IMPUTATION OF COMPLEX DATA WITH R-PACKAGE VIM: TRADITIONAL AND NEW METHODS BASED ON ROBUST ESTIMATION.

Key Invited Paper

Matthias Templ[1,2], Alexander Kowarik[1], Peter Filzmoser[2]

[1] Department of Methodology, Statistics Austria
[2] Department of Statistics and Probability Theory, TU WIEN, Austria

UNECE Worksession on data editing
Ljubljana, Mai 10, 2011

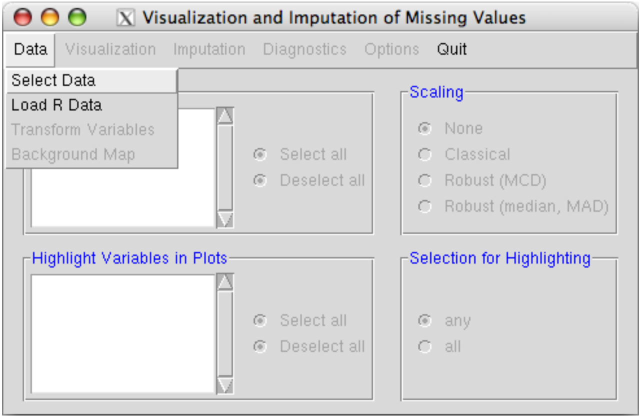
Content

1. Overview
2. Visualisation Tools
3. Robust Imputation: Motivation
4. Challenges
5. IRMI
6. Simulation results
7. Results from real-world data

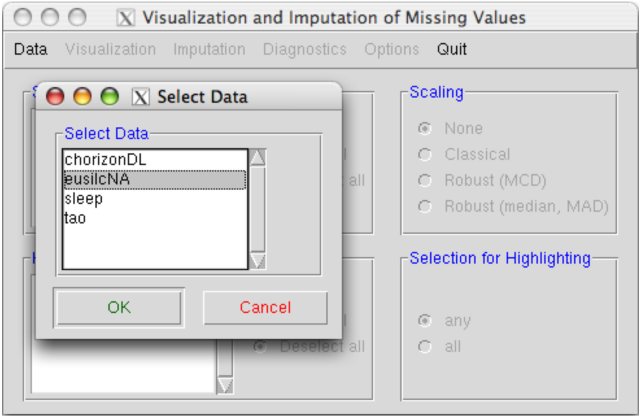
R-package VIM

- **VIM** = **V**isualization and **I**mputation of **M**issings
- Univariate, bivariate, multiple and multivariate plot methods to highlight missing values in complex data sets to learn about their structure (MCAR, MAR, MNAR). Comes with a GUI as well.
- Hot-deck, k-NN and EM-based (robust) imputation methods for complex data sets. Due to time reasons we mostly concentrate on EM-based imputation. For hot-deck and k-NN, please have a look at the paper.
- VIM-book in 2012

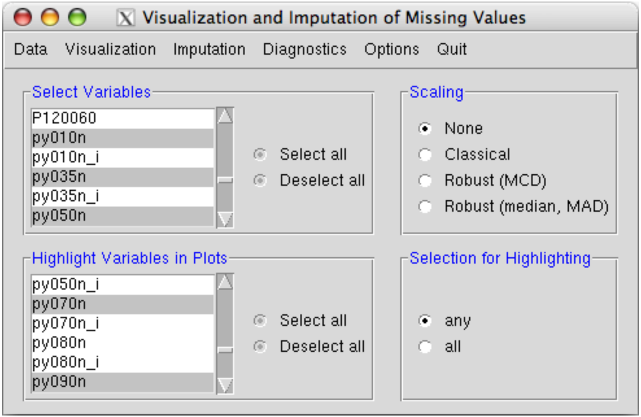
The GUI . . .


Visualization and Imputation of Missing Values
Data
Visualization
Imputation
Diagnostics
Options
Quit
Select Data
Load R Data
Transform Variables
Background Map
Select all
Deselect all
Scaling
None
Classical
Robust (MCD)
Robust (median, MAD)
Highlight Variables in Plots
Select all
Deselect all
Selection for Highlighting
any
all

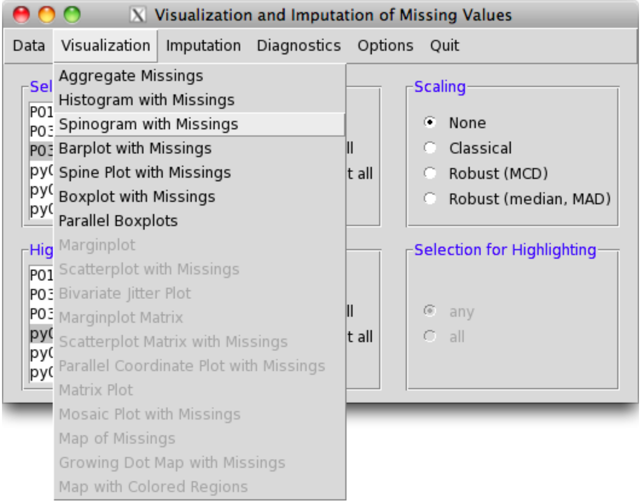
The GUI . . .


Visualization and Imputation of Missing Values
Data
Visualization
Imputation
Diagnostics
Options
Quit
Select Data
Select Data
chorizonDL
eusilcNA
sleep
tao
OK
Cancel
Scaling
None
Classical
Robust (MCD)
Robust (median, MAD)
Selection for Highlighting
any
all
Deselect all

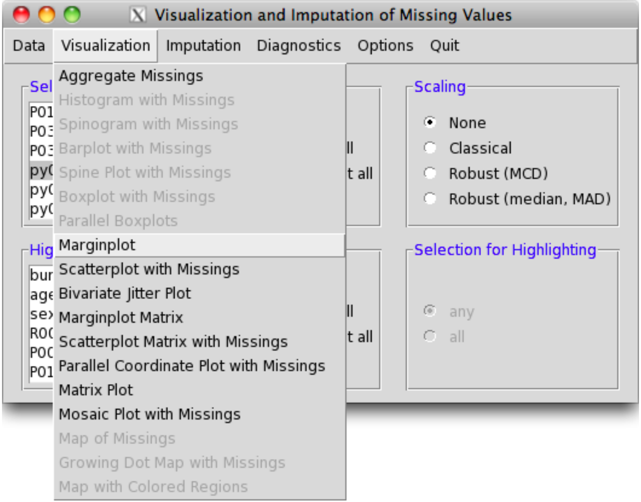
The GUI . . .


Visualization and Imputation of Missing Values
Data
Visualization
Imputation
Diagnostics
Options
Quit
Select Variables
P120060
py010n
py010n_i
py035n
py035n_i
py050n
Select all
Deselect all
Scaling
None
Classical
Robust (MCD)
Robust (median, MAD)
Highlight Variables in Plots
py050n_i
py070n
py070n_i
py080n
py080n_i
py090n
Select all
Deselect all
Selection for Highlighting
any
all

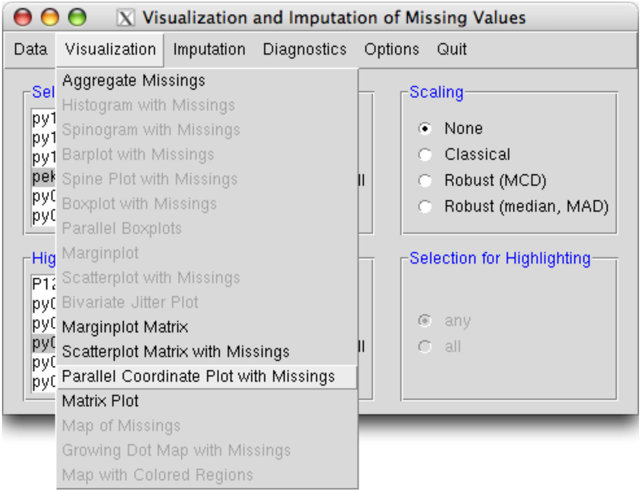
The GUI . . .


Visualization and Imputation of Missing Values
Data
Visualization
Imputation
Diagnostics
Options
Quit
Aggregate Missings
Histogram with Missings
Spinogram with Missings
Barplot with Missings
Spine Plot with Missings
Boxplot with Missings
Parallel Boxplots
Marginplot
Scatterplot with Missings
Bivariate Jitter Plot
Marginplot Matrix
Scatterplot Matrix with Missings
Parallel Coordinate Plot with Missings
Matrix Plot
Mosaic Plot with Missings
Map of Missings
Growing Dot Map with Missings
Map with Colored Regions
Scaling
None
Classical
Robust (MCD)
Robust (median, MAD)
Selection for Highlighting
any
all

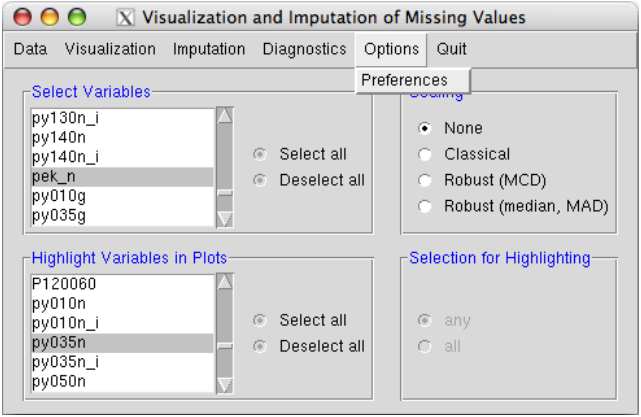
The GUI . . .


Visualization and Imputation of Missing Values
Data
Visualization
Imputation
Diagnostics
Options
Quit
Aggregate Missings
Histogram with Missings
Spinogram with Missings
Barplot with Missings
Spine Plot with Missings
Boxplot with Missings
Parallel Boxplots
Marginplot
Scatterplot with Missings
Bivariate Jitter Plot
Marginplot Matrix
Scatterplot Matrix with Missings
Parallel Coordinate Plot with Missings
Matrix Plot
Mosaic Plot with Missings
Map of Missings
Growing Dot Map with Missings
Map with Colored Regions
Scaling
None
Classical
Robust (MCD)
Robust (median, MAD)
Selection for Highlighting
any
all

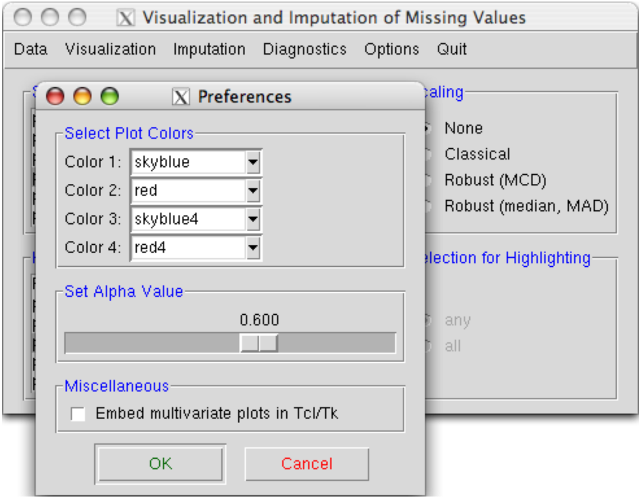
The GUI . . .


Visualization and Imputation of Missing Values
Data
Visualization
Imputation
Diagnostics
Options
Quit
Aggregate Missings
Histogram with Missings
Spinogram with Missings
Barplot with Missings
Spine Plot with Missings
Boxplot with Missings
Parallel Boxplots
Marginplot
Scatterplot with Missings
Bivariate Jitter Plot
Marginplot Matrix
Scatterplot Matrix with Missings
Parallel Coordinate Plot with Missings
Matrix Plot
Map of Missings
Growing Dot Map with Missings
Map with Colored Regions
Scaling
None
Classical
Robust (MCD)
Robust (median, MAD)
Selection for Highlighting
any
all

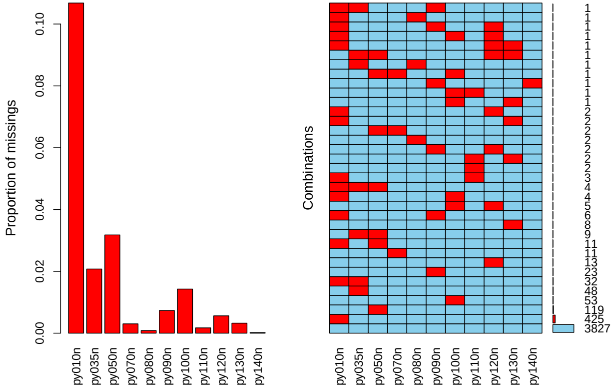
The GUI . . .


Visualization and Imputation of Missing Values
Data
Visualization
Imputation
Diagnostics
Options
Quit
Preferences
Select Variables
py130n_i
py140n
py140n_i
pek_n
py010g
py035g
Select all
Deselect all
None
Classical
Robust (MCD)
Robust (median, MAD)
Highlight Variables in Plots
P120060
py010n
py010n_i
py035n
py035n_i
py050n
Select all
Deselect all
Selection for Highlighting
any
all

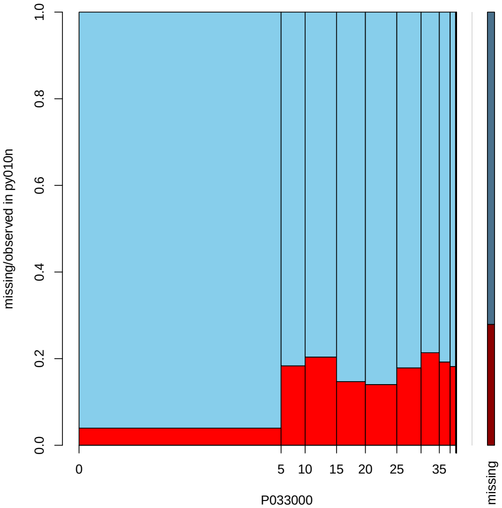
The GUI . . .


Visualization and Imputation of Missing Values
Data
Visualization
Imputation
Diagnostics
Options
Quit
Preferences
Select Plot Colors
Color 1: skyblue
Color 2: red
Color 3: skyblue4
Color 4: red4
Set Alpha Value
0.600
Miscellaneous
Embed multivariate plots in Tcl/Tk
OK
Cancel
None
Classical
Robust (MCD)
Robust (median, MAD)
any
all

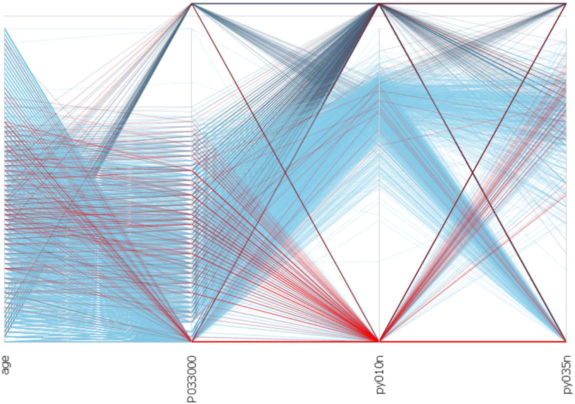
The GUI . . .


Proportion of missings
0.00
0.02
0.04
0.06
0.08
0.10
py010n
py035n
py050n
py070n
py080n
py090n
py100n
py110n
py120n
py130n
py140n
Combinations
1
1
1
1
1
1
1
1
1
1
1
2
2
2
2
2
2
2
3
4
4
5
6
8
9
11
11
13
23
32
48
53
119
425
3827

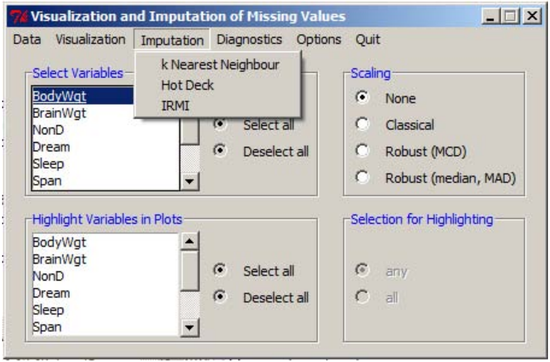
The GUI . . .


missing/observed in py010n
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
35
P033000
missing

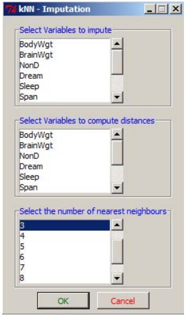
The GUI . . .



The GUI - Imputation

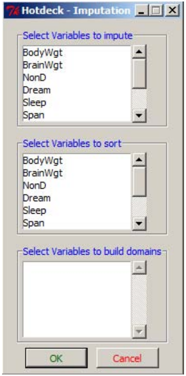


The GUI - Imputation



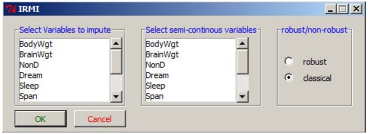

```
kNN(data, variable=colnames(data), metric=NULL, k=5,
  dist_var=colnames(data),weights=NULL,numFun = median,
  catFun=maxCat,makeNA=NULL,NAcond=NULL, impNA=TRUE,
  donorcond=NULL,mixed=vector(),trace=FALSE,
  imp_var=TRUE,imp_suffix="imp",addRandom=FALSE)
```

The GUI - Imputation



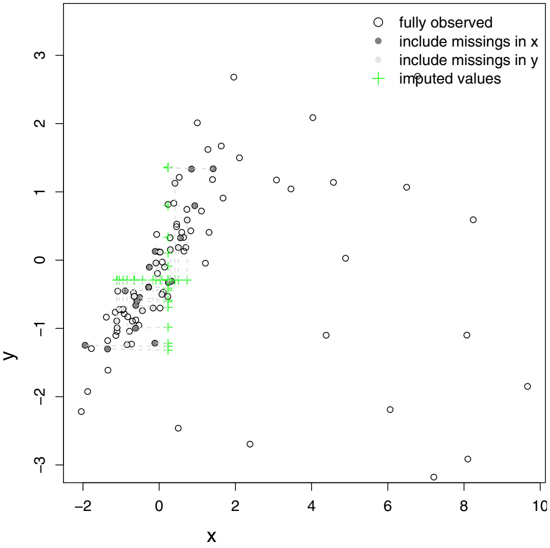

```
hotdeck(data, variable=colnames(data), ord_var=NULL,
  domain_var=NULL,makeNA=NULL,NAcond=NULL,impNA=TRUE,
  donorcond=NULL,imp_var=TRUE,imp_suffix="imp")
```

The GUI - Imputation

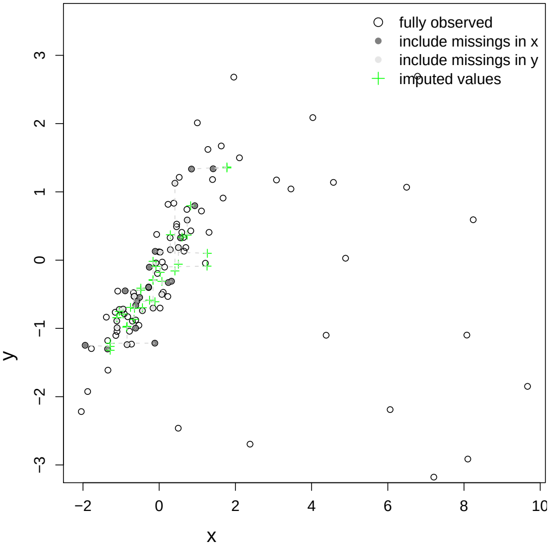



```
irmi(x, eps = 0.01, maxit = 100, mixed = NULL,
  step = FALSE, robust = FALSE, takeAll = TRUE,
  noise = TRUE, noise.factor = 1, force = FALSE,
  robMethod = "lmrob", force.mixed = TRUE,
  mi = 1, trace=FALSE)
```

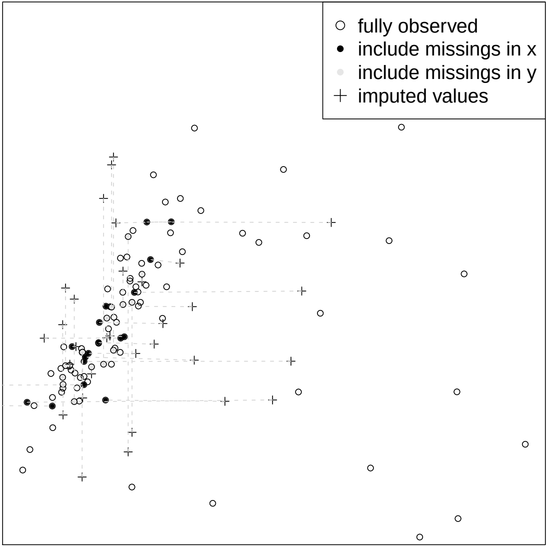
Median Imputation



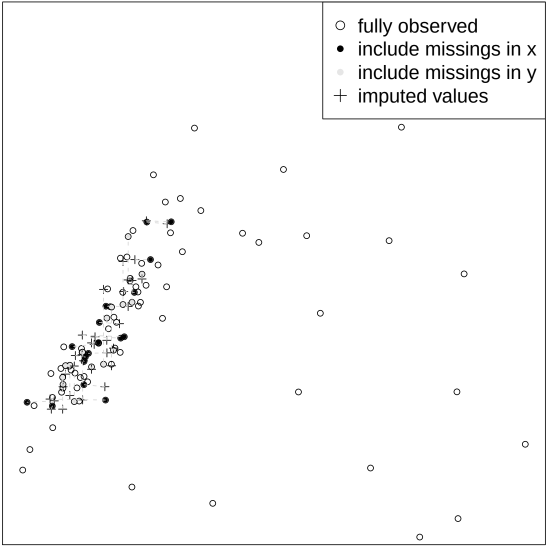
*k*NN Imputation



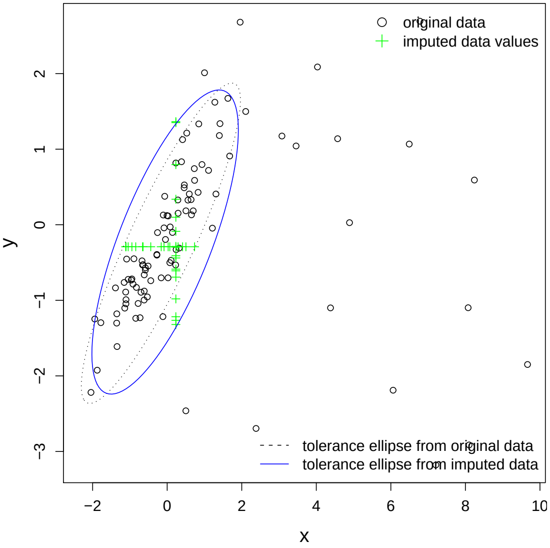
IVEWARE


fully observed
include missings in x
include missings in y
imputed values

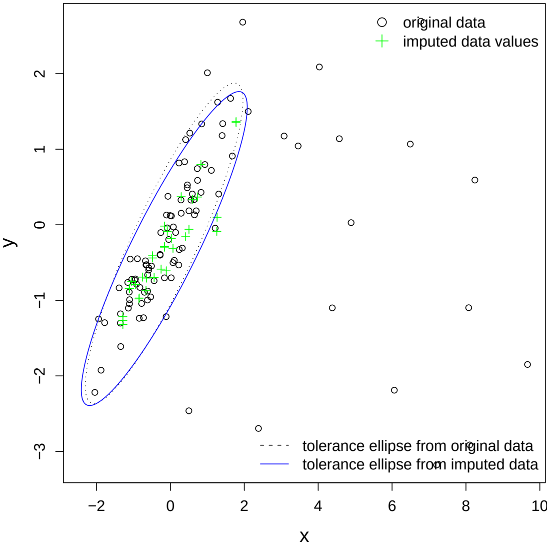
IRMI


fully observed
include missings in x
include missings in y
imputed values

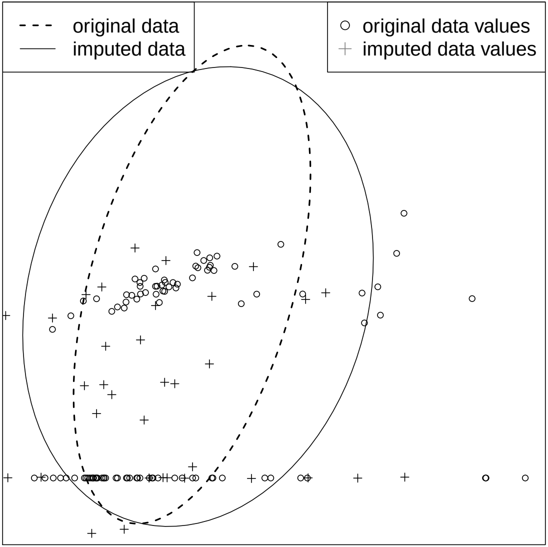
Median Imputation



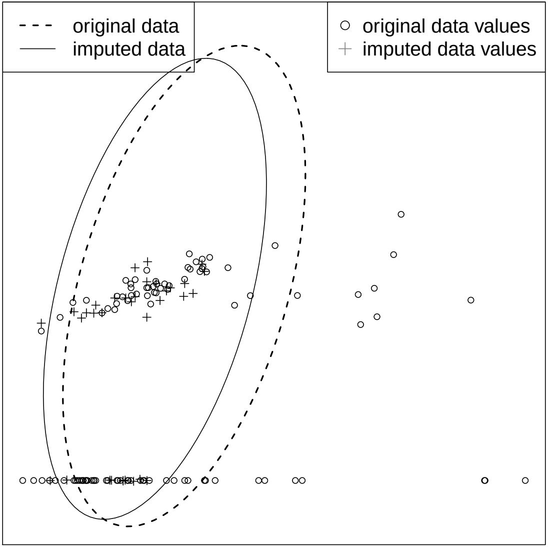
kNN Imputation



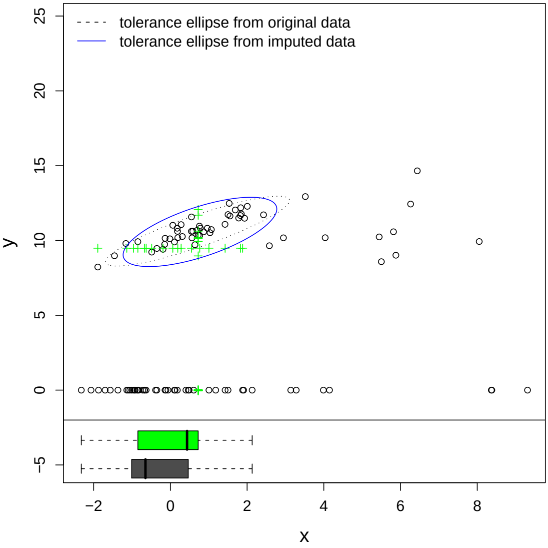
IVEWARE


original data
imputed data
original data values
imputed data values

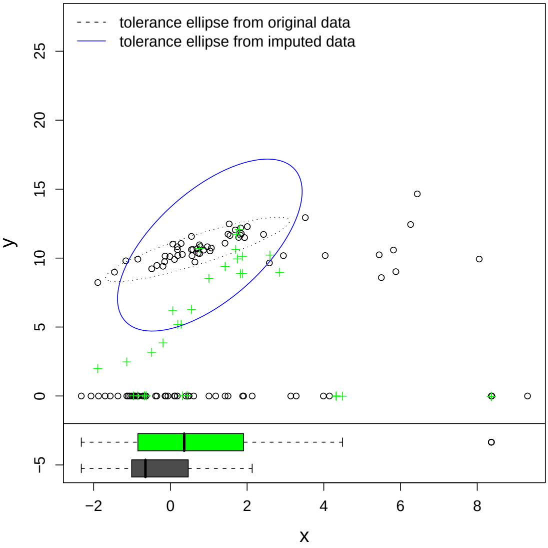
IRMI


original data
imputed data
original data values
imputed data values

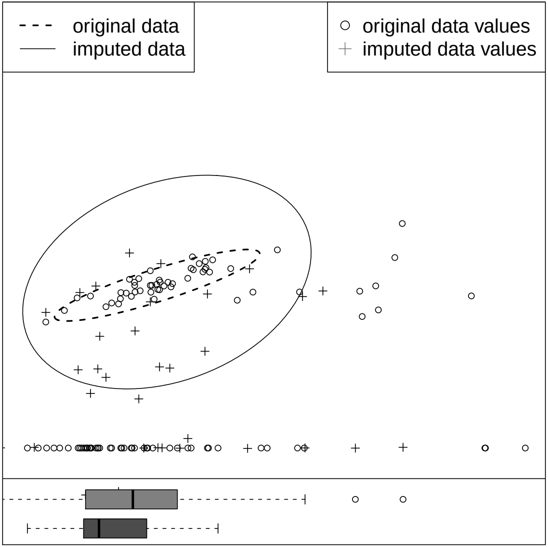
Median Imputation



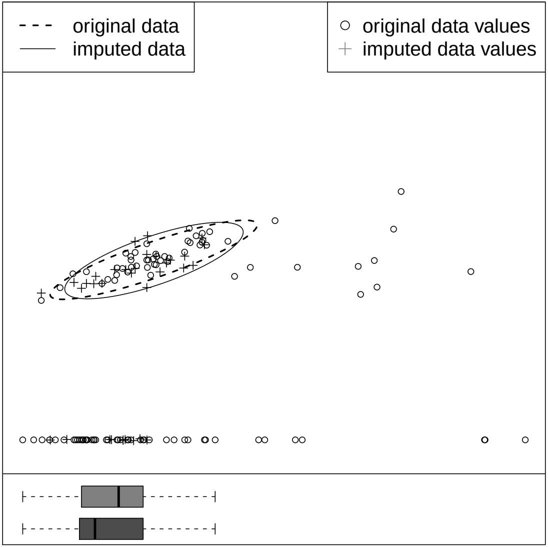
kNN Imputation



IVEWARE


original data
imputed data
original data values
imputed data values

IRMI


original data
imputed data
original data values
imputed data values

Some Challenges

Mixed type of variables: various variables being **nominal** scaled, some variables might be **ordinal** and some variables could be determined to be of **continuous** scale.

Semi-continuous variables: "**semi-continuous**" distributions, i.e. a variable consisting of a continuous scaled part and a certain proportion of equal values.

Far from normality: Virtually always outlying observations included in real-world data.

multiple imputation: Imputated must be both, reflect the multivariate structure of the data and including "randomness".

Some Challenges

Mixed type of variables: various variables being **nominal** scaled, some variables might be **ordinal** and some variables could be determined to be of **continuous** scale.

Semi-continuous variables: **"semi-continuous"** distributions, i.e. a variable consisting of a continuous scaled part and a certain proportion of equal values.

Far from normality: Virtually always outlying observations included in real-world data.

multiple imputation: Imputated must be both, reflect the multivariate structure of the data and including "randomness".

Some Challenges

Mixed type of variables: various variables being **nominal** scaled, some variables might be **ordinal** and some variables could be determined to be of **continuous** scale.

Semi-continuous variables: **"semi-continuous"** distributions, i.e. a variable consisting of a continuous scaled part and a certain proportion of equal values.

Far from normality: Virtually always outlying observations included in real-world data.

multiple imputation: Imputated must be both, reflect the multivariate structure of the data and including "randomness".

MIX, MICE, MI, MITOOLS, . . .

- All missing values imputed with simulated values drawn from their predictive distribution given the observed data and the specified parameter.
- $\longrightarrow$ based on sequential regressions.
- EM-based
- In general, there are often problems when applied to complex data sets.

. . . and, of course, they are highly driven by influencial points, representative and non-representative outliers.

MIX, MICE, MI, MITOOLS, . . .

- All missing values imputed with simulated values drawn from their predictive distribution given the observed data and the specified parameter.
- $\longrightarrow$ based on sequential regressions.
- EM-based
- In general, there are often problems when applied to complex data sets.

. . . and, of course, they are highly driven by influencial points, representative and non-representative outliers.

MIX, MICE, MI, MITOOLS, . . .

- All missing values imputed with simulated values drawn from their predictive distribution given the observed data and the specified parameter.
- $\longrightarrow$ based on sequential regressions.
- EM-based
- In general, there are often problems when applied to complex data sets.

. . . and, of course, they are highly driven by influencial points, representative and non-representative outliers.

MIX, MICE, MI, MITOOLS, ...

- All missing values imputed with simulated values drawn from their predictive distribution given the observed data and the specified parameter.
- $\longrightarrow$ based on sequential regressions.
- EM-based
- In general, there are often problems when applied to complex data sets.

...and, of course, they are highly driven by influencial points, representative and non-representative outliers.

MIX, MICE, MI, MITOOLS, . . .

- All missing values imputed with simulated values drawn from their predictive distribution given the observed data and the specified parameter.
- $\longrightarrow$ based on sequential regressions.
- EM-based
- In general, there are often problems when applied to complex data sets.

. . . and, of course, they are highly driven by influencial points, representative and non-representative outliers.

IVEWARE

- Very popular software used in many applications in Official Statistics.
- Similar to the previous mentioned methods.
- The imputations are obtained by fitting a sequence of (Bayesian) regression models and drawing values from the corresponding predictive distributions.
- **Sequentially imputation**: in each step, one variable serve as **response** and certain other variables serves as **predictors**. **Fit** a certain model using the observed part of the response and **estimate** (update) the (former) missing values in the response.
  - **Initialization loop:** ...
  - **Second outer loop:**
    Estimates of missing values are updated sequentially using one variable as response and all other variables as predictors until convergency.
- Since missing values are drawn from their predictive distribution given the observed data and the specified parameter, the procedure allows **multiple imputation**.

IVEWARE

- Very popular software used in many applications in Official Statistics.
- Similar to the previous mentioned methods.
- The imputations are obtained by fitting a sequence of (Bayesian) regression models and drawing values from the corresponding predictive distributions.
- **Sequentially imputation**: in each step, one variable serve as **response** and certain other variables serves as **predictors**. **Fit** a certain model using the observed part of the response and **estimate** (update) the (former) missing values in the response.
  - **Initialization loop:** ...
  - **Second outer loop:**
    Estimates of missing values are updated sequentially using one variable as response and all other variables as predictors until convergency.
- Since missing values are drawn from their predictive distribution given the observed data and the specified parameter, the procedure allows **multiple imputation**.

IVEWARE

- Very popular software used in many applications in Official Statistics.
- Similar to the previous mentioned methods.
- The imputations are obtained by fitting a sequence of (Bayesian) regression models and drawing values from the corresponding predictive distributions.
- **Sequentially imputation**: in each step, one variable serve as **response** and certain other variables serves as **predictors**. **Fit** a certain model using the observed part of the response and **estimate** (update) the (former) missing values in the response.
  - **Initialization loop:** ...
  - **Second outer loop:**
    Estimates of missing values are updated sequentially using one variable as response and all other variables as predictors until convergency.
- Since missing values are drawn from their predictive distribution given the observed data and the specified parameter, the procedure allows **multiple imputation**.

IVEWARE

- Very popular software used in many applications in Official Statistics.
- Similar to the previous mentioned methods.
- The imputations are obtained by fitting a sequence of (Bayesian) regression models and drawing values from the corresponding predictive distributions.
- **Sequentially imputation**: in each step, one variable serve as **response** and certain other variables serves as **predictors**. **Fit** a certain model using the observed part of the response and **estimate** (update) the (former) missing values in the response.
  - **Initialization loop:** ...
  - **Second outer loop:**
    Estimates of missing values are updated sequentially using one variable as response and all other variables as predictors until convergency.
- Since missing values are drawn from their predictive distribution given the observed data and the specified parameter, the procedure allows **multiple imputation**.

IVEWARE

- Very popular software used in many applications in Official Statistics.
- Similar to the previous mentioned methods.
- The imputations are obtained by fitting a sequence of (Bayesian) regression models and drawing values from the corresponding predictive distributions.
- **Sequentially imputation**: in each step, one variable serve as **response** and certain other variables serves as **predictors**. **Fit** a certain model using the observed part of the response and **estimate** (update) the (former) missing values in the response.
    - **Initialization loop:** ...
    - **Second outer loop:**
      Estimates of missing values are updated sequentially using one variable as response and all other variables as predictors until convergency.
- Since missing values are drawn from their predictive distribution given the observed data and the specified parameter, the procedure allows **multiple imputation**.

IRMI

- Only the second outer loop is used (missing values are initialised in an other manner)
- In contradiction to IVEWARE we use quite **different regression methods** $\rightarrow$ **Robust methods** (Note: a lot of problems has to be solved when using robust methods for complex data like EU-SILC).
- Alternatively, **stepwise** model selction tools are integrated using AIC or BIC.
- **Multiple imputation** is provided.

IRMI

- Only the second outer loop is used (missing values are initialised in an other manner)
- In contradiction to IVEWARE we use quite **different regression methods** → **Robust methods** (Note: a lot of problems has to be solved when using robust methods for complex data like EU-SILC).
- Alternatively, **stepwise** model selction tools are integrated using AIC or BIC.
- **Multiple imputation** is provided.

IRMI

- Only the second outer loop is used (missing values are initialised in an other manner)
- In contradiction to IVEWARE we use quite **different regression methods** $\rightarrow$ **Robust methods** (Note: a lot of problems has to be solved when using robust methods for complex data like EU-SILC).
- Alternatively, **stepwise** model selction tools are integrated using AIC or BIC.
- **Multiple imputation** is provided.

Selection of Regression Models

If the **response** is

- *continuous*, robust (IRMI) or ols (IMI, IVEWARE) regression methods are used.
- *categorical*, generalized linear regression is applied (IRMI: robust or non-robust).
- *binary*, logistic linear regression is applied (IRMI: robust but non-robust is prefered).
- *mixed*, a two-stage approach is used whereas in the first stage logistic regression is applied in order to decide if a missing value is imputed with zero or by applying robust regression based on the continuous part of the response.
- *count*, robust generalized linear models (family: Poisson) is used.

Selection of Regression Models

If the **response** is

- *continuous*, robust (IRMI) or ols (IMI, IVEWARE) regression methods are used.
- *categorical*, generalized linear regression is applied (IRMI: robust or non-robust).
- *binary*, logistic linear regression is applied (IRMI: robust but non-robust is prefered).
- *mixed*, a two-stage approach is used whereas in the first stage logistic regression is applied in order to decide if a missing value is imputed with zero or by applying robust regression based on the continuous part of the response.
- *count*, robust generalized linear models (family: Poisson) is used.

Selection of Regression Models

If the **response** is

- *continuous*, robust (IRMI) or ols (IMI, IVEWARE) regression methods are used.
- *categorical*, generalized linear regression is applied (IRMI: robust or non-robust).
- *binary*, logistic linear regression is applied (IRMI: robust but non-robust is prefered).
- *mixed*, a two-stage approach is used whereas in the first stage logistic regression is applied in order to decide if a missing value is imputed with zero or by applying robust regression based on the continuous part of the response.
- *count*, robust generalized linear models (family: Poisson) is used.

Selection of Regression Models

If the **response** is

- *continuous*, robust (IRMI) or ols (IMI, IVEWARE) regression methods are used.
- *categorical*, generalized linear regression is applied (IRMI: robust or non-robust).
- *binary*, logistic linear regression is applied (IRMI: robust but non-robust is prefered).
- *mixed*, a two-stage approach is used whereas in the first stage logistic regression is applied in order to decide if a missing value is imputed with zero or by applying robust regression based on the continuous part of the response.
- *count*, robust generalized linear models (family: Poisson) is used.

Selection of Regression Models

If the **response** is

- *continuous*, robust (IRMI) or ols (IMI, IVEWARE) regression methods are used.
- *categorical*, generalized linear regression is applied (IRMI: robust or non-robust).
- *binary*, logistic linear regression is applied (IRMI: robust but non-robust is prefered).
- *mixed*, a two-stage approach is used whereas in the first stage logistic regression is applied in order to decide if a missing value is imputed with zero or by applying robust regression based on the continuous part of the response.
- *count*, robust generalized linear models (family: Poisson) is used.

Selection of Regression Models

If the **response** is

- *continuous*, robust (IRMI) or ols (IMI, IVEWARE) regression methods are used.
- *categorical*, generalized linear regression is applied (IRMI: robust or non-robust).
- *binary*, logistic linear regression is applied (IRMI: robust but non-robust is prefered).
- *mixed*, a two-stage approach is used whereas in the first stage logistic regression is applied in order to decide if a missing value is imputed with zero or by applying robust regression based on the continuous part of the response.
- *count*, robust generalized linear models (family: Poisson) is used.

Errors from Categorical and Binary Variables

This error measure is defined as the proportion of imputed values taken from an incorrect category on all missing categorical or binary values:

$$err_c = \frac{1}{m_c} \sum_{j=1}^{p_c} \sum_{i=1}^{n} \mathbb{I}(x_{ij}^{orig} \neq x_{ij}^{imp}) \quad , \tag{1}$$

with $\mathbb{I}$ the indicator function, m_c the number of missing values in the p_c categorical variables, and n the number of observations.

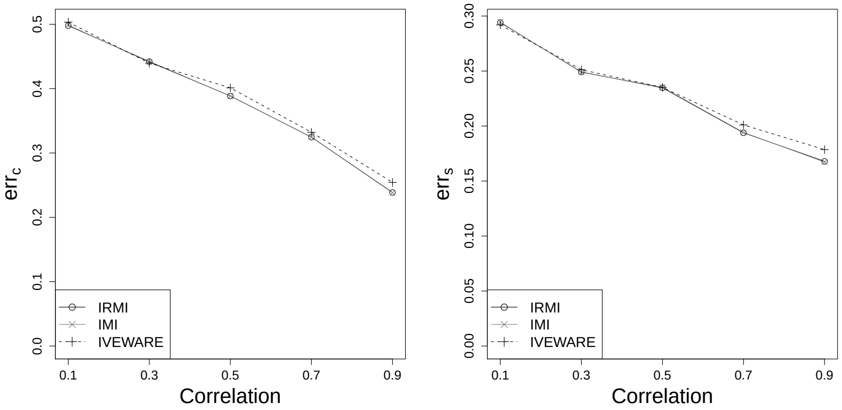
Errors from Continuous and Semi-continuous Variables

Here we assume that the constant part of the semi-continuous variable is zero. Then, the joint error measure is

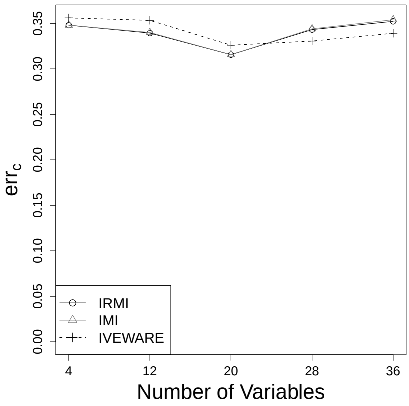
$$err_s = \frac{1}{m_s}\sum_{j=1}^{p_s}\sum_{i=1}^{n}\left[\left|\frac{(x_{ij}^{orig} - x_{ij}^{imp})}{x_{ij}^{orig}}\right| \cdot \mathbb{I}(x_{ij}^{orig} \neq 0 \ \wedge \ x_{ij}^{imp} \neq 0) + \mathbb{I}((x_{ij}^{orig} = 0 \ \wedge \ x_{ij}^{imp} \neq 0) \ \vee \ (x_{ij}^{orig} \neq 0 \ \wedge \ x_{ij}^{imp} = 0))\right] \quad (6)$$

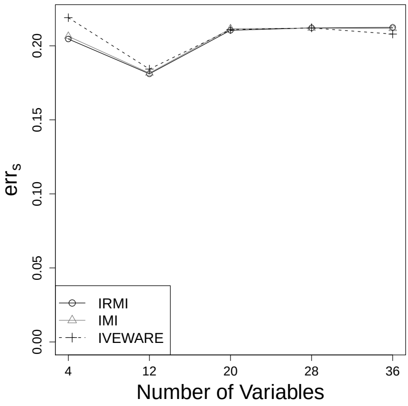
with m_s the number of missing values in the p_s continuous and semi-continuous variables.

Simulation Results: Varying the Correlation

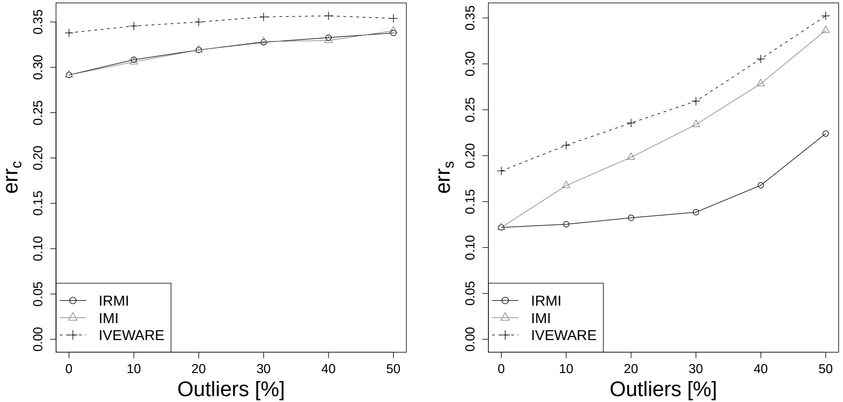


Simulation Results: Varying the Amount of Variables

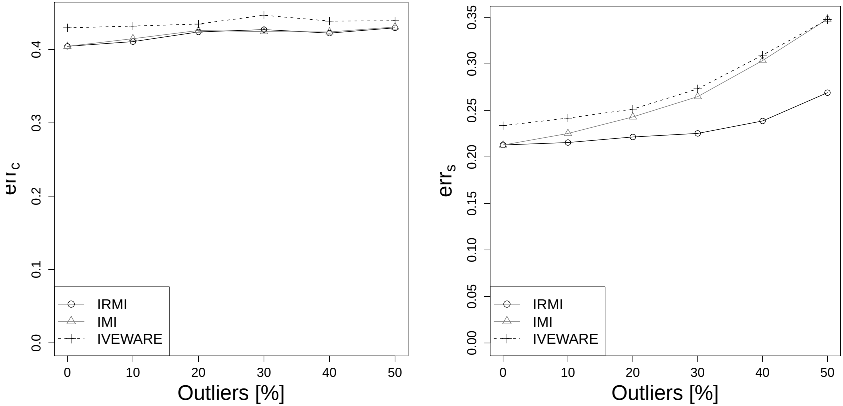




Including (moderate) Outliers and Varying their Amount, high Correlation



Including (moderate) Outliers and Varying their Amount, low Correlation

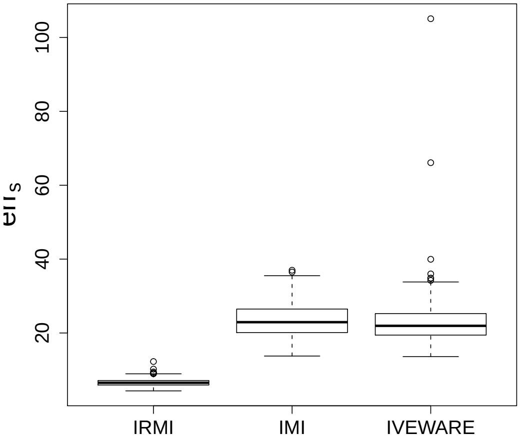


Imputation in EU-SILC

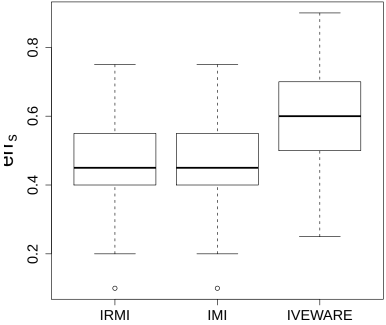
We considered certain HH-components, but also some nominal variables, such as *household size*, *region* and *htype3*.

1. $R = 0$
2. Set missing values in HH-components randomly (MCAR). $R++$
3. Impute the missing values.
4. Evaluate the imputations using certain information loss measures.
5. Go to (2) until $R = 100$.

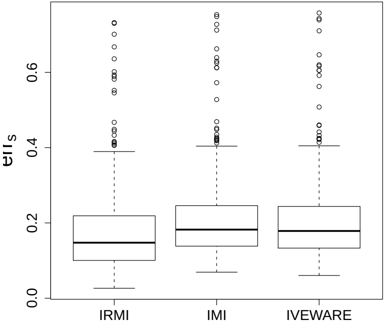
Imputation in EU-SILC, Results



CENSUS Data - no outliers

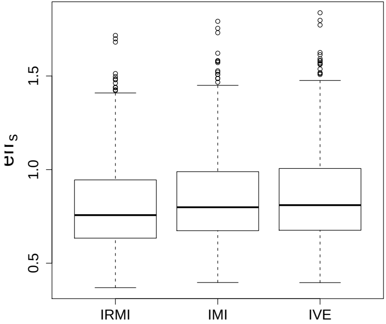


(i) Error for categorical variables

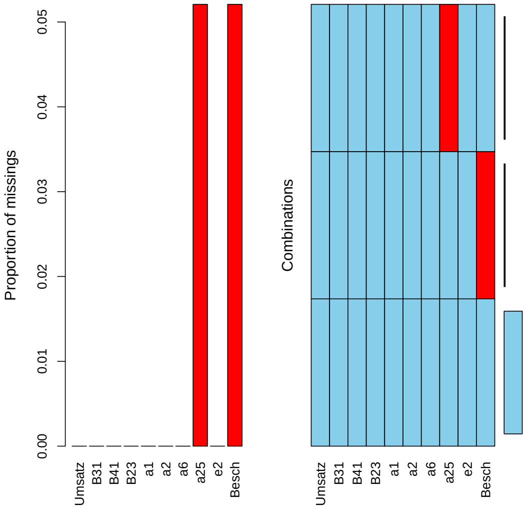


(j) Error for numerical variables

Airquality Data



Example Data: SBS data



Most important functionality for imputation

Listing 1: Hotdeck imputation.

```
hotdeck(x, ord_var=c("Besch","Umsatz"), imp_var=FALSE)
```

Listing 2: *k*-nearest neighbor imputation.

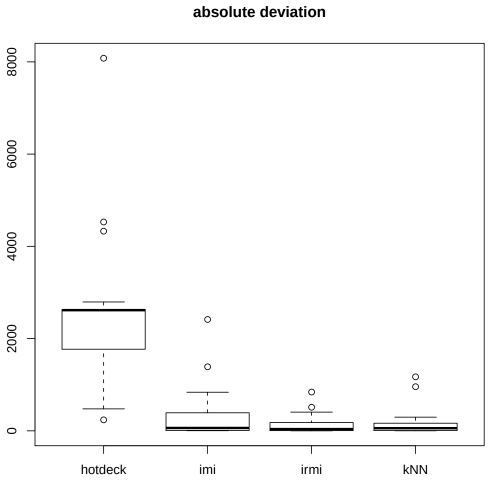
```
kNN(x)
```

Listing 3: Application of robust iterative model-based imputation.

```
imp <- irmi(x)
```

...sensible defaults!

SBS data: Simulation Results



Conclusion

- We proposed the system VIM for visualization and imputation of missing values.
- IRMI performs almost always best, but hot-deck methods have it's advantages as well (they are very fast and easy understandable)
- VIM is an free and open-source project. It can be freely downloaded at

  http://cran.r-project.org/package=VIM
- Joint development and contributions are warmly welcome.

References

M. Templ, A. Alfons, and A. Kowarik. VIM: Visualization and Imputation of Missing Values, 2011a. URL http://cran.r-project.org. R package version 2.0.1.

Matthias Templ, Alexander Kowarik, and Peter Filzmoser. Iterative stepwise regression imputation using standard and robust methods. Computational Statistics Data Analysis, In Press, Uncorrected Proof, 2011b. ISSN 0167-9473. doi: DOI:10.1016/j.csda.2011.04.012. URL http://www.sciencedirect.com/science/article/B6V8V-52R7YYH-2/2/2c6c9ed7138d50c4197e991d8ffb8a1f.