WP.21
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

(iv): Micro editing – methods and software

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

Key Invited Paper

Submitted by Department of Statistics and Probability Theory, Vienna University of Technology &
Methods Unit, Statistics Austria &
data-analysis OG (http://www.data-analysis.at)[1]

ABSTRACT

Imputation of item non-responses in complex surveys is of major interest for data providers. Hot and cold deck methods are still popular for this purpose because they are fast and easy to understand. Especially for larger data sets, k-nearest neighbor methods might be the preferable choice in terms of quality of the imputation. The R-Package VIM comes with both techniques. For the latter procedure, the implementation is done not by building the whole distance matrix between the observations, but by searching for neighbors individually for each missing value out of a candidate of possible neighbors which makes the procedure also applicable to large data sets. Moreover, the distribution of each variable can differ, i.e. a variable could be nominal, ordered, continuous but also semi-continuous distributed which made modifications of the methods necessary.
In addition, VIM comes with EM-based regression imputation algorithms using robust methods for statistical estimation of missing values. This algorithm again is able to deal with all data challenges like representative and non-representative outliers and a mixture of different distributions of variables, for example. Using real data sets and a synthetic close-to-reality population where we sample and then model missing values in a realistic manner, we have shown that this method outperforms other popular implementations of EM-based imputation methods in SAS and R.

I. INTRODUCTION

1. Imputation is a research topic present over the last few decades, and especially in the last years many developments have been made due to increasing computing power. The techniques for

[1] Prepared by Matthias Templ (matthias.templ@data-analysis.at), Alexander Kowarik (alexander.kowarik@data-analysis.at) and Peter Filzmoser (p.filzmoser@tuwien.ac.at).

imputation may be divided into univariate methods such as column-wise (conditional) mean imputation, and multivariate imputation. In the latter case there are basically three approaches: distance-based imputation methods such as hot- or cold-deck imputation or k-nearest neighbor imputation, covariance-based methods such as the approaches by Verboven, Branden, and Goos 2007 or Serneels and Verdonck 2008, and model-based methods such as (EM-based) regression imputation.

2. Candidates (donors) for imputation are searched by using hot- or cold-deck imputation methods. Several strategies are possible to choose the donor (randomly from the candidates, one after another, etc.). To decrease the computation time and to be able to impute large data sets, the implementation is done not by building the whole distance matrix between the observations beforehand, but by searching for neighbors individually for each missing value out of a candidate of possible neighbors. In addition, the distribution of each variable can differ, i.e. a variable could be nominal, ordered, continuous but also semi-continuous distributed.

3. If an imputation method is proper (see, e.g., Rubin 1987), i.e. able to deal with the randomness inherent in the data adequately, it can be used for multiple imputation, generating more than one candidate for each missing cell. However, the sampling variability can also be reflected by adding a certain noise to the imputed values, and valuable inference can also be obtained by applying bootstrap methods (Little and Rubin 1987, Alfons, Templ, and Filzmoser 2009).

4. Existing model-based methods assume that the data originate from a multivariate normal distribution (e.g. the MCMC methods of the imputation software MICE (van Buuren and Oudshoorn 2005), Amelia (Honaker, King, and Blackwell 2009), mi (Yu-Sung, Gelman, Hill, and Yajima 2009) or IVEWARE (Raghunathan, Lepkowski, and Hoewyk 2001)). This assumption becomes inappropriate as soon as there are outliers in the data, or in case of skewed or multimodal distributions. However, such challenges in the data come with almost all real-world data sets, and therefore imputation methods based on robust estimates should be applied. These methods should give approximatively the same results when the data originate from a multivariate normal distribution, and should give reliable estimates when certain assumptions like multivariate normality are violated.

5. The basic procedure behind most model-based imputation methods is the EM-algorithm (Dempster, Laird, and Rubin 1977), which can be thought of as a guidance for the iterative application of estimation, adaption and re-estimation. For the estimation, usually regression methods are applied in an iterative manner, which is known under the names regression switching, chain equations, sequential regressions, or variable-by-variable Gibbs sampling (see, e.g., van Buuren and Oudshoorn 2005, Muennich and Rässler 2004).

6. The rest of the paper is organized as follows:
Section II introduces the graphical user interface of the R-package `VIM`. The implemented distance-based imputation methods are outlined in Section III, followed by a brief description of the implemented model-based methodology in Section IV. In Section V a small application is given to show how the implemented imputation methods can be used (with their sensible default parameters) whereas also few visualisation tools to highlight missing values are shown. The final Section VI concludes.

II. THE R-PACKAGE VIM

7. The package `VIM` (Templ, Alfons, and Kowarik 2011, Templ and Filzmoser 2008b) allows to explore and to analyse the structure of missing values in data, to produce high-quality graphics for publications and to impute missing values using various methods.

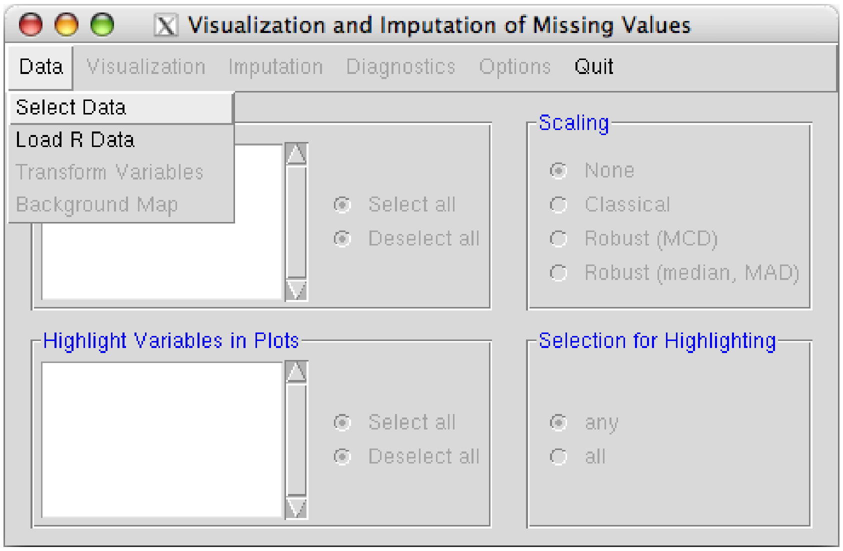

FIGURE 1. The `VIM` GUI and the *Data* menu.

8. The graphical user interface (GUI) has been developed using the R package `tcltk` (R Development Core Team 2011). It allows easy handling of the functions included in the package `VIM`. Figure 1 shows the GUI, which pops up automatically after loading the package, see Listing 1.

LISTING 1. Loading package VIM in R.

```
library(VIM)
```

If the GUI has been closed, it can be re-opened with the command `vmGUImenu()`. All selections and settings from the last session are thereby recovered.
For visualisation, the most important menus are the *Data*, the *Visualization* and the *Options* menus, while for imputation the *imputation* menu is of major importance.

9. The *Data* menu allows to select a data frame from the R workspace (see Figure 2). In addition, a data set in `.RData` format can be imported from the file system into the R workspace, which is then loaded into the GUI directly.
Transformations of variables are available via `Data` → `Transform Variables`. The transformed variables are thereby appended to the data set in use.

10. After a data set has been chosen, variables can be selected in the main dialog (see Figure 3). An important feature is that the variables will be used in the same order as they were selected, which is especially useful for parallel coordinate plots.
For visualisation purposes, variables for highlighting and imputation are distinguished from the plot variables and can be selected separately (see the lower left frame in Figure 3).
With the selected variables, different imputation methods can be applied to the data, either by clicking on the *imputation* button at the graphical user interface or by command line. To avoid a large amount of snapshots from the GUI, we only list the command line input to R in the following.

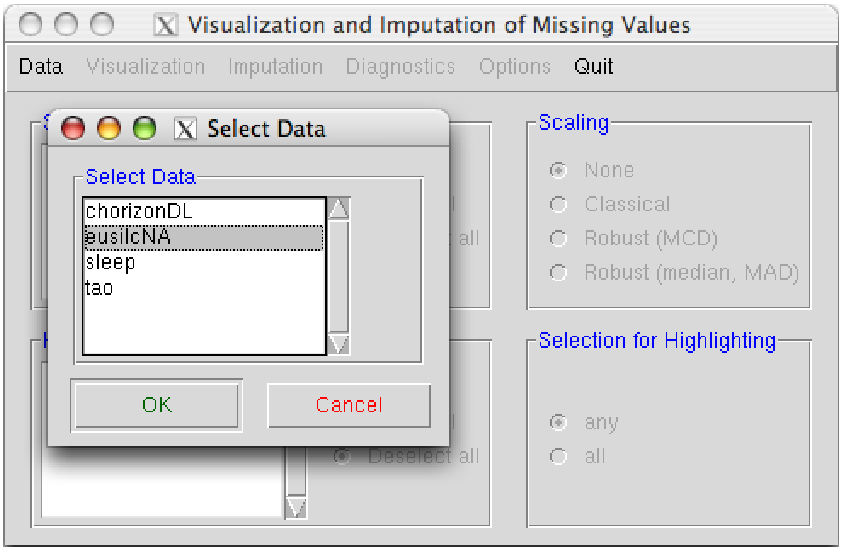

FIGURE 2. The dialog for data selection.

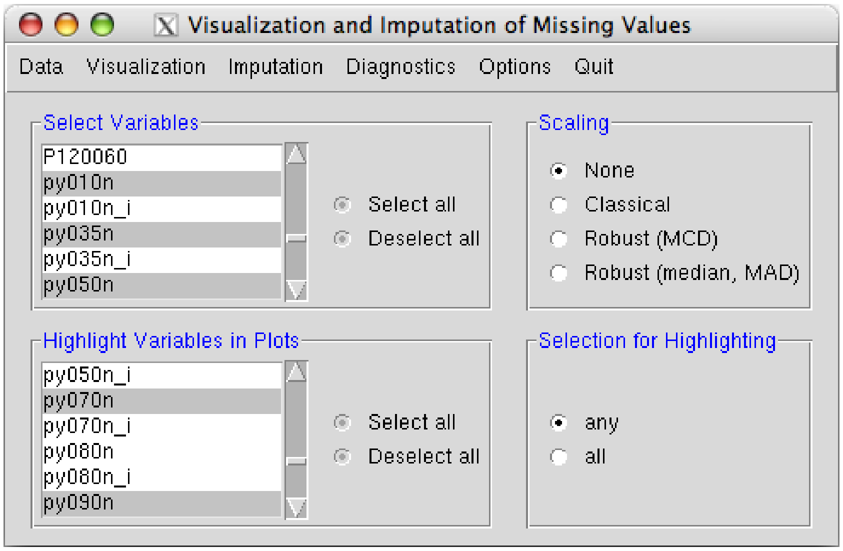

FIGURE 3. Variable selection with the `VIM` GUI.

III. DISTANCE-BASED IMPUTATION

11. The function `hotdeck()` incorporates different popular hot-deck imputation schemes such as *sequential hot deck* and *random (within a domain) hot deck*. Edits to exclude non-confirm donors can also be specified.

12. The function `knn()` can be used for imputing a data set with the popular method k-nearest neighbour. The distances are computed as weighted Gower distances (Box and Cox 1971), which can be calculated for continuous, semi-continuous, binary and categorical variables. Several functions for dealing with the k-nearest neighbours are proposed, e.g. (weighted) median and mean for continuous variables and different methods for selecting a category for imputing a categorical value (e.g. probabilities representing the occurrence of the different categories).

IV. MODEL-BASED IMPUTATION

13. A detailed description of iterative model-based imputation is given in Templ and Filzmoser 2008a and Templ, Kowarik, and Filzmoser 2011. The algorithm implemented in `VIM` is called `IRMI` which stands for iterative robust model-based imputation. The corresponding function is named `irmi()`. Basically, it mimics the functionality of `IVEWARE` (Raghunathan, Lepkowski, and Hoewyk 2001), but there are several improvements like better robustness properties of the estimation procedures. In each step of the iteration, one variable is used as a response variable and the remaining variables serve as the regressors. Thus the "whole" multivariate information will be used for imputation in the response variable. The proposed iterative algorithm can be summarized as follows:

(1) Initialisation of the missing values.
(2) Choose one variable as response and the others as predictors, and update the former missing values in the response.
(3) Go to the next variable and repeat the procedure.
(4) Repeat the whole procedure starting from 2 until convergence.
(5) Add noise to the final estimates in a proper way to allow for multiple imputation.

14. In Templ and Filzmoser 2008a and Templ, Kowarik, and Filzmoser 2011 it is shown that `irmi()` outperforms `IVEWARE` by graphical evaluation of the imputed values and by numerical evaluation of simulated data. `irmi()` also shows better performance for real-world data such as the European Union Statistics of Income and Living Conditions (EU-SILC) survey 2006 from Statistics Austria, the Austrian structural business statistics data (SBS) from 2006, a census data set from 1994 provided by the University of California (for details, see `http://www.ics.uci.edu/~mlearn/MLRepository.html`), and daily air quality measurements in New York, May to September 1973 (see also Chambers, Cleveland, Kleiner, and Tukey 2008).

V. DEMONSTRATION

15. For a small demonstration how to use the package `VIM` we choose the Austrian structural business statistics data (SBS) from 2006 which covers NACE sections C-K for enterprises with 20 or more employees (NACE C-F) or above a specified turnover (NACE G-K) (Eurostat 2008). For these enterprises more than 90 variables are available. Only limited administrative information is available for enterprises below these thresholds. The raw unedited data consist of 21669 observations including 3891 missing values.
Imputation should be made in reasonable subgroups of the data. A detailed data analysis of the raw data has shown that homogeneous subgroups are based on NACE 4-digits level data. Broader categories imply that the data consist of different kinds of sub-populations with different characteristics. For the sake of this study we have chosen the NACE 4-digits level *47.71 - "Retail sale of clothing in specialized stores"* - (Nace Rev. 1.11 *52.42*, ISIC Rev.4 *4771*). This typical NACE 4-digits level data set consists of 199 observations with 7 missing values and various outliers. In order to be able to apply imputation methods reasonably, specific variables were chosen, namely *turnover* (continuous), *number of white-collar employees*, *number of blue-collar workers*, *part-time employees*, *number of employees* (all discrete variables but considered as continuous), *wages*, *salaries*, *supply of trade goods for resale*, *intermediate inputs* and *revenues from retail sales* (considered as continuous variables).
For testing purposes we use the 192 complete observations of this data set. From those observations we set 5% of the values in variables *number of employees* and *intermediate inputs* (these are those variables that include missing values in the original data set) to be missing completely at random. Let x denote this data set.

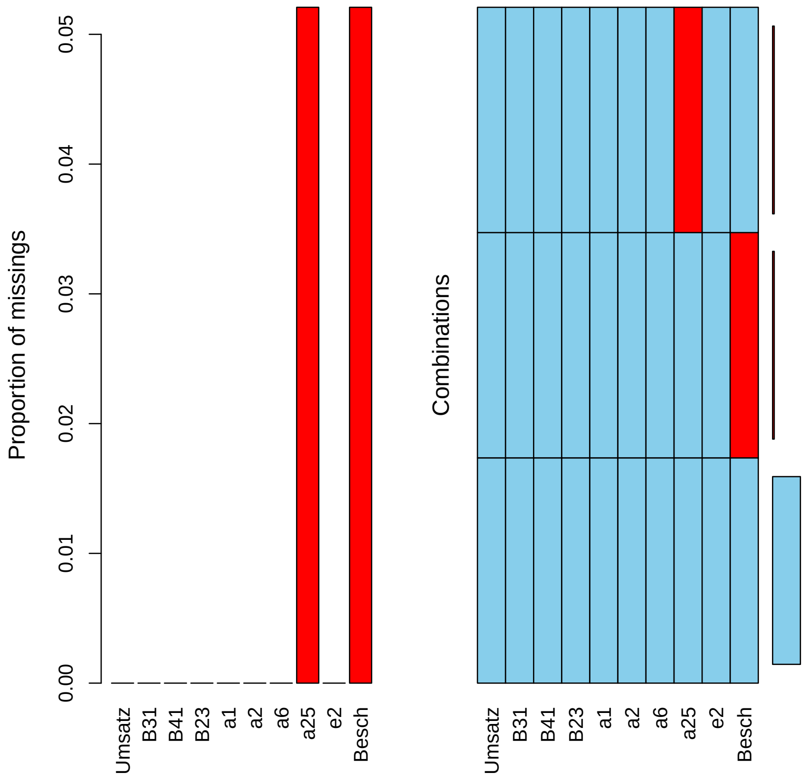

FIGURE 4. Percentage of missing values in the SBS data and combinations of missing values in the variables.

16. Figure 4 shows the percentage of missing values (left graphic) and the combinations of missing values in the variables (right graphic). It has been executed with

LISTING 2. Simple summary of missing values.

```
aggr(x)
```

17. Figure 5 highlights the missing values (in red) for the pairwise bivariate scatterplot of the Box-Cox transformed variables.

18. Figure 6 highlights missing values in a parallel coordinate plot. Both Figure 5 and Figure 6 show the situation missing completely at random. Note, that within the package VIM also other diagnostic plots for missing values may be implemented. Diagnostic plots are especially useful for analysing the structure of missing values.

19. Hotdeck imputation is shown by Listing 3 whereas the ord_var function argument specifies the variables (variable names) to be used to order the dataset. In addition to the function call shown in Listing 3, domains can be specified for imputation as well as a vector of values which should be imputed too, e.g. 8,9 or 98,99 in SPSS-data sets, and a list of conditions for a donor, like ≤ 10000, can be set (see the help file of function hotdeck for a detailed description).

LISTING 3. Hotdeck imputation.

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

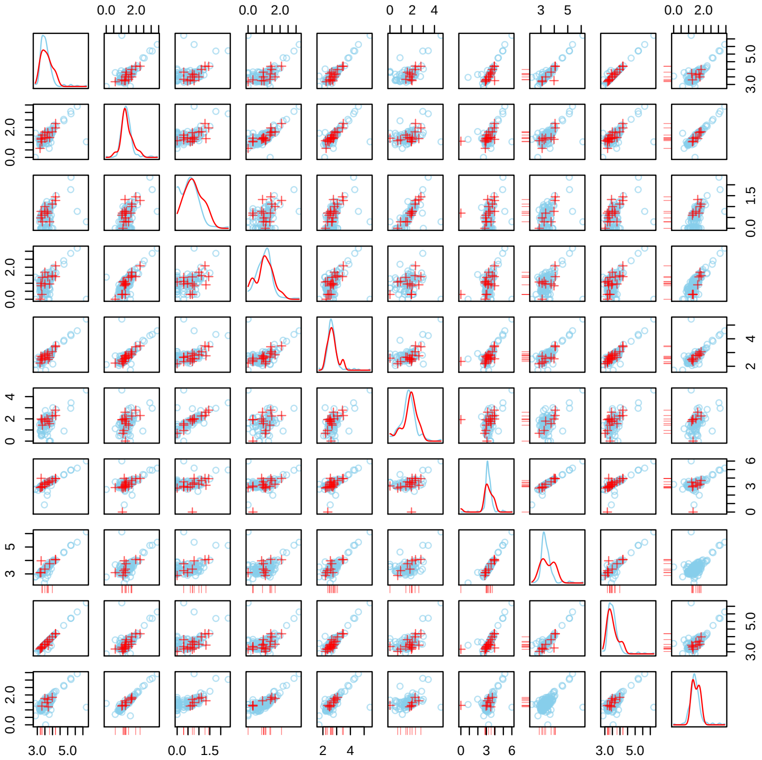

Figure 5. Scatterplot matrix with missing values in employment and intermediate inputs highlighted in red.

20. Listing 4 shows the application of k-nearest neighbor imputation with the R-package VIM. The data set must again be specified. Sensible defaults are given for other function parameters such as variables to be imputed, a method for computing the distances, the number of neighbors, the variables which should be used for distance computation, the vector of sampling weights, functions for evaluating the numerical and the categorical variables, a vector of values which should be imputed too (e.g., 999, 9998 in SPSS data sets) and a list of conditions for a donor (e.g., ≤ 10000).

Listing 4. k-nearest neighbor imputation.

```
kNN(x)
```

21. Iterative model-based imputation using robust methods can easily be applied by (here: the result is stored in object imp).

Listing 5. Application of robust iterative model-based imputation.

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

22. The absolute distances from the original data values to the imputed ones are shown in Figure 7. It is easy to see that hotdeck imputation is not the preferable choice here. This comes also from the fact that hotdeck is more useful when only categorical variables are available. However,

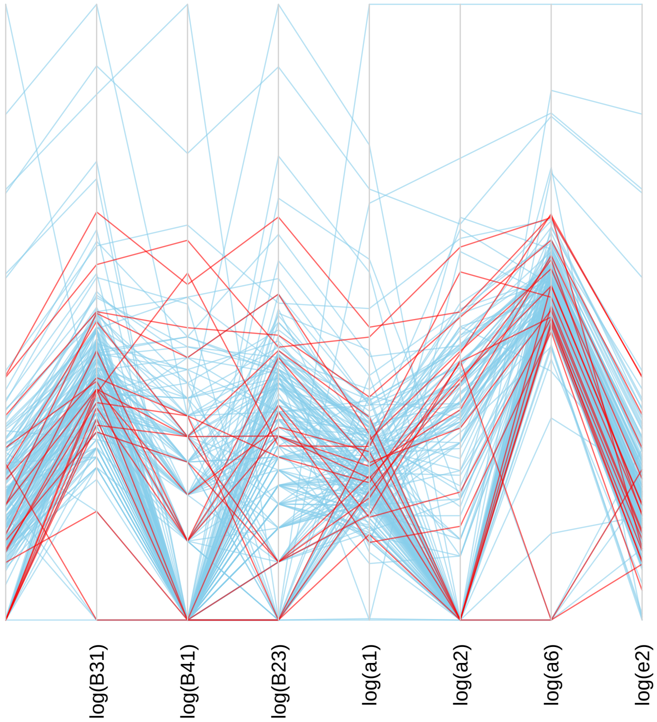

FIGURE 6. Parallel coordinate plot with missing values in employment and intermediate inputs highlighted in red.

k-nearest neighbor imputation should always outperform hotdeck. The best result is obtained by `irmi()` which applies robust methods. Using iterative model-based imputation with classical regression methods (method *imi*), the results become poor.

VI. CONCLUSIONS AND OUTLOOK

23. We showed that the visualisation of missing values is extremely simple with the package `VIM`, either by using the GUI or by typing code on the `R` command line. With the visualisation techniques in `VIM`, it is possible to increase insight into the structure of the data. This is necessary when dealing with missing values, e.g., before imputation is performed.
All real-world data sets we have seen so far, especially in official statistics, include outlying observations, and they often include different types of distributions. All these challenges are considered in the implemented methods, such as hotdeck, k-nn imputation, and iterative robust model-based imputation using robust methods for automatic imputation of missing values and their application.

24. All functions have sensible defaults and various user-friendly tools, like an automatic detection of the distribution of the variables. The application is straightforward and explained in the manual of the package. The package `VIM` can be freely downloaded from the comprehensive R archive network (see `http://cran.r-project.org/package=VIM`).

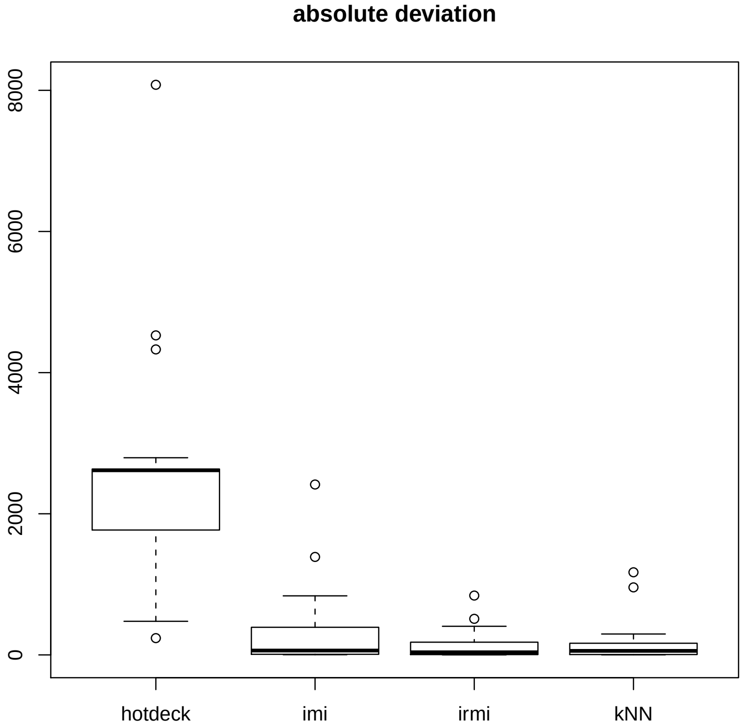

Figure 7. Boxplots of absolute distances from the original values to the imputed ones by different methods.

The application of some imputation methods was shown. A detailed description of the algorithms was not in focus by this contribution, but it can be found in Templ and Filzmoser 2008a and Templ, Kowarik, and Filzmoser 2011.

References

Alfons, A., M. Templ, and P. Filzmoser (2009). On the influence of imputation methods on laeken indicators: Simulations and recommendations. In *UNECE Work Session on Statistical Data Editing; Neuchatel, Switzerland*, pp. 10.

Box, G. and D. Cox (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*, 623–637.

Chambers, J., W. Cleveland, B. Kleiner, and P. Tukey (2008). *Graphical Methods for Data Analysis*. CA: Wadsworth, Belmont.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussions). *Journal of the Royal Statistical Society 39*, 1–38.

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

Honaker, J., G. King, and M. Blackwell (2009). *Amelia: Amelia II: A Program for Missing Data*. R package version 1.2-2.

Little, R. and D. Rubin (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Muennich, R. and S. Rässler (2004). Variance estimation under multiple imputation. In *Proceedings of Q2004 European Conference on Quality in Survey Statistics, Mainz*, pp. 19.

R Development Core Team (2011). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Raghunathan, T., J. Lepkowski, and J. Hoewyk (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*(1), 85–95.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Serneels, S. and T. Verdonck (2008). Principal component analysis for data containing outliers and missing elements. *Computational Statistics & Data Analysis 52*(3), 1712–1727.

Templ, M., A. Alfons, and A. Kowarik (2011). *VIM: Visualization and Imputation of Missing Values.* R package version 1.4.4.

Templ, M. and P. Filzmoser (2008a). EM-based stepwise regression imputation using standard and robust methods. Research report cs-2010-3, Department of Statistics and Probability Theory, Vienna University of Technology.

Templ, M. and P. Filzmoser (2008b). Visualization of missing values using the R-package VIM. Research report cs-2008-1, Department of Statistics and Probability Theory, Vienna University of Technology.

Templ, M., A. Kowarik, and P. Filzmoser (2011). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics and Data Analysis*. under minor revision.

van Buuren, S. and C. Oudshoorn (2005). Flexible multivariate imputation by MICE. Tno/vgz/pg 99.054, Netherlands Organization for Applied Scientific Research (TNO).

Verboven, S., K. Branden, and P. Goos (2007). Sequential imputation for missing values. *Computational Biology and Chemistry 31*, 320–327.

Yu-Sung, S., A. Gelman, J. Hill, and M. Yajima (2009). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*. to appear.