Interpretation of multivariate outliers for compositional data

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

a Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria
b Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Faculty of Science, 17. listopadu 12, CZ-77146 Olomouc, Czech Republic
c Geological Survey of Norway (NGU), P.O.Box 6315 Sluppen, N-7491 Trondheim, Norway

A R T I C L E   I N F O

Article history:
Received 12 April 2011
Received in revised form 31 May 2011
Accepted 22 June 2011
Available online 13 July 2011

Keywords:
Compositional data
Log-ratio transformations
Outlier detection
Compositional biplot

A B S T R A C T

Compositional data—and most data in geochemistry are of this type—carry relative rather than absolute information. For multivariate outlier detection methods this implies that not the given data but appropriately transformed data need to be used. We use the isometric logratio (ilr) transformation, which seems to be generally the most proper one for theoretical and practical reasons. In this space it is difficult to interpret the outliers, because the reason for outlyingness can be complex. Therefore we introduce tools that support the interpretation of outliers by representing multivariate information in biplots, maps, and univariate scatterplots.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

In many practical applications from geosciences one has to deal with compositional data, i.e., with multivariate observations describing quantitatively the parts of some whole. Thus, their components carry exclusively relative information about the parts (Aitchison, 1986). Typically these observations are expressed as data with a constant sum constraint such as proportions, percentages, or mg/kg. Standard statistical methods usually fail when they are applied directly to compositional data (Filzmoser and Hron, 2008; Filzmoser et al., 2009; Hron et al., 2010). Many authors appear to be under the impression that the main reason lies in a nonnormal distribution of the samples (for example, of chemical elements in a rock) and thus recommend applying a logarithmic transformation in order to achieve normality of chemical elements in a rock) and thus recommend applying a logarithmic transformation (Egozcue and Pawlowsky-Glahn, 2006) on the sample space of compositions. In this space it is difficult to interpret the outliers, because the reason for outlyingness can be complex. Therefore we introduce tools that support the interpretation of outliers by representing multivariate information in biplots, maps, and univariate scatterplots.

The positive constant $\kappa$ stands for 1 in the case of proportions, 100 for percentages, or $10^6$ for mg/kg.

Due to the fact that geochemical data follow the Aitchison geometry, standard statistical methods that rely mostly on the Euclidean geometry cannot be used for raw compositional data. Whether or not the data follow a normal distribution is of no importance at all. To transform the data to the Euclidean space, the family of logratio transformations from the simplex $S^D$ to the Euclidean real space was proposed. Only by following such a transformation is the use of the standard statistical methods possible. The three main types in this family of transformations are the additive logratio (alr), the centered logratio (clr), and the isometric logratio (ilr) transformation. The alr (Aitchison, 1986) is simple and could be used in the context of outlier detection. However, it is not recommended because it does not result in an orthogonal basis system, which is necessary for diagnostic tools following outlier detection. The clr (Aitchison, 1986) results in data singularity, which is in conflict with the usual tools for outlier detection. The ilr (Egozcue et al., 2003) is recommended because it forms a one-to-one relation between the Aitchison geometry on the simplex and the standard Euclidean geometry, with excellent geometrical properties.

The $D-1$ ilr variables are coordinates of an orthonormal basis on the simplex (with respect to the Aitchison geometry); thus a proper choice of the basis seems to be crucial for their interpretation. Here the big step ahead is the sequential binary partition procedure (Egozcue and Pawlowsky-Glahn, 2005), which enables interpretation of the orthonormal coordinates in the sense of balances between groups of compositional parts. Additionally, each ilr variable explains all the logratios, i.e., terms
of type $\ln(x_i/x_j)$, $i, j = 1, \ldots, D$, between parts of the corresponding groups (Filzmoser and Hron, in press); conversely, each logratio in the composition is exclusively explained by one balance. This point of view seems to be meaningful, because the definition of compositions implies that the only relevant information is contained in (log)ratios of compositional parts. Although the sequential binary partition can also be made to measure for the concrete geochemical problems (Buccianti et al., 2008), in practice the following $D$ choices of the orthonormal bases seem to be the most useful (Egozcue et al., 2003; Hron et al., 2010).

Explicitly, we obtain $(D-1)$-dimensional real vectors $z_l = (z_{l1}, \ldots, z_{lD-1})'$, $l = 1, \ldots, D$,

$$z_l^l = \sqrt{\frac{D - i}{D - i + 1}} \ln \frac{x_i^l}{\sqrt{\prod_{j=1}^{D-1} y_j^l}}, \quad i = 1, \ldots, D - 1,$$

where $(x_1^l, x_2^l, \ldots, x_i^l, x_{i+1}^l, \ldots, x_D^l)$ stands for a permutation of the parts $(x_1, \ldots, x_D)$ such that the $l$th compositional part always fills the first position, $(x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_D)$. In such a configuration, the first ilr variable $z_{11}$ explains all the relative information (logratios) about the original compositional part $x_1$; the coordinates $z_{21}', \ldots, z_{D-11}'$ then explain the remaining logratios in the composition (Filzmoser and Hron, in press). Note that the only important position is that of $x_1$ (because it can be fully explained by $z_{11}$), the other parts can be chosen arbitrarily, because different ilr transformations are orthogonal rotations of each other (Egozcue et al., 2003). Of course, we cannot say that $z_{i1}$ is the original compositional part $x_i$ but it explains all the information concerning $x_i$; thus, it stands for $x_i$.

An interesting consequence follows for the known clr transformation from $x^0$ to $\mathbb{R}^D$, resulting in

$$y = (y_1, \ldots, y_D)' = \left( \frac{\ln x_1}{\sqrt{\prod_{i=1}^{D} y_i}}, \ldots, \frac{\ln x_D}{\sqrt{\prod_{i=1}^{D} y_i}} \right)'.$$ 

It is easy to see that there exists a linear relationship between $y_i$ and $x_{i1}^l$, namely

$$y_i = \sqrt{\frac{D}{D - 1}} x_{i1}^l.$$ 

Thus, up to a constant, the single clr variables have the same interpretation as the corresponding ilr coordinates: they explain all logratios concerning the $l$th compositional part. However, as a consequence, some of the logratios are explained more than once by the $D$ clr variables (in contrast to the ilr transformation). This is also an intuitive reason for the resulting singularity $y_1 + \cdots + y_D = 0$ of clr variables, which makes, e.g., the use of robust multivariate statistical methods not possible. On the other hand, the clr transformation is a cornerstone of the compositional biplot (Aitchison and Greenacre, 2002), which will be employed further in the paper.

From the above-mentioned properties of logratio transformations, it is visible that the ordered $D$-tuple of the ilr coordinates, $z_{i1}^l$, $l = 1, \ldots, D$, can be obtained from clr-transformed data as $\sqrt{(D-1)/D} y_i$. Nevertheless, note that it would be not meaningful to interpret the relations between the clr variables or even between the variables $z_{i1}^l$ using their correlation structure; here

---

**Fig. 1.** Illustration of the determination of plot symbols and colors: (A) shows simulated three-part compositional data, (B) is the representation in the ilr space, and (C) is univariate scatterplots of the single univariate ilr variables. The symbols are determined by certain quantiles of the robust Mahalanobis distances, visualized by the ellipses in (B). The colors are determined by computing the distance of the points to the medians of the univariate ilr variables in (C). The medians of the three resulting distances determine the color for each observation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
the subcompositional incoherence (which means that the results of statistical modeling might be incompatible if only a subset of the parts were used; see, e.g., Aitchison, 1986; Filzmoser et al., 2010, for details) could lead to wrong conclusions. Some kind of "incompatibility" is obtained also for the single \( z_i^{(l)} \), \( l = 1, \ldots, D \) variables; however, here it is as a natural consequence of the fact that the information available in \( x \) was reduced just to a subcomposition.

In contrast to univariate outliers, whose identification as extreme observations is straightforward, for multivariate outliers the covariance structure of the data set needs to be considered as well (Filzmoser et al., 2005). Moreover, when working with compositional data, one has to consider the data structure in view of the Aitchison geometry; see, e.g., Hron et al. (2010) and Filzmoser and Hron (in press). For example, an elliptical point cloud arising from a multivariate normal distribution in the usual Euclidean geometry can look very different in the Aitchison geometry, depending on its position in space (see, for instance, the back-transformed ellipses in Fig. 1B to the Aitchison geometry in Fig. 1A). This is important for multivariate outlier detection methods, which are usually based on distances from an elliptically symmetric distribution. Moreover, each compositional data point can be shifted along the line from the origin through the point without changing the ratios of the compositional parts. Formally, an observed composition \( x = (x_1, \ldots, x_p) \) is defined as a member of the corresponding equivalence class of \( x \).

\[
X = \{c x, \quad c \in \mathbb{R}^+ \}.
\]

Thus, two compositions that are elements of the same equivalence class \( X \) (we call them also compositionally equivalent; see Egozcue and Pawlowsky-Glahn, 2006) contain the same information and have zero Aitchison (1986) distance. From this point of view, the "extremeness" of the outliers can be even more misleading than in case of standard (Euclidean) multivariate outliers.

The methods for outlier detection of compositional data will be discussed in Section 2, where both theoretical aspects and possibilities for graphical representations will be considered. Section 3 proposes several tools for the interpretation of multivariate outliers that have been implemented in the statistical software environment R (R Development Core Team, 2011). In Section 4 we show how the tool is used for real problems, and how results can be interpreted. Section 5 concludes.

2. Methods for multivariate outlier detection and graphical representation

As with the other multivariate methods applied to compositional data, it is important to use an appropriate data transformation first. Either the clr transformation or a proper choice of the ilr transformation can be used for this purpose; see Filzmoser and Hron (2008).

2.1. Theoretical aspects of outlier detection for compositional data

A representation of the compositional data in coordinates (i.e., the representation following an ilr transformation) makes it possible to apply all methods devised for an unconstrained sample space. In particular, multivariate outlier detection can be based on Mahalanobis distances, defined for a sample \( z_1, \ldots, z_n \) of \( (D-1) \)-dimensional observations (resulting for instance from an ilr transformation of the corresponding compositional sample) as

\[
MD(z_i) = \{z_i - T \} C^{-1} (z_i - T)^1/2, \quad i = 1, \ldots, n.
\]

Here, \( T = T(z_1, \ldots, z_n) \) and \( C = C(z_1, \ldots, z_n) \) are location and covariance estimators, respectively. The choice of the estimators is crucial for the quality of multivariate outlier detection. Taking the classical estimators of the arithmetic mean and sample covariance matrix often leads to useless results, because these estimators themselves are influenced by deviating data points. For this reason, robust counterparts need to be taken that downweight the influence of outliers on the resulting location and covariance estimation statistics. For this purpose several approaches have been proposed; see, e.g., Maronna et al. (2006). A popular choice is the MCD (minimum covariance determinant) estimator (Rousseeuw and Van Driessen, 1999), which also has the affine equivariance property. This means that for any nonsingular \((D-1) \times (D-1)\) matrix \( A \) and for any vector \( b \in \mathbb{R}^{D-1} \) the conditions

\[
T(A z_i + b, \ldots, A z_n + b) = A T(z_1, \ldots, z_n) + b,
\]

\[
C(A z_i + b, \ldots, A z_n + b) = A C(z_1, \ldots, z_n) A'
\]

are fulfilled. Thus, the estimators transform accordingly, as in the case of the arithmetic mean and sample covariance matrix, and it can easily be seen that the Mahalanobis distances remain unchanged under regular affine transformations; i.e.,

\[
MD(A z_i + b) = MD(z_i), \quad i = 1, \ldots, n.
\]

The identified outliers will thus be the same, independent of the choice of \( A \) and \( b \) for the affine transformation. This is important in the context of logratio transformations, because there exists an orthogonal relationship between different isometric logratio transformations. Affine equivariance thus guarantees that the identified outliers are invariant with the choice of such a transformation; see Filzmoser and Hron (2008). Under the assumption of multivariate normal distribution on the simplex, i.e., normal distribution of the orthonormal coordinates (Mateu-Figueras and Pawlowsky-Glahn, 2008), the (classical) squared Mahalanobis distances follow a \( \chi^2 \) distribution with \( D - 1 \) degrees of freedom; see, e.g., Maronna et al. (2006). This distribution might also be considered for the robust case, and a quantile, e.g., 0.975, can be used as a cutoff value separating regular observations from outliers. A more advanced approach to the cutoff value was introduced by Filzmoser et al. (2005). This method accounts for the actual numbers of observations and variables in the data set, and it tries to distinguish among extremes of the data distribution and outliers coming from a different distribution. This approach will be used for the graphical tools introduced in the next section.

2.2. Graphical representations

The procedure for outlier detection would not be comprehensive without displaying the results graphically. Because compositional data are multivariate observations by definition, it seems not to be possible to display their parts in univariate plots as is usual for standard multivariate observations. Nevertheless, the isometric logratio transformation makes it possible to display univariately relative information coming from all logratios to the 1st compositional part through the variables \( z_i^{(l)} \), \( l = 1, \ldots, D \) (Fišerová and Hron, in press). Although such plots will differ from plots displaying “raw” parts (in mg/kg or even in percentages), they represent the only meaningful way to visualize the relative information on single compositional parts. One should carefully decide which logratios will be covered up by the variable \( z_i^{(1)} \) because they can counteract and thus influence the overall statement on the behavior of the compositional part in the corresponding context. Then we can determine through values of \( z_i^{(l)} \) where the part \( x_l \) dominates in the corresponding logratios and where its influence is suppressed in the study area.
A popular graphical tool for visualizing patterns in the multivariate data structure is the biplot (Gabriel, 1971), which is based on a rank-2 approximation of the observations in terms of loadings and scores of a principal component analysis. The biplot was adapted to compositional data by Aitchison (1997), where the clr transformation was favored; however, to robustify the compositional biplot, an ilr transformation needs to be utilized (Filzmoser et al., 2009). The compositional biplot differs from the standard one in the interpretation of rays coming from loadings of the principal component analysis. While usually the main interest is devoted just to rays, and their length and in particular the angles between them that represent an approximation to the Pearson correlation coefficient, here we need to be careful because the clr space was used as the starting point for the principal component analysis. For this reason, the main interest in the compositional case is devoted to links (distances between rays) connected to single clr variables. Here the link between rays of the $i$th and $j$th clr variables (i.e., distances between the corresponding vertices) approximates the variance of the logratio $\ln(x_i/x_j)$. Consequently, if the vertices coincide, or nearly so, the ratio $x_i/x_j$ is constant, or nearly so. Additionally, from the linear relationship between $z_{ij}^l$ and $y_i$ it follows that the directions of the rays signal where the corresponding compositional parts dominate in the compositions (represented by scores on the PCA).

3. Tools for interpreting multivariate outliers

The tools discussed in this section are implemented and freely available in the R package mvoutlier; see Filzmoser and Gschwandtner (2011). Mainly, two functions are relevant for the user:

- `mvoutlier.CoDa()` requires an untransformed input data matrix with at least three compositional parts. Robust location and covariance estimations are derived using the adaptive approach of Filzmoser et al. (2005) (with sensible default values) for the ilr-transformed data. These are used for computing robust Mahalanobis distances, for grouping the data into regular observations and outliers, and for robust PCA. The latter loadings and scores are back-transformed to the clr space for the compositional biplot (Aitchison and Greenacre, 2002). Moreover, the univariate ilr variables $z_{ilj}^l$, for $l=1,\ldots,D$ are computed (see Eq. (1)), for univariate presentations. Finally, symbol colors and grayscale values are derived that can be optionally used in all preceding plots. The colors and gray levels should reflect the magnitude of the median element concentration of the observations; compare Filzmoser et al. (2005). This is done by computing for each observation the distances to the medians along the single ilr directions. The median of all distances determines the color (or grayscale): a high value, resulting in a red (or dark) symbol, means that most univariate parts have higher values than the average, and a low value, resulting in a blue (or light) symbol, refers to an observation with mainly low values. This characterization helps to interpret multivariate outliers. The output of this routine is an object of the class “mvoutlierCoDa,” and it can be visualized by the plot function below.

An example for simulated data with three parts (Fig. 1A) shows in more detail how the colors are determined. The original compositional data in Fig. 1A result in the data structure in Fig. 1B after an ilr transformation. The univariate scatterplots in Fig. 1C refer to the univariate ilr variables, where the color for the symbols is determined. The symbols with large “+,” which are in fact the multivariate outliers, are on average (median) far away from the origin (univariate median) and thus receive a red color. The large open circles are close to the origin in all three univariate presentations, and thus their color is blue or dark green.

- `plot.mvoutlierCoDa()` makes plots of the object resulting from the function `mvoutlier.CoDa()`. The available plots can be selected via the parameter “which,” and the options are as follows:
  - `which='biplot'` shows a compositional biplot (Aitchison and Greenacre, 2002) by using the robust PCA loadings and scores from the ilr-transformed data, and back-transformation to the clr space.
  - `which='map'` represents the symbols in the map at the geographical coordinates of the sample locations, and plots a background map, if available.
  - `which='uni'` plots all univariate ilr variables $z_{ilj}^l$, $l=1,\ldots,D$, as univariate scatterplots; i.e., the variables are shown in parallel vertical plots, and the horizontal position of the observations in each plot is random in order to make the symbols better distinguishable.
  - `which='parallel'` draws a parallel coordinate plot (Reimann et al., 2008), with the univariate ilr variables as parallel vertical axes, and the multivariate observations as a line combining the values of the axes.

The representation of the observations/symbols in the plots is controlled in the same way:

- `onlyout=TRUE` shows only the outlying observations; otherwise, if `onlyout=FALSE`, all observations are shown.
- `lmw=TRUE` shows all symbols (or lines for the parallel coordinate plot) in grayscale; otherwise the colors computed from the function `mvoutlier.CoDa()` are used.
- `symb=TRUE` represents the plot symbols according to their outlyingness (except for the parallel coordinate plot), as done in Filzmoser et al. (2005). The example in Fig. 1 illustrates the choice of plot symbols. The original compositional data in Fig. 1A are ilr-transformed, resulting in Fig. 1B. Here the robust squared Mahalanobis distances are computed and split by four values: the quantiles 0.25, 0.5, and 0.75 and the outlier cutoff mentioned in Section 2.1. By default, the symbols for the resulting five groups (in the above order) are large open circle, small open circle, point, small “+,” and large “+.” If `symb=FALSE`, the symbols are according to the definition of `obj.cex`; if this is not provided, a default symbol is used.
- `symbtxt=TRUE` presents the object number rather than symbols in the plots. For the parallel coordinate plot the numbers are shown in the left and right plot margins.

4. Examples

In this section we demonstrate the use of the outlier tools for two data sets from geochemistry.

4.1. Example 1: GEMAS

We consider a data set from the GEMAS project (Reimann et al., 2009). For illustration purposes we focus on the 473 observations that are available from Scandinavia, and we use the concentrations of the elements As, Au, Bi, Cu, Mo, Sb, and Sn. These elements can be considered indicative for a variety of mineral deposits that occur in the area. Suppose the data set is available in R as object `x`, a matrix (or data frame) with 473 rows and 7 columns. After the package is loaded with `library(mvoutlier)`, the procedure for multivariate...
outlier detection is applied with
> res <- mvoutlier.CoDa(x)

using all default parameters (see help(mvoutlier.CoDa)). The results are stored in the object res, which can be now used for plotting.

A compositional biplot is generated by
> plot(res, which = "biplot", onlyout = FALSE, symb = TRUE, symbtxt = FALSE)

and the result is shown in Fig. 2 (left). Here, all observations (not only the outliers) are plotted with special symbols: the color of the symbols corresponds to the size of the median element concentration, and the symbol itself corresponds to the outlying-ness (see above for details). If the focus is on the outliers, the command
> plot(res, which = "biplot", onlyout = TRUE, symb = TRUE, symbtxt = FALSE)

shows the same biplot projection, but only outliers are shown with an identification number; see Fig. 2 (right). Here the number is printed in the same color as the symbol in the previous plot.

The biplots in Fig. 2 explain about 60% of the total variability, and thus further principal components could be of interest for the inspection (they can be selected by the plot parameter choice).

---

Fig. 2. Compositional biplot for selected GEMAS data. Left: all data are shown by the special symbols; right: only outliers are shown by identification numbers using the symbol color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

---

Fig. 3. Univariate scatterplots for selected elements of the GEMAS data. Top: all data are shown by the special symbols; bottom: only outliers are shown by identification numbers using the symbol color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
The elements Sn, Bi, and Sb show a strong relationship (i.e., their ratios are nearly constant). The directions of the rays signal where observations with dominance of the corresponding compositional part are located. For example, Au is dominant in a relative sense parts of the survey area, revealing that good exploration targets. Observation “39” is very exceptional for its ratio of Au to the other elements, and this location would definitely be of interest to exploration geochemists. Observation “1” is rather exceptional for As, and number “2” has the lowest value for “ilr(Au)” and the highest for “ilr(Cu)”.

A final graphical representation is the parallel coordinate plot, generated by

```r
> plot(res, which = "parallel", onlyout = FALSE, symb = TRUE, symbtxt = FALSE)
> plot(res, which = "parallel", onlyout = TRUE, symb = TRUE, symbtxt = TRUE)
```

and shown in Fig. 5. The parameter `trans` makes it possible to change the transparency of the colors (default to 1 for nontransparent).

Compared to the univariate plots of Fig. 3, the parallel coordinate plots allow a better grasp of the multivariate information of the observations via the connecting lines. Fig. 5 (top) shows the more general data trends, while Fig. 5 (bottom) makes the details visible. Certain streams of lines are visible, such as for the outliers in the northern part of the survey area, revealing that these observations have similar characteristics. A comparison of the multivariate outlier map with the map of mineral deposits in Scandinavia (Eilu et al., 2008) shows the power of the approach even with the very low-density sampling used for the GEMAS project. Many of the most important mineralized areas are clearly indicated by multivariate outliers.

### 4.2. Example 2: Kola

The Kola data set has been studied in many publications; it is also explained in detail and used in Reimann et al. (2008). The data come from the Kola Peninsula in northern Europe, and the concentrations of more than 50 chemical elements have been measured in four soil layers. The data set is available in the R package `mvoutlier`, and an updated version in the package `mvoutliert`.

The multivariate outliers are not necessarily at the extremes of the univariate variables, but can be found in the whole range. By definition of the colors, there are mainly blue symbols in the central parts of the ilr variables, and red symbols in the extremes.

One can now go into more detail for the interpretation. The outliers in the northern part of the survey area have high values at “ilr(Au)”; i.e., Au is a dominating compositional part. Observation “39” is very exceptional for its ratio of Au to the other elements, and this location would definitely be of interest to exploration geochemists. Observation “1” is rather exceptional for As, and number “2” has the lowest value for “ilr(Au)” and the highest for “ilr(Cu)”.

...
StatDA. Here we consider the concentration of the elements As, Cd, Co, Cu, Mg, Pb, and Zn in the O-horizon (organic surface soil). Co and Cu are typical elements emitted from Ni smelters. As, Cd, and Pb are elements that are emitted in minor amounts. Mg and Zn are not in this emission spectrum but they may be influenced by other processes. The concentrations of Mg, for instance, are affected by the steady input of marine aerosols from the coast; see also Filzmoser et al. (2005).

After `mvoutlier.CoDa()` is applied to the selected data, an output object, say `res1`, is created, and we start the visual inspection with the parallel coordinate plot

```r
plot(res1, which = "parallel", bw = TRUE,
    onlyout = FALSE, symb = FALSE,
    symbtxt = FALSE)
```

which produces Fig. 6. Here we show all data by the lines, and the graylevel corresponds to regular observations (light) and outliers (dark).

The multivariate structure of the outliers is quite different from that of the other observations. It is also possible to see common streams corresponding to groups of observations with similar data structure.

For further analysis we focus exclusively on the outliers. The R commands for the following plots are in analogy to Example 1, and are thus not shown. Fig. 7 presents the outliers with the special symbols in a biplot and in the map. The biplot of the first two principal components now explains more than 70% of the total variability, but some elements are poorly represented (short rays). Co and Cu are closely related (in fact they are overplotted in the figure), and a number of outliers are dominated by these two elements. These outliers are observed at the locations of the Ni-smelters (Nikel/Zapolyarnij, Monchegorsk). Some further outliers near the coast are dominated by Mg. The remaining outliers are difficult to interpret with these plots, but they can be further inspected with the univariate scatterplot in Fig. 8.
et al., 2006). Such tools have also been developed in the context of robust statistics (Maronna, 2006).

5. Conclusions

Multivariate outliers are often the most interesting data points because they show atypical phenomena. Several methods have been proposed for the identification of multivariate outliers, making use of the technology of robust statistics (Maronna et al., 2006). Such tools have also been developed in the context of compositional data (Filzmoser and Hron, 2008). As a result, the data investigator gets the information on the samples that are potential multivariate outliers, but not the information because of which these samples are identified as being atypical. The tools proposed in this paper help to better understand the multivariate compositions of these outliers. Several exploratory tools have been developed for this purpose: representations in maps, in a compositional biplot, in univariate scatterplots, and in parallel coordinate plots. In all plots, the same special colors and symbols can be selected, referring to the relative position of the outliers in the multivariate data cloud and thus supporting interpretation of these observations. Since the considered data are compositional data, a relation to the single compositional parts can be established only by special ilr transformations. The new ilr variables are in fact related to the variables resulting from a clr transformation. Analysis and interpretations in the clr space (see, e.g., Grunsky, 2010) are thus useful in this context.

The developed tools are freely available in the R package mvoutlier. The function mvoutlier.CoDa() computes the multivariate outliers, and it prepares the information for the symbols and colors. The resulting object can then be used for the plot function. The argument which makes it possible to select among the four types of graphical presentations. All other arguments are consistent for the presentations, which makes it possible to see the same symbol and color choices in different views, revealing the structure of the multivariate outliers. The help page for the plot function contains several examples of how to generate the different plots. The examples can easily be executed by example(plot.mvoutlierCoDa).

Acknowledgments

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

References
