ELSEVIER

Computers & Geosciences 34 (2008) 320–338

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

"compositions": A unified R package to analyze compositional data

K. Gerald van den Boogaart[a,*], R. Tolosana-Delgado[b]

[a]*Institut für Mathematik und Informatik, Ernst-Moritz-Arndt-Universität Greifswald, D-17487 Greifswald, Germany*
[b]*Geowissentschaftliches Zentrum, Sedimentologie & Umweltgeologie, Georg-August-Universität Göttingen, D-37077 Göttingen, Germany*

Received 5 October 2006; received in revised form 21 November 2006; accepted 23 November 2006

Abstract

This contribution presents a new R package, called "compositions". It provides tools to analyze amount or compositional data sets in four different geometries, each one associated with an R class: rplus (for amounts, or open compositions, in a real, classical geometry), aplus (for amounts in a logarithmic geometry), rcomp (for closed compositions in a real geometry) and acomp (for closed compositions in a logistic geometry, following a log-ratio approach). The package allows to compare results obtained with these four approaches, since an analogous analysis can be performed according to each geometry, with minimal and straightforward modifications of the instructions. Beside these grounding classes, the package also includes: the most-basic features such as data transformations (e.g. logarithm, or additive logistic transform), basic statistics (both the classical ones, and those developed in the log-ratio framework of compositional analysis), high-level graphics (like ternary diagram matrix and scatter-plots) and high-level analysis (e.g. principal components or cluster analysis). Results of these functions and analysis are also provided in a consistent way among the four geometries, to ease their comparison.

Keywords: Aitchison geometry; Compositional data; Euclidean space; Multivariate data analysis; Software

1. Introduction

This paper aims at presenting a new package for R, the open-source statistical environment, devised to analyze compositional data. This paper describes its structure, and shows many of the available functions, but it is not intended as a guide to its use. Nevertheless, the appendix includes a recommended draft sequence of analysis. If more guidance on its basic usage is wanted, the package itself carries a hands-on instruction, *UsingCompositions.pdf*: it is installed in the /doc subdirectory of the package (available through the html help of R), and can be downloaded from the package home page, http://stat.boogaart.de/compositions/. Also, van

*Corresponding author.
E-mail addresses: boogaart@uni-greifswald.de (K.G. van den Boogaart), raimon.tolosana@geo.uni-goettingen.de (R. Tolosana-Delgado).
URL: http://www.iamg.org/CGEditor/index.htm

0098-3004/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2006.11.017

den Boogaart and Tolosana-Delgado (2006) present a detailed guide on the use of the package "`compositions`" within the scope of Aitchison (1986) analysis of compositions.

In the mathematical geology community, there is a strong disagreement on how to do a statistical analysis of compositional data. Though several warnings on the spurious effects of the so-called closure operation on the covariance matrix (Chayes, 1960), most users of statistics in geosciences simply ignore the problem, and continue using classical statistical methods. Aitchison (1982) put forward several considerations on compositions and their sample space (the so-called *Simplex*), which he argued that should be honored by any statistical method for compositions: results should be independent of the measurement units, as well as of permutations of the parts involved, and subcompositions should behave as marginals in classical statistics. Attending to these considerations, he suggested a methodology, which avoided the closure effect by taking log-ratios: shortly, he proposed to transform the data, apply standard statistical procedures on the transformed scores and back-transform the results, when it was sensible. His approach was complemented with a series of operations to measure change and distance between compositions. Afterwards, the Simplex (equipped with these operations) was identified as an Euclidean space on itself (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001). This opened the door to the concept of coordinates, and allowed Pawlowsky-Glahn (2003) to reformulate Aitchison's *recipe* in algebraic terms: take coordinates of your observations with respect to a chosen basis, use standard statistical techniques on them (being real unbounded numbers), and apply the results to the basis used. This is what is called the *principle of working on coordinates*: statistical analysis and probabilistic theory are applied to the coordinates of the observations in an Euclidean space structure.

On the other side of the spectrum, Shurtz (2003) suggests that the most important consideration on compositions is mass balance, and that any statistical analysis should honour this fundamental law: consequently, the closure effect on correlation should not be considered spurious anymore. In the same direction, Rehder and Zier (2002) interpret compositions as *standardized* (closed) *mixtures of components*, and put forward the convex mixture as the fundamental operation on compositions. These two considerations are mutually consistent, and conceptually support the sense of a classical Euclidean geometry and of linear statistical techniques in compositional data. It is nevertheless worth mentioning that these arguments to justify the suitability of an Euclidean approach have arisen as an argument against Aitchison's solution.

The number of applications found in the references using a classical approach is enormously higher than those using Aitchison's, although no argument or conscious decision is done on most of them to support their choice. In the authors' opinion, two reasons explain this. First, there is a lack of available software to analyze a data set under Aitchison's postulates: to the authors' knowledge, there are: a set of Matlab routines (*CoDa*, programmed by Aitchison, available on request), an MSExcel-based package (Thió-Henestrosa and Martín-Fernández, 2005, *CoDaPack*) and two sets of `R` routines (Beardah and Baxter, 2005; Bren and Batagelj, 2005). The second reason is the inertia of analysts, who classically consider that raw data statistics are "purely descriptive", i.e. there was no prior choice of a model. This idea is rather naive, as applying no transformation carries an implicit choice of a distance model.

In this context, the presented package is essentially built with the twofold aim of: (a) providing an easy way to analyze compositions using a classical approach *and* using the Aitchison one, so that (b) *the analyst* can compare results and ground a decision on which geometry to choose.

The fundamental structure of `compositions` follows Pawlowsky's (2003) principle of working on coordinates: the user chooses a geometry to represent the compositional data set, and then the package automatically computes its coordinates in an adequate reference system, conducts any desired analysis on the coordinates, and (when needed) applies the obtained results to the system basis before representing or returning results to the user. The procedure may be summarized as