Science of the Total Environment xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Science of the Total Environment

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

Univariate statistical analysis of environmental (compositional) data: Problems and possibilities

Peter Filzmoser [a,*], Karel Hron [b], Clemens Reimann [c]

[a] Institute of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Wien, Austria
[b] Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, CZ-77100 Olomouc, Czech Republic
[c] Geological Survey of Norway, N-7491 Trondheim, Norway

ARTICLE INFO

Article history:
Received 8 May 2009
Received in revised form 24 July 2009
Accepted 5 August 2009
Available online xxxx

Keywords:
Compositional data
Closure problem
Univariate statistical analysis
Exploratory data analysis
Log transformation

ABSTRACT

For almost 30 years it has been known that compositional (closed) data have special geometrical properties. In environmental sciences, where the concentration of chemical elements in different sample materials is investigated, almost all datasets are compositional. In general, compositional data are parts of a whole which only give relative information. Data that sum up to a constant, e.g. 100 wt.%, 1,000,000 mg/kg are the best known example. It is widely neglected that the "closure" characteristic remains even if only one of all possible elements is measured, it is an inherent property of compositional data. No variable is free to vary independent of all the others.

Existing transformations to "open" closed data are seldom applied. They are more complicated than a log transformation and the relationship to the original data unit is lost. Results obtained when using classical statistical techniques for data analysis appeared reasonable and the possible consequences of working with closed data were rarely questioned. Here the simple univariate case of data analysis is investigated. It can be demonstrated that data closure must be overcome prior to calculating even simple statistical measures like mean or standard deviation or plotting graphs of the data distribution, e.g. a histogram. Some measures like the standard deviation (or the variance) make no statistical sense with closed data and all statistical tests building on the standard deviation (or variance) will thus provide erroneous results if used with the original data.

1. Introduction

A classical example for a closed array or closed number system is a data set in which the individual variables are not independent of each other but are related by being expressed as a percentage or parts per million—as almost all environmental data are. Compositional data have been historically defined as summing up to a constant, but nowadays they have a broader definition, as they are considered to be parts of a whole which only give relative information (see Buccianti and Pawlowsky-Glahn, 2005, for an example). This definition thus also includes data that do not sum up to a constant. The problems of undertaking statistical analyses with "closed number systems" have been discussed much in specialized literature for more than 30 years, mostly in connection with multivariate data analysis (e.g. Chayes, 1960; Butler, 1976; Le Maitre, 1982; Woronow and Butler, 1986; Aitchison, 1986, 2008). However, the mathematical formalism is difficult and the consequences of using classical statistics for compositional data have thus never reached the wider environmental community. Data closure has often been treated as a topic for mathematical freaks, and intuitively it has been stated that this issue might have consequences only for multivariate data analysis or if major elements are considered in data analysis. Using practical examples, this paper demonstrates what happens if classical statistical methods are applied indiscriminately to environmental data in the simple univariate case.

The first step in statistical data analysis of environmental data should be to "look" at the data with appropriate graphical tools (Reimann et al., 2008). Typically, a histogram is inspected in order to obtain an idea about the data distribution, or a boxplot is drawn to show the median, skewness and tailedness of the distribution and to identify data outliers. In addition it will be of interest to estimate mean, variance, and probably further statistical data summary measures that characterize the observed data.

One basic question when performing these standard tasks is whether the original data or transformed data should be used. In environmental sciences many data are strongly right-skewed, a histogram of the original data may be almost uninformative due to the presence of some extreme outliers. Calculating the arithmetic mean for right-skewed data will result in a biased (too high) estimate

[*] Corresponding author. Tel.: +43 1 58801 10733; fax: +43 1 58801 10799.
E-mail addresses: P.Filzmoser@tuwien.ac.at (P. Filzmoser), hronk@seznam.cz (K. Hron), Clemens.Reimann@ngu.no (C. Reimann).

0048-9697/$ – see front matter ©2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2009.08.008

of the central value. If a log transformation is used for such data the histogram of a right-skewed distribution will be much closer to symmetry or may even show the shape of a normal distribution. Since the normal distribution plays an important role in classical statistical estimation, it will be tempting to now compute the arithmetic mean for the log-transformed data, and transform the results back to the original data scale to obtain a better-suited estimate of the central value than the arithmetic mean of the original data.

For univariate data it is always possible to use a transformation, like the Box–Cox transformation (Box and Cox, 1964), which brings the data (majority) as close as possible to normality. Location estimates can then be computed for the transformed data, and back-transformed to the original scale. Even outlier boundaries (Reimann and Garrett, 2005) can be calculated using, for example, a log-scale and can then be back-transformed to the original data scale.

However, although statistical requirements of data symmetry or normality may appear to be fulfilled following such a transformation, it is questionable whether this approach is meaningful given the compositional nature of the data. If each sample was analyzed for all possible chemical elements, the concentrations would sum up to 100%, or to 1,000,000 mg/kg. Thus the constraint of constant sum has to have certain consequences, even for the statistical analysis of a single element. This implies that compositional data can never be seen as truly univariate data, even if only one component is measured. The scientist always observes one component and the remainder, being the composition of all remaining components. If all variables were measured one could omit one data dimension (variable) without any loss of information due to the constant-sum constraint. As a consequence, the data are not represented in the Euclidean space, and thus Euclidean geometry is inappropriate for such data. In fact, this kind of data is known under the name *compositional data* or *closed data* (Aitchison, 1986), and the constant-sum constraint implies a special geometry, the so-called Aitchison geometry on the simplex, being appropriate for the simplex sample space of the data.

Euclidean geometry plays an important role, even in univariate statistical data analysis. Already for the apparently simple construction of a histogram one counts the number of data points falling into certain intervals with equal length, measured by the Euclidean distance. Therefore, a histogram may not reveal the true distribution as inherent in the data. Another typical example of the use of Euclidean geometry in statistical calculations is the arithmetic mean. It is the value with the smallest overall sum of squared Euclidean distances to each data point. If Euclidean geometry is not valid, the arithmetic mean is quite likely to be a poor estimate of the data center.

Generally speaking, the decision to transform the available data, and which transformation to use, should be based on the assumed geometry inherent to the data, and not on the shape of the histogram. The geometry inherent to the data should be chosen using criteria based on data scale and interpretability. If in that scale the shape of the histogram clearly deviates from normal distribution, more appropriate statistical estimators and models than classical (Gaussian) statistics need to be used.

This paper investigates the question of how to approach the univariate analysis of compositional data and what effect different transformations will have on the results. The log transformation is of special interest because it is most frequently used for analyzing environmental data (Reimann et al., 2008).

2. Transformations for compositional data

Compositional data are surprisingly frequent, e.g. in environmental sciences, geochemistry, chemistry, biology, technical sciences, or in official statistics. They are defined as compositional parts (variables, elements) that are positive and sum up to a constant c (after rescaling the data or if all possible parts are measured), usually chosen as 1 or 100, in case of percentages. More formally, an observation $\mathbf{x}=(x_1,\ldots,x_D)$ is, by definition, a D-part composition if, and only if, all its D components are strictly positive real numbers, and if all the relevant information is contained in the ratios between them (Aitchison, 1986). Then the former definition is a consequence of this more formal one. It is tempting to argue that for univariate data analysis, there is only $D=1$ part, and thus that this concept can be ignored. However, the values of the considered element are parts of the whole, and the information that has to be analyzed is the proportion of the element in the complete sample, forming 100%. The data unit (mg/kg, ppm—parts per million, etc.) expresses this fraction, and the values would not change if the sample had twice the volume (assuming equal distribution of the element in the sample). The above definition must also be considered for univariate data. Even a single variable carries, in fact, all the relevant information in the ratio on the whole.

If the values of only one variable are available or of interest, it is actually unclear what “ratio” means. An observation $\mathbf{x}=(x_1,\ldots,x_D)$ would thus have to be considered as a composition with $D=2$ parts, namely the part that has been observed and the remaining part, including the contributions of all the other variables. Summarizing the information of several variables in a single part is called amalgamation (Aitchison, 1986). Amalgamation is a non-linear projection in the Aitchison geometry, and the result will in general depend on which variables are summarized (for more details see Discussion). For this the notation, $\mathbf{x}=(x,1-x)$ can thus be used, assuming that the part x is given as a fraction of the whole with constant sum $c=1$ (this can be easily adjusted if the unit is %, mg/kg, etc.).

Various possibilities for data transformation of compositional data have been introduced in the literature; the most widely used is the family of one-to-one logratio transformations (Aitchison, 1986). The additive logratio (alr) transformation considers log transformations of the ratios formed by the compositional parts. For a D-part composition $\mathbf{x}$ the alr transformation is defined as:

$$alr(\mathbf{x}) = \left(\ln\frac{x_1}{x_j},\ldots,\ln\frac{x_{j-1}}{x_j},\ln\frac{x_{j+1}}{x_j},\ldots,\ln\frac{x_D}{x_j}\right) \tag{1}$$

where j is an element of the set $\{1,\ldots,D\}$. Thus, one part with index j is selected as the denominator for building the logratios. The alr transformation has often been criticized as being subjective, since the results depend on the choice of the part that is used as denominator (Aitchison, 1986). This subjectivity of alr is avoided by the centered logratio (clr) transformation, defined for a D-part composition $\mathbf{x}$ as:

$$clr(\mathbf{x}) = (y_1,\ldots,y_D) = \left(\ln\frac{x_1}{\sqrt[D]{\prod_{i=1}^{D}x_i}},\ldots,\ln\frac{x_D}{\sqrt[D]{\prod_{i=1}^{D}x_i}}\right). \tag{2}$$

In the case of $D=2$ parts the formula for alr is:

$$alr(\mathbf{x}) = \ln\left(\frac{x}{1-x}\right) = \ln(x)-\ln(1-x). \tag{3}$$

Thus, the choice of the denominator in alr is unique. The clr transformation would even yield two variables, the logratio of each part with the geometric mean—a further problem for univariate data analysis. Isometric logratio (ilr) transformations (Egozcue et al., 2003) are another useful class of logratio transformations with good theoretical properties. For a D-part composition $\mathbf{x}$, an ilr transformation can be chosen as $\mathbf{z}=(z_1,\ldots,z_{D-1})=\text{ilr}(\mathbf{x})$ with:

$$z_i = \sqrt{\frac{D-i}{D-i+1}}\ln\frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D}x_j}},\quad \text{for } i=1,\ldots,D-1. \tag{4}$$

Please cite this article as: Filzmoser P, et al, Univariate statistical analysis of environmental (compositional) data: Problems and possibilities, Sci Total Environ (2009), doi:10.1016/j.scitotenv.2009.08.008

Note that **z** has only $D-1$ components. For $D=2$ parts the result is:

$$z = ilr(\mathbf{x}) = \sqrt{\frac{1}{2}}\ln\left(\frac{x}{1-x}\right), \tag{5}$$

resulting in a univariate variable.

Here alr and ilr transformation gives a very similar result, differing only by the factor $\sqrt{1/2}$. Note that the result is also similar to the logit-transformation, which is defined with the factor 1/2, and which is mainly used for data consisting of proportions (Johnson and Wichern, 2007).

The factor is not important for univariate data analysis: the form of the histogram does not change, boxplot boundaries remain unchanged, and the back-transformed arithmetic mean will be exactly the same. The major difference is in the theoretical properties. The most important property of the ilr transformation is its *isometry*, meaning that it relates the geometry on the simplex directly to the usual Euclidean geometry (Egozcue and Pawlowsky-Glahn, 2006).

Fig. 1 shows four compositional data points, drawn on the left vertical axis. The distance between the points 0.1 and 0.2 is the same as the distance between 0.5 and 0.6, namely 0.1. This, however, is the Euclidean distance which is not meaningful for compositional data: the proportion 0.2 is twice as much as 0.1, while 0.6 is greater than 0.5 only by a factor of 1.2. This difference has to be reflected by an appropriate distance measure. Using the information of the remaining parts leads to the horizontal axis "$1-x$" in Fig. 1. The data points can be projected to the bivariate space with the constant-sum constraint, i.e. to the dashed-dotted line. This is, in fact, the correct presentation, because all the information on data points and constraints is included. Working in this space (along the dashed-dotted line) corresponds to working in the Aitchison geometry, the simplex sample space. Clearly, the Euclidean distance is still not appropriate, because boundary points behave differently from points in the center. An appropriate distance measure in this geometry is the Aitchison distance (Aitchison et al., 2000), defined for two compositions $\mathbf{x}=(x_1,\dots, x_D)$ and $\mathbf{y}=(y_1,\dots, y_D)$ as:

$$d_A(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(\ln\frac{x_i}{x_j}-\ln\frac{y_i}{y_j}\right)^2}. \tag{6}$$

In the example discussed, two compositions are considered, $\mathbf{x}=(x,1-x)$ and $\mathbf{y}=(y,1-y)$, and the Aitchison distance is:

$$d_A(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{2}}\left|\ln\left(\frac{x}{1-x}\right)-\ln\left(\frac{y}{1-y}\right)\right|, \tag{7}$$

Accordingly, the original values 0.1 und 0.2 lead to an Aitchison distance of 0.57, while the values 0.5 and 0.6 result in exactly half this distance.

The log transformation (axis "$\ln(x)$" in Fig. 1) is a first approach to emphasizing the different behaviors of the two pairs of data points. Thus, the transformed points 0.5 and 0.6 appear to be much closer than the other data pair. However, the distances 0.18 and 0.69 (measured in the Euclidean sense) differ substantially from the Aitchison distances.

Building the ratio "$x/(1-x)$" is a further step: the logarithm of this ratio is especially interesting and is visualized in Fig. 1 (right). It can be seen that – aside from the factor $\sqrt{1/2}$ – the Euclidean distance between two points corresponds exactly to formula (7) for the Aitchison distance. It would be exactly the same if the ilr transformation (5) was used instead of "$\ln(x/(1-x))$". This equality is called the isometric property of the ilr transformation, namely that for any two compositions **x** and **y** the relation:

$$d_A(\mathbf{x},\mathbf{y}) = d_E(ilr(\mathbf{x}), ilr(\mathbf{y})) \tag{8}$$

holds, where d_E denotes the Euclidean distance. Thus, of the indicated transformations only the ilr transformation yields an appropriate transformation from the Aitchison geometry to Euclidean space, where the Euclidean distance measure reflects the distances from the Aitchison geometry.

Omitting the scaling factor $\sqrt{1/2}$ from the ilr transformation (5) results in formula (3) for the alr transformation. Although the isometric property, represented by Eq. (8), is no longer fulfilled with alr, this has no consequences for univariate data analysis. However, for bivariate or multivariate data analysis this issue becomes much more important. The ilr transformation can be easily extended according to Eq. (4) to the multivariate case, and it keeps all the advantageous properties, while an extension of the alr transformation according to Eq. (1) has severe shortcomings (Pawlowsky-Glahn et al., 2008).

Thus the ilr transformation yields a correct representation of compositional data in Euclidean space where standard statistical methods can be applied. In the univariate case, however, the alr or the logit-transformation can also be used: a scaling factor is the only difference in relation to the ilr transformation. The log transformation leads to a geometrical representation that is inconsistent with the Aitchison geometry, and statistical methods applied in this space can be expected to provide erroneous results.

3. Consequences for univariate statistical graphics

3.1. Histogram

One of the most popular statistical graphics is the histogram. The histogram provides knowledge about the statistical distribution of the investigated variable. This knowledge is important for selecting appropriate estimators for center and spread, or for the choice of appropriate statistical tests (parametric or nonparametric).

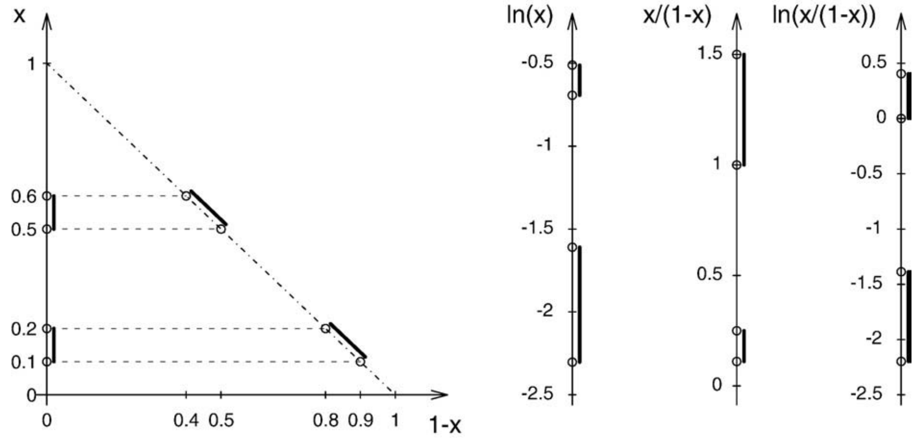

Fig. 1. Four data points on axis "x" with indicated distance between two pairs of points. Different transformations are used, which change the distance between the pairs.

The histogram is usually constructed by selecting equidistant intervals via the Euclidean distance measure, and counting the number of data points falling into each interval. Since only the ilr transformation allows for a correct geometrical presentation of compositional data in Euclidean space, the histogram has to be constructed with the ilr-transformed data. Alternatively, the alr transformation is suitable – although not consistent with the Aitchison geometry – since the form of the resulting histogram remains unaltered.

Fig. 2 shows the concentration of SiO_2 in European subsoils (Salminen, 2005). The left column of the figure presents a collection of plots from Exploratory Data Analysis (EDA), including the histogram. The figure shows three different representations of the data: the original SiO_2 data (top), the log-transformed data (middle), and the ilr-transformed data (bottom). For the transformed data the original data scale is provided on top of the plots. While the original data are left-skewed, the log-transformed data are even more left-skewed. The ilr transformation results in a more symmetric image of the data, but the form of the histogram is clearly different from a normal distribution. Several data points far away from the center might be considered as outliers. Thus, when estimating center and spread of the data it is advisable to use robust estimates.

3.2. Density trace

The density trace can be considered as a smooth form of the histogram (Scott, 1992). For its construction the selection of a certain band-width or window width and a weight function is necessary. When estimating the density at a certain point, the window is centered at this point and all values falling in this range are considered for the estimation, with influence according to the weight function. The weight function is symmetric, in the easiest case it is a rectangular function, and thus the whole concept of density estimation is based on symmetric distances relying on Euclidean geometry.

The histograms shown in Fig. 2 are combined with density traces. The ilr-transformed data allow a realistic view of the underlying density function, the two other presentations are in an inappropriate geometry.

3.3. Boxplot

Several different definitions of the boxplot exist in the literature. Here the original definition given by Tukey (1977) will be used. The boxplot is mainly based on sorted data (median, medians of the halves), and their position remains unchanged under log transformation or

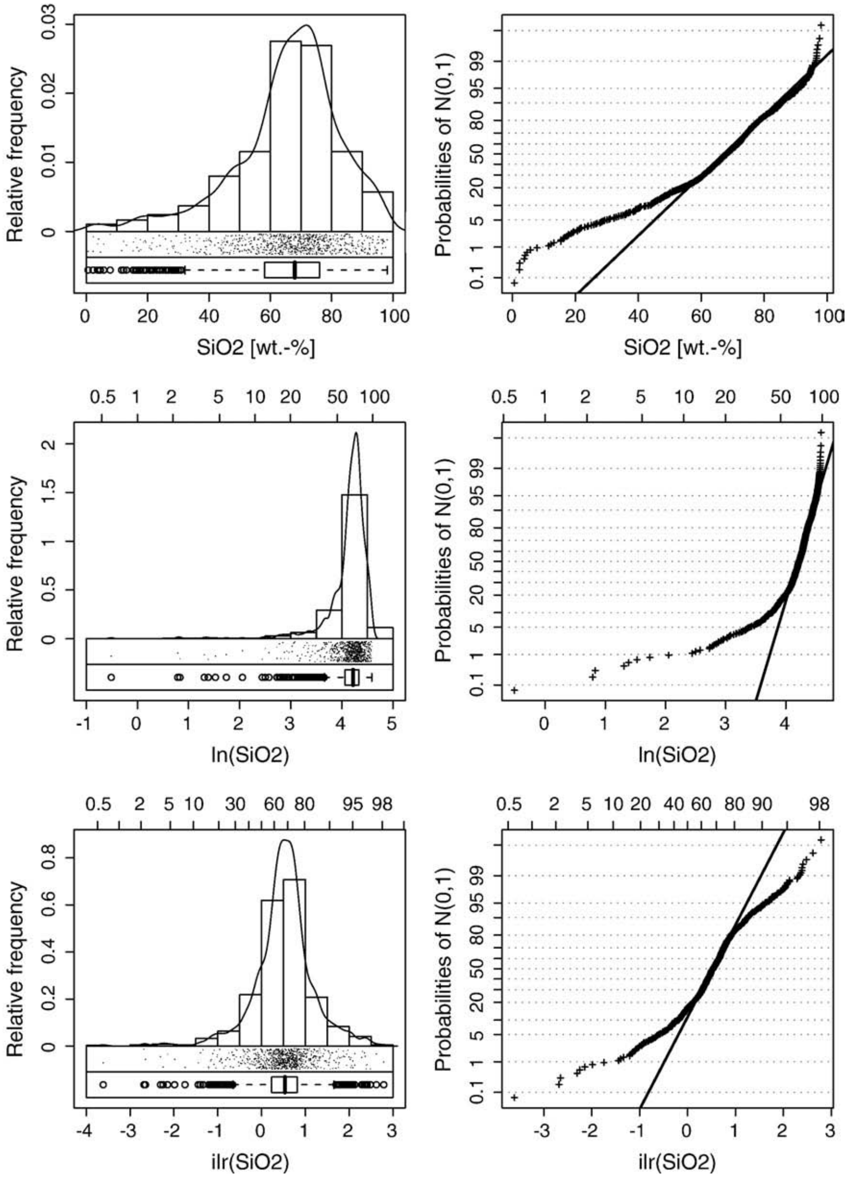

Fig. 2. EDA plots (left column) and CP-plots (right column) of SiO_2 (measured in wt.%) in subsoils of Europe. Upper row: original data; middle row: log-transformed data; bottom row: ilr-transformed data.

logratio transformations. However, the definition of the inner fence is based on the concept of symmetry, relying on Euclidean geometry. Therefore, a correct indication of potential outliers is only possible following an ilr (or alr) transformation. The boxplots shown in Fig. 2 for the original, the log-transformed and the ilr-transformed data differ considerably in the position of the whiskers, and accordingly, in the number of indicated outliers. Only the ilr transformation reveals potential upper outliers. Although the position of the median and the values defining the box do not change, their representation in the correct geometry is important. For the ilr-transformed data the median appears in the middle of the box, indicating the symmetry of the inner half of the data, while in an inappropriate geometry the non-centered median in the box would be interpreted as asymmetric data behavior.

3.4. Plots of the distribution function

The empirical cumulative distribution function (ECDF) is a step function with jumps of height $1/n$ at each of the n data points. This function is thus easy to construct. It has favorable statistical properties, because it converges (for n to infinity) to the theoretical underlying cumulative distribution function (e.g., Ross, 2002). Convergence, however, is only guaranteed in Euclidean geometry but not in the simplex. Thus, as in the previous graphical representations, the ilr-transformed variable, yielding a correct presentation of the data in Euclidean space, has to be used for the ECDF plot.

The ECDF plot of normally distributed data shows an S-shape. It is therefore difficult to judge if they deviate from this typical S-shape for the data at hand. Thus it is more advisable to transform the probability scale in such a way that normally distributed data would lie on a straight line in the plot. Deviations from a line are much easier to detect than deviations from an S-shape. The requested transformation takes quantiles of the hypothetical normal distribution, and one can directly scale either according to the quantiles (quantile–quantile (QQ-) plot) or according to the probabilities (cumulative probability (CP-) plot), see Reimann et al. (2008).

The comparison with probabilities of the (standard) normal distribution in the CP-plots in Fig. 2 (right) shows that none of the plots indicates normal distribution because of systematic deviations from the straight line. Even the ilr-transformed data show severe deviations in the tails of the distribution. Use of statistical methods or tests that assume normal distribution should thus be avoided.

4. Consequences for univariate summary statistics

Summary statistics are the key characteristics that describe the distribution of a variable. Usually measures of center and spread of a variable are most important. Whether additional information, like skewness and kurtosis, certain quantiles or a test for normality is needed, depends on the context.

4.1. Center of the distribution

The most common way of estimating the center of a distribution is to use the arithmetic mean. This estimator is known to be the best linear unbiased estimator (BLUE) of the "center" of the underlying theoretical distribution in terms of Euclidean geometry. The arithmetic mean is not only easy to compute but is also a meaningful approximation of the "true" center of the distribution. However, using the arithmetic mean needs care:

(a) The arithmetic mean should not be directly applied to compositional data because it relies on Euclidean geometry.
(b) In environmental sciences it is rarely possible to assume a "true" underlying data distribution, rather a mixture of several distributions caused by several sub-populations, and/or artefacts will govern the data distribution.

With respect to (a), a popular strategy is to compute the arithmetic mean for the log-transformed data, and to back-transform the result. This approach corresponds to computing the geometric mean for the original compositions. Although the direct use of the geometric mean in the simplex sample space appears to be meaningful, the remaining part(s) is ignored. The resulting estimate does not form a composition because it is not adjusted to the remaining part(s) (the geometric mean of the variable under consideration and the geometric mean of the rest have to sum up to 1 in order to form an appropriate composition).

These problems are avoided when using a logratio transformation, preferably the ilr transformation (5). After the computation of the arithmetic mean, say $\bar{z}$, of the ilr-transformed data, it is possible to back-transform the result to the original space with:

$$\bar{x} = \frac{\exp(\sqrt{2}\,\bar{z})}{\exp(\sqrt{2}\,\bar{z}) + 1}. \quad (9)$$

This corresponds to the geometric mean of the original data sample x, but adjusted to the remaining part, $1-x$. The result is again a BLUE—here in the sense of the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2002). Note that in the univariate case the same result would be obtained by using the alr- or logit-transformation.

Referring to (b) the task is not to estimate the location parameter of an underlying distribution, but rather to estimate the center in a robust way with less influence of outliers. Using the median for this purpose has the advantage of obtaining a highly robust estimate which, in this special case, is also not affected by the mentioned logratio transformations (the position of the median remains the same under strictly monotone transformations). Note that again, for the computation of the median, the remaining part $1-x$ is ignored. The correct procedure would consider both parts for determining the median, and transform the result back. However, since in this case the sum of both parts is 1, the result is the same as for computing the median directly. In a more general situation (more compositional parts, not constant sum), the direct computation of the median can be misleading. Other robust estimators for location, like M-estimators (Maronna et al., 2006), have to be computed for the ilr-transformed data, and back-transformed to the original space. Such estimators can also be used with more than two compositional parts.

As an example, three selected elements (Na_2O, SiO_2, Ni) in the Kola C-horizon data (Reimann et al., 2008) are considered. The data are visualized in Fig. 3 in the form of EDA plots. The distribution of Na_2O (left column) is left-skewed, and it remains left-skewed when using the log or the ilr transformation. SiO_2 and both of its transformations (middle column) are relatively symmetric. In contrast, Ni is heavily right-skewed, but both ln(Ni) and ilr(Ni) look symmetric.

Estimations of the center of the three elements are provided in Table 1. "mean" is the arithmetic mean of the original data, while "mean-log" and "mean-ilr" stand for the arithmetic mean of the log- and ilr-transformed data, respectively, back-transformed to the original space. "median" is for the median of the (original) data.

The results in Table 1 reflect the non-robustness of the arithmetic mean, because outliers causing skewness of the distribution heavily attract the estimation. This is visible especially for the log- and ilr-transformed values of Na_2O, and for the original values of Ni, where the results differ substantially from the median. If the distribution is close to symmetry, as for SiO_2 and its transformations, all estimators give about the same result. This once more underlines that graphical inspection of the data distribution – preferably of the ilr-transformed data – is required before selecting an appropriate estimator.

4.2. Variance of the distribution

The variance, or its square root, the standard deviation or spread, characterizes the concentration of the data around the central value. The traditional estimator of the variance is the well-known sample

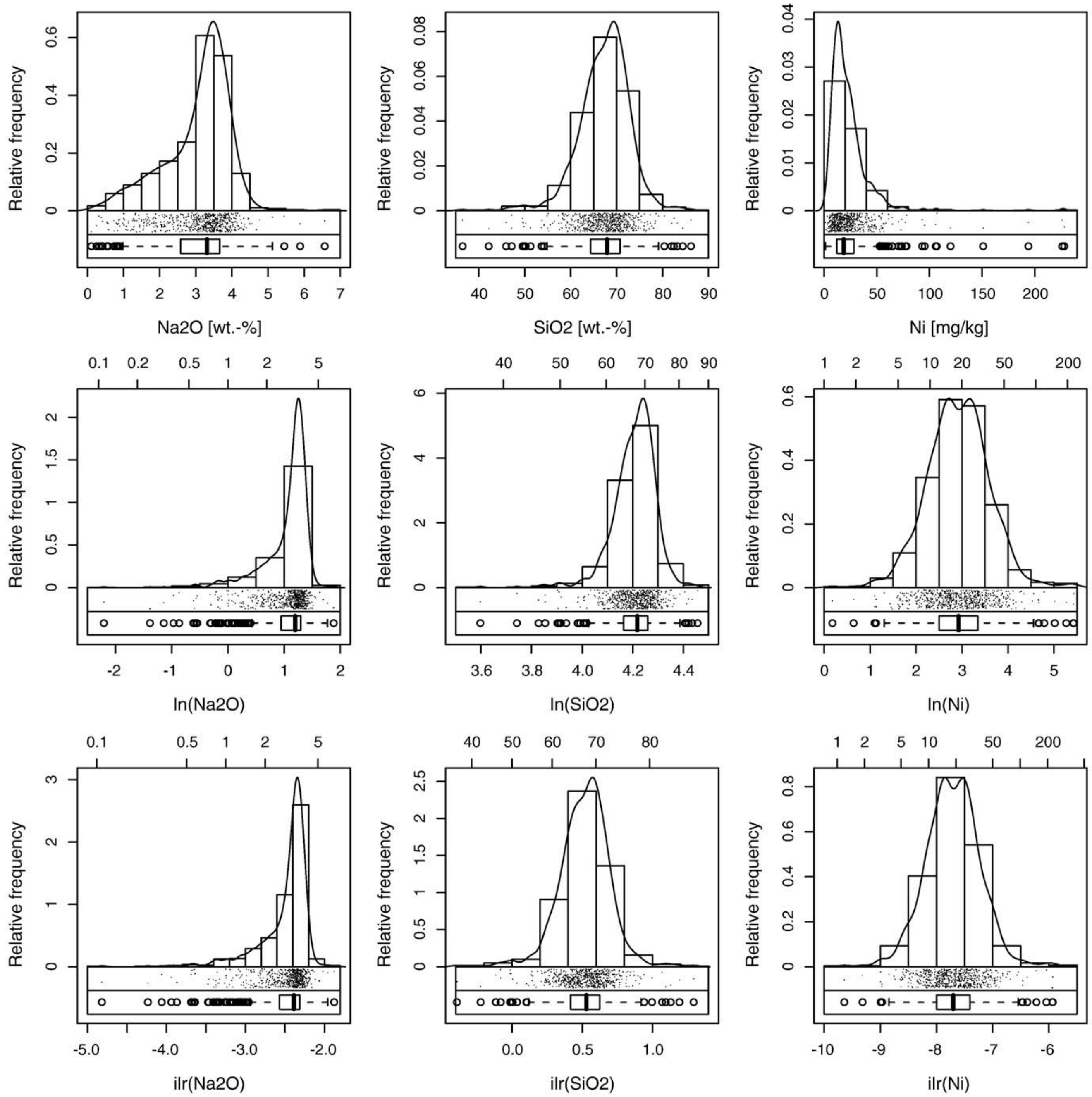

Fig. 3. EDA plots (Reimann et al., 2008) of Na_2O, SiO_2, and Ni in C-horizon soils of the Kola Peninsula, for the original, log-transformed, and ilr-transformed data.

variance which has good theoretical properties but relies on Euclidean geometry. A further difficulty in the context of compositional data is that it is no longer possible to compute the sample variance, e.g. for the log-transformed data, and to back-transform the result, because the log transformation changes the distances of the observations from the center asymmetrically.

Table 1
Results for three selected elements from the Kola C-horizon data: arithmetic mean of the original ("mean"), the log-transformed ("mean-log"), and the ilr-transformed ("mean-ilr") data.
Results for the transformed data are back-transformed to the original space.

	Mean	Mean-log	Mean-ilr	Median
Na_2O [wt.%]	3.05	2.84	2.85	3.31
SiO_2 [wt.%]	67.31	67.07	67.53	67.89
Ni [mg/kg]	23.40	18.46	18.46	18.60

With compositional data, any measure of spread should inform about the stability of the part x relative to the remainder $1-x$, i.e. about the variability of the ratio $x/(1-x)$. Here again the ilr transformation leads to a meaningful result, where the sample variance of the ilr-transformed data $z = ilr(\mathbf{x}) = \sqrt{\frac{1}{2}}\ln(\frac{x}{1-x})$ is computed.This concept is, in fact, a special case for $D=2$ compositional parts. In the more general case of $D \geq 2$ compositional parts, this estimator is known under the name *total variance estimator*, defined for a random composition $\mathbf{x} = (x_1,..., x_D)$ as:

$$totvar(\mathbf{x}) = \frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D} var\left(\ln\frac{x_i}{x_j}\right) \tag{10}$$

(Pawlowsky-Glahn et al., 2008). The total variance estimator is unbiased and converges in probability to the true variance of compositions around the center of their distribution (Hron and Kubáček, 2009). Instead of using the sample variance for "var" in Eq. (10) one can also take more robust alternatives, like the squared median absolute deviation (MAD). Note that the result does not depend on a data unit, and thus it can be considered as a measure of data homogeneity (stability). Small values indicate higher stability. In the case $D=2$ this refers to an approximately constant ratio between x and $1-x$.

The coefficient of variation, which is defined as standard deviation divided by the arithmetic mean, is often considered as a measure of precision as it expresses variability independent of the data measurement units. For compositional data such a measure is not needed (it would not even make sense), because instead the variance of the ilr-transformed data can be directly used.

For the above example (elements Na_2O, SiO_2, Ni of the Kola C-horizon, see Fig. 3), different measures of spread are computed. Table 2 shows the results for the empirical standard deviation (SD) and the MAD for the original (columns 1 and 2), the log-transformed (columns 3 and 4), and the ilr-transformed (columns 5 and 6) data. As already observed in Table 1, SD and MAD lead to very different results, especially for skewed (transformed) data. Although it is possible to apply SD and MAD to the original and log-transformed data, the results do not account for the compositional nature of the data. Thus the focus should be on the last two columns of Table 2, where the result of SD-ilr for Na_2O is unreliable due to lower outliers. Classical and robust estimators lead to comparable results in the other cases. Accordingly, the homogeneity (stability) of Na_2O and SiO_2 in the survey area is considerably higher than that of Ni.

4.3. Quantiles, percentiles

Quantiles and percentiles are based on order statistics, and as for the median, the order of the data values does not change under a log transformation or the considered logratio transformations. One can thus directly compute the desired quantiles or percentiles from the original compositional data. The interpretation is as usual, namely that the α-quantile (α %-percentile) is the value where a fraction α of the data is below and the fraction $1-\alpha$ is above this value. Note that the direct computation of quantiles and percentiles without using any transformation is only possible in this special case where x and the remainder is assumed to sum up to 1. In a more general situation (more than two parts) this procedure can be misleading (see discussion above for the median).

5. Accounting for other compositional parts

5.1. A second approach for univariate data analysis

The approach discussed so far assumes that for the variable of interest, say x_1, only the relationship to the remainder shall be investigated. If the values of additional variables $x_2,\ldots, x_D$ are known, and if the relationship of one variable to each of the other existing variables needs to be considered, an alternative approach to univariate data analysis can be used. This second approach can be seen as a generalization of the first approach from above for two reasons: (a) For a considered variable x_1 the remainder was assumed to be $1-x_1$ in the first approach, since nothing is known about other parts, or because the interest is only in a single part and its relative relationship to the rest of the whole. Now the remainder is subdivided into all known parts. (b) For the second approach it is not necessary that x_1 and the remaining parts sum up to 1, i.e. the information of all considered parts is incomplete. In this case, x_D can be defined as the remainder $1-(x_1+x_2+\ldots+x_{D-1})$ which provides the completed information to the whole.

Table 2
Results for three selected elements from the Kola C-horizon data: empirical standard deviation (SD) and median absolute deviation (MAD) of the original ("SD", "MAD"), the log-transformed ("SD-log", "MAD-log"), and the ilr-transformed ("SD-ilr", "MAD-ilr") data.

	SD	MAD	SD-log	MAD-log	SD-ilr	MAD-ilr
Na_2O	0.91	0.68	1.55	1.22	0.32	0.15
SiO_2	5.51	4.70	1.09	1.07	0.18	0.15
Ni	21.10	11.56	1.96	1.86	0.48	0.44

Due to the definition of compositional data, all the relevant information about x_1 is contained in the ratios to *each* of the remaining parts $x_2,\ldots, x_D$. Accordingly, this relative information for all remaining parts needs to be considered also for univariate data analysis. The ilr transformation from Eq. (4) can be used for this purpose. The ilr variable:

$$z_1 = \sqrt{\frac{D-1}{D}}\ln\frac{x_1}{\sqrt[D-1]{\prod_{j=2}^{D}x_j}}, \tag{11}$$

contains all the relative information between x_1 and $x_2,\ldots, x_D$, because none of $z_2,\ldots, z_{D-1}$ includes x_1. Note that this separation of the information into a single ilr variable is not possible for x_2 or any of the remaining parts, because there the relative information to x_1 is missing. If, on the other hand, x_2 is of interest for univariate analysis, one can use Eq. (11) by replacing x_1 with x_2. In this way, each compositional part can be expressed by a single ilr variable which can be used for univariate analysis. The resulting ilr variables can, however, not be used for multivariate analysis, because they do not correctly represent the multivariate data information.

This approach of constructing single ilr variables for the compositional parts seems to be similar to the result of the clr transformation, defined in Eq. (2). However, here a univariate analysis of the single clr variables is not reasonable. The clr variables together are, by definition, singular, $y_1+\ldots+y_D=0$, and thus any regular selection of the variables is not meaningful. Furthermore, when considering e.g. part x_1, the variable y_1 does not contain *all* the relative information on the remaining parts, since $y_2,\ldots, y_D$ also contain information about x_1 (they include x_1 in the denominator). This argument is valid for any of the parts.

5.2. Univariate graphics and data distribution

The ilr variable constructed according to Eq. (11) can be directly used for plotting the histogram, density trace, boxplot, or CP-plot, and it can also be directly used for tests referring to the data distribution. The results will, in general, be different from the first approach because now the relative information on each of the other observed parts is included. A further complication arises when it comes to visualizing the original data scale in the plots, compare Figs. 2 and 3 (legends on the top). For this purpose the inverse ilr transformation to Eq. (11) is needed, which is given by:

$$x_1 = \sqrt[D-1]{\prod_{j=2}^{D}x_j}\exp\left\{\sqrt{\frac{D-1}{D}}z_1\right\}. \tag{12}$$

Thus, as for the construction of z_1, information of all parts $x_2,\ldots, x_D$ is required for the back-transformation to the original scale.

5.3. Univariate summary statistics

The estimation of the univariate center is not straightforward. It is possible to compute the arithmetic mean of z_1, but it is not clear how to use Eq. (12) for back-transformation.

The problem is to find an appropriate adjustment of the result in order to accommodate the sum of the original parts which can, in general, be different from 1. One solution is to directly compute geometric means $g_1,..., g_D$ for the parts $x_1,..., x_D$, and adjust the result to a desired constant c by:

$$\bar{x}_1 = \frac{c}{\sum_{j=1}^{D} g_j} g_1. \tag{13}$$

A further approach is to take the median, but also then it is necessary to adjust the result. This can be done in a similar way to Eq. (13), which requires the computation of the medians for each variable. It should be noted that the median could be problematic because it is solely based on ordering of the variable of interest, and it does not account for the relationships to other variables. This is also the case for other order-based measures, like quantiles or percentiles.

Another procedure, accounting for even more information, is to estimate the mean of a part by using the complete ilr transformation of all parts $x_1,..., x_D$, see Eq. (4). For this multivariate data representation there is a unique inverse transformation, which can be used for the back-transformation of the arithmetic mean. The resulting value of this rather complex procedure, adjusted to a constant c, is the same as given by Eq. (13). A robust counterpart for estimating the center should also be based on the complete ilr space, see Filzmoser and Hron (2008). Robust location estimation (e.g. Maronna et al., 2006) can be done in this space, and the result needs to be back-transformed to the original space, adjusted by a constant.

As for the first approach for univariate data analysis, the variance can be estimated from the ilr variable z_1 from Eq. (11), and it serves as a stability measure. Small values of var(z_1) indicate approximately constant ratios to the parts $x_2,..., x_D$. However, although the total variance is the sum of the variances of the ilr variables, $\sum_{i=1}^{D} var(z_i) = totvar(\mathbf{x})$ (Pawlowsky-Glahn and Egozcue, 2001), it is not possible to relate it directly to the contributions of x_1 from the total variance,

$$totvar(\mathbf{x})|x_1 = \frac{1}{D}\sum_{i=2}^{D} var\left(\ln\frac{x_1}{x_i}\right).$$

For example, for $D=3$ it holds that:

$$var(z_1) = \frac{1}{3}var\left(\ln\frac{x_1}{x_2}\right) + \frac{1}{3}var\left(\ln\frac{x_1}{x_3}\right) - \frac{1}{6}var\left(\ln\frac{x_2}{x_3}\right)$$

Thus, for a small value of var(z_1) we need stability of x_1 to both x_2 and x_3, but at the same time instability between x_2 and x_3. The variance var(z_1) therefore accounts in some sense for multivariate stability.

6. Discussion

The problem of working with compositional data has been widely discussed (though often neglected) for multivariate data analysis (e.g. Chayes, 1960; Butler, 1976; Le Maitre, 1982; Woronow and Butler, 1986; Aitchison, 1986, 2008). In environmental sciences the problematic aspects of working with closed data even when using univariate data analysis have been ignored. Examples presented above demonstrate that the compositional nature of environmental data must be considered for practically all aspects of statistical data analysis. The inherent problem is that a compositional variable cannot be treated separately from the rest of the total composition of a sample. The relevant information is hidden in the ratios to the remainder and not in the analytical values themselves. An interpretation and statistical evaluation of the observed values are thus only meaningful if the relationship to the values of the remaining variables is taken into account. For univariate data analysis it is not quite clear how the remaining parts need to be considered, and which ratios should be used.

Two different approaches were introduced, both are based on the ilr transformations that allow work in standard Euclidean geometry which is the very basis for most statistical techniques. In the first approach the “remainder” to an observed variable x, is taken as $1-x$, while in the second approach the ratios to all remaining observed variables are considered. Thus, for the first approach, it is implicitly assumed that the remainder consists of all other variables that could be measured in a sample (therefore their sum with x is 1), even though these variables are not of direct interest. In the second approach there is also interest in the relationship to several other variables, which do not necessarily need to lead to a constant sum 1. Due to the above definition of compositional data, the second approach includes more detailed information but is mathematically more difficult to handle. The results from both approaches are, in general, different because summarizing the information of several variables in a single part, i.e. the amalgamation, is not coherent with the Aitchison geometry on the simplex. Therefore it is also important to note that one will not get compatible results when using the information given by a variable x compared to the remainder $1-x$, or compared to other parts, say y or z. For more discussion on this topic, see Mateu-Figueras and Daunis-i-Estadella (2008) and Egozcue and Pawlowsky-Glahn (2005).

For the practitioner it is difficult to judge which of the two approaches is more useful for the data at hand. The first approach can even be simplified if the data values of the variable of interest are small, say smaller than 0.1. In this case, $\ln(x/(1-x)) \approx \ln(x)$ because the denominator is close to 1. Accordingly, univariate graphical representations would be very similar for log-transformed and ilr-transformed data. Also the arithmetic mean, or a test for normality would give essentially the same result. This may thus present the preferable (easier) approach for univariate data analysis.

In practice, several variables have been measured, and often they do not sum up to 1. For the first approach the information on the additional measured variables is simply ignored. This loss of information has not only disadvantages, because other aspects need to be considered as well.

Many data sets – and this is especially true for environmental data – contain outliers, missing values, and values below the detection limit. The more variables measured, the higher is the chance that a particular (multivariate) observation is plagued with one of these data problems.

Thus, the more variables used as ratio variables in the second approach, the greater is the likelihood for inclusion of erroneous information. In this sense, additional information is not always an advantage, and it is recommended to rather remove variables including severe data problems. Because additional variables are used in the denominator of the relation there is a danger that unreliable small values can cause huge errors. Although the second approach accounts for more information, there are more technical difficulties for the univariate analysis due to the more complex procedure. The second approach may have to wait until we are able to analyze a sample completely, for all its components with sufficient precision and low enough detection limits. Even then, it will remain a challenge to interpret the results from a univariate statistical analysis, because all the other considered variables will have an influence that may not be easy to unravel. Thus it may be a better approach to use multivariate data analysis directly, to understand the relationships between the variables in multivariate space, based on the correct geometry (e.g. ilr-transformed data).

Plotting histograms or calculating arithmetic mean and standard deviation, and applying statistical tests to the raw data will lead to erroneous results. All these familiar procedures can however, be carried out with the ilr-transformed data, and afterwards back-transformed into the original data scale. The standard deviation, however, cannot be back-transformed. Thus, computing the standard deviation (or variance) for the original compositional data, or for transformed data (and back-transformed result) does not make any sense. The standard deviation is replaced by a unitless number, resulting from the ilr-transformed data, and it represents stability or homogeneity of the measurements.

Statistical data analysis techniques for environmental data often include the standard deviation as a central element (statistical tests, etc.). All results relying on the standard deviation can be misleading when applied in an inappropriate geometry.

For multivariate data analysis the ilr transformation from Eq. (4) should be used. This transformation has good theoretical properties and it represents the multivariate data structure in a new geometry. A simple log transformation, variable by variable, or any other transformation of the single variables is no longer sufficient.

With modern computing possibilities at hand (van den Boogaart et al., 2009; Templ et al., 2009) there is no longer any excuse for using mathematically wrong techniques in the analysis of environmental data.

Acknowledgement

The authors are grateful to the referees for helpful comments and suggestions. This work was supported by the Council of the Czech Government MSM 6198959214.

References

Aitchison J. The statistical analysis of compositional data. London, UK: Chapman and Hall; 1986. p. 416.

Aitchison J. The single principle of compositional data analysis, continuing fallacies, confusions and misunderstanding and some suggested remedies, http://hdl.handle.net/10256/706, 2008.

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V. Logratio analysis and compositional distance. Math Geol 2000;32:271–5.

Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Ser B 1964;26:211–52.

Buccianti A, Pawlowsky-Glahn V. New perspectives on water chemistry and compositional data analysis. Math Geol 2005;37:703–27.

Butler JC. Principal components analysis using the hypothetical closed array. Math Geol 1976;8:25–36.

Chayes F. On correlation between variables of constant sum. J Geophys Res 1960;65: 4185–93.

Egozcue JJ, Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. Math Geol 2005;37:795–828.

Egozcue JJ, Pawlowsky-Glahn V. Simplicial geometry for compositional data. In: Buccianti A, Mateu-Figueros G, Pawlowsky-Glahn V, editors. Compositional data analysis in the geosciences: from theory to practice. Bath, UK: Geological Society Publishing House; 2006. p. 67–77.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras F, Barceló-Vidal C. Isometric logratio transformations for compositional data analysis. Math Geol 2003;35:279–300.

Filzmoser P, Hron K. Outlier detection for compositional data using robust methods. Math Geosci 2008;40:233–48.

Hron K, Kubáček L. Statistical properties of the total variance estimator for compositional data. Technical Report 8/2009, Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 2009.

Johnson RA, Wichern DW. Applied multivariate statistical analysis, 6th ed. Upper Saddle River, New Jersey: Prentice Hall; 2007. p. 767.

Le Maitre RW. Numerical petrology. Amsterdam, The Netherlands: Elsevier Scientific Publishing Company; 1982. p. 281.

Maronna R, Martin D, Yohai V. Robust statistics: theory and methods. Toronto, ON: John Wiley & Sons Canada Ltd; 2006. p. 436.

Mateu-Figueras G, Daunis-i-Estadella J. Compositional amalgamations and balances: a critical approach, http://hdl.handle.net/10256/738, 2008.

Pawlowsky-Glahn V, Egozcue JJ. Geometric approach to statistical analysis on the simplex. Stoch Envir Res and Risk Ass 2001;15:384–98.

Pawlowsky-Glahn V, Egozcue JJ. BLU estimators for compositional data. Math Geol 2002;34:259–74.

Pawlowsky-Glahn V, Egozcue JJ, Tolosano-Delgado R. Lecture notes on compositional data analysis, http://hdl.handle.net/10256/297, 2008.

Reimann C, Garrett RG. Geochemical background—concept and reality. The Science of the Total Environment 2005;350:12–27.

Reimann C, Filzmoser P, Garrett RG, Dutter R. Statistical data analysis explained. Applied environmental statistics with R. Wiley, Chichester, UK, 2008, 343 pp.

Ross S. A first course in probability, 6th ed. Upper Saddle River, NJ: Prentice Hall; 2002. p. 528.

Salminen R (Chief editor). Geochemical atlas of Europe. Part 1: Background information, methodology and maps. EuroGeoSurveys, Espoo, Finland, 2005.

Scott DW. Multivariate density estimation. Theory, practice and visualization. New York: John Wiley & Sons, Inc; 1992. p. 317.

Templ M, Hron K, Filzmoser P. robCompositions: Robust estimation for compositional data, http://www.r-project.org, R package version 1.2, 2009.

Tukey J. Exploratory data analysis. Reading, Massachusetts: Addison-Wesley; 1977. p. 506.

van den Boogaart KG, Tolosana R, Bren M. compositions: Compositional data analysis, http://www.r-project.org, R package version 1.01-1, 2009.

Woronow A, Butler JC. Complete subcompositional independence testing of closed arrays. Comput. Geosci. 1986;12:267–79.

Please cite this article as: Filzmoser P, et al, Univariate statistical analysis of environmental (compositional) data: Problems and possibilities, Sci Total Environ (2009), doi:10.1016/j.scitotenv.2009.08.008