Modeling and Simulation 建模与仿真, 2023, 12(2), 1069-1082
Published Online March 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.122101

Hans汉斯

玻璃文物化学成分体系的统计分析

周冒伟，黄佳怡，孙雪婷，刘　芳

南京理工大学，数学与统计学院，江苏 南京

收稿日期：2023年2月4日；录用日期：2023年3月9日；发布日期：2023年3月16日

摘　要

玻璃的起源历史悠久，作为中西方早期贸易、文化交流的宝贵物证之一，玻璃极大地丰富了人类的物质文化。分析玻璃文物类型和化学成分之间的关系，对玻璃制作工艺、来源等的研究具有非常重要的科学指导意义。本文对我国古代玻璃的化学成分进行了研究，首先使用线性回归分析和相关性分析确立了玻璃文物表面的风化化学成分，得到玻璃文物表面风化与类型的关系；然后使用K-means算法对玻璃文物进行亚类划分；最后综合均值法和偏最小二乘回归模型预测玻璃文物风化前化学成分含量，并基于决策树法鉴别未知玻璃文物所属的类型。本文对于古代玻璃文物的成分分析与类别鉴定具有一定的指导价值。

关键词

玻璃文物，成分体系，聚类分析，决策树法，热力图

Statistical Analysis of the Chemical Composition System of Ancient Glass

Maowei Zhou, Jiayi Huang, Xueting Sun, Fang Liu

School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing Jiangsu

Received: Feb. 4th, 2023; accepted: Mar. 9th, 2023; published: Mar. 16th, 2023

Abstract

Glass has a long history of origin. As one of the precious material evidence of the early trade and cultural exchanges between China and the West, glass has greatly enriched the material culture of human beings. The analysis of the relationship between the type and chemical composition of glass cultural relics is of great scientific guiding significance to the research of glass production technology and source. This paper studies the ancient Chinese glass. Firstly, we use the linear regression analysis and the correlation analysis to establish the weathering chemical composition of the glass artifact surface and obtain the relationship between surface weathering and type of glass

文章引用: 周冒伟, 黄佳怡, 孙雪婷, 刘芳. 玻璃文物化学成分体系的统计分析[J]. 建模与仿真, 2023, 12(2): 1069-1082. DOI: 10.12677/mos.2023.122101

cultural relics. Then, we use the K-means algorithm to subclassify the glass artifacts. Finally, we combine the mean method and partial least squares regression model to predict the pre-weathering chemical content of glass relics, and identify the type of unknown glass relics based on the decision tree method. This paper has certain guiding value for the composition analysis and category identification of ancient glass cultural relics.

Keywords

Glass Relics, Composition System, Cluster Analysis, Decision-Making Tree, Heat Map

1. 引言

古代玻璃的起源历史悠久，作为中西方早期贸易、文化交流的宝贵物证之一，玻璃极大地丰富了人类的物质文化，同时，大量的古代玻璃制品存留至今。玻璃的主要原料是石英砂，化学成分为二氧化硅。纯石英砂的熔点较高，为了降低熔化温度，需要添加助熔剂。根据所添加助熔剂的不同，可以将我国早期的玻璃体系大致分为铅钡玻璃和高钾玻璃，又可以根据文物表面有无风化，继续划分为有风化铅钡玻璃、无风化铅钡玻璃、有风化高钾玻璃和无风化高钾玻璃。分析玻璃文物类型和化学成分之间的关系，对玻璃制作工艺、来源等的研究具有非常重要的科学指导意义，同时有利于为古代中西方物质文化交流提供科学依据。

公元前 4000 年，美索不达米亚和埃及地区就开始生产玻璃的雏形——费昂斯[1]，中国古代的玻璃体系较之晚了近 2000 年，玻璃制品在本土地区经历了从舶来品到自主生产的过程，但无论美索不达米亚、埃及还是中国，早期的玻璃制品皆以珠饰为主[2]。因此探究我国古代玻璃文物的化学成分体系和种类划分，是研究玻璃技术发展、来源的必经之路，在此基础上深入分析，也有利于探讨古代中西方之间的文化演进、贸易交流、人群迁徙等。

化学成分分析是研究古代玻璃制品成分体系不可或缺的重要一环。国内外对古代玻璃文物制品的研究，主要是围绕玻璃发展史和化学技术层面对玻璃文物进行化学成分检测展开的[3]。而对于数学统计层面，预测玻璃文物风化前后的类型和化学成分的文献较少。因此本文希望能够建立合适的数学模型，对玻璃文物风化前后的类型和化学成分进行预测。

我国出土的玻璃文物众多，对古代玻璃的发展史和化学成分检测的研究一直是重点问题。沈从文结合史料记载和对出土文物的考察提出“中国古代玻璃制作技术从颗粒装饰品逐渐转化成小件玻璃制品，至晚在战国晚期完成”[4]。安佳瑶对玻璃的整体发展史进行概述，将中国古代玻璃进行阶段性分析[1]。干福熹带领团队从化学技术的层面对玻璃文物进行了大量检测技术的研究，主要采用 PIXE 和 EDXRF 的方法，证明了古代最早的高钾玻璃和铅钡玻璃产自中国，提出了中国玻璃“自创说”理论[5] [6] [7] [8] [9]。成倩等人利用微损的激光剥离技术对文物进行化学成分检测，对文物的损害很小[10]。随着科技的不断发展，化学成分分析向着快速、准确、无损的方向发展。

为了在数学统计层面建立更精准的模型对玻璃文物进行化学成分分析和类型鉴别，本文对高教社杯数学建模 C 题中提供的实验数据进行建模，使用 SPSS、Python，通过回归分析、相关性分析、K-means 算法和决策树等，确立了玻璃文物表面的风化化学成分，得到玻璃文物表面风化与类型的关系，预测风

化前的化学成分含量；同时对高钾、铅钡玻璃进行亚类划分，进而鉴别未知玻璃文物所属的类型。

2. 成分体系

本章给出我国古代玻璃文物风化与化学成分之间的相关性分析，并对高钾、铅钡玻璃进行亚类划分。

我们收集到一批古代玻璃文物的相关数据在表单 2 中，进行数据预处理：1) 表单中空白处表示未检测到该化学成分，将其用 0 填充；2) 将成分比例累加之和不在 85%~105%的数据删除；3) 将表单剔除异常值后的化学成分数据进行归一化处理，使其累加和为 100%；4) 将“文物采样点”分开成“文物编号”和“采样点”，并对表单 1 和 2 进行合并。部分结果见如下图 1：

	文物编号	采样点	是否风化	是否高钾	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
0	01		0	1	71.03	0.00	10.23	6.47	0.89	4.03	1.78	3.96	0.00	0.00	1.20	0.00	0.00	0.40
1	02		1	0	36.32	0.00	1.05	2.34	1.18	5.74	1.86	0.26	47.48	0.00	3.57	0.19	0.00	0.00
2	03	部位1	0	1	87.05	0.00	5.19	2.01	0.00	4.06	0.00	0.78	0.25	0.00	0.66	0.00	0.00	0.00
3	03	部位2	0	1	62.41	0.00	12.51	5.94	1.12	5.56	2.18	5.15	1.43	2.89	0.71	0.10	0.00	0.00
4	04		0	1	68.58	0.00	10.07	7.41	1.62	6.70	2.14	2.27	0.00	0.00	0.82	0.00	0.00	0.37
5	05		0	1	63.81	0.00	11.35	7.62	1.83	7.77	2.71	3.39	0.00	0.00	0.97	0.06	0.00	0.49
6	06	部位1	0	1	68.39	0.00	7.45	0.00	2.00	11.27	2.42	2.54	0.20	1.40	4.23	0.11	0.00	0.00
7	06	部位2	0	1	60.51	0.00	7.77	5.47	1.75	10.17	6.11	2.21	0.35	0.98	4.55	0.12	0.00	0.00
8	07		1	1	92.91	0.00	0.00	1.07	0.00	1.99	0.17	3.25	0.00	0.00	0.61	0.00	0.00	0.00
9	08		1	0	20.18	0.00	0.00	1.48	0.00	1.34	0.00	10.43	28.73	31.29	3.60	0.37	0.00	2.58
10	08	严重风化点	1	0	4.69	0.00	0.00	3.25	0.00	1.13	0.00	3.20	33.03	31.17	7.70	0.54	0.00	15.30
11	09		1	1	95.24	0.00	0.59	0.62	0.00	1.32	0.32	1.55	0.00	0.00	0.35	0.00	0.00	0.00
12	10		1	1	96.95	0.00	0.92	0.21	0.00	0.81	0.26	0.84	0.00	0.00	0.00	0.00	0.00	0.00
13	11		1	0	35.21	0.00	0.22	3.68	0.74	2.82	0.00	5.17	26.62	15.32	9.83	0.39	0.00	0.00
14	12		1	1	94.70	0.00	1.01	0.72	0.00	1.47	0.29	1.66	0.00	0.00	0.15	0.00	0.00	0.00
15	13		0	1	60.13	2.91	12.77	8.86	0.00	6.28	2.93	4.82	0.00	0.00	1.29	0.00	0.00	0.00
16	14		0	1	63.10	3.41	12.40	8.31	0.67	9.32	0.51	0.47	1.64	0.00	0.16	0.00	0.00	0.00
18	16		0	1	66.23	2.13	14.75	8.40	0.53	6.28	0.43	1.09	0.11	0.00	0.00	0.04	0.00	0.00
20	18		0	1	81.71	0.00	9.69	0.00	1.57	3.14	0.00	0.00	0.00	0.00	1.40	0.07	2.43	0.00
21	19		1	0	30.01	0.00	0.00	2.97	0.60	3.61	1.35	3.55	43.36	5.42	8.94	0.19	0.00	0.00
22	20		0	0	42.26	0.00	0.80	0.00	0.00	6.16	1.71	5.41	10.52	26.64	6.50	0.00	0.00	0.00
23	21		0	1	77.83	0.00	0.00	4.78	1.24	6.28	2.41	3.33	1.02	2.00	1.12	0.00	0.00	0.00
24	22		1	1	92.35	0.00	0.74	1.66	0.64	3.50	0.35	0.55	0.00	0.00	0.21	0.00	0.00	0.00
25	23	未风化点	1	0	55.74	8.21	0.00	0.52	0.74	1.47	0.00	3.10	17.60	12.29	0.00	0.34	0.00	0.00
26	24		0	0	32.30	0.00	0.00	0.48	0.00	1.61	0.00	8.56	29.47	26.53	0.14	0.92	0.00	0.00
27	25	未风化点	1	0	52.14	2.38	0.00	0.65	0.00	1.96	1.60	1.15	32.87	6.85	0.20	0.21	0.00	0.00
28	26		1	0	19.83	0.00	0.00	1.44	0.00	0.70	0.00	10.59	29.58	32.31	3.14	0.45	0.00	1.96
29	26	严重风化点	1	0	3.72	0.00	0.40	3.01	0.00	1.18	0.00	3.60	29.95	35.49	6.05	0.62	0.00	15.97
30	27		1	1	93.84	0.00	0.00	0.95	0.55	2.54	0.20	1.56	0.00	0.00	0.36	0.00	0.00	0.00

Figure 1. Part of the preprocessed experimental data
图 1. 部分预处理后的实验数据

2.1. 相关性分析

要分析玻璃表面有无风化化学成分，采用多元线性回归分析与相关性分析相结合的方式。根据处理后的数据，以 14 个化学成分作为输入，是否风化作为输出，建立多元线性回归模型，计算 14 个指标的权重，规定大于某一阈值的为风化化学成分；同时进行相关性分析，验证我们所得结果的合理性。

2.1.1. 确定回归系数

令输入变量 $x_i\left(i=1,2,\cdots,14\right)$ 分别表示十四种化学成分的百分含量，输出变量 y 表示是否风化。y = 1 表示风化；y = 0 表示没有风化。

建立多元线性回归分析模型：

$$\begin{cases} y=\beta_0+\beta_1x_1+\cdots+\beta_0\beta_0+\varepsilon \\ \varepsilon\sim N\left(0,\sigma^2\right) \end{cases}$$

其中，$\beta_0,\beta_1,\cdots,\beta_m,\sigma^2$ $(m=14)$都是与 $x_1,x_2,\cdots,x_m$ 无关的未知参数，$\beta_0,\beta_1,\cdots,\beta_m$ 称为回归系数。

得到独立观测数据：

$$\begin{cases} b_i=\beta_0+\beta_1\beta_{i1}+\cdots+\beta_m\beta_{im}+\beta_i \\ \varepsilon_i\sim N\left(0,\sigma^2\right),i=1,\cdots,n \end{cases}$$

其中，b_i 为 y 的观察值，$a_{i1},\cdots,a_{im}$ 分别为 $x_1,x_2,\cdots,x_m$ 的观察值，i = 1~n $(n>m)$。

记

$$X=\begin{pmatrix}1 & a_{11} & \cdots & a_{1m}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & a_{n1} & \cdots & a_{nm}\end{pmatrix},\ Y=\begin{pmatrix}b_1\\ \vdots\\ b_n\end{pmatrix}$$

$$\varepsilon=\left(\varepsilon_1,\cdots,\varepsilon_n\right)^{\mathrm{T}},\ \beta=\left(\beta_0,\beta_1,\cdots,\beta_m\right)^{\mathrm{T}}$$

可得

$$\begin{cases}Y=X\beta+\varepsilon\\ \varepsilon\sim N\left(0,\sigma^2E_n\right)\end{cases}$$

其中，E_n 为 n 阶单位矩阵(n 为数据的样本数)。

利用 Python，将处理后的数据代入模型，进行模型的求解。根据以上模型建立过程，我们得到这 14 个化学成分对风化程度的权重及影响见如下表 1：

Table 1. Regression coefficients of the multiple linear regression model
表 1. 多元线性回归模型的回归系数

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)
−0.0557	0.0331	−0.8732	−0.1787	−0.0272	−0.0309	−0.2892
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)
−0.0403	0.1809	0.0354	0.1806	0.0540	−0.2316	−0.0061

由表 1，设置阈值为 0.1，取绝对值大于 0.1 的化学元素即氧化钾(K_2O)，氧化钙(CaO)，氧化铁(Fe_2O_3)，氧化铅(PbO)，五氧化二磷(P_2O_5)，氧化锡(SnO_2)，并认为这些化学成分对风化程度的影响相对较大，其余可以近似忽略。

2.1.2. 基于 Pearson 相关系数的风化相关性热力图

在实际情况下，化学成分对风化程度的影响很难是完全线性的，且各成分之间必然存在互相影响的情况，因此我们在多元线性回归分析的基础上，进行了相关性分析来更全面地分析化学成分对风化程度的影响。

假设有两组数据 $X\left\{X_1,X_2,\cdots,X_n\right\}$ 和 $Y\left\{Y_1,Y_2,\cdots,Y_n\right\}$ $(n=15)$。X 和 Y 代表是否风化和 14 种化学成分。当表面风化时取 1，没有发生风化时取 0。

样本均值为：

$$\bar{X}=\frac{\sum_{i=1}^{n}X_i}{n},\ \bar{Y}=\frac{\sum_{i=1}^{n}Y_i}{n}$$

样本协方差：

$$Cov\left(X,Y\right)=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{n-1}$$

样本 Pearson 相关系数[11]：

$$r_{XY}=\frac{Cov\left(X,Y\right)}{S_XS_Y}$$

其中，S_X，S_Y为样本标准差，公式如下：

$$S_X=\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}},S_Y=\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}{n-1}}$$

采用 SPSS 绘制出各化学成分的 Pearson 相关系数，见如下图 2：

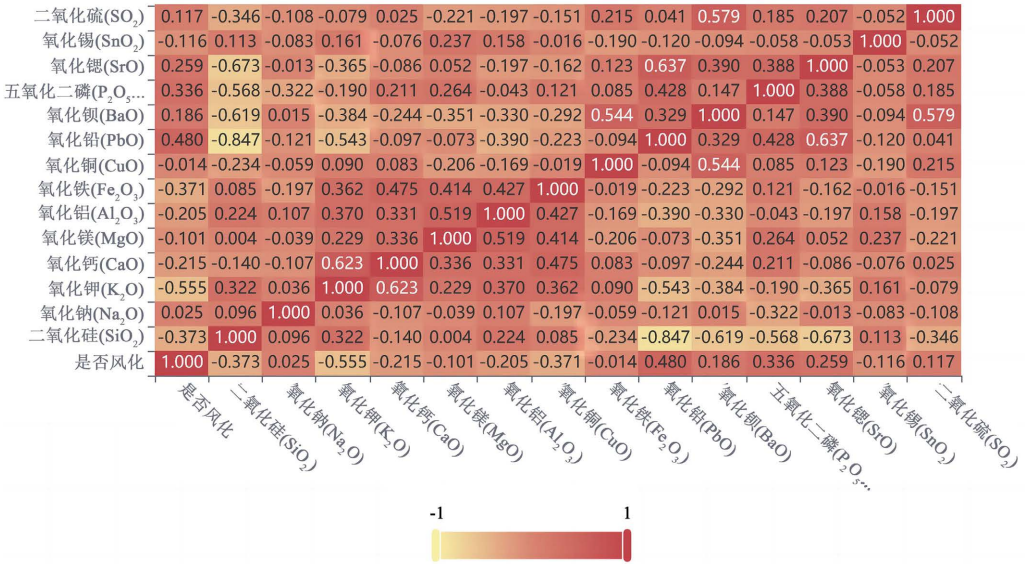

Figure 2. Heat map of chemical composition and weathering correlation
图 2. 化学成分与风化相关性热力图

由图 2，我们取阈值为 0.2，满足风化系数大于 0.2 的有 8 种化学成分，结果见如下表 2：

Table 2. Pearson Chemical composition with a correlation coefficient greater than 0.2
表 2. Pearson 相关系数大于 0.2 的化学成分

化学成分	二氧化硅	氧化钾	氧化钙	氧化铝
Pearson 系数	−0.373	−0.555	−0.215	−0.205
化学成分	氧化铁	氧化铅	五氧化二磷	氧化锶
Pearson 系数	−0.371	0.480	0.336	0.259

2.1.3. 总结

综合考虑，取两种的交集作为对风化程度有较大影响的化学成分：氧化钾(K_2O)，氧化钙(CaO)，氧化铁(Fe_2O_3)，氧化铅(PbO)，五氧化二磷(P_2O_5)。

最后，我们对五个风化化学成分和“类型是否为高钾玻璃”作风化相关性热力图，结果见如下图 3：

由图 3，可以发现：

1) 高钾玻璃风化程度与氧化钾、氧化钙、氧化铁含量关系成正相关，而铅钡玻璃的风化程度与氧化铅，五氧化二磷关系成正相关。

2) 在风化情况下，玻璃文物是高钾玻璃还是铅钡玻璃与氧化钾和氧化铅的含量有关。

3) 高钾玻璃是否风化与氧化钾的占比含量相关性较强；铅钡玻璃是否风化与氧化铅的占比含量相关性较强。

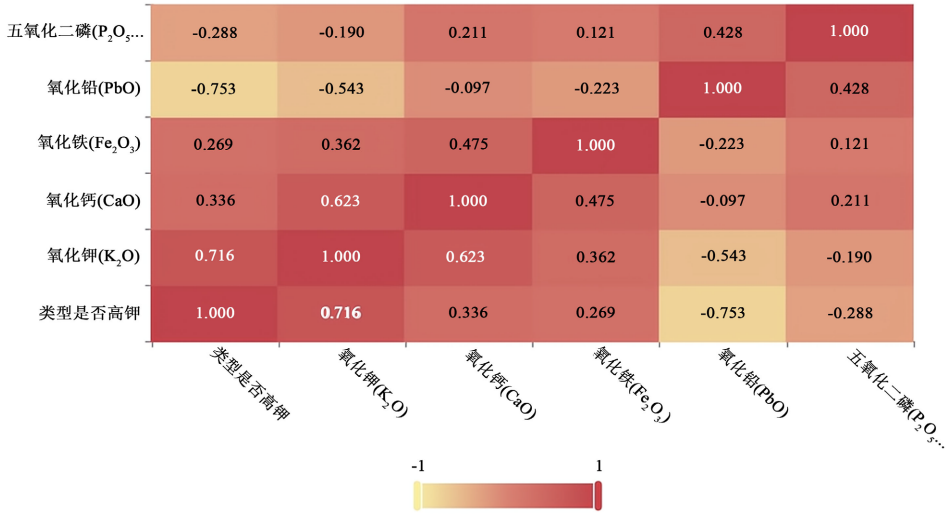

Figure 3. Heat map of weathering chemical composition correlation
图 3. 风化化学成分相关性热力图

2.2. 亚类划分

2.2.1. 确定分类中心

为了对高钾玻璃和铅钡玻璃进行更加细类的划分，我们可以采用聚类分析法。聚类分析是一种无监督学习，用于对未知类别的样本进行划分将它们按照一定的规则划分成若干个类簇，把相似(相关的)的样本聚在同一个类簇中，把不相似的样本分为不同类簇，从而分析样本之间内在的性质以及相互之间的联系规律。

K-means 算法是基于划分的聚类算法，计算样本点与簇质心的距离，与簇质心相近的样本点划分为同一类簇。两个样本距离越远，则相似度越低。

基本流程：

1) 设定 k 值，表示需要将数据分成 k 个簇，从样本点中随机选择 k 个点作为初始簇中心。

2) 分别计算每个样本点到各个初始簇中心的距离，将每个样本点划分到距离它最近的中心点。

3) 用各簇中所有样本的质心(即为均值，向量的各个维度分别取平均)代替原有的中心点。

4) 重复步骤 2 和 3，直到中心点不变或达到预定迭代次数时，算法终止。

迭代计算就是一个优化目标的过程，这是机器学习算法必不可少的一步，这里优化的是各个数据点到中心点的距离，距离越小越好。

根据预处理后的数据，利用 python，我们可以得到高钾和铅钡玻璃分类的中心见如下表 3：

Table 3. High potassium glass and lead barium glass classification center
表 3. 高钾玻璃和铅钡玻璃分类的中心

	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
铅钡中心	28.06	0.33	0.15	2.72	0.59	2.70	0.69
高钾中心	68.90	1.21	3.38	2.50	0.80	5.37	1.03
	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
铅钡中心	2.32	43.79	12.19	4.84	0.42	0.04	1.16
高钾中心	1.73	9.27	4.42	1.01	0.13	0.11	0.14

由表 3 确定的类中心，利用实验数据验证，对比结果发现，67 组数据中心正确判断了 58 组数据，正确率达 86.6%。故根据上表和化学成分的占比含量，可以对高钾玻璃和铅钡玻璃进行较为准确的分类。

2.2.2. 对高钾、铅钡玻璃进行亚类划分

在 2.2.1 的基础上，我们先设定 k = 3，使用 python 运行程序，得到玻璃亚类中心见如下表 4、表 5：

Table 4. High-potassium glass subclass center table
表 4. 高钾玻璃亚类中心表

	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
亚类 1	94.33	0.00	0.54	0.87	0.20	1.94	0.27
亚类 2	64.91	0.94	11.03	6.50	1.16	7.49	2.36
亚类 3	82.20	0.00	4.96	2.26	0.94	4.49	0.80
	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
亚类 1	1.57	0.00	0.00	0.28	0.00	0.00	0.00
亚类 2	2.88	0.41	0.59	1.55	0.05	0.00	0.14
亚类 3	1.37	0.42	0.67	1.06	0.02	0.81	0.00

Table 5. Lead barium glass subclass center table
表 5. 铅钡玻璃亚类中心表

	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
亚类 1	27.82	0.22	0.18	3.03	0.76	3.04	0.85
亚类 2	58.72	1.93	0.19	1.16	0.71	5.16	0.64
亚类 3	16.14	0.00	0.08	1.93	0.00	1.19	0.00
	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
亚类 1	1.42	48.18	8.68	5.32	0.43	0.06	0.00
亚类 2	1.21	20.35	8.26	1.18	0.23	0.07	0.17
亚类 3	7.27	30.15	31.36	4.12	0.58	0.00	7.16

表 4、表 5 对高钾玻璃和铅钡玻璃进行了具体的划分。

2.2.3. 合理性和敏感性分析

我们可以采用计算误差平方和 SSE、轮廓系数、CH 指标来评估聚类分析的好坏，本文采用轮廓系数。公式如下：

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

其中，$a(i)$ 表示化学成分的内聚度，计算公式如下：

$$a(i)=\frac{\sum_{j\neq 1}^{n}\text{distance}(i,j)}{n-1}$$

其中 j 表示与样本 i 在同一亚类内的其他样本点，distance 表示 i 与 j 的距离。$a(i)$ 越小说明该类越紧密。$b(i)$ 计算公式与 $a(i)$ 相类似。

$$S(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases}$$

当 $a(i)<b(i)$ 时，类内距离小于类间距离，聚类更紧凑，S 的值趋近于 1。轮廓系数 S 的取值范围为 [−1, 1]，轮廓系数越大，聚类效果越好[12]。

我们分别令 $k=1$，2，3，4，5，计算得到了各个情形下的轮廓系数，见如下图 4：

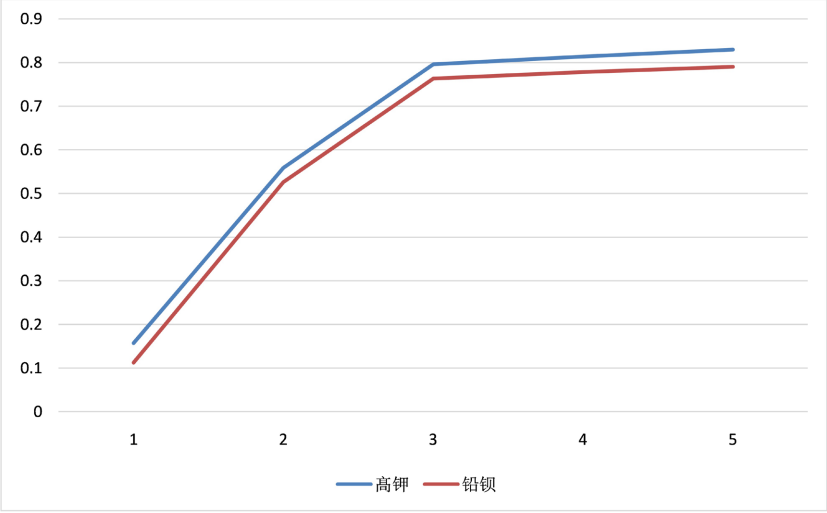

Figure 4. Plot of the contour coefficient versus k
图 4. 轮廓系数与 k 的关系图

由图 4 可知：

1) 高钾玻璃的聚类效果要始终优于铅钡玻璃，分类结果更加合理；

2) 当 $k<3$ 时，轮廓系数随着 k 的增大而明显增大，当 $k>3$ 时，增大速度放缓且趋近于一定值，K-means 聚类分析效果变化不大，分类结果已经较为良好，故 k 取 3 具有一定的合理性；

3) K-means 聚类分析模型对 k 的取值的敏感度较大，故选择一个合适的 k 对聚类分析十分重要。

保持 k = 3，改变化学成分个数，分析对聚类效果的影响。化学成分分别取 14 种，8 种(二氧化硅，氧化钾，氧化钙，氧化铝，氧化铁，氧化铅，五氧化二磷，氧化锶)，5 种(氧化钾，氧化钙，氧化铁，氧化铅，五氧化二磷)，结果见如下图 5：

由图 5 可知：

1) 化学成分个数越多，聚类效果越好。

2) 利用第一问求得的与风化相关性较强的化学成分，基本能够对高钾玻璃和铅钡玻璃进行分类。

3) 三种情况下，高钾玻璃的聚类效果都要优于铅钡玻璃。

4) 轮廓系数随化学成分个数的变化幅度较大，故聚类分析对化学成分个数的敏感度较大。

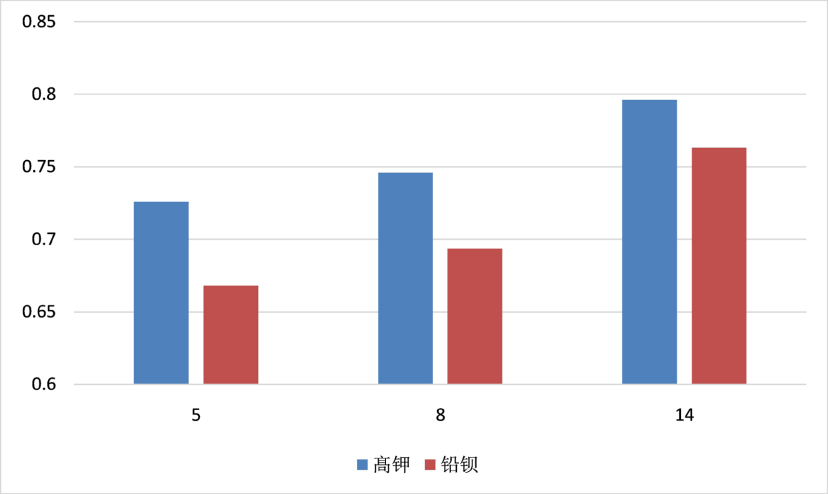

Figure 5. Diagaph of contour coefficient and number of chemical compositions
图 5. 轮廓系数与化学成分个数的关系图

3. 模型建立

本章综合均值法和偏最小二乘回归模型预测玻璃文物风化前化学成分含量，并基于决策树法鉴别未知玻璃文物所属的类型。

3.1. 预测玻璃文物风化前化学成分含量

3.1.1. 均值法预测风化前化学成分含量

设不同类型的玻璃文物风化前的化学成分为 c_{ij}，具体表示第 i 次校验的第 j 个化学成分。

那么，

$$C_j = \frac{1}{n_1}\sum_{i=1}^{n} c_{ij},\ (j = 1,2,\cdots,m)$$

其中，m 表示 m 种化学成分。

同理，设不同类型的玻璃文物风化后的化学成分为 d_{ij}，具体表示第 i 次校验的第 j 个化学成分。

那么，

$$D_j = \frac{1}{n_2}\sum_{i=1}^{n} d_{ij},\ (j = 1,2,\cdots,m)$$

易得知，风化过程的各化学成分的变化率

$$E_j = \frac{C_j - D_j}{D_j} \times 100\%,\ (j = 1,2,\cdots,m)$$

根据以上公式对实验数据进行处理，分别得到高钾玻璃和铅钡玻璃风化前后各个化学成分的均值以及变化率，结果见如下表 6。

由表 6 我们对给定风化后数据的部分进行还原，使用公式

$$C_j = D_j(1 + E_j)$$

即可得到不同类型的玻璃文物风化前的化学成分含量。

Table 6. Mean value and rate of change of each chemical composition before and after weathering
表 6. 风化前后各化学成分的均值及变化率

	高钾/风化后	高钾/风化前	高钾/变化率	铅钡/风化后	铅钡/风化前	铅钡/变化率
二氧化硅	94.3310	69.2315	−0.3625	34.5180	54.6896	0.3688
氧化钠	0.0000	0.7052	1.0000	0.9757	0.7929	−0.2305
氧化钾	0.5446	9.5150	0.9428	0.1451	0.2692	0.4610
氧化钙	0.8732	5.4404	0.8395	2.4311	1.2533	−0.9397
氧化镁	0.1978	1.1025	0.8206	0.7248	0.4996	−0.4508
氧化铝	1.9378	6.7386	0.7124	3.9297	3.2951	−0.1926
氧化铁	0.2659	1.9688	0.8649	0.5735	0.9617	0.4037
氧化铜	1.5684	2.5003	0.3727	2.0435	1.6191	−0.2621
氧化铅	0.0000	0.4163	1.0000	38.1319	24.1208	−0.5809
氧化钡	0.0000	0.6057	1.0000	10.7877	10.8813	0.0086
五氧化二磷	0.2813	1.4260	0.8028	4.3104	0.9683	−3.4515
氧化锶	0.0000	0.0424	1.0000	0.3770	0.3017	−0.2493
氧化锡	0.0000	0.2022	1.0000	0.0569	0.0657	0.1346
二氧化硫	0.0000	0.1051	1.0000	0.9949	0.2816	−2.5330

3.1.2. 偏最小二乘回归分析模型预测风化前化学成分含量

本文我们建立偏最小二乘回归分析模型预测风化前化学成分含量[13]：

设风化后 14 种化学成分含量为 $x_i\left(i=14\right)$，风化前 14 种化学成分含量为 $y_i\left(i=14\right)$。自变量的观测数据矩阵记为 $A=\left(a_{ij}\right)_{n*n}$，因变量的观测数据矩阵记为 $B=\left(B_{ij}\right)_{n*n}$。$n$ 为训练集的样本个数。

1) 数据标准化。将各指标值 a_{ij} 转化为标准化指标值 $\tilde{a}_{ij}$，有

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j^{(1)}}{s_j^{(1)}},\ i=1,2,\cdots,n.j=1,2,\cdots,14.$$

其中：$\mu_j^{(1)}=\frac{1}{n}\sum_{i=1}^{n}a_{ij}$；$s_j^{(1)}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(a_{ij}-\mu_{ij}^{(1)}\right)^2}$，即 $\mu_j^{(1)}$，$s_j^{(1)}$ 为第 j 个自变量 x_j 的样本均值和样本标准差。

对应地，称

$$\tilde{x}_j=\frac{x_j-\mu_j^{(1)}}{s_j^{(1)}},\ j=1,2,\cdots,14.$$

为标准化指标变量。

类似的，将 b_{ij} 转化为标准化指标值 $\tilde{b}_{ij}$，有

$$\tilde{b}_{ij}=\frac{b_{ij}-\mu_j^{(2)}}{s_j^{(2)}},\ i=1,2,\cdots,n.j=1,2,\cdots,14.$$

其中 $\mu_j^{(2)}=\frac{1}{n}\sum_{i=1}^{n}b_{ij}$；$s_j^{(2)}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(b_{ij}-\mu_{ij}^{(2)}\right)^2}$，即 $\mu_j^{(2)}$，$s_j^{(2)}$ 为第 j 个因变量 y_j 的样本均值和样本标准差。

对应地，称

$$\tilde{y}_j=\frac{y_j-\mu_j^{(2)}}{s_j^{(2)}},\ j=1,2,\cdots,14.$$

为标准化指标变量。

利用 corrcoef 函数求得 28 个变量之间的相关系数矩阵。进而分别提出标准化后自变量组 $\tilde{x}_j$ 和因变量组 $\tilde{y}_j$ 的数据。

求两个成分对时标准化指标变量与成分变量之间的回归方程。求得自变量组 $\tilde{x}_j$ 与因变量组 $\tilde{y}_j$ 之间的回归方程。最后将标准化变量 $\tilde{x}_j$、$\tilde{y}_j$ $(j = 1, 2)$分别还原为原始变量 x_j、y_j，得到回归方程。

3.1.3. 模型的求解

根据求得的风化前后各化学成分的均值及变化率，可以分析得到高钾、铅钡玻璃的分类规律有如下几点：

1) 对于高钾玻璃，风化后只有二氧化硅的含量大幅上升，其他化学成分含量都显著降低，大部分化学成分含量可以忽略不计。

2) 对于铅钡玻璃，风化后氧化钠、氧化钙、氧化镁、氧化铝、氧化铜、氧化铅、五氧化二磷、氧化锶、二氧化硫占比含量变多，其余变低。

通过 Matlab，求得偏最小二乘回归分析部分回归系数直方图，见如下图 6、图 7：

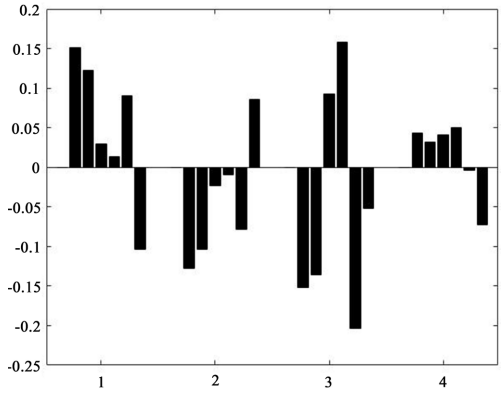

Figure 6. Histogram of partial least squares regression
图 6. 偏最小二乘回归部分回归系数直方图

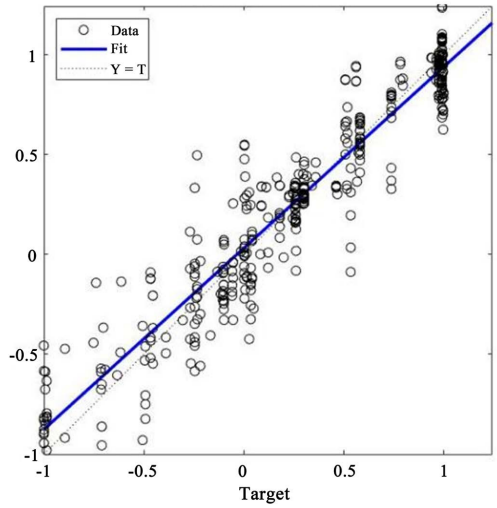

Figure 7. Fitting results of the regression equation
图 7. 回归方程的拟合结果

由图 6、图 7 可知，利用得到的回归方程去检验均值法预测出来的风化前化学成分含量的结果，拟合优度 $R^2 = 0.8143$，拟合效果较高，说明均值法所得结果较为合理。能够根据风化点的检测数据，预测风化前 14 种化学成分的含量。

3.2. 鉴别未知玻璃文物所属类型

3.2.1. 决策树法鉴别未知玻璃文物的类型

决策树是通过对训练样本进行归纳学习生成决策树或决策规则，然后使用决策树或决策规则对新数据进行分类的一种数学方法。决策树是一个树型结构，它由一个根结点、一系列内部结点及叶结点组成，每一结点只有一个父结点和两个或多个子结点，结点间通过分支相连。决策树的每个内部结点对应一个非类别属性或属性的集合(也称为测试属性)，每条边对应该属性的每个可能值。决策树的叶结点对应一个类别属性值，不同的叶结点可以对应相同的类别属性值[14]。

决策树的生成通常采用自顶向下的递归方式，在构造过程中，采用信息增益度量。信息增益最大表明了数据集中在分类过程中能够最大化减少不确定性，因此具有更好的分类效果。信息熵(H)以及信息增益(G)可定义如下：

$$H_{(p)} = -\sum p \times \log p$$

$$H_{(Y|X)} = \sum_{i=1}^{n} p_i H_{(Y|X=X_i)}$$

$$G_{(D,A)} = H_{(D)} - H_{(D|A)}$$

其中 p 表示随机变量的概率，A 表示特征，D 表示数据集，$H_{(D)}$ 表示经验熵，$H_{(Y|X)}$ 表示条件熵，$H_{(D|A)}$ 表示特征 A 在数据集 D 的条件下的经验条件熵。

3.2.2. 模型求解

利用 SPSS，将预处理后的数据，以化学成分和玻璃文物表面是否风化作为输入，类型作为模型的输出，训练决策树模型。最后将实验数据附件表单 3 [15]中未知玻璃类别的化学成分和表面风化作为输入，预测最终结果见如下表 7：

Table 7. Identification results of glass relics by decision tree method
表 7. 决策树法对玻璃文物的鉴别结果

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

3.2.3. 准确性和灵敏性分析

利用 SPSS，分别训练逻辑回归、向量机、随机森林模型，并与决策树法所预测结果相比较。结果见如下表 8：

Table 8. Identification results of glass artifacts by the four models
表 8. 四种模型对玻璃文物的鉴别结果

文物编号	逻辑回归	决策树	向量机	随机森林
A1	高钾	高钾	高钾	高钾
A2	铅钡	铅钡	铅钡	铅钡
A3	铅钡	铅钡	铅钡	铅钡
A4	铅钡	铅钡	铅钡	铅钡
A5	铅钡	铅钡	铅钡	铅钡
A6	高钾	高钾	高钾	高钾
A7	高钾	高钾	高钾	高钾
A8	铅钡	铅钡	铅钡	铅钡

由图可知，四种模型对实验数据附件表单 3 [15]中未知玻璃文物所属类型的鉴别结果完全一致。

在前面求得高钾玻璃和铅钡玻璃类中心的基础上，求未知玻璃的化学成分与类中心的欧氏距离，比较他们匹配程度的差异性，进而进行灵敏度分析。

设玻璃文物的化学成分为 $P_i\left(i=1,2,\cdots,14\right)$，高钾玻璃化学成分的类中心为 Q_{1i}，铅钡玻璃化学成分的类中心为 Q_{0i}。

则要求欧式距离之和尽量的小，即

$$\min d=\sum_{1}^{14}\left|P_i-Q_i\right|。$$

代入 Matlab 计算结果见如下表 9：

Table 9. Sensitivity analysis of the classification results
表 9. 分类结果的灵敏度分析

文物编号	铅钡中心	高钾中心	属性
A1	122.54	36.46	高钾
A2	51.78	90.56	铅钡
A3	34	87.76	铅钡
A4	48.84	72.5	铅钡
A5	25.06	99.76	铅钡
A6	133.52	49.44	高钾
A7	132.28	44.22	高钾
A8	57.86	61.88	铅钡

由表 9 可知：

1) 欧氏距离分析得到的结果与逻辑回归、决策树、向量机、随机森林相一致。

2) 鉴别未知玻璃文物属性对二氧化硅、氧化钾、氧化铅的敏感性较大，对其他化学成分的敏感性较小。

4. 结论

本文对所给的实验数据进行数据预处理，分析并建立数学模型，研究了玻璃文物表面有无风化化学成分含量的统计规律和不同类别玻璃文物样品化学成分之间的关联关系，并根据风化点检测数据预测玻璃文物风化前的化学成分含量，鉴别给定的玻璃文物的类别。

首先，我们将多元线性回归分析和相关性分析相结合，探讨了玻璃文物化学成分和表面风化之间的统计关系，发现对风化程度有较大影响的化学成分有五种：氧化钾(K_2O)，氧化钙(CaO)，氧化铁(Fe_2O_3)，氧化铅(PbO)，五氧化二磷(P_2O_5)。

然后，基于玻璃文物风化与化学成分存在相关性的结论，采用 K-means 聚类分析算法确定类中心，并对高钾玻璃和铅钡玻璃进行更加细致的亚类划分。

最后，基于不同类别玻璃文物样品化学成分之间的关联关系，将均值法和偏最小二乘回归分析模型相结合，建立了玻璃文物化学成分预测模型，根据风化点的检测数据，预测了风化前 14 种化学成分的含量。并基于决策树法建立了玻璃类型鉴别模型，利用欧式距离分析发现，所建立的鉴别模型准确率较高。本文对于古代玻璃文物的成分分析与类别鉴定具有一定的指导价值。

参考文献

[1] 安佳瑶. 玻璃史话[M]. 北京: 社会科学文献出版社, 2011: 7-11.

[2] 赵志强. 新疆巴里坤石人子沟遗址群出土玻璃珠的成分体系与制作工艺研究[D]: [硕士学位论文]. 西安: 西北大学, 2016.

[3] 刘淑娜. 中国古代北方游牧民族玻璃器研究[D]: [硕士学位论文]. 呼和浩特: 内蒙古师范大学, 2022.

[4] 沈从文. 玻璃史话[M]. 沈阳: 万卷出版公司, 2005: 1-5.

[5] 干福熹. 中国古代玻璃技术的发展[M]. 上海: 上海科学技术出版社, 2005.

[6] 干福熹. 中国古代玻璃的起源和发展[J]. 自然杂志, 2006(4): 192.

[7] 干福熹, 赵虹霞, 李青会, 李玲, 承焕生. 湖北省出土战国玻璃制品的科技分析与研究[J]. 江汉考古, 2010(2): 108-116.

[8] 干福熹, 黄振发, 肖炳荣. 我国古代玻璃的起源问题[J]. 硅酸盐学报, 1978(Z1): 99-104.

[9] 干福熹, 承焕生, 李青会. 中国古代玻璃的起源——中国最早的古代玻璃研究[J]. 中国科学(E 辑: 技术科学), 2007(3): 382-391.

[10] 成倩, 王博, 郭金龙, 崔剑锋. 丝绸之路且末古国墓地出土玻璃器成分特点研究[J]. 玻璃与搪瓷, 2012, 40(2): 21-29.

[11] 张宇镭, 党琰, 贺平安. 利用 Pearson 相关系数定量分析生物亲缘关系[J]. 计算机工程与应用, 2005(33): 83-86.

[12] 朱连江, 马炳先, 赵学泉. 基于轮廓系数的聚类有效性分析[J]. 计算机应用, 2010, 30(S2): 139-141.

[13] 司守奎, 孙玺菁. 数学建模算法与应用[M]. 第 3 版. 北京: 国防工业出版社, 2021: 358-364.

[14] 夏慧维. 基于决策树集成和宽度森林的网络流量分析与预测研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2020.

[15] 中国工业与应用数学学会. 2022 高教社杯全国大学生数学建模竞赛赛题[EB/OL]. http://www.mcm.edu.cn/html_cn/node/5267fe3e6a512bec793d71f2b2061497.html, 2022-09-15.