教学与竞赛

基于成分数据的古代玻璃制品的成分分析与鉴别

这样可以吗？

黄慧婷[1]，李春明[1]，刘思语[2]，毛　睿[1]

（1. 桂林电子科技大学 数学与计算科学学院，广西 桂林 541004；
2. 桂林电子科技大学 计算机与信息安全学院，广西 桂林 541004）

摘　要：根据一批我国古代玻璃制品的相关检测数据，对这些古代玻璃制品进行风化规律的研究，并根据风化规律进行分类，同时给出玻璃制品风化前的预测模型. 首先，利用卡方检验分析风化与颜色、玻璃类型以及纹饰间的关系，并通过回归方程建立化学成分趋势变换模型以预测风化前的化学成分含量；然后，采用决策树法和以 R 型聚类法得到特征变量为基础的 Q 型聚类对玻璃制品进行亚类分类，并进行了合理性和敏感性分析；第三，利用前述分类模型对未知类别玻璃文物的化学成分进行分析并得到其所属类型；最后，应用灰色关联分析，针对不同的玻璃文物，分析其化学成分之间的关联关系.

关键词：中心化对数变换；卡方检验；决策树法；聚类分析；灰色关联分析

中图分类号：O29　　**文献标志码**：A　　**文章编号**：2095-3070(2023)02-0052-11

DOI：10.19943/j.2095-3070.jmmia.2023.02.05

0　引言

玻璃是丝绸之路早期贸易交流的重要物证，其主要化学成分为 SiO_2，而纯石英砂的熔点较高，因此在炼制时需通过添加助熔剂来降低熔化温度. 古代常用硝石、铅矿石、草木灰及天然泡碱等作为助熔剂，且用石灰石作为稳定剂. 对不同类型玻璃制品的化学成分的分析将可以判断其发源或流行地[1]. 现有研究中，赵匡华[2]指出我国古代玻璃制品按照化学成分分为 3 种类型：铅基玻璃、钾基玻璃和钙钠玻璃，并重点探讨了铅基玻璃和钾基玻璃的来源和演进. 铅基玻璃中最主要的铅钡玻璃，是在烧制玻璃的过程中加入铅矿石作为助熔剂，其氧化铅(PbO)、氧化钡(BaO)含量较高，通常被认为是我国自己发明的玻璃品种，汉楚文化的玻璃以铅钡玻璃为主[3]；而以含钾量高的物质如草木灰作为助熔剂烧制而成的高钾玻璃，主要流行于我国岭南以及东南亚和印度等区域[4]. 由于古代玻璃很容易因为埋藏环境的影响而导致风化，且在风化的过程中，环境元素会与玻璃内部元素进行大量的交换，导致玻璃的成分比例发生改变，继而会影响对玻璃类别的正确判断.

利用科学合理的数据分析方法，为大批量样品数据的有效分析判断提供了理论依据. 伏修峰、干福熹[5]利用聚类分析和因子分析对我国南方和西南地区的古代玻璃成分进行研究，给出了铅钡玻璃和高钾玻璃的 5 个分类系别，并得出在汉代我国就已经拥有了自主生产玻璃的能力的结论. 使用类似的

收稿日期：2023-01-20

通讯作者：李春明，E-mail：Axieyun@163.com

引用格式：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.

HUANG H T，LI CH M，LIU S Y，et al. Composition analysis and identification of ancient glass products based on compositional data(in Chinese)[J]. Mathematical Modeling and Its Applications，2023，12(2)：52-62+124.

多元统计分析方法，已有大量研究成果运用于医学、环境、教育和地质等各个领域[6-8]. 对古代玻璃检测所获得的数据为成分数据，而传统的主成分分析方法在处理成分数据时产生了不适应性[9]，文献[10]中指出其不适应性主要包括两个方面：1)无法再利用欧式空间方法进行解释；2)传统方法得到成分数据的协方差矩阵是奇异的. 因此利用由 Aitchison[11]提出的中心化对数比变换(centered log-ratio，CLR)进行处理将是一种有效的做法.

根据 2022 年全国大学生数学建模竞赛 C 题[12]所提供的附件相关数据，在进行 CLR 处理后，利用 Pearson 和 Yates 卡方检验对这两种玻璃制品的表面风化和颜色、玻璃类型以及纹饰的关系进行分析，并利用回归模型对风化前后的变化规律进行研究；然后根据规律，使用决策树法对所给的古代玻璃制品进行初分类，使用 Q 型聚类法进行亚分类；再利用所建数学模型对未知类型的古代玻璃制品进行鉴别；最后分别分析不同玻璃类型的化学成分之间的关联程度.

1 数据预处理

赛题提供了高钾和铅钡两种古代玻璃制品相关的成分数据，包括这些文物的分类信息和相应的主要成分所占比例.

1.1 空缺数据处理

表单 1 中的“颜色”存在 4 组缺失数据. 经观察，缺失颜色的玻璃为铅钡风化玻璃，在该类玻璃中考虑不同颜色分布，根据分布规律对这 4 组缺失颜色进行填充后可得 19 号文物和 40 号文物为浅蓝、48 号文物为蓝绿、58 号文物为深绿. 值得说明的是，由于在其他 3 个因素，即表面风化、类型以及纹饰相同的情况下，其颜色会出现多种情况，同时表单 2 的文物采样点只是文物表面的随机部位，并不能代表整个文物，因此填充的偏差对后续数据分析影响不大.

针对表单 2、3 中的空缺数据和 0 值，由题目可知，考虑空缺数据为未检测到该成分，并不代表该文物不存在该化学成分，可能是因为含量极少，因此对该空缺数据以 0.000 001 替代；对于 0 值，其原因可能是含量极小，也用 0.000 001 进行填充.

1.2 剔除无效数据

因检测手段等原因可能导致成分比例累加和不等于 100%的情况，有效数据的各成分比例累加和应介于 85%～105%. 统计发现，文物采样点 15 与 17 成分的比例累加和均低于 85%，属于无效数据，故剔除 15 与 17 这两组数据.

1.3 成分数据[13]

成分数据的定和限制是其基本性质. 由题目背景可知附件中数据为各化学成分占比，为了更好地分析数据统计规律，本文对有效数据进行了转换. 设原始数据为 $\boldsymbol{x}_0=(x_1^0, x_2^0, \cdots, x_n^0)$，经过转换后的数据为 $\boldsymbol{x}=(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$，其累计和为 100%，转换公式如下：

$$\boldsymbol{x}_i=x_i^0/\sum_{j=1}^{n}x_j^0,\ i=1, 2, \cdots, n. \tag{1}$$

1.4 中心化对数比变换

n 元成分数据所处的向量空间称为单形空间. 由于单形空间需满足定和约束，因此针对普通数据的传统统计学分析方法对于成分数据不再适用[9-10]. 通过查阅文献[11, 14]得知，在单形空间上分析往往具有以下 3 个问题：

1)数据的直观形态在单形空间和欧式空间上不同，无法跨空间进行解释；

2)在单形空间上计算得到的成分数据的协方差矩阵有明显偏负性，与欧式空间上的内涵截然不同；

3)单形空间上的成分数据缺乏参数分布，使得对数据的变异模式进行分析时参数建模困难.

基于上述存在的问题，本文对数据进行中心对数比变换(CLR)处理，经过变换后的数据可以更加体现成分特性，使得成分数据的可解释性更强. CLR 的计算公式如下：

$$y_i=\ln\frac{x_i}{g(x)},\ i=1, 2, \cdots, n, \tag{2}$$

其中：$g(x)=[x_1, x_2, \cdots, x_n]^{\frac{1}{n}}$；$x_i$ 表示经过转换后第 i 个化学成分数据.

1.5 量化处理

为更好地进行相关性分析，分别对表单1中纹饰、类型、颜色和表面风化4个指标进行量化处理，结果如表1所示.

表1 量化处理

量化前	量化后
A，B，C	0，1，2
高钾，铅钡	0，1
黑，蓝绿，绿，浅蓝，浅绿，深蓝，深绿，紫	0，1，2，3，4，5，6，7
未风化，风化	0，1

2 问题1

2.1 期望计数

卡方检验是一种利用样本数据的实际值与理论值的吻合度来判断接受还是拒绝原假设的方法，常用于分析两个分类变量之间的相关性. 由于玻璃文物表面风化、纹饰、玻璃类型和颜色均为分类变量，因此本文设计表面风化-纹饰、表面风化-颜色和表面风化-玻璃类型3组卡方检验用于分析指标间的相关性. 期望值计算公式如下：

$$E_{i,j}=\sum_{i=1}^{r}O_{i,j}\sum_{j=1}^{c}O_{i,j}/\sum_{j=1}^{c}\sum_{i=1}^{r}O_{i,j}\ , \tag{3}$$

其中：$O_{i,j}$ 表示实际值；$E_{i,j}$ 表示期望值.

本文考虑到使用卡方检验分析相关性时，针对不同列联表及期望计数应采用不同的卡方检验方式，因此首先求出3组数据的期望计数. 表面风化-玻璃类型组属于2×2列联表，其全部单元格期望计数大于5且总样本量大于40，因此采用Pearson卡方检验；表面风化-纹饰与表面风化-颜色组均属于 $R\times C$ 列联表($R=2$ 且 $C>2$)，均不满足使用Pearson卡方检验的前提，因此对该两组采用Yates校正卡方检验.

2.2 卡方检验[15]

本文对表面风化-玻璃类型组采用Pearson卡方检验，检验步骤如下.

Step1 提出假设

原假设 H_0：表面风化与玻璃类型相互独立不相关；备择假设 H_1：表面风化与玻璃类型有关联.

Step2 构造检验统计量

卡方检验统计量 χ^2 可用于估计期望值与实际值的偏离程度，计算公式如下：

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}\ . \tag{4}$$

对表面风化-纹饰与表面风化-颜色组采用Yates校正卡方检验，Yates校正卡方检验统计量 χ^2_{Yates} 计算公式如下：

$$\chi^2_{\text{Yates}}=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(|O_{i,j}-E_{i,j}|-0.5)^2}{E_{i,j}}\sim\chi^2((r-1)(c-1))\ . \tag{5}$$

Pearson卡方检验及Yates校正卡方检验结果如表2所示.

表2 卡方检验结果表

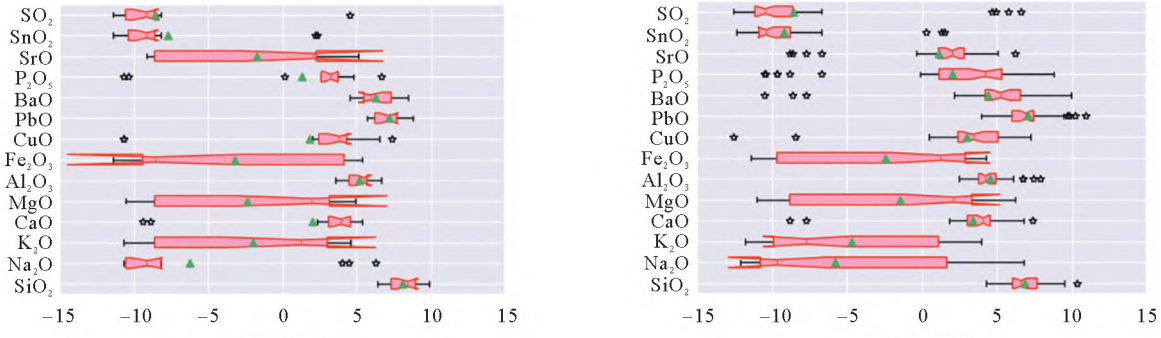
组别	检验统计量	P 值
表面风化-玻璃类型	6.880	0.009
表面风化-纹饰	4.957	0.084
表面风化-颜色	6.952	0.434

由表2，表面风化-玻璃类型组的Pearson卡方检验的 P 值为0.009，小于0.05，故拒绝原假设，接受备择假设，即认为表面风化与玻璃类型相关；表面风化-纹饰组与表面风化-颜色组的Yates校正卡方检验 P 值均大于0.05，因此在显著性水平5%上，不能拒绝原假设，拒绝备择假设，即认为表面风化与纹饰、颜色相互独立不相关.

2.3 描述统计规律

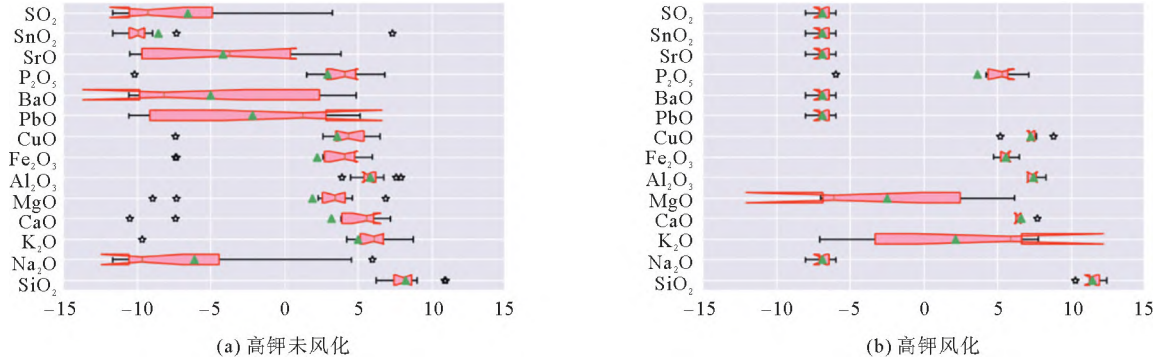
题目要求结合玻璃类型，研究有无风化样品化学成分含量的统计规律. 依据表单1将表单2中的数据分为4大类：高钾风化、高钾未风化、铅钡风化和铅钡未风化，分别讨论此4类玻璃文物表面化学成分含量的统计规律.

箱线图可反映多组连续型数据分布的散布范围以及中心位置，箱子的宽度在一定程度上可以反映样本数据的波动程度，因此本文通过箱线图来统计各个化学成分指标的数值分布特征. 经过中心化对数比变换后的数据的箱线图如图 1 和图 2 所示，其中横坐标表示转换后的欧式空间的值.



(a) 铅钡未风化　　(b) 铅钡风化

图 1　铅钡玻璃风化前后各成分的箱线图



(a) 高钾未风化　　(b) 高钾风化

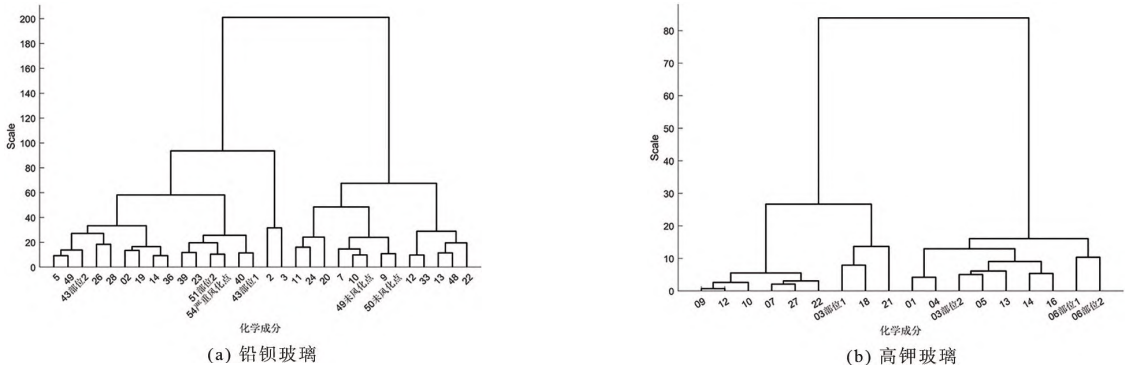
图 2　高钾玻璃风化前后各成分的箱线图

由图 1 和图 2 可以看出，铅钡玻璃与高钾玻璃的风化进程不同. 铅钡玻璃在风化过程中 CuO、K_2O 和 SiO_2 的含量会减少，风化前后 Fe_2O_3 数据分布离散程度都较大，说明风化对 Fe_2O_3 影响较小；而高钾玻璃在风化过程中 SrO、SO_2、PbO、BaO 和 K_2O 含量减少，SiO_2、Al_2O_3、CaO、Fe_2O_3 和 CuO 含量增多. 由此可以说明，K_2O 是风化流失产物，Al_2O_3 和 CaO 是风化产物.

2.4　确定风化点：Q 型聚类

题目要求根据风化点检测数据预测其风化前的化学成分含量. 由题目可知，表单 2 的检测数据除个别表明数据来源，其余均为随机采样而来. 根据箱线图求解可知，铅钡玻璃和高钾玻璃的化学成分变化规律不一致，因此本文先将数据根据玻璃类型进行分类，再对各自的两个类别进行 Q 型聚类分析，分别将两组数据分化为取自风化点和取自非风化点两类，从而确定需要预测的数据.

分别对高钾和铅钡玻璃未经过中心对数比转换[5]的数据进行 Q 型聚类，结果如图 3 所示.



(a) 铅钡玻璃　　(b) 高钾玻璃

图 3　铅钡玻璃和高钾玻璃各成分 Q 型聚类结果图

选定 $k=2$，分别对高钾和铅钡玻璃数据进行聚类分析. 基于表单1中给出的表面风化结果，可以对高钾玻璃的两类聚类进行区分，该聚类结果也可以反向证明本文箱线图结论正确. 对于铅钡玻璃，通过观察聚类结果发现:明确采样点为未分化点的数据均聚于一类，因此可以对铅钡玻璃的两类聚类进行区分.

2.5　建立时序关系

本题需要根据风化点检测数据预测风化前的化学成分含量，首先对成分数据通过聚类建立时序关系. 选定 $k=4$，分别对高钾和铅钡玻璃进行聚类. 本文以未风化、轻度风化、中度风化和严重分化分别表示通过Q型聚类得到的4个类别. 判定聚类所属类别方式与前文两类聚类方式一致. 具体聚类结果见表3和表4.

表3　高钾玻璃时序表

类别	组别
未风化点	06部位1，06部位2
轻度风化	01，03部位2，04，05，13，14，16
中度风化	03部位1，18，21
严重风化	07，09，10，12，22，27

表4　铅钡玻璃时序表

类别	组别
未风化点	28未风化点，29未风化点，31，32，33，35，37，44未风化点，45，48，53未风化点
轻度风化	11，20，23未风化点，24，25未风化点，42未风化点1，42未风化点2，46，47，49未风化点，50未风化点，55
中度风化	02，19，30部位1，30部位2，34，36，38，39，40，41，43部位1，43部位2，49，50，51部位1，51部位2，52，54，54严重风化点，56，57，58
严重风化	08，08严重风化点，26，26严重风化点

根据表单1数据，高钾玻璃03部位1、18、21点取自未风化玻璃，而聚类将其分为中度风化点；铅钡玻璃20、24、46、47、55点取自未风化玻璃，而聚类将其分为轻度风化点；铅钡玻璃30部位1、30部位2点取自未风化玻璃，而聚类将其分为中度风化点. 观其化学成分猜其原因，可能是取自未风化玻璃的风化区域. 铅钡玻璃28、29、48点取自风化玻璃，而聚类将其分为未风化点，观其化学成分猜其原因，可能是取自风化玻璃的未风化区域. 如表5和表6所示.

表5　高钾玻璃混淆矩阵表

高钾混淆矩阵		分类状态(时序)				
		未风化	轻度风化	中度风化	严重风化	合计
真实状态	风化	0	7	0	6	13
	未风化	2	0	3	0	5
合计		2	7	3	6	18

表6　铅钡玻璃混淆矩阵表

铅钡混淆矩阵		分类状态(时序)				
		未风化	轻度风化	中度风化	严重风化	合计
真实状态	风化	0	11	5	20	36
	未风化	9	0	2	2	13
合计		9	11	7	22	49

由高钾玻璃和铅钡玻璃混淆矩阵可清晰了解各个风化状态的点数.

2.6　基于中心化对数比变换的成分数据预测建模

根据聚类结果，取每个类的中心建立回归方程，如表7和表8所示，其数据均经过中心化对数比变换.

由表7和表8可知，拟合优度 R^2 相对较低可认为该化学成分不受风化影响，因此在后续预测过程中仅对拟合优度 R^2 较高的化学成分进行预测，对高钾玻璃的 P_2O_5、铅钡玻璃的 SnO_2 采取不变策略.

表 7　高钾玻璃不同化学物质的回归方程表

化学成分	拟合方程	R^2
SiO_2	$y_1=0.041\,785x^2-0.026\,844x+0.985\,56$	0.990
Na_2O	$y_2=-0.240\,93x^2+1.3672x-2.6554$	0.591
K_2O	$y_3=-0.035\,097x^2+0.001\,024\,5x+0.7911$	0.675
CaO	$y_4=-0.145\,68x^2+1.0762x-1.2343$	0.619
MgO	$y_5=-0.066\,418x^2+0.077\,537x+0.4481$	0.953
Al_2O_3	$y_6=0.017\,201x^2+0.008\,217\,6x+0.7064$	0.925
Fe_2O_3	$y_7=0.301\,81x^2-1.5603x+2.0066$	0.454
CuO	$y_8=0.172\,15x^2-0.769\,86x+1.2058$	0.523
PbO	$y_9=-0.059\,735x^2+0.025\,141x+0.082\,022$	0.444
BaO	$y_{10}=0.248\,32x^2-1.579x+1.5812$	0.592
P_2O_5	$y_{11}=0.046\,08x^2-0.214\,55x+0.691$	0.055
SrO	$y_{12}=0.113\,17x^2-0.854\,33x+0.748\,77$	0.821
SnO_2	$y_{13}=-0.181\,31x^2+1.2184x-2.7902$	0.712
SO_2	$y_{14}=-0.211\,35x^2+1.2311x-2.5667$	0.649

表 8　铅钡玻璃不同化学物质的回归方程表

化学成分	拟合方程	R^2
SiO_2	$y_1=-0.034\,471x^2+0.064\,375x+1.0677$	0.997
Na_2O	$y_2=-0.123\,71x^2+0.314\,32x-0.695\,58$	0.784
K_2O	$y_3=0.048\,235x^2-0.5324x+0.440\,46$	0.961
CaO	$y_4=0.109\,29x^2-0.474\,72x+0.796\,06$	0.476
MgO	$y_5=-0.340\,64x^2+1.3663x-1.2926$	0.944
Al_2O_3	$y_6=-0.026\,986x^2+0.062\,906x+0.674\,45$	0.983
Fe_2O_3	$y_7=-0.198\,17x^2+0.764\,26x-1.02\,14$	0.346
CuO	$y_8=0.002\,771\,4x^2+0.125\,05x+0.091\,312$	0.432
PbO	$y_9=-0.043\,799x^2+0.2391x+0.685\,24$	0.792
BaO	$y_{10}=0.074\,567x^2-0.339\,61x+1.0462$	0.290
P_2O_5	$y_{11}=-0.002\,747\,5x^2+0.279\,07x-0.391\,97$	0.997
SrO	$y_{12}=0.007\,187\,8x^2+0.159\,55x-0.368\,53$	0.999
SnO_2	$y_{13}=0.006\,677x^2-0.057\,856x-1.1683$	0.091
SO_2	$y_{14}=0.521\,79x^2-1.9704x+0.136\,99$	0.924

2.7　风化点未风化时的化学含量预测

假设回归方程为 $f(x)=ax^2+bx+c$，一个风化点的某化学成分为(t, y)，其中，t 表示该风化点所处的时序，y 表示该风化点的某化学成分含量检测值，距离 $d=|f(t)-y|$. 根据前文表述，当 $t=1$ 时表示未风化，因此本文认为对拟合曲线进行平移，使得该拟合曲线经过需要预测的点，即可反向预测出未风化前的数据. 该方法示意图如图 4 所示.

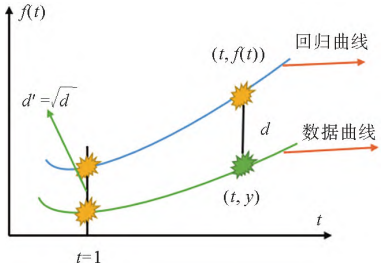


图 4　回归方程预测化学成分方法图

由于预测未风化前的化学成分含量取决于距离 d 和回归方程，当回归方程确定时距离 d 是唯一决定变量，为了减小异常数据对预测的影响，令 $d'=\sqrt{d}$，再使用 d'作为距离用于预测未风化前的化学成分.

1)中心化对数比逆变换

对于回归建模后求解出的未风化时各化学成分数据 $\boldsymbol{V}=(v_1, v_2, \cdots, v_p)$，需要通过中心化对数比逆变换回到实际值表示. 中心化对数比逆变换计算公式如下：

$$\begin{cases} w_j=v_j-v_p,\ j=1,2,\cdots,p-1, \\ x_j=\mathrm{e}^{w_i}/\left(1+\sum\limits_{i=1}^{p-1}\mathrm{e}^{w_j}\right),\ j=1,2,\cdots,p-1, \\ x_p=\left(1+\sum\limits_{i=1}^{p-1}\mathrm{e}^{w_i}\right)^{-1}. \end{cases} \tag{6}$$

2)预测结果

通过前文阐述的预测方法对风化点进行预测，结果如表 9 和表 10 所示.

表 9　高钾玻璃风化点预测表

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
07	42.53	0	0.24	0.01	0.07	7.56	14.25	6.98	0.01	0.06	28.20	0.04	0	0
09	17.04	0	47.70	0.01	0.08	2.19	18.21	4.90	0.01	0.05	9.79	0.01	0	0
10	40.17	0	38.40	0.01	0.06	3.01	14.68	3.61	0.01	0.06	0	0.04	0	0

续表 9

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
12	16.97	0	53.80	0.01	0.08	2.36	17.39	5.10	0.01	0.05	4.21	0.01	0	0
22	11.99	0	52.30	0.01	14.96	2.85	14.03	0.90	0.01	0.05	2.86	0.01	0	0
27	25.20	0	0.46	0.02	20.24	9.24	22.02	7.47	0.02	0.08	15.20	0.02	0	0
03	48.84	0	27.55	0.01	0.03	7.74	0.22	7.07	0.40	0.05	8.09	0.01	0	0
18	43.43	0	29.49	0	4.86	6.20	0.21	0.05	0	0.05	15.30	0.42	0	0
21	10.58	0	0.17	0.01	5.79	2.21	61.34	10.51	0.66	6.07	2.66	0.01	0	0

注：03为03部位1.

表 10　铅钡玻璃风化点预测表

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
08	77.52	0	0	0.13	0	3.90	0	0.34	9.76	8.34	0.01	0	0	0
08	38.63	0	0.01	1.13	0	8.77	0	0.41	29.00	21.90	0.04	0.02	0	0
26	78.66	0	0	0.13	0	1.40	0	0.34	10.70	8.74	0.01	0.01	0	0
26	54.88	0	1.54	0.80	0	6.23	0	0.32	19.00	17.20	0.02	0.01	0	0
02	67.14	0	0.78	3.81	0.07	13.8	0.03	0.11	13.90	0.26	0.09	0.03	0	0
19	63.57	0	0.01	5.12	0.08	5.65	0.04	1.19	14.70	9.51	0.15	0.03	0	0
30	54.37	0	0.79	4.53	0.07	5.81	0.03	0.01	10.70	23.10	0	0.03	0.5	0
30	56.83	0	0	4.36	0.06	4.79	0.04	0.01	10.30	23.00	0.05	0.04	0.6	0
34	69.54	0	0.34	1.37	0	1.82	0.02	0.37	9.70	16.80	0.01	0.02	0	0
36	65.29	0.1	0.67	0.62	0	3.11	0.03	0.17	12.60	17.40	0.01	0.03	0	0
38	57.44	0	0	1.78	0	3.59	0.02	0.15	13.50	23.40	0.01	0.03	0	0
39	67.40	0	0	1.11	0	1.26	0	0.26	18.80	11.10	0.02	0.02	0	0
40	62.79	0	0	1.44	0	1.24	0.01	0	22.30	12.20	0.03	0.02	0	0
41	56.92	0	1.23	7.18	0.15	5.30	0.06	0.15	16.00	12.80	0.16	0.05	0	0
43	31.73	0	0	3.92	0.05	7.33	0.02	1.02	34.40	21.50	0	0.04	0	0
43	60.47	0	0.01	7.76	0.10	6.05	0.05	0.37	16.90	7.99	0.20	0.06	0	0
49	61.67	0	0.01	6.15	0.10	8.58	0.05	0.20	12.90	10.10	0.17	0.05	0	0
50	43.16	0	0	4.63	0.06	3.33	0.03	0.24	13.50	34.90	0.12	0.05	0	0
51	60.46	0	0.01	5.77	0.12	6.25	0.06	0.28	14.70	11.60	0.16	0.02	0.5	0
51	71.22	0	0	1.99	0.03	5.74	0.01	0.28	20.60	0.08	0.05	0	0	0
52	62.79	0.1	0.01	4.74	0.10	3.47	0.04	0.22	16.20	12.20	0.14	0.05	0	0
54	55.30	0	0.82	5.37	0.1	6.98	0	0.22	18.10	12.90	0.12	0.07	0	0
54	70.79	0	0	0.01	0.02	6.71	0	0.35	22.00	0.07	0.05	0.02	0	0
56	66.48	0	0	1.01	0	4.21	0	0.23	15.40	12.60	0.03	0	0	0
57	65.38	0	0	1.00	0	4.59	0	0.27	16.20	12.60	0	0	0	0
58	63.09	0	1.08	5.54	0.11	5.01	0.05	0.43	14.10	10.50	0.16	0.01	0	0

表10中，第二个08、26、54均为严重风化点，两个30、43、51分别为部位1和部位2，预测数据中出现为0的数据表示该成分含量很少，并不是不存在.

3　问题 2

3.1　分析分类规律

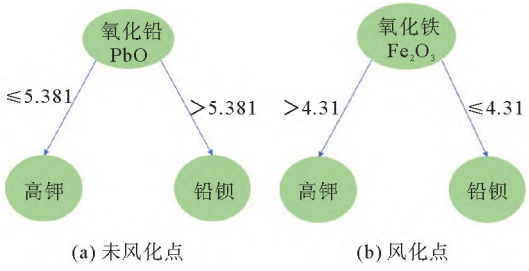
题目要求分析高钾玻璃和铅钡玻璃的分类规律. 由于表单2中每个数据所属玻璃类别都已知，因此可以采用监督学习进行分类. 由前文得知玻璃的主要化学成分含量会随着风化过程而变化，因此认为不可将所有数据混为一谈，在求解此问时选择将原始数据拆分成风化点数据与未风化点数据，再分别对两类数据使用决策树进行分类.

信息熵(H)以及信息增益(G)可定义如下：

$$H(p)=-\sum p\times\lg p,\quad H(Y\mid X)=\sum_{i=1}^{n}p_iH(Y\mid X=X_i),\quad G(D,A)=H(D)-H(D\mid A),$$

其中：$H(D)$表示经验熵；$H(D|A)$表示特征 A 在数据集 D 的条件下的经验条件熵.

针对风化点及未风化点数据集，取 70%数据作为训练集，30%数据作为测试集，得到如图 5 所示的决策树.



(a) 未风化点　　(b) 风化点

图 5　未风化点和风化点的未风化类型决策树

由图 5 可知，未风化点的玻璃类型分类规律是主要由 PbO 含量决定，当该玻璃中 PbO 含量小于或等于 5.381 时，将其归于高钾玻璃类，反之归为铅钡玻璃类；风化点的玻璃类型分类规律是主要由 Fe_2O_3 含量决定，当该玻璃中 Fe_2O_3 含量小于或等于 4.31 时，将其归为铅钡玻璃类，反之归为高钾玻璃类. 该模型的评估结果如表 11 所示.

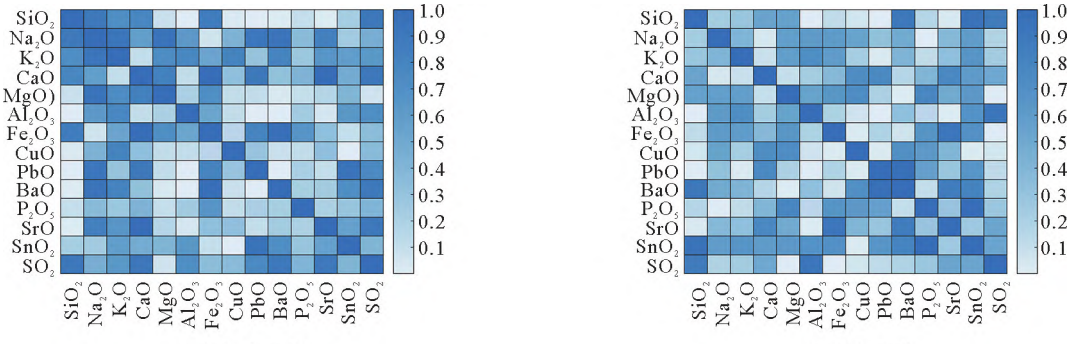
表 11　未风化类型决策树评估表

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

由表 11 可知，该模型的精确率、召回率、准确率和 F1 均为 1，表示该模型性能良好.

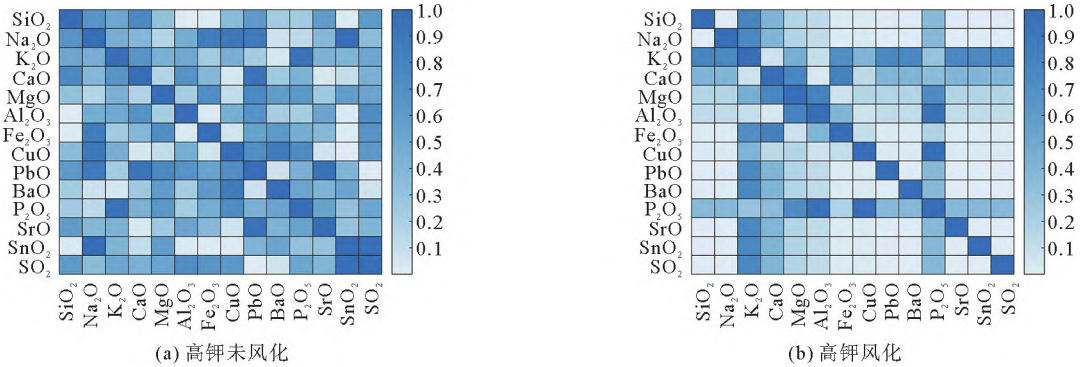
3.2　亚分类：R 型聚类分析

题目要求对每个类别选择合适的化学成分进行亚分类，基于前文得到的风化过程会对化学成分含量产生影响的结论，将原始数据分成铅钡风化、铅钡未风化、高钾风化和高钾未风化 4 个亚分类. 为了选择合适的化学成分进行亚分类，对玻璃的 14 种主要化学成分进行 R 型聚类分析，将 14 种化学成分聚成 3 类，再取每类具有代表性的成分进行 Q 型聚类，进而得到亚分类结果. 首先确定变量的相似性度量，求解出 14 种化学成分的皮尔逊相关系数矩阵，结果如图 6 和图 7 所示.



(a) 铅钡未风化　　(b) 铅钡风化

图 6　铅钡玻璃风化前后皮尔逊相关系数矩阵图



(a) 高钾未风化　　(b) 高钾风化

图 7　高钾玻璃风化前后皮尔逊相关系数矩阵图

通过相关系数矩阵可知，某些变量之间具有强相关性，因此可以对 14 个化学成分根据相关性进行 R 型聚类，再从每个类中选取具有代表性的特征变量.

本文使用相关系数用于度量变量间的相似性，使用最短距离法度量类间相似性进行 R 型聚类，分析结果如表 12 和表 13 所示，其中，* 代表特征变量.

表 12　未风化 R 型聚类表

类别	化学成分	
	高钾未风化	铅钡未风化
第一类	MgO^*，PbO，BaO，SrO	Na_2O，CaO^*
第二类	$SiO_2{}^*$，Na_2O，K_2O，Al_2O_3，P_2O_5，SnO_2	MgO^*，Fe_2O_3，SrO，SnO_2
第三类	CaO^*，Fe_2O_3，CuO，SO_2	$SiO_2{}^*$，K_2O，Al_2O_3，CuO，PbO，BaO，P_2O_5，SO_2

表 13　风化 R 型聚类表

类别	化学成分	
	高钾风化	铅钡风化
第一类	K_2O^*	$SiO_2{}^*$，Al_2O_3，CuO，PbO
第二类	$SiO_2{}^*$，Na_2O，CaO，Al_2O_3，Fe_2O_3，CuO，PbO，BaO，SrO，SnO_2，SO_2	CaO^*，BaO，P_2O_5，SrO，SO_2
第三类	MgO^*，P_2O_5	Na_2O，K_2O，MgO^*，Fe_2O_3，SnO_2

代表特征变量即选用的合适的化学成分，在后文将以 R 型聚类得到的特征变量为基础，进行 Q 型聚类. Q 型聚类结果如图 8 所示. 观察各类的数据特征，将各类划分为如下玻璃系统[5]：

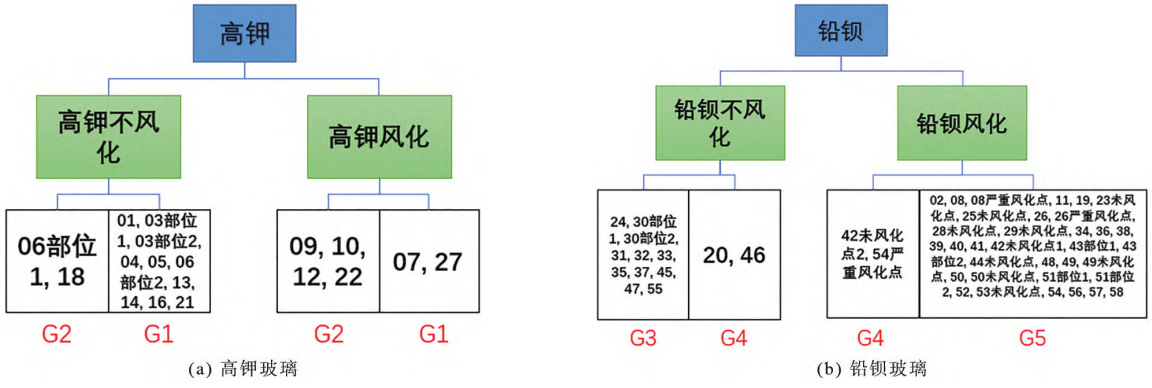
G1：K_2O-CaO(～10wt%)-SiO_2；

G2：K_2O-SiO_2；

G3：PbO-BaO-SiO_2；

G4：PbO(～25wt%)-BaO-SiO_2；

G5：CaO-PbO(～40wt%)-BaO-SiO_2.



(a) 高钾玻璃　　(b) 铅钡玻璃

图 8　高钾玻璃与铅钡玻璃 Q 型聚类结果图

3.3　合理性与敏感性分析

本文提出的划分亚类的方法充分利用了前问求解的答案，同时利用了 R 型和 Q 型聚类. 为了能够更好地划分亚类，在使用 R 型聚类得到合适的化学成分后，进行 Q 聚类对数据点进行划分，相比于直接进行 Q 型聚类划分亚类有更好的划分性能和划分依据.

对 R 型聚类得到的 3 个特征变量进行扰动处理，通过分析扰动比例对模型分类结果的影响对模型的敏感性进行分析. 对 CLR 变换后的数据进行扰动，通过给代表性特征变量随机进行扰动，可以得到

如表 14 所示结果.

从表 14 可以看出，扰动范围在[0.1，0.2]内，在一定程度上可以说明，小范围的数据变化不会对分类结果产生影响，由此可以说明本文使用的模型敏感性良好.

表 14 代表性特征变量扰动范围表

组别	代表性特征变量	扰动范围	组别	代表性特征变量	扰动范围	组别	代表性特征变量	扰动范围	组别	代表性特征变量	扰动范围
高钾未风化	MgO	0.10	铅钡未风化	CaO	0.20	高钾风化	K_2O	0.15	铅钡风化	SiO_2	0.15
	SiO_2	0.15		MgO	0.20		SiO_2	0.15		CaO	0.10
	CaO	0.15		SiO_2	0.05		MgO	0.15		MgO	0.15

4 问题 3

题目要求对未知类别玻璃文物的化学成分进行分析. 通过观察数据发现，所给出的数据为未分类玻璃文物的化学成分比例. 与表单 2 不同的是，表单 3 中的化学成分比例为文物的化学成分比例，不是通过进行随机表面采样检测到的化学成分比例，因此本文认为不需要再对表单 3 中数据进行分类(风化或未风化)处理.

在前文中分别对风化与未风化数据构建决策树，用于区分高钾和铅钡玻璃，认为表单 3 中数据与前文数据类型相似，因此在解决此问题时使用通过表单 2 数据训练好的决策树进行分类. 分类结果如表 15 所示.

通过 Q 型聚类方式对结果进行检验，分别对表面有风化和表面无风化玻璃进行聚类，聚类汇总结果如表 16 所示.

表 15 决策树分类结果表

玻璃类别	文件编号
高钾玻璃	A1，A6，A7
铅钡玻璃	A2，A3，A4，A5，A8

表 16 聚类结果表

玻璃类别	文件编号
高钾玻璃	A1，A6，A7
铅钡玻璃	A2，A3，A4，A5，A8

基于前文叙述，无论风化与否，对玻璃类型进行分类的唯一指标为 PbO. 对于风化类型，当 PbO 的含量大于 5.381 时，即可判定为铅钡玻璃，而在表单 3 中风化玻璃 A2、A5、A6 和 A7 的 PbO 的成分含量分别为 34.3%、12.23%、0%和 0%，它们与 5.381 相较差异较大，模型能接受的摆动范围也高，所以本文使用的模型在对风化玻璃进行分类时的敏感性较高. 未风化玻璃 A1、A3、A4 和 A8 的 PbO 的含量分别为 0%、39.58%、24.28%和 21.24%，它们相较于 5.381 的差异较大，模型能接受的数据摆动范围较高，因此本文使用的模型在对无风化玻璃进行分类时的敏感性较高.

5 问题 4：灰色关联分析

考虑到成分数据的协方差矩阵具有明显的偏负性，而灰色关联分析不假定数据的线性或者正态性，且是通过考虑系统中所有其他变量的影响来分析两个变量之间关系的方法[16]. 针对该问题，采用灰色关联分析来解决.

针对不同的玻璃类型，分析其化学成分之间的关联性，即分别研究高钾玻璃与铅钡玻璃的化学成分指标间的关联. 以 SiO_2 作为母序列，其余 13 组指标为子序列，建立灰色关联分析模型，遍历 SiO_2，Na_2O，…，SO_2 作为母序列，并对关联系数进行加权处理得到关联度值，以关联度值大小来衡量相关性大小.

分别以高钾玻璃和铅钡玻璃中 SiO_2 为母序列，对关联系数进行加权处理得到关联度值，求解出灰色关联度，结果如表17所示.

表17　高钾和铅钡 SiO_2 灰色关联度对比表

化学成分	高钾类关联度	铅钡类关联度	化学成分	高钾类关联度	铅钡类关联度
Na_2O	0.947	0.938	PbO	0.905	0.992
K_2O	0.924	0.913	BaO	0.947	0.975
CaO	0.949	0.957	P_2O_5	0.936	0.878
MgO	0.504	0.811	SrO	0.944	0.652
Al_2O_3	0.996	0.993	SnO_2	0.959	0.975
Fe_2O_3	0.943	0.863	SO_2	0.952	0.970
CuO	0.973	0.946			

当两类玻璃类型分别以 BaO 和 SrO 为母序列时，均与 PbO 的关联度最高；对于高钾玻璃而言，SiO_2 与 P_2O_5 互为最大关联性；对于铅钡玻璃而言，SiO_2 与 PbO 互为最大关联性.

6　结论

研究发现，玻璃文物表面风化程度与玻璃类型有关联，与纹饰和颜色相互独立不相关. 对铅钡玻璃及高钾玻璃来说，K_2O 是风化流失产物，Al_2O_3 和 CaO 是风化产物.

在主观分析层面，通过比对预测数据和采自未风化点的数据，发现预测数据与未风化点数据具有相似性，且使用经过中心化对数比变换后的数据进行回归分析，所以在预测风化前各化学成分含量的时候，可以保证预测的各化学成分累计和为100%，符合客观事实. 在客观分析层面，对混合了真实检测数据 D1 和预测数据 D2 的数据 D3 进行了聚类分析. 在二聚类情况下，结果表明预测数据 D2 全部与真实检测数据 D1 中的已知未风化检测数据聚为一类，即表面预测数据 D2 与未风化数据更为相似. 综上，认为本文提出的预测方法适用于预测风化前的化学成分含量.

高钾玻璃与铅钡玻璃的区分，主要取决于玻璃中 PbO 的含量，此含量为区分高钾玻璃与铅钡玻璃的主要指标. 对于3类代表性特征变量，其扰动范围处于[0.1，0.2]内，在一定程度上可以说明小范围的数据变化不会对分类结果产生影响，证明本文使用的模型敏感性良好. 本文针对问题3使用的分类模型具有较高的分类性能，在准确性和敏感性上取得了较好表现.

参考文献

[1]张斌. PIXE在古陶瓷、古玻璃产地中的应用研究[D]. 上海：复旦大学，2004.
[2]赵匡华. 试探中国传统玻璃的源流及炼丹术在其间的贡献[J]. 自然科学史研究，1991，10(2)：145-156.
[3]李晓岑. 关于中国铅钡玻璃的发源地问题[J]. 自然科学史研究，1996，15(2)：144-150.
[4]李青会，干福熹，顾冬红. 关于中国古代玻璃研究的几个问题[J]. 自然科学史研究，2007，26(2)：234-247.
[5]伏修锋，干福熹. 基于多元统计分析方法对一批中国南方和西南地区的古玻璃成分的研究[J]. 文物保护与考古科学，2006，18(4)：6-13.
[6]石军，熊苡. 多元统计、聚类分析法在自然资源开发中的应用[J]. 山东理工大学学报：自然科学版，2003，1：81-83.
[7]操群，周永正，余绍为. 一种基于主成分聚类分析的古陶瓷分类方法[J]. 中国陶瓷，2011，47(7)：48-51.
[8]王小娟. 中国古代白陶化学组成的多元统计分析[J]. 考古与文物，2017，5：124-138.
[9]罗纯. 基于成分数据若干分析方法的研究[D]. 长沙：中南大学，2011.
[10]郭丽娟，关蓉. 基于空间等价性的成分数据变换方法比较研究[J]. 统计学与应用，2018，7(2)：9-12.
[11]Aitchison J. Principal component analysis of compositional data[J]. Biomertrika，1983，70(1)：57-65.
[12]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.
[13]付亚龙. 成分数据处理方法研究[D]. 西安：长安大学，2019.
[14]蔡毅，邢岩，胡丹. 敏感性分析综述[J]. 北京师范大学学报：自然科学版，2008，44(1)：9-16.
[15]魏雨晨，武斌，宋怡昕，等. 基于卡方分析的大学生疫情后心理健康问题[J]. 中阿科技论坛(中英文)，2021，4：191-193.
[16]Fang H，Huang C，Zhao H，et al. CCLasso：correlation inference for compositional data through Lasso[J]. Bioinformatics，2015，31(19)：3172-3180.

(下转第124页)

Dynamic Production Optimization Model Based on Time Series Prediction

YANG Jingyun[1], LIAN Feihao[1], LU Jiyi[1], LING Weiwei[2]

(1. School of Information Engineering, Jiangxi Vocational College of Applied Technology, Ganzhou, Jiangxi 341000, China;
2. School of Social Management, Jiangxi Vocational College of Applied Technology, Ganzhou, Jiangxi 341000, China)

Abstract: In the production of multi-variety and small-batch materials, the demand is often unstable. Aiming at the problem of future demand forecasting, this paper establishes the weekly forecast model of material demand based on time series forecasting method and comprehensive use of mathematical software, and evaluates the model accuracy. In this paper, the concept of safety inventory is introduced, and the material production planning model based on safety inventory is established. At the same time, the production planning scheme of 6 kinds of materials is given. Considering that the inventory of materials needs to occupy funds, this paper discusses the relationship between safety inventory and service level, and the relationship between capital and service level, and selects the best safety inventory of each material based on the two relationships, and re-optimizes the production plan of 6 kinds of materials.

Key words: time series; safety inventory; service level; production planning

作者简介

杨静芸(2003—), 女, 2021 级专科生.
连斐豪(2002—), 男, 2021 级专科生.
卢继一(2003—), 男, 2021 级专科生.
凌巍炜(1983—), 男, 博士, 教授, 主要研究方向为数学建模、深度学习及地球物理正反演.

(上接第 62 页)

Composition Analysis and Identification of Ancient Glass Products Based on Compositional Data

HUANG Huiting[1], LI Chunming[1], LIU Siyu[2], MAO Rui[1]

(1. School of Mathematics and Computing Science, Guilin University of Electronic Science and Technology, Guilin, Guangxi 541004, China; 2. School of Computer and Information Security, Guilin University of Electronic Science and Technology, Guilin, Guangxi 541004, China)

Abstract: Based on a group of ancient glass objects in China, we studied the weathering pattern of these ancient glass objects, classified them according to the weathering pattern, and gave a prediction model for the weathering of glass objects before weathering. First, the relationship between weathering and color, glass type and ornamentation was analyzed by chi-square test, and the chemical composition trend transformation model was established by regression equation to predict the chemical composition content before weathering; then, the decision tree method and Q-type clustering based on R-type clustering method were used to classify the glass artifacts into subcategories, and the rationality and sensitivity analysis were conducted; third, the chemical composition of the un Third, the chemical composition of glass artifacts in the middle category was analyzed and their genus types were obtained using the aforementioned classification model; finally, grey correlation analysis was applied to analyze the correlations between the chemical composition of different glass artifacts.

Key words: centralized logarithmic transformation; chi-square test; decision tree method; cluster analysis; grey correlation analysis

作者简介

黄慧婷(2002—), 女, 2020 级信息与计算科学专业本科生.
李春明(2001—), 男, 2020 级应用统计学专业本科生.
刘思语(2002—), 男, 2020 级计算机科学与技术专业本科生.
毛　睿(1977—), 男, 硕士, 副教授, 主要从事最优控制理论与算法的研究.