Matthias Templ
TU WIEN & Statistics Austria
UNIDO, Study Visit OMAN, 2015

Outlier Detection

Overview

- Outlier detection
  - why outliers are such important
  - influence of outliers on analysis

And for later discussions we touch

- Imputation
- Statistical Disclosure Control

Outliers, Missing Values, SDC, ...

Effect of outliers: Regression

- outlier *Grand Prix*: Influence of outliers in regression

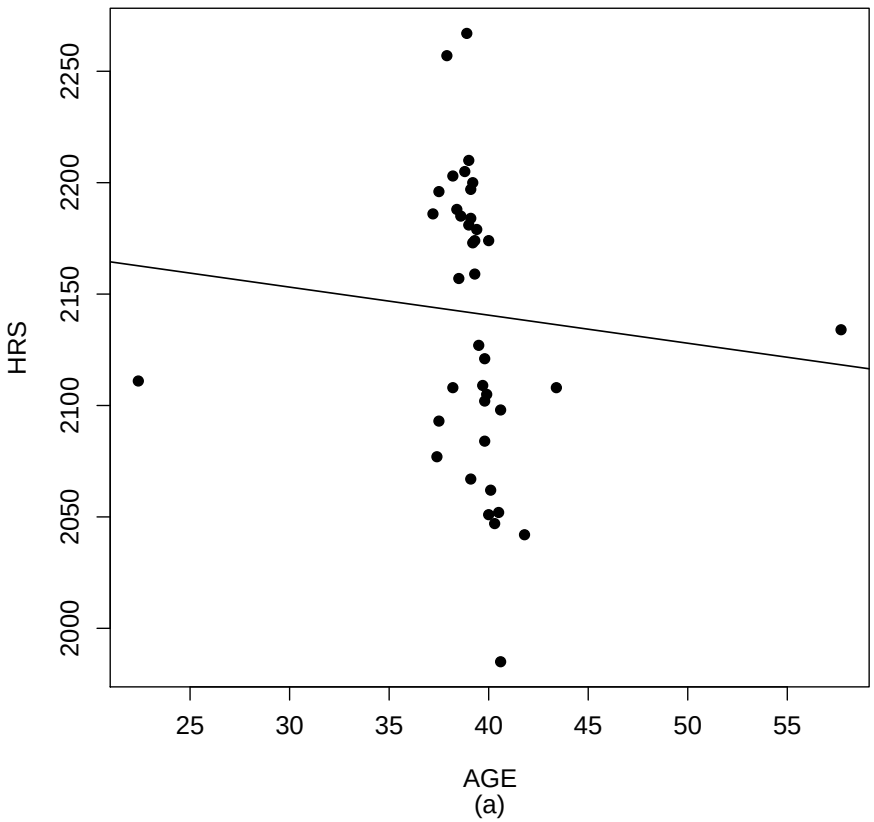
Effect of outliers: Regression

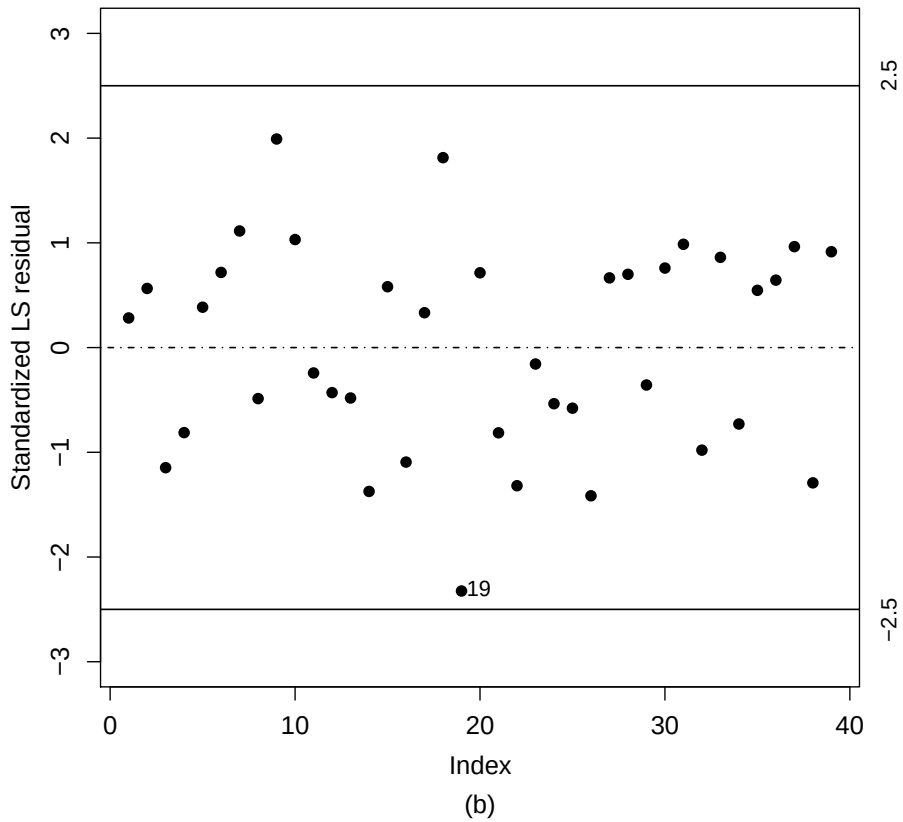
- Wages and Hours - available from http://lib.stat.cmu.edu/DASL/
- a national sample of 6000 households with a male head earning less than $15,000 annually in 1966
- 9 independent variables
- classified into 39 demographic groups
- estimate $y =$ the labor supply (average hours) from the available data (for the sake of example we will consider only one regressor variable: $x =$ average age of the respondents)

$$y = \beta_0 + \beta_1 x$$

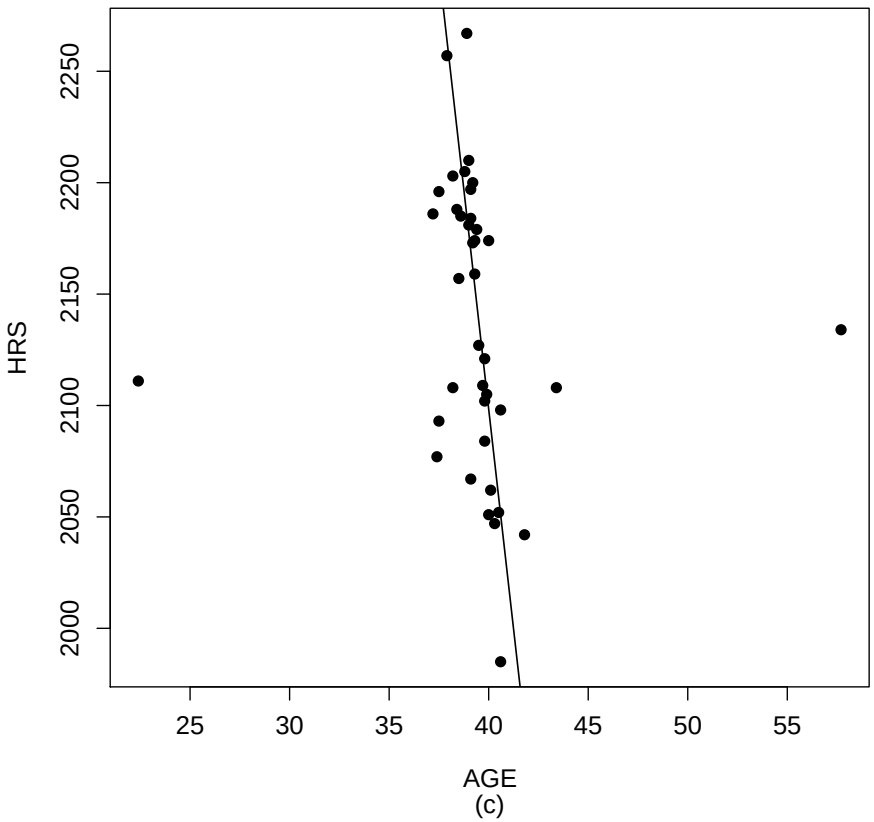
- We will fit an Ordinary Least Squares (OLS) and a robust Least Trimmed Squares (LTS) model

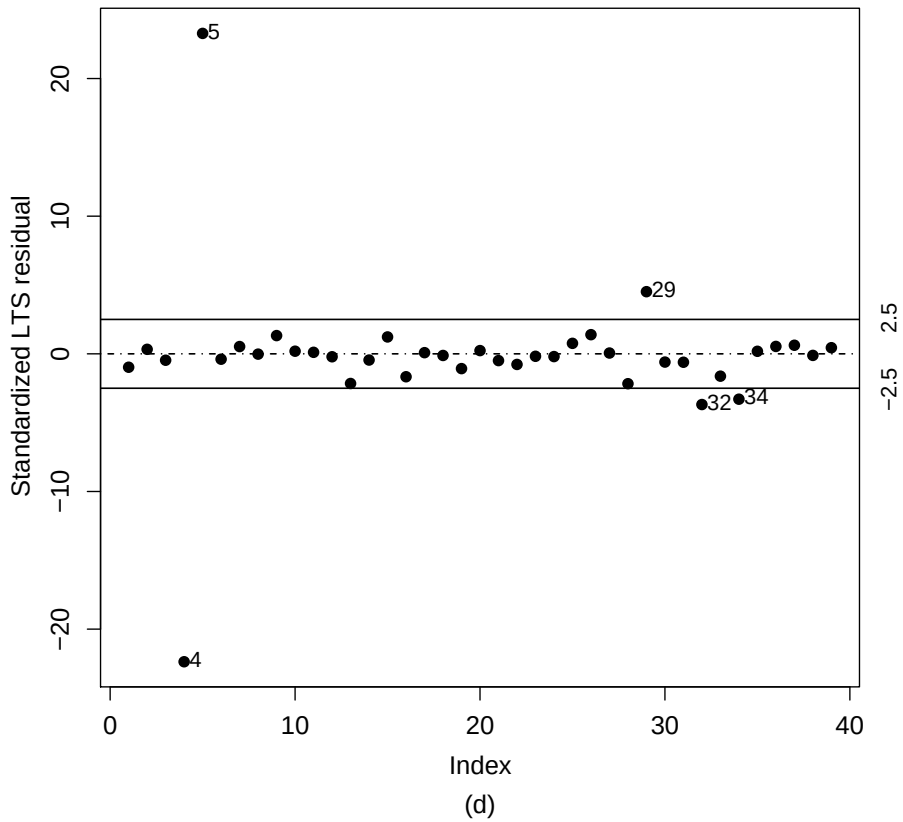
Effect of outliers: Regression - OLS





Effect of outliers: Regression - LTS





Effect of outliers: mean

STATISTIK AUSTRIA
Die Informationsmanager

(Loading Video...)

Effect of outliers: Gini

(Loading Video...)

Effect of outliers: Gini

(Loading Video...)

Example Worldbank

- World Bank has standardized a large collection of about 200 survey datasets from 90 developing countries.
- Purchasing Power Parity (PPP) estimates
- 2006 with complete other ranking than 2010 $\rightarrow$ only because of outliers
- Automated outlier detection and imputation

Example: Structural Business Statistics

- Editing data is highly time-consuming and resource intensive and often requires a significant part of the costs of the whole survey
- In case of the SBS data in Statistics Austria: a dozen of people check the data for errors
- Usually they look at thousands of tables to find errors

In the following, we briefly discuss two alternatives that may save a lot of manual work.

Deterministic approach: constrained automatic editing

- Automatic editing to edit the data in an optimized manner (automatically), given defined constraints (e.g. energy consumption + raw materials consumption $\leq$ intermediate consumption; ISIC XY + size 5: energy consumption $\leq$ c)
- Leads to overcorrection. Small errors from small enterprises have almost no effect on the ultimately published aggregated results but corrections can lead to biases in the estimates
- editing rules determined by subject matter specialists may destroy the multivariate structure of the data

edits (in R)

```
## Edit matrix:
##       x y  z Ops CONSTANT
## num1  1 3 -2  ==        0
## num2 -1 0  1  <=        0
##
## Edit rules:
## num1 : x + 3*y == 2*z
## num2 : z <= x
## Summary of normalized editmatrix
##      count
## block edits equalities inequalities variables
##     1     2          1            1         3
```

violatedEdits

```
##   x y z
## 1 0 0 0
## 2 2 0 1
## 3 1 1 1
##       edit
## record  num1  num2
##      1 FALSE FALSE
##      2 FALSE FALSE
##      3  TRUE FALSE
```

Distinguishment between

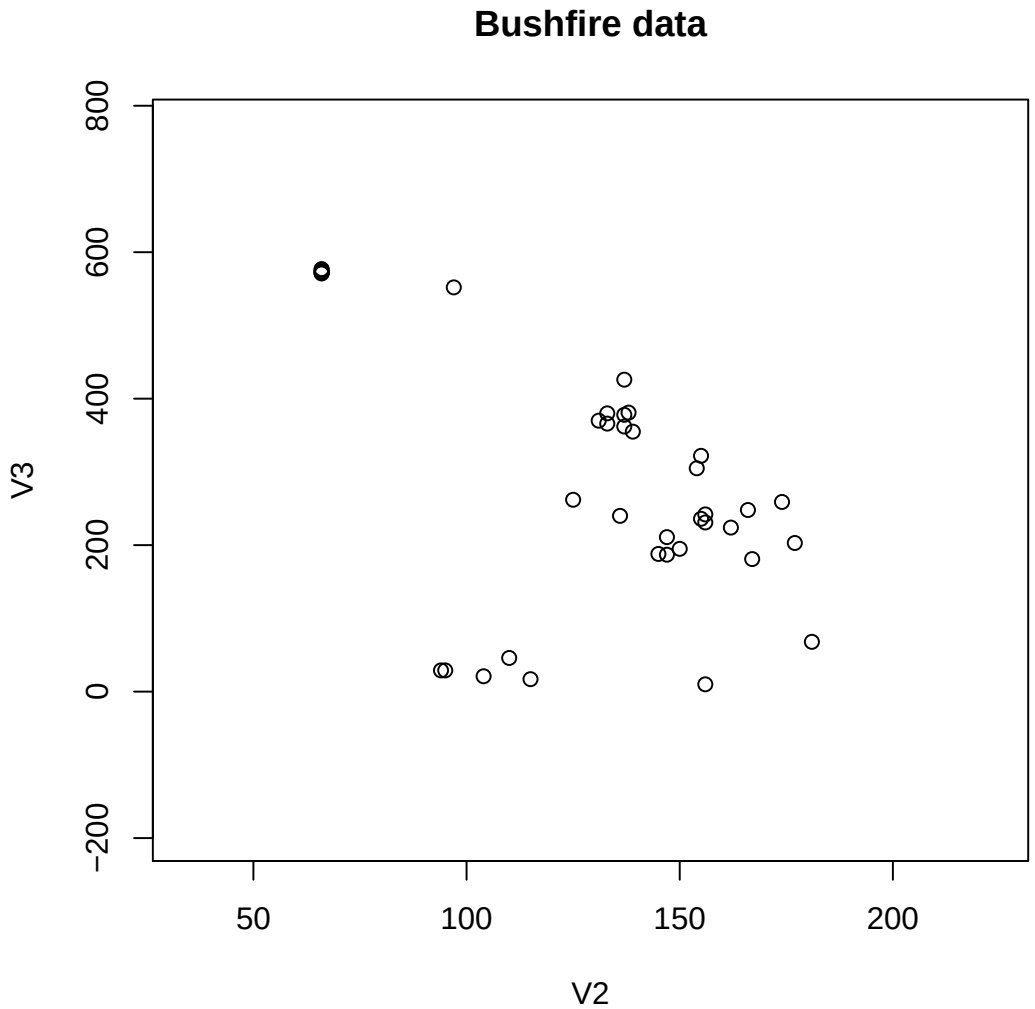
- univariate methods
- multivariate methods

Example: Bushfire data

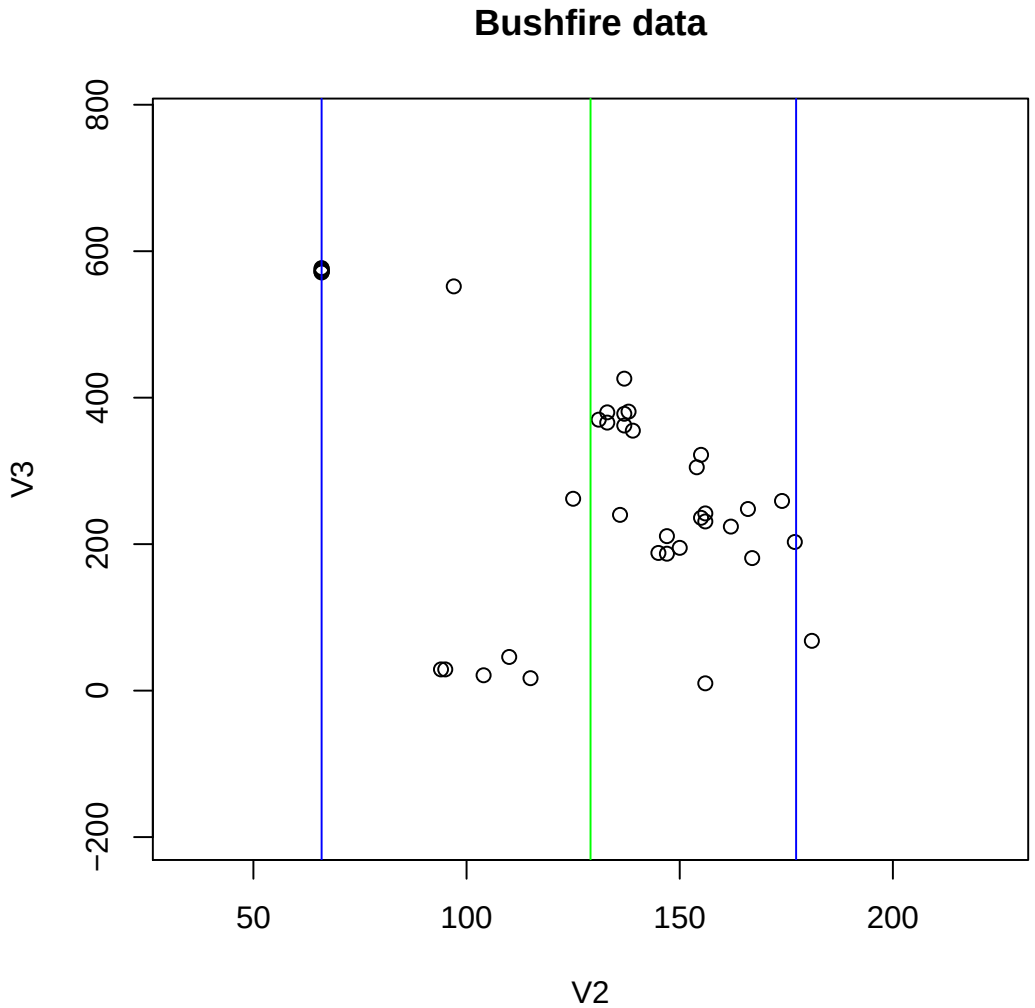
STATISTIK AUSTRIA
Die Informationsmanager

- 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`

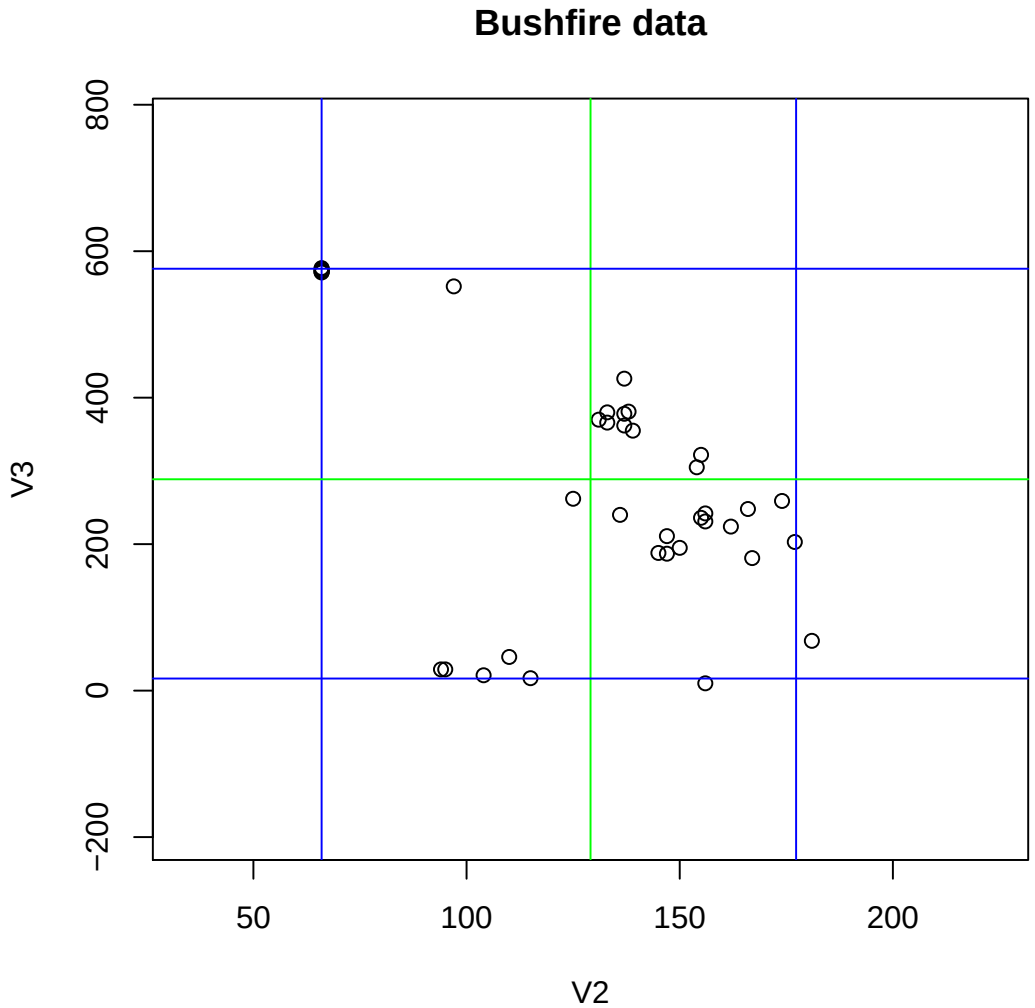
Example: Bushfire data



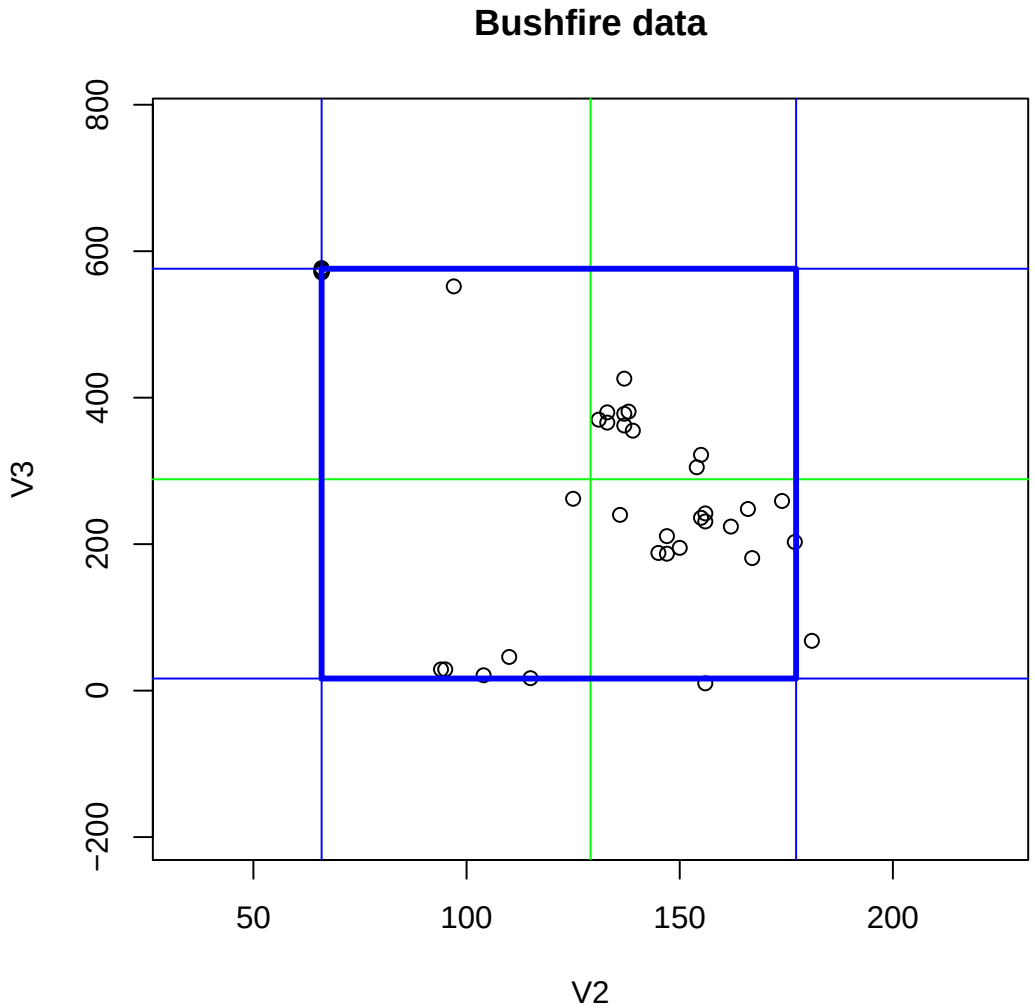
Example: Bushfire data



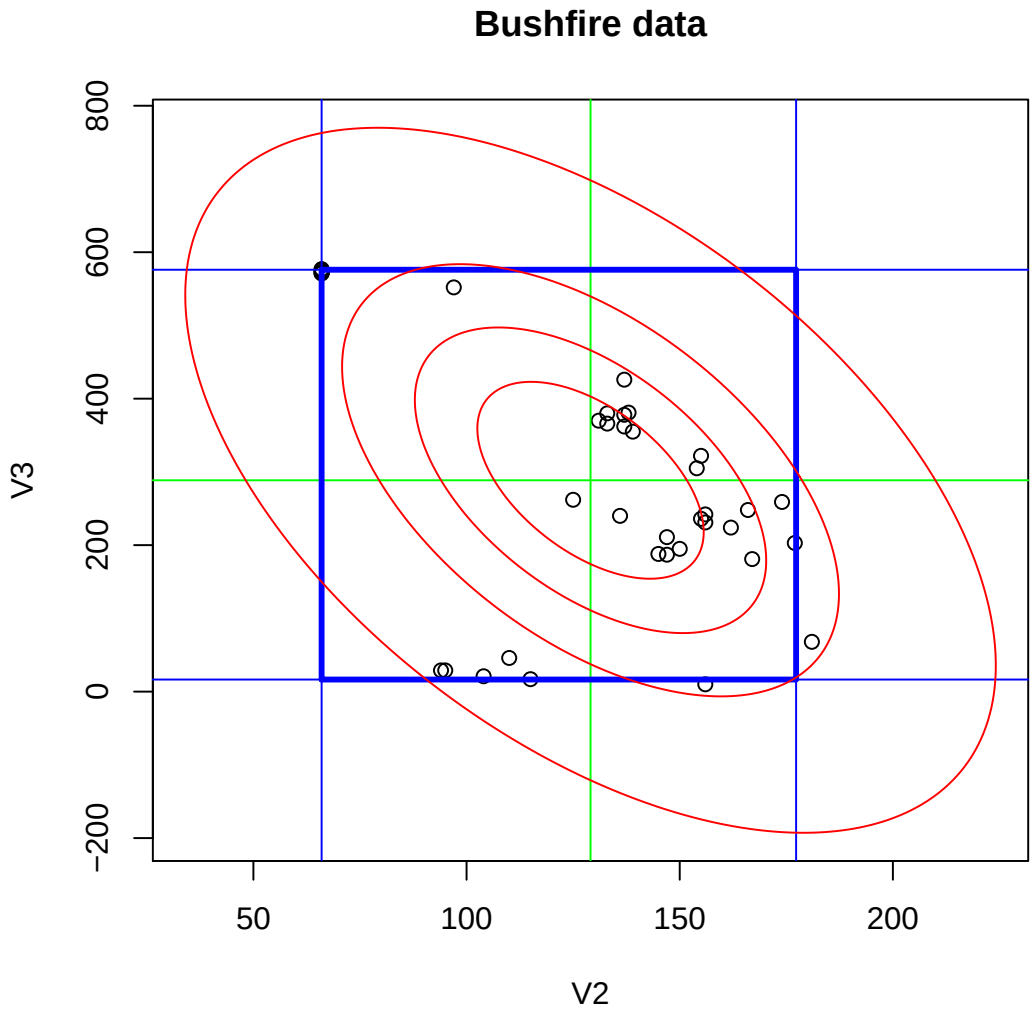
Example: Bushfire data



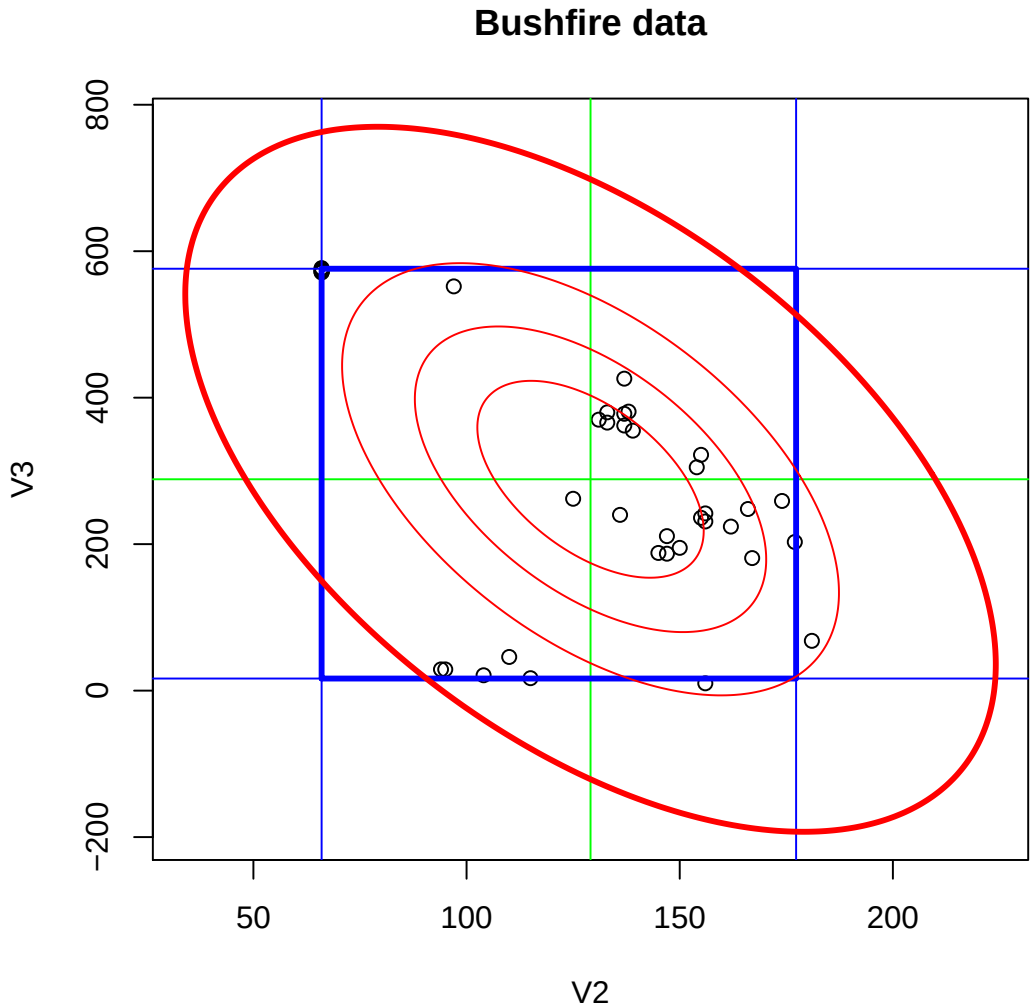
Example: Bushfire data



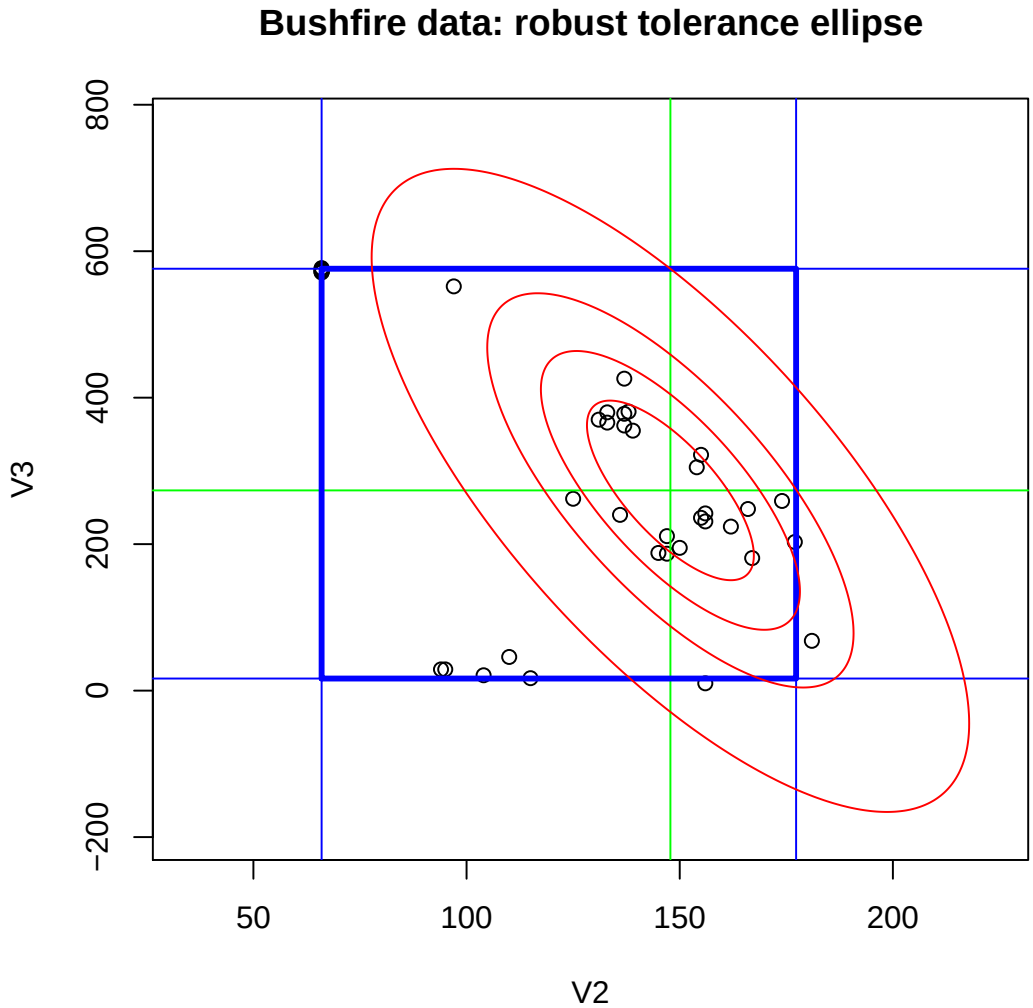
Example: Bushfire data



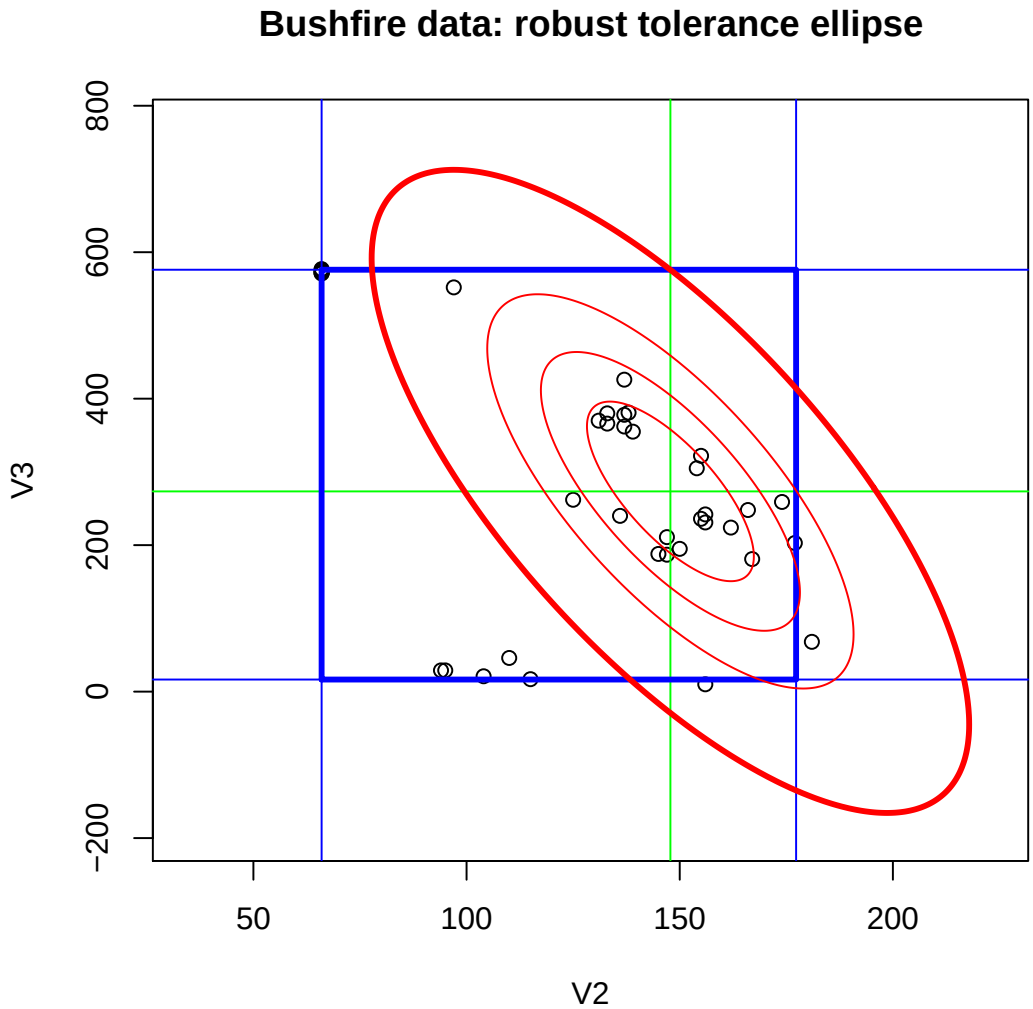
Example: Bushfire data



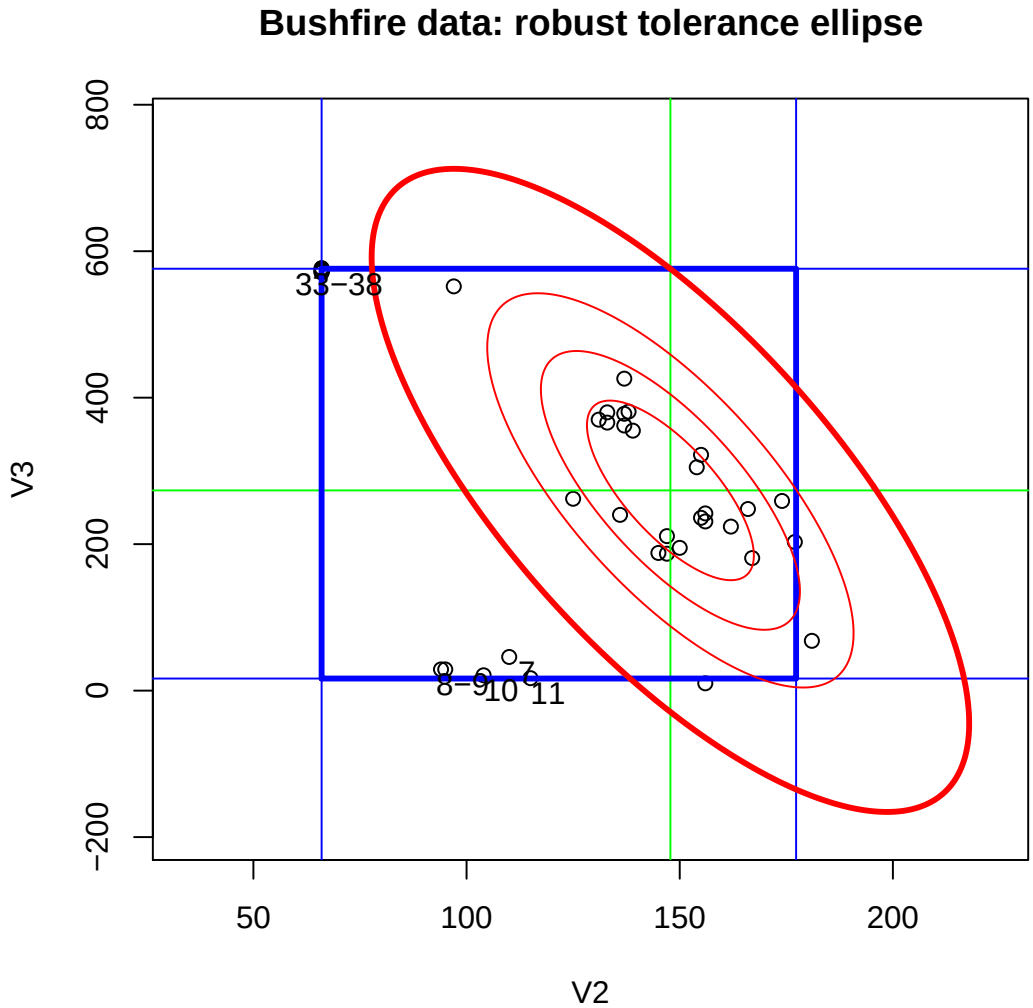
Example: Bushfire data



Example: Bushfire data



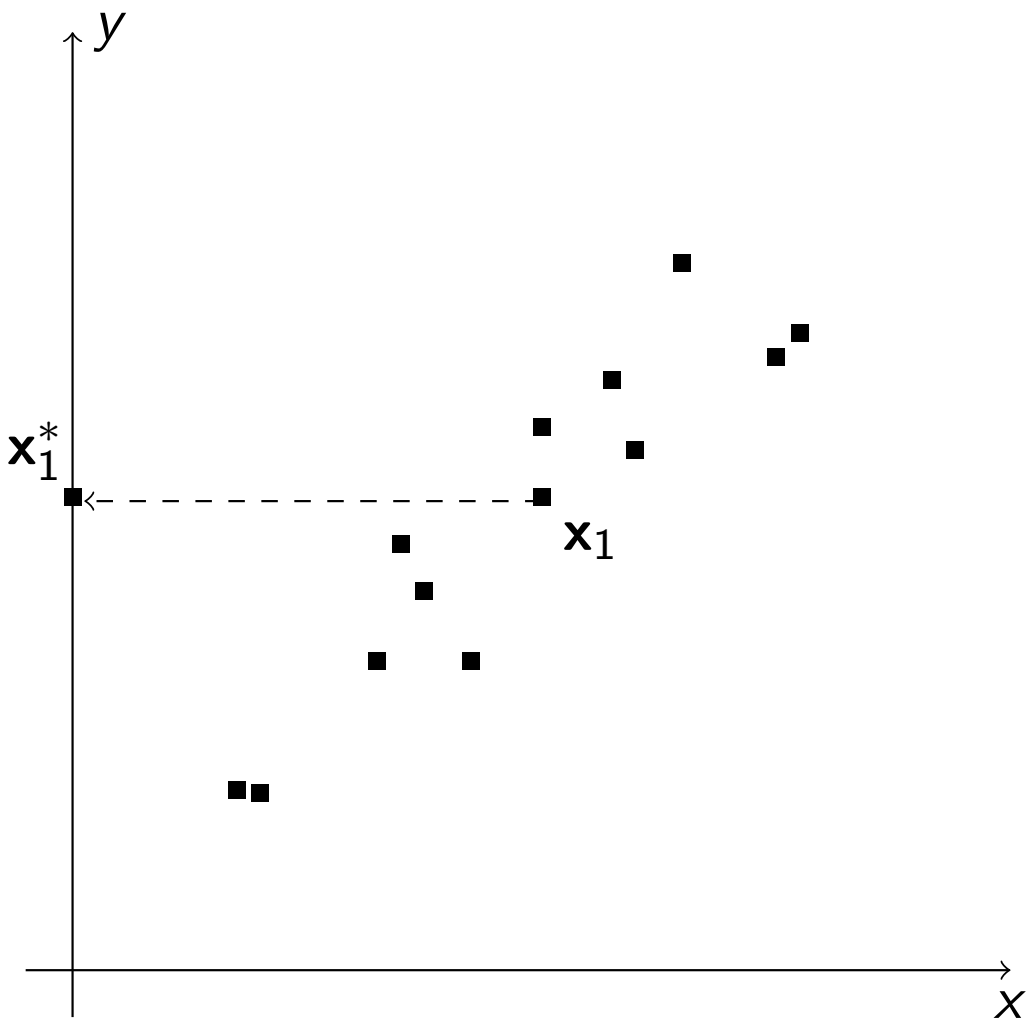
Example: Bushfire data



Additional Problems: Missing Values

- Missing values are imputed within the main bulk of the data, i.e. imputations should be non-outliers.
- Afterwards the outlier detection is applied on the imputed data.

Additional Problems: Zeros



00

$\mathbf{x}_1$: it's value on the x-axis is replaced by a zero ($\mathbf{x}_1^*$). The observation become immediately a multivariate outlier.

Ratios and Subgroups

- Not only absolute values are of interest to check for errors
- (log-) Ratios between variables might also be checked
- E.g. ratio of Energy consumption and materials consumption

Outlier detection may be applied in subgroups

Correction of Outliers

- Outlier detection methods flags observations/values $\sim$ potential outliers
- A potential outlier might be true
  - but may have large influence on indicators/estimations and variances
  - if corrected, bias is introduced
  - trade-off between bias and variance
- A potential outlier might be a measurement error
  - contact the enterprise and update values or
  - impute the measurement error

Imputation

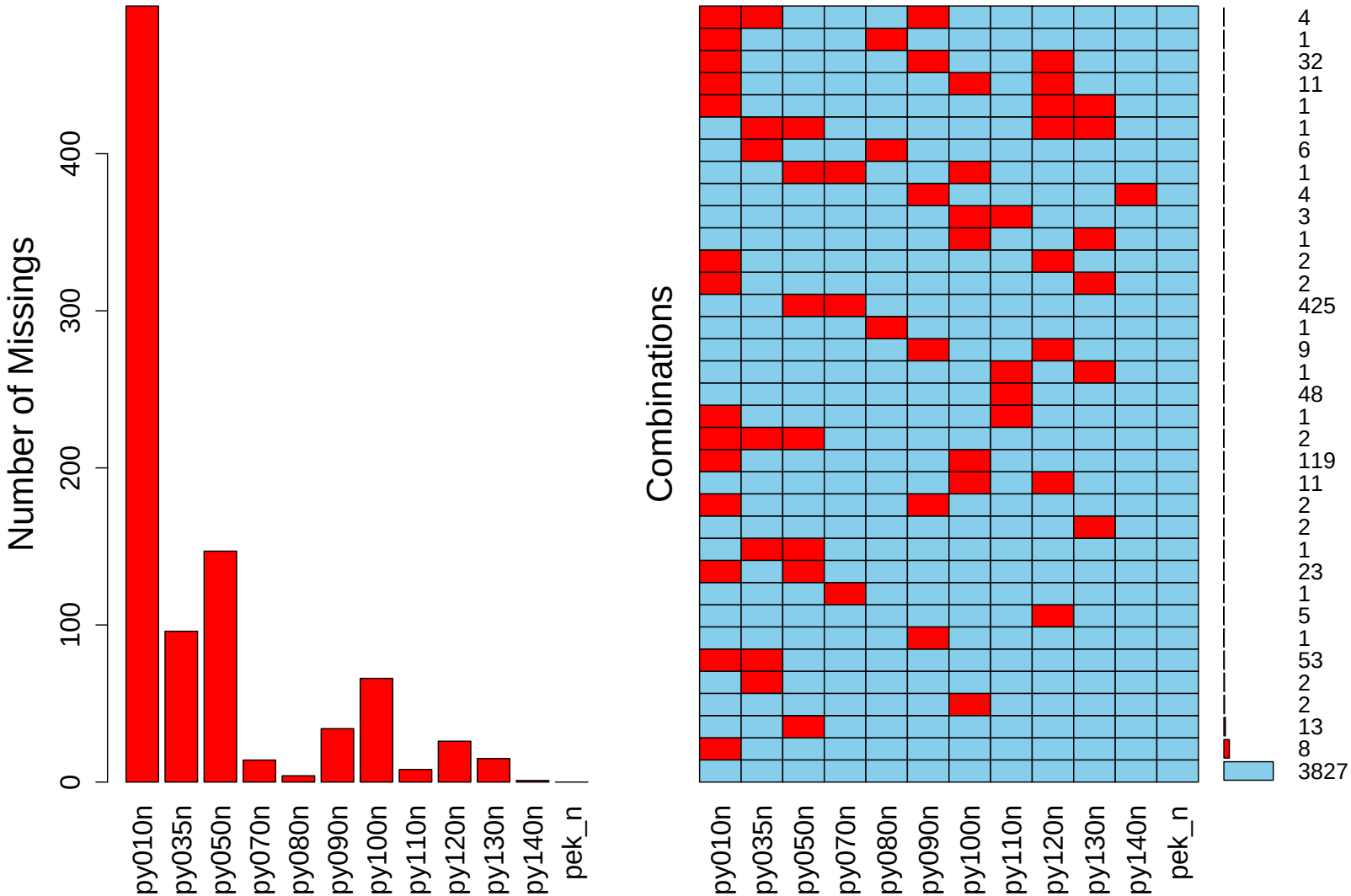
Imputation

- To replace item non-responses
- To impute the values of small enterprises/establishements
- To possible impute measurement errors

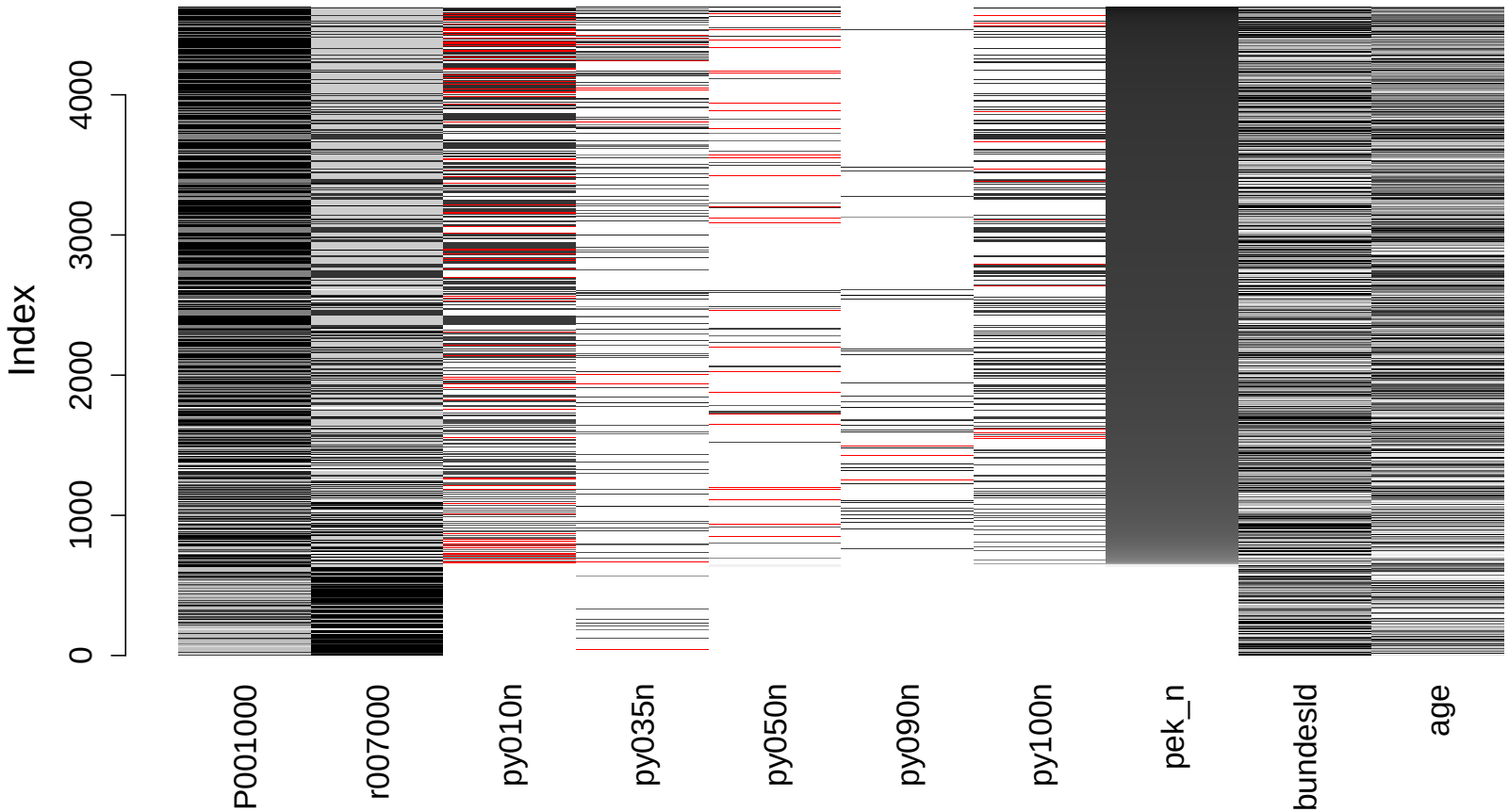
Motivation

- **Visualization** before and after the imputation helps to understand the mechanisms behind the specific missing data problems
- **efficient** implementations for different imputation methods needed (R package **VIM**)
- To impute data which include binary, nominal, categorical, ordered categorical, continuous, count, semi-continous variables

Aggregation



Missing values in matrix form



Note that another approx. 15 plot methods for missing values are available in R package VIM.

Overview of imputation methods 1/2

Single vs. multiple imputation

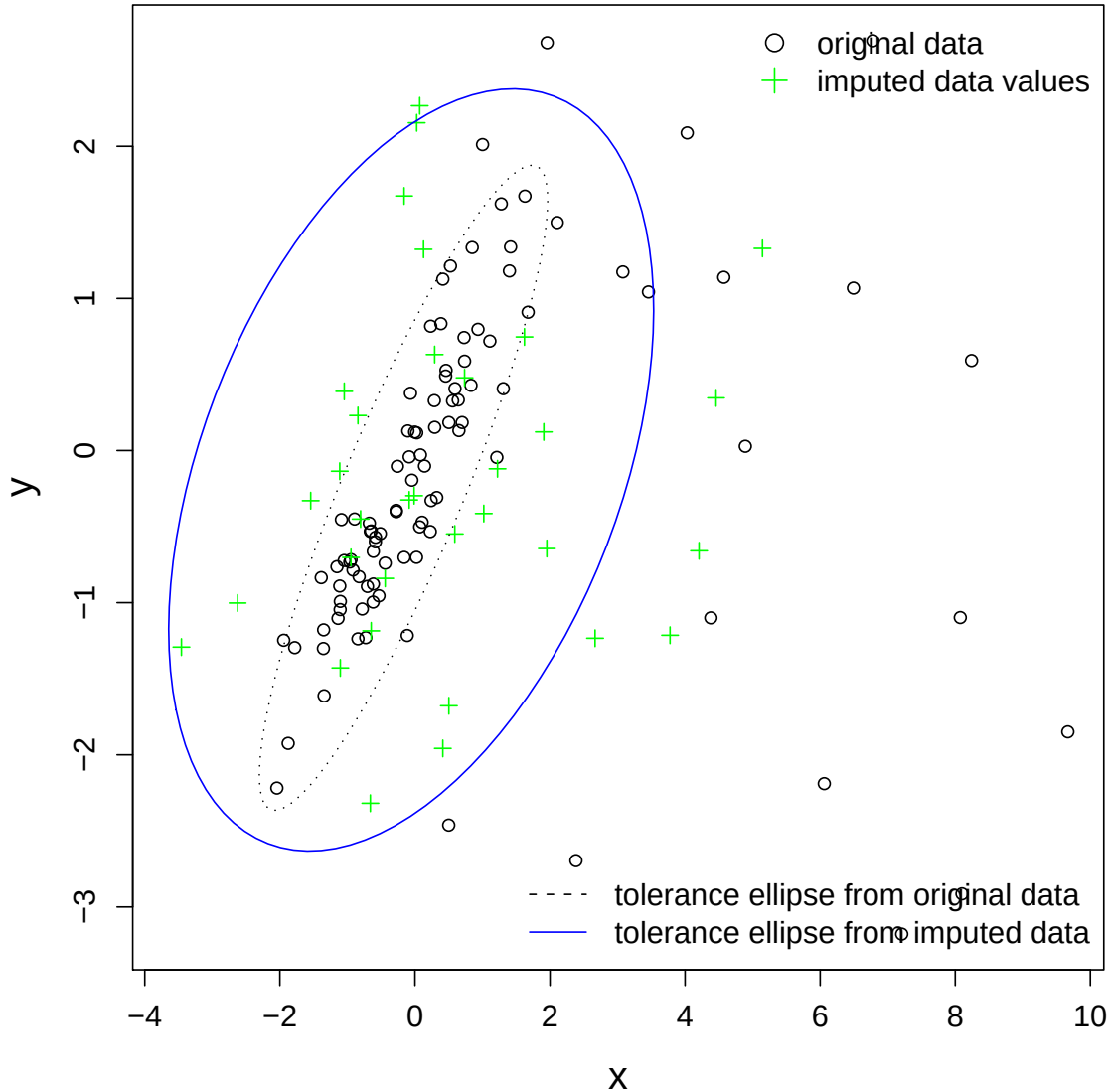
- **Single Imputation**: One (best) imputed value for each missing value
- **Multiple Imputation**: Every missing value is imputed multiple times (e.g. drawn from a predictive distribution)

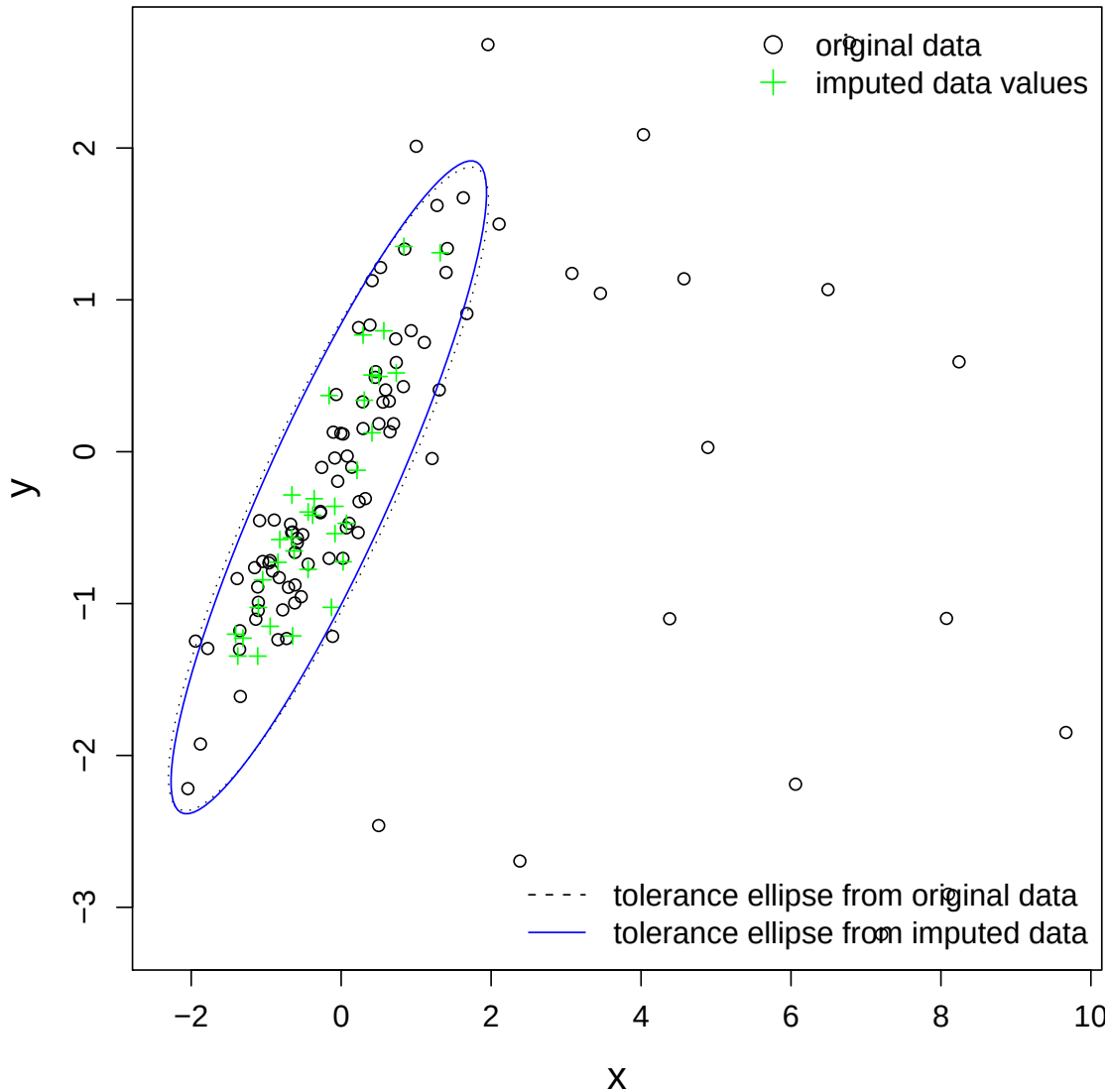
Overview of imputation methods 2/2

The methods can be grouped in:

- **Deductive imputation**: Imputation based on deductive rules
- (**Mean imputation)**: Imputation with the mean/median of a domain or the whole sample.
- **Random imputation**: Imputation by drawing from a random distribution
- **Donor imputation**: The imputed value is taken from another unit
  - random hot deck (from the same domain)
    example: Use a random enterpris from the same region, ISIC code and size class and use its turnover to impute a missing turnover value.
  - sequential hot deck
    sort the data set first.
  - *k*-nearest neighbor imputation
- **Model-based imputation**: A model is used to impute the missing values

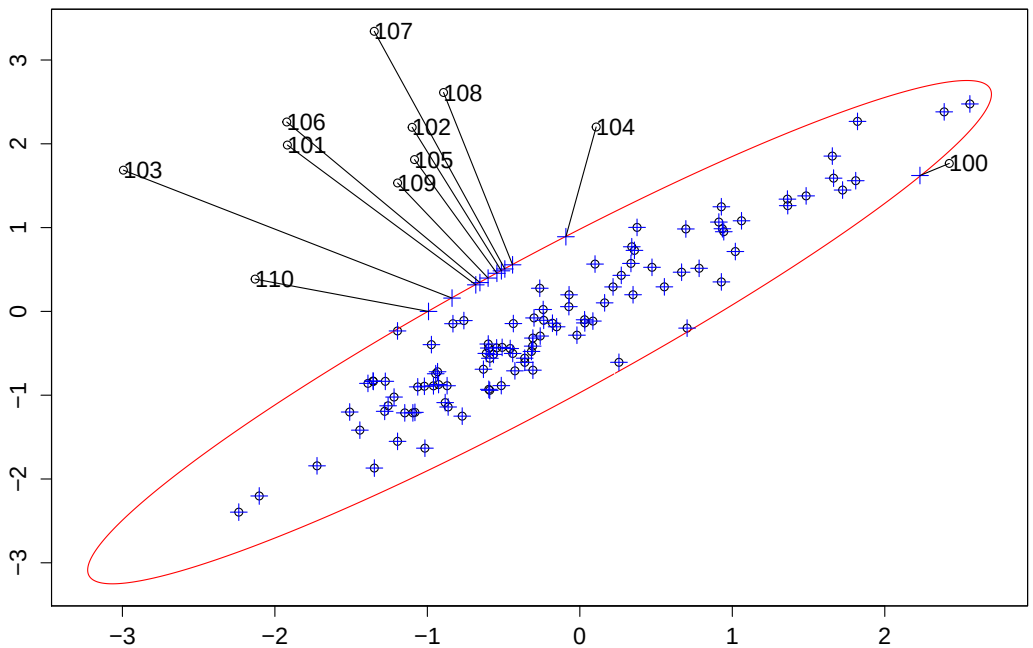
Non-robust Imputation




original data
imputed data values
tolerance ellipse from original data
tolerance ellipse from imputed data
y
x
−3
−2
−1
0
1
2
−2
0
2
4
6
8
10

Winsorizing

- Outliers might be not imputed with the mentioned imputation methods, but they might be winsorized.
- Winsorising them onto the boundaries of a 0.975% tolerance ellipse



Depending on national laws on privacy, microdata might not be delivered to researchers, other institutions or the public.

- Remote execution
- Remote access
- Safe data centres
- Anonymization of Microdata for Scientific-Use-Files
- (Heavy) anonymization of Microdata for Public-Use-Files
- Perturbation or synthetic data

Key Variables and Frequencies

- deletion of direct identifiers like enterprise name and address is not enough
- enterprise might be identified based on a combination of variables, such as ISIC code $\times$ size class of enterprise $\times$ ...
- after re-identification, the user knows every value of the enterprise in the data, thus also confidential information
- Frequencies of keys $\geq k$
- Large enterprises might always be identified

Conclusion

- Outlier detection to find possible outliers
- Correction of outliers to reduce the influence of true outliers or to replace measurement errors
- Multivariate methods are preferable, for outlier detection and imputation
- Confidentiality is a big issue. If violated, the trust in statistical institutions is lost.

Time for more details within a discussion

Thank you for your attention.

Thanks to Valentin Todorov for few slides taken from our tutorial on *R in the Statistical Office*

Contact:
matthias.templ@tuwien.ac.at