# Robust Methods for Compositional Data

Peter Filzmoser[1] and Karel Hron[2]

[1] Vienna University of Technology
Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria, *P.Filzmoser@tuwien.ac.at*
[2] Palacký University, Faculty of Science
tř. 17. listopadu 12, CZ-77146 Czech Republic, *hronk@seznam.cz*

**Abstract.** Many practical data sets in environmental sciences, official statistics and various other disciplines are in fact compositional data because only the ratios between the variables are informative. Compositional data are represented in the Aitchison geometry on the simplex, and for applying statistical methods designed for the Euclidean geometry they need to be transformed first. The isometric logratio (ilr) transformation has the best geometrical properties, and it avoids the singularity problem introduced by the centered logratio (clr) transformation. Robust multivariate methods which are based on a robust covariance estimation can thus only be used with ilr transformed data. However, usually the results are difficult to interpret because the ilr coordinates are formed by non-linear combinations of the original variables. We show for different multivariate methods how robustness can be managed for compositional data, and provide algorithms for the computation.

**Keywords:** Aitchison geometry, logratio transformations, robustness, affine equivariance, multivariate statistical methods

## 1 Compositional data and logratio transformations

Practical data sets are frequently characterized by multivariate observations containing relative contributions of parts on a whole. Examples are concentrations of chemical elements in a rock, household expenditures on various commodities from the monthly salary, or representations of various animal species in a study area in percentages. Often just percentages are used to express the mentioned relative magnitudes of the parts of the data and thus the simplex is usually referred to be the sample space. However, the situation is a more general one, because only the relevant information in the data is contained in the ratios between the parts. From this point of view, the percentages represent only a proper representation of the information, contained in the multivariate observations. These considerations led John Aitchison at the beginning of the eighties of the 20th century to introduce the term *compositional data* (or compositions for short) to characterize such kind of data and to propose possibilities for their statistical analysis using so-called logratio transformations.

The geometry of compositions, later denoted as the Aitchison geometry, follows their special properties and is based on special operations of perturbation, power transformation and the Aitchison inner product. In more detail,

for $D$-part compositions $\mathbf{x} = (x_1, \ldots, x_D)'$ and $\mathbf{y} = (y_1, \ldots, y_D)'$ and a real number $\alpha$, this results in compositions

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D),\ \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)$$

and a real number

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j},$$

respectively. Using usual Hilbert space properties, the Aitchison inner product also leads to the definitions of the Aitchison norm and distance. Moreover, the symbol $\mathcal{C}$ denotes a closure operation that moves the sum of the compositional parts to any chosen constant $\kappa$ without loss of information. As mentioned above, the constant $\kappa$ is usually chosen as 1 or 100 in order to represent the compositions on the $D$-part simplex (of dimension $D-1$),

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, \ldots, x_D)',\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa\}.$$

From the geometrical properties of compositional data it is easy to see that using standard statistical methods like principal component analysis, factor analysis or correlation analysis, designed for Euclidean space properties of standard multivariate data with absolute scale, can lead (and frequently does) to meaningless results. This has been demonstrated in various examples, e.g. the book Aitchison (1986), and further Aitchison et al.(2000), Filzmoser et al. (2009a), Filzmoser et al. (2009b), Pearson (1897).

Although the Aitchison geometry on the simplex has the usual properties that are known from the Euclidean geometry (Hilbert space), it is more natural to directly work in the Euclidean space. This means that a transformation of the compositional data from the simplex sample space to the Euclidean space is performed. In the transformed space the standard multivariate methods can be used. The main idea that leads to such transformations is to find a basis (or a generating system) and to express compositions in coefficients of such a basis (coordinate system). This class of mappings is widely known under the term logratio transformations. Nowadays three main approaches using the logratio family are used: additive, centered and isometric logratio transformations (coordinates). All of them move the operations of perturbation and power transformation to the usual vector addition and scalar multiplication. However, only the latter two transformations move the whole Aitchison geometry to the Euclidean one, i.e. including the Aitchison inner product. As the proposed transformations are one-to-one transformations, the obtained results are usually back-transformed to the simplex in order to simplify the interpretation.

The additive logratio transformation follows the idea to construct a (non-orthonormal) basis which is very easy to interpret. Thus, for a composition $\mathbf{x}$,

a special case of the *additive logratio (alr) transformations* (Aitchison, 1986) to $\mathbf{R}^{D-1}$, is defined as

$$alr(\mathbf{x}) = \left( \ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right)'.$$

It is easy to see that also another part can be used as ratioing part in the denominator. It is usually chosen in such a way that the interpretation of the result is facilitated. Note that different alr transformations are related by linear transformations (see, e.g., Filzmoser and Hron, 2008).

Taking a generating system on the simplex leads to the *centered logratio (clr) transformation* (Aitchison, 1986) to $\mathbf{R}^D$,

$$clr(\mathbf{x}) = \left( \ln \frac{x_1}{\prod_{i=1}^{D} x_i}, \ldots, \ln \frac{x_D}{\prod_{i=1}^{D} x_i} \right)'.$$

This transformation has also a good interpretability, and the compositional biplot (Aitchison and Greenacre, 2002), nowadays a very popular exploratory tool, takes advantage of this property. However, as the dimension of the simplex is only $D-1$, the clr transformation is singular, namely, the sum of the obtained coordinates is equal to zero. As a consequence, this makes the use of the robust statistical methods mentioned in the following section impossible.

The last proposal refers to the *isometric logratio (ilr)* transformations (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2005) from the simplex to $\mathbf{R}^{D-1}$, where the main idea is to express the coordinates in an orthonormal basis on the simplex. However, the corresponding coordinates are often not easy to interpret; one such choice of the orthonormal basis leads to

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

Thus, in spite of their advantageous geometrical properties, the ilr transformations are preferably used for methods where the interpretation is focused on the objects rather than on the single compositional parts, because in the latter case a consequent interpretation of the results in coordinates would be necessary. From the definition it is easy to see that all the ilr coordinates are mutually joined with orthogonal relations. An intuitive relation can be found also between clr and ilr transformations. Namely, the ilr coordinates are in fact coordinates of an orthonormal basis on the hyperplane $\mathcal{H}$, formed by the clr transformation. Thus also the relation $ilr(\mathbf{x}) = \mathbf{U} clr(\mathbf{x})$ holds, where the $(D-1) \times D$ matrix $\mathbf{U}$ contains in its rows the mentioned orthonormal basis on $\mathcal{H}$, and $\mathbf{U}\mathbf{U}' = \mathbf{I}_{D-1}$ (identity matrix of order $D-1$) is fulfilled.

Even more general, it has been shown that all three mentioned logratio transformations are mutually joined with linear relations (see, e.g., Filzmoser and Hron, 2008). This property is crucial for the robustification of statistical methods for compositional data.

## 2 Robustness for compositional data

Outliers and data inhomogeneities are typical problems of real data sets. This can severely affect classical multivariate statistical methods that ignore these problems, and the results might then even become meaningless. For this reason, robust statistical approaches were developed that reduce the influence of outliers and focus on the main data structure. An example is the estimation of multivariate location and covariance. The classical estimators, arithmetic mean and sample covariance matrix, are sensitive to outlying observations in the data set while robust estimators can resist a certain proportion of contamination. Among the various proposed robust estimators of multivariate location and covariance, the MCD (Minimum Covariance Determinant) estimator (see, e.g., Maronna et al., 2006) became very popular because of its good robustness properties and a fast algorithm for its computation (Rousseeuw and Van Driessen, 1999).

Besides robustness properties the property of affine equivariance of the estimators of location and covariance plays an important role. The location estimator $T$ and the covariance estimator $C$ are called affine equivariant, if for a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of $n$ observations in $\mathbf{R}^{D-1}$, any nonsingular $(D-1) \times (D-1)$ matrix $\mathbf{A}$ and for any vector $\mathbf{b} \in \mathbf{R}^{D-1}$ the conditions

$$T(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}) = \mathbf{A}T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{b},$$
$$C(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}) = \mathbf{A}C(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{A}'$$

are fulfilled. The MCD estimator shares the property of affine equivariance for both the resulting location and covariance estimator.

Since robust methods are usually designed for the Euclidean geometry and not for the simplex, a transformation of the raw compositional data is required. As mentioned earlier, the clr transformation is not useful for this purpose because robust estimators cannot deal with singular data. The alr and ilr transformations are thus possible transformations prior to robust estimation. However, it depends on the multivariate method and on the purpose of the analysis which of the alr and ilr transformations are useful. This issue will be discussed in more detail in the following sections for multivariate outlier detection, principal component analysis, and factor analysis.

### 2.1 Multivariate outlier detection

The Mahalanobis distance, defined for regular $(D-1)$-dimensional data as

$$\mathrm{MD}(\mathbf{x}_i) = \left[(\mathbf{x}_i - T)' C^{-1} (\mathbf{x}_i - T)\right]^{1/2},$$

is a popular tool for outlier detection (Maronna et al., 2006; Filzmoser et al., 2008). Here, the estimated covariance structure is used to assign a distance to each observation indicating how far the observation is from the center

of the data cloud with respect to the covariance structure. The choice of the location estimator $T$ and the scatter estimator $C$ is crucial. In case of multivariate normal distribution, the (squared) Mahalanobis distances based on the classical estimators arithmetic mean and sample covariance matrix follow approximately a $\chi^2$ distribution with $D-1$ degrees of freedom. In presence of outliers, however, only robust estimators of $T$ and $C$ lead to a Mahalanobis distance being reliable for outlier detection. Usually, also in this case a $\chi^2$ distribution with $D-1$ degrees of freedom is used as an approximate distribution, and a certain quantile (e.g. the quantile 0.975) is used as a cut-off value for outlier identification: observations with larger (squared) robust Mahalanobis distance are considered as potential outliers.

Compositional data need to be transformed prior to computing Mahalanobis distances. The linear relations between the logratio transformations can be used to prove that the Mahalanobis distances are the same for all possible alr and ilr transformations. This, however, is only valid if the location estimator $T$ and the covariance estimator $C$ are affine equivariant (Filzmoser and Hron, 2008). If the arithmetic mean and the sample covariance matrix are used, this holds also for the alr, clr and ilr transformations, where the inverse of the covariance matrix in the clr case is replaced by its Moore-Penrose inverse.

### 2.2 Principal component analysis (PCA)

PCA is one of the most popular tools for multivariate data analysis. Its goal is to explain as much information contained in the data as possible using as few (principal) components as possible (see, e.g., Reimann et al., 2008). In the case of compositional data, it is very popular to display both loadings and scores of the first two principal components by means of biplots. The compositional biplot is usually constructed for clr transformed data, where the resulting loadings and scores have an intuitive interpretation corresponding to the nature of compositions (Aitchison and Greenacre, 2002). However, it is not possible to robustify it because of the mentioned singularity of the clr transformation. As a way out, the ilr transformation can be used to compute the robust loadings and scores, which are then back-transformed to the clr space (Filzmoser et al., 2009a). In more detail, the $n \times (D-1)$ matrix $\mathbf{Z}_{ilr}$ of scores and $(D-1) \times (D-1)$ matrix $\mathbf{G}_{ilr}$ of loadings are transformed to $n \times D$ and $D \times D$ matrices

$$\mathbf{Z}_{clr} = \mathbf{Z}_{ilr}\mathbf{U} \text{ and } \mathbf{G}_{clr} = \mathbf{U}'\mathbf{G}_{ilr}\mathbf{U},$$

respectively. Thanks to the properties of the matrix $\mathbf{U}$ defined earlier in Section 1, and the affine equivariance of the MCD estimator, the resulting principal components correspond to the same nonzero eigenvalues for both ilr and clr. Thus, such a transformation of the loadings and scores can be used to obtain a robust compositional biplot of compositional data.

### 2.3 Factor analysis

Both PCA and factor analysis are based on the same objection, namely on reduction of the dimensionality, and the main principle is to decompose the multivariate data into loadings and scores. However, the more strict definition of factor analysis implies that the number of factors to be extracted is defined at the beginning of the procedure. In addition, an estimate of the proportion of variability has to be provided for each variable, which is not to be included in the factors but is considered unique to that variable (Reimann et al., 2008). This often leads to a better interpretation of the (rotated) factors than (rotated) principal components. However, the definition of factor analysis and the uniquenesses induce problems in case of compositions, where the treatment of the single compositional parts seems to be questionable. Here the clr transformation offers again a reasonable solution, however, the procedure to estimate the 'clr-uniquenesses' and loadings must be performed in an iterative manner and also the estimation of scores has to overcome the singularity of the clr transformed data, see Filzmoser et al. (2009b), for details. The key to robustness is again contained in the estimation of the covariance matrix where the same approach as for PCA can be used.

## 3 Real data example

The methods are demonstrated in the following using a data example of mean consumption expenditures of households from 2008 in the countries of the European Union. The data set is available at `http://epp.eurostat.ec.europa.eu` at `statistics_explained/index.php/Household_consumption_expenditure`. The expenditures on food, alcohol and tobacco, clothing, housing, furnishings, health, transport, communications, recreation, education, restaurants and hotels, and on other goods and services are reported for the countries Austria (A), Belgium (B), Bulgaria (BG), Cyprus (CY), Czech Republic (CZ), Denmark (DK), Estonia (EST), Finland (FIN), France (F), Germany (D), Greece (GR), Hungary (H), Ireland (IRL), Italy (I), Latvia (LV), Lithuania (LT), Luxembourg (L), Malta (M), Netherlands (NL), Poland (PL), Portugal (P), Romania (R), Slovakia (SK), Slovenia (SLO), Spain (ES), Sweden (S), and United Kingdom (GB). These are compositional data because the expenditures are parts of the overall household incomes. For example, if more money is devoted to one part, typically less money will be left for the other parts, and thus not the absolute number but only their ratios are informative.

At first we apply multivariate outlier detection using Mahalanobis distances. Location and covariance are estimated in a classical way but also robustly using the MCD estimator. Both variants are applied to the original untransformed data, and to the ilr transformed data. The results are presented with distance-distance plots (Rousseeuw and Van Driessen, 1999) in Figure 1. The dashed lines correspond to the outlier cut-offs using the

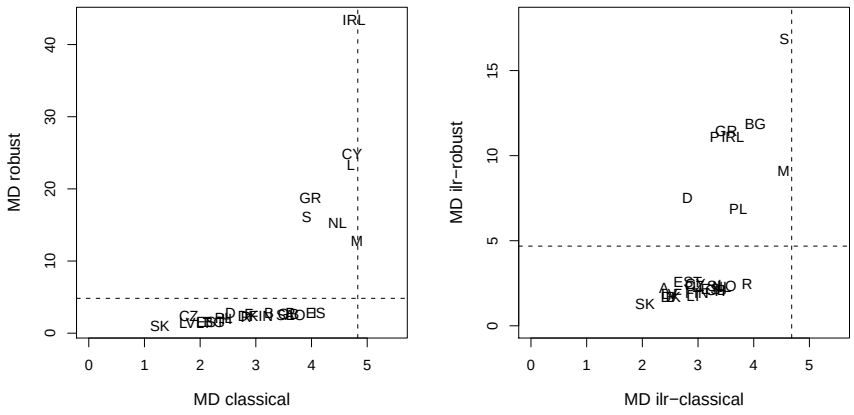


**Fig. 1.** Distance-distance plots for outlier detection using the untransformed (left) and the ilr transformed (right) data.

0.975 quantile of the corresponding $\chi^2$ distributions. Both figures reveal several masked outliers, but they do not indicate the same outliers. Without any further inspection it would be difficult to interpret the results. Since the data are of compositional nature, only the distance-distance plot of the ilr transformed data is reliable.

A deeper insight into the multivariate data structure can be achieved by a PCA. We want to compare the classical and the robust (MCD) approach, as well as PCA for the untransformed and the ilr transformed data (back-transformed to the clr space). The resulting biplots are shown in Figure 2. We can see a typical phenomenon when analyzing compositional data with inappropriate methods: all variables are positively correlated for the untransformed data which is an artifact resulting from the underlying geometry. Note that a robust analysis cannot 'repair' this geometrical artifact. In contrast, the biplots based on the ilr ($\rightarrow$ clr) transformed data show quite different relations between the variables. Using the biplots, it is also possible to explain the multivariate outliers identified in Figure 1. The robust biplot for the ilr ($\rightarrow$ clr) transformed data shows striking differences for expenditures on health. It is also interesting to see that on the right-hand side of the plot we find potentially richer countries (with a higher GDP) whereas on the left-hand side the poorer countries are located. The latter devote a larger part of their expenditures to food, communication, and alcohol and tobacco.

Similar to PCA we apply factor analysis to the original and to the ilr ($\rightarrow$ clr) transformed data, and perform in both cases a classical and a robust analysis. The biplots for the untransformed data show again a degenerated behavior. The robust biplot for the ilr ($\rightarrow$ clr) transformed data shows almost contrasting priorities for the expenditures: Positive values on factor 1 are referring to education, while negative values refer to recreation, trans-

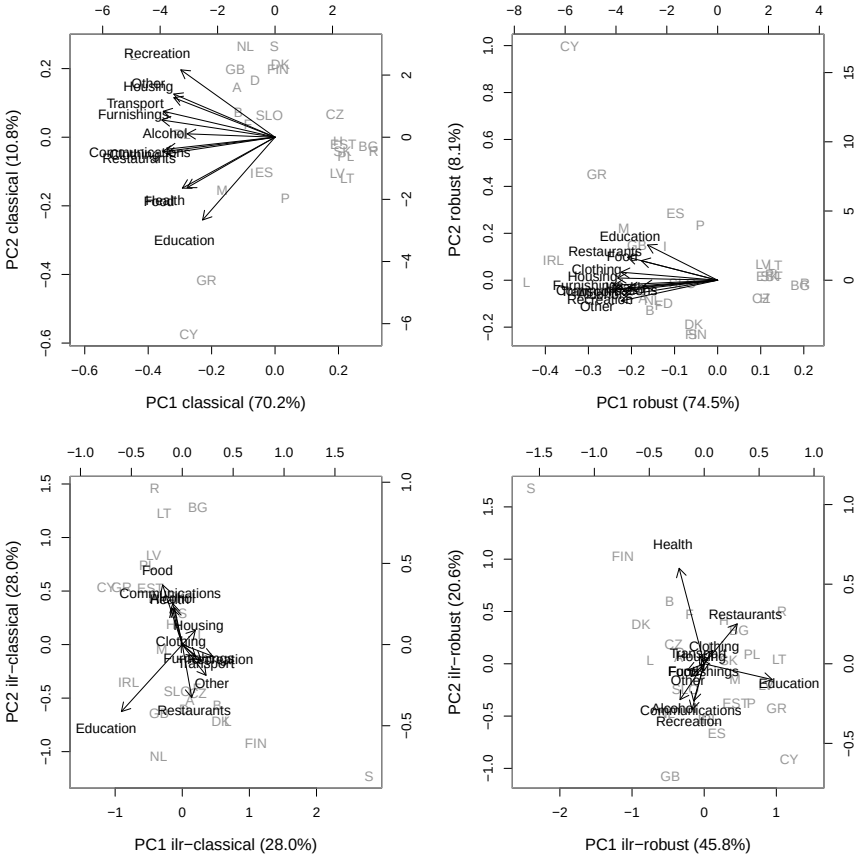


**Fig. 2.** Principal component analysis classical (left column) and robust (right column) using the original scaled (upper row) and the ilr (→ clr) transformed data.

port, housing, furnishings, and other goods and services. Of course, expenses for education are usually set by the political system. Factor 2 reflects the differences in expenditures for restaurants and hotel, versus expenditures for more basic needs like food, alcohol and tobacco, communications, and health. The poorer countries devote a larger proportion of their expenses to these basic needs.

## 4 Conclusions

For compositional data an appropriate transformation is crucial prior to performing any multivariate data analysis. In environmental sciences, like typically in geochemistry, compositional data are frequently simply logarithmically transformed. This transformation, however, can only achieve symmetry

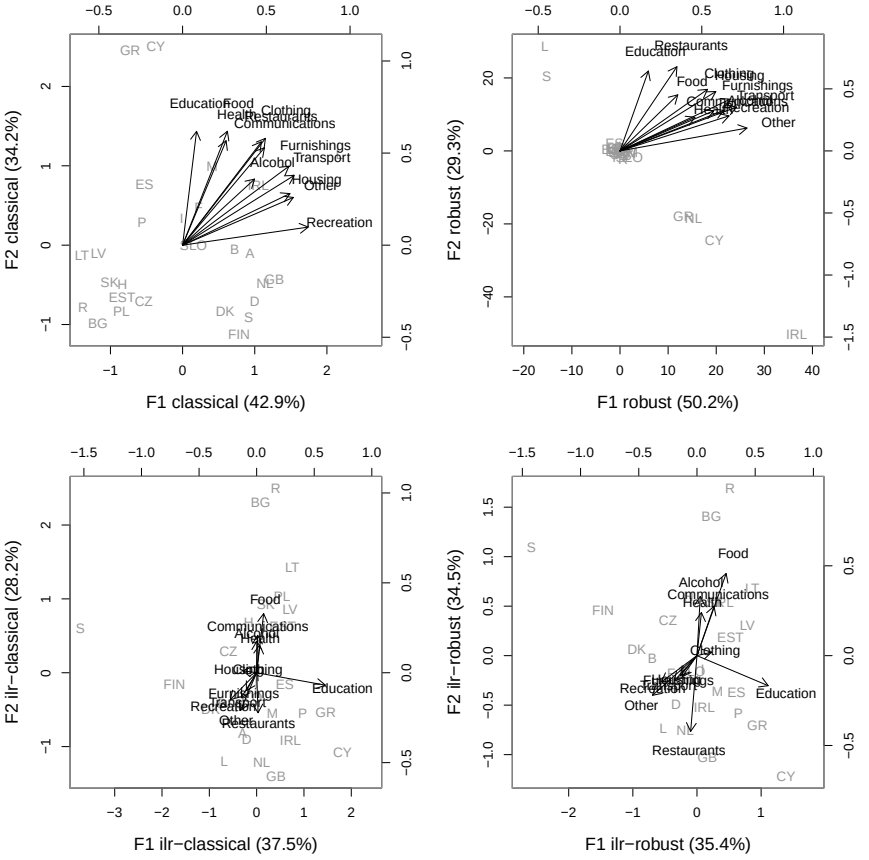


**Fig. 3.** Factor analysis classical (left column) and robust (right column) using the original scaled (upper row) and the ilr (→ clr) transformed data.

of the single variables, but it does not transform the data from the simplex to the Euclidean space. Phenomena like positive variable relations as shown in the upper rows of Figure 2 and 3 are typical outcomes of such an approach.

Generally, the ilr transformation shows the best properties. For a robust analysis, one important property is that the ilr transformed data are in general non-singular, which allows for the application of robust covariance estimators. However, since the ilr variables are difficult to interpret, they usually need to be back-transformed to the clr space, like it has been demonstrated for the compositional biplot.

The transformations, the adapted multivariate methods, and various other representations and statistical methods for compositional data have been implemented in the R library `robCompositions` (Templ et al., 2009).

## References

AITCHISON, J. (1986): *The statistical analysis of compositional data.* Chapman and Hall, London.

AITCHISON, J., BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J.A. and PAWLOWSKY-GLAHN, V. (2000): Logratio analysis and compositional distance. *Mathematical Geology 32 (3), 271-275.*

AITCHISON, J. and GREENACRE, M. (2002): Biplots of compositional data. *Applied Statistics 51, 375-392.*

EGOZCUE, J.J., PAWLOWSKY-GLAHN, V. and MATEU-FIGUERAS, G., BARCELÓ-VIDAL, C. (2003): Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35 (3), 279-300.*

EGOZCUE, J.J.and 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.*

FILZMOSER, P., HRON, K. and REIMANN, C. (2009a): Principal component analysis for compositional data with outliers. *Environmetrics 20, 621-632.*

FILZMOSER, P., HRON, K., REIMANN, C. and GARRETT, R. (2009b): Robust factor analysis for compositional data. *Computers & Geosciences 35 (9), 1854-1861.*

FILZMOSER, P., MARONNA, R. and WERNER, M. (2008): Outlier identification in high dimensions. *Computational Statistics & Data Analysis 52, 1694-1711.*

MARONNA, R., MARTIN, R.D. and YOHAI, V.J. (2006): *Robust statistics: theory and methods.* Wiley, New York.

PEARSON, K. (1897): 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. *Proceedings of the Royal Society of London 60, 489-502.*

REIMANN, C., FILZMOSER, P., GARRETT, R. and DUTTER, R. (2008): *Statistical data analysis explained: Applied environmental statistics with R.* Wiley, Chichester.

ROUSSEEUW, P. and VAN DRIESSEN, K. (1999): A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41, 212-223.*

TEMPL, M., HRON, K. and FILZMOSER, P. (2009): robCompositions: Robust estimation for compositional data, http://www.r-project.org, R package version 1.2, 2009.