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. Tel +43 1 58801 10733, FAX +43 1 58801 10799*
[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*

Abstract

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 that 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 the outliers by representing the multivariate information in biplots, maps, and univariate scatter plots.

Key words: compositional data, log-ratio transformations, outlier detection, compositional biplot

2010 MSC: 62F35, 62H25, 62H30

URL: P.Filzmoser@tuwien.ac.at (Peter Filzmoser), hronk@seznam.cz (Karel Hron), Clemens.Reimann@ngu.no (Clemens Reimann)

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 between the parts (Aitchison, 1986). Typically these observations are expressed as data with a constant sum constraint like 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 non-normal distribution of the samples (for example, of chemical elements in a rock) and thus recommend to apply a logarithmic transformation in order to achieve normality (Reimann et al., 2008) of the data set. Depending on the transformation chosen, normality can often be reached so that the data pass a statistical test. However, the reality is more severe. The original data follow in fact another geometry (called usually the Aitchison geometry, see, e.g., Egozcue and Pawlowsky-Glahn (2006) for details) on the sample space of compositions, the simplex, defined for a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)'$ as

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

The positive constant κ stands for 1 in case of proportions, 100 (percentages) or 10^6 (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 log-ratio transformations from the simplex S^D to the Euclidean real space was proposed. Only following such a transformation the use of the standard statistical methods is possible. The three main types of this family of transformations are: the additive logratio (alr), the centered logratio (clr), and the isometric logratio (ilr) transformation. 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. clr (Aitchison, 1986) results in data singularity, which is in conflict with the usual tools for outlier detection. 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 represents the sequential binary partition procedure (Egozcue and Pawlowsky-Glahn, 2005) that enables an interpretation of the orthonormal coordinates in the sense of balances between groups of compositional parts. Additionally, each ilr variable explains all the log-ratios, i.e. terms of type $\ln(x_i/x_j)$, $i, j = 1, \ldots, D$, between parts of the corresponding groups (Fišerová and Hron, 2010); conversely, each log-ratio 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 con-

tained in (log-)ratios of compositional parts. Although the sequential binary partition can also be made-to-measure for the concrete geochemical problem (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). Explicitely, we obtain $(D-1)$-dimensional real vectors $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$, $l = 1, \ldots, D$,

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \; i = 1, \ldots, D-1, \tag{1}$$

where $(x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$ stands for such a permutation of the parts $(x_1, \ldots, x_D)$ that always the l-th compositional part fills the first position, $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$. In such a configuration, the first ilr variable $z_1^{(l)}$ explains all the relative information (log-ratios) about the original compositional part x_l, the coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ then explain the remaining log-ratios in the composition (Fišerová and Hron, 2010). Note that the only important position is that of $x_1^{(l)}$ (because it can be fully explained by $z_1^{(l)}$), 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 $z_1^{(l)}$ *is* the original compositional part x_l, but it *explains all the information concerning* x_l, thus, it *stands for* x_l.

An interesting consequence follows for the known clr transformation from $\mathcal{S}^D$ to $\mathbf{R}^D$, resulting in

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

It is easy to see that there exists a linear relationship between y_l and $z_1^{(l)}$, namely $y_l = \sqrt{\frac{D}{D-1}} z_1^{(l)}$. Thus, up to a constant, the single clr variables have

the same interpretation as the corresponding ilr coordinates, they explain all log-ratios concerning the l-th compositional part. However, as a consequence, some of the log-ratios 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 $y_1 + \ldots + y_D = 0$ of clr variables, what makes, e.g., the use of robust multivariate statistical methods not possible. On the other hand, the clr transformation is a corner stone of the compositional biplot (Aitchison and Greenacre, 2002) that will be employed further in the paper.

From the above mentioned properties of log-ratio transformations it is visible that the ordered D-tuple of the ilr coordinates, $z_1^{(l)}$, $l = 1, \ldots, D$, can be obtained from clr-transformed data as $\sqrt{\frac{D-1}{D}}\mathbf{y}$. Nevertheless, note that it would be not meaningful to interpret the relations between the clr variables or even between the variables $z_1^{(l)}$ using their correlation structure; here the subcompositional incoherence (it means that the results of the statistical modeling might be incompatible if only a subset of the parts is 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 $z_1^{(l)}$, $l = 1, \ldots, D$, variables, however, here as a natural consequence of the fact that the information available in $\mathbf{x}$ was reduced just to a subcomposition.

In contrast to univariate outliers, where their 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 the view of the Aitchison geometry, see, e.g., Hron et al. (2010); Filzmoser and Hron (2011). 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 Figure 1(B) to the Aitchison geometry in Figure 1(A)). 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 $\mathbf{x} = (x_1, \ldots, x_D)'$ is defined as a member of the corresponding equivalence class of $\mathbf{x}$,

$$\underline{\mathbf{x}} = \{c\mathbf{x},\ c \in \mathbf{R}^+\}.$$

Thus, two compositions which are elements of the same equivalence class $\underline{\mathbf{x}}$ (we call them also compositionally equivalent, see Egozcue and Pawlowsky-Glahn, 2006) contain the same information and have zero Aitchison distance (Aitchison, 1986). 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 the following 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. The final Section 5 concludes.

2. Methods for multivariate outlier detection and graphical representation

As well as for the other multivariate methods applied to compositional data, it is important to use an appropriate data transformation first. Both 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) allows 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 $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of $(D-1)$-dimensional observations (resulting for instance from an ilr transformation of the corresponding compositional sample) as

$$\mathrm{MD}(\mathbf{z}_i) = \left[(\mathbf{z}_i - T)' C^{-1} (\mathbf{z}_i - T)\right]^{1/2}, \; i = 1, \ldots, n. \tag{2}$$

Here, $T = T(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and $C = C(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ are location and covariance estimators, respectively. The choice of the estimators is crucial for the quality of multivariate outlier detection. Taking the classical estimators 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 were proposed, see, e.g., Maronna et al. (2006). A popular choice is the MCD (Minimum Covariance Determinant)

estimator (Rousseeuw and Van Driessen, 1999) which has also the affine equivariance property. This means that for any nonsingular $(D-1)\times(D-1)$ matrix $\mathbf{A}$ and for any vector $\mathbf{b} \in \mathbf{R}^{D-1}$ the conditions

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

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

$$\mathrm{MD}(\mathbf{A}\mathbf{z}_i + \boldsymbol{b}) = \mathrm{MD}(\mathbf{z}_i),\ i = 1, \ldots, n. \tag{3}$$

The identified outliers will thus be the same, independent of the choice of $\mathbf{A}$ and $\mathbf{b}$ for the affine transformation. This is important in the context of log-ratio transformations, because there exists an orthogonal relationship between different isometric log-ratio 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-square 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 cut-off value separating regular observations from outliers. A more advanced approach for the cut-off 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 of outlier detection would not be comprehensive without displaying the results graphically. Because compositional data are multivariate observations by definition, it seems to be not possible to display their parts in univariate plots as it is usual for standard multivariate observations. Nevertheless, the isometric log-ratio transformation enables to display univariately relative information coming from all log-ratios to the l-th compositional part through the variables $z_1^{(l)}$, $l = 1, \ldots, D$ (Fišerová and Hron, 2010). Although such plots will differ from 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 log-ratios will be covered up by the variable $z_1^{(l)}$, 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 $z_1^{(l)}$, where the part x_l dominates in the corresponding log-ratios and where its influence is suppressed in the study area.

A popular graphical tool to visualize patterns in the multivariate data structure is the biplot (Gabriel, 1971), which is based on a rank-two 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, their length and in particular angles between them that represent an approximation of 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. distance between the corresponding vertices) approximates the variance of the log-ratio $\ln(x_i/x_j)$. 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 $z_1^{(l)}$ and y_l it follows that the directions of the rays signalize where the corresponding compositional parts dominate in the compositions (represented by scores of the PCA).

3. Tools for interpreting multivariate outliers

The tools discussed in the following are implemented and freely available in the R package `mvoutlier`, see Filzmoser and Gschwandtner (2011). Mainly, there are two functions that 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 $z_1^{(l)}$, for $l = 1, \ldots, D$ are computed, see Equation (1), for univariate presentations. Finally, symbol colors and gray scale values are derived that can be optionally used in all preceeding 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 along the single ilr variables the distances to the medians. The median of all distances determines the color (or grey scale): 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 (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 class "mvoutlierCoDa", and it can be visualized by the plot function below.

An example for simulated data with three parts (Figure 1) shows in more detail how the colors are determined. The original compositional data in Figure 1(A) result in the data structure Figure 1(B) after an ilr transformation. The univariate scatter plots in Figure 1(C) 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 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:
  - `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 $z_1^{(l)}$, $l = 1, \ldots, D$ as univariate scatter plots, 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, see, e.g.), with the univariate ilr variables as parallel vertical axes, and the multivariate observations as line combining the values of the axes.

FIGURE 1

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.
- `bw=TRUE` shows all symbols (or lines for the parallel coordinate plot) in gray scale; 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 Figure 1 illustrates the choice of the plot symbols. The original compositional data in Figure 1(A) are ilr-transformed, resulting in Figure 1(B). Here the robust squared Mahalanobis distances are computed and split by four values: the quantiles 0.25, 0.5, 0.75, and the outlier cut-off 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 "+", 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 as being indicative for the 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 loading the package 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 Figure 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 outlyingness (see above for details). If the focus is on the outliers, the command

```
> plot(res,which="biplot",onlyout=TRUE,symb=TRUE,symbtxt=TRUE)
```

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

FIGURE 2

The biplots in Figure 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`). The elements Sn, Bi and Sb show a strong relation (i.e. their ratios are nearly constant). The directions of the rays signalize where observations with dominance of the corresponding compositional part are located. For example, Au is dominant in relative sense for many of the outliers plotted in red. Because of the rather low explained variability we will provide a more rigorous interpretation for the observations in the univariate scatter plot, see Figure 4.

The observations, and in particular the multivariate outliers, can be presented in a map:

```
> plot(res,which="map",coord=XY,map=coo,onlyout=FALSE,symb=TRUE,
      symbtxt=FALSE)
```

The object `XY` contains the geographical coordinates of the observations, and `coo` includes the information of the background map in form of a matrix with two columns. An example is included in the help page of the function. The result in Figure 3 (left) shows the locations of the observations in the survey area, and the symbols are in the same style as in Figure 2 (left). Focusing only on the outliers is possible with

```
> plot(res,which="map",coord=XY,map=coo,onlyout=TRUE,symb=TRUE,
      symbtxt=TRUE)
```

and the resulting representation in the map is in Figure 3 (right), using the same numbers and colors as in Figure 2 (right).

FIGURE 3

The multivariate outliers are concentrated in certain areas. Especially in the northern part of the survey area several outliers are visible (red large "+"). According to the biplot in Figure 2 (right), these outliers are dominated by the element Au. Not surprisingly they mark areas where known gold deposits occur. On the other hand, also single outliers are identified, like the observations numbered by "1" and "2", which seem to be rather unusual when compared to the observations in the local neighborhood.

A more detailed interpretation of the multivariate outliers with respect to the single elements is possible in the plot of the univariate ilr variables.

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

generates Figure 4 (top), and

```
> plot(res,which="uni",onlyout=TRUE,symb=TRUE,symbtxt=TRUE)
```

results in Figure 4 (bottom), where the symbols and colors are the same as in the previous plots.

FIGURE 4

The multivariate outliers are not necessarily in the extremes of the univariate variables, but they 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 details 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)".

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

```
> plot(res,which="parallel",onlyout=FALSE,symb=TRUE,
      symbtxt=FALSE,transp=0.3)
```

or

```
> plot(res,which="parallel",onlyout=TRUE,symb=TRUE,
      symbtxt=TRUE,transp=0.5)
```

and shown in Figure 5. The parameter `trans` allows to change the transparency of the colors (default to 1 for non-transparent).

FIGURE 5

Compared to the univariate plots of Figure 4, the parallel coordinate plots allow to better grasp the multivariate information of the observations via the connecting lines. Figure 5 (top) shows the more general data trends, while

Figure 5 (bottom) makes the details visible. Certain streams of lines are visible, like 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 concentration of more than 50 chemical elements has been measured in four soil layers. The data set is available in the R package `mvoutlier`, and an updated version in the package `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 the 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 applying `mvoutlier.CoDa()` to the selected data, an output object, say `res1`, is created, and we start the visual inspection with the parallel coordinate plot:

```
> plot(res1,which="parallel",bw=TRUE,onlyout=FALSE,symb=FALSE,
```

```
        symbtxt=FALSE)
```

produces Figure 6. Here we show all data by the lines, and the grey level corresponds to regular observations (light) and outliers (dark).

FIGURE 6

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

For the further analysis we exclusively focus on the outliers. The R commands for the following plots are in analogy to Example 1, and are thus not shown. Figure 7 presents the outliers with the special symbols in a biplot and in the map. The biplot of the first two principal components explains now more than 70% of the total variability, but some elements are poorly represented (short rays). Co and Cu are highly 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 scatter plot in Figure 8.

FIGURE 7

The univariate scatter plot shown in Figure 8 confirms the above findings: The outlier group located at the Ni-smelters indeed have high values for

"ilr(Co)" and "ilr(Cu)", and Magnesium dominates the outliers on the coast. Outlier "20" is located at the sea harbor Murmansk, its ratio of As to the other elements is very low, for Zn it is rather high, and for the remaining elements the ratio is more in the average. Obviously, this behavior makes this observation special and thus outlying. Very exceptional is observation "33", with low ratios for As, Co and Cu, high ratio for Mg, but very high ratio for Pb, which is rather surprising.

FIGURE 8

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). Also in the context of compositional data such tools have been developed (Filzmoser and Hron, 2008). As a result, the data investigator gets the information of the samples being potential multivariate outliers, but not the information, for which reason 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 scatter plots, 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 an 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` allows 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 to the plot function contains several examples how to generate the different plots. The examples can be easily 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

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London, 416pp.

Aitchison, J., 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn, V. (ed.), Proceedings of IAMG '97: The third annual conference of the International Association for Mathematical Geology. Barcelona, CIMNE, 3-35.

Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. Applied Statistics, 51 (4), 375-392.

Buccianti, A., Egozcue, J.J., Pawlowsky-Glahn, V., 2008. Another look at the chemical relationships in the dissolved phase of complex river systems. Mathematical Geosciences, 40 (5), 475-488.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, 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.

Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264, 145-160.

Eilu, P., Hallberg, A., Berman, T., Feoktistov, V., Korsakova, M., Krasotkin, S., Kitosmanen, E., Lampio, E., Litvinenko, V. Nurmi, P.A., Often, M., Philippov, N., Sandstad, J.S. Stromov, V., Tontii, M. (comp.), 2008. Metallic mineral deposit map of the Fennoscandian shield, 1:2.000.000.

Geological Survey of Finland, Geological Survey of Sweden, The Federal Agency of Use of Mineral Resources of the Ministry of Natural Resources of the Russian Federation. ISBN 978-952-217-008-8.

Filzmoser, P., Garrett, R.G., Reimann, C., 2005. Multivariate outlier detection in exploration geochemistry. Computers & Geosciences, 31 (5), 579-587.

Filzmoser, P., Gschwandtner, M., 2011. mvoutlier: Multivariate outlier detection based on robust methods. *Manual and package*, version 1.9.1. *http://cran.r-project.org/package=mvoutlier*

Filzmoser, P., Hron, K., 2008. Outlier detection for compositional data using robust methods. Mathematical Geosciences, 40 (3), 233–248.

Filzmoser, P., Hron, K., Reimann, C., 2009. Principal component analysis for compositional data with outliers. Environmetrics, 20 (6), 621-632.

Filzmoser, P., Hron, K., Reimann, C., 2010. The bivariate statistical analysis of environmental (compositional) data. Science of the Total Environment, 408 (19), 4230-4238.

Filzmoser, P., Hron, K., 2011. Robust statistical analysis of compositional data. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. Wiley, New York, in press.

Fišerová, E., Hron, K., 2010. On interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences, accepted for publication.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58 (3), 453-467.

Grunsky, E.. 2010. The interpretation of geochemocal survey data. Computers and Geosciences, 10 (1), 27-74.

Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis, 54 (12), 3095-3107.

Maronna, R., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley, New York, 436pp.

Mateu-Figueras, G., Pawlowsky-Glahn, V., 2008. A critical approach to probability laws in geochemistry. Mathematical Geosciences, 40 (5), 489-502.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. Wiley, Chichester, 362pp.

Reimann, C., Demetriades, A., Eggen, O.A., Filzmoser, P., and the EuroGeoSurveys Geochemistry expert group, 2009. The EuroGeoSurveys geochemical mapping of agricultural and grazing land soils project (GEMAS) - Evaluation of quality control results of aqua regia extraction analysis. Technical Report 2009.049, Geological Survey of Norway (NGU), Trondheim, Norway.

491 Rousseeuw, P., Van Driessen, K., 1999. A fast algorithm for the minimum
492 covariance determinant estimator. Technometrics, 41 (3), 212–223.

Figure captions

Figure 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) are univariate scatter plots 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 median of the three resulting distances determines the color for each observation.

Figure 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.

Figure 3: Maps 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.

Figure 4: Univariate scatter plots 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.

Figure 5: Parallel coordinate plots for selected elements of the GEMAS data. Top: all data are shown by lines with the special colors; bottom: only outliers are shown, and identification numbers using the symbol colors are in the margins.

Figure 6: Parallel coordinate plots for the selected Kola data set. Regular observations are in light gray, multivariate outliers in dark gray.

Figure 7: Biplot (left) and map (right) for the selected Kola data set. Only the outliers are shown, and special symbols are used.

Figure 8: Univariate scatter plot for the selected Kola data set. Only the outliers are shown, and special symbols are used.

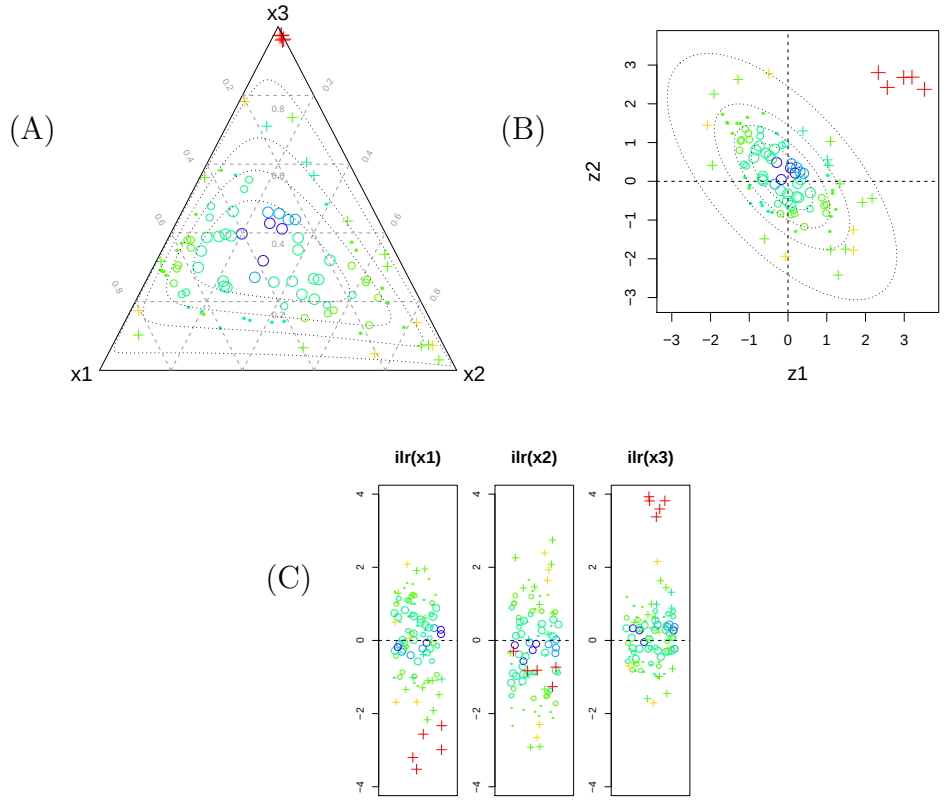

Figure 1: Illustration of the determination of plot symbols and colors: (A) shows simulated three-part compositional data, (B) is the representation in the ilr space, and (C) are univariate scatter plots of the single univariate ilr variables. The symbols are determined by certain quantiles of the robust 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 median of the three resulting distances determines the color for each observation.

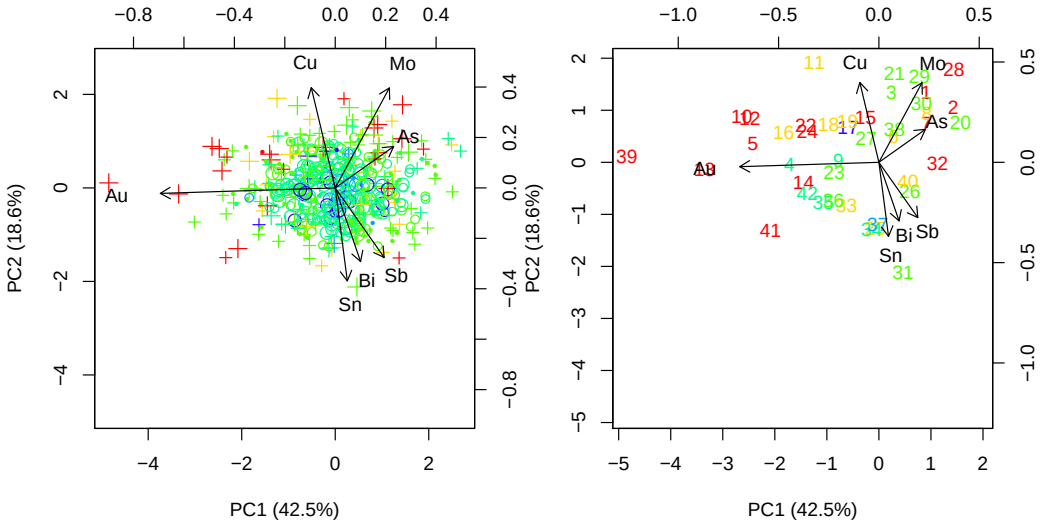

Figure 2: Compositional biplot for selected GEMAS data. Left: all data are shown by the special symbols; right: only outliers are shown by identification numbers using the symbol color.

521

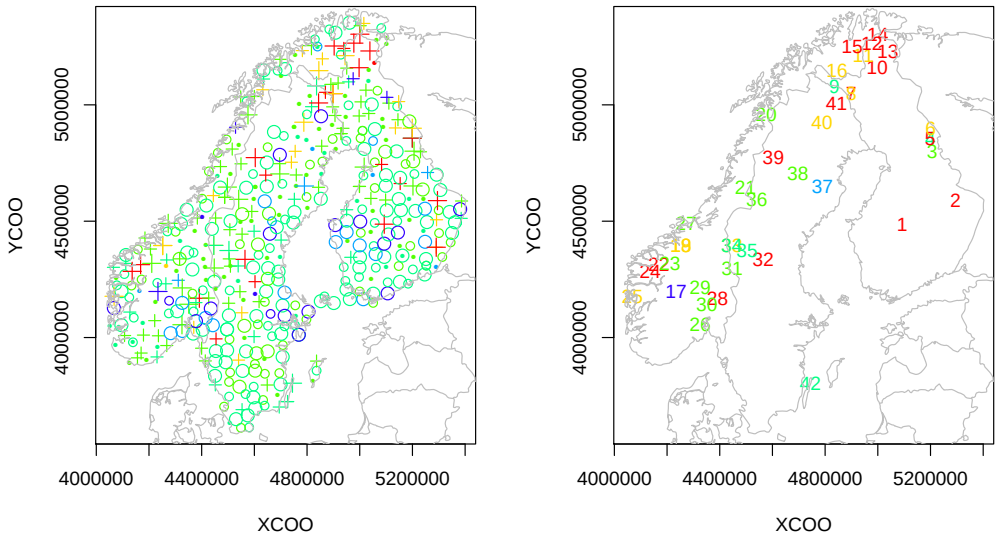

Figure 3: Maps for selected GEMAS data. Left: all data are shown by the special symbols; right: only outliers are shown by identification numbers using the symbol color.

522

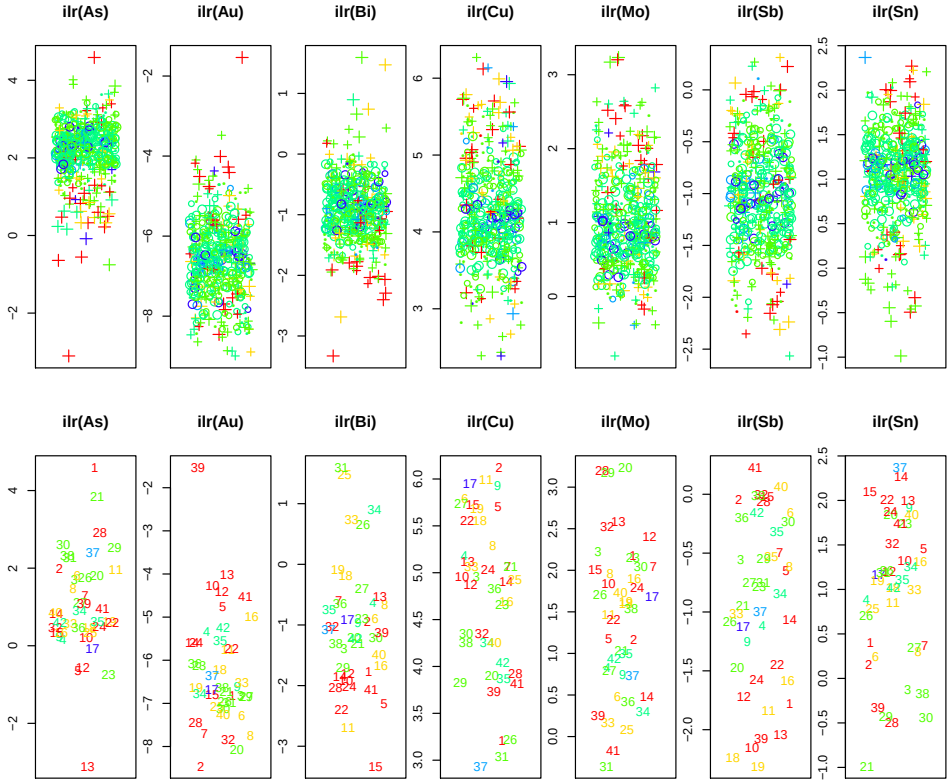

Figure 4: Univariate scatter plots for selected elements of the GEMAS data. Top: all data are shown by the special symbols; bottom: only outliers are shown by identification numbers using the symbol color.

523

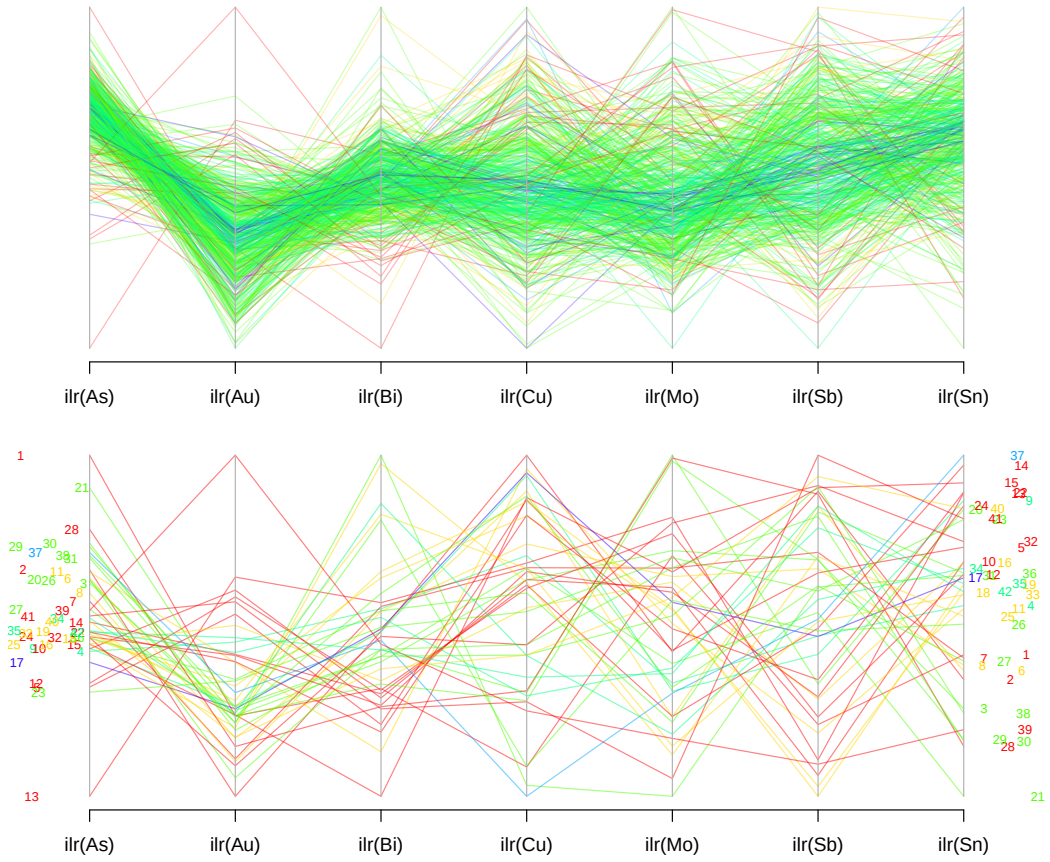

Figure 5: Parallel coordinate plots for selected elements of the GEMAS data. Top: all data are shown by lines with the special colors; bottom: only outliers are shown, and identification numbers using the symbol colors are in the margins.

524

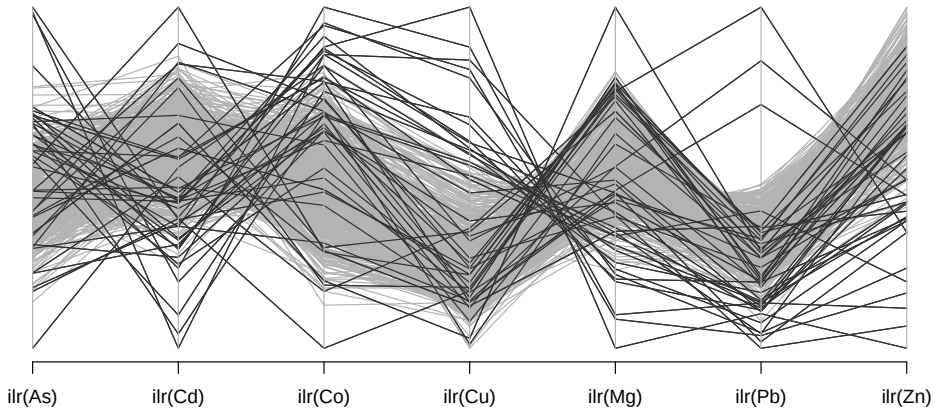

Figure 6: Parallel coordinate plots for the selected Kola data set. Regular observations are in light gray, multivariate outliers in dark gray.

525

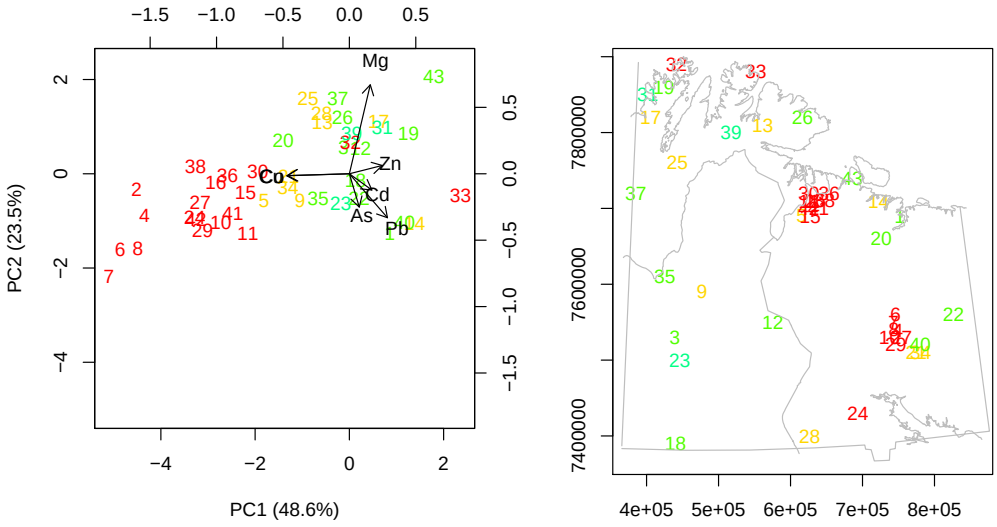

Figure 7: Biplot (left) and map (right) for the selected Kola data set. Only the outliers are shown, and special symbols are used.

526

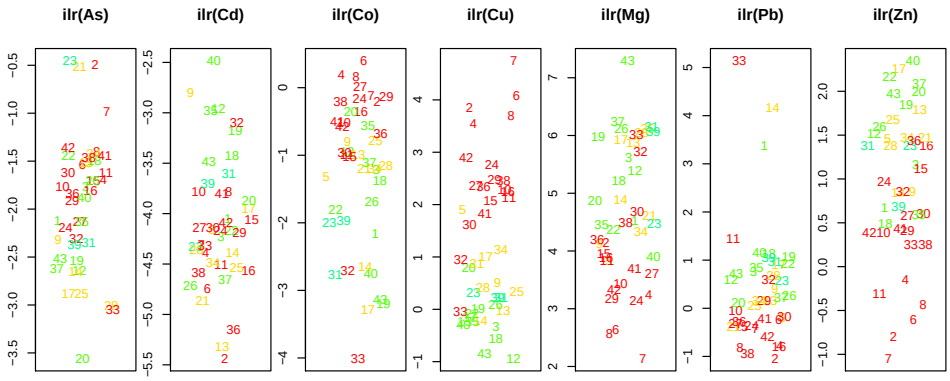

Figure 8: Univariate scatter plot for the selected Kola data set. Only the outliers are shown, and special symbols are used.

527