Modeling and Simulation 建模与仿真, 2023, 12(2), 1691-1704
Published Online March 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.122157

Hans汉斯

古代玻璃制品的成分分析与鉴别

李怡霖，王梦佳，张　琦，谢素霞

上海理工大学机械工程学院，上海

收稿日期：2023年2月5日；录用日期：2023年3月24日；发布日期：2023年3月31日

摘　要

本文针对玻璃制品易受到环境因素而风化，探究了玻璃文物的表面风化与其玻璃类型、纹饰和颜色之间的关系；并结合玻璃的类型，对文物样品表面有无风化化学成分含量进行统计并分析其内在规律；根据风化点的检测数据，建立模型预测玻璃文物其风化前的各个化学成分的含量；分析了高钾玻璃、铅钡玻璃的分类规律，以及探讨分类结果的合理性和敏感性；对不同类别的玻璃文物样品，研究分析了其化学成分之间的关联关系，比较了不同类别之间的化学成分关联关系存在的差异性。

关键词

古代玻璃制品，鉴别，随机森林算法，多元线性回归方程，聚类算法

Composition Analysis and Identification of Ancient Glass Products

Yilin Li, Mengjia Wang, Qi Zhang, Suxia Xie

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 5th, 2023; accepted: Mar. 24th, 2023; published: Mar. 31st, 2023

Abstract

Aiming at the weathering of glass products by environmental factors, this paper probes into the relationship between the surface weathering of glass relics and the type, pattern and color of glass. Combined with the type of glass, the contents of weathering chemical components on the surface of cultural relics samples were counted and their internal rules were analyzed. According to the test data of weathering point, a model was established to predict the contents of each chemical component of glass relics before weathering. The classification rules of high potassium glass and lead barium glass are analyzed, and the rationality and sensitivity of the classification results are discussed. The correlation between the chemical components of different types of glass cultural

文章引用: 李怡霖, 王梦佳, 张琦, 谢素霞. 古代玻璃制品的成分分析与鉴别[J]. 建模与仿真, 2023, 12(2): 1691-1704.
DOI: 10.12677/mos.2023.122157

relics samples was studied and analyzed, and the difference of the existence of the correlation between different types of chemical components was compared.

Keywords

Ancient Glass, Identification, Random Forest Algorithm, Multiple Linear Regression Equation, Clustering Algorithm

1. 引言

科学数据表明，玻璃的主要原料是石英砂，主要化学成分是二氧化硅(SiO_2)。在历史长期的演化中，这些玻璃制品易受到空气湿度、日照、风向风速等环境的干扰，导致这些玻璃制品出现不同程度的风化，在风化过程中，内部元素与环境元素进行大量交换，其成分比例发生变化，从而影响对其类别的正确判断。因此对风化程度进行分析处理和数据挖掘，一方面可以为玻璃制品的类别判断提供数据参考，另一方面对玻璃制品风化的预防和修护提供依据[1]。

目前已有电子探针、光谱分析仪等各类高精仪器能对古代文物的成分进行定量分析。本文研究内容是古代玻璃制品的属性及分类。

2. 玻璃属性相关联模型构建

2.1. 数据处理

玻璃文物纹饰、类型、颜色和表面风化这四个指标都是定类变量，并非连续变量，因此在进行差异化分析时，采用卡方检验对三组数据中的两个分组变量的显著性差异进行分析，比较计算得到的检验统计量 χ^2，最终确定各个变量与有无风化之间的关系，并且画出卡方交叉热力图。其次，对呈现显著性的因素，根据效应指标对其差异进行深入量化分析，得到其对玻璃表面风化的确切差异程度，为了研究呈现显著性的两个变量的正负向关系，且考虑到两个定类变量的因素，故用斯皮尔曼相关系数对呈现显著性的变量和表面风化之间进行分析[2]。

根据表 1 现有玻璃化学成分，属性及分类情况数据的统计分析，预处理后的数据分为有无风化两组，本文仅考虑表面风化程度对化学成分含量统计规律的影响，并对其中检测到的同种化学成分累加，得到的各种成分占比，数值饼状图如图 1 所示。

由图 1 成分分析饼图得出重要变量为二氧化硅，氧化铅，氧化钾，使用多元线性回归模型对成分预测，将二氧化硅，氧化铅，氧化钾分别作为因变量，有无风化，种类，纹饰，颜色作为自变量，对于定性变量使用 stata 对其转化为虚拟变量，其对应表如表 2 所示。

玻璃文物的化学成分众多，样本的特征维度很高，采用随机森林模型仍然能够高效地训练，在选取特征数 M 时，因为高钾玻璃、铅钡玻璃中有部分化学成分含量较大，在分析特征数时含量较大的因子会淡化其他化学成分之间的差异性，以及不能忽略风化因素对化学成分改变所带来的两者之间的差异性，故需对数据进行处理。处理后得到的柱形图如图 2 所示。

Table 1. Glass chemical composition
表 1. 玻璃化学成分

文物采样点	二氧化硅 (SiO_2)	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)	文物采样点	二氧化硅 (SiO_2)	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)
1	69	6	1	4	2	4	0	0	1	31	66	2	1	3	5	0	17	3	2
2	36	2	1	6	2	0	47	0	4	32	70	0	0	2	1	0	20	5	0
03 部位 2	62	6	1	6	2	5	1	3	1	34	36	1	0	2	0	2	47	10	0
4	66	7	2	6	2	2	0	0	1	35	66	0	0	1	0	0	22	6	0
5	62	7	2	8	3	3	0	0	1	36	40	0	0	2	0	1	42	11	0
06 部位 1	68	0	2	11	2	3	0	1	4	37	60	1	0	3	0	3	17	10	1
7	93	1	0	2	0	3	0	0	1	39	26	1	0	1	0	1	61	7	1
8	20	1	0	1	0	10	29	31	4	40	17	2	0	0	0	0	70	7	2
08 严重风化点	5	3	0	1	0	3	32	31	8	41	18	5	3	3	2	0	44	10	7
10	97	0	0	1	0	1	0	0	0	42 未风化点 2	51	0	1	6	0	3	20	11	0
11	34	4	1	3	0	5	25	15	9	43 部位 1	12	5	1	2	1	5	60	7	0
12	94	1	0	1	0	2	0	0	0	43 部位 2	22	6	1	3	1	2	45	3	13
13	59	9	0	6	3	5	0	0	1	44 未风化点	61	2	0	13	1	0	14	5	0
14	62	8	1	9	1	0	2	0	0	45	61	1	1	5	0	1	16	11	0
15	62	0	1	3	1	1	0	0	0	46	55	0	2	5	0	1	25	10	0
16	65	8	1	6	0	1	0	0	0	47	52	1	1	3	0	1	25	9	0
17	61	0	1	0	1	1	0	0	0	48	53	3	2	14	1	0	16	7	1
20	37	0	0	5	2	5	9	24	6	50	18	3	0	2	0	1	44	14	6
21	77	5	1	6	2	3	1	2	1	50 未风化点	45	3	1	4	0	1	31	6	6
23 未风化点	54	1	1	1	0	3	17	12	0	51 部位 2	21	5	1	3	0	1	51	0	9
24	32	0	0	2	0	8	29	26	0	52	26	2	1	1	0	1	47	9	6
25 未风化点	51	1	0	2	2	1	32	7	0	53 未风化点	64	1	1	6	0	1	14	9	0
26	20	1	0	1	0	11	30	32	3	54	22	3	1	4	0	1	55	7	4
26 严重风化点	4	3	0	1	0	4	30	35	6	54 严重风化点	17	0	1	4	0	1	58	0	14
28 未风化点	68	1	1	5	0	0	17	4	1	56	29	1	0	2	0	1	41	15	3
29 未风化点	63	3	1	14	1	1	12	2	0	57	25	1	0	2	0	1	45	17	0
30 部位 1	34	4	1	4	2	0	39	10	0	31	66	2	1	3	5	0	17	3	2

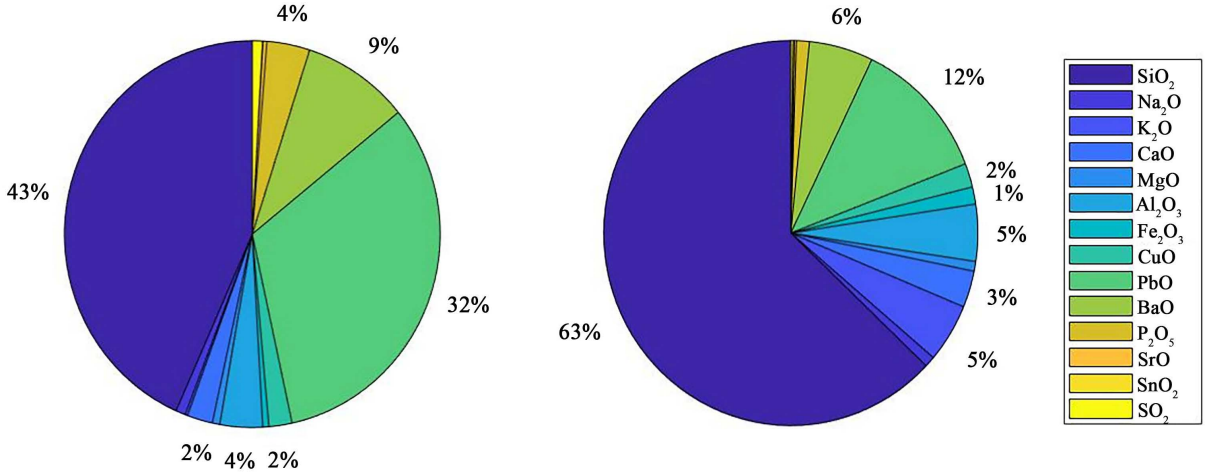
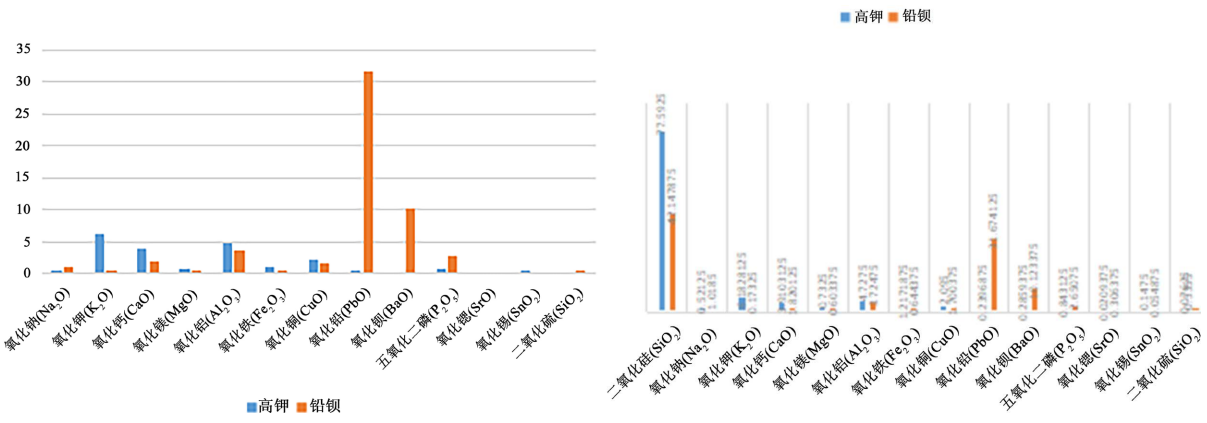


Figure 1. Pie chart of cultural relics composition
图 1. 文物成分占比饼状图

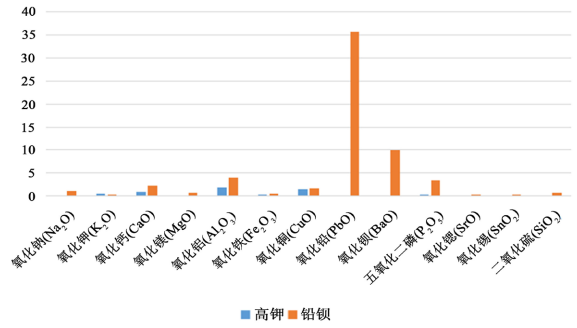
Table 2. Table of virtual variable correspondence
表 2. 虚拟变量对应表

虚拟变量	定性变量	虚拟变量	定性变量
efflo 1	表面风化==无风化	color 2	颜色==浅蓝
efflo 2	表面风化==风化	color 3	颜色==深绿
type 1	类型==铅钡	color 4	颜色==深蓝
type 2	类型==高钾	color 5	颜色==紫
cela 1	纹饰==A	color 6	颜色==绿
cela 2	纹饰==B	color 7	颜色==蓝绿
cela 3	纹饰==C	color 8	颜色==黑
color 1	颜色==浅绿		

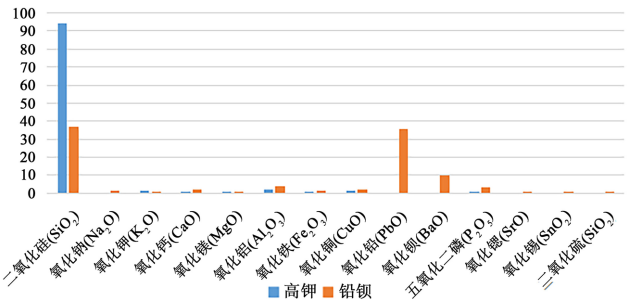


(a) 去除二氧化硅，高钾玻璃与铅钡玻璃化学成分比较

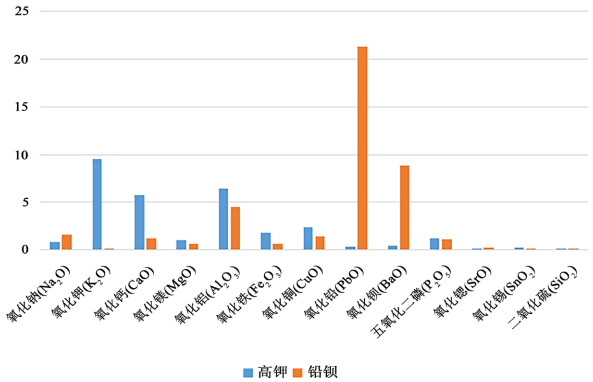
(b) 高钾玻璃与铅钡玻璃化学成分比较



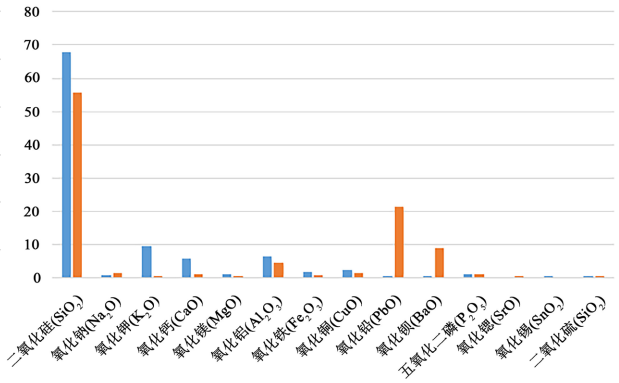
(c) 去除二氧化硅，风化的高钾玻璃与铅钡玻璃化学成分比较



(d) 风化的高钾玻璃与铅钡玻璃化学成分比较



(e) 去除二氧化硅，无风化的高钾玻璃与铅钡玻璃化学成分比较



(f) 无风化的高钾玻璃与铅钡玻璃化学成分比较

Figure 2. Glass chemical composition comparison histogram
图 2. 玻璃化学成分比较柱形图

减小特征选择个数 m，树的相关性和分类能力也会相应地降低；增大 m，两者也会随之增大，因此选择合适的 M 对于算法分类效果很重要，而通过两组图的比较得出两种玻璃文物之间的化学成分主要是二氧化硅，氧化钾，氧化钙，氧化铅，氧化钡，五氧化二磷之间有差异，因此将这六种化学成分作为变量 X 代入随机森林算法中[3]。

2.2. 模型流程图

以下为本文在进行建模的流程图(图 3、图 4)：

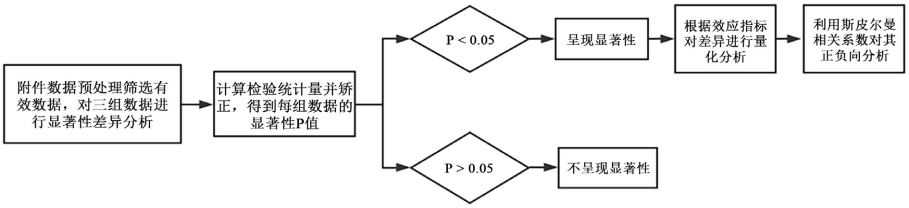


Figure 3. Flow chart of glass type and property analysis
图 3. 玻璃类型与属性分析流程图

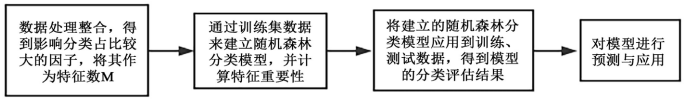


Figure 4. Flow chart of glass classification analysis
图 4. 玻璃分类分析流程图

3. 模型检验及应用

3.1. 玻璃类型与属性模型

3.1.1. 卡方检验分析

本文首先通过卡方检验，对纹饰、类型、颜色与表面风化进行差异化分析。在公式(1)中，f_0代表纹饰，类型，颜色，f_1代表表面风化。

$$\chi^2 = \sum \frac{(f_0 - f_e)}{f_e} \sim \chi^2(n) \tag{1}$$

计算结果详见下表 3。

卡方检验分析的结果显示：

基于表面风化和纹饰，显著性 P 值为 0.084*，基于表面风化和颜色，显著性 P 值为 0.405，水平上不呈现显著性，接受原假设，因此对于表面风化和纹饰、颜色数据不存在显著性差异；基于表面风化和类型，显著性 P 值为 0.009***，水平上呈现显著性，拒绝原假设，因此对于表面风化和类型数据存在显著性差异(图 5)。

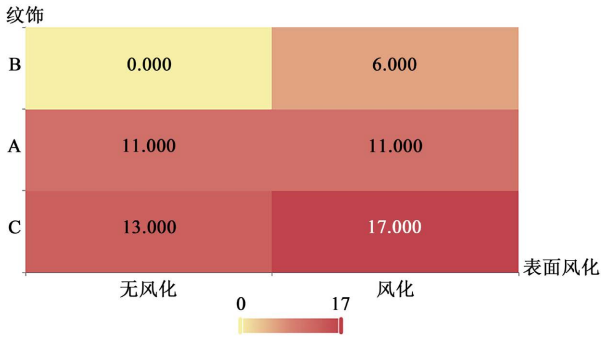
题目	名称	表面风化		总计	X^2	校正 X^2	P
		无风化	风化				
纹饰	C	13	17	30	4.957	4.957	0.084*
	A	11	11	22			
	B	0	6	6			
合计		24	34	58			
类型	高钾	12	6	18	6.880	5.452	0.009***
	铅钡	12	28	40			
合计		24	34	58			
颜色	蓝绿	6	9	15	7.234	7.234	0.405
	浅蓝	8	16	24			
	紫	2	2	4			
	深绿	3	4	7			
	深蓝	2	0	2			
	浅绿	2	1	3			
	黑	0	2	2			
	绿	1	0	1			
合计		24	34	58			

注：***、**、*分别代表 1%、5%、10%的显著性水平

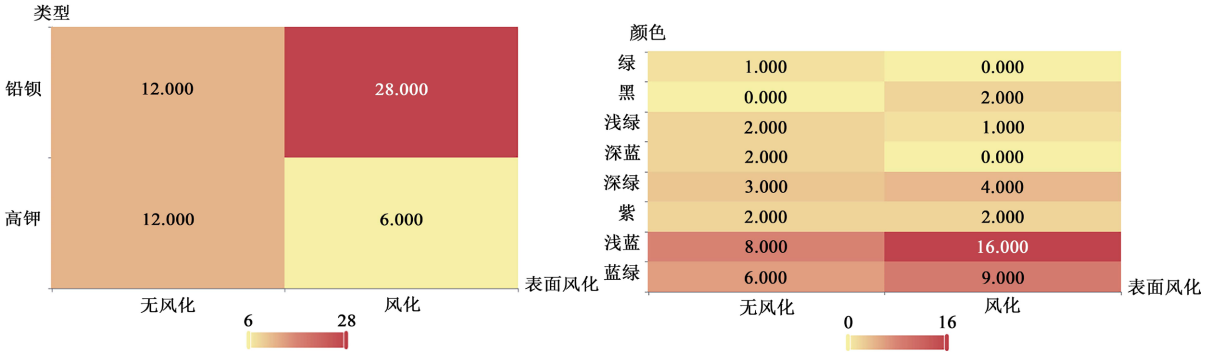
Figure 5. Chi-square analysis of ornamentation, type, color and surface weathering

图 5. 纹饰、类型、颜色与表面风化的卡方分析结果

计算出卡方分析结果后，为了更加直观看出显著性差异，用热力图的形式展示交叉列联表的值，并且通过颜色深浅反应表示值的大小(图 6、表 3)。



(a) 纹饰热力图



(b) 类型热力图　　(c) 颜色热力图

Figure 6. Surface weathering heat map
图 6. 表面风化热力图

Table 3. Quantitative analysis of effects
表 3. 效应量化分析

字段名/分析项	Phi	Crammer's V	列联系数	lambda
纹饰	0.292	0.292	0.281	0.000
类型	0.344	0.344	0.326	0.000
颜色	0.353	0.353	0.333	0.000

由上述分析可知，表面风化与类型之间有着显著性差异，然后通过 phi、Crammer's V.列联系数、lambda 对样本的相关程度进行分析，得到纹饰和表面风化的差异程度为中等程度差异；类型和表面风化的差异程度为中等程度差异；颜色和表面风化的差异程度为中等程度差异[4]。

3.1.2. 建立多元线性回归

利用多元线性回归模型对未风化前的成分预测，其一般公式为(2)所示，其中自变量和因变量已在 step 1 中转化为虚拟变量。

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_kx_k+\varepsilon \tag{2}$$

公式(2)中，β_1、β_2、$\beta_3\cdots\beta_k$ 是偏回归系数，与 x_1、x_2、$x_3\cdots x_k$ 无相关性，ε 为随机误差项。

假设，因变量与自变量之间存在线性关系，那他们之间的线性总体回归方程可以表示为：

$$\begin{cases} y_1=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon \\ y_2=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon \\ y_3=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon \end{cases} \tag{3}$$

其中，ε 为随机误差项 $\varepsilon\sim N\left(0,\sigma^2\right)$。

回归系数分析如表 4 所示。

并且得到的多元线性回归模型可以表示为：

$$\begin{cases} y_1=119.350+18.314x_1+\left(-13.857\right)x_2+\left(-38.879\right)x_3+\left(-56.483\right)x_4 \\ y_2=1.361+\left(-9.332\right)x_1+\left(-10.971\right)x_2+7.597x_3+7.971x_4 \\ y_3=\left(-16.863\right)+12.772x_1+26.019x_2+5.279x_3+23.764x_4 \end{cases}$$

对回归模型进行怀特检验(表 5)：

Table 4. Regression coefficient analysis
表 4. 回归系数分析

	二氧化硅 SiO_2		氧化钾 K_2O		氧化铅 PbO
efflo 1	0.000	type 1	−9.332***	efflo 1	0.000
	(.)		(−10.971)		(.)
efflo 2	−18.314***	type 2	0.000	efflo 2	12.772***
	(−3.775)		(.)		(3.284)
type 1	−13.857**	cela 1	7.597***	type 1	26.019***
	(−2.254)		(6.654)		(5.279)
type 2	0.000	cela 2	0.000	type 2	0.000
	(.)		(.)		(.)
cela 1	−38.879***	cela 3	7.971***	cela 2	0.000
	(-4.712)		(7.099)		(.)
cela 2	0.000	_cons	1.361	cela 3	23.764***
	(.)		(0.659)		(3.653)
cela 3	−56.483***	N	69	_cons	−16.863
	(−6.960)				(−1.409)
_cons	119.350***			N	69
	(7.992)				
N	69				

注：***p < 0.01 **p < 0.05 *p < 0.1。

Table 5. BP test results
表 5. BP 检验结果

化学成分种类	P 值
二氧化硅	0.1784
氧化钾	0.2258
氧化铅	0.1777

得到结论二氧化硅，氧化钾，氧化铅回归进行 BP 检验 P 值分别为 0.1784，0.2258，0.1777，P 值均大于 0.1，故不存在异方差。

3.2. 玻璃分类模型

3.2.1. 计算特征重要性

根据袋外误差率，对于特征 M，首先用训练好的随机森林在对 oob 数据集 D (二氧化硅，氧化钾，氧化钙，氧化铅，氧化钡，五氧化二磷)进行预测并求出误差率 Error 1。然后对数据 D (二氧化硅，氧化钾，氧化钙，氧化铅，氧化钡，五氧化二磷)中每个样本的特征 m 上加上随机噪音，然后再将 m 特征上带噪音的样本送入训练好的 RF 模型中训练得到新的误差率 Error 2，则 Error 2 - Error 1 越大说明该特征越重要。计算得到的特征值重要性如图 7 所示，其中氧化铅的重要性最大，并且得到混淆矩阵热力图如图 8 所示：

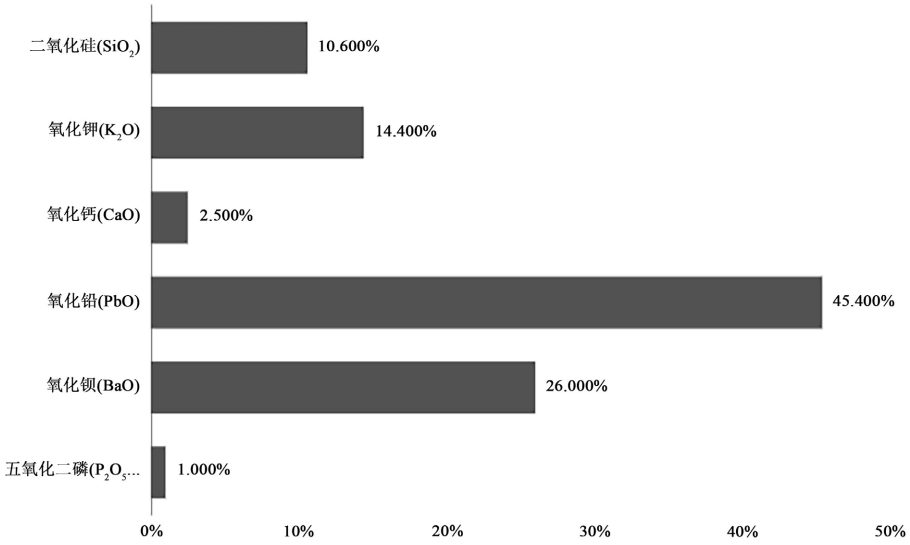


Figure 7. Feature importance
图 7. 特征重要性

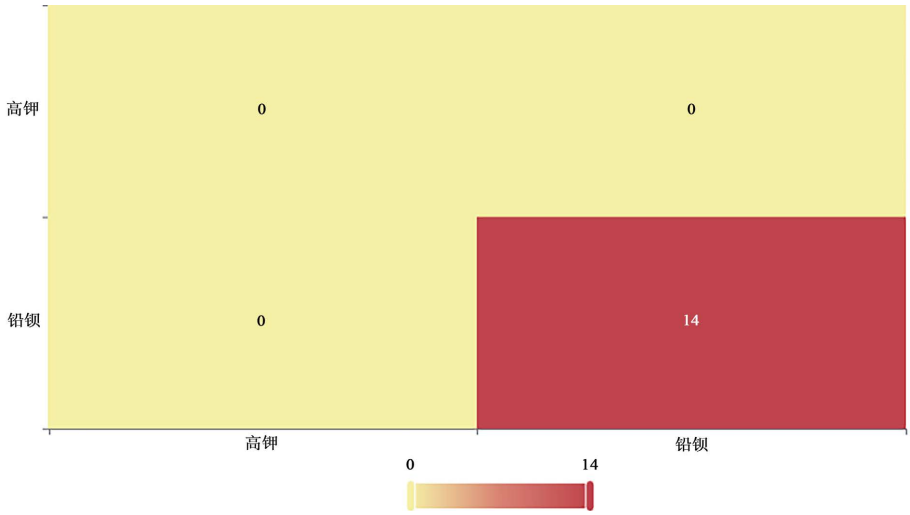


Figure 8. Confused rectangular heat map
图 8. 混淆矩形热力图

将数据集分为训练集和测试集，由于本数据集样本过少，在训练时将数据的 80%作为训练集进行分类的依据。剩下的 20%作为测试集来验证模型的准确性，自然形成一个对照数据集，用于模型的验证，所以随机森林不需要另外预留部分数据做交叉验证。将通过量化指标来衡量随机森林对训练、测试数据的分类效果，得到的结果如表 6、表 7 所示：

Table 6. Model evaluation results
表 6. 模型评估结果

	准确率	召回率	精确率	F1
训练集	1	1	1	1
测试集	1	1	1	1

DOI: 10.12677/mos.2023.122157

Table 7. Test data predicts evaluation results
表 7. 测试数据预测评估结果

预测结果 Y	类型	预测结果概率_铅钡	预测结果概率_高钾	二氧化硅 (SiO_2)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P_2O_5)	预测结果 Y	类型	预测结果概率_铅钡	预测结果概率_高钾	二氧化硅 (SiO_2)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P_2O_5)
铅钡	铅钡	1	0	28.79	0	4.58	34.18	6.1	11.1	铅钡	铅钡	1	0	63.66	0.11	0.78	13.66	8.99	0
铅钡	铅钡	1	0	54.61	0.3	2.08	23.02	4.19	4.32	铅钡	铅钡	1	0	22.28	0.32	3.19	55.46	7.04	4.24
铅钡	铅钡	1	0	17.98	0	3.19	44	14.2	6.34	铅钡	铅钡	0.86	0.14	17.11	0	0	58.46	0	14.13
铅钡	铅钡	1	0	45.02	0	3.12	30.61	6.22	6.34	铅钡	铅钡	0.99	0.01	49.01	0	1.13	32.92	7.95	0.35
铅钡	铅钡	1	0	24.61	0	3.58	40.24	8.94	8.1	铅钡	铅钡	1	0	29.15	0	1.21	41.25	15.45	2.54
铅钡	铅钡	0.87	0.13	21.35	0	5.13	51.34	0	8.75	铅钡	铅钡	1	0	25.42	0	1.31	45.1	17.3	0
铅钡	铅钡	1	0	25.74	0	2.27	47.42	8.64	5.71	铅钡	铅钡	1	0	30.39	0.34	3.49	39.35	7.66	8.99

3.2.2. 玻璃类型亚分类

本文采用的是 K-Means 聚类方法对文物进行亚分类[3]，但其受初始值和异常点影响，聚类结果可能不是全局最优而是局部最优，所以在开始前需要对数据预处理，先对采样数量半数以上的化学成分进行方差处理，方差大说明不稳定，可以挑选出主要变化的化学成分，经过分析将方差值以 2 为界限，大于 2 的化学成分将进行聚类分析(图 9)。

文物采样点	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)
平均值	76.64389	0.463333	6.401667	3.845	0.785	5.056667	1.376111	2.155556	0.274444	0.398889	1.028333	0.027778	0.131111	0.067778
方差	197.6592	1.119433	26.60711	10.33571	0.478514	8.939967	2.316213	2.103058	0.249658	0.668988	1.548836	0.001817	0.292232	0.023328

Figure 9. Variance value and mean
图 9. 方差值和平均值

经过数据分析将二氧化硅、氧化钾、氧化钙、氧化镁、氧化铝、氧化铁、氧化铜、五氧化二磷这八种化学成分进行下一步处理。

K-MEANS 聚类法：

算法输入：统计聚类个数 K，以及包含 n 个数据对象的数据样本集 U；

算法输出：满足方差要求且最小标准的聚类 k 个；

1) 选取初始聚类中心：从 n 个数据对象中任意选择 k 个对象将其作为初始聚类中心；

2) 欧氏距离划分：将每个聚类中所有对象的均值计算样本数据中每个对象与这些中心对象的欧氏距离，并根据最小距离重新对相应对象进行划分；

3) 重新计算每个聚类的中心对象(均值)；

4) 循环步骤 2 到步骤 3，直到每个聚类不再发生变化为止。

在 K-MEANS 聚类中，为了显现数据之间的相似度，常用的方法是通过欧氏距离表示，公式如(4)所示：

$$d_{ij}=\sqrt{\left|x_{1i}-x_{1j}\right|^2+\left|x_{2i}-x_{2j}\right|^2+\left|x_{3i}-x_{3j}\right|^2+\left|x_{4i}-x_{4j}\right|^2} \tag{4}$$

高钾玻璃模型求解与分析如表 8 所示：

Table 8. Differential analysis of high-potassium glass fields
表 8. 高钾玻璃字段差异性分析

	聚类类别(平均值 ± 标准差)			F	P
	类别 2 (n = 9)	类别 3 (n = 6)	类别 1 (n = 3)		
二氧化硅(SiO_2)	63.624 ± 3.558	93.963 ± 1.734	81.063 ± 5.368	145.911	0.000***
氧化钾(K_2O)	10.818 ± 2.37	0.543 ± 0.445	4.87 ± 4.718	32.212	0.000***
氧化钙(CaO)	6.363 ± 2.64	0.87 ± 0.488	2.24 ± 2.363	12.979	0.001***
氧化铝(Al_2O_3)	7.349 ± 2.346	1.93 ± 0.964	4.433 ± 1.603	14.929	0.000***
氧化铁(Fe_2O_3)	2.312 ± 1.643	0.265 ± 0.069	0.79 ± 1.368	4.827	0.024**
氧化铜(CuO)	2.819 ± 1.565	1.562 ± 0.935	1.353 ± 1.714	2.012	0.168

注：***、**、*分别代表 1%、5%、10%的显著性水平。

上表展示了定量字段差异性分析的结果，包括均值 ± 标准差的结果、F 检验结果、显著性 P 值，对于每个项分析是否小于 0.05 或者 0.01，检验其是否符合标准，若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据均值±标准差的方式对差异进行分析，反之则表明数据不呈现差异性。对于本文而言，呈现显著性差异的有二氧化硅(SiO_2)、氧化钾(K_2O)、氧化钙(CaO)、氧化铝(Al_2O_3)、氧化铁(Fe_2O_3)，仅有氧化铜(CuO)未显示出显著性差异。

通过聚类分析，用频数和百分比的形式显示聚类汇总结果，如表 9 所示：

Table 9. Summary results of high-potassium glass clustering
表 9. 高钾玻璃聚类汇总结果

聚类类别	频数	百分比%
聚类类别_1	3	16.667%
聚类类别_2	9	50.0%
聚类类别_3	6	33.333%
合计	18	100.0%

结果分为三类，用可视化的形式展示模型聚类的结果如图 10 所示：

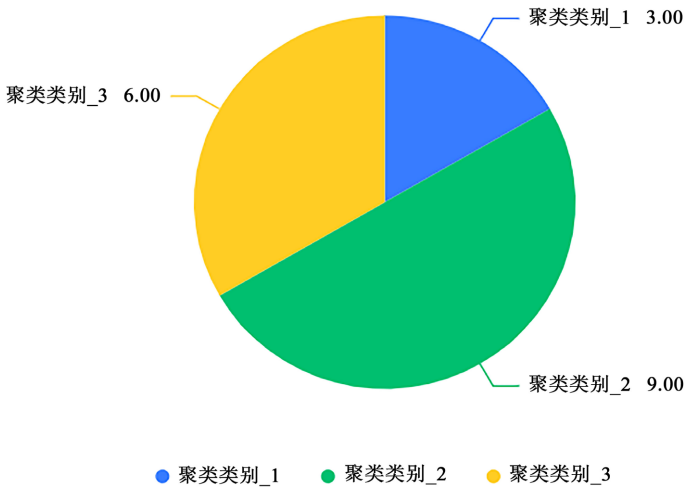


Figure 10. High-potassium glass clustering summary pie chart
图 10. 高钾玻璃聚类汇总饼状图

铅钡玻璃模型求解与分析如表 10 所示：

Table 10. Field difference analysis of lead-barium glass
表 10. 铅钡玻璃的字段差异性分析

	聚类类别(平均值 ± 标准差)				F	P
	类别 2 (n = 19)	类别 3 (n = 14)	类别 1 (n = 10)	类别 4 (n = 6)		
二氧化硅(SiO_2)	59.206 ± 7.644	33.031 ± 5.562	19.999 ± 4.284	19.593 ± 13.747	82.64	0.000***
氧化钙(CaO)	1.098 ± 0.861	2.616 ± 1.488	3.336 ± 2.062	1.598 ± 1.296	6.666	0.001***

Continued

氧化铝(Al_2O_3)	5.091 ± 4.092	3.452 ± 1.413	2.328 ± 1.322	1.895 ± 1.766	3.176	0.033**
氧化铜(CuO)	0.999 ± 1.017	1.391 ± 1.455	1.268 ± 1.507	6.827 ± 3.397	20.351	0.000***
氧化铅(PbO)	19.876 ± 5.812	40.057 ± 6.59	53.664 ± 8.825	26.503 ± 8.53	56.899	0.000***
氧化钡(BaO)	7.248 ± 3.085	9.492 ± 4.466	6.41 ± 4.357	29.888 ± 4.299	57.026	0.000***
五氧化二磷(P_2O_5)	0.61 ± 1.034	4.368 ± 4.208	6.239 ± 4.743	4.368 ± 2.642	7.6	0.000***

注：***、**、*分别代表 1%、5%、10%的显著性水平。

上表展示了定量字段差异性分析的结果，包括均值 ± 标准差的结果、F 检验结果、显著性 P 值，对于每个项分析是否小于 0.05 或者 0.01，检验其是否符合标准，若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据均值±标准差的方式对差异进行分析，反之则表明数据不呈现差异性。对于本文而言，所有的化学成分在聚类划分的类别中均存在显著性差异。

通过聚类分析，用频数和百分比的形式显示聚类汇总结果，如表 11 所示：

Table 11. Summary results of lead-barium glass clustering
表 11. 铅钡玻璃聚类汇总结果

聚类类别	频数	百分比%
聚类类别_1	10	20.408%
聚类类别_2	19	38.776%
聚类类别_3	14	28.571%
聚类类别_4	6	12.245%
合计	49	100.0%

结果分为三类，用可视化的形式展示模型聚类的结果如图 11 所示：

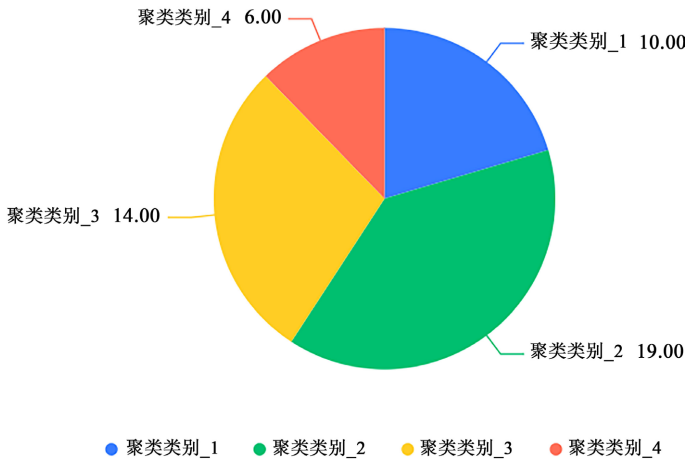


Figure 11. Lead-barium glass cluster summary pie chart
图 11. 铅钡玻璃聚类汇总饼状图

本文中在进行聚合分析时选用数据是附件已测的，为了更加清楚研究该模型对玻璃文物的亚类划分的敏感程度，我们将其中一种化学成分含量进行干扰，将氧化钙的含量每个增加 10%到 50%，继续对高

钾和铅钡玻璃聚合分类，观察并分析得到的数据，且与未做干扰的原始数据进行比较。

本文对未知类别玻璃文物的化学成分进行分析，鉴别其所属类型，在基于建立的随机森林模型下，将表单三中未知文物的化学成分数据代入模型中，由于建立的模型选取的特征数仅有六个，故在对其类型鉴别忽略其他化学成分的影响，这个影响因子较低对于整体预测的准确性干扰不大。

对其类型的预测如表 12 所示：

Table 12. Prediction results for the type of unknown artifacts
表 12. 对未知文物的类型预测结果

预测结果_Y	预测结果概率_铅钡	预测结果概率_高钾	二氧化硅(SiO_2)	氧化钾(K_2O)	氧化钙(CaO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)
高钾	0.04	0.96	78.45	0	6.08	0	0	1.06
铅钡	0.82	0.18	37.75	0	7.63	34.3	0	14.27
铅钡	0.91	0.09	31.95	1.36	7.19	39.58	4.69	2.68
铅钡	1	0	35.47	0.79	2.89	24.28	8.31	8.45
铅钡	0.81	0.19	64.29	0.37	1.64	12.23	2.16	0.19
高钾	0	1	93.17	1.35	0.64	0	0	0.21
高钾	0	1	90.83	0.98	1.12	0	0	0.13
铅钡	1	0	51.12	0.23	0.89	21.24	11.34	1.46

结果显示，通过比较预测结果的概率 A1、A6、A7 为高钾，其余文物类型为铅钡玻璃。

在进行随机森林的样本抽取中，对数据进行编号，并且使用随机抽样，抽取 80%的数据集作为训练集；如果训练集大小为 N，对于每棵树而言，采取随机且有放回地从训练集中的抽取 N 个训练样本，这两个随机性对随机森林的分类性能至关重要。由于它们的引入，使得随机森林不容易陷入过拟合，并且具有很好地抗噪能力，同时对缺省值不敏感。

4. 结论

本文中的特征很多，对于这种高维稠密型的数据在进行成分预测时，选用随机森林模型，无需做特征选择，本文中对古代玻璃制品的风化因素的研究，同样可以推广到更多非遗文物的保护和预防中，为建立数字文物保护机制提供了依据，也极大地促进了不同文化之间的交流和互动。

参考文献

[1] 阚颖浩. 山东博山元末明初玻璃作坊出土玻璃科技分析[D]: [硕士学位论文]. 济南: 山东大学, 2022. https://doi.org/10.27272/d.cnki.gshdu.2022.002831

[2] 王祉皓, 赵芗澈, 李智群, 郭明, 肖琬玥, 刘志坚. 基于机器学习的风化硅酸盐玻璃原成分预测及亚分类方法[J/OL]. 硅酸盐学报: 1-11, 2023-02-05. https://doi.org/10.14062/j.issn.0454-5648.20220985

[3] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO (Version 1.0.11), Online Application Software. https://www.spsspro.com

[4] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.