# A Computational and Methodological Framework for Visualisation and Imputation of Missing Values: the R-package VIM

Matthias Templ, Alexander Kowarik, Peter Filzmoser, Andreas Alfons
Statistik Austria, TU Wien, Univ. Wien

**Abstract:**

Data providers or data analysts have to impute item non-responses in complex surveys before analysis. However, first it is necessary to explore the data with missing values and learn about the structure of the missing values. Visualisation methods such as modified parallel coordinate plots, mosaic plots, scatterplot matrices, etc., have to be modified to deal with missing values and to show their structure.

For imputation of missing values, hot and cold deck methods are still popular because they are fast and easy understandable. Especially for larger data sets, k-nearest neighbor methods might be the preferable choice over hot deck methods. For the k-nearest neighbor approach, not the whole distance matrix between the observations is built, but neighbors are searched individually for each missing value out of possible candidate neighbors which makes the procedure also applicable to large data sets. Moreover, the distribution for variables differ, i.e. a mixture of variables of nominal, ordered, continuous but also semi-continuous scale, which make modifications of the methods necessary.

EM-based regression imputation algorithms are mainly used to impute missing values automatically, i.e. such methods are very helpful in hand of subject matter specialists who are not statistical experts. Since data virtually always comes with outlying observations, robust methods for statistical estimation of missing values should be used here. The implemented algorithm (see Templ et al., 2011) again is able to deal with all data challenges like representative and non-representative outliers and a mixture of different distributed 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.

The free and open-source R package VIM provides a graphical user interface for users having no experience with R and it includes all methods which mentioned above.

REFERENCES:

Matthias Templ, Alexander Kowarik, Peter Filzmoser: Iterative stepwise regression imputation using standard and robust methods, Computational Statistics & Data Analysis, Volume 55, Issue 10, 1 October 2011, Pages 2793-2806, ISSN 0167-9473, DOI: 10.1016/j.csda.2011.04.012.