Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FB Psychologie
Psychologische Methoden

Discussion Paper

No. 3

A Comparison of Multiple Imputation Techniques

October 2015

Daniel Salfran & Martin Spiess

Psychologische Methoden, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany

A Comparison of Multiple Imputation Techniques

Daniel Salfran

and

Martin Spiess*

*Corresponding author: Martin Spiess, Psychological Methods and Statistics, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany; e-mail: martin.spiess@uni-hamburg.de. The authors gratefully acknowledge financial support from the German Science Foundation via grant SP 930/8-1.

A Comparison of Multiple Imputation Techniques

Abstract

Most datasets are affected by item nonresponse. Software packages offering multiple imputation (MI) to compensate for missing values often provide several techniques to generate MIs like Parametric Bayesian imputation or k-nearest-neighbor (kNN) imputation. Given the variety of imputation techniques, a question in applications is which of the techniques to choose. In this paper we perform a comparison of various standard imputation techniques as provided by software packages and two newly proposed robust imputation techniques under different scenarios. The comparisons are based on simulations, but some aspects are illustrated using data collected to estimate functions fulfilled by music. The results imply that none of the imputation techniques works well under all scenarios considered: If the imputation model is correctly specified, then parametric methods allow valid inferences. However, if the (conditional) distribution of the variable to be imputed is misspecified, then neither these nor kNN techniques may lead to valid inferences. Robust, semiparametric techniques may work better in these situations, but fail if the variable to be imputed follows a heavy tailed distribution. However, results depend on sample size: In small samples the self-correcting property of MI seem to be at work, i.e., point estimators seem to be biased but coverage rates are acceptable due to overestimated variances, whereas structural problems of these techniques become obvious in larger samples.

Keywords: Missing Data; Multiple Imputation; Misspecified Imputation Models; Robustness

1 Introduction

Item nonresponse, i.e. not observing the reactions to items of otherwise observed units, is an obstacle in every statistical analysis. If item nonresponse is selective with respect to the research question, then simply ignoring the incompletely observed cases or applying other 'ad hoc' methods usually leads to invalid inferences (e.g. Little and Rubin, 2002; Rubin, 1987; Schafer and Graham, 2002). But even if the observed part of the sample is not selective due to nonresponse, it is often wise to compensate for missing values to avoid a big loss of information if only completely observed cases are analyzed (complete case analysis, CCA), which is the default option in many standard software packages for incompletely observed data sets.

Multiple imputation (MI; Rubin, 1987) is one established method to compensate for missing data. The idea behind MI is to generate several (M) predictions for each missing value, analyze the multiply imputed data set M times with standard software for completely observed data sets and then combine the M estimation results according to simple rules given in Rubin (1987).

Various imputation techniques are available in statistical software packages (see e.g., Horton and Kleinman, 2007), but not all of them are theoretically justified by Rubin (1987). Available techniques are based on strong assumptions, like Bayesian imputation as proposed by Rubin (1987), less strict assumptions, like techniques that impute donor values from the same data set (k-nearest-neighbor or kNN techniques; e.g., Little, 1988), which are believed to be more robust with respect to a misspecification of distributional assumptions (cf. Andridge and Little, 2010), or techniques based on weak assumptions like so-called robust imputation methods (e.g., Templ, Kowarik and Filzmoser, 2011). Hence, in applications it is often not obvious which of the techniques allow valid inferences under which conditions, and which do not.

Therefore, in this paper we compare various imputation techniques which are provided by leading software packages. More precisely, we are interested in: (1) how robust are

inferences given a misspecified imputation model, (2) do kNN techniques and robust imputation techniques generally lead to valid results and, (3), is it always better to adopt some imputation method as compared to ignoring the missing data problem and analyze only the completely observed cases? In addition, we will investigate if results of kNN imputation depend on the number of possible donors k. Dahl (2007) considers this topic, which seems not to have been picked up in the literature. As robust methods may not work well in small samples, we also vary the sample sizes to separate structural problems of the methods from sample size issues.

Evaluations and comparisons are first done via simulations and the different techniques are then applied to a real data example. The imputation techniques are evaluated with respect to the finite sample properties of parameter estimators in a regression model with missing values in a continuous predictor variable, and with respect to the frequency properties of confidence intervals under various conditions. One factor varied is the distribution of the predictor variable with missing values. Although distributions of predictor variables in the models of scientific interest are usually not of interest because statements are conditional on the realized values, they become relevant if missing predictor values are imputed. Another factor varied is the model that generates the missing data.

This paper is organized as follows: In Section 2 we describe the predominantly used classification of missing values. This section also provides a short introduction into the method of multiple imputation. Section 3 describes and discusses the various imputation techniques considered. With one exception, they are all – although not exclusively – available in the software package R. The setup of the simulation experiment is described in Section 4, and the results are presented in Section 5. Section 6 provides a real world example. Finally the results and their implications are discussed in Section 7.

2 Multiple Imputation

Based on Rubin (1976, 1987), missing data are either missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). Missing data are MCAR if the probability of the observed pattern of observed and missing data does not depend on any of the other variables relevant to the analysis of scientific interest, observed or not. They are MAR if this probability does depend on observed values of the relevant variables but not additionally on unobserved values of relevant variables. Most of the methods available in software packages are based on the assumption of missing data being MCAR or MAR. If the missing data are MCAR, then the observed part of the sample is just a non-selective random subsample of the intended complete sample. Finally, the unobserved data are MNAR if the probability of the pattern of observed and missing data does depend not only on observed but in addition on unobserved values of variables relevant to the research question. In this latter case, strong assumptions or external knowledge is usually necessary to compensate for the missing data. This paper deals with missing values being MAR.

The idea behind MI (Rubin, 1987) is to generate multiple (M) predictions ('imputations') for each missing datum, thus generating M completed versions of the incompletely observed data set each of which can be analyzed using standard methods. The M results are then combined according to simple rules to allow inferences. MI is developed based on a Bayesian approach and for the final analyses to allow valid inferences, the analysis method applied to the data set without missing values should be valid and the imputation method has to be proper. According to Rubin (1987) an imputation method is proper for the estimators of interest if, given the assumed response mechanism and an appropriate model for the data, the imputations are independent draws from the predictive posterior distribution of the variables with missing values given all other relevant observed variables. If the multiple imputation method is proper for the estimators of interest, then the analysis using the multiply imputed data set tends to be valid even in the frequentist

sense. Besides Bayesian methods, other methods like the approximate Bayesian Bootstrap can be proper as well (for a more extensive discussion, see Rubin, 1987, 1996; Schafer and Graham, 2002; Meng, 1994; Robins and Wang, 2000; Nielsen, 2003). Thus, to generate MIs, several imputation methods could be adopted.

The analysis of a multiply imputed data set is straightforward following Rubin's (1987) combining rules: Let $\hat{\boldsymbol{\theta}}_m$ be the estimator of scientific interest based on the mth imputed data set ($m = 1, \ldots, M$) and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_m)$ be the variance estimator of $\hat{\boldsymbol{\theta}}_m$, then the final estimator and its variance estimator are given by

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \sum_m \hat{\boldsymbol{\theta}}_m / M \quad \text{and} \\ \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) &= \sum_m \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_m)/M + \frac{(1+M^{-1})}{M-1} \sum_m (\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}})', \end{aligned} \tag{1}$$

where $\sum_m (\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}})'/(M-1)$ is the 'between' variance reflecting the amount of uncertainty in the estimator due to nonresponse. Generally, inference can then proceed like in situations without missing values.

3 Methods to Generate Imputations

Throughout we will assume that the model of scientific interest is a homoscedastic linear regression model

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{2}$$

where $\mathbf{x}$ is a vector of predictor variables $x_1, x_2, \ldots$ including the constant term. The first predictor variable, x_1, is not completely observed but missing data are MAR. Note that if only y were affected by missing values being MAR, then, as can be shown, imputations would not be beneficial, but may lead to invalid inferences if the imputation models are grossly misspecified. On the other hand, we consider this simple model since it is comparatively easy to study effects of misspecified imputation models on the inferences of

scientific interest. Further, MI techniques that do not allow valid inferences in this simple scenario, will very likely fail in more complex situations as well.

If all variables are completely observed, estimation of model (2) is straightforward with standard software. It is important to note that in situations without missing values, the distribution of the predictor variables is usually irrelevant. This is due to the fact that inference is justified and made conditional on the values of the predictors, random or not, realized in the sample. If the values of one or more of the predictors are missing, however, the distribution of these predictors become important if missing values are to be compensated by imputations. Then an imputation regression model for this variable on all other relevant variables has to be estimated and the careful specification of this model becomes relevant, which may not be straightforward. For example, if the regression model for the dependent variable given all predictors is a homoscedastic linear regression model and assuming that the reverse regression imputation model of the predictors to be imputed given the dependent variable and the completely observed predictors is linear, is equivalent to assuming that the dependent variable and the predictors to be imputed are multivariate normally distributed conditional on the completely observed predictors (Spanos, 1995). Conditional normality of the dependent variable in a homoscedastic linear model with incompletely observed metric predictors alone is not sufficient to justify a linear imputation model. Thus, linear imputation models would not in general be compatible with the true data generating process (DGP). Although it has been proposed to transform variables to make the assumption of multivariate normality more plausible (e.g. Honaker, King and Blackwell, 2011; Schafer, 1997), this technique does not work in general, because the distribution of variables in the observed part of the data set may be very different from the distribution of the same variables if there were no missing values.

However, according to Little and Rubin (2002), as long as the imputation model differs only slightly from an imputation model that would be compatible with the model of scientific interest, MI based inference tends to be conservative i.e., variances of estimators

tend to be overestimated such that the actual coverage of confidence intervals tend to be larger than the nominal coverage ('self correcting property' of MI). On the other hand, if the imputation model is grossly misspecified, then estimators of regression parameters may be biased and variances may be underestimated, leading to confidence intervals which are systematically too short and to actual rejection rates of true null hypotheses being too large.

3.1 Bayesian Normal Linear Regression

Imputation by a parametric Bayesian homoscedastic linear regression model is described in Rubin (1987; see Example 5.1, p. 166–167). Let $\mathbf{w}$ be a vector containing a constant term and predictor variables including y_i for the incompletely observed x_1. Note that by, e.g. including interaction terms, the number of elements in $\mathbf{w}$ can be different from the number of variables entering the model of scientific interest (2). For all units it is assumed that

$$x_1 = \mathbf{w}'\boldsymbol{\vartheta} + v \quad \text{where} \quad v \sim N(0, \sigma_v^2). \tag{3}$$

The unknown parameters of this model $(\boldsymbol{\vartheta}', \sigma_v^2)'$ are estimated based on the subsample for which $\mathbf{w}$ and x_1 are both observed. Then, assuming an appropriate prior distribution, the posterior distribution of $(\boldsymbol{\vartheta}', \sigma_v^2)'$ is derived. Random draws $\boldsymbol{\vartheta}^\star$ and $\sigma_v^{2\star}$ from this posterior distribution are then used to generate imputations $\widehat{x}_{\text{mis}}$ for those x_1 whose values have not been observed,

$$\widehat{x}_{\text{mis}} = \mathbf{w}'\boldsymbol{\vartheta}^\star + z\sigma_v^\star \,, \tag{4}$$

where z are draws from a standard normal distribution. To generate M imputations, the steps of drawing $\boldsymbol{\vartheta}^\star$ and $\sigma_v^{2\star}$ from their posterior distribution and generating predictions $\widehat{x}_{\text{mis}}$ are repeated M times.

This imputation model is justified by Rubin's work (e.g. Rubin, 1987; Schenker and Welsh, 1988; but see also Robins and Wang, 2000) and may be expected to allow valid inferences not only if the variables to be imputed are linearly dependent on all other

variables involved but to a certain extent, due to the 'self-correcting' property of MI (Little and Rubin, 2002), even in more general situations, like non-linear or non-normal models.

3.2 AMELIA

Honaker, King and Blackwell (2011; see also Honaker and King, 2010) propose an imputation method called 'Amelia II' which is based on a mixed approach. In a first step, a bootstrap sample from the data set with missing values is selected which replaces draws of parameters from a posterior distribution, and in a second step the EM-Algorithm (Dempster, Laird and Rubin, 1977) is adopted to estimate the parameters necessary to generate imputations by predicting the unobserved values using the data set with missing values. The EM algorithm successively calculates the expectation of the log-likelihood function with respect to the variables with missing values conditional on those with observed values and then maximizes this function with respect to the unknown parameters. This algorithm is usually fast and stable in linear regression problems. To generate M imputations, the bootstrap, EM- and prediction steps are repeated M times. The underlying assumption is that all variables are multivariate normally distributed.

3.3 Imputation based on Generalized Additive Models for Location, Scale and Shape

A robust imputation technique, denoted as gamlss, has been proposed by de Jong, van Buuren and Spiess (2014). It is based on a class of generalized additive models for location, scale and shape (GAMLSS) proposed by Rigby and Stasinopoulos (2005) and allows flexible modeling of the location (e.g. the mean), the scale (e.g. variance), and the shape (e.g., skewness, kurtosis) of the distribution of the dependent variable, which in the case considered here is the incompletely observed predictor variable, given all other variables.

The sub-model considered in in this paper is

$$g_1(\mu_{x_1}) = \alpha_1 + \sum_{j=1}^{J_1} h_{1j}(v_j) \quad \text{and} \quad g_2(\sigma_{x_1}) = \alpha_2 + \sum_{j=1}^{J_2} h_{2j}(v_j), \tag{5}$$

where $g_p(\cdot)$, $p = 1, 2$, are monotone link functions that relate the parameters of the conditional distribution, i.e. the mean μ_{x_1} and the standard deviation σ_{x_1}, to predictor variables v_j, $h_{pj}(\cdot)$ represent the effects of v_j, and α_p are constant terms. Note that this formulation allows for heteroscedasticity. To generate imputations, we allow the predictors to have non-linear effects. Thus, the unknown functions h_{pj} of v_j are smoothing terms.

The R implementation of the imputation method for our simulations uses the GAMLSS package (Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007) in R to fit model (5) based on (penalized) maximum likelihood estimation and adopting the default link functions. Rigby and Stasinopoulos (2005) and Stasinopoulos and Rigby (2007) provide a description of the algorithms used by this package. The functions h_{pj} are chosen to be smoothing terms, more specifically, we use penalized B-splines with 20 knots, a piecewise polynomial of second degree, a second order penalty and automatic selection of the smoothing parameter using the Local Maximum Likelihood criterion (for a discussion, see Eilers and Marx, 1996). For high amounts of smoothing, the fit of this smoother approaches linearity. The conditional distribution of the dependent variable given all predictors is assumed to be the normal distribution.

Imputations are generated as follows (c.f. de Jong, van Buuren and Spiess, 2014). Firstly, fit the imputation model based on the completely observed part of the data set. Secondly, replace the observed values of the variable to be imputed by predictions from the fitted model in step one ('bootstrap sample'). Thirdly, refit the model based on this bootstrap sample and predict the unobserved values. Repeating steps 2 and 3 M times generates M imputations for each missing value.

It should be noted that the approach proposed by Rigby and Stasinopoulos (2005) is much more general. For example, the dependent variables may be discrete, truncated, or count data with corresponding link functions. Models may include linear and interaction

terms, polynomials, random effects and/or spline terms. The conditional distributions of the dependent variable can be chosen from a large class of distributions.

3.4 Iterative Robust Model-based Imputation

Templ, Kowarik and Filzmoser (2011) propose an algorithm called 'Iterative Robust Model-based Imputation' (IRMI) implemented in the R package VIM (Templ, Alfons, Kowarik and Prantner, 2014). The essence of the method is an imputation procedure like the one described in Section 3.1, but adopting one of several robust estimation methods to reduce the influence of outlying observations on the regression parameter estimates. Instead of drawing parameters from their posterior distribution, however, they are fixed at their posterior mean. The added error term is multiplied by a non-justified factor larger then one to account for the additional uncertainty in the imputations due to the need of estimating the model. Multiple imputations are generated by repeating the above steps M times (for a description, see Templ, Kowarik and Filzmoser, 2011).

The default option for continuous dependent variables in IRMI is an estimator proposed by Yohai (1987) which is efficient in linear regression models with normally distributed errors but at the same time largely ignores outliers. The principal problem of such an automatic method, however, is that it does not differentiate between valid and invalid outliers. Thus, e.g., if the conditional distribution of a variable to be imputed is skewed, valid values which are in a sparsely populated region may be ignored when fitting the imputation model. This would lead to estimating the imputation model using systematically selective samples and thus to adopting an improper imputation method.

A small simulation study presented by Templ, Kowarik and Filzmoser (2011) is intended to show the good properties of the technique. However, coverage rates of the true values in this limited study range between 0.882 and 0.906 given $\alpha = 0.05$. In fact, this imputation method seems not to be proper. In an additional study, imputation techniques are evaluated based on comparisons of true but unobserved and imputed values. With

respect to these error measures, the technique proposed by Templ, Kowarik and Filzmoser (2011) performs better than an imputation method following the procedure described in Section 3.1. However, for an imputation method to be proper it is neither required nor implied that some measure of distances between true and imputed values is minimal (see Rubin, 1987, 1996, 2003).

3.5 k-Nearest-Neighbor Imputation

An alternative to fully parametric methods are k-nearest-neighbor (kNN) techniques. The idea is to find, for each case with missing x_1 but observed (or already imputed) $\mathbf{w}$, say, k completely observed neighbors that are somehow close with respect to $\mathbf{w}$ to the case with a missing value. From this pool of neighbors, one donor is usually randomly selected and its value x_1 is taken as an imputation for the case with a missing value.

The main advantages of kNN imputation is that it is simple, that it seems to avoid strong parametric assumptions, that it can easily be applied to various types of variables to be imputed and that only eligible and observed values are imputed (e.g., Andridge and Little, 2010; Little, 1988; Schenker and Taylor, 1996).

Closeness is usually expressed as a distance measure, one popular being based on the (estimated) conditional mean of $(x_1|\mathbf{w})$,

$$d_{i,i'}^{PMM} = |\hat{E}(x_{i1,\text{mis}}|\mathbf{w}_i) - \hat{E}(x_{i'1,\text{obs}}|\mathbf{w}_{i'})|,$$

where $x_{i1,\text{mis}}$ denotes variable x_1 of unit i whose value has not been observed, and $x_{i'1,\text{obs}}$ denotes variable x_1 of unit i' whose value has been observed ($i, i' = 1, \ldots, n$). The imputation technique is also called 'predictive mean matching' (PMM) imputation.

In many cases, the predictive mean is estimated using a linear regression of $x_{1,obs}$ on $\mathbf{w}$. The idea underlying this PMM version goes back to Rubin (1986; see also Rubin, 1987, Metric-Matching Hot-deck Method, p. 158 and Example 5.2, p. 168) and Little (1988) who coined the name. Let coefficient $\hat{\boldsymbol{\vartheta}}$ be the estimated regression coefficient of a linear

model, then the measure function simplifies to

$$d_{i,i'}^{PMM} = |(\mathbf{w}_i - \mathbf{w}_{i'})'\hat{\boldsymbol{\vartheta}}|. \tag{6}$$

However, highly nonlinear effects of elements from $\mathbf{w}$ on x_1 may not appropriately be incorporated, as their weighted difference can be zero simply because the corresponding elements in $\hat{\boldsymbol{\vartheta}}$ are zero. It can also be zero due to weighted differences summing up to zero.

To generate M imputations for a Bayesian version, one would repeat the following steps M times: Estimate the imputation model based on the completely observed cases. Randomly select parameters from their posterior distribution, use these parameter values to calculate the distance measure, search for neighbors and randomly select for each $x_{1,\text{mis}}$ one value $x_{1,\text{obs}}$ from the corresponding set of neighbors. For a Bootstrap version one would replace drawing parameters from their posterior distribution by selecting a Bootstrap sample from the completely observed cases, estimate the conditional mean model and calculate the distances using these parameter estimates.

Finally, selection of the imputations from the pool of possible donors at each of the M steps can proceed either unweighted or weighted, in which case close neighbors usually receive a higher probability of being selected as compared to more remote units.

Two notes are worth mentioning. First, by using observed x_1 values from some donors as imputations, it is implicitly assumed that they are random independent draws from an approximate posterior distribution of $x_{1,\text{mis}}$ given $\mathbf{w}_{\text{mis}}$, the vector of predictors for this unobserved x_1. Thus, the assumption is that the probability of observing x_1 given $\mathbf{w}_{\text{mis}}$ is independent from differences between $\mathbf{w}_{\text{mis}}$ and $\mathbf{w}_{\text{obs}}$, the values of $\mathbf{w}$ of completely observed neighbors. This is equivalent to assuming that the missing data are MCAR within the cells implicitly defined by the k neighbors. Strictly speaking, the assumption is that the missing data are neither MCAR nor MAR, but missing locally completely at random (MLCAR).

Second, a special case of kNN imputation is $k = 1$, i.e. the closest neighbor is the donor.

In this case there is no random selection of the values to be imputed and even appropriately taking into account the uncertainty in the parameter estimates of the imputation model does not make this method proper.

A slightly different distance measure was proposed by van Buuren and Groothuis-Oudshoorn (2011)

$$d_{i,i'}^{MICE} = |\mathbf{w}_i \dot{\boldsymbol{\vartheta}} - \mathbf{w}_{i'} \tilde{\boldsymbol{\vartheta}}|, \tag{7}$$

where $\tilde{\boldsymbol{\vartheta}}$ is the posterior mean of the parameters of the reverse regression model, and $\dot{\boldsymbol{\vartheta}}$ is a draw from the corresponding posterior distribution (for a description, see Vink, Frank, Pannekoek and van Buuren, 2014).

Although simulation results imply that PMM versions of kNN imputation seem to work well (e.g., Andridge and Little, 2010; Yu, Burton and Rivero-Arias, 2007; Vink et al., 2014), it is not clear if kNN imputation techniques are proper imputation methods. In fact, Schenker and Taylor (1996) state that if the number of possible donors is too small, the M imputations will be correlated leading to a higher variance of the estimator of interest. On the other hand, increasing the number of neighbors of a case to be imputed (the query point), may lead to biased estimators due to a violation of the MLCAR assumption. In a simulation study using fixed (three and ten) possible donors they found a slight undercoverage of the interesting parameter of two to three percent. The missing data in their study are MCAR. Similar results are reported in a simulation study of de Jong, van Buuren and Spiess (2014) with missing data being MAR, who found no (obvious) bias but mild to moderate undercoverage using the kNN imputation method with $k = 3$.

Dahl (2007) shows that under some mild conditions and using a distance measure which is topologically equivalent to Euclidian distance, imputations based on kNN techniques can be interpreted as draws from the conditional distribution of the incompletely observed variable given observed values, with decreasing correlations if $n \to \infty$, $k := k(n) = n^r$ and $r \in (0, 1)$. Dahl (2007) proposes $k(n) = \sqrt{n}$ as this is 'canonical in the sense of representing the mid-point of the interval' defined by $r \in (0, 1)$ (Dahl, 2007, p.

5915).

However, convergence rates to the true distribution may vary at different query points, depending on whether regions are thinned out by the response mechanism or not, which is not the case if the missing data are MCAR, as in the simulation study of Schenker and Taylor (1996).

A distance measure proposed by Dahl (2007) is

$$d^D_{i,i'} = |(\mathbf{w}_i - \mathbf{w}_{i'})'||\hat{\boldsymbol{\vartheta}}|. \tag{8}$$

Obviously, $d^D_{i,i'}$ is zero if $(\mathbf{w}_i - \mathbf{w}_{i'})' = \mathbf{0}$ or when $\hat{\boldsymbol{\vartheta}}$ is zero. Differences in variables with strong predictive effects are given higher weight than differences in variables with small effects. Compensatory effects or differences are not taken into account.

A non-parametric version of kNN imputation provided by function aregImpute as part of the R package HMISC (Harrell, 2015) is based on first drawing a bootstrap sample from the completely observed part of the sample and then fitting a flexible additive model while finding the transformation of the variable with missing values and of all functions of the predictors, $f_j(\mathbf{w}_j)$ $(j = 1, 2, \ldots)$, that maximizes the coefficient of determination. For a description, see Harrell (2015) and the literature cited therein. At least two ways to proceed are possible: Either draw another bootstrap sample from the predictions for x_1 in the completely observed part of the sample, adopt a distance measure, calculate distances in the predicted values of this Bootstrap sample and the predicted values for the not observed values to draw a donor (denoted as aI-PMM in Section 5). Or just use the predictions from the first Bootstrap step and the predictions for the not observed values to generate imputations (denoted as aI-boot in Section 5). The distance measure adopted is similar to (6) with w_i and $w_{i'}$ replaced by their corresponding functions. Unlike in the above described kNN techniques, donors are now randomly drawn from multinomial distributions with probabilities inversely proportional to the distances. Repeating these steps M times generates M imputations.

A difference between aregImpute and the GAMLSS imputation method described in

Section 3.3 is that the former fixes the number of knots of the transformation functions to a default fixed value, while GAMLSS optimizes the smoothing parameter of h_{ij} using cross-validation. The latter strategy may be superior as the performance of a smoother is extremely sensitive to the appropriateness of the chosen smoothing parameter.

Most standard analysis software packages or functions offer one of these or a similar kNN technique, often with a default value for k, like $k = 5$ (e.g., SAS Institute Inc., 2013, p. 5074; or MICE, Vink, Frank, Pannekoek and van Buuren, 2014).

4 Simulation Experiment: Description

The multiple imputation techniques described in Section 3 differ with respect to the generality of situations for which they are proposed. However, they are all multiple imputation techniques taking advantage of Rubin's (1987) work. They all assume missing data to be MCAR or MAR, most are implemented in available software and they are all – the exception being gamlss – easily available from within the software R. Although the persuasiveness of the justifications of MI techniques varies, they all share the property that no proof exists showing that inferences based on the multiply imputed data sets and the combining rules (1) are valid in all situations of potential interest.

Due to the lack of theoretical results, the properties of scientifically interesting estimators based on multiply imputed data sets can systematically be studied only in simulation experiments.

In our simulation experiment, we consider the situation of incompletely observed predictor variables in a regression model of scientific interest with misspecified imputation models. Validity of subsequent analyses actually means (approximate) unbiasedness or consistency of point and variance estimators as well as that the actual coverage of true values by confidence intervals is close to the nominal coverage. A secondary criterion is the relative efficiency or precision of (approximately) unbiased estimators, which amounts to comparing their variances.

Throughout, we consider a multiple linear regression problem with missing values in one predictor for simplicity. The 'true' DGP realizes a linear regression model with normally distributed homoscedastic errors. This is the usual regression model adopted in applications, although the normality assumption is not necessary in larger samples. All simulations were run within R (R Core Team, 2015).

Each simulation consists of several steps. In a first step, data are generated according to the linear regression model

$$y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + x_{i,3}\beta_3 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \text{ for all } i = 1, \ldots, n \,, \tag{9}$$

with sample sizes $n = 40$, $n = 100$ or $n = 1000$. The true values of the parameters weighting the predictors are $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 = \beta_3 = \sqrt{2/3}$. The error variance σ^2 was chosen so that the coefficient of determination, R^2, was between 0.3 and 0.7.

An important condition with respect to the imputation model is the distribution of the predictor variable with missing values. The only variable with missing values was x_1. Thus, values for this variable were generated either from a normal distribution, a χ^2- or a t-distribution, in the two latter cases with three degrees of freedom. Note that the true imputation regression model of x_1 on y and the other predictors can not be a linear model in the latter two cases (Spanos, 1995). Hence, a corresponding normal linear imputation model is a misspecified model. All other predictor variables were generated mutually independent according to normal distributions in all but the case with x_1 following a heavy-tailed t-distribution: In that experiment covariates were generated according to a multivariate t-distribution.

As the golden standard, a homoscedastic linear regression model was estimated based on the data set without missing values. In the results section, the corresponding condition will be denoted as COM.

In the second step, missing values were generated. Throughout the missing values are MAR and were generated according to one of two discrete or two continuous missing data mechanisms (MDM). More precisely, the probability of observing a value of x_1,

$\mathrm{P}(r = 1|\mathbf{v})$, r being the response indicator with $r = 1$ if x_1 is observed and $r = 0$ otherwise, is a function of $\kappa = \mathbf{v}'\boldsymbol{\gamma}$, where $\mathbf{v} = (x_2\ x_3\ y\ yx_3)'$ and $\boldsymbol{\gamma}$ reflects the influence of x_2, x_3, y and the interaction of y and x_3 on the probability of observing x_1. Note that, conditional on y, x_1 had no effect on the response probability, so the missing values are MAR if y is included in the imputation models. A set of thresholds (discrete MDM) or a constant (continuous MDM) was chosen such that the fraction of observed values was approximately 0.65. The categorical functions realized three and four response classes respectively, simulating the areas with lowest response probabilities either on one side or in the center of the range of x_1. Figure 1 shows the two categorical mechanisms.

In case of the discrete missing data mechanisms, $\boldsymbol{\gamma} = (-.4\ -.4\ .5\ .1)'$. The continuous response functions were smooth versions of their categorical counterparts. Since the general results are virtually the same for the categorical and the continuous missing data mechanisms, results based on the latter are omitted to save space.

As a standard ad hoc technique, complete case analysis is adopted, i.e., only the completely observed cases are used to estimate a linear regression model. Results under this strategy will be denoted as CCA.

Subsequent to generating the missing values in each simulation and conducting CCA, as a third step, the different imputation techniques described in Section 3 are adopted to generate multiply ($M = 10$) completed data sets. In the last step, the multiply imputed data sets were used to estimate model (9).

Multiple imputations generated according to the Bayesian linear model as described in Section 3.1 were generated via the package *MICE* (version 2.22) (van Buuren and Groothuis-Oudshoorn, 2011) available from within R. The corresponding results will be denoted as LM.

Results based on generating imputations with AMELIA (version 1.7.3) (Honaker, et al., 2011; see Section 3.2) are denoted as amelia, and results based on the robust imputation function IRMI included in package VIM (version 4.3.0) (Templ et al., 2014; see

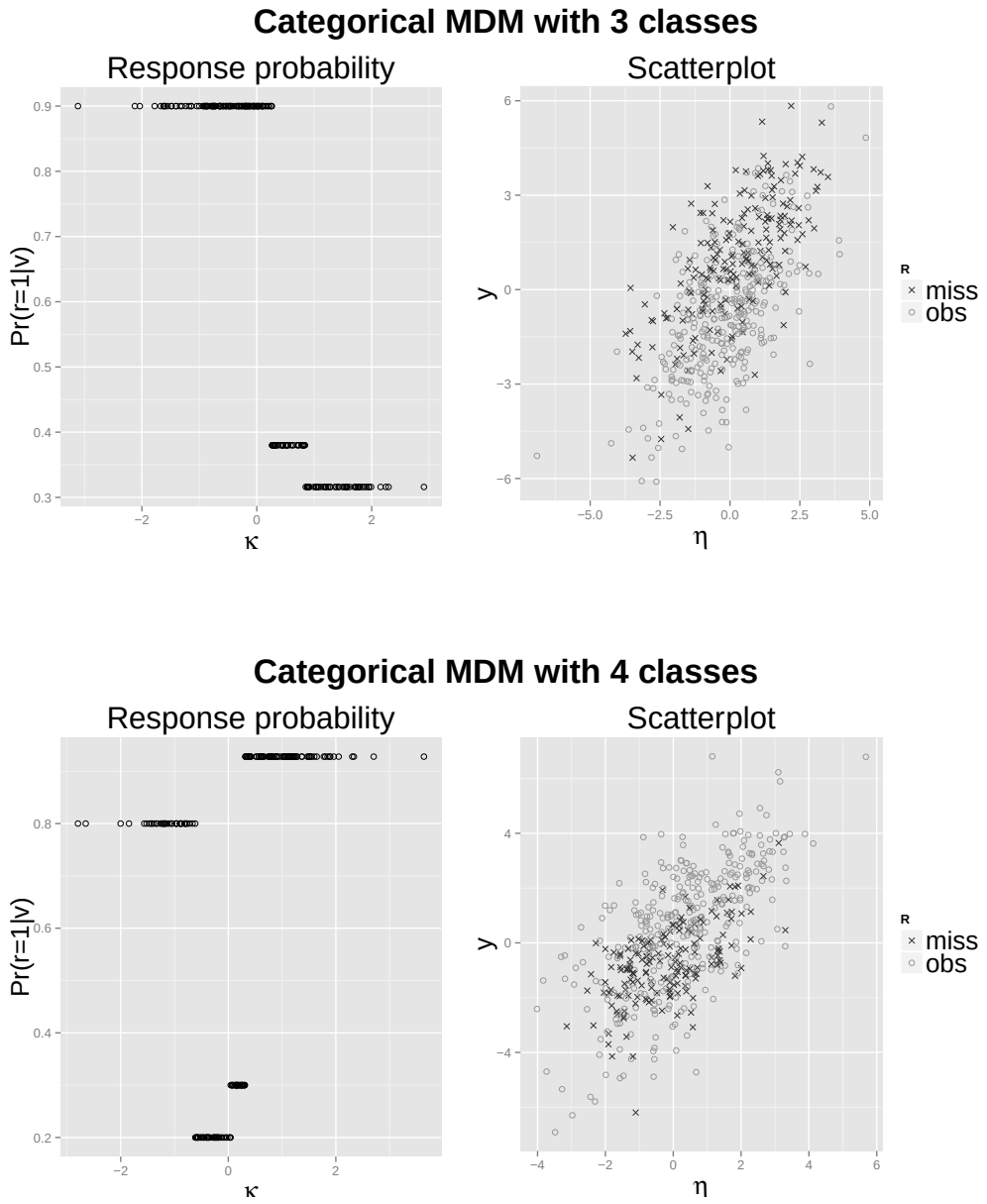

Figure 1: Response probability as a function of $\kappa = \mathbf{v}'\boldsymbol{\gamma}$ (left) and scatterplot of y against $\eta = \mathbf{x}'\boldsymbol{\beta}$ with observed ($\bullet$) and missing ($\times$) values of x_1 (right), n=200. The upper mechanism consists of three classes that generate most missing values above the median value of x_1; the bottom mechanism, with four classes, systematically generates missing values in a central region within the range of x_1.

Section 3.4) are denoted as irmi. Imputation by generalized additive models (Section 3.3) is currently not available in any R library, but was implemented for the current simulation study. Results are denoted by gamlss.

In Section 3.5 different kNN techniques are described, based on different distance measures. Results based on $d_{i,i'}^{PMM}$ are denoted as dpmm, results based on $d_{i,i'}^{D}$ are denoted as dD and those based on the distance measure as implemented in *MICE* (van Buuren and Groothuis-Oudshoorn, 2011) are denoted as dmice. The former two techniques were implemented for this simulation experiment in R, imputations according to the latter technique were generated using *MICE*. For each of the different kNN versions, different sizes of the donor pools were adopted, namely $k = 1, 3, 5, 10, 20$ and $k = \sqrt{n}$.

In addition, the two versions of a more robust kNN technique available in function aregImpute (version 3.16-0), as part of the R-package HMISC (Harrell, 2015) and described in Section 3.5 were tested as well. The Bayesian version will be denoted as aI-PMM, the bootstrap version as aI-boot. Both versions worked only with the default $k = 3$. Every attempt to set a different number for k resulted in a warning and termination of the program.

The evaluation of the imputation methods is based on 1000 simulated data sets under each condition. It should be noted that not all imputation techniques combined with the estimation of the model converged. The maximum number of failures was 15, i.e., 1.5% of the simulations in case of a t-distributed variable with missing values, $n = 1000$ adopting gamlss. In general, gamlss failed most often (six or seven times under most of the conditions), but it was not the only technique: Several kNN techniques also resulted in termination of the program under some conditions, usually in two out of 1000 simulations.

To save space, we only report the results with respect to β_1 weighting the predictor with missing values. The results for each condition considered are the mean of the estimates for the regression parameters ($m(\hat{\beta})$), the square root of the mean of their estimated variances ($m(\widehat{sd})$), their standard deviation over the simulations (sd) and the actual coverage of the

true values by 0.95 confidence intervals (COV). As each simulation is an independent Bernoulli experiment with respect to the confidence interval covering the true value or not, a symmetric 0.95 interval around the nominal coverage of 0.95 is $[0.936, 0.964]$. Thus the actual coverage should be roughly within this interval if all assumption are met.

5 Results

Since the robustness of imputation techniques is one of the main topics, we simulated one scenario in which all the variables were multivariate normally distributed, and two scenarios in which they were not. In the former case, the parametric Bayesian linear imputation model (LM) and the parametric imputation model underlying amelia are correctly specified models and thus we expect that these two models are the best choice for generating imputations with respect to bias and efficiency. In the non-normal cases, these imputation models are misspecified. Hence one would expect that imputation techniques based on semi- or non-parametric techniques like kNN, aregImpute or gamlss outperform LM and amelia.

5.1 Normal covariates

Table 1 presents the results using the three class MDM (upper panel of Figure 1) for three increasing sample sizes based on the complete data set (COM), the completely observed cases (CCA) and the multiply imputed data sets.

A first result to note from Table 1 is that with but one exception, the estimator for β weighting x_1 seems to be downward biased for $n = 40$, the exception being the estimator under the amelia condition which seems to be unbiased.

As expected, without missing values (condition COM), the results are acceptable with respect to coverage for all sample sizes, and with respect to bias if $n \geq 100$. In addition, estimates for β under this condition have the smallest variance. Since the missing data

	n=40				n=100				n=1000			
Technique	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV
COM	0.987	0.288	0.288	0.954	0.991	0.180	0.176	0.939	1.003	0.056	0.054	0.940
CCA	0.843	0.362	0.348	0.929	0.853	0.224	0.210	0.891	0.863	0.067	0.064	0.449
LM	0.918	0.373	0.372	0.958	0.972	0.224	0.221	0.948	0.997	0.068	0.066	0.942
amelia	1.000	0.392	0.374	0.929	1.010	0.226	0.218	0.941	1.002	0.068	0.066	0.942
dpmm, $k = 1$	0.924	0.406	0.384	0.929	0.966	0.252	0.224	0.917	1.007	0.073	0.064	0.920
dpmm, $k = 3$	0.890	0.393	0.382	0.938	0.950	0.247	0.220	0.925	0.996	0.072	0.062	0.918
dpmm, $k = 5$	0.862	0.384	0.383	0.931	0.930	0.246	0.223	0.922	0.993	0.073	0.062	0.903
dpmm, $k = 10$	0.796	0.358	0.391	0.938	0.898	0.241	0.227	0.924	0.990	0.072	0.062	0.910
dpmm, $k = 20$	0.669	0.314	0.401	0.916	0.837	0.232	0.233	0.904	0.987	0.072	0.062	0.913
dpmm, $k = \sqrt{n}$	0.848	0.378	0.386	0.936	0.899	0.243	0.227	0.924	0.970	0.073	0.063	0.893
dD, $k = 1$	0.913	0.411	0.344	0.901	0.957	0.249	0.197	0.884	0.988	0.077	0.056	0.843
dD, $k = 3$	0.857	0.381	0.370	0.929	0.919	0.237	0.212	0.925	0.972	0.073	0.060	0.881
dD, $k = 5$	0.816	0.368	0.376	0.925	0.892	0.233	0.218	0.925	0.975	0.073	0.062	0.887
dD, $k = 10$	0.739	0.340	0.391	0.933	0.842	0.229	0.227	0.896	0.951	0.072	0.063	0.866
dD, $k = 20$	0.627	0.294	0.402	0.906	0.773	0.219	0.234	0.851	0.931	0.072	0.064	0.828
dD, $k = \sqrt{n}$	0.795	0.360	0.383	0.938	0.840	0.230	0.226	0.902	0.924	0.071	0.065	0.765
dmice, $k = 1$	0.946	0.423	0.376	0.910	0.985	0.251	0.219	0.909	1.004	0.075	0.064	0.912
dmice, $k = 3$	0.922	0.394	0.374	0.927	0.968	0.243	0.219	0.919	0.999	0.073	0.064	0.920
dmice, $k = 5$	0.885	0.389	0.378	0.938	0.948	0.243	0.219	0.922	0.990	0.073	0.064	0.922
dmice, $k = 10$	0.818	0.367	0.386	0.943	0.912	0.243	0.222	0.930	0.997	0.072	0.064	0.919
dmice, $k = 20$	0.682	0.311	0.399	0.919	0.851	0.231	0.229	0.904	0.986	0.072	0.064	0.910
dmice, $k = \sqrt{n}$	0.871	0.387	0.378	0.938	0.912	0.240	0.222	0.931	0.979	0.072	0.064	0.902
aI-PMM	0.857	0.396	0.400	0.936	0.947	0.243	0.226	0.938	0.994	0.074	0.065	0.918
aI-boot	0.819	0.375	0.403	0.943	0.910	0.245	0.234	0.934	0.992	0.073	0.063	0.918
irmi	0.778	0.355	0.411	0.946	0.788	0.212	0.250	0.917	0.808	0.066	0.076	0.267
gamlss	0.852	0.404	0.464	0.941	1.008	0.243	0.244	0.929	1.001	0.070	0.068	0.939

Table 1: Estimation results based on the complete data set (COM), completely observed cases (CCA) and multiply imputed datasets. The fraction of observed x_1 values is 0.65 and the true value of β is 1. Multivariate Normal data and discrete MDM with three classes.

are MAR, complete case analysis (condition CCA) leads to invalid results as expected, i.e., severely biased estimates and unacceptable low coverage rates, which gets even worse with increasing sample size.

On the other hand, imputations based on the parametric Bayesian model (condition LM) or the mixed Bootstrap/EM approach implemented in amelia work well for sample sizes $n \geq 100$. If $n = 40$ coverage is acceptable under condition LM, although the estimates seem to be biased, and under condition amelia, in which case estimates seem to be unbiased, but the coverage is too low due to an underestimation of the variance.

There is a general pattern of results common to all kNN techniques: With an increasing number of potential donors, the downward bias in the estimates seems to increase while at the same time variances are getting smaller and the estimated variances tend to be larger. In all cases if $n \geq 100$, coverage rates are too low. As a consequence, a true null hypothesis would be rejected far too often. If $n = 40$, coverage rates are in an acceptable range in a few cases, but there is a general tendency of confidence intervals being too short. With just a few exceptions, the estimates are downward biased. This bias decreases with increasing sample size and decreasing number of potential donors.

Comparing the results under conditions dpmm, dD and dmice one can conclude, that they are rather similar, with a small advantage of dpmm and dmice over dD with respect to bias of the estimates and coverage rates. Results based on aI-PMM and aI-boot are within the range of results of the other kNN methods. Note that for both kNN methods, the number of neighbors is fixed at $k = 3$.

Finally, estimates under condition irmi are severely biased for all sample sizes while variances are overestimated. However, overestimation masks the bias only in the small sample case, leading to an acceptable coverage rate. For $n \geq 100$ the coverage rate is too small and for $n = 1000$ it is unacceptable low.

On the other hand, results under condition gamlss are mixed. If $n = 40$, the estimator for β is biased but this seems to be masked by an overestimation of its variance so that

the coverage is acceptable. If $n = 100$, the bias vanished, but coverage is too low. If $n = 1000$, the results are acceptable.

In general, it seems not necessarily to be better to adopt some imputation method than to just ignore the missing data. In fact, choosing an inappropriate imputation technique may lead to estimators with larger bias and coverage rates which are too low.

Table 2 presents results based on the four class MDM (lower panel of Figure 1). This mechanism generates most missing values in the center of the range of x_1. Again all variables are multivariate normally distributed and thus results under conditions LM and amelia are expected to be acceptable. The different kNN techniques should lead to results which are closer to being acceptable, as in contrast to the three class MDM now the distributions of potential donors tend to be closer to being symmetric around the true value to be estimated.

The results in Table 2 show a similar pattern as in Table 1, although the estimates seem not to be systematically downward biased.

Again, under condition COM the results are acceptable with respect to coverage and now also with respect to bias for all sample sizes, and the estimates for β have the smallest variance. CCA leads to severely biased estimates and coverage rates which, although acceptable for $n = 40$, are too low if $n = 100$ and disastrous if $n = 1000$. This seems not to be due to an underestimation of the variance but rather to biased estimates.

On the other hand, imputations under conditions LM or amelia work well, although estimates are slightly biased in small samples. This bias vanishes with increasing n, however, and the coverage under condition LM is too low for $n = 100$ and for $n = 40$ under condition amelia. The slight undercoverage seems to be due to a combined effect of a small bias and a slight underestimation of variances.

According to the kNN techniques, a similar pattern as observed in Table 1 emerges, although now estimates are upward biased for small k and downward biased for large k. Hence there seems to exist an optimal number of possible donors, which is, however, not

	n=40				n=100				n=1000			
Technique	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV
COM	0.999	0.302	0.288	0.949	0.990	0.180	0.176	0.944	1.000	0.057	0.056	0.941
CCA	1.164	0.394	0.392	0.940	1.152	0.230	0.233	0.917	1.158	0.069	0.071	0.405
LM	0.976	0.343	0.339	0.942	0.987	0.209	0.203	0.936	0.999	0.063	0.063	0.939
amelia	1.073	0.366	0.334	0.922	1.025	0.214	0.202	0.941	1.004	0.063	0.062	0.942
dpmm, $k = 1$	1.014	0.370	0.338	0.916	1.000	0.217	0.204	0.931	1.002	0.067	0.063	0.935
dpmm, $k = 3$	1.003	0.361	0.336	0.915	0.996	0.215	0.201	0.930	1.002	0.067	0.061	0.924
dpmm, $k = 5$	0.988	0.357	0.338	0.931	0.994	0.214	0.200	0.921	1.001	0.067	0.061	0.928
dpmm, $k = 10$	0.950	0.343	0.342	0.940	0.985	0.214	0.201	0.928	1.000	0.067	0.061	0.922
dpmm, $k = 20$	0.854	0.304	0.350	0.947	0.958	0.208	0.204	0.936	0.999	0.067	0.060	0.926
dpmm, $k = \sqrt{n}$	0.984	0.353	0.337	0.930	0.986	0.213	0.201	0.930	0.998	0.067	0.060	0.929
dD, $k = 1$	1.014	0.369	0.315	0.895	1.000	0.218	0.186	0.905	1.000	0.071	0.055	0.877
dD, $k = 3$	0.983	0.350	0.328	0.926	0.985	0.211	0.194	0.923	0.996	0.067	0.058	0.916
dD, $k = 5$	0.957	0.341	0.334	0.936	0.974	0.208	0.196	0.933	0.994	0.066	0.059	0.921
dD, $k = 10$	0.904	0.320	0.342	0.950	0.951	0.205	0.199	0.930	0.989	0.065	0.060	0.927
dD, $k = 20$	0.818	0.287	0.351	0.952	0.913	0.197	0.204	0.927	0.982	0.064	0.060	0.927
dD, $k = \sqrt{n}$	0.949	0.340	0.335	0.941	0.952	0.205	0.199	0.929	0.975	0.064	0.060	0.922
dmice, $k = 1$	1.034	0.369	0.331	0.917	1.008	0.217	0.197	0.922	1.002	0.070	0.060	0.919
dmice, $k = 3$	1.019	0.367	0.333	0.923	1.007	0.215	0.198	0.928	1.002	0.069	0.060	0.923
dmice, $k = 5$	1.009	0.365	0.334	0.925	1.004	0.212	0.198	0.933	1.002	0.068	0.060	0.920
dmice, $k = 10$	0.967	0.347	0.339	0.932	0.994	0.212	0.200	0.930	1.001	0.068	0.060	0.922
dmice, $k = 20$	0.860	0.309	0.350	0.952	0.967	0.208	0.202	0.929	1.000	0.068	0.060	0.924
dmice, $k = \sqrt{n}$	0.997	0.357	0.335	0.928	0.993	0.213	0.199	0.930	0.999	0.067	0.060	0.924
aI-PMM	0.980	0.359	0.340	0.930	0.991	0.214	0.200	0.929	1.000	0.068	0.060	0.922
aI-boot	0.959	0.348	0.346	0.941	0.973	0.214	0.205	0.935	0.999	0.067	0.062	0.932
irmi	1.026	0.361	0.364	0.947	1.021	0.209	0.219	0.954	1.035	0.065	0.067	0.925
gamlss	0.942	0.369	0.440	0.949	1.034	0.220	0.216	0.939	1.007	0.064	0.063	0.941

Table 2: Estimation results based on the complete data set (COM), completely observed cases (CCA) and multiply imputed datasets. The fraction of observed x_1 values is 0.65 and the true value of β is 1. Multivariate Normal data and discrete MDM with four classes.

$\sqrt{n}$ as proposed by Dahl (2007). If $n = 40$, coverage rates are in an acceptable range for large k, which seems to be due to an overestimation of the variance, such that biases are masked. In general however, variances seem to be underestimated leading to intervals which are too short in large samples.

Estimates under condition irmi are slightly biased for all sample sizes which seems to be masked by a small upward bias of the variance estimates if $n \leq 100$. As a consequence, coverage rates are acceptable for $n \leq 100$, but not for $n = 1000$. Under condition gamlss estimates for β are slightly biased which seems to be masked by an overestimation of its variance so that the coverage is acceptable for $n = 40$. If $n \geq 100$, the bias diminishes and the coverage is acceptable.

The results of this section imply that, not surprisingly, complete case analysis is invalid in general if missing values in a predictor variable are MAR. But the results also show that kNN methods can not be recommended either as they tend to lead to undercoverage and thus falsely rejecting true null hypotheses too often. Surprisingly, the results with respect to coverage rates are less pronounced in small samples. In these cases biases of the estimates for β tend to be masked by overestimated variances, which corresponds to the 'self correcting' property of MI (Little and Rubin, 2002).

Fully parametric imputation techniques like those based on LM or amelia seem to be acceptable for the situations simulated in this section. This, however, is expected as these methods follow Rubin's (1987) theory and all models are correctly specified. In small samples they tend to lead to biased estimates, the bias being masked by an overestimation of variances.

The robust methods (conditions irmi and gamlss) behave rather differently. Results under condition irmi can be acceptable or disastrous with respect to bias and coverage in large samples. Under condition gamlss coverage rates are acceptable in most cases, but the estimates for β tend to be biased in small samples.

Thus, CCA analyses, all kNN techniques and the technique implemented in irmi seem

to have structural problems which are partly masked in small samples but become obvious in larger samples. On the other hand, LM, amelia or gamlss also seem to have problems in small samples, which tend to be masked by overestimated variances, and vanish if the sample size increases. Hence these problems seem not to be structural but rather small sample problems.

Adopting the continuous MDMs, the general pattern of results did not change. Hence these results are omitted. Since the kNN technique using Dahl's (2007) distance measure was found to be inferior to the other kNN techniques, we omit these results in the next section.

5.2 Non-normal covariates

As a first example of non-normal distributed predictor variables, x_1 was generated from a χ^2 distribution with three degrees of freedom. Missing data were generated using the three-class MDM. Results are presented in Table 3.

As expected, analysis based on the data set without missing values leads to valid inferences, whereas CCA leads to biased estimates and coverage rates which are too low. However, the results in Table 3 also show that with only a few exceptions all imputation techniques fail to allow valid inferences. Notably, LM and amelia lead to unacceptable coverage rates getting worse with increasing sample size. In this case, there is obviously no 'self-correcting' property at work.

In contrast, kNN techniques seem to work better. Although in the $n = 1000$ case coverage rates are too small in general due to underestimated variances, in case of $n = 40$ and $n = 100$ kNN techniques with large k may lead to acceptable coverages. If $n = 40$, this is mainly due to overestimated variances, whereas for $n = 100$ results are acceptable with respect to bias and coverage if $k = 10$, for both techniques dpmm and dmice. Imputation technique irmi leads to invalid inferences with respect to coverage for $n = 100$ and completely breaks down for $n = 1000$. For all samples sizes, the estimates for β are

	n=40				n=100				n=1000			
Technique	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV
COM	0.995	0.123	0.122	0.958	1.001	0.072	0.072	0.949	0.999	0.022	0.022	0.950
CCA	0.947	0.220	0.188	0.928	0.965	0.118	0.108	0.931	0.974	0.035	0.032	0.866
LM	1.090	0.219	0.195	0.908	1.105	0.124	0.102	0.819	1.097	0.038	0.029	0.139
amelia	1.140	0.231	0.199	0.859	1.127	0.126	0.105	0.782	1.100	0.038	0.031	0.160
dpmm, $k = 1$	1.065	0.261	0.199	0.849	1.052	0.130	0.097	0.836	1.009	0.037	0.027	0.853
dpmm, $k = 3$	1.034	0.251	0.226	0.921	1.050	0.129	0.106	0.871	1.012	0.036	0.027	0.839
dpmm, $k = 5$	0.994	0.241	0.245	0.949	1.040	0.128	0.114	0.900	1.016	0.036	0.027	0.831
dpmm, $k = 10$	0.907	0.229	0.273	0.958	1.004	0.129	0.131	0.944	1.021	0.036	0.028	0.833
dpmm, $k = 20$	0.750	0.201	0.297	0.938	0.929	0.121	0.151	0.961	1.027	0.037	0.030	0.820
dpmm, $k = \sqrt{n}$	0.976	0.242	0.250	0.954	1.004	0.126	0.131	0.946	1.027	0.037	0.031	0.827
dmice, $k = 1$	1.077	0.258	0.190	0.862	1.054	0.134	0.097	0.843	1.009	0.037	0.028	0.854
dmice, $k = 3$	1.046	0.254	0.216	0.911	1.051	0.131	0.106	0.873	1.013	0.037	0.029	0.860
dmice, $k = 5$	1.011	0.248	0.235	0.936	1.043	0.131	0.113	0.902	1.016	0.036	0.029	0.857
dmice, $k = 10$	0.919	0.234	0.268	0.963	1.007	0.131	0.129	0.937	1.021	0.037	0.030	0.849
dmice, $k = 20$	0.754	0.199	0.298	0.941	0.935	0.125	0.150	0.960	1.027	0.037	0.031	0.821
dmice, $k = \sqrt{n}$	0.992	0.242	0.245	0.951	1.008	0.129	0.130	0.942	1.028	0.037	0.032	0.842
aI-PMM	1.014	0.264	0.242	0.932	1.044	0.132	0.110	0.882	1.007	0.037	0.030	0.889
aI-boot	0.987	0.249	0.252	0.948	1.035	0.129	0.112	0.900	1.006	0.037	0.027	0.853
irmi	0.752	0.233	0.318	0.944	0.762	0.127	0.182	0.839	0.714	0.041	0.054	0.000
gamlss	0.878	0.302	0.371	0.950	0.994	0.172	0.180	0.927	1.008	0.040	0.040	0.946

Table 3: Estimation results based on the complete data set (COM), completely observed cases (CCA) and multiply imputed data sets. The fraction of observed x_1 values is 0.65 and the true value of β is 1. x_1 is χ^2-distributed with three degrees of freedom and discrete MDM with three classes.

severely biased.

The only imputation function that seems to lead to acceptable results with respect to coverage even if $n = 1000$ is gamlss. It leads to biased estimates if $n = 40$, but this bias vanishes with increasing n. However, the robustness of this imputation technique comes at the price of relatively large variances in the small and medium sample sizes.

In a simulation study of He and Raghunathan (2009), all tested imputation methods broke down when the distribution of the variable to be imputed was strongly heavy tailed (see also de Jong, van Buuren and Spiess, 2014). Therefore, in a next step, all the covariates were generated according to a t-distribution with three degrees of freedom. Again adopting the three-class MDM to generate missings in x_1, the results are presented in Table 4.

	n=40				n=100				n=1000			
Technique	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV	$m(\hat{\beta})$	sd	$m(\widehat{sd})$	COV
COM	0.987	0.354	0.349	0.942	0.995	0.196	0.199	0.955	0.999	0.058	0.056	0.950
CCA	0.825	0.482	0.437	0.924	0.865	0.255	0.242	0.914	0.893	0.075	0.068	0.636
LM	0.926	0.526	0.488	0.935	1.003	0.272	0.264	0.953	1.040	0.079	0.070	0.877
amelia	0.990	0.557	0.508	0.921	1.036	0.275	0.267	0.940	1.042	0.080	0.071	0.881
dpmm, $k = 1$	0.888	0.542	0.500	0.920	0.948	0.299	0.276	0.929	1.004	0.089	0.073	0.896
dpmm, $k = 3$	0.860	0.511	0.495	0.936	0.928	0.288	0.273	0.936	0.999	0.087	0.071	0.894
dpmm, $k = 5$	0.828	0.497	0.495	0.936	0.912	0.287	0.273	0.937	0.995	0.087	0.072	0.897
dpmm, $k = 10$	0.782	0.464	0.494	0.949	0.885	0.275	0.276	0.942	0.989	0.087	0.072	0.900
dpmm, $k = 20$	0.685	0.416	0.496	0.940	0.842	0.261	0.276	0.937	0.979	0.086	0.072	0.900
dpmm, $k = \sqrt{n}$	0.823	0.487	0.492	0.935	0.882	0.276	0.275	0.939	0.970	0.086	0.073	0.889
dmice, $k = 1$	0.912	0.578	0.493	0.901	0.962	0.311	0.266	0.916	1.005	0.092	0.069	0.863
dmice, $k = 3$	0.880	0.526	0.487	0.928	0.946	0.297	0.264	0.922	1.001	0.087	0.070	0.880
dmice, $k = 5$	0.853	0.511	0.488	0.936	0.930	0.289	0.264	0.926	0.997	0.087	0.071	0.894
dmice, $k = 10$	0.796	0.477	0.491	0.947	0.898	0.274	0.267	0.937	0.991	0.085	0.071	0.893
dmice, $k = 20$	0.690	0.416	0.495	0.941	0.852	0.263	0.270	0.931	0.981	0.085	0.072	0.901
dmice, $k = \sqrt{n}$	0.846	0.492	0.490	0.937	0.899	0.273	0.267	0.938	0.972	0.086	0.072	0.885
aI-PMM	0.787	0.510	0.522	0.946	0.880	0.292	0.287	0.942	0.959	0.090	0.082	0.906
aI-boot	0.763	0.476	0.519	0.944	0.844	0.279	0.296	0.944	0.954	0.089	0.085	0.913
irmi	0.828	0.477	0.502	0.953	0.879	0.251	0.280	0.953	0.911	0.075	0.079	0.809
gamlss	0.744	0.538	0.562	0.886	0.919	0.330	0.331	0.924	0.895	0.190	0.168	0.948

Table 4: Estimation results based on complete data set (COM), completely observed cases (CCA) and multiply imputed data sets. The fraction of observed x_1 values is 0.65 and the true value of β is 1. Multivariate t-distributed covariates with three degrees of freedom and discrete MDM with three classes.

Obviously, none of the imputation techniques leads to acceptable results if $n = 1000$. In that case, all the imputation techniques lead to biased estimates or underestimated variances. An outstanding example is gamlss which leads to biased estimates but at the same time to underestimated variances. A closer look on single simulation results reveals that this is due to some extreme cases where estimates of β are far to small, which is the consequence of very bad model fits.

For smaller samples sizes, coverages may be acceptable even with biased estimates because variances are overestimated, as for aI-PMM, aI-boot with $k = 3$ and irmi. On the other hand, gamlss leads to biased estimates and low coverages. The parametric imputation techniques LM and amelia lead to biased estimates and slight undercoverage for $n = 40$ but work well if $n = 100$. The pattern for dpmm and dmice is similar to what we observed before: With increasing k, the bias and the relation of true to estimated variance increases, eventually leading to acceptable coverages in single cases, like dmice or dpmm with $k = 10$ and $n = 40$.

6 The effect of functions of music on music preferences.

As an illustration, we estimated a linear regression model of the degree of preferences for ones own favourite music style on the degree to which the corresponding kind of music serves the needs of the listener ('functions of music') following Schäfer and Sedlmeier (2009) and based on the data provided by these authors.

More precisely, the dependent variable was the mean over six statements measured via 10-point Likert scales ('I like this music', 'I couldn't live without this music', 'I just need this music', 'I'm a passionate listener of this music', 'I regularly visit clubs or concerts to listen to this music', 'I usually spend a lot of money to purchase this music') with the poles 'not agree at all' and 'completely agree' and representing different aspects of the same

'music preference' construct (see Schäfer and Sedlmeier, 2009, for details). Independent variables were functions of music, again measured on 10-point Likert scales with extreme values 'not agree at all' and 'completely agree'. The statements ('functions of music') used in the following analysis are 'Expresses my values' (values), 'Gives me information' (inform), 'Expresses my identity' (identity), 'Puts me in a good mood' (mood), 'Makes me feel ecstatic' (ecstatic), 'Helps me meet people' (contact) and 'Is music I can appreciate as art' (art). Additionally we included age as a covariate.

Data were collected via the internet and the questionnaires were completed online (for more details, see Schäfer and Sedlmeier, 2009). In our analysis, we used the data of $n = 476$ completely observed units. Since we want to illustrate the effects of the different imputation techniques we first estimated the regression parameters based on the full sample, and then generated 167 missing values in the covariate 'values', according to the three-class missing mechanism described in Section 4. Thus the missing values are MAR.

The results for COM in Table 5 imply that all the variables included with but two exceptions seem to have a positive effect on the degree of music preference ($\alpha = 0.05$). The exceptions are the function 'Gives me information' which seems to have no (significant) effect and age which seems to have a negative effect. Not surprisingly, the results under the CCA strategy would imply a different pattern: Instead of the function of music expressing ones values, now the function that music provides information seems to have an effect.

If the data set is multiply imputed with the Bayesian linear model, we would conclude that – in contrast to the analysis under COM – the function 'values' has no effect. The data set imputed with AMELIA would lead to the same conclusions as under the data set without missing values.

If the incomplete data set is imputed with a kNN technique, then the conclusions vary even within one technique, depending on the number of neighbours. They can be the same as for the complete data set (dpmm with $k = 1, 5, 20$ or $k = \sqrt{n}$ or dmice with

Method	Intercept	values	inform.	identity	mood	ecstatic	contact	art	age
COM	2.606*	0.090*	0.051	0.114*	0.170*	0.107*	0.110*	0.098*	-0.042*
CCA	2.441	0.071	0.049	0.097	0.134	0.110	0.154	0.081	-0.023
LM	2.682	0.074	0.051	0.117	0.165	0.110	0.112	0.099	-0.042
amelia	2.689	0.075	0.049	0.118	0.163	0.110	0.113	0.100	-0.042
dpmm, $k = 1$	2.642	0.084	0.048	0.117	0.164	0.108	0.113	0.099	-0.041
dpmm, $k = 3$	2.641	0.071	0.054	0.120	0.169	0.110	0.112	0.100	-0.042
dpmm, $k = 5$	2.635	0.092	0.044	0.115	0.165	0.109	0.113	0.098	-0.042
dpmm, $k = 10$	2.639	0.075	0.053	0.119	0.169	0.110	0.111	0.100	-0.042
dpmm, $k = 20$	2.642	0.080	0.052	0.118	0.166	0.109	0.112	0.099	-0.042
dpmm, $k = \sqrt{n}$	2.630	0.074	0.054	0.120	0.170	0.109	0.110	0.100	-0.042
dmice, $k = 1$	2.651	0.087	0.046	0.115	0.165	0.109	0.112	0.098	-0.042
dmice, $k = 3$	2.667	0.082	0.049	0.117	0.162	0.110	0.113	0.099	-0.042
dmice, $k = 5$	2.670	0.087	0.047	0.116	0.163	0.110	0.111	0.099	-0.042
dmice, $k = 10$	2.650	0.079	0.052	0.118	0.166	0.110	0.111	0.099	-0.042
dmice, $k = 20$	2.639	0.074	0.055	0.119	0.167	0.110	0.111	0.100	-0.041
dmice, $k = \sqrt{n}$	2.607	0.083	0.051	0.118	0.169	0.109	0.111	0.099	-0.041
irmi	2.669	0.058	0.061	0.123	0.172	0.108	0.111	0.100	-0.042
gamlss	2.653	0.082	0.050	0.116	0.168	0.107	0.112	0.100	-0.042

Table 5: Regression results based on the complete data set (COM), complete case analysis (CCA), nearest neighbour techniques based on predictive mean matching (pmm) and and multiply imputed data sets. The fraction of observed values of variable values is 0.65 and the missing values are generated according to the discrete MDM with three classes. For the analysis COM, $*$ marks effects where a 95% confidence interval does not cover 0. Significance results differing from the complete data analysis are marked by a surrounding box; $n = 476$ and number of imputations $m = 10$.

$k = 1, 3, 5$), as under the CCA strategy (dpmm with $k = 10$ and dmice with $k = 20$), like under the imputation technique with the linear Bayesian model (dpmm with $k = 3$ and dmice with $k = 20$) or even completely different in that in addition to all other variables 'information' is also significant (dmice with $k = \sqrt{n}$).

As a general result and in light of the results in Section 5, one may conclude from Table 5 that for an increasing number of neighbours, the estimator of the parameter weighting the imputed covariate 'values' seems to be increasingly biased. Imputation with irmi results in inferences similar to CCA. In contrast to CCA, however, $\hat{\beta}_1$ and its standard error is smaller, whereas the effect of 'Gives me information' and it's standard error are larger. Imputation with gamlss leads to very similar inferences as under the complete data condition.

Further, the different imputation techniques are all techniques based on simulated values. Thus, for one single data set, the conclusions may differ even if we assume that some of the techniques do not misspecify the true data generating model and the conclusions even for the same technique may differ if we run the imputation step with other starting values for the pseudo random number generators. In addition, results may not change for covariates which are uncorrelated with the imputed variable, but may differ even for completely observed covariates which are correlated with the imputed one.

7 Discussion

In this paper we compared several versions of kNN imputation, model based Bayesian imputation as proposed by Rubin (1987) and semi-/nonparametric approaches like irmi (Templ et al., 2011) and gamlss (de Jong, van Buuren and Spiess, 2014) in different scenarios via simulations.

The results show that the parametric imputation methods following the suggestions of Rubin (1987) work in many situations. However, if the imputation models are misspecified, then estimates themselves but also their estimated variances tend to be biased. This may lead to severe undercoverage and thus to rejecting true null hypotheses far too often.

Usually in regression models, the (conditional) distribution of the dependent variable is modelled. If one or more of the covariates are affected by missing values, however, then

a (conditional) distribution of the variables with missing values has to be assumed to generate imputations. This distribution may be much harder to justify as we are usually not well prepared for this task. Thus, imputations have to be generated very carefully, instead of just adopting standard or default techniques.

Several popular software packages offer kNN imputation as one or even the default option for generating multiple imputations. Advantages of kNN imputation often claimed are its simplicity, the avoidance of strong parametric assumptions, the ease with which this technique can be applied to other types of variables and that only eligible and observed imputations are generated. However, kNN techniques can not be recommended in general: The optimal k is unknown and it seems not to be possible to clearly identify situations that allow valid inferences with kNN imputation. Looking at the coverage and bias, there are small differences between the techniques considered. Further, the strategy proposed by Dahl (2007) of taking the number of donors as the square root of the sample size does not lead to better results as compared to the other methods, and the non-parametric kNN version of Harrell (2015) seems not offer an improvement over standard parametric versions. Further research could pick up the idea of considering kNN techniques that optimize k locally, depending on the density of completely observed cases close to the query point (Schenker and Taylor, 1996). This would make this technique, however, more complicated and costly.

Imputation technique irmi (Templ et al., 2011) can not be recommended either - it leads to invalid inferences under most scenarios. This may be due to the automatic outlier detection: If the missing mechanism thins out certain regions, then this algorithm may identify 'outliers' which are in fact valid values leading to improper imputation models. This may even be worse if the variable with missing data is not symmetrically distributed.

A technique that allowed valid inferences in most cases considered in this work is based on the semi-parametric technique gamlss (de Jong, van Buuren and Spiess, 2014). However, it fails in case of heavy tailed distributions and although promising in general,

this technique is not implemented yet in a stable version.

The main message of this study is that some available parametric methods, if based on Rubin's theoretical arguments, work if misspecification is not severe. Surprisingly, the self-correcting property claimed, e.g. in Little and Rubin (2002) works to some extent in small samples, but fails to work in several cases considered here. Then, even CCA may be closer to acceptable results with respect to coverage than most of the parametric, robust or semi-parametric techniques.

References

Andridge, R.R. & Little, R.J.A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review, 78,* 40–64.

Dahl, F.A. (2007). Convergence of random k-nearest-neighbour imputation. *Computational Statistics & Data Analysis, 51*, 5913–5917.

de Jong, R., van Buuren, S. & Spiess, M. (2014). Multiple imputation of predictor variables using generalized additive models. To appear in: *Communications in Statistics – Computation and Simulation.*
DOI: 10.1080/03610918.2014.911894

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Ser. B, 39(1),* 1–22.

Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11(2),* 89–121.

Harrell, F.E. (2015). Package 'Hmisc'. Version 3.15-0.
http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf

He, Y. & Raghunathan, T. (2009). On the Performance of Sequential Regression Multiple

Imputation Methods with Non Normal Error Distributions. *Communications in Statistics – Simulation and Computation, 38(4),* 856–883.

Honaker, J. & King, G. (2010). What to Do about Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science, 54(2),* 561–581.

Honaker, J., King, G. & Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software, 45(7),* 1–47.

Horton, N.J. & Kleinman, K.P. (2007). Much Ado About Nothing. *The American Statistician, 61(1),* 79–90.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics, 6,* 287–296.

Little, R.J.A. & Rubin, D.B. (2002). *Statistical analysis with missing data (2nd ed.).* New York: John Wiley.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science, 9 (4)*, 538–558.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing: Vienna, Austria. (http://www.R-project.org)

Rigby, R.A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3),* 507–554.

Robins, J.M. & Wang, N. (2000). Inference for Imputation Estimators. *Biometrika, 87(1)*, 113–124.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika, 63(3),* 581–590.

Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4,* 87–94.

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

Rubin, D.B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association, 91(434),* 473–489.

Rubin, D.B. (2003). Discussion on Multiple Imputation. *International Statistical Review, 71(3)*, 619–625.

SAS Institute Inc. (2013). SAS/STAT User's Guide, 13.1. Cary, NC: SAS Institute Inc.

Schäfer, T. & Sedlmeier, P. (2009). From the functions of music to music preference. *Psychology of Music, 37*, 279–300. (Original DOI: 10.1177/0305735608097247)

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall.

Schafer, J.L. & Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7(2),* 147–177.

Schenker, N. & Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis, 22,* 425–446.

Schenker, N. & Welsh, A.H. (1988). Asymptotic Results For Multiple Imputation. *The Annals of Statistics, 16(4),* 1550–1566.

Spanos, A. (1995). On normality and the linear regression model. *Econometric Reviews, 14(2),* 195–203.

Stasinopoulos, D.M. & Rigby, R.A. (2007). Generalized additive models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software, 23(7),* 1–46.

Templ, M., Alfons, A., Kowarik, A. & Prantner, B. (2014). *VIM: Visualization and*

Imputation of Missing Values. R package version 4.1.0. http://CRAN.R-project.org/package=VIM.

Templ, M., Kowarik, A. & Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics and Data Analysis, 55,* 2793–2806.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software 45(3),* 1–67.

Vink, G., Frank, L.E., Pannekoek, J. & van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica, 68(1),* 61–90.

Yohai, V. (1987). High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics 15,* 642–665.

Yu, L.-M., Burton, A. & Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research, 16,* 243–258.

1 de Jong, R., van Buuren, S. & Spiess, M. (2013). *Multiple imputation of predictor variables using generalized additive models.*

2 Jordan, P. & Spiess, M. (2013). *Fundamentale Probleme beim Einsatz testtheoretischer Modelle zur Diagnose von Individuen.*

3 Salfran, D. & Spiess, M. (2015). *A Comparison of Multiple Imputation Techniques.*