Detection and visualization of outliers in establishment surveys

V. Todorov[1]

[1]United Nations Industrial Development Organization, Vienna

uRos 2024, Athens, 27-29 December 2024

Outline

The Industrial Sector

In general, industrial statistics are statistics reflecting characteristics and economic activities of the units engaged in a class of industrial activities that are defined in terms of the *International Standard Industrial Classification of All Economic Activities* (ISIC)
(IRIS 2008)

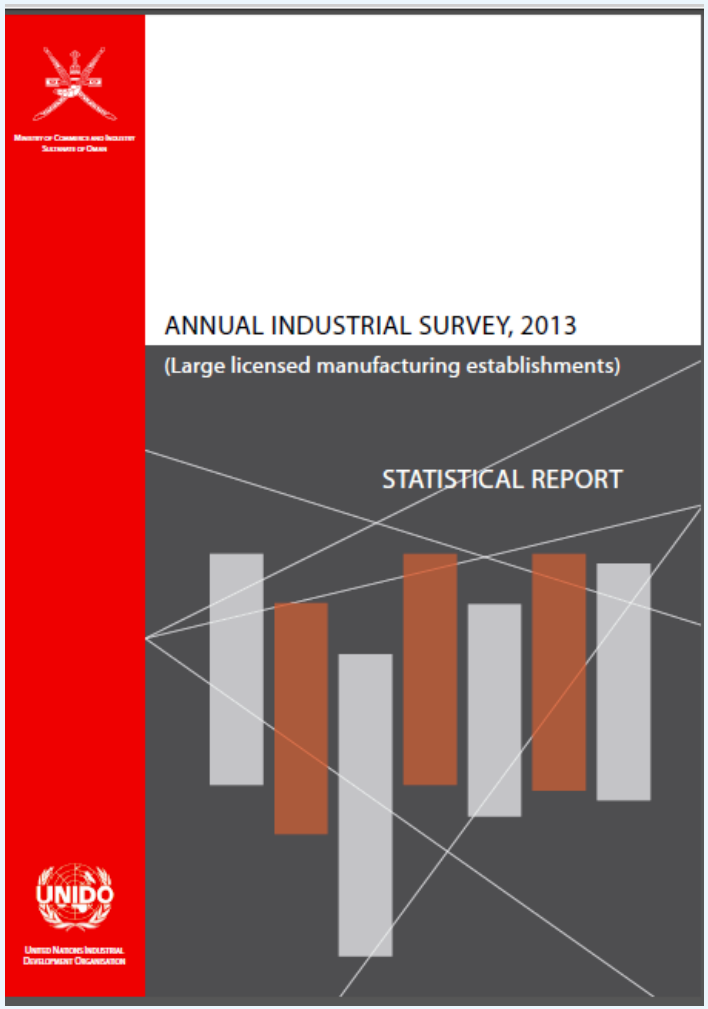
The industrial sector corresponds to:

ISIC Revision 3		ISIC Revision 4	
C	Mining and quarrying	B	Mining and quarrying
D	Manufacturing	C	Manufacturing
E	Electricity, gas and water supply	D	Electricity, gas, steam and air conditioning supply
		E	Water supply; sewerage, waste management and remediation activities

Divisions, Groups and Classes in Manufacturing: Example

- **Section C**—Manufacturing
- **Division 20**—Manufacture of chemicals and chemical products
- **Group 201**—Manufacture of basic chemicals, fertilizers and nitrogen compounds, plastics and synthetic rubber in primary forms
  - 2011—Manufacture of basic chemicals
  - 2012—Manufacture of fertilizers and nitrogen compounds
  - 2013—Manufacture of plastics and synthetic rubber in primary forms

Annual Industrial Survey in Oman: 2012-2019



- Data collected first in 2013-2014 with reference year 2012
- Data collection continued (at least) until 2019
- Statistical unit: establishment
- Scope: covered all large manufacturing establishments licensed with the Ministry of Commerce and Industry and operating in the Sultanate of Oman.
- Large=employing 10 or more persons engaged

- Initial frame: around 900 establishments
- Response - nearly 95% in terms of employment (701 establishments)
- The questionnaire:
  - Following the "International Recommendations for Industrial Statistics" of the United Nations
  - Activity classification: ISIC Revision 4
  - Product classification: CPC 2.0
  - 8 pages, more that 300 fields
  - Data entry: at MOCI, by trained staff, since 2019 online by the establishments
  - Strict formal validation

- Collected data still had some inconsistencies
  - Stem from insufficient understanding of the terms and concepts applied
  - There is a belief that information supplied would be transmitted to the income tax authorities:hiding information related to output and over-report on inputs.
  - A number of establishments are engaged in several equally important but dissimilar activities
  - Difficulties providing data on consumption of electricity, water and fuels separately; purchase of raw materials and sales from own production by main product

2. EMPLOYMENT
2.1 Number of persons engaged

Employment as at	Omani		Non-Omani		Total	
	Male	Female	Male	Female	Male	Female
2.11 30th June 2012	* 80	* 20	*	*	* 120	*
2.12 31st December 2012	*	*	*	*	*	*

2. EMPLOYMENT
2.1 Number of persons engaged

Employment as at	Omani		Non-Omani		Total	
	Male	Female	Male	Female	Male	Female
2.11 30th June 2012	* 80	* 20	*	*	* 12080	* 20
2.12 31st December 2012	*	*	*	*	*	*

4.1 Electricity generated, purchased and consumed

Description	Unit	Quantity	Value (R.O)
4.11 Electricity purchased	KWH	*	*
4.12 Electricity generated, if any	KWH	*	*
4.13 Electricity sold	KWH	*	*
4.14 Electricity consumed (= 4.11 + 4.12 - 4.13)	KWH	*	*

The data set selected for the example:

- One 4-digit ISIC Revision 4 class: 2395="Articles of concrete, cement and plaster", 84 establishments
- 6 variables

EMP	Number of employees
COMP	Gross wages and salaries paid to the employees
GO	Gross Output
IC	Intermediate consumption
EPV	Electricity purchased (value)
WPV	Water purchased (value)

- For 15 establishments there are missing values in at least one variable

Note: Value Added (VA) = GO - IC

Outliers in Sample Surveys

- "Rule based" approach - identification by data specific edit rules developed by subject matter experts followed by deletion and imputation ← strictly deterministic, ignore the probabilistic component, extremely labor intensive
- Univariate methods - favored for their simplicity. These are informal graphical methods like histograms, box plots, dot plots; quartile methods to create allowable range for the data; robust methods like medians, Winsorized means, etc.
- Multivariate methods - rarely used although most of the business surveys collect multivariate qualitative data

The Challenges

- The methods must be able to work with moderate to large data sets (hundreds of variables and tens of thousands of observations) - therefore we consider computational speed a very important criterion
- Survey data often contain missing values, therefore the methods must be able to work with incomplete data
- The survey data are often skewed - use appropriate transformations (Raymaekers and Rousseeuw, 2021; Atkinson *et al.*, 2024) or special robust methods for skewed data (Hubert *et al.*, 2008)
- The methods must be able to cope with the complex sample design of a survey using sampling weights

Difficult set-up: Large multivariate incomplete sample survey data

Outliers and Robustness

Outlier detection and Robust estimation are closely related

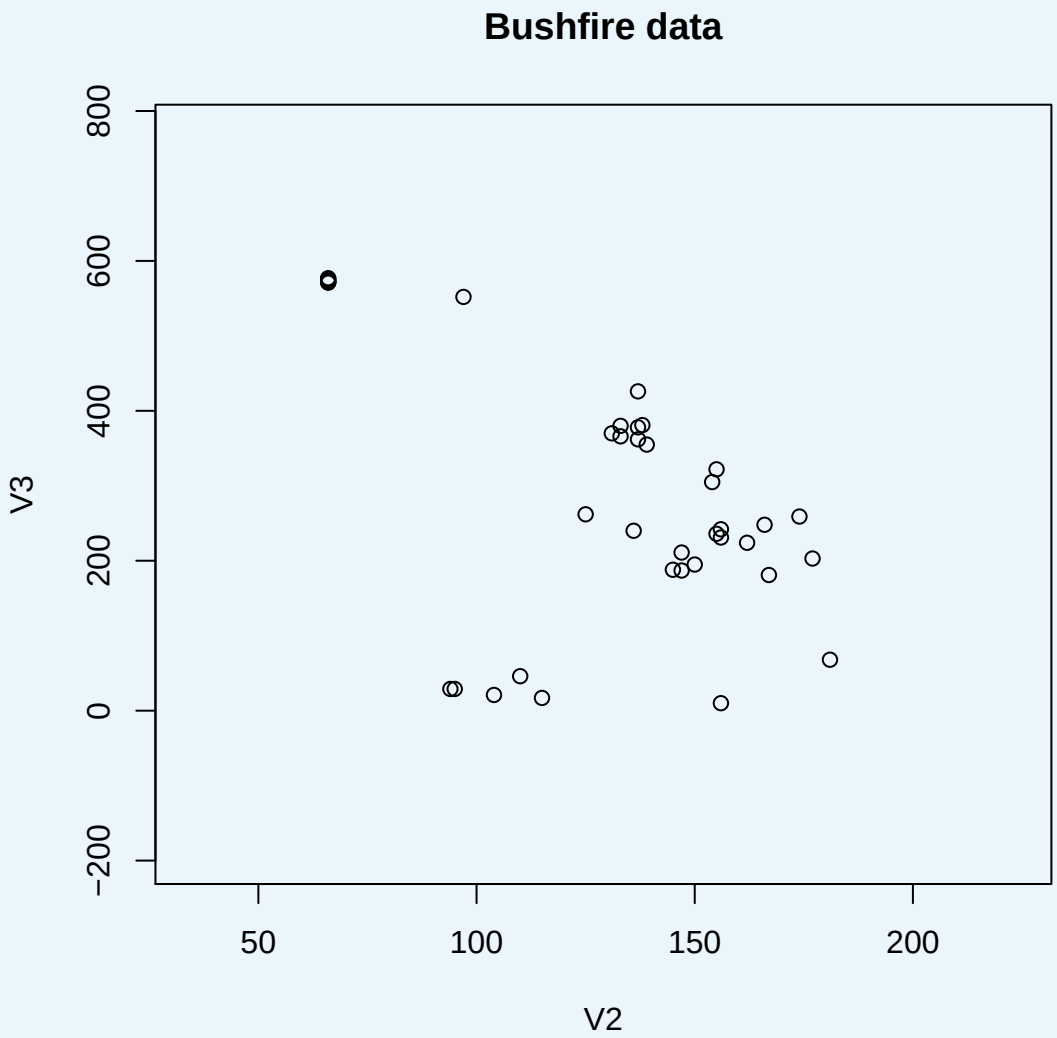
1. Robust estimation: find an estimate which is not influenced by the presence of outliers in the sample
2. Outlier detection: find all outliers, which could distort the estimate

- If we have a solution to the first problem we can identify the outliers using robust residuals or distances
- If we know the outliers we can remove or downweight them and use classical estimation methods
- For the purposes of official statistics the second approach is more appropriate

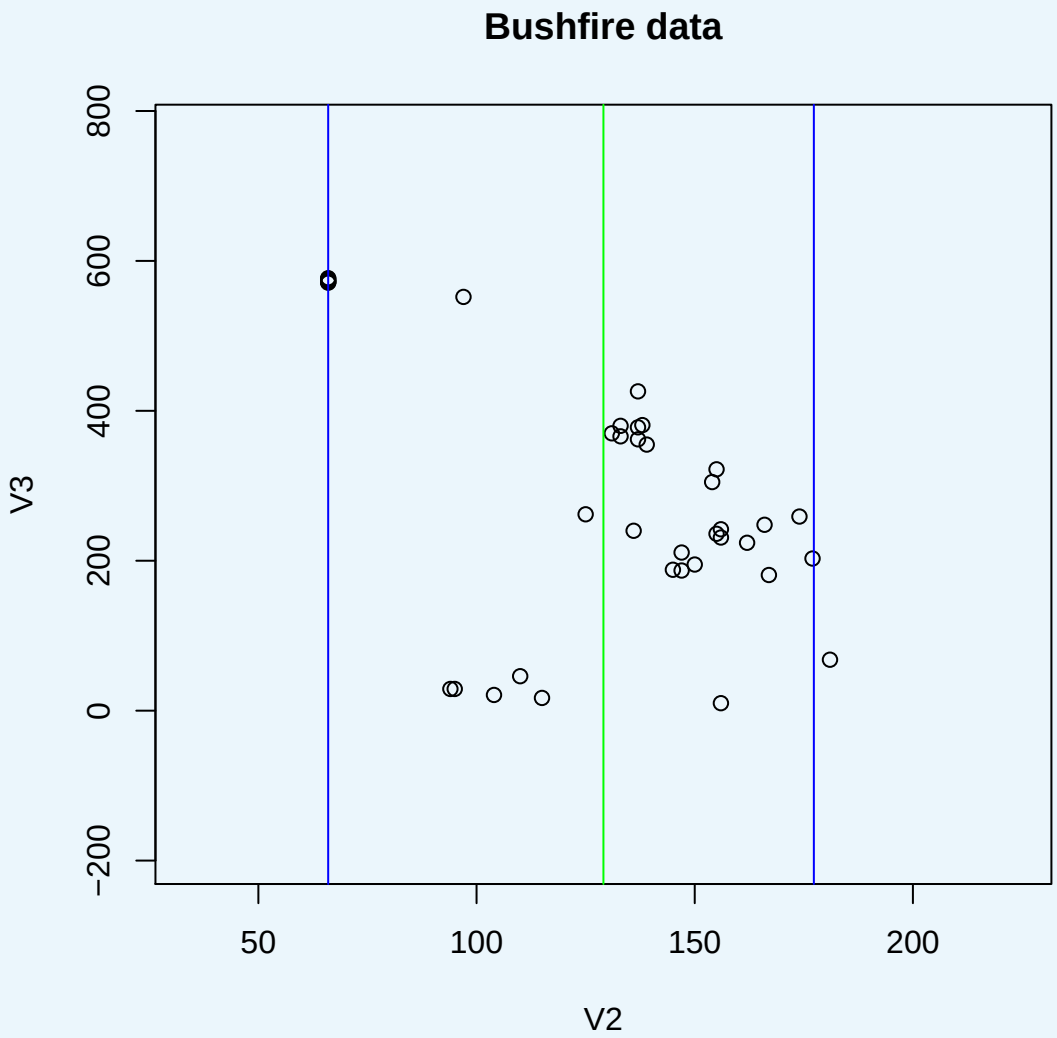
Example: Bushfire data

- A data set with 38 observations in 5 variables - Campbell (1989)
- Contains satellite measurements on five frequency bands, corresponding to each of 38 pixels
- Used to locate bushfire scars
- Very well studied (Maronna and Yohai, 1995; Maronna and Zamar, 2002)
- 12 clear outliers: 33-38, 32, 7-11; 12 and 13 are suspect
- Available in the `R` package `robustbase`

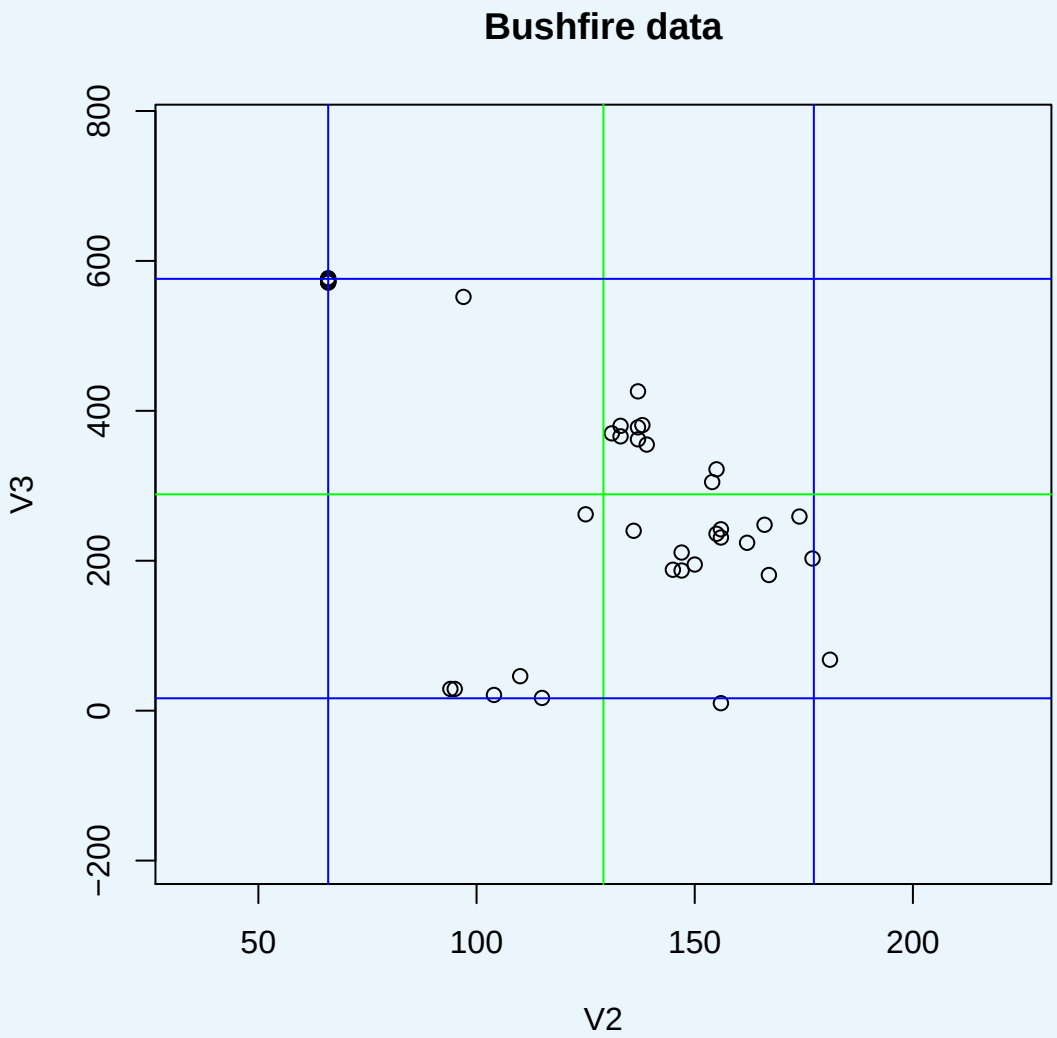
Example: Bushfire data



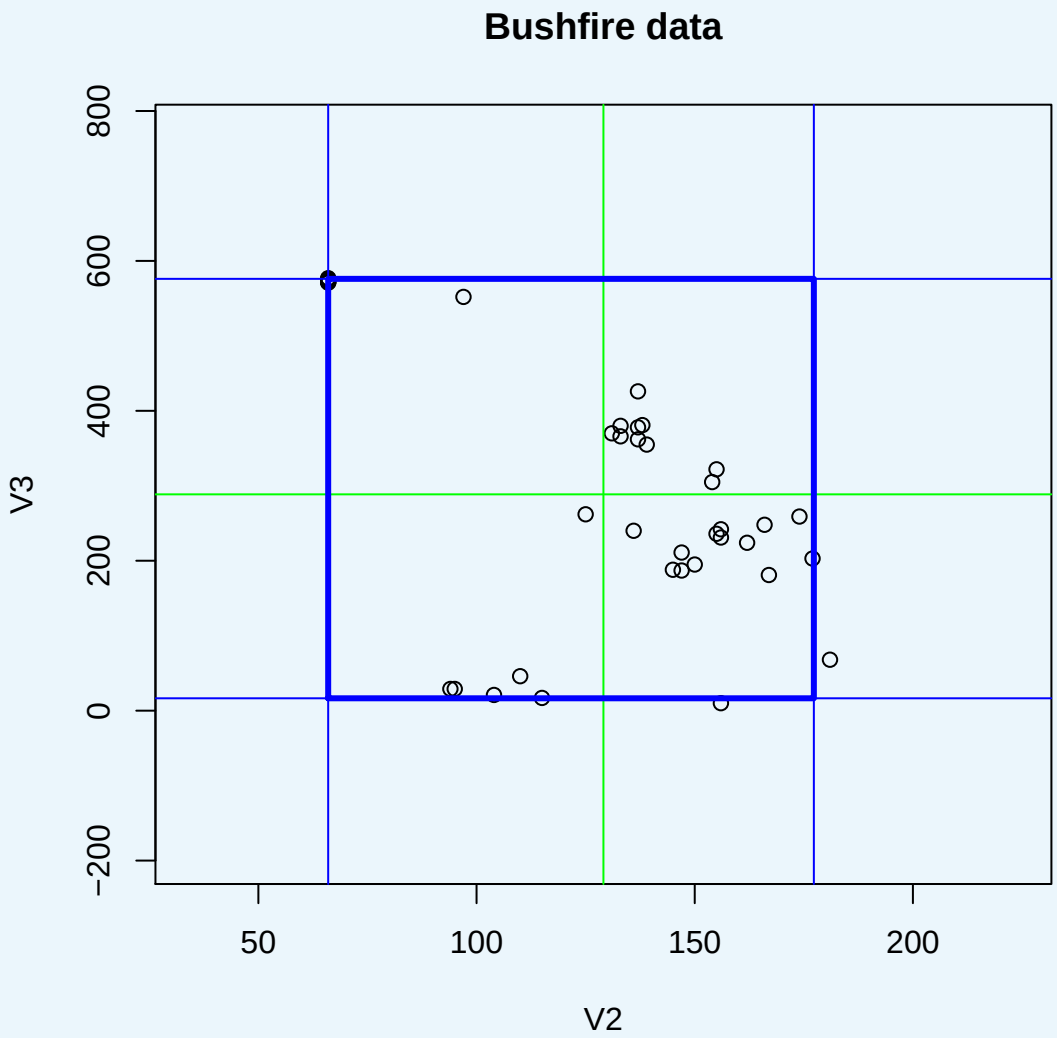
Example: Bushfire data



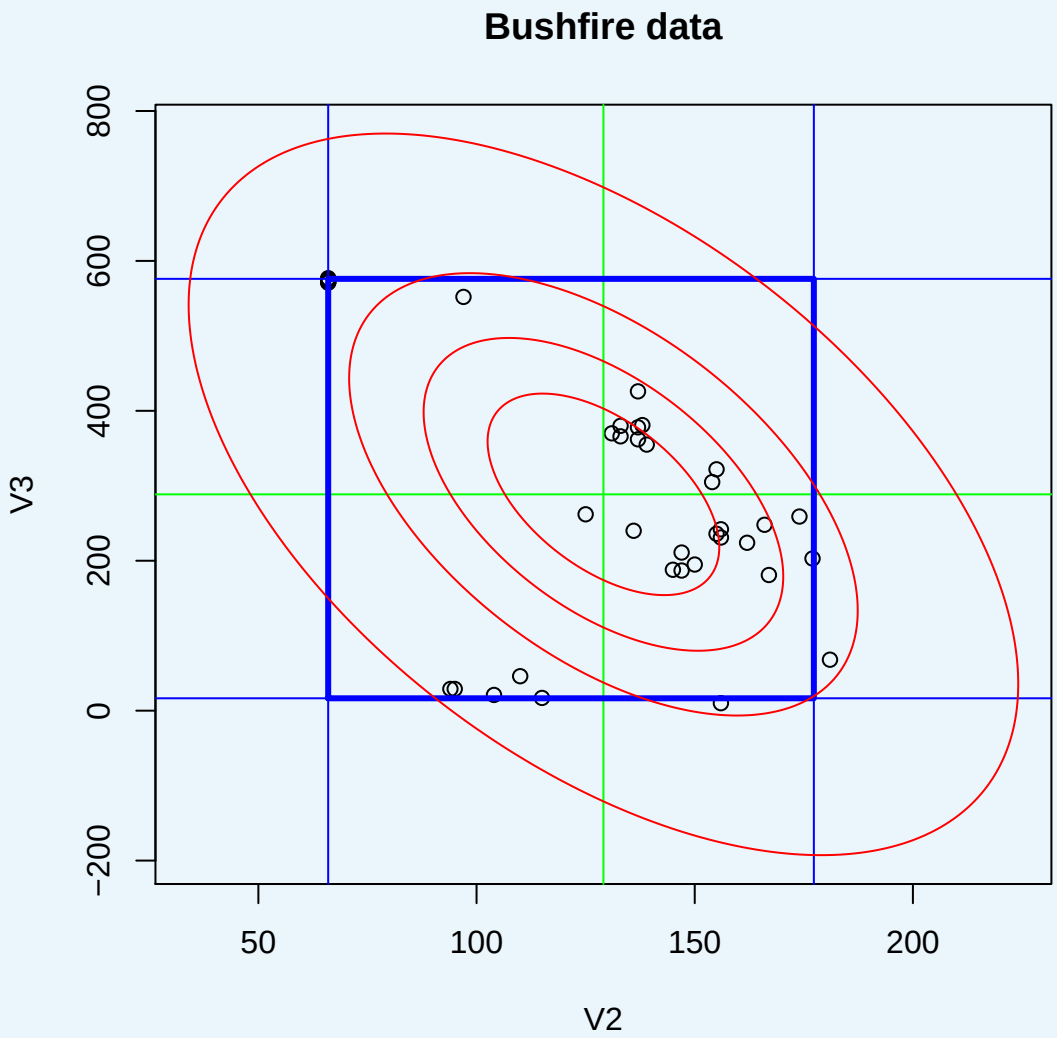
Example: Bushfire data



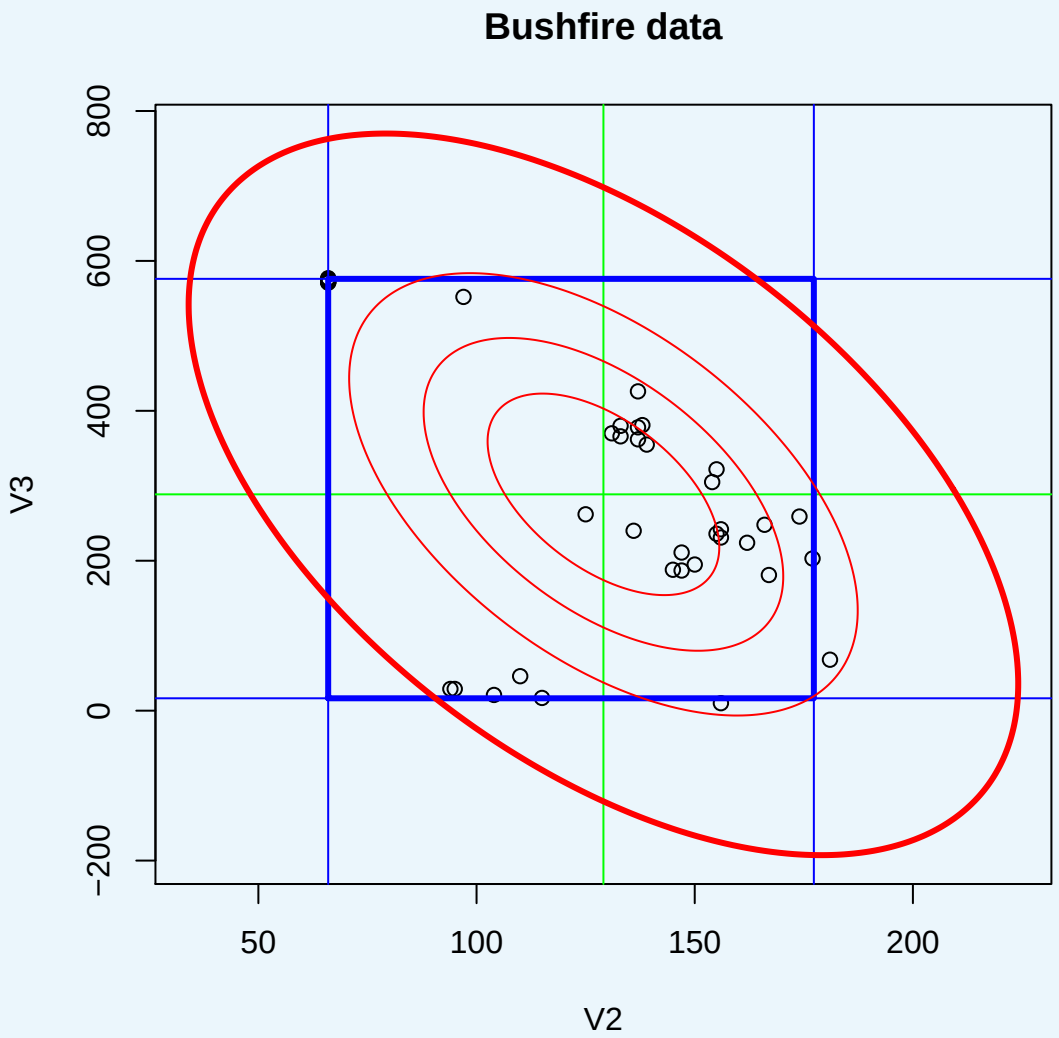
Example: Bushfire data



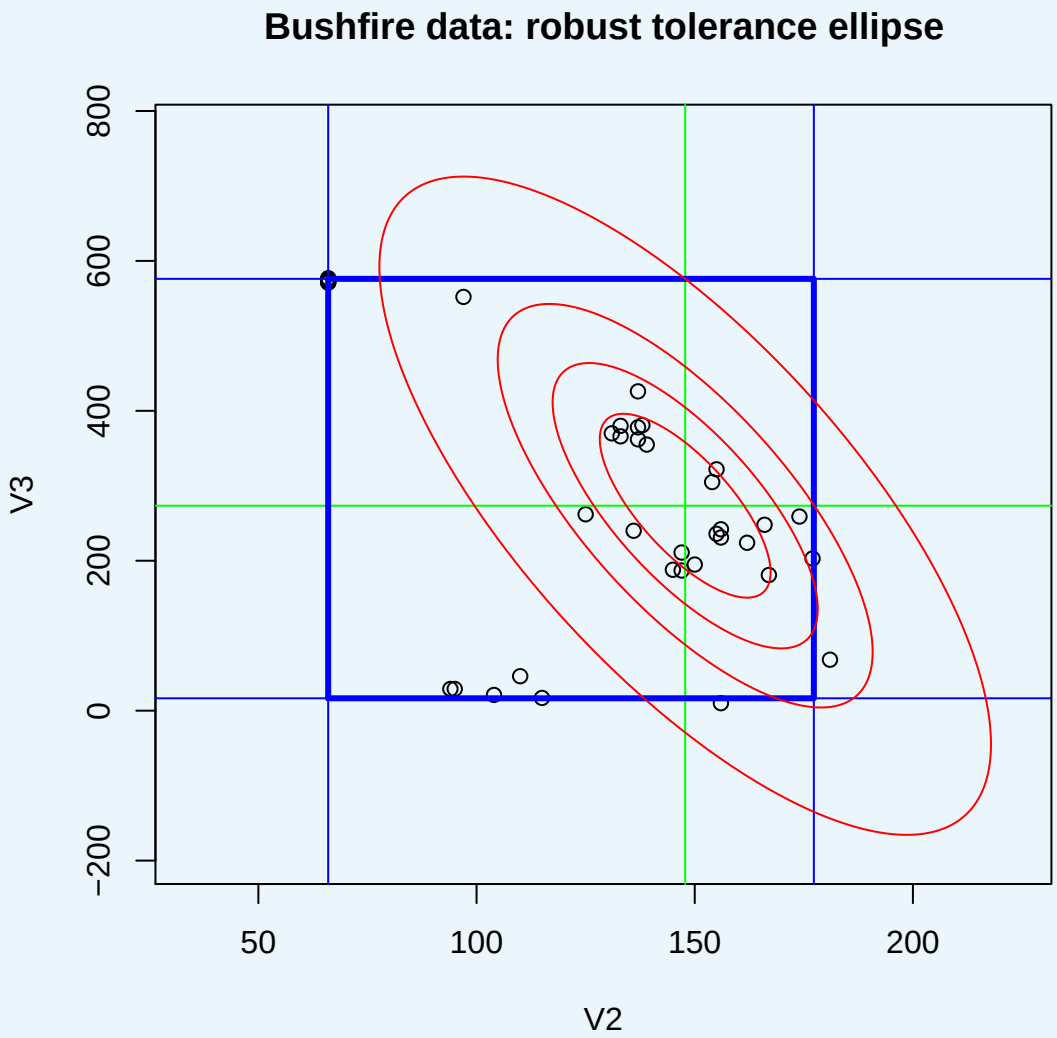
Example: Bushfire data



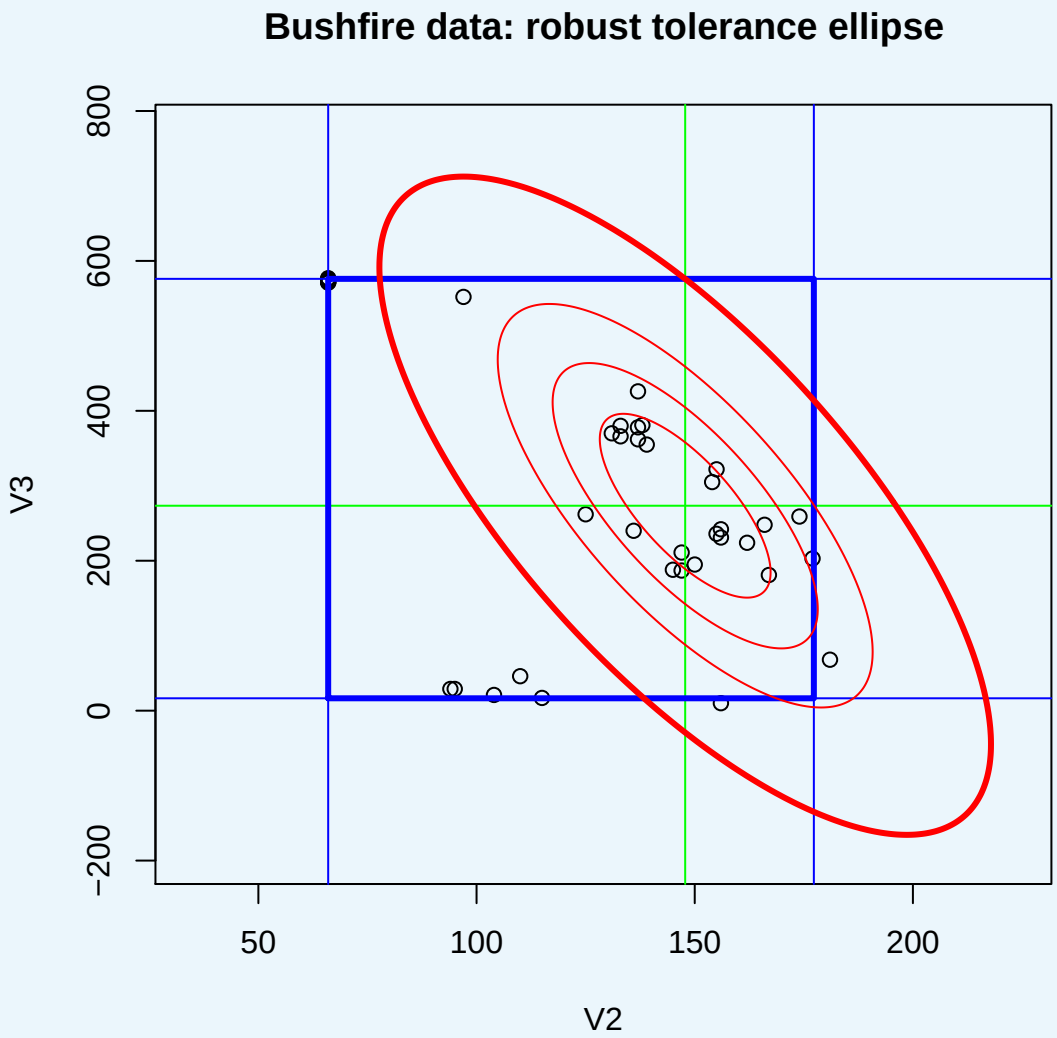
Example: Bushfire data



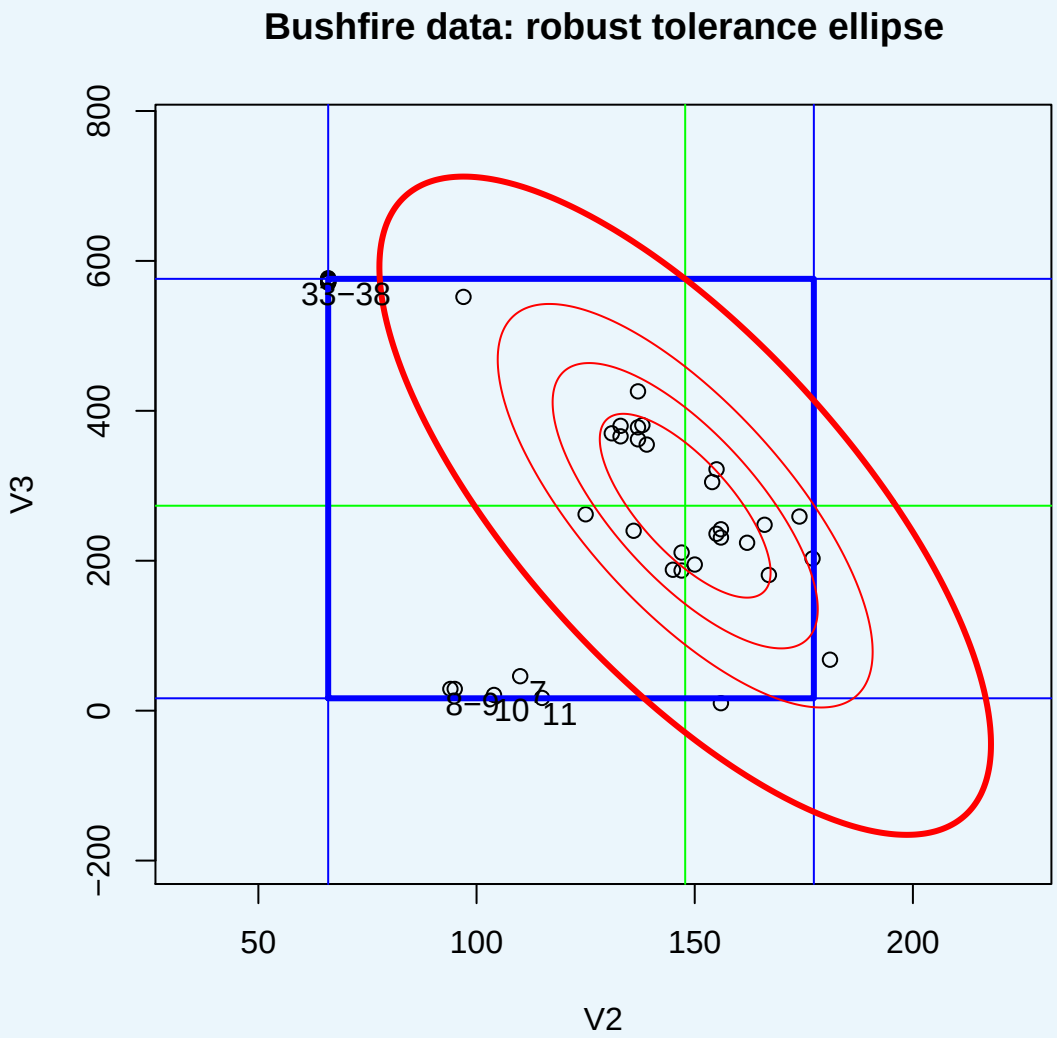
Example: Bushfire data



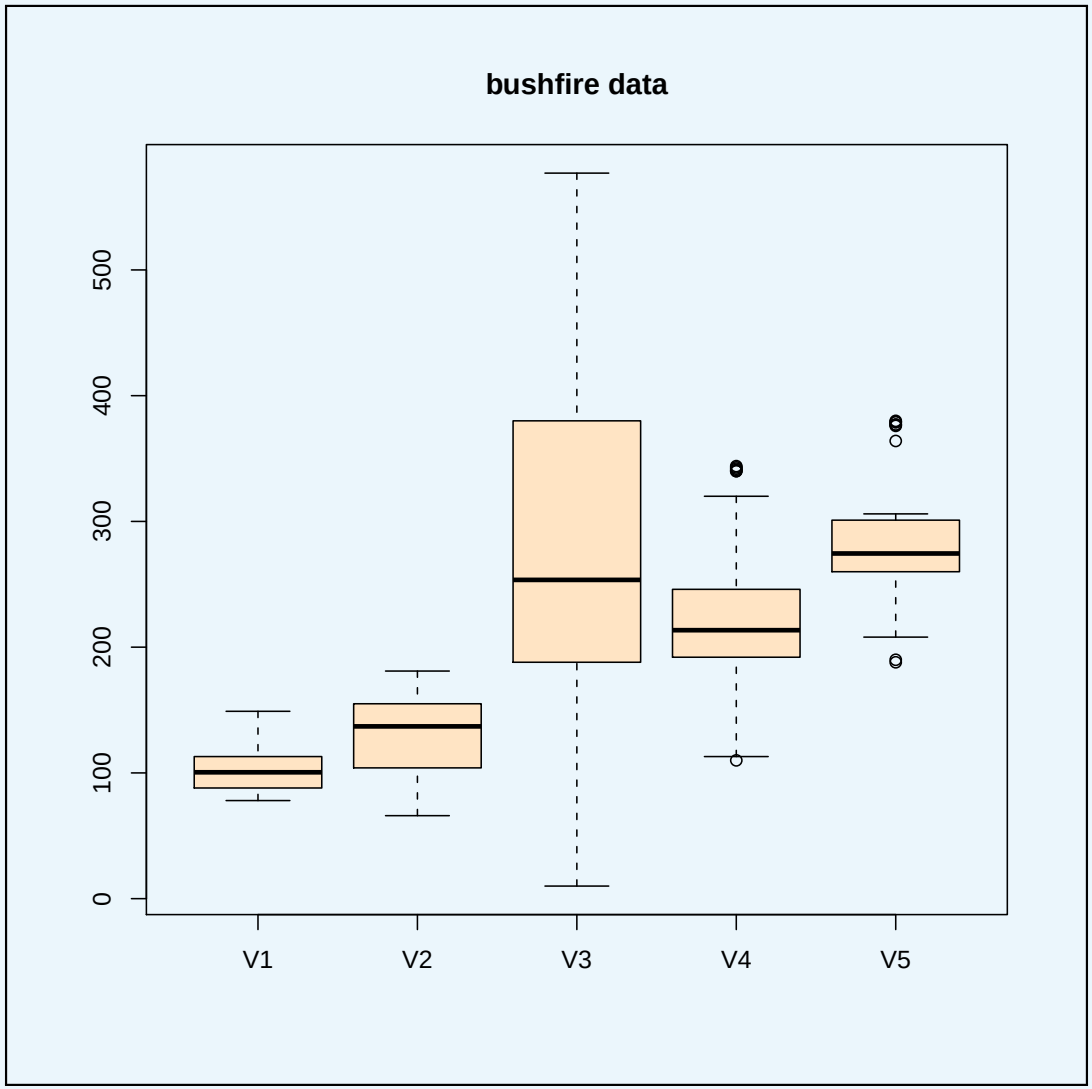
Example: Bushfire data



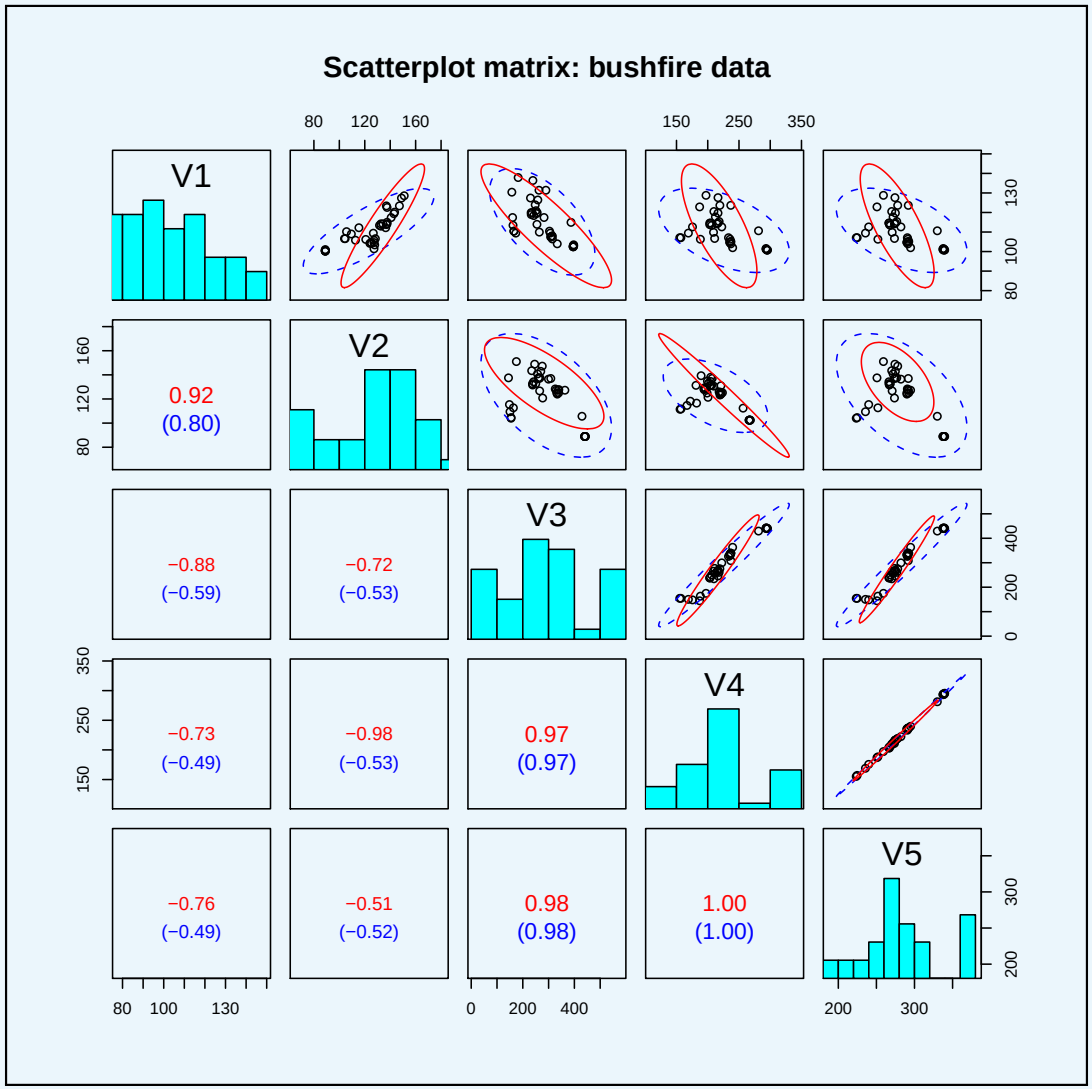
Example: Bushfire data



Example: Bushfire data - Boxplots



Example: Bushfire data - Scatterplot matrix



General Framework for Multivariate Outliers

Two phases (Rocke and Woodruff, 1996)

1. Calculate Robust Distances
   - Obtain robust estimates of location T and scatter C
   - Calculate robust Mahalanobis-type distance

$$RD_i = \sqrt{((\mathrm{x}_i - \mathrm{T})^t \mathrm{C}^{-1} (\mathrm{x}_i - \mathrm{T}))}$$

2. Cutoff point: Determine separation boundary Q.

   Declare points with $RD_i > Q$, i.e. points which are sufficiently far from the robust center as outliers.

   Usually $Q = \chi^2_p(0.975)$ but see also Hardin and Rocke (2005), Filzmoser, Garrett, and Reimann (2005), Cerioli, Riani, and Atkinson (2008).

Outliers in Sample Surveys: Multivariate methods

- Statistics Canada (Franklin *et al.*, 2000) - Annual Wholesale and Retail Trade Survey (AWRTS)
- The EUREDIT project of the EU (Charlton 2004)
- Todorov *et al.* (2011): R package **rrcovNA**
- Bill and Hulliger (2016): R package **modi**
- Wada *et al.* (2020): R packages **RMSD** and **RMSDp**
- D'Orazio (2023)
- Todorov (2024): The R Package Ecosystem for Robust Statistics

The **modi** package

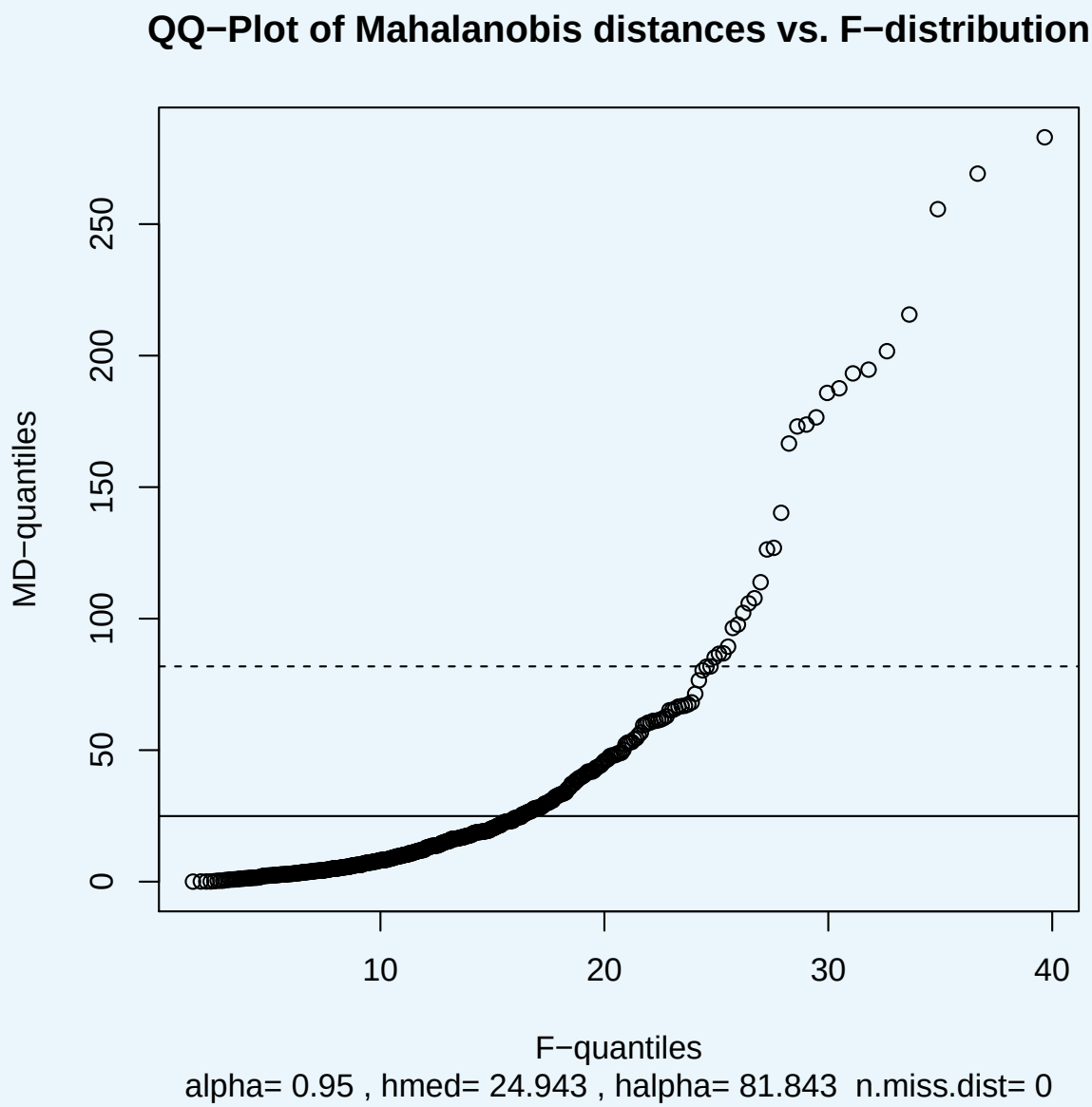
- Bill and Hulliger (2016)
- Available at CRAN
  - TRC - Transformed Rank Correlations - Béguin and Hulliger (2004)
  - EA - Epidemic Algorithm - Béguin and Hulliger (2004)
  - BEM - Béguin and Hulliger (2008) - a combination of BACON algorithm (Billor, Hadi and Vellemann 2000) and EM
- Data set sepe: anonymized sample of a pilot survey on environment protection expenditures of the Swiss private economy (1993).

All three algorithms can handle sampling weights

```
> library(modi)
> library(car)
> data(sepe)
> vlist <- c(3:5, 8:11, 14)
> colnames(sepe)[vlist]
[1] "totinvwp" "totinvwm" "totinvap" "totinvto"
[5] "totexpwp" "totexpwm" "totexpap" "totexpto"
> sepex <- recode(as.matrix(sepe), "0=NA")
> logsepe <- log(sepex[, vlist] + 1)

## decrease the cutoff quantile for good observations
n <- nrow(logsepe)
> res <- BEM(logsepe, sepe$weight, c0=5, alpha=0.01/n)
BEM has detected 89 outlier(s) in 3.12 0 3.2 NA NA seconds.
> res$cutpoint
[1] 37.14862
```

```
> PlotMD(res$dist,  ncol(logsepe), alpha=0.95)
```



Handling of the detected outliers

```
> imp <- Winsimp(data, res$center, res$scatter, outind)
> sum(imp$imputed.data < 0)
[1] 99
```

	original.mean	mean.norm	mean.before	mean.after
totinvwp	0.71	0.73	0.79	0.87
totinvwm	0.47	0.56	0.72	0.71
totinvap	0.88	1.00	1.05	1.14
totinvto	1.51	1.81	1.69	1.92
totexpwp	0.99	1.05	1.07	1.09
totexpwm	1.53	1.62	1.46	1.70
totexpap	0.48	0.47	0.48	0.55
totexpto	2.01	2.12	1.98	2.21
Determinant	4.86	16.01	4.32	11.05

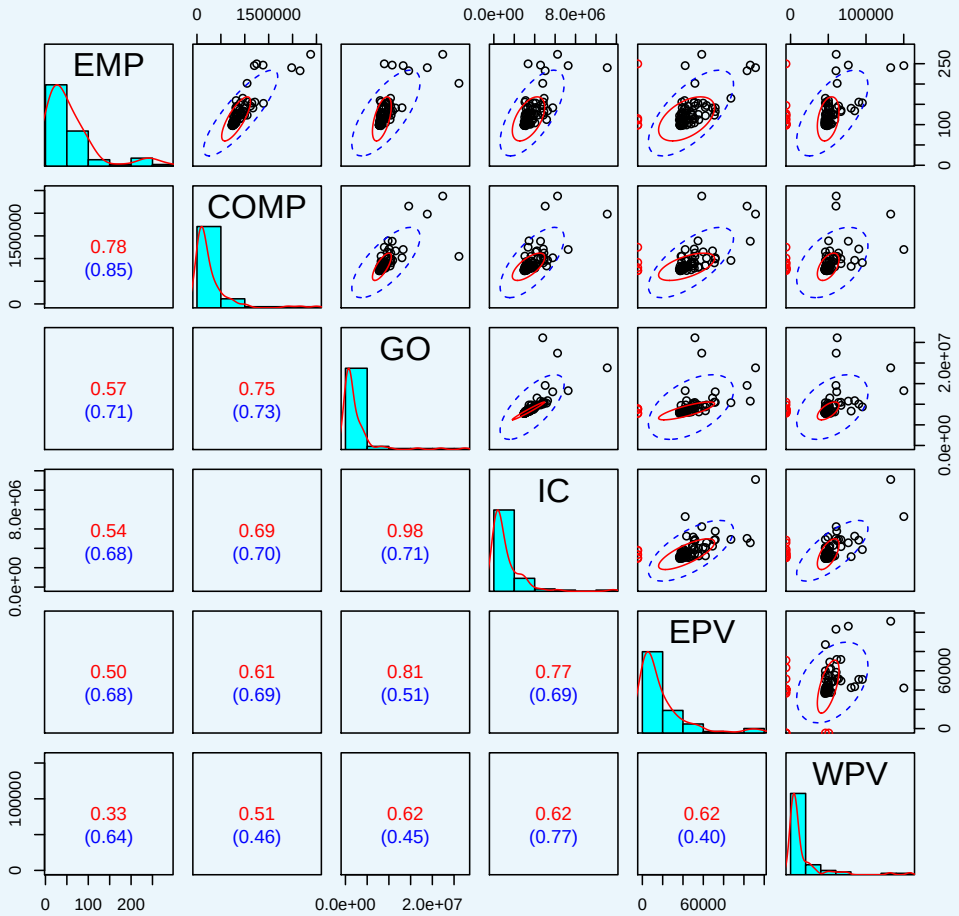
Table: Means and determinant of the covariance matrix for original, normally imputed and robustly imputed after re-introduction of zeros in different steps

The **rrcovNA** package

- Similar structure to **rrcov**: S4 classes with a number of diagnostic and visualization functions
- Available at CRAN
- MCD, OGK, S - following an MVN imputation with an EM algorithm
- Classical and robust PCA for incomplete data
- NEW
  - Deterministic MCD DETMCD: Hubert *et al.* (2012)
  - Deterministic S and MM estimates (DETS and DETMM): Hubert *et al.* (2015)
  - Generalized S estimates GSE: Danilov *et al.* (2012)

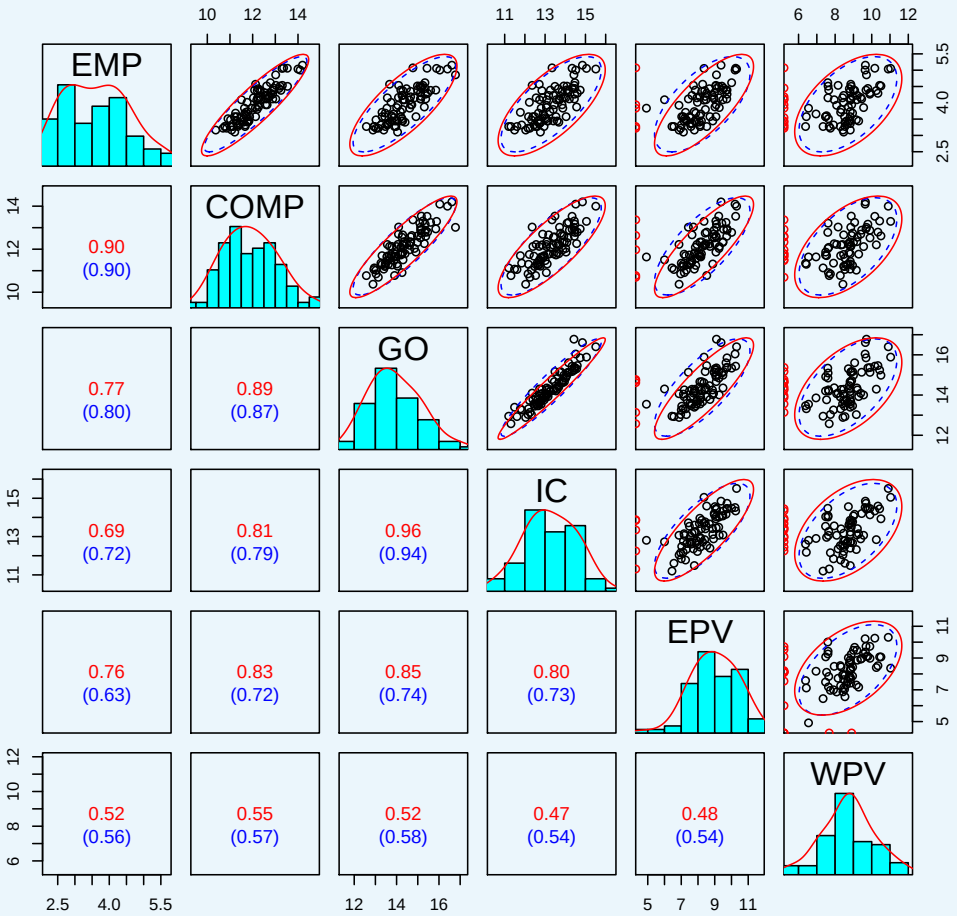
Example 1: AIS (ISIC=2395)

```
> library(rrcovNA)
> cv <- CovNAMcd(ais); plot(cv, which="pairs")
```



The example: AIS (ISIC=2395)

```
> library(rrcovNA)
> lcv <- CovNAMcd(log(ais)); plot(lcv, which="pairs")
```



The example: AIS (ISIC=2395)

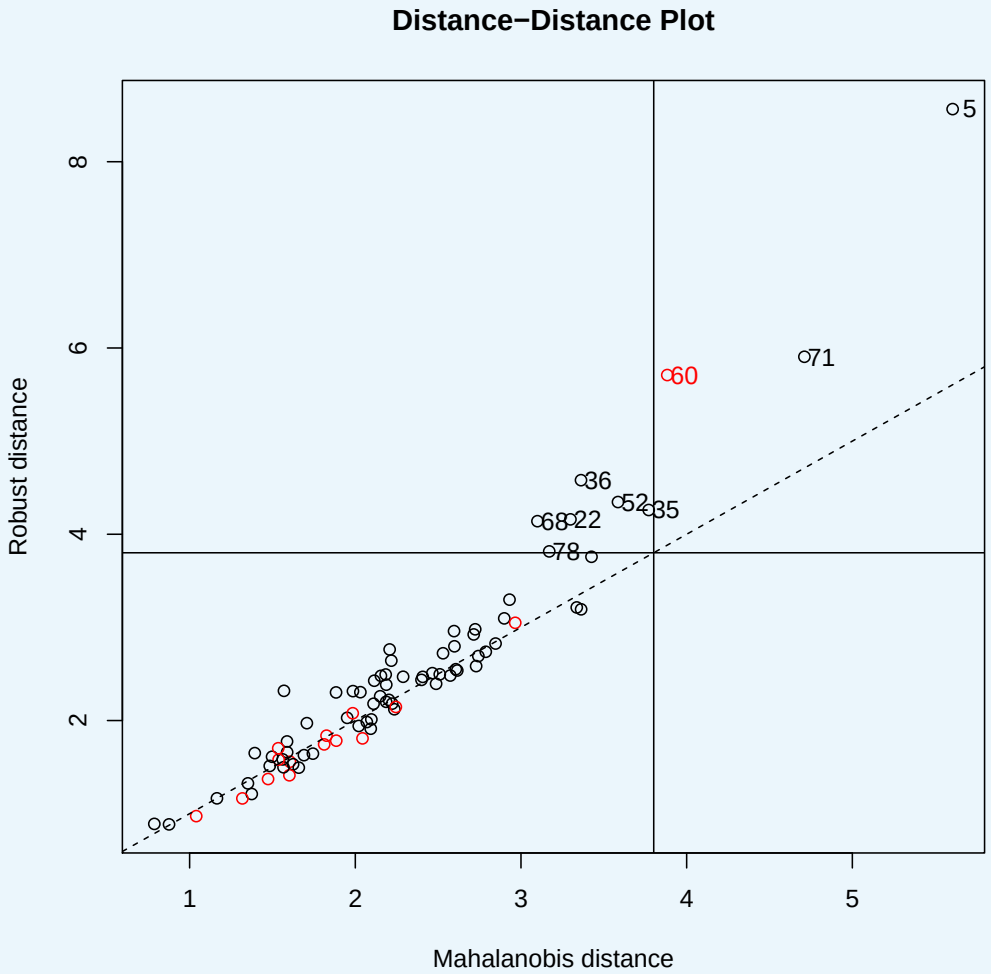
```
> (lcv <- CovNASest(log(ais), method="GSE"))
Call:
CovNASest(x = log(ais), method = "GSE")
-> Method:  Generalized S-Estimator

Robust Estimate of Location:
[1]   3.6633   3.6335  -0.7825  11.5271   9.0910   7.6720   9.1587

Robust Estimate of Covariance:
       PER     EMP     FEM     GWS     ECV     WCV     FUEL
PER    2.019   2.070   4.242   2.539   2.570   2.281   1.836
EMP    2.070   2.128   4.398   2.621   2.655   2.337   1.895
FEM    4.242   4.398  16.871   5.870   5.800   4.583   3.909
GWS    2.539   2.621   5.870   3.558   3.655   3.048   2.291
ECV    2.570   2.655   5.800   3.655   4.195   3.237   2.449
WCV    2.281   2.337   4.583   3.048   3.237   3.878   1.945
FUEL   1.836   1.895   3.909   2.291   2.449   1.945   3.301
```

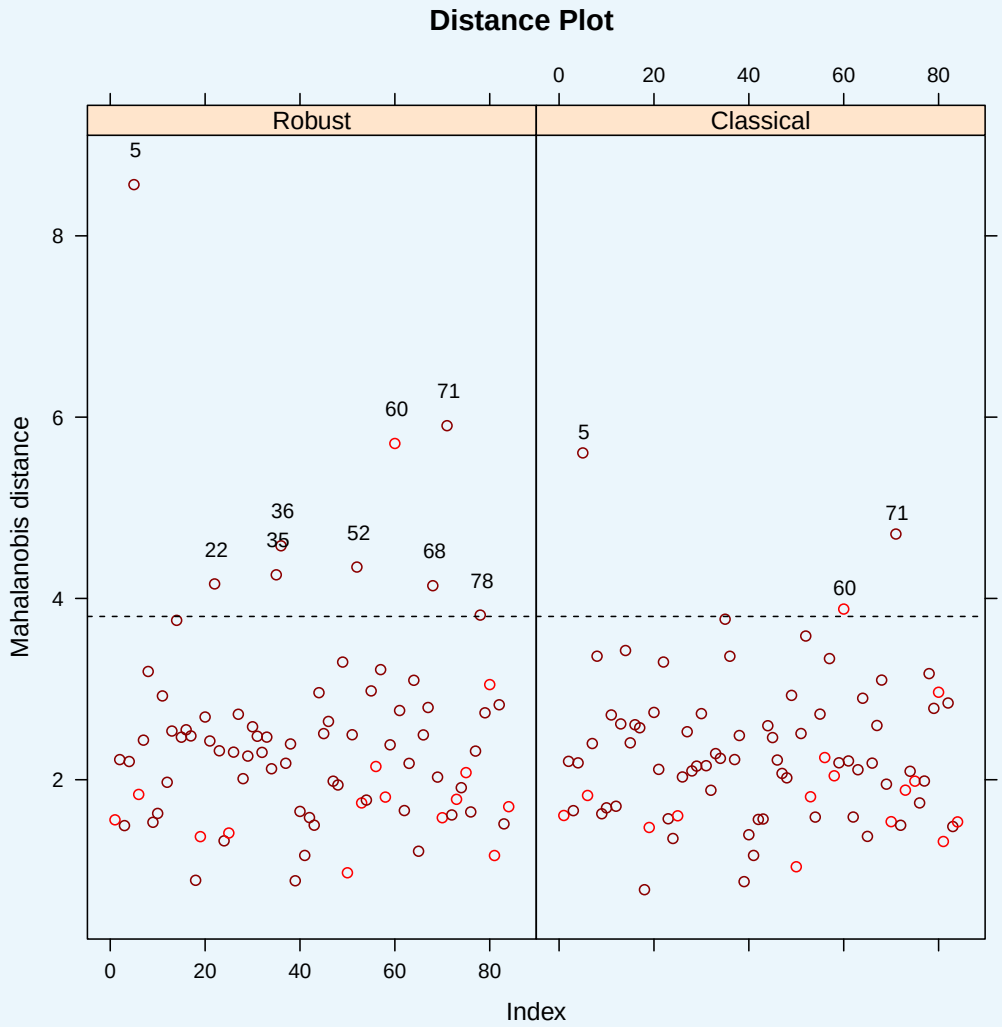
The example: AIS (ISIC=2395)

```
> plot(lcv, which="dd")
```



The example: AIS (ISIC=2395)

```
> plot(lcv, which="xydistance")
```



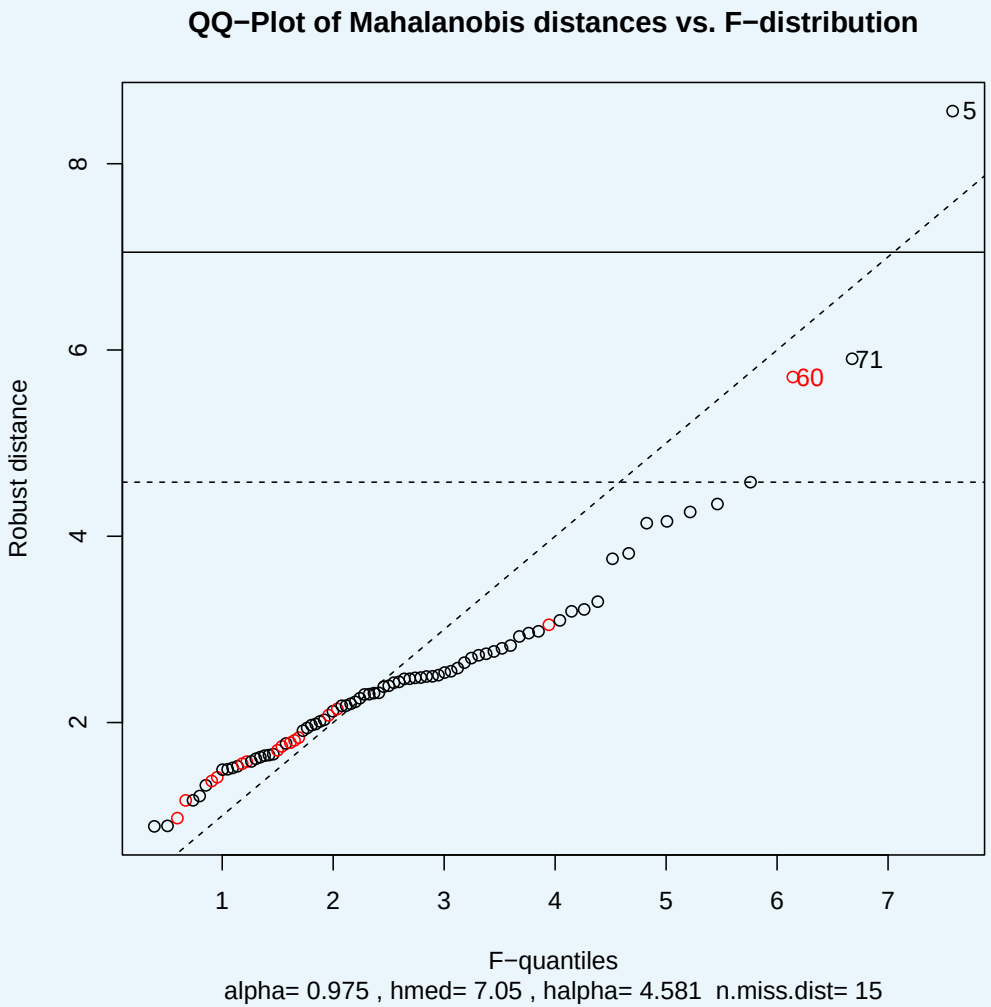
AIS (ISIC=2395): the influence of the variables

- Observation 5: has very high share of value added in gross output (VA/GO), more than 90% and at the same time very low electricity consumption
- Observation 60: has very low wages per employee (WS/EMP) as well as very low electricity consumption (compared to output)
- Observations 71: has very low value added per employee (VA/EMP) and very low electricity consumption (compared to output)
- ...

Note: Value Added (VA) = GO - IC

The example: AIS (ISIC=2395)

```
> plot(lcv, which="Fdist")
```

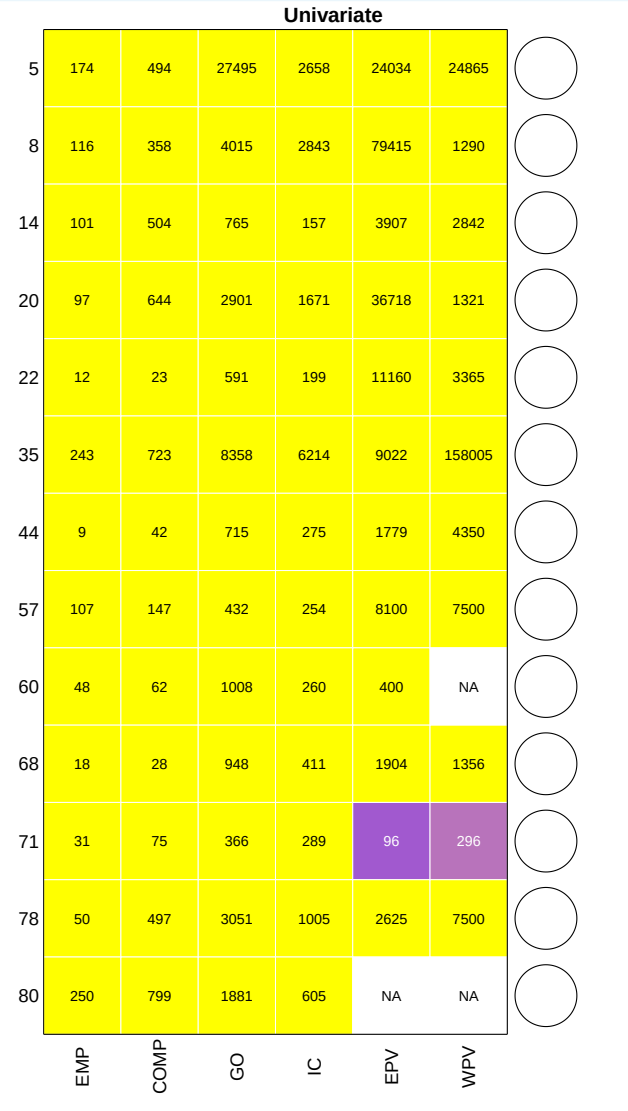


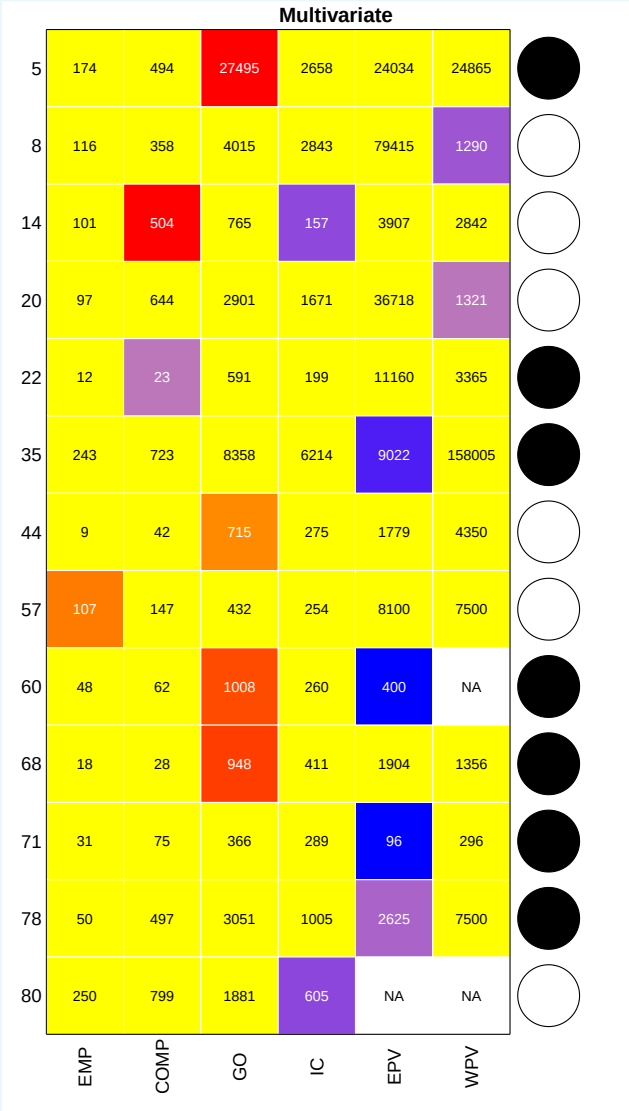
Detecting Deviating Data Cells (DDC)

Detecting Deviating Data Cells (Rousseeuw and Van Den Bossche, 2018): functions `DDC()` and `cellMap()` in package **cellWise**

- Often many rows have a few contaminated cell values: may not be visible by looking at each variable (column) separately (Alqallaf, *et al.*, 2009).
- Preprocessing and standardizing (robustly) the data; Apply univariate outlier detection and flag outlying cells
- Find connected variables (with strong robustly computed bivariate correlation)
- Compute predicted values for all cells and using these predictions compute the standardized cell residuals
- Flag cell-wise outliers
- Flag row-wise outliers

AIS (ISIC=2395): cellMap (package **cellWise**)



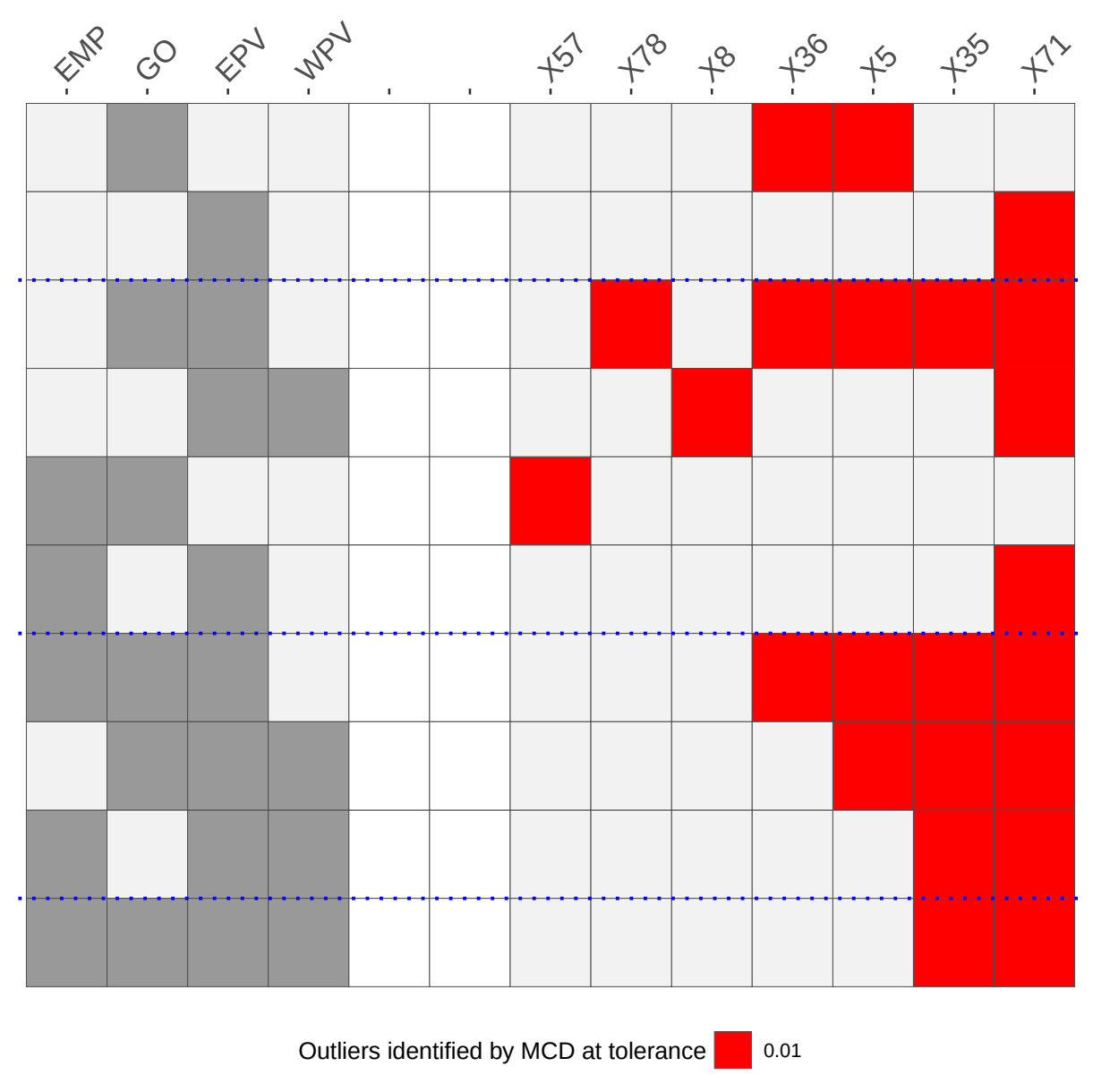


Visualizing multivariate outliers

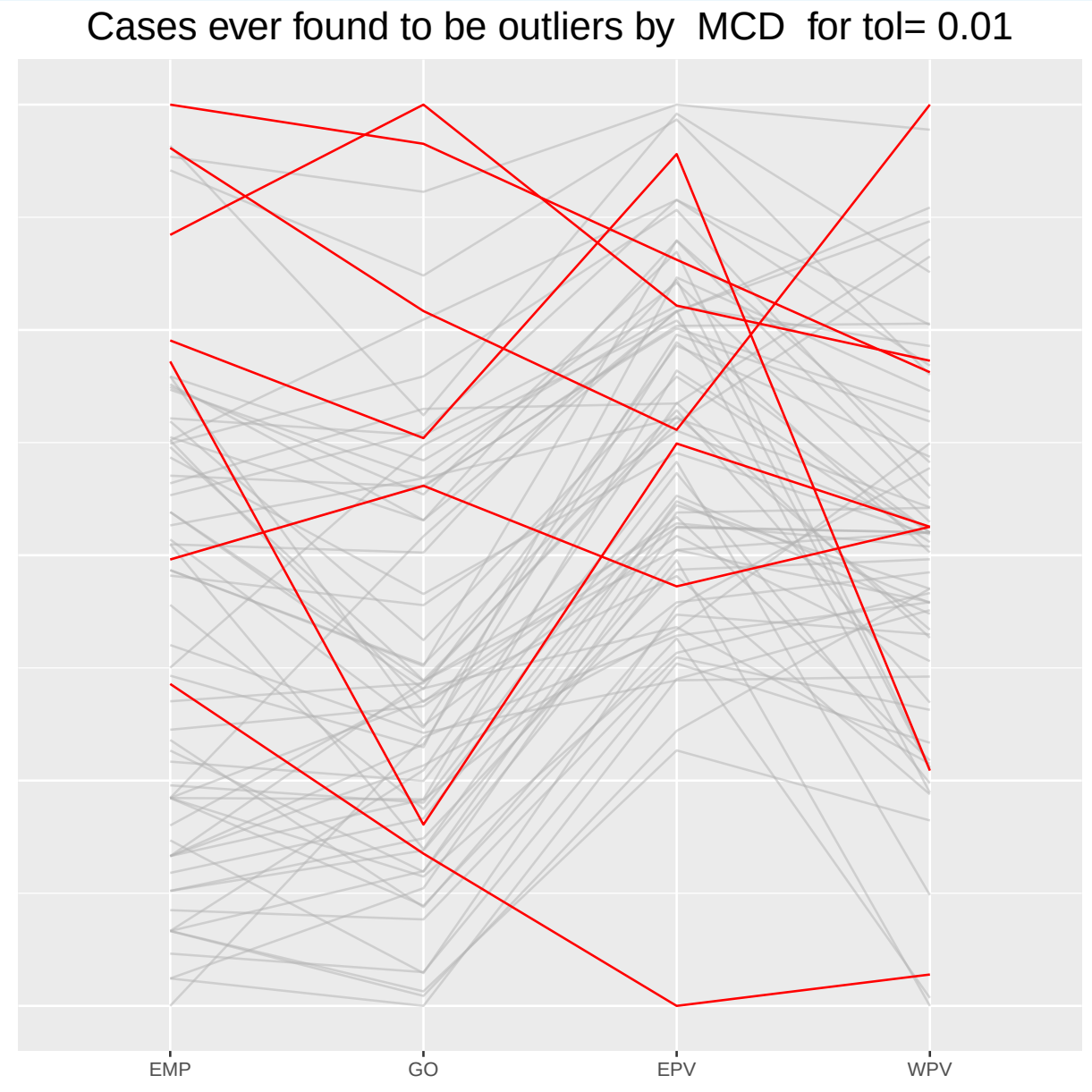
Outliers on different dimensions of a dataset

- Look first at the lower dimensions, checking the individual variables first, then pairs of variables, then possibly sets of three...
- How to summarise and visualise this information to support analysis?
- A new visualisation tool, the O3 plot in the R package **OutliersO3**
- The methods used are: HDoutliers (**HDoutliers** package), mvBACON in **robustX**, adjOutlyingness and covMcd both in **robustbase**, FastPCS in **FastPCS** and DDC in **cellWise**
- Display outlyingness using two or more methods simultaneously

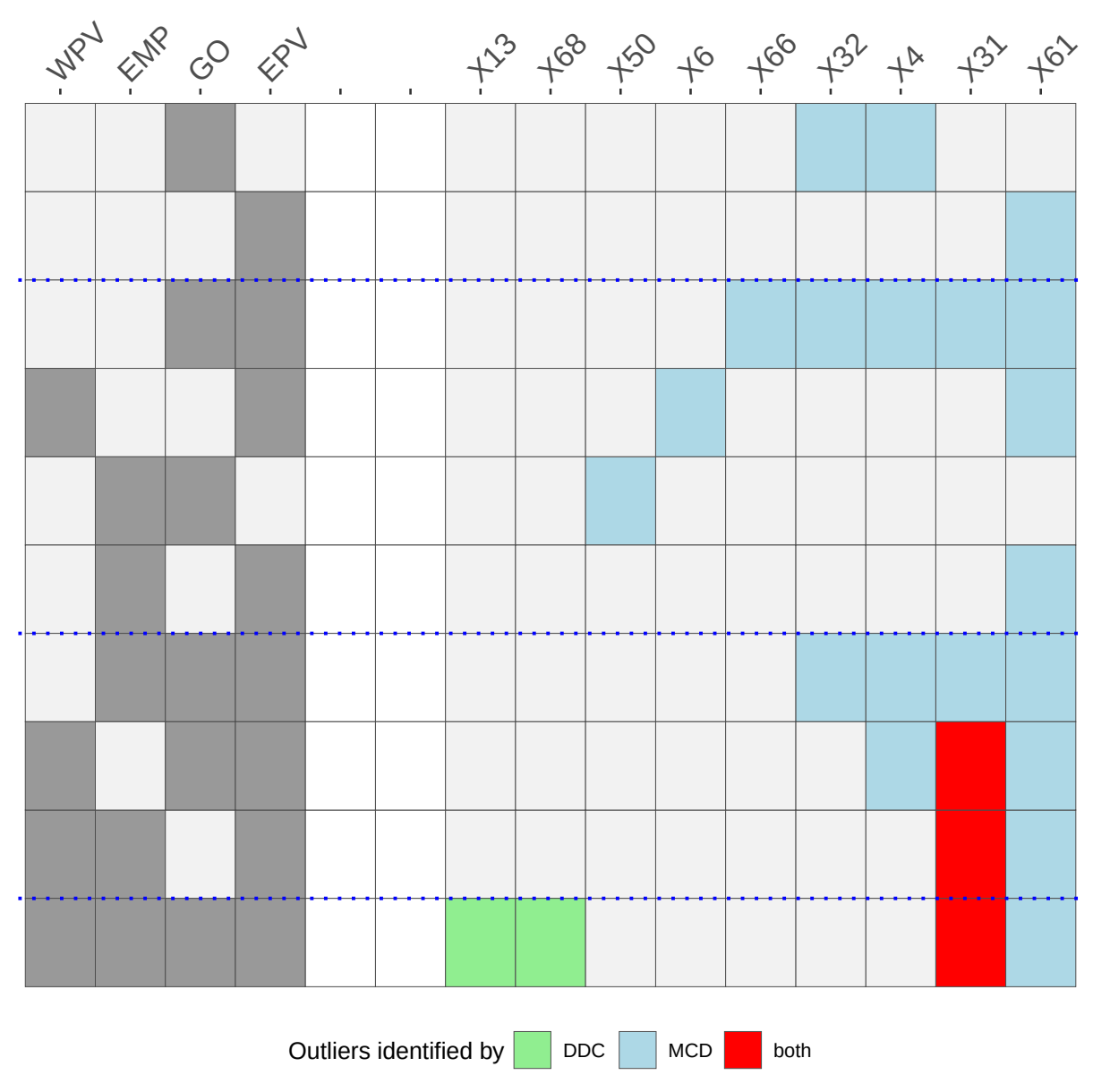
Overview Of Outliers (package **OutliersO3**)



Parallel Coordinates plot



Overview Of Outliers (package **OutliersO3**)



Example 2: Bushfire data

Simple experiment with the Bushfire data

- 12 outliers: 33-38, 32, 7-11; 12 and 13 are suspect
- Missing values added with an MCAR mechanism
- Created 4 data sets: with 10%, 20%, 30% and 40% missing data
- For each method and data set the known outliers are indicated as detected or not
- Non-outliers that were classified as outliers, or swamped non-outliers are given too (FP=false positives)

- Repeat $m = 100$ times for each method and missingness rate
- Average the number of non-identified outliers and the number of regular observations declared outliers

Average percentage of outliers that were not identified

	0	10	20	30	40
MCD	0.00	3.46	9.46	18.15	28.69
DETMCD	0.00	1.62	5.85	12.85	24.62
S	0.00	36.46	54.38	64.77	77.77
DETS	0.00	15.46	31.77	52.15	68.23
GSE	0.00	4.00	10.85	15.46	31.46
BEM	7.69	10.08	11.31	11.92	13.92

Average percentage of non-outliers that were classified as outliers

	0	10	20	30	40
MCD	12.00	4.28	2.84	2.44	2.08
DETMCD	12.00	3.92	1.92	1.60	1.16
S	0.00	2.48	2.80	2.12	1.44
DETS	0.00	0.16	0.56	1.44	1.08
GSE	0.00	2.00	3.40	2.12	2.92
BEM	4.00	4.76	4.24	6.12	6.88

Simulation study: Austrian Structural Business Statistics Data

- In Todorov et al. (2011) we evaluated different outlier detection algorithms on generated synthetic (but close-to-reality) data sets based on a real structural business statistics data set.
- The following algorithms were compared:
  - MCD, S, OGK, SIGN1 from package **rrcovNA**
  - TRC, EA and BEM from package **modi**
- For an outlier fraction of 10% all estimators except EA perform excellent in terms of outlier error rate (FN) and identify all outliers independently of the percentage of missing values
- 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%).
- Now we will evaluate the deterministic MCD (DETMCD), deterministic S (DETS) and generalized S (GSE), under the same conditions.

Austrian Structural Business Statistics Data

- More than 320.000 enterprizes. Available raw data set: 21669 observations in 90 variables, structured according NACE revision 1.1 with 3891 missing values
- We investigate the following 10 variables of NACE 52.42 - "Retail sale of clothing"

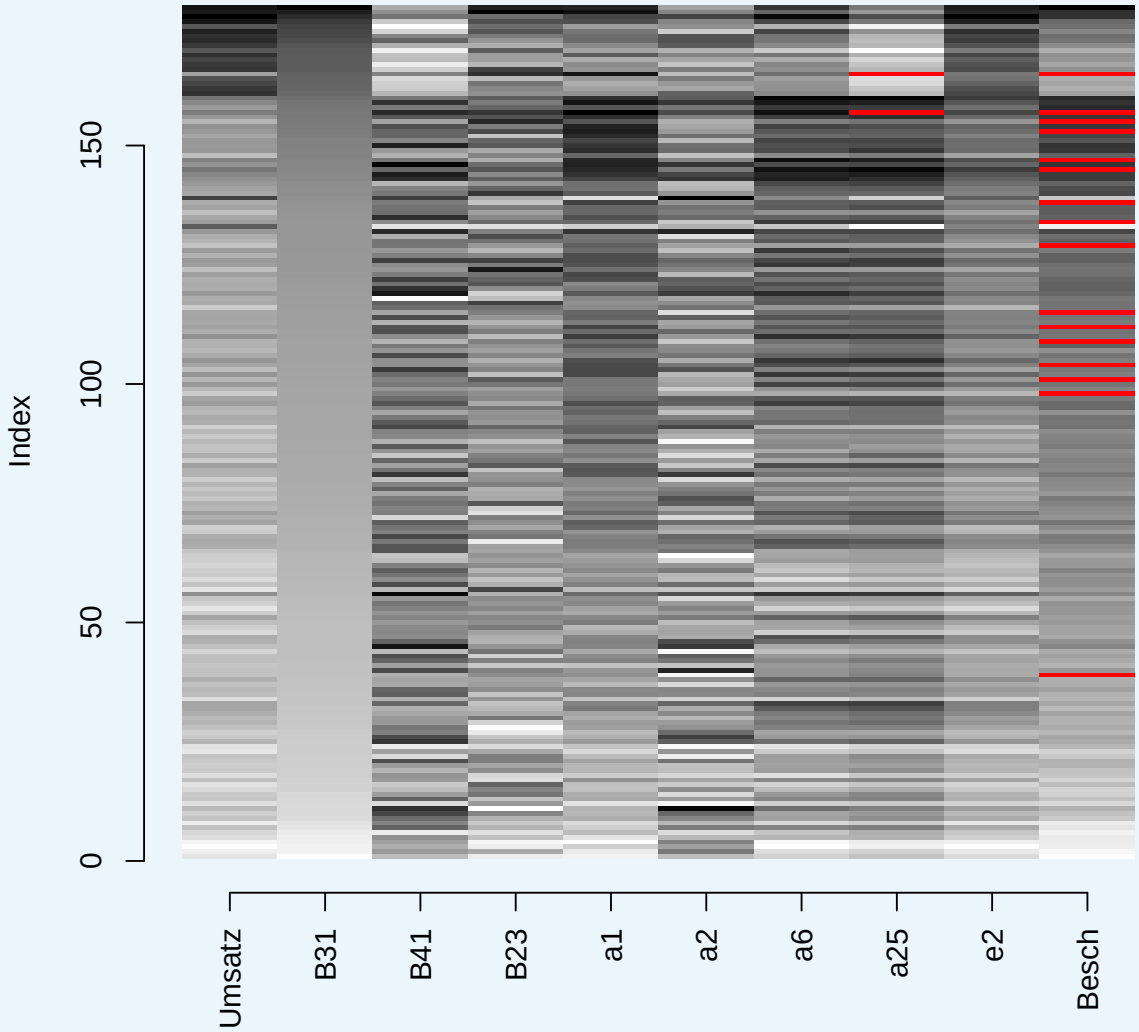
TURNOVER	Total turnover
B31	Number of white-collar employees
B41	Number of blue-collar workers
B23	Part-time employees
EMP	Number of employees
A1	Wages
A2	Salaries
A6	Supply of trade goods for resale
A25	Intermediate inputs
E2	Revenues from retail sales

Synthetic SBS data, NACE 5244

Missing value patterns analyzed with the R package VIM.

Data matrix Plot:

- Missing values are red colored
- The darker a line the higher the value of an observation

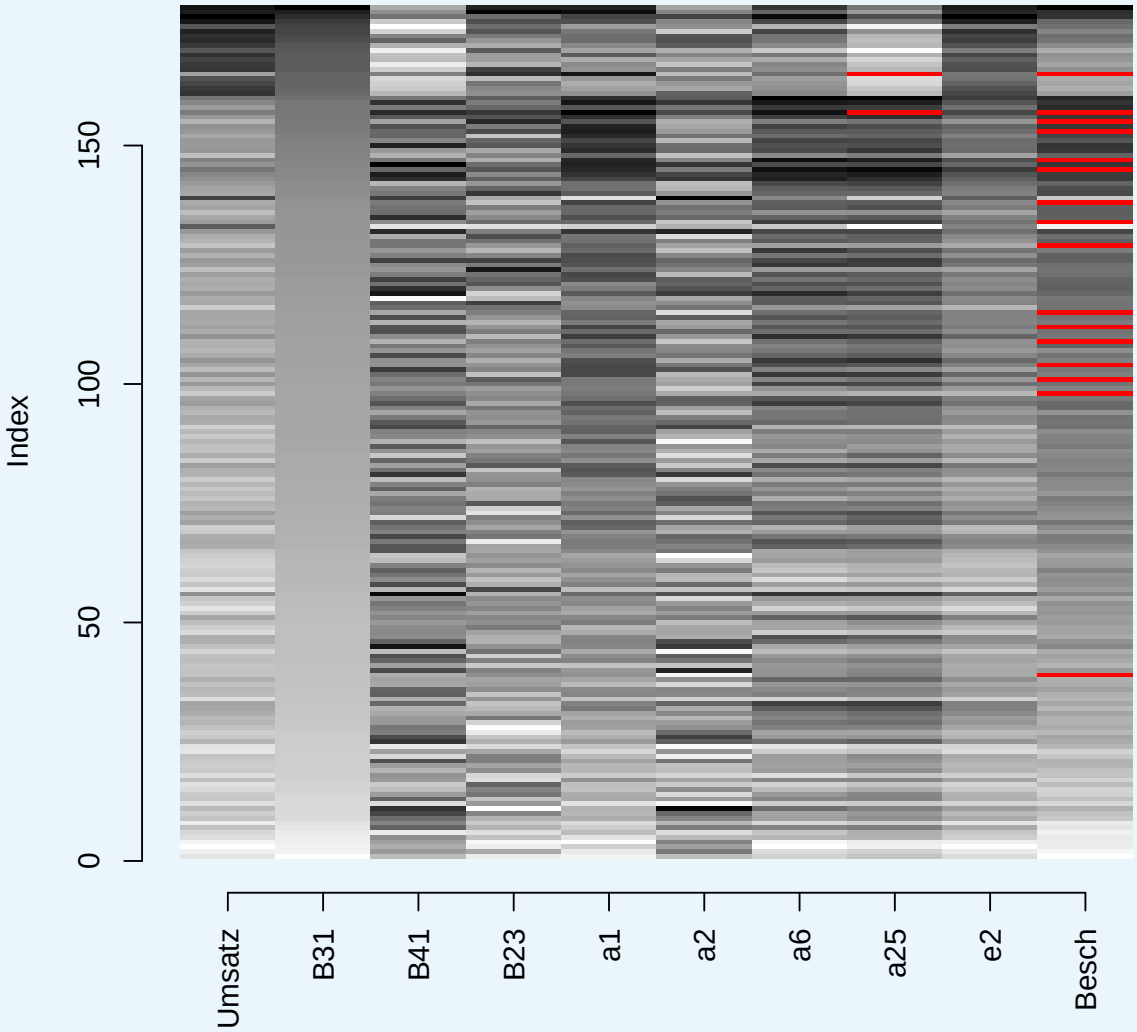


Synthetic SBS data, NACE 5244

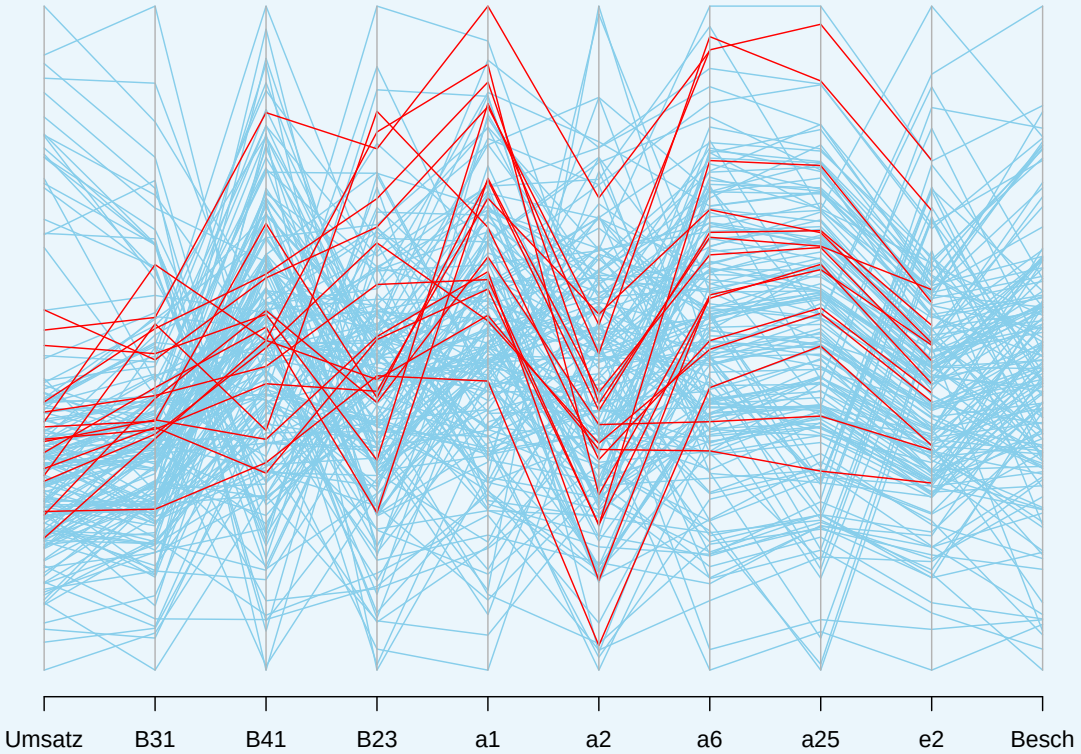
Missing value patterns analyzed with the R package VIM.

Data matrix Plot:

- Missing values are red colored
- The darker a line the higher the value of an observation



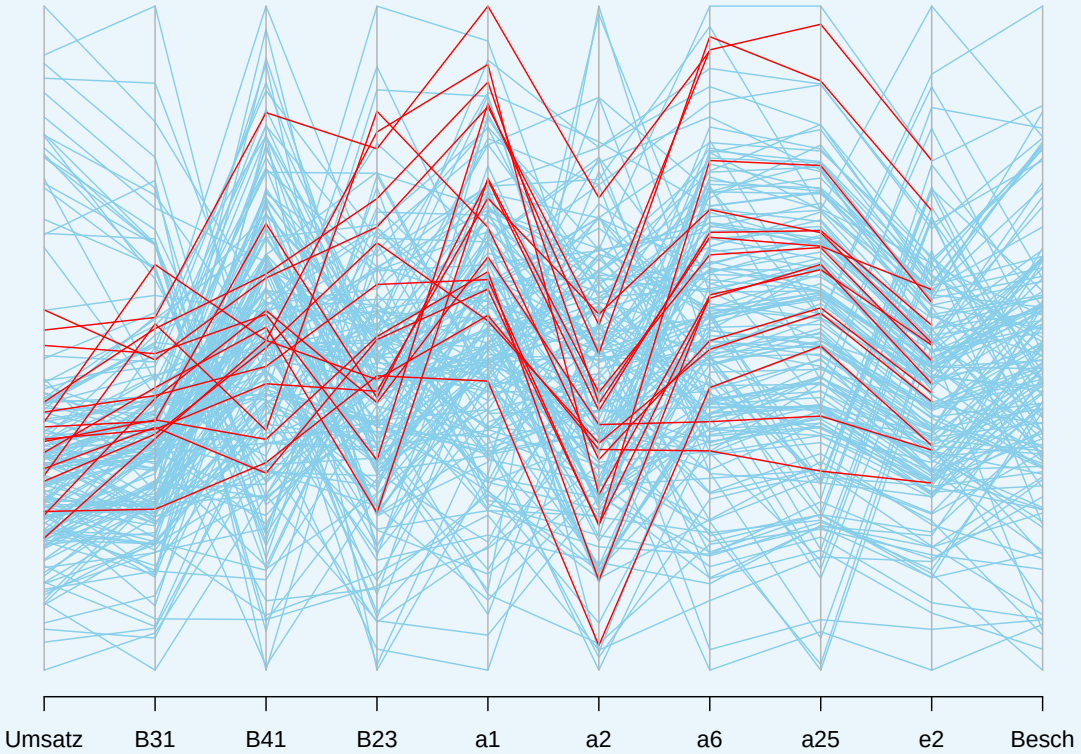
Synthetic SBS data, NACE 5244



Parallel Coordinate Plot:

- Observations with Missing values in *EMP* are red colored
- → MAR situation.

Synthetic SBS data, NACE 5244



Parallel Coordinate Plot:

- Observations with Missing values in *EMP* are red colored
- → MAR situation.

Simulation Setup I

Simulation settings

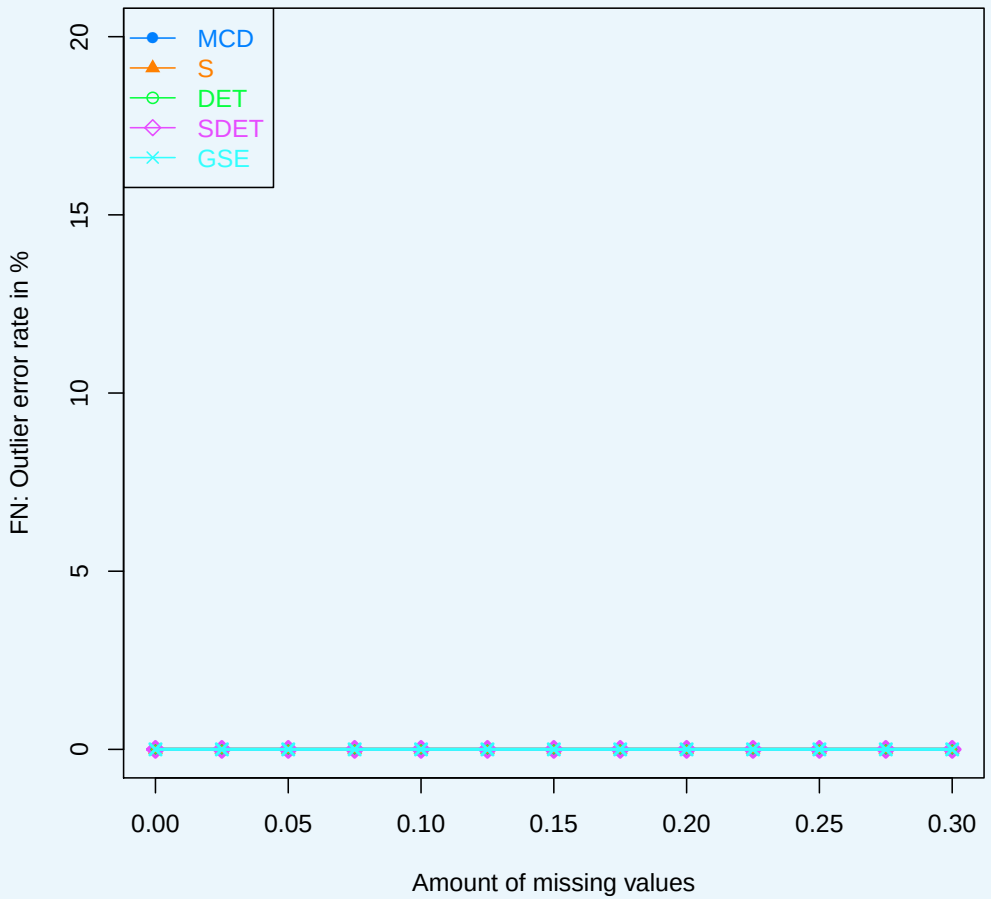
- Log-normal data generated according to the structure (T, C) and size of the original data.
- Two experiments:
  1. Fixed fraction of outliers $= 0.1$ and missing rates $= 0.0, \ldots, 0.3$ with step 0.025
  2. Fixed missing rate $= 0.1$ and fractions of outliers $= 0.0, \ldots ..., 0.25$ with step 0.025
- Methods: MCD and S (as a benchmark), DETMCD, DETS, GSE.
- m=400 repeated for all data sets and methods

Simulation Setup II

We compare

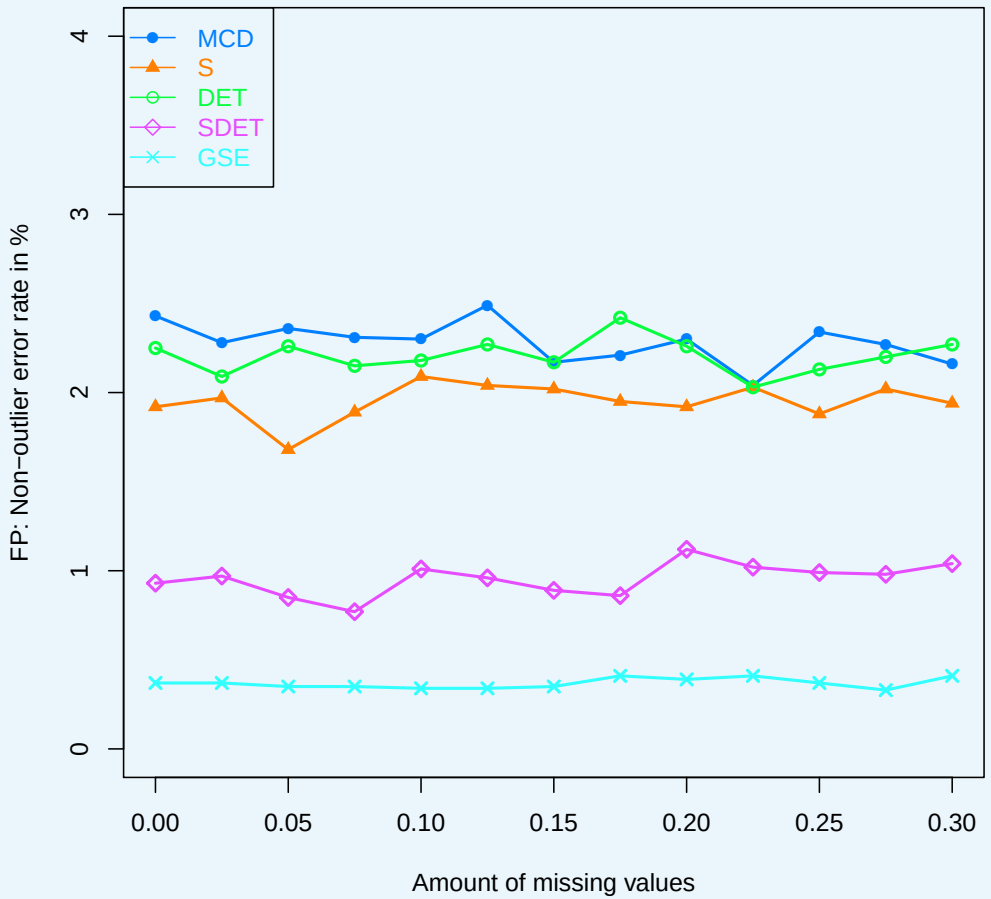
- The average percentage of false negatives (FN) - the outliers that were not identified, or masked outliers (outlier error rate)
- The average percentage of false positives (FP) - non-outliers that were classified as outliers, or swamped non-outliers (inlier error rate)
- Average computation time

Simulation results I



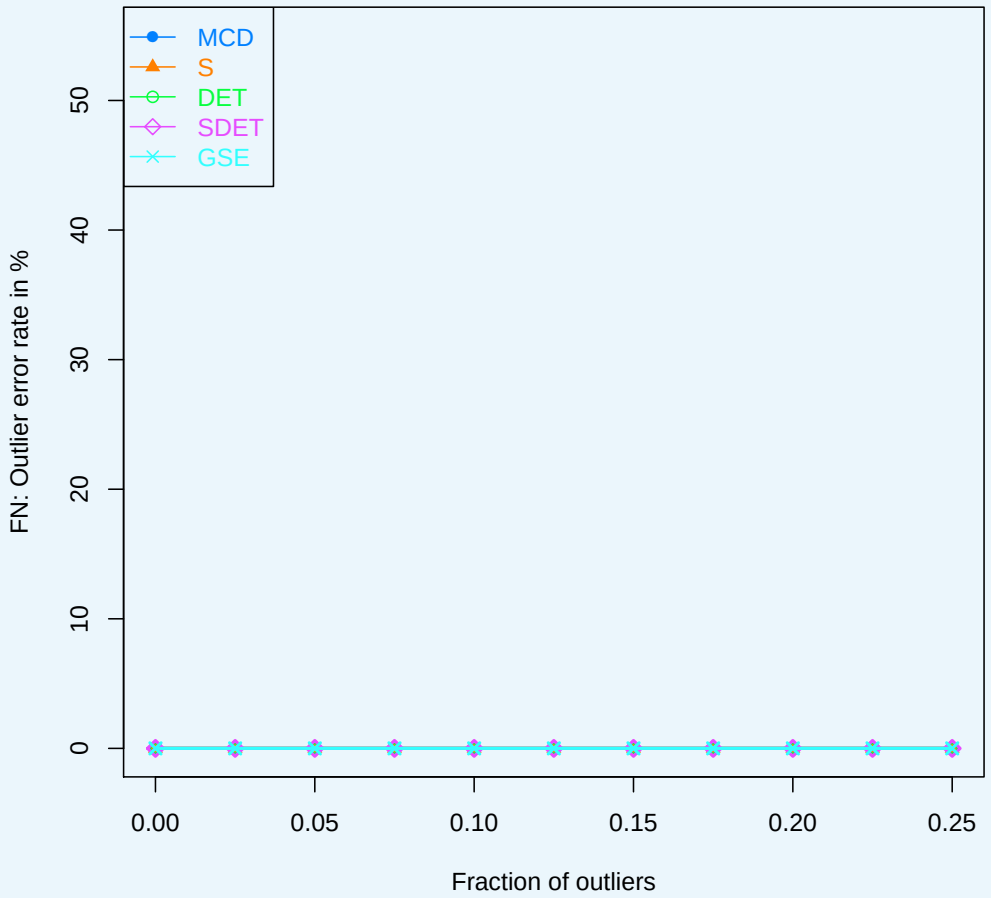
- False Negatives (FN) or outlier error rate
- Fixed fraction of outliers: 10%
- Varying percent of missingness
- Average over 400 runs
- All methods perform excellent and identify all outliers

Simulation results II



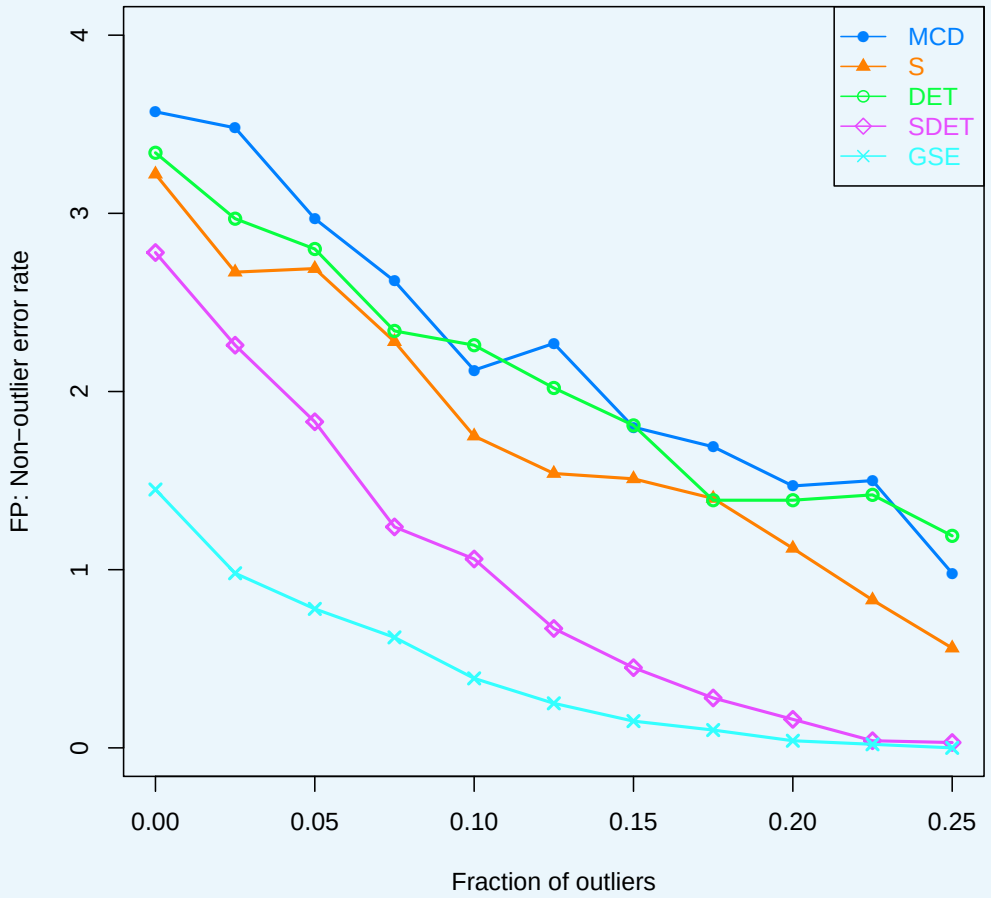
- False Positives (FP) or non-outlier error rate
- Fixed fraction of outliers: 10%
- Varying percent of missingness
- All methods perform well (less than 3%)
- GSE performs best followed by SDET (below 1%).

Simulation results III



- False Negatives (FN) or outlier error rate
- Fixed missingness rate: 10%
- Varying fraction of outliers
- Again nothing to compare: all estimators identify all outliers.

Simulation results IV



- False Positives (FP) or inlier error rate
- Fixed missingness rate: 10%
- Varying fraction of outliers
- In terms of non-outlier error rate GSE and SDET perform best (uniformly less than 3%) followed by the rest (less than 4%).

Conclusions and Outlook

- We considered methods for identification of outliers in large multivariate incomplete establishment survey data
- Two previous studies were reviewed, new methods and their implementation in R were presented: DETMCD, SDET and GSE
- The methods were compared in terms of identification performance on examples and simulation study based on real data
- Several R packages were presented: **modi**, **rrcov**, **cellWise** and **OutlierO3** and illustrated on real data examples.

- Outlook
  - Sampling weights for MCD and S estimators.
  - What to do after the outliers are found? $\Rightarrow$ Development of a practical procedure for handling of multivariate outliers.

References I

Bill, M. and Hulliger, B.
Treatment of multivariate outliers in incomplete business survey data.
Austrian Journal of Statistics, **45** 3–23, 2016

D'Orazio, M.
Some approaches to outliers' detection in R.
Romanian Statistical Review, **2023**(1), 2023

Rousseeuw, P.J., Van Der Bossche, W.
Detecting deviating data cells.
Technometrics, **60**:2 135–145, 2018

Todorov V., Templ M. and Filzmoser P.
Detection of multivariate outliers in business survey data with incomplete information.
Advances in Data Analisys and Classification, **5**, 37–56, 2011.

Todorov, V.
The R package ecosystem for robust statistics.
Wiley Interdisciplinary Reviews: Computational Statistics, **16**:6 e70007, 2024

References II

Unwin, A.
Multivariate outliers and the O3 plot.
Journal of Computational and Graphical Statistics, **28**:3 635–643, 2019

Wada, K., Kawano, M. and Tsubaki, H.
Comparison of multivariate outlier detection methods for nearly elliptical distributions.
Austrian Journal of Statistics, **49**:2 1–17, 2020