教学与竞赛

基于成分数据的古代玻璃制品分析与分类

马佩莹，韩雁来，李德兰，陈佳佳
（山西财经大学 统计学院，山西 太原 030006）

摘 要：古代玻璃制品的化学成分属于成分数据，基于成分数据分析方法可以对玻璃制品的化学成分进行分析，研究其分类规律，并对未知玻璃文物鉴别其所属类型. 首先，基于 Spearman 相关系数以及卡方检验分析玻璃文物表面风化与其类型、纹饰、颜色的关系；通过单形空间均值来分析玻璃表面有无风化化学成分含量的统计规律；构建 Dirichlet 回归模型来预测风化点风化前的化学成分含量. 其次，构建决策树、偏最小二乘判别分析两种模型对两类玻璃进行初分类特征选择；进一步，用 K-means 聚类对两类玻璃进行亚分类，并通过偏最小二乘判别分析对两类玻璃进行亚分类特征选择；进而，利用所得分类规律对未知类别玻璃鉴别其所属类型. 最后，运用灰色关联分析分别探究两类玻璃化学成分之间的关联关系及其差异性.

关键词：成分数据；决策树；偏最小二乘判别分析；K-means 聚类；灰色关联

中图分类号：O29　　**文献标志码**：A　　**文章编号**：2095-3070(2023)02-0063-11

DOI：10.19943/j.2095-3070.jmmia.2023.02.06

0 引言

玻璃的发展历史悠久，是人类最早发明的人造材料之一，在中国经历了从舶来品到自主生产的过程. 玻璃外观虽相似，但根据其添加的助熔剂不同，其主要化学成分也不同[1-2]. 根据一批我国古代玻璃制品的相关数据研究玻璃制品的化学成分及鉴别，对古代玻璃文物的保护具有重要意义[3].

有学者根据中国古代玻璃的发展，将玻璃成分的演变分为 5 个阶段：从春秋到战国前期的 K_2O-CaO-SiO_2 系统，其中 $K_2O/Na_2O>1$；从战国到东汉时期的 BaO-PbO-SiO_2 系统和 K_2O-SiO_2 系统；从东汉到唐代时期的 PbO-SiO_2 系统；从唐代到元代时期的 K_2O-PbO-SiO_2 系统；从元代到清代时期的 K_2O-CaO-SiO_2 系统[4]. 主成分分析方法已经被大量应用于古代文物的具体成分研究中，特别是古陶瓷、古玻璃器化学成分的分析处理[5]. 此外，一些学者采用外束质子激发 X 荧光技术(PIXE)、电感耦合等离子体原子发射光谱分析(ICP AES)方法，对新疆、湖北等地区出土的一批战国时期的玻璃珠(包含镶嵌玻璃珠)、玻璃璧样品进行检测[6]. 史美光等[7]应用扫描电镜-能量色散 X 射线分析 (SEM-EDX)、电感耦合等离子体发射光谱分析方法、密度测定和偏光显微镜观察等方法，对十多件有代表性的钾硅玻璃样品进行了研究. 有学者采用光谱分析、同位素射线荧光分析和比重测试 3 种方法对西汉到北宋的 52 件玻璃样品进行了分析[8].

对玻璃化学成分的研究大多基于物理、化学方法对其成分进行判别并研究其演化历程，仅有少数

收稿日期：2023-01-19
基金项目：山西省高等学校教学改革创新项目(J20220570)
通讯作者：陈佳佳，E-mail：chenjiajia0401@163.com
引用格式：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.
MA P Y，HAN Y L，LI D L，et al. Analysis and classification of ancient glass products based on compositional data(in Chinese)[J]. Mathematical Modeling and Its Applications，2023，12(2)：63-73.

研究采用统计方法. 本文从统计视角出发，基于成分数据[9-10]，运用统计方法建立成分预测、玻璃分类及进一步亚分类的模型，进而将这些模型推广，运用至新发现玻璃的成分研究与分类，有助于促进玻璃文物的保护和玻璃制品的生产制造.

1　模型假设

1)化学成分指采样点处的化学成分；

2)未检测到化学成分的原因是仪器精度受限，故暂时将化学成分缺失值记为0，后续进行插补处理；

3)检测到的化学成分0值是由于四舍五入得到的，后续进行插补处理.

2　数据预处理

2.1　附表1数据预处理

附表1给出了玻璃文物编号、纹饰、类型、颜色和表面风化的基本信息. 由于数据量较大且存在一定的缺失值，故对数据进行缺失值处理，防止对后续建模产生不利影响. 其中缺失值均为玻璃文物颜色，按其他信息可分为“纹饰A、表面风化、铅钡”和“纹饰C、表面风化、铅钡”两类，按此分别进行填补.

在所给文物中，符合“纹饰C、表面风化、铅钡”条件并已知颜色的15件文物颜色分布如表1所示.

表1　颜色分布Ⅰ

颜色	蓝绿	浅蓝	浅绿	深绿	紫
对应文物编号	56，57	11，25，43，51，52，54	41	34，36，38，39	08，26

因数据为定性数据，故采用热卡来填充缺失值. 对15件文物及待填充的40、58号文物的化学成分进行比较，用成分相似的文物颜色进行填充. 方法是：一件文物在两个部位进行采样时取均值代表，在一个部位和严重分化点采样时按1∶2的权重计算后作为代表，在一个部位和未分化点采样时将部位的化学成分作为代表. 经比较可知，40号与39号文物成分相似，58号与51号成分相似，故将40和58号文物颜色分别填充为深绿和浅蓝.

符合“纹饰A、表面风化、铅钡”条件并已知颜色的9件文物颜色如表2所示.

表2　颜色分布Ⅱ

颜色	黑	蓝绿	浅蓝
对应文物编号	49，50	23	02，28，29，42，44，53

由附表2可知，9件文物中有6件文物采样点为未风化点，所得到的成分含量可参考性较低，故对此采用众数进行填补，即19和48号文物颜色填充为浅蓝.

2.2　附表2、3数据预处理

累加和介于85%～105%之间的数据视为有效数据，故15号和17号文物为无效数据，将其剔除. 假设表单2、3中出现的空白为仪器精度受限未检测到该成分，可看作近似零值；另外，假设表单2、3中的0值是由于四舍五入得到，也可看作近似零值. 采用乘法替换对近似零值进行插补，具体过程为如下.

考虑成分数据集

$$\boldsymbol{X}=[x_{ij}]_{n\times D}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nD} \end{pmatrix}, \tag{1}$$

其中：$\boldsymbol{X}$ 中每一行为一个成分数据，每个成分数据含有 D 个部分，如第 i 行表示第 i 个玻璃文物的化学成分数据；$x_{i,j}$ 表示第 i 个玻璃文物的第 j 个化学成分的含量. 假定成分数据中有近似零值，且不同成分数据相同部分对应的探测范围是相同的. 记探测范围向量为 $\boldsymbol{e}=(e_1, e_2, \cdots, e_D)^{\mathrm{T}}$，其中，$e_j$ 为成分数据集 $\boldsymbol{X}$ 的第 j 个部分对应的探测范围. 运用乘法简单替换法，对 x_{ij} 替换后的数据为

$$\hat{x}_{ij}=\begin{cases}\delta_{ij}, & x_{ij}=0,\\ x_{ij}\left[1-\left(\sum\limits_{k|x_{ik}=0}\delta_{ik}\right)/c\right], & x_{ij}>0,\end{cases}\tag{2}$$

其中：δ_{ij} 为一个小于 e_j 的数；c 为成分数据的常数和约束，即 $\sum\limits_{j=1}^{D}x_{ij}=c$.

通过实验发现，当成分数据集中近似零值比例不高，δ_{ij} 等于探测范围的 65%时插补效果最好，即 $\delta_{ij}=0.65e_j$. 记附表 2、3 中空白处所在列的最小值为临界值，并将其乘以 0.65，得到插补值，如表 3 和表 4 所示(仅列出部分插补值).

表 3　表单 2 成分数据插补缺失值表(部分)　　%

文物采样点	SiO_2	Na_2O	K_2O	SnO_2	SO_2
01	69.3300	0.5275	9.9900	0.1517	0.3900
03 部位 1	87.0500	0.5287	5.1900	0.1520	0.0727
03 部位 2	61.7100	0.5239	12.3700	0.1506	0.0720
57	25.4200	0.5264	0.0724	0.1513	0.0724
58	30.3900	0.5239	0.3400	0.1506	0.0720

表 4　表单 3 成分数据插补缺失值表(部分)　　%

文物编号	表面风化	SiO_2	Na_2O	K_2O	SnO_2	SO_2
A1	无风化	78.4500	0.5277	0.0726	0.1517	0.5100
A2	风化	37.7500	0.5298	0.0728	0.1523	0.0728
A3	无风化	31.9500	0.5239	1.3600	0.1506	0.0720
A4	无风化	35.4700	0.5240	0.7900	0.1507	0.0721
A5	风化	64.2900	1.2000	0.3700	0.4900	0.0716
A6	风化	93.1700	0.5278	1.3500	0.1517	0.0726
A7	风化	90.8300	0.5281	0.9800	0.1518	0.1100
A8	无风化	51.1200	0.5248	0.2300	0.1509	2.2600

因成分数据相加需为 100，故利用相对信息对数据进一步处理. 其中相对信息是指成分数据仅有的信息反映在成分间的比率中，与每个成分的绝对数据是无关的. 因为成分数据的每个成分乘以相同的正常数，成分间的比率是不变的，因此成分数据可以看作等价类，且类中数据含有相同的信息，可以通过适合的尺度因子表示为相同比例向量. 闭合运算是对初始向量乘以合适的尺度因子，使得闭合后的成分和为常数 k(这里为 100)，定义为

$$C(\boldsymbol{x})=C(x_1,x_2,\cdots,x_D)^{\mathrm{T}}=\left(k\cdot x_1/\sum_{i=1}^{D}x_i,\ k\cdot x_2/\sum_{i=1}^{D}x_i,\cdots,k\cdot x_D/\sum_{i=1}^{D}x_i\right)^{\mathrm{T}}.\tag{3}$$

对于任意两个向量 $\boldsymbol{x}$，$\boldsymbol{y}\in\mathbf{R}_+^D$，如果 $C(\boldsymbol{x})=C(\boldsymbol{y})$，则 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 是成分等价的. 故将每个玻璃文物化学成分乘以计算出的相应因子得到最终数据.

3　模型建立与求解

3.1　问题 1

3.1.1　玻璃表面风化与纹饰、玻璃类型、颜色的相关性分析

首先进行 Spearman 相关系数分析. 在分析之前对定性数据进行虚拟变量处理，见表 5. 利用 SPSS 软件分别对纹饰、玻璃类型、颜色与表面风化进行 Spearman 相关性检验，结果见表 6. 表 6 结果表明，纹饰、颜色与表面是否风化无相关关系，而玻璃类型与其存在相关关系.

表 5　虚拟变量处理表

变量	处理结果	变量	处理结果
纹饰 A	1	浅蓝色	1
纹饰 B	2	深蓝色	2
纹饰 C	3	蓝绿色	3
铅钡	1	浅绿色	4
高钾	2	深绿色	5
无风化	0	绿色	6
风化	1	紫色	7
黑色	8		

表6　其他变量与表面风化相关性表

	纹饰	表面风化		玻璃类型	表面风化		颜色	表面风化
纹饰	1.000	0.080	玻璃类型	1.000	−0.301	颜色	1.000	0.088
表面风化	0.080	1.000	表面风化	−0.301	1.000	表面风化	0.088	1.000

其次，为保证结果的可信度，利用卡方检验对数据进行分析，在分析之前对数据进行频数统计，见图1—图3.

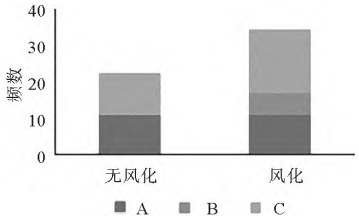


图1　纹饰与表面风化的频数统计

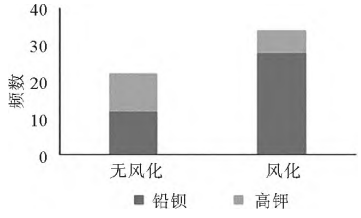


图2　玻璃类型与表面风化的频数统计

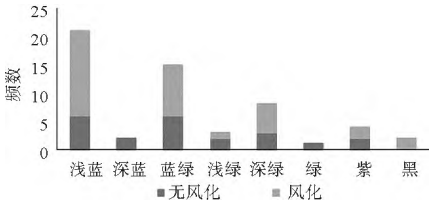


图3　颜色与表面风化的频数统计

由图1可知，在无风化与风化两种情况下，纹饰类型A、B、C变化差异较小，故可初步推断表面风化与纹饰类型无关；由图2可知，风化与无风化相比，铅钡玻璃类型占比增加，高钾玻璃类型占比减少，且变化幅度较大，故可初步推断表面风化与玻璃类型有关；由图3可知，风化与无风化相比，各种颜色变化较小，故可初步推断表面风化与颜色无关.

为进一步验证推断是否合理，利用SPSS软件进行卡方检验，结果见表7.

表7　卡方检验结果表

变量	卡方检验 p 值
纹饰与表面风化	0.085
玻璃类型与表面风化	0.024
颜色与表面风化	0.325

因只有玻璃类型与表面风化的卡方检验 p 值小于0.05，故玻璃类型与表面风化有相关关系，纹饰、颜色与表面风化的卡方检验 p 值均大于0.05，故纹饰、颜色与表面风化无相关关系.

3.1.2　不同类型玻璃表面有无风化时化学成分含量的统计规律分析

运用附表2处理后的成分数据进行分析，分别计算出铅钡、高钾玻璃的无风化与风化检测点数据均值. 具体计算方法如下.

因成分数据为几何结构，故可对其进行扰动运算(类似实数空间上的加法运算). 对于任意成分数据 $\boldsymbol{x}=(x_1, x_2, \cdots, x_D)^{\mathrm{T}}$，$\boldsymbol{y}=(y_1, y_2, \cdots, y_D)^{\mathrm{T}}\in S^D$，$\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的扰动运算定义为：

$$\boldsymbol{x}\oplus\boldsymbol{y}=C(x_1y_1, x_2y_2, \cdots, x_Dy_D)^{\mathrm{T}}\in S^D. \tag{4}$$

由扰动运算可进一步计算均值. 运用R语言对成分数据进行均值计算，结果如表8所示(仅列出部分化学成分).

表8　高钾玻璃、铅钡玻璃风化与无风化均值表(部分)　　%

类型	SiO_2	Na_2O	K_2O	SnO_2	SO_2
高钾未风化	74.9529	0.8818	7.2326	0.2110	0.1234
高钾风化	93.6615	0.5268	0.3579	0.1515	0.0724
铅钡未风化	59.5046	1.3090	0.1851	0.1871	0.0953
铅钡风化	27.0317	0.7432	0.1502	0.2105	0.1745

为更直观地观察化学成分的变化，将结果可视化，如图4和图5所示. 其中，外圈为风化，内圈为无风化.

对于高钾玻璃，由图4可知风化后 SiO_2 占比增加较大，K_2O 显著减少，Al_2O_3 和CaO占比均有所下降；对于铅钡玻璃，由图5可知风化与无风化相比 SiO_2 占比降低，PbO、BaO、P_2O_5 和CaO占比升高.

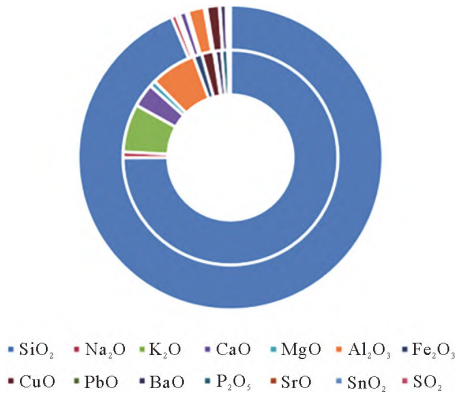


图 4　高钾玻璃风化前后元素对比图(彩图见封三)

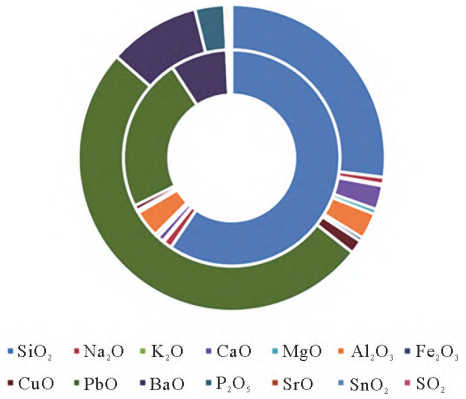


图 5　铅钡玻璃风化前后元素对比图(彩图见封三)

3.1.3　风化前化学成分预测

首先对附表 1 中的数据进行取虚拟变量处理，见表 5；再将附表 1 中的变量作为自变量，附表 2 中的化学成分作为因变量，作 Dirichlet 回归. 得到回归模型后，将自变量中的风化(1)替换为未风化(0)，得到预测结果，如表 9 所示(因预测结果数据量较大，故此处只给出了部分预测结果).

表 9　预测风化前的化学成分含量表(部分)　　%

文物编号	SiO_2	Na_2O	K_2O	SnO_2	SO_2
02	32.812	2.525	1.359	1.016	0.883
07	72.503	1.816	2.830	0.944	0.813
08	46.822	2.567	1.677	1.234	0.964
57	43.538	2.406	1.253	1.099	0.972
58	41.464	2.301	1.070	1.025	0.963

3.2　问题 2

3.2.1　玻璃的分类规律分析

决策树是根据已有样本的信息与现有分类，基于信息熵下降最快的原则，通过不断训练样本选择分类特征，最终得出包含节点与有向边的树状分类模型，可根据此模型对样本类别进行判断[11]. 该方法易于理解且便于结果可视化，同时对于小样本具有很好的分类效果. 考虑到本文样本量较小且需要选择不同类型玻璃的分类特征，故选用决策树模型.

在进行建模之前首先要对成分数据进行中心对数比(centered log-ratio，CLR)变换. 对于任意成分数据 $\boldsymbol{x}=(x_1, x_2, \cdots, x_D)^{\mathrm{T}}\in S^D$，clr 变换将 $\boldsymbol{x}\in S^D$ 变换为 $\mathbf{R}^D$ 上的系数，clr 系数为

$$\mathrm{clr}(\boldsymbol{x})=\left(\log\frac{x_1}{g_{\mathrm{m}}(\boldsymbol{x})}, \log\frac{x_2}{g_{\mathrm{m}}(\boldsymbol{x})}, \cdots, \log\frac{x_D}{g_{\mathrm{m}}(\boldsymbol{x})}\right)^{\mathrm{T}}. \tag{5}$$

记 clr 变换后的数据为 $\mathrm{clr}(\boldsymbol{x})=\boldsymbol{\xi}=(\xi_1, \xi_2, \cdots, \xi_D)^{\mathrm{T}}$，则 clr 逆变换为

$$\boldsymbol{x}=\mathrm{clr}^{-1}(\boldsymbol{\xi})=C(\exp\{\xi_1\}, \exp\{\xi_2\}, \cdots, \exp\{\xi_D\})^{\mathrm{T}}, \tag{6}$$

用转化后的成分数据建立分类模型. 为使分类结果直观明了，本文使用决策树模型对其进行特征的选择，进一步探究主要分类特征，从而推断两种玻璃的分类规律. 在此模型中，将风化、SiO_2、Na_2O、K_2O、CaO、MgO、Al_2O_3、Fe_2O_3、CuO、PbO、BaO、P_2O_5、SrO、SnO_2 和 SO_2 作为分类特征，将玻璃的类型作为类别构建决策树模型，且将 67 个样本划分为训练集与测试集，其中 47 个样本为训练集，剩余则为测试集. 最后采用 R 语言进行软件实现，结果如图 6 所示. 再用测试集进行预测，正确率为 100%.

由图 6 可知，最终以 PbO 为特征进行两类玻璃分类，若 PbO 值大于 1.3 则为铅钡玻璃，反之为高钾玻璃.

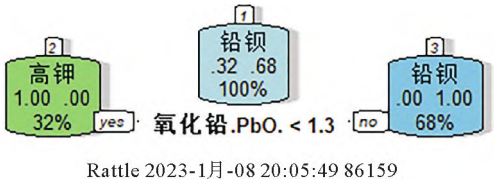


图 6　决策树结果

偏最小二乘判别分析是一种有监督的判别分析方法，可视为回归模型的扩展，主要用于多变量分析技术中的判别分析[12]. 与偏最小二乘回归相比，偏最小二乘判别分析仅需 1 个数据集，但需要提前知道数据的标签，其余与偏最小二乘回归原理相同[13]. 在此模型中，因变量为给定的类标签，自变量与传统回归模型相同，用此回归结果可推断出对因变量分类有较大影响的特征，从而可根据该模型对新给定的样本进行所属类别判断. 通过计算不同变量的投影重要性(VIP)可反映每个变量的加载权重以及其对因变量的解释程度，从而通过该方法选择出对分类有较大影响的变量[14].

下面采用偏最小二乘判别分析对高钾玻璃和铅钡玻璃进行分类. 由于 clr 变换后数据求和为 0，以玻璃类型为分类变量，风化、化学成分为特征变量，建立偏最小二乘判别分析，结果如图 7 所示，分析模型验证结果如图 8 所示. 由图 7 可知，高钾玻璃和铅钡玻璃有明显的分离趋势. 由图 8 可知，Q2 左侧交于 Y 轴负半轴，说明模型构建成功.

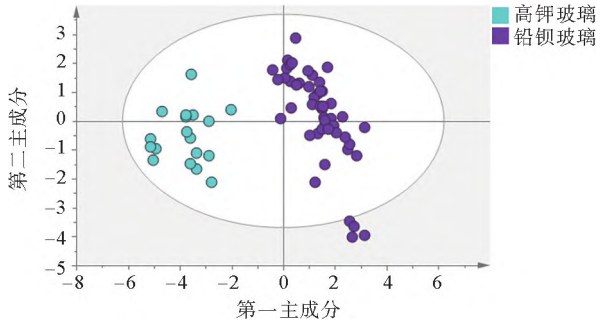


图 7 偏最小二乘判别分析结果

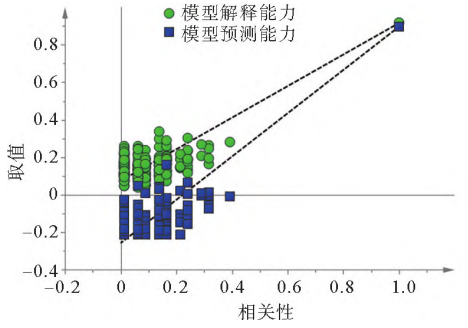


图 8 偏最小二乘判别分析模型验证

计算不同变量的投影重要性 VIP 值，结果如表 10 所示. 筛选 VIP 值大于 1 的特征，对于分类有影响的特征为 PbO、K_2O、BaO 和 SrO. 上面图表结果是在 simca 软件操作完成.

表 10 偏最小二乘判别分析的不同特征的 VIP 值

变量	VIP	变量	VIP
风化	0.2098	CuO	0.5694
SiO_2	0.8907	PbO	**2.4614**
Na_2O	0.1203	BaO	**1.3472**
K_2O	**1.8000**	P_2O_5	0.1226
CaO	0.5095	SrO	**1.1113**
MgO	0.3871	SnO_2	0.3548
Al_2O_3	0.6157	SO_2	0.1247
Fe_2O_3	0.7284		

3.2.2 亚类划分方法结果及敏感性分析

K-means 是一种根据样本距离将样本分类的无监督学习方法，首先把样品随机分成 k 个初始类；再不断进行修改迭代最终将样品分到其最近均值的类中去. 其中样本间的距离测算采用欧氏距离. 欧式距离公式为：

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 . \qquad (7)$$

将两类玻璃分别进行更进一步的亚类划分. 本次划分并不知亚类划分的结果，该批数据并未有明确的类型，因此该类问题为无监督学习问题，对其进行聚类分析.

首先需要选择所分类别的个数，采用 R 语言中的 fviz_nbclust 函数进行判断，结果如图 9 和图 10 所示. 根据其间断点可知两种玻璃的亚分类最终均选取 3 类.

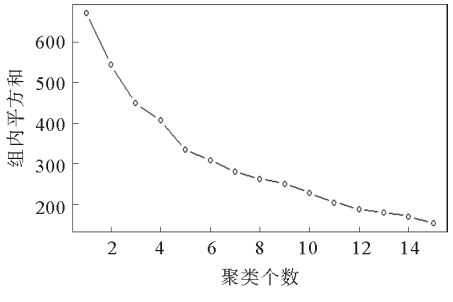


图 9 高钾分类个数确定

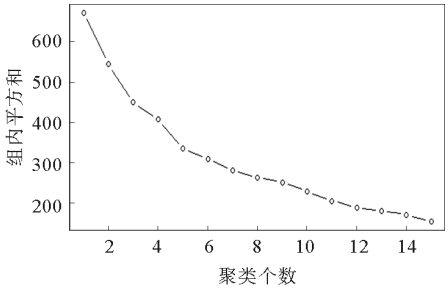


图 10 铅钡分类个数确定

其次，在确定聚类个数的情况下用 K-means 进行聚类，结果如图 11 和图 12 所示.

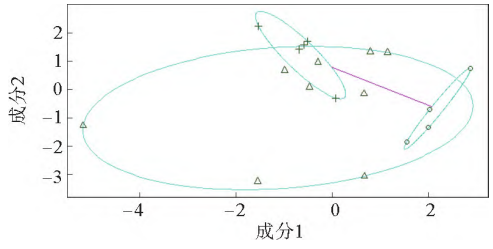


图 11　高钾聚类情况

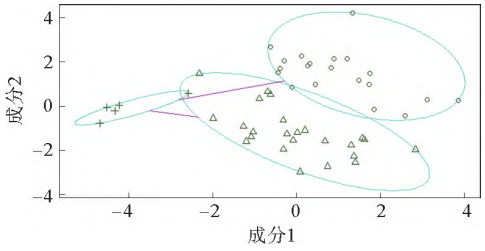


图 12　铅钡聚类情况

由图 11、图 12 可知铅钡玻璃可被明显地分为 3 类，故其亚分类包含 3 类；高钾玻璃 3 类划分并未有非常明显的类别，仅可粗略将其分为 3 类，如表 11 所示.

表 11　亚分类

玻璃类型	亚分类	文物编号
高钾玻璃	1	06 部位 1，18
	2	21，07，09，10，12，22，27
	3	01，03 部位 1，03 部位 2，04，05，06 部位 2，13，14，16
铅钡玻璃	1	20，37，50 未风化点，08，08 严重风化，11，19，26，26 严重风化，39，40，43 部位 2，50，51 部位 1，52，54，54 严重风化，56，58
	2	28 未风化点，29 未风化点，30 部位 1，30 部位 2，31，32，35，49 未风化点，02，41，48，49，51 部位 2
	3	23 未风化点，24，25 未风化点，33，42 未风化点 1，42 未风化点 2，44 未风化点，45，46，47，53 未风化点，55，34，36，38，43 部位 1，57

在确定分类个数时，不同的选择会出现不同的结果，故应对此模型的敏感性进行分析. 因高钾玻璃并未有明显的分类效果，故尝试多种分类个数进行聚类，从而选择出较优的聚类情况.

下面对每类玻璃的亚类划分选择合适的化学成分.

对于高钾玻璃，亚类划分为 3 类. 由于其中一类只有两种文物玻璃，因此删除这一类文物，以其余类为分类变量，化学成分为特征，构建偏最小二乘判别分析，如图 13 和图 14 所示. 由图 13 可知，高钾玻璃的亚类有明显的分离趋势. 由图 14 可知，Q2 左侧交于 Y 轴负半轴，说明模型构建成功.

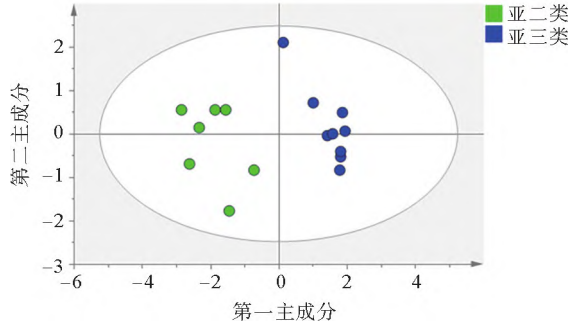


图 13　高钾玻璃亚类偏最小二乘判别分析结果

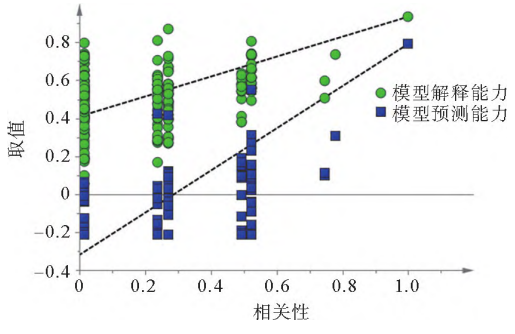


图 14　高钾玻璃亚类偏最小二乘判别分析模型验证

计算不同变量的 VIP 值，结果见表 12. 筛选 VIP 值大于 1 的特征，表明对于高钾玻璃亚类分类有影响的特征分别为 K_2O、SiO_2、CaO、SnO_2 和 BaO.

表 12　高钾玻璃亚类偏最小二乘判别分析的不同特征的 VIP 值表

变量	VIP	变量	VIP	变量	VIP	变量	VIP
SiO_2	**1.5267**	CuO	0.7509	MgO	0.0216	SrO	0.4147
Na_2O	0.3815	PbO	0.2068	Al_2O_3	0.4654	SnO_2	**1.2763**
K_2O	**2.3016**	BaO	**1.2158**	Fe_2O_3	0.5235	SO_2	0.3721
CaO	**1.3078**	P_2O_5	0.0160				

对于铅钡玻璃，亚类划分为 3 类. 以这 3 类为分类变量，化学成分为特征，构建偏最小二乘判别分析，结果如图 15 和 16 所示. 由图 15 可知，铅钡玻璃的 3 类有明显的分离趋势. 由图 16 可知，Q2 左侧交于 Y 轴负半轴，说明模型构建成功.

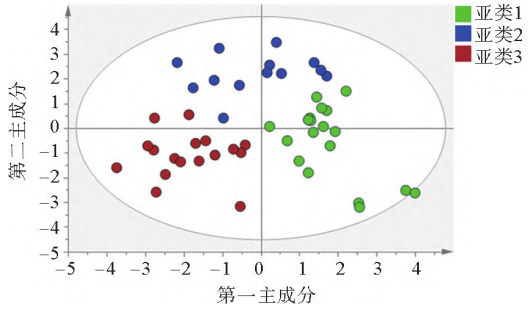


图 15 铅钡玻璃亚类偏最小二乘判别分析结果

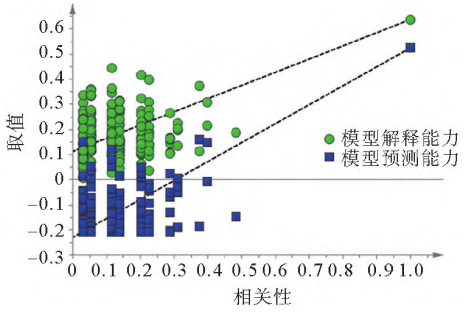


图 16 铅钡玻璃亚类偏最小二乘判别分析模型验证

计算不同变量的 VIP 值，结果见表 13. 筛选 VIP 值大于 1 的特征，对于铅钡玻璃亚类分类有影响的特征分别为 P_2O_5、CuO、Na_2O 和 Fe_2O_3.

表 13 铅钡玻璃亚类偏最小二乘判别分析的不同特征的 VIP 值表

变量	VIP	变量	VIP
SiO_2	0.9550	CuO	**1.3711**
Na_2O	**1.3046**	PbO	0.2813
K_2O	0.5825	BaO	0.8605
CaO	0.6696	P_2O_5	**2.0512**
MgO	0.7716	SrO	0.4435
Al_2O_3	0.8339	SnO_2	0.4663
Fe_2O_3	**1.2119**	SO_2	0.7193

3.3 问题 3

方法一：表单 3 中未知玻璃类型的类别划分. 基于决策树的结果，以 PbO 为特征进行分类，若 PbO 值大于 1.5 则为铅钡玻璃，反之为高钾玻璃. 结果表明，A1、A6、A7 为高钾玻璃，A2、A3、A4、A5、A8 为铅钡玻璃.

方法二：选取问题 2 中对于高钾玻璃和铅钡玻璃筛选的特征 PbO、K_2O、BaO 和 SrO，建立这 4 个特征与玻璃类型的偏最小二乘回归分析，其中因变量玻璃类型中高钾玻璃取值为 1，铅钡玻璃取值为 0.

基于交叉验证方法计算预测均方根误差，使用所有主成分进行回归. 得到的结果如表 14 所示.

表 14 不同主成分个数下偏最小二乘回归结果表

评价准则	1 个	2 个	3 个	4 个
交叉验证的预测均方根误差	0.1471	0.1389	0.1371	0.1372
调整的交叉验证的预测均方根误差	0.1470	0.1381	0.1365	0.1365
累计贡献率/%	80.19	87.15	93.55	100

从回归结果可以看出，主成分个数为 3 时，模型在经交叉验证后得到的预测均方根误差最小，同时 3 个主成分对各变量的累计贡献率已经达到 93%，因此将偏最小二乘回归的主成分个数设定为 3.

主成分个数确定后，计算得到 PbO、K_2O、BaO 和 SrO 的偏最小二乘回归系数分别为 0.0134、−0.1582、−0.0288 和−0.0206.

将表单 3 中 PbO、K_2O、BaO 和 SrO 的数据代入偏最小二乘回归模型，得到不同文物的预测值，若预测值接近 1，为高钾玻璃；预测值接近 0，为铅钡玻璃. 预测结果见表 15.

通过上述分析，可以看出两种方法预测结果一致.

表 15 表单 3 未知玻璃文物的类型预测

文物编号	预测值	预测玻璃类型
A1	1.0408	高钾
A2	0.0177	铅钡
A3	0.0980	铅钡
A4	0.1898	铅钡
A5	0.2558	铅钡
A6	0.9615	高钾
A7	0.9631	高钾
A8	0.1122	铅钡

为了更进一步分析高钾玻璃与铅钡玻璃的亚类划分，基于上述亚类划分结果：

1)对于高钾玻璃，以 3 个亚类为因变量，化学成分 K_2O、SiO_2、CaO、SnO_2 和 BaO 为自变量，建立偏最小二乘回归，通过交叉验证确定主成分个数为 2，对表单 3 中 A1、A6 和 A7 文物进行预测，结果表明 A1、A6 和 A7 都是高钾玻璃的同一亚类；

2)对于铅钡玻璃，以 3 个亚类为因变量，化学成分 P_2O_5、

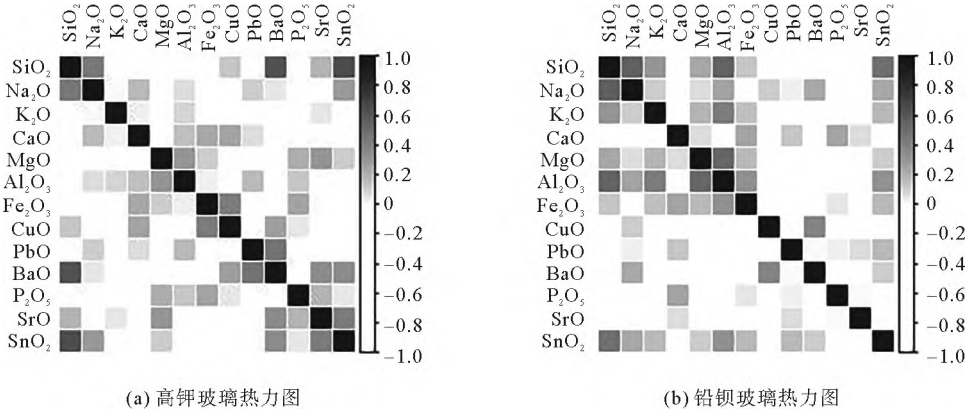
CuO、Na_2O 和 Fe_2O_3 为自变量，建立偏最小二乘回归，通过交叉验证确定主成分个数为 2，对表单 3 中 A2、A3、A4、A5 和 A8 文物进行预测，结果表明 A2 和 A4 是铅钡玻璃的同一亚类，A3 和 A8 是铅钡玻璃的同一亚类，A5 是铅钡玻璃的另一亚类.

3.4 问题 4

3.4.1 相关性分析

灰色系统着重内涵不明确的对象，将不确定的“灰”转化为可解读的“白”信息[15]. 其中灰色关联通过序列曲线集合形状可直观地对关联度进行分析.

将玻璃按照高钾玻璃与铅钡玻璃分类讨论. 因为题中要求分析不同类别文物化学成分之间的关联关系，且变量较多，热力图更可直观地体现两两化学成分之间的相关关系，故分别制作高钾玻璃与铅钡玻璃的化学元素热力图，如图 17 所示.



(a) 高钾玻璃热力图　　(b) 铅钡玻璃热力图

图 17　高钾与铅钡玻璃热力图

由图 17(a)可以看出，按照热力图上方从左到右元素顺序，每个元素与左右相邻元素相关性较强，SiO_2 与 BaO 关联性较强，其余两两成分之间关联性相对较弱. 由图 17(b)可以看出，SnO_2、SiO_2、Na_2O、Fe_2O_3、K_2O、MgO、Al_2O_3 这 7 个成分两两之间均存在相关性(Fe_2O_3 和 Na_2O 之间除外). CuO、BaO、SO_2 两两之间均存在较为强烈的相关性.

考虑到玻璃的主要化学成分为 SiO_2，故将 SiO_2 作为母序列，其他化学成分作为子序列，运用 Matlab 进行灰色关联分析，最终得出关联度值. 使用关联度值对 49 个评价对象进行排序，结果如表 16 所示.

表 16　两类玻璃化学成分关联度对比表

评价项(铅钡)	关联度	排名	评价项(高钾)	关联度	排名
Na_2O	0.7904	11	Na_2O	0.6291	11
K_2O	0.9639	5	K_2O	0.8991	8
CaO	0.8227	10	CaO	0.9132	7
MgO	0.9205	9	MgO	0.8571	9
Al_2O_3	0.9603	6	Al_2O_3	0.9765	2
Fe_2O_3	0.9278	8	Fe_2O_3	0.5685	12
CuO	0.6742	12	CuO	0.9204	6
PbO	0.9785	1	PbO	0.9577	4
BaO	0.9694	4	BaO	0.5655	13
P_2O_5	0.5400	13	P_2O_5	0.8341	10
SrO	0.9587	7	SrO	0.9798	1
SnO_2	0.9729	2	SnO_2	0.9492	5
SO_2	0.9698	3	SO_2	0.9673	3

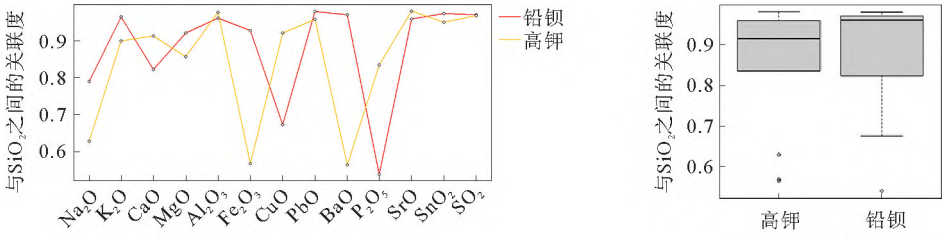
关联度值介于 0～1 之间，该值越大代表其与母序列 SiO_2 之间的相关性越强. 从表 16 可以看出：本次 13 个评价项中，铅钡玻璃中，PbO 的综合评价最高(关联度为 0.9798)，P_2O_5 的综合评价最低(关联度为 0.5400)；高钾玻璃中，SrO 的综合评价最高(关联度为 0.9785)，BaO 的综合评价最低(关联度为 0.5655).

3.4.2 差异性分析

比较不同类别之间的化学成分关联关系的差异性. 由于热力图中相关系数为对称矩阵，因此仅选取高钾玻璃相关系数上三角矩阵与铅钡玻璃相关系数上三角矩阵. 由于每个化学成分之间的相关关系

是配对的，因此选取配对样本的非参数 wilcoxon 检验，所得 p 值大于 $\alpha(0.05)$，因此不拒绝原假设，即两种类型玻璃的化学成分之间的关联关系没有显著差异.

通过 R 作出关于铅钡玻璃与高钾玻璃的关联度折线图和箱线图，如图 18 所示. 由图 18 可知，高钾玻璃与铅钡玻璃相比分散程度较小，且通过表 16 的排名可知，两者的化学成分之间的关联关系没有显著差异，进一步验证了上述结论.



(a) 折线图　(b) 箱线图

图 18　铅钡玻璃与高钾玻璃的关联度折线图和箱线图

4　模型检验及可靠性分析

4.1　针对问题 1 的检验

在解答问题 1 之前，对表单中数据首先进行预处理. 对于玻璃的化学成分，近似零值插补后转换为成分数据，化学成分比例和为 100%，数据预处理合理. 对于问题 1，考虑到成分数据的特殊结构，在成分数据单形空间上计算均值，选择适用于成分数据的 Dirichlet 回归模型，分析方法相比传统分析方法更加合理.

4.2　针对问题 2 的检验

对于问题 2，选择两种方法分析高钾玻璃与铅钡玻璃的分类规律，两种方法结果一致，进一步验证了分类模型的合理性. 对每类玻璃亚类划分时，基于 K-means 聚类分析确定了最优聚类个数，分析结果真实可靠.

4.3　针对问题 3 的检验

偏最小二乘回归方法基于交叉验证确定了最优主成分个数. 基于两种方法对未知玻璃文物进行类别预测，结果一致，因此预测结果合理.

4.4　针对问题 4 的检验

采用 Pearson 相关系数分析化学成分之间的关联关系，基于 wilcoxon 检验比较两种类型玻璃的化学成分之间关联关系的差异性. 结果真实可靠.

5　模型评价与展望

5.1　模型的优点

本文基于成分数据对玻璃化学成分进行分析，并对不同玻璃类型进行分类，主要优点如下：

1)采用 Dirichlet 回归模型对风化前的化学成分进行预测；

2)对于玻璃类型的分类，考虑到成分数据求和为 100%的约束，在模型构建前，首先对化学成分进行 clr 变换，使得成分数据变换为欧式空间上的普通数据；

3)采用不同分类模型对玻璃类型进行分类，不同模型结果一致.

5.2　模型的缺点

对于问题 4，不同玻璃类型的化学成分之间的关联关系，采用 Pearson 相关系数来说明. 但 Pearson 相关系数只能度量变量之间的线性相关关系，在使用之前未进行线性相关关系检验.

5.3　模型的展望

对于问题 2，分析不同玻璃之间的分类规律时，由于两种类型玻璃的样本量不是很接近，因此后续可以考虑不平衡样本分类模型，通过对训练集样本重采样或方法修正来进行分类. 对于问题 4，可

考虑利用其他相关系数来度量化学成分之间的关联关系，例如最大距离相关系数、互信息等.

参考文献

[1]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]. 西安：西北大学，2016.
[2]安家瑶. 玻璃器史话[M]. 北京：社会科学文献出版社，2011：7-11.
[3]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.
[4]干福熹. 中国古代玻璃的起源和发展[J]. 自然杂志，2006，28(4)：187-193.
[5]Baxter M J. Exploratory multivariate analysis in archaeology[M]. Edinburgh：Edinburgh University Press，1994：48-52.
[6]李青会，黄教珍，李飞，等. 中国出土的一批战国古玻璃样品化学成分的检测[J]. 文物保护与考古科学，2006，18(2)：8-13.
[7]史美光，何欧里，周福征. 一批中国汉墓出土钾玻璃的研究[J]. 硅酸盐学报，1986，14(3)：307-313.
[8]史美光，曲长芝，张日清，等. 中国早期玻璃器检验报告[J]. 考古学报，1984，4：449-457.
[9]Pawlowsky-Glahn V，Buccianti A. Compositional data analysis：theory and applications[M]. Chichester：Wiley，2011.
[10]Pawlowsky-Glahn V，Egozcue J J，Tolosana-Delgado R. Modeling and analysis of compositional data[M]. Chichester：Wiley，2015.
[11]李航. 统计学习方法[M]. 2 版. 北京：清华大学出版社，2019.
[12]Lee L C，Liong C Y，Jemain A A. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data：a review of contemporary practice strategies and knowledge gaps[J]. Analyst，2018，143(15)：3526-3539.
[13]崔耀耀，孔德明，孔令富，等. 基于重构三维荧光光谱结合偏最小二乘判别分析的油类识别方法研究[J]. 光谱学与光谱分析，2020，40(12)：3789-3794.
[14]邱丰，俞艳文，魏宇锋，等. 拉曼光谱法结合偏最小二乘法快速测定柴油十六烷值[J]. 理化检验(化学分册)，2021，57(10)：885-889.
[15]刘小彩，韩宗霖，付宁，等. 基于灰色关联度的河南省高校 R&D 影响因素分析[J]. 管理工程师，2022，27(5)：58-65.

Analysis and Classification of Ancient Glass Products Based on Compositional Data

MA Peiying，HAN Yanlai，LI Delan，CHEN Jiajia
(School of Statistics，Shanxi University of Finance and Economics，Taiyuan，Shanxi 030006，China)

Abstract：The chemical composition of ancient glass products belongs to the compositional data. Therefore，using the compositional data analysis method，the chemical composition of glass products is analyzed and its classification rule is studied，thus the type of unknown glass relics is identified. Firstly，based on Spearman correlation coefficient and Chi-square test，the relationship between surface weathering of glass relics and its type，decoration and color were analyzed；the statistical rule of weathering chemical composition on glass surface was analyzed by means in simplex space；the Dirichlet regression model was constructed to predict the chemical composition before weathering. Secondly，two models of decision tree and partial least squares discriminant analysis were constructed to select the features of the initial classification of the two types of glass；further，K-means clustering was used to sub-classify the two types of glass，and partial least squares discriminant analysis was used to sub-classify the two types of glass；then，the classification rules were used to identify the unknown type of glass. Finally，grey correlation analysis was used to explore the correlation and difference between the two types of glass chemical components.

Key words：compositional data；decision tree；partial least squares discriminant analysis；K-means clustering；grey relation

作者简介

马佩莹(2002—)，女，山西财经大学统计学院 2020 级本科生.
韩雁来(2001—)，女，山西财经大学统计学院 2020 级本科生.
李德兰(2001—)，女，山西财经大学统计学院 2020 级本科生.
陈佳佳(1991—)，女，博士，副教授，主要从事数理统计方面的研究.