基于多元统计分析的玻璃文物特征规律的挖掘

张子恒[1]，王舒熠[1]，封江浩[2]

[1]华中科技大学经济学院 湖北武汉
[2]华中科技大学软件学院 湖北武汉

【摘要】古代玻璃是东西方经济文化交流的重要见证，其成分，类别上的差异反映东西方不同的工艺特色与地质特征。本文基于多元统计分析方法，从定性和定量两种视角，对不同类别的古代玻璃开展多层次的关联性分析，采取定性的卡方检验模型，发现其中类别对表面风化的影响较为显著；通过分析文物采样点化学成分的统计规律，使用定性的方差分析分别挖掘各化学成分与类别，表面风化之间关系，通过量化统计规律并据此预测风化点风化前的化学成分含量挖掘各个特征及其关系的统计规律，为古代玻璃样本的类型判断，成分分析提供统计依据。通过分析玻璃文物表面风化与三个定性变量之间关系可知其中二氧化硅，氧化钾等 7 种化学成分与类别和表面风化显著关联。

【关键词】玻璃文物；方差分析；卡方检验
【收稿日期】2022 年 12 月 26 日 【出刊日期】2023 年 1 月 21 日 【DOI】10.12208/j.aics.20230006

Mining the characteristic patterns of glass artifacts based on multivariate statistical analysis

Ziheng Zhang[1], Shuyi Wang[1], Jianghao Feng[2]

[1]China, School of Economics, Huazhong University of Science and Technology, Wuhan, Hubei
[2]School of Software, Huazhong University of Science and Technology, Wuhan, Hubei, China

【**Abstract**】Historic glass is an important witness to the economic and cultural exchange between the East and the West, and the differences in composition and type reflect the different technological and geological characteristics of the East and the West. Based on multivariate statistical analysis, this paper presents a multi-level correlation analysis of different categories of ancient glass from both qualitative and quantitative perspectives, and adopts a qualitative chi-square test model to find that category has a more significant effect on surface weathering. The statistical patterns were quantified and used to predict the pre-weathering chemical content of the weathering sites, and the statistical patterns of each feature and its relationship were explored to provide a statistical basis for determining the type and composition of ancient glass samples. By analysing the relationships between surface weathering and three qualitative variables, seven chemical constituents, including silica and potassium oxide, were found to be significantly associated with category and surface weathering.

【**Keywords**】Glass artefacts; Analysis of variance; Chi-square test

1 引言

早期玻璃通过丝绸之路由西方地区传入我国，我国古代玻璃吸收其技术后采用本土材料制作玻璃，因此中西方玻璃制品的化学成分不尽相同。古代玻璃埋藏时易受环境影响而风化，造成其内部成分比例的改变，影响对其类别的判断。文物表面风化情况，纹饰，类型，颜色都是定类数据，彼此间构成列联表，显然不适用连续型数据分析方法。在列联分析中，卡方检验可用于分类资料相关分析和比较分类数据的差异性；线性对数模型可将列联表中的单元格频数的对数表示为各个变量及其之间交互效应的线性模型[1]，用于纯粹定类变量间关系的评价。在这两种方法之前，可以使用关联规则挖掘算法先验地找到类型数据间的关联关系，再加以检验。根据玻璃文物的采样点的数据，结合其类型，是否风化，对于化学成分分布的统计规律进行分析，

本质上是一个定性数据与定量数据的关联性分析问题。研究内容是玻璃文物采样点的化学成分与类别，风化条件的关系，研究对象是玻璃文物采样点。

本文利用多因素方差分析和多重比较的思想。在估计风化对于化学成分的偏效应，偏效应是指在其他因素不变的情况下（比如类别，纹饰，颜色）未风化样本被风化导致的化学成分的水平变化[2]。鉴于不同的类别，纹饰，颜色分组下，偏效应有一定的差异，所以我们针对上述三个变量进行聚类分析，聚类而得来的每一小类有着较为集中的类别，纹饰，颜色，所以我们针对每一小类的偏效应进行估计，预测数据较为有效。在对此类问题的定性分析中，如果定性数据划分的子样本较少，子样本之间趋于均衡，可以采用参数方法下的双样本 t 检验或者非参数方法下的双总体 Wilcoxon 符号秩检验的方法做差异性分析；如果定性数据维度较高，可以采用多因素方差分析，研究不同变量分组下因变量（所研究的定量数据）的水平差异[3]。在对该问题的定量分析中，可以采取回归分析，多重比较等方法量化子样本之间的差异。该问题的研究内容是玻璃文物采样点的化学成分与类别，风化条件的关系，研究对象是玻璃文物采样点，所以玻璃文物采样点的化学成分为唯一依据，玻璃文物采样点的类别以该采样点所属的玻璃文物的类别为唯一依据[4]。在未特殊注明下，玻璃文物采样点是否风化的判断以所属文物的是否风化为唯一依据。但对于已风化文物的未风化点，基于文献结果和模型假设，本文认为文物采样点是否风化的判断最终以其化学成分作为唯一依据，所以更正为“未风化”。

2 数据预处理

为方便后续模型的建立，精准挖掘附件数据的统计规律与分布特征，首先需要对原始数据进行清洗和预处理，去除噪音和无效数据，发掘数据规律，避免在后续统计分析中由于数据不准导致的错误。

2.1 数据完整性判断

数据的完整与否，首要在于是否有缺失值：附件数据中缺失值较为集中的区域在于各个样本点的化学成分，样本点中某化学成分存在缺失，意味着该成分未被检测到，所以缺失值的零值替换有一定的合理性；不难发现，氧化锡，二氧化硫缺失值分布达到了 91.0%和 89.5%，一般而言，如果某个特征缺失值比例如果超过 75%～80%，那么该指标的观测数据随机性较大，极易得到错误结论，所以在后续的统计分析中可以忽略氧化锡，二氧化硫这两个特征。有 4 组玻璃文物的颜色观测是缺失的，为了保证模型的合理性，并与现实情况相符合，可以将该类缺失值替换为单独一类的观测，称为“颜色无法识别”[5]。

数据的完整性还在于，数据之间的观测是否是正交的；不难发现缺失了纹饰为“B”并且类别为“铅钡”的样本数据。对于这一样本缺失，可以认为是由纹饰，类别，两个定性变量之间的相关性导致的。

2.2 数据有效性判断

首先，根据成分比例累加和在 85%以下或在 105%以上视为无效数据，出现了数据的谬误观测，所以文物采样点为 15，17 的两个样本可以删去；待预测类别的数据均是正常的；其次，数据的有效性还在于数据的集中分布，不存在离群值。最后发现玻璃文物采样点的若干化学成分氧化钾，氧化钡的分布呈现正偏态，离群数据较多；为保证信息的完整性，可以对偏态数据取对数处理，使得数据具有分布的正态性。

2.3 样本均衡判断

由于样本数据存在较多定类变量，比如类别，是否风化这类指标，所以有必要考虑样本不均衡问题；样本不均衡问题是指，各个类别之间的样本数据体量差异过大，导致数据挖掘中出现潜在的过拟合问题或者欠拟合问题。类别为“铅钡”的玻璃文物样本量明显多于类型为”高钾“的样本量，对此可以采用下采样和过采样相结合的方法，可以对体量较多的数据随机选取一定的样本，对体量较小的数据随机取样并加入原数据，作为数据的重复观测。最终均衡后的样本中两种类别体量均为 30 组数据。

2.4 标准化处理

由于数据的数量级不同，离散程度不同，所以有必要对数据进行 z—score 标准化处理。处理后的数据大致服从于标准正态分布，例如“二氧化硅”这一指标的分布情况。

3 模型建模与求解

研究文物表面风化和纹饰，类型和颜色之间的关系，为了严谨性，也防止由于文物数据体量较小，

单一的数据分析方法可能无法发掘数据内在规律或被偶然现象所误导，故我们采用数据可视化与描述统计，先验分析，定性检验和定量分析相结合的方法，发掘数据特点并进行检验。

3.1 定性分析：数据可视化与描述性统计

可以明显看出两种玻璃的特征纹饰，铅钡玻璃只有 A，C 两种纹饰类型，C 纹饰数量较高，可能与类型相关程度较高；高钾玻璃有 A，B，C 三种纹饰类型，每种类型数量表现为均衡分布，可以知道 B 是高钾玻璃的特征纹饰，结合题目，可以解释为古代中西方制作材料或文化及审美差异所致。可以明显看出两种玻璃的颜色特征，高钾玻璃主要为“蓝绿”，“浅蓝”，铅钡玻璃以“浅蓝”，“深绿”为主。玻璃颜色共有八种类型，其中有四个数据点的颜色无法识别，均为铅钡玻璃。高钾玻璃的颜色情况只有四种，分别为“蓝绿”，“深绿”，“浅蓝”和“深蓝”，且集中表现为“蓝绿”色；铅钡玻璃各个颜色均有，其中“浅蓝色”最多，“深绿色”次之。可以初步推断，高钾玻璃和铅钡玻璃的化学成分及含量不同，导致其表现出一定的特有的颜色。可以比较直观地看出两种玻璃各自的风化比例，高钾玻璃中风化个体与未风化个体的比例为 6：12，铅钡玻璃中风化个体与未风化个体比例为 28：12，铅钡玻璃风化比例明显高于高钾玻璃。

3.2 定性分析：基于 Apriori 算法的关联规则挖掘

Apriori 关联规则挖掘属于机器学习范畴下数据关联性分析的一种，适用于发掘“某一事件发生而引起的另外一件事件发生”的统计规律。可以结合案例“啤酒和尿布的故事”来说明算法的原理。

（1）基于统计数据发现，67%的顾客在购买啤酒的同时也会购买尿布；在这一观测下，“购买啤酒”称为规则的“前项”，购买尿布称为规则的“后项”。所有前项和后项构成一个项集，每一前项和后项之间构成项集中的一个规律[7]，记作：

$$\{Diaper\} \rightarrow \{Beer\}$$

（2）介绍项集的支持度这一概念，表示所有的事务（在本例下，等同于所有的交易）下出现该项集的概率，可以通过前项后项概率的加和来计算：

$$Support\ (\{Diaper,beer\}) = P\ (Diaper\ \cup\ Beer)$$

其中 Support 表示该项集的支持度，表示该类规律的频繁程度；

（3）对各个项集支持度进行排序，若某项集支持度高于频繁程度，称之为频繁项集。

（4）最后对频繁项集做显著性检验；置信度为前项发生，则后项也发生的概率，记作：

$$Cofidence\left(Diaper \Rightarrow Beer\right) = P\left(Beer \mid Diaper\right)$$

研究对象为是否风化与颜色，纹饰，类别之间的关系，这四个变量均为定性变量，并存在着一定的关联性，如：

（1）纹饰为 B 类型玻璃文物均属于“高钾”类别，“铅钡”类别玻璃文物没有观察到 B 纹饰；

（2）已被风化的玻璃文物不会出现深蓝色和绿色，且颜色无法识别的样本均被风化的；由于定类变量较多，每个定类变量的取值也比较多，所以人工发掘类似的规律较为困难；所以本文选择将 Apriori 关联规则挖掘算法作为正式研究之前的先验判断和样本依据，精准研究问题，节省人力物力。

通常设置阈值为 0.5，得到频繁项集及其显著性检验（部分结果）可见，纹饰 A 风化规律在一定程度上都是有章可循的。这些规律性的结果将会作为接下来卡方检验（定性数据的相关性分析）的起点和依据，将会逐步引入纹饰，类型，颜色三个变量，与是否风化这一变量做相关性分析和检验。

3.3 模型建立：基于卡方检验和对数线性模型的关联性定量分析模型

根据 Aprioi 算法得到的先验结果，得知一些变量之间存在关联关系或联合关联关系，首先需要对全体定类数据进行列联分析，定性地判断其关联关系。

卡方检验：

在列联分析中，卡方检验可用于分类资料相关分析和比较分类数据的差异性，下面简单介绍一下卡方检验模型。一个简单实现卡方检验的方法是 SPSS 软件。SPSS 中默认使用的统计量为皮尔逊卡方，其中 *Oi* 代表观测频数，*Ei* 代表期望频数。

卡方检验的原假设为 *H0：Oi=Ei*，因此若观测值对于期望值出现了较大程度的偏离，则倾向于有较大的卡方，说明数据在不同组别中有较大的差异性。进行卡方检验时，大部分检验结果只需要看皮尔逊卡方输出值即可，只有一类数据情况较为特殊，

即 2×2 的列联表，该表的自由度为 1，根据卡方分布曲线自由度小于 2 的曲线与其他曲线分布形状有很大差异。

4 结果分析

本文通过卡方检验进行分析，首先借助 SPSS 中描述统计功能探究了三类关系：纹饰*表面风化类型*表面风化颜色*表面风化，得到输出结果中有两类显著，然后分析第一步中结果，引入“类型”作为分类基础，探究纹饰*表面风化*类型，得到的卡方检验共得到三个显著结果，其余均不显著，认为无关联关系。最后得出皮尔逊卡方显著，说明类型和表面风化之间存在关联关系，高钾类型下的皮尔逊卡方值显著，铅钡不显著，说明在高钾类型下，纹饰和表面风化之间存在 一定的关联关系。通过定性分析作为定量分析的先导，能够节省物力、精准把握问题。

参考文献

[1] Żabiński Grzegorz,Gramacki Jarosław,Gramacki Artur, Miśta-Jakubowska Ewelina,Birch Thomas,Disser Alexan -dre. Multi-classifier majority voting analyses in provenan -ce studies on iron artefacts[J]. Journal of Archaeological Science,2020,113(C).

[2] 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(05):124-138.

[3] 李清临,朱君孝,秦颍,毛振伟,王昌燧,陈建立.微量元素示踪法在古代青铜器铜矿料来源研究中的应用[J].光谱学与光谱分析,2005(10):166-168.

[4] 苗建民,陆寿麟.运用中子活化分析和多元统计方法鉴定唐三彩陶器的真伪[J].故宫博物院院刊,2001(06):70-85.

[5] 陈建立.数学分析方法在考古学中的应用[J].中原文物,2000(01):48-52.

[6] 郭喜浩,徐昉昊,黄晓波,江涛,梁浩然,李长志,李智超.基于多元统计分析的油-源对比——以渤海湾盆地渤东凹陷为例[J].石油与天然气地质,2022,43(05):1259-1270.

[7] 常宏杰,樊温泉.新郑汉墓大泉五十的多元统计分析研究[J].华夏考古,2022(02):97-102.