WP. 40
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchatel, Switzerland, 5-7 October 2009)

Topic (viii): Selective and macro editing

OUTLIER DETECTION IN SURVEY DATA USING ROBUST METHODS

Invited paper

Prepared by Valentin Todorov (UNIDO), Matthias Templ (Statistics Austria) and
Peter Filzmoser (Vienna University of Technology)

I. ABSTRACT

Many different methods for macro 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 Bèguin 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, macro editing is essential to check the multivariate data for possible data problems and it is not deterministic like usual editing using certain rules/constraints. First we review the available multivariate outlier detection methods. 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 will be soon available under GPL at `CRAN`.

II. INTRODUCTION

1. 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 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 side, if we know the outliers we can 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 after that can be treated in the traditional way. The concept of representative and non-representative outliers (Chambers, 1986) will not be considered.

2. 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 [Little and Rubin, 1987], i.e. the missing values depend only on observed data. 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 way that the units with large size are selected with high probability (often selected in "take-all" strata) while the small sized ones are sampled with low probability. 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.

3. Maybe these challenges are the reason that multivariate outlier detection methods are rarely applied to sample survey data. One of the exceptions is Statistics Canada [Franklin et al., 2000] which is based on a robust version of principal component analysis and Stahel-Donoho estimator of multivariate location and covariance matrix [Stahel, 1981, Donoho, 1982]. Several robust methods are investigated in the EUREDIT project and are further developed by Bèguin and Hulliger [2004, 2008] which will be considered also in the present study.

4. The rest of the paper is organized as follows. In Section III 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. Section IV presents an example based on an well-known complete data set with inserted simulated missing data and continues with a simulation study which mimics a real sample survey data set from *Statistics Austria*. Section V presents the availability of the software discussed so far and Section VI concludes.

III. ALGORITHMS FOR OUTLIER DETECTION

5. A general framework for multivariate outlier identification in a p-dimensional data set $\boldsymbol{X}$ is to compute some measure of the distance of a particular point from the center of the data and declare outliers as 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 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 $\boldsymbol{X}$:

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

6. 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 MLE estimates. 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 of Mahalanobis distance.

7. 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 is suitable to use Maronna et al. [2006]. Most of these methods are implemented in the R statistical environment and are available in the object-oriented framework for robust multivariate analysis Todorov and Filzmoser [2009].

8. 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 $D_0 = \chi^2_p(0.975)$, assuming that $\boldsymbol{X}$ follows multivariate normal distribution. Then 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. Maronna and Zamar [2002] propose to use a transformation of the cutoff value which should help the distribution of the robust distances RD_i to resemble χ^2 for non-normal original data:

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

A. HANDLING OF INCOMPLETE DATA

9. 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. 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 call this procedure ER-estimator. Little and Smith [1987] propose to use as a starting point for the ER algorithm an MLE estimate 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 is 0 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.

10. *Normal imputation followed by HBDP estimation*
A straightforward strategy for adapting estimators of location and covariance to work with missing data is to perform one step of Gaussian (non-robust) imputation and then run any of the above described algorithms, like for example MCD, OGK, S and Stahel-Dooho (SDE) on the already complete data. This class of methods we call e.g. PM-MCD (poor man's MCD), etc. In this way we can adapt also projection based algorithms like SIGN1 [Filzmoser et al., 2008]. The next step after estimating reliably the location and covariance matrix 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 $\boldsymbol{C}_{ooi}$ as the parts of $\boldsymbol{T}$ and $\boldsymbol{C}$ which correspond to the observed part of $\boldsymbol{x}_i$. Then

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

is the robust distance computed only from the observed part of $\boldsymbol{x}_i$. If $\boldsymbol{x}_i$ is uncontaminated, the data are multivariate normal distributed and missing values are missing at random, then the 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].

11. *Transformed Rank Correlation (TRC).*
This is one of the algorithms proposed by Bèguin 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 covariance matrix. The initial matrix is calculated using bivariate Spearman Rank correlations 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.

12. *Epidemic Algorithm (EA).*
The second algorithm proposed by Bèguin 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 not at all infected are considered outliers. A disadvantage of this method is the number of tuning parameters which have to be set in order to attain best performance.

13. *BACON-EEM (BEM).*
The third algorithm [Bèguin 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.

14. The last three algorithms - TRC, EA and BEM take into account also the sampling weights.

B. COMPUTATION TIMES

15. 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 constructed. Contaminated data sets with different size and dimensions, varying from (100,5) to (50000,30) were generated. The data follow the model of *shift outliers*, i.e. all observations come from multivariate normal distribution $N_p(\mathbf{0},\mathbf{I}_p)$, but the "bad" ones have their mean shifted by $b = 10$. The percentage of good observations is over 50%. Then 20% missing values created with a simple data missing completely at random (MCAR) mechanism were added to the generated data sets. Each test run was repeated a number of times (100) to have consistent error on the timing.

16. The experiment was performed on a 3Ghz PC with 2Gb Memory running Windows XP Professional. All computations were performed in R 2.9.0. The MCD, OGK, S and SDE algorithms from `rrcov` 0.5.2 were used and the imputation was performed by function `imp.norm()` from package `norm`. The R code for the EA, TRC and BEM algorithms was provided by the authors. Figure 1 displays graphically the results of the experiment in logarithmic scale. OGK is fastest, followed closely by MCD. 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.

IV. EXAMPLES AND EXPERIMENTS

17. In this section we illustrate the behavior of the considered algorithms in the presence of missing values on one artificial example - a complete real data set in which we introduce missing data with varying percentage. Then the algorithms are compared in a simulation study based on real data sets.

A. BUSHFIRE DATA

18. As a first example we consider the `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. The classical outlier detection methods based on the sample mean and covariance matrix do not recognize any outlier but most of the robust high breakdown point methods identify the outlying clusters 32-38 and 7-11 and to a lesser extend the single outlying observation 31. Observations 29 and 30 lie on the boundary. Similarly as Bèguin and Hulliger [2004] we simulate item-non-response by applying a simple MCAR mechanism which generates $q = \{0.0, 0.1, 0.2, 0.3, 0.4\}$ fraction of missing values. For each of these five cases we generate $m = 100$ 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. The problem with EA and TRC could be the number of tuning parameters which we may not have been set properly. The non-outlier error rate for the complete data set can for most of the procedures (MCD, OGK, S, SDE) be explained by the observations lying on the border - 12, 28, 29

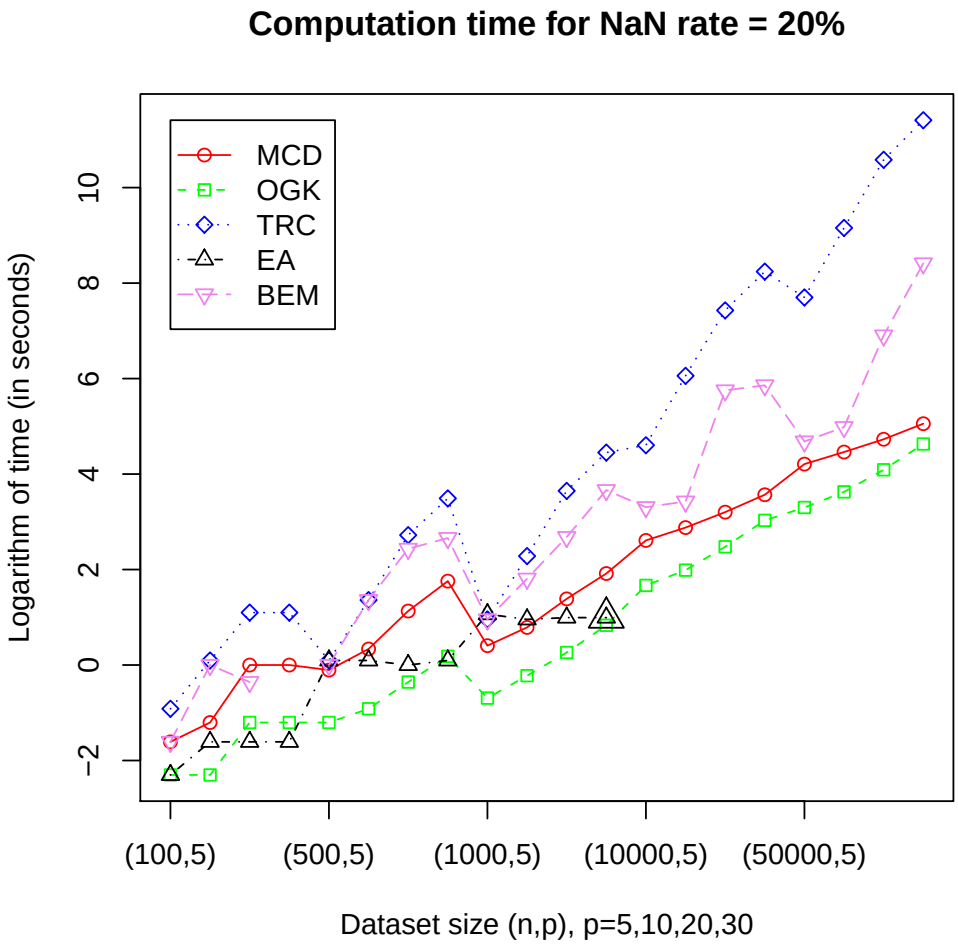

FIGURE 1. Comparing the computation time of the algorithms in the presence of missing data.

and 30 - which in many cases are declared outliers. With 10% of missing data MCD performs best followed by BEM, SIGN1 and SDE. In terms of the non-outlier error rate, S is best followed by SDE and BEM. When more missing data are added to the data set the outlier identification power decreases gradually with BEM being the best, followed by MCD. Only BEM copes with 40% of missing data - all other procedures break down, failing to identify 30 and more percent of the outliers.

B. STRUCTURAL BUSINESS STATISTICS DATA SET

19. The Austrian structural business statistics data (SBS) from 2006 covers NACE sections C-K for enterprizes with 20 or more employees (NACE C-F) or above a specified turnover (NACE G-K) [Eurostat, 2008]. For these enterprizes more than 90 variables are available. Only limited administrative information is available for enterprizes 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 includes 3891 missing values. Small enterprizes are imputed (about 233429 enterprizes) using robust regression models by including information from administrative data sources. The data source used in this work is based on the raw unedited data set. The highest amount of missing values is

	Outlier error rate (FN)					Non-outlier error rate (FP)				
	0	10	20	30	40	0	10	20	30	40
MCD	0.00	3.46	9.46	18.15	31.15	12.00	4.28	2.84	2.44	2.32
OGK	0.00	8.46	18.77	36.00	37.77	16.00	5.72	4.88	3.68	2.72
S	0.00	9.23	22.15	27.08	43.23	0.00	3.96	4.08	3.92	2.92
EA	67.83	74.25	76.50	86.92	87.83	1.88	1.27	1.58	1.08	1.73
TRC	25.00	16.92	22.58	21.75	21.08	3.85	10.81	10.96	8.42	7.54
BEM	0.00	4.46	6.23	8.31	8.62	4.00	9.56	9.12	12.32	16.52
SIGN1	0.00	5.15	19.00	35.23	45.38	20.00	13.08	8.00	5.68	5.32
SDE	0.00	6.23	16.15	25.23	45.38	1.72	3.28	1.68	1.72	1.60

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

TURNOVER	Total turnover
B31	Number of white-collar employees
B41	Number of blue-collar workers
B23	Part-time employees
EMP	Number of employees
A1	Wages
A2	Salaries
A6	Supply of trade goods for resale
A25	Intermediate inputs
E2	Revenues from retail sales

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

presented in the variable `EMP`. The higher the amount of white-collar employees (`B31`) the higher the probability of missing 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.

20. As mentioned above, outlier detection should be made in reasonable 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. 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. The correlations between these variables can be interpreted in a reasonable way. A systematic mechanism for missingness could rarely be detected, the probability of a value to be missing in EMP might depend slightly on other variables such as B31.

21. A synthetic data set is needed because of two reasons. Firstly, synthetic (but close-to-reality) data sets are needed for simulation. Secondly, the disclosure of information must be avoided. The structure of the main bulk of the data is estimated using a robust covariance estimation with the MCD estimator of the log-transformed raw data. Synthetic data are generated using the covariance structure of these "good" data points. Zeros are included in specific variables (and in specific combinations) in the data with equal fraction with respect of the structure of missing values of the raw data set. The moderate MAR situation is respected when generating missing values in the synthetic data file. Missing

values are generated using the information of the raw data. Therefore, the probability of missing in EMP slightly depends on B31 as well in the synthetic data set.

22. In our tests, different outlier specifications are generated. For the results presented in this paper we generate shifted outliers with the same covariance structure. Outliers of moderate size are generated with mean $\boldsymbol{\mu}' = (\mu_1 + 3.5, \mu_2 + 2, \mu_3 - 1, \mu_4, \mu_5, \mu_6 + 1, \mu_7, \mu_8 + 1, \mu_9, \mu_{10} - 1)^t$, with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{10})$ estimated from the raw data set using the MCD estimator. 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. The structure of the synthetic close-to-reality subset of the raw SBS data is shown in Figure 2. Missing values are colored in red. The darker a line in the left pabel of Figure 2 the higher the value of an observation. Missing values occur in two variables, A25 and EMP, as for the raw data set.

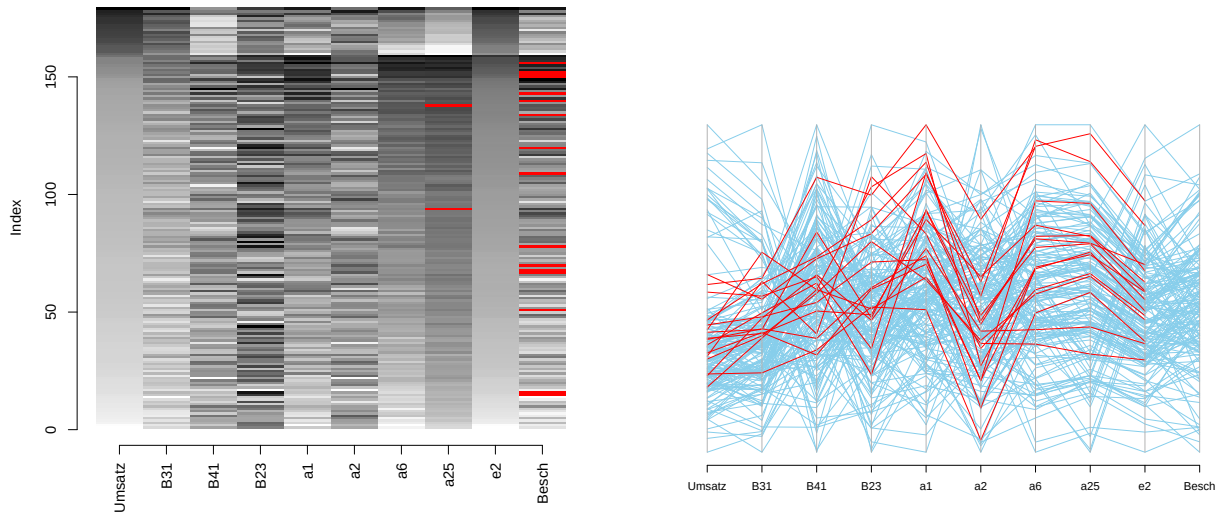

FIGURE 2. Missing value patterns analyzed with the R package `VIM` using the Matrix plot (left panel) and Parallel coordinate plot (right panel)

C. SIMULATION EXPERIMENT

23. Based on the 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 with a step 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 with a step 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). We exclude from the study the EA algorithm because of its erratic results as well as the Stahel-Donoho estimator because of its long computation time. The results for the rest of the estimators are shown in Figure 3 and Figure 4.

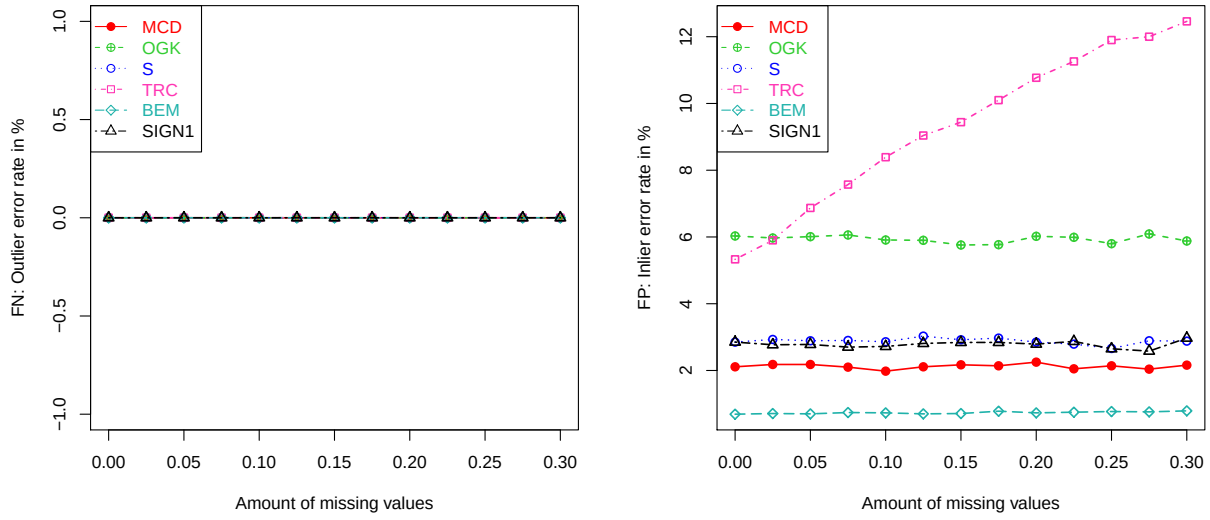

FIGURE 3. Average outlier error rate (left) and average non-outlier error rate (right) for fixed fraction of 10% outliers and varying percentage of missings.

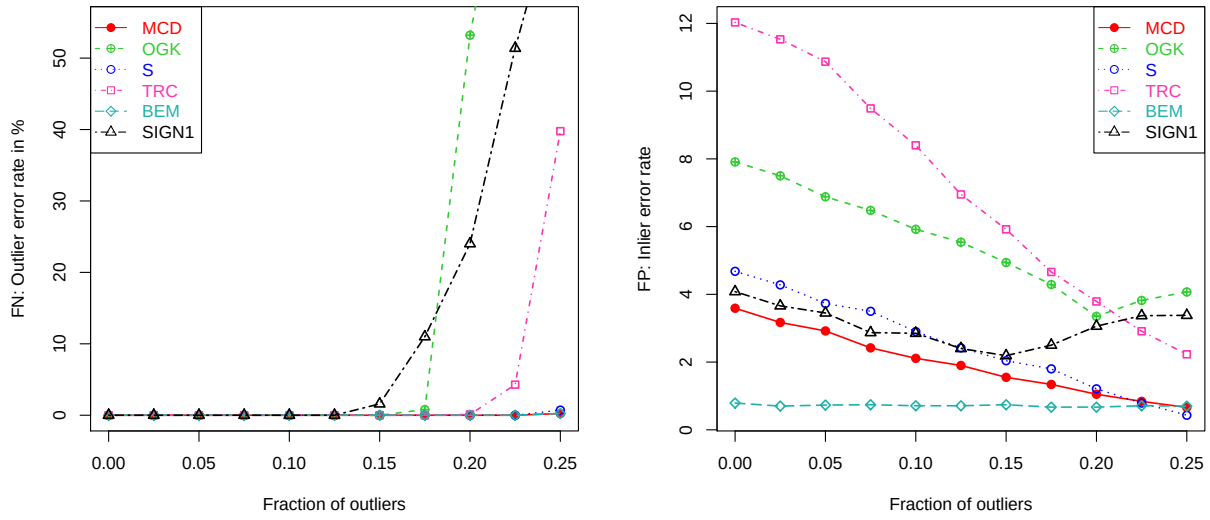

FIGURE 4. Average outlier error rate (left) and average non-outlier error rate (right) for fixed fraction of 10% missing values and varying fraction of outliers.

24. 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 3. The average percentage of non-outliers that were declared outliers (FP) differ and BEM performs best, followed closely by MCD, S and SIGN1 (below 3%). OGK declares about 6% of regular observations as outliers, independently of the proportion of missings, and TRC performs worst with the average non-outlier error rate increasing with increase of the fraction of missing values.

25. 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 and OGK are the first followed by TRC. BEM, MCD and S cope with all investigated fractions of outliers without any problems. In terms of the non-outlier error rate BEM performs best followed by MCD, SIGN1 and S. OGK comes next with error rate between 4% and 8% and TRC is again last.

V. AVAILABILITY

26. Most of the algorithms discussed in this paper are available in the R packages `robustbase`, `rrcov`, `mvoutlier` and the forthcoming `rrcovNA`.

VI. CONCLUSION AND OUTLOOK

27. In this paper we tested several approaches for identifying outliers in data sets including missing values. Two aspects seems 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 behavior 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.

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

ACKNOWLEDGEMENTS

The contribution of Matthias Templ was partly funded by the European Union (represented by the European Commission) within the 7th 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

V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, 1994.

Cèdric Bèguin and Beat Hulliger. 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, 2004.

Cèdric Bèguin and Beat Hulliger. The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. *Survey Methodology*, 34(1):91–103, 2008.

N. Billor, A. S. Hadi, and P. F. Vellemann. Bacon: Blocked adaptative computationally-efficient outlier nominators. *Computational Statistics and Data Analysis*, 34(3):279–298, 2000.

N. A. Campbell. Bushfire maping using NOAA AVHRR data. Technical report, CSIRO, 1989.

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

A. P. Dempster, M. N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–22, 1977.

D. L. Donoho. Breakdown properties of multivariate location estimators. Technical report, Harvard University, Boston, 1982.

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

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

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

F. R. Hampel, E. M. Ronchetti, Rousseeuw P. J., and W. A. Stahel. *Robust Statistics. The Approach Based on Infuence Functions*. John Wiley & Sons, 1986.

M. Hubert, P. J. Rousseeuw, and S. van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, 23:92–119, 2008.

R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, International, 2002. fifth edition.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1987.

R. J. A. Little and P. J. Smith. Editing and imputation for quantitative data. *Journal of the American Statistical Association*, 82:58–69, 1987.

R. A. Maronna and V. J. Yohai. The behavoiur of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90:330–341, 1995.

R. A. Maronna and R. H. Zamar. Robust estimation of location and dispersion for high-dimensional datasets. *Technometrics*, 44:307–317, 2002.

R. A. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, 2006.

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

W. A. Stahel. Breakdown of covariance estimators. Research Report 31, ETH Zurich, 1981. Fachgruppe für Statistik.

M. Templ and P. Filzmoser. Visualization of missing values using the R-package VIM. Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology, 2008. URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2008-1complete.pdf.

V. Todorov and P. Filzmoser. An object oriented framework for robust multivariate analysis. 2009. submitted for publication.