MULTIVARIATE OUTLIER DETECTION IN EXPLORATION GEOCHEMISTRY

Peter Filzmoser[1,*], Robert G. Garrett[2] & Clemens Reimann[3]

[1]Institute of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Wien, Austria. E-mail: P.Filzmoser@tuwien.ac.at, Tel.: +43 1 58801 10733, Fax +43 1 58801 10799

[2]Geological Survey of Canada, Natural Resources Canada, 601 Booth Street, Ottawa, Ontario, Canada, K1A 0E8. E-mail: garrett@gsc.NRCan.gc.ca

[3]Geological Survey of Norway, N-7491 Trondheim, Norway. E-mail: Clemens.Reimann@ngu.no

[*]Corresponding author

ABSTRACT

A new method for multivariate outlier detection able to distinguish between extreme values of a normal distribution and values originating from a different distribution (outliers) is presented. To facilitate visualising multivariate outliers spatially on a map, the multivariate outlier plot, is introduced. In this plot different symbols refer to a distance measure from the centre of the distribution, taking into account the shape of the distribution, and different colours are used to signify the magnitude of the values for each variable. The method is illustrated using a real geochemical data set from far-northern Europe. It is demonstrated that important processes such as the input of metals from contamination sources and the contribution of sea-salts via marine aerosols to the soil can be identified and separated.

KEYWORDS: Multivariate outliers, Robust statistics, Exploration geochemistry, Background.

1. INTRODUCTION

The detection of data outliers and unusual data structures is one of the main tasks in the statistical analysis of geochemical data. Traditionally, despite the fact that geochemistry data sets are almost always multivariate, outliers are most frequently sought for each single variable in a given data set (Reimann et al., 2005). The search for outliers is usually based on location and spread of the data. The higher (lower) the

analytical result of a sample, the greater is the distance of the observation from the central location of all observations; outliers thus, typically, have large distances. The definition of an outlier limit or threshold, dividing background data from outliers, has found much attention in the geochemical literature and to date no universally applicable method of identifying outliers has been proposed (see discussion in Reimann et al., 2005). In this context, background is defined by the properties, location and spread, of geochemical samples that represent the natural variation of the material being studied in a specific area that are uninfluenced by extraneous and exotic processes such as those related to rare rock types, mineral deposit forming processes, or anthropogenic contamination. In geochemistry, outliers are generally observations resulting from a secondary process and not extreme values from the background distribution. Samples where the analytical values are derived from a secondary process – be it mineralisation or contamination – do not need to be especially high (or low) in relation to all values of a variable in a data set, and thus attempts to identify these samples with classical univariate methods commonly fail. However, this problem often may be overcome by utilising the multivariate nature of most geochemical data sets.

In the multivariate case not only the distance of an observation from the centroid of the data has to be considered but also the shape of the data. To illustrate this, 2 variables with normal distributions having a defined correlation (Figure 1) are simulated. The estimated central location of each variable is indicated by dashed lines (their intersection marks the multivariate centre or centroid of the data).

In the absence of a prior threshold (Rose et al., 1979) a common practice of geochemists is to identify some fraction, often 2%, of the data at the upper and lower extremes for further investigation. Today this is achieved by direct estimation of the percentiles and visual (EDA) inspection of the data. In previous time when computers were not widely available an approximation of the 97.5[th] percentile was obtained by estimating the mean and standard deviation (sdev) for each variate and computing the value of mean $\pm 2 \cdot$sdev. The 2% limits are indicated by dotted lines on Figure 1. If candidates for outliers are defined to be observations falling in the extreme 2% fractions of the univariate data for each variable, the rectangle visualised with bold

dots separates potential outliers from non-outliers. This procedure ignores the elliptical shape of the bivariate data and therefore it is not effective.

The shape and size of multivariate data are quantified by the covariance matrix. A well-known distance measure which takes into account the covariance matrix is the Mahalanobis distance. For a p-dimensional multivariate sample $x_1,\ldots,x_n$ the Mahalanobis distance is defined as:

$$\text{MD}_i := \left((x_i - t)^{\text{T}} C^{-1} (x_i - t) \right)^{1/2} \quad \text{for} \quad i = 1,\ldots,n, \qquad (1)$$

where t is the estimated multivariate location and C the estimated covariance matrix. Usually, t is the multivariate arithmetic mean, the centroid, and C is the sample covariance matrix. For multivariate normally distributed data the values MD_i^2 are approximately chi-square distributed with p degrees of freedom (χ_p^2). By setting the (squared) Mahalanobis distance equal to a certain constant, i.e. to a certain quantile of χ_p^2, it is possible to define ellipsoids having the same Mahalanobis distance from the centroid (e.g, Gnanadesikan, 1977).

Figure 1 illustrates this for the bivariate normally distributed data. The ellipses correspond to the quantiles 0.25, 0.50, 0.75 and 0.98 of χ_2^2. Points lying on an ellipse thus have the same distance from the centroid. This distance measure takes the shape of the data cloud into account and has potential for more reliably identifying extreme values.

Multivariate outliers can now simply be defined as observations having a large (squared) Mahalanobis distance. As noted above for the univariate case, when no prior threshold is available a certain proportion of the data or quantile of the normal distribution is selected for identifying extreme samples for further study. Similarly, in the multivariate case a quantile of the chi-squared distribution (e.g., the 98% quantile $\chi_{p;0.98}^2$) could be considered for this purpose. However, this approach has several shortcomings that will be investigated in this paper. The Mahalanobis distances need to be estimated by a robust procedure in order to provide reliable measures for the recognition of outliers. In the geochemical context what is required is a reliable estimate of the statistical properties of natural background. Using robust

estimates that remove (trim) or downweight extreme values in a population is an effective, if conservative, solution. It is conservative to the extent that if there are in fact no outliers the only consequence is that the true variability (variance-covariance) of the data will be underestimated. Furthermore, by selecting a fixed quantile for outlier identification there is no adjustment for different sample sizes. To address this situation an adaptive outlier identification method has been developed. Finally, the *multivariate outlier plot* is introduced as a helpful tool for the interpretation of multivariate data.

2. THE ROBUST DISTANCE

The Mahalanobis distance is very sensitive to the presence of outliers (Rousseeuw and Van Zomeren, 1990). Single extreme observations, or groups of observations, departing from the main data structure can have a severe influence on this distance measure. This is somewhat obscure because the Mahalanobis distance should be able to detect outliers, but the same outliers can heavily affect the Mahalanobis distance. The reason is the sensitivity of arithmetic mean and sample covariance matrix to outliers (Hampel et al., 1986). A solution to this problem is well-known in robust statistics: t and C in equation (1) have to be estimated in a robust manner, where the expression 'robust' means resistance against the influence of outlying observations. Many robust estimators for location and covariance have been introduced in the literature, for a review see Maronna and Yohai (1998). The minimum covariance determinant (MCD) estimator (Rousseeuw, 1985) is probably most frequently used in practice, partly because it is a computationally fast algorithm (Rousseeuw and Van Driessen, 1999).

The MCD estimator is determined by that subset of observations of size h which minimises the determinant of the sample covariance matrix, computed from only these h points. The location estimator is the average of these h points, whereas the scatter estimator is proportional to their covariance matrix. As a compromise between robustness and efficiency, a value of $h \approx 0.75\ n$ (n is the sample size) will be employed in this study.

The choice of h also determines the robustness of the estimator. The breakdown value of the MCD estimator is approximately $(n-h)/n$, with $h \approx 0.75\,n$ the breakdown is approximately 25%. The breakdown value is the fraction of outliers that when exceeded will lead to completely biased estimates (Hampel et al., 1986).

Using robust estimators of location and scatter in the formula for the Mahalanobis distance (1) leads to so-called *robust distances* (RD). Rousseeuw and Van Zomeren (1990) used these RDs for multivariate outlier detection. If the squared RD for an observation is larger than, say, $\chi^2_{2;0.98}$, it can be declared a candidate outlier.

This procedure is illustrated using real data from the Kola project (Reimann et al., 1998). Figure 2 shows the plot of Be and Sr determined in C-horizon soils. Using the arithmetic mean and the sample covariance matrix in equation (1) it is possible to construct the ellipse corresponding to the squared Mahalanobis distance equal to $\chi^2_{2;0.98}$. This ellipse (often called a *tolerance ellipse*) is visualised as a dotted line in Figure 2. It identifies the extreme members of the bivariate population and its shape reflects the structure of the covariance matrix. By computing the RDs with the MCD estimator another tolerance ellipse (solid line in Figure 2) can be constructed using the same quantile, $\chi^2_{2;0.98}$. It is clearly apparent that many more points in the upper right of Figure 2 are identified as candidate outliers. These outliers cause the elongated orientation and shape of the dotted ellipse through their influence on the classical non-robust computation. This influence is also reflected in the resulting correlation coefficients. Whereas the Pearson correlation based on the classical estimates is 0.66, the robust correlation based on the MCD estimator is only 0.18. The next step would be an appropriate visualisation of the outliers in a map in order to support the geochemical interpretation of the observations. This will be demonstrated later for other examples. The high correlation of Be and Sr in Figure 2 is due to a few samples of soil developed on alkaline rocks that display unusually high concentrations of both these elements. The high non-robust correlation coefficient is thus an inappropriate estimate for the majority of the data as it is unduly influenced by true outliers (due to completely different geology).

3. MULTIVARIATE OUTLIERS OR EXTREMES?

In the univariate case, Reimann et al. (2005) pointed out the difference between extremes of a distribution and true outliers. Outliers are thought to be observations coming from one or more *different* distributions, and extremes are values that are far away from the centre but which belong to the *same* distribution. In an exploratory univariate data analysis it is convenient to start with simply identifying all extreme observations as extreme. It is an important aim of data interpretation to identify the different geochemical processes that influence the data. Only in doing so can the true outliers be identified and differentiated from extreme members of the one or more background populations in the data. This distinction should also be made in the multivariate case.

In the previous section the assumption of multivariate normality was implicitly used because this led to chi-square distributed Mahalanobis distances. Also for the RD this assumption was used, at least for the majority of data (depending on the choice of h for the MCD estimator). Defining outliers by using a fixed threshold value (e.g., $\chi^2_{p;0.98}$) is rather subjective because:

1) If the data should indeed come from a single multivariate normal distribution, the threshold would be infinity because there are no observations from a different distribution (only extremes);

2) There is no reason why this fixed threshold should be appropriate for every data set; and

3) The threshold has to be adjusted to the sample size (see Reimann et al., 2005; and simulations below).

A better procedure than using a fixed threshold is to adjust the threshold to the data set at hand. Garrett (1989) used the chi-square plot for this purpose, by plotting the squared Mahalanobis distances (which have to be computed at the basis of robust estimations of location and scatter) against the quantiles of χ^2_p, the most extreme points are deleted until the remaining points follow a straight line. The deleted points are the identified outliers, the multivariate threshold corresponds to the distance of the closest outlier, the farthest background individual, or some intermediate distance. Alternately, the cube root of the squared Mahalanobis distances may be plotted against normal quantiles (e.g., Chork, 1990). This procedure (Garrett, 1989) is not

automatic, it needs user interaction and experience on the part of the analyst. Moreover, especially for large data sets, it can be time consuming, and also to some extent it is subjective. In the next section a procedure that does not require analyst intervention, is reproducible and therefore objective, and takes the above points, 1) to 3), into consideration is introduced.

4. ADAPTIVE OUTLIER DETECTION

The chi-square plot is useful for visualising the deviation of the data distribution from multivariate normality in the tails. This principle is used in the following. Let $G_n(u)$ denote the empirical distribution function of the squared robust distances RD_i^2, and let $G(u)$ be the distribution function of χ_p^2. For multivariate normally distributed samples, G_n converges to G. Therefore the tails of G_n and G can be compared to detect outliers. The tails will be defined by $\delta = \chi_{p;1-\alpha}^2$ for a certain small α (e.g., $\alpha = 0.02$), and

$$p_n(\delta) = \sup_{u \geq \delta}(G(u) - G_n(u))^+ \tag{2}$$

is considered, where "+" indicates the positive differences. In this way, $p_n(\delta)$ measures the departure of the empirical from the theoretical distribution only in the tails, defined by the value of δ. $p_n(\delta)$ can be considered as a measure of outliers in the sample. Gervini (2003) used this idea as a reweighting step for the robust estimation of multivariate location and scatter. In this way, the efficiency (in terms of statistical precision) of the estimator could be improved considerably.

$p_n(\delta)$ will not be directly used as a measure of outliers. As mentioned in the previous section, the threshold should be infinity in case of multivariate normally distributed background data. This means, that if the data are coming from a multivariate normal distribution, no observation should be declared as an outlier. Instead, observations with a large robust distance should be seen as extremes of the distribution. Therefore a critical value p_{crit} is introduced, which helps to distinguish between outliers and extremes. The measure of outliers in the sample is then defined as

$$\alpha_n(\delta) = \begin{cases} 0 & \text{if } p_n(\delta) \le p_{crit}(\delta, n, p) \\ p_n(\delta) & \text{if } p_n(\delta) > p_{crit}(\delta, n, p). \end{cases} \tag{3}$$

The threshold value is then determined as $c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta))$.

The critical value p_{crit} for distinguishing between outliers and extremes can be derived by simulation. For different sample sizes n and different dimensions (numbers of variables) p data from a multivariate normal distribution are simulated. Then equation (2) is applied for computing the value $p_n(\delta)$ for a fixed value δ (in the simulations $\delta = \chi^2_{p;0.98}$ is used). The procedure is repeated 1000 times for every considered value of n and p.

To directly compute the limiting distribution of the statistic defined by equation (2) would be a more elegant way for determining the critical value. However, even for related simpler problems Csörgő and Révész (1981, Chapter 5) note that this is analytically extremely difficult and they recommend simulation.

The resulting values give an indication of the differences between the theoretical and the empirical distributions, $G(u) - G_n(u)$, if the data are sampled from multivariate normal distributions. To be on the safe side, the 95% percentile of the 1000 simulated values can be used for every n and p, and these percentiles are shown for p=2, 4, 6, 8, 10 by different symbols in Figure 3. By transforming the x-axis by the inverse of $\sqrt{n}$ it can be seen that - at least for larger sample size - the points lie on a line (see Figure 3). The lines in Figure 3 are estimated by LTS (least trimmed sum of squares) regression (Rousseeuw, 1984). Using LTS regression the less precise simulation results for smaller sample sizes have less influence. The slopes of the different lines (the intercept is 0 because for n tending to infinity the difference between empirical and theoretical distribution is 0) are shown in Figure 4. The resulting points can again be approximated by a straight line, which allows definition of the critical value as a function of n and p:

$$p_{crit}(\delta, n, p) = \frac{0.24 - 0.003 \cdot p}{\sqrt{n}} \quad \text{for} \quad p \le 10. \tag{4}$$

For larger dimension ($p>10$) the same procedure can be applied. The 95% percentiles of 1000 simulated values for different sample sizes and dimensions are shown in Figure 5. The linear dependency becomes worse for high dimension and low sample size. The estimated slopes form a linear trend (Figure 6) and the resulting approximative formula is:

$$p_{crit}(\delta,n,p) = \frac{0.252 - 0.0018 \cdot p}{\sqrt{n}} \quad \text{for} \quad p > 10 . \qquad (5)$$

5. EXAMPLE

To test the procedure data from the Kola project (Reimann et al., 1998) are again used. The objective is to identify outliers in the O-horizon (organic surface soil) data caused by industrial contamination from Ni-smelters. A combination of two typical contaminant elements (Co and Cu), three minor contaminants (As, Cd and Pb) and two elements that are not part of the emission spectrum of the Ni-smelters (Mg and Zn) are used as a test data set. Magnesium is influenced by a second major process in the study area, the steady input of marine aerosols near the Arctic coast. This leads to a build-up of Mg in the O-horizon, and this process can be detected for more than 100 km inland (Reimann et al., 2000). Thus the test-task is to detect outliers in the 7-dimensional space at the basis of 617 observations. The procedure for adaptive outlier detection is illustrated in Figure 7. The solid line is the distribution function of χ_7^2. Robust squared distances RD_i^2 on the basis of the MCD estimator are computed, and their empirical distribution function, G_n, is represented by small circles. According to equation (2) the task is to find the supremum of the difference between these two functions in the tails. With $\delta = \chi_{7;0.98}^2 = 16.62$ (dotted line in Figure 7) a supremum of $p_n(\delta) = 0.1026$ is obtained. Equation (4) gives a critical value $p_{crit}(\delta,n,p) = 0.0088$, which is clearly lower than the above supremum. For this reason it can be assumed that large robust distances come from at least one different distribution. From equation (3) the measure of outliers is 10.26%, corresponding to 65 outliers. The resulting threshold value $c_n(\delta) = 18.64$ is slightly larger than δ, and presented in Figure 7 as a dashed line. This new threshold value is called the *adjusted quantile*.

6. VISUALISATION OF MULTIVARIATE OUTLIERS

An important issue is the visualisation of multivariate outliers, in the simplest case it is possible to plot them on a map. On a map clusters of outliers would indicate that some regions have a completely different data structure than others. Figure 8 shows the multivariate outliers for the above example on such a map, using the symbol "+" for outliers. Two clusters of outliers occur in Russia. As expected, they mark the two large industrial centres at Monchegorsk and Nikel with neighbouring Zapoljarnij. There are a number of outliers in the northwestern, Norwegian, part of the region. This is an almost pristine area with little industry and a low population density (see Reimann et al., 1998). At a first glance it is perhaps surprising to find outliers in this area. The detection of outliers due to contamination was the prime objective of the investigation. However, multivariate outliers are not only observations with high values for every variable, more importantly they are observations departing from the dominant data structure. In the case of a data set of contamination related variables, outliers also could be observations with very low values for the contamination related elements, indicating extremely clean (less-contaminated) regions. The reality is that Mg is highly enriched in marine aerosols and thus enriched in the O-horizon of podzols along the Norwegian coast, and in this remote near-pristine area the levels of the contamination related elements are within normal background ranges or low. Thus the reason for the Norwegian coast outliers is apparent, but Figure 8 makes no distinction between contamination and pristine coastal multivariate outliers.

The above demonstrates the necessity for developing a more effective way of visualising multivariate outliers. Firstly, it should be possible to provide a better visualisation of the distribution of the robust distances, and secondly it is desirable to distinguish between outliers with extremely low values and outliers having very high values of the variables.

Both features are fulfilled with the visualisation in Figure 9, the *multivariate outlier plot*. The simulated two-dimensional data set in Figure 9 represents a background and an outlying population. The robust distances were computed and – similar to Figure 1 – three inner tolerance ellipses (dotted lines) are shown for the 0.25, 0.5, and 0.75

quantiles of χ_2^2. The outer ellipse corresponds to the threshold $c_n(\delta)$ with $\delta = \chi_{2;0.98}^2$ of the adaptive outlier detection method. Values in the inner ellipse, which are at the center of the main mass of the data, are represented by a small dot. Observations between the 0.25 and 0.5 tolerance ellipses are shown by a larger dot. Going further outwards, a small circle is used as a symbol, and the most distant non-outliers are plotted as a small plus. Finally, multivariate outliers that are outside the outer tolerance ellipse are represented by a large plus.

For the second feature, i.e. distinguishing between different types of outliers, a colour (heat) scale is used that depends on the magnitude of the values for each variable. Low values are depicted in blue, and high values in red. More specifically, the colour scale is chosen according to the Euclidean distances (dashed lines) of the scaled observations from the coordinate-wise minimum, such that all coordinates have the same influence on the symbol colour. This procedure is illustrated in Figure 9 for the Euclidean distances of the simulated data.

Applying the above visualisation technique to the O-horizon soil data gives the multivariate outlier plot in Figure 10. Indeed, the spatial distribution of the robust distances becomes much clearer with the different symbols, and the colour scale is very helpful in distinguishing the different types of multivariate outliers. Two outlier clusters are proximal to the industrial centres at Monchegorsk and Nikel. Obviously, high values for most of the variables occur there, and hence give an indication of heavy contamination. The northern region of the investigated area also includes many multivariate outliers, but the symbols are in blue or green. This region is not at all contaminated and exhibits low values of the contaminant elements, and this combined with the input of sea spray (Mg) as a locally important process results in the outliers. The proposed visualisation permits discrimination between these very different families of outliers.

7. FROM MULTIVARIATE BACK TO UNIVARIATE

With the help of good visualisation for multivariate outliers it is easier to explain their structure and interpret the geochemical data. To support interpretation it is useful to

visualise the multivariate outliers for every single variable. Highlighting the multivariate outliers on the maps for every single element could achieve this. It is possible to use the same symbols as in the multivariate outlier plot to provide important information about the structure of these outliers.

For exploratory investigations, however, it is informative to have an overview of the position of the multivariate outliers within the distribution of the single elements. To achieve this we can simply plot the values of the elements and use the same symbols and colours as in the multivariate outlier plot. See Figure 11 for the Kola O-horizon data. All variables are presented as a series of vertically scaled parallel bars, where the values are scattered randomly in the horizontal direction (one-dimensional scatter plot). Since the original values of the variables have very different data ranges, the data were first centered and scaled for this presentation by using the robust multivariate estimates of location and scatter. In this way the different variables can be easily compared. This visualisation provides insight into the data structure and quality. As in the multivariate outlier plot, the multivariate outliers are presented by large symbols "+" for every variable. Not unsurprisingly in the light of the previous discussion, the multivariate outliers occur over the complete univariate data ranges, and not only at the extremes. Moreover, extremely low values, e.g., for Pb, which seem to be univariate outliers are not necessarily multivariate outliers. The explanation can be found by looking at the simulation example, Figure 9, again, where the lowest values for the x-axis are not multivariate outliers but members of the main data structure.

8. CONCLUSIONS

An automated method to identify outliers in multivariate space was developed and demonstrated with real data. In the univariate case it is often very difficult to identify data outliers originating from a second or other rare process, rather than extreme values in relation to the underlying data of the more common process(es). Extreme values can be easily detected due to their distance from the core of the data. If they originate from the underlying data they are of little interest to the exploration or environmental geochemist because they will neither identify mineralisation nor contamination. In contrast, in the multivariate case it is necessary to also consider the

shape of the data, its structure, in the multivariate space and all the dependencies between the variables. Thus the really interesting data outliers, caused by additional, rare processes, can be easily identified.

Not surprisingly the identified multivariate outliers in the test data set consisting of 7 variables and 617 samples are often not the univariate extreme values. In the context of Figure 1, they are equivalent to the distant off-axis individuals in the middle of the data range, e.g., the individual at (-1,1). The map of the multivariate outliers clearly identifies contaminated sites and those affected by the input of marine aerosols near the coast as regionally important processes causing different data outlier populations.

Although multivariate outlier identification is important for thorough data analysis, the task of interpretation goes beyond that first step as the researcher is also interested in identifying the geochemical processes leading to the data structure. A crucial point, however, is that multivariate outliers are not simply excluded from further analysis, but that after applying robust procedures which reduce the impact of the outliers the outliers are actually left in the data set. Working in this way permits the outliers to be viewed in the context of the main mass of the data, which facilitates an appreciation of their relationship to the core data. In this context, the data analyst should use a variety of procedures, often graphical, to gain as great an insight as possible into the data structure and the controlling processes behind the observations. For example, since factor analysis (like many other multivariate methods) is based on the covariance matrix, a robust estimation of the covariance matrix will reduce the effect of (multivariate) outlying observations (Chork and Salminen, 1993; Reimann et al., 2002) and lead to a data interpretation centred on the dominant process(es). Furthermore, when a single dominant process is present the factor loadings may be interpretable in the context of that process. When non-robust procedures are used in the presence of multiple processes factor analysis often behaves more like a cluster analysis procedure. In such cases the factor loadings provide little or no information on the internal structure of the processes, but define a framework for differentiating between them. Both applications have merit, the latter in exploratory data analysis, and the former in more detailed studies. Unfortunately, the EDA approach is often misused for a detailed process study, leading to questionable conclusions.

We conclude that proper exploratory data analysis and outlier recognition plays an essential part in the interpretation of geochemical data, and we suggest, data from other geoscience and physical science studies.

The method has been implemented in the free statistical software package *R* (see http://cran.r-project.org/). It is available as a contributed package called "mvoutlier", and it contains all the programs to the proposed methods and additionally valuable data sets from geochemistry, like the Kola data (Reimann et al., 1998) and data from Northern Europe (Reimann et al., 2003).

REFERENCES

Chork, C.Y., 1990. Unmasking multivariate anomalous observations in exploration geochemical data from sheeted-vein tin mineralisation near Emmaville, N.S.W., Australia. Journal of Geochemical Exploration 37 (2), 205-223.

Chork, C.Y., Salminen, R., 1993. Interpreting exploration geochemical data from Outukumpu, Finland: A MVE-robust factor analysis. Journal of Geochemical Exploration 48 (1), 1-20.

Csörgő, M., Révész, P., 1981. Strong Approximations in Probability and Statistics. Academic Press, New York, 284 pp.

Garrett, R.G., 1989. The chi-square plot: a tool for multivariate outlier recognition. Journal of Geochemical Exploration 32 (1/3), 319-341.

Gervini, D., 2003. A robust and efficient adaptive reweighted estimator of multivariate location and scatter. Journal of Multivariate Analysis 84, 116-144.

Gnanadesikan, R., 1977. Methods for the Statistical Data Analysis of Multivariate Observations. John Wiley and Sons, New York, 311 pp.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W., 1986. Robust Statistics. The Approach Based on Influence Functions. John Wiley & Sons, New York, 502 pp.

Maronna, R.A., Yohai, V.J., 1998. Robust estimation of multivariate location and scatter. In: Kotz, S., Read, C., Banks, D., (Eds.), Encyclopedia of Statistical Sciences Unpdate Volume 2. John Wiley & Sons, New York, pp. 589-596.

Reimann, C., Äyräs, M., Chekushin, V., Bogatyrev, I., Boyd, R., Caritat, P. De, Dutter, R., Finne, T.E., Halleraker, J.H., Jæger, Ø., Kashulina, G., Lehto, O., Niskavaara, H., Pavlov, V., Räisänen, M.L., Strand, T., Volden, T., 1998. Environmental Geochemical Atlas of the Central Barents Region. NGU-GTK-CKE Special Publication, Geological Survey of Norway, Trondheim, Norway, 745 pp. ISBN 82-7385-176-1.

Reimann, C., Banks, D., Kashulina, G., 2000. Processes influencing the chemical composition of the O-horizon of podzols along a 500 km north-south profile from the coast of the Barents Sea to the Arctic Circle. Geoderma 95, 113-139.

Reimann, C., Filzmoser, P., Garrett, R.G., 2002. Factor analysis applied to regional geochemical data: Problems and possibilities. Applied Geochemistry 17 (2), 185-206.

Reimann, C., Filzmoser, P., Garrett, R.G., 2005. Background and threshold: critical comparison of methods of determination. Science of the Total Environment, in press.

Reimann, C., Siewers, U., Tarvainen, T., Bityukova, L., Eriksson, J., Gilucis, A., Gregorauskiene, V., Lukashev, V.K., Matinian, N.N., Pasieczna, A., 2003. Agricultural Soils in Northern Europe: A Geochemical Atlas. Geologisches Jahrbuch, Sonderhefte, Reihe D, Heft SD 5, Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, 279 pp. ISBN 3-510-95906-X.

Rose, A.W., Hawkes, H.E., Webb, J.S., 1979. Geochemistry in Mineral Exploration, 2nd ed. Academic Press, London, 657 pp.

Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79 (388), 871-880.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W., (Eds.), Mathematical Statistics and Applications, Volume B. Akadémiai Kiadó, Budapest, pp. 283-297.

Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212-223.

Rousseeuw, P.J., Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85 (411), 633-651.
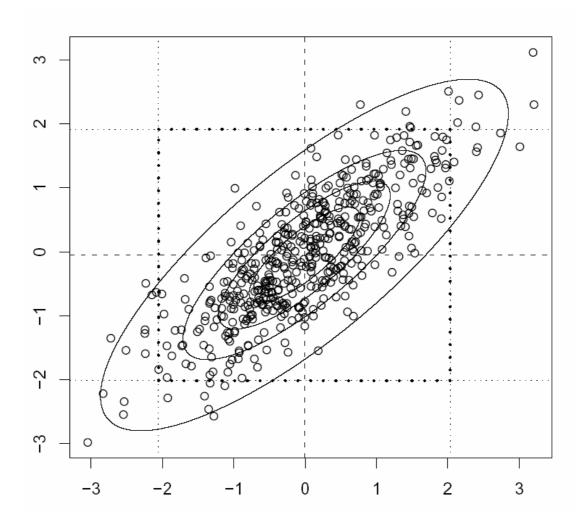
FIGURE CAPTIONS

Figure 1. Simulated standard normally distributed data with a predetermined correlation. The dashed lines mark the locations (means) of the variates, the ellipses correspond to the 0.25, 0.50, 0.75 and 0.98 quantiles of the chi-squared distribution, and the bold dotted lines to the 2nd and 98th empirical percentiles for the individual variables. Hence, the inner rectangular (bold dotted lines) can be considered for univariate outlier recognition, the outer ellipse for multivariate outlier identification.

Figure 2. Scatterplot of $\log_e$(Be) and $\log_e$(Sr). The covariance is visualised by tolerance ellipses. The non-robust estimation (dotted ellipse) leads to a Pearson correlation coefficient of 0.66, the robust procedure (solid ellipse) estimates a Pearson correlation of 0.18 for the core population, i.e. weight of 1, identified by the MCD procedure.

Figure 3. Simulated critical values according to equation (2) for multivariate normal distributions with different sample sizes (x-axis) and dimensions p. The linear trends for the dimensions plotted, and increasing sample size, are indicated by the lines.

Figure 4. The slopes of the lines from Figure 3 plotted against the dimension p. The line is an estimation of the linear trend, and leads to equation (4).

Figure 5. Simulated critical values analogous to Figure 3, but for higher dimensions ($p>10$).

Figure 6. The slopes of the lines from Figure 5 plotted against the dimension p. The line is an estimation of the linear trend, and leads to equation (5).

Figure 7. Adaptive outlier detection rule for the Kola O-horizon data: In the tails of the distribution (chosen as $\chi^2_{7;0.98}$ and indicated by a dotted line) we search for the supremum of the positive differences between the distribution function of χ^2_7 (solid line) and the empirical distribution function of RD^2_i (small circles). The resulting value is the adjusted quantile (dashed line) that separates outliers from non-outliers.

Figure 8. Map showing the regular observations (circles) and the identified multivariate outliers (+).

Figure 9. Preparation for the multivariate outlier plot: Five different symbols are plotted depending on the value of the robust distance. The five classes are defined by tolerance ellipses (dotted lines) for the chi-squared quantiles 0.25, 0.5, and 0.75, and the outlier threshold of the adaptive outlier detection method. The (heat) colour of the symbols varies continuously from the smallest to the largest values for every variable. Thus, observations lying on one dashed curve have the same colour.

Figure 10. Multivariate outlier plot with symbols according to Figure 9 provides an alternative presentation to Figure 8.

Figure 11. Plot of the single elements for the Kola O-horizon data, with the same symbols as used in Figure 10.

Figure 1. Simulated standard normally distributed data with a predetermined correlation. The dashed lines mark the locations (means) of the variates, the ellipses correspond to the 0.25, 0.50, 0.75 and 0.98 quantiles of the chi-squared distribution, and the bold dotted lines to the 2nd and 98th empirical percentiles for the individual variables. Hence, the inner rectangular (bold dotted lines) can be considered for univariate outlier recognition, the outer ellipse for multivariate outlier identification.
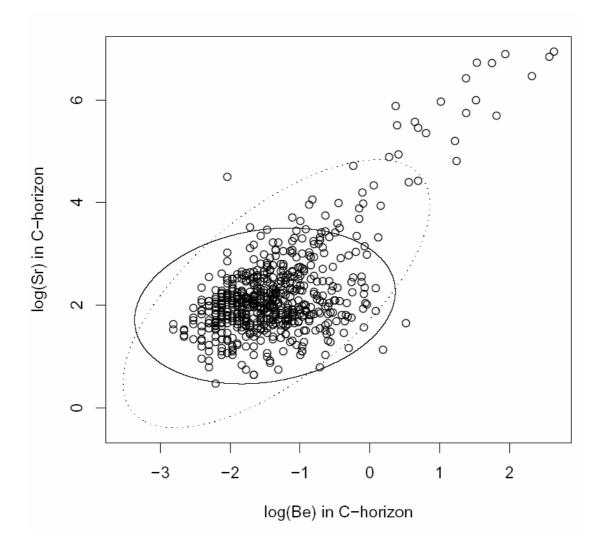
Figure 2. Scatterplot of $\log_e$(Be) and $\log_e$(Sr). The covariance is visualised by tolerance ellipses. The non-robust estimation (dotted ellipse) leads to a Pearson correlation coefficient of 0.66, the robust procedure (solid ellipse) estimates a Pearson correlation of 0.18 for the core population, i.e. weight of 1, identified by the MCD procedure.

Figure 3. Simulated critical values according to equation (2) for multivariate normal distributions with different sample sizes (x-axis) and dimensions p. The linear trends for the dimensions plotted, and increasing sample size, are indicated by the lines.
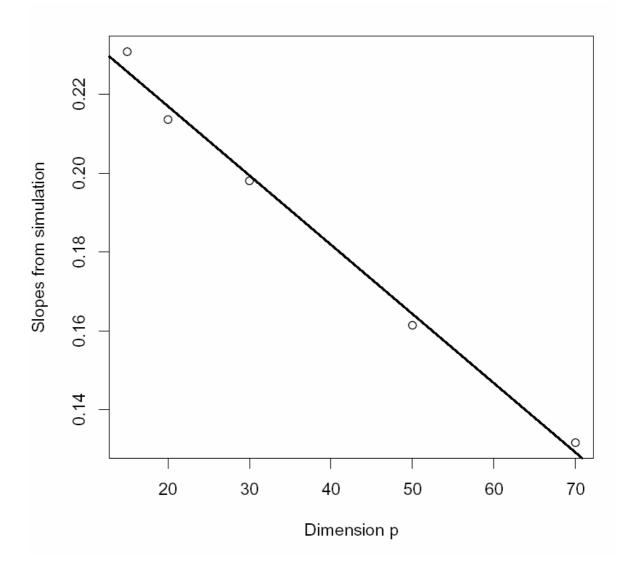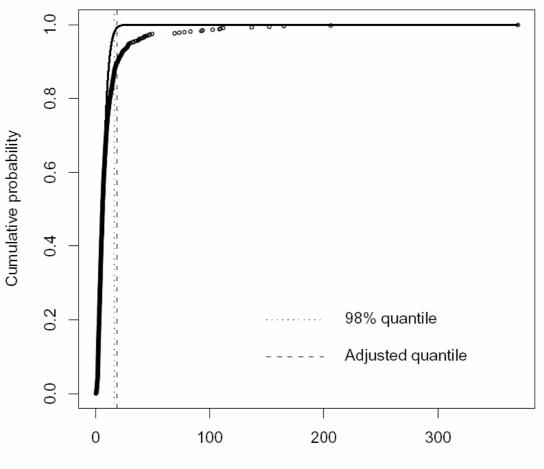
Figure 4. The slopes of the lines from Figure 3 plotted against the dimension *p*. The line is an estimation of the linear trend, and leads to equation (4).

Figure 5. Simulated critical values analogous to Figure 3, but for higher dimensions (*p*>10).

Figure 6. The slopes of the lines from Figure 5 plotted against the dimension *p*. The line is an estimation of the linear trend, and leads to equation (5).

Figure 7. Adaptive outlier detection rule for the Kola O-horizon data: In the tails of the distribution (chosen as $\chi^2_{7;0.98}$ and indicated by a dotted line) we search for the supremum of the positive differences between the distribution function of χ^2_7 (solid line) and the empirical distribution function of RD^2_i (small circles). The resulting value is the adjusted quantile (dashed line) that separates outliers from non-outliers.
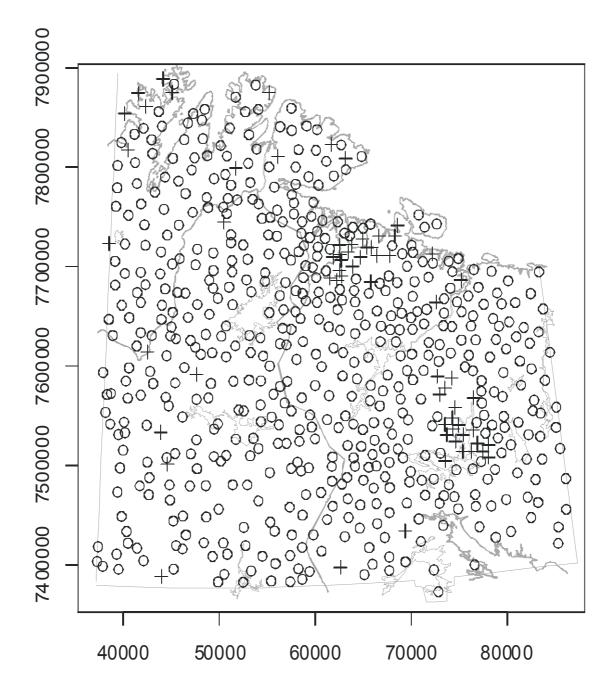
Figure 8. Map showing the regular observations (circles) and the identified multivariate outliers (+).
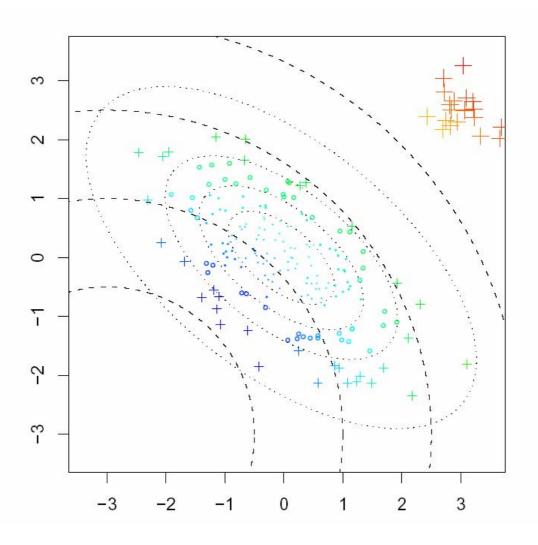
Figure 9. Preparation for the multivariate outlier plot: Five different symbols are plotted depending on the value of the robust distance. The five classes are defined by tolerance ellipses (dotted lines) for the chi-squared quantiles 0.25, 0.5, and 0.75, and the outlier threshold of the adaptive outlier detection method. The (heat) colour of the symbols varies continuously from the smallest to the largest values for every variable. Thus, observations lying on one dashed curve have the same colour.
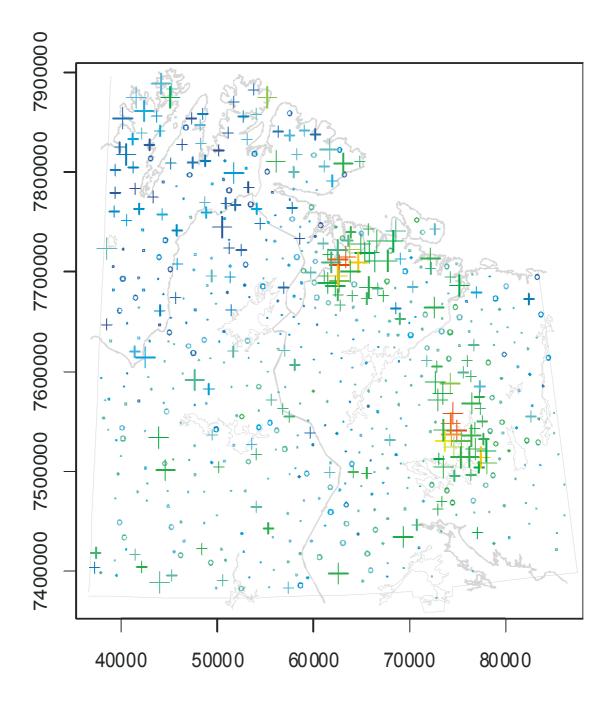
Figure 10. Multivariate outlier plot with symbols according to Figure 9 provides an alternative presentation to Figure 8.
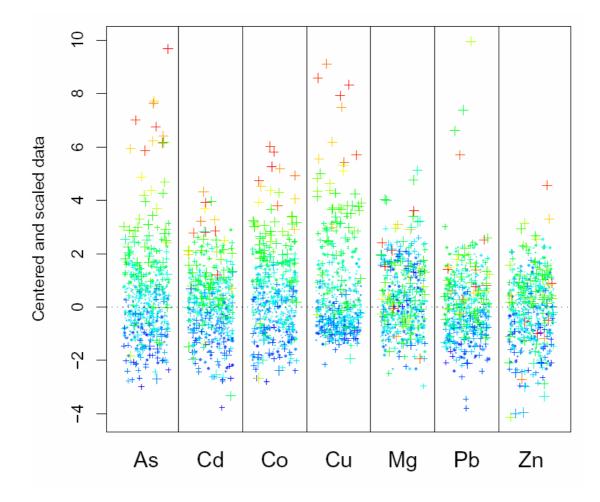
Figure 11. Plot of the single elements for the Kola O-horizon data, with the same symbols as used in Figure 10.