BIAS REDUCTION OF FINITE POPULATION IMPUTATION BY KERNEL METHODS

NICKLAS PETTERSSON

Department of Statistics, Stockholm University

nicklas.pettersson@stat.su.se

Abstract

Missing data is a nuisance in statistics. Real donor imputation can be used with item nonresponse. A pool of donor units with similar values on auxiliary variables is matched to each unit with missing values. The missing value is then replaced by a copy of the corresponding observed value from a randomly drawn donor. Such methods can to some extent protect against nonresponse bias. But bias also depends on the estimator and the nature of the data. We adopt techniques from kernel estimation to combat this bias. Motivated by Pólya urn sampling, we sequentially update the set of potential donors with units already imputed, and use multiple imputations via Bayesian bootstrap to account for imputation uncertainty. Simulations with a single auxiliary variable show that our imputation method performs almost as well as competing methods with linear data, but better when data is nonlinear, especially with large samples.

Keywords Bayesian bootstrap; Boundary and nonresponse bias; Missing data; Multiple imputation; Pólya urn models; Real donor imputation.

1. INTRODUCTION

In sample surveys missing data often has to be dealt with. Imputation is a standard treatment for sporadically missing values in the sample data due to item nonresponse. Given observed auxiliary variable(s) *X* related to the incomplete study variable *Y*, an imputation model is usually estimated from units where both *x* and *y* values are observed, modelled by a missing at random (MAR) mechanism which assumes that the probability of missingness only depends on observed values. The missing *y* values are then replaced by imputed values, and multiple imputation can account for the fact that imputed values differs from the true ones, so that standard methods can be used (Rubin, 1987). Imputed values may be non-observable values derived from a model, or real-donor values derived from observed values (Laaksonen, 2000). Donors to each donee (or recipient) are usually found by selecting units close to the donee according to some distance measure on *X*.

Imputation methods employing parametric models may be effective (Schafer, 1997), but their benefits diminish with sample size and can lead to severe bias if the underlying assumptions are violated. Methods based on nonparametric models can then provide robustness to nonresponse bias at the cost of some efficiency. Bias of methods based on nonparametric models also depends on the derivation of the imputed values, and the nature of the bounded data. The bias of a mean estimate of *y* is related to the individual unit bias of *x*, the expectation over donor *x*'s minus the actual *x*, through individual unit bias of *y*. When *X* is continuous, the asymptotic bias of *x* for an interior donee can easily be set to zero. This is more difficult for donees that lie on the boundary of the data. By viewing imputation as pointwise kernel smoothing, and adopting bias reduction techniques from that area, we propose a real donor method which aims at mitigating such bias of individual *x* as to implicitly reduce bias of the mean estimator of *y*.

Our method starts out from the popular hot deck imputation; see Little and Andridge (2010) for a review. For each donee unit where *y* is missing, a pool consisting of *k* potential donor units with observed *y*-values is identified. The missing *y* value of the donee is then filled in by a copy of the observed *y* value from a unit in the donor pool. Adjustment cells imputation bring together all zero distance donors and donees, having the same categorized *x*, creating an illusion that individual *x*'s are unbiased. Cells may therefore only contain donees. This is avoided by non-categorizing distance measures, which produce donor pools that can be better matched to the donee, but the number of *k* nearest neighbour (*k*NN) donors has to be decided. Justified by Bayesian exchangeability through Pólya sampling (Feller, 1971), we extend the set of potential donors to include previously imputed donees, and handle imputation uncertainty through multiple imputation.

Individual bias in *x* is first addressed by relating distances between the donee and the donors to the donor selection probabilities, giving closer donors higher donation probability. Siddique and Belin (2008) set selection probabilities inversely proportional to the distance between predictive means of donor and donee units, while Conti, Marella and Scanu (2008) let a Gaussian kernel decide the selection probabilities. We propose to use an Epanechnikov (1969) kernel, which asymptotically can minimize mean squared error of an estimate. We expect reduction of variance in general and boundary donee bias of *x*.

Boundary bias can also be reduced by letting the selection probabilities be found from local linearization (Simonoff, 1996). Aerts, Claeskens, Hens and Molenberghs (2002) use non-negative constrained weights asymptotically equivalent to kernel weights as selection probabilities. We calibrate our selection probabilities by a Lagrange function, similar to calibration of design weights (Deville and Särndal, 1992), but on a pointwise level.

Our third bias reduction method is inspired by Rice (1984), who tightened the kernel at the boundary. By reducing *k* for boundary donees, on average closer but fewer donors are obtained compared to interior donees, which contribute to the bias reduction of *x*.

The paper is structured as follows; Section 2 presents real donor imputation with Pólya urn sampling and multiple imputation. Our proposed methods are described in Section 3, and further studied by simulations in Section 4. The paper is then concluded in Section 5.

2. BACKGROUND ON REAL DONOR AND MULTIPLE IMPUTATION

A simple random sample (SRS) of i=1, …, n units from a population of N units is drawn with the aim to estimate the mean $\bar{y} = \sum y_i / N$ of the study variable Y, and the value y_i is observed in the sample. The indicator R_i=1 for the r units where y_i is observed, while R_i=0 for nonresponding units. In real donor imputation, each donee i should have a donor pool of k_i units. Denote by q_i the number of units that possibly could enter pool i. Given our SRS design, we simply set k_i=q_i=r for all i, and use all respondents as potential donors. Later we allow k_i, q_i, and r to differ, and may omit index i when it is dispensable.

For each donee i, a donor j is selected with probability λ_{ij}, and the imputed value $\hat{y}_i$ is a copy of y_j. When all n-r missing values have been imputed, an estimate of $\bar{y}$ is

$$\hat{\bar{y}} = \frac{1}{n}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_i \right). \tag{1}$$

Since the expectation of an imputed value is

$$E(\hat{y}_i) = \sum_{j=1}^{q} \lambda_{ij} y_j \ , \tag{2}$$

the individual bias of $\hat{y}_i$ is

$$B(\hat{y}_i) = E(\hat{y}_i) - y_i \, . \tag{3}$$

Due to the SRS design it follows that $E(y_i) = \bar{y}$. The bias of (1) is therefore

$$B(\hat{\bar{y}}) = E(\bar{\bar{y}}) - \bar{y} = \frac{1}{n}\left\{\sum_{i=1}^{r} E(y_i) + \sum_{i=r+1}^{n} E(\hat{y}_i)\right\} - \bar{y} = \frac{1}{n}\sum_{i=r+1}^{n} B(\hat{y}_i) \, . \tag{4}$$

Now assume a known auxiliary variable *X* and a MAR mechanism, so that the response probability does not depend on *y*; *P*(*R*=1|*Y*,*X*)=*P*(*R*=1|*X*). We further assume that the expected value of *Y* does not depend on *R*, *E*(*Y*|*X*)=*g*(*X*), which is another consequence of MAR. Denote the *x*-value of the donor selected for donee *i* by $\hat{x}_i$. Its expectation is $E(\hat{x}_i) = \sum_{j=1}^{q} \lambda_{ij} x_j$. We may expect to reduce (3), and thereby (4), by reducing the bias of x_i

$$B(\hat{x}_i) = E(\hat{x}_i) - x_i = \sum_{j=1}^{q} \lambda_{ij} x_j - x_i \, . \tag{5}$$

2.1 ADJUSTMENT CELLS AND K-NEAREST NEIGHBOUR IMPUTATION

As a background we first describe two common methods for imputation, adjustment cells and nearest neighbour imputation. Our suggested method in Subsection 3.3 is based on the latter. All methods are illustrated on the simple dataset in Table 1, where *x* is observed on all *n*=7 units, while *y* is only observed on *r*=5 units. Table 1 is ordered after *x*. The cut off between the two adjustment cells is set to *x*=0. Let $\lambda_{ij} = 1/k_i$ for donee *i*=3, 6 and donor *j*. Since donor pools are determined from *x*, we usually have that $k_i < q_i$.

Example 1. Imputation within adjustment cells. Only units within the same adjustment cell may be used as donors. So although q_3=*r*=5, the k_3=4 potential donors for Unit 3 are Units 1, 2, 4 and 5, and $B(\hat{x}_3) \approx -0.013$. We randomly draw one of them, say Unit 4, and impute the missing *y*-value as $\hat{y}_3 = 0.022$. Unit 7 is the only (k_6=1) potential donor to Unit 6, so $B(\hat{x}_6) = 0.231$ and we impute $\hat{y}_6 = -0.099$. If single donor situations are not allowed, a common solution is to collapse adjustment cells. Units 1, 2, 4, 5 and 7 are then the (k_3=q_3=5*)* potential donors to Unit 3, and $B(\hat{x}_3) \approx -0.107$.

Assume again we draw Unit 4. Unit 6 has the same donors, so $B(\hat{x}_6) \approx -0.247$. If the imputed Unit 3 also had been allowed to act as a donor (so that $k_6=q_6=r+1=6$) we would have had $B(\hat{x}_6) = -0.265$.

Table 1. Data in Examples 1-5, with *x* and *y* generated by model *NO* in Subsection 4.1.

Unit no	1	2	3	4	5	6	7
x-cat.	1	1	1	1	1	2	2
x	-0.413	-0.381	-0.255	-0.152	-0.125	0.099	0.330
y	-0.555	-0.476	Missing (-0.136)	0.022	0.349	Missing (0.335)	-0.099

Note: (the true but unknown value in parenthesis is given here for illustrative purposes.)

Example 2. Imputation by kNN. We now discard the categorization of *x*, and use 4NN imputation (i.e. $k_3=k_6=4$). Since Units 1, 2, 4 and 5 are the closest (among the $q_3=5$) units to donee Unit 3, $B(\hat{x}_3) \approx -0.013$ as in Example 1. Assume unit 4 was drawn. Unit 6 then has Units 2, 4, 5 and 7 as donors with $B(\hat{x}_6) = -0.181$. By allowing the imputed Unit 3 as a donor (so that $q_6=r+1=6$) we get $B(\hat{x}_6) \approx -0.150$ based on Units 3, 4, 5, and 7.

Adjustment cells imputation effectively matches donors to a donee and is widely used. But having a single donor can severely affect variances, as explained in Subsection 2.3. Collapsing cells is a simple solution, but *k*NN can provide better matching. Since Unit 3 has half of its donors on each side (as $k_3=4$) we call it an interior unit, while Unit 6 with only a single donor on the right is called a boundary unit. We will make use of this distinction in Subsection 3.3, where we suggest how to further improve *k*NN matching and try to reduce bias. Allowing imputed donees to act as donors for subsequent donees differs from usual donor imputation, but a Bayesian justification based on Exchangeability and Pólya urns is given in Subsection 2.2.

2.2 IMPUTATION BY PÓLYA URN SAMPLING AND BAYESIAN BOOTSTRAP

Descriptions of imputation methods which use previously imputed values in subsequent imputations can be found in Rubin (1987) and Kong, Liu and Wong (1994). These methods attempt to impute the missing values by draws from their posterior predictive distributions, and rely on a Bayesian motivation going back to de Finetti's (1931) theorem on exchangeable sequences. If the probability distribution for any finite sequence of *n* random variables drawn from an infinite series of random variables is the same, then any such infinite series is exchangeable. A sequence of independent and identically distributed (iid) random variables is always exchangeable, but the opposite is not true. But under some assumptions any exchangeable sequence is distributed as a sequence that is iid given some parameters which in turn have a prior distribution. Hewitt and Savage (1955) generalized de Finetti's theorem to non-binary variables, and Diaconis and Freedman (1980) showed that it is approximately true for long but finite sequences of variables, implying finite exchangeability.

Pólya urn sampling produces an exchangeable but non-iid series, see Feller (1971). Assume a sample of *n* units where we have observed either the value 0 or 1 on variable *Y*. Then 1) draw a single unit at random from the sample, 2) duplicate the drawn unit, and 3) replace both the drawn and the duplicated unit into the sample. The procedure is then repeated, but now with the updated sample of size n+1. By repeating the procedure ad infinitum, the generated sequence of values on the units is then an infinite exchangeable sequence. Blackwell and MacQueen (1973) generalized Pólya urn sampling to allow for more than two categories, and Ferguson (1973) extended to continuous variables.

Finite population Bayesian bootstrap (FPBB) (Lo, 1988) is based on Pólya urn sampling from a sample (of size *n*) to a large finite population (of size *N*). If a sample is drawn by SRS and the observed units are randomly drawn from the sample itself by SRS, then the observed units may be treated as a part of an exchangeable series of variables. In our example (Table 1) we may treat the sample as the population, and the five observed units as our sample. Pólya sampling may then be applied to reconstruct the remaining n-r units from the r observed ones, corresponding to imputation within the collapsed adjustment cells using Unit 3 as potential donor to Unit 6 in Example 1 (where k_6=q_6= r+1=6). Knowing the full population size, Pólya sampling can be done to the whole population, starting from the *r* observed units, and sequentially impute all N-r units. An estimate of $\bar{y}$ is then simply the mean of the bootstrap population.

As $N \to \infty$, FPBB approaches the model based Bayesian bootstrap by Rubin (1981). They raise two objections to bootstrap methods in connection to the exchangeability assumption. First they ask whether it is reasonable to assume that all possible distinct values of a variable have been observed in a sample. The objection is definitely valid with the continuous and very small sample in Table 1. Assuming unlimited precision all realized values of a continuous variable are unique, so we will not observe all values until we have observed the whole population. But our ability to grasp the data distribution should improve with the sample size, unless data is censored or if missingness in other ways is concentrated to certain regions of the data. This (strong) dependence on sample size is a characteristic common to nonparametric methods, simply because they refrain from parametric assumptions.

Assuming all possible distinct values are observed, Rubin’s second objection is that the probabilities of occurrences for similar values might be dependent. This calls for smoothing of probabilities, but bootstrapping assumes strict independence. If the distribution of realized or bootstrap samples differs much from the true population, some estimators might perform poorly. As for the first objection, the larger the sample, the more likely we are to observe the distribution of the true data, so benefits from smoothing should, in general, diminish.

2.3 BAYESIAN BOOTSTRAP AND MULTIPLE IMPUTATION

Imputation by FPBB basically corresponds to multiple imputation (Rubin, 1987). A general overview of variance estimation with single imputation is given in Little and Rubin (2002), and an overview for hot deck imputation in Andridge and Little (2010).

Assume a sample from a finite population of exchangeable units with *n*-*r* missing values on variable *Y* imputed *d*=1, …, *D* times. The distribution of the estimates

$$\hat{\bar{y}}_{d,n} = \frac{1}{n}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_{di}\right), \; d=1,\ldots,D, \qquad (6)$$

then reflects the imputation uncertainty due to that imputed values for the same unit differs between the imputed datasets. A point estimate of $\bar{y}$ is given by

$$\hat{\bar{\bar{y}}}_n = \frac{1}{D}\sum_{d=1}^{D} \hat{\bar{y}}_{d,n} \;, \qquad (7)$$

and the variance of $\hat{\bar{\bar{y}}}_n$ is estimated as

$$\hat{V}\left(\hat{\bar{\bar{y}}}_n\right) = \frac{D+1}{D} B_n + \overline{W_n}\,. \qquad (8)$$

Component $B_n = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\bar{y}}_{d,n} - \hat{\bar{\bar{y}}}_n\right)^2$ accounts for imputation uncertainty, and sampling uncertainty is covered by the variance component $\overline{W_n} = \frac{1}{D}\sum_{d=1}^{D} W_{d,n}$, where

$$W_{d,n} = \left(\frac{N-n}{N-1}\right)\left(\frac{1}{n-1}\right)\left[\left(\sum_{i=1}^{r} y_{di} - \hat{\bar{y}}_{d,n}\right)^2 + \left(\sum_{i=r+1}^{n} \hat{y}_{di} - \hat{\bar{y}}_{d,n}\right)^2\right], d=1,\ldots,D, \qquad (9)$$

is the estimated variance within a bootstrap set. The term $\frac{N-n}{N-1}$ is the finite population correction. If both the *n-r* non-responding and the *N-n* non-sampled units in each bootstrap set had been imputed, then a population estimate similar to (7) would have been

$$\hat{\bar{\bar{y}}}_N = \frac{1}{D}\sum_{d=1}^{D}\hat{\bar{y}}_{d,N} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{N}\hat{y}_{di}\right). \qquad (10)$$

Sampling uncertainty vanishes with a completely imputed population, so (8) simplifies to

$$\hat{V}\left(\hat{\bar{\bar{y}}}_N\right) = \frac{D+1}{D}B_N = \left(\frac{D+1}{D}\right)\left(\frac{1}{D-1}\right)\sum_{d=1}^{D}\left(\hat{\bar{y}}_{d,N} - \hat{\bar{\bar{y}}}_N\right)^2 . \qquad (11)$$

With missing values deterministically imputed, as in the uncollapsed cell in Example 1 with a single donor (k_6=1), all imputed bootstrap sets will have the same value imputed, so B_N (or B_n) will be underestimated. In particular, if all values are deterministically imputed, then $\hat{\bar{y}}_{1,N} = \ldots = \hat{\bar{y}}_{D,N} = \hat{\bar{\bar{y}}}_N$, implying that $B_N = 0$, so that $\hat{V}\left(\hat{\bar{\bar{y}}}_N\right) = 0$ in (11).

3. KERNEL ESTIMATION AND KERNEL IMPUTATION

One may look at donor imputation from the view of kernel estimation. We give a brief introduction to the area, describe the connections to imputation, and suggest how to improve estimation and achieve bias reduction of (7) or (10) using auxiliary variable *X*.

3.1 SHORT BACKGROUND ON KERNEL ESTIMATION

Kernel estimation is a method to estimate a density. Assume that *q* values are observed on *x* and a density *f*(*x*) at a point x_i is to be estimated.

Denote the distance x_j-x_i by $\tilde{x}_{ij}$. Given a kernel function *K*, the pointwise kernel estimate of *f* at x_i is then

$$\hat{f}(x_i) = \frac{1}{qh}\sum_{j=1}^{q} K\left(\frac{\tilde{x}_{ij}}{h}\right) = \frac{1}{q}\sum_{j=1}^{q} K_h(\tilde{x}_{ij}),$$

where *K* is typically symmetric, unimodal and integrates to 1. We restrict to situations where *K* is proportional to the indicator function $I(|\tilde{x}_{ij}| < h)$, which is 1 if the statement is true. Function K_h is *K* scaled by the bandwidth (or smoothing) parameter *h*, which determines that *K* is positive if $|\tilde{x}_{ij}| < h$, and zero if $|\tilde{x}_{ij}| \geq h$. The choice of *h* is usually more important than *K*. If *h* is fixed for all *i*, the number of units $k_i \geq 0$ within the range $x_i \pm h$ is random. Instead, if the number of units k_i is fixed at *k*, the bandwidth h_i will be random. Methods to select a fixed *h* or *k* range from subjective judgement of plots and simple automatic rules of thumb, to more sophisticated methods based on cross-validation and plug-in estimates (Wand and Jones, 1995). Fixing *h* is more frequent, and a fixed *k* is best used when the exact size is noncritical, typically with $k \approx q^{1/2}$ (Silverman, 1986).

A commonly used measure of accuracy is the mean integrated squared error (MISE)

$$MISE(\hat{f}) = \int E(\hat{f}(x) - f(x))^2 dx = \int (B\{\hat{f}(x)\})^2 dx + \int V\{\hat{f}(x)\} dx, \qquad (12)$$

where a pointwise approximation of the bias component is given by

$$B\{\hat{f}(x_i)\} = E\{\hat{f}(x_i)\} - f(x_i) = \frac{1}{q}\sum_{j=1}^{q} E\{K_h(\tilde{x}_{ij})\} - f(x_i), \qquad (13)$$

and an approximation of the variance with independent x_j is given by

$$V\{\hat{f}(x_i)\} = \frac{1}{k_i} V\{K_h(\tilde{x}_{ij})\}. \qquad (14)$$

Given that *K* is symmetric and *h* (or *k*) is reduced, bias in (13) will decrease while variance in (14) will increase. The variance goes to zero as $qh \to \infty$ (or $k \to \infty$), while bias depends on the curvature of *f* and is asymptotically unrelated to *q*, unless $h \to 0$ (or $k/q \to 0$) as $q \to \infty$. Bias then converge to zero if x_i lies in the interior (unbounded) part of *x*, while if x_i lies within a bandwidth *h* from the boundary of *x*, the bias will not vanish. Given an optimal choice of *h*, MISE in (12) is approximately minimized if *K* is set to the unimodal Epanechnikov (1969) function

$$K_h^{Ep}(\tilde{x}_{ij}) = \frac{3}{4}\{1 - (\tilde{x}_{ij})^2\} I(|\tilde{x}_{ij}| < h). \qquad (15)$$

3.2 KERNEL IMPUTATION

Assume K_h is a positive function scaled so that $\sum_{j=1}^{q} K_h(\tilde{x}_{ij}) = 1$, where $\tilde{x}_{ij} = x_i - x_j$ and the sum is over the donor pool described in Subsection 2.1. When the selection probabilities are given by $\lambda_{ij} = K_h(\tilde{x}_{ij})$ we call the technique kernel imputation. The expectation of $\hat{y}_i$ in (2) thus becomes the Nadaraya-Watson (1964) estimator

$$E(\hat{y}_i) = \sum_{j=1}^{q} \lambda_{ij} y_j = \sum_{j=1}^{q} K_h(\tilde{x}_{ij}) y_j \ .$$

With a uniform kernel $K_h^{Un}(\tilde{x}_{ij}) \propto I(|\tilde{x}_{ij}| < h)$, the donee *i* has *k* potential donor units within the range $x_i \pm h$ with selection probabilities $\lambda_{ij} = K_h^{Un}(\tilde{x}_{ij}) = 1/k$, and *q*-*k* units outside the range with $\lambda_{ij} = 0$. When donor data at x_i is sparse, fixing *k* instead of *h* will cover more distant donors, which avoids situations with no or few donors. With donors densely located in a vicinity of x_i, using an adaptable parameter h_i (caused by the fixed *k*) will in general result in donor pools that are better matched to donee *i*.

3.3 KERNEL IMPUTATION WITH BIAS REDUCTION

We suggest the use of multiple kernel imputation but also add three special devices, mainly to decrease imputation bias, but also to decrease the random errors. The bias $B(\hat{\bar{y}}_i)$ in (4) is related to $B(\hat{x}_i)$ in (5) and $B\{\hat{f}(x_i)\}$ in (13) through $B(\hat{y}_i)$ in (3) and $\lambda_{ij} \propto K_h(\tilde{x}_{ij})$. Given a model *E*(*Y*|*X*)=*g*(*X*) and a response mechanism *P*(*R*=1|*X*), we will probably reduce $B(\hat{\bar{y}}_i)$ by reducing $B\{\hat{f}(x_i)\}$ or $B(\hat{x}_i)$. Examples 3 to 5 are in line with this, and each presents one of our three proposed devices.

Example 3. Imputation with Epanechnikov selection probabilities. It is easy to believe that giving donors close to the donee higher probabilities is better than using a uniform kernel function. This is the idea behind this example. Due to the optimality properties shown by the non-negative Epanechnikov function in Kernel estimation, we suggest to use it here. In Example 2, donee 3 had *k*=4 donors, with $B(\hat{x}_3) \approx -0.013$ and $E(\hat{y}_3) = -0.176$. With Epanechnikov probabilities $\lambda_{3j}^{Ep} = K_h^{Ep}(\tilde{x}_{3j})$ from (15), the closer (furthest) donor is more (less) likely to donate. With h_3=0.3715, Units 1, 2, 4 and 5 are assigned probabilities 0.238, 0.252, 0.260 and 0.250, so $B(\hat{x}_3) \approx -0.010$ and $E(\hat{y}_3) = -0.170$. Suppose that we draw Unit 4. If h_6=0.417 units 3, 4, 5 and 7 will get the probabilities λ_{6j}^{Ep} at 0.125, 0.274, 0.304 and 0.297, so that $B(\hat{x}_6) = -0.113$ and $E(\hat{y}_6) = 0.068$, compared to $B(\hat{x}_6) \approx -0.150$ and $E(\hat{y}_6) \approx 0.052$ in Example 2.

Given a symmetric kernel function the expected bias of interior donees is zero, so we only expect a reduction of variance by the change from K^{Un} to K^{Ep}. But given the same bandwidth *h* (or *k*), we do expect some reduction of bias for boundary donees since we switch from K^{Un} to the parabolic shaped K^{Ep}.

Example 4 Imputation with adjusted selection probabilites. A technique which fully eliminates $B(\hat{x}_i)$ is to adjust the probabilities given by the kernel so that the expectation over the *x*-values equals the donee x_i. More technically we propose to replace λ_{ij} by λ_{ij}' as close as possible but such that $E(\hat{x}_i) = x_i$ holds. λ_{ij}' is easily found by Lagrange minimisation as the solution to

$$\min_{\Lambda_1, \Lambda_2, \lambda_{ij}, j=1,\ldots,k} \sum_{j=1}^{k} L\left(\lambda_{ij} - \lambda_{ij}'\right) + \Lambda_1 \left\{ \sum_{j=1}^{k} \lambda_{ij}' \left(\tilde{x}_{ij}\right) \right\} + \Lambda_2 \left(\sum_{j=1}^{k} \lambda_{ij} - 1 \right), \qquad (16)$$

where $L\left(\lambda_{ij} - \lambda_{ij}'\right)$ is a distance function and Λ_1 and Λ_2 are Lagrange multipliers.

For the data in Table 1 and using Euclidean distances we get $\lambda_{31}^{Ep}{}' \approx 0.217$, $\lambda_{32}^{Ep}{}' \approx 0.235$, $\lambda_{34}^{Ep}{}' \approx 0.277$ and $\lambda_{35}^{Ep}{}' \approx 0.272$, with $E(\hat{y}_3) \approx -0.143$. Assuming Unit 4 is drawn, we get $\lambda_{6j}^{Ep}{}' \approx 0.011$, $\lambda_{6j}^{Ep}{}' \approx 0.217$, $\lambda_{6j}^{Ep}{}' \approx 0.263$ and $\lambda_{6j}^{Ep}{}' \approx 0.508$, with $E(\hat{y}_6) \approx 0.036$. Both $B(\hat{x}_i)$ are zero.

By solving (16) it is possible to obtain λ_{ij}' that results in $B(\hat{x}_i) = 0$ for both interior and boundary donees, as long as there are possible donors at both sides of x_i. (Other restrictions, for example, deterministic situations, may also prohibit unbiased solutions). The proposed adjustment of selection probabilities resembles the use of approximate kernel regression weights in imputation (Aerts, Claeskens, Hens, and Molenberghs, 2002), or calibration of design weights (Deville and Särndal, 1992) but on a pointwise level.

Example 5. Imputation with fewer donors at the boundary. Problems occur at the boundaries since there may be none or only few possible donor *x*-values at one side of x_i. We suggest that the width of the kernel then should be decreased. With multidimensional *x* one could also use an oblong donor pool instead of a spherical (quadratic) one.

Consider only boundary Unit 6. Setting k=2 shrinks the bandwidth from h_6=0.417 to h_6=0.241, which results in selection probabilities $\lambda_{65}^{Ep} \approx 0.624$ and $\lambda_{67}^{Ep} \approx 0.376$ for donors 5 and 7, with $B(\hat{x}_6) \approx -0.053$ and $E(\hat{y}_6) \approx 0.181$, compared to $B(\hat{x}_6) \approx -0.150$ and $E(\hat{y}_6) \approx 0.052$ from Example 3. Applying the Lagrange adjustment in (16) results in $\lambda_{65}^{Ep}{}' \approx 0.508$ and $\lambda_{67}^{Ep}{}' \approx 0.492$, with $B(\hat{x}_6) = 0$ and $E(\hat{y}_6) \approx 0.128$.

The expected bias of boundary units is directly related to the bandwidth and the reduction of $|B(\hat{x}_6)|$ from shrinking k is in line with this. But this bias reduction is expected to come at the cost of higher $V(\hat{x}_6)$ since we use fewer possible donors.

4. SIMULATION STUDY

Here we use our suggested bias reduction methods from Subsection 3.3 in a design-based simulation study with simulated data, and compare with other imputation methods.

4.1. SETUP OF SIMULATION STUDY

We construct two related populations. First *N=1 600* values are simulated from a *Un(0,1)* distribution (u) and a standard normal distribution (e) using R (R Development Core Team, 2009). The populations are then constructed, one with a linear (*LI*) relationship ($x^{LI}=u-1/2; y^{LI}=u+e/7-1/2$) and one with a nonlinear (*NO*) relationship ($x^{LI}=u-1/2; y^{NO}=sin(u\pi)+e/7-2/\pi$). From each population we draw *1 000* samples of size n=*100*, *400* and *900*. In each sample we create 50 % nonresponse on $y_\cdot$, using the MAR mechanism $P(y_\cdot \text{ is observed}) \propto 1-u^{1/4}$.

Table 2. Bias correction in kernel imputation

ID for kernel imputation methods	*U*	*E*	*L*	*S*	*EL*	*ES*	*LS*	*ELS*
Epanechnikov selection probabilities	No	Yes	No	No	Yes	Yes	No	Yes
Lagrange adjustment of biased units	No	No	Yes	No	Yes	No	Yes	Yes
Shrinkage to k=$k^{5/6}$ at boundary	No	No	No	Yes	No	Yes	Yes	Yes

The missing data in the sample or the population were imputed by all combinations of the three bias correction methods: Epanechnikov (*E*) selection probabilities, Lagrange (*L*) adjustment, and shrinkage (*S*) of the donor pool for boundary biased units. The methods' initial letters are used for notation as displayed in Table 2. The *k* potential donors were found using Euclidian distance and a square root rule $k = q^{½}$, where *q* is the number of eligible (observed and imputed) donor units.

Mean estimates of $\bar{y}^{LI}$ and $\bar{y}^{NO}$ from our methods are compared to estimates based on complete data (*CD*) and complete cases (*CC*). Estimates $\hat{\hat{\bar{y}}}_n^{-}$ based on imputed samples are also compared to estimates from ten single imputation methods, *SI*$_i$ *i*=1,…,10, and thirteen multiple imputation methods, *MI*$_i$ *i*=1,…,13. Estimates $\hat{\hat{\bar{y}}}_N^{-}$ based on fully imputed populations are only compared to the *MI*$_i$ methods. All *MI*$_i$ and *SI*$_i$ methods are derived from the R-packages described in Appendix 1. Appendix 2 and 3 contain results for estimates of $\bar{y}^{LI}$ and $\bar{y}^{NO}$ with the comparison methods.

The *SI*$_i$ point and variance estimates $\hat{\bar{y}}_n^{-}$ and $\hat{V}\left(\hat{\bar{y}}_n^{-}\right)$ are calculated as in (6) and (9), while all multiple imputation estimates $\hat{\hat{\bar{y}}}_n^{-}$ and $\hat{\hat{\bar{y}}}_N^{-}$ are calculated as in (7) and (10), with variance estimates $\hat{V}\left(\hat{\hat{\bar{y}}}_{-}^{-}\right)$ given by (8) and (11). We used either *D*=5 or D=20 replicates for all multiple imputation methods. To simplify the description, we henceforth replace $\hat{\hat{\bar{y}}}_{-}^{-}$ by $\hat{\bar{y}}_{-}^{-}$. Empirical averages from simulations, with *M* representing *n* or *N*, are calculated as $\bar{G}_M^{-} = \frac{1}{1000}\sum_{g=1}^{1000} G_{g,M}^{-}$, where $G_{g,M}^{-}$ is some function based on the *g*:th data, such as a point estimate $\hat{\bar{y}}_{g,M}^{-}$, the empirical mean squared error $MSE\left(\hat{\bar{y}}_{g,M}^{-}\right) = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\bar{y}}_{gi,M}^{-} - \bar{y}^{-}\right)^2$, bias $B\left(\hat{\bar{y}}_{g,\mathrm{M}}^{-}\right) = \hat{\bar{y}}_{g,\mathrm{M}}^{-} - \bar{y}^{-}$ or variance $V\left(\hat{\bar{y}}_{g,M}^{-}\right) = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\bar{y}}_{gi,M}^{-} - \hat{\bar{y}}_{g,M}^{-}\right)^2$, the average estimated variance $\hat{V}\left(\hat{\bar{y}}_{g,\mathrm{M}}^{-}\right)$, or the average double sided confidence interval length $CIL = 2t_{(1-\alpha,df)}\left\{\hat{V}\left(\hat{\bar{y}}_{g,\mathrm{M}}^{-}\right)\right\}^{1/2}$ and coverage $CIC = I\left\{\left|\hat{\bar{y}}_{g,M}^{-}\right| \le CIL/2\right\}$. The significance level of the *t*-statistic is always set to α=0.05, and the degrees of freedom $v = (D+1)\left(1 + \frac{1}{D+1}\frac{\overline{W}_M}{B_M}\right)^2$ where $\overline{W}$ and *B* are the variances components of (8) as described in Subsection 2.3 (Rubin, 1987). We always multiply $\bar{G}_M^{-}$ by 100 (100^2) if $G_{g,M}^{-}$ is a first (second) moment function.

4.2 RESULTS FROM SIMULATION STUDY

Results for $\hat{\bar{y}}^{LI}$ ($\hat{\bar{y}}^{NO}$) are presented in Table 3 (4), and for comparison methods in Appendix 2 (3). We only show results for sample sizes 100 and 900, and 20 imputed datasets for multiple (including kernel) imputation. Results using *n*=400 ended up in between *n*=100 and *n*=900 with kernel imputation. This was mostly the case for multiple imputation comparison methods as well, except for bias (and sometimes for MSE dominated by bias) which tended to be highest with *n*=400. Comparing D=20 and D=5, most simulation results were up to 15 % lower for kernel imputation with *D*=20 compared to *D*=5. Confidence coverage was only slightly smaller, but interval lengths were down to 30 % shorter. Bias was rather unaffected by *D*, with $B\left(\hat{\bar{y}}_n^{LI}\right)$ as an exception which almost halved but from a low level. Results for multiple imputation comparison methods had the same tendencies, but were more mixed.

Table 3. Simulation results for estimates of $\bar{y}^{LI}$, including 95 % confidence intervals.

		Sample size n=100, nonresponse r=50						*Sample size n=900, nonresponse r=450*					
M	**ID**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**
n; sample imputed	*U*	14.4	0.69	13.9	11.9	93.2	14.1	0.85	0.17	0.82	0.79	95.5	3.7
	E	13.8	0.47	13.6	11.8	92.9	14.0	0.84	0.13	0.82	0.80	95.8	3.7
	L	14.0	0.60	13.6	12.0	92.9	14.1	0.84	0.15	0.81	0.85	96.3	3.8
	S	14.2	0.58	13.8	11.9	93.5	14.1	0.84	0.14	0.81	0.80	95.2	3.7
	EL	13.7	0.40	13.5	11.8	93.0	14.0	0.85	0.12	0.83	0.84	95.7	3.8
	ES	13.6	0.41	13.5	11.7	92.9	13.9	0.84	0.12	0.83	0.80	95.8	3.7
	LS	13.9	0.52	13.7	12.0	93.6	14.1	0.85	0.13	0.83	0.85	95.6	3.8
	ELS	13.6	0.36	13.5	11.8	93.6	14.0	0.84	0.11	0.83	0.84	95.6	3.8
N; population imptued	*U*	6.2	0.78	5.6	4.4	91.0	8.6	0.44	0.15	0.42	0.40	93.6	2.6
	E	6.0	0.58	5.7	4.0	88.9	8.2	0.45	0.13	0.43	0.40	93.7	2.6
	L	6.1	0.66	5.7	4.4	90.0	8.6	0.48	0.17	0.45	0.45	93.7	2.8
	S	6.1	0.70	5.6	4.3	90.0	8.5	0.44	0.13	0.42	0.39	94.9	2.6
	EL	6.1	0.52	5.8	4.0	89.5	8.2	0.47	0.13	0.45	0.43	93.7	2.7
	ES	5.9	0.51	5.7	3.7	88.6	7.9	0.44	0.12	0.43	0.41	94.4	2.6
	LS	6.1	0.59	5.8	4.2	88.6	8.4	0.48	0.15	0.46	0.45	94.0	2.8
	ELS	6.0	0.45	5.8	3.9	89.4	8.0	0.47	0.12	0.45	0.44	94.1	2.7

With the sample imputed in Table 3, bias decreased with increased sample size and added bias corrections (*E*, *S* or *L*). Variance dominated mean squared error, and seemed to decrease slightly with bias corrections and *n*=100. Average estimated variance was below the true value for *n*=100 and 400, but the underestimation was ameliorated by the added bias correction and it almost disappeared for *n*=900. Confidence interval coverage (CIC) was slightly below the stated 95 % for *n*=100 and 400, but slightly above for *n*=900. Confidence interval lengths (CIL) decreased with sample size. Patterns were similar for the whole population imputed but all figures were lower. An exception is $B\left(\hat{\bar{y}}_{\mathrm{n}}^{U}\right)$, which was smaller than $B\left(\hat{\bar{y}}_{\mathrm{N}}^{U}\right)$, but became more alike with increased sample size.

Single imputation methods (in Appendix 2) had similar or slightly better MSE compared to *ELS*, except SI_3- SI_6 which also had large bias. They always underestimated variance, and interval coverage decreased with sample size. Many multiple imputation methods behaved as well or somewhat better than *ELS*. Exceptions were MI_5 and MI_{13} (and mostly MI_{12}) with underestimated variance and poor coverage. MI_{13} also had huge bias. With the whole sample imputed MI_6 also underestimated variance severely, and MI_9 and MI_{10} had extremely large bias for *n*=100.

Table 4. Simulation results for estimates of $\bar{y}^{NO}$, including 95 % confidence intervals.

		Sample size n=100, nonresponse r=50						*Sample size n=900, nonresponse r=450*					
M	**ID**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**
n; sample imputed	*U*	20.9	2.29	15.6	13.6	89.2	15.1	1.30	0.68	0.84	0.86	91.3	3.8
	E	17.6	1.65	14.9	13.0	92.1	14.8	1.12	0.53	0.84	0.87	93.4	3.8
	L	18.1	1.83	14.7	14.0	91.6	15.3	1.14	0.55	0.84	0.94	94.4	4.0
	S	19.0	1.90	15.3	13.4	90.8	15.0	1.18	0.58	0.84	0.86	92.8	3.8
	EL	16.3	1.37	14.4	13.3	92.3	14.9	1.05	0.45	0.85	0.92	94.8	4.0
	ES	16.8	1.40	14.8	12.9	91.5	14.7	1.05	0.46	0.84	0.87	93.6	3.8
	LS	17.2	1.60	14.6	13.6	92.1	15.1	1.07	0.49	0.84	0.93	94.3	4.0
	ELS	15.9	1.23	14.4	13.1	92.9	14.8	1.00	0.40	0.83	0.91	95.2	3.9
N; population imptued	*U*	14.0	2.45	8.0	6.5	83.1	10.4	0.88	0.65	0.46	0.43	82.6	2.7
	E	10.4	1.81	7.1	5.4	84.8	9.5	0.73	0.51	0.46	0.43	86.6	2.7
	L	11.2	1.96	7.4	6.6	87.3	10.5	0.77	0.54	0.48	0.49	87.0	2.9
	S	11.5	1.97	7.6	6.0	86.2	10.0	0.76	0.55	0.46	0.42	86.7	2.7
	EL	9.4	1.54	7.0	5.4	86.6	9.5	0.66	0.43	0.47	0.46	90.3	2.8
	ES	9.0	1.48	6.8	4.9	86.3	9.0	0.66	0.46	0.45	0.42	88.8	2.7
	LS	10.1	1.66	7.4	6.0	87.4	10.0	0.70	0.47	0.48	0.48	89.5	2.9
	ELS	8.5	1.30	6.8	5.0	87.2	9.1	0.62	0.38	0.47	0.46	90.9	2.8

In Table 4, both $MSE\left(\hat{\bar{y}}_{n}^{NO}\right)$and $B\left(\hat{\bar{y}}_{n}^{NO}\right)$ decreased in all cases with added bias correction and increasing sample size when the sample was imputed.

Variance fell with sample size and somewhat with bias corrections for n=100. The underestimation of variance lessened with sample size, and $\hat{v}\left(\hat{\bar{y}}_{\mathrm{n}}^{NO}\right)$ was even somewhat higher then $V\left(\hat{\bar{y}}_{n}^{NO}\right)$ with n=900. Confidence interval coverage increased with sample size and added bias corrections, but was always below the stated 95 % except for *ELS* with n=900. Confidence interval lengths decreased with sample size. The patterns were similar when the whole population was imputed, but all figures were lower except for bias, which was somewhat higher with n=100, about the same with n=400, and slightly lower with n=900.

With only the sample imputed, nearest neighbour methods SI_7- SI_{10} and predictive mean matching methods MI_5-MI_6 in Appendix 3 had MSE similar to *ELS*, but with lower bias and higher variance. Their underestimation of variance also increased with sample size, with worsening confidence interval coverage. With the whole population imputed, MI_5-MI_6 gave small or zero estimates of variance. Method MI_{12} gave better coverage rate than *ELS*, both with the sample and population imputed, but overestimated the high variance severely and gave very wide confidence intervals. All other methods had much larger MSE than *ELS*, due to larger bias or variance. Several methods that rely on regression models had MSE similar to complete cases, with bias dominating the MSE.

5. CONCLUSIONS

Our proposed imputation method for missing value of a study variable assumes a relationship to a fully observed continuous auxiliary variable. Common to other methods based on nonparametric models, our method relies on having observed the data dispersion, which is more probable with larger samples. The noninformative Bayesian approach with Pólya urn sampling only using the sample as a prior and with multiple imputation can effectively address uncertainty with minimal assumptions. Given a missing at random mechanism, the real donor approach with imputed values selected among already observed (and thus presumably realistic) values, can also effectively remove nonresponse bias even with nonlinearities in the data. The use of kernel methods addresses the bias caused by having sparse and bounded finite sample data.

As expected, the simulation study with linear data demonstrated a small loss of efficiency compared to methods utilizing parametric assumptions, but with the nonlinear data the improvement by bias corrections was relatively larger, and comparison methods were generally outperformed. In both cases, our three suggested devices (Epanechnikov kernel, Lagrange adjustment, and shrinkage at the boundary) always reduced bias. Properties seemed to improve with increasing the sample size, which agrees with the nonparametric reliance on the sample size. Many of the multiple imputation comparison methods managed to give at least 95 % coverage with linear data, which kernel imputation only did for the largest sample imputed. However, except for one extremely inefficient comparison method, kernel imputation with all bias corrections and the largest sample was the only method which reached 95 % coverage with the nonlinear data. Since the response probabilities were strongly related to the study variable through the auxiliary, imputation methods with linear parametric assumptions displayed bias (and hence MSE) sometimes even larger than for complete cases when imputing the nonlinear data.

Variance (and hence MSE) went down when the whole population was imputed instead of just the sample. The effect is similar to what would have been expected from applying (post-) stratification weights based on the auxiliary. Since the bias share of MSE increased when sample was imputed the confidence interval coverage rates fell. A similar but weaker effect was seen when the number of imputed datasets was increased.

Several extensions of the proposed method could be explored, including multivariate auxiliary and study variables, use of more or other prior information, estimators other than means, alternative distance metrics, more elaborate ways of choosing the number of donors, including the degree of shrinkage, or other aspects related to boundary donees.

REFERENCES

Aerts, M. Claeskens, G. Hens, N. and Molenberghs G., (2002). Local multiple imputation. *Biometrika*, 89(2), pp.375-388.

Andridge, R.R. and Little, R.J.A., (2010). A review of Hot Deck Imputation for Survey Non-response. *International Statistical Review,* 78(1), pp.40-64.

van Buuren, S. and Groothuis-Oudshoorn, K., (2010). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software,* (in press).

Conti, P.L. Marella, D. and Scanu, M., (2008). Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. *Computational Statistics and Data Analysis*, 53(2), pp.354-365.

Deville, J-C. and Särndal, C-E., (1992), Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association,* 87(418), pp.376-382.

Diaconis, P. and Freedman, D., (1980). Finite exchangeable sequences. *Annals of Probability,* 8(4), pp.745-764.

Epanechnikov, V.A., (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14(1), pp. 153–158.

Feller, W., (1971). An Introduction to Probability Theory and Its Applications, 2nd ed. Wiley, New York.

Ferguson, T.S., (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics,* 1(2), pp. 209-230.

de Finetti, B., (1931). Funzione caratteristica di un fenomeno aleatorio, *Atti della R. Academia Nazionale dei Lincei, Classe di Scienze Fisiche, Mathematice e Naturale,* 6(4), pp.251-299.

Gelman, A. Hill, J. Su, Y-S. Yajima, M. and Pittau, M.G., (2010). mi: Missing Data Imputation and Model Checking. R package version 0.09-11.

Gramacy, R.B., (2010). monomvn: Estimation for multivariate normal and Student-t data with monotone missingness. R package version 1.8-3.

Gross, K. and Bates, D., (2008). mvnmle: ML estimation for multivariate normal data with missing values. R package version 0.1-8.

Harrell, F.E., (2010). Hmisc: Harrell Miscellaneous. R package version 3.8-3.

Hewitt, E. and Savage, L.J., (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society,* 80(2), pp.470-501.

Hoff, P., (2010). sbgcop: Semiparametric Bayesian Gaussian copula estimation and imputation. R package version 0.975.

Honaker, J. King, G. and Blackwell, M., (2011) Amelia: Amelia II: A Program for Missing Data. R package version 1.5-4.

Kim, K-Y. and Yi, G-S., (2008). SeqKnn: Sequential KNN imputation method. R package version 1.0.1.

Kong, A. Liu, J.S. and Wong, W.H., (1994) Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American statistical association,* 89(425), pp.278-288.

Laaksonen, S. (2000). Regression-based nearest neighbour hot decking, *Computational Statistics,* 15(1), pp. 65–71.

Little R.J.A. and Rubin, D.B., (2002). *Statistical analysis with missing data*. Hoboken: Wiley.

Lo, A.Y., (1988). A Bayesian bootstrap for a finite population. *The Annals of Statistics*, 16(4), pp.1684-1695.

Nadaraya, E.A., (1964). On estimating regression. *Theory of Probability and its Applications*, 9(1), pp.141-142.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rice, J., (1984). Boundary modification for kernel regression. *Communications in statstistcs- Theory and methods*, 13(7), pp.893-900.

Rubin, D.B., (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9(1), pp.130–134.

Rubin, D.B., (1987). *Multiple imputation for nonresponse in surveys*. Hoboken; Wiley.

Schafer, J.L., (1997). *Analysis of incomplete multivariate data*. London; Chapman and Hall.

Siddique, J. and Belin, T.R., (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine,* 27(1), pp.83-102.

Silverman, B.W., (1986). *Density estimation for statistics and data analysis*. London; Chapman and Hall.

Simonoff, J.S., (1996). *Smoothing methods in statistics*. New York; Springer-Verlag.

Stacklies, W. Redestig, H. and Wright, K., (2011). pcaMethods: A collection of PCA methods. R package version 1.24.0.

Templ, M. Hron, K. and Filzmoser, P., (2010). robCompositions: Robust Estimation for Compositional Data. R package version 1.4.3.

Wand, M.P. and Jones, M.C., (1995). *Kernel smoothing*. London; Chapman and Hall.

Watson, G.S., (1964). Smooth regression analysis. *Sankhya Series A*, 26(4), pp.359-372.

APPENDIX 1. R PACKAGES AND CODE FOR ALTERNATIVE ESTIMATORS

R-Package	ID	R-code
monomvn. Gramacy (2010)	SI_1	monomvn(data)
mvnmle. Gross (2008)	SI_2	mlest(data)
pcaMethods.* Stacklies, Redestig and Wright (2011)	SI_3	llsImpute(data,k=1,center=T,correlation="pearson",verbose=F,allVariables=T)
	SI_4	pca(data,method="nipals")
	SI_5	pca(data,method="ppca")
	SI_6	pca(data,method="svdImpute")
robCompositions. Templ, Hron and Filzmoser (2010)	SI_7	impKNNa(data,k=1,metric="Euclidean",agg="median",primitive=T)
	SI_8	impKNNa(data,k=5,metric="Euclidean",agg="median",primitive=T)
SeqKnn. Kim and Yi (2008)	SI_9	SeqKNN(data,k=1)
	SI_{10}	SeqKNN(data,k=5)
Amelia. Honaker, King and Blackwell (2011)	MI_1	amelia(data,m = D)
Hmisc. Harrell (2010)	MI_2	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='closest',nk=0,curtail=T,boot.method="approximate bayesian")
	MI_3	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='closest',nk=0,curtail=F,boot.method="approximate bayesian")
	MI_4	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='weighted',nk=0,curtail=T,boot.method="approximate bayesian")
	MI_5	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='pmm',match='closest',nk=0,curtail=T,boot.method="approximate bayesian")
	MI_6	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='pmm',match='weighted',nk=0,curtail=T,boot.method="approximate bayesian")
	MI_7	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='closest',nk=c(0,3:5),B=10,curtail=T,boot.method= "approximate bayesian")
	MI_8	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='closest',nk=c(0,3:5),B=10,tlinear=F,curtail=T,boot.method="approximate bayesian")
mi. Gelman (2010)	MI_9	mi(data.frame(data),n.imp=D,add.noise=noise.control(method="reshuffling",K=1,post.run.iter=20),n.iter=30)
	MI_{10}	mi(data.frame(data),n.imp=D,add.noise=noise.control(method="fading",pct.aug=10,post.run.iter=20),n.iter=30)
mice. van Buuren and Groothuis-Oudshoorn (2010)	MI_{11}	mice(data,m=D,method="norm")
	MI_{12}	mice(data,m=D,method="pmm")
sbgcop. Hoff (2010)	MI_{13}	sbgcop.mcmc(data,nsamp=D)

R-packages for single (SI) and multiple (MI) imputation methods are available at *http://cran.r-project.org/web/packages/* and (*) *http://www.biocondoctor.org/biocLite.R.*

The object ‘data’ is created as ‘data <- cbind(x,y)’ in R, where ‘x’ is the fully observed auxiliary variable vector, and ‘y’ is the partly observed study variable vector. Object ‘D’ is the number of imputed datasets.

APPENDIX 2. SIMULATION RESULTS, ALTERNATIVE $\bar{y}^{LI}$ -ESTIMATORS

		Sample size n=100, nonresponse r=50						*Sample size n=900, nonresponse r=450*					
M	**ID**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**
n	*CD*	9.7	-.03	9.7	10.0	94.2	12.4	0.49	-.01	0.49	0.51	95.1	2.8
	CC	32.7	3.5	20.4	10.7	70.2	12.8	13.4	3.43	1.67	0.87	11.6	3.7
n; sample imputed	*SI_1*	12.1	-.15	12.1	10.1	92.3	12.4	0.73	-.02	0.73	0.51	90.8	2.8
	SI_2	12.1	-.15	12.1	9.9	92.3	12.3	0.73	-.02	0.73	0.51	90.8	2.8
	SI_3	20.9	2.77	13.2	8.7	79.6	11.6	11.3	3.24	0.85	0.45	1.8	2.6
	SI_4	39.5	4.81	16.4	5.0	42.6	8.7	31.6	5.51	1.18	0.25	0.0	1.9
	SI_5	23.7	3.09	14.2	8.1	73.2	11.1	37.0	5.98	1.30	0.23	0.0	1.9
	SI_6	51.2	5.72	18.4	4.6	32.4	8.3	44.3	6.55	1.40	0.23	0.0	1.9
	SI_7	14.1	-.06	14.1	9.9	89.1	12.3	1.21	-.02	1.21	0.51	78.4	2.8
	SI_8	13.3	0.08	13.3	9.2	88.7	11.8	0.92	-.06	0.92	0.48	83.8	2.7
	SI_9	14.3	-.03	14.3	9.9	88.6	12.3	1.24	-.03	1.24	0.51	78.5	2.8
	SI_{10}	13.3	0.01	13.3	9.4	89.4	12.0	0.95	-.01	0.95	0.49	83.5	2.7
	MI_1	12.5	-.54	12.2	12.4	95.4	14.4	0.75	0.11	0.73	0.90	97.9	3.9
	MI_2	12.3	0.04	12.3	12.2	95.7	14.3	0.74	0.03	0.73	0.82	96.8	3.7
	MI_3	12.2	-.14	12.1	12.8	95.5	14.7	0.75	-.03	0.75	0.85	97.4	3.8
	MI_4	12.2	0.05	12.2	12.2	95.2	14.3	0.73	0.03	0.73	0.82	97.3	3.7
	MI_5	14.1	-.06	14.1	9.9	89.3	12.3	1.20	-.02	1.20	0.51	80.3	2.9
	MI_6	12.9	0.11	12.9	10.6	93.0	13.3	0.80	0.25	0.73	0.63	93.1	3.2
	MI_7	12.3	0.04	12.3	12.1	94.9	14.3	0.74	0.03	0.74	0.82	96.7	3.7
	MI_8	12.2	0.06	12.2	12.2	95.3	14.3	0.74	0.03	0.74	0.82	97.0	3.7
	MI_9	12.1	-.03	12.1	13.2	96.1	14.9	0.74	-.10	0.73	0.87	97.4	3.8
	MI_{10}	12.4	-.25	12.3	12.6	95.0	14.5	0.75	-.03	0.75	0.86	97.3	3.8
	MI_{11}	12.2	-.26	12.2	12.9	96.0	14.7	0.73	0.07	0.73	0.92	98.1	4.0
	MI_{12}	12.7	0.19	12.7	12.2	94.9	14.3	1.12	0.09	1.11	0.76	90.4	3.6
	MI_{13}	28.5	3.71	14.7	12.1	81.7	14.3	14.5	3.67	0.99	0.70	2.4	3.4
N; population imptued	*MI_1*	5.3	0.12	5.3	4.9	94.1	9.0	0.39	-.05	0.39	0.36	93.8	2.5
	MI_2	5.4	0.44	5.2	8.7	98.6	12.1	0.41	0.03	0.41	0.68	98.9	3.4
	MI_3	5.2	0.11	5.2	10.7	98.8	13.4	0.40	-.01	0.39	0.71	98.9	3.5
	MI_4	5.7	0.46	5.5	9.5	98.2	12.7	0.41	0.03	0.41	0.69	98.2	3.4
	MI_5	7.5	0.29	7.4	0	0	0	0.85	-.06	0.84	0.00	6.5	0.2
	MI_6	6.9	0.62	6.5	0.1	21.8	1.3	0.45	0.24	0.39	0.10	65.8	1.3
	MI_7	5.6	0.46	5.4	9.0	98.4	12.3	0.39	0.04	0.39	0.69	99.5	3.4
	MI_8	5.3	0.45	5.1	8.6	97.9	12.0	0.40	0.03	0.40	0.69	99.0	3.4
	MI_9	52.8	6.01	16.7	14.4	68.5	15.9	0.39	0.12	0.38	0.49	96.8	2.9
	MI_{10}	65.9	7.03	16.5	8.3	38.1	12.0	0.39	0.08	0.38	0.46	97.1	2.8
	MI_{11}	5.5	0.82	4.9	4.2	93.0	8.5	0.37	0.00	0.37	0.60	99.3	3.2
	MI_{12}	6.3	0.56	6.0	4.3	90.0	8.6	0.76	0.27	0.68	0.22	71.0	1.9
	MI_{13}	59.0	6.29	19.5	0.5	9.5	3.1	26.3	5.02	1.10	0.25	0.0	2.1

Estimators are based on complete data (*CD*), complete cases (*CC*), multiply imputed (*MI*) and singly imputed (*SI*) datasets. Confidence interval coverage (CIC) and length (CIL) are from double-sided intervals with 5 % significance level.

APPENDIX 3. SIMULATION RESULTS, ALTERNATIVE $\bar{y}^{NO}$ -ESTIMATORS

		Sample size n=100. nonresponse r=50						*Sample size n=900. nonresponse r=450*					
M	**ID**	***MSE***	***B***	***V***	***V^***	***CIC***	***CIL***	***MSE***	***B***	***V***	***V^***	***CIC***	***CIL***
n	*CD*	10.6	0.05	10.6	11.2	95.4	13.1	0.6	0.02	0.55	0.57	95.8	3.0
	CC	63.9	6.6	19.9	9.2	42.8	11.8	44.0	6.51	1.55	0.75	0.0	3.4
n; sample imputed	SI_1	40.3	4.4	21.1	10.7	65.8	12.8	40.2	6.20	1.75	0.54	0.0	2.9
	SI_2	40.3	4.4	21.1	10.5	65.3	12.6	40.2	6.20	1.75	0.54	0.0	2.9
	SI_3	24.7	-.5	24.5	9.4	76.1	12.0	2.1	0.35	2.00	0.48	65.0	2.7
	SI_4	29.1	3.1	19.3	5.5	57.4	9.1	22.8	4.61	1.49	0.28	0.4	2.1
	SI_5	26.0	2.4	20.0	5.2	60.6	8.9	25.0	4.83	1.58	0.29	0.2	2.1
	SI_6	25.9	2.4	20.1	5.2	60.7	8.9	15.0	3.66	1.56	0.26	2.1	2.0
	SI_7	15.2	0.4	15.1	10.7	89.3	12.8	1.4	0.16	1.39	0.57	79.2	2.9
	SI_8	14.9	0.8	14.2	9.7	87.5	12.2	1.1	0.19	1.09	0.53	83.1	2.9
	SI_9	15.5	0.4	15.4	10.7	88.2	12.8	1.5	0.15	1.43	0.57	78.3	2.9
	SI_{10}	14.8	0.6	14.4	10.1	88.8	12.4	1.2	0.20	1.13	0.54	83.4	2.9
	MI_1	36.3	3.8	22.1	23.0	89.4	19.7	44.3	6.52	1.77	2.38	1.4	6.4
	MI_2	44.1	4.8	21.1	24.7	85.7	20.5	40.8	6.24	1.88	2.28	2.0	6.2
	MI_3	46.9	5.0	21.7	27.6	87.0	21.5	41.1	6.26	1.84	2.31	1.6	6.3
	MI_4	43.8	4.7	21.5	24.7	86.5	20.4	40.7	6.23	1.82	2.27	1.9	6.2
	MI_5	15.2	0.4	15.1	10.7	89.3	12.8	1.4	0.16	1.39	0.57	79.6	3.0
	MI_6	13.9	0.7	13.4	11.5	93.1	13.8	1.7	0.95	0.84	0.68	79.9	3.4
	MI_7	43.7	4.8	21.0	25.3	86.7	20.7	40.4	6.21	1.84	2.27	1.6	6.2
	MI_8	43.5	4.8	20.9	25.0	86.3	20.6	40.5	6.22	1.83	2.29	2.0	6.3
	MI_9	42.7	4.6	21.7	22.3	85.4	19.5	39.9	6.17	1.84	1.90	1.1	5.7
	MI_{10}	38.8	4.0	22.8	21.8	86.6	19.2	40.4	6.22	1.76	2.03	1.5	5.9
	MI_{11}	38.3	4.1	21.1	23.4	88.1	20.0	42.8	6.40	1.75	2.31	1.3	6.3
	MI_{12}	59.5	-2.2	54.6	74.6	93.7	35.1	2.1	0.85	1.37	18.6	100	17.6
	MI_{13}	36.6	4.2	19.0	14.5	79.4	15.7	28.2	5.17	1.49	1.00	1.0	4.1
N; population imputed	MI_1	97.9	8.5	26.3	24.2	61.0	20.1	36.9	5.91	1.95	1.93	1.7	5.7
	MI_2	92.6	8.3	23.1	42.0	80.1	26.6	37.5	5.95	2.03	3.57	9.2	7.8
	MI_3	105	8.9	25.9	53.9	84.4	30.1	38.1	6.00	2.05	3.65	8.8	7.9
	MI_4	98.6	8.7	23.1	45.5	81.0	27.7	37.6	5.97	2.00	3.58	8.4	7.8
	MI_5	9.1	1.0	8.0	0.0	0	0	0.8	0.06	0.84	0.00	7.8	0.2
	MI_6	9.8	1.6	7.2	0.1	19.6	1.4	1.3	0.93	0.41	0.12	36.5	1.4
	MI_7	94.2	8.4	23.4	44.5	79.8	27.4	37.5	5.96	2.01	3.55	8.1	7.8
	MI_8	95.3	8.5	22.9	42.2	79.8	26.6	37.8	5.97	2.07	3.61	9.7	7.9
	MI_9	45.3	5.1	19.2	18.0	79.6	17.7	35.2	5.77	1.89	2.20	3.5	6.1
	MI_{10}	58.7	6.3	18.9	10.7	55.3	13.7	34.2	5.68	2.01	1.83	2.9	5.6
	MI_{11}	110	9.3	23.5	17.3	46.3	17.4	37.3	5.95	1.88	2.66	3.6	6.8
	MI_{12}	75.4	-1.3	73.6	146	96.9	47.9	2.3	-.90	1.49	24.7	100	20.3
	MI_{13}	38.6	4.3	19.9	0.7	20.9	3.4	18.4	4.10	1.54	0.53	2.3	3.0

Estimators are based on complete data (*CD*), complete cases (*CC*), multiply imputed (*MI*) and singly imputed (*SI*) datasets. Confidence interval coverage (CIC) and length (CIL) are from double-sided intervals with 5 % significance level.