WP. 36
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Neuchatel, Switzerland, 5-7 October 2009)

Topic (vii): Indicators for measuring the quality impact of data editing and imputation

# ON THE INFLUENCE OF IMPUTATION METHODS ON LAEKEN INDICATORS: SIMULATIONS AND RECOMMENDATIONS

**Invited paper**

Submitted by the Department of Statistics and Probability Theory, Vienna University of Technology, and the Department of Methodology, Statistics Austria[1]

## ABSTRACT

The aim of this paper is to estimate the influence of selected imputation methods on the set of Laeken indicators that are computed from EU-SILC data. While the procedures to measure the additional uncertainty from imputation via bootstrap and jackknife are relatively simple, complex simulation studies need to be performed to obtain reliable results for complex data. The R package `simFrame` is a general framework for statistical simulation and is well-suited for this purpose. Using `simFrame`, different imputation methods and their effect on Laeken indicators are evaluated. New multivariate methods based on robust estimates are investigated for future consideration. Extensive close-to-reality simulation will be used in the AMELI project for further studies of the most promising methods.

## I. INTRODUCTION

1. The *Laeken* indicators are a set of indicators for measuring social cohesion, used by European countries as a benchmarking system for the Lisbon Strategy to turn the European Union into the most dynamic and competitive economy in the world (cf. Atkinson, Cantillon, Marlier, and Nolan 2002). Most of these indicators are based on the *European Union Statistics on Income and Living Conditions* (EU-SILC).

2. EU-SILC is an annual survey conducted in EU member states and other European countries that provides comparable cross-sectional and longitudinal data on income and living conditions. Generally speaking, EU-SILC data consists of personal data and household data. More precisely, EU-SILC is composed of four data files, two on the household level and two on the personal level. Income data is split into many personal and household components as well, e.g., *employee cash or near cash income*,

[1]Prepared by Andreas Alfons (alfons@statistik.tuwien.ac.at), Matthias Templ (templ@tuwien.ac.at) and Peter Filzmoser (p.filzmoser@tuwien.ac.at).

*cash benefits or losses from self-employment* or *imputed rent*. These variables are of particular interest for this paper.

3. The subset of Laeken indicators based on EU-SILC is described in EU-SILC 2004. Some of these indicators are poverty indicators such as the *at-risk-of-poverty rate* (ARPR), which is defined as the percentage of individuals with an equivalized disposable income[2] below 60% of the national median equivalized disposable income. Others are inequality indicators, i.e., they measure the inequality of the income distribution rather than the level of income. The *quintile share ratio* (QSR) is an important example for these inequality indicators. It is defined as the ratio of the total equivalized disposable income received by the 20% of a country's population with the highest income to that received the 20% with the lowest income. Given the definition of the QSR, it is clear that it is highly non-robust. A sensitivity analysis and first robustifications are presented in Hulliger and Schoch 2009. Closely related to the QSR, but more complex, is the well-known *Gini coefficient*.

4. Income distributions have been widely studied in the literature with special respect to extreme incomes and inequality. Many of these papers follow a parametric or semi-parametric approach (e.g., Cowell and Victoria-Feser 2007; Vandewalle, Beirlant, Christmann, and Hubert 2007), but non-parametric approaches are discussed as well (e.g., Cowell and Victoria-Feser 2003). Van Kerm 2007 applied many of these techniques to income data in EU-SILC and studied their impact on Laeken indicators. A summary and open-source implementations of semi-parametric methods are given by Holzer 2009.

5. The aim of the research project AMELI (*Advanced Methodology of the European Laeken Indicators*) is to improve the quality of the Laeken indicators. This includes developing suitable imputation methods for EU-SILC data in order to enhance the data quality. Many variables of EU-SILC data are on a nominal, ordinal, or semi-continuous scale. Some income components contain zeros in up to 80% of the observations while the rest of the observations are on a continuous scale, which causes serious problems for the imputation of missing values. This paper is focused on the influence of imputation methods on the Laeken indicators and describes how to account for the additional uncertainty from imputation. Ideas for new imputation methods based on robust estimations are presented as well. Moreover, a simulation study based on real EU-SILC data is performed.

## II. MISSING VALUE MECHANISMS

6. There are three important cases of processes responsible for generating missing values (Rubin 1976; Schafer 1997; Little and Rubin 2002). Let $X = (X_{obs}, X_{miss})$ denote the data, where $X_{obs}$ and $X_{miss}$ are the observed and missing parts, respectively. The missing values are *missing at random* (MAR) if

$$P(X_{miss}|X) = P(X_{miss}|X_{obs}), \tag{1}$$

i.e., the probability of missingness does not depend on the missing part $X_{miss}$.

7. If the probability of missingness does not depend on the observed part $X_{obs}$, the important special case of MAR called *missing completely at random* (MCAR) is obtained, given by

$$P(X_{miss}|X) = P(X_{miss}). \tag{2}$$

[2]The *equivalized disposable income* of a household is defined as its total disposable income divided by its *equivalized size* (according to the modified OECD scale), to take the size and composition of the household into account. By definition, the equivalized disposable income is assigned to every household member, including children.

8. On the other hand, if Equation (1) is violated and missingness is in some way related to the outcome variables, i.e., the probability of missingness depends on $X_{miss}$, the missing values are said to be *missing not at random* (MNAR). This translates to the equation

$$P(X_{miss}|X) = P(X_{miss}|(X_{obs}, X_{miss})). \tag{3}$$

Hence the missing values can not be fully explained by the observed part of the data in this case.

9. A motivational example is given in (Little and Rubin 2002). Consider two variables *age* and *income*, with missing values in income. If the probability of missingness is the same for all individuals, regardless of their age or income, then the data are MCAR. If the probability that income is missing varies according to the age of the respondent, but does not vary according to the income of respondents with the same age, then the data are MAR. If the probability that income is recorded varies according to income for those with the same age, then the data are MNAR. Note that MNAR *cannot* be detected in practice, as this would require knowledge of the missing values themselves.

10. In order to select an appropriate imputation method (especially for model-based imputation), it is necessary to know the multivariate structure of the missing values beforehand. The R package `VIM` (Templ and Filzmoser 2008; Templ and Alfons 2009) contains visualization tools that allow not only to detect the missing value mechanisms (MAR or MCAR), but also to gain insight into the quality and various other aspects of the underlying data.

## III. ESTIMATING IMPUTATION UNCERTAINTY

11. Different approaches for estimating standard errors that account for the additional uncertainty from imputation are presented in Little and Rubin 2002. Modified bootstrap or jackknife procedures may be applied in order to obtain consistent estimates for the standard errors for imputed data. It is well known that the jackknife requires smooth statistics, therefore it not suitable for estimating the variability of Laeken indicators. Instead, the bootstrap approach is used in this paper. One replication of the modified bootstrap procedure consists of three steps:

(1) Draw a bootstrap sample from the raw data (i.e., the original unimputed data).
(2) Impute the missing values in the bootstrap sample from step (1).
(3) Estimate the quantity of interest (for this paper, one of the Laeken indicators) using the imputed bootstrap sample from step (2).

From these bootstrap replicates, standard errors and confidence intervals can be obtained. Various methods for computing confidence intervals exist, e.g., Efron's percentile method (see, e.g., Efron and Tibshirani 1993). They are in general slightly larger than their counterparts obtained by the standard bootstrap with already imputed data. As an example, Templ, Filzmoser, and Hron 2009 showed that mean imputation – a simple method still frequently used – may lead to higher uncertainty and bias.

12. Multiple imputation constitutes another possibility to incorporate the additional variability due to imputations, which is less computationally expensive than resampling methods. However, the imputations must be *proper* in order to lead to consistent standard errors (see, e.g., Rubin 1987; Nielson 2003). While a mathematical analysis whether imputations are proper in Rubin's sense is virtually impossible for complex methods based on robust estimation, the problem can be addressed with Monte Carlo simulation studies. Rässler 2004 gives a detailed description on how to use simulations to investigate if a multiple imputation method is proper or at least approximate proper.

## IV. IMPUTATION METHODS

13. Clearly, univariate imputation methods such as mean imputation fail to reflect the multivariate structure of the observed data in the filled-in data set. Even simplified distance-based methods such as single donor imputation suffer from the same problem. Moreover, as mentioned in Section I, EU-SILC data contain many variables on a nominal, ordinal, or semi-continuous scale. When dealing with such complex data sets, most of the existing imputation methods cannot be applied due to these mixed scale variables.

14. Regression-based (multiple) imputation of mixed continuous and categorical data is described in Schafer 1997. However, the corresponding routines implemented in the R package `mix` (Schafer 2009) cannot handle the case that all observations of one category are missing. In addition, these methods are not suitable for semi-continuous variables. This is also true for several other similar multiple imputation methods.

15. Raghunathan, Lepkowski, Van Hoewyk, and Solenberger 2001 developed an iterative (multiple) imputation method that uses a sequence of regression models. It is implemented in the software `IVEware` and frequently used in official statistics. In every iteration, the data are split into a response and predictor variables. Based on the distribution of the response, a specific regression model is chosen. The missing values are thereby updated by drawing a value from the corresponding prediction interval. An advantage of this method is that it can be applied to mixed scale data, including semi-continuous variables. Nevertheless, it is based on non-robust estimates and might thus be seriously influenced by outlying observations.

16. An iterative imputation procedure for compositional data based on (robust) regression models is presented in Hron, Templ, and Filzmoser 2008 and implemented in the R package `robCompositions` (Templ, Hron, and Filzmoser 2009). Even though this method is designed for the application to compositional data, it can be modified for non-compositional data containing mixed scale variables by using different (robust) regression models according to the distribution of the response. Furthermore, this method is suitable for multiple imputation, e.g., to incorporate the additional uncertainty from imputation into variance estimation.

17. $k$-nearest neighbor ($k$nn) imputation methods (e.g., Troyanskaya, Cantor, Sherlock, Brown, Hastie, Tibshirani, Botstein, and Altman 2001) follow an entirely different approach. The idea of $k$nn imputation is that the value for an empty data cell depends on the $k$ most similar observations with available information in the corresponding variable. However, in the case of EU-SILC data, appropriate distance measures for nominal, ordinal and (semi-)continuous variables must be chosen. Reasonable choices are the Hamming distance for the nominal variables, the Manhattan distance for the ordinal variables and the Euclidean distance for the (semi-)continuous variables. Note that the ordinal and the (semi-)continuous variables should be on a comparable scale, respectively. Nevertheless, another crucial question is how to combine these different distances into a single distance measure.

## V. SIMULATIONS USING THE R PACKAGE `simFrame`

18. In research projects involving many scientists or institutions, it is absolutely necessary to agree beforehand upon common guidelines for simulation studies regarding simulation designs, outlier or missing data models, evaluation criteria, etc. Otherwise the results obtained by different participants will be incomparable, thus making it impossible to draw meaningful conclusions from the simulations.

19. To simplify using such common guidelines for simulation experiments, an object-oriented framework for statistical simulation has been implemented in the R package `simFrame` (Alfons 2009). The open-source statistical environment R (R Development Core Team 2009) was chosen for two main reasons. First, it has become the main framework for computing in statistics research. Second, it includes a well-developed programming language and provides interfaces to many others, including the fast low-level languages C and Fortran.

20. With `simFrame`, researchers can make use of a wide range of simulation designs with a minimal amount of programming. Its object-oriented implementation based on S4 classes and methods (Chambers 1998; Chambers 2008) gives maximum control over input and output, while at the same time providing clear interfaces for extensions by user-defined classes and methods.

21. Due to the possibility of inserting missing values into specified variables in every simulation run, `simFrame` is well-suited for evaluating imputation methods. Different missing value rates may thereby be used for different variables. In addition, it is possible to generate MCAR, MAR or MNAR situations. In order to investigate robust methods, a certain proportion of the data may be contaminated. The existing framework may also be extended with user-defined missing data or contamination models. Furthermore, an appropriate plot method is selected automatically depending on the structure of the simulation results.

## VI. PRELIMINARY RESULTS

22. The most promising imputation methods, along with newly developed robust methods, will be evaluated in the AMELI project using extensive close-to-reality simulations with an artificial population (see, e.g., Münnich, Schürle, Bihler, Boonstra, Knotterus, Nieuwenbroek, Haslinger, Laaksonen, Wiegert, Eckmair, Quatember, Wagner, Renfer, and Oetliker 2003). However, as an artificial population containing all necessary variables is not yet available, this paper is confined to preliminary results based on the public use sample of the Austrian EU-SILC data from 2004 (Statistics Austria 2006; Statistics Austria 2007). The Laeken indicators computed from this data set are treated as the true values and reference variances are obtained via bootstrap with 100 repetitions. Note that in this simulation experiment, bootstrapping is done by households rather than individuals, because when the additional uncertainty from imputation is incorporated into variance estimation, the equivalized household income needs to be recomputed after imputation for every bootstrap sample. A simple weight adjustment is used to ensure that the weights of every bootstrap sample sum up to the population total. Moreover, the same bootstrap samples that are used for obtaining the reference variances are used in every simulation run for maximum comparability.

23. In the simulation study presented in this section, the influence of imputation of the personal income components is of interest. To keep it simple, missing values are only generated in two variables, *PY010N* (employee cash or near cash income) and *PY050N* (cash benefits or losses from self-employment). This is not a major restriction, as these two variables are responsible for a large part of the personal income. In order to investigate a realistic situation, 10% of *PY010N* and 5% of *PY050N* are set to missing values in every simulation run. The probability of missingness thereby depends on age and personal net income, hence the missing value mechanism is MNAR. Of course, missing values are only inserted for adults[3], as well as imputation is only performed on the subset of adults.

[3]Individuals of 14 years or more are considered adults on the modified OECD scale.

24. EU-SILC data typically contains a small proportion of outliers. Therefore, 1% of the data were selected randomly to be contaminated. It should be noted that the contamination is fixed throughout the whole simulation experiment. For one half of the selected observations, the values in *PY010N* were replaced by values from a normal distribution with mean $\mu = 1\,000\,000$ and standard deviation $\sigma = 50\,000$. For the other half, the values in *PY050N* were replaced by values from the same normal distribution. This ensures that the contaminated observations are true outliers and not leverage points for regression-based imputation methods. As for inserting missing values, only the income of adults is contaminated.

25. Point estimates are obtained after performing weighted mean imputation, as well as modifications of the IMI (iterative model-based imputation) and IRMI (iterative robust model-based imputation) methods by Hron, Templ, and Filzmoser 2008, respectively. Variance estimates for these three imputation methods are computed both excluding and including the additional uncertainty (see Section III). Furthermore, only households that do not contain outliers are used for computing the point estimates and the bootstrap replicates for variance estimation. A comparison with the regression-based method by Raghunathan, Lepkowski, Van Hoewyk, and Solenberger 2001 and a suitable $k$nn approach (see Section IV) is of high interest and is future work.

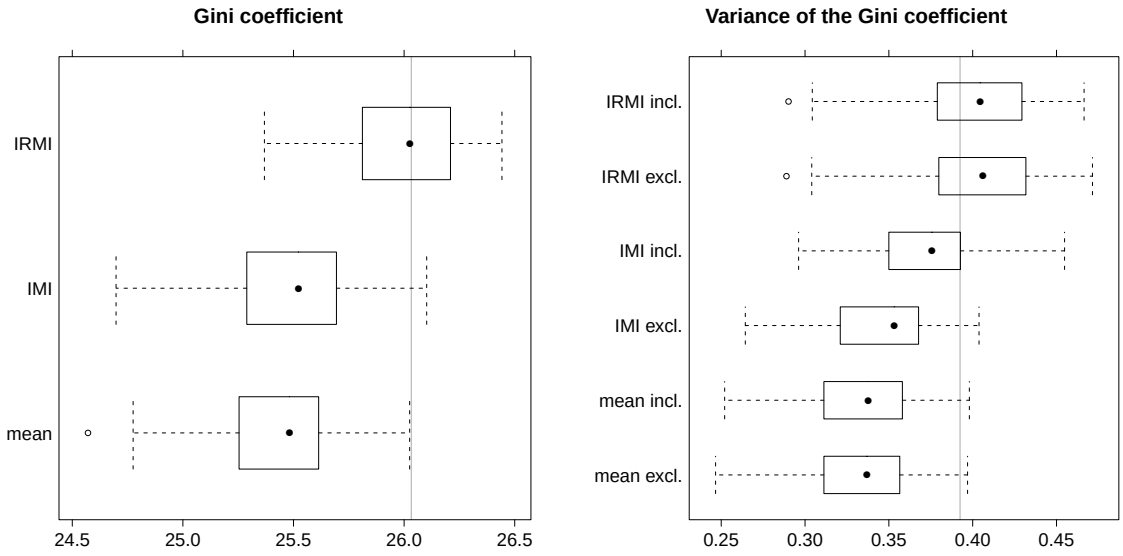


FIGURE 1. Point and variance estimates for the Gini coefficient from 100 simulation runs. The vertical lines represent the "true" values, respectively.

26. Figure 1 shows point and variance estimates for the Gini coefficient from 100 simulation runs. Clearly, weighted mean imputation and IMI result in a significant negative bias, whereas the IRMI approach gives unbiased results for the Gini coefficient. While the reference variance is underestimated with weighted mean imputation and IMI, it is slightly overestimated with IRMI. However, only IMI imputation results in additional uncertainty, which is also the case for the at-risk-of poverty rate (ARPR) and the quintile share ratio (QSR). The latter results are not shown in this paper due to the limited space.

27. To further investigate this phenomenon, a different type of estimator is considered in the following. Point and variance estimates for the weighted mean of the equivalized disposable income are displayed in Figure 2. Weighted mean imputation and IMI result in a large amount of additional

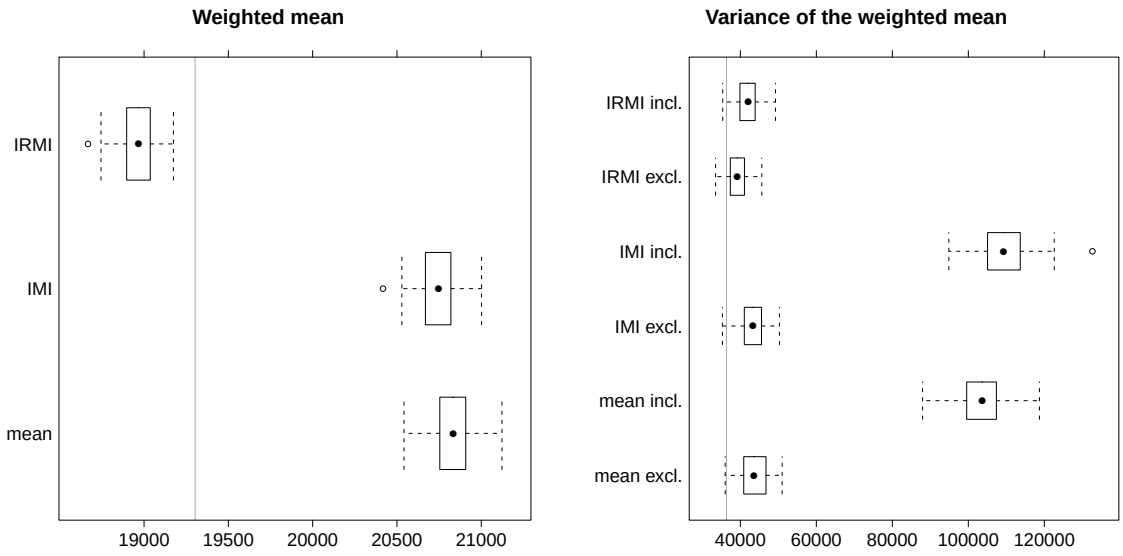


FIGURE 2. Point and variance estimates for the weighted mean of the equivalized disposable income from 100 simulation runs.

uncertainty. The amount of additional uncertainty is much less with IRMI, but still significant. Hence it depends on both the imputation method and the estimator whether additional uncertainty arises from imputation.

## VII. CONCLUSIONS AND OUTLOOK

28. While unsuitable imputation methods lead to a bias in almost any estimator, the effect of missing values/imputation on variance estimation also depends on the type of estimator. Concerning the Laeken indicators, little to no additional uncertainty is introduced by various imputation methods. On the other hand, imputation results in larger variances for estimators such as the weighted mean. Bootstrapping allows to measure the additional uncertainty from imputation and considers the imputation method used and the (type of) estimator automatically. The results show that imputation using robust methods leads to much better results than non-robust methods.

29. In addition, we mentioned our R package `simFrame`, which could act as a simulation framework to carry out simulations on complex data sets from official statistics, but also for simulations on data sets related to other research fields. Further simulations will be set up within the AMELI project to compare more imputation methods on samples from close-to-reality artificial populations.

## ACKNOWLEDGMENTS

This work was partly funded by the European Union (represented by the European Commission) within the 7$^{th}$ framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI (Advanced Methodology for European Laeken Indicators), Grant Agreement No. 217322). Visit http://ameli.surveystatistics.net for more information.

## References


Alfons, A. (2009). *simFrame: Simulation Framework*. R package version 0.1.

Atkinson, T., B. Cantillon, E. Marlier, and B. Nolan (2002). *Social Indicators: The EU and Social Inclusion*. New York: Oxford University Press. ISBN 0-19-925349-8.

Chambers, J. (1998). *Programming with Data*. New York: Springer. ISBN 0-387-98503-4.

Chambers, J. (2008). *Software for Data Analysis: Programming with R*. New York: Springer. ISBN 978-0-387-75935-7.

Cowell, F. and M.-P. Victoria-Feser (2003). Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality 1*(3), 191–219.

Cowell, F. and M.-P. Victoria-Feser (2007). Robust stochastic dominance: A semi-parametric approach. *Journal of Economic Inequality 5*(1), 21–37.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

EU-SILC (2004). Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. EU-SILC 131-rev/04, Working group on Statistics on Income and Living Conditions (EU-SILC), Eurostat, Luxembourg.

Holzer, J. (2009). Robust methods for the estimation of selected Laeken indicators. Master's thesis, Vienna University of Technology, Vienna, Austria.

Hron, K., M. Templ, and P. Filzmoser (2008). Imputation of compositional data using classical and robust methods. Research Report SM-2008-4, Department of Statistics and Probability Theory, Vienna University of Technology.

Hulliger, B. and T. Schoch (2009). Robustification of the quintile share ratio. In *Proceedings of the International Conference on New Techniques and Technologies in Statistics (NTTS 2009), Brussels*.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). New York: Wiley. ISBN 0-471-18386-5.

Münnich, R., J. Schürle, W. Bihler, H.-J. Boonstra, P. Knotterus, N. Nieuwenbroek, A. Haslinger, S. Laaksonen, R. Wiegert, D. Eckmair, A. Quatember, H. Wagner, J.-P. Renfer, and U. Oetliker (2003). Monte Carlo simulation study of European surveys. DACSEIS Deliverables D3.1 and D3.2, University of Tübingen.

Nielson, S. (2003). Proper and improper multiple imputation. *International Statistical Review 71*(3), 593–627. With discussion.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

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

Rässler, S. (2004). The impact of multiple imputation for DACSEIS. DACSEIS research paper 5, DACSEIS research project.

Rubin, D. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

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

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC. ISBN 0-412-04061-1.

Schafer, J. (2009). *mix: Estimation/multiple Imputation for Mixed Categorical and Continuous Data*. R package version 1.0-7.

Statistics Austria (2006). Einkommen, Armut und Lebensbedingungen 2004: Ergebnisse aus EU-SILC 2004. Technical report, Statistics Austria, Vienna, Austria. ISBN 3-902479-59-0.

Statistics Austria (2007). EU-SILC 2004: Erläuterungen: Mikrodaten-Subsample für externe Nutzer. Technical report, Statistics Austria, Vienna, Austria.

Templ, M. and A. Alfons (2009). *VIM: Visualization and Imputation of Missing Values*. R package version 1.3.

Templ, M. and P. Filzmoser (2008). 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., P. Filzmoser, and K. Hron (2009). Robust imputation of missing values in compositional data using the R-package robCompositions. In *Proceedings of the International Conference on New Techniques and Technologies in Statistics (NTTS 2009), Brussels*.

Templ, M., K. Hron, and P. Filzmoser (2009). *robCompositions: Robust Estimation for Compositional Data*. R package version 1.2.2.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics 17*(6), 520–525.

Van Kerm, P. (2007). Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. IRISS Working Paper 2007-01, IRISS at CEPS/INSTEAD.

Vandewalle, B., J. Beirlant, A. Christmann, and M. Hubert (2007). A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics & Data Analysis 51*(12), 6252–6268.