

面向不确定性与系统复杂性的农作物种植策略优化研究

摘要

关键字： 混合整数线性规划 鲁棒优化 仿真优化 修复式遗传算法 蒙特卡洛模拟 微观经济学 市场动态反馈机制

一、问题背景

古代丝绸之路不仅是商贸通道，也是文化与技术交流的重要桥梁，其中，玻璃制品是早期东西方物质文化交流的重要物证。早期玻璃制品由西亚和埃及地区传入，其技术与风格影响了中国本土的玻璃制造业。中国古代工匠在吸收外来技术的基础上，利用本土原料进行生产，制造出外观相似但化学成分体系相异的玻璃器物。这种成分上的差异为鉴别古代玻璃制品的产地与技术来源提供了客观依据。

玻璃的主要成分为二氧化硅 (SiO_2)。为降低其熔点，制造过程中需加入助熔剂。古代中西方采用的助熔剂体系不同，形成了成分各异的玻璃类别。例如，以草木灰为助熔剂的钾玻璃 (K_2O 含量较高) 和以铅矿石为助熔剂的铅钡玻璃 (PbO 、 BaO 含量较高)，后者被普遍认为是古代中国独立发展的玻璃品种。然而，玻璃制品在长期埋藏过程中，其表面易与环境发生元素交换，导致化学成分发生改变，这一风化过程给准确的成分分析与类型鉴别带来了挑战。因此，需要建立一套系统的数据分析方法，以消除或减弱风化作用的干扰，准确识别玻璃文物的化学成分规律，并对其进行科学分类与鉴别。



图 1 问题背景

二、问题重述

问题一：分析玻璃文物表面风化状态与其玻璃类型、纹饰和颜色等物理属性的统计关系。在此基础上，结合玻璃类型，研究表面风化对化学成分含量的影响规律，并建立数学模型，根据风化样品的化学成分数据，预测其风化前的成分含量。

问题二：根据已分类的高钾玻璃与铅钡玻璃的化学成分数据，建立有效的分类判据。进而，在每个大类中，选择合适的化学成分作为指标，对该类别进行亚类划分，并给出具体的划分方案。最后，对分类与划分结果的合理性及稳定性进行分析。

问题三：利用已建立的分类模型，对一批未知类别的玻璃文物样品的化学成分数据进行分析，鉴别其所属的玻璃类型。同时，需要对鉴别结果的敏感性进行评估，以考察分类结果的稳健程度。

问题四：针对高钾玻璃与铅钡玻璃两个类别，分别探究其内部各化学成分之间的关联关系。通过比较两个类别在化学成分关联模式上的异同，表现不同玻璃体系在原料构成与制造工艺上可能存在的差异。

三、 问题分析

对于问题一，该问题包含两个递进的部分。第一部分要求分析风化状态与玻璃类型、纹饰、颜色等多个定性变量之间的关系，可采用列联表分析与卡方检验等统计方法，检验这些变量之间是否存在显著的相依性。第二部分旨在建立风化前后化学成分的映射关系。此过程可视为一个回归或预测问题，可以通过分析同一文物上风化点与未风化点的成分差异，建立多元回归模型，从而实现对风化前成分的定量估计。

对于问题二，其核心是分类与聚类任务。首先，区分高钾与铅钡玻璃是一个监督学习分类问题。由于类别标签已知，可利用逻辑回归、支持向量机或决策树等分类算法，建立基于化学成分的分类器。其次，在已确定的类别内部进行亚类划分，是一个无监督学习的聚类问题。因缺乏亚类的先验标签，可采用 K-均值聚类或层次聚类等算法，依据关键化学成分的分布特征进行探索性划分。对结果的合理性分析可通过交叉验证评估分类器性能，通过轮廓系数等指标评价聚类效果；敏感性分析则可通过扰动数据来检验模型输出的稳定性。

对于问题三，该问题是问题二所建分类模型的直接应用。需要将表单三中未分类样本的化学成分数据输入已训练好的分类器，以获得其预测类别。其敏感性分析旨在评估分类决策的可靠性，可以通过计算样本点到分类边界的距离或在样本成分数据上施加微小扰动，观察分类结果是否发生改变，来衡量分类的稳健性。

对于问题四，该问题要求探究不同类别玻璃内部化学成分的相互关系。此分析可通过计算各类别样本的协方差矩阵或相关系数矩阵来实现。皮尔逊相关系数是衡量两个连续变量间线性关系强度的常用指标。通过为高钾和铅钡玻璃分别构建相关系数矩阵，并利用热力图等可视化手段，可以直观地展示不同类别玻璃内部各元素间的协同或拮抗关系，比较其模式差异，为探究其工艺与原料来源提供数据支持。

四、 数据预处理与分析

为了后续模型的建立与求解，本文进行如下数据预处理，即数据清洗、数据量化等。具体机制如图2所示。

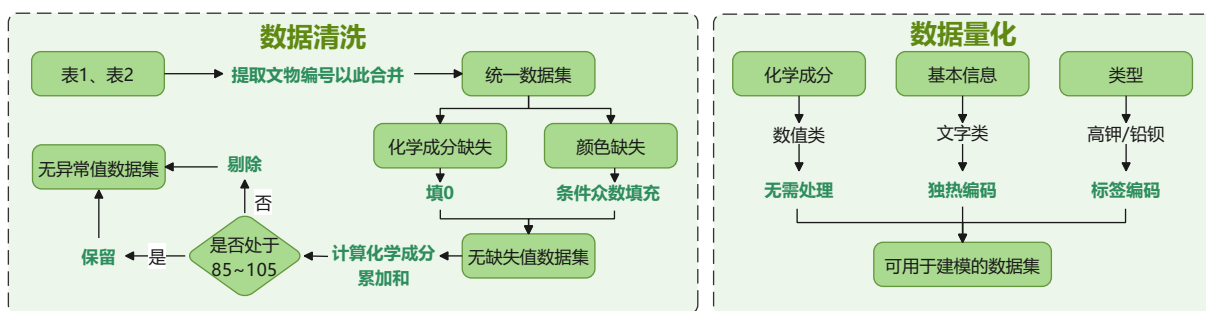


图2 数据预处理机制

4.1 数据清洗

本文首先进行了数据清洗。题中给出的附件包含三个表单，表单1包含文物编号、纹饰、类型、颜色和表面风化情况，表单2则包含了文物采样点和化学成分比例。通过观察发现，表单2的“文物采样点”列前缀的数字部分与表单1的文物编号相对应，因此我们通过正则表达式提取出其数字部分，创建出“文物编号”列，并基于此将表单1内容合并进来，形成包含完整信息的数据集。

随后我们对数据集中各变量进行缺失值统计，发现颜色和化学成分比例存在显著的缺失问题。题目中提到表单2中化学成分比例若存在空白处表示未检测到该成分，因此我们将这些缺失值填充为0，处理前后的化学成分频数如图3所示。

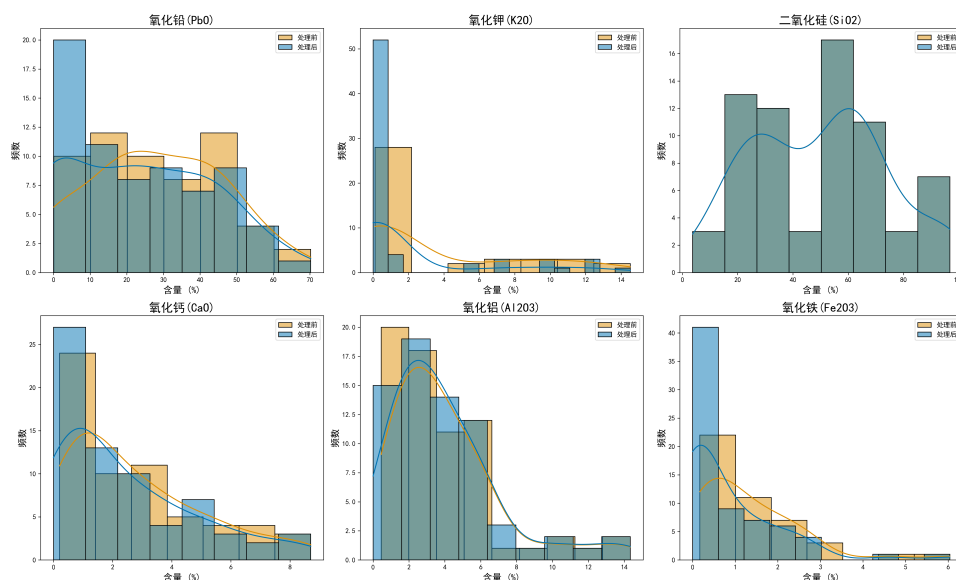


图3 化学成分缺失值填充

而对于颜色缺失，我们首先想到通过全局众数填充，但是其误差可能较大，于是我们采用了一种条件众数填充法。具体而言，对于一个颜色信息缺失的样本，系统会自动寻找数据集中所有与该样本在类型、表面风化和纹饰上完全一致的其他样本，然后用这个小群体中出现次数最多的颜色来填充缺失值。其填充内容如图??所示。

表 1 按类型和纹饰划分的颜色众数及对应文物编号

类型	纹饰	颜色众数	对应文物编号
铅钡	A	浅蓝	2, 19, 20, 23, 28, 29, 30, 42, 44, 45, 46, 47, 48, 49, 50, 53
铅钡	C	浅蓝	8, 11, 24, 25, 26, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 43, 51, 52, 54, 55, 56, 57, 58
高钾	A	蓝绿	3, 4, 5, 6, 18, 21
高钾	B	蓝绿	7, 9, 10, 12, 22, 27
高钾	C	浅蓝	1, 13, 14, 15, 16, 17

针对各种化学成分比例，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，因此我们对其进行成分性异常值检测。我们计算每个样本所有化学成分的累加和，并筛选出总和介于 85% 到 105% 之间的数据作为有效样本，得到 67 条有效已分类样本、8 条有效未分类样本以及 2 条异常样本，然后剔除异常样本。

最后我们检测了高钾玻璃和铅钡玻璃以及它们中风化和未风化的占比，如图4所示：如图4所示，在有效样本中，高钾与铅钡玻璃的比例约为 1:2.7，风化与无风化的比例接

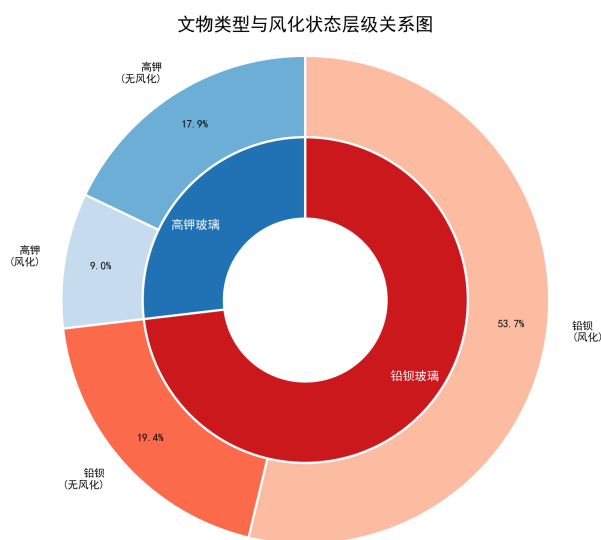


图 4 类型与风化嵌套饼图

近 1:1.4。总体来看，各类样本数量分布相对均衡，不存在极端情况，因此在后续建模中无需采用欠采样或过采样等特殊处理。

数据清洗的整体效果如图5所示。

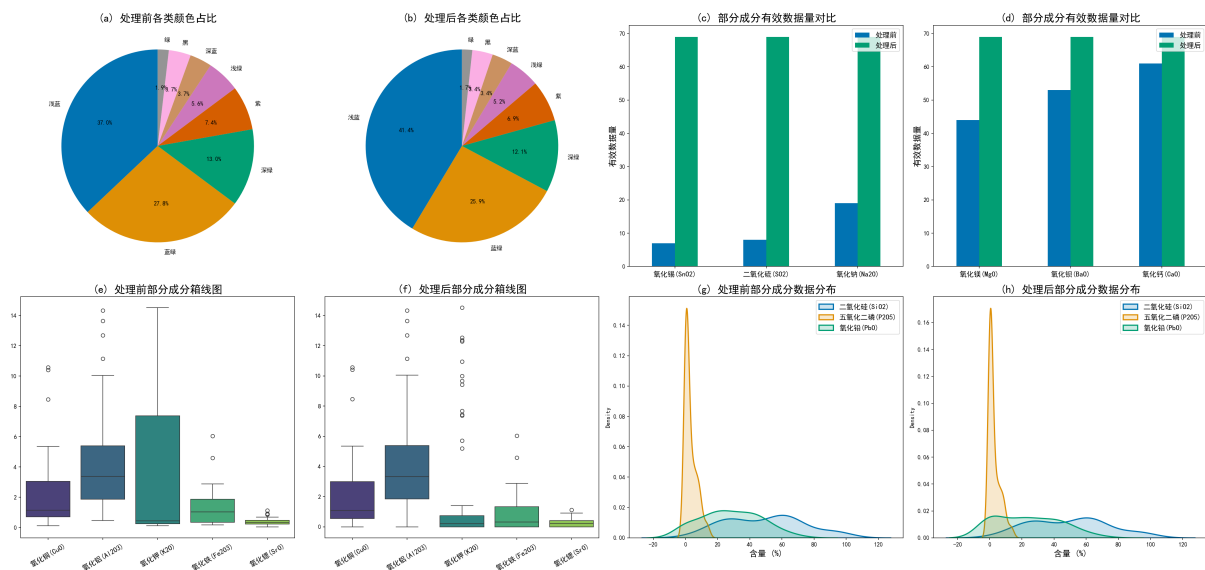


图 5 数据清洗效果

4.2 数据量化

为了便于后续相关数学模型的建立与求解，增强可理解性，本文进一步将合并后的文件中的名称进行数据量化。对于作为因变量的“类型”列，我们使用标签编码，将“铅钡”映射为 0，“高钾”映射为 1。对于作为自变量的各项分类特征列，为了避免引入虚假的、不存在的顺序和距离关系，我们排除了顺序编码和 RGB 向量的方案，最终选择了最忠实于数据本身特性的独热编码。数据量化的具体细节如表2所示：

表 2 数据量化方法示例

列名	原始数据形态	量化方法	量化后形态
类型	[高钾，铅钡]	标签编码	[1,0]
表面风化	[风化，无风化]	独热编码	[1,0]
颜色	[蓝绿，浅蓝，...]	独热编码	[00000001,00000010,...]
纹饰	[A,B,C]	独热编码	[001,010,100]
各类化学成分比例	如 69.33	无需处理	69.33

五、 问题一：风化影响分析与成分恢复模型

古代玻璃文物在长期埋藏过程中会发生风化作用，导致其表面化学成分发生改变。为探究风化作用对不同类别玻璃文物的影响，并尝试恢复其风化前的成分含量，本章建立了分析与预测模型。我们首先检验了表面风化与文物其他物理属性之间的统计关联，

然后分析了风化对高钾和铅钡两类玻璃化学成分含量的影响规律，最后构建了一个基于风化系数的预测模型，用于推断风化样品的原始化学成分，其框架如图6所示。

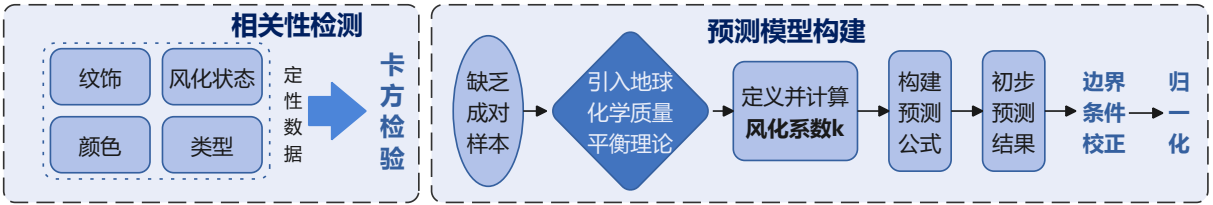


图 6 风化影响分析与成分恢复模型框架

5.1 表面风化与文物物理属性的关联性检验

为了探究文物表面风化现象是否与其物理属性存在关联，我们首先需要分析表面风化状态与玻璃类型、纹饰及颜色这几个变量之间的关系。这些变量的共同特征是它们均为分类变量，其取值为离散的类别而非连续的数值。

这一数据特性决定了用于衡量连续变量间线性关系的皮尔逊相关系数或用于比较组间均值差异的方差分析等方法在此并不适用，因为对“浅蓝”、“纹饰 A”等类别进行数值运算不具备实际意义，需要采用一种能够处理定性数据频数的非参数检验方法来分析它们之间的关联性。因此，针对此问题，我们使用了卡方检验 (χ^2 test)。

卡方检验 (χ^2 test) 是一种专门用于判断两个或多个分类变量之间是否存在关联的经典统计方法，其核心思想在于比较观测频数与期望频数之间的差异。其中，观测频数是样本数据中各类组合的实际计数值，而期望频数则是在“变量间相互独立”这一零假设下，根据边际概率计算出的理论计数值。检验过程通过计算两者差异的卡方统计量，并将其转换为 P 值来进行判断。在本研究中，我们设定显著性水平为 0.05，若计算所得的 P 值小于该阈值，则拒绝变量间相互独立的零假设，认为它们之间存在显著的统计学关联。

为直观分析文物表面风化现象与其物理属性的关联，我们制作了一组关系分析的可视化图，如图7所示。图7b展示了风化状态在两类玻璃中的分布情况，数据显示，铅钡玻璃的风化样本占其总数的 73.5%，这一比例远高于高钾玻璃的 33.3%。此外，图7a与图7c进一步展示了风化样本在不同纹饰和颜色类别下的数量分布。所有纹饰为 B 的样本均为风化样本，而纹饰为 A 的样本中风化与未风化数量相同。在不同颜色中，浅蓝色样本的风化数量为 20，远超其未风化数量 6，而蓝绿色样本中两者的数量则基本持平。这些在不同类别下风化比例与数量的显著差异，直观地表明表面风化与文物的物理属性并非相互独立。

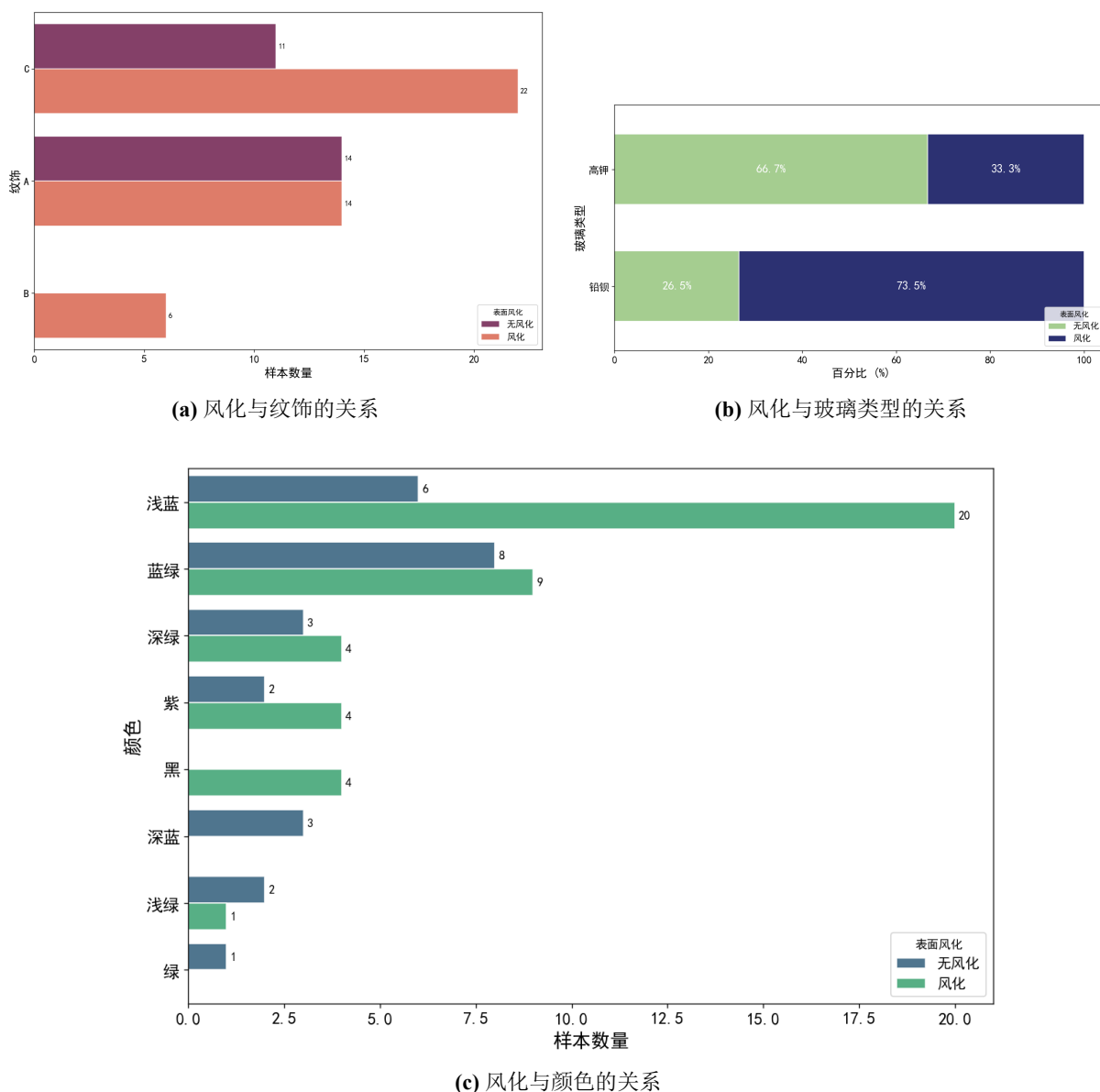


图 7 表面风化与文物物理属性的关联性可视化

5.2 风化对两类玻璃化学成分含量的影响规律

基于风化与玻璃类型存在关联的结论，我们进一步对风化在高钾和铅钡两类玻璃中引起的化学成分变化规律进行分析。我们将样本数据分为高钾未风化、高钾风化、铅钡未风化、铅钡风化四个组别，并对各组样本的化学成分含量分布进行了比较。

我们采用分面箱线图对两类玻璃在风化前后的化学成分分布进行可视化，如图8与图9所示。箱线图展示了数据的中位数、四分位距和离散程度。从图中可以观察到，对于高钾玻璃，风化作用导致氧化钾 K_2O 的含量中位数显著下降，而氧化硅 SiO_2 的含量则有上升趋势。对于铅钡玻璃，风化作用主要表现为氧化铅 PbO 与氧化钡 BaO 含量的大幅降低，同时氧化硅 SiO_2 含量相应增加。这种变化说明风化过程中发生了元素的选择性流失与富集。

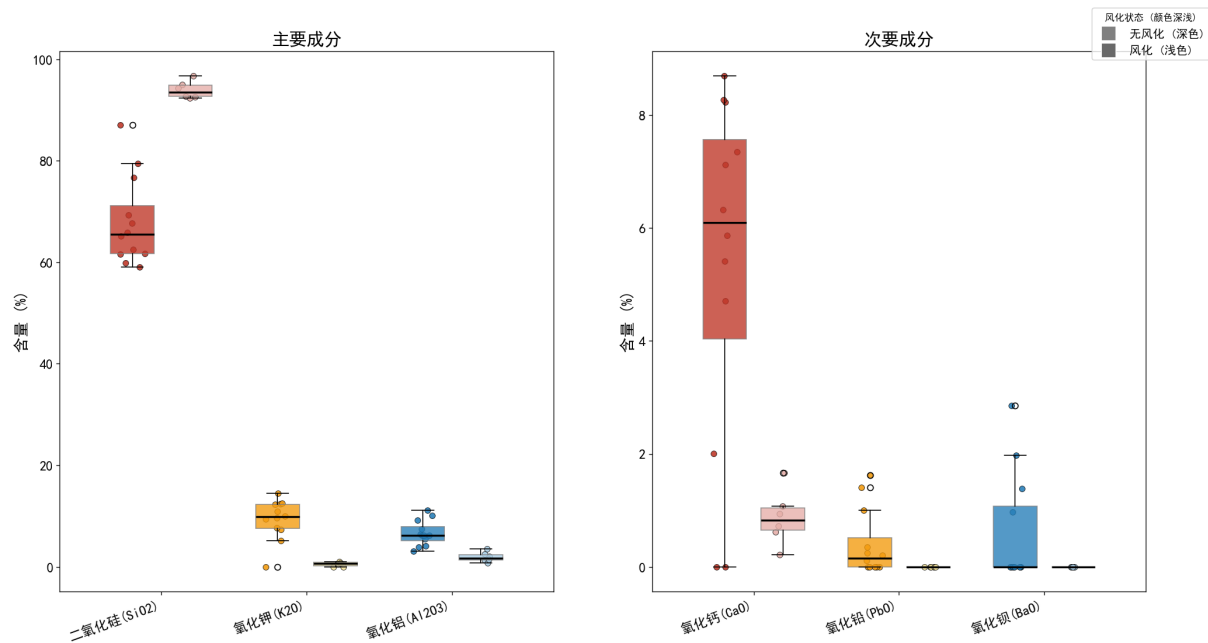


图 8 高钾玻璃在风化前后各化学成分含量分布

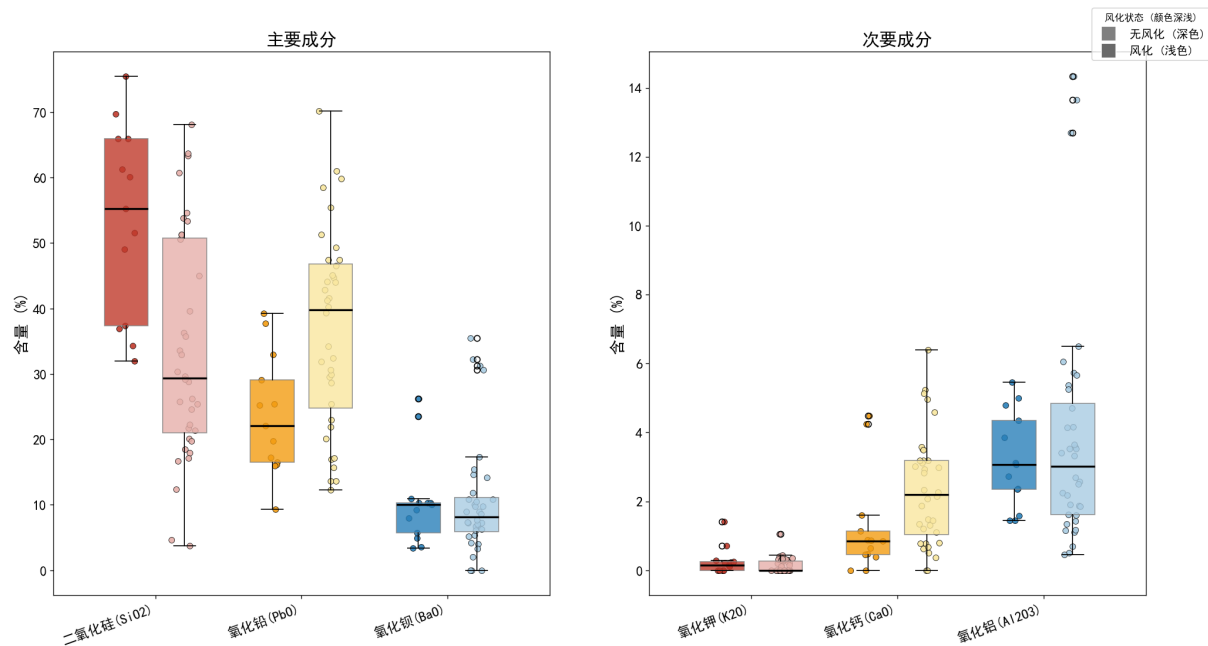


图 9 铅钡玻璃在风化前后各化学成分含量分布

为了更细致地观察关键化学成分的分布形态变化，我们绘制了部分核心化学成分，包括氧化铅 PbO 、氧化钾 K_2O 、氧化钡 BaO 以及二氧化硅 SiO_2 的分布图，如图10至图13所示。图中包含直方图与核密度估计曲线，它们共同描述了数据分布的集中趋势和形态。分析这些分布图可以发现，高钾玻璃在风化后，其氧化钾 K_2O 的含量分布从一个较宽的区间转化至接近零值的极低水平。对于铅钡玻璃，风化作用使其特征成分氧化铅 PbO 与氧化钡 BaO 的含量分布整体向低值区移动。与此相反，作为玻璃基体的二氧

化硅 SiO_2 ，其含量分布在两类玻璃中均表现出向高值区偏移的趋势，说明在风化过程中其他元素的流失导致了二氧化硅的相对富集。

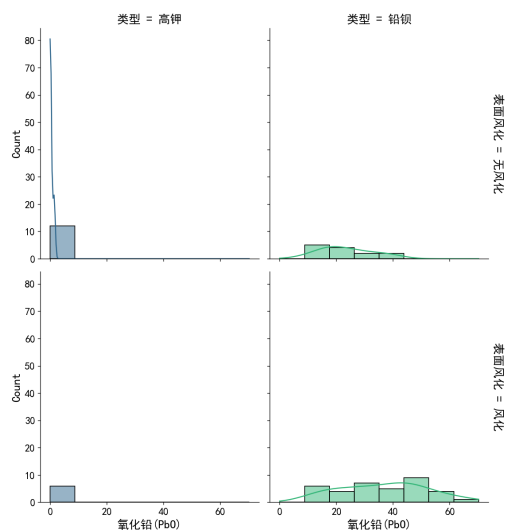


图 10 氧化铅 PbO 含量分布

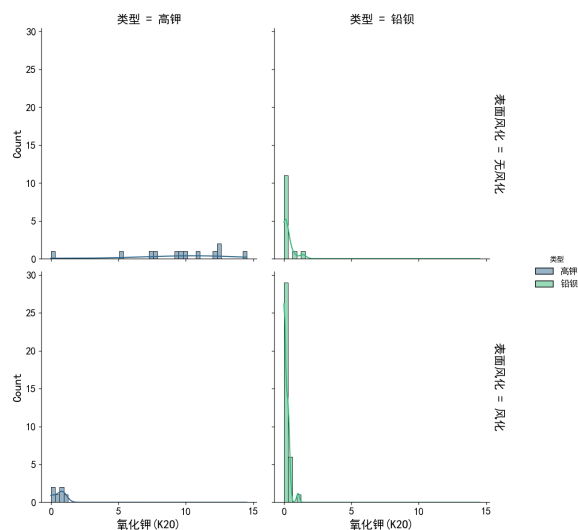


图 11 氧化钾 K_2O 含量分布

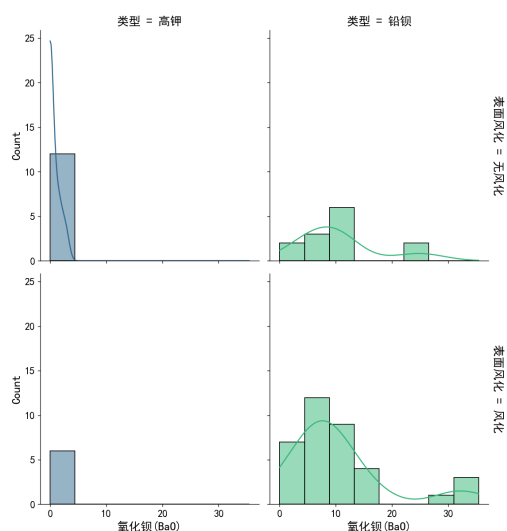


图 12 氧化钡 BaO 含量分布

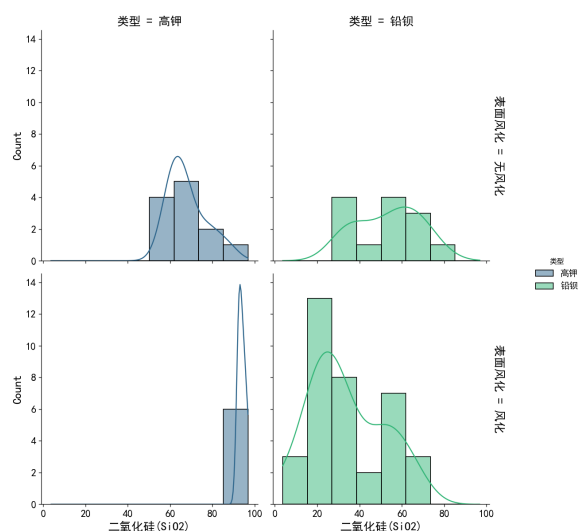


图 13 二氧化硅 SiO_2 含量分布

5.3 基于风化系数的化学成分含量预测模型

预测风化前化学成分的主要困难在于数据中缺乏源自同一文物的风化与未风化的观测样本。这使得依赖大量成对样本进行训练的传统监督学习模型，例如回归分析，难以直接应用且存在较高的过拟合风险。

于是，我们转向从风化过程的物理化学机理中寻求建模依据。玻璃的风化过程与地质学中岩石的蚀变过程在原理上具有相似性，均为长期化学环境作用下的元素迁移过程。因此，我们引入了地球化学领域成熟的质量平衡分析理论来构建预测模型，该理论能够在缺乏直接演变过程数据时，对成分变化进行有效推断。

我们的模型的核心假设为：特定化学成分在风化过程中的流失或富集比例，与其在未风化状态下的原始含量相关。此假设参考了地质学中分析交代蚀变作用的艾索康图法的原理，它将风化视为一个系统性的化学变化过程而非随机过程，从而建立了基于风化系数的预测方法。

我们首先为每种玻璃类型 t 和每种化学成分 j 定义一个风化系数 $k_{t,j}$ 。该系数由该类型玻璃中所有风化样本与未风化样本的平均含量计算得出，其数学表达式如下：

$$k_{t,j} = 1 - \frac{\bar{C}_{t,j,\text{weathered}}}{\bar{C}_{t,j,\text{unweathered}}} \quad (1)$$

其中， $\bar{C}_{t,j,\text{weathered}}$ 表示 t 类玻璃风化样本中 j 成分的平均含量， $\bar{C}_{t,j,\text{unweathered}}$ 表示 t 类玻璃未风化样本中 j 成分的平均含量。

利用计算得到的风化系数，我们可以对任意一个已知风化后成分含量 $C_{\text{weathered},j}$ 的样本，进行其风化前含量 $C'_{\text{unweathered},j}$ 的初步预测，其预测公式为：

$$C'_{\text{unweathered},j} = \frac{C_{\text{weathered},j}}{1 - k_{t,j}} \quad (2)$$

考虑到测量误差和模型的局限性，初步预测得到的各成分总和可能偏离 100%。因此，我们对预测结果进行了边界条件校正。我们计算所有预测成分的总和，并进行条件归一化处理，以确保最终得到的预测成分总和落入题目要求的 85% 至 105% 的有效区间内，从而使预测结果在化学上更具合理性。模型的最终预测结果，以原始含量和预测含量并列对比的形式，被完整地保存至 Result/问题一 _ 预测结果.xlsx 文件中。