地球化学成分数据对数比值转换的若干问题思考

谭亲平[1]，夏勇[1]，王学求[2,3]

(1. 中国科学院 地球化学研究所 矿床地球化学国家重点实验室，贵州 贵阳 550081；2. 国土资源部 地球化学探测技术重点实验室，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；3. 联合国教科文组织 全球尺度地球化学国际研究中心，河北 廊坊 065000）

地球化学数据是典型的成分数据，它的“闭合效应”已被广泛熟知（Aitchison，1986）。闭合效应既所有成分（元素含量）的总量等于 1（100%），成分之间都是相互制约的，呈现出负或正的相关性，这些相关都是虚假相关，没有任何地质意义。最典型的虚假相关就是硅与其它元素的负相关性。地球化学家认为在进行数学处理前应先“打开”数据（Zuo et al., 2013）。目前，最流行的方法就是进行对数转换，转换的方法包括：直接对数变换（ln or log transformation），非对称对数比值变换（additive−logratio transformation, alr），中心对数比值变换（centered−logratio transformation, clr）和等距对数比值变换（isometric−logratio transformation, ilr）（Aitchison, 1986; Egozcue et al., 2003），前三种是一对一的变换，后一种是非一对一的，即每一个成分的计算都是随着成分的排序的不同而不同。

由于成分数据的实质是占总量的百分比，单位常常是 100%，ug/g, ng/g 等，无论取样量有多大，某个成分占总量的比值是不变的。数学地质学家们自然而然的想到了用比值的方法来“打开”成分数据，这就提出了公式：

（1）$\mathrm{alr(x)} = \left\{ln\frac{x_1}{x_D}, ln\frac{x_2}{x_D}, \cdots, ln\frac{x_{D-1}}{x_D}\right\}$，其中 x（D）为任意成分或元素（Aitchison, 1986）。

但是公式（1）选择任意一个成分来做分母，具有人为的随意性，同时会牺牲掉被选择的那个成分，因此进一步又提出了公式，就是用几何平均值来代替任意值：

（2）$\mathrm{clr(x)} = \left\{ln\frac{x_1}{g(x)}, ln\frac{x_2}{g(x)}, \cdots, ln\frac{x_D}{g(x)}\right\}$，其中 g（x）为几何平均值（Aitchison, 1986）。

对于 ilr 来说，定义 D 个成分的两个样品为 x=（x_1, ……, x_D）和 y=（y_1, ……, y_D），计算两个样品的距离需要投射到 D 维变量空间，它包含了所有的成分信息和约束条件，这时的空间结构就是艾奇逊几何结构。很显然，欧几里德距离（样品在单一成分上的距离）应用在这里不合适，正确的距离计算应是艾奇逊距离（Filzmoser et al., 2009），公式为：

（3）$d_A(x,y) = \sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(ln\frac{x_i}{x_j} - ln\frac{y_i}{y_j}\right)^2}$。

假如只考虑由两个成分组成的样品，即 $x_j=1-x_i$，那么艾奇逊距离变为：

（4）$d_A(x,y) = \sqrt{\frac{1}{2}}\left|ln\left(\frac{x}{1-x}\right) - ln\left(\frac{y}{1-y}\right)\right|$。

很显然，它的距离公式和公式（1）相比，仅仅差一个$\sqrt{1/2}$。如果不考虑$\sqrt{1/2}$，那么 ilr 转换就可以用 alr 来代替，这就是所谓的几何属性的改变，由艾奇逊距离转变为欧几里德距离。数学地质学家们认为两个成分相关性的实质应该是艾奇逊几何空间的距离问题。假如一个样品只有两个成分组成，那么 ilr 的转换只剩下一个成分，当 D=2 时，ilr=$\sqrt{1/2}$alr，他与 alr 转换的差别就是多了一个$\sqrt{1/2}$，这差别对于相关分析没有影响，主要的差别就是它的理论属性，ilr 转换重要的属性就是等距（Isometry），也就是说将艾奇逊距离直接转换为欧几里德距离，最终提出了 ilr 转换的公式（Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2006）：

（5）$\mathrm{ilr(x)} = \sqrt{\frac{D-i}{D-i+1}}ln\frac{x_i}{\sqrt[D-i]{\prod_{j=i+i}^{D}x_j}}$，其中 i=1, 2, ……, D−1，$\sqrt[D-i]{\prod_{j=i+i}^{D}x_j}$为 j 到 D 成分几何平均值。

基金项目：国家重点研发计划（2016YFC0600607）； 全球地球化学填图计划（DD20160116）

作者简介：谭亲平，男，1986 年生，博士后，主要从事矿床学研究，E-mail: 565310821@qq.com

* 通讯作者，E-mail: wangxueqiu@igge.cn

对于中心对数比值变换，如果将公式（2）分解就可以得到：$clr(x_i)=ln(x_i)-ln(g(x))$。它与对数变换之间的差别就是每个 $ln(x_i)$都减去一个常数，即 $ln(g(x))$。地球化学数据经过 clr 转换之后会出现大量的负相关，造成这种负相关的根本原因还是 clr 的计算过程，即：$Sum=[ln(x_1)+ln(x_2)+\ldots+ln(x_n)]-ln(g(x))\times n=0$，即所有样品成分的总和为零。clr 转换并没有改变地球化学成分数据的本质，只是将所有成分的总和从 1 变为 0。更为致命的是，仅针对微量元素而言，由于在地质样品中含量微乎其微，受总体加和效应的影响微弱，在 clr 转换之后造成人为的加和效应，造成微量元素之间大量的负相关。因此，clr 的转换方法不适合“打开”地球化学成分数据。

ilr 是一种有效的“打开”地球化学成分数据的方法，因为它具有优越的理论属性，但是当成分之和不等于 1 时，它的使用需要慎重。数学地质学家们对 ilr 的推算过程是假设成分之和等于 1，它适用于样品各个成分之间加和等于 1（100%）的数据，比如地学中所涉及的主量元素数据。成分之和等于 1 的假设要求各成分的单位必须统一，特别是当主微量元素地球化学数据一起使用时。ilr 不是一对一的变换，变换后的数据不能用于单元素的研究，只能用于那些基于相关系数矩阵或协方差矩阵的研究，比如多元统计分析。由于 ilr 不能用于单元素，ilr 在多元统计分析后期还要转换到原始对数空间才能进一步使用。但是，当成分之和不等于 1 时，整个推算过程也就不存在，即不存在 y=1-x。这种变换并不适用于仅仅涉及微量元素的地球化学数据，因为仅涉及微量元素不可能存在成分之和等于 1 的假设。另外，微量元素顾名思义就是样品中含量微乎其微的成分，它含量的多少受主量元素的影响较弱。如果将微量元素和主量元素的单位统一为 100%，微量元素的量级仅为：$n \times 0.0001\%$，与主量元素相比，相差 4 个数量级。因此，当仅仅涉及微量元素或各成分加和不等于 1 的地球化学数据时，直接对数转换应为最合适的方法。

参 考 文 献：

Aitchison, J. 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London, UK.

Egozcue, J.J., Pawlowsky-Glahn, V. 2006. Simplicial geometry for compositional data. Geological Society, London, Special Publications, 264(1), 145-159.

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.

Filzmoser P, Hron K, Reimann C, et al. 2009. Robust factor analysis for compositional data. Computers & Geosciences, 35(9):1854-1861.

Zuo, R.G., Xia, Q.L., Wang, H.C. 2013. Compositional data analysis in the study of integrated geochemical