Noname manuscript No.
(will be inserted by the editor)

Detection of Multivariate Outliers in Business Survey Data with Incomplete Information

Valentin Todorov · Matthias Templ · Peter Filzmoser

Received: date / Accepted: date

Abstract Many different methods for statistical data editing can be found in the literature but only few of them are based on robust estimates (for example such as BACON-EEM, Epidemic algorithms (EA) and Transformed rank correlation (TRC) methods of Béguin and Hulliger). However, we can show that outlier detection is only reasonable if robust methods are applied, because the classical estimates are themselves influenced by the outliers. Nevertheless, data editing is essential to check the multivariate data for possible data problems and it is not deterministic like the traditional micro editing where all records are extensively edited manually using certain rules/constraints. The presence of missing values is more a rule than an exception in business surveys and poses additional severe challenges to the outlier detection. First we review the available multivariate outlier detection methods which can cope with incomplete data. In a simulation study, where a subset of the Austrian Structural Business Statistics is simulated, we compare several approaches. Robust methods based on the Minimum Covariance Determinant (MCD) estimator, S-estimators and OGK-estimator as well as BACON-BEM provide the best results in finding the outliers and in providing a low false discovery rate. Many of the discussed methods are implemented in the R package `rrcovNA` which is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org` under the GNU General Public License.

Keywords Multivariate outlier detection · Robust statistics · Missing values

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization.

V. Todorov
United Nations Industrial Development Organization (UNIDO), Vienna International Centre, P.O. Box 300, A-1400 Vienna, Austria
E-mail: v.todorov@unido.org

M. Templ
Department of Methodology, Statistics Austria
Department of Statistics and Probability Theory, Vienna University of Technology,

P. Filzmoser
Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria

1 Introduction

Outliers are present in virtually every data set in any application domain, and the identification of outliers has a hundred years long history. Much research has been done and many definitions for an outlier exist (see for example Barnett and Lewis, 1994) but it is sufficient to say that outliers are observations (not necessary errors) that are found far away from the main body of the data and do not follow an assumed model. The multivariate aspect of the data collected in surveys makes the task of outlier identification particularly challenging. The outliers can be completely hidden in one or two dimensional views of the data. This underlines that univariate outlier detection methods are useless, although they are favored in the National Statistical Offices (NSO) because of their simplicity. Outlier detection and robust estimation are closely related (see Hampel et al, 1986; Hubert et al, 2008) and the following two problems are essentially equivalent:

- Robust estimation: find an estimate which is not influenced by the presence of outliers in the sample.
- Outlier detection: find all outliers, which could distort the estimate.

A solution of the first problem allows us to identify the outliers using their robust residuals or distances while on the other hand, if we know the outliers we could remove or downweight them and then use the classical estimation methods. In many research areas the first approach is the preferred one but for the purposes of official statistics the second one is more appropriate. Therefore the focus in the present work is on using robust methods to identify the outliers which can be treated in the traditional way afterwards.

NSOs often correct the collected data in detail which is usually not necessary, because small errors from small enterprises have almost no effect on the ultimately published aggregated results (see Granquist, 1997). This process is highly time-consuming and resource intensive and often requires a significant part of the costs of the whole survey. Furthermore, this process is error-prone, because editing rules determined by subject matter specialists may destroy the multivariate structure of the data and in case of survey data, sampling errors may cause that correct data entries are marked as uncorrect by deterministic rules (see also, De Waal, 2009).

To improve the editing process, techniques such as selective editing (see, e.g., Lawrence and McKenzie, 2000), automatic editing (De Waal, 2003; Fellegi and Holt, 1976) and macroediting (see, e.g., Granquist, 1990) can be applied instead of the traditional microediting approach, where all records are extensively edited manually (see also, De Waal, 2009). Automatic editing is done by formulating a framework for both, the automatic construction of implicit edits and a linear programming algorithm to edit the data in an optimized manner automatically (Fellegi and Holt, 1976) using heuristic approaches for large data sets. While this approach is perfectly suited for editing categorical data, outlier detection seems to be preferable when editing continuous or semi-continuous scaled variables. Selective editing is closely related to the concept of the influence function in robust statistics (in both concepts the effect of each observation on an estimator is measured), i.e. selective editing detects those statistical units which highly influence one (or a few) pre-defined measure(s). However, usually many different estimates are obtained from the same data set and those outliers which do only have high influence on the measures as used in the selective editing process are not detected.

The aim of outlier detection in the survey context is to detect outliers in a multivariate space in general. The detected outliers may be representative or non-representative (Chambers, 1986), i.e. the procedure can flag as outliers not only measurement errors but also correct entries which do have an abnormal behaviour compared to the main bulk of the data. In other words, outliers do not need to be errors but, if they are, they can have a large influence on the estimates and it is reasonable to detect them. Fortunately great reduction of costs in the statistical editing process is possible by outlier detection (see also Luzi et al, 2007) which helps for understanding the quality of the data and contributes to ending up with acceptable estimations and aggregates. After the outliers are identified the statistical agencies have information which statistical units should be analysed and possibly corrected. In business surveys they usually contact those outlying enterprises where they cannot explain the abnormality of the behaviour of these enterprises in the underlying data set.

In contrast to univariate outliers, multivariate outliers are not necessarily extreme along a single coordinate. They could deviate from the multivariate structure formed by the majority of the observations. To illustrate this we will consider the well-known `bushfire` data set which was used by Campbell (1989) to locate bushfire scars and was studied in detail by Maronna and Yohai (1995). It is available from the R package **robustbase** and is a complete data set consisting of 38 observations in 5 dimensions. In the left panel of Figure 1 a scatter-plot of the variables `V2` and `V3` is shown which reveals most of the outliers - the two clusters 33-38 and 7-11. The estimated central location of each variable is indicated by dashed-dotted lines and their intersection represents the multivariate centroid of the data. The dotted lines are at the 2nd and 98th empirical percentiles for the individual variables. A univariate outlier detection would declare as candidates the observations falling outside the rectangle visualized with bold dotted lines. Such a procedure would ignore the elliptical shape of the bivariate data. The bivariate data structure can be visualized by Mahalanobis distances, which depend on the center and the covariance (see Equation (1) below). Certain quantiles (e.g. 0.25, 0.50, 0.75 and 0.98) will result in tolerance ellipses of the corresponding size. It is, however, crucial how location and covariance are estimated for this purpose. Both the univariate and the multivariate procedures illustrated in the left panel of Figure 1 are based on classical estimates of location and covariance and therefore they fail to identify the outliers in the data. The right panel of Figure 1 shows the same picture but robust estimates of location and covariance are used (here we used the MCD estimator, see below). All outliers lie clearly outside the ellipse corresponding to the 0.98th quantile.

In sample survey data the outliers do not come alone, but are almost always ”accompanied“ by missing values, large data sets, sampling weights, mixture of continuous and categorical variables, to name some of the additional challenges and these pose severe requirements on the estimators. Survey data are usually incomplete and we assume that the data are missing at random (MAR) (Little and Rubin, 1987), i.e. the missing values depend only on observed data. This is a weaker assumption than the missing completely at random (MCAR) mechanism when the missingness is unrelated to the observed data. In the setting of business surveys we do not know for certain if MAR holds and in general it is not possible to test if MAR holds for a particular data set, nevertheless it is possible to use exploratory data analysis tools like these provided in the R package **VIM** to investigate the nature of the process generating the missing values (see Section 4.2). We could expect departures from MAR, but the pragmatic approach would be to expect at the same time that these departures would not be serious enough to cause the performance of MAR-based methods to be seriously degraded as pointed out by Schafer and Graham (2002), see also Luzi et al (2007). Schafer

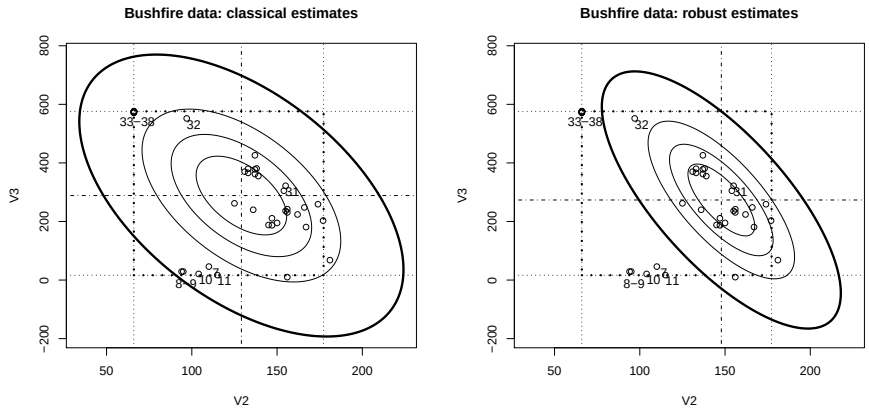

Fig. 1 Example of multivariate outliers: the variables `V2` and `V3` of the bushfire data set. Dashed-dotted lines mark the centers of the variables, and ellipses represent the 0.25, 0.50, 0.75 and 0.98 quantiles of the Mahalanobis distances. The bold dotted lines are the 2nd and 98th empirical percentiles of the individual variables. Location and covariance in the left panel are estimated by the sample mean and covariance matrix while in the right panel a robust alternative is used.

and Graham (2002), page 173, recommend, although other procedures could be occasionally useful too, to apply likelihood-based procedures, which are appropriate under general MAR conditions.

One of the unique features of the outlier detection problem in the analysis of survey data is the presence of sampling weights. For example, in business surveys the sample design is defined in such a way that the sampling units (e.g. enterprises or establishments) with large size (large number of employees or high turnover) are selected with high probability (often selected in "take-all" strata) while the small sized ones are sampled with low probability (see for example Hidiroglou and Lavallée, 2009). The sampling weights are (calibrated) inverses of these (inclusion) probabilities. These sampling weights will be used in the estimation procedure and therefore cannot be left unaccounted for in the phase of data cleaning and outlier detection.

These challenges are probably the reason why multivariate outlier detection methods are rarely applied to sample survey data. One of the exceptions is Statistics Canada (Franklin et al, 2000; Franklin and Brodeur, 1997) where a robust method for multivariate outlier detection was used in the Annual Wholesale and Retail Trade Survey (AWRTS). This method is based on a robust version of principal component analysis using robust estimates of multivariate location and covariance obtained from the Stahel-Donoho estimator (SDE) (Stahel, 1981a; Donoho, 1982). Several robust methods are investigated in the EUREDIT project (EUREDIT Project, 2004) which aims at improving the efficiency and the quality of automatic methods for statistical data editing and imputation at NSOs. These robust methods are further developed by Béguin and Hulliger (2004, 2008) and will be considered also in the present study.

This paper is organized as follows. In Section 2 the general outlier detection framework

is presented and the available algorithms are briefly reviewed. Their applicability to incomplete data is discussed and their computational performance is compared in Section 3. Section 4 presents an example based on a well-known complete data set with inserted simulated missing data and continues with a simulation study which mimics a real data set from *Statistics Austria*. Section 5 presents the availability of the software discussed so far and Section 6 concludes.

2 Algorithms for outlier detection

2.1 General principles

A general framework for multivariate outlier identification in a p-dimensional data set $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is to compute some measure of the distance of a particular data point from the center of the data and declare as outliers those points which are too far away from the center. Usually, as a measure of "outlyingness" for a data point $\boldsymbol{x}_i, i = 1, \ldots, n$, a robust version of the (squared) Mahalanobis distance RD_i^2 is used, computed relative to high breakdown point robust estimates of location $\boldsymbol{T}$ and covariance $\boldsymbol{C}$ of the data set $\boldsymbol{X}$:

$$RD_i^2 = (\boldsymbol{x}_i - \boldsymbol{T})^t \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{T}) \tag{1}$$

The most common estimators of multivariate location and scatter are the sample mean $\bar{\boldsymbol{x}}$ and the sample covariance matrix $\boldsymbol{S}$, i.e. the corresponding ML estimates (when the data follow a normal distribution). These estimates are optimal if the data come from a multivariate normal distribution but are extremely sensitive to the presence of even a few outliers in the data. The outlier identification procedure based on $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ will suffer from the following two problems (Rousseeuw and Leroy, 1987):

1. *Masking:* multiple outliers can distort the classical estimates of mean $\bar{\boldsymbol{x}}$ and covariance $\boldsymbol{S}$ in such a way (attracting $\bar{\boldsymbol{x}}$ and inflating $\boldsymbol{S}$) that they do not get necessarily large values of the Mahalanobis distance, and
2. *Swamping:* multiple outliers can distort the classical estimates of mean $\bar{\boldsymbol{x}}$ and covariance $\boldsymbol{S}$ in such a way that observations which are consistent with the majority of the data get large values for the Mahalanobis distance.

In the last several decades much effort was devoted to the development of affine equivariant estimators possessing a high breakdown point. The most widely used estimators of this type are the Minimum Covariance Determinant (MCD) estimator and the Minimum Volume Ellipsoid (MVE) estimator, S-estimators and the Stahel-Donoho estimator. These estimators can be configured in such a way as to achieve the theoretically maximal possible breakdown point of 50% which gives them the ability to detect outliers even if their number is as much as almost half of the sample size. If we give up the requirement for affine equivariance, estimators like the orthogonalized Gnanadesikan-Kettenring (OGK) estimator are available and the reward is an extreme gain in speed. For definitions, algorithms and references to the original papers it is suitable to use Maronna et al (2006). Most of these methods are implemented in the R statistical environment (R Development Core Team, 2009) and are available in the object-oriented framework for robust multivariate analysis (Todorov and Filzmoser, 2009).

After having found reliable estimates for the location and covariance matrix of the data set, the second issue is to determine how large the robust distances should be in order to

declare a point an outlier. The usual cutoff value is a quantile of the χ^2 distribution, like $D_0 = \chi^2_p(0.975)$. The reason is that if $\boldsymbol{X}$ follows a multivariate normal distribution, the squared Mahalanobis distances based on the sample mean $\bar{\boldsymbol{x}}$ and sample covariance matrix $\boldsymbol{S}$ follow χ^2_p distribution (see for example Johnson and Wichern, 2002, p. 189). This will no more be valid if robust estimators are applied and/or if the data have other than multivariate normal distribution. In Maronna and Zamar (2002) it was proposed to use a transformation of the cutoff value which should help the distribution of the squared robust distances RD_i^2 to resemble χ^2 for non-normal original data:

$$D_0 = \frac{\chi^2_p(0.975) med(RD_1^2, \ldots, RD_n^2)}{\chi^2_p(0.5)}. \tag{2}$$

For other alternatives which could lead to more accurate cutoff value see Filzmoser et al (2005); Hardin and Rocke (2005); Cerioli et al (2009); Riani et al (2009).

A drawback of all so far considered methods is that they work only with complete data which is not a usual case when dealing with sample surveys. In the next two subsections we describe and introduce methods that are able to cope with missing values.

2.2 Algorithms based on imputation

Robustifying the EM algorithm. Little and Smith (1987) were the first to propose a robust estimator for incomplete data by replacing the MLE in the M-step of the EM algorithm (see Dempster et al, 1977) by an estimator belonging to the general class of M-estimates Huber (1981) and called this procedure ER-estimator. They suggested to use as a starting point for the ER algorithm ML estimation where the missing values were replaced by the median of the corresponding observed data. Unfortunately, the breakdown point of this estimator, as of all general M-estimates cannot be higher than $1/(p+1)$ (see for example Maronna et al, 2006, p. 186) which renders it unusable for the purpose of outlier detection. Copt and Victoria-Feser (2004) constructed a high breakdown point estimator of location and covariance for incomplete multivariate data by modifying the MCD estimator and using it as a starting point not for an ER algorithm but for an S-estimator, adapted to work with incomplete data. They call this estimator `ERTBS`. An implementation of this procedure was available by the authors in the form of a compiled shared library but it did not perform as well as expected and was excluded from further investigations in this work.

Normal imputation followed by high-BP estimation. A straightforward strategy for adapting estimators of location and covariance to work with missing data is to perform one preliminary step of imputation and then run any of the above described algorithms, like for example MCD, OGK, S and Stahel-Donoho (SDE) on the complete data. Many different methods for imputation have been developed over the last few decades and here we will consider a likelihood-based approach such as the before mentioned expectation maximization (EM) imputation method (Dempster et al, 1977) assuming the underlying model for the observed data is Gaussian. This method is able to deal with MCAR and MAR missing values mechanism. For Gaussian data the EM algorithm starts with some initial values for the mean and the covariance matrix and iterates through imputing missing values (imputation step) and re-estimating the mean and the covariance matrix from the complete data set (estimation step). The iteration process stops when the maximum relative difference in

all of the estimated means, variances or covariances between two iterations is less than or equal to a given value. The estimated in this way parameters are used to draw the missing elements of the data matrix under the multivariate normal model (all further necessary details about the Gaussian imputation can be found in Schafer, 1997, Sections 5.3 and 5.4). We are tempted to call this class of methods for example PM-MCD (poor man's MCD), etc. but for the sake of simplicity we will call them simply MCD, S, SDE, etc. meaning the high breakdown method applied to complete data and at the same time this method applied to normally (non-robustly) imputed data. Whenever another method for imputation precedes the high breakdown estimation method, like the robust sequential imputation described in the next Section, the corresponding notation will be used. In this way we can adapt also projection based algorithms like SIGN1 (Filzmoser et al, 2008).

The next step after estimating reliably the location $\boldsymbol{T}$ and covariance matrix $\boldsymbol{C}$ is to compute the robust distances from the incomplete data. For this purpose we have to adapt Equation (1) to use only the observed values in each observation $\boldsymbol{x}_i$ and then to scale up the obtained distance. We rearrange the variables if necessary and partition the observation $\boldsymbol{x}_i$ into $\boldsymbol{x}_i = (\boldsymbol{x}_{oi}, \boldsymbol{x}_{mi})$ where $\boldsymbol{x}_{oi}$ denotes the observed part and $\boldsymbol{x}_{mi}$ - the missing part of the observation. Similarly, the location and covariance estimates are partitioned, so that we have $\boldsymbol{T}_{oi}$ and $\boldsymbol{C}_{oi}$ as the parts of $\boldsymbol{T}$ and $\boldsymbol{C}$ which correspond to the observed part of $\boldsymbol{x}_i$. Then

$$RD_{oi}^2 = (\boldsymbol{x}_{oi} - \boldsymbol{T}_{oi})^t \boldsymbol{C}_{oi}^{-1} (\boldsymbol{x}_{oi} - \boldsymbol{T}_{oi}) \tag{3}$$

is the squared robust distance computed only from the observed part of $\boldsymbol{x}_i$. If $\boldsymbol{x}_i$ is uncontaminated, follow a multivariate normal distribution, and if the missing values are missing at random, then the squared robust distance given by Equation (3) is asymptotically distributed as $\chi^2_{p_i}$ where p_i is the number of observed variables in $\boldsymbol{x}_i$ (see Little and Smith, 1987).

The MCD estimator is not very efficient at normal models, especially if h is selected so that maximal breakdown point (BP) is achieved (Croux and Haesbroeck, 1999), and the same is valid for the OGK estimator (Maronna et al, 2006, p. 193, 207). To overcome the low efficiency of these estimators, a reweighted version can be used (see Lopuhaä and Rousseeuw, 1991; Lopuhaä, 1999). For this purpose a weight w_i is assigned to each observation $\boldsymbol{x}_i$, defined as $w_i = 1$ if $RD_{oi}^2 \leq \chi^2_{p_i,0.975}$ and $w_i = 0$ otherwise, relative to the raw estimates $(\mathbf{T}, \mathbf{C})$ and using Equation (3). Then the reweighted estimates are computed as

$$\begin{aligned} \mathbf{T}_R &= \frac{1}{\nu} \sum_{i=1}^{n} w_i \boldsymbol{x}_i, \\ \mathbf{C}_R &= \frac{1}{\nu - 1} \sum_{i=1}^{n} w_i (\boldsymbol{x}_i - \mathbf{T}_R)(\boldsymbol{x}_i - \mathbf{T}_R)^t, \end{aligned} \tag{4}$$

where ν is the sum of the weights, $\nu = \sum_{i=1}^{n} w_i$. Since the underlying data matrix is incomplete, the EM algorithm is used to compute $\mathbf{T}_R$ and $\mathbf{C}_R$. These reweighted estimates $(\mathbf{T}_R, \mathbf{C}_R)$ which have the same breakdown point as the initial (raw) estimates but better statistical efficiency are computed and used by default for the methods MCD and OGK.

Robust sequential imputation followed by high-BP estimation. Since we assume that outliers are present in the data we could expect an improvement of the performance of the previously described methods if the non-robust Gaussian imputation is substituted by a robust imputation technique that can handle simultaneously missing and outlying values. One

such method was proposed by Vanden Branden and Verboven (2009) (RSEQ), extending the sequential imputation technique (SEQimpute) of Verboven et al (2007) by robustifying some of its crucial steps. SEQimpute starts from a complete subset of the data set $\mathbf{X}_c$ and estimates sequentially the missing values in an incomplete observation, say $\mathbf{x}^*$, by minimizing the determinant of the covariance of the augmented data matrix $\mathbf{X}^* = [\mathbf{X}_c; (\mathbf{x}^*)^t]$. Since SEQimpute uses the sample mean and covariance matrix it will be vulnerable to the influence of outliers and it is improved by plugging in robust estimators of location and scatter. One possible solution is to use the outlyingness measure as proposed by Stahel (1981b) and Donoho (1982) and successfully used for outlier identification in Hubert et al (2005). We can compute the outlyingness measure for the complete observations only but once an incomplete observation is imputed (sequentially) we could compute the outlyingness measure for it too and use it to decide if this observation is an outlier or not. If the outlyingness measure does not exceed a predefined threshold the observation is included in the further steps of the algorithm. After obtaining a complete data set we proceed by applying a high breakdown point estimation method in the same way as described in the previous section.

2.3 Algorithms based on other strategies

Transformed Rank Correlation (TRC). This is one of the algorithms proposed by Béguin and Hulliger (2004) and is based, similarly as the OGK algorithm of Maronna and Zamar (2002) on the proposal of Gnanadesikan and Kettenring for pairwise construction of the covariance matrix. The initial matrix is calculated using bivariate Spearman Rank correlations $\rho(\boldsymbol{x}_k, \boldsymbol{x}_h), 0 \geq i, j \leq p$ which is symmetric but not necessarily positive definite. To ensure positive definiteness of the covariance matrix the data are transformed into the space of the eigenvectors of the initial matrix and univariate estimates of location and scatter are computed which are used then to reconstruct an approximate estimate of the covariance matrix in the original space. The resulting robust location and covariance matrix are used to compute robust distances for outlier identification. In case of incomplete data the Spearman rank correlations are computed only from observations which have values for both variables involved and thus the initial covariance matrix estimate can be computed using all available information. From this matrix the transformation matrix can be computed but in order to apply the transformation complete data matrix is needed which is the most difficult problem in this algorithm. To obtain complete data each missing item x_{ik} is imputed by a simple robust regression of $\boldsymbol{x}_k$ on $\boldsymbol{x}_h$ where $\boldsymbol{x}_h$ is the variable with highest rank correlation $\rho(\boldsymbol{x}_k, \boldsymbol{x}_h)$. In order to be able to impute x_{ik}, the item x_{ih} must be observed and the quality of the regression on a given variable is controlled by an overlap criterion, i.e. in order to choose a variable as a regressor, the number of observations in which both variables are observed must be greater than some γn where $0 < \gamma < 1$ is a tuning constant. After all missing items have been imputed (or some observations with too few observed items have been removed) the complete data matrix can be used to perform the transformation and compute the final robust location and covariance matrix as in the case of complete data.

Epidemic Algorithm (EA). The second algorithm proposed by Béguin and Hulliger (2004) is based on data depth and is distribution free. It simulates an epidemic which starts at a multivariate robust center (sample spatial median) and propagates through the point cloud. The infection time is used to judge on the outlyingness of the points. The latest infected points or those not infected at all are considered outliers. The adaption of the algorithm to missing values is straightforward by leaving out missing values from the calculations of

the univariate statistics (medians and median absolute deviations) as well as from the distance calculations. If too many items are missing in a pair of observations the corresponding distance is set to infinity (in the practical implementation a simplified criterion is applied excluding observations with less than $p/2$ observed items from the epidemic). EA is the only method studied here which has no assumptions about the form of the distribution but the tuning of the infection in order to attain best performance may be problematic. Without going into detail about the choice of parameters we will note that in general the recommendations given by the authors (Béguin and Hulliger, 2004, p.279) and the defaults in the R programs were followed. It seems that of crucial importance is the choice of the transmission function which determines the probability that a point is infected given another point and the distance between them, the reach of the infection and the selection of the deterministic mode of infection. In the examples in Section 4 we used the root function with maximal reach and deterministic infection.

BACON-EEM (BEM). The third algorithm (Béguin and Hulliger, 2008) developed in the framework of the EUREDIT project is based on the algorithm proposed by Billor et al (2000) which in turn is an improvement over an earlier "forward search" based algorithm by one of the authors. It is supposed to be a balance between affine equivariance and robustness - it starts from a data set which is *supposed* to be outlier free and moves forward by inspecting the rest of the observations - good points are added as long as possible. The adaptation for incomplete data consists in replacing the computation of the location and covariance at each step by an EM-algorithm.

The last three algorithms - TRC, EA and BEM take into account also the sampling context in which the data are assumed to be a random sample s of size n from a finite population U with size N and the sample is drawn with the sample design $p(s)$. The sample design $p(s)$ defines the inclusion probabilities and the corresponding sampling weights. With these weights the classical mean and covariance are estimated using the Hàjek estimators and the weighted median and MAD are estimated as described in Béguin and Hulliger (2004) where further details about the adaptation of the algorithms to sampling weights can be found.

It is rather difficult to provide general recommendations on the appropriate algorithm to use in a certain data configuration. Most algorithms rely on elliptical symmetry of the data majority; an exception is the *Epidemic Algorithm*. Affine equivariance of estimators like MCD or S is usually lost if a normal imputation is done a-priori within the algorithm. The combination of several routines makes it also impossible to determine the overall breakdown point of the procedure. For these reasons, a comparison is best made at the basis of the computation time (Section 3), and within real and simulated data sets (Section 4).

3 Computation times

It is important to consider the computational performance of the different algorithms in case of large data sets (with $n \geq 100$). To evaluate and compare the computational times a simulation experiment was carried out. Contaminated data sets with different sizes and dimensions, (n, p), varying from (100,5) to (50000,30) were generated. The generated data followed the model of *mean-shift outliers*, i.e. in each data set $(1 - \varepsilon)n$ of the observations were generated from standard multivariate normal distribution $N_p(\mathbf{0}, \mathbf{I}_p)$ and the remaining εn were generated from $N_p(\boldsymbol{\mu}, \mathbf{I}_p)$ where $\boldsymbol{\mu} = (b, \ldots, b)^t$ with $b = 10$. The contamination

proportion ε was chosen to be equal to 0.25, i.e. 25% of the generated observations in each data set were outliers. Then 20% missing values, created with a missing completely at random (MCAR) mechanism were included in the generated data sets. For this purpose all np observation items were subjected to independent Bernoulli trials, realized with a probability of success (i.e. the item is set to missing) $q = 0.20$. For each (n, p) combination 100 data sets were generated and the runtime of each method was averaged across them.

The experiment was performed on a 2.4Ghz Intel®Core™2 Quad PC with 3.5Gb RAM running Windows XP Professional. All computations were performed in R 2.11.1. The MCD, OGK, S and SDE algorithms from **rrcov** 1.0-1 were used and the imputation was performed by functions `em.norm()` and `imp.norm()` from the package **norm** (Schafer, 1997). The R code for the EA, TRC and BEM algorithms was provided by the authors. Figure 2 displays graphically the results of the experiment in logarithmic scale. Fastest is SIGN1, followed closely by OGK and MCD with MCD being faster than OGK when n and p are large due to the nesting and partitioning steps of the FAST-MCD algorithm. The computation time of the S algorithm is similar to that of MCD but S is slower for higher problem sizes because as initial estimate is used MVE and not MCD. It is important to note that although MCD has better statistical properties and is faster to compute, the recommended initial estimate in the S estimator algorithm is MVE because of its smallest maximum bias (see Maronna et al, 2006, p.199). EA is also fast, but it cannot cope with larger problems because of very high memory requirements. Next comes BEM which suffers from the fact that the available implementation is in pure R as compared to the native implementation of MCD (FORTRAN), OGK (C) and S estimates (the most computationally intensive part - the initial MVE estimates in C). With RSEQ we denote here for brevity the robust sequential imputation algorithm followed by MCD. It would be slightly faster if the imputation is followed by SIGN1 or OGK instead of MCD. The Stahel-Donoho estimator is much slower for large data sets than the other algorithms and therefore we show only the results for $p = 5$ and $p = 10$ (for comparison, with $p = 20$ it took 1 hour for $n = 1000$ and 13 hours for $n = 10000$ while MCD did the same in 1 and 2 seconds, respectively).

4 Examples and Experiments

In this section we illustrate the behaviour of the considered algorithms in the presence of missing values on examples. The first example is based on a well known in the literature complete real data set in which we introduce missing data with varying percentage. The advantage of this data set is that it is very well studied and the outliers are known in advance. In the second example we evaluate the algorithms on generated synthetic (but close-to-reality) data sets based on a real structural business statistics data set. This is a standard procedure in official statistics providing a way to avoid the disclosure of information (see also Rubin, 1993). This type of data sets include more difficulties for outlier detection algorithms than we could cover in the current study and give a promising area of future work.

4.1 Bushfire data

As a first example we consider the `bushfire` data set which was briefly presented in Section 1. Although this data set is not specific for survey data, it is well known in the literature, and it may give first insights into the performance of the algorithms. The classical outlier

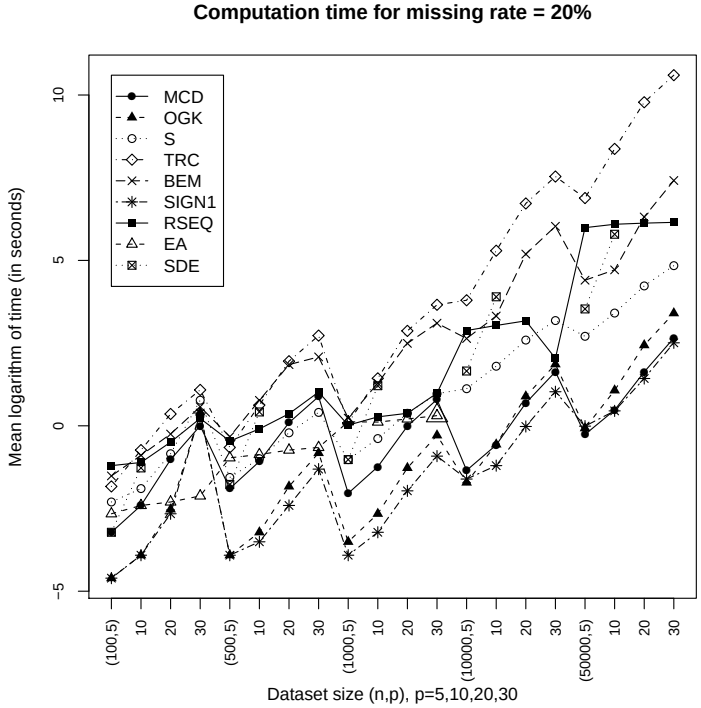

Fig. 2 Comparing the computation time of 9 robust estimation algorithms in the presence of 20% missing data.

detection method using the sample mean and covariance matrix in Equation 1 (see the illustration in the left panel of Figure 1) does not recognize any outlier but most of the robust high breakdown point methods (MCD, OGK, S, BEM, SDE) identify the outlying clusters 32-38 and 7-11 and to a lesser extent the single outlying observation 31 (cf. Maronna and Yohai, 1995, p.337). Observations 29 and 30 lie on the boundary. The outliers are clearly identified in the distance-distance plot (Rousseeuw and van Zomeren, 1990) shown in Figure 3. Thus the list of outliers in our example will consist of the 13 observations 7-11 and 31-38. We simulate item-non-response by applying MCAR mechanism which generates a fraction $q = \{0.0, 0.1, 0.2, 0.3, 0.4\}$ of missing values. For each of these five cases we generate $m = 400$ data sets, perform the outlier detection with each of the considered methods and calculate the following two measures:

- FN - Average outlier error rate: the average percentage of outliers that were not identified - false negatives or masked outliers
- FP - Average non-outlier error rate: the average percentage of non-outliers that were classified as outliers - false positives or swamped non-outliers

The results are shown in Table 1. When no missing values are present, all methods but EA and TRC behave properly and identify all outliers. Both EA and TRC miss one outlier and declare one good observation as an outlier. The non-outlier error rate for the complete data set can be explained for most of the procedures (MCD, OGK, S, SDE) by the observations lying on the border - 12, 28, 29 and 30 - which in many cases are declared outliers. With

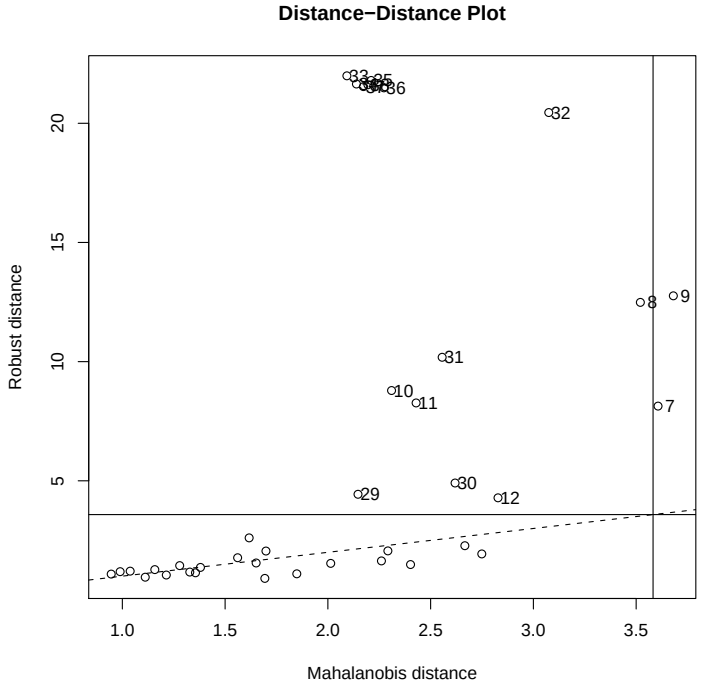

Fig. 3 Distance-distance plot of the `bushfire` data set computed with the FAST-MCD algorithm (using function `CovMcd()` from package **rrcov**).

Table 1 Bushfire data with simulated missing values: average percentage of outliers that were not identified and average percentage of regular observations that were declared outliers

	Outlier error rate (FN)					Non-outlier error rate (FP)				
	0	10	20	30	40	0	10	20	30	40
MCD	0.00	2.52	9.15	17.12	30.83	12.00	3.88	2.91	2.14	2.08
OGK	0.00	5.75	19.60	32.15	41.15	16.00	6.26	4.30	3.36	3.68
EA	7.69	13.79	27.85	43.00	53.21	4.00	12.14	11.58	10.37	9.08
SIGN1	0.00	4.37	20.00	35.71	46.94	20.00	12.49	8.05	5.54	4.58
SDE	0.00	7.37	17.12	26.73	41.29	2.29	3.15	1.90	1.56	1.44
BEM	0.00	4.88	5.69	9.31	14.04	4.00	8.71	11.37	9.78	7.76
S	0.00	9.08	20.88	30.79	45.19	0.00	3.81	4.49	3.92	3.52
TRC	7.69	16.46	16.79	14.50	13.69	4.00	13.64	11.47	10.13	11.57
RSEQ_MCD	0.00	1.71	8.73	19.12	34.46	12.00	9.34	9.44	3.50	1.94
RSEQ_OGK	0.00	2.04	6.25	28.98	48.71	16.00	9.02	9.57	7.75	3.98
RSEQ_SDE	0.00	1.88	9.79	21.37	31.21	7.53	9.10	7.75	2.95	1.97
RSEQ_S	0.00	2.92	14.15	28.85	37.63	0.00	5.65	8.80	4.79	2.88
RSEQ_SIGN1	0.00	2.08	9.67	47.65	78.12	20.00	13.07	10.49	17.39	22.41

10% of missing data the RSEQ_XXX methods based on robust sequential imputation are best with respect to the average outlier error rate, followed by MCD, SIGN1 and BEM. In terms of the non-outlier error rate, SDE, S and MCD are the best followed by RSEQ_S, OGK and BEM. In general all methods based on robust imputation have higher non-outlier error rate than the corresponding methods preceded by normal imputation which is the price paid for the higher robustness. When more missing data are added to the data set the outlier identification power decreases gradually with BEM being the best, followed by MCD and TRC. Most of the RSEQ methods perform still well with 20% of missing data but their performance declines with 30%. Only BEM and TRC cope with 40% of missing data - all other procedures break down, failing to identify 30 or more percent of the outliers. However, in the case of BEM and TRC the price is a higher non-outlier error rate.

4.2 Structural business statistics data set

The Austrian structural business statistics data (SBS) from 2006 covers NACE sections C-K for enterprises with 20 or more employees (NACE C-F) or above a specified turnover (NACE G-K) (Eurostat, 2008). For these enterprises more than 90 variables are available. Only limited administrative information is available for enterprises below these thresholds and for those which belong to other economic branches (NACE sections L to O). The raw unedited data consist of 21669 observations which include 3891 missing values. The small enterprises (about 230.000) are not present in the raw data and therefore not considered for our purpose. The highest amount of missing values is presented in the variable `EMP` (number of employees). The higher the amount of white-collar employees (`B31`) the higher the probability of missingness in variable `EMP`. This can be easily derived by applying the R package **VIM** (Templ and Filzmoser, 2008) for interactive exploration of the missing values mechanism.

Rather than performing the outlier detection with the complete data set, it is often reasonable to identify outliers in subgroups of the data. A detailed data analysis of the raw data has shown that ideal subgroups are based on NACE 4-digits level data (see also Dinges and Haitzmann, 2009). Broader categories imply that the data consist of different kinds of sub-populations with different characteristics. For the sake of this study we have chosen the NACE 4-digits level *47.71 - "Retail sale of clothing in specialized stores"* - (Nace Rev. 1.11 *52.42*, ISIC Rev.4 *4771*). This is a typical NACE 4-digits level data set and consists of 199 observations with 7 missing values and various outliers. In order to be able to apply outlier detection reasonably, specific variables were chosen, namely the ten variables shown in Table 2. We are not using the original data set for outlier detection, but only synthetic data sets. This is a standard procedure in official statistics, and it is a way to avoid the disclosure of information. Thus, one generates synthetic (but close-to-reality) data sets from the original data, which can then be used for simulation. In order to generate outlier-free synthetic data sets, the structure of the main bulk of the data is estimated using a robust covariance estimation with the MCD estimator applied to the log-transformed raw data (without missing values). The logarithmic scale was chosen to make the data distributions more symmetric. It should be noted that some variables (number of employees in different categories B31, B41 and B23) are discrete valued but their range goes from 0 to approximately 20000 which makes it reasonable to assume a continuous (log-normal) distribution. Synthetic data are generated using the covariance structure of these "good" data points. Afterwards, missing values are included in the synthetic data sets, where the missing values are generated using

Table 2 Variables of the Austrian SBS data which are considered in our simulation study.

1	TURNOVER	Total turnover
2	B31	Number of white-collar employees
3	B41	Number of blue-collar workers
4	B23	Part-time employees
5	A1	Wages
6	A2	Salaries
7	A6	Supply of trade goods for resale
8	A25	Intermediate inputs
9	E2	Revenues from retail sales
10	EMP	Number of employees

the information of the raw data. This especially refers to the probability of missing values in variable EMP which increases with increasing values of variable B31, as seen from the original data. As for the raw data, missing values occur in two variables (EMP and A25).

In our tests, different outlier specifications are generated. In the next section will be presented the results for shifted outliers with the same covariance structure. Outliers of moderate size are generated with mean $\boldsymbol{\mu}' = \boldsymbol{\mu} + (3.5, 2, -1, 0, 0, 1, 0, -1, 1.5, 0)^t = (\mu_{TURNOVER} + 3.5, \mu_{B31} + 2, \mu_{B41} - 1, \mu_{B23}, \mu_{A1}, \mu_{A2} + 1, \mu_{A6}, \mu_{A25} - 1, \mu_{E2} + 1.5, \mu_{EMP})^t$, with $\boldsymbol{\mu}$ estimated using the MCD estimator. This robust estimation was performed on the raw log-transformed data set taking into account only the complete observations. The data shifted in this way look close to reality when comparing the raw data and the synthetic data with the help of exploratory tools. In Figure 4 a parallel coordinate plot (which is a visualization technique for multivariate data used to plot individual data elements across many dimensions, see Wegman, 1990; Venables and Ripley, 2003) of one generated data set with 10% of outliers is shown. The outliers are drawn with darker lines. When varying the amount of outliers and/or missing values in a simulation study it is necessary to fully control the rate of missing values and/or outliers. This also means that no outliers should be replaced by missing values, otherwise the outliers generated beforehand are then masked by missing values.

It is clearly seen that the variable EMP has the highest amount of missing values but at the same time it is not contaminated (we do not shift its mean when generating the simulated data) which corresponds to the real data we mimic. However a question could arise what would happen if outliers in this variable were also present. To answer this question we performed an additional experiment following exactly the same conditions described above but the variable EMP was shifted by 1.5. The performance of all estimators on all comparison criteria remained stable (changed by less than 1%).

If the sampling context is taken into account, the data are assumed to be a random sample s of size n from a finite population U with size N and the sample is drawn with the sample design $p(s)$ which defines the inclusion probabilities and the corresponding sampling weights. In business statistics often the inclusion probabilities and the corresponding sampling weights are taken equal in each strata (e.g. depending on the employee size class - large enterprises have inclusion probability one while small enterprises have a lower inclusion probability). Since outlier detection should be done in reasonable subgroups these subgroups should ideally not differ from the stratification used for sampling (assuming simple random stratified design with equal inclusion probabilities in each strata). For unequal

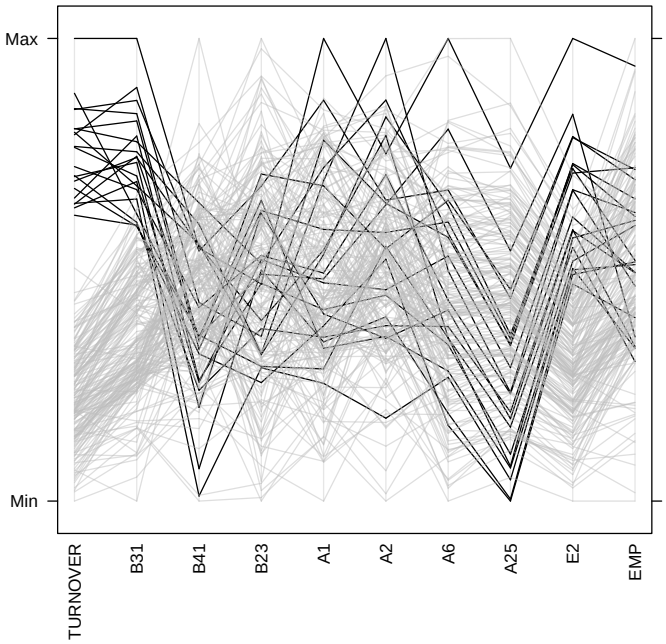

Fig. 4 Parallel coordinate plot of a typical data set with 10% of outliers (represented by the dark lines).

inclusion probabilities in a stratum the algorithms EA ,TRC and BEM (Béguin and Hulliger, 2004, 2008) might be the preferable choice. However, Béguin and Hulliger (2008), page 102 advice to use also a non-weighted version of the algorithms since it is possible that outliers are masked by large sampling weights which could dominate the model estimate. Furthermore not all algorithms which we are considering in this study are adapted to handling sampling weights, therefore we assume a design with equal inclusion probabilities and the EA, TRC and BEM will be run with all sampling weights set equal to one.

4.3 Simulation Experiments

Based on the Austrian Structural Business Statistics (SBS) data set we perform two experiments. In the first case we fix the fraction of outliers to 0.1 and vary the missing rates from 0.0 to 0.3 by 0.025. Alternatively, in the second case we fix the missing rate to 0.1 and vary the fraction of outliers from 0.0 to 0.25 by 0.025. For each configuration $m = 400$ data sets are generated and for each of them the outliers are identified using all available procedures. As in the previous example two measures are calculated: (i) the average percentage of outliers that were not identified (FN) and (ii) the average percentage of non-outliers that were classified as outliers (FP). The methods based on robust sequential imputation (RSEQ) identify all outliers in all configurations as the methods based on normal imputation do, so in terms of outlier error rate there is no space for improvement. The non-outlier error rate

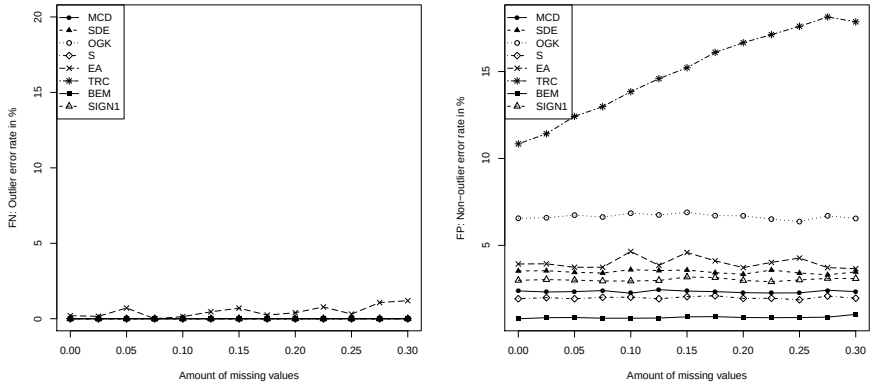

Fig. 5 Average outlier error rate (left) and average non-outlier error rate (right) for fixed fraction of 10% outliers and varying percentage of missings.

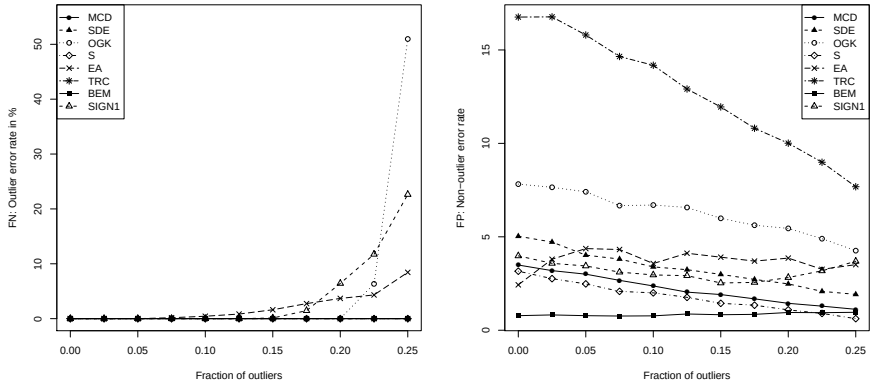

Fig. 6 Average outlier error rate (left) and average non-outlier error rate (right) for fixed fraction of 10% missing values and varying fraction of outliers.

of these methods is slightly higher than that of the corresponding methods based on normal imputation as we already noted in the previous example. Thus the results of the methods based on robust sequential imputation are quite similar to the corresponding methods based on normal imputation and therefore we do not show them in the following presentation (but these results are available from the authors upon request). The results for the rest of the estimators are presented in Figure 5 and Figure 6. For an outlier fraction of 10% all estimators perform excellent in terms of outlier error rate (FN) and identify all outliers independently of the percentage of missing values as seen from the left panel of Figure 5. An exception is EA which in some cases misses one or two outliers (below 2%). The average percentage of non-outliers that were declared outliers (FP) differ and BEM performs best, followed closely by S, MCD, SIGN1 and SDE (below 3%). With less than 5% EA comes next and

OGK declares somewhat more than 6% of regular observations as outliers, independently of the proportion of missings. TRC performs worst with the average non-outlier error rate increasing with increase of the fraction of missing values.

When the percentage of missing values is kept fixed at 10% and the fraction of outliers is varied, some of the estimators break down quite early - SIGN1 is the first followed by OGK. EA remains below 6%. The other estimators cope with all investigated fractions of outliers without any problems. In terms of the non-outlier error rate BEM performs best followed by S, MCD, SIGN1, SDE and EA (less than 5%). OGK comes next with error rate between 5% and 8% and TRC is again last.

The bushfire example with its severe outliers and with added high percentage of missing values was a serious difficulty for most of the methods which broke down by 40% missing data. On the other hand the SBS data was easier to handle in both cases (10% outliers and high percentage of missings or only 10% of missing data and high percentage of outliers) by most of the algorithms. In general in both examples the best performer were the BEM and MCD.

5 Software Availability

The complete data algorithms discussed in this paper are available in the R packages **robustbase**, **rrcov**, **mvoutlier** and their counterparts for incomplete data are implemented in the package **rrcovNA**. These packages are available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org under the GNU General Public License. The three algorithms from the EUREDIT project (`TRC`, `EA` and `BEM`), as already mentioned, were kindly provided by the authors.

6 Conclusions and outlook

In this paper we tested several approaches for identifying outliers in data sets including missing values. Two aspects seem to be of major importance: the computation time and the accuracy of the outlier detection method. For the latter we used the fraction of false negatives - outliers that were not identified - and the fraction of false positives - non-outliers that were declared as outliers. Overall, the preference of the ”optimal” algorithm will depend on the data structure. Most of the used algorithms are designed for elliptically symmetric distribution, and their behaviour could be completely different in case of very skewed distributions. The simulations and examples have shown that the BACON-EEM algorithm is generally a good choice in terms of precision. If computation time is an important issue (because the data set is very large and of high dimension), the algorithm based on the MCD estimator might be the preferable choice. The missing data mechanism will in general affect the performance of the outlier detection methods. For example, if algorithms based on imputation are used (Section 2.2), the quality of the imputed values can be poor in case of missing not at random, and this will also have consequences for the subsequent outlier detection algorithm.

Structural business statistics data include more difficulties for outlier detection algorithms, like zero values, binary or categorical variables, skewness of the data as well as possible

complex sampling designs of the survey. In our future work we will concentrate on these issues.

Acknowledgements The careful review and constructive comments of the associate editor and the anonymous reviewers helped us to substantially improve the manuscript. We thank B. Hulliger and C. Béguin for making their R code available to us. The contribution of Matthias Templ was partly funded by the European Union (represented by the European Commission) within the 7^{th} framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI (Advanced Methodology for European Laeken Indicators), Grant Agreement No. 217322). Visit `http://ameli.surveystatistics.net` for more information.

References

Barnett V, Lewis T (1994) Outliers in Statistical Data. John Wiley & Sons

Béguin C, Hulliger B (2004) Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 127(2):275–294

Béguin C, Hulliger B (2008) The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. Survey Methodology 34(1):91–103

Billor N, Hadi AS, Vellemann PF (2000) Bacon: Blocked adaptative computationally-efficient outlier nominators. Computational Statistics & Data Analysis 34(3):279–298

Campbell NA (1989) Bushfire maping using NOAA AVHRR data. Technical report, CSIRO

Cerioli A, Riani M, Atkinson AC (2009) Controlling the size of multivariate outlier tests with the MCD estimator of scatter. Statistics and Computing 19(3):341–353

Chambers RL (1986) Outlier robust finite population estimation. Journal of the American Statistical Association 81:1063–1069

Copt S, Victoria-Feser MP (2004) Fast algorithms for computing high breakdown covariance matrices with missing data. In: Hubert M, Pison G, Struyf A, Van Aelst S (eds) Theory and Applications of Recent Robust Methods, Statistics for Industry and Technology Series, Birkhauser Verlag, Basel

Croux C, Haesbroeck G (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis 71:161–190

De Waal T (2003) Processing of erroneous and unsafe data. Ph.d. thesis, Erasmus University, Rotterdam

De Waal T (2009) Statistical data editing. In: Peffermann D, Rao C (eds) Handbook of Statistics 29A. Sample Surveys: Design, Methods and Applications, Elsevier B. V., Amsterdam, The Netherlands, pp 187–214

Dempster AP, Laird MN, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 39:1–22

Dinges G, Haitzmann M (2009) Modellbasierte Ergänzung der Konjunkturstatistik im Produzierenden Bereich; Darstellung der statistischen Grundgesamtheit im Produzierenden Bereich. Statistische Nachrichten 9:1153–1166, URL `www.stat.at/web_de/downloads/methodik/kjp.pdf`

Donoho DL (1982) Breakdown properties of multivariate location estimators. Tech. rep., Harvard University, Boston, URL `http://www-stat.stanford.edu/~donoho/Reports/Oldies/BPMLE.pdf`

EUREDIT Project (2004) Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project, vol 1 and 2. EUREDIT consortium, URL http://www.cs.york.ac.uk/euredit/results/results.html

Eurostat (2008) NACE Rev. 2. Statistical classification of economic activites in the European Community. Eurostat, Methodologies and Working papers, ISBN 978-92-79-04741-1

Fellegi I, Holt D (1976) A systematic approach to automatic edit and imputation. Journal of the American Statistical Association 71:17–35

Filzmoser P, Garrett RG, Reimann C (2005) Multivariate outlier detection in exploration geochemistry. Computers & Geosciences 31:579–587

Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Computational Statistics & Data Analysis 52(3):1694–1711

Franklin S, Brodeur M (1997) A practical application of a robust multivariate outlier detection method. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 186–191, URL http://www.amstat.org/sections/srms/proceedings

Franklin S, Brodeur M, Thomas S (2000) Robust multivariate outlier detection using Mahalanobis' distance and Stahel-Donoho estimators. In: ICES - II, International Conference on Establishment Surveys - II

Granquist L (1990) A review of some macro-editing methods for rationalizing the editing process. In: Proceedings of the Statistics Canada Symposium, Ottawa, Canada, pp 225–234

Granquist L (1997) The new view on editing. International Statistical Review 65:381–387

Hampel FR, Ronchetti EM, J RP, Stahel WA (1986) Robust Statistics. The Approach Based on Infuence Functions. John Wiley & Sons

Hardin J, Rocke DM (2005) The distribution of robust distances. Journal of Computational and Graphical Statistics 14:910–927

Hidiroglou MA, Lavallée P (2009) Sampling and estimation in business surveys. In: Pefferman D, Rao C (eds) Handbook of Statistics 29A. Sample Surveys: Design, Methods and Applications, Elsevier B. V., Amsterdam, The Netherlands, pp 441–470

Huber PJ (1981) Robust Statistics. John Wiley & Sons

Hubert M, Rousseeuw PJ, Vanden Branden K (2005) Robpca: A new approach to robust principal component analysis. Technometrics 47:64–79

Hubert M, Rousseeuw PJ, van Aelst S (2008) High-breakdown robust multivariate methods. Statistical Science 23:92–119

Johnson RA, Wichern DW (2002) Applied Multivariate Statistical Analysis. Prentice Hall, International, fifth edition

Lawrence D, McKenzie R (2000) The general application of significance editing. Journal of Official Statistics 16:243–253

Little RJA, Rubin DB (1987) Statistical Analysis with Missing Data. John Wiley & Sons, New York

Little RJA, Smith PJ (1987) Editing and imputation for quantitative data. Journal of the American Statistical Association 82:58–69

Lopuhaä HP (1999) Asymptotics of reweighted estimators of multivariate location and scatter. The Annals of Statistics 27:1638–1665

Lopuhaä HP, Rousseeuw PJ (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics 19:229–248

Luzi O, De Waal T, Hulliger B, Di Zio M, Pannekoek J, Kilchmann D, Guarnera U, Hoogland J, Manzari A, Tempelman C (2007) Recommended practices for editing and imputation in cross-sectional business surveys. Report

Maronna RA, Yohai VJ (1995) The behaviour of the Stahel-Donoho robust multivariate estimator. Journal of the American Statistical Association 90:330–341

Maronna RA, Zamar RH (2002) Robust estimation of location and dispersion for high-dimensional datasets. Technometrics 44:307–317

Maronna RA, Martin D, Yohai V (2006) Robust Statistics: Theory and Methods. John Wiley & Sons, New York

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/, ISBN 3-900051-07-0

Riani M, Atkinson AC, Cerioli A (2009) Finding an unknown number of multivariate outliers. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71(2):447–466

Rousseeuw PJ, Leroy AM (1987) Robust Regression and Outlier Detection. John Wiley & Sons, New York

Rousseeuw PJ, van Zomeren BC (1990) Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85:633–651

Rubin DB (1993) Discussion: Statistical disclosure limitation. Journal of Official Statistics 9:462–468

Schafer J (1997) Analysis of Incomplete Multivariate Data. Chapman and Hall, London

Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. Psychological Methods 7:147–177

Stahel WA (1981a) Breakdown of covariance estimators. Research Report 31, ETH Zurich, Fachgruppe für Statistik

Stahel WA (1981b) Robuste schätzungen: Infinitesimale optimalität und schätzungen von kovarianzmatrizen. Ph.d. thesis no. 6881, Swiss Federal Institute of Technology (ETH), Zürich, URL http://e-collection.ethbib.ethz.ch/view/eth:21890

Templ M, Filzmoser P (2008) Visualization of missing values using the R-package VIM. Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology

Todorov V, Filzmoser P (2009) An object oriented framework for robust multivariate analysis. Journal of Statistical Software 32(3):1–47, URL http://www.jstatsoft.org/v32/i03/

Vanden Branden K, Verboven S (2009) Robust data imputation. Computational Biology and Chemistry 33(1):7–13

Venables WN, Ripley BD (2003) Modern Applied Statistics with S, 4th edn. Springer

Verboven S, Vanden Branden K, Goos P (2007) Sequential imputation for missing values. Computational Biology and Chemistry 31(5-6):320–327

Wegman E (1990) Hyperdimensional data analysis using parallel coordinates. Journal of the American Statistical Association 85:664–675