Informed kernel imputation

Nicklas Pettersson[1]

[1]Stockholm University, e-mail: nicklas.pettersson@stat.su.se

Abstract

Some multiple real donor imputation methods are proposed which maintain the structure of a dataset with nonresponse. They utilize unit- and aggregate-level information which may be known for the full population, for a subpopulation, for the sample or only for the response set. Only imputed datasets which satisfy the constraints derived from the aggregate information are accepted. The imbalances of real donor unit pools is offset by letting the distribution of the imputed values approximately equal the distribution of the true values given what is known. The methods are intended for a general use e.g. for multi-purpose surveys and surveys with various types of estimands on different levels, and are exemplified by simulations with a realistic dataset.
Keywords: Missing data; Real donor multiple imputation; Boundary bias, Multiple imputation.

1 Introduction

Missing data due to e.g. item and unit nonresponse are usually unavoidable when conducting a survey. Good estimation methods should be flexible enough to counteract this by taking all relevant and accessible information into account. Methods which are able to fit most estimation techniques are also desirable. Many methods are not sufficiently general to handle item and unit nonresponse simultaneously and may be tedious to customize in order to fit complex situations while taking various forms of background information into account. We will explore whether kernel imputation (Pettersson, 2012; 2013) can be used for this purpose.

Kernel imputation has been used for missing data imputation when estimating a population mean from a simple random sample where the expectation of the study variable is a continuous function of the auxiliary variables. By making use of ideas in Nelson and Meeden (1998) and Meeden (2003) we will explore how kernel imputation can be modified to:

- constrain the imputed data to known (sub)population aggregates of auxiliary (or study) variables, and
- utilize data from other similar studies.

When such statistical information, which does not originate from the data itself, is utilized we denote our method as informed kernel imputation. In addition we will consider:

- estimates of regression coefficients in addition to mean estimates. With several values missing for the same unit we will use a common donor approach,
- applications to a realistic dataset. Previous simulations (Pettersson, 2012; 2013) examining kernel imputation features were undertaken on simulated data,
- other sampling designs in addition to simple random sampling, e.g. with varying selection probabilities,
- different kinds of nonresponse causes, and
- informative designs and the probability (or propensity) to obtain response.

In order to analyze a dataset with missing data one always makes assumptions on the missing data mechanism. Delimiting an analysis to the complete case (CC) implies the strong assumption that that the values of incomplete observed units are missing completely at random (MCAR). In the absense of further knowledge, assuming a missing at random (MAR) mechanism, stating that the reason for missing can be explained by the observed data, is often a good compromise between realism and simplicity. Examples of explicit modelling and analyzes through maximum likelihood or regression imputation are covered by Little and Rubin (2002), and calibration,

which implicitly entails a form of MAR and linear models, by Lundström and Särndal (2007). Andridge and Little (2010) give an overview of hot deck imputation, which also implies some form of MAR assumption.

Kernel imputation is primarily intended for MAR mechanisms. On simulated data the method has been shown to perform almost as well as competing methods when imputing a study variable which is a linear function of the continuous auxiliaries, but usually much better when this functional form is nonlinear (Pettersson, 2012; 2013). It is to a great extent nonparametric and employs a real donor approach (Laaksonen, 2000) which has the attractive property that the imputed values are copies of actually observed realistic values. It reduces error by resolving the real donor weakness that the set (or pool) of potential donor values often is imbalanced, through adaption of features from kernel estimation. The inherent dependence on the number of donors or size of donor pools in kernel estimation is also reduced. The method handles both item and unit nonresponse and other forms of missing data. To be able to estimate the imputation random error it uses a multiple response technique. Each imputed dataset is analyzed as if it were complete, but inference is drawn under a multiple imputation combination rule (Rubin, 1987). To preserve relationships (both with observed values and among missing values) the imputations are conditional on observed variables and may be multivariate in a common donor manner.

In Section 2 we describe how the kernel imputation method of Pettersson (2012; 2013) is used to construct balanced real donor pools such that the distribution of the imputed values approximately equals the distribution of the true values. We then present the methods which we use to utilize different kinds of statistical information in combination with imputation of nonresponse in Section 3 and discuss theoretical considerations. Simulation based on a realistic dataset in order to exemplify kernel imputation with utilization of various statistical information is set up in Section 4 and the results are presented in Section 5. The paper is then concluded in Section 6.

2 Kernel imputation

2.1 Notation

In a population U with N units each unit i is characterized by three related variables X_i, Y_i and Z_i, which may all be vectorvalued. Our interest is in estimating an arbitrary population estimand $T(Z_i, X_i, Y_i, i \in U)$. Variable Z_i is an auxiliary which is assumed to be known for all units in the population and may e.g. include both register and design variables. Given a sample S with n units, X_i is fully observed except for unit nonrespondents, while Y_i is exposed both to item and unit nonresponse. For notational convenience we ignore unit nonresponse unless otherwise indicated. A response indicator R_i is defined for $i \in S$, such that $R_i = 1$ if Y_i is observed and $R_i = 0$ if Y_i is missing.

In kernel imputation each unit i with missing (or unobserved) values is matched to a donor pool $S_{i,k}$ containing the k_i nearest available neighbour donor units with complete data, where nearness is measured by a distance function on (X_i, Z_i), e.g. $d_{i,j} = \frac{1}{h_i} \begin{pmatrix} x_j - x_i \\ z_j - z_i \end{pmatrix} G_i^{-1}(x_j z_j - x_i z_i)$ which is a quadratic distance measure where h_i is a scalar and G_i is a non-negative matrix.

A probability measure $\lambda_i = (\lambda_{i,1}, ..., \lambda_{i,k_i})$ is assigned to each pool $S_{i,k}$ where $\lambda_{i,j}$ are the donor selection probabilities of the k_i units. The missing value y_i is imputed by randomly drawing donor j with probability λ_i and imputing its value y_j. The expected value of the imputed unit corresponds to a pointwise kernel smoother $\sum_j \lambda_{i,j} Y_j = \widetilde{y}_i$ (Härdle, 1990).

2.2 Reduction of individual bias

A disadvantage of using real donors is that the matching between auxiliary values of a donee i and its pool is biased, in the sense that $(x_i, z_i) \neq \sum_{j \in S_{i,k}} \lambda_{i,j}(x_j, z_j) = E(x_j, z_j) = (\widehat{x}_i, \widehat{z}_i)$. The individual bias is accentuated when (x_i, z_i) lies close to the boundary (i.e. convex hull) of the possible donors' auxiliary values, i.e. when most values lies at one side of (x_i, z_i). Such donor pools might introduce a bias in the imputed study value $\widetilde{y}_i$.

By reducing the individual bias in $(\widehat{x}_i, \widehat{z}_i)$, the bias in $\widetilde{y}_i$ will hopefully be reduced, and thereby also the possible bias in $T(x, y, z)$. Several methods are given in Pettersson (2012, 2013). An optimal choice of λ_i which approx-

imately minimizes MSE in kernel estimation is by letting them be decided by an Epanechnikov (E) kernel (Silverman, 1986). The selection probabilities may often be calibrated (L) to eliminate most of the bias in $(\widehat{x}_i, \widehat{z}_i)$. If calibration is not feasible for some donees at the boundary of (x_i, z_i), one may at least attempt to reduce it by shrinking (S) the size h_i parameter so that the furthest donors are removed, or by reorienting (R) G_i so that the donors which contribute most to the bias are substituted by donors which contributes to the bias to a lesser extent. A more thorough description of these methods can be found in Pettersson (2013).

3 Informed imputation

3.1 Stepwise using previously imputed units

Given a sample and a correctly specified structure of the full data model a direct estimate would be both quick and efficient. But in realistic situations robustness to model misspecification is desirable. Less immediate estimators such as those derived from stepwise Pólya sampling techniques may then be preferable.

In the simplest case with no auxiliary variable (x or z) and a sample of n exchangeable units, e.g. a simple random sample, the basic idea is to 1) draw a single donor unit at random from the response set, 2) duplicate the drawn donor, and 3) replace the subsequent unit to be imputed in the response set by the duplicated unit. This procedure is repeated until all missing values in the sample have been imputed. Each donee being imputed utilizes not only the complete cases but also the previously imputed units. The wider distribution of the estimates from such imputed datasets reflects the imputation variance as it appears in the multiple imputation combination rule (Rubin, 1987).

The imputation process can be extended to impute the remaining $N - n$ units in the population. Estimates from such imputed sample or population datasets may be treated as draws from a Pólya posterior. The Pólya posterior estimates have been shown to possess several desirable (frequentist and Bayesian) properties (Ghosh and Meeden, 1997).

With several related variables either completely or partially observed, it is more reasonable not to treat the missing values as directly exchangeable. In that case a modification of the Pólya sampling procedure will probably

become more efficient if donees and their donors are conditionally/almost exchangeable. Kernel imputation is such a modified Pólya sampling. The selection probabilities in step 1 must then be increased for units which are better matched on the observed variables among the fully observed (or previously imputed) units and the donee unit to be imputed. In what follows we refer to this imputation method as "modified Pólya sampling".

3.2 Conditioning on the design

Imputation should make use of auxiliary variables which are predictive of the missing values (Little, 1988) and the propensity to respond (Rosenbaum and Rubin, 1983). A general imputation method should therefore be able to appropriately incorporate features of an informative design, e.g. unequal probabilities, πps-sampling, stratification, clustering and multi-stage sampling.

We suggest that inclusion probabilities and other design variables (e.g. stratum indicators, cluster units, or size measurements) therefore can be treated like other auxiliary variables and may be used in the construction of donors pools by including them in the distance metric and for finding good donor selection probabilities λ_{ij}. Nonlinear design variables may improve the choice of the donor pools and the λ_{ij}. But if the design variables already are included in the quadratic distance, there is no need to also include the inclusion probabilities.

The expectation of a calculated distance should be unaffected by inclusion of uninformative design variables, so such variables could as well be left out of the distance calculation. Features of the design with limited significance to the matching should be included in the metric so that a donee with relatively few or bad matches within the same stratum may get potential donors from neighbouring strata. The effect of giving a vaguely informative design variable excessive influence in a metric may result in reduced efficiency, but it may be better than not giving it any influence at all and risking bias due to overestimated similarities between disjoint design subsets.

3.3 Constraining to known quantities

We consider each imputed (population) dataset as independently drawn by modified Pólya sampling. But in many cases there is external knowledge

about population aggregates. If the population is pinned or bounded by other knowledge, then one solution is to reject imputed datasets that do not comply with this knowledge. Assume that a population mean is known to lie within a certain bound. One may then use an acceptance-rejection rule to filter out the datasets whose estimated mean lies within the bound.

If the constraints are extensive, this type of rule may become computationally demanding since few population draws would be accepted. With very detailed or exact constraints, an alternative is to accept a dataset which approximately meats the requirements or to pick the one that closest meats the constraints. If there is complete failure to accept any draws and the possibility of unfulfillable constraints can be ruled out, this might be a signal of a badly fitted or incomplete imputation model, and might be used as a means for improving the imputation method. The distribution of all imputed datasets should give a hint on this issue, as well as on the (potential) importance of applying the constraints.

3.4 Using external unit level information

Although a survey is conducted in order to gather new information there may exist similar units in other studies (which may be used as proxies) e.g. from registers or a similar survey at a previous occasion or of a similar population. If data is sparse, e.g. if there exists no or only one close donor, an obvious option is to use such external units as potential donors in our donor pools. This borrowing of units conforms to the traditional cold deck imputation approach, and might involve modifying the data, e.g. by restating wages with inflation. It may be particularly attractive when the data in a conducted survey is sparse or believed not to be exhaustive enough, e.g. due to insufficient coverage by the sampling frame, but risks introducing bias due to discrepancies between the survey and the external data. The degree of exchangeability to external units may be accounted for in the distance calculation in the same way as design features e.g. by an indicator variable for the unit origin.

Since values in the external data already may have been imputed, an interesting question is whether previously imputed values on units should be used or not. Such considerations would depend on the similarity between the missing data generating processes and the necessary missing data assumption. Given a reasonable common MAR assumption for all the data,

an alternative may be to borrow all unimputed data, simultaneously impute the survey and the external units, and then discard the latter group. This may be an optimal method, but if one wants consistency, i.e. if one wants already published statistics not to change, a better choice might be to use all of the previously imputed set.

3.5 Preserving associations between variables

When estimating $T(x, y, z)$ from imputed datasets it is desirable that the relevant associations have been preserved between the variables, both within the imputed variables and between the imputed and observed variables. If T is the mean of y in a subdomain defined by z, this can be accomplished by including z in the distance metric when imputing y, possibly without the use of x. But if T is the mean of y in a subdomain defined by x, or a coefficient from a regression involving all of x,y and z, then the association between the two variables with missing values x and y becomes important, in these two cases the mean and the covariance structure respectively.

Most imputation models impute each variable separately, possibly utilizing previously imputed variables when imputing the remaining variables with missing values or reimputing the imputed dataset using previously imputed datasets in a MCMC manner (van Buuren and Groothuis-Oudshoorn, 2010)). Due to their univariate character, such methods can be efficient but may fail to preserve associations between variables.

Real donor imputation offers the possibity of imputing several missing values on a unit simultaneously by copying them from the same donor in a common donor approach, and thereby transfer the observed associations to the unobserved values. Associations among variables being imputed and variables included in the distance calculations can then be preserved. This common donor imputation strategy may be limited. The full data distribution of many variables may not be covered by a sample without increasing the number of observations. With multiple values missing in the same units an alternative is then to impute variables in blocks where relations are most needed to be preserved.

4 Simulation

4.1 Population and external sample

In this simulation set, we study the effect from incorporating different kinds of prior information in kernel imputation. The simulations focus on estimating three variables, one population mean and two regression coefficients, as illustrated by simulations in a realistic dataset. The dataset originates from a study on juvenile delinquency during 1959-62 (SOU, 1971), which was complemented with population register data in the eighties. We will use this data as a base for our simulation study, but with some changes in order to make it suitable for our purpose. Our primary goal is to illustrate the method and even though we believe our results to be fairly close to the truth they should not be used for criminological conclusions. (An exact description of the data preparation can be obtained from the author).

Our data consists of two samples. The data from 1960-62 is used to form a population, $U60$, with 3650 boys. This is about 10 % of all 11-15 year old boys in Stockholm during that period. We also obtained an external sample, $S59$, of size 200 from the year 1959 with the same variables. This sample can be considered as coming from a similar study performed two years earlier. It will be utilized in a cold deck manner as described in Section 4.3.

For each boy in $U60$ and $S59$ the data contains the following items:

Assumed fully observed from population register (z in previous notation):

- $A\_$, Geographical area (in $U60/S59$); $A0$=highly congested (1033/84 boys), $A1$=modestly congested (2617/116 boys) .
- SG, Social Group, 0=high (merged from the two highest Social groups, I and II), 1=low (corresponding to Social group III).
- SC, School credits at the age of 15, summed.

Assumed observed on unit responders in the sample (x in previous notation):

- FT, Family type, 0=living with both parents, 1=split family.
- PR, Number of prosecutions at legal courts until the age of 29. Collected from the Swedish legal registers. (One prosecution may correspond to many crimes. The figure does not include crimes with "nolle prosequi").

Assumed observed on item responders in the sample (y in previous notation):

- PS, Average pension savings as income deductions by tax authorities until the age of 34. Collected from pension registers. This is a good indicator of how much they worked in regular jobs during that period.

Table 1: Categorical varibles in $U60$ and $S59$. Both variables refer to ages 11-15y. ρ is the correlation between a variable and pension savings.

Variable name	Description, (categories)	Area	$U60$		$S59$	
			Share(%)	ρ	Share(%)	ρ
FT	Family type,	$A0$	76/24	0.04	73/27	-0.01
	(Whole/Split)	$A1$	78/22	0.43	59/41	0.46
SG	Social group,	$A0$	64/36	0.10	69/31	0.23
	(High/Low)	$A1$	56/44	0.12	49/51	0.14

Table 2: Continuous varibles in $U60$ and $S59$. $q\_$ denotes a quantile. ρ is the correlation to pension savings.

Variable name	Description (age)	Set	Area	Mean	Std	q_0	q_{50}	q_{100}	ρ
SC	School credits (11-15y)	$U60$	$A0$	31.5	8.6	15	30	46	.08
			$A1$	31.1	7.8	11	30	50	.02
		$S59$	$A0$	29.6	8.9	15	27	46	.03
			$A1$	29.6	8.1	15	29	45	.04
PR	Number of prosecutions (16-29y)	$U60$	$A0$	5.3	15.4	0	0	125	-.61
			$A1$	2.2	11.4	0	0	178	-.38
		$S59$	$A0$	6.3	16.4	0	0	57	-.64
			$A1$	2.0	9.5	0	0	80	-.28
PS	Average pension savings (16-34y)	$U60$	$A0$	289	119	1	304	467	1
			$A1$	313	94	2	332	524	1
		$S59$	$A0$	274	134	1	331	464	1
			$A1$	285	110	26	325	459	1

The numeric aspects of the binary variables are described in Table 1. In some imputation simulations we will assume that only the distribution of SG is known in the two areas $A0$ (64/36 %) and $A1$ (56/44 %) or that the marginal distribution (58/42 %) in $U60$ is known for SG (see Section 4.3). The continuous variables are described in Table 2. We will sometimes assume that the correlation ρ between pension savings and the number of prosecutions (PR) is known separately in areas $A0$ (-.61) and $A1$ (-.38) or only totally ($-.48$) in $U60$ (see Section 4.3).

4.2 Sampling and response mechanisms

From the population $U60$ we drew a sample ($S60$) of 400 units, with stratified random sampling. Exactly 100 boys were drawn from each of the four possible combinations of area and social group.

Table 3: Response parameters, probabilities and probabilty ranges for response mechanisms. The parameters denoted $S59$ or $S60$ are used for unit nonresponse and $R59$ or $R60$ for item nonresponse.

Mechanism	Set	α	α_{FT}	α_{SG}	α_{PR}	α_{SC}	$E[\pi_i]$	$min[\pi_i]$	$max[\pi_i]$
mar1	$S60/S59$	0	0	-2	0	.1	.82/.80	.55/.60	.97/.95
	$R60/R59$	1	2	1	-.05	0	.71/.69	.53/.55	.75/.75
mar2	$S59$	0	0	2	0	-0.1	.80	.67	.90
	$R59$	1	2	1	.05	-0.1	.69	.54	.94

Nonresponse was constructed using the following logistic model for response

$$\pi_i = \frac{exp(\alpha + \alpha_{SC}SC_i + \alpha_{SG}SG_i + \alpha_{PR}PR_i + \alpha_{FT}FT_i)}{1 + exp(\alpha + \alpha_{SC}SC_i + \alpha_{SG}SGi + \alpha_{PR}PR_i + \alpha_{FT}FT_i)} \quad (1)$$

and the parameters in Table 3. In samples $S60$ and $S59$ we created unit nonresponse, i.e. FT, PR and PS were set to missing. This resulted in response sets $R60$ and $R59$ in which we created item nonresponse (on PS). Two different models were used. The reason was that we wanted to see the effect of having either the same or having different response mechanisms in the main sample ($S60$ and $R60$) and the external sample ($S59$ and $R59$). In the first scenario the same unit and item response mechanisms (mar1) were used in both samples. The other scenario also used mar1 in the main sample, while the external sample had a response mechanism (mar2) where the impact of SG, PR and SC on the response probability were different.

4.3 Imputation and estimation

Both unit and item nonresponse were imputed with several setups. We used a similar form as in Pettersson (2012; 2013) and calculated the number of donors k in a donor pool as

$$k = [2^K p^{4/(4+q)}] \quad (2)$$

where K is a size parameter, p is the number of eligible (including both fully observed and previously imputed) donors, q is the number of continuous observed variables, and [ ] denotes the integer part. Given, e.g. a size parameter $K = -1$ and $p = 100$ fully observed or imputed units, for imputation of unit nonresponse with $q = 1$ (item nonresponse with $q = 2$) we would have $k = 20$ ($k = 11$) potential donors.

In simulations reported in Section 5 we study the effects of nonresponse on a population mean and one regression coefficient. We estimated the mean population pension savings (μ) and the family type coefficient β_{FT} of pension savings from a linear regression equation

$$PS \sim \beta_0 + \beta_{FT} FT + \beta_{PR} PR \tag{3}$$

using the Survey package in R (Lumley, 2012). β_{FT} was estimated within areas $A0$ and $A1$, and were denoted Ω and $\mho$ respectively.

Table 4: Mean and coefficients in population ($U60$) and external sample ($S59$). Standard errors are given within paranthesis.

	μ (Std)	Ω (Std)	$\mho$ (Std)
Population U60	5816 (0)	319 (0)	1625 (0)
Sample S59	5579 (157)	645 (530)	1767 (357)

The mean in the population ($U60$) and the coefficients in the areas ($A0$ and $A1$) are found in table 4 together with estimates based on the external sample ($S59$). The process of drawing samples, imposing nonresponse, imputing and estimating is repeated 1000 times, and all values are imputed $B = 25$ times in each setup.

We use superscripts in the right position to denote estimates based on the complete data for sample $S60$ without nonresponse (CD), the response set $R60$ after removing also the partial nonresponse (CC), or the different kernel imputation features of Pettersson (2012; 2013): Epanechnikov (E) selection probabilities, Lagrange (L) adjustment, reorientation (R) and shrinkage (S)

(see Section 2.2), e.g. $\widehat{\theta}^{U}$ when no features are used, and $\widehat{\theta}^{ELRS}$ when all four features are used.

In addition to the potential donors units in $S60$, as discussed in Section 3.4, external donors units may also be used from $R59$, $S59$ after it has been imputed, or simultaneously while imputing $S59$, $S59+$. The preimputation of $S59$ is done with the same setup as the imputation of $S60$. This is noted in the lower right corner, where e.g. $\widehat{\theta}_{\emptyset}$ means no external units are potential donors.

A superscript in the left position denotes that the estimator is subject to some constraint. The superscript is μ or μ_A when the estimator is constrained to the population mean or area means of SG, and by ρ or ρ_A when it is constrained to the correlation between PS and PR in the population or the areas, e.g. ${}^{\rho_A}\widehat{\theta}$ when constrained to the correlation in areas $A0$ and $A1$. Any constraints are applied by imputing 200 datasets, and then rejecting all but the 25 which are closest to fulfill the constraints.

For each estimator we calculate root mean squared error

$$RMSE = \sqrt{\frac{1}{1000}\sum_{g=1}^{1000}(\widehat{\theta}_g - \theta)^2} \tag{4}$$

where

$$\widehat{\theta}_g = \sum_{b=1}^{B}\widehat{\theta}_{b,g} \tag{5}$$

is the overall estimated mean in the B imputed datasets, bias

$$BIAS = \frac{1}{1000}\sum_{g=1}^{1000}(\widehat{\theta}_g - \theta), \tag{6}$$

variance

$$VAR = \frac{1}{1000}\sum_{g=1}^{1000}(\widehat{\theta}_g - \frac{1}{1000}\sum_{f=1}^{1000}\widehat{\theta}_f)^2, \tag{7}$$

standard error

$$SE = \sqrt{VAR}, \tag{8}$$

and relative error of estimated variance

$$REEV = \frac{\widehat{VAR} - VAR}{VAR}, \tag{9}$$

where

$$\widehat{VAR} = \frac{1}{1000}\sum_{g=1}^{1000}(\widehat{\sigma}_{y_g} + \frac{B+1}{B(B-1)}\sum_{b=1}^{B}(\widehat{\theta}_g - \widehat{\theta}_{b,g})^2) \tag{10}$$

is the average estimated variance. The estimated SE is

$$\widehat{SE} = \sqrt{\widehat{VAR}}. \tag{11}$$

We always divide $BIAS$ and SE by $RMSE$ of the complete data estimator $\widehat{\theta}^{CD}$, and multiply all figures by 100.

5 Simulation results

5.1 Donor pool size

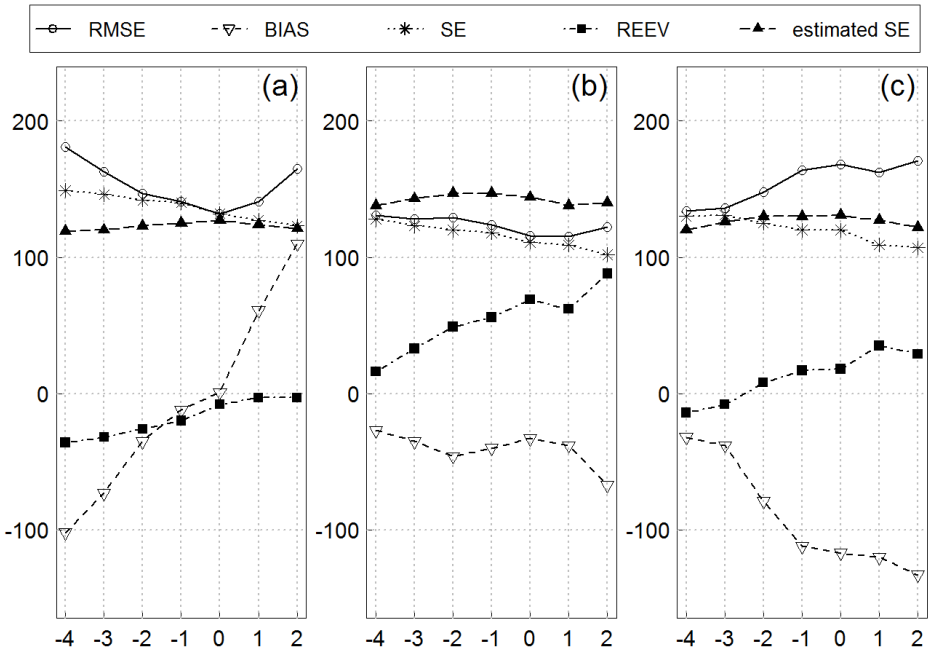

Figure 1: Estimates from 1000 simulations for (a) $\widehat{\mu}_{\emptyset}^{U}$ (b) $\widehat{\Omega}_{\emptyset}^{U}$ (c) $\widehat{\mho}_{\emptyset}^{U}$ with mar1 response mechanism. The x-axis is the size parameter K for number of donors $k = [2^K p^{4/(4+q)}]$.

The donor pool sizes were relatively important. This can be seen in Figure 1. With an increasing donor pool, SE of all three estimators fell, while estimated SE was less affected, causing REEV to increase. REEV was closest to zero with a large donor pool for $\widehat{\mu}_{\emptyset}^{U}$, constantly overestimated for $\widehat{\Omega}_{\emptyset}^{U}$, and close to zero with a small donor pool for $\widehat{\mho}_{\emptyset}^{U}$.

RMSE of $\widehat{\mu}_{\emptyset}^{U}$ is lowest where BIAS is about zero, approximately as donor size parameter $K = 0$. A larger donor pool causes the estimated SE to be

less biased.

BIAS of $\widehat{\Omega}_{\emptyset}^{U}$ and $\widehat{\mho}_{\emptyset}^{U}$ are always negative and the absolute bias increased with the increase of the donor pool . BIAS had strong influence on $\widehat{\mho}_{\emptyset}^{U}$ causing RMSE to be smallest for the smallest donor pool, while RMSE was smallest for $\widehat{\Omega}_{\emptyset}^{U}$ for larger donor pools. SE were always overestimated for $\widehat{\Omega}_{\emptyset}^{U}$ while REEV for $\widehat{\mho}_{\emptyset}^{U}$ were approximately zero as $K = -2.5$.

5.2 Kernel features

Due to the relatively strong dependence on donor pool size, we display results for a smaller pool with $K = -2.5$ (which was relatively favourable to $\widehat{\Omega}$ and $\widehat{\mho}$) and a larger pool with $K = 1.5$ (which was relatively favourable to $\widehat{\mu}$).

Table 5: Estimates from 1000 simulations with no kernel features (U), Epanechnikov selection probabilities (E), Lagrangean calibration of selection probabilities (L), reorientation R or shrinkage (S) for boundary units, or all features combined (ELRS).

	RMSE			BIAS			SE			REEV		
*	$\widehat{\mu}^*_\emptyset$	$\widehat{\Omega}^*_\emptyset$	$\widehat{\mho}^*_\emptyset$	$\widehat{\mu}^*_\emptyset$	$\widehat{\Omega}^*_\emptyset$	$\widehat{\mho}^*_\emptyset$	$\widehat{\mu}^*_\emptyset$	$\widehat{\Omega}^*_\emptyset$	$\widehat{\mho}^*_\emptyset$	$\widehat{\mu}^*_\emptyset$	$\widehat{\Omega}^*_\emptyset$	$\widehat{\mho}^*_\emptyset$
CD	100	100	100	-4	7	4	100	100	100	-1	-0	6
CC	147	158	160	57	53	58	135	149	149	-43	-4	1
K=-2.5												
U	151	128	142	-50	-41	-61	143	122	128	-29	43	1
E	150	126	140	-52	-36	-59	141	121	127	-29	42	2
L	149	125	141	-45	-36	-60	142	120	128	-29	44	0
R	151	128	142	-48	-41	-61	143	122	128	-29	42	1
S	158	127	140	-61	-38	-49	146	122	131	-32	42	-5
ELRS	154	126	138	-56	-36	-47	143	121	129	-31	39	-4
K=1.5												
U	131	117	161	17	-44	-113	130	108	115	-7	64	28
E	129	116	166	-2	-32	-121	129	111	114	-1	70	33
L	131	117	155	15	-34	-108	130	112	111	-3	69	37
R	132	117	163	29	-46	-117	129	107	114	-5	69	28
S	133	116	160	11	-41	-110	132	108	116	-7	68	25
ELRS	130	115	156	9	-28	-111	130	112	110	-2	74	45

Results for a smaller ($K = -2.5$) and a larger ($K = 1.5$) donor pool are seen in Table 5. The effects of applying the kernel features were relatively small and strongly related to donor pool size. For the mean $\widehat{\mu}_\emptyset$ most effects were similar to those reported by Pettersson (2012;2013) with model simulated datasets. We summarize the results for the mean $\widehat{\mu}_\emptyset$ as:

- A large donor pool was preferred. This was mainly due to the fact that shrinkage (S) of a small pool worsened BIAS, SE and REEV. Due to lower SE the estimator even had smaller RMSE of $\widehat{\mu}_\emptyset$ for CC than for kernel imputation with smaller donor pools.

- With larger donor pools, applying all kernel features (ELRS) simultaneously improved RMSE and BIAS. However, reorientation (R) increased BIAS.
- SE was relatively unaffected by the kernel features, though shrinkage (S) seemed to worsen SE.
- With a larger donor pool REEV was slightly improved.

For the coefficient $\widehat{\Omega}_{\emptyset}$ we summarize results as:

- RMSE was slightly reduced or unaffected by all the kernel features.
- Adding kernel features with a small donor pool reduce either or both of BIAS and SE, while with a large donor pool there was more of a trade-off between BIAS and SE.
- BIAS had the opposite direction of the CC estimator, was relatively large, but was reduced by all kernel features except for reorientation (R).
- As seen in REEV the overestimation of SE was exacerbated as the donor pool increased. With a small (large) donor pool the overestimation was smallest with all (no) kernel features added.

For the coefficient $\widehat{\mho}_{\emptyset}$ we summarize results as:

- RMSE was relatively unaffected or slightly reduced by all the kernel features. Two larger effects with a larger donor pool were Epanechnikov (E) which increased RMSE, and Lagrange (L), which decreased RMSE.
- SE fell while BIAS and RMSE increased with the size of donor pool.
- As for the CC estimator, with a small donor pool SE was well estimated, while it was overestimated with a larger donor pool.
- BIAS was improved considerably by shrinkage (S) with a small donor pool, and worsened by Epanechnikov (E) with a larger donor pool.
- There seemed to be a trade-off between BIAS and SE when adding kernel features.

5.3 Using external units

Table 6: Estimates from 1000 simulations with mar1 response mechanism when utilizing external units for imputation.

	RMSE			BIAS			SE			REEV		
$*$	$\widehat{\mu}_*^U$	$\widehat{\Omega}_*^U$	$\widehat{\mho}_*^U$	$\widehat{\mu}_*^U$	$\widehat{\Omega}_*^U$	$\widehat{\mho}_*^U$	$\widehat{\mu}_*^U$	$\widehat{\Omega}_*^U$	$\widehat{\mho}_*^U$	$\widehat{\mu}_*^U$	$\widehat{\Omega}_*^U$	$\widehat{\mho}_*^U$
K=-2.5												
$\emptyset$	151	128	142	-50	-41	-61	143	122	128	-29	43	1
$R59$	147	119	132	-73	-50	-64	128	108	115	-13	87	24
$S59$	146	116	127	-86	-58	-69	118	101	106	1	118	45
$S59+$	143	120	133	-70	-53	-72	125	108	112	-9	86	32
K=1.5												
$\emptyset$	131	117	161	17	-44	-113	130	108	115	-7	64	28
$R59$	113	108	147	-16	-41	-110	112	100	98	22	109	69
$S59$	107	105	140	-33	-45	-108	101	95	88	49	144	104
$S59+$	115	110	151	-19	-43	-114	113	101	99	20	107	64

The use of external units from $R59$ in our simulations is summarized with the following points; see also Table 6.

- In all situations the use of external units reduced SE and also RMSE.
- Using external units from the preimputed $S59$ had the largest effect on all measures. The effects of $R59$ or $S59+$ where the external sample was not preimputed were relatively similar.
- With small donor pools the BIAS increased when external units were used.
- The effect (not shown here) from using the mar2 response mechanisms on $S59$ and $R59$ instead of mar1 was relatively small on all measures.

5.4 Constraining to known quantities

Table 7: Estimates from 1000 simulations with mar1 response mechanism with different constraints of known quantities.

	RMSE			BIAS			SE			REEV		
*	$*\widehat{\mu}^U_\emptyset$	$*\widehat{\Omega}^U_\emptyset$	$*\widehat{\mho}^U_\emptyset$	$*\widehat{\mu}^U_\emptyset$	$*\widehat{\Omega}^U_\emptyset$	$*\widehat{\mho}^U_\emptyset$	$*\widehat{\mu}^U_\emptyset$	$*\widehat{\Omega}^U_\emptyset$	$*\widehat{\mho}^U_\emptyset$	$*\widehat{\mu}^U_\emptyset$	$*\widehat{\Omega}^U_\emptyset$	$*\widehat{\mho}^U_\emptyset$
K=-2.5												
none	151	128	142	-50	-41	-61	143	122	128	-29	43	1
μ	145	131	144	-42	-42	-61	139	124	130	-27	-30	-62
μ_A	145	129	143	-43	-41	-60	139	122	130	-27	-26	-61
ρ	156	122	136	-58	-38	-44	145	116	129	-31	-21	-60
ρ_A	156	122	136	-58	-38	-44	145	116	129	-31	-21	-60
K=1.5												
none	131	117	161	17	-44	-113	130	108	115	-7	64	28
μ	129	118	160	16	-45	-110	128	110	116	-4	-14	-42
μ_A	129	116	160	14	-43	-109	128	108	117	-4	-12	-42
ρ	139	112	150	-4	-42	-91	139	104	119	-17	-12	-41
ρ_A	139	112	150	-4	-42	-91	139	104	119	-17	-12	-41

The use of constraints had several significant effects, as seen in Table 7. The difference between constraining to the population level (μ or ρ) and to the area level (μ_A or ρ_A) was marginal. Constraining to the the known mean (μ) of FT

- improved all measures for $\widehat{\mu}_\emptyset$ irrespective of donor pool size and
- significantly reduced estimates of SE for $\widehat{\Omega}_\emptyset$ and $\widehat{\mho}_\emptyset$, causing REEV to fall.

Constraining to the the known correlation (ρ) between PR and PS

- significantly reduced RMSE of $\widehat{\Omega}_\emptyset$ mainly through BIAS (with a small donor pool) or SE (with a large donord pool), and $\widehat{\mho}_\emptyset$ mainly through BIAS,

- significantly reduced estimates of SE for $\widehat{\Omega}_\emptyset$ and $\widehat{\upsilon}_\emptyset$, causing REEV to fall and
- significantly increased RMSE of $\widehat{\mu}_\emptyset$, either through BIAS with a small donor pool or through SE with a large donor pool. REEV of $\widehat{\mu}_\emptyset$ was slightly worsened irrespective of donor pool size.

5.5 Combining kernel features, external units and constraints.

In this section we combine the kernel features, utilization of external units and constraints, following up each of the subfigures of Figure 1 with similar figures.

5.5.1 Population mean μ

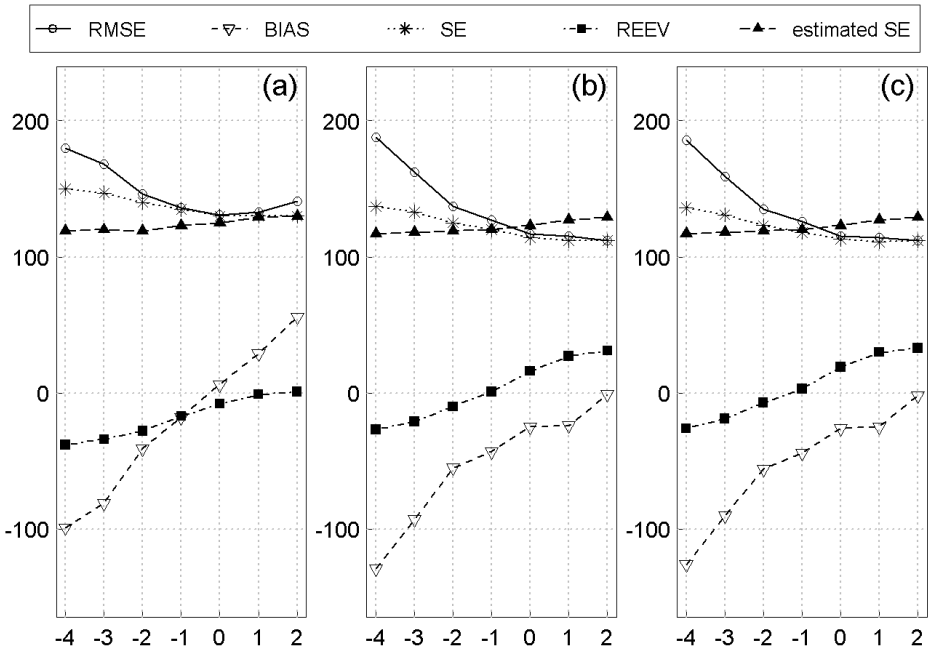

Figure 2: Estimates from 1000 simulations for (a) $\widehat{\mu}_{\emptyset}^{ELRS}$ (b) $\widehat{\mu}_{R59}^{ELRS}$ (c) ${}^{\mu}\widehat{\mu}_{R59}^{ELRS}$. The x-axis is the size parameter K used for calculating the number of donors $k = [2^K p^{4/(4+q)}]$.

When moving from (a) $\widehat{\mu}^{elrs}$ to (b) $\widehat{\mu}_{R59}^{ELRS}$ in Figure 2 we see that BIAS and SE are shifted downwards. RMSE also shifts somewhat downwards except for a small donor pool. When moving further to (c) ${}^{\mu}\widehat{\mu}_{R59}^{ELRS}$ RMSE is shifted slightly further downwards.

5.5.2 Regression coefficient Ω

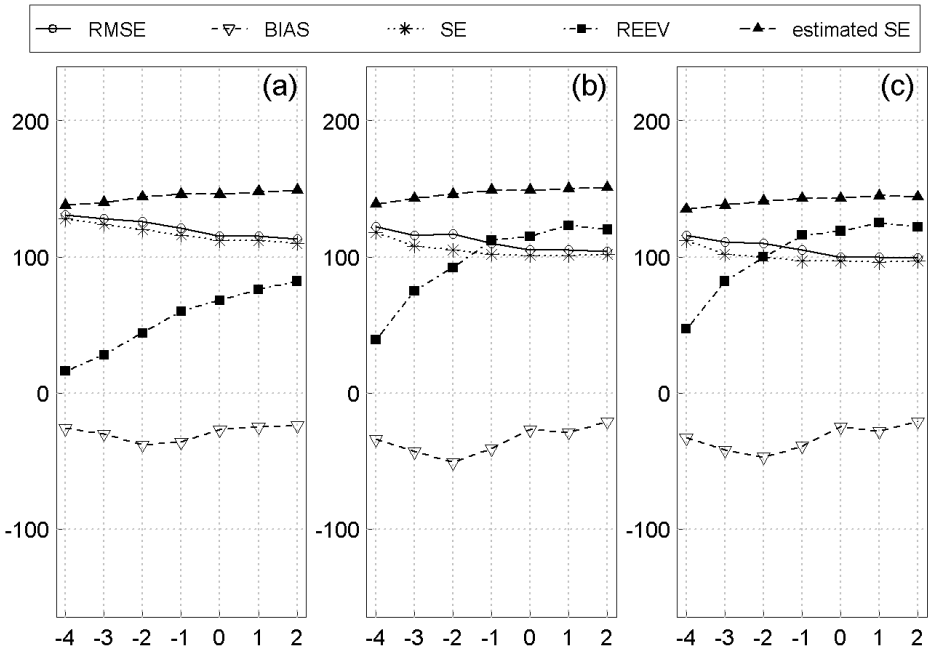

Figure 3: Estimates from 1000 simulations for (a) $\widehat{\Omega}_{\emptyset}^{ELRS}$ (b) $\widehat{\Omega}_{R59}^{ELRS}$ (c) ${}^{\rho}\widehat{\Omega}_{R59}^{ELRS}$. The x-axis is the size parameter K used for calculating the number of donors $k = [2^K p^{4/(4+q)}]$.

When moving from (a) $\widehat{\Omega}^{ELRS}$ to (b) $\widehat{\Omega}_{R59}^{ELRS}$ in Figure 3 we see that SE and RMSE are shifted downwards, while estimated SE is relatively unchanged, causing REEV to increase. The same things are seen when moving further to (c) ${}^{\Omega}\widehat{\Omega}_{R59}^{ELRS}$ but the effects are smaller.

5.5.3 Regression coefficient υ

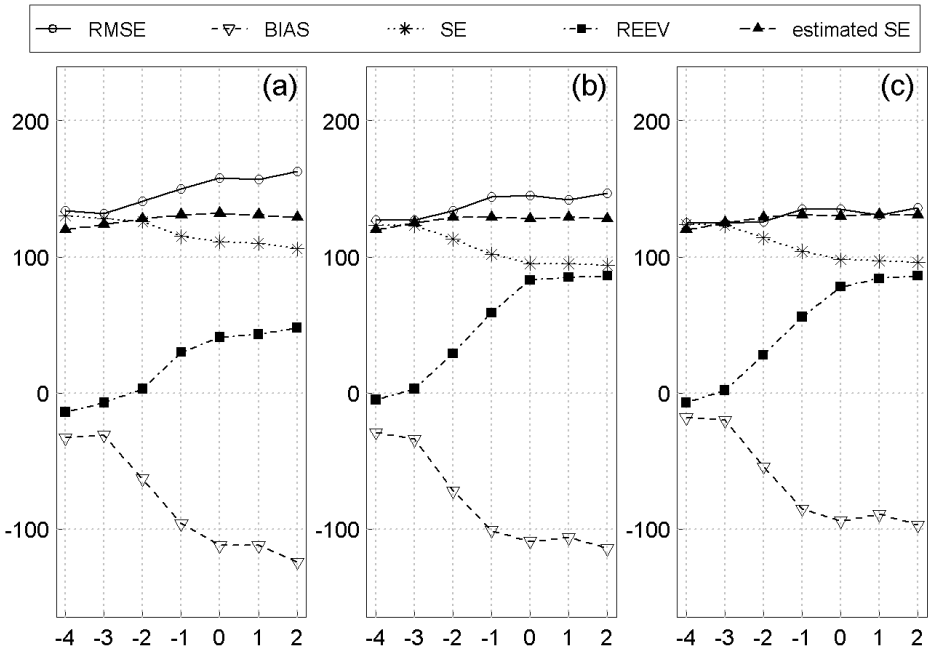

Figure 4: Estimates from 1000 simulations for (a) $\widehat{\upsilon}_{\emptyset}^{ELRS}$ (b) $\widehat{\upsilon}_{R59}^{ELRS}$ (c) ${}^{\rho}\widehat{\upsilon}_{R59}^{ELRS}$. The x-axis is the size parameter K used for calculating the number of donors $k = [2^K p^{4/(4+q)}]$.

When moving from (a) $\widehat{\upsilon}^{ELRS}$ to (b) $\widehat{\upsilon}_{R59}^{ELRS}$ in Figure 4 we see that SE and RMSE is shifted downwards, while estimated SE is relatively unchanged, causing REEV to increase a little. When moving further to (c) ${}^{\Omega}\widehat{\upsilon}_{R59}^{ELRS}$ BIAS improves so that RMSE is shifted downwards.

6 Discussion

There is an apparent potential in informing kernel imputation by utilizing external units (in a cold deck manner) or constraints from known quantities

(on auxiliaries). Both of these features are relatively easily incorporated. In our example with relatively sparse data, the use of external units helped in reducing the variance, but complicated the variance estimation and also introduced bias in estimates. In other situations it is recommended to use any available validation measures to minimize the risk of bias, including the plotting of the data. On the issue whether to borrow a respondent sample, to borrow the preimputed sample, or to impute simultaneously, it seemed as if the preimputed case had the strongest influence while the differences was smaller between the other two approaches. Using different kinds of nonresponse causes worked well in our case with a common MAR assumption, but this needs to be further explored.

Utilizing constraints reduced error of the estimates. The correlation constraints which were irrelevant to the mean estimate were also harmful to it. But the mean constraint did not seem harmful to the coefficients. It was particularly difficult to estimate SE of regression coefficients, but using correlation constraints was an improvement. It should be possible to enhance the constraining effects simply by imputing more datasets to select among. This is also probably needed if one would be fulfilling several constraints simultaneously.

It was not our purpose here, but a major challenge is to decide the size of donor pool, a topic which is discussed in Pettersson (2012; 2013) which needs further exploration. Our simulations indicate that one may need different donor pool sizes for different estimates. It may be possible to find a compromise which can be improved by combining several constraints, but there is probably seldom a single best donor pool size.

To account for the design we included the stratum variables directly in the distance measure. An extension could be to also include the response propensities. The common donor approach could also be compared to a single donor approach.

Taken together the kernel features improved estimation as in Pettersson (2012; 2013), but not all features contributed in all situations. More investigation is warranted to discover whether this were specific to the data or to the context used in this study.

7 References

Andridge, R.R., and Little, R.J.A. (2010). A review of hot deck imputation for survey non-response. International Statistical Review, 78, 40-64.

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

Lundström, S., and Särndal, C.E. (2007). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. Methodology Reports from Statistics Sweden 2007:2. Statistics Sweden, Research and Development

Epanechnikov, V. (1969). Nonparametric estimation of a multivariate probability density. Theory of Probability & Its Applications. 14, 153-158.

Ghosh, M., and Meeden, G. (1997). Bayesian methods for finite population sampling. London: Chapman & Hall.

Härdle, W. (1990). Applied nonparametric regression. New York: Cambridge university press.

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

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

T. Lumley (2012). survey: analysis of complex survey samples. R package version 3.28-2.

Meeden, G. (2003). A noninformative Bayesan approach to Small Area Estimation. Survey Methodology, 29, 19-24.

Nelson, D., and Meeden, G. (1998). Using prior information about population quantiles in finite population sampling. Sakhyã A, 60, 426-445.

Pettersson, N., (2012). Bias reduction of finite population imputation by kernel methods. *To appear in Statistics in Transitions new series.*

Pettersson, N. (2013). Kernel imputation with multivariate auxiliaries. *submitted*

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

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

SOU, (1971). 1956 års klientelundersökning rörande ungdomsbrottslingar. Unga lagöverträdare. 1, Undersökningsmetodik. Brottsdebut och återfall. (in Swedish). Stockholm.