Computers & Geosciences 35 (2009) 1854–1861

Contents lists available at ScienceDirect

Computers & Geosciences

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

Robust factor analysis for compositional data

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

[a] Vienna University of Technology, Department of Statistics and Probability Theory, Wiedner Hauptstr. 8-10, Vienna 1040, Austria
[b] Palacký University, Faculty of Science, Tomkova 40, Olomouc 77100, Czech Republic
[c] Geological Survey of Norway, Trondheim 7491, Norway
[d] Geological Survey of Canada, 601 Booth St., Ottawa, Canada K1A 0E8

ARTICLE INFO

Article history:
Received 22 April 2008
Received in revised form
16 December 2008
Accepted 29 December 2008

Keywords:
Centred logratio transformation
Isometric logratio transformation
Factor analysis
Compositional biplot
Robust estimation

ABSTRACT

Factor analysis as a dimension reduction technique is widely used with compositional data. Using the method for raw data or for improperly transformed data will, however, lead to biased results and consequently to misleading interpretations. Although some procedures, suitable for factor analysis with compositional data, were already developed, they require pre-knowledge of variable groups, or are complicated to handle. We present an approach based on the centred logratio (clr) transformation that does not build on this pre-knowledge, but still recognizes the specific character of compositional data. In addition, by using the isometric logratio transformation it is possible to robustify factor analysis using a robust estimation of the covariance matrix. A back-transformation of the results to the clr space allows an interpretation of the results with compositional biplots. The method is demonstrated with data from the Kola project, a large ecogeochemical mapping project in northern Europe.

1. Introduction

The goal of factor analysis is to extract a few directions in the data space, called the factors or latent variables, that are not directly measurable but represent certain features inherent in the data (see, e.g., Basilevsky, 1994, or Johnson and Wichern, 2007). Thus factor analysis reduces the data dimensionality to these few representative factors, and therefore aims at summarizing the multivariate information in a compact form. Although factor analysis has been criticized frequently as a rather subjective method where the results strongly depend on the chosen parameters, it is still successfully used in many applications. Generally speaking, when using factor analysis as an exploratory method, the results will show properties inherent in the multivariate data, which should, however, be carefully checked with other methods, preferably with less complicated visualization tools (see Reimann et al., 2008).

When applying factor analysis to compositional data, it is crucial to apply an appropriate transformation. A log-transformation will often reduce data skewness, but does not accommodate the compositional nature of the data. Aitchison (1986) suggested several possible transformations from the family of logratio transformations. The centred logratio (clr) transformation treats the variables symmetrically and thus allows for a more concise interpretation. The disadvantage of the subcompositional incoherence of the clr transformation, and the singularity of the clr transformed data has led to the development of the isometric logratio (ilr) transformation (Egozcue et al., 2003). Although this approach prevents numerical problems, results are not interpretable because one has to deal with new $D-1$ variables that do not relate in a simple way with the D original ones. An even more sophisticated procedure to avoid subcompositional incoherence was suggested by Egozcue and Pawlowsky-Glahn (2005), which is based on an ilr transformation for groups of variables. However, since finding meaningful variable groups is exactly the goal of factor analysis, this approach seems to be inappropriate here.

Other approaches for factor analysis with compositional data have been suggested (e.g., Tolosana-Delgado et al., 2005), but here we will present an approach that is in analogy to the standard procedure for factor analysis. It is based on the clr transformation which leads to meaningful biplots. They, however, have to be interpreted appropriately using the basic properties of compositional biplots (Aitchison and Greenacre, 2002). The basic property, that distinguish them from the standard biplots (Gabriel, 1971), is focusing on the distance between vertices of the rays (links) that approximate the dispersion of the ratio of two compositional variables. Thus, if the corresponding vertices coincide, this means that the ratio is constant, or nearly so. We extend the classical approach of factor analysis to robust factor analysis for compositional data. Robust factor analysis can be obtained via a robust estimation of the covariance matrix (Pison et al., 2003). However, robust covariance estimation is not directly possible in the clr

* 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), garrett@NRCan.gc.ca (R. Garrett).

0098-3004/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2008.12.005

space because of the singularity problem. Therefore, the robust covariance estimation is performed in the ilr space. In contrast to outlier detection for compositional data, where transformed variables need no interpretation (Filzmoser and Hron, 2008), the results have to be back-transformed here to the clr space which allows for an interpretation.

In the next section we briefly summarize the necessary elements of compositional data analysis as well as the limitations of factor analysis in the context of compositional data. Section 3 is devoted to the robust estimation of the factor analysis model, and in Section 4 the procedure is applied to the moss layer of the Kola data set (Reimann et al., 1998). The final section summarizes the new findings.

2. Factor analysis and compositional data

The direct application of multivariate statistical methods to raw D-parts compositional data can lead to improper results. For example, the problem of so called spurious correlations (Chayes, 1960) occurs, namely that one can obtain different results of correlation analysis, depending on whether the whole composition or only a subcomposition is taken. Thus an appropriate transformation from the logratio family is necessary before the analysis can be carried out. Especially for factor analysis the choice of the actual transformation is important because of the various parameters that are to be estimated in this procedure. The centred logratio transformation (Aitchison, 1986) seems to be best suitable with respect to the interpretation of the resulting values, for example, by using compositional biplots. It is a transformation from the simplex sample space to the D-dimensional real space, and for a compositional random (column) vector $\boldsymbol{x}=(x_1,\ldots,x_D)^T$ it is defined by

$$\boldsymbol{y}=(y_1,\ldots,y_D)^T=\left(\ln\frac{x_1}{\sqrt[D]{\prod_{i=1}^{D}x_i}},\ldots,\ln\frac{x_D}{\sqrt[D]{\prod_{i=1}^{D}x_i}}\right)^T.$$

The definition can be written in matrix form,

$$\boldsymbol{y}=\left(\boldsymbol{I}_D-\frac{1}{D}\boldsymbol{J}_D\right)\ln\boldsymbol{x}=\boldsymbol{H}\ln\boldsymbol{x}\quad\text{with }\boldsymbol{I}_D=\begin{pmatrix}1&&0\\&\ddots&\\0&&1\end{pmatrix},$$

$$\boldsymbol{J}_D=\begin{pmatrix}1&\ldots&1\\\vdots&\ddots&\vdots\\1&\ldots&1\end{pmatrix},\qquad(1)$$

both being matrices of dimension $D\times D$. However, it is easily seen that the clr transformation results in singular data ($y_1+\cdots+y_D=0$), and many multivariate methods will not be applicable in case of singularity. A further problem is the so-called subcompositional incoherence (Aitchison, 1986), where the clr covariance matrix of a full composition and of a given subcomposition is related through a linear but quite complex relationship (Aitchison, 1986). Moreover, the matrix $\mathrm{Cov}(\boldsymbol{y})$ is singular summing up to zero by rows and columns. On the other hand, the property that the clr variables preserve some link to the original (compositional) parts can be used in special cases for a visualization of the compositional covariance structure by means of compositional biplots. These are biplots with focus on links instead of rays, which approximate the dispersion of the ratio of two compositional parts (for details, see Aitchison and Greenacre, 2002).

In many practical applications the data are simply log-transformed,

$$\ln\boldsymbol{x}=(\ln x_1,\ldots,\ln x_D)^T.$$

Frequently, the results from a statistical analysis of the log-transformed data are even more easily interpretable than those from an analysis of the clr transformed data. However, the log-transformation does not account for the compositional nature of the data. For example, the dimensionality of D-part compositional data is only $D-1$, but the log-transformation results in D-dimensional data and thus destroys the basic properties of the data. As a consequence, a multivariate method like factor analysis will be forced to work in an inappropriate geometrical space, and biased results can be expected.

Factor analysis is traditionally used to discover a number of factors (new variables) that cannot be observed directly. The factors are supposed to lead to a better interpretation and understanding (in contrast to principal components) of the original data after dimension reduction. They focus attention on possible underlying processes controlling the data distribution rather than the responses (measurements). In spite of the problems with the clr transformation mentioned above, it is a preferable choice for factor analysis, for example, because the resulting factors and loadings can then be interpreted using compositional biplots (Tolosana-Delgado et al., 2005). However, some restrictive aspects discussed below need to be considered.

For the random vector $\boldsymbol{y}$, the factor analysis model is defined as

$$\boldsymbol{y}=\boldsymbol{\Lambda}\boldsymbol{f}+\boldsymbol{e}\qquad(2)$$

with the factors $\boldsymbol{f}$ of dimension $k<D$, the error term $\boldsymbol{e}$, and the loadings matrix $\boldsymbol{\Lambda}$. Using the usual model assumptions (see, e.g., Basilevsky, 1994), the factor analysis model (2) can be written as

$$\mathrm{Cov}(\boldsymbol{y})=\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T+\boldsymbol{\Psi},\qquad(3)$$

where $\boldsymbol{\Psi}=\mathrm{Cov}(\boldsymbol{e})$ has a diagonal form. The diagonal elements are called uniquenesses (or unique variances) and they include the part of the variance of the components of $\boldsymbol{y}$ that is not explained by the factors.

In the case of compositional data, the vector $\boldsymbol{y}$ is the clr transformed random vector (see (1)). As mentioned above, $\mathrm{Cov}(\boldsymbol{y})$ is singular, and this is in conflict with a diagonal form of $\boldsymbol{\Psi}$ in (3) (a diagonal matrix with strictly positive diagonal elements is regular, and the sum on the right-hand side in (3) thus must yield a regular matrix). The problem can be solved by projecting the diagonal matrix $\boldsymbol{\Psi}$ onto the hyperplane $y_1+\cdots+y_D=0$ formed by the clr transformation. The new model is then

$$\mathrm{Cov}(\boldsymbol{y})=\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T+\boldsymbol{H}\boldsymbol{\Psi}\boldsymbol{H}^T,\qquad(4)$$

where $\boldsymbol{H}$ comes from the definition of the clr transformation in Eq. (1). Since $\boldsymbol{H}^T=\boldsymbol{H}$, the last term in (4) can be written as $\boldsymbol{\Psi}^*=\boldsymbol{H}\boldsymbol{\Psi}\boldsymbol{H}$. The matrix $\boldsymbol{\Psi}^*$ has no longer a diagonal form. Therefore the interpretation is different from $\boldsymbol{\Psi}$, and the unique variances cannot be directly assigned to the single clr variables. However, if the number D of compositional parts is large, the off-diagonal elements of $\boldsymbol{\Psi}^*$ will become small, and one could interpret the diagonal elements in the usual way as unique variances, being assigned essentially to the single variables.

For estimating the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}^*$ in (4) we assume that $\boldsymbol{y}$ is only centred but not scaled in order to allow for an appropriate interpretation of the compositional biplot. The first step is the estimation of the covariance matrix of $\boldsymbol{y}$, which is traditionally done by the sample covariance matrix. In Section 3 we will discuss alternative approaches. Similar to the principal factor analysis (PFA) procedure (see, e.g., Basilevsky, 1994) we propose the following iterative algorithm for parameter estimation.

Step 1: initialize the diagonal elements of $\boldsymbol{\Psi}$ by the largest off-diagonal elements of the estimated covariance matrix $\boldsymbol{C}(\boldsymbol{y})$ of $\mathrm{Cov}(\boldsymbol{y})$;

Step 2: compute $\boldsymbol{\Psi}^* = \boldsymbol{H}\boldsymbol{\Psi}\boldsymbol{H}$;
Step 3: estimate $\boldsymbol{\Lambda} = (\lambda_{ij})$ from the relation $\boldsymbol{C}(\boldsymbol{y}) - \boldsymbol{\Psi}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$;
Step 4: update the diagonal elements ψ_i of $\boldsymbol{\Psi}$ by $\psi_i = \{\boldsymbol{C}(\boldsymbol{y})\}_{i,i} - \sum_{j=1}^{k}\lambda_{ij}^2$ (here $\{\boldsymbol{C}(\boldsymbol{y})\}_{i,i}$ denotes the i-th diagonal element of $\boldsymbol{C}(\boldsymbol{y})$);
Step 5: go to Step 2 and iterate until the elements in $\boldsymbol{\Psi}^*$ stabilize.

There is no guarantee that the above algorithm converges (note that there is also no proof that the original iterative algorithm converges). However, in our practical experiments the convergence was always reached. For a better interpretation of the estimated loadings matrix $\boldsymbol{\Lambda}$ an orthogonal or oblique rotation can be performed.

The resulting loadings are visualized graphically in a loading plot (e.g., see Figs. 2 and 3), and in a compositional biplot (see, e.g., Fig. 4). The loading plot shows loadings for all factors, while in the biplot we concentrate on pairs of factors. High (absolute) values of the loadings represent high influence of the corresponding clr variables on the factor. The pattern of the highly influential variables determines the interpretation of the factor. Since we analyzed the clr transformed variables, the interpretation of the relations between the original variables is different. Similar values of the loadings for clr variables y_i and y_j refer to an approximatively constant ratio x_i/x_j of the original (compositional) parts (Aitchison, 1986). Since the clr variables are linear combinations of the factors, see (2), we can go even further: similar loadings for y_i and y_j on one factor refer to an approximatively constant (log-)ratio x_i/x_j along this factor. This means that the factor where we have similar loadings will not change the ratio x_i/x_j, i.e. this factor is not related to any relative enrichment or depletion of x_i with respect to x_j.

Fig. 1 shows an example of this property. The left picture shows that the logratio of Ni and Co along the robust Factor F1 changes (see Fig. 2 (bottom) or Fig. 4 (left) in Section 4). This means that the loadings of Ni and Co differ along this factor. On the other hand, the right picture with the logratio of Cu and Co for the same factor shows no systematic trend (since the factor has been computed robustly, the influence of the outliers visible in the plot is reduced). Consequently, the loadings of Cu and Co are very similar for this factor.

In addition to the loadings, the estimation of the factor scores is required, and again care has to be taken because of the singularity of Cov($\boldsymbol{y}$). It is possible to use the regression method (see, e.g., Basilevsky, 1994) by taking the generalized inverse of the covariance matrix. Here, the Bartlett estimation is employed, which directly considers model (2) as a regression model, and where the factors are estimated by weighted least squares regression,

$$\widehat{\boldsymbol{f}} = (\boldsymbol{\Lambda}^T(\boldsymbol{\Psi}^*)^+\boldsymbol{\Lambda})^+\boldsymbol{\Lambda}^T(\boldsymbol{\Psi}^*)^+\boldsymbol{y}.$$

Here, $(\boldsymbol{A})^+$ denotes the Moore–Penrose pseudo-inverse of a singular matrix $\boldsymbol{A}$. For this procedure, the estimated matrices $\boldsymbol{\Lambda}$ (after possible orthogonal rotation) and $\boldsymbol{\Psi}$ are treated as the true values. In the n samples case, the estimated random vector $\boldsymbol{f}$ corresponds to the $n \times k$ matrix $\widehat{\boldsymbol{F}}$ of factor values. In the case of two factors, the rows of this matrix can be displayed together with rows of the estimated loadings matrix $\boldsymbol{\Lambda}$ in a biplot. The factors represent the objects and loadings represent the variables. As mentioned above, an appropriate interpretation of the results with respect to the original compositions is important.

3. Robust estimation

The estimation of loadings and scores is based on the estimation of the covariance matrix Cov($\boldsymbol{y}$). Let us consider a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from the distribution of the random vector $\boldsymbol{x}$. The clr transformation results in a sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Traditionally, the estimation is done with the sample covariance matrix $\boldsymbol{S} = 1/(n-1)\sum_{i=1}^{n}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T$, where $\overline{\boldsymbol{y}}$ denotes the arithmetic mean. However, it is well known that in case of outlying observations in the data set this estimation may lead to very unreliable results. In this case a robust estimation is required, and a popular choice is the MCD (minimum covariance determinant) estimator, for which also a fast algorithm is available (Rousseeuw and Van Driessen, 1999). The MCD estimator looks for a subset h out of n observations with the smallest determinant of their sample covariance matrix. The robust estimator of covariance is the sample covariance matrix of the h observations, multiplied by a factor for consistency at normal distribution. The subset size h determines the robustness of the estimator, and it can be varied between half the sample size and n.

Unfortunately, robust procedures cannot deal with singular data, which exactly is the case for clr transformed compositions. A way out is to use the ilr transformation (Egozcue et al., 2003), which expresses the clr images in an orthonormal basis on the hyperplane $y_1 + \cdots + y_D = 0$, formed by this transformation. For a composition $\boldsymbol{x}$ the ilr transformation can be defined as

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

for $i = 1, \ldots, D-1$.

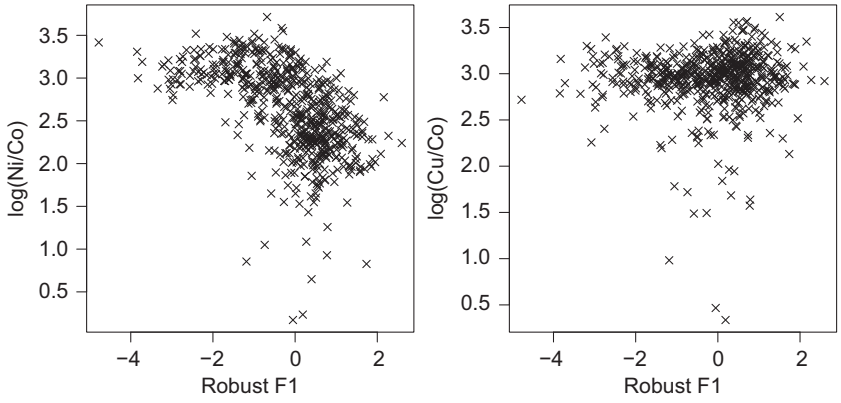

Fig. 1. Inspection of logratio of some variables along robust Factor F1 for example used in Section 4. Left plot: logratio of Ni and Co shows a clear trend meaning that their loadings differ along this factor. Right plot: logratio of Cu and Co is almost constant, i.e. loadings are very similar along this factor.

An equivalent representation is

$$\boldsymbol{z} = \boldsymbol{V}^T\boldsymbol{y} = \boldsymbol{V}^T\left(\boldsymbol{I}_D - \frac{1}{D}\boldsymbol{J}_D\right)\ln\boldsymbol{x} \tag{5}$$

with the $D \times (D-1)$ matrix $\boldsymbol{V}$ with orthonormal basis vectors in its columns (i.e. $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_{D-1}$)

$$\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{D-1}), \quad \boldsymbol{v}_i = \sqrt{\frac{i}{i+1}}\left(\frac{1}{i}, \ldots, \frac{1}{i}, -1, 0, \ldots, 0\right)^T,$$
$$i = 1, \ldots, D-1.$$

The composition $\boldsymbol{x}$ results in the required nonsingular data $\boldsymbol{z}$, but any interpretation in the sense of original compositional parts is not possible.

A way of using the advantageous properties of the ilr transformation for robust factor analysis is the following. The ilr transformation can be utilized to obtain a robust estimation of the covariance matrix of the random vector $\boldsymbol{z}$ from the sample $\boldsymbol{z}_1 = \mathrm{ilr}(\boldsymbol{x}_1), \ldots, \boldsymbol{z}_n = \mathrm{ilr}(\boldsymbol{x}_n)$. Using Eq. (5), the obtained robust covariance matrix $\mathrm{Cov}(\boldsymbol{z})$ is then back-transformed to the clr space by

$$\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{V}\mathrm{Cov}(\boldsymbol{z})\boldsymbol{V}^T.$$

The resulting robust version of $\mathrm{Cov}(\boldsymbol{y})$ can now be used for the parameter estimation in factor analysis, as described above, and the interpretation refers to the clr space.

Note that the same results would be obtained by using the additive logratio transformation (Aitchison, 1986), independent from the denominator chosen. Since there is a linear relation between all logratio transformations, any robust estimator which is affine equivariant will lead to the same robust estimation of $\mathrm{Cov}(\boldsymbol{y})$ (Filzmoser and Hron, 2008).

4. Illustrative example

For illustrating the theoretical results we will use the Kola data (Reimann et al., 1998). The data set is available in the R package *StatDA* (R development core team, 2008). This data set resulted from a large ecogeochemical mapping project carried out from 1993 to 1998 by the Geological Surveys of Finland and Norway and the Central Kola Expedition, Russia. In total, samples in five different layers were collected from more than 600 sites and analyzed for the concentration of various chemical elements. In the following we will use the data from the moss layer. Elements with problems due to censoring were not used.

Because several different processes (e.g., sea spray, contamination) are present in the project area which lead to inhomogeneities in the data, a robust analysis can be more stable than a classical analysis. A robust analysis will thus focus more on the major data structure, and it will be less influenced by artefacts and inhomogeneities. In the following we will thus compare results from classical and robust factor analyses. Moreover, a comparison

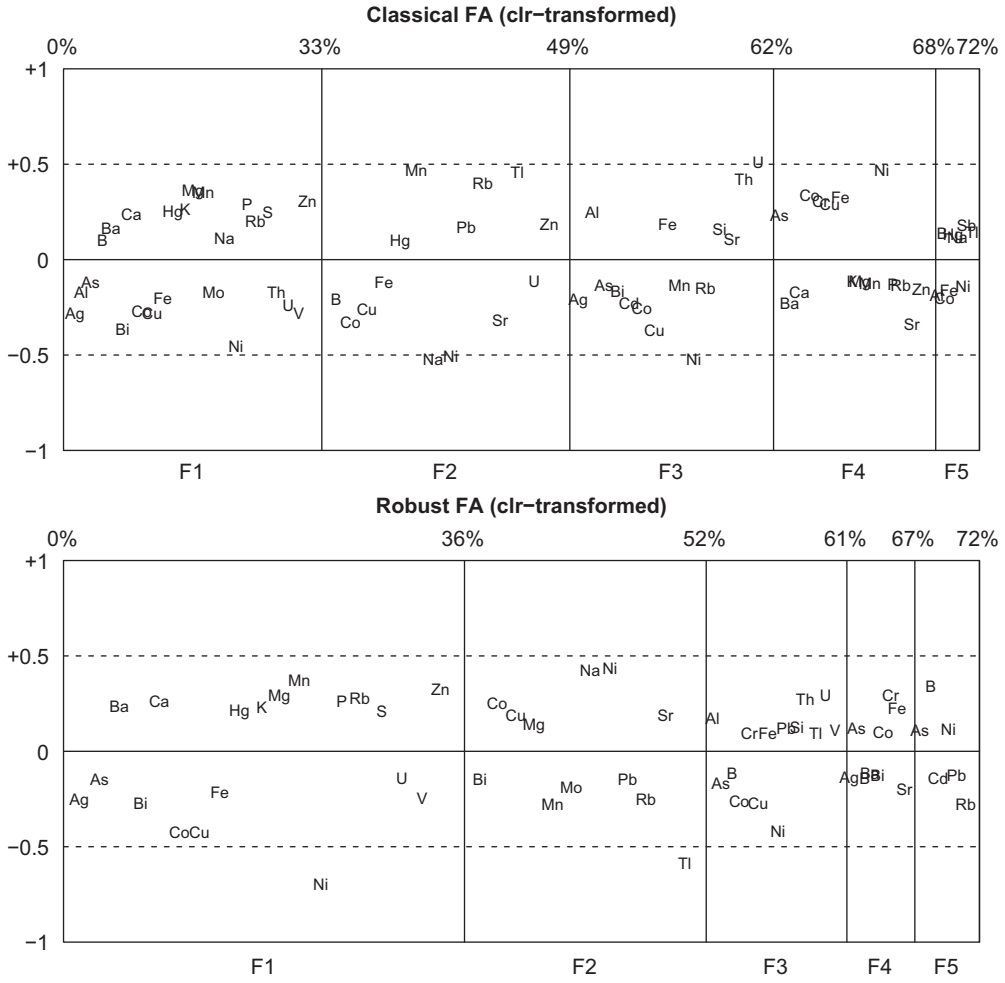

Fig. 2. Factor loading plots for classical (top) and robust (bottom) factor analysis based on clr transformed data of Kola project moss layer. Elements with loadings between −0.1 and 0.1 are not plotted.

is made for the suggested factor analysis based on the clr transformed data with a factor analysis on the log-transformed data.

We used principal factor analysis (PFA) with a varimax rotation of the factors in order to achieve a better interpretation of the resulting factors (Reimann et al., 2002). In all analyses the number of factors was 5, which resulted in a reasonable percentage of explained variability, and allowed a better comparison among the different strategies. Fig. 2 shows the resulting loading plots of classical (top) and robust (bottom) factor analysis for the clr

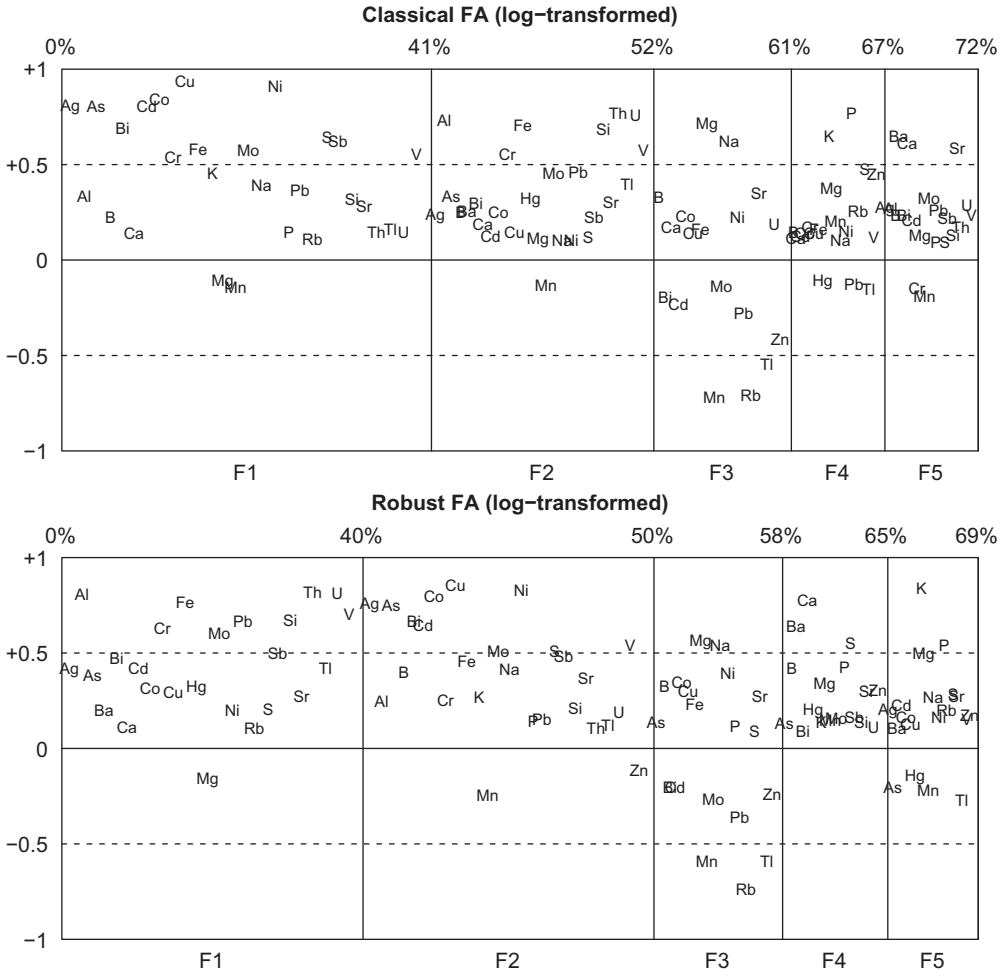

Fig. 3. Factor loading plots for classical (top) and robust (bottom) factor analysis based on log-transformed data of Kola project moss layer.

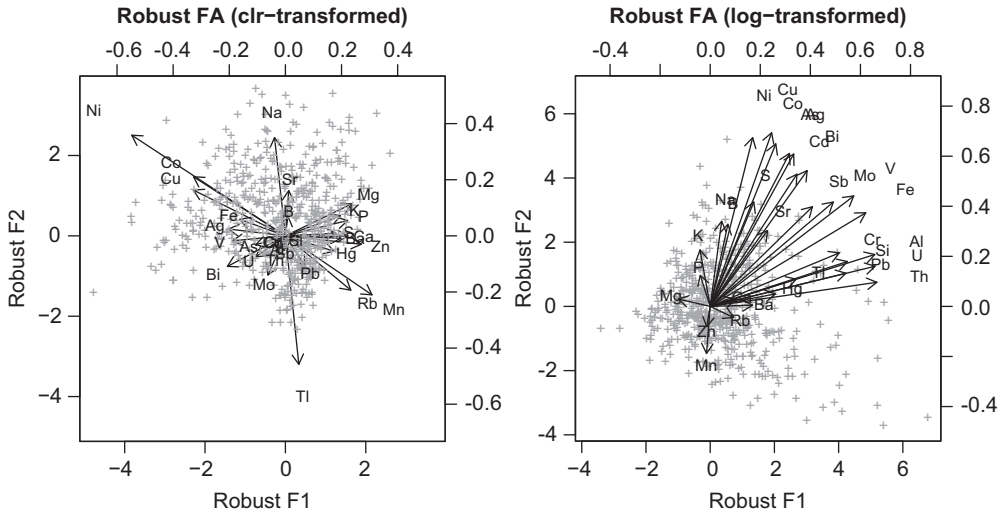

Fig. 4. Biplots for first and second factor of robust factor analysis for clr transformed (left) and log-transformed (right) data of Kola project moss layer.

transformed data. The position of the element names in the plot reflects the loading of the element on the different factors. In addition, the percentages at the top of the plots display the cumulative explained percentage of total variability. The scale on the horizontal axis is according to the relative amount of variability explained by each single factor (see Reimann et al., 2008). In both cases, five factors explain about 70% of the total variability, and an addition of a further factor would result only in marginal improvement of the explanation. The percentages explained by each single factor change from the classical to the robust analysis, especially for the first factor. The loadings do not change dramatically, but some shifts in their magnitudes are clearly visible. Factor 2 has changed the sign. The percentages explained by each factor are generally similar, but the percentages explained by Factors 1 and 2 are increased in the robust analysis. This is to be expected as the robust analysis focuses on the underlying processes influencing the majority of the data by reducing the influence of outliers. The first factor, explaining about a third of the variability, reflects contaminant releases from Russian industrial sites at Monchegorsk, Nikel and Zapoljarnij—negative loadings, while positive loadings reflect natural biological processes in the mosses upon which the contaminants have been deposited. In Factor 2 robust positive loadings reflect the deposition of sea spray transported inland close to the north coast in contrast to the dominant moss biological elemental association (negative loadings). Factor 3 reflects the influence of airborne particulates from two different sources: a smelter source in the negative loadings and a natural aluminosilicate mineral and apatite mining, Apatity and Kovdor, source in the positive loadings. More details on the interpretation of the factors are provided in Reimann et al. (2008). There are also details provided concerning the Kola project. A map of the survey area is shown in Fig. 5 (upper left).

As a comparison to the factor analyses of the clr transformed data, Fig. 3 provides the resulting loading plots for classical (top) and robust (bottom) factor analysis of the log-transformed data.

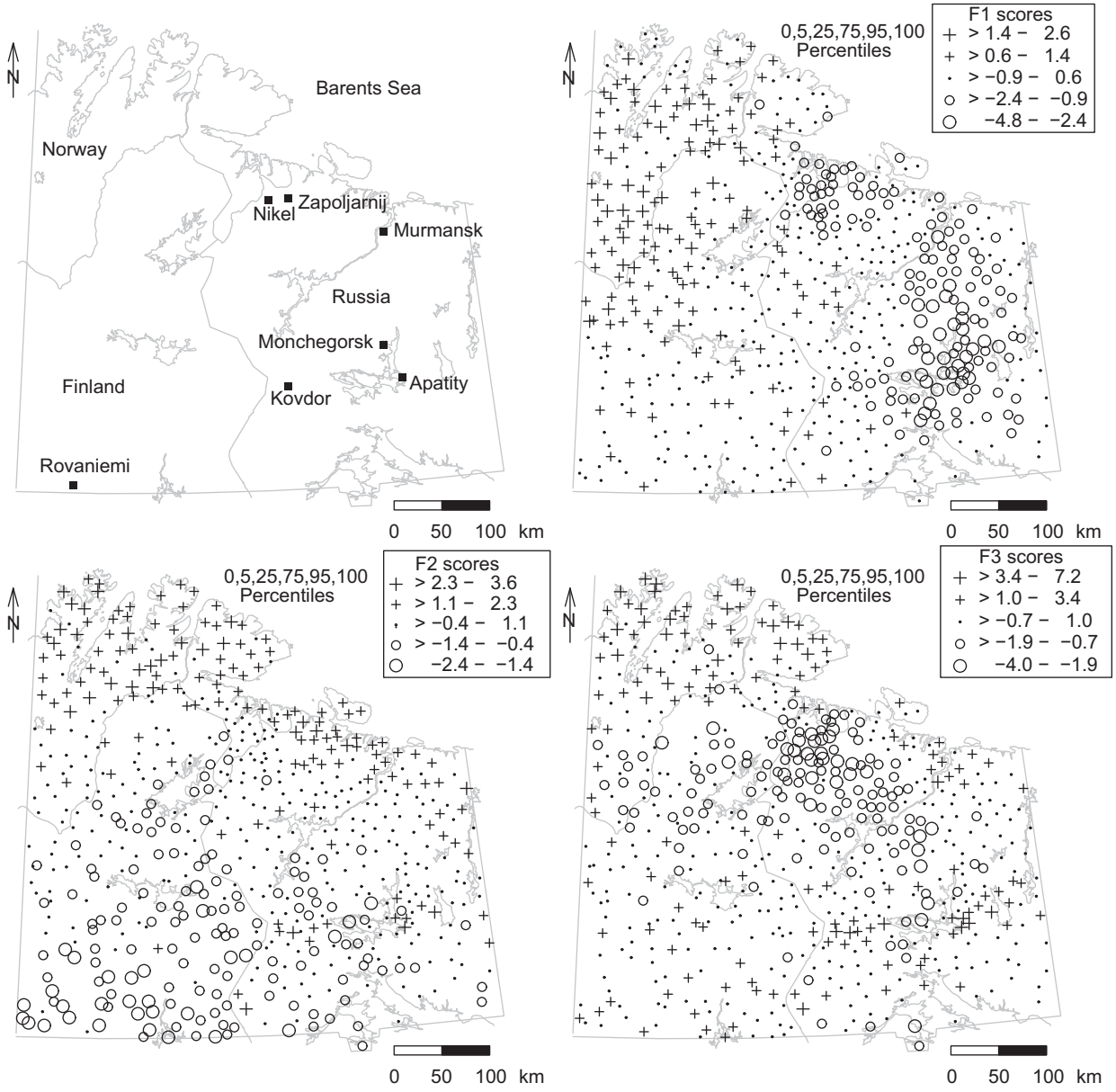

Fig. 5. Map of survey area (upper left), and factor scores for first (upper right), second (lower left), and third (lower right) factor of robust factor analysis for clr transformed data of Kola project moss layer.

The inappropriate geometry resulting from the log-transformation becomes visible by the absence of high negative loadings (with a few exceptions). The interpretation of the factors changes completely relative to the previous analysis. Moreover, the analysis seems to be quite unstable because the relevance of the factors changes from classical in comparison with robust analysis, and thus also the order of the factors.

The resulting loadings and scores for Factors 1 and 2 from the robust analyses are shown in Fig. 4 in the form of biplots. The biplot for the log-transformed data (right) shows the artefact that the loadings are essentially concentrated in the positive quadrant of the plot. This figure best reflects the major differences between using an appropriate transformation or using simply the log-transformed data. In the latter case (Fig. 4, right) the relation between the variables would indicate the dominance of industrial contamination. The left biplot, however, reveals the presence of several different underlying processes, like sea spray (Na, Sr), a variable group related to plant nutrients (Mg, K, P, S, Ca, Ba, Zn, Hg), and contamination (Cu, Co, Ni).

The factor scores relate to the values at the sample locations in the project area, and each factor can thus be presented by a map. We use the so-called EDA symbol set for coding the values of the factor scores (Reimann et al., 2008). The maps shown in Fig. 5 are the location map (upper left), and the scores for the first three factors of robust PFA for the clr transformed data. Considering the left biplot in Fig. 4 and the loading plot in Fig. 2 (bottom), low scores on Factor F1 indicate industrial contamination from Monchegorsk and Nikel-Zapoljarnij, and to a lesser extent Apatity and Kovdor. Factor F2 is dominated by the effects of sea spray blown inland, and is clearly visible in the high scores south of the Barents Sea coast. The larger area of influence in the northwest is a result of the severity of Atlantic storms, topography and lack of forest cover. The effects of open-pit mining of alkalic rocks at Apatity and Kovdor are visible in positive Factor F3 scores. This factor has a geochemical association indicative of silicate rock particulates and dust. In the Apatity and Kovdor areas these are of industrial origin, in the far northwest high scores of this factor are also present in a mountainous area lacking forest cover; this

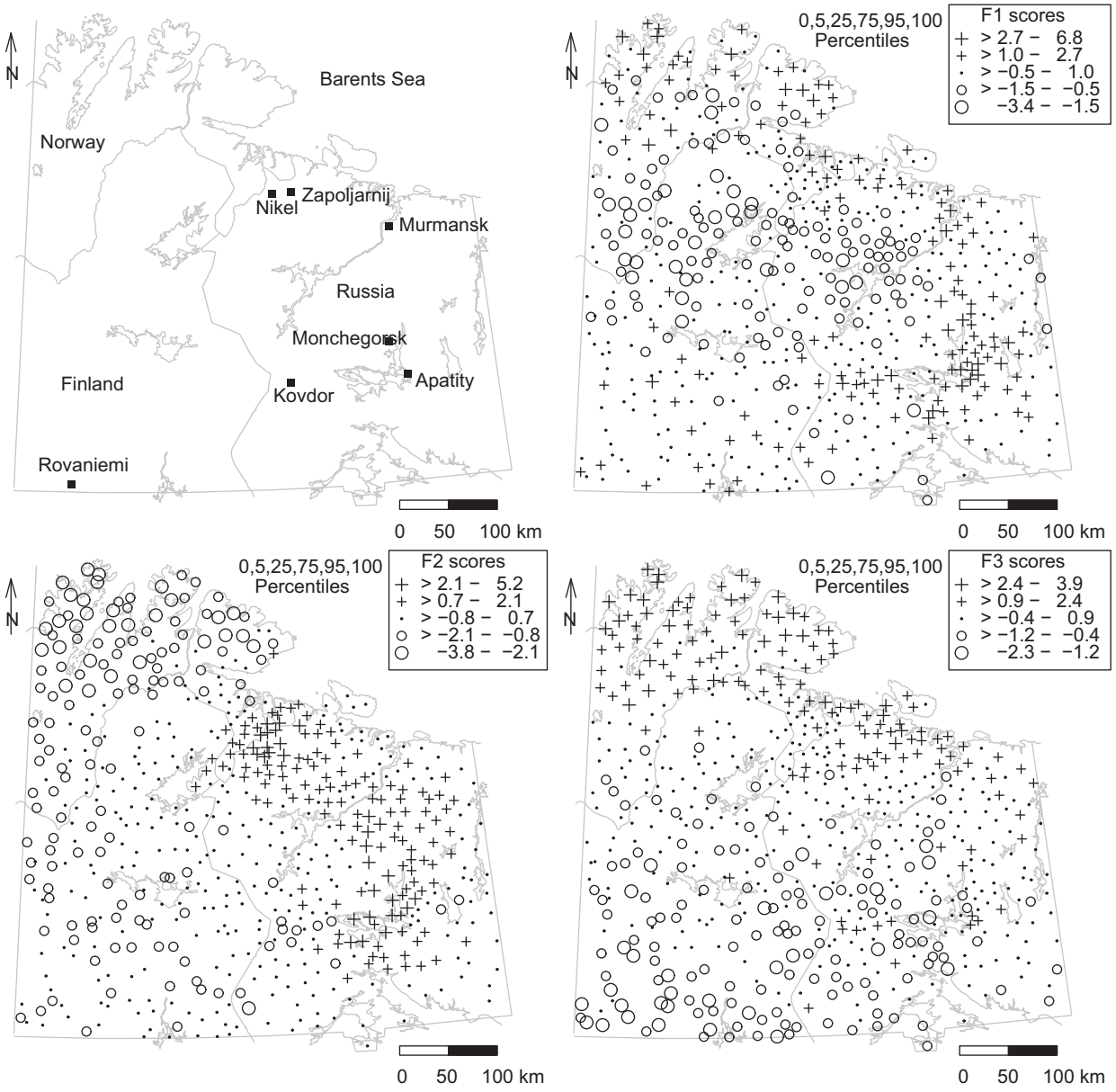

Fig. 6. Map of survey area (upper left), and factor scores for first (upper right), second (lower left), and third (lower right) factor of robust factor analysis for log-transformed data of Kola project moss layer.

facilitates the movement of natural rock weathering products as fine particulates. The combination of the maps of Factors F1 and F3 permits a discrimination between the industrial contamination from the two major sources in the area. Firstly, the processing of Cu–Ni ores and matte at Monchegorsk and Nikel-Zapoljarnij, and, secondly, the open-pit mining of alkalic rocks at Apatity and Kovdor.

In comparison to Fig. 5, the scores from robust factor analysis for the log-transformed data are shown in Fig. 6 for the first three factors. As it can be anticipated from the loading plots, the maps show completely different structures. Interestingly, the maps still show clear regional patterns that relate to certain phenomena in the project area. However, it could be very misleading to interpret these scores, they fail to reveal the true complexity of the different processes active in the environment. Factor F1 shows a combination of sea spray along the coast and contamination from mining and processing of alkaline rocks at Apatity and Kovdor. Factor F2 shows the extent of contamination from the Russian nickel industry at Monchegorsk and Nikel-Zapoljarnij, and the extent of the relatively pristine environment in the west and northwest remote from industrial sources and under the influence of North Atlantic air masses. The map of Factor F3 dominantly reflects the effects of sea spray, and to a lesser extent the presence of mining at Kovdor and Apatity.

The improved ability to identify and differentiate between the different processes active in the Kola environment with the robust clr transformed data demonstrates in a practical application the superiority of undertaking the PFA following an appropriate transformation.

5. Conclusions

We have introduced an approach to robust factor analysis for compositional data that uses the centred logratio transformation. Robust estimation is performed in the space of the isometric logratio transformed data where problems of singularity no longer occur. Since an interpretation of the results is not possible in the ilr space, the results are back-transformed to the clr space. They can be presented as usual in the form of biplots, but it has to be considered that the resulting biplots are compositional biplots which have to be interpreted appropriately (Aitchison and Greenacre, 2002). Using the moss layer of the Kola data (Reimann et al., 1998), this approach is compared with classical factor analysis based on clr transformed data, and with robust and classical factor analysis for log-transformed data. The results for the log-transformed data are completely different, which reflects the improper geometry of the log-transformed space. Robust factor analysis carried out in the ilr space is successful in identifying interesting relevant processes active in the Kola environment.

The programs to generate the figures in this paper are based on R, and they are made available at http://www.statistik.tuwien.ac.at/public/filz/programs/.

Acknowledgements

The authors would like to thank Dr. Raimon Tolosana-Delgado and an anonymous reviewer for their constructive comments, which helped to greatly improve this paper. This work was supported by the Council of the Czech Government MSM 6198959214.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London, 416pp.
Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. Applied Statistics 51, 375–392.
Basilevsky, A., 1994. Statistical Factor Analysis and Related Methods. Theory and Applications. Wiley, New York, 737pp.
Chayes, F., 1960. On correlation between variables of constant sum. Journal of Geophysical Research 65 (12), 4185–4193.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueraz, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35 (3), 279–300.
Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Mathematical Geology 37 (7), 795–828.
Filzmoser, P., Hron, K., 2008. Outlier detection for compositional data using robust methods. Mathematical Geosciences 40 (3), 233–248.
Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58 (3), 453–467.
Johnson, R., Wichern, D., 2007. Applied Multivariate Statistical Analysis, sixth ed. Prentice-Hall, London, 816pp.
Pison, G., Rousseeuw, P.J., Filzmoser, P., Croux, C., 2003. Robust factor analysis. Journal of Multivariate Analysis 84, 145–172.
R development core team, 2008. R: a language and environment for statistical computing, Vienna. ⟨http://www.r-project.org⟩.
Reimann, C., Äyräs, M., Chekushin, V., Bogatyrev, I., Boyd, R., de Caritat, P., Dutter, R., Finne, T., Halleraker, J., Jæger, O., Kashulina, G., Lehto, O., Niskavaara, H., Pavlov, V., Räisänen, M., Strand, T., Volden, T., 1998. 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, 745pp.
Reimann, C., Filzmoser, P., Garrett, R.G., 2002. Factor analysis applied to regional geochemical data: problems and possibilities. Applied Geochemistry 17, 185–206.
Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. Wiley, Chichester, 362pp.
Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.
Tolosana-Delgado, R., Otero, N., Pawlowsky-Glahn, V., Soler, A., 2005. Latent compositional factors in the Llobregat river basin (Spain) hydrochemistry. Mathematical Geology 37 (7), 681–702.