Survey Methodology

June 2004
How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line 1 800 263-1136
National telecommunications device for the hearing impaired 1 800 363-7629
 Depository Services Program inquiries 1 800 700-1033
 Fax line for Depository Services Program 1 800 889-9734
 E-mail inquiries infostats@statcan.ca
 Website www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.
Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.
Variance Estimation with Hot Deck Imputation Using a Model

J. Michael Brick, Graham Kalton and Jae Kwang Kim

Abstract

When imputation is used to assign values for missing items in sample surveys, naïve methods of estimating the variances of survey estimates that treat the imputed values as if they were observed give biased variance estimates. This article addresses the problem of variance estimation for a linear estimator in which missing values are assigned by a single hot deck imputation (a form of imputation that is widely used in practice). We propose estimators of the variance of a linear hot deck imputed estimator using a decomposition of the total variance suggested by Särndal (1992). A conditional approach to variance estimation is developed that is applicable to both weighted and unweighted hot deck imputation. Estimation of the variance of a domain estimator is also examined.

Key Words: Missing data; Model-assisted approach; Conditional variance estimation.

1. Introduction

The important practical problem of estimating the variance of an estimate computed from a data set in which some of the items are missing and values are assigned by imputation has been addressed in a number of different ways (e.g., see Rubin 1987 and Rao and Shao 1992). The approach used in this article is based on the model-assisted approach introduced by Särndal (1992). In the initial application, Särndal used the model-assisted approach with a simple random sample in which the missing data were imputed using deterministic ratio imputation. Subsequently, the approach has been extended to other imputation methods and sample designs (e.g., Deville and Särndal 1994; Rancourt, Särndal and Lee 1994; and Gagnon, Lee, Rancourt and Särndal 1996). This article extends the model-assisted approach to general forms of linear estimators in which missing values have been assigned by hot deck imputation within imputation cells. This form of hot deck imputation, which replaces a missing item by the value observed for a responding unit in the same cell, is one of the most frequently used methods of imputing for missing items in household sample surveys (Brick and Kalton 1996). This paper employs a conditional approach to develop a variance estimator for hot deck imputed estimators that is valid for general sample designs and a variety of estimation strategies.

In the model-assisted approach, the difference between an imputed estimator (the term used here to denote an estimator based in part on imputed values), \( \hat{\theta}_I \), and the corresponding finite population parameter, \( \theta_N \), is written as

\[
\hat{\theta}_I - \theta_N = (\hat{\theta}_n - \theta_N) + (\hat{\theta}_I - \hat{\theta}_n),
\]

where \( \hat{\theta}_n \) is the usual, approximately design unbiased, estimator of \( \theta_N \) with complete response. The first term on the right hand side of (1) is called the sampling error and depends only on the sampling distribution of the estimator based on the sample design used to select the full sample, denoted by \( p \). The second term is the imputation error; it depends on the sampling distribution, the response mechanism \( R \) that generates the respondents from the full sample, and the imputation mechanism \( I \) for filling in the missing values. This paper is restricted to estimators \( \hat{\theta}_I \) that involve only one variable subject to missing data.

We use a model-assisted approach that makes assumptions about the distribution of the variable of interest in the population. We refer to these assumptions as a superpopulation model, denoted by \( \xi \). In general, the aim of imputation is to create a multi-purpose data set that can be validly analyzed in many different ways, potentially involving the associations of a variable subject to imputation with any of the other variables in the data set. Since a superpopulation model is needed to impute for item non-responses in a way that preserves such associations, it is natural to use that approach also in variance estimation.

Under the superpopulation model, the total variance for an imputed estimator is given by

\[
V_{TOT} = E_\xi E_p E_R E_I (\hat{\theta}_I - \theta_N)^2,
\]

where \( E_\xi, E_p, E_R, \) and \( E_I \) refer to expectations with respect to the superpopulation model, the sampling mechanism, the response mechanism, and the imputation mechanism, respectively. We assume that the sample design, response mechanism, and the imputation mechanism are unconfounded as described by Rubin (1987) and used by Särndal (1992) and all of the other literature cited above.
on the model-assisted approach. Essentially, unconfounded mechanisms allow the order of the expectations to be changed so that the expectation with respect to the model can be taken first. Thus, the total variance can be re-written as \( V_{\text{TOT}} = E_p E_R E_I (\hat{\theta}_j - \theta_N)^2 \). Roughly speaking, unconfounded sampling, response, and imputation mechanisms imply that the mechanisms are independent of the distribution of the \( y \)-value being analyzed after conditioning on auxiliary variables (e.g., stratification variables for sampling or imputation cells for imputing). Thus, for example, we assume the value of the variable being imputed is independent of the probability of response within each hot-deck cell. Rubin (1987, pages 36–39) has a more detailed discussion of unconfounded mechanisms.

Using the decomposition given in equation (1), Särndal (1992) expressed the total variance for the imputed estimator as

\[
V_{\text{TOT}} = E_\xi E_p E_R E_I (\hat{\theta}_j - \theta_N)^2 = V_{\text{SAM}} + V_{\text{IMP}} + 2V_{\text{MIX}},
\]

(3)

where \( V_{\text{SAM}} = E_\xi E_p (\hat{\theta}_n - \theta_N)^2 \) is the sampling variance, \( V_{\text{IMP}} = E_\xi E_p E_R E_I (\hat{\theta}_j - \theta_N)^2 \) is the imputation variance, and \( V_{\text{MIX}} = E_\xi E_p E_R E_I ((\hat{\theta}_n - \theta_N) (\hat{\theta}_j - \theta_N)) \) is a mixed component. In this formulation, the total variance and its components are more aptly described as anticipated variances because they incorporate the added expectation with respect to the superpopulation model.

The model-assisted approach to variance estimation with imputed data used in this paper should be distinguished from model-assisted sampling (Särndal, Swensson and Wretman 1992). With model-assisted sampling, models are used to guide the choice of efficient sample designs and estimators, but the validity of statistical inferences is not dependent on the validity of the models. In contrast, when some data are missing, reliance on models for inferences is essential, both for point estimators and for variance estimators for them. In this paper, the general approach to inference employs the imputation model assumptions (i.e., superpopulation model and unconfoundedness assumptions) only to the extent necessary to account for imputed data. Both the point estimators and the variance estimators are the standard design-based estimators when no data are missing. Whether the variance estimators are approximately unbiased for \( V_{\text{SAM}} \) depends on the validity of the imputation model. Also, the estimators for \( V_{\text{IMP}} \) and \( V_{\text{MIX}} \) rely completely on the imputation model. Thus the validity of the model is much more critical with model-assisted variance estimation with imputed data than it is with model-assisted sampling. Särndal (1992) argues that if we are willing to accept the validity of the model in point estimation with imputed data, we should also be willing to accept its validity for variance estimation.

Variance estimators are obtained by conditioning on the realized set of sampled units, responding units, and imputations. We develop estimators of \( V'_{\text{SAM}} = E_\xi [\hat{\theta}_n - \theta_N]^2 \mid A, A_R, d] \), \( V'_{\text{IMP}} = E_\xi [\hat{\theta}_j - \theta_N]^2 \mid A, A_R, d] \), and \( V'_{\text{MIX}} = E_\xi [(\hat{\theta}_n - \theta_N) (\hat{\theta}_j - \theta_N)] \mid A, A_R, d] \), where \( A \) and \( A_R \) denote matrices of indices for the sampled and responding units, respectively, and \( d \) is the set of indices for the imputations. The conditioning is on the set of indices, not on the values of the units. The matrix \( d \) is an \( r \times (n-r) \) matrix in which the rows refer to respondents and the columns to nonrespondents. In this paper, we consider only single imputation methods, in which case all but one of the \( d_{ij} = 0 \) in every column. The exception occurs in the row of the donor respondent when \( d_{ij} = 1 \).

By considering the conditional expectations of \( V'_{\text{IMP}} \) and \( V'_{\text{MIX}} \), the estimators reflect the number of times responding units are used as donors in the given application rather than taking the expectation over all possible imputation outcomes. We argue below that these are the appropriate variances to estimate in a given application. If the variance estimators are conditionally unbiased, they are also, of course, unconditionally unbiased.

A conditional approach is useful for two reasons. First, when an estimator is conditionally unbiased and consistent (as \( \hat{\theta}_j \) is assumed to be for \( \hat{\theta}_n \) ), the conditional variance is generally a more appropriate estimator for making inferences from a realized sample than an unconditional variance (Holt and Smith 1979, Rao 1999, Kalton 2002). Thus, a variance estimator that conditions on the actual number of times each donor is used is to be preferred to a variance estimator that conditions on the actual number of times each donor is used as a donor in the given application rather than taking the expectation over all possible donor selections.

Second, the results apply to any unconfounded imputation scheme that substitutes observed values for missing ones and for which \( E_\xi (\hat{\theta}_j) = E_\xi (\hat{\theta}_n) \).

2. Hot Deck Imputation

We consider a simple model for which hot deck imputation is appropriate. Assume that the finite population \( (U) \) is composed of \( G \) classes or cells. Within cell \( g (g = 1, \ldots, G) \), the elements in \( U \) are realizations of independently and identically distributed random variables with mean \( \mu_g \) and variance \( \sigma_g^2 \). This cell mean model can be written as

\[
Y_{g,i}^{(i)} (\mu_g, \sigma_g^2), \quad i \in U_g.
\]

Statistics Canada, Catalogue No. 12-001
where \( \sim \) is an abbreviation for independently and identically distributed.

A linear estimator of \( \theta \) with complete item response from a complex sample survey can be written as

\[
\hat{\theta}_n = \sum_{i \in A} w_i y_i, \tag{5}
\]

where \( w_i \) is the weight that accounts for unequal selection probabilities and the estimation strategy. When the cell mean model holds, a more efficient estimator of \( \theta \) uses the unweighted group means, i.e., \( \hat{\theta}_u = \sum \sum w_{ig} y_{ig} \) where \( y_{ig} = \sum_j y_{igj} / n_g \). However, the model-assisted approach does not place complete reliance on the model; rather, it uses the standard design-based approach to the extent possible and the model is used only for the missing data.

The weights in (5) can be the inverse of the probability of selection weights or calibration adjusted weights, as described below.

The hot deck imputed value for \( y_j \) is \( y_j^* = \sum_{i \in A_g} d_{ij} y_i \) and the imputed estimator is

\[
\hat{\theta}_j = \sum_{i \in A} w_i \tilde{y}_i = \sum_{i \in A_g} w_{ig} y_{ig} + \sum_{j \in A_g} w_{ij} \sum_{i \in A_g} d_{ij} y_i, \tag{6}
\]

where \( \tilde{y}_i = y_i \) for \( i \in A_g \) and \( \tilde{y}_j = y_j^* \) for \( i \in A_g \). We assume throughout that imputed values are selected from respondents in the same imputation cells, and that each cell contains at least one respondent.

This imputation formulation does not specify the way in which donors are selected. It thus covers both unweighted hot deck imputation in which donors are selected with equal probabilities within each cell and weighted hot deck imputation. Weighted hot decks are typically used when assumptions are made only about the response distribution.

The form (6) also covers with and without replacement imputation methods. For example, it covers the common hot deck procedure in which a respondent is randomly selected to be a donor within a cell, and then that respondent is not used as a donor again until every other respondent in the cell has been used.

While not explicitly considered here, nearest neighbor imputation procedures that use continuous variables to identify a small set of the most similar respondents and then randomly select one as the donor, satisfy the above requirements. Furthermore, researchers often use hot deck methods even when continuous variables are available. Little (1986) discusses strategies for forming imputation cells using variables that are predictive of the \( y \) variable and notes that imputation within cells and regression imputation should produce similar results in many circumstances. Cochran (1968) and Aigner, Goldberger and Kalton (1975) show that a relatively small number of well-constructed cells formed from a continuous variable can capture a large proportion of the predictive power of the variable.

The conditional bias of the imputed estimator under the cell mean model is

\[
E_x (\hat{\theta}_j - \hat{\theta}_n | A, A_R, d) =
\]

\[
E_x \left[ \sum_{i \in A_g} w_{ij} (y_j^* - y_j) \right] = 0,
\]

since \( E_x (y_j^* ) = E_x (\sum_{i \in A_g} d_{ij} y_i ) = \sum_{i \in A_g} d_{ij} E_x (y_i) = \sum_{i \in A_g} d_{ij} \mu_i = \mu_j \) for \( j \) in cell \( g \). This expectation is conditioned on the indices of the sampled units, the responding units, and the donors. However, since the estimator is conditionally unbiased for any sample, it is also unconditionally unbiased. Kim and Fuller (1999) also use this conditioning argument. Estimators for each component of the variance of the hot deck imputed estimator are given in the next section.

3. Estimation of the Components of the Total Variance

This section contains the main results about estimators of the three components of the total variance of a linear hot deck imputed estimator. Throughout, we assume unconditional sampling, response, and imputation mechanisms and a linear complete sample estimator of the form (5). The results require that the cell mean model holds and that there is at least one respondent in each imputation cell. We begin with the variance due to sampling, \( V_{SAM} \).

We assume that there exists a complete sample variance estimator, \( \hat{V}_n \), that is design unbiased for the sampling variance of \( \hat{\theta}_n \), is a quadratic in the \( y \) variable, and is of the form

\[
\hat{V}_n = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} y_i y_j = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} y_i^2 + 2 \sum_{i \in A} \sum_{j \in A} \Omega_{ij} y_i y_j, \tag{7}
\]

for known coefficients \( \Omega_{ij} \). This formulation covers the Horvitz-Thompson estimator, where the \( \Omega_{ij} \) are determined by the single and joint probabilities of selection. It also covers the linearized variance estimator for the generalized regression (GREG) estimator. Rao, Yung and Hidiroglou (2002) show that the linearized variance estimator for the GREG estimator can be written by substituting \( g_{ix} e_i \) for \( y_i \) in the variance estimator for the Horvitz-Thompson estimator of a total. Here, \( g_{ix} \) is the sample-dependent \( g \) weight and \( e_i = y_i - x_i^T \hat{B} \), where \( x_i \) is the vector of auxiliary variables and \( \hat{B} \) is the vector of estimated regression coefficients. Since \( g_{ix} \) is not a function of \( y \) and \( \hat{B} \) linear in the \( y \) variable, \( g_{ix} e_i \) is linear in \( y \). Therefore, the linearized variance estimator for the GREG estimator is quadratic in \( y \) and can be expressed in the form given by
equation (7). Note that in this case the \( \Omega_{ij} \) may be dependent on the specific sample as well as on the selection probabilities. Deville and Särndal (1992) show that any calibration estimator has the same asymptotic variance as the GREG. Thus, asymptotic variance estimators for calibration estimators in general have the required quadratic form.

The naïve variance estimator treats imputed values as if they were observed values and can be written as

\[
\hat{V}_0 = \sum_{i \in A} \left( \sum_{j \in A} \Omega_{ij} \gamma_{ij} \frac{y_i - \bar{y}_j}{\sigma_J} \right).
\]

Lemma 1 gives the bias of the naïve variance estimator as an estimator for \( V_{SAM}' \). As noted earlier, the naïve variance estimator is proposed as the estimator of \( V_{SAM}' \) to be as consistent as possible with design-based inference. An additional practical reason for using the naïve variance estimator is to take advantage of existing software programs that estimate the sampling variance under complex samples.

**Lemma 1.** Under the cell mean model with unconfounded sampling, response, and imputation mechanisms and the assumptions that \( \hat{\Theta}_i \) is an unbiased hot deck imputed linear estimator given by (6) and \( \hat{V}_n \) is an unbiased complete sample variance estimator given by (7), then the bias of the naïve variance estimator, \( \hat{V}_0 \), as an estimator of \( V_n \) is

\[
2 \sum_{g=1}^{G} \sum_{j \in A_g} \sum_{i \in A_g} \Omega_{ij} d_{ij} \sigma_J^2 + 2 \sum_{g=1}^{G} \sum_{i \in A_g} \Omega_{ii} \gamma_{ii} \sigma_J^2,
\]

where \( A_R = A_R \cap U_g, A_M = A_M \cap U_g \) and

\[
\gamma_{ij} = \sum_{k \in A_g} d_{ik} d_{kj}.
\]

For any two nonrespondents, \( i \) and \( j \), that have the same donor, \( \gamma_{ij} = 1; \gamma_{ij} = 0 \) otherwise. By definition, \( \gamma_{ii} = 1 \).

**Proof.** We begin by noting that the difference between \( \hat{V}_0 \) and \( \hat{V}_n \) can be written as

\[
\hat{V}_0 - \hat{V}_n = \sum_{i \in A} \left( \sum_{j \in A} \Omega_{ij} (\bar{y}_i^2 - \bar{y}_j^2) \right)
+ 2 \sum_{i < j} \sum_{i \in A} \sum_{j \in A} \Omega_{ij} (\bar{y}_i - \bar{y}_j)(y_i - y_j)
= \sum_{i \in A} \Omega_{ii} (y_i^2 - \bar{y}_i^2)
+ 2 \sum_{i < j} \sum_{i \in A} \sum_{j \in A} \Omega_{ij} (y_i y_j - \bar{y}_j y_j)
+ 2 \sum_{i < j} \sum_{i \in A} \sum_{j \in A} \Omega_{ij} (y_i y_j - y_i y_j).
\]

Under the cell mean model, the conditional expectation of the first term of (11) is zero. The conditional expectation

\[
E_{\xi}(y_i y_j - y_i y_j) = E_{\xi}[y_i (y_j - y_j)] = 0
\]

unless respondent \( i \) is the donor for nonrespondent \( j \); it is thus zero when units \( i \) and \( j \) are in different cells and is only nonzero for one \( i \) and \( j \) in the same cell \( g \). It may be represented by \( E_{\xi}[y_i (y_j - y_j)] = d_{ij} \sigma_J^2 \). The conditional expectation

\[
E_{\xi}(y_i y_j - y_i y_j) = d_{ij} \sigma_J^2
\]

is zero unless nonresponding units \( i \) and \( j \) have the same donor, which can occur only if these units are in the same cell. It can be represented by \( E_{\xi}(y_i y_j - y_i y_j) = d_{ij} \sigma_J^2 \) for \( i \neq j \). Applying these results in equation (11) gives

\[
E_{\xi}(\hat{V}_0 - \hat{V}_n | A, A_R, d) = 2 \sum_{g=1}^{G} \sum_{i \in A_g} \sum_{j \in A_g} \Omega_{ij} d_{ij} \sigma_J^2
+ 2 \sum_{g=1}^{G} \sum_{i < j} \sum_{i \in A_g} \sum_{j \in A_g} \Omega_{ij} \gamma_{ij} \sigma_J^2.
\]

The proof is completed by noting that since \( \hat{V}_n \) is unbiased under the design, it is also unbiased for \( \hat{V}_{SAM}' \). Substituting a model unbiased estimator for \( \sigma_J^2 \), say \( \hat{\sigma}_J^2 \), gives an unbiased estimator of the bias of the naïve variance estimator. Note that whenever respondents donate their values to more than one nonrespondent, the last term in equation (12) is positive; otherwise, it is zero.

Two simple examples illustrate applications of these results. Consider first the estimation of a population mean from a simple random sample selected with replacement. In this case, \( \Omega_{ii} = n^{-2} \) and \( \Omega_{ij} = -n^{-2} (n-1)^{-1} \) for \( i \neq j \). Assume that the cell mean model holds with hot deck imputation and that no donor is used more than once. By Lemma 1, the bias of \( \hat{V}_0 \) is \( -2n^{-2} (n-1)^{-1} \sum m_g \sigma_J^2 \), where \( m_g = \sum_{i \in A_g} \sum_{j \in A_g} \sum_{i \in A_g} d_{ij} \) is the number of imputed values in cell \( g \). In this case, the bias of the naïve variance estimator is \( O_p(n^{-2}) \) and hence is negligible for large \( n \). Now suppose that every missing value in each cell is imputed from the same donor. In this case, with \( \sum_{i \in A_g} \sum_{j \in A_g} \gamma_{ij} = m_g (m_g - 1)/2 \), the bias of \( \hat{V}_0 \) is \( -n^{-2} (n-1)^{-1} \sum m_g (m_g + m_g) \sigma_J^2 \), which is \( O_p(n^{-3}) \) and is of the same order as the sampling variance.

As the second example, consider a simple two-stage sample of size \( n = ab \), in which all clusters are selected from a population of \( A \) equal-sized clusters by simple random sampling and \( b \) of \( B \) elements are selected by simple random sampling within each sampled cluster. Let \( y_{ai} \) be the value for \( y \) for sampled unit \( i \) in cluster \( \alpha \). Assume that the first stage sampling fraction is small enough to ignore. The estimate of the variance of the sample mean is of the form given by equation (7) where \( \Omega_{wA_{bij}} = a^{-2} b^{-2} = n^{-2} \) for \( \alpha = \beta \), and \( \Omega_{wA_{bij}} = -a^{-2} (a-1)^{-1} \) for \( \alpha \neq \beta \). These
values can now be inserted into equation (9) to compute an estimate of the bias. For example, suppose that all missing values are imputed using donors from the same cluster (the cells are the clusters) and no donor is used more than once. In this case, the bias of the naïve variance estimator is \(2n^{-2} \sum_m m_a \sigma_g^2\), where \(m_a\) is the number of nonrespondents in cluster \(a\). Now, suppose an overall mean model hot deck is used and no donor can donate more than once, but that donors are always chosen from different clusters than their missing values. In this case, the bias of the naïve variance estimator is \(-2n^{-2} (a-1)^{-1} \sigma^2 \sum_m m_a\). This two-stage example shows the naïve variance estimator can be biased in either direction. In both of the cases considered, the bias is of lower order than the variance, and if \(a\) is large the bias will be negligible.

The second component of the total variance is the variance due to imputation, \(V_{IMP}\). Lemma 2 gives an unbiased estimator for this component with hot deck imputation.

**Lemma 2.** Under the assumptions used in Lemma 1, an unbiased estimator of \(V_{IMP}\) is

\[
\hat{V}_{IMP} = 2 \sum_{g=1}^{G} \left\{ \sum_{i \in A_{U_{g}}} w_i^2 \hat{\sigma}_g^2 + \sum_{i,j \in A_{U_{g}}} w_i w_j \gamma_{ij} \hat{\sigma}_g^2 \right\},
\]

where \(\hat{\sigma}_g^2\) is an unbiased estimator for \(\sigma_g^2\).

**Proof.** Since the variance due to imputation involves the squared difference between the imputed and complete response estimates, we begin by writing

\[
(\hat{\theta}_j - \hat{\theta}_n)^2 = \left[ \sum_{i \in A} w_i (y_i^* - y_i) \right]^2
\]

\[
= \sum_{i \in A} w_i^2 (y_i^* - y_i)^2
\]

\[
+ 2 \sum_{i,j} w_i w_j (y_i^* - y_i) (y_j^* - y_j).
\]

Noting that \(E_{\xi}(y_i^* - y_i)^2 = 2\sigma_g^2\) for \(i\) in cell \(g\) and, from above, \(E_{\xi}(y_i^* - y_i) (y_j^* - y_j) = \gamma_{ij} \sigma_g^2\), it follows that

\[
\hat{V}_{IMP} = 2 \sum_{g=1}^{G} \left\{ \sum_{i \in A_{U_{g}}} w_i^2 \sigma_g^2 + \sum_{i,j \in A_{U_{g}}} w_i w_j \gamma_{ij} \sigma_g^2 \right\}.
\]

Substituting \(\hat{\sigma}_g^2\), a model unbiased estimator for \(\sigma_g^2\), establishes the lemma.

Equation (14) shows that the imputation variance has positive contributions from each imputed value and also from using donors more than once. For example, suppose that the weights for all sample cases are equal. The contribution to the imputation variance from cell \(g\) is then proportional to the sum of the number of missing cases in the cell and the number of pairs of nonrespondents that receive values from the same donors. Limiting the number of times donors are re-used can reduce the imputation variance.

The third term in the total variance is \(V_{MIX}\), which previous research often considered small or negligible (e.g., Särndal 1992; Deville and Särndal 1994). Lemma 3 gives an unbiased estimator for \(V_{MIX}\).

**Lemma 3.** Under the assumptions used in Lemma 1, an unbiased estimator for \(V_{MIX}\) is

\[
\sum_{g=1}^{G} \left[ \sum_{i \in A_{U_{g}}} \sum_{j \in A_{U_{g}}} w_i w_j d_{ij} - \sum_{j \in A_{U_{g}}} w_j^2 \right] \sigma_g^2.
\]

**Proof.** Begin by writing \((\hat{\theta}_j - \hat{\theta}_n) (\hat{\theta}_n - \hat{\theta}_a) = \hat{\theta}_a (\hat{\theta}_j - \hat{\theta}_a) - \theta_N (\hat{\theta}_j - \hat{\theta}_a).\) Let \(\theta_N\) be the finite population total, which can be written as \(\sum_{i \in U - A} y_i + \sum_{i \in A_{g}} y_i + \sum_{i \in A_{R}} y_i\). Using this expression, the second component can be expanded as

\[
\theta_N (\hat{\theta}_j - \hat{\theta}_a) = \left[ \sum_{i \in U - A} y_i + \sum_{i \in A_{g}} y_i + \sum_{i \in A_{R}} y_i \right] \left[ \sum_{j \in A_{U_{g}}} w_j (y_j^* - y_j) \right].
\]

In taking the conditional expectation of this product, the only nonzero contributions occur either when unit \(i\) in \(A_{R}\) is the donor for \(y_j^*\), or when unit \(i\) in \(A_M\) in the first set of parentheses is unit \(j\) in the second set. In the first case, \(E_{\xi}(y_i^* - y_j) = d_{ij} \sigma_g^2\) for \(i \in A_{g}, j \in A_{g}\). In the second case, if nonrespondent unit \(i\) in \(A_{Mg}\) is the same as unit \(j\) in the second term, \(i = j\), \(E_{\xi}(y_i^* - y_j) = -\sigma_g^2\), and this expectation is 0 otherwise. Thus,

\[
E_{\xi}(\theta_N (\hat{\theta}_j - \hat{\theta}_a) | A, A_R, d) = \sum_{g} \sum_{j \in A_{U_{g}}} w_j \sigma_g^2
\]

\[
- \sum_{g} \sum_{j \in A_{U_{g}}} w_j \sigma_g^2 = 0.
\]

The first term can be expressed as

\[
\hat{\theta}_a (\hat{\theta}_j - \hat{\theta}_a) = \left[ \sum_{i \in A_{g}} w_i y_i + \sum_{i \in A_{R}} w_i y_i \right] \left[ \sum_{j \in A_{U_{g}}} w_j (y_j^* - y_j) \right].
\]

Using the results for \(E_{\xi}(y_i^* - y_j)\) given above,

\[
V_{MIX} = \sum_{g=1}^{G} \left[ \sum_{i \in A_{U_{g}}} \sum_{j \in A_{U_{g}}} w_i w_j d_{ij} - \sum_{g=1}^{G} \sum_{j \in A_{U_{g}}} w_j^2 \right] \sigma_g^2.
\]
The estimator of $V'_{\text{MIX}}$ is zero when the weights are constant, or more generally when the weights of the donors are equal to the weights of the missing cases to which they are assigned. Most of the simulations in the literature (e.g., Särndal 1992; Lee, Rancourt and Särndal 1995) have used simple random samples so that the estimates of the mixed term from the simulations are approximately equal to zero.

To illustrate the effect of unequal weights, consider a stratified simple random sample selected from two equal size strata with replacement, and suppose that the sampling rate in stratum 2 is $k$ times the rate in stratum 1. Let the imputation model be the overall cell mean model and let the hot deck procedure select donors with simple random sampling without replacement. For this simple situation, $V'_{\text{MIX}}$ can be derived algebraically. Table 1 shows the percentage contribution of the mixed term to the total variance $(100 \times 2V'_{\text{MIX}} / V'_{\text{TOT}})$ for various combinations of strata response rates. The table illustrates the fact that when the sampling weights are unequal, the contribution of the mixed term may be important and can be either positive or negative. The mixed term may also be important in domain estimation, as discussed in the next section.

### Table 1

<table>
<thead>
<tr>
<th>Response rate</th>
<th>Oversampling rate in stratum 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum 1</td>
<td>Stratum 2</td>
</tr>
<tr>
<td>----------------</td>
<td>------------</td>
</tr>
<tr>
<td>100%</td>
<td>80%</td>
</tr>
<tr>
<td>100</td>
<td>60</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>60</td>
<td>80</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>60</td>
<td>20</td>
</tr>
</tbody>
</table>

Now consider estimating the total variance using the three lemmas for the hot deck estimator under the cell mean model. To estimate $V'_{\text{SAM}}$, we can either use the naïve variance estimator, with its bias as given in Lemma 1, or correct for the bias with a procedure similar to that recommended by Särndal (1992). For a single stage sample, the bias correction given by Lemma 1 is easy to apply. However, with multi-stage sampling the correction involving $\Omega$ may be complicated and difficult to implement in practice. In this case, the naïve variance estimator should produce an adequate approximation provided that the number of sampled clusters is large, that no donor is used too often, and that the percentage of missing data in each cell is not extremely large.

For the other two components, the only unknown quantities that must be estimated from the sample are the cell variances, $\sigma^2_g$. These parameters could be estimated using either unweighted observations or weighted observations, where the weights are the selection weights. Fuller (2002) recommends the use of weighted observations to provide more robust estimates. Unbiased estimators of the conditional variance due to imputation and the mixed component are computed by substituting unbiased estimates of the cell variances, $\hat{\sigma}^2_g$. Then, adding $\hat{V}_0$, $\hat{V}'_{\text{IMP}}$, and $2\hat{V}'_{\text{MIX}}$ gives an estimator of the total variance

$$\hat{V}'_{\text{TOT}} = \hat{V}_0 + 2\sum_{g=1}^G \sum_{i,j} w_{ij} \gamma_y \hat{\sigma}^2_g + 2\sum_{g=1}^G \sum_{i,j} w_{ij} d_{ij} \hat{\sigma}^2_g. \quad (17)$$

To examine this estimator, we give a few simple examples with known solutions. All of these examples involve samples with equal weights so the mixed component is zero. First, assume simple random sampling with replacement, hot deck imputation under the overall cell mean model, and no donor used more than once. Using the naïve variance estimator for $V'_{\text{SAM}}$, the estimated total variance is $n^{-1} s_y^2 + 2n^{-1} \hat{\sigma}^2 (1 - rm^{-1})$, where $s_y^2 = (n-1)^{-1} \sum_{i,t} (\tilde{y}_i - \bar{y}_j)^2$, $r$ is the number of respondents, and $m$ is the number of missing cases. If we use $\hat{\sigma}^2$ instead of $s_y^2$ (where $\hat{\sigma}^2$ is model unbiased while $s_y^2$ has a small sample bias), then this simplifies to $r^{-1} \hat{\sigma}^2 [1 + m(m-r)m^{-2}]$. Taking the expectation of this estimator gives the unconditional variance of the without-replacement hot deck estimator given by Kalton (1983, page 25, 2.3.1.7).

If a multiple cell mean model rather than an overall cell mean model is used, then the estimated total variance is $n^{-1} s_y^2 + 2n^{-2} \sum_{g=1}^G \hat{\sigma}^2_g (n_g - r_g)$, which is similar to the result given by Tollefson and Fuller (1992).

Continuing with the simple random sampling example, now allow donors to be used more than once with the overall cell mean model. Again using $\hat{\sigma}^2$ instead of $s_y^2$, the estimated total variance is approximately

$$n^{-2} \hat{\sigma}^2 \left[ n + m + \sum_{i,j} \gamma_y \right]. \quad (18)$$

For fixed $m$, the variance in equation (18) is minimized when no donor is used more often than any other donor, to the extent possible (thereby minimizing $\sum_{i,j} \gamma_y$).

Therefore, an imputation scheme that uses any donor at most once more than any other donor minimizes the total variance.
If donors are selected by simple random sampling with replacement, then \( E_i [\gamma_{ij}] = r^{-1} \) and the expected value of (18) is \( r^{-1} \sigma^2_w [1 + n^{-2} m (r-1)] \). This is the expected variance of the with-replacement hot deck estimator given by Kalton (1983, page 26, 2.3.1.9).

These examples show that the approach produces reasonable estimates for the total variance in simple cases and highlights the conditional nature of the variance estimates. For example, (18) is conditional on the actual number of times donors are used rather than on the expected number of times they are used (the unconditional result). The approach is flexible enough to allow a variety of imputation methods, including with- and without-replacement and weighted and unweighted versions of the hot deck.

4. Domain Estimation

This section considers the important problem of domain estimation under the cell mean model with hot deck imputed data. Previous research on this topic is limited (Lee et al. 1995). The standard estimator with complete response for a population total for domain \( v \) is \( \hat{\theta}_v = \sum_{i \in A_v} w_i y_i \), which may be alternatively expressed as \( \hat{\theta}_v = \sum_{i \in A} w_i \delta_v y_i \) where \( w_i = \delta_v w_i \) with \( \delta_v = 1 \) if \( i \in A_v \) and \( \delta_v = 0 \) otherwise. The hot deck imputed estimator is \( \hat{\delta}_v = \sum_{i \in A} w_i \delta_w \tilde{y}_i = \sum_{i \in A} w_i \tilde{y}_i \). Throughout we assume that \( \delta_w \) is known for all \( i \in A_v \).

The cell mean model assumes that all the elements in a cell have the same distribution. In general, some elements in a cell may be in the domain and others not. One version of the model assumes a separate cell mean model for the domain alone and then applies an appropriate imputation scheme. The theory given in the previous section covers this case, and it will, therefore, not be discussed further here. While it is feasible to account for key domains in the imputation stage, it is impossible to consider all possible domains analysts may wish to study. Thus, the focus in this section on domains that cut across imputation cells has important practical implications, especially for analysis of public use data files.

We now discuss the estimation of the three components of \( V'_{\text{TOT}} \), the variance of an imputed domain total. Consider first the estimation of \( V'_{\text{SAM}} \). In the case of complete response, by setting \( y_i = 0 \) for elements outside the domain, the estimated sampling variance can be expressed in the form of equation (7) as \( \hat{V}_{ai} = \sum_{i \in A_v} \Omega_{ii} y_i^2 + 2 \sum_{i < j \in A_v} \Omega_{ij} y_i y_j \). With domain membership known for all sample elements, the conditional bias of the imputed variance estimator \( \hat{V}_{0i} \), following the developments in section 3 is:

\[
E_i (\hat{V}_{0i} - \hat{V}_{ai}) = \frac{1}{2} \sum_{g=1}^G \sum_{i \in A_g} \sum_{j < i \in A_{g'}} \Omega_{ij} d_{ij} \sigma^2_i + \frac{1}{2} \sum_{g=1}^G \sum_{i \in A_g} \sum_{j < i \in A_{g'}} \Omega_{ij} y_i \sigma^2_i.
\]

(19)

As discussed in section 3, with large samples \( \hat{V}_{0i} \) may be conveniently employed to estimate \( V'_{\text{SAM}} \), using standard survey sampling variance estimation software. It is interesting to note that the naïve variance estimator would be unbiased if all the donors were from outside the domain (thus, \( d_{ij} = 0 \)) and no donor was used more than once (\( y_i = 0 \)).

The derivation of \( V'_{\text{IMP}} \) follows directly from Lemma 2, where the weights are treated as constants in the conditional expectation. Replacing \( w_i' \) for \( w_i \) in equation (13) gives

\[
\hat{V}'_{\text{IMP}} = 2 \sum_{g=1}^G \left\{ \sum_{i < j \in A_{g'}} \Omega_{ij} w_i' y_i' \sigma^2_i + \sum_{i \in A_{g'}} w_i' \sigma^2_i \right\}.
\]

(20)

Note that the mixed component is not zero for a domain total, even if all the original weights are equal. With equal weights \( w \) (but not equal \( w' \)), the contribution to \( \hat{V}'_{\text{IMP}} \) is negative when the donor is from inside the domain whereas it is negative when the donor is from outside the domain. As a result \( \hat{V}'_{\text{IMP}} = -w^2 \sum_{g} l_{gv} \sigma^2_i \), where \( l_{gv} \) is the number of donors from outside the domain in cell \( g \). In this case, ignoring the mixed component with domain estimation results in an overestimate of the total variance. With unequal weights, the bias due to ignoring the mixed component can be either positive or negative.

The total variance of a (linear) imputed domain estimator under the cell mean model is then estimated by
\[
\hat{V}_{TOT}' = \hat{V}_0 + 2\sum_{g=1}^N \sum_{i,j \in \mathcal{A}_{ij}} w_i w_j y_{ij} \hat{\sigma}_g^2 + 2\sum_{g=1}^N \sum_{i \in \mathcal{A}_i} \sum_{j \in \mathcal{A}_j} w_i w_j \hat{\sigma}_g^2.
\]

As an illustration, consider the case of equal weights within the domain \((w_{hi} = w_{ai})\) and no donor used more than once. In this case, the second term on the right in (21) is zero and the third term reflects the variance increase from imputation. If all the missing values are imputed using donors from the domain, then the third term is \(2 w_i^2 \sum m_{gi} \hat{\sigma}_g^2\) where \(m_{gi}\) is the number of missing items in cell \(g\) and domain \(i\). On the other hand, if no units are imputed from within the domain, then this term is zero. Thus, the total variance is minimized when the donors are selected from outside the domain rather than from within the domain. This result occurs because imputing from outside the domain in effect substitutes a new value for a missing value for domain estimation, thus maintaining the original domain sample size. On the other hand, imputing from within the domain does not increase domain sample size and there is also a penalty to the variance from reusing a domain respondent’s value for the nonrespondent.

If the distribution of \(y\) varies by domain (i.e., the imputation model is misspecified), then choosing donors from outside the domain results in biased estimates. Since all models are misspecified to some degree, it is therefore generally unwise to intentionally select donors from outside the domain in order to minimize the variance.

## 5. Simulation Study

A small simulation study was performed to examine the model-assisted variance estimates for estimating an overall total and a domain total. A sample of 40 clusters with exactly 5 units in each cluster was selected from an infinite superpopulation, where \(y_{ai}\) is the study variable for unit \(i\) in cluster \(a\). The \(y\) values were generated from \(y_{ai} = \tau a_i + e_{ai}\), where \(a_i\) and \(e_{ai}\) are independent random draws from the standard normal distribution. Thus, the \(y\) values have mean zero, variance \((\tau^2 + 1)\), and correlation \(\rho = \tau^2 / (1 + \tau^2)\) if the units are from the same cluster and \(\rho = 0\) otherwise. Values of \(\tau = 0\) and \(\tau = 0.5\) were chosen, giving correlations of 0 and 0.2, respectively. The value, \(\rho = 0.2\), was chosen to illustrate the effect of a high intraclass correlation. In addition to the \(y\) variable, an indicator variable for domain \(v\) was generated by independent sampling with the probability of being in the domain of 0.25. Respondents were selected from the full sample using a uniform response probability of 0.6 and missing values were imputed using a single-cell-with-replacement hot deck. A total of 5,000 Monte Carlo samples was selected.

The simulated point estimators for the overall total and the domain total are unbiased. The means and biases of the model-assisted variance estimators \((\hat{V}_{TOT}', \hat{V}_0, \hat{V}_{IMP}', \hat{V}_{MIX}')\) are given in Table 2 (the tabulated values are divided by \(N^2 10^{-4}\)). When \(\rho = 0\), the relative biases of the variance estimators for the overall and domain totals are very small. On the other hand, when \(\rho = 0.2\), the variance estimators have negative relative biases that are not negligible (a relative bias of \(-13\%\) for the overall total and \(-5\%\) for the domain total). To identify the source of the bias, Table 2 also gives the means and biases of the three variance components. The tabled values show that \(\hat{V}_{IMP}'\) and \(\hat{V}_{MIX}'\) are approximately unbiased, and it is only \(\hat{V}_0\) that has a non-negligible bias.

When \(\rho = 0\) the cell mean model holds and \(\hat{V}_0\) is unbiased as expected under the theory. When \(\rho = 0.2\), the correlation of the \(y\) values within clusters implies that the cell mean model assumption does not hold. The imputation procedure replaces some missing values using donors from outside the cluster, causing \(\hat{V}_0\) to underestimate the sampling variance due to the underestimation of the intraclass correlation. In this particular situation, the model failures do not result in biased estimates for the other two components. However, these components could be biased under other types of model failure. The simulation illustrates the dependence of the model-assisted estimators on the model assumptions and this is discussed further in the next section.

### Table 2

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Mean</th>
<th>Bias</th>
<th>Mean</th>
<th>Bias</th>
<th>Mean</th>
<th>Bias</th>
<th>Mean</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{y})</td>
<td>0.0</td>
<td>104</td>
<td>-0.5</td>
<td>50</td>
<td>-1.9</td>
<td>54</td>
<td>-0.9</td>
<td>0</td>
</tr>
<tr>
<td>(\hat{y}_v)</td>
<td>0.2</td>
<td>126</td>
<td>-19.6</td>
<td>86</td>
<td>-20.7</td>
<td>61</td>
<td>-0.1</td>
<td>0</td>
</tr>
</tbody>
</table>

\(\hat{V}_{TOT}', \hat{V}_0, \hat{V}_{IMP}', \hat{V}_{MIX}'\) are given in Table 2 (the tabulated values are divided by \(N^2 10^{-4}\)). The values in the table are actual values divided by \(N^2 10^{-4}\).

## 6. Discussion

This paper describes a method for estimating the variance of a survey estimate when some of the values are imputed using hot deck imputation. The method uses a model-assisted approach and conditions on indices for sample members, respondents, and hot deck donors. The approach extends the work of Deville and Särndal (1994) to variance estimation for hot deck imputation, probably the most widely used method of imputation in household surveys. The proposed variance estimator is valid for a general

Statistics Canada, Catalogue No. 12-001
sample design and for a variety of estimation procedures under the superpopulation model and unconfounded assumptions. The paper also extends the previous work by handling stochastic rather than deterministic imputation and giving conditions for the bias of the naive variance estimator as an estimator of $V_{SAM}$ to be small.

The results focus attention on the need to take the mixed component into account when the sample elements have unequal weights. In particular, since domain estimates can be treated by assigning adjusted weights of zero for sample elements not in the domain, the mixed term needs to be taken into account in estimating the variance of domain imputed estimates even if the original weights were equal. Other statistics can also be covered by the approach used for domain estimates. For example, for the simple regression of $y$ on $x$, with $y$ including hot deck imputed values and $x$ complete, the regression coefficient can be expressed as a weighted linear combination of the $y$'s: $b = \sum w_i (x_i - \bar{x})/\sum w_i (x_i - \bar{x})^2 = \sum w_i' y_i$, where $w_i' = w_i (x_i - \bar{x})/\sum w_i (x_i - \bar{x})$. The difference between two domain estimates, $\theta_{12}$ and $\theta_{22}$, can be expressed as $\theta_{1i} - \theta_{2i} = \sum_{i \in v1} w_i y_i - \sum_{i \in v2} w_i y_i = \sum w_i y_i$, where $w_i' = w_i$ for $i \in v1$, $w_i' = -w_i$ for $i \in v2$, and $w_i = 0$ for $i \notin v1 \cup v2$.

The last example, involving the difference between domain estimates where imputation cells cut across domains, highlights the importance of the model in the imputation process. In this example, the analytic interest in the difference between the domain statistics is incompatible with an imputation model that assumes no difference in the $y$ distribution across domains within imputation cells. By imputing across domains with a hot deck cell imputation scheme, the sample domain means for $y$ will be brought closer together, thus decreasing the estimate of the difference. Thus, a good imputation model is crucial for producing valid point estimates.

The model-assisted approach to variance estimation with imputed data described here assumes a linear estimator, but smooth nonlinear functions can also be included using a Taylor series approximation. Like the Rao and Shao (1992) adjusted jackknife method, the model-assisted method is applicable with general sample designs and estimation schemes. However, the adjusted jackknife method is applicable only with a weighted hot-deck whereas, as a result of its model assumptions, the model-assisted method can be employed with a variety of hot deck methods, including choosing donors with equal probability and with probabilities proportional to their weights. The model-assisted method of variance estimation could also be extended to other imputation schemes such as nearest neighbor imputation and fractional hot deck imputation (Kalton and Kish 1984; Fay 1996; Kim 2000), a technique which reduces the variance due to imputation.

Implementation of the model-assisted method with hot deck imputation requires the availability of the information needed to compute the three components of the total variance. Standard survey sampling variance estimation software can be used to compute an estimate of $V_0$ that is approximately unbiased with large samples, but as the simulation study illustrates the estimate may be biased if the cell mean model does not hold. The computations of the other components require information on the identity of the donor for each imputed value and of the imputation cell membership of all sample members. From this information, $d_{ij}$ and $\gamma_{ij}$ can be determined. In addition, an estimate of $\sigma^2$ is required.

While the theory given above applies to variance estimation with many sample designs, including multi-stage samples, there are serious concerns about the validity of the imputation model in many cases. In the case of multi-stage sampling, the means of many survey variables differ across PSUs, yet hot deck cells are seldom formed within PSUs. Rather they are constructed in terms of other variables that cut across PSUs. Even within these cells there may be differences in means between PSUs. These differences may be offsetting to some extent and not introduce substantial biases for point estimation. However, their effect on variance estimation may be more significant. As indicated in the simulation, failure of the assumptions may have a greater impact on second order statistics than first order statistics. This issue merits more detailed investigation.

Imputation is more difficult when the goal is estimating a function of more than one variable with missing values. To produce an unbiased estimate of a parameter that involves several variables subject to imputation requires the development of an appropriate multivariate model and an imputation procedure consistent with that model. Given an appropriate model and hot deck imputation that is consistent with it, the model-assisted approach to variance estimation can then be implemented. However, estimating the variance becomes considerably more complex with multivariate estimates. The development of practical methods of imputation and variance estimation for this situation is much needed.

Acknowledgements

The authors would like to acknowledge the National Center for Education Statistics, Institute for Education Sciences for funding this research, and in particular the support of the Project Officer, Marilyn Seastrom. We also would like to thank the Associate Editor and referees for their constructive comments.
References


