 data.sort_values(by='Date',inplace=True)
17 data=data.reset_index(drop=True)
18 data
19
20
21 # In[3]:
22
23
24 import matplotlib.pyplot as plt
25 plt.rcParams['font.sans-serif'] = ['Times New Roman']
26 plt.rcParams['axes.unicode_minus'] = False
27 ax=data.plot(x='Date', y=['Number of reported results', 'Number in hard mode'])
28 plt.xticks(fontsize=12)
29 plt.yticks(fontsize=12)
30 plt.xlabel('Date', fontsize=14)
31 plt.ylabel('Person Quantity', fontsize=14)
32 plt.legend()
33 plt.tight_layout()
34 plt.savefig("figures\\报告结果每日变化.pdf")
35
36
37 # In[4]:
38
39
40 data['WordLength'] = data['Word'].apply(len)
41 data['SumRate']=data.loc[:,['1 try','2 tries','3 tries','4 tries','5 tries','6 tries','7 or more tries (X)']].sum(axis=1)
42 data['HardRate']=data['Number in hard mode']/data['Number of reported results']
43 data
```

```
44
45
46 # In[5]:
47
48
49 data[data['WordLength']!=5]
50
51
52 # In[6]:
53
54
55 ax=data.plot.scatter(x='Date', y='HardRate')
56 plt.xticks(fontsize=12)
57 plt.yticks(fontsize=12)
58 plt.xlabel('Date',fontsize=14)
59 plt.ylabel('Hard mode frequency',fontsize=14)
60 plt.tight_layout()
61 plt.savefig('figures\\每日选择困难模式人数频率变化.pdf')
62
63
64 # In[7]:
65
66
67 data['HardRateDiff']=data['HardRate'].diff()
68
69
70 # In[8]:
71
72
73 data.plot.scatter(x='Date', y='HardRateDiff',color='g')
74 plt.xticks(fontsize=12)
75 plt.yticks(fontsize=12)
76 plt.xlabel('Date',fontsize=14)
77 plt.ylabel('Hard mode frequency gradient',fontsize=14)
78 plt.tight_layout()
79 plt.savefig('figures\\每日选择困难模式人数频率变化率.pdf')
80
81
82 # In[9]:
83
84
85 data=data.fillna(0)
86 data
87
88
89 # In[10]:
```

```
90  
91  
92 data[abs(data['HardRateDiff'])>=0.02]
```
