成分数据理论和无监督聚类 K-means 方法提取背景和异常信息

——以安徽省兆吉口铅锌矿床为例

刘艳鹏[1)],朱立新[2)],马生明[3)],段吉琳[4)],弓秋丽[3)]

1) 东华理工大学核资源与环境国家重点实验室,江西南昌,330013;
2) 中国地质调查局发展研究中心,北京,100037;
3) 中国地质科学院地球物理地球化学勘查研究所,河北廊坊,065000;
4) 东华理工大学地球科学学院,江西南昌,330013

内容提要:地球化学数据是应用地球化学研究的重要组成部分,是化学勘查工作的基础成果。勘查地球化学数据基本上以元素的质量百分浓度(简称浓度)的形式表达,是典型的成分数据。其表达的是"组分/总体"相对质量贡献信息,而不是绝对的质量变化信息。浓度数据分布在单纯形空间,而不是整个欧式空间。对成分数据进行处理之前,进行适当的对数比值转换处理可以提高其信息表达。本文以安徽省兆吉口铅锌矿床土壤中 Pb 数据为示范案例,通过对数比值转换方法优化浓度数据的结构以提高相对信息的表达,并利用无监督学习 K-means 聚类方法根据对数比值转换数据分布空间质心的距离识别背景和异常信息,最后对 K-means 聚类方法识别的背景和异常与迭代 2 倍标准差法和浓度-面积分形分析法进行比较以衡量其表现。结果表明:浓度数据表达的是相对质量信息,而不是绝对质量关系,不同样品间不能通过比较浓度高低推断出质量的多寡关系。对数比值方法可以有效地提高浓度数据的结构和信息表达,K-means 方法能够准确识别对数比值转换数据的背景和异常信息,其效果类似浓度-面积分形分析方法,比迭代 2 倍标准差法好。

关键词:机器学习;无监督分类;成分数据;K-means 聚类;背景异常

地球化学数据是应用地球化学研究的重要组成部分,是化学勘查工作的基础成果。当样品化学分析结果出来后,应该及时将其以图件形式展示出来(Levinson, 1974),以便优选进一步工作部署的决策。地球化学图件为矿产勘查、土壤质量评价和风险评估、环境安全和人类健康、地球化学基准等研究提供着重要参考信息(Xie Xuejing et al., 1997; Morris et al., 2003; Smith et al., 2011)。随着社会经济的发展,不同尺度的地球化学调查项目陆续在全球范围内开展,比如中国多目标地球化学调查(Li Min et al., 2014)、中国多尺度地球化学填图(Xie Xuejing et al., 2008)、UNESCO 全球地球化学数据库(Darnley et al., 2005)、英国地球化学基准填图(Johnson et al., 2005)、欧洲农牧地土壤地球化学填图(Reimann et al., 2018)、北美土壤地球化学景观计划(Smith et al., 2011)。这些项目的开展,推动着样品采集和分析的标准化(Salminen et al., 1998; de Caritat et al., 2009; Smith et al., 2009)。通常每个样品会分析 50~70 多个元素/氧化物,进而产生了海量的地球化学数据。对这些数据进行合理的分析、解释和推断是地球化学勘查工作的基本需求。数据解释的关键步骤是识别物质分布的背景和异常。通常认为背景是"未矿化地质体中元素的正常丰度",异常是"元素对正常地球化学

注:本文为国家自然科学基金委青年科学基金项目(编号 41902071)、国际(地区)合作与交流项目(编号 42011530173)和东华理工大学博士科研启动基金项目(编号 DHBK2019313)联合资助的成果。
收稿日期:2022-07-25;改回日期:2022-09-22;网络发表日期:2022-11-07;责任编委:范宏瑞;责任编辑:潘静。
作者简介:刘艳鹏,男,1990 年生。博士,助理研究员,主要从事数学地球科学研究。E-mail: liuyanpeng@ecut.edu.cn。通讯作者:朱立新,男,1963 年生。研究员,主要从事勘查地球化学研究。E-mail: lixinz@cags.ac.cn。

引用本文:刘艳鹏,朱立新,马生明,段吉琳,弓秋丽. 2022. 成分数据理论和无监督聚类 K-means 方法提取背景和异常信息——以安徽省兆吉口铅锌矿床为例. 地质学报, 96(11): 4038~4055, doi: 10.19762/j.cnki.dizhixuebao.2022088.
Liu Yanpeng, Zhu Lixin, Ma Shengming, Duan Jilin, Gong Qiuli. 2022. Identification of background and anomaly information via compositional data theory and unsupervised K-means clustering: a case study of Zhaojikou Pb-Zn ore deposit, Anhui Provicne. Acta Geologica Sinica, 96(11): 4038~4055.

分布模式的偏离”(Hawkes and Webb, 1963)。异常下限是背景和异常的阈值,它的选择受元素的分布模式影响。目前,确定背景和异常的方法主要有:① 相关地区的参考值;② 平均值+n 倍标准偏差(n 通常为 1,2,2.5,3);③ 直方图;④ 累积频率图;⑤ 定向调查;⑥ 地质统计;⑦ 分形(Hawkes and Webb, 1963; Lepeltier, 1969; Cheng Qiuming et al., 1994; Matschullat et al., 2000)。其中,平均值法和分形方法是目前使用最广泛的方法。平均值法的理论基础是“元素在地球中的分布服从正态或对数正态分布”,通过观察样本数据的频率分布模式、期望和偏差,确定背景和异常。由于 n 的取值由专家的知识和经验决定,导致平均值法确定的异常下限具有一定主观性。平均值法存在的另一个问题是对离群值(outlier)的清洗。为了使数据满足正态分布的假设,通常在统计分析时会将偏离平均值 4 倍标准差的样本视为离群值而剔除掉。存在的问题是,没有足够的证据表明这些被视为“离群值”的样本不具有地质统计意义。通过剔除离群值得到的背景有时会相对偏低,比如迭代 2 倍离差法。分形方法的理论基础是“元素在地球中的分布服从尺度法则,具有分形或多重分形模式”,通过对分形维度的研究,确定背景和异常的阈值(Cheng Qiuming et al., 1994; Allegre and Lewin, 1995)。

元素的质量百分浓度(下文简称浓度)数据是典型的成分数据。成分数据的定义是:假设存在向量 $X=[x_1, x_2, \cdots, x_n]$,如果 x_n 均为正实数且加和为常数,则 X 称为成分数据(Aitchison, 1982)。$x_1, x_2, \cdots\cdots, x_n$ 等变量称为成分数据的组分。组分携带的是关于“组分与总体”相对信息,而不是单个组分的绝对信息。理想条件下,成分数据最多有 $n-1$ 个自由维度。成分数据是多个独立的成分相互混合后闭合的产物,具有多元变量属性。受闭合的影响,组分 x_m 与 $x_i(m\neq i)$ 之间是相互关联纠缠的,这种相互关联通常与相关性和协方差的解释相矛盾,因此大部分多元统计方法不适合处理成分数据(Chayes, 1960)。比如地质样品中元素 Pb 的浓度,代表的信息是元素 Pb 的质量在样品总质量中所占的比例,与其相对应的是非 Pb 组分的浓度。虽然 Pb 和非 Pb 物质质量之间不一定具有相关性,但如果只考虑浓度关系,Pb 浓度的升高/降低会导致非 Pb 物质浓度的降低/升高,二者成伪相关(spurious correlation)。识别闭合效应导致的伪相关和误解是成分数据解释的一个主要难点(Chayes, 1960)。因此,对成分数据进行的统计分析应满足尺度不变性、排列不变性和子成分一致性的原则(Aitchison, 1986)。元素的浓度由所有成分的活动结果共同决定。元素在地质作用中的活动可以分为富集、亏损和不活动三类,通常与其电荷/半径比(离子势)有关。溶液中具有低离子势的元素倾向于形成水合阳离子优先离开,具有高离子势的元素倾向于形成水合氧阳离子离开,而具有中等离子势的元素则倾向于在固体沉淀物中保持不动(Pearce, 2014)。

机器学习是挖掘数据模式的有效方法。机器学习通常分为监督学习、无监督学习和半监督学习。监督学习需要人类根据研究的问题对部分数据进行标记,以帮助人工智能代理学习到有效的数据特征,从而实现对新的未知数据的预测而解决特定的问题。无监督学习则是通过无标签数据学习隐藏在数据中的模式。学习的数据有没有打标签是监督学习和无监督学习的区别。半监督学习则是介于监督学习和无监督学习之间的一类方法(Russell et al., 2010)。虽然无监督学习比监督学习困难,但更强大。在无监督学习中,没有标签可以利用,因此,该类人工智能的任务不会特别明确,其表现也没有明确的度量标准。通常,人们可以根据最终的结果和任务目标之间的完成情况进行判断。无监督学习最主要的方法是降维和聚类(Figueiredo and Jain, 2002)。通过对降维和聚类的应用,可以实现异常检测和群组分割。降维可以有效地应对维数灾难,极大地缓解了人们认识数据在高维空间分布模式的困难。聚类是通过利用数据分布的底层结构和定义对具有相似特征的数据的分组规则进行分类(Jain et al., 1999)。聚类过程无需任何关于数据集的先验知识,仅根据定义的规则把高维空间 R^n 的数据样本按照相同的特征属性分离成合适的类群。理想的分类结果是每一类数据只包含相似的样本数据,并且与其他类中的样本数据有着明显的差异。这种差异的度量由数据的底层结构和算法的目标决定。根据聚类的模式,可以分为连通性聚类(比如分层聚类)、质心聚类(比如 K-means)、分布聚类(比如高斯混合模型聚类)、密度聚类(比如 DBSCAN)和网格聚类(比如 STING)。聚类在地球化学异常识别中有着大量应用(Kirkwood et al., 2016)。其中,K-means 是一种矢量量化广泛使用的无监督聚类方法,旨在将若干个观测样本划分为 K 个簇,其中每个样本属于簇质心与其距离最近的簇(Lloyd,

1982)。这将把样本的数据空间分离到 Voronoi 单元,并通过最小化样本与簇质心的距离平方实现聚类。K-means 方法度量的样本与簇质心的距离,而不是频率。进行 K-means 聚类时可以不用剔除所谓的"离群值",K-means 算法会自动按照距离将"离群值"归为一类。背景和异常在频率分布上具有相应的分布中心,因此,背景和异常的识别问题可以转换成样本与分布质心距离的问题,进而通过 K-means 算法实现。K-means 在地球化学中的使用通常为综合异常的识别(Zhou Shuguang et al., 2018; Ghezelbash et al., 2020)。

本文的研究目的是如何通过组分比值提取成矿过程的成分质量演化信息,利用对数比值转换方法优化浓度数据相对信息的结构表达,最后使用无监督学习 K-means 方法根据对数比值转换数据与分布空间质心的距离识别背景和异常信息,以阐明成分数据的信息由数据的内部结构承载。示范案例为安徽省兆吉口铅锌矿床土壤中 Pb 数据。文章首先介绍了成分数据理论对数比值、组分比值的地质意义和利用 K-means 原理识别背景和异常的原理,然后以兆吉口铅锌矿床的 Pb 元素土壤数据进行演示。

1 理论与方法

1.1 成分数据理论

假设某地质样品有 n 种组分,其所有组分的浓度值应该分布在 $S^n=\{X=[x_1,x_2,\cdots,x_n]|x_m>0,\ m=1,2,\cdots\cdots,n;\sum_{m=1}^{n}x_m=k\}$,其中,$k$ 为常量,一般为 1 或 100%,则 X 为成分数据。由于常合约束效应(即组分加和恒为 100%),成分数据分布在单纯形空间(simplex space),而不是整个欧式空间。常和约束只是信息的一种表达方式,承载了多个组分的相对于总体的信息。在原始信息不变的情况下,可以通过组分之间的比值表达具体的信息。比如 Pb 对样品总质量的质量贡献可以用 Pb 与非 Pb 物质之间的浓度比值表达。多数元素在地球中的分布近似服从对数正态分布(Ahrens, 1953, 1954a, 1954b),因此可以将组分比值进行对数运算来改善数据结构,以满足对成分数据的统计需求和提高信息的表达。这种方法称为对数比值(log-ratio)转换方法。目前主要有加性对数比值(additive log-ratio, alr)、中心对数比值(centred log-ratio, clr)、等距对数比值(isometric log-ratio, ilr; Aitchison, 1986; Egozcue et al., 2003)。三种方法存在各自的优点和劣势,需要根据研究的问题和相应的知识来选择相应的转换方法,以使得进行对数比值转换后表达的相对信息保持不变(McKinley et al., 2016)。对成分数据进行统计处理之前,需要进行适当的转换处理以优化其分布结构,以便将数据从单纯形空间投射到欧式空间,减少数据之间的扭曲纠缠程度。alr、clr 和 ilr 是目前的主流方法(Aitchison, 1986; Egozcue et al., 2003)。三种转换方法公式如下:

$$\mathrm{alr}(x_i)=\ln\frac{x_i}{x_n}(i=1,2,\cdots\cdots,n-1) \quad (1)$$

$$\mathrm{clr}(x_i)=\ln\frac{x_i}{g(X)} \quad (i=1,2,\cdots\cdots,n-1,n) \quad (2)$$

$$\mathrm{ilr}(X)=[<X,e_1>_a,<X,e_2>_a,\cdots\cdots,<X,e_{n-1}>_a] \quad (3)$$

$$\text{令 } e_i=\left[\exp\left(\frac{1}{i},\cdots\cdots,\frac{1}{i},-\frac{1}{j},\cdots\cdots,-\frac{1}{j}\right)\right]$$

(括号内有 i 个 $\frac{1}{i}$,j 个 $-\frac{1}{j}$) (4)

$$\text{则 } y_i=\sqrt{\frac{ij}{i+j}}\ln\left(\frac{g(x_1,x_2,\cdots\cdots,x_i)}{g(x_{i+1},\cdots\cdots,x_{i+j})}\right)$$

($i+j=n$, $i=1,2,\cdots\cdots,n-1$) (5)

其中,$g(X)$ 为向量的几何平均值。从公式可以看出,alr 转换表达的是数据内的任意 $n-1$ 个组分相对于第 n 个组分的比值信息,转换后的数据是不等距的。n 维成分数据经过 alr 转换后只能得到 $n-1$ 维数据。clr 转换表达的是数据内的所有组分相对于几何平均值的比值信息,转换后的数据是等形等距的。n 维成分数据经过 clr 转换后能得到 n 维数据。但由于 clr 相对数据中心进行转换,其转换结果会产生一个奇异的协方差矩阵。alr 和 clr 的变换结果没有与成分数据在单纯形空间的分布正交(Aitchison, 1986)。ilr 实际上是成分数据在正交坐标系的关联表达,其在 S^n 和 R^{n-1} 空间是等距的,由此避免了 alr 和 clr 两种转换的缺点。n 维成分数据经过 ilr 转换后能得到 $n-1$ 维数据,但是,在同一个正交基转换中,只有一个 ilr 转换数据能够与对应的单元素浓度数据直接地关联起来,其他的数据很难让人理解(Egozcue et al., 2003)。

1.2 组分比值的意义

假设地质作用发生前的地质系统是均匀的,任意成分 m 和 i 经过地质作用后的质量比值为:

$$\frac{M_m^{\mathrm{A}}}{M_i^{\mathrm{A}}}=\frac{M_m^{\mathrm{O}}+\Delta M_m}{M_i^{\mathrm{O}}+\Delta M_i}=\frac{w_m^{\mathrm{O}}+\Delta w_m}{w_i^{\mathrm{O}}+\Delta w_i}=\frac{w_m^{\mathrm{A}}}{w_i^{\mathrm{A}}} \quad (6)$$

其中,M 表示质量,下角标 i、m 表示组分,上角标 O 表示地质作用发生前的初始状态,A 表示地质地质作用发生后的结果状态,w 表示质量百分数。从公式(6)可以看出,经过地质作用后的组分比值由原岩的含量及其质量变化率两个因素决定。由于地质作用发生前的地质系统是均匀的,原岩含量可以视为常量,则地质作用发生后的组分比值仅由成分各自的质量变化率决定。

令 i 为地质过程中的不活动组分(比如 Zr、Hf、Ti),m 为任意组分,则:

$$\frac{M_m^A}{M_i^A}=\frac{M_m^O+\Delta M_m}{M_i^O}=\frac{1}{M_i^O}\Delta M_m+\frac{M_m^O}{M_i^O}=\frac{1}{w_i^O}\Delta w_m+\frac{w_m^O}{w_i^O}=\frac{w_m^A}{w_i^A} \tag{7}$$

公式(7)实际上是 Gant 方程(Grant, 1986)的变形,其表明任意组分 m 与不活动组分 i 的含量比值 $\frac{w_m^A}{w_i^A}$ 是关于 m 质量变化率 Δw_m 的线性函数,其斜率为不活动成分 i 的原岩含量的倒数 $\frac{1}{w_i^O}$,截距为初始状态时 m 与 i 的质量比值 $\frac{w_m^O}{w_i^O}$。Δw_m 与 $\frac{w_m^A}{w_i^A}$ 之间的图形关系如图 1a 所示。

当 m 质量不发生变化时,即 $\Delta w_m \to 0$ 时,

$$\frac{M_m^A}{M_i^A}=\frac{w_m^A}{w_i^A}=\lim_{\Delta w_m\to 0}\left(\frac{1}{w_i^O}\Delta w_m+\frac{w_m^O}{w_i^O}\right)=\frac{w_m^O}{w_i^O} \tag{8}$$

公式(8)表明,如果在地质过程中存在两种以上的组分不发生活动,则任意两种不活动组分的质量比值 $\frac{w_m^A}{w_i^A}$ 为常量,且等于原岩中的质量比值 $\frac{w_m^O}{w_i^O}$。

将公式(7)变形为:

$$w_m^A=\frac{\Delta w_m+w_m^O}{w_i^O}w_i^A \tag{9}$$

公式(9)表明经过地质作用后,任意组分 m 的含量是关于不活动组分 i 的含量的线性函数,其斜率为 $\frac{\Delta w_m+w_m^O}{w_i^O}$,其图像为一条经过原点的直线。对 w_i^A 和 w_m^A 进行投图,并连接该点和原点,其斜率差值 $\frac{\Delta w_m}{w_i^O}$ 可以反映出地质过程不同阶段中 m 的质量变化率的演化过程。如果 m 带入富集,则 $\Delta w_m>0$,在 w_i^A-w_m^A 图上为一组经过原点,斜率从 $\frac{w_m^O}{w_i^O}$ 不断增加的直线簇。增长的斜率反映了在地质过程不同阶段中 m 的质量富集情况。如果 m 带出亏损,则 $\Delta w_m<0$,在 w_i^A-w_m^A 图上为一组经过原点,斜率从 $\frac{w_m^O}{w_i^O}$ 不断降低的直线簇。亏损的斜率反映了在地质过程不同阶段中 m 的质量亏损情况。如果 m 不活动,则 $\Delta w_m=0$,公式(9)变为:

$$w_m^A=\frac{w_m^O}{w_i^O}w_i^A \tag{10}$$

公式(10)是公式(8)的变形,表明任意两种不活动组分在 w_i^A-w_m^A 图上的投点为一条经过原点的直线,斜率为原岩中 m 与 i 的质量比值 $\frac{w_m^O}{w_i^O}$。在实际研究中,可以通过系统采样的方法,通过对不活动组分 i 和其他组分 m 的 w_i^A-w_m^A 散点图(图 1b)研究组分质量的演化关系。

1.3　质量百分浓度、组分比值、对数浓度、对数组分比值的关系

浓度数据的信息解译需要依据问题—知识驱动的模式进行。元素的质量百分浓度数据是样本中某种组分 m 的质量分数,是经过地质作用后该物质的质量 $M_m^A=(M_m^O+\Delta M_m)$ 与样本总质量 $M^A=(M_m^O+\Delta M_m)+(M_m^O+\Delta M_{other})$ 的比值 w_m。质量百分浓度无量纲,表达的是组分 m 相对总质量的百分比贡献,而不是以 kg 为单位的绝对质量变化。w_m 的信息由成对的共轭数组$[w_m, w_{other}]$组成,一是 m 的初始质量及其演化信息,二是非 m 组分的初始质量及其演化信息。二者共同表达了地质作用后单位质量的样本中物质成分的质量百分比。w_m 与 w_{other} 呈负相关,即方向相反,分布形态关于点(0.5,0.5)对称。这种负相关,不是由 m 和非 m 的质量演变关系决定的,而是由于闭合作用导致的。为了去掉负相关性,可以将该信息转载到 m 的质量与非 m 组分的质量比值 $M_m/M_{other}=w_m/w_{other}=w_m/(1-w_m)$。$w_m$、$w_{other}$ 和 w_m/w_{other} 实际表达的信息是一致的,都是关于样品中 m 的质量贡献及其演化信息。w_m、w_{other} 是闭合的共轭数据,而 w_m/w_{other} 则是开放的自由数据。由公式 7 可知 $w_m^A/w_{other}^A=(w_m^O+\Delta w_m)/(w_{other}^O+\Delta w_{other})$,如果非 m 组分为活动组分($\Delta w_{other}\neq 0$),则 w_m^A/w_{other}^A 同时携带着物质 m 和非 m 物质的质量变化信息,而不仅仅是 m 的质量变化信息。因此,在多元成分数据中不是所有的组分比值都携带着与 w_m 相同的信息。组分比值需要根据具体研究的问题和背景知识进行选择。

与浓度 w_m 信息等同的组分比值是 $w_m/w_{other}=w_m/(1-w_m)$,其中,$w_m+w_{other}=1$。这是一个组分

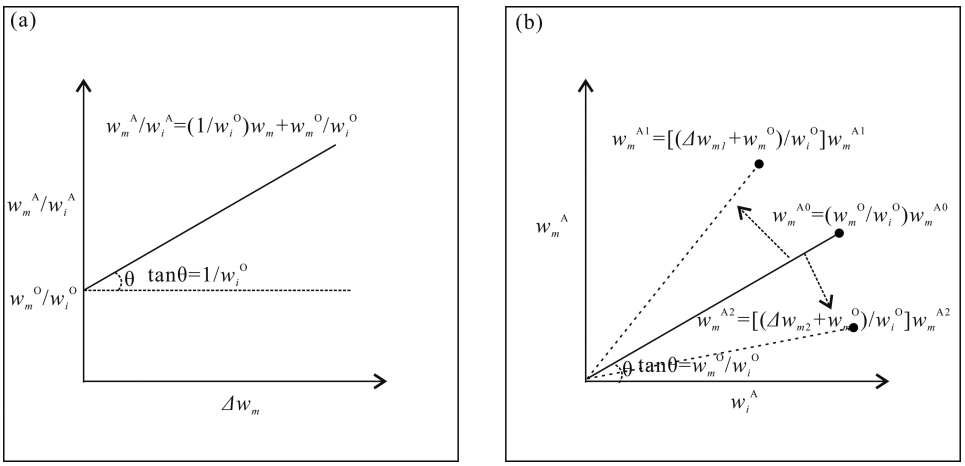

图 1 组分比值关系图

Fig. 1 Schematic of component ratios

(a)—Δw_m-$\frac{w_m^A}{w_i^A}$关系图；(b)—w_i^A-w_m^A 关系图

(a)—Plot of Δw_m-$\frac{w_m^A}{w_i^A}$ relationship; (b)—plot of w_i^A-w_m^A relationship

$n=2$ 的成分数据。此时，对应的对数浓度为 $\ln(w_m)$，将 $n=2$ 带入公式(1)、(2)、(5)进行对数比值转换，得到 $\mathrm{alr}(w_m)=\ln\left(\frac{w_m}{w_{\mathrm{other}}}\right)$，$\mathrm{clr}(w_m)=\ln\left(\frac{w_m}{\sqrt{w_m w_{\mathrm{other}}}}\right)=\frac{1}{2}\ln\left(\frac{w_m}{w_{\mathrm{other}}}\right)$，$\mathrm{ilr}(w_i)=\sqrt{\frac{1}{2}}\ln\left(\frac{w_m}{w_{\mathrm{other}}}\right)$。对数浓度可以看成 $\ln\left(\frac{w_m}{1}\right)$ 的形式，即物质 m 相对于整体样本的质量贡献的对数比值表达。$\mathrm{alr}(w_m)$、$\mathrm{clr}(w_m)$ 和 $\mathrm{ilr}(w_m)$ 都是关于 $\ln\left(\frac{w_m}{w_{\mathrm{other}}}\right)$ 的不同系数的表达，其数据结构是一样的。三种对数比值转换数据与对数浓度转换数据携带的信息是一致的，都是关于样本总质量中物质 m 和非 m 物质的经过相应的质量演化后的相对质量贡献。与非 m 物质的浓度 w_{other} 信息等同的比值是 w_{other}/w_m，其对数比值为 $\ln(w_{\mathrm{other}}/w_m)=-\ln(w_m/w_{\mathrm{other}})$。此时，物质 m 与非 m 物质之间的浓度关系清晰可见，即数据分布结构相同，方向相反。由此可见，对数比值转换数据的绝对值的数据结构表达了整个样品成分的质量演化关系，正负表示相对多寡。

1.4 K-means 聚类与背景异常识别

1.4.1 K-means 聚类原理

K-means 聚类方法可以将含有 N 个样本，每个样本有 P 个变量数据 X 划分为 K 类(C_1，C_2，……，C_k)，其中 C_k 表示簇 K 中 n_k 个样本的集合，K 是给定的。令 $X_{N\times P}=\{x_{ij}\}_{N\times P}$ 为 $N\times P$ 的数据矩阵，K-means 算法将对 $X_{N\times P}$ 进行分类，以使得每类中行向量(样本)与各自类的质心向量之间距离的平方最少与到其他类的质心向量的距离一样小。簇 C_k 的质心是 P 维空间的一个点，通过对簇内样本上的 P 个变量的值进行平均而得到。簇 C_k 的第 j 个变量的质心值为：

$$\bar{x}_j^{(k)}=\frac{1}{n_k}\sum_{i\in C_k}x_{ij} \tag{11}$$

完整的质心向量为：

$$\bar{X}^k=(\bar{x}_1^{(k)},\bar{x}_2^{(k)},\cdots\cdots,\bar{x}_P^{(k)})' \tag{12}$$

聚类是一个计算强度非常复杂的任务(Gentle，2002)，典型的 K-means 算法主要包括以下步骤：

(1)初始值 K 由随机 P 维向量 $S=(s_1^{(k)},s_2^{(k)},\cdots\cdots,s_P^{(k)})$ 决定，$1\leqslant k\leqslant K$。第 i 个样本和第 k 个中心的欧式距离平方为：

$$d^2(i,k)=\sum_{j=1}^{P}(x_{ij}-s_j^{(k)})^2 \tag{13}$$

样本按 $d^2(i,k)$ 值最小聚类。

(2)初始聚类后，由公式(12)计算簇质心，然后按(6)比较样本与每个簇质心的距离，并将样本归入距离最小的类。

(3)更新分类的成员，并重新计算簇质心(12)。

(4)重复步骤 2 和 3，直至各分类间无样本可以移动。

在聚类过程中，尽量使残差平方和最小，残差平方和为：

$$SSE=\sum_{j=1}^{P}\sum_{k=1}^{K}\sum_{i\in C_k}(x_{ij}-\bar{x}_j^{(k)})^2 \tag{14}$$

在 K 聚类初始化后,检查所有样本和质心的距离,如果簇 C_k 中的某个样本存在:

$$\frac{n_k}{n_k-1}d^2(i,k)>\frac{n_{k^*}}{n_{k^*}-1}d^2(i,k^*) \tag{15}$$

则将 C_k 中的第 i 个样本移动到 C_{k^*} 中,以减少 SSE 值(Späth, 1980)。SSE 值在 K-means 中通常叫做惯性(inertia)。惯性可以用来评价不同簇之间的连贯性,进而用来寻找最优的聚类数。

K-means 是一种稳健的聚类算法,其收敛结果会受到初始启动条件的影响。K-means 的初始启动方法主要有随机法(RANDOM)、Forgy 法、Macqueen 法和 Kaufman 法(MacQueen, 1967; Anderberg and MEX, 1973; Steinley, 2006; Kaufman and Rousseeuw, 2009)。随机法和 Kaufman 法效果优于其他方法,但随机法较 Kaufman 法更常用些。

肘方法(elbow method)是确定 K-means 聚类数目常用方法(Bholowalia and Kumar, 2014)。肘方法是一种依据方差解释百分比确定聚类数量的方法。其算法的思想原则是"当选择的聚类数目模型再次增加一个聚类时,不会提供更好的数据解释"。该方法将簇解释的数据量相对于簇数量绘制成折线图。通常第一个簇会解释很多信息,随着簇数量的增加,在某些时候边际增益会急剧下降并在图中形成一个像手肘一样的拐角。正确的簇数量(K 值)一般选择在该点,因此称为"肘部标准"(Marutho et al., 2018)。该方法从 $K=2$ 开始测试,并在每一步中不断增加 1,计算簇数和解释的数据量。在某个 K 值,新增加的数据解释量会急剧下降,然后当进一步增加 K 值时,数据解释量会达到一个平台期。该值则被认为是合理的 K 值。在此之后,尽管增加了簇的数量,但新增的簇非常接近现有的一些簇(Liu Fan and Deng Yong, 2021)。肘方法的主要步骤为:① 初始化赋值:$K=1$;② 启动;③ 增加 K 值:$K=K+1$;④ 衡量最优质量解决方案的数据解释量;⑤ 如果在某个时候解决方案的数据解释增量大幅下降;⑥ 这是真正的 K 值;⑦ 结束。

K-means 聚类的目标是使同一簇内的样本具有高相似性,不同簇间的样本具有低相似性。聚类表现优劣的评价方法主要有外部评估(external evaluation)、内部评估(internal evaluation)和聚类趋势(cluster tendency)。外部评估是指在真实标签已知的情况下对聚类结果的好坏进行评估,主要包括纯度(准确率)、精确率、召回率、兰德系数和 F 值等(Lever et al., 2016)。内部评估是指不需要借助于外部信息(比如真实标签),只依靠聚类结果和样本本身的属性来进行评估的方法。常见的内部评估方法有轮廓系数(Silhouette Coefficient)、Calinski-Harabasz Index 和 Davies-Bouldin 系数(Strehl and Ghosh, 2002)。聚类趋势是衡量待聚类数据中可能存在的聚类程度,通常在聚类操作之前用霍普金斯统计量(Hopkins Statistic)进行检验。外部评估需要知道真实标签,适用于监督学习。内部评估则适合无监督学习。轮廓是一种解释和验证数据集群内一致性的方法,该技术提供了一个简洁的图形表示每个对象的分类情况。轮廓系数将相同簇中元素的平均距离与其他簇中元素的平均距离进行对比,以衡量一个样本与其自己所属的簇(内聚)相比与其他簇(分离)的相似程度(Rousseeuw, 1987)。轮廓系数描述的簇内外差异与 K-means 方法的基本原理一致,因此本文选则轮廓系数对 K-means 聚类结果进行评价。

对任意样本 $i\in C_K$,定义 $a(i)=\dfrac{1}{|C_K|-1}\sum\limits_{j\in C_K,j\neq i}d(i,j)$ 为样本 i 到相同簇中其他样本的平均距离,以衡量把样本 i 分配到簇 K 的好坏。$a(i)$ 越小,表示分配的越好。其中,$|C_K|$ 为簇 K 中的样本数量,$d(i,j)$ 为样本 i 到同一簇中样本 j 的距离。定义 $b(i)=\min_{J\neq K}\dfrac{1}{|C_J|}\sum\limits_{j\in C_J}d(i,j)$ 为样本 i 到其他不同簇中所有样本的最小平均距离,以衡量样本 i 与其他簇的相异性(Aranganayagi and Thangavel, 2007)。据此,可以定义任意样本 i 的轮廓系数 $s(i)$ 为:

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}},|C_K|>1 \tag{16}$$

且 $s(i)=0,|C_K|=1$。

$s(i)$ 可以表示为:

$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases} \tag{17}$$

由公式(17)可知,轮廓系数的值域为[−1,+1],轮廓系数越接近 1,表示簇内样本之间紧凑,簇间距离大,聚类效果越好;轮廓系数越接近−1,表示聚类效果越差。

1.4.2 背景异常的分类问题

背景和异常的识别问题本质上是分类问题。分类是将相关样本分组到不同类中,以对样本进行识别、区分和理解的过程。主要分为二分类、多分类和等级分类。分类是机器学习的基础任务,已经被广泛应用在新闻分类、图片分类、视频分类、广告过滤、内容审核、推荐系统等领域。机器学习分类通过训练学习,建立一个从输入空间 X 到输出空间 Y(离散值)的映射(Goodman and Kruskal, 1979)。元素含量具有明显不同的分布模式,如果具有双重分布模式,则可以归纳为二分类问题,比如背景异常问题;如果具有多重分布模式,则可以归纳为多分类问题,比如岩石的地球化学分类、多重分形分布问题。机器学习分类的主要方法有贝叶斯网络、K-means、决策树、多层感知器、逻辑回归、随机森林、XGBoost、k-NN、Gradient Boost、SVM 等。K-means 方法能将样本分为 $K(K\geqslant 2)$ 个类别,既适合二分类,也适合多分类(Ahmed et al., 2020)。元素的含量分布如果服从简单的双重分布模式,则背景和异常识别问题可以看是二分类问题;如果服从多重分布模式,则可以看作是多分类问题。当是二分类问题时,则是一维数据 K-means 聚类的二分类应用特例,即 $X_{N\times 1}=\{x_{ij}\}_{N\times 1}$,$N$ 为样本数量,P 为元素数量($P=1$),$K=2$。公式(11)为质心计算公式,第 i 个样本与质心的距离为 $d^2(i,k)=(x_{ij}-s_j^{(k)})^2$。

2 研究案例

本文以安徽省兆吉口铅锌矿床地表 Pb 元素数据为案例,首先以最简单的二元成分数据展示了 Pb 浓度和非 Pb 物质浓度之间的伪相关性及如何使用成分数据理论提高元素的频率分布和信息表达,通过不活动元素 Zr 与 Pb 的散点图阐明 Pb 的质量演化关系,然后利用 K-means 聚类方法对转换后的数据进行质心和距离的计算,识别元素的背景和异常识别。最后将 K-means 方法识别的背景和异常与迭代 2 倍标准差法和分形方法识别背景和异常的结果进行比较,以评价 K-means 方法的性能。

2.1 地质概况

实验矿床位于安徽省东至县西南方向,处在长江中下游成矿带与皖南多金属成矿带之间。矿区出露地层主要为中元古代蓟县系溪口岩群环沙组(Pt_2h)、新元古代青白口系历口群葛公镇组(Qbg)、新生代第四系下蜀组(QP_3x),是一套由浅海-滨海相碎屑岩通过区域变质作用发展过来的低绿片岩相浅变质岩。岩性主要为变粉砂岩和变细砂岩,存在少量变粉砂质泥岩。矿区内发育断裂和褶皱构造。断裂构造主要为东至断裂及其次级裂隙带,褶皱构造主要有兆吉口倒转背斜和官港倒转背斜(图 2)。矿体主要以透镜体状赋存在东至断裂的次级裂隙中。破碎带内发育碎裂岩和构造角砾岩(乐成生等, 2011)。矿体与上盘葛公镇组(Qbg)和下盘环沙组(Pt_2h)的界线清晰,围岩蚀变不发育。矿石结构主要为半自形粒状结构和不等粒变晶镶嵌结构。矿石矿物主要为闪锌矿、方铅矿和黄铁矿,脉石矿物主要为石英和方解石(Liu Yanpeng et al., 2016, 2019)。

2.2 数据处理

实验数据为兆吉口铅锌矿床的 352 个土壤样品的 Pb 数据(刘艳鹏, 2017)。样品于 2012 年底在中国地质科学院地球物理地球化学勘查研究所中心实验室完成化学分析,Pb 元素分析方法为 X 射线荧光光谱法,检出限为 2 μg/g。样品采集和分析方法详见 Liu Yanpeng et al.(2019)。由于 Zr 基本保存在锆石中,其化学性质稳定,不受风化作用影响,本文选择 Zr 作为不活动元素研究 Pb 的质量变化关系。

首先根据 Pb 的浓度 w_{Pb},计算出非 Pb 物质的浓度 w_{other}、Pb 组分比值 w_{Pb}/w_{other}、对数浓度 $\ln(w_{Pb})$、对数比值 $\mathrm{alr}(w_{Pb})$、$\mathrm{clr}(w_{Pb})$ 和 $\mathrm{ilr}(w_{Pb})$ 等转换数据。然后对数据进行基本描述统计分析,以展示其基本的统计特征。统计的参量有平均值、标准差、中位数、截尾平均值、绝对中位差、最小值、最大值、极差、偏度、峰度和标准误差,并将相关统计结果以箱形图展示。为展示数据的频率分布特征,绘制了直方图和 Q-Q 图。绘制元素分布以展示 Pb 浓度 w_{Pb},非 Pb 物质浓度 w_{other}、组分比值 w_{Pb}/w_{other}、对数浓度 $\ln(w_{Pb})$、对数比值 $\mathrm{alr}(w_{Pb})$、$\mathrm{clr}(w_{Pb})$ 和 $\mathrm{ilr}(w_{Pb})$ 的空间分布特征。对 Zr 与 Pb 绘制散点图研究 Pb 的质量演化关系,然后利用 K-means 聚类方法对对数浓度转换和对数比值转换结果进行聚类分析,迭代 2 倍标准差法和浓度-面积分形方法对提取背景和异常信息,绘制异常分布图展示将 K-means 提取的异常结果与迭代 2 倍标准差法和浓度-面积分形得到的异常信息进行对比,评价 K-means 识别背景和异常的效果。

3 结果

3.1 描述统计

Pb 浓度 w_{Pb}、非 Pb 物质浓度 w_{other}、组分比值

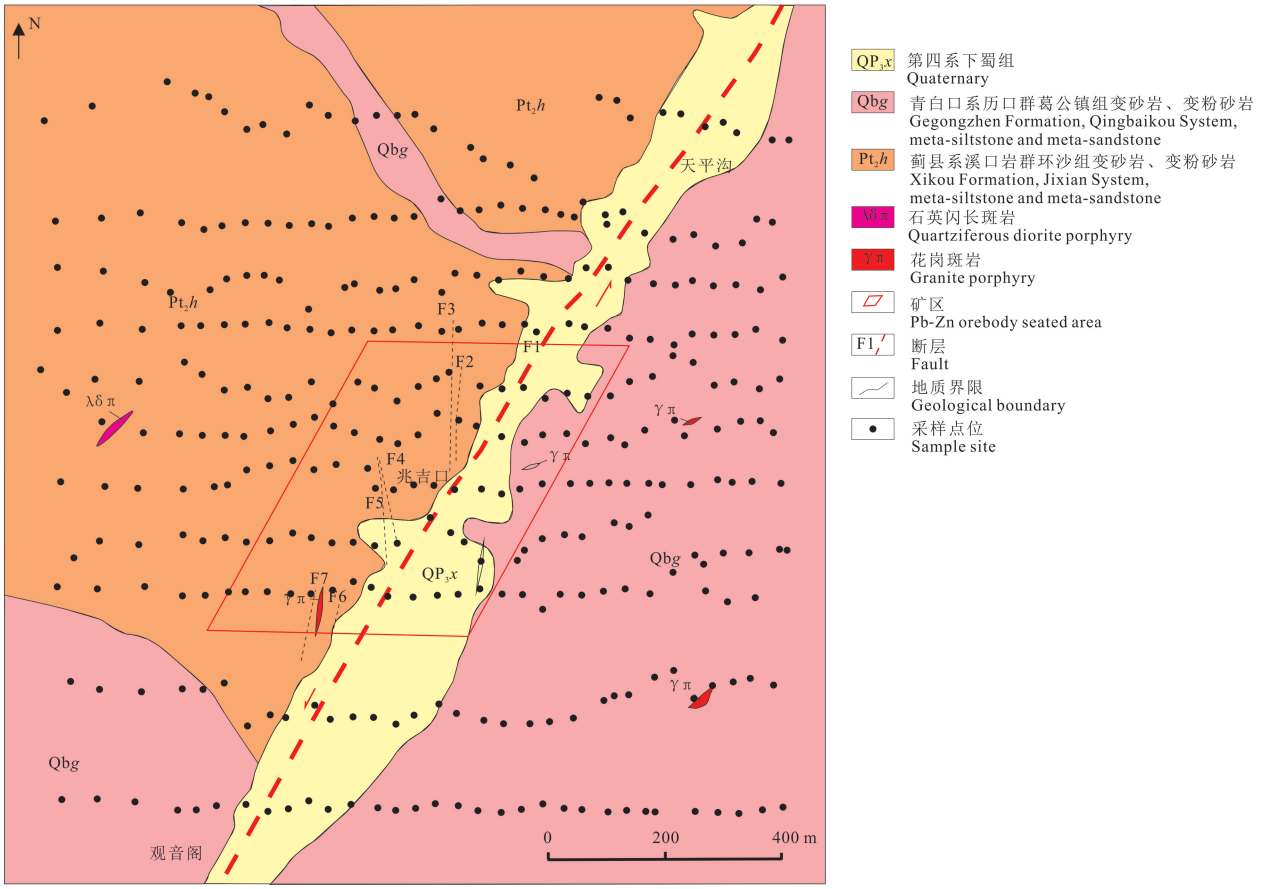

图 2　安徽省兆吉口铅锌矿床地质概况和采样位置图(据乐成生等,2011;Liu Yanpeng et al.,2016,2019 修改)

Fig. 2　Map of geologic schematic and sampling locations of the Zhaojikou Pb-Zn ore deposit, Anhui Province (modified after Le Chengsheng et al., 2011; Liu Yanpeng et al.,2016,2019)

表 1　安徽省兆吉口铅锌矿床 Pb 浓度及其转换数据描述统计表

Table 1　Descriptive statistics of Pb concentration and corresponding transformation data in the Zhaojikou Pb-Zn ore deposit, Anhui Province

项目	w_{Pb}	w_{others}	w_{Pb}/w_{others}	$\ln(w_{Pb})$	$alr(w_{Pb})$	$clr(w_{Pb})$	$ilr(w_{Pb})$
平均值	79.31×10^{-6}	999920.7×10^{-6}	7.94×10^{-5}	3.67	−10.15	−5.07	−7.18
标准差	268×10^{-6}	268×10^{-6}	2.69×10^{-5}	0.77	0.77	0.39	0.55
中位数	30.91×10^{-6}	999969.1×10^{-6}	3.09×10^{-5}	3.43	−10.38	−5.19	−7.34
截尾平均值	34.56×10^{-6}	999965.4×10^{-6}	3.46×10^{-5}	3.5	−10.31	−5.16	−7.29
绝对中位差	8.44×10^{-6}	8.44×10^{-6}	0.85×10^{-5}	0.27	0.27	0.14	0.19
最小值	15.03×10^{-6}	996675×10^{-6}	1.50×10^{-5}	2.71	−11.11	−5.55	−7.85
最大值	3325×10^{-6}	999985×10^{-6}	334×10^{-5}	8.11	−5.70	−2.85	−4.03
极差	3309.97×10^{-6}	3309.97×10^{-6}	332×10^{-5}	5.4	5.4	2.7	3.82
偏度	9.1	−9.1	9.11	2.88	2.88	2.88	2.88
峰度	92.9	92.9	93.08	9.78	9.79	9.79	9.79
标准误差	14.28×10^{-6}	14.28×10^{-6}	1.43×10^{-5}	0.04	0.04	0.02	0.03

w_{Pb}/w_{other}、对数浓度 $\ln(w_{Pb})$和对数比值 $alr(w_{Pb})$、$clr(w_{Pb})$和 $ilr(w_{Pb})$转换数据的平均值、标准偏差、中位数、截尾均值、绝对中位差、最小值、最大值、偏度、峰度和标准误差列于表 1 中,并在箱形图(图 3)中展示。从表 1 和图 3 可以看出,Pb 的浓度 w_{Pb} 和非 Pb 物质浓度 w_{other} 的标准偏差、绝对中位差、极差是一样的,只是集中分布的方向相反。组分比值 w_{Pb}/w_{other} 与 w_{Pb} 基本相同。对数浓度和对数比值转换数据的离散程度远远小于浓度 w_{Pb}。其频率分布特征展示在直方图(图 4)和 Q-Q 图(图 5)上。由图 4 可以看出,浓度数据的频率分布过于分散,没有展示出较好的分布特征,而经过转换后的对数浓度

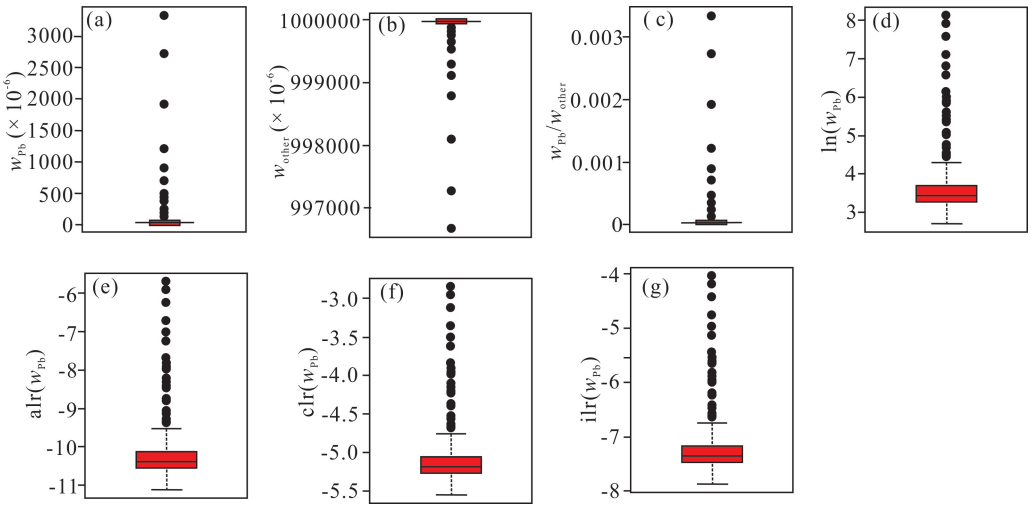

图 3 安徽省兆吉口铅锌矿床 Pb 浓度和对应转换数据箱形图

Fig. 3 Boxplots of Pb concentration and its corresponding transformation data in the Zhaojikou Pb-Zn ore deposit, Anhui Province

(a)—Pb 浓度箱形图；(b)—非 Pb 物质浓度箱形图；(c)—Pb 比值箱形图；(d)—对数 Pb 浓度箱形图；(e)—alr 转换数据箱形图；(f)—clr 转换数据箱形图；(g)—ilr 转换数据箱形图

(a)—Boxplot of Pb concentration; (b)—boxplot of non Pb material concentration; (c)—boxplot of the ratio of Pb to other material; (d)—boxplot of logarithm of Pb concentration; (e)—boxplot of alr transformation data; (f)—boxplot of clr transformation data; (g)—boxplot of ilr transformation data

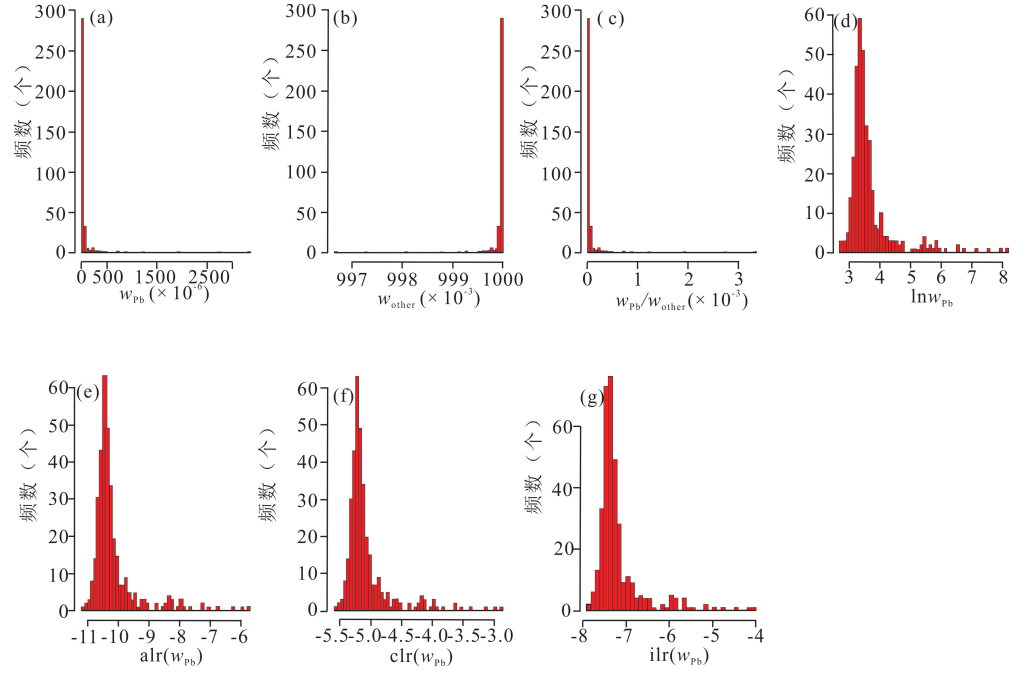

图 4 安徽省兆吉口铅锌矿床 Pb 浓度及对应转换数据直方图

Fig. 4 Histograms of Pb concentration and corresponding transformation data of the Zhaojikou Pb-Zn ore deposit, Anhui Province

(a)—Pb 浓度直方图；(b)—非 Pb 物质浓度直方图；(c)—Pb 比值直方图；(d)—对数 Pb 浓度直方图；(e)—alr 转换数据直方图；(f)—clr 转换数据直方图；(g)—ilr 转换数据直方图

(a)—Histogram of Pb concentration; (b)—histogram of non Pb material concentration; (c)—histogram of the ratio of Pb to other material; (d)—histogram of logarithm of Pb concentration; (e)—histogram of alr transformation data; (f)—histogram of clr transformation data; (g)—histogram of ilr transformation data

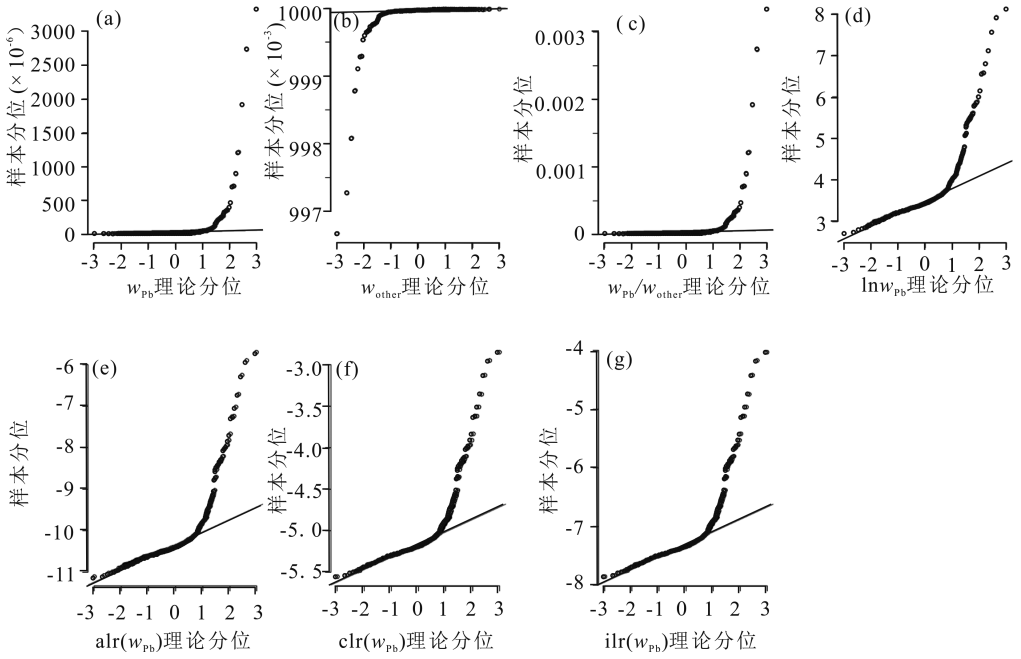

图 5　安徽省兆吉口铅锌矿床 Pb 浓度及对应转换数据 Q-Q 图

Fig. 5　Quantile-quantile (Q-Q) plots of Pb concentration and corresponding transformation data of the Zhaojikou Pb-Zn ore deposit, Anhui Province

(a)—Pb 浓度 Q-Q 图;(b)—非 Pb 物质浓度 Q-Q 图;(c)—Pb 比值 Q-Q 图;(d)—对数 Pb 浓度 Q-Q 图;(e)—alr 转换数据 Q-Q 图;(f)—clr 转换数据 Q-Q 图;(g)—ilr 转换数据 Q-Q 图

(a)—Q-Q plot of Pb concentration; (b)—Q-Q plot of non Pb material concentration; (c)—Q-Q plot of the ratio of Pb to other material; (d)—Q-Q plot of logarithm of Pb concentration; (e)—Q-Q plot of alr transformation data; (f)—Q-Q plot of clr transformation data; (g)—Q-Q plot of ilr transformation data

ln(w_{Pb})和对数比值 alr(w_{Pb})、clr(w_{Pb})和 ilr(w_{Pb})数据的分布特征比较清楚,呈现出“单峰”和“右偏”的特征。表 1 结果表明,ln(w_{Pb})和 alr(w_{Pb})、clr(w_{Pb})和 ilr(w_{Pb})的峰度和偏度是一致。图 4 和图 5 清楚地显示 w_{Pb}、ln(w_{Pb})、alr(w_{Pb})、clr(w_{Pb})和 ilr(w_{Pb})数据基本可以分为两种数据群落。

3.2　w_{Pb}/w_{Zr} 表达 Pb 的质量演化关系

Pb 在成矿过程中的质量演化关系可以由 w_{Pb}/w_{Zr} 表达。w_{Pb}/w_{Zr} 可由 Zr-Pb 的散点图表达(图 5),公式(9)表明,在原岩含量 w_{Pb}^{O} 和 w_{Zr}^{O} 不变的情况下,各点与原点连线的斜率 $\frac{\Delta w_{Pb}+w_{Pb}^{O}}{w_{Zr}^{O}}$ 反映了 Δw_{Pb} 的相对变化。图 6 结果表明,Pb 在成矿作用下大致可以分为两类,一类与原点连线的斜率小,近乎平行于 x 轴,表明其质量迁移率接近于 0;另一类与原点的联系斜率大,与 x 轴斜交,其质量迁移率明显大于 0。两类样本之间的连续过渡。

3.3　w_{Pb}、w_{other} 和 w_{Pb}/w_{other} 对数比值与对数浓度的信息

为展示元素 Pb 的浓度 w_{Pb}、非 Pb 物质的浓度 w_{other} 和浓度比值 w_{Pb}/w_{other} 的空间结构和信息,绘制了 w_{Pb}、w_{other} 和 w_{Pb}/w_{other} 的空间分布图(图 7a～c)。从图 7a～c 中可以看出,w_{Pb} 和 w_{Pb}/w_{other} 数据所代表的空间分布信息相同,均表达了 Pb 在矿区位置含量相对较高,在周边相对较低的分布特点。w_{other} 所代表的空间分布信息的方向与 w_{Pb}、w_{Pb}/w_{other} 相反,即非 Pb 物质在矿区位置含量相对较低,在周围相对较高的特点。三种数据表达的信息是一致的。为展示对数比值与对数浓度之间的信息关系,绘制了对数浓度、对数比值转换数据的空间分布图(图 7d～g)。由图 7d～g 可以看出,ln(w_{Pb})、alr(w_{Pb})、clr(w_{Pb})和 ilr(w_{Pb})数据的空间结构信息相同,均表达了 Pb 在矿区中心的含量相对较高、周边含量相对较低的特点。对数数据(图 7d～g)从矿区周边到矿区中心,Pb 含量的分布变化是渐进的,反映了 Pb 的质量演化关系的连续性,与图 6 一致,而浓度数据 w_{Pb} 的则是突变的(图 7a)。

3.4　背景与异常

对转换化后的对数浓度 ln(w_{Pb})、对数比值 alr(w_{Pb})、clr(w_{Pb})和 ilr(w_{Pb})进行 K-means 聚类识

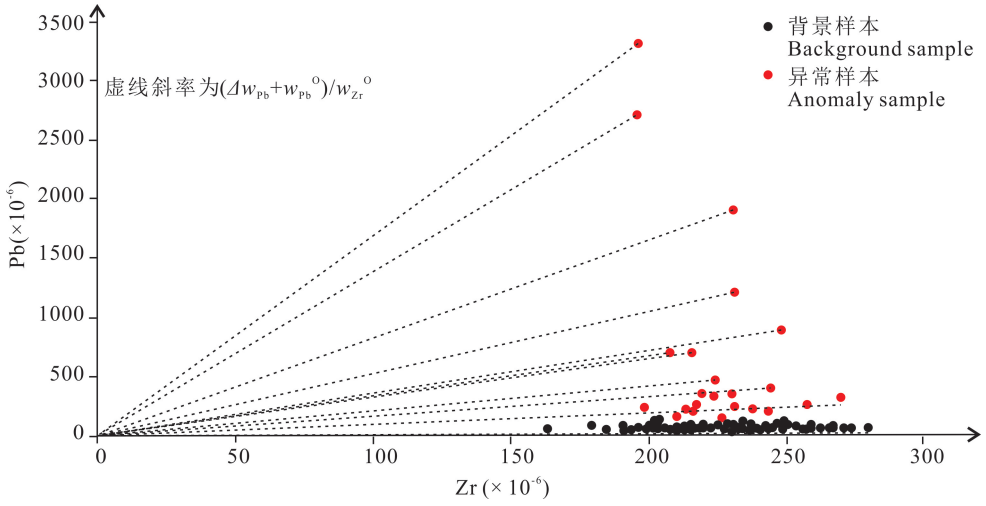

图 6 安徽省兆吉口铅锌矿床 Zr-Pb 的质量演化关系散点图

Fig. 6 Scatter plot of Zr-Pb showing Pb's mass involution route in the Zhaojikou Pb-Zn ore deposit, Anhui Province

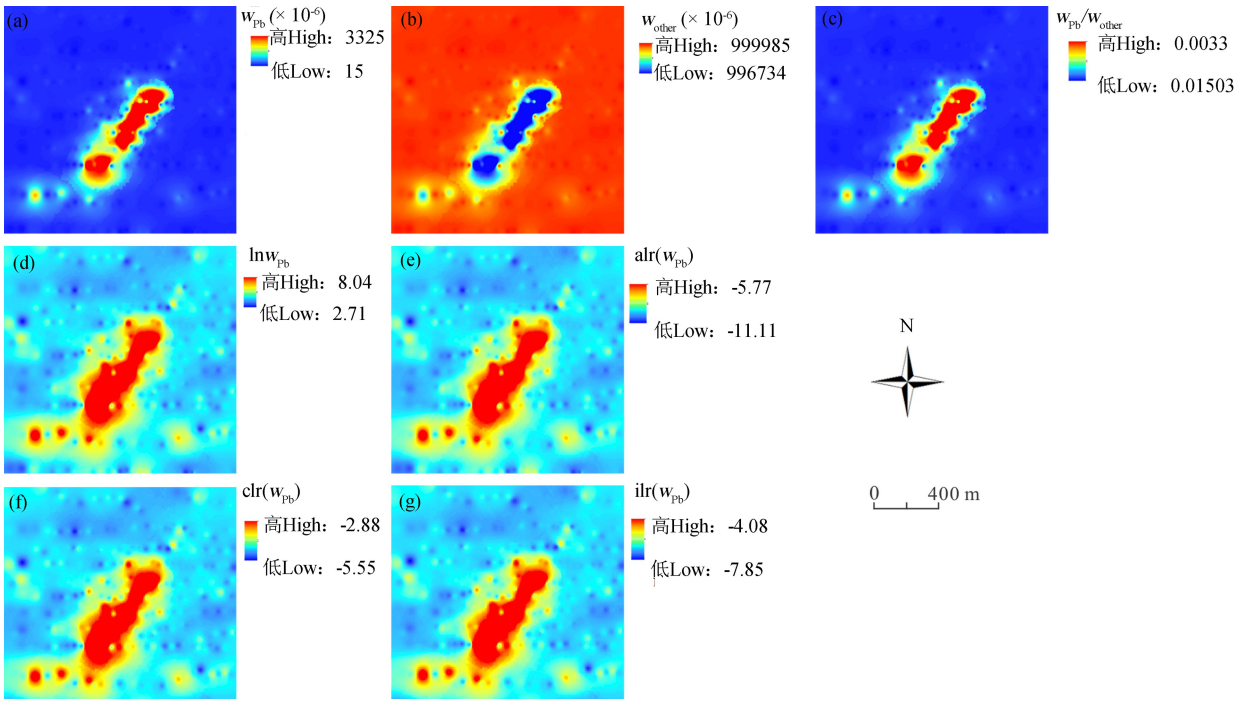

图 7 安徽省兆吉口铅锌矿床 Pb 浓度及对应转换数据空间分布图

Fig. 7 Spatial distribution maps of Pb concentration and corresponding transformation data of the Zhaojikou Pb-Zn ore deposit, Anhui Province

(a)—Pb 浓度分布图;(b)—非 Pb 物质浓度分布图;(c)—Pb 比值分布图;(d)—对数 Pb 浓度分布图;(e)—alr 转换数据分布图;(f)—clr 转换数据分布图;(g)—ilr 转换数据分布图

(a)—Distribution map of Pb concentration; (b)—distribution map of non Pb material concentration; (c)—distribution map of the ratio of Pb to other material; (d)—distribution map of logarithm of Pb concentration; (e)—distribution map of alr transformation data; (f)—distribution map of clr transformation data; (g)—distribution map of ilr transformation data

别背景和异常信息,结果如图 8、图 9 和表 2 所示。从肘方法聚类数图(图 8)和轮廓系数(表 2)中可以看出,元素 Pb 的 ln(w_{Pb})、alr(w_{Pb})、clr(w_{Pb})和 ilr(w_{Pb})转换数据最优分类方案均是 2 类。对这四种转换数据进行聚类数为 2 的 K-means 聚类分析,将 K-means 分类的临界值作为异常下限,分类结果的空间位置关系见图 9,聚类结果的质心统计在表 3 中。由图 9 可以看出,K-means 聚类的样本位置与

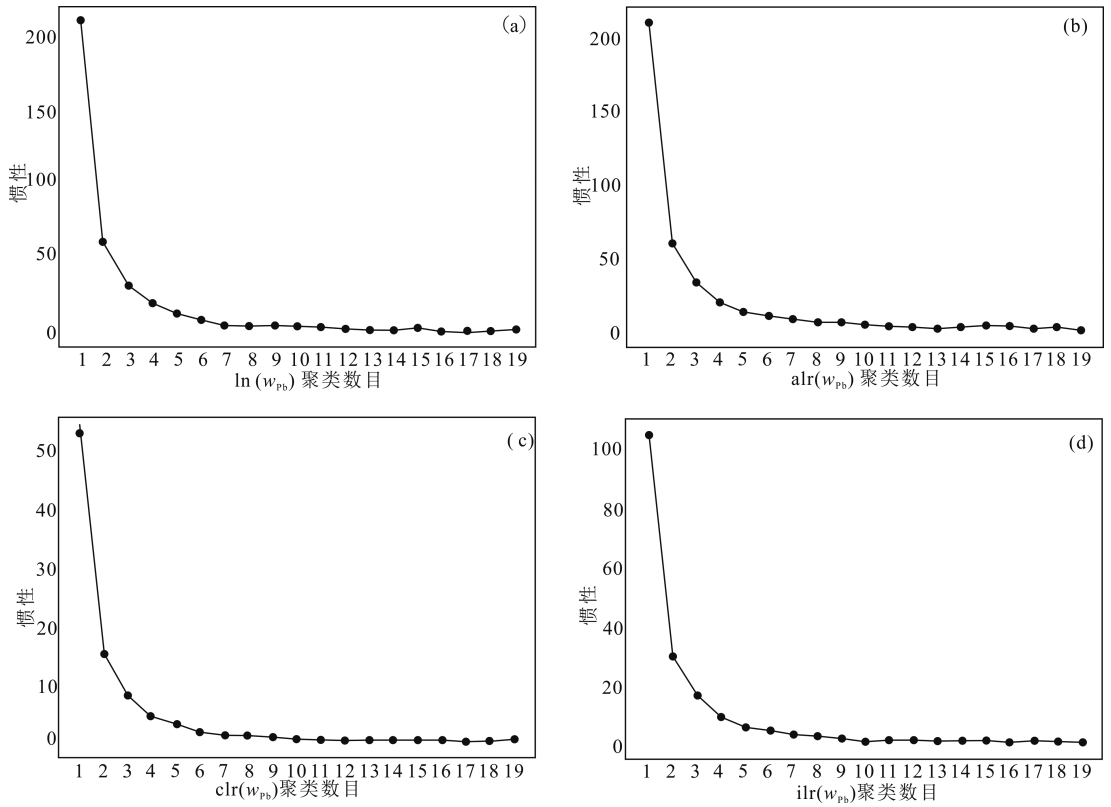

图 8　安徽省兆吉口铅锌矿床 K-means 聚类肘方法图

Fig. 8　Elbow method plots of K-means clustering in the Zhaojikou Pb-Zn ore deposit, Anhui Province

(a)—对数 Pb 浓度肘方法图;(b)—alr 转换数据肘方法图;(c)—clr 转换数据肘方法图;(d)—ilr 转换数据肘方法图

(a)—Elbow method plot of logarithm Pb concentration; (b)—elbow method plot of alr transformation data; (c)—elbow method plot of clr transformation data; (d)—elbow method plot of logarithm of ilr transformation data

表 2　安徽省兆吉口铅锌矿床 Pb 元素 K-means 分类轮廓系数表

Table 2　Silhouette coefficients of logarithm Pb concentration and logarithm ratios in the Zhaojikou Pb-Zn ore deposit, Anhui Province

数据类型	聚类数目					
	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
$\ln w_{Pb}$	0.821	0.672	0.644	0.529	0.527	0.525
alr(w_{Pb})	0.821	0.672	0.644	0.529	0.525	0.525
clr(w_{Pb})	0.821	0.672	0.647	0.529	0.525	0.524
ilr(w_{Pb})	0.821	0.672	0.644	0.529	0.527	0.525

表 3　安徽省兆吉口铅锌矿床 Pb 元素 K-means 分类质心表($K=2$)

Table 3　Centroids of K-means classification ($K=2$) of Pb in the Zhaojikou Pb-Zn ore deposit, Anhui Province

项目	$\ln(w_{Pb})$	clr(w_{Pb})	clr(w_{Pb})	ilr(w_{Pb})
质心 1	3.48	−10.34	−5.17	−5.57
质心 2	5.94	−7.87	−3.96	−7.31

矿体就位空间具有高度耦合性。同时利用迭代 2 倍标准差法和浓度-面积分形方法对浓度数据 w_{Pb} 提取异常,并将与 K-means 方法的结果进行对比。迭代 2 倍标准差法计算得到的异常下限值为 38.59×10^{-6},浓度-面积分形方法得到的异常下限为 120.66×10^{-6}(图 10)。K-means 聚类、迭代 2 倍标准差法和浓度-面积分形分析方法得到的异常下限统计在表 4 中。从表 4 中可以看出,K-means 聚类得到的异常下限对应的浓度数据和浓度-面积分形分析方法的结果相似,均在 $w_{Pb}=120\times10^{-6}$ 左右,而迭代 2 倍标准差法得到的异常下限值为 38.59×10^{-6},明显偏低。K-means 聚类与浓度-面积分形方法的共同点是对数据进行了对数处理,从而提取了“元素在空间上服从对数分布”的规律信息,而迭代 2 倍标准差法仅从频率角度分析,不能有效识别对数规律信息。根据表 4 绘制了 Pb 元素的异常图(图 11)。其中,图 11a～d 为对数浓度 $\ln(w_{Pb})$、对数比值 alr(w_{Pb})、clr(w_{Pb})、ilr(w_{Pb})等对数数据的异常图,图 11e 为浓度-面积分形方法 $w_{Pb}=120\times10^{-6}$ 时的异常图,图 11f 为迭代 2 倍标准差法 $w_{Pb}=38\times10^{-6}$ 时的异常图。从图 11 中可以看出,迭代 2 倍标准差法的异常面积分布最广,富含的噪音

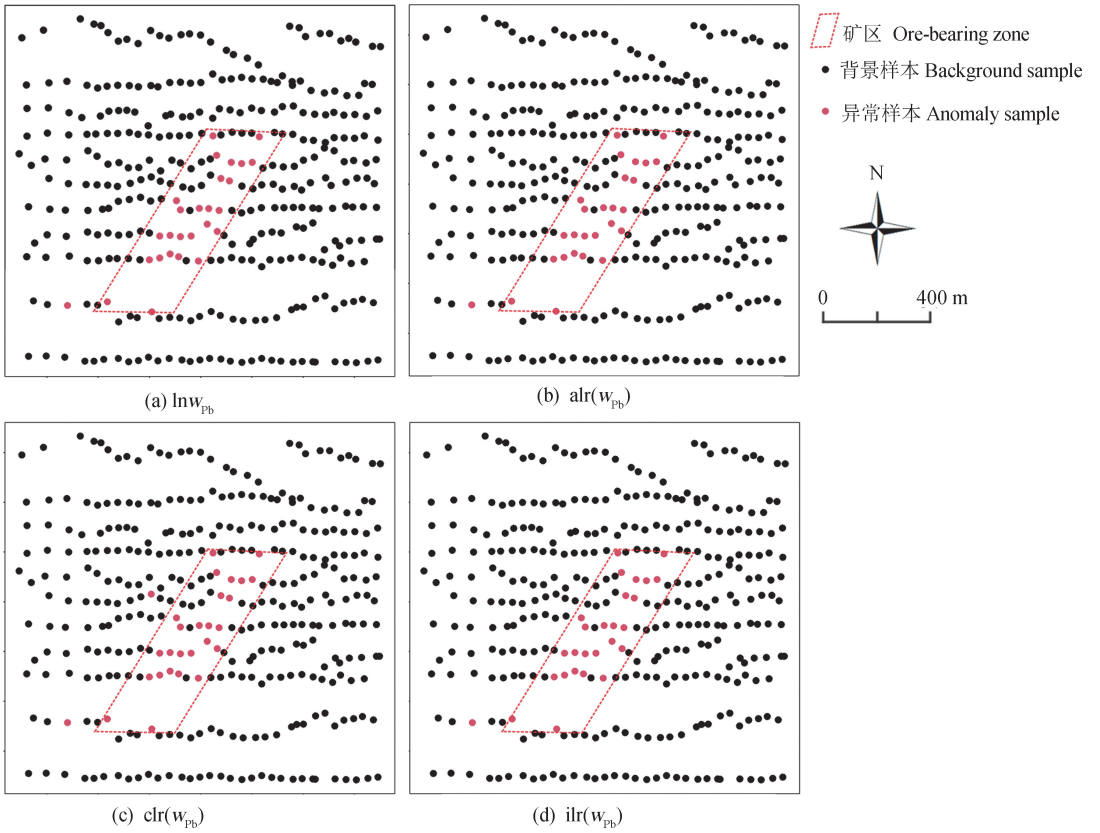

图 9 安徽省兆吉口铅锌矿床 Pb 对数浓度(a)、alr 数据(b)、clr 数据(c)和 ilr 数据(d) K-means 背景和异常分类与已知矿耦合图

Fig. 9 Locations maps of K-means cluster samples on logarithmic Pb concentration (a), alr data (b), clr data (c), and ilr data (d) in the Zhaojikou Pb-Zn ore deposit, Anhui Province

表 4 兆吉口地区 K-means、迭代二倍标准差法、浓度-面积分形方法异常下限表

Table 4 Thresholds of K-means, iterative 2δ method, and the concentration-area fractal analysis in the Zhaojikou Pb-Zn ore deposit, Anhui Province

项目	$\ln(w_{Pb})$	$\mathrm{alr}(w_{Pb})$	$\mathrm{clr}(w_{Pb})$	$\mathrm{ilr}(w_{Pb})$	迭代 2 倍标准差	C-A 分形
K-means 异常下限	4.76	−9.05	−4.56	−6.40	-	-
对应浓度($\times 10^{-6}$)	117.23	117.23	110.01	117.23	38.59	120.66

注:"-"表示无此项值。

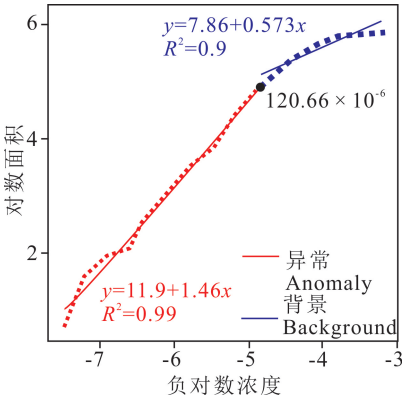

图 10 安徽省兆吉口铅锌矿床 Pb 浓度数据面积-浓度分形图

Fig. 10 Plot of area-concentration fractal analysis on Pb data in the Zhaojikou Pb-Zn ore deposit, Anhui Province

最多,而浓度-面积分形分析和 K-means 聚类方法识别的异常信号较少,基本都与矿体就位空间对应。对数浓度和对数比值数据的异常图比浓度数据的异常图噪音最少,对数转换数据的信噪比最高。

4 讨论

4.1 组分比值携带质量演化信息

组分比值携带着组分的质量演化信息。公式(6)表明经过地质作用后的组分比值由原岩的组分含量及其质量变化率两个因素决定。单个元素的浓度数据是成分数据,携带元素与整体质量变化的相对变化信息,而不是绝对质量变化信息。公式(9)表明元素浓度携带的信息由单位质量的样本中元素的

原岩浓度、元素质量变化率和总体质量变化率共同决定。比如 Pb 经过地质作用后的浓度数据反映的是元素 Pb 的原始质量、Pb 在地质作用中的质量变化与系统总体质量及其变化的相对比值,而不是以 kg 为单位的绝对质量的多少。这种变量被约束在单纯形空间内分布。总体质量变化是所有组分共同活动的结果,闭合操作使得组分之间具有伪相关性。比如尽管没有任何理论表明在成矿作用中 Pb 质量的增加一定会伴随非 Pb 物质的质量亏损,但在图 7 中 w_{Pb} 和 w_{other} 的空间浓度分布呈明显的"此高彼低"的负相关。公式(6)表明组分比值可以消除总体质量变化的约束,进而展示组分之间的相对质量变化率。如果构成组分比值分母的元素为不活动元素,则可反映分子组分的绝对质量变化率。公式(7)表达了任意组分 m 与不活动组分 i 的含量比值 w/w_i 是关于 m 质量变化率 Δw_m 的线性函数,其斜率为不活动成分 i 的原岩含量的倒数 $\frac{1}{w_i^{\mathrm{O}}}$,截距为初始状态时 m 与 i 的质量比值 $\frac{w_m^{\mathrm{O}}}{w_i^{\mathrm{O}}}$。在地质研究中,常常对不同样品的浓度数据进行对比,通过浓度的高低推断质量的多寡。这种分析推断隐含着一个被忽略的假设前提,即样品所代表地质系统的体积和质量在地质过程中保持不变。然而,大量的研究(Helgeson et al., 1970; Grant, 1986; Bohrson and Spera, 2001; Bjørlykke and Jahren, 2012; 马生明和朱立新, 2014; 马生明等, 2016)表明地质系统是开放的,在地质过程中(比如岩浆分异、成矿作用、风化作用),系统质量和体积是变化的,而不是恒定不变的。因此,在系统质量不守恒的情况下,不同样品中元素的浓度变化不能代表质量的变化。在对浓度数据进行分析时,需要弄清楚其所代表的相对信息含义。

4.2　对数组分比值可以优化浓度数据的信息表达

元素分布服从对数正态分布,经过对数转换后的对数浓度 $\ln(w_{\mathrm{Pb}})$、对数组分比值 $\mathrm{alr}(w_{\mathrm{Pb}})$、$\mathrm{clr}(w_{\mathrm{Pb}})$、$\mathrm{ilr}(w_{\mathrm{Pb}})$数据则具有较清晰的结构,更能反映对数尺度的规律。从元素浓度分布图(图 7)中可以看出,浓度数据 w_{Pb} 反映的 Pb 分布规律和 $\mathrm{alr}(w_{\mathrm{Pb}})$、$\mathrm{clr}(w_{\mathrm{Pb}})$、$\mathrm{ilr}(w_{\mathrm{Pb}})$大体相似,但存在一定程度的不同。它们都反映出 Pb 在矿体就位空间浓度相对升高的特点,不同之处在于浓度数据 w_{Pb} 的空间分布从低浓度到高浓度是突变的,即仅在断裂带的含矿位置呈现出高值,而在两侧的青白口系和蓟县系突兀地降成低值,这种现象,隐含着成矿作用只在矿体就位空间对 Pb 的分布有影响,而对矿体外部的空间没有影响或影响极小的可能性。而 $\mathrm{alr}(w_{\mathrm{Pb}})$、$\mathrm{clr}(w_{\mathrm{Pb}})$、$\mathrm{ilr}(w_{\mathrm{Pb}})$反映出 Pb 在整个研究区的分布具有渐变特征,隐含着成矿作用对整个矿区的 Pb 分布都有影响的可能性。这两种可能,后者更加符合经验认知。这种"突变分布"和"渐变分布"的区别表明对数组分比值转换数据要比浓度数据更能表达成矿作用对元素分布的影响。公式(1)、(2)、(5)表明,在二元成分数据($n=2$)中,$\mathrm{alr}(w_{\mathrm{m}})$、$\mathrm{clr}(w_{\mathrm{m}})$、$\mathrm{ilr}(w_{\mathrm{m}})$三者是关于 $\ln\left(\frac{w_m}{1-w_m}\right)$ 不同系数的表达,其数据结构是一致的。因此在单元素信息表达方面,$\mathrm{alr}(w_m)$、$\mathrm{clr}(w_m)$、$\mathrm{ilr}(w_m)$并无区别。表 1 中可以看出,尽管 $\mathrm{alr}(w_m)$、$\mathrm{clr}(w_m)$、$\mathrm{ilr}(w_m)$三者的极差、平均值等有区别,但其峰度和偏度是一致的。图 3～图 5 中也可以看到,尽管 $\mathrm{alr}(w_m)$、$\mathrm{clr}(w_m)$、$\mathrm{ilr}(w_m)$的值不同,但其形状及携带的信息是一致的。对数组分比值数据携带的相对信息都是从浓度数据中继承过来的,反映了在成矿过程中 Pb 元素质量相对于其他非 Pb 物质的原始含量及浓度变化的信息。这种变化信息是尺度不变和子成分连续的(Aitchison, 1986),表现在 w_{Pb} 和 $w_{\mathrm{Pb}}/w_{\mathrm{other}}$,$\ln(w_{\mathrm{Pb}})$、$\mathrm{alr}(w_m)$、$\mathrm{clr}(w_m)$和 $\mathrm{ilr}(w_m)$代表的信息是一致的。与原始浓度数据相比较,$\mathrm{alr}(w_m)$、$\mathrm{clr}(w_m)$、$\mathrm{ilr}(w_m)$数据去掉了负相关性,具有和对数浓度相同的数据结构,表达的信息更加符合对数尺度的规律。

4.3　K-means 聚类方法可以有效识别背景和异常

地球化学背景是元素在地球分布的正常丰度,其地质意义是反映元素在成矿作用发生前的正常分布。异常是对正常的偏离,即在给定尺度的地球环境中元素相对正常分布模式的偏离,其地质意义是反映成矿作用发生后元素的分布(Hawkes and Webb, 1963)。背景和异常与尺度有关,把地球化学背景的取值视为一个分布范围要比视为一个具体的值更加合理,所以背景通常会以"均值±n 倍标准差"的形式表达(Matschullat et al., 2000; Reimann and Garrett, 2005)。K-means 方法将数据分配到不同的簇中,计算各个簇的"质心"及每个样本到质心的"距离",按照距离的大小,将样本分配到最近的类中。通过循环以上步骤,直至找到分配到类别的数据点不会变化的理想质心。计算"质心"的过程可以理解为寻找元素在不同地质作用下的分

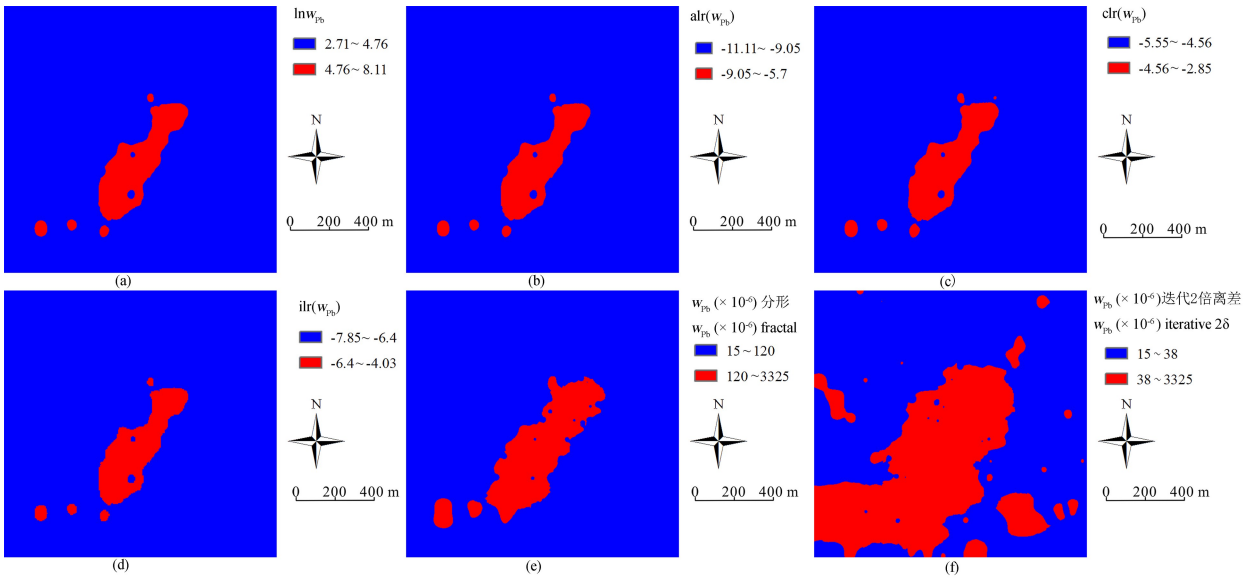

图 11 安徽省兆吉口铅锌矿床 Pb 元素 K-means、分形和迭代 2 倍标准差法异常图

Fig. 11 Maps of anomaly distributions of Pb concentration based on K-means, fractal analysis, and iterative 2δ method in the Zhaojikou Pb-Zn ore deposit, Anhui Province

(a)—对数 Pb 浓度 K-means 异常图；(b)—alr 数据 K-means 异常图；(c)—clr 数据异常图；(d)—ilr 数据 K-means 异常图；(e)—Pb 浓度面积-浓度分形异常图；(f)—Pb 浓度迭代 2 倍标准差法异常图

(a)—K-means anomaly map of logarithmic Pb concentration; (b)—K-means anomaly map of alr data; (c)—K-means anomaly map of clr data; (d)—K-means anomaly map of ilr data; (e)—area-concentration anomaly map of Pb concentration; (f)—iterative 2δ method anomaly map of Pb concentration

布“中心”,“距离”则是衡量元素分布受作用的影响大小。图 8 和表 2 结果显示,Pb 元素最佳的聚类数为 2,表明 Pb 拥有 2 个分布“质心”,主要由两种地质作用控制。这两个“质心”,实际上代表着 Pb 的浓度数据可以分为“背景”和“异常”两个沃罗诺伊原胞。在每个原胞内部的数据到质心的距离是最近的,离对方的“质心”要比离内部“质心”的距离都要远。表 4 表明了 K-means 聚类识别的异常效果和浓度-面积分形分析方法一致。从图 2、图 9 和图 11 可以看出,K-means 识别出的“异常”样品与矿体就位空间吻合度高,且异常分布的噪音少,信噪比高。

5 结论

本文通过成分数据理论对元素的质量百分浓度数据所携带的信息进行研究,并结合 K-means 聚类方法识别背景和异常,得到以下结论:

(1)组分比值携带着质量演化信息。经过地质作用后的组分比值由原岩的组分含量及其质量变化率两个因素决定。任意组分 m 与不活动组分 i 的含量比值 $\frac{w_m^A}{w_i^A}$ 是关于 m 质量变化率 Δw_m 的线性函数,其斜率为不活动成分 i 的原岩含量的倒数 $\frac{1}{w_i^O}$,截距为初始状态时 m 与 i 的质量比值 $\frac{w_m^O}{w_i^O}$。

(2)任意成分 m 的含量是关于不活动成分 i 的含量投图并与原点连线,其斜率可以反映出地质过程不同阶段中 m 的质量变化率 Δw_m 的相对情况。如果 m 带入富集,则 $\Delta w_m > 0$,在 w_i^A-w_m^A 图上为一组经过原点,斜率从 $\frac{w_m^O}{w_i^O}$ 不断增加的直线簇。增长的斜率反映了在地质过程不同阶段中 m 的质量富集情况。如果 m 带出亏损,则 $\Delta w_m < 0$,在 w_i^A-w_m^A 图上为一组经过原点,斜率从 $\frac{w_m^O}{w_i^O}$ 不断降低的直线簇。亏损的斜率反映了在地质过程不同阶段中 m 的质量亏损情况。

(3)浓度数据表达的是组分和系统总质量的相对变化信息,而不是质量的绝对变化信息。浓度 w_m 携带着物质 m 和非 m 物质的两种质量变化关系。在对元素浓度数据解译时需要了解清楚其携带的信息含义,结合研究问题和已有知识进行解读。使用对数浓度、对数组分比值方法转换后的数据要比浓度数据具有更加清楚的分布特征。对数浓度 $\ln(w_m)$、对数组分比值 $\mathrm{alr}(w_m)$、$\mathrm{clr}(w_m)$、$\mathrm{ilr}(w_m)$

的频率分布和空间特征要比浓度数据 w_m 的好。alr(w_m)、clr(w_m)和 ilr(w_m)三种转换方法的单元素信息表达效果是一样的。

(4) K-means 方法可以有效提取对数浓度 ln(w_m)、对数比值 alr(w_m)、clr(w_m)、ilr(w_m)转换数据的背景和异常。其效果和分形方法一致,优于平均值法。对数数据的空间分布结构优于浓度数据,表达的信息更加清晰。

References

Ahmed M, Seraj R, Islam S M S. 2020. The K-means algorithm: a comprehensive survey and performance evaluation. Electronics, 9(8): 1295～1936.

Ahrens L H. 1953. A fundamental law of geochemistry. Nature, 172(4390): 1148.

Ahrens L H. 1954a. The lognormal distribution of the elements (2). Geochimica et Cosmochimica Acta, 6(2): 121～131.

Ahrens L H. 1954b. The lognormal distribution of the elements (a fundamental law of geochemistry and its subsidiary). Geochimica et Cosmochimica Acta, 5(2): 49～73.

Aitchison J. 1982. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44: 139～160.

Aitchison J. 1986. The Statistical Analysis of Compositional Data. Dordrecht: Springer, 1～460.

Alle'gre C J, Lewin E. 1995. Scaling laws and geochemical distributions. Earth and Planetary Science Letters, 132(1): 1～13.

Anderberg M R. 1973. Cluster Analysis for Applications: Probability and Mathematical Statistics: a Series of Monographs and Textbooks. Cambridge: Academic Press.

Aranganayagi S, Thangavel K. 2007. Clustering categorical data using silhouette coefficient as a relocating measure. International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007): 13～17.

Bholowalia P, Kumar A. 2014. EBK-means:a clustering technique based on elbow method and K-means in WSN. International Journal of Computer Applications, 105(9): 17～24.

Bjørlykke K, Jahren J. 2012. Open or closed geochemical systems during diagenesis in sedimentary basins: constraints on mass transfer during diagenesis and the prediction of porosity in sandstone and carbonate reservoirs. AAPG Bulletin, 96(12): 2193～2214.

Bohrson W A, SPERA F J. 2001. Energy-constrained open-system magmatic processes ii: application of energy-constrained assimilation-fractional crystallization (ec-afc) model to magmatic systems. Journal of Petrology, 42(5): 1019～1041.

Chayes F. 1960. On correlation between variables of constant sum. Journal of Geophysical Research, 65(12): 4185～4193.

Cheng Qiuming, Agterberg F P, Ballantyne S B. 1994. The separation of geochemical anomalies from background by fractal methods. Journal of Geochemical Exploration, 51(2): 109～130.

Darnley A, Bjorklund A, Bolviken B, Gustavsson N, Koval P, Plant J, Steenfelt A, Tauchid M, Xie Xuejin. 2005. A global geochemical database for environmental and resource management: recommendations for international geochemical mapping. Final Report of IGCP Project 259, Paris, France, UNESCO, 1～122.

de Caritat P, Cooper M, Lech M, McPherson A, Thun C. 2009. National geochemical survey of Australia: sample preparation manual. Canberra, Geoscience Australia Record, 1～28.

Egozcue J J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35(3): 279～300.

Figueiredo M A T, Jain A K. 2002. Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3): 381～396.

Gentle M. 2002. The CRM Project Management Handbook: Building Realistic Expectations and Managing Risk. London: Kogan Page Publishers, 1～227.

Ghezelbash R, Maghsoudi A, Carranza E J M. 2020. Optimization of geochemical anomaly detection using a novel genetic K-means clustering (gkmc) algorithm. Computers & Geosciences, 134: 104335.

Goodman L A, Kruskal W H. 1979. Measures of association for cross classifications. In: Measures of Association for Cross Classifications. New York: Springer New York, 2～34.

Grant J A. 1986. The isocon diagram; a simple solution to gresens' equation for metasomatic alteration. Economic Geology, 81(8): 1976～1982.

Hawkes H E, Webb S J. 1963. Geochemistry in mineral exploration. Soil Science, 95(4): 283.

Helgeson H C, Brown T H, Nigrini A, Jones T A. 1970. Calculation of mass transfer in geochemical processes involving aqueous solutions. Geochimica et Cosmochimica Acta, 34(5): 569～592.

Jain A K, Murty M N, Flynn P J. 1999. Data clustering: a review. ACM Computing Surveys (CSUR) 31(3): 264～323.

Johnson C C, Breward N, Ander E L, Ault L. 2005. G-BASE: baseline geochemical mapping of Great Britain and Northern Ireland. Geochemistry: Exploration, Environment, Analysis, 5(4): 347～357.

Kaufman L, Rousseeuw P J. 2009. Finding Groups in Data: an Introduction to Cluster Analysis. New York: John Wiley & Sons, 1～342.

Kirkwood C, Cave M, Beamish D, Grebby S, Ferreira A. 2016. A machine learning approach to geochemical mapping. Journal of Geochemical Exploration, 167: 49～61.

Le Chengsheng, Liu Huihua, Zhong Zhaohui, He Jinhua, Wang Yiwei, Shi Chunwang, Zhang Xianchao, Chen Jinglong, Quan Pinggui, Jiang Zhilin, Zhao Yongli, Hong Yuming. 2011. The reconnaissance survey report of the Zhaojikou Pb-Zn ore deposit in the Dongzhi County, Anhui Province. Anhui Provincial Institute of Nuclear Resource Exploration Technology, 125 (in Chinese).

Lepeltier C. 1969. A simplified statistical treatment of geochemical data by graphical representation. Economic Geology, 64(5): 538～550.

Lever J, Krzywinski M, Altman N. 2016. Classification evaluation. Nature Methods, 13(8): 603～604.

Levinson A A. 1974. Introduction to Exploration Geochemistry (2 Edition). Maywood: Applied Publishing, 1～924.

Li Min, Xi Xiaohuan, Xiao Guiyi, Cheng Hangxin, Yang Zhongfang, Zhou Guohua, Ye Jiayu, Li Zhonghui. 2014. National multi-purpose regional geochemical survey in China. Journal of Geochemical Exploration, 139: 21～30.

Liu Fan, Deng Yong. 2021. Determine the number of unknown targets in open world based on elbow method. IEEE Transactions on Fuzzy Systems, 29(5): 986～995.

Liu Yanpeng. 2017. Metallogenic geochemical mechanism of Zhaojikou epithermal Pb-Zn ore deposit in Anhui Province. Doctoral dissertation of Chinese Academy of Geological Sciences (in Chinese with English abstract).

Liu Yanpeng, Ma Shengming, Zhu Lixin, Sadeghi M, Doherty A L, Cao Dawang, Le Chengsheng. 2016. The multi-attribute anomaly structure model: an exploration tool for the Zhaojikou epithermal Pb-Zn deposit, China. Journal of Geochemical Exploration, 169: 50～59.

Liu Yanpeng, Zhu Lixin, Ma Shengming, Guo Fusheng, Gong Qiuli, Tang Shixin, Gopalakrishnan G, Zhou Yongzhang. 2019. Constraining the distribution of elements and their controlling factors in the Zhaojikou Pb-Zn ore deposit, SE

China, via fractal and compositional data analysis. Applied Geochemistry, 108: 104379.

Lloyd S. 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2): 129～137.

Ma Shengming, Zhu Lixin. 2014. Multidimensional anomaly system for hydrothermal nonferrous metal deposits: taking the Matou porphyry molybdenum copper mine in Anhui Province as an example. Journal of Jilin University (Earth Science Edition), 44(1): 134～144 (in Chinese with English abstract).

Ma Shengming, Zhu Lixin, Su Lei, Tang Lilin, Liu Yanpeng. 2016. Mineralizing agent sulfur and metallogenic process. Acta Geologica Sinica, 90(9): 2427～2436 (in Chinese with English abstract).

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 281～297.

Marutho D, Handaka S H, Wijaya E. 2018. The determination of cluster number at K-mean using elbow method and purity evaluation on headline news. 2018 International Seminar on Application for Technology of Information and Communication, 533～538.

Matschullat J, Ottenstein R, Reimann C. 2000. Geochemical background—can we calculate it? Environmental Geology, 39(9): 990～1000.

McKinley J M, Hron K, Grunsky E C, Reimann C, de Caritat P, Filzmoser P, van den Boogaart K G, Tolosana-Delgado R. 2016. The single component geochemical map: fact or fiction? Journal of Geochemical Exploration, 162: 16～28.

Morris P A, Pirajno F, Shevchenko S. 2003. Proterozoic mineralization identified by integrated regional regolith geochemistry, geophysics and bedrock mapping in western Australia. Geochemistry: Exploration, Environment, Analysis, 3(1): 13～28.

Pearce J A. 2014. Immobile element fingerprinting of ophiolites. Elements, 10(2): 101～108.

Reimann C, Garrett R G. 2005. Geochemical background—concept and reality. Science of the Total Environment, 350(1): 12～27.

Reimann C, Fabian K, Birke M, Filzmoser P, Demetriades A, Négrel P, Oorts K, Matschullat J, de Caritat P, Albanese S, Anderson M, Baritz R, Batista M J, Bel-Ian A, Cicchella D, De Vivo B, De Vos W, Dinelli E, Ďuriš M, Dusza-Dobek A, Eggen O A, Eklund M, Ernsten V, Flight D M A, Forrester S, Fügedi U, Gilucis A, Gosar M, Gregorauskiene V, De Groot W, Gulan A, Halamić J, Haslinger E, Hayoz P, Hoogewerff J, Hrvatovic H, Husnjak S, Jähne-Klingberg F, Janik L, Jordan G, Kaminari M, Kirby J, Klos V, Kweć ko P, Kuti L, Ladenberger A, Lima A, Locutura J, Lucivjansky P, Mann A, Mackovych D, McLaughlin M, Malyuk B I, Maquil R, Meuli R G, Mol G, O'Connor P, Ottesen R T, Pasnieczna A, Petersell V, Pfleiderer S, Poňavič M, Prazeres C, Radusinović S, Rauch U, Salpeteur I, Scanlon R, Schedl A, Scheib A, Schoeters I, Šefčik P, Sellersjö E, Slaninka I, Soriano-Disla J M, Šorša A, Svrkota R, Stafilov T, Tarvainen T, Tendavilov V, Valera P, Verougstraete V, Vidojević D, Zissimos A, Zomeni Z, Sadeghi M. 2018. GEMAS: establishing geochemical background and threshold for 53 chemical elements in European agricultural soil. Applied Geochemistry, 88: 302～318.

Rousseeuw P J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20: 53～65.

Russell S, Russell S J, Norvig P, Davis E. 2010. Artificial intelligence: a modern approach. Prentice Hall, 90: 33～48.

Salminen R, Tarvainen T, Demetriades A, Duris M, Fordyce F, Gregorauskiene V, Kahelin H, Kivisilla J, Klaver G, Klein H. 1998. FOREGS geochemical mapping field manual. Espoo, Geological Survey of Finland, 1～38.

Smith D B, Woodruff L G, O'Leary R M, Cannon W F, Garrett R G, Kilburn J E, Goldhaber M B. 2009. Pilot studies for the North American soil geochemical landscapes project—site selection, sampling protocols, analytical methods, and quality control protocols. Applied Geochemistry, 24(8): 1357～1368.

Smith D B, Cannon W F, Woodruff L G. 2011. A national-scale geochemical and mineralogical survey of soils of the conterminous United States. Applied Geochemistry, 26: S250～S255.

Späth H. 1980. Cluster Analysis Algorithms for Data Reduction and Classification of Objects. Chichester: E. Horwood Halsted Press, 1～226.

Steinley D. 2006. K-means clustering: a half-century synthesis. British Journal of Mathematical and Statistical Psychology, 59(1): 1～34.

Strehl A, Ghosh J. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3: 583～617.

Xie Xuejing, Mu Xuzhan, Ren Tianxiang. 1997. Geochemical mapping in China. Journal of Geochemical Exploration, 60(1): 99～113.

Xie Xuejing, Wang Xueqiu, Zhang Qin, Zhou Guohua, Cheng Hangxin, Liu Dawen, Cheng Zhizhong, Xu Shanfa. 2008. Multi-scale geochemical mapping in China. Geochemistry: Exploration, Environment, Analysis, 8(3): 333～341.

Zhou Shuguang, Zhou Kefa, Wang Jinlin, Yang Genfang, Wang Shanshan. 2018. Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. Frontiers of Earth Science, 12(3): 491～505.

参 考 文 献

乐成生, 刘惠华, 钟朝晖, 何金华, 王一伟, 史春旺, 张咸超, 陈经龙, 全平贵, 蒋智林, 赵永利, 洪玉明. 2011. 安徽省东至县兆吉口铅锌矿普查报告. 安徽省核工业勘查技术总院, 125.

刘艳鹏. 2017. 安徽省兆吉口浅成低温热液型铅锌矿床成矿地球化学机制研究. 中国地质科学院博士学位论文.

马生明, 朱立新. 2014. 热液成因有色金属矿多维异常体系——以马头斑岩型钼铜矿为例. 吉林大学学报(地球科学版), 44(1): 134～144.

马生明, 朱立新, 苏磊, 汤丽玲, 刘艳鹏. 2016. 矿化剂元素硫(S)与成矿. 地质学报, 90(9): 2427～2436.

Identification of background and anomaly information via compositional data theory and unsupervised K-means clustering: a case study of Zhaojikou Pb-Zn ore deposit, Anhui Provicne

LIU Yanpeng[1)], ZHU Lixin[*2)], MA Shengming[3)], DUAN Jilin[4)], GONG Qiuli[3)]

1) *State Key Laboratory of Nuclear Resources and Environment, East China University of Technology, Nanchang, Jiangxi* 330013, *China*;

2) *Development Research Centre, China Geological Survey, Beijing* 100037, *China*;

3) *Institute of Geophysical and Geochemical Exploration, CAGS, Langfang, Hebei* 065000, *China*;

4) *School of Earth Sciences, East China University of Technology, Nanchang, Jiangxi* 330013, *China*

* *Corresponding author: lixinz@cags.ac.cn*

Abstract

Geochemical data are important part of applied geochemical research and basic achievements of geochemical exploration survey. Exploration geochemical data are mainly expressed in the form of "percentages of element mass concentration (abbreviated to "concentration")", which are typical compositional data. It expresses information on relative mass contribution about the ratio of "parts to whole", rather than information on absolute mass change. The concentration data are distributed in the simplex space, rather than the entire Euclidean space. Before data processing, application of appropriate logarithmic ratio transformation would improve the structure and information representation of compositional data. In this paper, the Pb concentration data in the soil of the Zhaojikou Pb-Zn deposit in the Anhui Province is taken as a case study. The logarithmic ratio transformation methods were used to optimize the data structure of the Pb concentration to improve the expression of relative information. Then the K-means clustering of unsupervised learning methods was adopted to identify the background and abnormal information according to the distances of the centroids of the distribution space of the logarithmic ratio transformation data. Finally, the background and anomaly identified by the K-means clustering method was compared with the results of iterative 2δ method and the concentration-area fractal analysis method to evaluate its performance. The results show that: ① the log-ratio method can effectively improve the structure and information expression of concentration data; ② K-means clustering method can effectively identify the background and abnormal information of log-ratio transformed data; its performance is similar to the concentration-area fractal analysis method, and better than the iterative 2δ method.

Key words: machine learning; unsupervised classification; compositional data; K-means clustering; background and anomaly

《地质学报》(中文版)征稿简则

《地质学报》是中国地质学会主办的地质科学学术刊物。《地质学报》反映地质科学各分支学科及边缘学科中最新、最高水平的基础理论研究和基本地质问题研究成果。《地质学报》(中文版)和《地质学报》(英文版)分别独立刊载论文。

一、《地质学报》编辑部与作者约定如下:

1. 作者应保证稿件不一稿两投,并对所投稿件拥有无可争议的著作权。

2. 所有文章均需通过网上办公系统投稿,《地质学报》中文版请投 **http://www.geojournals.cn/dzxb/ch/index.aspx**;《地质学报》英文版:**https://onlinelibrary.wiley.com/journal/17556724**;《地质论评》:**http://www.geojournals.cn/georev/ch/index.aspx**。

网上投稿,请将文、图、表放入同一个 Microsoft Word 文件中(请作者自留原始文件,以备修改,详细投稿办法见网站说明)。投稿被接收与否以编辑部网上收妥回信为准。

3. 不得将投向本编辑部的稿件同时投至其他刊物,否则视为一稿两投。

4. 编辑部承诺一般在 90 日内给出刊用与否的通知。作者在 90 日内,不应将稿件另投他刊。

5. 对决定录用的稿件,作者应根据编辑部提供的修改意见修改后,向编辑部提交论文 Word 文档、清绘好的 CorelDRAW 图件。

6. 稿件文责自负,若做实质性修改,须征得作者同意。

7. 稿件刊出后,将按规定支付稿酬。

二、对投稿内容的要求:

每篇文章需包含下列要素:文章题目(不多于 25 个汉字)、作者、作者单位、内容提要(不少于 400 个汉字)、关键词(5 个左右)、引言(本刊不标"引言"字样,但必须有引言节)、正文、图表、致谢、参考文献、注释、英文摘要(同中文摘要)、作者简介。重要内容说明如下:

1. **正文**:长度不限。应有地质背景、研究方法、研究结果、讨论、结论等几部分。

投向《地质学报》(英文版)的稿件,行文必须规范、通顺,请附相应的中文稿,以备准确理解原文含意。

2. **图件**:① 凡涉及国界的图件必须绘制在地图出版社公开出版的最新地理底图上。② 图件请用 CorelDraw X4 版本格式最好(且不是导入的)。若为其他软件编成的图件,请提供 600 dpi 的 TIFF 格式的文件。彩色照片(包括图版)请提供 600 dpi 以上的 JPG 格式文件。③ 图件大小:竖排图件宽度<165 mm(通栏图)和<80 mm(半栏图),高度都<230 mm。横排图件宽度<240 mm,高度<150 mm。图内的中文全部用宋体,英文和数字用 Times New Roman,希腊字符用幼圆字体;字号 8 pt,个别字体可用 7 pt。④ 图件不同区域可用通用地质花纹(或符号)区分,除照片外,一般不用灰度图。若必须用灰度图表示不同区域时,灰阶应尽量少,阶差应尽量地大。⑤ 图件若为彩色照片者,可选择集中制成图件。⑥ 图名、图说明、图例注释都应有相应的英文说明。

3. **参考文献**:本刊采用著者-年制,在正文及其图表中:如文献有两个作者或两个作者以上,用"等"或"et al."。举例说明:××××(Whalen et al., 1987; Gilder et al., 1991; 许志琴等,2002; Hu Ruizhong et al., 2008)。或"Song Biao et al.(2002)指出采样方法……";"许志琴等(2003)指出……"(同时列出多篇文献时,按年代先后排列)。

中文文献均需提供英译,所有英文文献均放在"References"标题之下。英文文献按"第一作者字母序+年代"排列,其后其他文种放在"参考文献标题之下,按中文、日文、西文、俄文、其他文排列。中文文献按"第一作者姓名汉语拼音字母+年代"先后排列。其他文均按各自第一作者姓名字母顺序排列。若是投《地质学报(英文版)》无须考虑上述情况。

文章请列出全部作者。但专著可按原书封面样式给出,其中的论文写"见:XXX 主编."项时,指明主编一人即可("见:XXX 等主编.")。每一条文献的列出格式请参照我刊 2016 年以来的文章。书籍的引用,分两种情况:a、书籍本身有相应英文名的,引用按正常要求;b、书籍本身没有相应英文名的,尽量不引用此文献,若必须引用,则仅在中文参考文献中列出,正文中的出现用中文引用。

4. **注释**:引用非公开出版物时在文后单列注释一栏,格式与参考文献相同。参考文献及注释详细格式可见《地质学报》修改注意事项。

5. **英文摘要**:在《地质学报》中文版和《地质论评》上发表的论文必须提交英文摘要,包括题名、作者、作者单位、内容提要和关键词。作者和作者单位均应为全名,内容提要与相应中文提要一致,最好更为详细。

6. **作者简介**:主要介绍作者的学术经历,包括资助项目、姓名、性别、学历、职称、研究方向、E-mail 等,具体格式可参考我刊已发表的最新文章。

《地质学报》(中文版)编辑部