Variance Estimation When Donor Imputation is Used to Fill in Missing Values

Jean-François Beaumont[1] and Cynthia Bocci[2]
Statistics Canada, Statistical Research and Innovation Division (Jean-Francois.Beaumont@statcan.ca)[1]
Statistics Canada, Business Survey Methods Division (Cynthia.Bocci@statcan.ca)[2]

Abstract

Donor imputation is frequently used in surveys. However, very few variance estimation methods that take into account donor imputation have been developed in the literature. This is particularly true for surveys with high sampling fractions using nearest donor imputation, often called nearest-neighbour imputation. In this paper, we develop a variance estimator for donor imputation based on the assumption that the imputed estimator of a domain total is approximately unbiased under an imputation model. Our variance estimator is valid irrespective of the magnitude of the sampling fractions and the complexity of the donor imputation method. We evaluate its performance in a simulation study when nearest-neighbour imputation is used. We also show empirically that nonparametric estimation of the conditional model mean and variance via smoothing splines brings robustness with respect to imputation model misspecifications.

Keywords: Edit rules; Hierarchical imputation classes; Hot-Deck Imputation, Imputation Model, Nearest-Neighbour Imputation, Smoothing Splines.

1. Introduction

Donor imputation is defined as any imputation method for which the missing values for one or more variables of a nonresponding unit, often called a recipient, are replaced by the corresponding values of some donor; i.e., a responding unit with no missing value for these variables. There are several types of donor imputation methods that are used in practice. In household surveys, Random Hot-Deck (RHD) imputation is often the method of choice. With this method, the missing values of a recipient are imputed by randomly choosing a donor among the set of potential donors. In business surveys, Nearest-Neighbour (NN) imputation is more common. With NN imputation, the missing values of a recipient are replaced by the corresponding values of the closest potential donor with respect to a vector of quantitative auxiliary variables. These auxiliary variables are usually first standardized so that they all have a comparable scale.

The popularity of donor imputation in practice is mostly due to its convenience. In particular, it can be used to impute simultaneously more than one variable and it leads to plausible (observed) values, which is especially important if some variables to be imputed are categorical. It is also sometimes considered for the following two statistical reasons: i) it preserves the marginal distribution of the variables being imputed (see Chen and Shao, 2000 in the case of NN imputation), and ii) it helps preserving relationships between variables, especially if a common donor is used to impute simultaneously all the variables with missing values. As shown in section 3, NN imputation has also the additional feature of being a nonparametric imputation method as it leads to an imputed estimator with negligible bias without requiring the specification of a parametric model.

It is worth noting that donor imputation may not be the most efficient imputation method in any specific scenario. Nevertheless, it is quite a popular imputation method in surveys due to its practical advantages. Therefore, it remains useful to develop variance estimation methods that take donor imputation into account. This was actually the goal of this research with the ultimate objective of its implementation in SEVANI, the System for Estimation of Variance due to Nonresponse and Imputation that is being developed at Statistics Canada.

Often, the donor imputation process is complicated by post-imputation edit rules and hierarchical imputation classes. Post-imputation edit rules are constraints that restrict the set of potential donors for a given recipient to those that make the imputed recipient satisfy these constraints. In many cases, they involve relationships between variables that must be satisfied.

Hierarchical imputation classes occur in the situation where it is desirable to perform imputation independently within small imputation classes. In such case, it may happen that the number of potential donors in some classes is too small and, therefore, imputation is not performed within these classes. This problem is aggravated by the use of post-imputation edit rules, which restrict even more the number of potential donors available for a given recipient. After the first round of imputation, there may thus be some

recipients that have not been imputed. To solve this problem, classes are usually collapsed and imputation is repeated a second time for the non-imputed recipients. This process of collapsing classes followed by imputation is repeated until every recipient has found a suitable donor.

Post-imputation edit rules and hierarchical imputation classes have an effect on which donor can be chosen to impute a recipient and, thus, they have an effect on the properties of the resulting imputed estimator of a population total or mean. As shown later, the variance estimation method that we consider can handle naturally these two practical considerations.

If the sampling fraction can be assumed to be negligible so that a without-replacement sampling design can be approximated by a with-replacement sampling design then resampling variance estimation methods can be considered (e.g., Rao and Shao, 1992; Rancourt, 1999; Chen and Shao, 2001; and Kim, 2002). The latter three papers dealt with NN imputation. For NN imputation, an alternative to resampling variance estimation is the method of Chen and Shao (2000). For non-negligible sampling fractions, the literature on variance estimation is more limited. Three notable exceptions are Fay (1999), Rancourt, Särndal and Lee (1994) and Brick, Kalton and Kim (2004). Fay (1999) considered a resampling variance estimation method for the U.S. Census while the latter two papers are based on the general method developed by Särndal (1992).

Rancourt, Särndal and Lee (1994) considered NN imputation under simple random sampling assuming that a ratio imputation model holds. Brick, Kalton and Kim (2004) considered RHD imputation under more general sampling designs assuming a one-factor analysis of variance model holds. Our work can be viewed as an extension of these two papers to general donor imputation methods (with possibly post-imputation edit rules and hierarchical imputation classes) under general sampling designs and more general imputation models. Our approach is also based on Särndal (1992). However, it differs from it in the way the sampling portion of the total variance is estimated.

2. Donor Imputation

We are interested in estimating the population domain total $T_{dy} = \sum_{k \in U} d_k y_k$, where *U* is the finite population of size *N*, *d* is the domain indicator variable indicating whether unit *k* is in the domain of interest ($d_k = 1$) or not ($d_k = 0$) and *y* is the variable of interest. A sample *s* of size *n* is taken from *U* according to a probability sampling design $p(s)$. In the absence of nonresponse, we assume that the Horvitz-Thompson estimator $\hat{T}_{dy} = \sum_{k \in s} w_k d_k y_k$ would be used, where $w_k = 1/\pi_k$ and π_k is the selection probability of unit *k*.

Variable *y* is only observed for a subset s_r of *s* according to a response mechanism $q(s_r \mid s)$. This subset of size n_r is called the set of respondents (or donors) while its complement $s_m = s - s_r$ of size $n_m = n - n_r$ is called the set of nonrespondents (or recipients).

To compensate for the missing *y*-values, donor imputation is performed. This leads to the imputed estimator

$$\hat{T}_{dy}^I = \sum_{k \in s_r} w_k d_k y_k + \sum_{k \in s_m} w_k d_k y_{l(k)}, \qquad (2.1)$$

where $l(k) \in s_r$ is the donor used to impute the recipient *k*. As pointed out in the introduction, a variety of strategies can be considered in practice in order to find donors for imputing recipients. Usually, a vector $\mathbf{x}_k$ of auxiliary variables, available for all the sample units $k \in s$, is used to determine a set s_m^* of selected donors that are "close" to the corresponding recipients in s_m; i.e., for each recipient $k \in s_m$, the corresponding close donor in s_m^* is $l(k)$. The meaning of "close" is given more precisely in section 3 (see equation 3.4). The vector $\mathbf{x}_k$ may contain imputation class indicator variables as in RHD imputation within classes, quantitative auxiliary variables as in NN imputation or a combination of both. Also, situations in which donors are selected according to a random imputation mechanism $o(s_m^* \mid s, s_r)$, as in RHD imputation, are common in practice. What is important to note is that, for any donor imputation method, the imputed estimator (2.1) can always be rewritten as

$$\hat{T}_{dy}^I = \sum_{k \in s_r} W_{dk} y_k, \qquad (2.2)$$

where

$$W_{dk} = w_k d_k + \sum_{i \in s_{m,k}} w_i d_i$$

and $s_{m,k} = \{i : i \in s_m \text{ and } l(i) = k\}$, for $k \in s_r$, is the subset of recipients in s_m that had their missing *y*-value imputed by the same donor *k*. In other words, (2.2) means that the imputed estimator $\hat{T}_{dy}^I$ is linear in

the respondent *y*-values, no matter how complicated the imputation process is, which includes the potential use of post-imputation edit rules and hierarchical imputation classes. This observation will be useful when developing a variance estimator in section 4.

3. Approach to inference

As in Särndal (1992), we decompose the total error, $\hat{T}_{dy}^I - T_{dy}$, of the imputed estimator $\hat{T}_{dy}^I$ as

$$\hat{T}_{dy}^I - T_{dy} = \left(\hat{T}_{dy} - T_{dy}\right) + \left(\hat{T}_{dy}^I - \hat{T}_{dy}\right). \qquad (3.1)$$

The first term on the right-hand side of (3.1) is called the sampling error while the second term is called the nonresponse error. To evaluate properties of the imputed estimator, we use the following imputation model *m*:

$$\begin{aligned} E_m(y_k \mid \mathbf{X},\mathbf{Z},\mathbf{D}) &= \mu(\mathbf{x}_k) \equiv \mu_k, \\ V_m(y_k \mid \mathbf{X},\mathbf{Z},\mathbf{D}) &= \sigma^2(\mathbf{x}_k) \equiv \sigma_k^2, \\ \mathrm{cov}_m(y_k, y_l \mid \mathbf{X},\mathbf{Z},\mathbf{D}) &= 0, \end{aligned} \qquad (3.2)$$

for $k \neq l$, where the subscript *m* indicates that the expectation, variance and covariance are evaluated with respect to the imputation model, **X** is the *N*-row matrix containing $\mathbf{x}_k'$ in its k^{th} row, **Z** is the matrix of design information (e.g., strata and cluster indicators, size measure, …), **D** is a *N*-element vector containing d_k as its k^{th} element, and $\mu(.)$ and $\sigma^2(.)$ are parametric or nonparametric smooth functions of $\mathbf{x}$. Note that **X** may contain information about the design or the domain of interest. Further, we make the following assumption:

A1) $F(\mathbf{Y} \mid s, s_r, s_m^*, \mathbf{X},\mathbf{Z},\mathbf{D}) = F(\mathbf{Y} \mid \mathbf{X},\mathbf{Z},\mathbf{D})$, where $F(.)$ denotes the distribution function and **Y** is a *N*-element vector containing y_k as its k^{th} element.

Assumption (A1) implies that the response mechanism must be ignorable with respect to the imputation model.

Under the imputation model *m* and assumption (A1), the overall bias can be written, using (3.1), as

$$E_{mpqo}\left(\hat{T}_{dy}^I - T_{dy}\right) = E_{pqo}\left\{B_m(\hat{T}_{dy}^I)\right\},$$

where

$$\begin{aligned} B_m(\hat{T}_{dy}^I) &= E_m\left\{(\hat{T}_{dy}^I - \hat{T}_{dy}) \mid s, s_r, s_m^*\right\} \\ &= \sum_{k \in s_m} w_k d_k \left(\mu_{l(k)} - \mu_k\right) \end{aligned} \qquad (3.3)$$

is the conditional model bias and where the subscripts *p*, *q* and *o* represents the sampling design, the response mechanism and the imputation mechanism, respectively. Thus, the conditional model bias and the overall bias vanish if $\mu_{l(k)} = \mu_k$ for all the recipients $k \in s_m$. For instance, this is the case if the imputation model is such that $\mu_k = \mu_c$ for all sample units *k* in imputation class *c* and if each recipient $k \in s_m$ is imputed using a donor in the same class as *k*. Brick, Kalton and Kim (2004) considered this model under RHD imputation within classes. When hierarchical imputation classes are needed due to some small initial imputation classes, the equality $\mu_{l(k)} = \mu_k$ will not hold exactly for all the recipients but may be assumed to hold asymptotically. For NN imputation, Fay (1999) used an imputation model making explicitly the assumption $\mu_{l(k)} = \mu_k$ for all $k \in s_m$. This is perhaps a somewhat strong assumption for NN imputation. Instead, we use the general imputation model given in (3.2) and make the weaker assumption that

$$\mu_{l(k)} - \mu_k = o_p\left(1/\sqrt{n_r}\right) \qquad (3.4)$$

holds. Note that this now requires viewing the vector $\mathbf{x}$ as being random, at least for the continuous variables in that vector. If we further assume

A2) $w_k = O(N/n)$ and $n/n_r = O(1)$

then

$$B_m(\hat{T}_{dy}^I) = (n_m/n)\, o_p\left(N/\sqrt{n}\right). \qquad (3.5)$$

We will see in the next section that this is sufficient to ignore the conditional model bias.

Since $\mu(.)$ is a smooth function of $\mathbf{x}$, assumption (3.4) is satisfied provided that each component of $(\mathbf{x}_{l(k)} - \mathbf{x}_k)$ is $o_p\left(1/\sqrt{n_r}\right)$. In the case of a single continuous auxiliary variable *x*, it is shown in the appendix that $(x_{l(k)} - x_k) = o_p\left(1/\sqrt{n_r}\right)$ for NN imputation, if the following condition is satisfied:

A3) x_k, for $k \in s_r$, are independent, given s, s_r, s_m^*, $\mathbf{Z}$, $\mathbf{D}$ and x_k, for $k \in s_m$, with a

probability density function $f_k(x) > 0$ over the entire range of *x*-values in the population.

Using somewhat different conditions, Chen and Shao (2000) also showed that the overall bias with NN imputation and a single continuous auxiliary variable is negligible.

In the rest of this paper, we assume that the conditional model bias is negligible. In practice, it is always possible to estimate this bias by noting from (2.2) that it can be rewritten as

$$B_m(\hat{T}_{dy}^I) = \sum_{k \in s_r} (W_{dk} - w_k d_k)\mu_k - \sum_{k \in s_m} w_k d_k \mu_k$$

and by replacing the unknown conditional model means μ_k in the above equation by consistent estimates $\hat{\mu}_k$. Then, it can be checked whether this bias estimate is negligible or not compared to the square root of the variance estimate.

4. Variance Estimation

In this section, we omit conditioning on *s*, s_r, s_m^* to simplify the notation. Using (3.1), the overall Mean Squared Error (MSE) can be written as

$$\begin{aligned} E_{mpqo}\left(\hat{T}_{dy}^I - T_{dy}\right)^2 &= E_{pqo}E_m\left(\hat{T}_{dy}^I - T_{dy}\right)^2 \\ &= E_{pqo}\left[V_m\left(\hat{T}_{dy}^I - T_{dy}\right) + \left\{E_m\left(\hat{T}_{dy}^I - T_{dy}\right)\right\}^2\right] \\ &= E_{pqo}\left[V_m\left(\hat{T}_{dy}^I - T_{dy}\right) + \left\{B_m(\hat{T}_{dy}^I) + E_m\left(\hat{T}_{dy} - T_{dy}\right)\right\}^2\right], \end{aligned}$$

where

$$E_m\left(\hat{T}_{dy} - T_{dy}\right) = \sum_{k \in s} w_k d_k \mu_k - \sum_{k \in U} d_k \mu_k$$

and

$$V_m\left(\hat{T}_{dy}^I - T_{dy}\right) = \sum_{k \in s_r} (W_{dk} - d_k)^2 \sigma_k^2 + \sum_{k \in U - s_r} d_k \sigma_k^2 \ .$$

Since $E_m(\hat{T}_{dy})$ is a Horvitz-Thompson estimator of $E_m(T_{dy})$ then $E_m(\hat{T}_{dy} - T_{dy})$ is typically assumed to be $O_p(N/\sqrt{n})$ under standard conditions. Also, $V_m(\hat{T}_{dy}^I - T_{dy}) = O_p(N^2/n)$ if the assumption

A4) $W_{dk} = O_p(N/n)$ and $\sigma_k^2 = O_p(1)$

holds. The first part of assumption (A4) means that the number of times the same donor can be used to impute recipients is bounded in probability. From (3.5), the conditional model bias $B_m(\hat{T}_{dy}^I)$ can thus be neglected in the expression for the overall MSE. This leads to the approximation:

$$\begin{aligned} E_{mpqo}\left(\hat{T}_{dy}^I - T_{dy}\right)^2 &\approx E_{pqo}\left[V_m\left(\hat{T}_{dy}^I - T_{dy}\right) + \left\{E_m\left(\hat{T}_{dy} - T_{dy}\right)\right\}^2\right] \\ &= E_{pqo}\left[\begin{array}{l} E_m\left(\hat{T}_{dy} - T_{dy}\right)^2 + V_m\left(\hat{T}_{dy}^I - \hat{T}_{dy}\right) \\ \qquad + 2\,\mathrm{cov}_m\left\{\left(\hat{T}_{dy}^I - \hat{T}_{dy}\right), \left(\hat{T}_{dy} - T_{dy}\right)\right\} \end{array}\right] \\ &= E_p E_m\left(\hat{T}_{dy} - T_{dy}\right)^2 \\ &\quad + E_{pqo}\left[V_m\left(\hat{T}_{dy}^I - \hat{T}_{dy}\right) + 2\,\mathrm{cov}_m\left\{\left(\hat{T}_{dy}^I - \hat{T}_{dy}\right), \left(\hat{T}_{dy} - T_{dy}\right)\right\}\right]. \end{aligned}$$

The second row is derived from (3.1). The first component of the last row is usually called the sampling variance while the second component is called the nonresponse component in this paper. In section 4.1, we discuss the estimation of the sampling variance while, in section 4.2, we discuss the estimation of the nonresponse component.

4.1 Sampling variance estimation

To estimate the sampling variance

$$V_{SAM} = E_p E_m\left(\hat{T}_{dy} - T_{dy}\right)^2 = E_m V_p\left(\hat{T}_{dy}\right),$$

let us first consider a design-unbiased full response sampling variance estimator $v(y)$ of $V_p(\hat{T}_{dy})$; i.e., $E_p\left(v(y)\right) = V_p(\hat{T}_{dy})$. For instance, the usual Horvitz-Thompson estimator

$$v(y) = \sum_{k \in s}\sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (w_k d_k y_k)(w_l d_l y_l) \quad (4.1)$$

is design-unbiased, where π_{kl} is the joint selection probability of sample units *k* and *l*. Obviously, this estimator cannot be used when there is nonresponse since it depends on some unobserved *y*-values. Särndal (1992) proposed the variance estimator $\hat{V}_{SAM}^S = v(y_\bullet) + \hat{V}_{DIF}$, where

$$y_{\bullet k} = \begin{cases} y_k & , k \in s_r \\ y_{l(k)} & , k \in s_m \end{cases},$$

and where $\hat{V}_{DIF}$ is an estimator of

$$V_{DIF} = E_m\left(v(y) - v(y_\bullet)\right).$$

Deriving an expression for V_{DIF} may be somewhat tedious in general for donor imputation. For RHD imputation within classes, Brick, Kalton and Kim (2004) suggested the simplified sampling variance estimator $\hat{V}_{SAM}^{BKK} = v(y_\bullet)$, which is simply the naive sampling variance estimator that treats imputed values as true values. They showed in two examples that, under some reasonable conditions, V_{DIF} is negligible. Their proposed sampling variance estimator is thus a natural one to choose with donor imputation since it is easy to compute. However, it seems difficult to show that it is approximately unbiased in the general case.

Instead, we consider

$$\tilde{V}_{SAM} = E_m\left(v(y) \mid \mathbf{Y}_r\right),$$

where $\mathbf{Y}_r$ is the portion of $\mathbf{Y}$ that contains only responding units. It is easy to show that $\tilde{V}_{SAM}$ is unbiased for V_{SAM}; i.e., $E_{mpqo}(\tilde{V}_{SAM}) = V_{SAM}$. Since $\tilde{V}_{SAM}$ will usually depend on the unknown quantities μ_k and σ_k^2, we replace them by consistent estimators $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ to obtain our proposed sampling variance estimator $\hat{V}_{SAM}$. For instance, if we take the sampling variance estimator $v(y)$ given in (4.1) then $\hat{V}_{SAM}$ becomes

$$\hat{V}_{SAM} = v(y_\bullet^{\hat{\mu}}) + \sum_{k\in s_m}(1-\pi_k)w_k^2 d_k \hat{\sigma}_k^2 \quad , \qquad (4.2)$$

where

$$y_{\bullet k}^{\hat{\mu}} = \begin{cases} y_k & , k \in s_r \\ \hat{\mu}_k & , k \in s_m . \end{cases}$$

Our sampling variance estimator is very easy to implement if a software package is already available to compute $v(y_\bullet^{\hat{\mu}})$. It can be obtained using Särndal's approach and conditioning on $\mathbf{Y}_r$ when taking the model expectation in the expression for V_{DIF}. This conditioning greatly simplifies the derivations as compared to the unconditional approach of Särndal. Also, it seems useful to condition on the observed y-values as is also done with multiple imputation. Indeed, $\hat{V}_{SAM}$ is similar to the sampling variance estimator under multiple imputation since (4.2) could be approximated by

i) randomly imputing the missing values from the imputation model m and using $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ instead of the unknown μ_k and σ_k^2;
ii) computing a full response sampling variance estimate by treating the imputed values in (i) as true values;
iii) repeating the steps (i) and (ii) a large number of times; and then
iv) taking the average of the sampling variance estimates obtained in step (ii) to compute the final sampling variance estimate.

As the number of repetitions increases, the sampling variance estimator resulting from the above procedure converges to estimator (4.2).

4.2 Nonresponse component estimation

The nonresponse component, denoted by C_{NR}, was given above as

$$C_{NR} = E_{pqo}\left[V_m\left(\hat{T}_{dy}^I - \hat{T}_{dy}\right) + 2\,\mathrm{cov}_m\left\{\left(\hat{T}_{dy}^I - \hat{T}_{dy}\right),\left(\hat{T}_{dy} - T_{dy}\right)\right\}\right].$$

Using expression (2.2), the nonresponse error can be written in the linear form

$$\hat{T}_{dy}^I - \hat{T}_{dy} = \sum_{k\in s_r}\left(W_{dk} - w_k d_k\right) y_k - \sum_{k\in s_m} w_k d_k y_k \ .$$

As a result, we have

$$V_m\left(\hat{T}_{dy}^I - \hat{T}_{dy}\right) = \sum_{k\in s_r}\left(W_{dk} - w_k d_k\right)^2 \sigma_k^2 + \sum_{k\in s_m} w_k^2 d_k \sigma_k^2 \qquad (4.3)$$

and

$$\mathrm{cov}_m\left\{\left(\hat{T}_{dy}^I - \hat{T}_{dy}\right),\left(\hat{T}_{dy} - T_{dy}\right)\right\} = \sum_{k\in s_r}\left(W_{dk} - w_k d_k\right)\left(w_k - 1\right) d_k \sigma_k^2 - \sum_{k\in s_m} w_k\left(w_k - 1\right) d_k \sigma_k^2 . \qquad (4.4)$$

An estimator $\hat{C}_{NR}$ of the nonresponse component C_{NR} can simply be obtained by adding (4.3) and (4.4) and by replacing the unknown σ_k^2 by $\hat{\sigma}_k^2$. Finally, the overall MSE is estimated by $\hat{V}_{SAM} + \hat{C}_{NR}$.

For RHD imputation within classes and a one-factor analysis of variance model, our estimator $\hat{C}_{NR}$ of the nonresponse component reduces to the one given in Brick, Kalton and Kim (2004). However, note that our

development is much simpler owing to the use of the linear form (2.2) of the imputed estimator $\hat{T}_{dy}^{I}$. Also, our estimator $\hat{C}_{NR}$ is very easy to compute in practice, once the weights W_{dk} have been obtained, as it does not involve any double summation.

Särndal (1992) suggested ignoring the covariance (4.4) to simplify variance estimation as it is zero in some cases. For RHD imputation within classes, Brick, Kalton and Kim (2004) showed that this covariance may be either positive or negative and may not always be negligible. Since an estimate of the covariance (4.4) is not more difficult to compute than an estimate of the variance (4.3), there does not seem to be any practical reason to ignore estimating this covariance.

Note that (4.4) can be rewritten as

$$\mathrm{cov}_m\left\{\left(\hat{T}_{dy}^{I}-\hat{T}_{dy}\right),\left(\hat{T}_{dy}-T_{dy}\right)\right\}= \sum_{k\in s_m} w_k d_k\left\{\left(w_{l(k)}-1\right)d_{l(k)}\sigma_{l(k)}^2-\left(w_k-1\right)d_k\sigma_k^2\right\}.$$

Therefore, the above covariance is small when each recipient *k* is in the same domain as its donor $l(k)$ and has a weight w_k and a model variance σ_k^2 close to the weight and the model variance of its donor. Also, this covariance can be quite negative if the weights $w_{l(k)}$ of the donors are small. A negative covariance reduces the nonresponse component and the overall MSE. This may thus suggest choosing donors with small weights provided that this imputation strategy does not introduce any model bias.

5. Simulation Study

We conducted a simulation study to evaluate the performance of our variance estimator in terms of Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE). In this section, we briefly describe the simulation set-up and a few results. More details will be given in a forthcoming paper that we are writing.

We first generated a population of size 1000 with two *y*-variables, a domain variable *d* and a single auxiliary variable *x*. The first variable of interest, y^{LIN}, is generated from a linear model between *y* and *x* while the second variable of interest, y^{NLIN}, is generated from a nonlinear model between *y* and *x*. From this population, we generated 1000 independent samples of size 500 by simple random sampling without replacement. This is a case with a large sampling fraction ($1/2$). Similarly, we also generated 10000 independent samples of size 50, which yields a small sampling fraction ($1/20$). For each selected sample, nonresponse was generated independently from one sample unit to another with a response probability that depends on *x* and an average response probability in the population of about 0.5. Missing values were imputed using NN imputation for each sample.

Our simulation study had two main objectives:

i) Compare parametric (PAR) and nonparametric (NPAR) estimation of μ_k and σ_k^2 for our proposed method;
ii) With μ_k and σ_k^2 being estimated nonparametrically, compare our proposed method, our proposed method but with the sampling variance estimated by the naïve estimator $\hat{V}_{SAM}^{BKK}=v(y_\bullet)$, which we denote by BKK, and the method of Chen and Shao (2000), which we denote by CS.

For the parametric estimation method, we estimated μ_k and σ_k^2 using the same linear model as the one used to generate y^{LIN}. Nonparametric estimation of μ_k and σ_k^2 was achieved using the procedure TPSPLINE of SAS (SAS Institute Inc., 1999), which performs smoothing splines based on penalized least-squares estimation. For the estimation of σ_k^2, we actually modeled $(y_k-\hat{\mu}_k)^2$ as a function of *x*. The predicted model variance $\hat{\sigma}_k^2$ for the recipients was occasionally very large or negative. To address this issue, we winsorized the largest and smallest $\hat{\sigma}_k^2$. We used the largest and the smallest positive $\hat{\sigma}_k^2$ among the respondents as the upper and lower winsorization thresholds respectively.

Some results are given in table 1 and 2. We focus here on the scenario with a large sampling fraction as the results were more striking than those obtained when the sampling fraction was small.

Table 1: Comparison of parametric and nonparametric estimation of μ_k and σ_k^2 for our proposed method.

Method	RB in %		RRMSE in %	
	y^{LIN}	y^{NLIN}	y^{LIN}	y^{NLIN}
PAR	-2.4	358.4	15.7	514.1
NPAR	-0.3	-18.8	21.5	54.6

First, we can observe in table 1 that nonparametric estimation of μ_k and σ_k^2 is clearly superior to

parametric estimation for variable y^{NLIN}, both in terms of RB and RRMSE. For the variable y^{LIN}, the nonparametric method is still good in terms of bias although it is slightly less efficient than the parametric method. These results show the robustness of the nonparametric method.

Table 2: Comparison of variance estimation methods when μ_k and σ_k^2 are estimated nonparametrically.

Method	RB in %		RRMSE in %	
	y^{LIN}	y^{NLIN}	y^{LIN}	y^{NLIN}
Proposed	-0.3	-18.8	21.5	54.6
BKK	-0.3	-12.0	21.8	69.1
CS	33.9	59.6	53.7	118.7

Results in table 2 show that the CS method is quite biased when the sampling fraction is large. This is not surprising as this method was designed to work only with negligible sampling fractions. Our proposed method and the BKK method are both effective in controlling the bias. In terms of efficiency, our method seems to be slightly better than the method of BKK, especially for the variable y^{NLIN}.

6. Conclusion

We have proposed a variance estimation method, which uses an imputation model and which is valid for any type of donor imputation. Our variance estimator is valid even in the presence of high sampling fractions and was shown to work well in a simulation study. One key aspect of any variance estimation method that relies on an imputation model is the estimation of the conditional model mean μ_k and variance σ_k^2. We have shown empirically that using nonparametric smoothing splines offers robustness with NN imputation; i.e., it gives variance estimates with small bias without needing to specify a parametric imputation model while it remains not too far away from the parametric alternative, in terms of efficiency, when the imputation model is correctly specified.

We have done all our theoretical development under the assumption that the Horvitz-Thompson estimator would be used if there was no missing *y*-value. The extension to calibration estimators (Deville and Särndal, 1992) does not introduce any major complication. A sampling variance estimator can be obtained as in section (4.1) with $v(y)$ being a full response sampling variance estimator under calibration. The estimation of the nonresponse component is simply obtained by replacing the Horvitz-Thompson weights w_k in the derivations by calibration weights.

Our overall MSE estimator remains valid for many other imputation methods as long as the imputed estimator can be written in the linear form (2.2), with some suitably defined weights W_{dk}, and is approximately unbiased under the imputation model. For instance, this is the case with fractional donor imputation (e.g., Kim and Fuller, 2004) or regression imputation (e.g., Deville and Särndal, 1994).

Acknowledgements

The authors would like to thank Joël Bissonnette from Statistics Canada and Begoña Martín from the UK Office for National Statistics for their useful comments and discussions.

Appendix:

Proof that $(x_{l(k)} - x_k) = o_p\left(1/\sqrt{n_r}\right)$ *for NN imputation*

In this appendix, all probability statements are conditional on *s*, s_r, s_m^*, **Z**, **D** and x_k, for $k \in s_m$. Let x_0 be any given fixed value within the range of *x*-values in the population. With NN imputation, a donor is chosen such that its *x*-value is the closest to x_0 among the donors $k \in s_r$. Let us denote by x_0^{NN}, the *x*-value of this nearest donor. Let us also denote by $F_k(x)$, the distribution function of x_k, for $k \in s_r$. Then, from assumption (A3) and for any constants $\varepsilon > 0$ and $\tau > 0$, we have

$$\begin{aligned} P\left(\left|x_0^{NN} - x_0\right| \geq \frac{\varepsilon}{n_r^\tau}\right) &= \prod_{k \in s_r} P\left(\left|x_k - x_0\right| \geq \frac{\varepsilon}{n_r^\tau}\right) \\ &= \prod_{k \in s_r}\left[1 - \left\{F_k\left(x_0 + \frac{\varepsilon}{n_r^\tau}\right) - F_k\left(x_0 - \frac{\varepsilon}{n_r^\tau}\right)\right\}\right] \\ &= \prod_{k \in s_r}\left(1 - \frac{2\varepsilon \tilde{f}_k(x_0)}{n_r^\tau}\right), \end{aligned}$$

where

$$\tilde{f}_k(x_0) = \frac{n_r^\tau}{2\varepsilon}\int_{x=x_0-n_r^{-\tau}\varepsilon}^{x=x_0+n_r^{-\tau}\varepsilon} f_k(x)\partial x .$$

From assumption (A3), $f_k(x) > 0$ and we thus have $\tilde{f}_k(x_0) > 0$ and $C_k = 2\varepsilon \tilde{f}_k(x_0) > 0$ for all positive integer n_r. Note that $\lim_{n_r \to \infty} \tilde{f}_k(x_0) = f_k(x_0) > 0$. Letting

$C_{\min} > 0$ be the smallest value of C_k over $k \in s_r$ and all positive integers n_r, we obtain

$$\begin{aligned} P\left(\left|x_0^{NN} - x_0\right| \geq \frac{\varepsilon}{n_r^{\tau}}\right) &= \prod_{k \in s_r}\left(1 - n_r^{-\tau} C_k\right) \\ &\leq \prod_{k \in s_r}\left(1 - n_r^{-\tau} C_{\min}\right) \\ &= \left(1 - n_r^{-\tau} C_{\min}\right)^{n_r} \\ &= \exp\left\{n_r \log\left(1 - n_r^{-\tau} C_{\min}\right)\right\} \\ &= \exp\left\{\frac{\log\left(1 - n_r^{-\tau} C_{\min}\right)}{n_r^{-1}}\right\}. \end{aligned}$$

Taking the limit as $n_r \to \infty$ on both sides and using l'Hospital's rule, we have

$$\lim_{n_r \to \infty} P\left(\left|x_0^{NN} - x_0\right| \geq \frac{\varepsilon}{n_r^{\tau}}\right) \leq \exp\left(-\tau C_{\min} \lim_{n_r \to \infty} n_r^{1-\tau}\right) .$$

As a result,

$$\lim_{n_r \to \infty} P\left(\left|x_0^{NN} - x_0\right| \geq \frac{\varepsilon}{n_r^{\tau}}\right) = 0 ,$$

provided that $\tau < 1$. Thus, $(x_0^{NN} - x_0) = o_p\left(1/n_r^{\tau}\right)$ when $\tau < 1$ and, in particular, when $\tau = 1/2$.

In practice, we may have imputation classes and post-imputation edit rules which restrict the set of potential donors so that the number of potential donors, say n_{r*}, may be smaller than n_r. Using a development similar to the above, we obtain

$$\lim_{n_r \to \infty} P\left(\left|x_0^{NN} - x_0\right| \geq \frac{\varepsilon}{n_r^{\tau}}\right) \leq \exp\left(-\tau C_{\min} \lim_{n_r \to \infty} n_{r*}^2 n_r^{-\tau-1}\right) .$$

Therefore, we still have $(x_0^{NN} - x_0) = o_p\left(1/n_r^{\tau}\right)$ provided that $n_r^{(\tau+1)/2} / n_{r*} = o(1)$. Now, suppose that n_{r*} is such that $n_r^{\alpha} / n_{r*} = O(1)$ for a constant $\alpha \geq 0$. Then, $(x_0^{NN} - x_0) = o_p\left(1/n_r^{\tau}\right)$ if $\alpha > (\tau+1)/2$. For instance, if we let $\tau = 1/2$, we need to have $\alpha > 3/4$.

References

Brick, J.M., Kalton, G., and Kim, J.K. (2004). Variance Estimation with Hot Deck Imputation Using a Model. *Survey Methodology*, 30, 57-66.

Chen, J., and Shao, J. (2000). Nearest Neighbour Imputation for Survey Data. *Journal of Official Statistics*, 16, 113-131.

Chen, J., and Shao, J. (2001). Jackknife Variance Estimation for Nearest-Neighbour Imputation. *Journal of the American Statistical Association*, 96, 260-269.

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

Deville, J.-C., and Särndal, C.-E. (1994). Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator. *Journal of Official Statistics*, 10, 381-394.

Fay, R.E. (1999). Theory and Application of Nearest Neighbour Imputation in Census 2000. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 112-121.

Kim, J.K. (2002). Variance Estimation for Nearest-Neighbour Imputation with Application to Census Long Form Data. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1857-1862.

Kim, J.K., and Fuller, W.A. (2004). Fractional Hot Deck Imputation. *Biometrika*, 91, 559-578.

Rancourt, E., Särndal, C.-E., and Lee, H. (1994). Estimation of the Variance in the Presence of Nearest Neighbour Imputation. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 888-893.

Rancourt, E. (1999). Estimation with Nearest Neighbour Imputation at Statistics Canada. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 131-138.

Rao, J.N.K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data under Hot-Deck Imputation. *Biometrika*, 79, 811-822.

Särndal, C.-E. (1992). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. *Survey Methodology*, 18, 241-252.

Sas Institute Inc. (1999). *SAS OnlineDoc, version eight*. SAS institute Inc., Cary, NC, U.S.A.