Science of the Total Environment 607–608 (2017) 965–971

Contents lists available at ScienceDirect

Science of the Total Environment

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

A new method for correlation analysis of compositional (environmental) data – a worked example

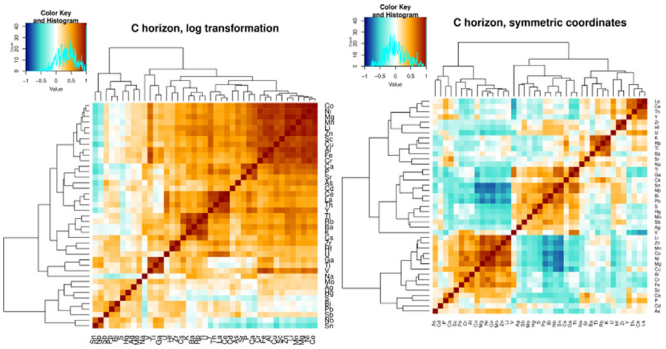
C. Reimann [a,*], P. Filzmoser [b], K. Hron [c], P. Kynčlová [b], R.G. Garrett [d]

[a] Geological Survey of Norway, P.O.·Box 6315, 7491 Sluppen, Norway
[b] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstr. 8–10, 1040 Vienna, Austria
[c] Department of Mathematical Analysis and Applications of Mathematics, Palacký University, 17. listopadu 12, 77146 Olomouc, Czech Republic
[d] Emeritus Scientist, Geological Survey of Canada, Natural Resources Canada, 601 Booth St., Ottawa, ON K1A 0E8, Canada

HIGHLIGHTS

- A new method for correlation analysis of compositional data is demonstrated.
- 'Classical' scatterplots can provide a wrong impression of the relations between two variables.
- Heat-maps provide a fast overview of the correlation structure of a data set.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 11 May 2017
Received in revised form 8 June 2017
Accepted 8 June 2017
Available online xxxx

Editor: D. Barcelo

Keywords:
Correlation
Scatterplot
Compositional data analysis
Log-ratio methodology
CoDa

ABSTRACT

Most data in environmental sciences and geochemistry are compositional. Already the unit used to report the data (e.g., µg/l, mg/kg, wt%) implies that the analytical results for each element are not free to vary independently of the other measured variables. This is often neglected in statistical analysis, where a simple log-transformation of the single variables is insufficient to put the data into an acceptable geometry. This is also important for bivariate data analysis and for correlation analysis, for which the data need to be appropriately log-ratio transformed. A new approach based on the isometric log-ratio (ilr) transformation, leading to so-called symmetric coordinates, is presented here. Summarizing the correlations in a heat-map gives a powerful tool for bivariate data analysis. Here an application of the new method using a data set from a regional geochemical mapping project based on soil O and C horizon samples is demonstrated. Differences to 'classical' correlation analysis based on log-transformed data are highlighted. The fact that some expected strong positive correlations appear and remain unchanged even following a log-ratio transformation has probably led to the misconception that the special nature of compositional data can be ignored when working with trace elements. The example dataset is employed to demonstrate that using 'classical' correlation analysis and plotting XY diagrams, scatterplots, based on the original or simply log-transformed data can easily lead to severe misinterpretations of the relationships between elements.

* Corresponding author.
E-mail addresses: clemens.reimann@ngu.no (C. Reimann), robert.garrett@canada.ca (R.G. Garrett).

http://dx.doi.org/10.1016/j.scitotenv.2017.06.063
0048-9697/

1. Introduction

Compositional data (CoDa) are characterized by the fact that they are part of a whole. They convey relative information and are usually reported in a relative unit like wt%, mg/kg or μg/L. The true information of compositional data is carried by the ratios between their components. For data analysis it does not matter whether the whole composition is known (analysed) or not. The compositional nature of the data is inherent in the aim of the analysis (relative structure of a whole), and demonstrated by their relative unit representation. In geochemistry practically all analytical results are thus compositional. The calculation of correlation coefficients is a very popular technique in the statistical analysis of geochemical (environmental) data as a first description of pairwise variable associations.

Correlation analysis estimates the strength of the relationship between any pair of variables. The covariance is a measure of this relationship and depends on the variability of each of the two variables. Because covariances can take any number only the sign (+ or −) is informative, the strength of the relation between the two variables, however, cannot be interpreted. To obtain standardised numbers, the correlation coefficients, it is necessary to eliminate the dependency on the variability of each variable.

Three widely used methods to calculate a correlation coefficient are named after the people that first proposed them: Pearson, Spearman and Kendall (Galton, 1889, 1890; Spearman, 1904; Kendal, 1938). All these methods result in a number between −1 and +1 that expresses how closely the two variables are related, ±1 shows a perfect 1:1 relation (positive or negative) and 0 indicates that no systematic relationship exists between the two variables. A correlation of ±0.5 often indicates a significant relationship, but this depends on the number of samples taken. Relationships between two variables are best visualised in a scatterplot. However, when working with many variables, scatterplots occupy a lot of physical space. Reducing the information of multiple scatterplots to one number per plot may simplify studying the relationships (similarly to using location and spread measures to characterise the data distribution). If correlation analysis results are presented without a name for the method, usually the Pearson correlation coefficient has been estimated, which is a measure for relationship.

The Spearman rank method (Spearman, 1904) provides a non-parametric (distribution free) measure of correlation between two variables. It does not measure linear relation, but estimates if the association is monotonically (steadily) increasing or decreasing. Searching for a monotonic relationship is far more general, and less restrictive, than searching for a linear one. However, in the (rare) case of a bi- or multivariate normal distribution the Pearson method performs better because it is more precise. Again a correlation coefficient of ±1 indicates a perfect monotonic relationship. In the Spearman coefficient it is the ranks of the sorted values that determine the result, not the actual data-values. Thus the data are first ranked (sorted), and the Pearson correlation of the ranks of the data is then computed. This is the reason that the Spearman rank correlation coefficient is relatively robust against data outliers. One of the important advantages of the Spearman rank correlation is that the results will be the same for the original data and for any strictly monotonic transformation, as such a transformation does not disrupt the order of the data-values from lowest to highest. Thus, log-transformation does not change the Spearman rank correlation.

The Kendall-tau method (Kendal, 1938) is quite similar to the Spearman rank method. It also measures the extent of monotonically increasing or decreasing relationships between the pairs of variables. However, it uses a different method of calculation (looking at the sign of the slope of the line connecting each existing pair of points, summing the signs, and dividing the result by the number of pairs). The method is relatively robust against data outliers – as long as the sign of the slope does not change the result will stay the same. Thus Kendall-tau is independent of the actual values of the data, and a strictly monotonic transformation will not alter the estimated correlation coefficient. The calculation of the Kendall-tau correlation cannot be visualised in an easy graphic. To plot the Kendall-tau correlation would require connecting all possible pairs of data points by lines and to study their slopes.

When dealing with compositional data it is a general question, does correlation analysis of the raw data makes sense or not? Problems with classical correlation analysis were described >100 years ago (spurious correlation: Pearson, 1897). It took the geoscience community a very long time (with one noteworthy exception: Chayes, 1960) to realize that correlation analysis of compositional data should not be based on the raw or log-transformed data. An appropriate statistical analysis of compositional data based on log-ratios was finally described by Aitchison (1986), unfortunately at a level that most geoscientists will not be able to follow. The central idea is to express compositional data in real variables (coordinates) that capture their specific features. From the geometrical perspective, it is indeed about different *coordinate representations*, however, the environmental (geochemical) communities are rather still used to refer to *transformations*. For multivariate data analysis many solutions have in the meantime been presented in the literature (e.g., Aitchison and Greenacre, 2002; Buccianti and Pawlowsky-Glahn, 2005; Buccianti et al., 2006; Egozcue and Pawlowsky-Glahn, 2011; Filzmoser and Hron, 2008; Filzmoser et al., 2009; Hron et al., 2010; Otero et al., 2005; Pawlowsky-Glahn and Buccianti, 2002, 2011; Pawlowsky-Glahn et al., 2015; Tolosana-Delgado and van den Boogaart, 2011; von Eynatten et al., 2003). The consequences of working with compositional data in univariate data analysis have been discussed by Filzmoser et al. (2009). Filzmoser et al. (2010) presented solutions for the bivariate analysis of compositional data – with the exception of a good solution for correlation analysis in the traditional sense of positive and negative associations between the elements. Reimann et al. (2012) provided a worked example of compositional data analysis for a given major element data set of European agricultural soil. These authors pointed out that while many solutions for multivariate data analysis already exist, a good replacement for the classical correlation analysis, taking care of the compositional nature of geochemical data, was still missing and that substantial problems related especially to the appropriate bivariate representation of compositional data still remain unsolved. Following the suggestions of Aitchison (1986), a method built around the variance of the pairwise log-ratios was suggested for measuring the strength of the proportionality between the elements (Filzmoser et al., 2010). This solution, however, could not really replace the classical correlation analysis in the eyes of the geochemical community because it does not allow to distinguish between positive and negative associations. Thus geochemists continue to use improper preprocessing of the data like a simple log-transformation in order to use the classical approach.

A better mathematical solution to the problem has recently been presented by Kynčlová et al. (2017) using an approach based on a log-ratio transformation, called symmetric coordinates. To demonstrate how the method works in practise, a recently published data set from a regional soil geochemical mapping program (Reimann et al., 2015a; Reimann et al., 2015b; Reimann et al., 2016) will be used to demonstrate and discuss the differences between a classical correlation analysis and results based on the use of this new technique. It will be shown how classical correlation analysis based on log-transformed data can lead to erroneous conclusions about the relation between any two variables in the data set because the influence of all other variables is neglected. It will further be demonstrated that using symmetrical coordinates the relations between any two variables in the two datasets can be more reliably interpreted in terms of geochemical processes.

In addition, visualizations in form of heat-maps allowing to easier grasp the results of correlation analysis will be presented.

2. Material and methods

2.1. Data set

For the purpose of this paper a well published dataset from Nord-Trøndelag, central Norway is used (Reimann et al., 2015a; Reimann