Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Variance estimation under composite imputation: The methodology behind SEVANI

by Jean-François Beaumont and Joël Bissonnette

December 2011

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 171-179
Statistics Canada, Catalogue No. 12-001-X

Variance estimation under composite imputation: The methodology behind SEVANI

Jean-François Beaumont and Joël Bissonnette [1]

Abstract

Composite imputation is often used in business surveys. The term "composite" means that more than a single imputation method is used to impute missing values for a variable of interest. The literature on variance estimation in the presence of composite imputation is rather limited. To deal with this problem, we consider an extension of the methodology developed by Särndal (1992). Our extension is quite general and easy to implement provided that linear imputation methods are used to fill in the missing values. This class of imputation methods contains linear regression imputation, donor imputation and auxiliary value imputation, sometimes called cold-deck or substitution imputation. It thus covers the most common methods used by national statistical agencies for the imputation of missing values. Our methodology has been implemented in the System for the Estimation of Variance due to Nonresponse and Imputation (SEVANI) developed at Statistics Canada. Its performance is evaluated in a simulation study.

Key Words: Auxiliary value imputation; Composite imputation; Donor imputation; Imputation model; Linear imputation; Regression imputation; SEVANI.

1. Introduction

Composite imputation is often used in business surveys. The term "composite" means that more than a single imputation method is used to impute missing values for a variable of interest. The choice of a method over another one depends on the availability of auxiliary variables. For instance, ratio imputation could be used to impute a missing value when an auxiliary value is available; otherwise, mean imputation could be an alternative.

The problem of estimating the variance in the presence of a single imputation method has been extensively studied in the literature; *e.g.*, two excellent reviews of this topic are: Lee, Rancourt and Särndal (2001) and Haziza (2009). Although the use of composite imputation occurs frequently in practice, there is little literature on estimating its variance. The literature includes a jackknife variance estimator that was proposed and evaluated empirically in Rancourt, Lee and Särndal (1993). Sitter and Rao (1997) developed further the theory and obtained design-consistent linearization and jackknife variance estimators. In both papers, two imputation methods were considered, with ratio imputation being one of the methods, simple random sampling was used and uniform nonresponse was assumed. Later, Felx and Rancourt (2001) extended the general methodology proposed in Särndal (1992) and Deville and Särndal (1994) to composite imputation using simplifying assumptions. Finally, Shao and Steel (1999) developed an interesting and general reverse approach to variance estimation to deal with composite imputation (see also Kim and Rao 2009). Shao and Steel (1999) claimed that their reverse approach leads to derivations that are less involved than those found in Deville and Särndal (1994). We do not fully agree with this statement. Our results indicate that, in general, our extension to Särndal's approach actually leads to simpler derivations than those obtained with the Shao and Steel approach. The reverse approach may however become quite attractive when the sampling fraction is negligible and a replication variance estimation technique is chosen (see section 7 for greater detail).

We consider the methodology proposed by Särndal (1992) as a starting point. It requires the validity of an imputation model; *i.e.*, a model for the variable being imputed. At first glance, the extension of this methodology to composite imputation seems to be quite tedious, as noted by Shao and Steel (1999), until we notice that most imputation methods used in practice lead to imputed estimators that are linear in the observed values of the variable of interest. This considerably simplifies the derivation of a variance estimator even when there is a single imputation method. For the estimation of the sampling portion of the overall variance, we use a methodology (see Beaumont and Bocci 2009) that is slightly different than the one proposed by Särndal (1992). This allows us to simplify the derivations further. This research has been implemented in version 2 of the System for the Estimation of Variance due to Nonresponse and Imputation (SEVANI), which is developed at Statistics Canada (see Beaumont, Bissonnette and Bocci 2010).

The paper is structured as follows. In section 2, some notation is introduced and composite imputation is explained. Linear imputation is defined in section 3. Our

1. Jean-François Beaumont, Statistics Canada, Statistical Research and Innovation Division, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6. E-mail: jean-francois.beaumont@statcan.gc.ca; Joël Bissonnette, Statistics Canada, Business Survey Methods Division, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6. E-mail: joel.bissonnette@statcan.gc.ca.

approach to inference and our main assumptions are described in section 4. In section 5, a number of results are stated regarding variance estimation under composite imputation. Section 6 presents the results of a simulation study that assesses the performance of our variance estimator. The reverse approach is briefly discussed in section 7 to highlight the differences with our approach. Finally, a short conclusion is given in section 8.

2. What is composite imputation?

Suppose that we are interested in estimating the population domain total $\theta = \sum_{k \in U} d_k y_k$, where U is the finite population of size N, y is the variable of interest and d is a domain indicator variable indicating whether population unit k is in the domain of interest ($d_k = 1$) or not ($d_k = 0$). A sample s of size n is selected from the finite population U according to a probability sampling design $p(s)$. In the absence of missing values, θ can be estimated by the Horvitz-Thompson estimator $\hat{\theta} = \sum_{k \in s} w_k d_k y_k$, where $w_k = 1/\pi_k$ is the design weight and π_k is the selection probability of unit k. Although it is possible to extend our results to calibration estimators, it is not considered in this paper to keep matters simple.

Variable y can be missing for some of the sampled units but we assume that the domain indicator variable d is always observed for those units. The set of sampled units with an observed y-value, called the set of respondents, is denoted by s_r. It is assumed to have been generated according to a nonresponse mechanism $q(s_r \mid s)$. The set of nonrespondents is denoted by $s_m = s - s_r$. It is further split into J mutually exclusive subsets, $s_m^{(j)}$, $j = 1, ..., J$, such that $s_m = \bigcup_{j=1}^{J} s_m^{(j)}$, if composite imputation with $J > 1$ imputation methods is used. All the missing y-values within a given subset $s_m^{(j)}$ are imputed with the same method j. However, different imputation methods are used to impute missing values in different subsets. The resulting imputed estimator can be expressed as

$$\hat{\theta}_I = \sum_{k \in s_r} w_k d_k y_k + \sum_{k \in s_m} w_k d_k y_k^* = \sum_{k \in s_r} w_k d_k y_k + \sum_{j=1}^{J} \sum_{k \in s_m^{(j)}} w_k d_k y_k^*, \qquad (2.1)$$

where y_k^* is the imputed y-value for unit k.

Composite imputation is quite frequent in business surveys. It is used because there are missing values in auxiliary variables used for imputation. To fix ideas, let $\mathbf{x}_k$ be the complete vector of auxiliary variables for unit k. Ideally, all the missing y-values would be imputed using a single imputation method based on the complete vector $\mathbf{x}_k$. Unfortunately, there may be missing values in the auxiliary variables so that, for some nonrespondents, we cannot use $\mathbf{x}_k$ to impute their missing y-value; we can only use a subset of $\mathbf{x}_k$. We denote as $\mathbf{x}_k^{\text{obs}}$, the vector of observed auxiliary variables for unit k. This vector does not necessarily contain the same observed variables from one unit to the next. To impute the missing y-value of a given unit k, an imputation method is chosen based on the available auxiliary variables $\mathbf{x}_k^{\text{obs}}$. Since there may be a number of nonresponse patterns in the complete vector of auxiliary variables, the imputation strategy may contain a number of imputation methods.

Example:

The variance estimation issues raised by composite imputation can be better understood by considering the following example. Suppose that the complete vector of auxiliary variables for unit k is $\mathbf{x}_k = (x_{1k}, x_{2k})$, where x_{1k} is strongly related to y_k but subject to missing values while x_{2k} is set to a constant for all sampled units ($x_{2k} = 1$, $k \in s$). Ideally, x_{1k} is used to impute y_k if it is missing. If x_{1k} is not available, only x_{2k} can be used. Table 1 summarizes the information available for the different subsets of the sample s.

Table 1
Available information when there is one auxiliary variable x_1 and a constant x_2

Subsets		y	x_1	x_2	$\mathbf{x}^{\text{obs}}$
s_r	$s_r^{(1)}$	O	O	O	(x_1, x_2)
	$s_r^{(2)}$	O	M	O	(M, x_2)
s_m	$s_m^{(1)}$	M	O	O	(x_1, x_2)
	$s_m^{(2)}$	M	M	O	(M, x_2)

O: Observed; M: Missing.

The set of nonrespondents s_m is divided into the subsets $s_m^{(1)}$ and $s_m^{(2)}$ depending on the availability of x_1. Similarly, the set of respondents is divided into subsets $s_r^{(1)}$ and $s_r^{(2)}$. In this example, we could use ratio imputation to impute missing y-values in $s_m^{(1)}$ and mean imputation to impute missing y-values in $s_m^{(2)}$. Note that simple linear regression imputation could be used instead of ratio imputation (if it better fits the data). We have chosen ratio imputation in this example for its simplicity and because it is frequently used in business surveys.

Only the respondents in $s_r^{(1)}$ can be used to impute missing y-values in $s_m^{(1)}$ through ratio imputation. The imputed value for a unit k in $s_m^{(1)}$ is $y_k^* = x_{1k} \sum_{l \in s_r^{(1)}} \omega_l^{(1)} y_l / \sum_{l \in s_r^{(1)}} \omega_l^{(1)} x_{1l}$, where $\omega_l^{(1)}$ is some weight used for ratio imputation (imputation method 1). Typical choices are: $\omega_l^{(1)} = w_l$ (design-weighted imputation) or $\omega_l^{(1)} = 1$ (unweighted imputation). For mean imputation, the respondents in $s_r^{(2)}$ as well as those in $s_r^{(1)}$ can be used to impute

missing y-values in $s_m^{(2)}$. In practice, it is common to use both sets of respondents to improve the stability of the imputed mean. The imputed value for a unit k in $s_m^{(2)}$ is

$$y_k^* = \sum_{l \in s_r} \omega_l^{(2)} y_l \Big/ \sum_{l \in s_r} \omega_l^{(2)},$$

where $\omega_l^{(2)}$ is a weight used for mean imputation (imputation method 2). (Typical choices of $\omega_l^{(2)}$ are the same as those for $\omega_l^{(1)}$; *i.e.*, $\omega_l^{(2)} = w_l$ or $\omega_l^{(2)} = 1$.) This implies that units in $s_r^{(1)}$ can be contributors to both imputation methods. This raises issues for variance estimation of the resulting composite imputation estimator. These issues will be addressed in section 5.

3. What is linear imputation?

The imputation method j is said to be linear if the imputed value y_k^* for a sample unit $k \in s_m^{(j)}$ can be written in the linear form

$$y_k^* = \varphi_{0k}^{(j)} + \sum_{l \in s_r} \varphi_{lk}^{(j)} y_l. \qquad (3.1)$$

The quantities $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$, for $l \in s_r$, are obtained without using y-values, but may depend on s and s_r. The linear form (3.1) is satisfied by several of the most common imputation methods in practice such as (weighted or unweighted) linear regression imputation, donor imputation and auxiliary value imputation. A nice review of these methods is found in Haziza (2009). Note that auxiliary value imputation does not use the y-values of respondents; *i.e.*, $y_k^* = \varphi_{0k}^{(j)}$ (see Beaumont, Haziza and Bocci 2011). For donor imputation, the imputed value y_k^* is equal to the y-value of a suitably chosen respondent (donor) so that $\varphi_{0k}^{(j)} = 0$ and $\varphi_{lk}^{(j)} = 0$ for all but one respondent $l \in s_r$. Detailed expressions for $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ are given in the Methodology Guide of SEVANI (Beaumont, Bissonnette and Bocci 2010), which is available on request from the authors.

Let $\Omega_I^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k^*$ be the contribution of imputation method j to the estimator $\hat{\theta}_I$. Using (3.1), $\Omega_I^{(j)}$ can be decomposed as follows:

$$\begin{aligned}\Omega_I^{(j)} &= \sum_{k \in s_m^{(j)}} w_k d_k y_k^* \\ &= \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{0k}^{(j)} + \sum_{l \in s_r} y_l \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{lk}^{(j)} \\ &= W_{0d}^{(j)} + \sum_{l \in s_r} W_{dl}^{(j)} y_l \,, \end{aligned} \qquad (3.2)$$

where $W_{0d}^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{0k}^{(j)}$ and $W_{dl}^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{lk}^{(j)}$. Using (3.2), the imputed estimator (2.1) can be expressed in the linear form:

$$\begin{aligned}\hat{\theta}_I &= \sum_{k \in s_r} w_k d_k y_k + \sum_{j=1}^{J} \Omega_I^{(j)} \\ &= W_{0d}^{(+)} + \sum_{k \in s_r} (w_k d_k + W_{dk}^{(+)}) y_k \,, \end{aligned} \qquad (3.3)$$

where $W_{0d}^{(+)} = \sum_{j=1}^{J} W_{0d}^{(j)}$ and $W_{dk}^{(+)} = \sum_{j=1}^{J} W_{dk}^{(j)}$.

Continuing with the example introduced at the end of section 2, we observe that, for ratio imputation, $\varphi_{0k}^{(1)} = 0$ and $\varphi_{lk}^{(1)} = \omega_l^{(1)} x_{1k} / \sum_{l \in s_r} \omega_l^{(1)} x_{1l}$, for $l \in s_r$, with $\omega_l^{(1)} = 0$, for $l \in s_r^{(2)}$. For mean imputation, we have $\varphi_{0k}^{(2)} = 0$ and $\varphi_{lk}^{(2)} = \omega_l^{(2)} / \sum_{l \in s_r} \omega_l^{(2)}$, for $l \in s_r$. Consequently, $W_{0d}^{(1)} = 0$, $W_{0d}^{(2)} = 0$,

$$W_{dl}^{(1)} = \omega_l^{(1)} \sum_{k \in s_m^{(1)}} w_k d_k x_{1k} \Big/ \sum_{k \in s_r} \omega_k^{(1)} x_{1k}$$

and $W_{dl}^{(2)} = \omega_l^{(2)} \sum_{k \in s_m^{(2)}} w_k d_k / \sum_{k \in s_r} \omega_k^{(2)}$. This implies that $W_{0d}^{(+)} = 0$ and $W_{dk}^{(+)} = W_{dk}^{(1)} + W_{dk}^{(2)}$.

4. Approach to inference and main assumptions

We consider three sources of variability when evaluating expectations and variances of the imputed estimator: the variability due to the imputation model, the sampling design and the nonresponse mechanism. Note that the use of an imputation model to make inference in the presence of imputation can be found in Rubin (1987), Hidiroglou (1989) and Särndal (1992). In what follows, we will use the subscripts m, p and q to denote the expectations, variances and covariances evaluated with respect to the imputation model, sampling design and nonresponse mechanism respectively.

We consider the following imputation model to describe the relationship between the y-variable and the vector $\mathbf{x}^{\text{obs}}$ of observed auxiliary variables:

$$\begin{aligned} E_m(y_k \mid \mathbf{X}^{\text{obs}}) &= \mu_k \\ V_m(y_k \mid \mathbf{X}^{\text{obs}}) &= \sigma_k^2 \\ \text{cov}_m(y_k, y_l \mid \mathbf{X}^{\text{obs}}) &= 0 \,, \end{aligned} \qquad (4.1)$$

for $k \neq l$ and $k, l \in U$. The population matrix $\mathbf{X}^{\text{obs}}$ contains the vectors of observed auxiliary variables, $\mathbf{x}_k^{\text{obs}}$, for $k \in U$, and μ_k and σ_k^2 are functions of $\mathbf{x}_k^{\text{obs}}$. Asymptotically m-unbiased and m-consistent estimators of μ_k and σ_k^2 are denoted by $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ respectively. Since we will always condition on $\mathbf{X}^{\text{obs}}$, we exclude this conditioning from the notation to simplify it. For instance, $E_m(y_k \mid \mathbf{X}^{\text{obs}})$ will be written as $E_m(y_k)$.

In model (4.1), we condition on the observed auxiliary variables. Since the nonresponse pattern in the vector $\mathbf{x}$ is not the same for all the nonrespondents, a separate conditional model must be validated and fitted for each nonresponse pattern. In principle, these conditional models should be used to determine the imputation methods chosen.

Note that model (4.1) reduces to the standard conditional model (*e.g.*, Särndal 1992) when the vector **x** of auxiliary variables is not subject to missing values.

Remark: The validity of the variance estimation method in section 5 requires μ_k and σ_k^2 to be correctly specified. Although a parametric form for μ_k may often be acceptable, it may be more difficult to determine a suitable parametric form for σ_k^2. To avoid this issue and obtain some robustness against misspecification of the model variance, σ_k^2 can be estimated non parametrically; see the empirical study of Beaumont, Haziza and Bocci (2011) for an illustration of this property under auxiliary value imputation. In the context of donor imputation, Beaumont and Bocci (2009) showed empirically that nonparametric estimation of both μ_k and σ_k^2, via penalized smoothing splines, reduced significantly the vulnerability of our variance estimator to misspecifications of the model mean and variance.

In addition to the imputation model (4.1), we also assume that:

$$F(\mathbf{Y} \mid s, s_r, \mathbf{X}^{\text{obs}}, \mathbf{Z}, \mathbf{D}) = F(\mathbf{Y} \mid \mathbf{X}^{\text{obs}}), \qquad (4.2)$$

where $F(\cdot)$ denotes the distribution function, **Y** and **D** are *N*-element vectors containing respectively y_k and d_k as their k^{th} element, and **Z** is a *N*-row matrix of design information, which implicitly or explicitly contains information about the selection probabilities π_k and joint selection probabilities π_{kl}, for $k, l \in U$. This assumption, often implicit in other papers, allows us to treat the response indicators, the domain indicators and the design information as fixed when taking model expectations and variances. A careful choice of the auxiliary variables is necessary to satisfy this assumption. For instance, the design information and the domain indicators should be considered as potential auxiliary variables.

The imputation strategy given in our example started in section 2 could be justified by a model with $\mu_k = \beta_1 x_{1k}$ and $\sigma_k^2 = \sigma_1^2 x_{1k}$, for $k \in s_r^{(1)}$ or $k \in s_m^{(1)}$, and $\mu_k = \beta_2$ and $\sigma_k^2 = \sigma_2^2$, for $k \in s_r^{(2)}$ or $k \in s_m^{(2)}$. The model parameters β_1, β_2, σ_1^2 and σ_2^2 are unknown. Note that if the x_{1k}'s are assumed to be identically distributed random variables with mean μ_x and variance σ_x^2, then $\beta_2 = \beta_1 \mu_x$ and $\sigma_2^2 = \beta_1^2 \sigma_x^2 + \sigma_1^2 \mu_x$. The imputed values $y_k^* = \hat{\mu}_k$, for $k \in s_m$, are obtained by estimating the model parameters β_1 and β_2 from the observed data. For instance, the *m*-unbiased estimators of β_1 and β_2 could be chosen as

$$\hat{\beta}_1 = \sum_{k \in s_r^{(1)}} \omega_k^{(1)} y_k \Big/ \sum_{k \in s_r^{(1)}} \omega_k^{(1)} x_{1k}$$

and

$$\hat{\beta}_2 = \sum_{k \in s_r^{(2)}} \omega_k^{(2)} y_k \Big/ \sum_{k \in s_r^{(2)}} \omega_k^{(2)}$$

respectively. This would lead to $\hat{\mu}_k = \hat{\beta}_1 x_{1k}$, for $k \in s_r^{(1)}$ or $k \in s_m^{(1)}$, and $\hat{\mu}_k = \hat{\beta}_2$, for $k \in s_r^{(2)}$ or $k \in s_m^{(2)}$. As in section 2, one could also consider the potentially more efficient estimator $\hat{\beta}_2^* = \sum_{k \in s_r} \omega_k^{(2)} y_k / \sum_{k \in s_r} \omega_k^{(2)}$ instead of $\hat{\beta}_2$. Unfortunately, $\hat{\beta}_2^*$ is biased under the model since

$$E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(2)} (x_{1k} \beta_1 - \beta_2)}{\sum_{k \in s_r} \omega_k^{(2)}}. \qquad (4.3)$$

As pointed out above, if the x_{1k}'s are assumed to be identically distributed random variables with mean μ_x and variance σ_x^2, $\beta_2 = \beta_1 \mu_x$ and equation (4.3) can be rewritten as

$$E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + \beta_1 \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(2)}}{\sum_{k \in s_r} \omega_k^{(2)}} \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(2)} (x_{1k} - \mu_x)}{\sum_{k \in s_r^{(1)}} \omega_k^{(2)}}. \qquad (4.4)$$

It can be shown under weak conditions that $E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + O_p(1/\sqrt{n})$ so that the model bias of $\hat{\beta}_2^*$ is asymptotically negligible. However, since $\text{var}_m(\hat{\beta}_2^* \mid s, s_r) = O_p(1/n)$, the squared model bias is not necessarily asymptotically negligible compared to the model variance of $\hat{\beta}_2^*$. At least, $\hat{\beta}_2^*$ is *m*-consistent for β_2. From (4.3) or (4.4), we can see that the model bias of $\hat{\beta}_2^*$ can be controlled by assigning a smaller weight $\omega_k^{(2)}$ to units $k \in s_r^{(1)}$ relative to units $k \in s_r^{(2)}$. For instance, one could consider using $\omega_k^{(2)} = w_k / n^{\alpha}$, for $k \in s_r^{(1)}$ and some $\alpha > 0$, and $\omega_k^{(2)} = w_k$, for $k \in s_r^{(2)}$. In the extreme case where $\omega_k^{(2)} = 0$, for $k \in s_r^{(1)}$, $\hat{\beta}_2^*$ is model-unbiased because it is equal to $\hat{\beta}_2$. Note that the model bias of $\hat{\beta}_2^*$ could be larger than $O_p(1/\sqrt{n})$ if x_{1k}, $k \in s_r^{(1)}$, have a mean different from x_{1k}, $k \in s_r^{(2)}$. In such case, controlling the model bias of $\hat{\beta}_2^*$ might be more important.

In the case of donor imputation, a fourth source of variability needs to be considered when donors are randomly selected among respondents to impute nonrespondents. In this paper, the subscript q will implicitly indicate that moments are evaluated with respect to the joint distribution induced by the nonresponse mechanism and the random donor selection mechanism. As a result, when conditioning on s_r, as in (4.2), it should be kept in mind that conditioning is not only on the set of respondents but also on the set of selected donors.

5. Variance estimation

Särndal (1992) expresses the total error of the imputed estimator as:

$$\hat{\theta}_I - \theta = (\hat{\theta} - \theta) + (\hat{\theta}_I - \hat{\theta}), \qquad (5.1)$$

where the first term on the right-hand side of (5.1) is called the sampling error and the second term is called the nonresponse error. Using the assumptions given in section 4 and $E_p(\hat{\theta} - \theta) = 0$, the overall bias of the imputed estimator reduces to $E_{mpq}(\hat{\theta}_I - \theta) = E_{pq}B_m$, where $B_m = E_m(\hat{\theta}_I - \hat{\theta} \mid s, s_r)$ is the (conditional) model bias of the imputed estimator. Using (2.1), the model bias can be expressed as

$$B_m = \sum_{j=1}^{J} \sum_{k \in s_m^{(j)}} w_k d_k E_m(y_k^* - y_k \mid s, s_r). \qquad (5.2)$$

This means that the model bias and the overall bias vanish if the model expectation of the imputation error, $y_k^* - y_k$, is zero, for $k \in s_m^{(j)}$ and $j = 1, ..., J$. In principle, an imputation strategy should be chosen so that this condition is satisfied (at least approximately). This is typically assumed in the literature (*e.g.*, Särndal 1992; Shao and Steel 1999).

In the example introduced in section 2, the model bias (5.2) reduces to

$$B_m = \left(\sum_{k \in s_m^{(2)}} w_k d_k \right) E_m(\hat{\beta}_2^* - \beta_2 \mid s, s_r).$$

An expression for $E_m(\hat{\beta}_2^* - \beta_2 \mid s, s_r)$ is given by (4.3) or (4.4). As noted in the paragraph that follows equation (4.4), the model bias, B_m, can be controlled by assigning a smaller weight $\omega_k^{(2)}$ to units $k \in s_r^{(1)}$ relative to units $k \in s_r^{(2)}$. It is also small if the number of nonrespondents imputed by method 2 is small. Note that our variance (or Mean Squared Error, MSE) estimation approach requires the slightly weaker assumption that $E_q(B_m \mid s)$ is negligible (see section 5.3).

Using (5.1), Särndal (1992) decomposed the overall MSE into three components:

$$E_{mpq}(\hat{\theta}_I - \theta)^2 = E_m \mathrm{var}_p(\hat{\theta}) + E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\} + 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\}. \qquad (5.3)$$

The overall MSE (5.3) becomes approximately equivalent to the overall variance, $\mathrm{var}_{mpq}(\hat{\theta}_I - \theta)$, when the overall bias is negligible. The first, second and third terms on the right-hand side of (5.3) are referred to as the sampling variance, the nonresponse variance and the mixed component respectively. The sum of the last two terms can be called the nonresponse component since these terms would disappear if there were no nonresponse. The nonresponse component is simply the difference between the overall MSE/variance and the sampling variance. In what follows, we develop an estimator for each of these three terms.

5.1 Estimation of the sampling variance

Let $v(y)$ be a *p*-unbiased estimator of $\mathrm{var}_p(\hat{\theta})$ that would be used under complete response. The typical Horvitz-Thompson estimator is

$$v(y) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (w_k d_k y_k)(w_l d_l y_l), \qquad (5.4)$$

where π_{kl} is the joint selection probability of units *k* and *l*. In the presence of nonresponse, $\hat{V}_{\mathrm{ORD}} = v(y_\bullet)$ is the naïve sampling variance estimator that treats the imputed values as true values, where $y_\bullet$ is the imputed *y*-variable; *i.e.*, $y_{\bullet k} = y_k$, for $k \in s_r$, and $y_{\bullet k} = y_k^*$, for $k \in s_m$.

Särndal (1992) proposed the following *mpq*-unbiased estimator of the sampling variance $V_{\mathrm{SAM}} = E_m \mathrm{var}_p(\hat{\theta})$:

$$\hat{V}_{\mathrm{SAM}} = \hat{V}_{\mathrm{ORD}} + \hat{V}_{\mathrm{DIF}},$$

where $\hat{V}_{\mathrm{DIF}}$ is an *m*-unbiased estimator of $V_{\mathrm{DIF}} = E_m(v(y) - \hat{V}_{\mathrm{ORD}} \mid s, s_r)$. Unfortunately, the expression for $\hat{V}_{\mathrm{DIF}}$ is usually tedious to derive, and it is even more so when composite imputation is used.

Beaumont and Bocci (2009) simplified Särndal's derivations by conditioning on $\mathbf{Y}_r$, the vector containing the responding *y*-values. More explicitly, let $V_{\mathrm{DIF}}^{\mathrm{C}} = E_m(v(y) - \hat{V}_{\mathrm{ORD}} \mid s, s_r, \mathbf{Y}_r)$ and $\hat{V}_{\mathrm{DIF}}^{\mathrm{C}}$ be an *m*-unbiased estimator of $V_{\mathrm{DIF}}^{\mathrm{C}}$; *i.e.*, $E_m(\hat{V}_{\mathrm{DIF}}^{\mathrm{C}} \mid s, s_r, \mathbf{Y}_r) = V_{\mathrm{DIF}}^{\mathrm{C}}$. Our *mpq*-unbiased sampling variance estimator is $\hat{V}_{\mathrm{SAM}}^{\mathrm{C}} = \hat{V}_{\mathrm{ORD}} + \hat{V}_{\mathrm{DIF}}^{\mathrm{C}}$. Since $\hat{V}_{\mathrm{ORD}}$ is a constant when conditioning on *s*, s_r and $\mathbf{Y}_r$, $\hat{V}_{\mathrm{SAM}}^{\mathrm{C}}$ can simply be obtained by estimating $E_m(v(y) \mid s, s_r, \mathbf{Y}_r)$. If (5.4) is used,

$$E_m(v(y) \mid s, s_r, \mathbf{Y}_r) = v(y_\bullet^{\mu}) + \sum_{k \in s_m} (1 - \pi_k) w_k^2 d_k \sigma_k^2, \qquad (5.5)$$

where $y_{\bullet k}^{\mu} = y_k$, for $k \in s_r$, and $y_{\bullet k}^{\mu} = \mu_k$, for $k \in s_m$. An estimator $\hat{V}_{\mathrm{SAM}}^{\mathrm{C}}$ of (5.5) is obtained by replacing the unknown mean μ_k and unknown variance σ_k^2 in (5.5) by *m*-unbiased (or at least *m*-consistent) estimators $\hat{\mu}_k$ and $\hat{\sigma}_k^2$. This estimator is easy to compute provided a software package that treats the complete response case is available to obtain the first term on the right-hand side of (5.5). The general formula (5.5) can be used for every imputation strategy. The only difference between different imputation strategies lies in the choice of the imputation model and the estimators $\hat{\mu}_k$ and $\hat{\sigma}_k^2$.

5.2 Estimation of the nonresponse variance

An *mpq*-unbiased estimator of the nonresponse variance $V_{\mathrm{NR}} = E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\}$ is obtained by finding an *m*-unbiased estimator of

$$E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\} = \mathrm{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} + B_m^2. \quad (5.6)$$

Using $\hat{\theta}_I$ defined in the first equation of (3.3), the nonresponse error with composite imputation can be decomposed into J components:

$$\hat{\theta}_I - \hat{\theta} = \sum_{j=1}^{J} (\Omega_I^{(j)} - \Omega^{(j)}),$$

where $\Omega^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k$. Each of these J components, $\Omega_I^{(j)} - \Omega^{(j)}$, is associated with a different imputation method. Since y_k^* only involves observed y-values, $\Omega_I^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k^*$ only involves observed y-values as well and thus $\Omega_I^{(j)}$ and $\Omega^{(j)}$ are independent under the model. Therefore, the model variance of the nonresponse error can be written as

$$\mathrm{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} = \sum_{i=1}^{J}\sum_{j=1}^{J} \mathrm{cov}_m(\Omega_I^{(i)}, \Omega_I^{(j)} \mid s, s_r) + \sum_{j=1}^{J} \mathrm{var}_m(\Omega^{(j)} \mid s, s_r). \quad (5.7)$$

Note that the covariances $\mathrm{cov}_m(\Omega_I^{(i)}, \Omega_I^{(j)} \mid s, s_r)$, for $i \neq j$, are not necessarily negligible because some observed y-values can be used for more than one imputation method.

The derivations of the model variance (5.7) could be quite involved when several imputation methods are used because of the non-negligible covariances. The algebra can be greatly simplified for linear imputation methods. By using the second equation given in (3.3), the nonresponse error can be expressed as

$$\hat{\theta}_I - \hat{\theta} = W_{0d}^{(+)} + \sum_{k \in s_r} W_{dk}^{(+)} y_k - \sum_{k \in s_m} w_k d_k y_k. \quad (5.8)$$

Since the nonresponse error is linear in the y-values, its model variance is given by

$$\mathrm{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} = \sum_{k \in s_r} (W_{dk}^{(+)})^2 \sigma_k^2 + \sum_{k \in s_m} w_k^2 d_k \sigma_k^2. \quad (5.9)$$

If the model bias B_m is negligible, an *mpq*-unbiased estimator $\hat{V}_{\mathrm{NR}}$ of the nonresponse variance V_{NR} is obtained by replacing σ_k^2 in (5.9) by an *m*-unbiased (and *m*-consistent) estimator $\hat{\sigma}_k^2$. If the model bias is not negligible, it can be estimated by an *m*-consistent estimator $\hat{B}_m$ and, using equation (5.6), the nonresponse variance estimator $\hat{V}_{\mathrm{NR}}$ can be replaced by $\hat{V}_{\mathrm{NR}} + \hat{B}_m^2$. Note that $\hat{B}_m^2$ is *m*-consistent for B_m^2 provided that $\hat{B}_m$ is *m*-consistent for B_m. The estimator $\hat{B}_m$ can be found by using (5.8) and writing the model bias as

$$B_m = E_m(\hat{\theta}_I - \hat{\theta} \mid s, s_r) = W_{0d}^{(+)} + \sum_{k \in s_r} W_{dk}^{(+)} \mu_k - \sum_{k \in s_m} w_k d_k \mu_k. \quad (5.10)$$

The estimator $\hat{B}_m$ is obtained by replacing μ_k in (5.10) by an *m*-consistent estimator $\hat{\mu}_k$.

5.3 Estimation of the mixed component

An *mpq*-unbiased estimator of the mixed component

$$V_{\mathrm{MIX}} = 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\}$$

is obtained by finding an *m*-unbiased estimator of

$$2E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\} = 2\,\mathrm{cov}_m\{(\hat{\theta}_I - \hat{\theta}), (\hat{\theta} - \theta) \mid s, s_r\} + 2B_m E_m\{(\hat{\theta} - \theta) \mid s, s_r\}. \quad (5.11)$$

Since both the nonresponse error and the sampling error are linear in the y-values, using (5.8) we obtain:

$$2\,\mathrm{cov}_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\} = 2\sum_{k \in s_r} W_{dk}^{(+)}(w_k - 1) d_k \sigma_k^2 - 2\sum_{k \in s_m} w_k (w_k - 1) d_k \sigma_k^2. \quad (5.12)$$

If the model bias B_m is negligible, an *mpq*-unbiased estimator $\hat{V}_{\mathrm{MIX}}$ of the mixed component V_{MIX} is obtained by replacing σ_k^2 in (5.12) by an *m*-unbiased (and *m*-consistent) estimator $\hat{\sigma}_k^2$. Note that the mixed component is not necessarily negligible (Brick, Kalton and Kim 2004) and, moreover, it has been found to often be negative in practice.

If the model bias B_m is not negligible, it may not be possible to easily estimate the second component on the right-hand side of (5.11). The reason is that $E_m\{(\hat{\theta} - \theta) \mid s, s_r\}$ involves knowing $\mathbf{x}_k^{\mathrm{obs}}$ as well as the domain indicator variable d for the nonsampled portion of the population; this information may not be available. This problem can be bypassed by changing the inferential framework. The full multivariate distribution between y, $\mathbf{x}$ and d can be modeled instead of conditioning on d and $\mathbf{x}^{\mathrm{obs}}$. We did not implement this idea in SEVANI because it leads to a more complex modeling task and makes it difficult to obtain a general variance expression that is easy to implement. Ignoring the second component on the right-hand side of (5.11) should not be of great concern in practice when the model bias is not too large. In section 5.4, we provide a diagnostic that can be helpful for determining whether the model bias is important or not.

The mixed component can also be written as

$$V_{\text{MIX}} = 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\}$$
$$= 2E_{pq}[\text{cov}_m\{(\hat{\theta}_I - \hat{\theta}), (\hat{\theta} - \theta) \mid s, s_r\}]$$
$$+ 2E_p[E_q(B_m \mid s)E_m\{(\hat{\theta} - \theta) \mid s\}].$$

Expression (5.12) can therefore be used to obtain an estimator of V_{MIX} provided that $E_q(B_m \mid s)$ is negligible. This is a weaker assumption than requiring B_m to be negligible since this assumption is satisfied when either B_m or $E_q(\hat{\theta}_I - \hat{\theta} \mid s)$ is negligible. For instance, in our earlier example, B_m may not be negligible but, if $d_k = 1$ and $\omega_k^{(1)} = \omega_k^{(2)} = w_k$, $E_q(\hat{\theta}_I - \hat{\theta} \mid s) \approx 0$ under uniform nonresponse (see Sitter and Rao 1997).

5.4 Estimation of the overall MSE/variance

The overall MSE, or overall variance if the overall bias is negligible,

$$V_{\text{TOT}} = E_{mpq}(\hat{\theta}_I - \theta)^2 = V_{\text{SAM}} + V_{\text{NR}} + V_{\text{MIX}}$$

can be estimated by $\hat{V}_{\text{TOT}} = \hat{V}_{\text{SAM}}^{\text{C}} + \hat{V}_{\text{NR}} + \hat{V}_{\text{MIX}}$ if the model bias, B_m, is negligible. The nonresponse component estimator is $\hat{V}_{\text{NR}} + \hat{V}_{\text{MIX}}$. From a user's perspective, the estimator $\hat{V}_{\text{TOT}}$ is of greater interest than its individual components. A user may nevertheless be interested in the estimator of the sampling variance, $\hat{V}_{\text{SAM}}^{\text{C}}$, or the ratio $\hat{V}_{\text{SAM}}^{\text{C}}/\hat{V}_{\text{TOT}}$. The latter estimates the contribution of the sampling variance to the overall variance.

As pointed out in section 5.2, if the model bias is not negligible, the nonresponse variance can be estimated by $\hat{V}_{\text{NR}} + \hat{B}_m^2$ instead of $\hat{V}_{\text{NR}}$. This leads to the overall MSE estimator $\hat{V}_{\text{TOT, ADJ}} = \hat{V}_{\text{SAM}}^{\text{C}} + (\hat{V}_{\text{NR}} + \hat{B}_m^2) + \hat{V}_{\text{MIX}}$.

A statistic that can be useful as a diagnostic to determine the magnitude of the model bias is either $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT}}}$ or $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT, ADJ}}}$. A large value of any of these two statistics may be an indication that the model bias is not negligible and that the composite imputation procedure should be questioned. The advantage of $|\hat{B}_m|/\sqrt{\hat{V}_{\text{TOT, ADJ}}}$ over $|\hat{B}_m|/\sqrt{\hat{V}_{\text{TOT}}}$ is that it is bounded; *i.e.*,

$$0 \le |\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT, ADJ}}} \le 1.$$

5.5 Random regression imputation

A random regression residual e_k is sometimes added to the regression imputed value y_k^* to preserve the natural variability of the *y*-variable. We suggest that the random residuals e_k be generated independently with $E_*(e_k \mid s, s_r) = 0$ and $\text{var}_*(e_k \mid s, s_r) = \hat{\sigma}_k^2$, where the subscript * indicates that the expectation and variance are taken with respect to the random imputation mechanism. This leads to the imputed value $y_k^{*R} = y_k^* + r_k e_k$, with $r_k = 1$ if unit k has been imputed with a random residual added and $r_k = 0$ otherwise. The imputed estimator (2.1) with y_k^* replaced by y_k^{*R} is denoted by $\hat{\theta}_I^* = \hat{\theta}_I + \sum_{k \in s_m} w_k d_k r_k e_k$. Since $E_*(e_k \mid s, s_r) = 0$, adding a random residual does not introduce any bias in the imputed estimator. The overall MSE of $\hat{\theta}_I^*$ can be expressed as

$$E_{mpq*}(\hat{\theta}_I^* - \theta)^2 = E_{mpq}(\hat{\theta}_I - \theta)^2 + E_{mpq}\text{var}_*(\hat{\theta}_I^* \mid s, s_r). \quad (5.13)$$

The first term on the right-hand side of (5.13) is estimated as in section 5.4. The second term is estimated by

$$\text{var}_*(\hat{\theta}_I^* \mid s, s_r) = \sum_{k \in s_m} w_k^2 d_k r_k \hat{\sigma}_k^2. \quad (5.14)$$

6. Simulation study

We conducted a Monte-Carlo simulation study to assess the methodology described in section 5. A bivariate population of $N = 400$ units was generated that contains an auxiliary variable *x* and a variable of interest *y*. For each population unit, the auxiliary variable was generated according to a gamma distribution with mean 48 and variance 768. The variable of interest *y* was generated conditionally on *x* from a gamma distribution with mean $1.5x$ and variance $16x$. Half of the population was randomly assigned a missing value to *x*. As no domain of interest was generated, θ is the overall population total of variable *y*.

Ten thousand samples were selected from this population using simple random sampling without replacement. We considered two sample sizes: $n = 100$ and $n = 250$. For each sample, nonresponse to variable *y* was generated independently from one unit to another with a nonresponse probability of 0.3. We used the same imputation strategy as in the example in section 2 with $\omega_l^{(1)} = 1$, for $l \in s_r^{(1)}$, and $\omega_l^{(2)} = 1$, for $l \in s_r$. Nonrespondents to variable *y* with an observed *x*-value were imputed by ratio imputation while those with a missing *x*-value were imputed by mean imputation.

The population *y*-values were kept fixed throughout the replications of the simulation experiment; each replication consisted of selecting a sample and then generating nonresponse to variable *y*. If we had strictly followed the theoretical development in section 5, we would have generated new *y*-values at each replication according to the imputation model. However, it is more common in the literature to fix the population *y*-values when conducting a simulation experiment. For instance, our simulation set-up is essentially the same as the one discussed in Rancourt, Lee and Särndal (1993), who also considered composite imputation.

We computed the Monte-Carlo sampling variance and overall MSE as $V_{\text{SAM}}^{\text{MC}} = \sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2 / R$ and $V_{\text{TOT}}^{\text{MC}} = \sum_{r=1}^{R}(\hat{\theta}_{I,r} - \theta)^2 / R$ respectively, where the subscript r indicates that estimates are computed using the r^{th} replicate and $R = 10{,}000$. The Monte-Carlo relative bias of any estimator of V_{SAM}, say v_{SAM}, is computed as $\text{RB}(V_{\text{SAM}}) = \sum_{r=1}^{R}(v_{\text{SAM},r} - V_{\text{SAM}}^{\text{MC}}) / (V_{\text{SAM}}^{\text{MC}} R)$. Similarly, we computed the Monte-Carlo relative bias of an estimator of V_{TOT}, denoted as $\text{RB}(V_{\text{TOT}})$, and the Monte-Carlo relative bias of an estimator of $V_{\text{SAM}} / V_{\text{TOT}}$, denoted as $\text{RB}(V_{\text{SAM}} / V_{\text{TOT}})$. Finally, we computed the Monte-Carlo coverage rates of confidence intervals for θ with a 95% confidence level assuming that $\hat{\theta}_I$ is normally distributed.

The results of our simulation study are given in table 2. In the columns labeled SEVANI, the sampling variance, V_{SAM}, and the overall MSE, V_{TOT}, are estimated for each sample by $\hat{V}_{\text{SAM}}^{\text{C}}$ and $\hat{V}_{\text{TOT, ADJ}}$ respectively (see section 5.4). We have also obtained results by replacing $\hat{V}_{\text{TOT, ADJ}}$ by $\hat{V}_{\text{TOT}}$. We do not report these additional results in table 2 as they were quite close to those obtained with $\hat{V}_{\text{TOT, ADJ}}$. This suggests that the model bias B_m is not important in this case. In the columns labeled Naïve, both the sampling variance and the overall MSE are estimated by $\hat{V}_{\text{ORD}}$ (see section 5.1).

Table 2
Results of the simulation study

	$n = 100$		$n = 250$	
	SEVANI	**Naïve**	**SEVANI**	**Naïve**
$\text{RB}(V_{\text{SAM}})$	2.82%	-17.59%	3.02%	-17.68%
$\text{RB}(V_{\text{SAM}} / V_{\text{TOT}})$	8.30%	-	5.84%	-
$\text{RB}(V_{\text{TOT}})$	-5.07%	-40.68%	-2.66%	-52.89%
Coverage Rate	93.38%	86.20%	94.42%	81.80%

These results show that the methodology described in section 5 and implemented in SEVANI is better than the naïve variance estimator for the estimation of the components of variance and the construction of confidence intervals. The use of SEVANI leads to small Monte-Carlo relative biases and coverage rates close to the targeted nominal rate (95%). Our methodology is also useful for users who would like to estimate the contribution of the sampling variance to the overall MSE; *i.e.*, $V_{\text{SAM}} / V_{\text{TOT}}$. Note that $V_{\text{SAM}}^{\text{MC}} / V_{\text{TOT}}^{\text{MC}}$ is 71.98% for $n = 100$ and 57.23% for $n = 250$. Since $V_{\text{SAM}}^{\text{MC}} / V_{\text{TOT}}^{\text{MC}}$ is not close to 100% even for $n = 100$, the effects of nonresponse and imputation cannot be systematically ignored when estimating the overall MSE.

7. The reverse approach

Shao and Steel (1999) proposed a reverse approach to variance estimation developed to deal with composite imputation. They assumed that the overall bias is negligible and suggested the following decomposition of the overall variance:

$$E_{mpq}(\hat{\theta}_I - \theta)^2 = E_{mq}\text{var}_p(\hat{\theta}_I \mid U_r) + E_{mq}\{E_p(\hat{\theta}_I \mid U_r) - \theta\}^2, \qquad (7.1)$$

where U_r is a conceptual population of respondents. The inner expectation and variance in the right side of (7.1) are taken with respect to the sampling design. Unfortunately, the imputed estimator $\hat{\theta}_I$ is generally not linear with respect to the sampling design even though it is linear with respect to the observed y-values. Therefore, the imputed estimator $\hat{\theta}_I$ is typically linearized (*e.g.*, Shao and Steel 1999; Kim and Rao 2009). More explicitly, the quantities $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ often depend on the sample in a nonlinear way; *e.g.*, this is true with linear regression imputation (see the example at the end of section 3) and donor imputation. It is not always straightforward to account for the sampling variability of $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ when using (7.1). For example, there is no literature on the use of the reverse approach to estimate the variance under nearest-neighbour imputation. Moreover, since each composite imputation strategy yields its own linearized imputed estimator, it is not an easy task to implement this methodology in a generalized software package.

Using our approach, the inner expectation in the expressions for the nonresponse variance,

$$V_{\text{NR}} = E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\},$$

and the mixed component,

$$V_{\text{MIX}} = 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\},$$

are taken with respect to the imputation model (conditionally on s and s_r). The imputed estimator is linear and the derivations are straightforward because the quantities $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ are constructed without using the y-values. The estimation of the sampling variance, $V_{\text{SAM}} = E_m \text{var}_p(\hat{\theta})$, does not involve these two quantities (see equation 5.5); thus, their possible non-linearity with respect to the sampling design does not cause any difficulty. This implies that nearest-neighbour imputation can be easily handled with our approach (see Beaumont and Bocci 2009).

It is for all the above reasons that we believe that the reverse approach might be more cumbersome to implement in a generalized software package than our approach. This

does not mean that the reverse approach is not useful. Indeed, both approaches lead to identical variance estimators when a census is conducted. Beaumont, Haziza and Bocci (2011) showed that they also lead to identical variance estimators under auxiliary value imputation (because $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ do not depend on s and s_r). Both approaches depend on the correct specification of the imputation model and no approach is expected to systematically outperform the other.

The reverse approach may have a practical advantage over our approach when the sampling fraction is negligible. In such case, Shao and Steel (1999) showed that the second component on the right side of (7.1) can be neglected. The first component is estimated by finding a design-based estimator of $\text{var}_p(\hat{\theta}_I \mid U_r)$. If a replication variance estimation technique (*e.g.*, the jackknife or the bootstrap) is chosen for the estimation of $\text{var}_p(\hat{\theta}_I \mid U_r)$, the whole approach becomes quite attractive and practical. Also, it does not depend on the validity of the imputation model; in particular, the correct specification of the model variance σ_k^2. The jackknife variance estimators of Rancourt, Lee and Särndal (1993) and Sitter and Rao (1997) can be justified by this approach.

8. Conclusion

Our methodology for composite imputation has been implemented in version 2 of SEVANI because of its ease of implementation and generality. It works for most imputation methods used in practice, as most imputation methods are linear. The variance computations are the same for every composite imputation strategy once the quantities $W_{0d}^{(+)}$, $W_{dk}^{(+)}$, $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ have been computed. This eases the development of a generalized system.

Although we have focused on the estimation of a domain total using the Horvitz-Thompson estimator, SEVANI can also deal with domain means and calibration estimators. Parametric and nonparametric methods of estimating μ_k and σ_k^2 are also available. Greater detail can be found in the Methodology Guide of SEVANI (Beaumont, Bissonnette and Bocci 2010) available upon request from the authors.

Acknowledgements

We would like to thank the reviewers for their comments. We would also like to thank Mike Hidiroglou, Eric Rancourt and Cynthia Bocci from Statistics Canada for their suggestions and discussions on the topic. All these comments contributed to improve the paper.

References

Beaumont, J.-F., Bissonnette, J. and Bocci, C. (2010). SEVANI, version 2.3, Methodology Guide. Internal report, Methodology Branch, Statistics Canada.

Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Beaumont, J.-F., Haziza, D. and Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515-537.

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

Deville, J.-C., and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Felx, P., and Rancourt, E. (2001). Applications of Variance due to Imputation in the Survey of Employment, Payrolls and Hours. Methodology Branch Working Paper, Statistics Canada, BSMD-2001-009E.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Sample Surveys*: *Theory, Methods and Inference*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: Elsevier BV, 29A, 215-246.

Hidiroglou, M.A. (1989). Unpublished handwritten notes kindly shared with us by the author.

Kim, J.-K., and Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Lee, H., Rancourt, E. and Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little). New-York: John Wiley & Sons, Inc., 315-328.

Rancourt, E., Lee, H. and Särndal, C.-E. (1993). Variance estimation under more than one imputation method. In *Proceedings of the International Conference on Establishments Surveys*, June 1993, Buffalo, American Statistical Association, 374-379.

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

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R.R., and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.