教学与竞赛

古代玻璃制品成分分析与鉴别的统计建模

宛 惠，邓明华

（北京大学 数学科学学院，北京 100871）

摘 要：给出 2022 年“高教社杯”全国大学生数学建模竞赛 C 题“古代玻璃制品的成分分析与鉴别”可行的解法，并对赛题评阅进行简要评述. 该题旨在通过对古代玻璃制品的化学成分数据分析，探索文物分类方法. 本文在充分考虑数据的成分性特点基础上，采用中心对数比变换将数据从单纯型映射到欧式空间，在变换后的空间上进行相应的统计分析.

关键词：成分数据；中心对数比变换；有监督分类；无监督聚类；关联分析

中图分类号：O29 **文献标志码**：A **文章编号**：2095-3070(2023)02-0027-14

DOI：10.19943/j.2095-3070.jmmia.2023.02.03

1 问题背景

丝绸之路是古代中西方文化交流的通道，其中玻璃是早期贸易往来的宝贵物证. 早期的玻璃在西亚和埃及地区常被制作成珠形饰品传入我国，我国古代玻璃吸收其技术后在本土就地取材制作，因此与外来的玻璃制品外观相似，但化学成分却不相同.

玻璃的主要原料是石英砂，主要化学成分是二氧化硅(SiO_2). 由于纯石英砂的熔点较高，为了降低熔化温度，在炼制时需要添加助熔剂. 古代常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等，并添加石灰石作为稳定剂，石灰石煅烧以后转化为氧化钙(CaO). 添加的助熔剂不同，制成的玻璃的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，氧化铅(PbO)和氧化钡(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.

古代玻璃极易受埋藏环境的影响而风化. 在风化过程中，内部元素与环境元素进行大量交换，导致其成分比例发生变化，从而影响对其类别的正确判断. 如图 1 的文物标记为表面无风化，表面能明显看出文物的颜色和纹饰，但不排除局部有较浅的风化；图 2 的文物标记为表面风化，表面大面积灰黄色区域为风化层，是明显风化区域，紫色部分是一般风化表面. 在部分风化的文物中，其表面也有未风化的区域[1].

本题目旨在不同的场景下对玻璃文物样品进行成分分析与鉴别. 限于篇幅，对赛题不再赘述，具体参考大学生数学建模竞赛公布的原题[1].

收稿日期：2023-01-21

基金项目：国家自然科学基金(31871342)

通讯作者：邓明华，E-mail：dengmh@math.pku.edu.cn

引用格式：宛惠，邓明华. 古代玻璃制品成分分析与鉴别的统计建模[J]. 数学建模及其应用，2023，12(2)：27-40.

WAN H，DENG M H. Statistical modeling and analysis of ancient glass composition(in Chinese)[J]. Mathematical Modeling and Its Applications，2023，12(2)：27-40.

图1　未风化的蜻蜓眼玻璃珠样品

图2　风化的玻璃棋子样品

2　问题简析

本题通过对古代玻璃制品化学成分的数据分析，研究有无风化玻璃制品成分的变化规律，以及高钾、铅钡两种玻璃类型的化学成分的统计规律，并探索亚分类的方法，进而可以依据未知分类的文物化学成分对文物进行准确的分类. 本题数据的主要特点是成分性，即各化学成分比例的累加和应100%，具有定和约束，在统计学上称为“成分数据”. 同时由于定和约束，成分数据各变量之间具有明显的共线性，使得常规的统计分析方法失效，通常需要通过适当的变换来解决这类问题，比如中心对数比变换(centered log-ratio，CLR)[2]等.

问题1　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，是相关性分析问题，首先对数据进行预处理，补全缺失值，再通过列联表、卡方检验等方式探究表面风化与这3个特征是否显著相关；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. 这需要首先对成分数据进行预处理，删除无效数据，并对数据进行中心对数变换，填补缺失值，其次可再通过多元方差分析探究对于不同的玻璃类型，表面风化引起哪些化学成分的显著变化，对于存在显著变化的化学成分，可从总体均值差异角度探究其变化规律，并实现对已风化样本在风化前的化学成分含量的预测.

问题2　分析高钾玻璃、铅钡玻璃的分类规律，首先需对数据进行预处理，消除风化因素的影响，再通过总体均值、关键成分识别等方法找出两类玻璃的重要特征，从而识别两者的差异和分类规律；对于亚类划分，首先筛选出同一类样本里具有区分度的成分特征，并选择合适的亚类数，再通过K均值聚类得到最终的划分结果. 使用轮廓系数、戴维森堡丁指数及卡林斯基-哈拉巴斯指数这3个常用的聚类评价指标，使用基于熵权法的Topsis模型综合评价划分方案的优劣及合理性，并通过添加不同程度的高斯噪声，评价亚类划分的敏感性.

问题3　鉴别未知类别玻璃文物所属类型，首先采用与之前相同的预处理和数据变换形式，并分为风化和未风化两类进行讨论. 由问题2得出的分类规律，可对文物类型作出初步预测，进一步以上一问中有标签的预处理后的无风化样本为训练集训练随机森林模型，以未知文物为测试集进行预测. 对于分类结果的敏感性，一方面可对风化的未知样本进行还原，使用针对未风化样本训练的随机森林模型重新预测标签；另一方面可对数据添加高斯噪声，判断预测结果的稳健性.

问题4　针对不同类别的玻璃文物样品，去除成分性影响后分别计算各化学成分之间的相关系数并进行相关性检验，判断哪些化学成分有显著线性相关性(也可进一步考虑非线性关联)，并比较不同类别之间的化学成分关联关系的差异性.

3　问题求解

3.1　问题1

3.1.1　数据预处理

可以观察到，表单1中编号为19、40、48和58的4个文物样本的颜色特征存在缺失，这里首先根据类型、纹饰、表面风化进行分组，采用对应的颜色众数对缺失值进行填补，均填补颜色浅蓝.

对于表单2，首先删除成分比例累加和不在85%～105%之间的数据. 对于剩余的数据，根据表单

1 标注上各采样点对应的类型及表面风化情况. 若采样点化学成分存在缺失，将其填充为 0.01(由于接下来成分数据要进行中心对数变换，故填充一个较小的非零正数). 最后对各化学成分按比例进行归一化，使得各样本的化学成分之和均为 100%.

3.1.2　表面风化与 3 种特征的关系

首先，列联表统计各类型、纹饰、颜色对应表面是否风化的文物个数，详细信息见表 1－3.

表 1　不同类型的文物表面风化情况表

表面风化	类型		合计
	高钾	铅钡	
风化	6	28	34
无风化	12	12	24
合计	18	40	58

表 2　不同纹饰的文物表面风化情况表

表面风化	纹饰			合计
	A	B	C	
风化	11	6	17	34
无风化	11	0	24	24
合计	22	6	58	58

表 3　不同颜色的文物表面风化情况表

表面风化	颜色								合计
	黑	蓝绿	绿	浅蓝	浅绿	深蓝	深绿	紫	
风化	2	9	0	16	1	0	4	2	34
无风化	0	6	1	8	2	2	3	2	24
合计	2	15	1	24	3	2	7	4	58

直观上看，文物的不同类型与表面是否风化关联较强，为验证这一观点，检验两个类别变量是否具有显著相关性. 接下来使用卡方独立性检验进一步探究表面风化与这 3 种特征是否有显著影响.

首先，对表面风化与文物类型进行卡方独立性检验.

H_0：表面风化与文物类型独立 v. s. H_1：表面风化与文物类型不独立.

可算得卡方统计量 $\chi^2=5.4518$，$p=0.0196<0.05$，因此拒绝原假设，说明表面风化与文物类型相关.

接下来，对表面风化与文物纹饰进行卡方独立性检验.

H_0：表面风化与文物纹饰独立 v. s. H_1：表面风化与文物纹饰不独立.

可算得卡方统计量 $\chi^2=4.9565$，$p=0.0839>0.05$. 但由于计算的理论值里有小于 5 的数值出现，进一步用 Fisher 精确检验进行验证，$p=0.0836>0.05$，因此接受原假设，说明表面风化与文物纹饰独立.

其次，对表面风化与文物颜色进行卡方独立性检验.

H_0：表面风化与文物颜色独立 v. s. H_1：表面风化与文物颜色不独立.

可算得卡方统计量 $\chi^2=7.2338$，$p=0.4050>0.05$. 但由于计算的理论值里有小于 5 的数值出现，进一步用 Fisher 精确检验进行验证，$p=0.4713>0.05$，因此接受原假设，说明表面风化与文物颜色独立.

3.1.3　化学成分含量的统计变化规律

由于定和约束，成分数据各变量之间具有明显的共线性，使得常规的统计分析方法失效，因此，这里首先对成分数据进行中心对数比变换. 假设 n 维成分数据为 $\boldsymbol{x}=(x_1, x_2, \cdots, x_n)$，则 CLR 变换公式为：

$$\mathrm{CLR}(\boldsymbol{x})=\left(\ln\frac{x_1}{g(x)}, \ln\frac{x_2}{g(x)}, \cdots, \ln\frac{x_n}{g(x)}\right), \tag{1}$$

其中，$g(x)=\sqrt[n]{x_1x_2\cdots x_n}$. 由于式中自变量表示表面是否风化，因变量是 14 个化学成分，因此这里对于 CLR 变换后的数据使用多元方差分析(multivariate analysis of variance，MANOVA)来探究表面风化对哪些化学成分具有显著影响. 分高钾和铅钡两类分别进行讨论.

对于高钾类文物，CLR 变换后各化学成分有无风化的均值如图 3 所示.

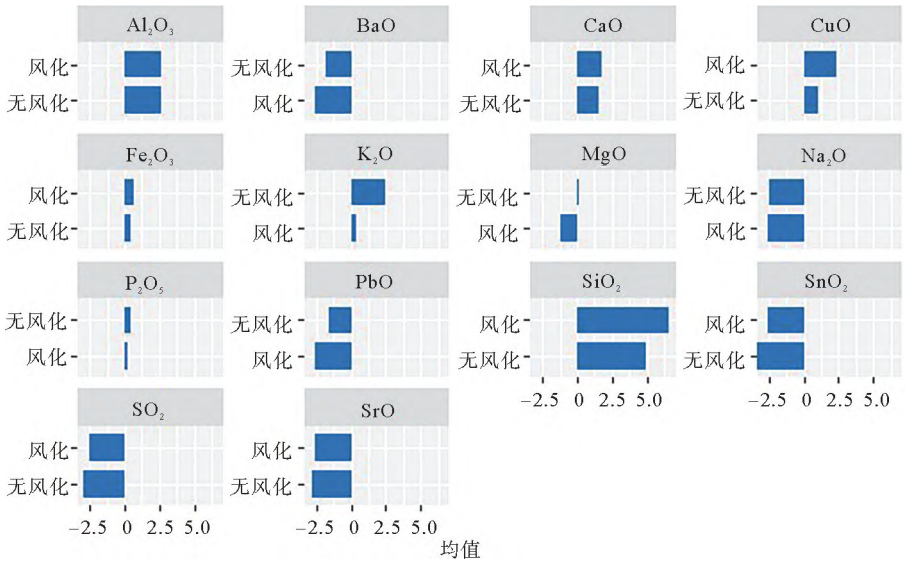


图 3　高钾类文物(CLR 变换后)有无风化时各化学成分的均值

由图 3 可以看出，SiO_2、CuO、K_2O 等化学成分均值差异较大，进一步进行多元方差分析.

H_0：高钾类文物 14 个化学成分在风化前后均值相等 v. s.

H_1：至少 1 个化学成分在风化前后均值存在显著差异.

各化学成分的 F 统计量以及对应的 P 值见表 4. 由结果可知，风化会引起高钾类样本 SiO_2 的显著增加，$p=1.0040\times10^{-5}<0.05$；并会一定程度上引起 CuO 的增加($p=0.0629$)，以及 K_2O 的降低($p=0.0627$)，其余化学成分的变化差异不大.

表 4　高钾类文物多元方差分析结果表

化学成分	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3
F 统计量	39.7760	0.0116	4.0028	0.0674	2.2053	0.0088	0.0711
p 值	1.0040×10^{-5}	0.9156	0.0627	0.7985	0.1570	0.9265	0.7931
化学成分	CuO	PbO	BaO	P_2O_5	SrO	SnO_2	SO_2
F 统计量	3.9956	1.2927	0.6120	0.0454	0.3730	1.1054	0.2620
p 值	0.0629	0.2723	0.4455	0.8340	0.5499	0.3087	0.6158

对于铅钡类文物，CLR 变换后各化学成分有无风化的均值见图 4.

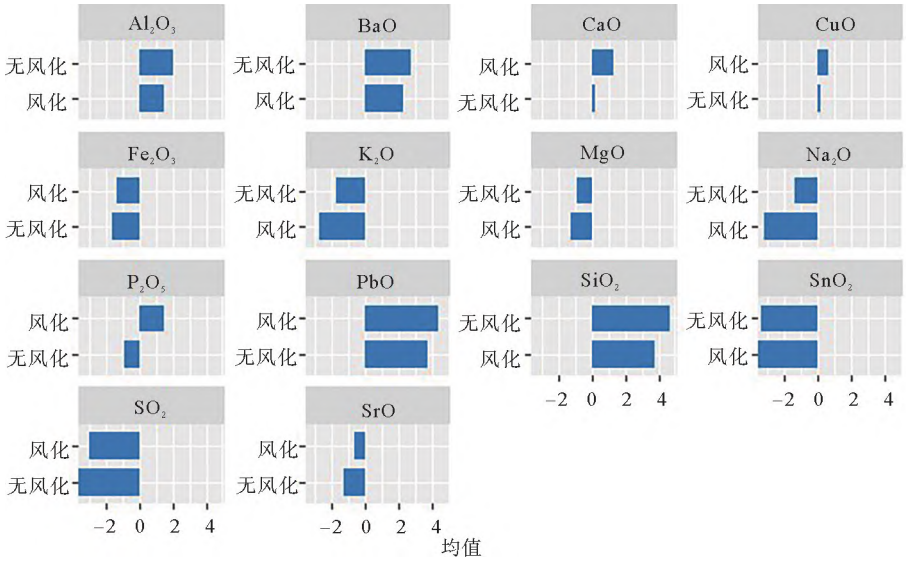


图 4　铅钡类文物(CLR 变换后)有无风化时各化学成分的均值

由图 4 可以看出，SiO_2、Na_2O、CaO、K_2O 和 P_2O_5 等化学成分均值差异较大. 进一步进行多元方差分析.

H_0：铅钡类文物 14 个化学成分在风化前后均值相等 v. s.

H_1：至少 1 个化学成分在风化前后均值存在显著差异.

各化学成分的 F 统计量以及对应 P 值见表 5. 由结果可知，风化会引起铅钡类样本 SiO_2 的显著降低($p=3.0850\times10^{-6}<0.05$)，$Na_2O$ 的显著降低($p=0.0097<0.05$)，K_2O 的显著降低($p=0.0363<0.05$)，CaO 的显著增加($p=0.0113<0.05$)，Al_2O_3 的显著降低($p=0.0026<0.05$)，PbO 的显著增加($p=0.0012<0.05$)，P_2O_5 的显著增加($p=0.0001<0.05$)，其余化学成分的变化差异不大.

表 5　铅钡类文物多元方差分析结果表

化学成分	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3
F 统计量	28.0340	7.2744	4.6444	6.9485	0.3484	10.1010	0.2149
p 值	3.0850×10^{-6}	0.0097	0.0363	0.0113	0.5578	0.0026	0.6451
化学成分	CuO	PbO	BaO	P_2O_5	SrO	SnO_2	SO_2
F 统计量	0.8920	11.9590	1.0489	17.2170	3.4536	0.3683	1.4260
p 值	0.3498	0.0012	0.3110	0.0001	0.0694	0.5468	0.2384

3.1.4　风化前化学成分含量预测

对于判断存在显著差异的化学成分，可以通过统计风化前后各化学成分改变的平均比例来预测已风化的样本在风化前的化学成分含量，最终使用 CLR 逆变换使得得到的预测值满足定和约束. 分高钾和铅钡两类进行讨论.

对于高钾类文物，认为风化前后含量发生变化的主要成分为 SiO_2、CuO 和 K_2O，这 3 种化学成分风化前后的均值和比值(CLR 变换后)见表 6. 据此，对于一个高钾类的风化点检测数据，假设其原始 SiO_2、CuO 和 K_2O 含量分别为 P_{SiO_2}、P_{CuO} 和 P_{K_2O}，CLR 变换后分别为 $CLR(P_{SiO_2})$、$CLR(P_{CuO})$ 和 $CLR(P_{K_2O})$，减去表 6 对应的差值后分别为 $CLR^*(P_{SiO_2})$、$CLR^*(P_{CuO})$ 和 $CLR^*(P_{K_2O})$，则相应的风化前化学成分含量预测公式为(以 SiO_2 为例)：

表 6　高钾类文物主要化学成分含量变化(CLR 变换后)表

化学成分	SiO_2	CuO	K_2O
无风化	4.9647	1.0730	2.4618
风化	6.5840	2.3369	0.3558
差值	1.6193	1.2639	−2.1060

$$\hat{P}_{SiO_2}=\frac{\exp\{CLR^*(P_{SiO_2})\}\times(P_{SiO_2}+P_{CuO}+P_{K_2O})}{\exp\{CLR^*(P_{SiO_2})\}+\exp\{CLR^*(P_{CuO})\}+\exp\{CLR^*(P_{K_2O})\}},\tag{2}$$

其余化学成分认为无明显变化. 由此得到高钾类的风化点检测数据风化前化学成分含量预测值，见表 7.

表 7　高钾类的风化点检测数据风化前化学成分含量预测表

文物采样点	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO	BaO	P_2O_5	SrO	SnO_2	SO_2
07	91.15	0.01	0.40	1.07	0.01	1.98	0.17	4.55	0.01	0.01	0.61	0.01	0.01	0.01
09	75.98	0.01	19.57	0.62	0.01	1.32	0.32	1.77	0.01	0.01	0.35	0.01	0.01	0.01
10	70.11	0.01	27.65	0.21	0.01	0.81	0.26	0.87	0.01	0.01	0.01	0.01	0.01	0.01
12	66.23	0.01	29.43	0.72	0.01	1.46	0.29	1.65	0.01	0.01	0.15	0.01	0.01	0.01
22	69.79	0.01	23.20	1.66	0.64	3.50	0.35	0.59	0.01	0.01	0.21	0.01	0.01	0.01
27	92.78	0.01	0.42	0.94	0.54	2.51	0.20	2.19	0.01	0.01	0.36	0.01	0.01	0.01

对于铅钡类文物，认为风化前后含量发生变化的主要成分为 SiO_2、Na_2O、K_2O、CaO、Al_2O_3、PbO 和 P_2O_5，这 7 种化学成分风化前后的均值和比值(CLR 变换后)见表 8. 类似于上述对高钾类文物的预测过程，可得到铅钡类的风化点检测数据风化前化学成分含量预测值，见表 9(限于篇幅，预测

无明显差异的化学成分未在表中列出).

表8　铅钡类文物主要化学成分含量变化(CLR变换后)表

化学成分	SiO_2	Na_2O	K_2O	CaO	Al_2O_3	PbO	P_2O_5
无风化	4.6951	−1.3600	−1.7645	0.2286	2.0101	3.7477	−0.9339
风化	3.7420	−3.2039	−2.7434	1.2711	1.4537	4.3748	1.4623
差值	−0.9531	−1.8438	−0.9789	1.0424	−0.5563	0.6272	2.3962

表9　铅钡类的风化点检测数据风化前化学成分含量预测表

文物采样点	SiO_2	Na_2O	K_2O	CaO	Al_2O_3	PbO	P_2O_5
02	68.04	0.05	2.02	0.60	7.23	18.32	0.24
08	40.80	0.05	0.02	0.41	1.83	11.96	0.26
08严重风化点	17.97	0.09	0.04	1.69	2.91	26.05	1.03
11	63.16	0.05	0.41	0.90	3.40	9.83	0.62
19	63.32	0.05	0.02	0.85	5.13	18.84	0.66
26	40.56	0.05	0.02	0.40	0.96	12.46	0.23
26严重风化点	14.05	0.09	1.55	1.55	3.00	23.27	0.80
34	66.79	0.05	0.48	0.20	2.03	17.89	0.02
36	63.25	8.65	0.23	0.08	1.72	13.70	0.00
38	60.41	6.17	0.02	0.17	3.17	18.63	0.03
39	60.76	0.06	0.02	0.35	0.78	29.09	0.09
40	48.47	0.07	0.03	0.74	0.88	41.93	0.18
41	49.73	0.07	1.22	1.82	6.03	24.47	0.71
43部位1	38.72	0.08	0.03	2.22	4.72	38.45	0.00
43部位2	57.85	0.06	0.03	2.32	6.11	24.57	1.20
48	68.92	2.52	0.42	0.50	11.86	4.18	0.05
49	62.55	0.05	0.02	1.35	7.86	15.29	0.85
50	50.47	0.07	0.03	1.22	3.53	25.43	0.62
51部位1	56.70	0.06	0.02	1.12	8.13	19.09	0.66
51部位2	59.83	0.07	0.03	1.95	4.73	29.63	0.86
52	57.44	6.63	0.02	0.69	1.74	21.79	0.45
54	53.52	0.06	0.79	1.04	6.70	27.43	0.36
54严重风化点	51.26	0.07	0.03	0.00	7.35	36.07	1.49
56	61.28	0.05	0.02	0.35	2.62	17.86	0.19
57	55.89	0.05	0.02	0.39	3.22	20.42	0.00
58	63.02	0.05	0.72	0.98	4.91	16.80	0.65

3.2　问题2

3.2.1　数据预处理

为避免表面风化这一因素的影响，首先通过表7和表9的预测结果将已风化采样点的化学成分含量修改为预测的未风化值，然后对数据进行CLR变换以消除定和约束.

3.2.2　分析高钾玻璃、铅钡玻璃的分类规律

高钾玻璃和铅钡玻璃各化学成分CLR变换后的均值见图5.

从图5可以看出，不同文物之间化学成分SnO_2和SO_2的差异不明显，因此去除了这两个相对不重要的成分，以保留更重要且有区分度的成分进行后续分类. 为更直观展示保留的重要成分是否能够明确划分不同类型的文物，这里使用UMAP将数据降到二维后进行可视化展示，见图6. 从图6可以看出，除个别异常点之外(文物编号21)，其余样本均能够根据保留的重要特征进行区分.

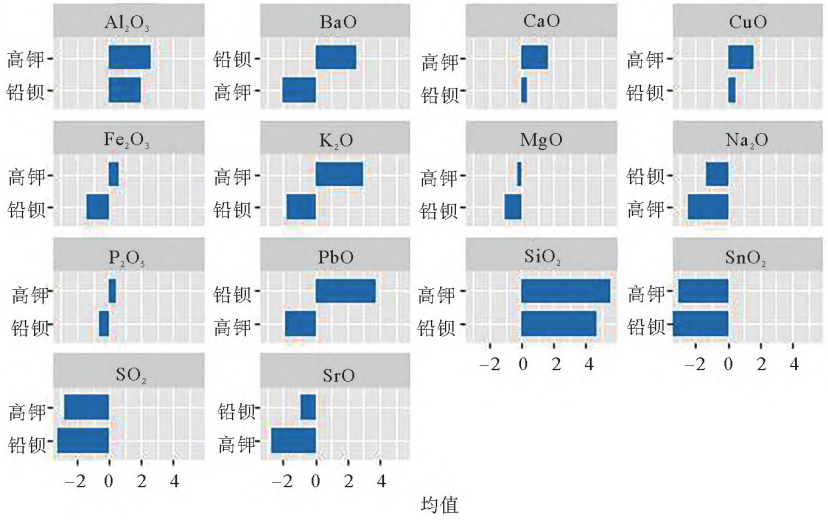


图 5　不同类型文物各化学成分均值(CLR 变换后)

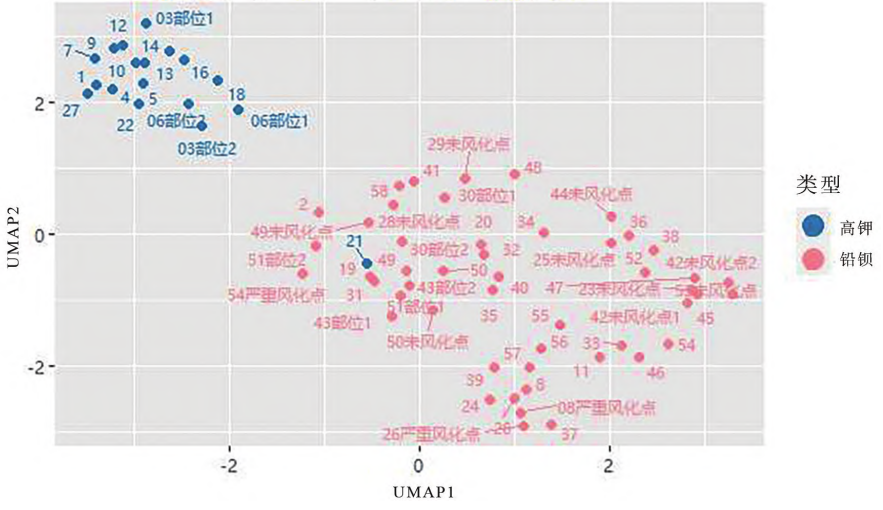


图 6　不同类型文物保留重要成分后数据可视化图(CLR 变换后)

为进一步探究文物的分类规律，对高钾、铅钡类型进行了关键成分识别. 使用随机森林模型中的两个重要指标“Mean Decrease Accuracy”和“Mean Decrease Gini”得分来衡量成分的重要性. 其中，“Mean Decrease Accuracy”表示随机森林预测准确性的降低程度，“Mean Decrease Gini”计算了每个变量对分类树每个节点上观测值的异质性的影响. 这两个指标的数值越大表示该成分对于划分文物类型的重要性越大，结果见图 7.

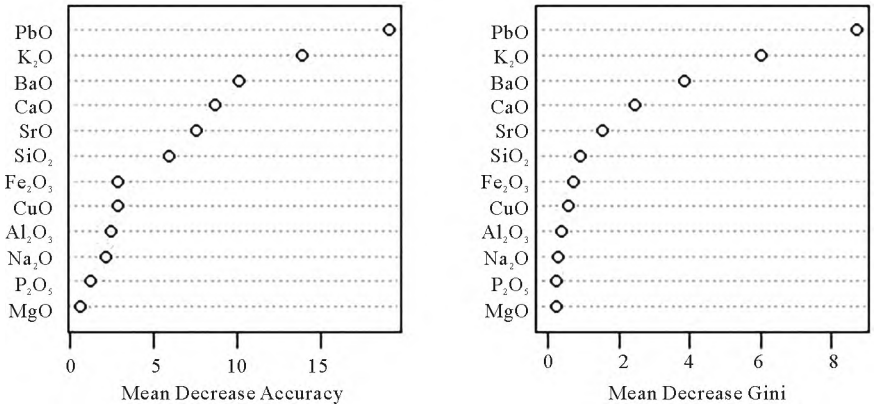


图 7　保留的 12 个化学成分重要性得分

从图 7 可以看出，重要性得分最高的 3 个化学成分为 PbO、K_2O 和 BaO. 据此，结合图 5，可得高钾和铅钡玻璃的分类规律为：相比于铅钡玻璃，高钾玻璃的 PbO、BaO 含量偏低，K_2O 含量偏高.

3.2.3 亚类划分

可以作为亚类划分依据的化学成分应当在同一类型的文物里存在较大的区分度，也即离散程度较高. 通过直方图统计各成分的分布情况，并给出了各成分的方差大小，从而筛选有区分度的成分特征用于亚类划分.

高钾类采样点各成分分布直方图及方差见图 8. 综合来看，K_2O、Na_2O 和 CaO 在高钾类样本间存在较高的区分度，可以作为划分亚类的依据. 分别考虑使用 1 种、2 种或全部 3 种上述化学成分进行亚类划分的结果，可以通过 K 均值聚类对高钾类文物进行亚类划分. 对于类别数的选择，主要依据以下两个原则：1)使簇内平方误差和尽量小；2)同一样本的不同采样点应被分到同一亚类. 若任何簇数选择都无法满足第 2 条要求，便舍去此亚类划分方案.

依此，可得不同划分依据的最佳簇数选择，见表 10. 最终亚类划分的可视化结果见图 9.

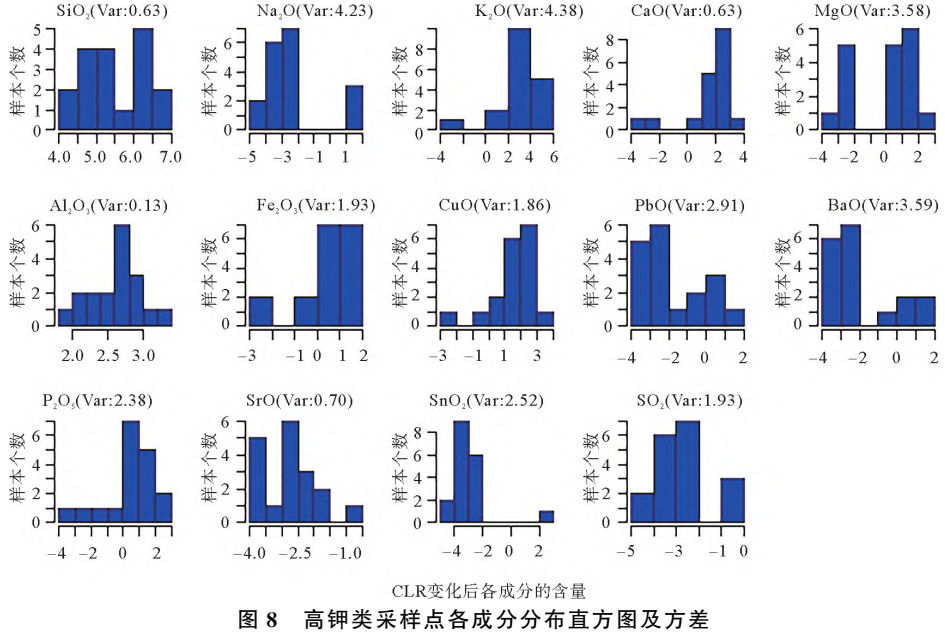


图 8 高钾类采样点各成分分布直方图及方差

表 10 高钾类文物亚类划分最佳簇数表

划分依据	K_2O	Na_2O	K_2O、Na_2O	Na_2O、CaO	K_2O、Na_2O、CaO
最佳簇数	3	2	3	2	2

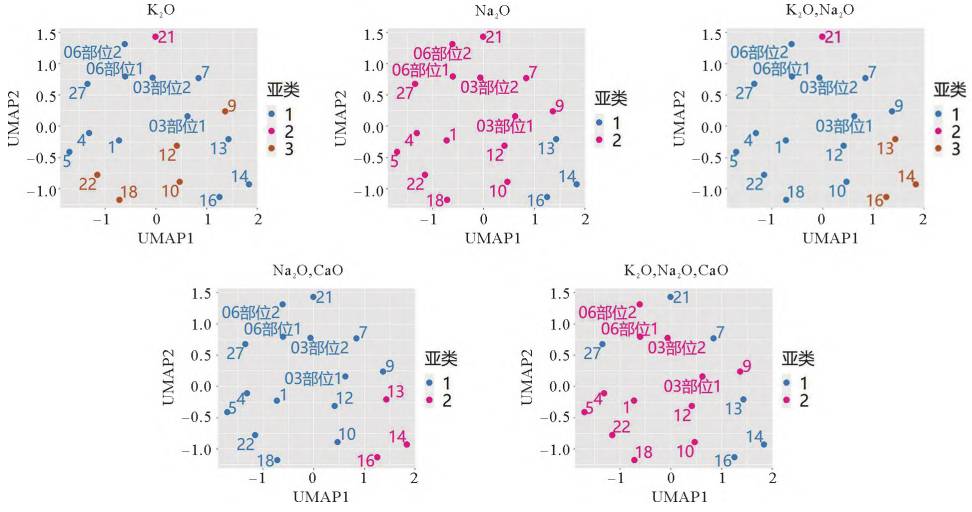


图 9 高钾类不同划分依据的亚类划分结果图

铅钡类采样点各成分分布直方图及方差见图 10. 综合来看，Na_2O、Fe_2O_3 和 MgO 在铅钡类样本间存在较高的区分度，可以作为划分亚类的依据. 不同的划分依据得到的最佳簇数选择见表 11，最终亚类划分的可视化结果见图 11.

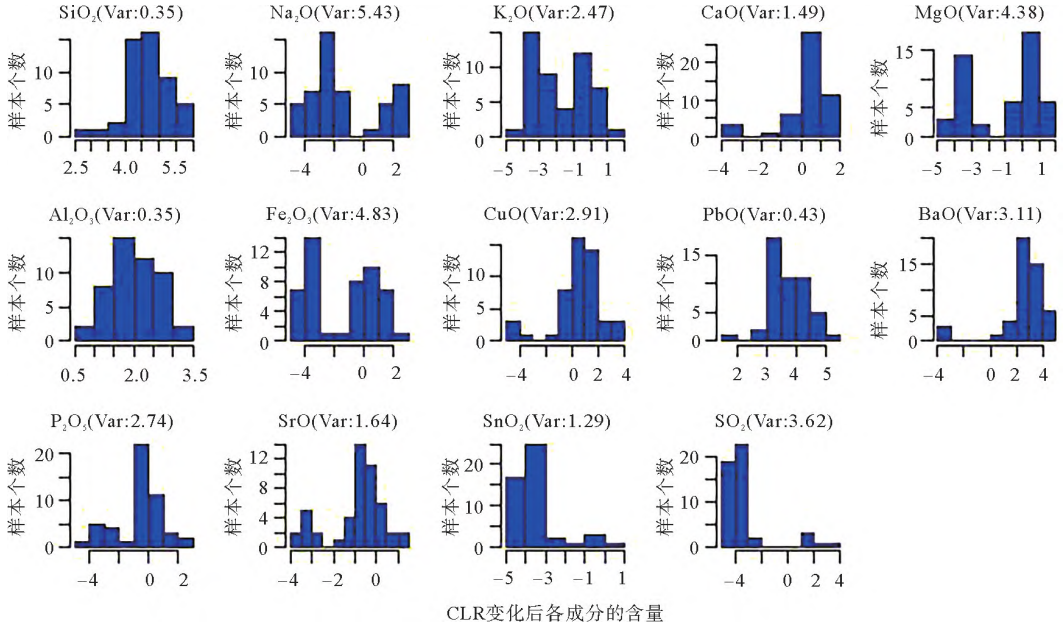


图 10　铅钡类采样点各成分分布直方图及方差

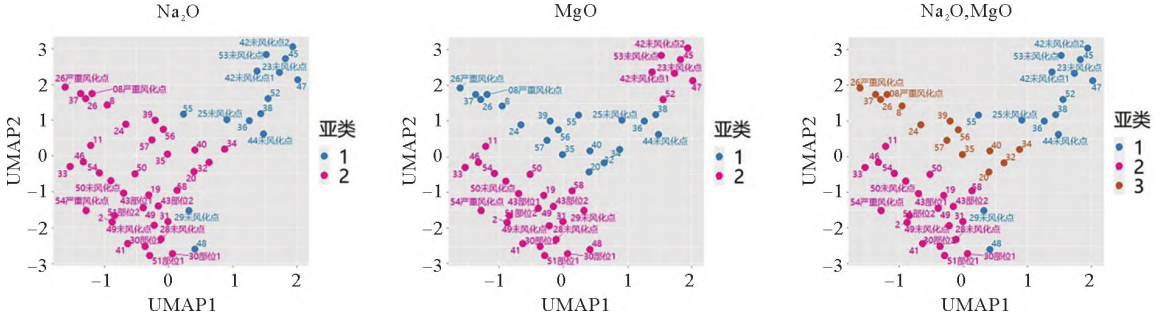


图 11　铅钡类不同划分依据的亚类划分结果图(彩图见封三)

1)合理性分析

对于高钾类样本不同特征分类依据给出的亚类划分结果，使用评价聚类的 3 个常用指标轮廓系数(silhouette coefficient，SI)、戴维森堡丁指数(Davies-bouldin index，DBI)以及卡林斯基-哈拉巴斯指数(Calinski-Harabasz index，CHI)来综合评价亚类划分结果. 这 3 个指标的定义如下.

表 11　铅钡类文物亚类划分最佳簇数表

划分依据	Na_2O	MgO	Na_2O、MgO
最佳簇数	2	2	3

· 轮廓系数　对于单个样本 i，设 x 是与它同类的其他样本的平均距离，y 是与它距离最近不同类别中样本的平均距离，则

$$\mathrm{SI}_i=(y-x)/\max\{x,\ y\},\tag{3}$$

而包含 N 个样本的总体轮廓系数为

$$\mathrm{SI}=\sum_{i=1}^{N}\mathrm{SI}_i/N.\tag{4}$$

轮廓系数越大，表示类的内部越紧凑，不同类的距离越大，聚类效果越好.

· 戴维森堡丁指数　该指标计算任意两类的类内距离平均之和除以这两类的中心距离，并求最大值. 假设一共有 K 个类别，则该指标具体定义如下：

$$\mathrm{DBI}=\frac{1}{K}\sum_{k=1}^{K}\max_{k\neq m,m\in[1,K]}\frac{\sigma_k+\sigma_m}{d(c_k,c_m)},\tag{5}$$

其中：σ_k 为第 k 类样本到其类中心的平均欧氏距离；$d(c_k,c_m)$为第 k 类和第 m 类的类中心之间的欧氏距离. DBI 值越小，说明分散程度越低，聚类效果越好.

·卡林斯基-哈拉巴斯指数 该指标通过评估类间方差和类内方差来计算聚类得分. 假设数据集有 N 个样本，共分为 K 个类别，则该指标具体定义为

$$\mathrm{CHI}=\frac{\mathrm{Tr}(\boldsymbol{B}_k)}{\mathrm{Tr}(\boldsymbol{W}_k)}\times\frac{N-k}{k-1},\tag{6}$$

其中：$\boldsymbol{B}_k$ 是组间离散矩阵，即不同簇之间的协方差矩阵；$\boldsymbol{W}_k$ 是簇内离散矩阵，即一个簇内数据的协方差矩阵；Tr 表示矩阵的迹. CHI 越高，表示组内离散程度越小，组间离散程度越大，聚类效果越好.

为更合理地结合这 3 个指标对聚类效果进行客观的评价，这里采用基于熵权法的 Topsis 模型来给各指标赋予不同的权重.

高钾类样本不同聚类方案基于熵权法的 Topsis 模型得到的最终得分见表 12. 从表 12 可以看出，对于高钾类样本，若选用单个特征，则当划分依据为 Na_2O 时，亚类划分得到的聚类效果最好. 该亚类划分方案将文物 13、14 及 16 划为一类，其他样本划为一类. 相比其他样本，13、14 及 16 的 Na_2O 含量明显较高. 若选用两个特征，则当划分依据为 K_2O 和 Na_2O 时，亚类划分得到的聚类效果最好. 该亚类划分方案将文物 21 作为单独一类，将文物 13、14 及 16 划为一类，其他样本划为一类. 从原始数据来看，相比其他样本，样本 13、14 及 16 的 Na_2O 含量明显较高，而样本 21 的 K_2O 含量明显较低. 这两个划分方案都是较为合理的.

表 12 高钾类文物基于 Topsis 熵权法得到的各聚类方案得分表

划分依据	K_2O	Na_2O	K_2O、Na_2O	Na_2O、CaO、K_2O	K_2O、Na_2O、CaO
最终得分	0.13	0.92	0.95	0.92	0.17
排名	4	2	1	2	3

铅钡类样本不同聚类方案基于熵权法的 Topsis 模型得到的最终得分见表 13. 从表 13 可以看出，使用 Na_2O 和 MgO 作为划分依据的优势比较明显. 从原始数据可以观察到，亚类 1 的 MgO 含量偏低，Na_2O 含量偏高；亚类 2 的 MgO 含量偏高，Na_2O 含量偏低；亚类 3 的 MgO 和 Na_2O 含量均偏低，具有较明显的区分度.

表 13 铅钡类文物基于 Topsis 熵权法得到的各聚类方案得分表

划分依据	Na_2O	MgO	Na_2O、MgO
最终得分	0.11	0.11	1.00
排名	2	2	1

2)敏感性分析

对于高钾类和铅钡类样本，分别给出了建议的划分依据和结果. 为验证划分结果的稳健性，分别给 CLR 变换后的数据添加均值为 0、方差不同的高斯噪声 ε_1、ε_2 和 ε_3，其中，$\varepsilon_1\sim\mathrm{Normal}(0,0.1^2)$，$\varepsilon_2\sim\mathrm{Normal}(0,0.5^2)$，$\varepsilon_3\sim\mathrm{Normal}(0,1^2)$，再次进行 K 均值聚类，观察划分结果是否发生变化.

对于高钾类样本，若以 Na_2O 为划分依据，加入前两种高斯噪声 ε_1 和 ε_2 均不会使亚类划分结果发生变化，加入高斯噪声 ε_3 会使得文物 9 和 18 划分到不同亚类，其余样本均不变，总体来说划分结果对噪声并不敏感，稳健性较强；若以 K_2O 和 Na_2O 为划分依据，加入高斯噪声 ε_1 不会使亚类划分结果发生变化，加入高斯噪声 ε_2 会使得文物 3、6、7 和 27 划分到不同亚类，而加入高斯噪声 ε_3 会使得文物 1、3、6、7 和 27 划分到不同亚类，其余样本均不变，总体来说相比只用 Na_2O 作为划分依据，划分结果对噪声更为敏感.

对于以 Na_2O 和 MgO 为划分依据的铅钡类样本，加入高斯噪声 ε_1 不会使亚类划分结果发生变化，加入高斯噪声 ε_2 会使得文物 29 划分到不同亚类，其余样本均不变，加入高斯噪声 ε_3 会使得文物 19、25、29 和 55 划分到不同亚类，其余样本均不变，总体来说划分结果对噪声并不敏感，稳健性较强.

3.3 问题 3

采用与问题 1 相同的预处理方式对未知文物数据的缺失值进行填充并归一化，且进行 CLR 变换以消除定和约束. 为避免表面风化这一因素的影响，将未知类别的文物分为风化和未风化两类分别进行讨论.

对于无风化的未知文物，CLR 变换后的结果见表 14. 由问题 2 得出的分类规律可以初步判断出，A1 更符合高钾的分类特征，A3、A4 和 A8 更符合铅钡的分类特征. 为进一步验证，以问题 2 中有标签的预处理后的无风化样本为训练集训练随机森林模型，并以这 4 个无风化的未知文物为测试集，预测结果与之前判断一致.

表 14　无风化的未知文物 CLR 变换后结果表

文物编号	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO	BaO	P_2O_5	SrO	SnO_2	SO_2
A1	5.57	−3.40	−3.40	3.01	1.83	3.18	1.97	1.95	−3.40	−3.40	1.26	−2.30	−3.40	0.53
A3	3.55	−4.52	0.40	2.06	−0.12	1.16	2.04	−1.47	3.77	1.64	1.08	−0.56	−4.52	−4.52
A4	3.53	−4.65	−0.28	1.02	0.01	1.91	1.82	−0.08	3.15	2.08	2.09	−1.31	−4.65	−4.65
A8	2.85	0.00	−2.55	−1.20	0.00	−0.33	0.00	1.12	1.98	1.35	−0.70	−2.25	0.00	−0.27

对于风化的未知文物，CLR 变换后的结果见表 15. 由问题 2 得出的分类规律可以初步判断出，A6 和 A7 更符合高钾的分类特征，A2 和 A5 更符合铅钡的分类特征. 类似地，以问题 2 中有标签的预处理后的风化样本为训练集训练随机森林模型，并以 4 个风化的未知文物为测试集，预测结果与之前判断一致，最终对未知文物类别的预测结果见表 16.

表 15　风化的未知文物 CLR 变换后结果

文物编号	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO	BaO	P_2O_5	SrO	SnO_2	SO_2
A2	5.68	−2.55	−2.55	4.09	−2.55	2.90	−2.55	−2.55	5.59	−2.55	4.71	−2.55	−2.55	−2.55
A5	4.04	0.06	−1.12	0.37	0.73	2.42	−0.33	−0.18	2.38	0.65	−1.78	−1.68	−0.84	−4.73
A6	6.44	−2.70	2.21	1.46	0.35	2.33	0.60	2.46	−2.70	−2.70	0.35	−2.70	−2.70	−2.70
A7	6.43	−2.68	1.90	2.04	−2.68	3.54	0.50	2.08	−2.68	−2.68	−0.12	−2.68	−2.68	−0.28

表 16　未知文物的类别预测结果

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

进一步，从两个方面来验证预测结果是否稳健. 首先，考虑将风化的文物数据 A6 和 A7 按问题 2 中高钾类文物的预测公式还原为未风化数据，将风化的文物数据 A2、A5 按问题 2 中铅钡类文物的预测公式还原为未风化数据，见表 17. 再用已训练好的无风化样本为训练集的随机森林模型，对这 4 个样本进行预测，分类结果保持不变，说明了预测结果是合理的.

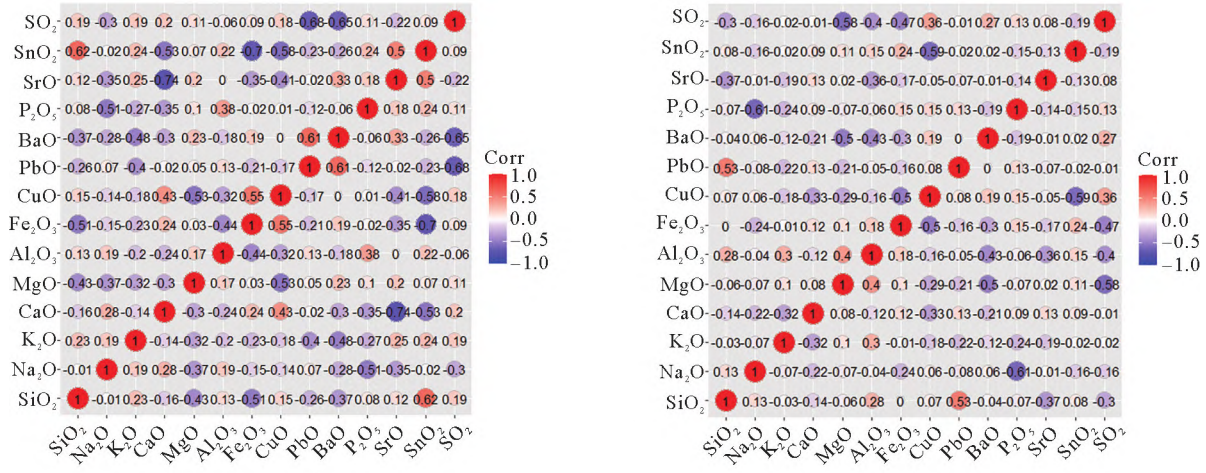
另一方面，同样给 CLR 变换后的未知数据添加和问题 2 分布相同的 3 种高斯噪声，发现分类结果均未发生任何变化，说明了模型预测的稳健性.

表 17　风化文物的风化前化学成分含量预测

文物编号	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO	BaO	P_2O_5	SrO	SnO_2	SO_2
A2	78.67	0.05	0.02	2.16	0.01	3.27	0.01	0.01	14.72	0.01	1.04	0.01	0.01	0.01
A5	75.78	3.45	0.45	0.26	2.35	10.11	0.81	0.94	2.97	2.17	0.01	0.21	0.49	0.01
A6	59.64	0.01	35.85	0.65	0.21	1.53	0.27	1.58	0.01	0.01	0.21	0.01	0.01	0.01
A7	63.62	0.01	28.47	1.12	0.01	5.08	0.24	1.17	0.01	0.01	0.13	0.01	0.01	0.11

3.4 问题 4

首先，将表单 2 中已预处理并进行 CLR 变换后的数据(且风化数据已调整为未风化数据)分为高钾和铅钡两类，分别计算各化学成分之间的相关系数并进行相关性检验，相关系数可视化结果见图 12.



(a) 高钾类文物　　(b) 铅钡类文物

图 12　高钾类文物与铅钡类文物各化学成分之间的相关系数图(彩图见封三)

由图 12 可以看出，若考虑线性关联，对于高钾类文物，SiO_2 与 SnO_2 之间存在显著正相关(相关性检验 p 值为 0.01)，与 Fe_2O_3 之间存在显著负相关(相关性检验 p 值为 0.03)；Na_2O 与 P_2O_5 之间存在显著负相关(相关性检验 p 值为 0.03)；K_2O 与 BaO 之间存在显著负相关(相关性检验 p 值为 0.04)；CaO 与 SrO 之间存在显著负相关(相关性检验 p 值小于 0.01)，且与 SnO_2 之间存在显著负相关(相关性检验 p 值为 0.02)；MgO 与 CuO 之间存在显著负相关(相关性检验 p 值为 0.02)；Fe_2O_3 与 CuO 存在显著正相关(相关性检验 p 值为 0.02)，与 SnO_2 存在显著负相关(相关性检验 p 值小于 0.01)；CuO 与 SnO_2 存在显著负相关(相关性检验 p 值为 0.01)；PbO 与 BaO 存在显著正相关(相关性检验 p 值为 0.01)；与 SO_2 存在显著负相关(相关性检验 p 值小于 0.01)；BaO 与 SO_2 存在显著负相关(相关性检验 p 值小于 0.01)；SrO 与 SnO_2 存在显著正相关(相关性检验 p 值为 0.03).

对于铅钡类文物，SiO_2 与 PbO 之间存在显著正相关(相关性检验 p 值小于 0.01)，与 SrO 之间存在显著负相关(相关性检验 p 值为 0.01)，与 SO_2 之间存在显著负相关(相关性检验 p 值为 0.03)；Na_2O 与 P_2O_5 存在显著负相关(相关性检验 p 值小于 0.01)；K_2O 和 Al_2O_3 存在显著正相关(相关性检验 p 值为 0.04)，和 CaO 存在显著负相关(相关性检验 p 值为 0.02)；CaO 与 CuO 存在显著负相关(相关性检验 p 值为 0.02)；MgO 与 Al_2O_3 存在显著正相关(相关性检验 p 值小于 0.01)，与 CuO 存在显著负相关(相关性检验 p 值为 0.04)，与 BaO 存在显著负相关(相关性检验 p 值小于 0.01)，与 SO_2 存在显著负相关(相关性检验 p 值小于 0.01)；Al_2O_3 与 BaO 存在显著负相关(相关性检验 p 值小于 0.01)，与 SrO 存在显著负相关(相关性检验 p 值为 0.01)，与 SO_2 存在显著负相关(相关性检验 p 值小于 0.01)；Fe_2O_3 与 CuO 存在显著负相关(相关性检验 p 值小于 0.01)，与 BaO 存在显著负相关(相关性检验 p 值为 0.03)，与 SO_2 存在显著负相关(相关性检验 p 值小于 0.01)；CuO 与 SO_2 存在显著正相关(相关性检验 p 值为 0.01)，与 SnO_2 显著负相关(相关性检验 p 值小于 0.01).

对于不同类别之间的化学成分关联关系的差异，发现高钾类样本 Fe_2O_3 与 CuO 是显著线性正相关的，而铅钡玻璃则相反.

4　评阅综述

4.1　评阅总体情况

总体来说，本题的答题情况不是太好，绝大部分论文都存在缺陷，在后面提到的几个典型问题上

存在或多或少的不足，甚至是逻辑上的错误. 幸运的是，最后遴选出来 3 篇优秀论文 C155[3]、C065[4] 和 C229[5]. 这 3 篇论文在几个典型问题上都有不错的表现，在“大学生在线”上公示过.

论文 C155 考虑到了数据的成分性，对数据进行 CLR 变换，对变换后的数据进行统计分析. 在分析风化程度与其他因素关系中，注意到了卡方检验的适用范围和 Yate 校正；在风化前预测中，他们利用风化程度引入时间概念(未风化、轻度风化、中度风化、重度风化)、聚类，对聚类中心点 CLR 变换后的数据建立趋势模型 $f(t)$，通过平移量预测分化前，CLR 逆变换保证了预测后的值满足定和限制，参见文献[3]的图 4. 在亚类聚类时，他们采用了所谓的 R 型聚类，即对 14 种化学成分进行聚类，每个类中选取代表变量，然后进行亚分类.

论文 C065 考虑到了数据的成分性，对数据进行 CLR 变换，在变换后的数据上进行统计分析；在分析风化程度与其他因素关系中，在不满足卡方检验的适用范围时采用 Fisher 检验代替；采用均值差进行分布匹配预测风化前，暗含了等方差的假设：

$$\xi_{\text{erode}}+\Delta\xi_{\text{mean}}=\xi_{\text{predict}}.$$

他们采用 SVM 分类得到比较简单的大类区分规律；采用层次聚类法分出亚类，然后用 SVM 再给出简单的亚类区分规律，并用 SVM 参数变化量度量样本对分类的影响.

论文 C229 考虑到了数据的成分性，对数据进行 CLR 变换，在变换后的数据上进行统计分析；他们采用狄氏(Dirichilet)回归预测风化前化学成分，用风化状态作为哑变量进行回归，本质上等价于等方差下的分布匹配；在分类中，他们采用偏最小二乘判别分析(PLS-DA)进行大类分类，计算变量的投影重要性(VIP)，筛选出对分类有影响的特征——PbO、K_2O、BaO 和 SrO；在亚类划分中，他们在聚类基础上再采用 PLS-DA 进行分类，筛选重要变量，发现对高钾玻璃亚类分类有影响的特征为 K_2O、SiO_2、CaO、SnO_2 和 BaO，而对铅钡玻璃亚类分类有影响的特征为 P_2O_5、CuO、Na_2O 和 Fe_2O_3.

4.2 几个典型问题

• 成分数据处理问题. 本题的最大特点就是成分数据. 尽管由于测量误差，每个样本的所有化学成分相加不是刚好为 100%，但根据题目中所强调的，各化学成分之和应该处理成 100%；成分数据是在单纯形空间上取值，一般采用中心对数比变换(CLR)转换到低一维的欧氏空间；由于要进行对数变换，没有观察数据的化学成分需要用一个小量(ε)加以填充，然后归一化成定和.

• 风化前化学成分预测问题. 数据中只有风化后的数据，缺少风化前的数据与之配对，没有配对的学习样本来建立普通的预测模型，只能通过分布匹配，在比较强的假设下给出预测. 有的同学将风化前后作为哑变量进行回归分析，本质上等价于等方差假设下的均值匹配方法；有的同学在上述模型的基础上加入其他协变量(颜色，纹理等)建立多元回归模型，也是可以的，相当于利用其他协变量进行细分，本质上还是等价于均值匹配方法. 但是在评阅中发现了大量论文简单套用机器学习方法进行预测，是完全错误的.

• 亚类区分中的变量选择问题. 选择变量后会影响样本之间距离的计算，从而会影响聚类结果. 这是一个典型的无监督变量选择问题，相比于有监督的变量选择问题，相关的文献比较少. 有同学先作主成分分析，然后用几个主成分作为变量进行聚类，这个不是真正的变量选择. 但根据主成分中变量的载荷系数选择主成分中载荷系数绝对值大的变量，有一定的道理.

• 成分数据相关分析问题. 成分数据上皮尔逊相关系数有偏负的倾向，不能准确反映化学成分绝对量之间的相关关系. 举一个极端的例子. 如果只有两个成分，绝对量为 y_1 和 y_2，归一化后得到成分 x_1 和 x_2，$x_1+x_2=1$，那么 $\text{PCC}(x_1, x_2)=-1$. 显然这个负相关与绝对量 y_1、y_2 之间的相关性相差甚远. 实际上，可以借助于中心对数比变换(CLR)对绝对量相关关系加以估计. 本人课题组对成分数据相关矩阵估计、精度矩阵估计(网络推断)和差异矩阵估计有一定的研究，提出了几个算法(CCLasso[6]，gCoda[7]，CDTrace[8]，CDTr[9]，CodaLoss[10])，有兴趣的老师和同学请参考.

参考文献

[1]全国大学生数学建模组委会. 2022“高教社杯”全国大学生数学建模竞赛赛题[EB/OL]. [2022-09-15]. http://

www.mcm.edu.cn/html_cn/node/5267fe3e6a512bec793d71f2b2061497.html.
[2]Mert M C, Filzmoser P, Hron K. Error propagation in isometric log-ratio coordinates for compositional data: theoretical and practical considerations[J]. Mathematical Geosciences, 2016, 48(8): 941-961.
[3]黄慧婷，李春明，刘思语，等. 基于成分数据的古代玻璃制品的成分分析与鉴别[J]. 数学建模及其应用，2023，12(2)：52-62+124.
[4]邓天宇，邱奕珲，池正昊. 基于CLR的玻璃文物成分分析与分类模型[J]. 数学建模及其应用，2023，12(2)：41-51.
[5]马佩莹，韩雁来，李德兰，等. 基于成分数据的古代玻璃制品分析与分类[J]. 数学建模及其应用，2023，12(2)：63-73.
[6]Fang H, Huang C, Zhao H, et al. CCLasso: correlation inference for compositional data through Lasso[J]. Bioinformatics, 2015, 31(19): 3172-3180.
[7]Fang H, Huang C, Zhao H, et al. gCoda: conditional dependence network inference for compositional data[J]. Journal of Computation Biology, 2017, 24(7): 699-708.
[8]Yuan H, He S, Deng M. Compositional data network analysis via lasso penalized D-trace loss[J]. Bioinformatics, 2019, 35(18): 3404-3411.
[9]He S, Deng M. Direct interaction network and differential network inference from compositional data via lasso penalized D-trace loss[J]. PLoS One, 2019, 14: e0207731.
[10]Chen L, He S, Zhai Y, et al. Direct interaction network inference for compositional data via codaloss[J]. Journal of Bioinformation and Computational Biology, 2020, 18(6): 2050037.

Statistical Modeling and Analysis of Ancient Glass Composition

WAN Hui, DENG Minghua
(School of Mathematical Sciences, Peking University, Beijing 100871)

Abstract: This paper presents a feasible solution to problem C "Ancient glass composition analysis and discrimination" in 2022 Higher Education Press Cup China Undergraduate Mathematical Contest (CUMCM 2022), and gives a brief summary of review process. The aim of this problem is analyzing the chemical composition and exploring classification method for ancient glasses. The most important characteristics of this problem is its compositional data. Considering the compositional characteristic of the data, we use the central log-ratio transformation, which transform the compositional data in simplex space to Euclidean space, then the corresponding statistical analysis can be done in Euclidean space.
Key words: compositional data; central log-ratio transformation; supervised classification; unsupervised clustering; correlation analysis

作者简介

宛　惠(1997—)，女，博士研究生在读，研究方向为生物信息学.
邓明华(1969—)，男，博士，教授，主要研究方向为计算生物学与数学建模.