Journal of Geochemical Exploration 141 (2014) 1–5

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/jgeoexp

Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes?

A. Buccianti [a,*], E. Grunsky [b]

[a] Department of Earth Sciences, University of Florence, Italy
[b] Geological Survey of Canada, Ottawa, Ontario K1A 0E8, Canada

ARTICLE INFO

Article history:
Received 11 March 2014
Accepted 17 March 2014
Available online 26 March 2014

Keywords:
Compositional data analysis
Geochemical data
Simplex geometry
Log-ratio approach
Environmental modelling

ABSTRACT

Geochemical data are typically reported as compositions, in the form of some proportions such as weight percents, parts per million, etc., subject to a constant sum (e.g. 100%, 1,000,000 ppm). This latter implies that such data are "closed"; that is, for a composition of *D*-components, only $D-1$ components are required. The statistical analysis of compositional data has been a major issue for more than 100 years. The problem of spurious correlation, introduced by Karl Pearson in 1897, affects all data measuring parts of some whole, which are by definition, constrained; and such type of measurements are present in all fields of geochemical research. The use of the log-ratio transform was introduced by John Aitchison to overcome these constraints by opening the data into the real number space, within which standard statistical methods can be applied. However, many statisticians and users of statistics in the field of geochemistry are unaware of the problems affecting compositional data, as well as solutions that overcome these problems. A look into the ISI Web of Science and Scopus databases shows that most papers where compositional data are the core of a geochemical research continue to ignore methods to correctly manage constrained data. A key question is how we can demonstrate that the interpretation of the behaviour of chemical species in natural environment and in geochemical processes is improved when the compositional constraint of geochemical data is taken into account through the use of new methods. In order to achieve this aim, this special issue of the Journal of Geochemical Exploration focuses on the correct statistical analysis of compositional data. Applications in exploration, monitoring and environments by considering several geological matrices are presented and discussed illustrating that several paths can be followed to understand how geochemical processes work.

1. Introduction

More than 100 years have elapsed since Karl Pearson wrote his paper on spurious correlation in 1897 (Pearson, 1897; Fig. 1) and more than 30 years from a first solution based on log-ratios proposed by John Aitchison (Aitchison, 1982; Fig. 2). Since then, the approach has been characterised by many studies on the natural geometry of the sample space where compositional data are positioned. In order to understand the meaning of these words, it is necessary to link *geochemistry* to *geometry*, two fields of research apparently distant but, in reality, closely linked.

The *geometry* of a composition is the metric of the *sample space.* For example, when we measure concentrations of some geological material in laboratory we do not expect to find negative values, only positive values from which an interpretation is based. The sample space is where compositional values are located. Compositions are compared by measuring their distance and translations along linear or non-linear trends. It is in this sample space where *random variables*, the mathematical rules to attribute probability to the occurrences of events, are defined for statistical inference.

Some sample spaces may be better than others to exploit the information contained in the data. This is the case of compositional data where the elements of the composition are non-negative and sum up to a constant, e.g. to unity since they have been scaled by the total of the components as a standardization practice. A consequence of this is that a composition of *D* parts, $[x_1, x_2,\ldots,x_D]$, can be identified with a closed vector

$$\mathbf{x} = C[x_1, x_2, \ldots, x_D] = \left[\frac{x_1 \cdot k}{\sum_{i=1Dx_i}}, \frac{x_2 \cdot k}{\sum_{i=1Dx_i}}, \ldots, \frac{x_D \cdot k}{\sum_{i=1Dx_i}}\right], \quad (1)$$

where *C* is called the closure operation to the constant *k* (Aitchison, 1986).

* Corresponding author. Tel.: +39 0552757493; fax: +39 055284571.
E-mail address: antonella.buccianti@unifi.it (A. Buccianti).

http://dx.doi.org/10.1016/j.gexplo.2014.03.022
0375-6742/

Fig. 1. Karl Pearson (27 March 1857–27 April 1936) the scientist who founded the discipline of mathematical statistics.

The set of real positive vectors closed to a constant k constitutes the constrained sample space called simplex of D parts, denoted by S^D and defined as

$$S^D = \{(x_1, x_2, \ldots, x_D) : x_1 > 0, x_2 > 0, \ldots, x_D > 0; x_1 + x_2 + \ldots + x_D = k\}. \quad (2)$$

Note that geochemical data are always non-negative and are restricted to the positive part of the real sample space, R^D_+. To be noticed here is that the previous approach gives importance to the sample space (sum constraint) but compositional data cannot be a priori closed. In most situations is the analyst who decides that the total of each sample is not relevant and then normalise the data to proportions. All the sets of data are equivalence classes from a mathematical point of view (Buccianti and Pawlowsky-Glahn, 2005).

The key in understanding compositional data relies on defining a correspondence between the simplex S^D and R^D, the real space governed by Euclidean geometry, through the use of a metric where classical statistics can be applied for an unbiased interpretation of the relationships and patterns of geochemical data.

The equivalence between S^D and R^D is obtained by defining equivalent operations in S^D. The definition of the operations of sum (difference) and product, called *perturbation* and *powering*, together with the definition of other properties (*norm*, *distance*, *inner product*) allow us to consider S^D as a space with a structure governed by the Euclidean geometry completely equivalent to the geometry of the correspondent unrestricted real space with $D - 1$ dimensions (Billheimer et al., 2001; Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2006).

Fig. 2. John Aitchison in occasion of the Codawork 2005 in Girona (E).

2. The algebraic–geometric structure of the simplex

Two fundamental operations of change govern the algebraic–geometric structure (the metric) of the simplex of D-parts, **perturbation**, denoted by the symbol $\oplus$, and **powering** denoted by the symbol $\odot$ (Aitchison, 1986). Consider two compositions $\mathbf{x}$ and $\mathbf{y}$. To perturb $\mathbf{x}$ by $\mathbf{y}$, first calculate the component-wise product and then close the result to 100 (C closure operator) to produce $\mathbf{z}$:

$$\mathbf{x} \oplus \mathbf{y} = C[x_1 \cdot y_1, x_2 \cdot y_2, \ldots, x_D y_D] = \mathbf{z}. \quad (3)$$

It is then clear that

$$\mathbf{z} \ominus \mathbf{y} = C\left[\frac{z_1}{y_1}, \frac{z_2}{y_2}, \ldots, \frac{z_D}{y_D}\right] = \mathbf{x}. \quad (4)$$

Consider now a real number a; the power-transformed composition is given by

$$a \odot \mathbf{x} = C\left[x_1^a, x_2^a, \ldots x_D^a\right]. \quad (5)$$

The used symbols $\oplus$ and $\odot$ emphasize the analogy with the operations of displacement or translation and scalar multiplication of vectors in R^D. Perturbation clearly corresponds to addition in R^D, while powering is the multiplication. The internal ($\oplus$) and external ($\odot$) operations define a vector or linear space structure on S^D, a structure that can be extended by the introduction of the simplicial metric or Aitchison distance:

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\left[\sum_{i=1}^{D}\left\{log\frac{x_i}{g(\mathbf{x})} - log\frac{y_i}{g(\mathbf{y})}\right\}^2\right]}, \quad (6)$$

where $g(\cdot)$ is the geometric mean of the parts of the composition, $g(\mathbf{x}) = (\prod_{i=1}^{D} x_i)^{1/D}$.

It is possible to demonstrate that this metric satisfies the usual metric axioms and other desirable properties. The definition of distance, together with that of norm (the distance of $\mathbf{x}$ from the origin of a linear space), the inner product (the cosine of the angle between compositional vectors) and the operations of perturbation and powering, provide a Euclidean structure to the simplex, now called Aitchison simplicial geometry (Pawlowsky-Glahn and Egozcue, 2001).

Within this framework it is possible to work inside the simplex to perform an analysis of compositional data free of inconsistencies.

An important consequence of this approach is how the concept of distance changes from S^D, with the Aitchison distance d_a reported in Eq. (6) to the Euclidean distance d_e in R^{D-1} defined as

$$d_e(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_{D-1} - y_{D-1})^2}. \quad (7)$$

Eqs. (6) and (7) measure the same magnitude but in different sample spaces. Applying Eq. (7) in a sample space that supports, for the same concept of distance, Eq. (6) is an error that can lead to misleading results. Similar scenario appears for spherical data, where people agree to use the angular distance.

However, compositional data can also be transformed to go out of the simplex S^D and into the unconstrained real space R^{D-1} by applying one of the proposed log-ratio transformations (Aitchison, 1982).

3. The log-ratio transformation

As explained in Aitchison et al. (2000) the log-ratio transformations allow us to make meaningful statements on compositional data only involving ratios of components. The first principle of compositional data