Statistica Sinica **21** (2011), 515-537

ON VARIANCE ESTIMATION UNDER AUXILIARY VALUE IMPUTATION IN SAMPLE SURVEYS

Jean-François Beaumont[1], David Haziza[2] and Cynthia Bocci[1]

[1]*Statistics Canada and* [2]*Université de Montréal*

Abstract: We study the problem of variance estimation for a domain total when auxiliary value imputation, sometimes called cold-deck or substitution imputation, has been used to fill in missing data. We consider two approaches to inference which lead to different variance estimators. In the first approach, the validity of an imputation model is required. Our proposed variance estimator is nevertheless robust to misspecification of the second moment of the model. Under this approach, we show the somewhat counter-intuitive result that the total variance of the imputed estimator can be smaller than the sampling variance of the complete-data estimator. We also show that the naïve variance estimator (i.e. the variance estimator obtained by treating the imputed values as observed values) is a consistent estimator of the total variance when the sampling fraction is negligible. In the second approach, the validity of an imputation model is not required but response probabilities need to be estimated. Our mean squared error estimator is obtained using robust estimates of response probabilities and is thus only weakly dependent on modeling assumptions. We also show that both approaches lead to asymptotically equivalent total mean squared errors provided that the imputation model underlying the imputed estimator is correctly specified and the sampling fraction is negligible. Finally, we propose a hybrid variance estimator that can be viewed as a compromise between the two approaches. A simulation study illustrates the robustness of our proposed variance (mean squared error) estimators.

Key words and phrases: Cold-deck imputation, imputation model, nonresponse model, response probability, robust variance estimator, self-efficiency.

1. Introduction

Auxiliary Value (AV) imputation, sometimes called cold-deck imputation (e.g., Shao (2000)) or substitution imputation (Chambers (2005)), is frequently used in surveys to compensate for item nonresponse. For a given nonresponding unit i, AV imputation consists of replacing the missing value of a variable of interest y using only reported values coming from other auxiliary variables of this unit i. Therefore, a unit with a missing y-value is never imputed using reported y-values of other units when AV imputation is used. A special case of this imputation method is historical imputation, which is particularly useful in repeated economic surveys for variables that tend to be stable over time (e.g., number of employees). A version of historical imputation, sometimes called previous value or carry-forward imputation, consists of replacing the missing value y_i for a given unit i by the value reported on a previous cycle of the survey by the same unit i and for the corresponding variable of interest. Another special case of AV imputation occurs in Statistics Canada's business surveys in the context of the tax replacement program. In an ongoing effort to not only reduce respondent burden and collection cost but also improve data quality, Statistics Canada has been working to increase the use of administrative data in its survey programs. The idea is to identify a group of units that will be surveyed and another group of units for which tax data will be used. For the latter group, AV imputation is currently used for its simplicity. That is, a missing value y_i is replaced by the value of a conceptually similar or identical variable available in a tax file for unit i. Although AV imputation is widely used in practice, the literature on the theoretical properties of this imputation method is quite limited. One notable exception is Shao (2000). Our goal is not to advocate the blind use of AVI; it is simply to study the properties of this method in greater depth because it is widely used in practice.

In Section 2, we introduce notation, assumptions, and the imputed estimator of a domain total under AV imputation. Two approaches to inference are considered: the Imputation Model (IM) approach, originally proposed by Särndal (1992), and the Nonresponse Model (NM) approach. Variance estimation under the IM approach is discussed in Section 3. This approach requires the validity of an imputation model. Our proposed variance estimator is nevertheless robust to misspecification of the second moment of the model. Under this approach, we show the somewhat counter-intuitive result that the total variance of the imputed estimator can be smaller than the sampling variance of the complete-data estimator. This can be explained by noting that the complete-data estimator, although quite useful, is not self-efficient. Self-efficiency, a concept introduced by Meng (1994), is not required for the validity of our method but is required for the validity of multiple imputation (see also Kim et al. (2006)). The use of multiple imputation may thus lead to a substantial overestimation of the total variance in this context. We also show that the naïve variance estimator (i.e., the variance estimator obtained by treating the imputed values as observed values) is a consistent estimator of the total variance when the sampling fraction is negligible. This result can be explained using the reverse approach of Fay (1991) (see also Shao and Steel (1999)). Note that it is well known that the naïve variance estimator may grossly underestimate the total variance for other imputation methods, especially when the response rate is low. The validity of the naïve variance estimator seems to be a property specific to AV imputation.

Mean Squared Error (MSE) estimation under the NM approach is discussed in Section 4. In this approach, the validity of an imputation model is not required but response probabilities need to be estimated. Our MSE estimator is obtained using robust estimates of response probabilities and is thus only weakly dependent on modeling assumptions, contrary to the variance estimator obtained using the IM approach. In Section 5, it is shown that both approaches lead to asymptotically equivalent total MSE provided that the imputation model underlying the imputed estimator is correctly specified and the sampling fraction is negligible. If both conditions are satisfied, the choice of one approach or the other should thus be determined from the validity/invalidity of the imputation model and perhaps personal tastes in the interpretation of confidence intervals, variances and so on. The results of a simulation study are described in Section 6 to illustrate the robustness of our proposed variance (MSE) estimators. Finally, Section 7 contains a discussion on alternative estimators to the AV imputed estimator.

2. Notation and Assumptions

Let U be a finite population of size N. Our goal is to estimate the population domain total, $t_{dy} = \sum_{i \in U} d_i y_i$, for a variable of interest y and a domain of interest d, where $d_i = 1$ if unit i is in the domain of interest and $d_i = 0$ otherwise. We select a random sample s of size n according to a probability sampling design $p(s)$. The complete-data estimator under consideration is the Horvitz-Thompson (HT) estimator

$$\hat{t}_{dy} = \sum_{i \in s} w_i d_i y_i, \tag{2.1}$$

where $w_i = 1/\pi_i$ is the sampling weight of unit i and π_i is its selection probability. The Horvitz-Thompson estimator is design-unbiased, i.e., $E_p(\hat{t}_{dy}) = t_{dy}$, where the subscript p indicates that the expectation is evaluated with respect to the sampling design. It is well known that a design-unbiased estimator of the variance of $\hat{t}_{dy}$, $V_{SAM} = V_p(\hat{t}_{dy})$, is

$$\hat{V}_{SAM} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} (d_i y_i)(d_j y_j), \tag{2.2}$$

where $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j)/\pi_{ij} \pi_i \pi_j$ and π_{ij} is the joint selection probability of units i and j.

Let r_i be the response indicator of unit i such that $r_i = 1$ if unit i responds to variable y and $r_i = 0$ otherwise. We assume that variable d is not subject to nonresponse. We now define the imputed estimator under AV imputation by

$$\hat{t}^I_{dy} = \sum_{i \in s} w_i d_i \widetilde{y}_i, \tag{2.3}$$

where $\widetilde{y}_i = r_i y_i + (1 - r_i) x_i$ and x_i is the imputed value used to replace the missing value y_i. We assume that the auxiliary variable x is available for all the sample units (respondents and nonrespondents). For instance, the following three forms of AV imputation are often used in business surveys: $x_i = z_i$, $x_i = y_i'$ or $x_i = y_i' z_i / z_i'$, where y_i' is the value of variable y for unit i observed at a previous cycle of the survey, z is an auxiliary variable observed at the current cycle or coming from an administrative source, and z' is the corresponding variable observed at a previous cycle.

To study the properties of the imputed estimator (2.3), we use the standard decomposition of the total error of $\hat{t}_{dy}^I$ (e.g., Särndal (1992));

$$\hat{t}_{dy}^I - t_{dy} = (\hat{t}_{dy} - t_{dy}) + (\hat{t}_{dy}^I - \hat{t}_{dy}). \tag{2.4}$$

The first term on the right-hand side of (2.4), $\hat{t}_{dy} - t_{dy}$, is called the sampling error of $\hat{t}_{dy}^I$ whereas the second term, $\hat{t}_{dy}^I - \hat{t}_{dy}$, is called the nonresponse error of $\hat{t}_{dy}^I$. Under AV imputation, the nonresponse error can be expressed as

$$\hat{t}_{dy}^I - \hat{t}_{dy} = -\sum_{i \in s} w_i d_i (1 - r_i)(y_i - x_i). \tag{2.5}$$

This expression is useful later when the bias and variance of the imputed estimator (2.3) are evaluated. Next, we describe two approaches to inference that are used to obtain variance/MSE estimators in Sections 3 and 4: the IM approach and the NM approach.

2.1. The nonresponse model approach

In the NM approach, inference is made with respect to the joint distribution induced by the sampling design and the nonresponse model. The nonresponse model is a set of assumptions about the unknown distribution of the response indicators $\mathbf{R}_s = \{r_i; i \in s\}$, often called the nonresponse mechanism. The probability that sample unit i responds is denoted by $p_i = P(r_i = 1 \mid s, \mathbf{X}_s, \mathbf{X}_s^*)$, where $\mathbf{X}_s = \{x_i; i \in s\}$, $\mathbf{X}_s^* = \{\mathbf{x}_i^*; i \in s\}$, and $\mathbf{x}^*$ is a (potential) vector of additional auxiliary variables available for all sample units. The probability that both sample units i and j respond is denoted by $p_{ij} = P(r_i = 1, r_j = 1 \mid s, \mathbf{X}_s, \mathbf{X}_s^*)$. Under the assumption that sample units respond independently, we have $p_{ij} = p_i p_j$, for $i \neq j$. Expectations and variances taken with respect to the nonresponse model are denoted by the subscript q to distinguish them from expectations and variances taken with respect to the sampling design, which are denoted by the subscript p.

In this approach, we assume that, after conditioning on s, $\mathbf{X}_s$ and (potentially) $\mathbf{X}_s^*$, the nonresponse mechanism is independent of (or unconfounded with) all other variables involved in the imputed estimator (2.3) as well as the joint selection probabilities. In other words, this assumption means that the distribution of $\mathbf{R}_s$ does not depend on $\mathbf{D}_s = \{d_i, i \in s\}$, $\mathbf{Y}_s = \{y_i; i \in s\}$, $\mathbf{W}_s = \{w_i; i \in s\}$ and $\mathbf{\Pi}_s = \{\pi_{ij}; i \in s, j \in s\}$, after conditioning on s, $\mathbf{X}_s$ and $\mathbf{X}_s^*$. As a result, except for the response indicators r_i, all variables involved in the imputed estimator (2.3), as well as the joint selection probabilities, can be treated as fixed when taking expectations and variances with respect to the nonresponse model.

From (2.5) and the above model assumptions, it is straightforward to see that the q-expectation of the nonresponse error is

$$E_q(\hat{t}_{dy}^I - \hat{t}_{dy} \mid s) = -\sum_{i \in s} B_{wi}, \tag{2.6}$$

where $B_{wi} = w_i d_i (1 - p_i)(y_i - x_i)$. If we further assume that the response probabilities p_i do not depend on s, then the pq-expectation of the total error of $\hat{t}_{dy}^I$ can be written as

$$E_{pq}(\hat{t}_{dy}^I - t_{dy}) = -\sum_{i \in U} B_i, \tag{2.7}$$

where $B_i = d_i(1 - p_i)(y_i - x_i)$. Note that the right-hand sides of (2.6) and (2.7) are not necessarily negligible, which implies that the imputed estimator $\hat{t}_{dy}^I$ is not necessarily unbiased under the NM approach.

2.2. The imputation model approach

In the IM approach, inference is made with respect to the joint distribution induced by the imputation model, the sampling design, and the nonresponse model. The imputation model is a set of assumptions about the unknown distribution of $\mathbf{Y}_U = \{y_i; i \in U\}$. With AV imputation, the following model m is quite natural:

$$m : y_i = x_i + \varepsilon_i, \tag{2.8}$$

where ε_i is a random error term uncorrelated with x_i such that $E_m(\varepsilon_i) = 0$, $E_m(\varepsilon_i \varepsilon_j) = 0$, for $i \neq j$, $V_m(\varepsilon_i) = E_m(\varepsilon_i^2) = \sigma_i^2$, and σ_i^2 is some unknown function of x_i (and, potentially, $\mathbf{x}_i^*$). The subscript m indicates that expectations and variances are evaluated with respect to model m. It is implicit in our notation that every expectation or variance with respect to model m is conditional on $\mathbf{X}_U = \{x_i; i \in U\}$ and $\mathbf{X}_U^* = \{\mathbf{x}_i^*; i \in U\}$. In this approach, we assume that the distribution of the model errors $\boldsymbol{\varepsilon}_U = \{\varepsilon_i; i \in U\}$ does not depend on s, s_r, $\mathbf{D}_U = \{d_i; i \in U\}$, $\mathbf{W}_U = \{w_i; i \in U\}$, and $\mathbf{\Pi}_U = \{\pi_{ij}; i \in U, j \in U\}$, after conditioning on $\mathbf{X}_U$ and $\mathbf{X}_U^*$, where $s_r = \{i : r_i = 1, i \in s\}$ is the set of respondents to variable y. As a result, except for the variable of interest y, all variables involved in the imputed estimator (2.3), as well as the joint selection probabilities, can be treated as fixed when taking expectations and variances with respect to the imputation model m.

From (2.5) and the model assumptions, it is straightforward to see that $E_m(\hat{t}^I_{dy} - \hat{t}_{dy} \mid s, s_r) = 0$. As a result, we also have $E_{mpq}(\hat{t}^I_{dy} - t_{dy}) = 0$ so that the imputed estimator $\hat{t}^I_{dy}$ is *mpq*-unbiased. We emphasize that this unbiasedness property requires the validity of the imputation model (2.8), that is quite restrictive and may not always hold in practice. If model (2.8) cannot be validated using available data, then regression models involving unknown parameters to be estimated may offer a more flexible alternative. For instance, Shao (2000) used ratio-type imputation models to obtain an MSE estimator. Nevertheless, we assume that model (2.8) holds when using the IM approach since this assumption leads to an *mpq*-unbiased imputed estimator and, thus, AV imputation is naturally justified under this model. Note that the NM approach is robust in the sense that it does not require the validity of model (2.8), or that of any other imputation model, unlike the IM approach.

It is worth clarifying the main difference between the two approaches to inference at this stage. In the NM approach, $\mathbf{Y}_U$ is treated as fixed and we are interested in estimating the fixed finite population parameter t_{dy}; this is a standard estimation problem. In the IM approach, $\mathbf{Y}_U$ is considered as random. As a result, t_{dy} is random as it involves the population y-values. In the IM approach, we are thus interested in predicting the unknown random variable t_{dy}; this is now a prediction problem rather than an estimation problem. Note that we are not interested in estimating the model expectation of t_{dy}, $E_m(t_{dy}) = t_{dx}$. This distinction between estimating a fixed finite population parameter and predicting a random variable is analogous to the distinction between the design-based approach to inference and the model-based approach to inference (e.g., Valliant, Dorfman and Royall (2000)) in the full response case.

3. Variance Estimation: The IM Approach

In this section, we apply the method proposed by Särndal (1992) to AV imputation. Using the decomposition (2.4) and the fact that $E_m(\hat{t}^I_{dy} - \hat{t}_{dy} \mid s, s_r) = 0$, it is straightforward to show that the total variance of the imputed estimator $\hat{t}^I_{dy}$ can be expressed as

$$V_{mpq}(\hat{t}^I_{dy} - t_{dy}) = E_{mpq}(\hat{t}^I_{dy} - t_{dy})^2 = V^m_{SAM} + V^m_{NR} + V^m_{MIX}, \tag{3.1}$$

where $V^m_{SAM} = E_m\{V_p(\hat{t}_{dy})\} = E_m(V_{SAM})$ is the (anticipated) sampling variance of the complete-data estimator $\hat{t}_{dy}$, $V^m_{NR} = E_{pq}V_m(\hat{t}^I_{dy} - \hat{t}_{dy} \mid s, s_r)$ is the nonresponse variance of the imputed estimator $\hat{t}^I_{dy}$, and $V^m_{MIX} = 2E_{pq}\mathrm{Cov}_m(\hat{t}^I_{dy} - \hat{t}_{dy}, \hat{t}_{dy} - t_{dy} \mid s, s_r)$ is a mixed component. Estimation of V^m_{SAM} is discussed in Section 3.1 while estimation of V^m_{NR} and V^m_{MIX} is discussed in Section 3.2. Section 3.3 adds some remarks about the estimation of the total variance.

3.1. Estimation of the sampling variance V^m_{SAM}

Let $\hat{V}_{ORD}$ be the naive variance estimator of $\hat{t}^I_{dy}$, i.e., the variance estimator obtained by treating the imputed values as observed values. The variance estimator $\hat{V}_{ORD}$ is thus obtained by replacing y_i by $\widetilde{y}_i$ in (2.2), which leads to

$$\hat{V}_{ORD} = \sum_{i\in s}\sum_{j\in s}\Delta_{ij}(d_i\widetilde{y}_i)(d_j\widetilde{y}_j). \tag{3.2}$$

It is well known that $\hat{V}_{ORD}$ usually underestimates V^m_{SAM} for deterministic regression imputation methods. We show below that this is also true for AV imputation. To overcome this difficulty, Särndal (1992) proposed to estimate $V_{DIF} = E_m(\hat{V}_{SAM} - \hat{V}_{ORD} \mid s, s_r)$ by an m-unbiased estimator $\hat{V}_{DIF}$, i.e., $E_m(\hat{V}_{DIF} \mid s, s_r) = V_{DIF}$. Then it is straightforward to show that an mpq-unbiased estimator of V^m_{SAM} is given by $\hat{V}^m_{SAM} = \hat{V}_{ORD} + \hat{V}_{DIF}$. For AV imputation, under (2.8), it is easily seen that

$$V_{DIF} = \sum_{i\in s} w_i(w_i - 1)d_i(1 - r_i)\sigma^2_i. \tag{3.3}$$

Note that $\hat{V}_{ORD}$ underestimates V^m_{SAM} since V_{DIF} in (3.3) is a nonnegative quantity. An estimator $\hat{V}_{DIF}$ can be simply obtained by estimating σ^2_i in (3.3), for $i \in s - s_r$. In the literature, an m-unbiased estimator of σ^2_i is typically obtained by parametrically modeling σ^2_i (e.g., by assuming that $\sigma^2_i = \sigma^2$ or that $\sigma^2_i = \sigma^2 x_i$, where σ^2 is an unknown model parameter to be estimated). Since σ^2_i may be difficult to model in practice, we prefer using an asymptotically m-unbiased nonparametric estimator $\hat{\sigma}^2_i$ of $E_m(\varepsilon^2_i) = \sigma^2_i$. This leads to an asymptotically mpq-unbiased estimator $\hat{V}^m_{SAM} = \hat{V}_{ORD} + \hat{V}_{DIF}$ of V^m_{SAM} that does not require one to specify the form of the unknown variance σ^2_i. It is thus robust to a misspecification of the second moment of the imputation model (2.8). In the simulation study in Section 6, $\hat{\sigma}^2_i$ for the nonrespondents is obtained using smoothing splines (e.g., Wegman and Wright (1983)). In practice, it may also be useful to estimate σ^2_i separately within imputation classes defined on the basis of $\mathbf{x}^*$, if additional auxiliary variables are available.

3.2. Estimation of the nonresponse variance V^m_{NR} and the mixed component V^m_{MIX}

An estimator $\hat{V}^m_{NR}$ of $V^m_{NR} = E_{pq}V_m(\hat{t}^I_{dy} - \hat{t}_{dy} \mid s, s_r)$ can be simply obtained by estimating $V_m(\hat{t}^I_{dy} - \hat{t}_{dy} \mid s, s_r)$. From (2.5), we have

$$V_m(\hat{t}^I_{dy} - \hat{t}_{dy} \mid s, s_r) = \sum_{i\in s} w^2_i d_i(1 - r_i)\sigma^2_i. \tag{3.4}$$

Then, V_{NR}^m is estimated by replacing σ_i^2 in (3.4) by an estimator $\hat{\sigma}_i^2$ as in Section 3.1.

Finally, we obtain an estimator $\hat{V}_{MIX}^m$ of $V_{MIX}^m = 2E_{pq}\text{Cov}_m(\hat{t}_{dy}^I - \hat{t}_{dy}, \hat{t}_{dy} - t_{dy} \mid s, s_r)$ by estimating $2\text{Cov}_m(\hat{t}_{dy}^I - \hat{t}_{dy}, \hat{t}_{dy} - t_{dy} \mid s, s_r)$. Again from (2.5), we have

$$2\text{Cov}_m(\hat{t}_{dy}^I - \hat{t}_{dy}, \hat{t}_{dy} - t_{dy} \mid s, s_r) = -2\sum_{i\in s} w_i(w_i - 1)d_i(1 - r_i)\sigma_i^2. \tag{3.5}$$

Similarly as above, V_{MIX}^m is estimated by replacing σ_i^2 in (3.5) by an estimator $\hat{\sigma}_i^2$. It is worth noting that, assuming $\hat{\sigma}_i^2 > 0$ for all $i \in s$, $\hat{V}_{MIX}^m$ is always negative and not negligible compared to $\hat{V}_{DIF}$ and $\hat{V}_{NR}^m$, even in the case of self-weighting designs. This is in contrast to random hot-deck imputation (Brick, Kalton, and Kim (2004)) and ratio imputation (Särndal (1992)), where $\hat{V}_{MIX}^m$ may be negligible. Indeed, we have $\hat{V}_{MIX}^m = -2\hat{V}_{DIF}$ and

$$\hat{V}_{MIX}^m = -2\hat{V}_{NR}^m + 2\sum_{i\in s} w_i d_i(1 - r_i)\hat{\sigma}_i^2.$$

The last term of the right-hand side of the previous equation is negligible compared to $2\hat{V}_{NR}^m$ if the following three conditions are satisfied, as both n and N increase:

Condition 1: $\sigma_i^2 = O(1)$ for $i \in U$, and $\hat{\sigma}_i^2 = \sigma_i^2 + o_p(1)$, for $i \in s$,

Condition 2: $w_i = O(N/n)$ for $i \in U$, and

Condition 3: $n/N = o(1)$.

Using the first two conditions, we have: $\hat{V}_{MIX}^m / 2\hat{V}_{NR}^m = -1 + O_p(n/N)$. As a result, $\hat{V}_{MIX}^m \approx -2\hat{V}_{NR}^m$ if Condition 3 also holds.

The fact that $\hat{V}_{MIX}^m$ is not negligible is a consequence of the self-inefficiency (Meng (1994)) of the complete-data estimator $\hat{t}_{dy}$, as pointed out in Meng and Romero (2003) and Kim et al. (2006). This apparent inefficiency can only be due to a failure to use the information contained in model (2.8) in the construction of the complete-data estimator. It does not imply that $\hat{t}_{dy}$ is useless. In fact, the estimator $\hat{t}_{dy}$ is widely used in surveys as the prototype estimator (i.e., the estimator that one would use in the complete response case). The sign and magnitude of $\hat{V}_{MIX}^m$ also has the implication that a multiple imputation procedure based on model (2.8), with the complete-data estimator $\hat{t}_{dy}$, may lead to a significantly biased multiple imputation variance estimator with resulting confidence intervals that are too long. This is due to the fact that the multiple imputation variance estimator completely ignores the mixed component (see Kim et al. (2006); Meng and Romero (2003); and Kott (1995)). The methods we discuss in this paper have the advantage over multiple imputation of being valid even if the mixed component is not negligible and the complete-data estimator is not self-efficient.

3.3. Estimation of the total variance

From (3.1)−(3.5), an estimator $\hat{V}^m_{TOT}$ of the total variance $V^m_{TOT} = V_{mpq}(\hat{t}^I_{dy} - t_{dy})$ is given by

$$\begin{aligned}\hat{V}^m_{TOT} &= \hat{V}_{ORD} + \hat{V}_{DIF} + \hat{V}^m_{NR} + \hat{V}^m_{MIX} \\ &= \hat{V}_{ORD} + \sum_{i\in s} w_i d_i (1 - r_i)\hat{\sigma}^2_i. \end{aligned} \tag{3.6}$$

Under Conditions 1 and 2 as well as

Condition 4: $\hat{V}_{ORD} = O_p(N^2/n)$,

it is straightforward to see that

$$\frac{\sum_{i\in s} w_i d_i (1 - r_i)\hat{\sigma}^2_i}{\hat{V}_{ORD}} = O_p(\frac{n}{N}).$$

This implies that $\hat{V}^m_{TOT}/\hat{V}_{ORD} = 1 + O_p(n/N)$. Hence, $\hat{V}^m_{TOT} \approx \hat{V}_{ORD}$ and the naïve variance estimator $\hat{V}_{ORD}$ is consistent when Condition 3 also holds (and provided that $\hat{V}^m_{TOT}$ is consistent for V^m_{TOT}). This also means that $\hat{V}^m_{TOT}$ may be smaller than $\hat{V}^m_{SAM}$ in many instances, which is a consequence of the self-efficiency issue noted in Section 3.2. It is interesting to note that the total variance estimator (3.6) is not dependent on the validity of $\hat{\sigma}^2_i$ as an estimator of σ^2_i if Conditions 1−4 hold, since $\hat{V}^m_{TOT} \approx \hat{V}_{ORD}$ and $\hat{V}_{ORD}$ does not involve $\hat{\sigma}^2_i$. The validity of the individual components V^m_{SAM}, V^m_{NR}, and V^m_{MIX} may however be quite dependent on the validity of $\hat{\sigma}^2_i$. Moreover, the validity of (3.6) remains model-dependent as the imputed estimator is biased if the first moment of model (2.8) does not hold. In such a case, the total MSE becomes a quantity to estimate that is more meaningful than the total variance.

The reverse approach (Fay (1991)) provides an alternative way of obtaining a variance estimator under the imputation model (2.8). Using the variance decomposition in Shao and Steel (1999), the variance estimator for V^m_{TOT} is given by $\hat{V}^{m,R}_{TOT} = v_1 + v_2$, where $v_1 = \hat{V}_{ORD}$ and $v_2 = \sum_{i\in s} w_i d_i (1 - r_i)\hat{\sigma}^2_i$. It is interesting to note that this approach leads to a variance estimator identical to (3.6). Also, using this approach, the naïve variance estimator $v_1 = \hat{V}_{ORD}$ can be interpreted as an estimator of the sampling variance of $\hat{t}^I_{dy}$ conditional on the response indicators r_i, for $i \in U$.

4. MSE Estimation: The NM Approach

When the imputation model (2.8) is not fully satisfactory, it may be desirable to use an alternative approach to variance estimation that is less dependent on it. The NM approach offers such an alternative. It uses a mean squared error decomposition similar to (3.1):

$$E_{pq}(\hat{t}^I_{dy} - t_{dy})^2 = V_{SAM} + V_{NR} + V_{MIX}, \tag{4.1}$$

where $V_{SAM} = V_p(\hat{t}_{dy})$ is the sampling variance of the complete-data estimator $\hat{t}_{dy}$, $V_{NR} = E_{pq}(\hat{t}^I_{dy} - t_{dy})^2$ is the nonresponse variance of the imputed estimator $\hat{t}^I_{dy}$, and $V_{MIX} = 2E_{pq}(\hat{t}^I_{dy} - \hat{t}_{dy})(\hat{t}_{dy} - t_{dy})$ is the mixed component. Estimation of V_{SAM} is discussed in Section 4.1, estimation of V_{NR} is discussed in Section 4.2 while estimation of V_{MIX} is discussed in Section 4.3.

4.1. Estimation of the sampling variance V_{SAM}

The sampling variance estimator $\hat{V}_{SAM}$, given in (2.2), cannot be used directly since it depends on missing y-values. We thus use instead the sampling variance estimator

$$\hat{V}^*_{SAM} = \sum_{i \in s} \sum_{j \in s} \frac{\Delta_{ij}}{\hat{p}_{ij}} (r_i d_i y_i)(r_j d_j y_j), \tag{4.2}$$

where $\hat{p}_{ij}$ is an estimator of p_{ij}. We now state a result that is useful to express more conveniently the estimator $\hat{V}^*_{SAM}$ under the assumption of independence of the response indicators. Under this assumption, we have $\hat{p}_{ij} = \hat{p}_i \hat{p}_j$, for $i \neq j$, where $\hat{p}_i$ is an estimator of p_i.

Result 1. Let $\hat{p}_{ij} = \hat{p}_i \hat{p}_j$, for $i \neq j$, $\hat{p}_{ii} = \hat{p}_i$ and

$$\hat{V}^* = \sum_{i \in s} \sum_{j \in s} \frac{\Delta_{ij}}{\hat{p}_{ij}} (r_i \alpha_i)(r_j \alpha_j),$$

for some α_i. Here $\hat{V}^*$ is an estimator of $V_p(\sum_{i \in s} w_i \alpha_i)$ if the α_i are fixed constants. Then, α_i being constants or not, $\hat{V}^*$ can be written as

$$\hat{V}^* = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \frac{(r_i \alpha_i)}{\hat{p}_i} \frac{(r_j \alpha_j)}{\hat{p}_j} - \sum_{i \in s} \frac{(1 - \pi_i)}{\pi_i^2} \frac{(1 - \hat{p}_i)}{\hat{p}_i^2} r_i \alpha_i^2. \tag{4.3}$$

The proof of *Resut* 1 is straightforward and is thus omitted. The variance estimator $\hat{V}^*_{SAM}$ in (4.2) can be obtained using (4.3) with $\alpha_i = d_i y_i$. The first term on the right-hand side of (4.3) is easy to obtain using standard software packages designed for the complete-data variance estimator (2.2). The second term can usually not be obtained using standard software packages, but is easy to compute as it does not involve a double summation. *Result* 1 is also useful when dealing with the mixed component in Section (4.3).

The estimates $\hat{p}_i$ in (4.3) can be obtained by modeling the probabilities p_i as a function of x and (potentially) $\mathbf{x}^*$. Like Da Silva and Opsomer (2004), we prefer using a nonparametric nonresponse model in order to get rid of additional assumptions. The advantage of a nonparametric approach over a parametric one is that it is weakly dependent on modeling assumptions. In the simulation study in Section 6, the $\hat{p}_i$ are obtained using smoothing splines.

4.2. Estimation of the nonresponse variance V_{NR}

To estimate unbiasedly the nonresponse variance $V_{NR} = E_{pq}(\hat{t}^I_{dy} - \hat{t}_{dy})^2$, it suffices to estimate unbiasedly the conditional nonresponse variance $E_q\{(\hat{t}^I_{dy} - \hat{t}_{dy})^2 \mid s\}$, which can be expressed using (2.6) as

$$\begin{aligned} E_q\{(\hat{t}^I_{dy} - \hat{t}_{dy})^2 \mid s\} &= V_q\{(\hat{t}^I_{dy} - \hat{t}_{dy}) \mid s\} + \Big[E_q\{(\hat{t}^I_{dy} - \hat{t}_{dy}) \mid s\}\Big]^2 \\ &= \sum_{i\in s} w_i^2 p_i(1-p_i)d_i(y_i - x_i)^2 + \Big(\sum_{i\in s} B_{wi}\Big)^2. \end{aligned} \tag{4.4}$$

An obvious estimator of (4.4) is

$$\hat{V}_{NR} = \sum_{i\in s} w_i^2(1-\hat{p}_i)r_i d_i(y_i - x_i)^2 + \sum_{i\in s}\sum_{j\in s} \frac{r_i r_j}{\hat{p}_{ij}} \hat{B}_{wi}\hat{B}_{wj}, \tag{4.5}$$

where $\hat{B}_{wi} = w_i d_i(1-\hat{p}_i)(y_i - x_i)$. Assuming independence of the response indicators so that $\hat{p}_{ij} = \hat{p}_i\hat{p}_j$, for $i \neq j$, it is straightforward to show, similarly as in *Result* 1, that (4.5) reduces to

$$\hat{V}_{NR} = \sum_{i\in s} w_i^2(1-\hat{p}_i)r_i d_i(y_i - x_i)^2 + \Big(\sum_{i\in s} \frac{r_i\hat{B}_{wi}}{\hat{p}_i}\Big)^2 - \sum_{i\in s} \frac{(1-\hat{p}_i)}{\hat{p}_i^2} r_i \hat{B}_{wi}^2. \tag{4.6}$$

4.3. Estimation of the mixed component V_{MIX}

The mixed component V_{MIX} can be expressed as

$$\begin{aligned} V_{MIX} &= 2E_{pq}(\hat{t}^I_{dy} - \hat{t}_{dy})(\hat{t}_{dy} - t_{dy}) = -2E_p\Big\{(\hat{t}_{dy} - t_{dy})\sum_{i\in s} B_{wi}\Big\} \\ &= -2\mathrm{Cov}_p\Big(\hat{t}_{dy}, \sum_{i\in s} B_{wi}\Big) = V_p\Big(\hat{t}_{dy} - \sum_{i\in s} B_{wi}\Big) - V_p(\hat{t}_{dy}) - V_p\Big(\sum_{i\in s} B_{wi}\Big) \\ &= V_p\Big(\sum_{i\in s} w_i a_i\Big) - V_p\Big(\sum_{i\in s} w_i b_i\Big) - V_p\Big(\sum_{i\in s} w_i c_i\Big), \end{aligned} \tag{4.7}$$

where $a_i = d_i y_i - B_{wi}/w_i$, $b_i = d_i y_i$, and $c_i = B_{wi}/w_i$. Assuming that the response probabilities p_i do not depend on s, each variance term in the last row of (4.7) can be easily estimated as in Section 4.1 using *Result* 1. Indeed, note that $V_p(\sum_{i\in s} w_i b_i)$ is actually the sampling variance V_{SAM}. Therefore, if we replace α_i in (4.3) by a_i, b_i, and c_i, we obtain an estimator of $V_p(\sum_{i\in s} w_i a_i)$, $V_p(\sum_{i\in s} w_i b_i)$ and $V_p(\sum_{i\in s} w_i c_i)$, respectively. It is then necessary to replace the unknown a_i and c_i by their estimates $\hat{a}_i = d_i y_i - \hat{B}_{wi} w_i$ and $\hat{c}_i = \hat{B}_{wi}/w_i$, respectively, to obtain the resulting estimator $\hat{V}^*_{MIX}$ of the mixed component.

Finally, following (4.1), an estimator of the total mean squared error $V_{TOT} = E_{pq}(\hat{t}_{dy} - t_{dy})^2$ is given simply by

$$\hat{V}_{TOT} = \hat{V}^*_{SAM} + \hat{V}_{NR} + \hat{V}^*_{MIX}. \tag{4.8}$$

5. Link Between the IM and NM Approaches

In this section, we study the link between the IM and the NM approaches. First, in Section 5.1, we show that the mean squared errors obtained under the two approaches are asymptotically equal provided the imputation model (2.8) holds and the sampling fraction becomes negligible as the population size N grows. Then, in Section 5.2, we propose a hybrid variance estimator that can be viewed as a compromise between the two approaches.

5.1. Asymptotic equality of the mean squared errors

In this section, we show, under mild regularity conditions, that the mean square error of $\hat{t}^I_{dy}$ obtained under the NM approach is asymptotically equivalent to that of $\hat{t}^I_{dy}$ obtained under the IM approach when the imputation model holds and the sampling fraction goes to 0 as the population size increases. In practice, this result is important because it confirms that, as long as the sampling fraction is negligible and the population size is large, the use of the NM approach or the IM approach leads to almost identical mean squared errors.

Result 2: $E_{pq}(\hat{t}^I_{dy} - t_{dy})^2 / E_{mpq}(\hat{t}^I_{dy} - t_{dy})^2 \xrightarrow{p} 1$ as $N \to \infty$ and $n/N \to 0$.

The proof of *Result* 2 is given in the Appendix.

5.2. A hybrid variance estimator

In this section, we propose a hybrid variance estimator that can be viewed as a compromise between the variance estimators obtained under the IM and the NM approaches. Recall from Section 2.1 that the conditional nonresponse bias of the imputed estimator $\hat{t}^I_{dy}$ under the NM approach is $B = -\sum_{i\in s} B_{wi}$. Since B depends on unknown quantities, we propose to estimate it by

$$\hat{B} = -\sum_{i\in s} \frac{r_i}{\hat{p}_i} \hat{B}_{wi}, \tag{5.1}$$

where $\hat{B}_{wi} = w_i d_i(1 - \hat{p}_i)(y_i - x_i)$. The estimator $\hat{B}$ is conditionally asymptotically q-unbiased for B, provided $\hat{p}_i$ is asymptotically q-unbiased for p_i. If the imputation model is valid, then we have $E_m(\hat{B} \mid s, s_r) = 0$. We propose to perform a test of hypothesis with $H_0 : E_m(\hat{B} \mid s, s_r) = 0$ as the null hypothesis and $H_A : \neg H_0$ as the alternative hypothesis. We use the test statistic

$$t = \frac{\hat{B}}{\sqrt{\hat{V}_m(\hat{B} \mid s, s_r)}},$$

where $\hat{V}_M(\hat{B} \mid s, s_r) = \sum_{i \in s} w_i^2 r_i[(1 - \hat{p}_i)^2/\hat{p}_i^2] d_i \hat{\sigma}_i^2$, and $\hat{\sigma}_i^2$ is a (parametric or non-parametric) estimator of σ_i^2 The hybrid variance estimator is then defined as

$$\hat{V}_{TOT}^{HYB} = \begin{cases} \hat{V}_{TOT}^{m} & \text{if } |t| \leq z_{\alpha/2}, \\ \hat{V}_{TOT} & \text{otherwise}, \end{cases} \tag{5.2}$$

where $z_{\alpha/2}$ is the value from the standard normal distribution for a given level α. In other words, if $|t|$ is large, then it is likely that the imputation model is not valid, in which case we use the variance estimator derived under the NM approach. On the other hand, if $|t|$ is small, then it is likely that the imputation model is valid, in which case we use the variance estimator derived under the IM approach. Note that the validity of the test relies on the fact that the asymptotic distribution of $\hat{B}$ is normal. This assumption is often satisfied in practice. As we discuss in Section 6, we can obtain $\hat{V}_{TOT}^{m}$ by using a parametric or a nonparametric estimator of σ_i^2. Similarly, we can obtain $\hat{V}_{TOT}$ by using a parametric or a nonparametric estimator of p_i.

6. Simulation Study

We performed a simulation study to evaluate the accuracy of our proposed robust variance/MSE estimators. In Section 6.1, the simulation experiment is described and, in Section 6.2, results are given.

6.1. Description of the simulation study

We first generated a population U of 400 units with an auxiliary variable x_i drawn from an exponential distribution with mean 1, a domain variable d_i drawn from a Bernoulli distribution with probability 0.5, and the variables $y_{1i} = x_i + \delta_i$, $y_{2i} = x_i + \sqrt{x_i}\delta_i$, and $y_{3i} = \exp(0.5x_i) + \delta_i$, for $i \in U$, where δ_i follows a standard normal distribution. In the case of the variables y_1 and y_2, note that the first moment, $E_m(y_{ki} \mid x_i) = x_i$ for $k = 1, 2$, satisfies the first moment of the model (2.8) underlying AV imputation. Also, note that for the variable y_1, the variance of the errors is constant, whereas it is function of x for the variable y_2. The model used to generate the variable y_3 is a significant departure from the model (2.8).

From the generated population, we selected $R = 10{,}000$ random samples, of size $n = 50$ and $n = 200$ according to simple random sampling without replacement. Finally, for each selected sample, the response indicators r_i were generated independently from a Bernoulli distribution with probability p_i. Specifically, non-response was missing at random based on the auxiliary variable x_i with the probability of response for unit i given by

$$p_i = 0.1 + 0.9\Big[1 + \exp(-0.75 - \lambda x_i)\Big]^{-1},$$

where λ was chosen so that the overall response probability is 0.5. Note that the minimum response probability is 0.1. Missing values for the three variables of interest were replaced by the values of variable x to obtain estimates of the three population domain totals for a particular set of respondents.

For each sample and variable of interest, we considered four variance estimators: the parametric imputation model estimator, the semi-parametric imputation model estimator, the non-parametric estimator, and a hybrid estimator, denoted $\hat{V}_{TOT}^{m,PAR}$, $\hat{V}_{TOT}^{m,SPAR}$, $\hat{V}_{TOT}^{NPAR}$ and $\hat{V}_{TOT}^{HYB}$, respectively. The first two variance estimators are obtained under the IM approach using (3.6) while the third variance estimator is obtained under the NM approach using (4.8). The hybrid variance estimator is given by (5.2). Details of each estimator, as developed for this simulation study, are discussed below.

Both $\hat{V}_{TOT}^{m,PAR}$ and $\hat{V}_{TOT}^{m,SPAR}$ require an estimator, $\hat{\sigma}_i^2$ of the unknown model variance σ_i^2 for the nonrespondents $i \in s - s_r$. For $\hat{V}_{TOT}^{m,PAR}$, the variance σ_i^2 is assumed to be constant for all population units ($\sigma_i^2 = \sigma^2$, for $i \in U$) and is estimated by an m-unbiased estimator

$$\hat{\sigma}_i^2 = \hat{\sigma}^2 = \sum_{j \in s_r} \frac{(y_i - x_i)^2}{n_r},$$

where n_r is the number of respondents. For $\hat{V}_{TOT}^{m,SPAR}$, we first note that the unknown model variance σ_i^2 can be expressed as $\sigma_i^2 = E_m(\varepsilon_i^2 \mid s, s_r, \mathbf{X}_s) = h(x_i)$, for some unknown function $h(\cdot)$. We propose to estimate it nonparametrically by $\hat{\sigma}_{PLS,i}^2$ using the respondents' y-values and penalized least squares estimation, as implemented in the procedure TPSPLINE of SAS (Sas Institute Inc. (1999)). Estimation of the smoothing parameter is obtained by minimizing the generalized cross validation function. For the nonrespondents, the estimates $\hat{\sigma}_{PLS,i}^2$ may occasionally be either very large or negative since their x-values may be outside the range of x-values of the respondents. To obtain more stability in the estimated model variances, the negative $\hat{\sigma}_{PLS,i}^2$, for $i \in s$, were winsorized to $\min\{\hat{\sigma}^2_{PLS,i} : \hat{\sigma}^2_{PLS,i} > 0, i \in s_r\}$ while the largest $\hat{\sigma}^2_{PLS,i}$, for $i \in s - s_r$, were winsorized to $\max\{\hat{\sigma}^2_{PLS,i} : i \in s_r\}$.

The nonparametric estimator, $\hat{V}^{NPAR}_{TOT}$, requires an estimator $\hat{p}_i$ of $p_i = E_q(r_i = 1 \mid s, \mathbf{X}_s) = g(x_i)$, for some unknown function $g(\cdot)$, and an estimator $\hat{p}_{ij}$ of p_{ij}. Again, we used penalized least squares estimation, as implemented in the procedure TPSPLINE of SAS, to obtain $\hat{p}_i$ and the assumption of independence of response indicators to obtain $\hat{p}_{ij} = \hat{p}_i\hat{p}_j$, for $i \neq j$. The estimated $\hat{p}_i$ could occasionally be outside the range $[0, 1]$. Thus, the estimated probabilities resulting from the TPSPLINE procedure were winsorized as follows: initial estimated probabilities larger than one were set equal to one. Those initial estimated probabilities smaller than 10% of the observed sample response rate were set equal to 10% of the observed sample response rate. Then the winsorized weights, $\hat{p}_i^{-1}$ were calibrated to yield final estimated response probabilities that had their sum among the respondents equal to the sample size. With these final estimated probabilities, the estimate $\hat{V}_{TOT}$ as in (4.8) was calculated. Ultimately, to obtain a more stable variance estimator, the nonparametric estimator $\hat{V}^{NPAR}_{TOT}$ was obtained by taking the maximum of $\hat{V}_{TOT}$ and the naïve variance estimator of a domain total (i.e., $\hat{V}^{NPAR}_{TOT} = \max(\hat{V}_{TOT}, \hat{V}_{ORD})$).

Finally, we calculated the hybrid variance estimator as

$$\hat{V}^{HYB}_{TOT} = \begin{cases} \hat{V}^{m,SPAR}_{TOT} & \text{if } |t| \leq z_{\alpha/2}, \\ \hat{V}^{NPAR}_{TOT} & \text{otherwise.} \end{cases}$$

In each simulated sample, we calculated four variance estimators. We took the Monte Carlo average of an estimator $\hat{\theta}$ to be

$$E_{MC}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}^{(r)}, \tag{6.1}$$

where $\hat{\theta}^{(r)}$ denotes the estimator $\hat{\theta}$ in the rth simulated sample, $r = 1, \ldots, R$. As a measure of bias of a variance estimator $\hat{V}$, we used the Monte Carlo percent relative bias given by

$$RB_{MC}(\hat{V}) = 100 \times \frac{E_{MC}(\hat{V}) - V_{MC}(\hat{t}^I_{dy})}{V_{MC}(\hat{t}^I_{dy})}, \tag{6.2}$$

where $E_{MC}(\hat{V})$ is obtained from (6.1) by replacing $\hat{\theta}$ with $\hat{V}$ and $V_{MC}(\hat{t}^I_{dy}) = E_{MC}(\hat{t}^I_{dy} - E_{MC}(\hat{t}^I_{dy}))^2$, which is obtained from (6.1) by replacing $\hat{\theta}$ by $(\hat{t}^I_{dy} - E_{MC}(\hat{t}^I_{dy}))^2$. As a measure of stability of a variance estimator $\hat{V}$, we used the Monte Carlo percent relative root mean square error (RRMSE) given by

$$RRMSE_{MC}(\hat{V}) = 100 \times \frac{[E_{MC}(\hat{V} - V_{MC}(\hat{t}^I_{dy}))^2]^{1/2}}{V_{MC}(\hat{t}^I_{dy})},$$

where $E_{MC}(\hat{V}-V_{MC}(\hat{t}_{dy}^{I}))^2$ is obtained from (6.1) by replacing $\hat{\theta}$ with $(\hat{V}-V_{MC}(\hat{t}_{dy}^{I}))^2$.

Finally, we computed the Monte Carlo average of each variance component (i.e., sampling variance, nonresponse variance, and the mixed term) in order to investigate their respective contribution to the total variance. The Monte Carlo sampling variance is $V_{SAM}^{MC}=E_{MC}(\hat{t}_{dy}-t_{dy})^2$, obtained from (6.1) by replacing $\hat{\theta}$ by $(\hat{t}_{dy}-t_{dy})^2$. The Monte Carlo nonresponse variance is $V_{NR}^{MC}=E_{MC}(\hat{t}_{dy}^{I}-\hat{t}_{dy})^2$, obtained from (6.1) by replacing $\hat{\theta}$ by $(\hat{t}_{dy}^{I}-\hat{t}_{dy})^2$, Finally, the Monte Carlo average of the mixed term was obtained as $V_{MIX}^{MC}=V_{MC}(\hat{t}_{dy}^{I})-V_{SAM}^{MC}-V_{NR}^{MC}$.

6.2. Results of the simulation study

Table 1 shows the Monte Carlo RB and Monte Carlo RRMSE in percentages for $n=200$. As expected, $\hat{V}_{TOT}^{m,PAR}$ and $\hat{V}_{TOT}^{m,SPAR}$ performed comparatively well for the variable y_1 in terms of RB since both the first and second moments of the model are correctly specified. In terms of RRMSE, we note a slightly larger value for $\hat{V}_{TOT}^{m,SPAR}$, which demonstrates that nonparametric procedures entail a loss of efficiency when the model is correctly specified. The variance estimator $\hat{V}_{TOT}^{NPAR}$ was moderately biased with a value of RB approximately equal to 22%. It is important to note that its RRMSE was considerably larger than that of $\hat{V}_{TOT}^{m,PAR}$ or $\hat{V}_{TOT}^{m,SPAR}$ with a value approximately equal to 88%. This result suggests that, when the imputation model is correctly specified, the variance estimators derived under the NM approach are inefficient in comparison with those derived under the IM approach. Finally, the hybrid variance estimator $\hat{V}_{TOT}^{HYB}$ is a compromise between $\hat{V}_{TOT}^{m,SPAR}$ and $\hat{V}_{TOT}^{NPAR}$, both in terms of RB and RRMSE. Note that as the level α of the test decreased, the RB and RRMSE of $\hat{V}_{TOT}^{HYB}$ decreased. This result can be explained by the fact that, as the level α decreases, the value $z_{\alpha/2}$increases. As a result, it becomes increasingly more difficult to reject the null hypothesis, in which case $\hat{V}_{TOT}^{m,SPAR}$ is used. Since the model for the variable y_1 satisfies the AV imputation model (2.8), we expect $\hat{V}_{TOT}^{m,SPAR}$ to be chosen in most samples, resulting in small biases.

For the variable y_2 it is clear that when the variance structure is not correctly specified, the estimator $\hat{V}_{TOT}^{m,PAR}$ is biased, here with a value of RB approximately equal to 18%. Using a nonparametric estimator of σ_i^2 helps in reducing the bias; the value of RB of $\hat{V}_{TOT}^{m,SPAR}$ was about 7% so a good bias reduction was achieved. Once again, the hybrid variance estimator is a compromise between $\hat{V}_{TOT}^{m,SPAR}$ and $\hat{V}_{TOT}^{NPAR}$, both in terms of RB and RRMSE.

For the variable y_3 the variance estimators $\hat{V}_{TOT}^{m,PAR}$ and $\hat{V}_{TOT}^{m,SPAR}$ were considerably biased with values of RB approximately equal to -78% and -84%, respectively. This result is not surprising since the first moment of the model used

Table 1. RB and RRMSE of variance estimators for a domain total using 10,000 samples each of size 200.

	RB (%)			RRMSE (%)		
	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{V}_{TOT}^{m,PAR}$	7.5	17.7	-78.1	16.3	27.3	78.7
$\hat{V}_{TOT}^{m,SPAR}$	8.1	7.0	-83.7	16.8	21.7	83.9
$\hat{V}_{TOT}^{NPAR}$	22.1	6.0	-9.4	87.7	35.5	46.8
$\hat{V}_{TOT,z_{\alpha/2}=1.96}^{HYB}$	16.3	9.1	-11.5	61.9	29.0	48.8
$\hat{V}_{TOT,z_{\alpha/2}=2.575}^{HYB}$	9.7	7.5	-21.0	34.5	24.5	54.8

to generate y_3 is highly mis-specified. On the other hand, $\hat{V}_{TOT}^{NPAR}$ performed well in terms of bias with a value approximately equal to -9%. This result shows that, when the imputation model is incorrectly specified, the use of $\hat{V}_{TOT}^{NPAR}$ may provide a reasonable solution. Again, the hybrid variance estimator leads to compromise values. The RRMSE of the estimators $\hat{V}_{TOT}^{m,PAR}$ and $\hat{V}_{TOT}^{m,SPAR}$ were large, which is mainly due to their large bias. The estimators $\hat{V}_{TOT}^{NPAR}$ and $\hat{V}_{TOT}^{HYB}$ had similar values of RRMSE.

Table 2 shows the Monte Carlo RB and Monte Carlo RRMSE in percentages for $n = 50$. For the variables y_1 and y_2, the estimators $\hat{V}_{TOT}^{m,PAR}$ and $\hat{V}_{TOT}^{m,SPAR}$ performed well in terms of RB. This result can be explained in light of (3.6), which shows that the relative contribution of the naïve variance estimator, $\hat{V}_{ORD}$ to the total variance is substantial when the sampling fraction is negligible and that this component is not affected by the fact that the second moment of the imputation model is correctly specified or not. It is interesting to note that this result still holds with a sampling fraction equal to 12.5%, which can be explained by the fact that the sampling fraction is relatively small and the imputation model has good predictive power. As a result, we expect the value of $\hat{\sigma}_i^2$ (estimated parametrically or nonparametrically) to be small. The RB of $\hat{V}_{TOT}^{NPAR}$ was relatively small for the variables y_1 and y_2 with values of 13.0% and 7.0%, respectively. Once again, the hybrid estimator $\hat{V}_{TOT}^{HYB}$ leads to compromise values. For the variable y_3 the RB of $\hat{V}_{TOT}^{m,PAR}$ and $\hat{V}_{TOT}^{m,SPAR}$ was relatively large, which can be explained by the misspecification of the imputation model, while the RB of $\hat{V}_{TOT}^{NPAR}$ remained relatively small. It is worth noting that all variance estimators showed a similar RRMSE for all three variables, with $\hat{V}_{TOT}^{NPAR}$ being slightly less efficient than the other variance estimators.

The relative contribution of the variance components is shown in Tables 3 and 4 for sample sizes 200 and 50, respectively. As expected, the component

Table 2. RB and RRMSE of variance estimators for a domain total using 10,000 samples each of size 50.

	RB (%)			RRMSE (%)		
	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{V}_{TOT}^{m,PAR}$	0.01	1.7	-43.0	40.7	52.9	70.2
$\hat{V}_{TOT}^{m,SPAR}$	0.2	0.2	-46.8	40.6	52.5	70.6
$\hat{V}_{TOT}^{NPAR}$	13.0	7.7	-10.3	59.5	59.2	76.9
$\hat{V}_{TOT,z_{\alpha/2}=1.96}^{HYB}$	3.3	2.4	-22.9	47.6	55.7	76.7
$\hat{V}_{TOT,z_{\alpha/2}=2.575}^{HYB}$	0.6	0.5	-38.7	41.8	52.9	72.7

Table 3. Contribution (in percentage) of variance components to the total variance of a domain total for 10,000 samples of size 200.

	$\frac{V_{SAM}^{MC}}{V_{TOT}^{MC}}(\%)$	$\frac{V_{NR}^{MC}}{V_{TOT}^{MC}}(\%)$	$\frac{V_{MIX}^{MC}}{V_{TOT}^{MC}}(\%)$
y_1	108.4	35.6	-44.0
y_2	109.0	24.7	-33.7
y_3	17.1	90.1	-7.2

Table 4. Contribution (in percentage) of variance components to the total variance of a domain total for 10,000 samples of size 50.

	$\frac{V_{SAM}^{MC}}{V_{TOT}^{MC}}(\%)$	$\frac{V_{NR}^{MC}}{V_{TOT}^{MC}}(\%)$	$\frac{V_{MIX}^{MC}}{V_{TOT}^{MC}}(\%)$
y_1	125.3	26.5	-51.8
y_2	119.7	17.9	-37.6
y_3	70.1	61.6	-31.7

$\hat{V}_{MIX}^{m}$ was negative in all scenarios. Also, its contribution to the overall total was not negligible in most cases. For example, for variable y_1 and $n = 200$ the relative contribution of $\hat{V}_{MIX}^{m}$ was -44%. For the variables y_1 and y_2 it is interesting to note that the estimated sampling variance $\hat{V}_{SAM}^{m}$ was larger than the total variance with a contribution larger than 100%.

7. Summary and Discussion

In this paper, we derived variance/MSE estimators under both the IM and the NM approaches for AV imputation. When the imputation model is correctly specified, the variance estimators derived under the IM approach, $\hat{V}_{TOT}^{m,PAR}$ (or $\hat{V}_{TOT}^{m,SPAR}$), are asymptotically unbiased and very efficient, whereas the variance estimator derived under the NM approach, $\hat{V}_{TOT}^{NPAR}$, is, in this case, somehow inefficient. However, when the imputation model is incorrectly specified, the variance estimators derived under the IM approach may be considerably biased and the use of $\hat{V}_{TOT}^{NPAR}$ may provide a reasonable solution. We proposed an hybrid variance estimator that can be viewed as a compromise between the NM approach and the IM approach. The results from the simulation study suggest that the hybrid variance estimator has promise.

In the case of the IM approach, we showed that AV imputation led to estimators whose properties are not typical with respect to other imputation methods such as deterministic regression imputation. The results can be summarized as follows: (i) when the sampling fraction n/N is negligible, the naïve variance estimator, $\hat{V}_{ORD}$ provides a consistent estimator of the total variance; (ii) the mixed component is always negative and its contribution to the total variance may be considerable, and (iii) the estimated sampling variance may be larger than the estimated total variance.

As was shown in Section 2.1, the imputed estimator $\hat{t}_{dy}^{I}$ is generally biased under the NM approach; its conditional nonresponse bias, B, is given by (2.6). An asymptotically pq-unbiased estimator of B, denoted by $\hat{B}$ is given by (5.1). Following Haziza and Rao (2006), we obtain a bias-adjusted estimator of the domain total t_{dy} by subtracting the estimated conditional nonresponse bias of the imputed estimator $\hat{t}_{dy}^{I}$, which leads to

$$\begin{aligned}\hat{t}_{dy}^{I(a)} &= \hat{t}_{dy}^{I} - \hat{B} \\ &= \sum_{i \in s} w_i d_i x_i + \sum_{i \in s} \frac{w_i}{\hat{p}_i} r_i d_i (y_i - x_i). \qquad (7.1)\end{aligned}$$

The bias-adjusted estimator (7.1) is similar to a difference estimator in the context of two-phase sampling. It is interesting to note that this bias-adjusted estimator is asymptotically pq-unbiased and mpq-unbiased. Hence the estimator $\hat{t}_{dy}^{I(a)}$ is valid under either the NM approach or the IM approach. Estimator (7.1) does not have the form of an imputed estimator, like (2.3), so its implementation may seem awkward for some users. To overcome this problem, calibrated imputation (Beaumont (2005)) could be used. In this context, calibrated imputation would consist of finding imputed values, for units $k \in s - s_r$ that are close to the corresponding values of x, but that are also constrained to yield an imputed estimator identical to (7.1) for specified domains.

Estimator (7.1) provides an alternative to the imputed estimator (2.3), which is unbiased under the NM approach. It thus offers some protection against a misspecification of the imputation model (2.8) since it remains asymptotically pq-unbiased no matter the validity of the imputation model. An alternative imputed estimator that could achieve the same goal would be obtained by determining an improved imputation model based on less restrictive assumptions, and to use this model to construct the imputed values. For instance, a linear or nonlinear regression imputed estimator could be considered. Evidently, robustness would be achieved at the expense of reduced efficiency if the imputation model (2.8) holds. This highlights the importance of careful modeling when determining an imputed estimator.

Acknowledgement

We thank a referee for constructive comments and suggestions. Work of David Haziza was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

Appendix: Proof of Result 2

We assume the following regularity conditions as n and N grow arbitrarily large.

C$'$1: $\max_{1\le i\le N} w_i = O(N/n)$.

C$'$2: $E_m(y_i^{\delta}) = O(1)$, $\delta = 1, \ldots, 4$.

C$'$3: $\widetilde{\Delta}_{ij} = (\pi_{ij} - \pi_i\pi_j)/(\pi_i\pi_j) = O(1/n)$, $i \neq j$.

C$'$4: The response probability p_i does not depend on s and satisfies $K_1 < p_i$ for some nonnegative constant K_1.

By Chebychev's Inequality, Result 2 is obtained by showing that

$$\frac{V_m[E_{pq}(\hat{t}_{dy}^I - t_{dy})^2]}{[E_{mpq}(\hat{t}_{dy}^I - t_{dy})^2]^2} \to 0 \text{ as } N \to \infty \text{ and } \frac{n}{N} \to 0. \tag{A.1}$$

To show (A.1), we show that

(i) $E_{mpq}(\hat{t}_{dy}^I - t_{dy})^2 = O\Big(N^2/n\Big)$,

(ii) $V_m E_{pq}(\hat{t}_{dy}^I - t_{dy})^2 = \max\Big\{O\Big(N^3/n^2\Big), O(N^2)\Big\}$.

To show (i), write $E_{pq}(\hat{t}_{dy}^I - t_{dy})^2$ as

$$\begin{aligned} E_{pq}(\hat{t}_{dy}^I - t_{dy})^2 &= V_p\Big(\sum_{i\in s} w_i d_i(y_i - E_i)\Big) + \sum_{i\in U} w_i \frac{p_i}{1-p_i} d_i E_i^2 + \Big(\sum_{i\in U} d_i E_i\Big)^2 \\ &= \sum_{i\in U}\sum_{j\in U} \widetilde{\Delta}_{ij} d_i d_j (y_i - E_i)(y_j - E_j) + \sum_{i\in U} w_i \frac{p_i}{1-p_i} d_i E_i^2 + \Big(\sum_{i\in U} d_i E_i\Big)^2, \end{aligned}$$

where $E_i = (1-p_i)(y_i - x_i)$. Noting that

$$E_m\{(y_i - E_i)(y_j - E_j)\} = \begin{cases} x_i^2 + p_i^2\sigma_i^2 & \text{if } i = j, \\ x_i x_j & \text{if } i \neq j, \end{cases}$$

it follows that, under conditions C′1−C′4,

$$\begin{aligned} E_{mpq}(\hat{t}_{dy}^I - t_{dy})^2 &= \sum_{i\in U}\sum_{j\in U}\widetilde{\Delta}_{ij} d_i x_i d_j x_j + \sum_{i\in U}(w_i - 1) d_i p_i^2 \sigma_i^2 \\ &\quad + \sum_{i\in U} w_i d_i p_i (1-p_i)\sigma_i^2 + \sum_{i\in U} d_i (1-p_i)^2 \sigma_i^2 \\ &= O\Big(\frac{N^2}{n}\Big). \end{aligned}$$

To show (ii), we express $V_m E_{pq}(\hat{t}_{dy}^I - t_{dy})^2$ as

$$\begin{aligned} &V_m E_{pq}(\hat{t}_{dy}^I - t_{dy})^2 \\ &= V_m\Big[\sum_{i\in U}\sum_{j\in U}\widetilde{\Delta}_{ij} d_i d_j (y_i - E_i)(y_j - E_j)\Big] \\ &\quad + V_m\Big[\sum_{i\in U} w_i \frac{p_i}{1-p_i} d_i E_i^2\Big] + V_m\Big[\Big(\sum_{i\in U} d_i E_i\Big)^2\Big] \\ &\quad + 2\text{Cov}_m\Big[\sum_{i\in U}\sum_{j\in U}\widetilde{\Delta}_{ij} d_i d_j (y_i - E_i)(y_j - E_j), \sum_{i\in U} w_i \frac{p_i}{1-p_i} d_i E_i^2\Big] \\ &\quad + 2\text{Cov}_m\Big[\sum_{i\in U}\sum_{j\in U}\widetilde{\Delta}_{ij} d_i d_j (y_i - E_i)(y_j - E_j), \Big(\sum_{i\in U} d_i E_i\Big)^2\Big] \\ &\quad + 2\text{Cov}_m\Big[\sum_{i\in U} w_i \frac{p_i}{1-p_i} d_i E_i^2, \Big(\sum_{i\in U} d_i E_i\Big)^2\Big]. \end{aligned} \tag{A.2}$$

The first term of the right-hand side of (A.2) can be written as

$$\begin{aligned} &V_m\Big[\sum_{i\in U}\sum_{j\in U}\widetilde{\Delta}_{ij} d_i d_j (y_i - E_i)(y_j - E_j)\Big] \\ &= \sum_{i\in U}\sum_{j\in U}\sum_{k\in U}\sum_{l\in U}\widetilde{\Delta}_{ij}\widetilde{\Delta}_{kl} d_i d_j d_k d_l \text{Cov}_m\{(y_i - E_i)(y_j - E_j), (y_k - E_k)(y_l - E_l)\}. \end{aligned} \tag{A.3}$$

Under conditions C′1−C′4, it is somewhat tedious but straightforward to show that the right-hand side of (A.3) is of order $O(N^3/n^2)$. The second term on the right-hand side of (A.2) is also $O(N^3/n^2)$. Next, the third term on the right hand side of (A.2) can be written as

$$\begin{aligned} V_m\Big[\Big(\sum_{i\in U} d_iE_i\Big)^2\Big] &= \sum_{i\in U}\sum_{j\in U}\sum_{k\in U}\sum_{l\in U} d_id_jd_kd_l\mathrm{Cov}_m(E_iE_j, E_kE_l) \\ &= 2\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} d_id_jE_m(E_i^2)E_m(E_j^2) + \sum_{i\in U} d_iV_m(E_i^2) \\ &= O(N^2). \end{aligned} \tag{A.4}$$

Finally, for the three covariance terms in (A.2), we argue as follows. Let X_1 and X_2 be two random variables and let ρ_{X_1,X_2} denote the coefficient of correlation between X_1 and X_2. Since $|\rho_{X_1,X_2}| \leq 1$, it follows that

$$|\mathrm{Cov}(X_1, X_2)| \leq \sqrt{V(X_1)}\sqrt{V(X_2)} \leq \max\{V(X_1), V(X_2)\},$$

where $V(X_1)$ and $V(X_2)$ denote the variance of X_1 and X_2 respectively. Therefore, $\mathrm{Cov}(X_1, X_2)$ is of order smaller than $\max\{O(V(X_1)), O(V(X_2))\}$.

The three covariance terms in (A.2) are thus of order smaller than max $\{O(N^3/n^2), O(N^2)\}$. Therefore, we have $V_mE_{pq}(\hat{t}^I_{dy} - t_{dy})^2$ is $\max\{O(N^3/n^2), O(N^2)\}$. Combining (i) and (ii), we have

$$\frac{V_mE_{pq}(\hat{t}^I_{dy} - t_{dy})^2}{[E_{mpq}(\hat{t}^I_{dy} - t_{dy})^2]^2} = \frac{\max\{O(N^3/n^2), O(N^2)\}}{O(N^4/n^2)} = \max\Big\{O\Big(\frac{1}{N}\Big), O\Big(\frac{n^2}{N^2}\Big)\Big\} \to 0,$$

$$\text{as } N \to \infty \text{ and } \frac{n}{N} \to 0.$$

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *J. Roy. Statist. Soc. Ser. B* **67**, 445-458.

Chambers, R. L. (2005). Substitution imputation vs. nearest neighbour imputation: An application to estimation of a distribution. *The Imputation Bulletin* **5**, 3-8.

Brick, J. M., Kalton, G. and Kim, J. K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology* **30**, 57-66.

Da Silva, D. N. and Opsomer, J. D. (2004). Properties of the weighting cell estimator under a nonparametric response mechanism. *Survey Methodology* **30**, 45-55.

Fay, B. E. (1991). A design-based perspective on missing data variance. *Proceedings of the Annual Research Conference*, 429-440, U.S. Bureau of the Census.

Haziza, D. and Rao, J. N. K. (2006), A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology* **32**, 53-64.

Kim, J. K., Brick, J. M., Fuller, W. A. and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *J. Roy. Statist. Soc. Ser. B* **68**, 509-521.

Kott, P. S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods*, 380-383, American Statistical Association.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9**, 538-573.

Meng, X.-L. and Romero, M. (2003). Discussion: Efficiency and self-efficiency with multiple imputation inference. *Internat. Statist. Rev.* **71**, 607-618.

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.

Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology* **26**, 79-85.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *J. Amer. Statist. Assoc.* **94**, 254-265.

Valliant, R., Dorfman, A. and Royall, R. M. (2000). *Finite Population Sampling: A Prediction Approach*. John Wiley, New-York.

Wegman, E. J. and Wright, I. W. (1983). Splines in Statistics. *J. Amer. Statist. Assoc.* **78**, 351-365.

Statistics Canada, Tunney's Pasture, R.H. Coats Building, 16th floor, Ottawa, K1A 0T6, Canada.

E-mail: Jean-Francois.Beaumont@statcan.gc.ca

Département de mathématiques et de statistique, Université de Montréal, Montreal, H3C 3J7, Canada.

E-mail: David.Haziza@umontreal.ca

Statistics Canada, Tunney's Pasture, R.H. Coats Building, 16th floor, Ottawa, K1A 0T6, Canada.

E-mail: Cynthia.Bocci@statcan.gc.ca

(Received January 2009; accepted January 2010)