Missing Values

A Computational and Methodological Framework for Visualisation and Imputation of Missing Values

Matthias Templ[1,2,3], Alexander Kowarik[1,3], Peter Filzmoser[2], Andreas Alfons[4]

[1] Department of Methodology, Statistics Austria
[2] Department of Statistics and Probability Theory, TU WIEN, Austria
[3] http://www.data-analysis.at/
[4] ORSTAT Research Center, Faculty of Business and Economics, K.U.Leuven, Belgium

Content

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).
- **C**ommand **L**ine **I**nterface and **G**raphical **U**ser **I**nterface.
- Hot-deck, *k*-NN and EM-based (robust) imputation methods for complex data sets including binary, nominal, ordered categorical, count, continuous and semi-continuous variables.

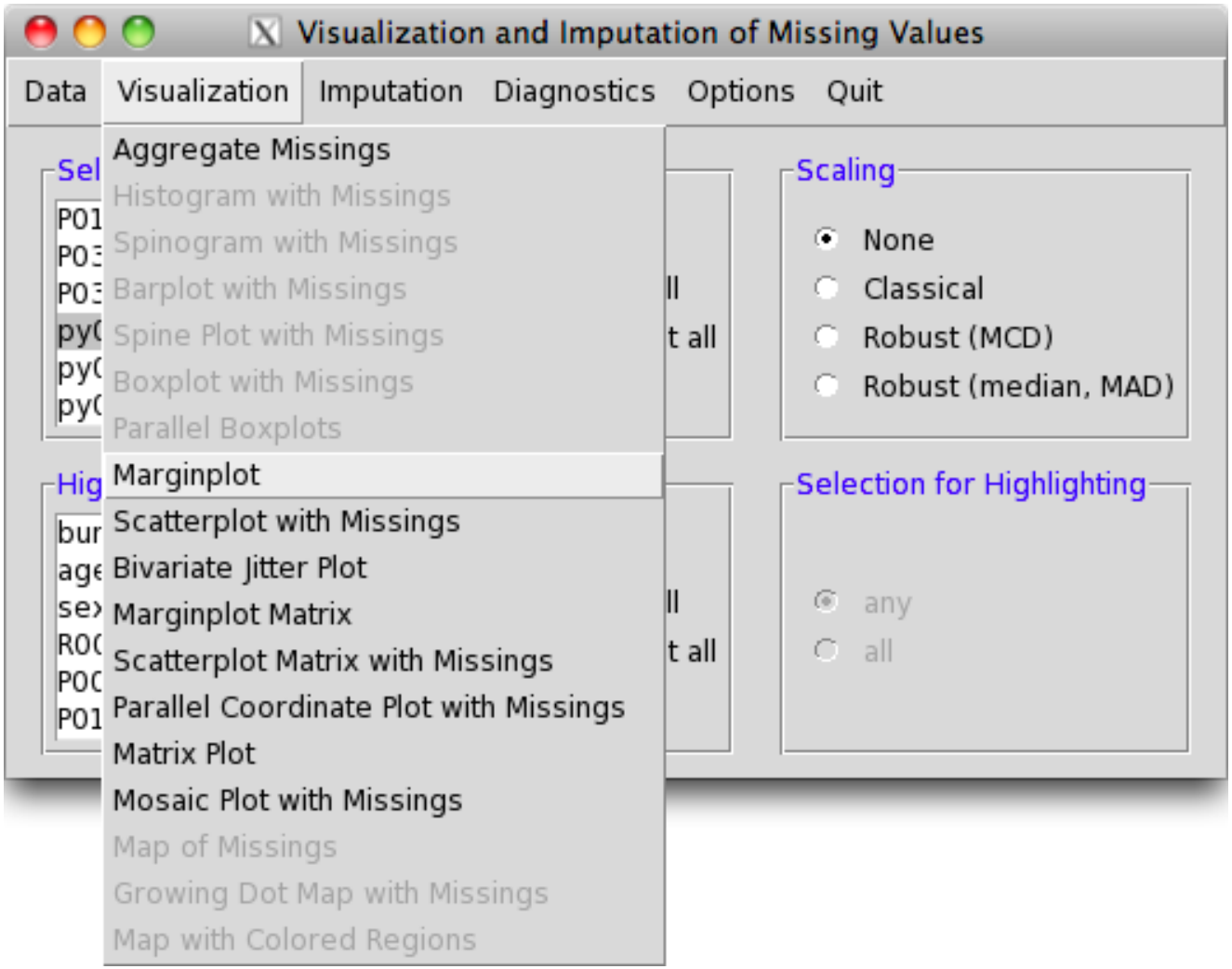
Exploring the Structure of Missing Values

- Presented two years ago at Statistische Tage
- Classical plot methods cannot deal with missing values
- VIM allows to explore the multivariate structure of missing values and their systematical appearance in the data
- VIM gives support to learn about the data with few clicks
- VIM allows to produce high-quality graphics for publication and supports interactivity during analysis
- Chosen variables (say X) are plotted (plotting variables) and information is used for colouring based on variables for highlighting (say Y). May X explain the structure of missing values in Y?

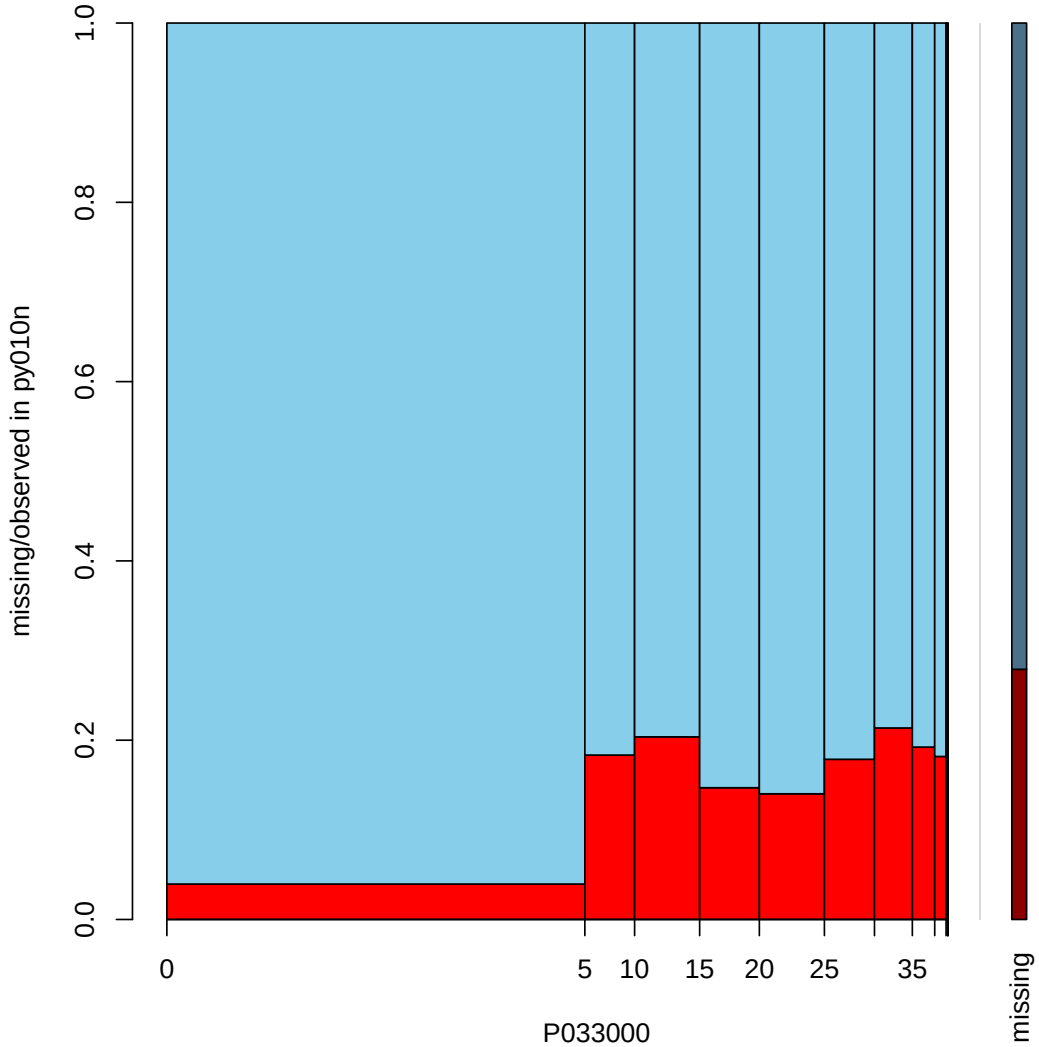
Exploring the Structure of Missing Values

- Presented two years ago at Statistische Tage
- Classical plot methods cannot deal with missing values
- `VIM` allows to explore the multivariate structure of missing values and their systematical appearance in the data
- `VIM` gives support to learn about the data with few clicks
- `VIM` allows to produce high-quality graphics for publication and supports interactivity during analysis
- Chosen variables (say X) are plotted (plotting variables) and information is used for colouring based on variables for highlighting (say Y). May X explain the structure of missing values in Y?

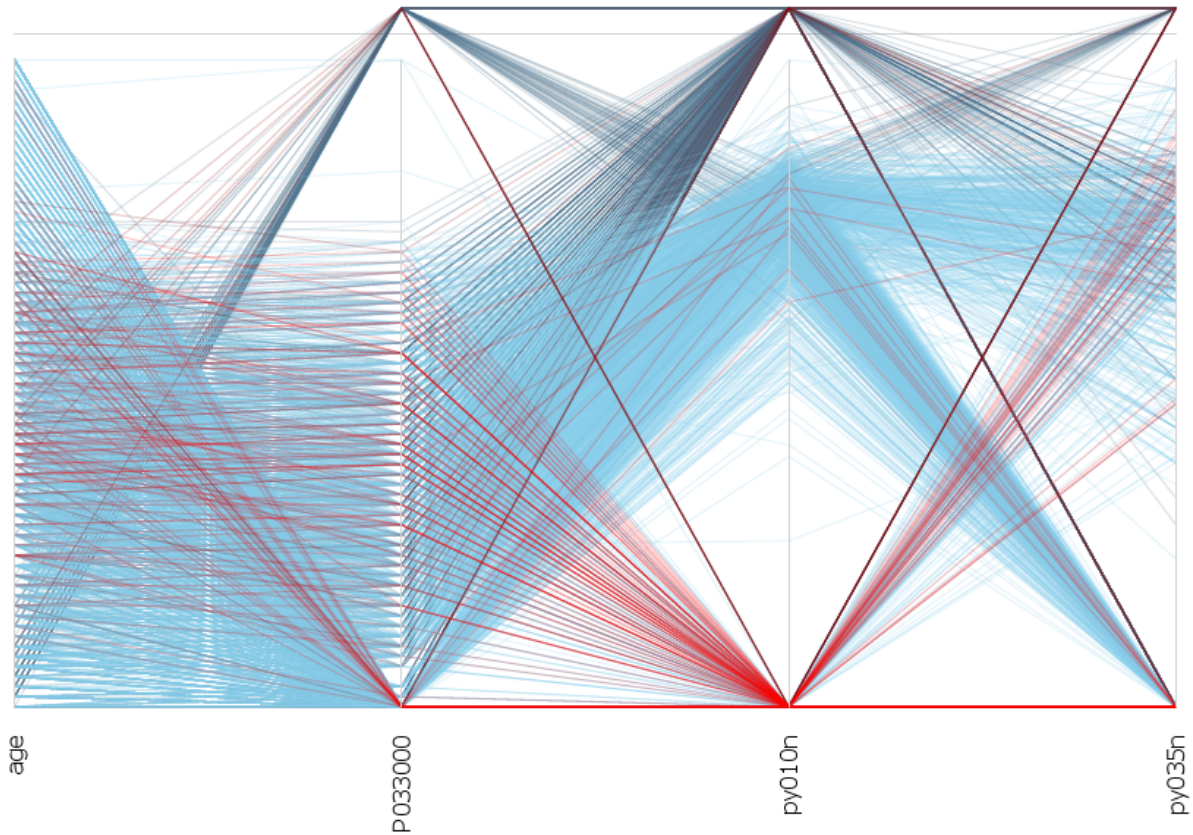
The GUI . . .


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

An Example . . .


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

Another Example ...



Some Challenges for Imputation of Complex Surveys

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 values have to be both, to reflect the multivariate structure of the data and including "randomness".

Time complexity Fast algorithms are required since we deal with relatively large data sets.

Some Challenges for Imputation of Complex Surveys

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 values have to be both, to reflect the multivariate structure of the data and including "randomness".

Time complexity Fast algorithms are required since we deal with relatively large data sets.

Some Challenges for Imputation of Complex Surveys

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.

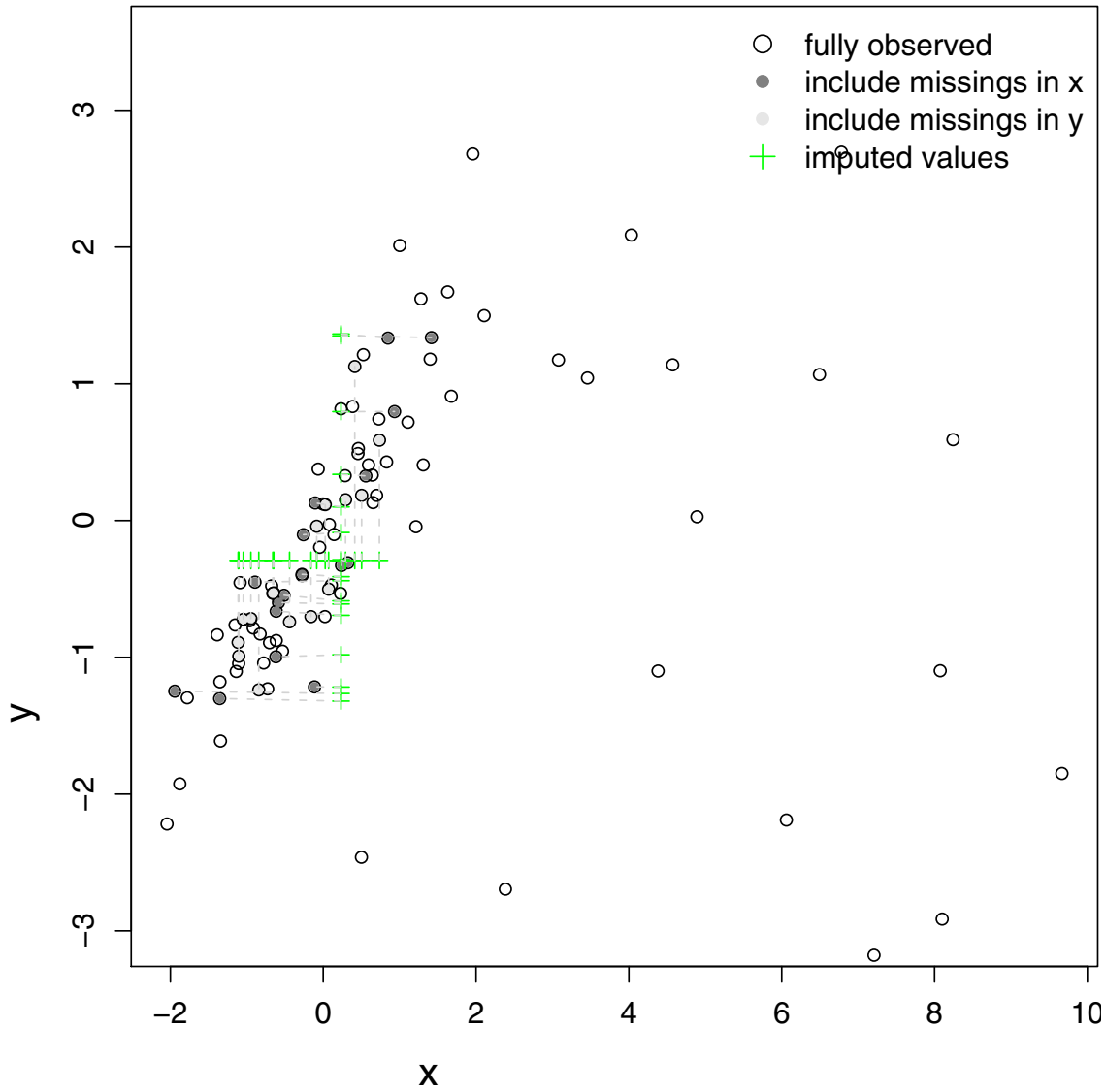
Semi-continuous variables: "**semi-continuous**" distributions, i.e. a variable consisting of a continuous scaled part and a certain proportion of equal values.

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

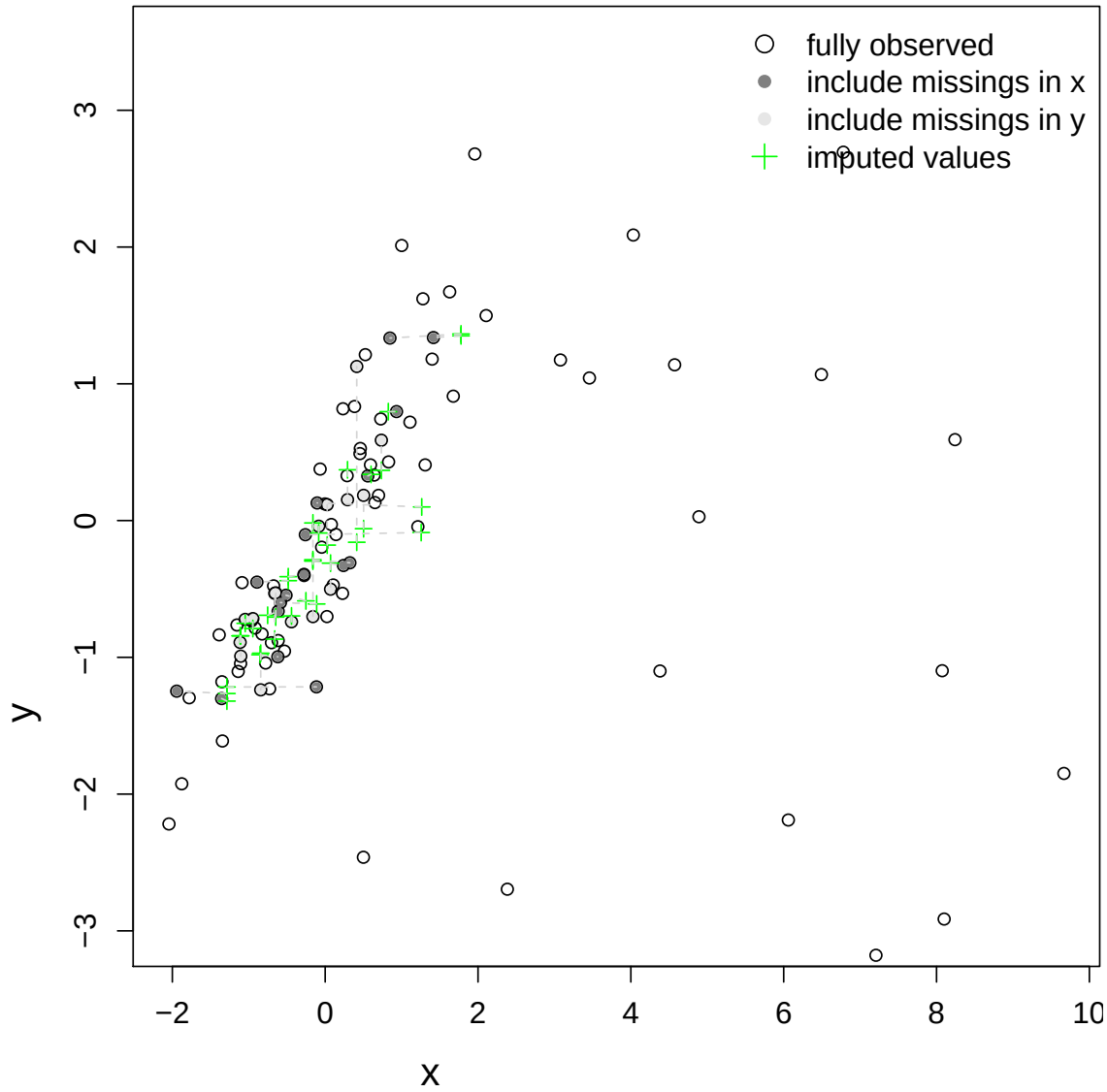
Multiple imputation: Imputated values have to be both, to reflect the multivariate structure of the data and including "randomness".

Time complexity Fast algorithms are required since we deal with relatively large data sets.

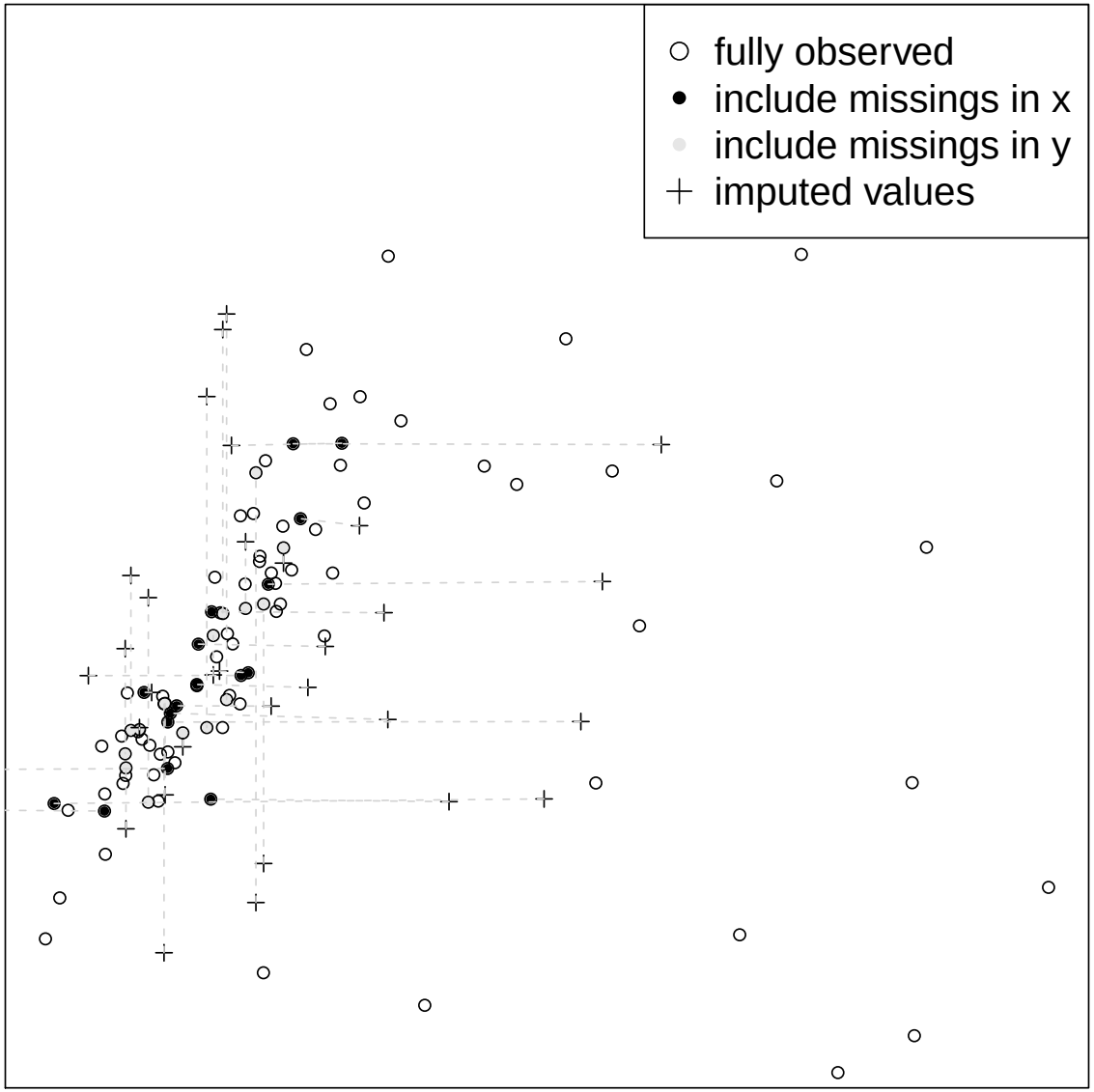
Median Imputation



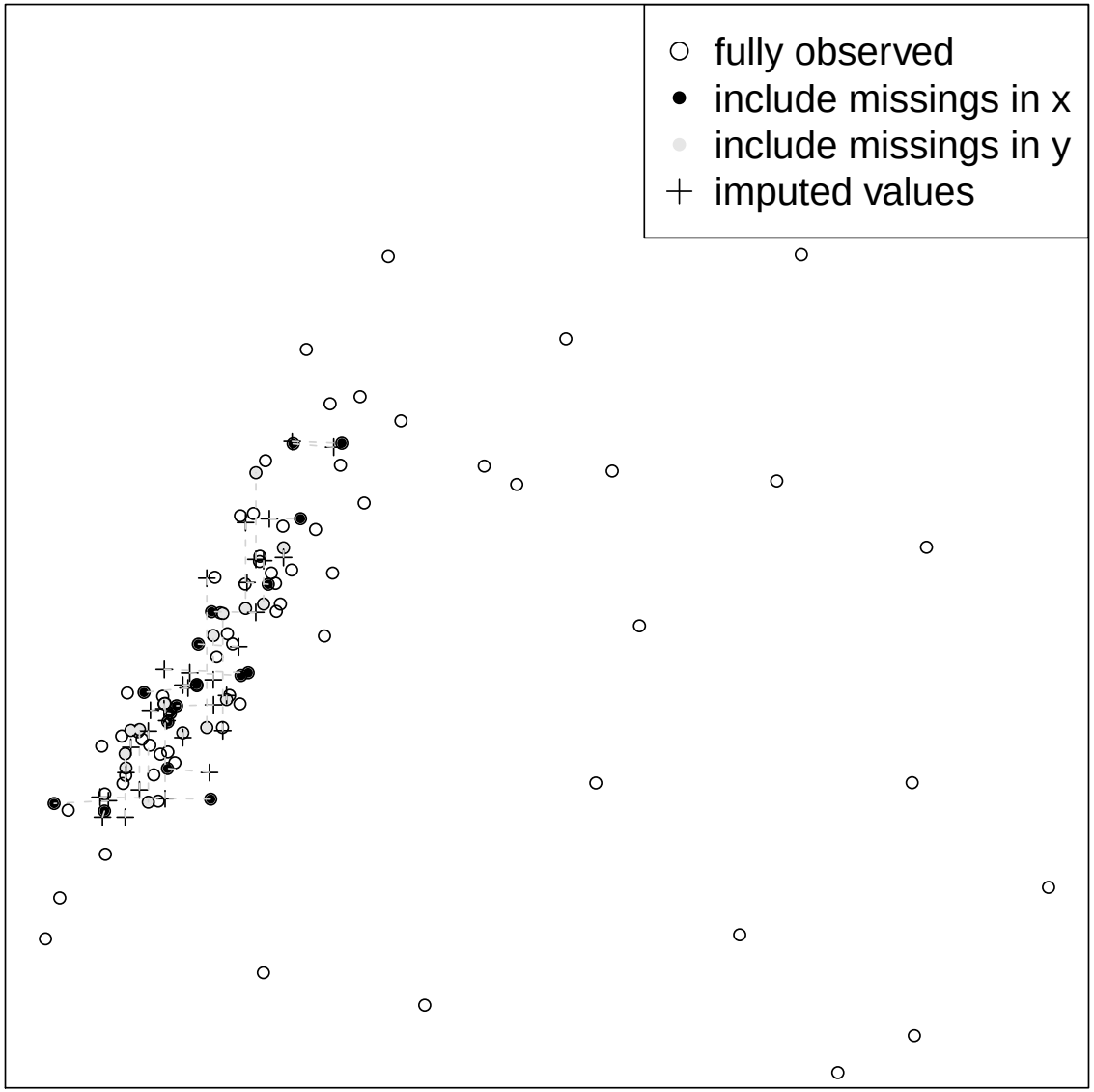
kNN Imputation



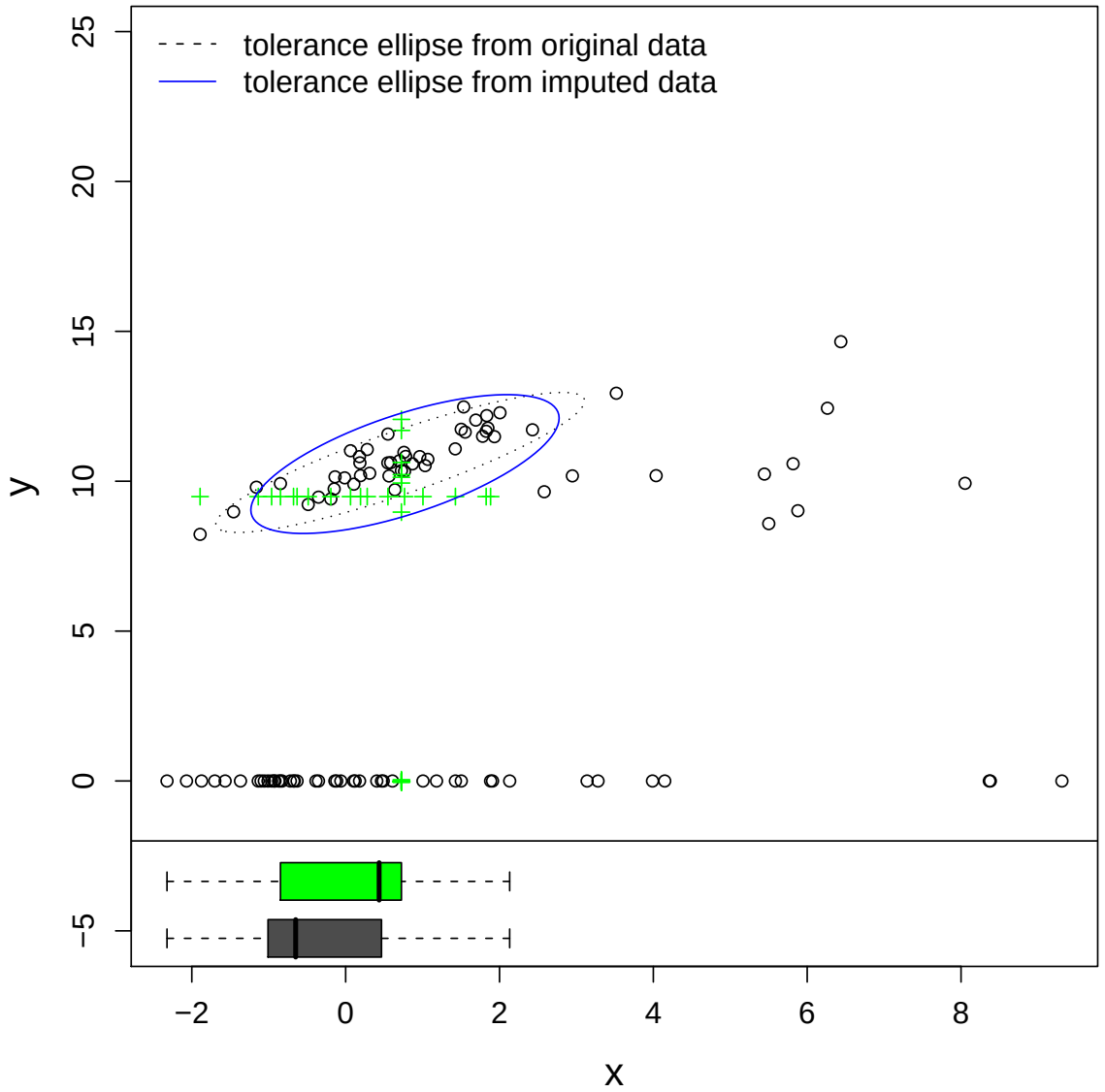
IVEWARE


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

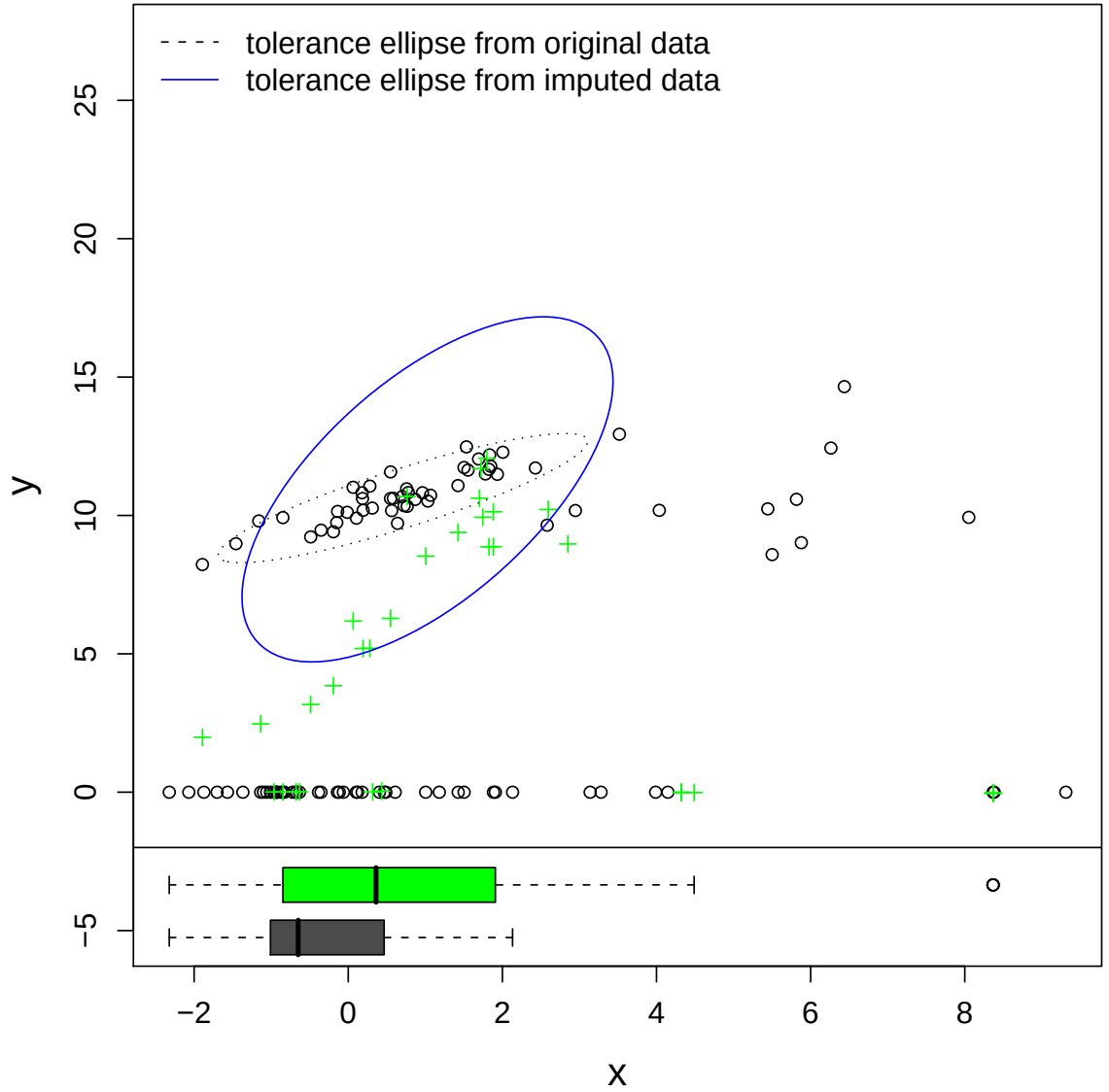
IRMI


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

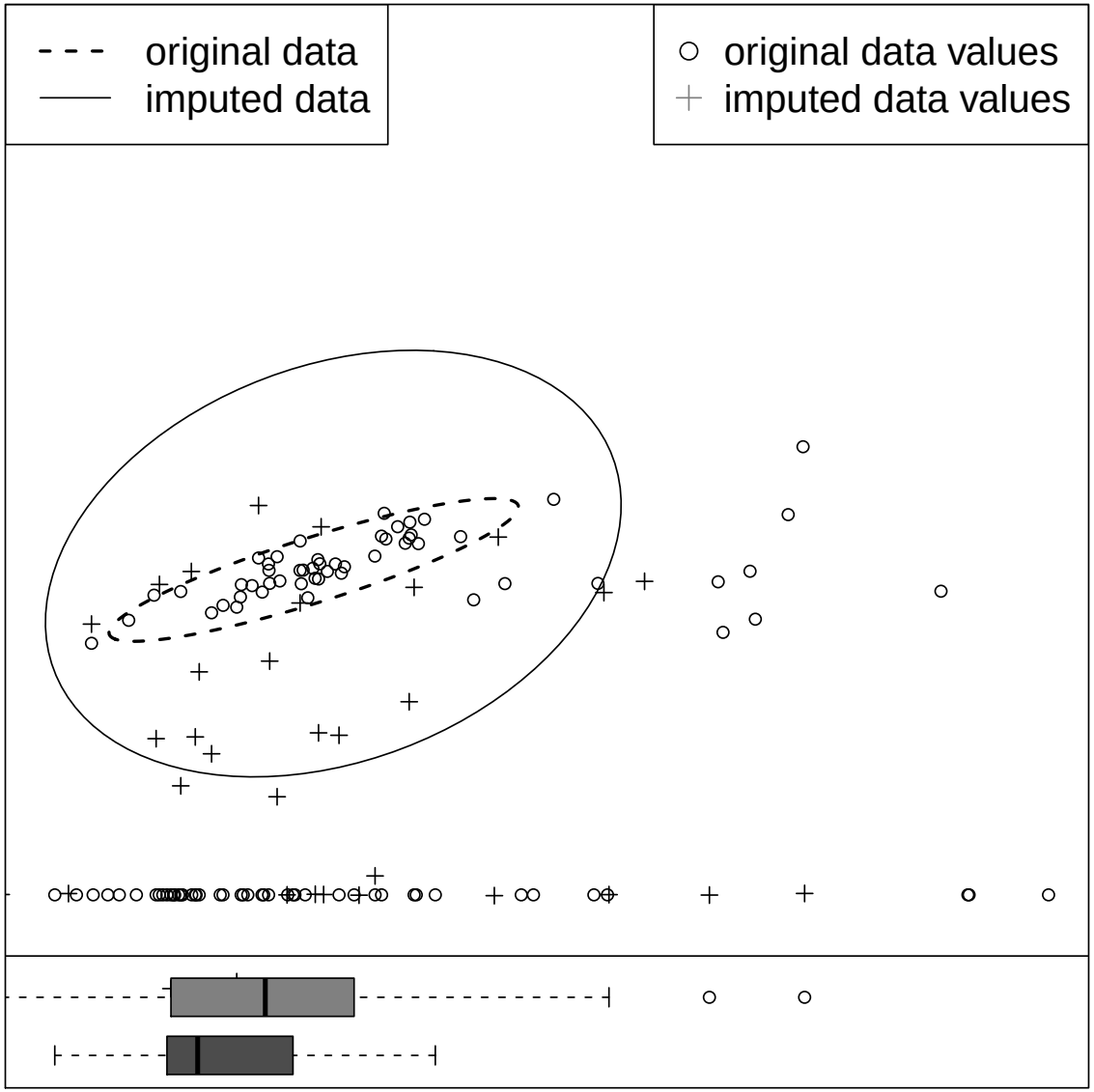
Median Imputation



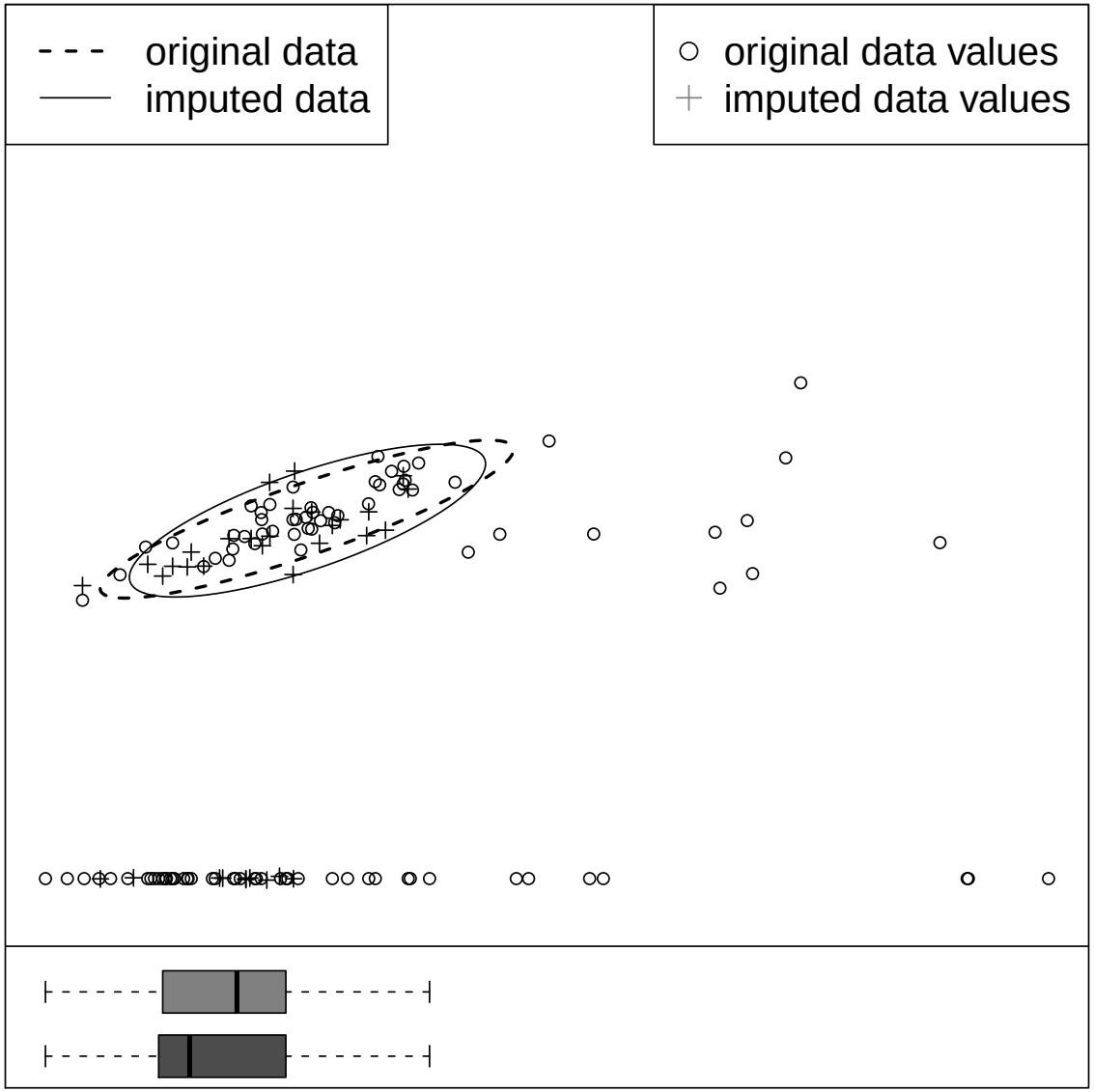
kNN Imputation



IVEWARE


original data
imputed data
original data values
imputed data values

IRMI


original data
imputed data
original data values
imputed data values

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 EM-based sequential regressions.
- In general, often problems occurs when applied to complex data sets (time complexity, influence of outliers, semi-continuous variables)

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 EM-based sequential regressions.
- In general, often problems occurs when applied to complex data sets (time complexity, influence of outliers, semi-continuous variables)

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 EM-based sequential regressions.
- In general, often problems occurs when applied to complex data sets (time complexity, influence of outliers, semi-continuous variables)

From http://CRAN.R-project.org/view=OfficialStatistics

Imputation

A distinction between iterative model-based methods, k-nearest neighbor methods and miscellaneous methods is made. However, often the criteria for using a method depend on the scale of the data, which statistics are typically a mixture of continuous, semi-continuous, binary, categorical and count variables.

EM-based Imputation Methods:

- Package mi provides iterative EM-based multiple Bayesian regression imputation of missing values and model checking of the regression models used. The regression models for each variable can als defined. The data set may consist of continuous, semi-continuous, binary, categorical and/or count variables.
- Package mice provides iterative EM-based multiple regression imputation. The data set may consist of continuous, binary, categorical and/or count variables.
- Package mitools provides tools to perform analyses and combine results from multiply-imputated datasets.
- Package Amelia provides multiple imputation where first bootstrap samples with the same dimensions as the original data are drawn, and then used for EM-based imputation. It is also possible to impu longitudial data. The package in addition comes with a graphical user interface.
- Package VIM provides EM-based multiple imputation (function `irmi()`) using robust estimations, which allows to adequately deal with data including outliers. It can handle data consisting of continu continuous, binary, categorical and/or count variables.
- Package mix provides iterative EM-based multiple regression imputation. The data set may consist of continuous, binary or categorical variables.
- Package pan provides multiple imputation for multivariate panel or clustered data.
- Package norm provides EM-based multiple imputation for multivariate normal data.
- Package cat provides EM-based multiple imputation for multivariate categorical data.
- Package MImix provides tools to combine results for multiply-imputed data using mixture approximations.
- Package robCompositions provides iterative model-based imputation for compositional data (function `impCoda()`).

k-Nearest Neighbor (knn) Imputation Methods

- Package yaImpute performs popular nearest neighbor routines for imputation of continuous variables where different metrics and methods can be used for determining the distance between observatio
- Function `SeqKNN()` in Package SeqKnn imputes the missing values in continuously scaled variables sequentially. First, it separates the dataset into incomplete and complete observations. The observa incomplete set are imputed by the order of missing rate. Once the missing values in an observations are imputed, the imputed observation is moved into the complete set.
- Package impute impute provides knn imputation of continuous variables.
- Package robCompositions provides knn imputation for compositional data (function `impKNNa()`) using the Aitchison distance and adjustment of the nearest neighbor.
- Package rrcovNA provides an algorithm for (robust) sequential imputation (function `impSeq()` and `impSeqRob()` by minimizing the determinant of the covariance of the augmented data matrix.

Miscellaneous Imputation Methods:

- Package missMDA allows to impute incomplete continuous variables by principal component analysis (PCA) or categorical variables by multiple correspondence analysis (MCA).
- Package mice (function `mice.impute.pmm()`) and Package Hmisc (function `aregImpute()`) allow predicitve mean matching imputation.
- Package VIM allows to visualize the structure of missing values using suitable plot methods. It also comes with a graphical user interface.

Imputation Methods that Fulfils our Requirements

Imputation

A distinction between iterative model-based methods, k-nearest neighbor methods and miscellaneous methods is made. However, often the criteria for using a method depend on the scale of the data, which statistics are typically a mixture of continuous, semi-continuous, binary, categorical and count variables.

EM-based Imputation Methods:

- Package mi provides iterative EM-based multiple Bayesian regression imputation of missing values and model checking of the regression models used. The regression models for each variable can als defined. The data set may consist of continuous, semi-continuous, binary, categorical and/or count variables.
- Package mice provides iterative EM-based multiple regression imputation. The data set may consist of continuous, binary, categorical and/or count variables.
- Package mitools provides tools to perform analyses and combine results from multiply-imputated datasets.
- Package Amelia provides multiple imputation where first bootstrap samples with the same dimensions as the original data are drawn, and then used for EM-based imputation. It is also possible to imp longitudinal data. The package in addition comes with a graphical user interface.
- Package VIM provides EM-based multiple imputation (function `irmi()`) using robust estimations, which allows to adequately deal with data including outliers. It can handle data consisting of continu continuous, binary, categorical and/or count variables.
- Package mix provides iterative EM-based multiple regression imputation. The data set may consist of continuous, binary or categorical variables.
- Package pan provides multiple imputation for multivariate panel or clustered data.
- Package norm provides EM-based multiple imputation for multivariate normal data.
- Package cat provides EM-based multiple imputation for multivariate categorical data.
- Package MImix provides tools to combine results for multiply-imputed data using mixture approximations.
- Package robCompositions provides iterative model-based imputation for compositional data (function `impCoda()`).

k-Nearest Neighbor (knn) Imputation Methods

- Package yaImpute performs popular nearest neighbor routines for imputation of continuous variables where different metrics and methods can be used for determining the distance between observatio
- Function `SeqKNN()` in Package SeqKnn imputes the missing values in continuously scaled variables sequentially. First, it separates the dataset into incomplete and complete observations. The observa incomplete set are imputed by the order of missing rate. Once the missing values in an observations are imputed, the imputed observation is moved into the complete set.
- Package impute imputes provides knn imputation of continuous variables.
- Package robCompositions provides knn imputation for compositional data (function `impKNNa()`) using the Aitchison distance and adjustment of the nearest neighbor.
- Package rrcovNA provides an algorithm for (robust) sequential imputation (function `impSeq()` and `impSeqRob()` by minimizing the determinant of the covariance of the augmented data matrix.

Miscellaneous Imputation Methods:

- Package missMDA allows to impute incomplete continuous variables by principal component analysis (PCA) or categorical variables by multiple correspondence analysis (MCA).
- Package mice (function `mice.impute.pmm()`) and Package Hmisc (function `aregImpute()`) allow predicitve mean matching imputation.
- Package VIM allows to visualize the structure of missing values using suitable plot methods. It also comes with a graphical user interface.

IVEWARE

- Very popular software used in many applications.
- 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.
- 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.
- 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.
- 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.
- 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 by *k*NN imputation)
- 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 by kNN imputation)
- 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 by kNN imputation)
- 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

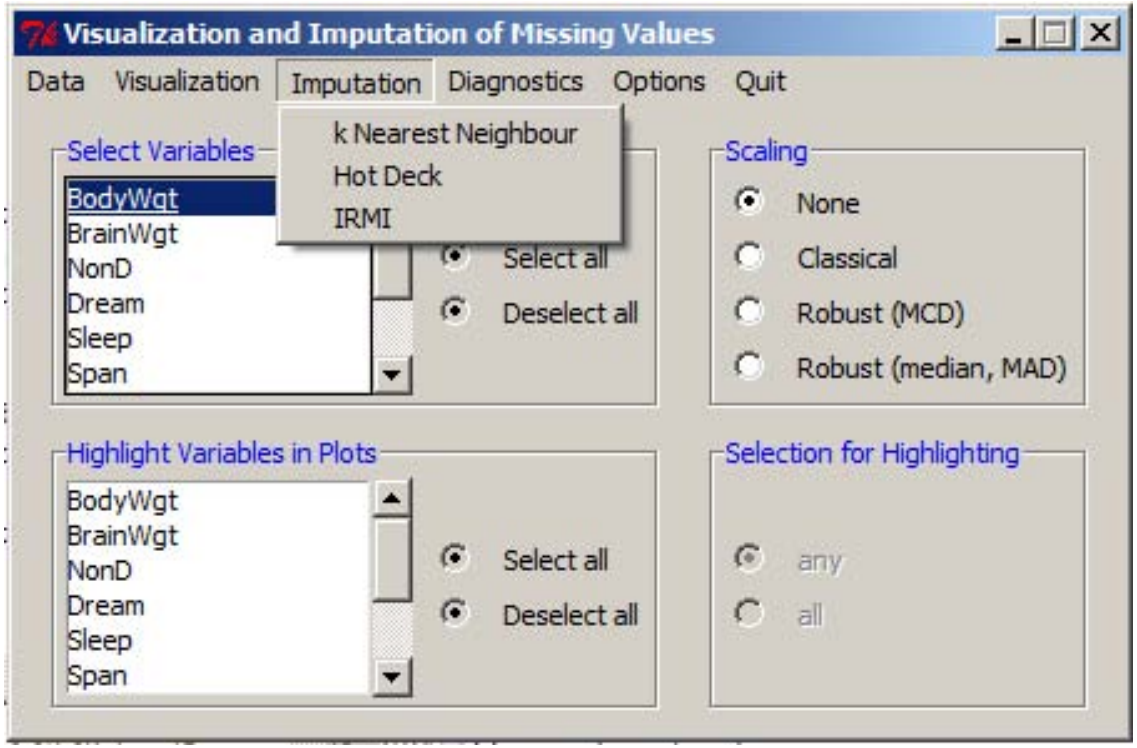
- *continuous*, robust (IRMI) or ols (IMI, IVEWARE) regression methods are used.
- *categorical*, generalized linear regression is applied (IRMI: robust or non-robust).
- *binary*, logistic linear regression is applied (IRMI: robust but non-robust is prefered).
- *mixed*, a two-stage approach is used 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

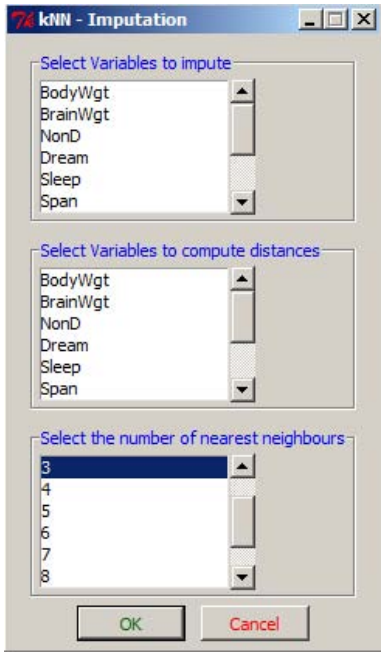
If the **response** is

- *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.

The GUI - Imputation

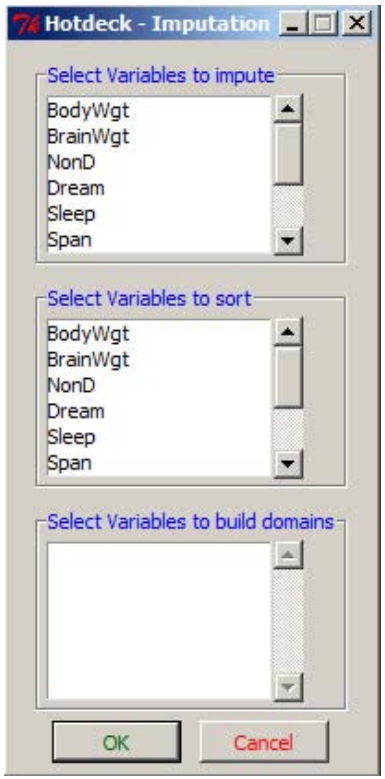


The GUI - Imputation



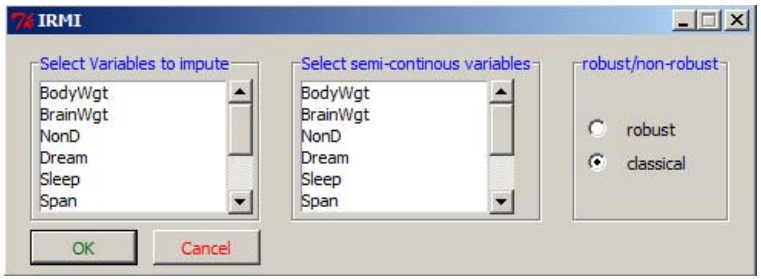

```
kNN(data, variable=colnames(data), metric=NULL, k=5,
  dist_var=colnames(data),weights=NULL,numFun = median,
  catFun=maxCat,makeNA=NULL,NAcond=NULL, impNA=TRUE,
  donorcond=NULL,mixed=vector(),trace=FALSE,
  imp_var=TRUE,imp_suffix="imp",addRandom=FALSE)
```

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)
```

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.

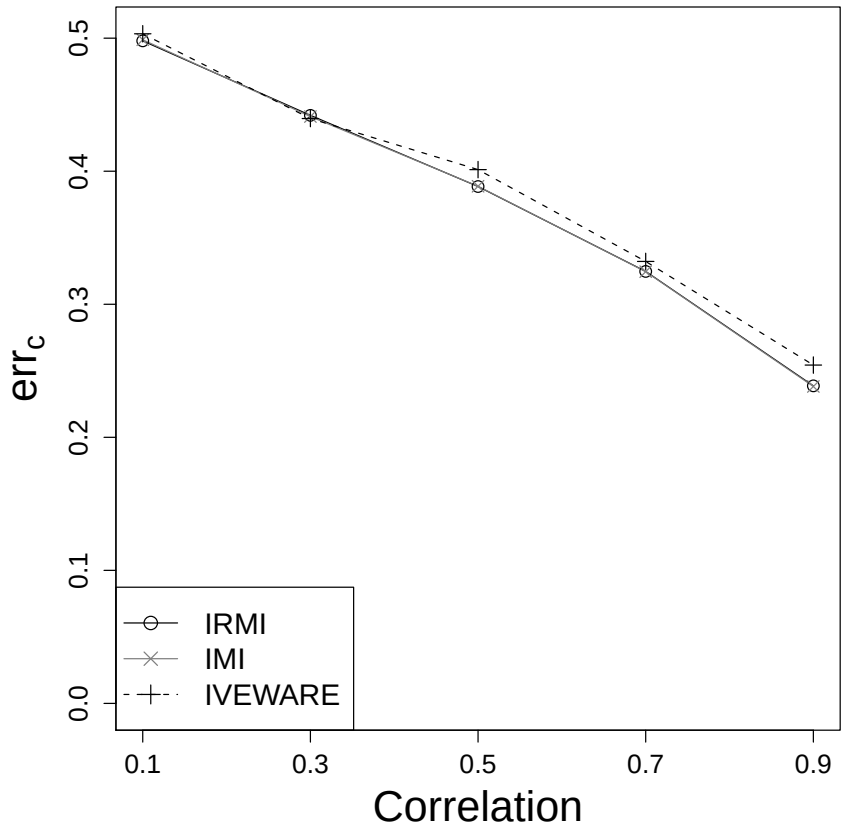
Errors from Continuous and Semi-continuous Variables

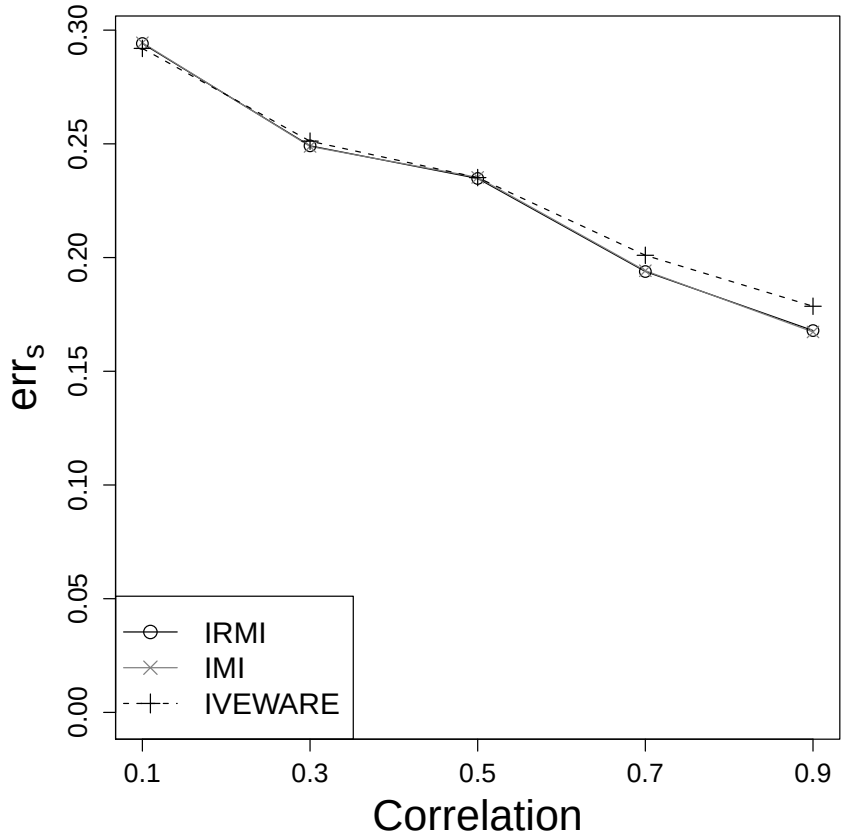
Here we assume that the constant part of the semi-continuous variable is zero. Then, the joint error measure is

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

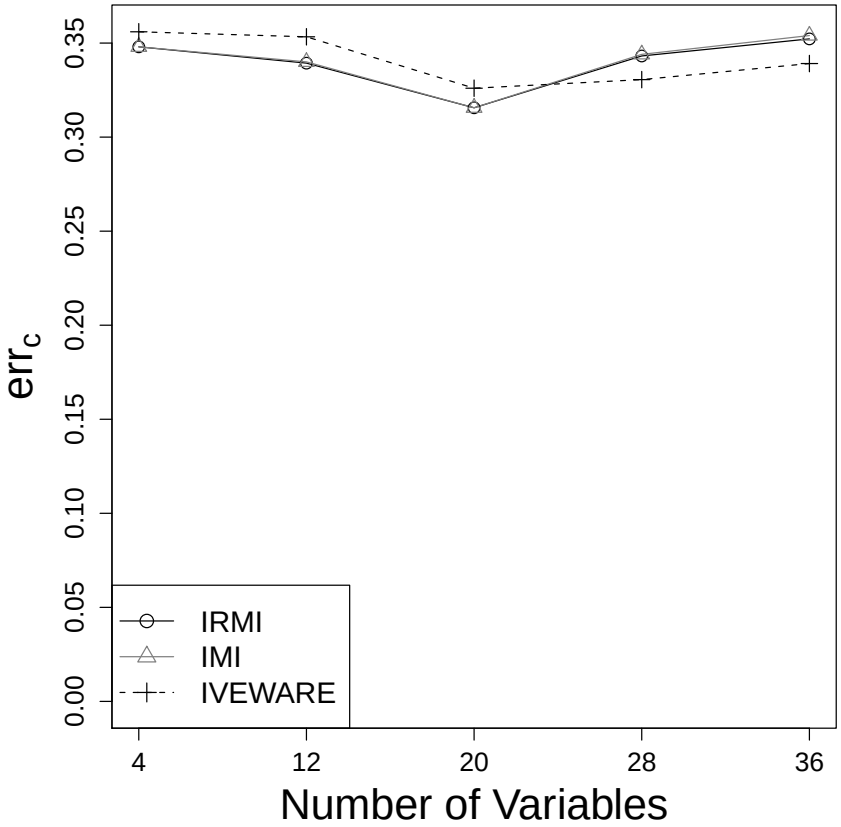
with m_s the number of missing values in the p_s continuous and semi-continuous variables.

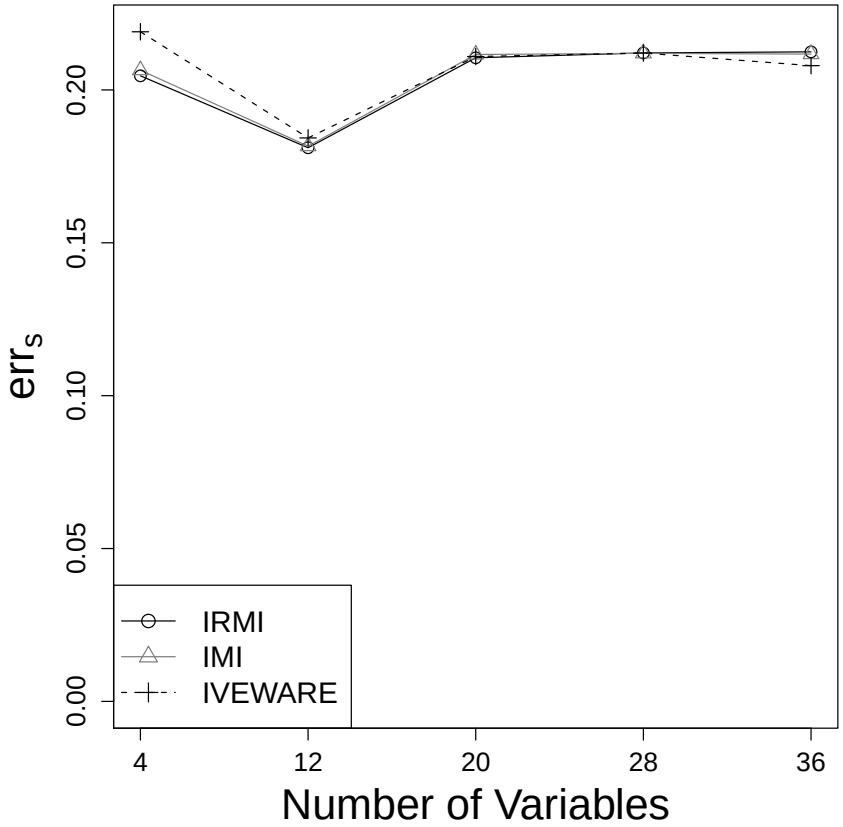
Simulation Results: Varying the Correlation



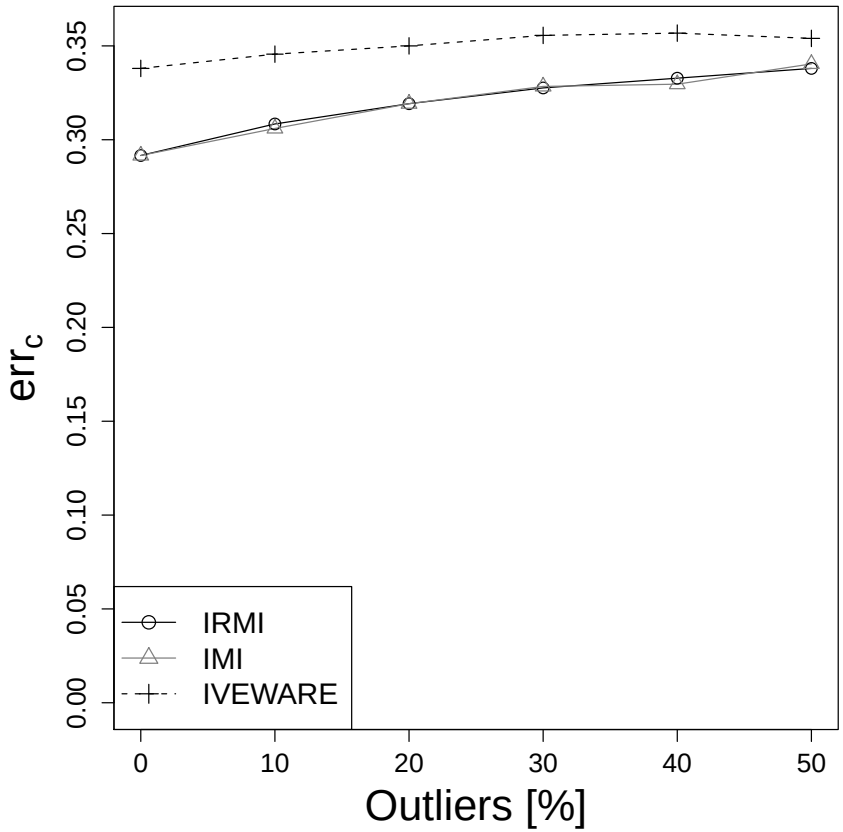


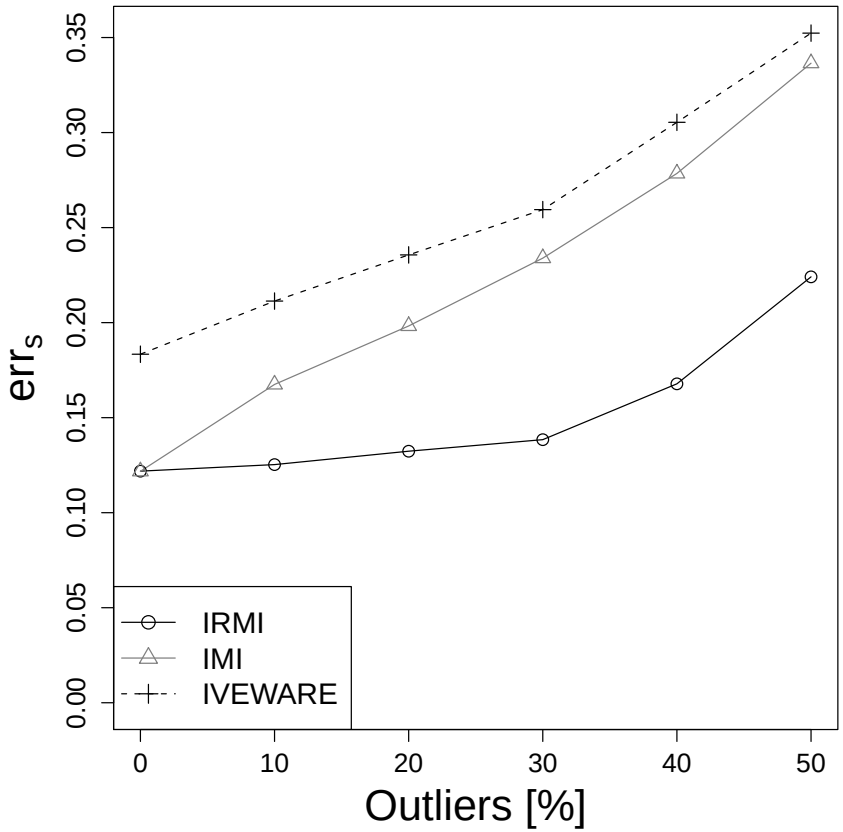
Simulation Results: Varying the Amount of Variables



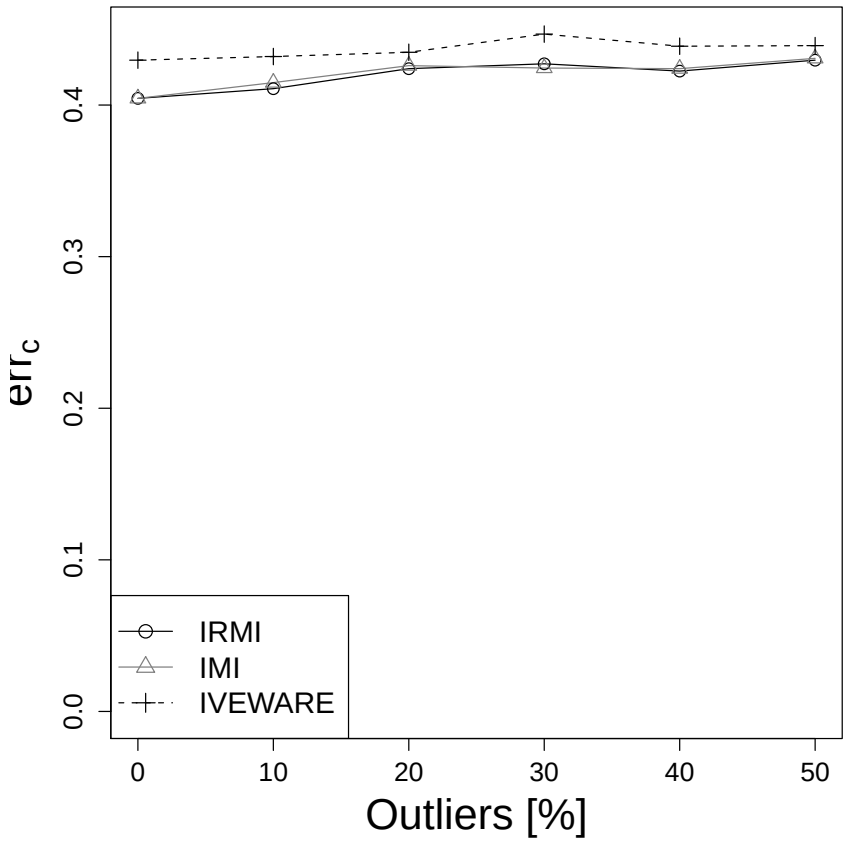


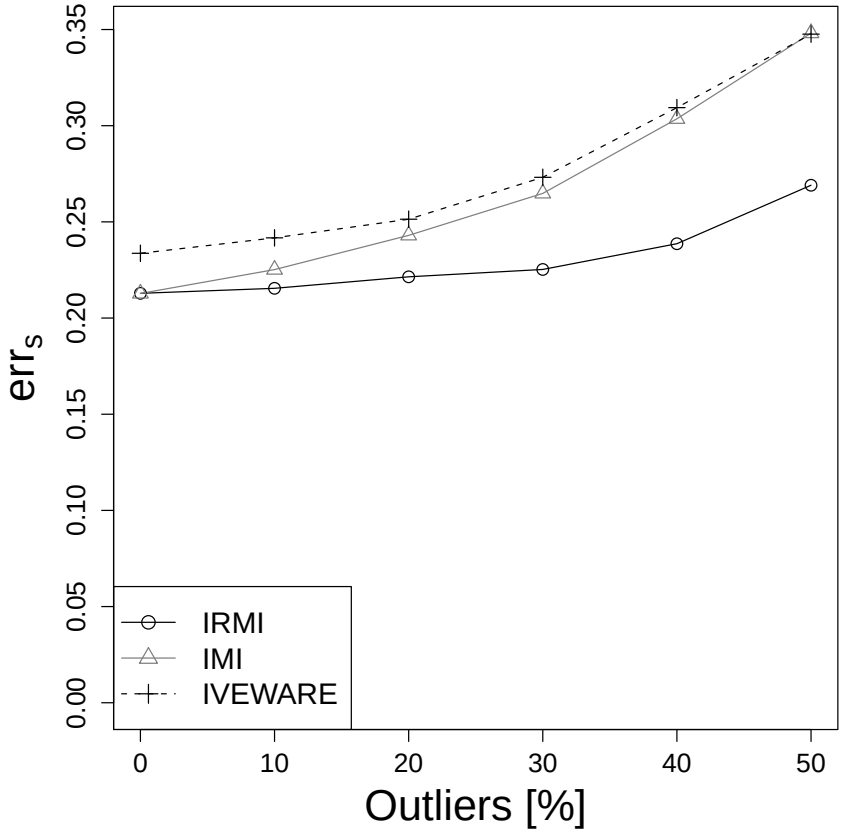
Including (moderate) Outliers and Varying their Amount, high Correlation





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



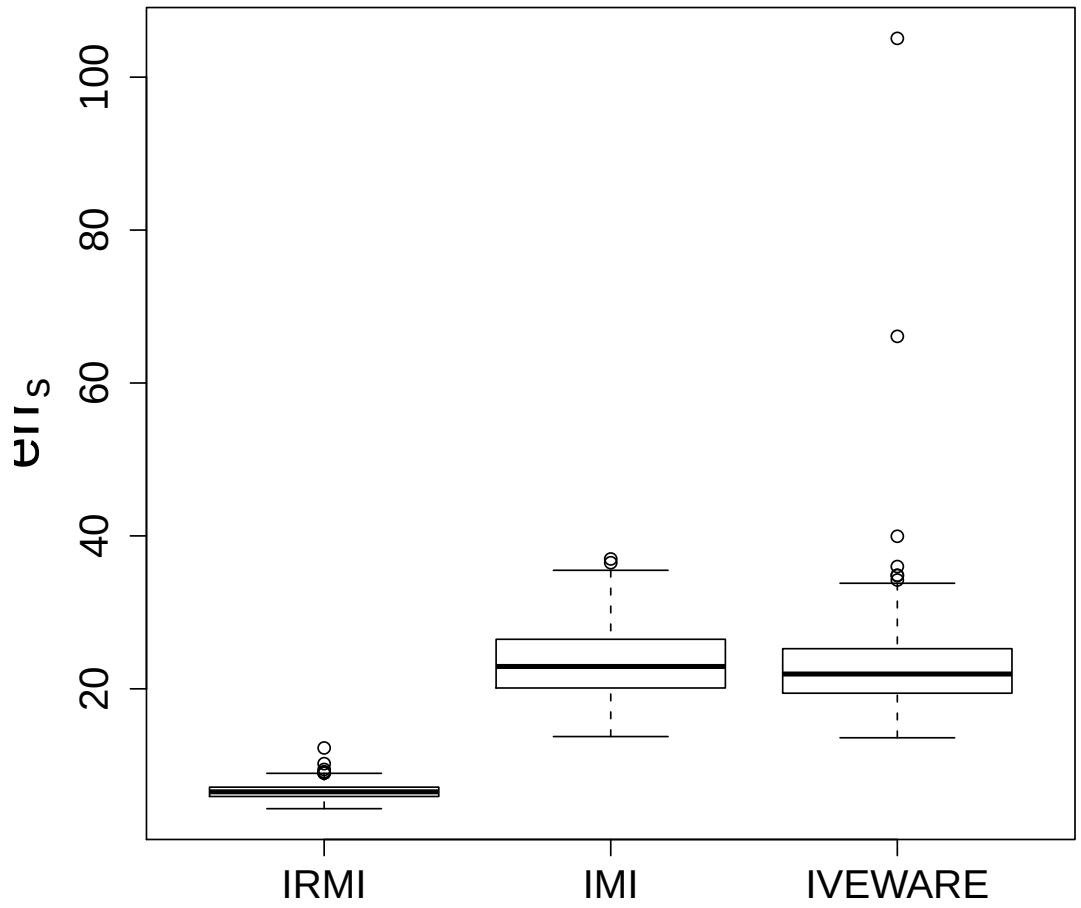


Imputation in EU-SILC

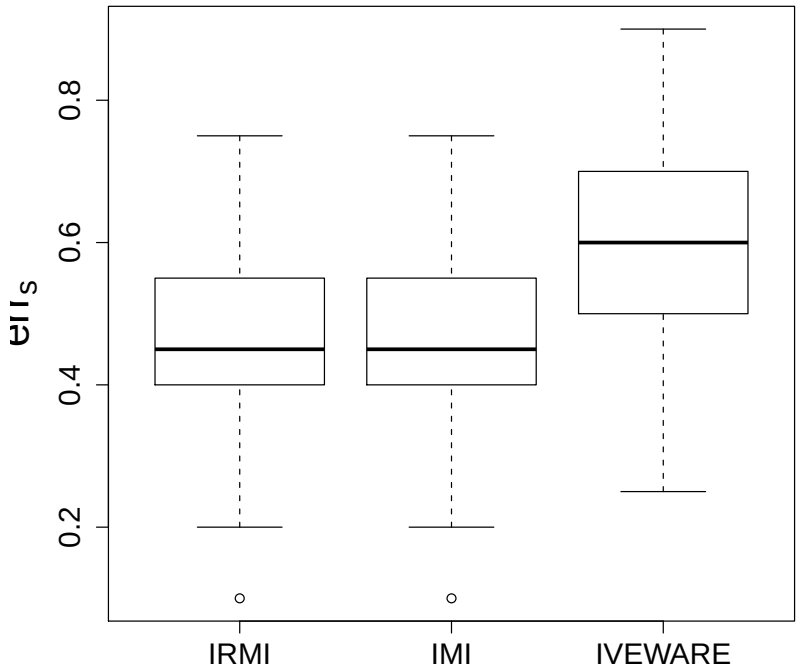
We considered certain HH-components, but also some nominal variables, such as *household size*, *region* and *htype3*.

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 $R = 100$.

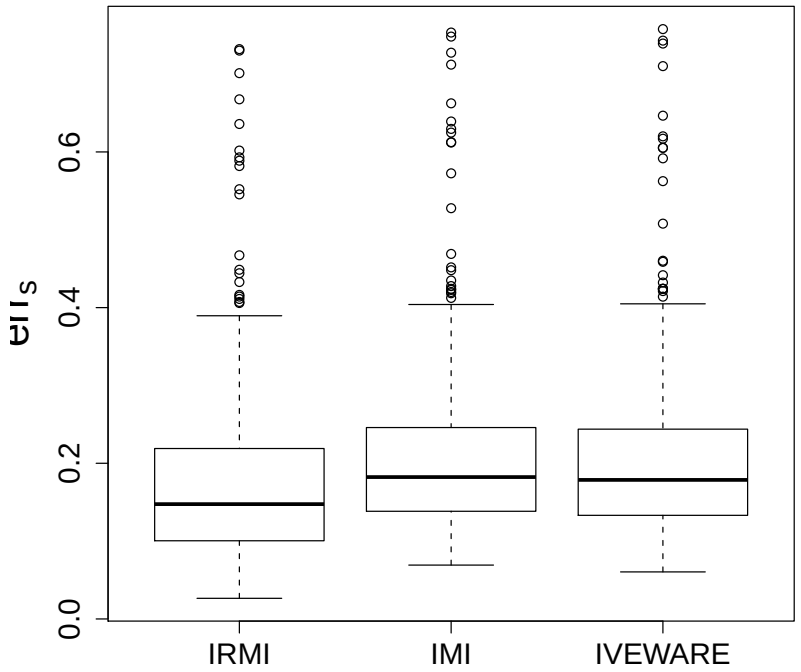
Imputation in EU-SILC, Results



CENSUS Data - no outliers

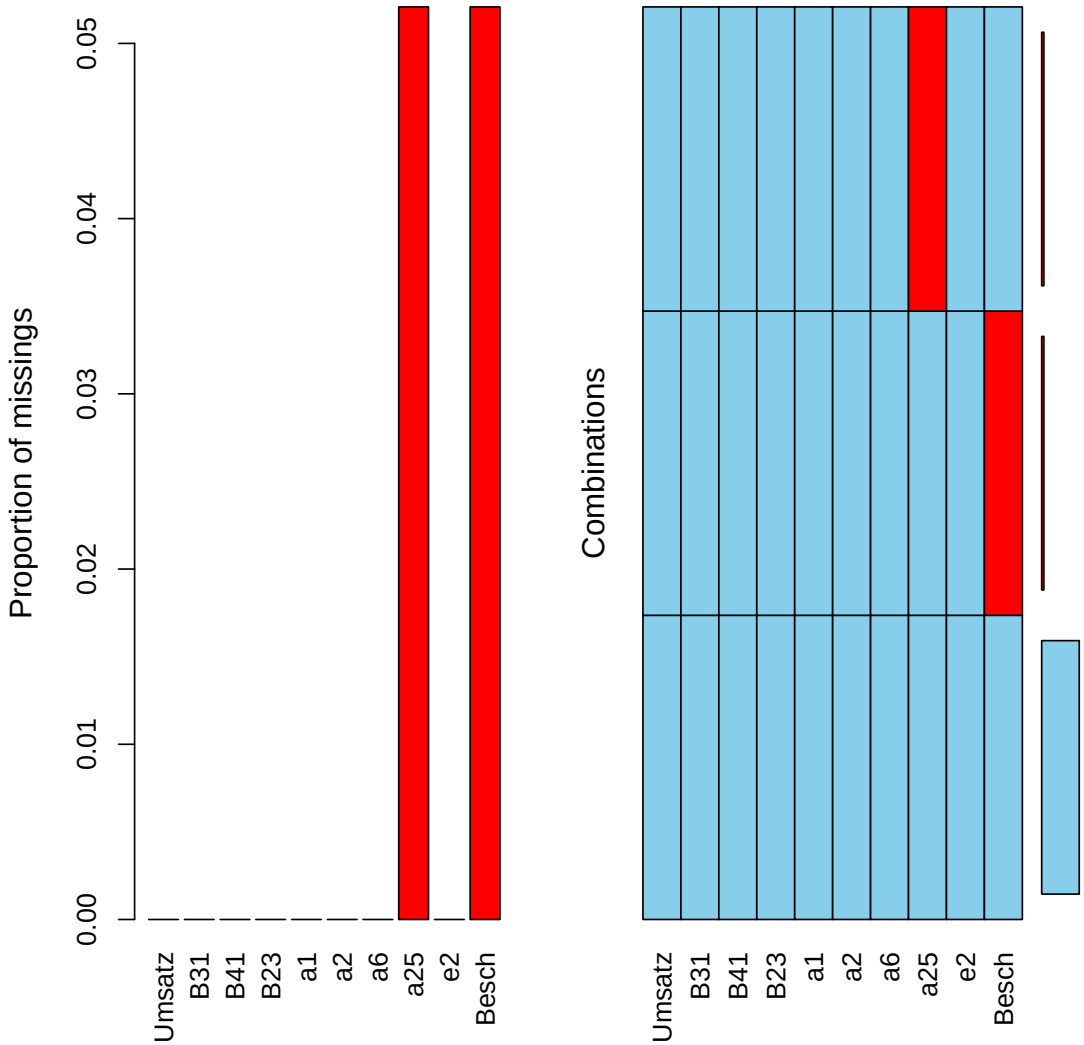


(i) Error for categorical variables



(j) Error for numerical variables

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.

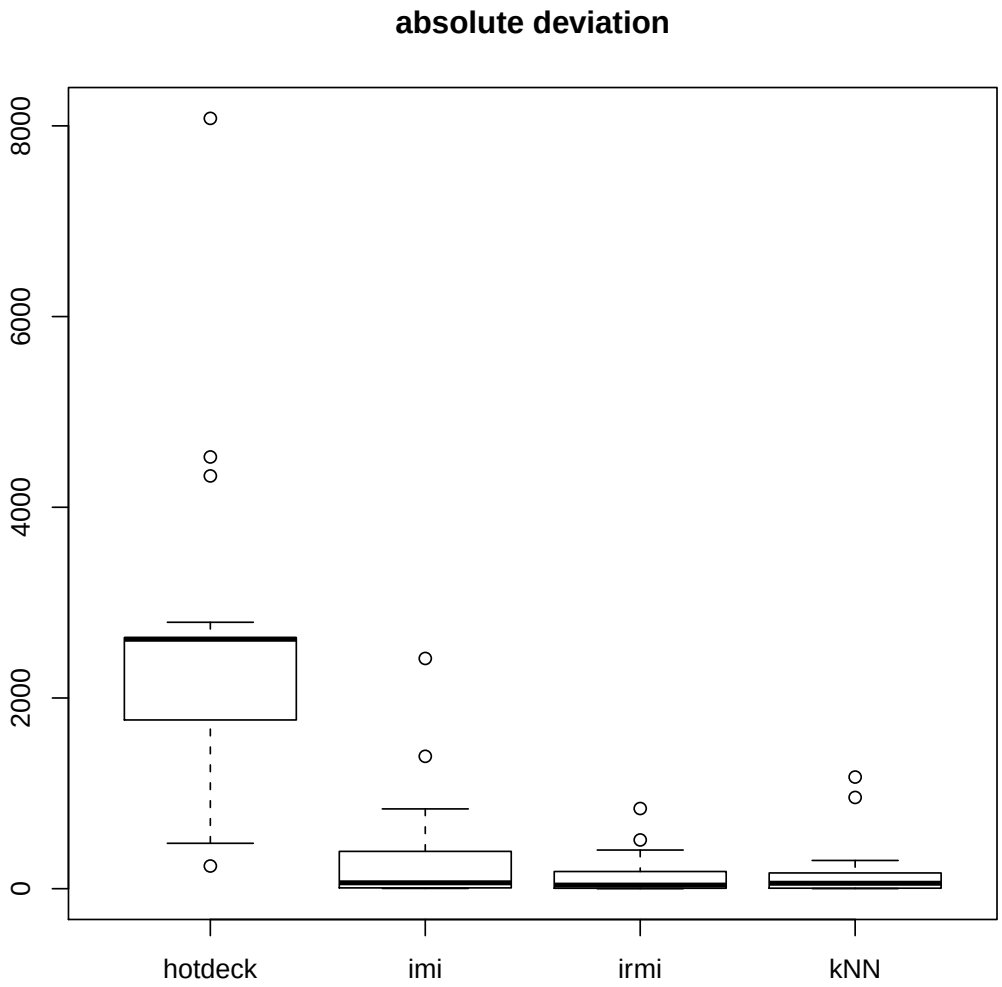
```
kNN(x)
```

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

```
irmi(x)
```

. . . sensible defaults!

SBS data: Simulation Results



Conclusion and Remarks

- `VIM` allows to visualise and impute complex data set.
- VIM is an free and open-source project. It can be freely downloaded at

  http://cran.r-project.org/package=VIM
- IRMI performs almost always best, but hot-deck methods have it's advantages as well (they are very fast and easy understandable)
- For high-dimensional data sets and large plots, the graphics can be simple embedded in TKRplots using sliders to navigate in the plots.
- Currently, highlighting of imputed values is included in `VIM`

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 2.0.4.

M. Templ, A. Kowarik, and P. Filzmoser. Iterative stepwise regression imputation using standard and robust methods. Computational Statistics & Data Analysis, 55:2793–2806, 2011b. ISSN 0167-9473.