Computers & Geosciences 50 (2013) 165–173

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

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

Is compositional data analysis a way to see beyond the illusion?

Antonella Buccianti

University of Florence (I), Department of Earth Sciences, Via G. La Pira 4, 50121 Florence (I), Italy

ARTICLE INFO

Article history:
Received 30 April 2012
Received in revised form
13 June 2012
Accepted 14 June 2012
Available online 28 June 2012

Keywords:
Compositional data analysis
Geochemistry
Worldwide lithology
River chemistry
Volcanic gases
Isometric log-ratio transformation

ABSTRACT

Notwithstanding the numerous contributions that have been published on theoretical and practical aspects of the management of compositional (constrained) data during the last thirty years, in geochemistry most of the scientific papers in international journals continue to ignore their peculiar features. In order to understand the reasons of the undervaluation of the effects of an incorrect choice of the sample space and, consequently, an incorrect application of the distance concept, case studies of comparison between methodologies will be presented and discussed. The aim is to evaluate the differences in interpretation of geochemical processes affecting rocks, water and gaseous samples, when the two different approaches, classical and compositional, are adopted. If we compare the results of case studies following the two paths it is possible to evaluate which type of error (and consequences) will affect our evaluations in geochemistry.

The presence of expected differences between the two approaches indicates that compositional data analysis can be a way to see beyond the illusion due to the constrained space. However, the possibility that the difference is tenuous in some situations, not revealed a priori, may be at the origin of the unconscious choice of the classical approach. Is this condition which some researchers call "common sense" frequently encountered in geochemistry? The paper is aimed to try to answer the proposed question, and to understand the difficulty of diffusion of compositional data analysis even if now simple tools of investigation, for different degrees of knowledge, are available.

1. Introduction

The development of graphical and numerical tools to perform compositional data analysis (CoDA) represents a benchmark problem in geological sciences and, in particular, in computational geochemistry. This discussion has long animated researchers working in different fields and, notwithstanding that appropriate tools are now available to correctly investigate the features of compositional data, most of the published papers in the international literature avoid facing this attractive question. Moreover, people that are interested in managing compositional data in a consistent[1] mathematical and statistical framework, often have to convince referees of the appropriateness of their methodology. Answers such as "nature is stronger than closure" or "a good geologist is able to recognise forced relationships from the natural one" may be typical. Why is it so difficult to convince the scientific community about the adoption of the CoDA approach? Are the expected differences between results obtained by using classical and compositional approaches able to convince us that a way to see beyond the illusion due to the constrained space is to take into account its geometry? Or the differences are usually so tenuous, that the unconscious choice of the classical approach is recognised as "common sense"? The paper is aimed to try to answer the proposed question and to understand the difficulty of diffusion of compositional data analysis even if now simple tools are available.

The first benchmark of the compositional data problem can be considered the presence of the *spurious correlations* recognised by Pearson (1897) affecting all data that measure parts of the same whole, such as percentages, proportions, ppm and so on. His work represented the first approach able to recognise that if X, Y and Z are uncorrelated, then X/Z and Y/Z ratios, will not be uncorrelated. Chayes (1960) found a mathematical demonstration of Pearson's work and showed that some of the correlations between the components of the composition must be negative due to the sum constraint, thus affecting interpretation of natural processes, biased by this effect.

Even if natural data are often non-negative, ranging in a sample space with a restriction to R_+^D (they are only positive, and variables move only on the positive parts of real space), compositional data have a further restriction since they have been

E-mail address: antonella.buccianti@unifi.it

[1] In the sense of coherent with the principles of compositional data analysis (Egozcue and Pawlowsky-Glahn, 2011) that is: (a) The relative character of the information carried by the data is taken into account, (b) The models that are compatible with the sample space and constraints do not need to be taken additionally into account.

0098-3004/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2012.06.012

scaled by the total of the components of the composition. This important operation of standardization is fundamental in interpreting and comparing results obtained by experimental measures in geochemistry, since data are related to the same weight (solid matrices) or volume (solutions and gaseous mixtures). The consequence is that compositional data with D components not only pertain to the positive part of R^D but occupy a restricted part of its axes, in general from 0 to the constant defined *a priori*. In mathematical terms compositional data are represented as pertaining to a sample space called the simplex S^D:

$$S^D = \left\{ \boldsymbol{x} = (x_1, x_2, x_D) : x_i > 0 \ (i = 1,2,D), \sum_{i=1}^{D} x_i = \kappa \right\} \quad (1)$$

where κ is a given positive constant, defined *a priori* and depending on how the parts are measured. The key question here is not to discuss the nature of compositional data, since they have to be defined exactly in this way if we want to make comparisons among cases, starting from the same baseline. The key question is whether standard statistical analysis which assumes that the sample space is R^D with D dimensions, where all the values from $\pm\infty$ are generable and have a probability to be found in a sampling process, is appropriate to represent the investigated phenomena. In other words, in an olivine, a silicate mineral represented by the formula (Mg, Fe)SiO_2, it is known that Mg and Fe substitute each other and that a rigorous stoichiometric law governs the process characterised by the competition of two ions for the same crystallographic site. It is also clear in this framework that when Mg decreases, Fe tends to increase (common sense). However, the classical approach, that simply represents this phenomenon in a binary diagram, where abundance of Fe and Mg are analysed in respect to each other, or versus Si considered a common base, does not represent a coherent geometry on which to base statistics, both descriptive and inferential, and to propose models able to indicate how natural phenomena work. In fact, considering compositional data as real data, the hypothesis that it is possible to obtain negative contents of Fe and Mg in a sampling process is considered as feasible. The formulation of this hypothesis is frequently performed, even if unconsciously, when a correlation coefficient is determined, or some modelling of the linear pattern of the data on binary diagrams is proposed. The probability of a negative concentration may be low, but it is not possible to know its value in advance and this also affects the determination of simple central tendency statistics (mean) and variability measures (variance). Consequently, as reported in Aitchison et al. (2000), it should be obvious that with compositional data only the statements about the ratios of the components are meaningful, since their use respects the fundamental principle of *scale invariance*. This item for a long time was indirectly recognised in geochemistry as is testified by the common use of ratio diagrams. However, often these diagrams were realised considering ratios with the same denominator, for example X/Z and Y/Z ratios that, as reported by Pearson (1897) were affected by spurious correlations.

In the statistical literature there is a long history of the search for a solution to the statistical analysis of compositional data (Aitchison and Egozcue, 2005). The main contribution to a solution is attributable to John Aitchison in the early 1980s (Aitchison, 1982) when he introduced the log-ratio approach using the intuitive concept of difference associated with the features of data. For example, the log-ratio approach was proposed to capture the difference between 5% and 10% and that between 45% and 50%, difference equal to 5 in both cases in the Euclidean real space. Following this approach, compositions are transformed to move them into real space using a log-ratio transformation, analysed by classical statistical methods, and results reported back to the simplex, by using the correspondent inverse transformation. A further key step was the recognition of the Euclidean space nature of the simplex (Pawlowsky-Glahn and Egozcue, 2001). In this framework compositions can be represented by their coordinates in the simplex with a suitable orthonormal basis, leading to the *ilr* (isometric log-ratio) transformation (Egozcue et al., 2003). Its use allow us to avoid the arbitrariness of denominator choice related to the *alr* (additive log-ratio) transformation and to the singularity of the *clr* (centered log-ratio) transformation, the two transformations originally proposed by Aitchison (1982).

In this paper the comparison of the results obtained for some interesting compositional cases investigated by using the classical and the log-ratio approach allows us to verify how the illusion to see compositional data as real data may compromise our understanding of natural phenomena. To achieve this aim, the *ilr* transformation was used to represent a composition as a real vector. Even if the computation of *ilr* coordinates appears to be complex, there are different rules on how to generate them (Egozcue et al., 2003). The identification of balances, a particular form of *ilr* coordinates (Egozcue and Pawlowsky-Glahn, 2005) may simplify the adoption of this transformation. Balances, reflecting the relative variation of two groups of parts, represent a powerful tool for researchers to prove their geochemical hypothesis, translating ideas on natural phenomena in numbers moving in a coherent geometry. Balances in fact define coordinates of the samples within an orthogonal system of axes, i.e., they are usual random variables in real space.

2. Working on coordinates: the *ilr* transformation of compositional data and the balances approach

In statistics the real space R^k (k=number of dimensions) is assumed to be the natural sample space for a given set of observations. Standard statistics have been developed in R^k using its particular algebraic–geometric structure, which is commonly known as Euclidean geometry. Linear algebra allows us to translate standard statistics into any sample space, different from R^k, if it has an Euclidean vector space structure. Definition of basic operations in the simplex such as perturbation and powering, with the associated norm and distance, permits us to analyse data (Aitchison, 2001; Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001). In this framework, the procedure of the sequential binary partition to identify orthonormal coordinates, can be adopted (Egozcue and Pawlowsky-Glahn, 2005). In a first step the parts of the composition are divided into two groups: the parts of the first group are coded by $+1$ and the parts of the second group are coded by -1. In this way the first coordinate describing the balance between the $+1$ and -1 parts is obtained. In the second and following steps a previous group of parts is divided into new groups, similarly coded by $+1$ and -1 while the components that are not involved are coded with a zero. The number of steps required for all the groups to contain a single component is exactly $D-1$, dimensions of S^D. The whole procedure can be summarised in a table as reported in Egozcue and Pawlowsky-Glahn (2005). From a general point of view, in the kth step the balance z_k (Eq. (2)) between two groups is obtained so that the r_k ($+1$) parts are balanced with the s_k (-1) parts:

$$z_k = \sqrt{\frac{r_k s_k}{r_k + s_k}} \ln \frac{(x_{i1} x_{i2} x_{ir_k})^{1/r_k}}{(x_{j1} x_{j2} x_{js_k})^{1/s_k}}, \quad k = 1, D-1 \quad (2)$$

or:

$$z_k = \sqrt{\frac{rs}{r+s}} \ln \frac{g_m(\boldsymbol{x}_+)}{g_m(\boldsymbol{x}_-)}, \quad (3)$$