Science of the Total Environment 408 (2010) 4230–4238

Contents lists available at ScienceDirect

Science of the Total Environment

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

The bivariate statistical analysis of environmental (compositional) data

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

[a] Vienna University of Technology, Department of Statistics and Probability Theory, A-1040 Vienna, Austria
[b] Palacký University Olomouc, Faculty of Science, Department of Mathematical Analysis and Applications of Mathematics, CZ-77146 Olomouc, Czech Republic
[c] Geological Survey of Norway (NGU), PO Box 6315 Sluppen, N-7491 Trondheim, Norway

ARTICLE INFO

Article history:
Received 1 February 2010
Received in revised form 5 May 2010
Accepted 8 May 2010
Available online 9 June 2010

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

ABSTRACT

Environmental sciences usually deal with compositional (closed) data. Whenever the concentration of chemical elements is measured, the data will be closed, i.e. the relevant information is contained in the ratios between the variables rather than in the data values reported for the variables. Data closure has severe consequences for statistical data analysis. Most classical statistical methods are based on the usual Euclidean geometry — compositional data, however, do not plot into Euclidean space because they have their own geometry which is not linear but curved in the Euclidean sense. This has severe consequences for bivariate statistical analysis: correlation coefficients computed in the traditional way are likely to be misleading, and the information contained in scatterplots must be used and interpreted differently from sets of non-compositional data. As a solution, the ilr transformation applied to a variable pair can be used to display the relationship and to compute a measure of stability. This paper discusses how this measure is related to the usual correlation coefficient and how it can be used and interpreted. Moreover, recommendations are provided for how the scatterplot can still be used, and which alternatives exist for displaying the relationship between two variables.

1. Introduction

The consequences of working with closed data in univariate data analysis have recently been demonstrated (Filzmoser et al., 2009b). Most importantly, the classical standard deviation should not be calculated for the original variables and used in connection with closed data. Histograms and density traces are frequently used to study the structure of univariate data and to provide an idea about their distribution. Due to the special geometrical properties of compositional data – they are not represented in the standard Euclidean space but rather have curvilinear properties (from the Euclidean point of view) and plot on the simplex (Aitchison, 1986; Egozcue and Pawlowsky-Glahn, 2006) – these graphics will invariably provide a wrong impression of the data distribution when used with the raw, untransformed data. The best available transformation for compositional data (almost all data presented in environmental sciences are compositional) is the isometric logratio (ilr) transformation as suggested by Egozcue et al. (2003). It allows the correct representation of compositional data in Euclidean space (Filzmoser et al., 2009b), which is appropriate and necessary whenever distances between observations are of importance because it reflects the standard geometry everybody is used to. Only when following such a transformation is it possible to get a realistic impression of the underlying distribution of a single variable and to choose the appropriate statistical tools for further data analysis (e.g., robust methods if there are still outliers). Because environmental data are often very right skewed, the graphical inspection of the data has most often been done with log-transformed data. Filzmoser et al. (2009b) could show that the log transformation will provide results that are almost equal to the ilr transformation up to element concentrations of 10 wt.%. This fortunate property of the log transformation is, however, lost when entering bivariate data analysis.

This paper will investigate the consequences of working with compositional data during bivariate data analysis. Here the user is most often interested in plots showing the strength of the relationship between two variables or in the correlation between a pair of variables in a dataset. However, the traditional interpretation of these plots assumes an underlying Euclidean geometry. Just as in the univariate case, the compositional data must first be transformed in order to get an unbiased impression of the relationship between two variables. In many cases a simple log transformation of each variable is used before proceeding with a scatterplot of a variable pair, or prior to calculating the Pearson correlation coefficient. This procedure is incorrect for compositional data, because the log transformation results in an incorrect representation of the bivariate data. The problem was recognised many years ago in classical geochemistry, studying the relation between major and trace elements in rocks. The problem has been discussed in some text books (e.g. Rock, 1988, — "the problem looms particularly large in geology"), however, without presenting a real solution. Some authors have more intuitively tried to

* Corresponding author.
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 © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2010.05.011

avoid closure via plotting scattergrams of ratios rather than of single variables (e.g. Miesch et al., 1966; Pearce, 1968) and in Russia correlation diagrams in which major elements are expressed as ratios to the total O content of a rock were used even earlier (Podolsky, 1962; Ivanov, 1963). Such ratios are unconstrained, however, their common denominators introduce other correlation problems (Rock, 1988). Aitchison (1986) lays the foundations of how to work with compositional data, however, this work is written at a level that a non-mathematician will have difficulties to understand. Due to the fact that classical statistics are all that is taught in most undergraduate, earth and environmental sciences classes, the consequences of continuing with data analysis in the wrong geometry have not been fully understood and have never reached the wider geo- and environmental science community.

1.1. Transformations for compositional data

The key reference to the statistical analysis of compositional data is Aitchison (1986). Thus, for more than 20 years the nature of this special type of data and how to deal with it has been known but little is done about it. Some expert groups are aware of the problem and are working on it, but have difficulties in reaching the wider scientific community. One problem may be that not only the statistical treatment but also the vocabulary is somewhat different from the standard case: the term "variables" is replaced by "compositional parts", or simply by "parts". Accordingly, a D-part composition is a row vector $\mathbf{x}=(x_1, \ldots, x_D)$ in which all components are positive real numbers carrying only relative information. The latter property distinguishes this type of data from ordinary multivariate data, in which the information is absolute. For example, a person's height or weight is usually given by absolute numbers, while the concentration of chemical elements in the soil is provided as proportion of a whole, e.g. as parts per million (ppm). When measuring all components of the soil sample, they would sum up to 1 or 100%, or 1 million ppm. Due to this constant sum constraint inherent in the data, the geometrical space is not the usual (Euclidean) space, but the so-called simplex sample space (Aitchison, 1986; Egozcue and Pawlowsky-Glahn, 2006). It is important to realise that this is also valid if not all possible compositional parts are available, i.e. even when the sum of the compositional parts is smaller than one. In this case one speaks about subcompositions. The geometry of compositional data is nowadays known under the name Aitchison geometry. As a consequence, the distance between two observations is not measured by the Euclidean distance that is used in daily life, but by the so-called Aitchison distance (Aitchison et al., 2000). Standard statistical procedures, like drawing a histogram, or computing the arithmetic mean, have thus to be based on the Aitchison geometry (Filzmoser et al., 2009b). Fortunately, there is a convenient way to transform compositional data from the simplex sample space to the usual Euclidean space, namely by the family of logratio transformations (Aitchison, 1986; Egozcue et al., 2003). There are two transformations that provide a one-to-one relationship from the simplex to the standard Euclidean space with good geometric properties: the centered logratio (clr) transformation (Aitchison, 1986) and the previously mentioned ilr transformation (Egozcue et al., 2003). The clr transformation results in a multivariate observation $\mathbf{y}=(y_1, \ldots, y_D)$ and is defined as

$$\mathbf{y} = \left(\ln \frac{x_1}{\sqrt{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt{\prod_{i=1}^{D} x_i}} \right). \quad (1)$$

Thus, each compositional part is divided by the geometric mean of all parts, and the logarithm provides the representation in the Euclidean space. Although this transformation is frequently used in practice, it has two severe drawbacks: (a) the sum of the resulting parts is zero, i.e. $y_1+\ldots+y_D=0$, and (b) the transformation is subcompositionally incoherent (Aitchison, 1986). The latter property means that when different subsets of variables (parts) are considered (e.g. because different sets of elements are reported by different laboratories), the clr-transformed results differ in general. This has serious consequences for bivariate data analysis, because usually a data set consists of more than two variables, but any chosen bivariate subset of interest would not reflect the original data. The ilr transformation, which chooses an orthonormal basis on the hyperplane formed by the clr transformation provides a solution to this problem. In other words, the ilr transformation results in a multivariate observation $\mathbf{z}=(z_1, \ldots, z_{D-1})$ in $(D-1)$-dimensional Euclidean space, and the new ilr variables are "technical" constructions according to a chosen basis, like

$$\mathbf{z} = (z_1, \ldots, z_{D-1}), z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, i = 1, \ldots, D-1, \quad (2)$$

based on the original data. If only two parts, e.g. the two-part subcomposition (x_1, x_2), are considered, Eq. (2) simplifies to

$$z = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}. \quad (3)$$

The ilr variable z is only univariate, but it includes all the relevant information between x_1 and x_2 which is in fact contained in their (log) ratio.

Fig. 1 visualises the functionality of the ilr transformation for two-part compositions. In both pictures compositional data summing up to a number between 0 and 1 (filled circles) are depicted. Since only the ratio between the two parts of an observation is relevant, it is possible to vary the points along the straight lines from the origin without any loss of information. Thus, the data points could also be projected on the 135° line where both parts sum up to 1. Applying the ilr transformation gives the new axis parallel to this line, and for each data the new position on the ilr axis can be traced. In the left picture the distances between the data points on the ilr axis are irregular, while in the right picture they are regular. The input data leading to these patterns are, in the first case, data for which the ratios between the parts are regular, i.e. 0.1, 0.2, …, 0.9, 1/0.1, 1/0.2, …, 1/0.9, and, in the second case they are multiples, i.e. 1, 2, 4, 8, 16, and 1/2, 1/4, 1/8, 1/16. The points on the ilr axis contain only univariate information, which can be summarised by the boxplot indicated in the figure. Specifically the length of the box representing the spread of the points is informative as a measure of the stability of the two compositional parts.

A disadvantage of the scale used in the ilr space (Fig. 1) is that the numbers are difficult to interpret. For example, a ratio of $x_1/x_2=10$ would plot at position $\left(1/\sqrt{2}\right)\ln(10)=1.63$. It will thus be more convenient to rescale the axis according to $log(x_1/x_2)$, compared to Eq. (3). The ratio $x_1/x_2=10$ will then plot at $log(10)=1$, indicating the orders of magnitudes (number of powers) by which x_1 and x_2 differ. This scale will be used in all subsequent plots for the ilr space.

1.2. Stability measure for compositional parts

Considering Fig. 1 for identifying the strength of the relationship between the two compositional parts, it is obvious that the correlation coefficient as usually defined is no longer appropriate. This problem was discussed many years ago (Pearson, 1897; Chayes, 1960), but it is still often neglected. In both data examples shown, the usual correlation coefficient would be negative, but because any of the data points could be shifted along the indicated lines without loss of information, this coefficient is meaningless. Instead, a measure for the univariate ilr-transformed data is required, and this has already been defined for logratios by Aitchison (1986) using the concept of the

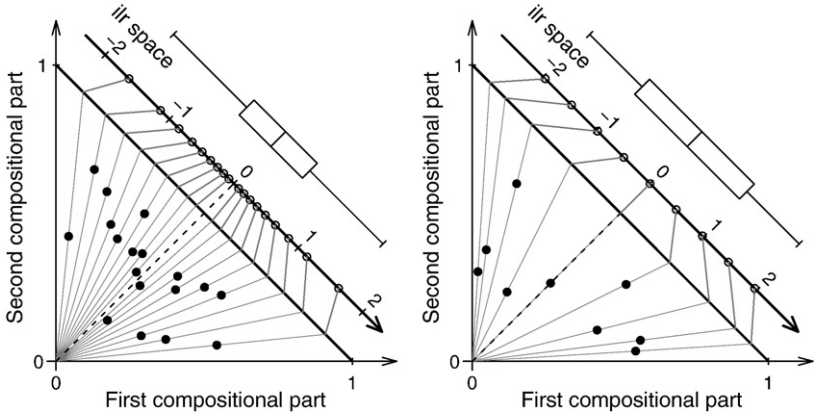

Fig. 1. Geometrical explanation of the ilr transformation for two-part compositions. Left picture: regular increase of the ratios; right picture: multiple increase of the ratios.

variation matrix as a measure of stability. Consider the parts x_1 and x_2, and their logratio expressed by the ilr variable z by Eq. (3). Then var(z) serves as a measure of stability of the two parts, where "var" stands for the variance. Low variability means that the ratio x_1/x_2 is nearly constant, i.e. for all observations the two parts show about the same ratio. On the other hand, high variability reflects very different ratios of the two parts among the observations. One could normalise the measure of variability to the interval [0,1] by defining

$$\mathrm{cor}(x_1,x_2) = \exp(-\mathrm{var}(z)), \tag{4}$$

see Buccianti and Pawlowsky-Glahn (2005). A variability tending to zero causes a value tending to 1, and large variability relates to a result approaching zero. The notation of the result refers to a correlation coefficient of the ilr variable, which will be denoted as ilr correlation in the following. However, it is not a correlation in the usual sense but just a normalization of the ilr variance.

Although these concepts were presented in the specialised literature several years ago, it is not yet clear how to deal with them in practice. Are the results generally different from those achieved by the traditional way of analyzing data, e.g., calculating the usual correlation coefficient for variable pairs? When will differences occur? The goal of this paper is to allow the reader to get a certain feeling for these relatively new concepts and for the consequences of the continued use of unsuitable techniques with the data at hand. Both approaches are compared for a number of practical examples and new plots are introduced that can help to give insight into the real relationship between variable pairs.

2. Correlation coefficients based on different transformations

Correlation coefficients result in numbers lying in the range -1 to $+1$, which express the relationship between two variables. The Pearson correlation is the traditional way to compute a correlation coefficient, and it measures the linear relationship between two variables (see, e.g., Reimann et al., 2008). Since environmental data frequently show a skewed distribution, the variables are often log-transformed prior to computing the Pearson correlation coefficient. The question is, however, whether such correlation coefficients will really provide useful and reliable information for compositional data, where only ratios contain the relevant information and not the values of the single variables themselves.

The correlation coefficient is considered as an indication of the causal relationship between two variables. It is well known that this is not necessarily true (e.g. the example of the spurious correlation between number of babies and storks). There may exist a third, "lurking", variable that the other two variables are independently related to and that causes a high correlation between otherwise unrelated variables. For compositional data it is, in principle, possible to take into account *all* variables, because if they can be measured they would sum up to 1. In this case, such artificial correlations as those described before, can never appear. However, since ratios between the parts need to be considered, the situation becomes difficult in another sense. For example, the traditional correlation coefficient can only be applied to balances, resulting from an ilr transformation (see Filzmoser and Hron, 2009).

Fig. 2 shows scatterplots of the log-transformed variables (upper row) with plots of the ilr-transformed variables (lower row). The ilr transformation results in univariate variables, plotted on the vertical axes (the horizontal axes correspond to the index of the observations). The underlying data are taken from the Kola moss layer, a data set with about 600 samples of terrestrial moss analyzed for the concentration of several (39) chemical elements (see Reimann et al., 2008). The data are available in the R-package StatDA as data frame moss (see Filzmoser and Steiger, 2009).

On top of each figure information about the variable relationships is printed: in the top row the Pearson correlation coefficients of the log-transformed data are provided, while in the bottom row the (empirical) variance of the ilr-transformed variables is given, together with the ilr correlation measure, see Eq. (4). The variable pairs were selected in order to observe different scenarios for the log-transformed data: high correlation (upper left), lack of correlation (inner two in upper row), and negative correlation (upper right). For the vertical axes of the ilr-transformed data the same scale was used in order to emphasize the differences in variability: the two figures on the lower left show ilr variables with low variability (high ilr correlation), while those in the two figures on the lower right show high variability (low ilr correlation). At first glance, the results for the ilr correlations are surprising and counter-intuitive: there seems to be no relationship to the correlations from the log-transformed data. Using the example plots in Fig. 2, it is possible to distinguish three situations from a non-compositional point of view:

- Positive correlation (upper left picture of Fig. 2): Co and Cu increase by the same order of magnitude, they differ approximately by a factor of 100. In other words, their ratio is almost constant which leads to an ilr variable with very small variance. However, high correlation of the log-transformed (or untransformed) data does not necessarily lead to small ilr variance. For instance, if both variables show a strong linear relationship, but their ranges are very different, the ilr variance will get higher, because the ratio between the variables is instable.
- Lack of correlation (both upper middle pictures in Fig. 2): both Ca and Zn vary by about one and a half orders of magnitude, leading to

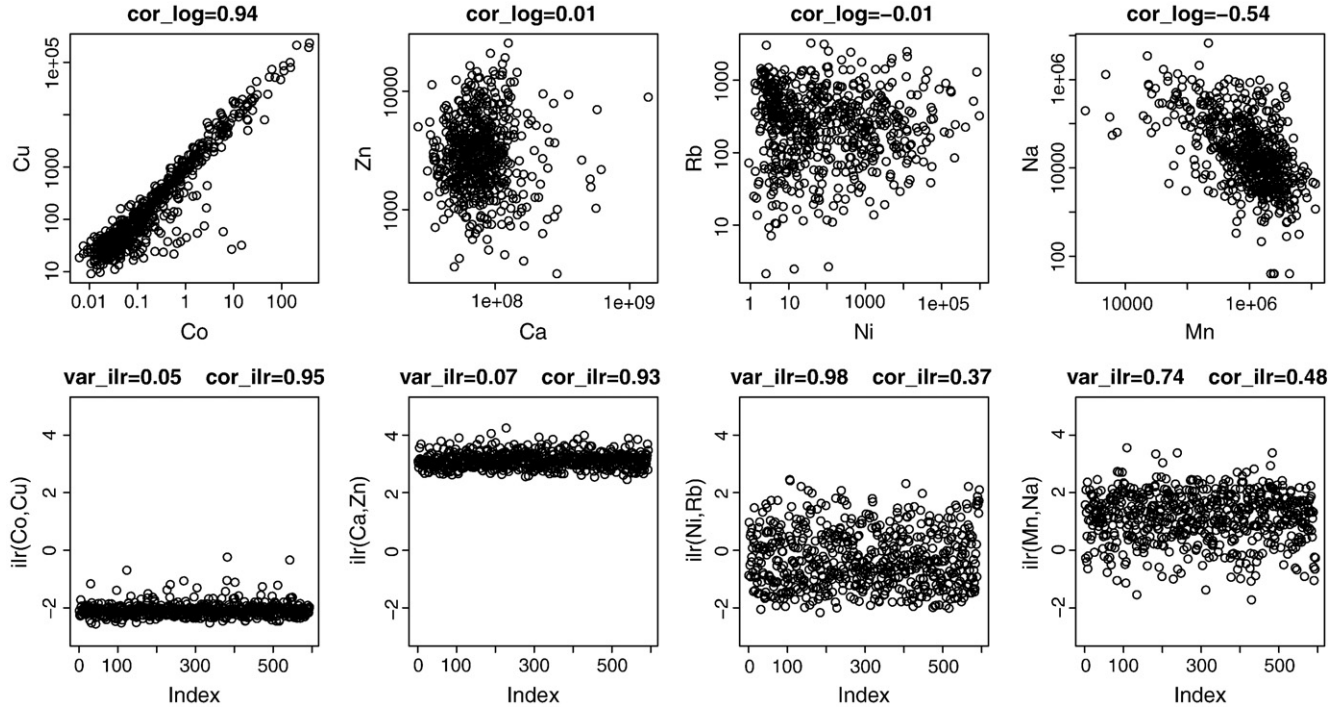

Fig. 2. Log-transformed (upper row) and ilr-transformed (lower row) variable pairs from the Kola moss layer, together with measures of association.

a small ilr variance. On the other hand, Ni varies by six orders of magnitude, while Rb varies much less. Consequently, their ratio leads to high variability.

- Negative correlation (upper right picture of Fig. 2): as in the other cases, only the stability of the ratio between the variables is a useful measure of relationship. The stability is determined by the ilr variance, which depends on the difference in orders of magnitude of the single variables.

Summarizing, the correlation coefficient computed from the closed data does not reflect the stability of the ratios of the variables (ilr variance). This, however, is the only relevant information when investigating the relationship between pairs of compositional parts. If both variables show only small variation (like Ca and Zn), the stability is automatically high (small ilr variance). If both variables show high variation (Co and Cu), then the stability can only be high (high ilr correlation) if the ratios are of similar size.

The 39 variables of the Kola moss data are now used in Fig. 3 to compare the correlation coefficients of the log-transformed, the clr-transformed, and the ilr-transformed data. In total there are 39*38/2 = 741 different variable pairs. The left-hand picture compares the Pearson correlation coefficients for the clr-transformed data with those of the log-transformed data. Each point in the plot refers to a variable pair, and thus there are 741 points in the plot. The point cloud is rather unstructured. Only the highest correlations appear to be high for both transformations, but generally the correlations for the log-transformed and clr-transformed data are quite different. The middle picture compares the Pearson correlation of the clr-transformed data with the correlation of the ilr-transformed data, see Eq. (4). Again there is no relationship between both measures. Finally, the right-hand picture compares the Pearson correlation of the log-transformed data with the correlation based on the ilr-transformed variables, with no clear relationship between the outcomes. This exercise has been carried out with other environmental science data sets, and the picture is essentially always the same: the correlations based on different versions of the transformation are generally very different. Even the correlations of the log-transformed and the clr-transformed data are, in general, not comparable. This demonstrates that working in the wrong geometry will usually give wrong answers, which will then lead to a completely different interpretation of the results of such a dataset.

Note that the use of other correlation coefficients for the original log-transformed data, like Spearman's rank or Kendall's tau, or even a robust correlation coefficient (see, e.g., Reimann et al., 2008), does not solve the problem. For example, because the data points can vary along the straight lines from the origin without any loss of information (Fig. 1), the ranks of the observations for the single variables become meaningless, and Spearman's rank correlation would yield just an arbitrary number, being close to the Pearson correlation in case of log-normal distribution.

3. Correlation replaced by stability

The stability measure mentioned in Section 1.2 is based on an estimation of the variance of the logratio of two compositional parts. Other than the classical estimation by the empirical variance, more robust versions which are less influenced by outlying observations, using the MAD (median absolute deviation) or the IQR (interquartile range), are possible (see, e.g., Reimann et al., 2008). The MAD is defined as

$$MAD(z) = 1.483 \cdot \text{median}_i |z_i - \tilde{z}| \quad (5)$$

where $\tilde{z} = \text{median}(z_1, \ldots, z_n)$ is the median of the observations, and the squared MAD is a very robust measure of variance.

Since ratios of parts account for the essential information of compositions, and since the variance is considered as the basic statistical information, the stability measure provides the key information about how and in which way the data vary. So far, however, the focus has been on bivariate data analysis and on the stability measure resulting from the ilr variable of the two compositional parts, even if the full data matrix included more than two variables. In a multivariate analysis, different procedures should be used (see Section 5).

In order to develop some “feeling” for what the stability measure may mean for real data, the Kola moss data set is used again, and an investigation of the pairwise relationships between Ca and all other available elements is carried out. Note that the focus is still on the

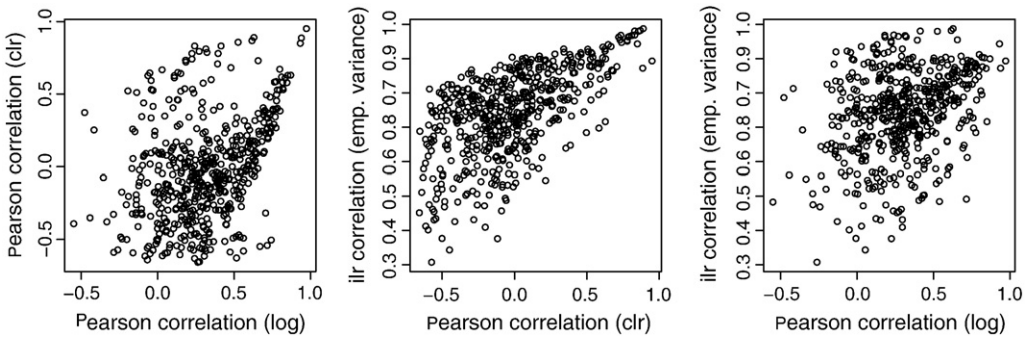

Fig. 3. Comparison of the correlation coefficients for different transformations of the Kola moss data. Each point represents the correlation of a variable pair.

relationship between the pairs, and not on the multivariate data. The resulting ilr-transformed variables are shown in Fig. 4 in form of boxplots, where the boxes are arranged by decreasing values of the medians. Each boxplot represents the relationship of a variable pair according to the ilr transformation from Eq. (3), and thus is constructed like the boxplots presented in Fig. 1. The scale on the vertical axis is already according to the orders of magnitudes of variation between the pairs, see end of Section 1.1. The numbers on top of the plot refer to the ilr correlation based on the squared MAD. Almost all values plot above zero, and thus the element Ca is more dominant in the composition, which is not surprising considering that Ca is a major element in plants. Compared to U, Ca is, on average, more than 5 orders of magnitude higher. The boxplots reveal many outliers which are informative with respect to the data quality and homogeneity. Some example pairs are shown in Fig. 5: the boxplot of the pair Ca–Sc, with many outliers on both sides, is problematic because of the data quality of Sc. Similarly, rounding effects of Y cause various boxplot outliers. On the other hand, the boxplot for Ca–Ni includes no outliers and shows a wide box, which results in a relatively homogeneous data cloud in Fig. 5. Since Ni covers several more orders of magnitude than Ca, the stability is small. Nickel is one of the main elements emitted by industry in the survey area which causes an unusual variability. In contrast, the elements Ca and K vary in about the same range of only one and a half orders of magnitude, which leads to a high stability (small box, few outliers). All major plant nutrients show this behaviour (note the order of elements on horizontal axis).

4. How to use scatterplots

A scatterplot is a bivariate plot presenting the observations of a variable pair. The scatterplot is very informative because it can reveal groups and outliers in the data, and it provides a graphical impression of the relationship between the variables. In the context of compositional data, however, care has to be taken because of the inappropriate geometry of scatterplots when using the original or the log-transformed data. According to the arguments provided in the previous sections, the shape of the point cloud in such a scatterplot does not refer to a measure of correlation between the variables. One could actually question whether it then still makes sense at all to draw and study scatterplots with closed data.

A plot of a univariate ilr variable includes the relevant information on the relation between the two variables. Is there any "loss of information" if the scatterplot of the closed data is completely omitted and ignored?

Fig. 6 shows, for selected elements of the Kola moss data set, a modified scatterplot matrix, where the upper right part shows scatterplots of the log-transformed variables, and the lower left part presents boxplots for the ilr-transformed pairs according to Fig. 4. Obviously, the "traditional" scatterplots show a lot of structure, which is lost in the boxplots. The structure might be caused by subpopulations in the data relating to various effects on the regional element distribution like contamination or sea spray (see, e.g., Reimann et al., 2008). The stability measure does not directly reveal such phenomena: they might just lead to a certain reduction of stability and thus to an increase of the boxes (see above, Ca/Ni in Fig. 5).

The scatterplot of the log-transformed data provides an impression of the bivariate data in the wrong geometry. Even worse, the single variables are already expressed in the wrong scale (see Filzmoser et al., 2009b), and the scatterplot does not account for the ratios between the parts, which is considered as the relevant information contained in compositional parts. Is it thus even "allowed" to inspect such a bivariate plot? When looking at the scatterplot Cu versus Ni in Fig. 6, it can be seen that samples showing high concentrations of

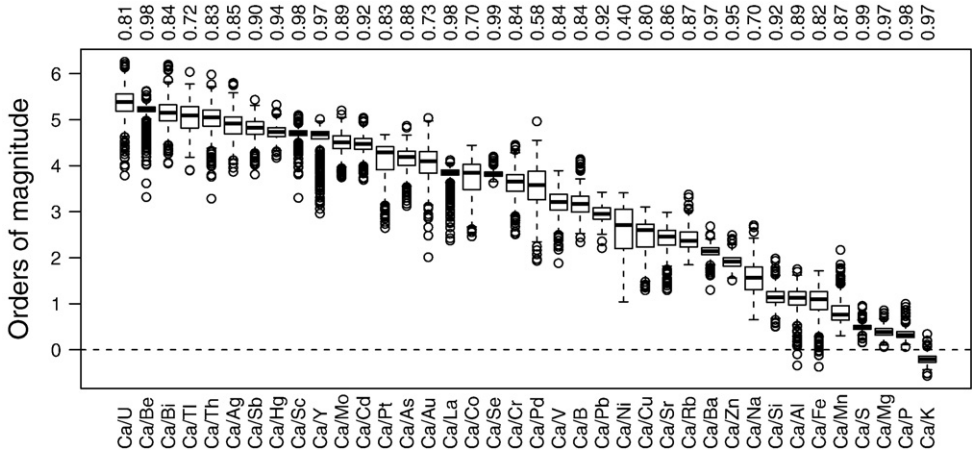

Fig. 4. Boxplot comparison of the pairwise relationships of Ca from the Kola moss data to all other available elements in this layer.

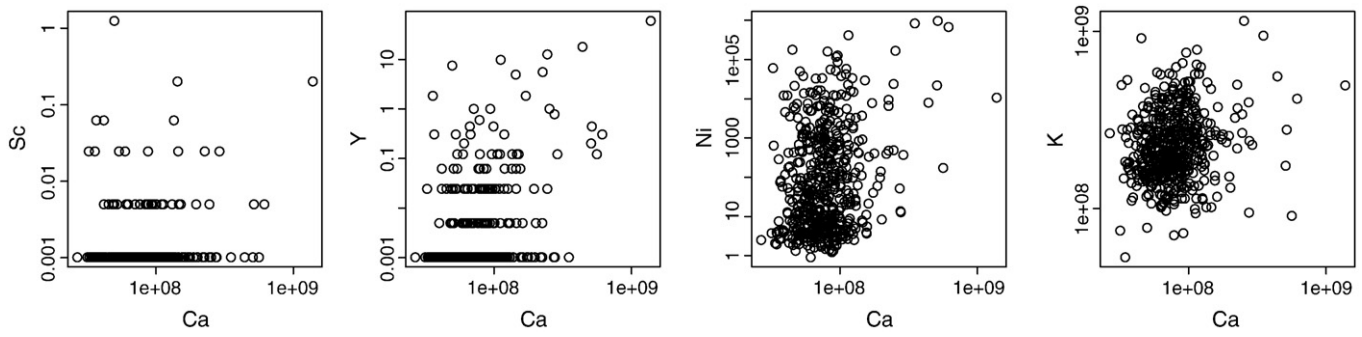

Fig. 5. Scatterplots of selected variable pairs (log-transformed) from Fig. 4.

Cu also have high Ni concentrations. If the maps for Cu and Ni would be compared, they would essentially give the same impression. This is important information for the practitioner who wants to know the distribution of the chemical elements in the survey area, and both elements are actually emitted in large amounts by the Russian nickel industry (Reimann et al., 1998). Thus the observation in the scatterplot as such obviously provides a correct result. Though it is true that the stability between Cu and Ni is quite high, this does not automatically mean that their maps show the same structure. The ilr correlation between Mo and Sb is even higher (0.94), but the scatterplot looks very different. In fact, it reveals a group of observations in the lower right part with a different behaviour. Is it "allowed" to focus on this group and, e.g., to go back to their locations in the maps of Mo and Sb? The answer is yes, because this observation is not related to a statistical correlation between the elements. Rather than looking at correlation, the scatterplot is used as a tool for exploratory data analysis here, in order

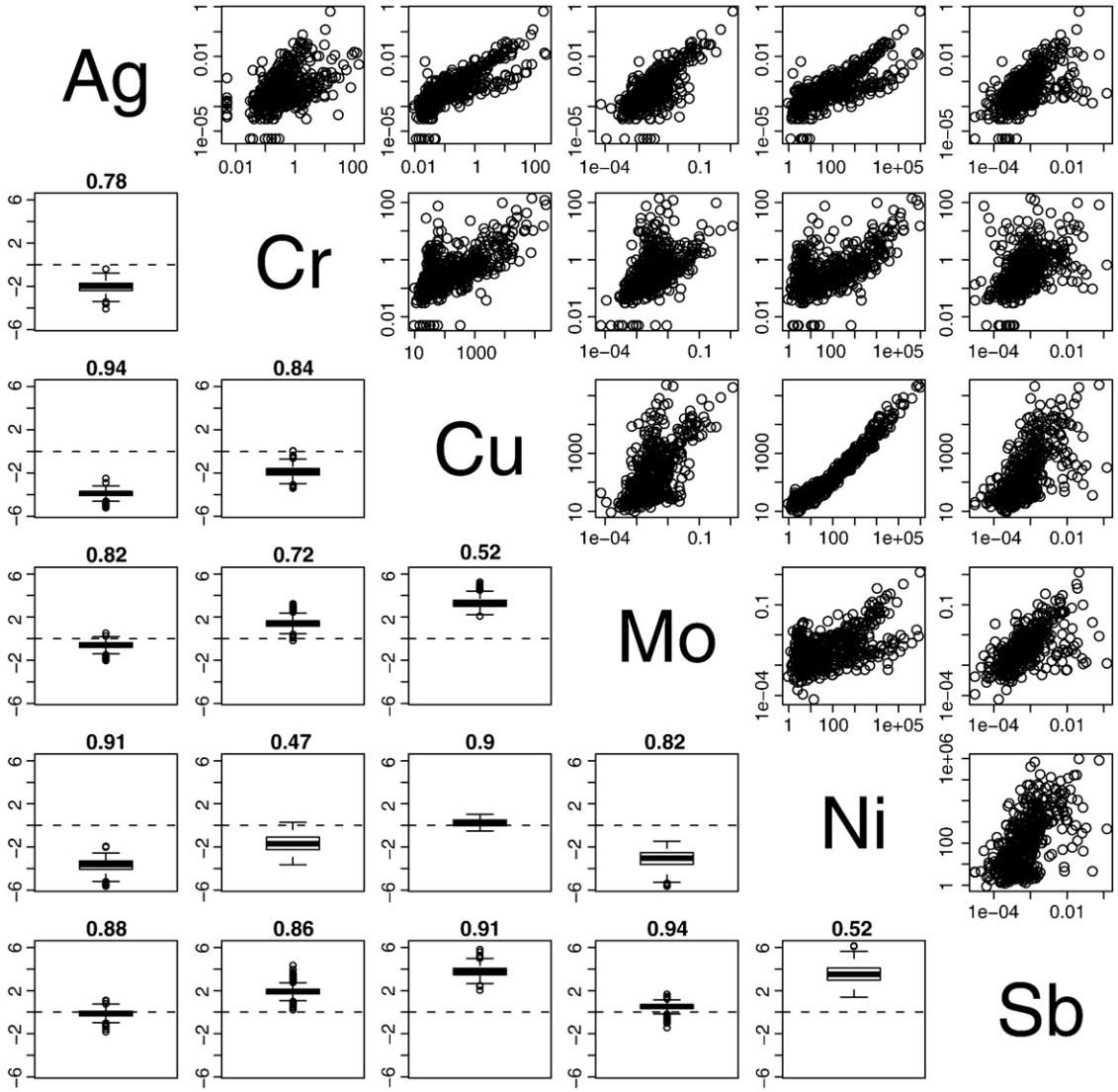

Fig. 6. Scatterplot matrix for six selected elements of the Kola moss data: traditional scatterplots of the log-transformed variables in the upper right part, and boxplots of the ilr variables in the lower left part. The numbers on top of the boxplots provide the ilr correlation.

to detect unusual data behaviour. It would thus be incorrect to discuss a high correlation between the log-transformed element concentrations of Mo and Sb, and to point out deviations from this correlation (e.g. the group of observations with unusually high Sb values). Such an interpretation would be misleading because the reported "correlation" ignores the remaining parts of the composition, which may be responsible for the relationship. An exploratory procedure discovers groups in the bivariate data, and usually aims to find an explanation by going back to the univariate information.

Whenever using a scatterplot one thus needs to be very aware that the points as such are presented in an incorrect geometry. Thus other grouping information could still be hidden or deviations can be incorrectly emphasized. On the other hand, the explorative approach only tries to discover atypical data behaviour. This can be done as long as no confirmatory procedure like a statistical test with model assumptions is used, or a regression line is added indicating the strength of the linear relation. In this exploratory sense, and only in this sense, scatterplots still contain useful information. This information should, however, never be interpreted in statistical terms. The plot needs to be verified using other graphics, e.g. maps, to understand and interpret the underlying geochemical processes.

5. Multivariate extensions

Bivariate statistical analyses–even when they are combined–cannot replace a multivariate data analysis. A typical example was shown in Fig. 6 with the scatterplot matrix. Although relationships between all pairs are investigated, it is difficult to generate an "overall picture" of the multivariate relationships. This is not surprising, because the same problem is encountered also in the normal case, using well-behaved (non-compositional) data. A standard tool for investigating and presenting the multivariate data structure is the biplot which is based on the overall correlation matrix (see, e.g., Reimann et al., 2008). It has just been demonstrated above that the bivariate correlation does not make sense when working with compositional data — can then a procedure like principal component analysis (PCA) or factor analysis, building on the multivariate correlation matrix make any sense?

For compositional data it is also possible to construct a biplot, but here the basic question is how to transform the data. It is not sufficient to put together all correlations resulting from the pairwise ilr transformations, i.e. the numbers on top of the plots in Fig. 6, and to apply the usual procedure for constructing the biplot. Here it is necessary to transform the complete data matrix appropriately, and this can be done by the ilr transformation shown in Eq. (2). The exact procedure for constructing the biplot is presented in Filzmoser et al. (2009a); it is important to note that the interpretation of the biplot differs from the usual interpretation (see Aitchison and Greenacre, 2002). This method was applied to all available elements (with reasonable data quality) from the Kola moss layer. The biplot is thus constructed with 31 elements, and the scatterplots presented in Fig. 6 form a subset of this complex data set. The results for the compositional biplot are shown in Fig. 7 (right). The first two principal components express about 51% of the total variability, and thus they cannot reveal all the multivariate information (one would have to inspect further principal components). Nevertheless, some characteristic processes in the data become immediately visible: Co, Cu and Ni are the main elements emitted by industry in the survey area, and also As and Ag belong to this association of smelter-related elements (lower right quadrant); an association of Pb, Tl, Th, U, Mo, Al, and Si (upper left quadrant) reflects the deposition of terrigenous dust on the mosses; and a Mg, K, P, Ca, S, Zn and Mn association is characteristic of biological processes in the mosses. The association Na, B and Sr indicates the input of marine aerosols along the coast. Thus, major spatial and biological processes in the survey area determine the first two principal components.

Since the stabilities of the ratios of the compositional parts form the input to the compositional biplot shown in Fig. 7 (right), it is possible to verify the element associations at least for the element selection presented in the scatterplot matrix in Fig. 6 (lower left part). For example, the ratios of the variable pairs Cu–Ni, Cu–Ag, and Ni–Ag show high stability and thus they are closely related in the biplot. In contrast, the pairs Mo–Cu and Cr–Ni refer to low stability, expressed by larger distances between the rays representing the variables in the biplot.

For comparison, also the (wrong) biplot based on the log-transformed moss data is shown (Fig. 7, left). This is the version one would see in the vast majority of environmental applications. The difference to the ilr version is obvious. Most importantly, in the "classical" biplot almost all the variables are arranged only in a half-plane. This is a typical indication of working in the wrong geometry with closed data: in the plot using the ilr-transformed data the effect of opening the data is immediately visible. Still, the classical procedure works as it is supposed to work: the configuration of the variables in this biplot approximates the correlation matrix of the log-transformed data in the best possible way. This can even be verified by the element subset shown in the upper right part of the scatterplot matrix (Fig. 6). However, due to using closed data the correlations based on the log-transformed data do not reflect the real underlying variable relationships, and thus the biplot is misleading. Without knowing the processes that become so clearly visible when using the ilr-transformed data, the user might arrive at a completely wrong

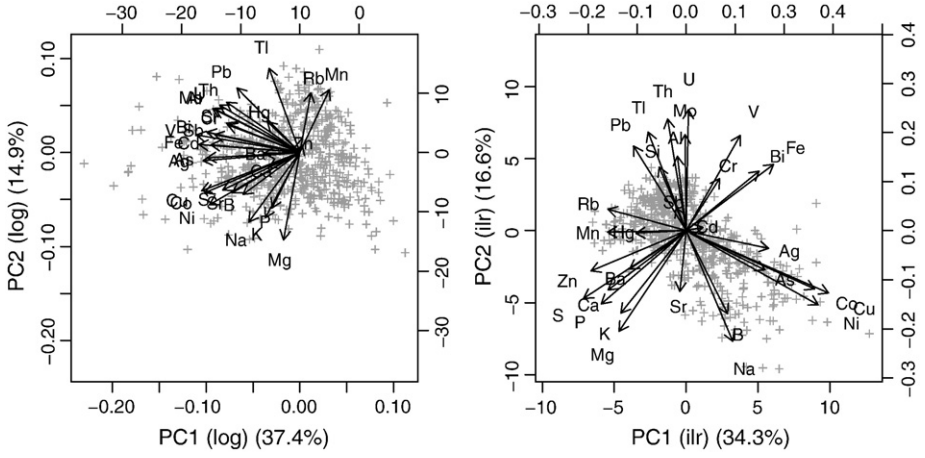

Fig. 7. Biplots for the Kola moss data set based on the log-transformed variables (left) and on the ilr-transformed data (right).

interpretation, e.g. that the majority of elements are co-emitted by industry.

Note that both biplots shown in Fig. 7 are robust, i.e. they have been constructed with the robust correlation matrix of the log-transformed and ilr-transformed data, respectively (for details, see Filzmoser et al., 2009a). This allows us to focus on the homogeneous data majority, and reduces the influence of data outliers. Nevertheless, even a robust analysis performed on the log-transformed data is not able to "repair" the artifacts of the geometry.

There are many other multivariate methods for presenting the multivariate data structure of compositional data. Most importantly, they all have to be based on an appropriate transformation, like the ilr transformation, applied to the complete data matrix. A difficulty with this transformation is the interpretation of the resulting ilr variables (Egozcue and Pawlowsky-Glahn, 2005) which can be considered as mathematical constructions, see Eq. (2). For an interpretation in terms of the original variables the results of the multivariate analysis have to be back-transformed, usually to the clr space. This was also the case in Fig. 7 (right), and thus the "usual" interpretation of compositional biplots (see Aitchison and Greenacre, 2002) is facilitated.

Filzmoser et al. (2009c) proposed factor analysis for compositional data. In contrast to principal components, the factors are supposed to be interpretable in terms of the original variables, and they can often be assigned to certain processes in the data. Maps of the factor scores will then show the regional distribution of these processes (for details, see Filzmoser et al., 2009c).

6. Discussion

The discussion about an appropriate transformation of compositional data needs to be based on the geometry inherent in the data, and not on properties of the statistical distribution of the observed data values. In other words, although a log transformation of the variables may often result in a more symmetric statistical distribution, sometimes even in a distribution that looks similar to a normal distribution, this does not mean that the log transformation is suitable for the subsequent statistical analysis of environmental data. The log transformation can and should be used only in the univariate case, and if the observed values for a variable are small enough (see Filzmoser et al., 2009b), but it is inappropriate for investigating bivariate or multivariate relationships. There are many papers describing the special geometry of compositional data (see, e.g., Egozcue and Pawlowsky-Glahn, 2006). The relatively new ilr transformation allows the representation of compositional data in the standard Euclidean space and does not result in singularity problems. The only reason why the data should be transformed to Euclidean space is because practically all the statistical "standard tools" are designed for this space, and they can lead to wrong conclusions when used in another space.

When using the ilr transformation to transfer the compositional data into Euclidean space, bivariate plots become univariate. Now the variability of the ilr-transformed variables provides information about the relationship between two variables. The ilr transformation can be viewed as a non-linear transformation resulting in univariate information. The distances between the data points correspond to the Aitchison distance reflecting the geometry of compositional data (Egozcue and Pawlowsky-Glahn, 2006), but the scale of the new univariate data is unfortunately not straightforward to interpret. Rescaling the axis according to the simple logratio, and using the logarithm to the basis 10, allows an interpretation in terms of the difference of the original variables expressed in orders of magnitude. The larger the variability, the smaller is the ilr correlation between the compositional parts. The smaller the variability, the more stable is one variable with respect to the other one, and thus the ilr correlation measure gets higher. Variability can therefore be expressed as the stability of one compositional part on the other one. This stability may also be observed spatially in the case of spatial data. Highly varying ratios of two elements in a survey area indicate low stability, and consequently they result in low ilr correlation.

One of the simplest visualization tools, the scatterplot, representing the data pairs measured on two variables in one plot, cannot be interpreted in the usual way. It should only be used as an exploratory tool in order to detect unusual data behaviour or data groups, but not in a confirmatory sense. The tendency of the point cloud in a scatterplot does not reflect the correlation as for traditional non-compositional data and the plots should not be used to find linear (or curve-linear) relationships. The traditional way of linear or non-linear regression relating a response variable with an explanatory variable is not meaningful for compositional data since the data points only include relative information.

It is important to note that the compositional nature is inherent in the data structure even if the sum of the considered elements is not constant or does not equal 100%. For example, when adding the element concentrations of all the variables used in the biplots in Fig. 7, the sum for the samples is on averageabout 12,000 mg/kg, i.e. the element concentrations form only about 1% of the complete sample. This is even more extreme in the bivariate case, where only two variables are of interest for the statistical analysis. The inappropriateness of the usual Euclidean geometry for compositional data, and the inappropriateness of statistical methods relying on this geometry have been demonstrated for both the bivariate and the multivariate cases.

7. Conclusions

The fact that compositional (i.e. practically all environmental) data are not represented in the standard Euclidean space leads to severe consequences during data analysis, which become especially serious in the bivariate case. The scatterplot cannot be interpreted in the usual way, though it can still be used in a truly exploratory data analysis sense. Note that bivariate plots of different ratios, the classical solution to the closure problem in petrology, do make sense. Correlation coefficients based on raw or log-transformed compositional data should not be calculated, the resulting values have no statistical meaning. As a consequence, tests for lack of correlation are not meaningful either for compositional data. It is not even clear how such a test could be adapted appropriately for compositional data since any such test should be based on the variance (or standard deviation) of the corresponding ilr variable. Multivariate data analysis can provide a solution to the loss of classical bivariate correlation analysis. Once the data are opened (i.e. ilr-transformed) standard techniques like PCA or factor analysis can again be used to study the relation between all variables in the multivariate space even though these techniques are based on correlations.

Acknowledgments

This work was supported by the Council of the Czech Government MSM 6198959214.

References

Aitchison J. The statistical analysis of compositional data. London, U.K: Chapman and Hall; 1986.
Aitchison J, Greenacre M. Biplots of compositional data. Appl Stat 2002;51:375–92.
Aitchison J, Barceló-Vidal C, Martín-Fernández J, Pawlowsky-Glahn V. Logratio analysis and compositional distance. Math Geol 2000;32:271–5.
Buccianti A, Pawlowsky-Glahn V. New perspectives on water chemistry and compositional data analysis. Math Geol 2005;37(7):703–27.
Chayes F. On correlation between variables of constant sum. J Geophys Res 1960;65:4185–93.
Egozcue J, Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. Math Geol 2005;37:795–828.
Egozcue J, 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 J, 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. Correlation analysis for compositional data. Math. Geosci. 2009;41: 905–19.
Filzmoser P, Steiger B. StatDA: statistical analysis for environmantel data; 2009. R package version 1.1. URL http://cran.at.r-project.org/web/packages/StatDA/index. html.
Filzmoser P, Hron K, Reimann C. Principal component analysis for compositional data with outliers. Environmetrics 2009a;20:621–32.
Filzmoser P, Hron K, Reimann C. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci Total Environ 2009b;407: 6100–8.
Filzmoser P, Hron K, Reimann C, Garrett RG. Robust factor analysis for compositional data. Comput Geosci 2009c;35:1854–61.
Ivanov D. Linear parageneses of the chief rock forming minerals in the granites of central kazakhstan. Dokl. Akad. Sci. USSR 1963;150:134–6.
Miesch A, Chao E, Cuttita F. Multivariate analysis of geochemical data on tectites. J Geol 1966;74:673–91.
Pearce T. A contribution to the theory of variation diagrams. Contrib. Mineral. Petrol. 1968;19:142–57.
Pearson K. Mathematical contributions to the theory of evolution. on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc London 1897;LX:489–502.
Podolsky Y. Linear parageneses of the main rock forming minerals of the alkalic rocks in the central part of the kola peninsula. Dokl. Acad. Sci. USSR 1962;146:155–8.
Reimann C, Äyräs M, Chekushin V, Bogatyrev I, Boyd R, Caritat Pd, et al. Environmental geochemical atlas of the Central Barents region. Geological Survey of Norway (NGU), Geological Survey of Finland (GTK), and Central Kola Expedition (CKE), Special Publication, Trondheim, Espoo, Monchegorsk; 1998.
Reimann C, Filzmoser P, Garrett R, Dutter R. Statistical data analysis explained. Applied Environmental Statistics with R. Wiley, Chichester, UK; 2008.
Rock N. Numerical geology. Lecture Notes in Earth Sciences 18. New York-Berlin-Heidelberg: Springer Verlag; 1988. 427 pp.