VARIANCE ESTIMATION IN THE PRESENCE OF IMPUTED DATA FOR THE GENERALIZED ESTIMATION SYSTEM

Hyunshik Lee and Eric Rancourt, Statistics Canada, Carl-Erik Särndal, Université de Montréal
Eric Rancourt, Statistics Canada, 11-I, R.H. Coats Bldg, Ottawa, ON, Canada, K1A 0T6

Key Words: Domain, Jackknife technique, Model-assisted approach, Multiple Imputation, Single Imputation.

1. Introduction

The Generalized Estimation System (GES) is a micro-computer estimation package for survey data that has been developed at Statistics Canada. It is designed to produce point estimates and the associated variance estimates for domain totals, sizes, means and ratios. Two options are available for estimation of the variance: the model-assisted theory with the Taylor linearization technique and the jackknife technique. The system can handle a variety of sample designs and types of auxiliary information. It has become an important tool in many surveys conducted at Statistics Canada.

Missing data is a common problem in sample surveys and imputation is frequently used to fill in substitute values for the missing data. At present however, GES does not provide a proper variance estimation when the data contain imputed values. It treats imputed values as if they were real observations. It is well known that this can cause severe under-estimation of the true variance. To address this problem, Rubin (1987) proposed the multiple imputation method. This procedure requires multiply-imputed values for each missing value, resulting in multiple completed data sets. The advantage of this method is that the standard variance formula for 100% response can be used. So a variance estimation package such as GES could be used to compute estimates for each of the completed data sets. The results would then be combined outside of GES as proposed by Rubin (1978, 1987). Note that some simple extra programming would have to be carried out by the user. However, multiple imputation has not been used at Statistics Canada so far, in contrast to single imputation which is extensively used. Therefore, provision of proper variance estimates for singly-imputed data sets is necessary. In response to this need, there has been a considerable amount of research done recently. It includes Särndal (1990), Rao (1990), Lee, Rancourt and Särndal (1991, 1994), Rao and Shao (1992), Deville and Särndal (1991, 1994), Rancourt, Lee and Särndal (1993), Rancourt, Särndal and Lee (1994), Kovar and Chen (1994), Rao and Sitter (1995). These papers focus on estimation at the population level while GES is designed to produce estimates for arbitrary domains. Therefore, there is a need to adapt the techniques for variance estimation with imputed data to the problem of domain estimation and this is the new feature of the paper.

The purpose of this paper is to summarize the requirements for GES for proper variance estimation in the case of single value imputation using the mean, hot-deck, ratio or nearest neighbour method. These imputation methods are the most frequently used at Statistics Canada. In Sections 3 and 4, we consider two techniques for variance estimation of singly-imputed data: the model-assisted approach and the jackknife technique. Both of these options are currently available in GES. In section 5, the implementation aspects are presented and in section 6 we give a summary of the areas where work remains to be done.

2. Notation and Assumptions

Let $U=\{1,\ldots,k,\ldots,N\}$ be the index set of the population. Simple random sampling without replacement is used to draw a sample s of size n from U. The set of respondents, denoted by r, is of size m, and the set of nonrespondents, denoted by o, is of size l. Then we have $s=r\cup o$. The variable of interest is denoted by y. We are interested in the estimation of the population total $Y_U=\sum_U y_k$ (or more generally the total for a specified domain). The nonresponse mechanism that generates missing data is assumed to be uniform.

If unit $k\in o$, then the missing value y_k is imputed and its imputed value is denoted by $\hat{y}_k$. The data set after imputation is given by $\{y_{\bullet k}: k\in s\}$ where

$$y_{\bullet k}=\begin{cases} y_k & \text{if } k\in r \\ \hat{y}_k & \text{if } k\in o. \end{cases}$$

The usual simple expansion estimator for the population total, Y_U is:

$$\hat{Y}_{\bullet s}=\frac{N}{n}\sum_s y_{\bullet k}=\frac{N}{n}\left(\sum_r y_k+\sum_o \hat{y}_k\right).$$

Now consider a domain U_d. Its intersections with s, r and o are denoted by $s_d=U_d\cap s$, $r_d=U_d\cap r$ and $o_d=U_d\cap o$, and their respective sizes are denoted by n_d, m_d and l_d. Let us define

$$y_k(d)=\begin{cases} y_k & \text{if } k \in U_d \\ 0 & \text{otherwise.}\end{cases}$$

Then the domain total

$$Y_U(d)=\sum_U y_k(d)$$

is estimated using the completed data set by

$$\hat{Y}_{\bullet s}(d)=\frac{N}{n}\sum_s y_{\bullet k}(d)=\frac{N}{n}\{\sum_r y_k(d)+\sum_o \hat{y}_k(d)\}$$

where

$$y_{\bullet k}(d)=\begin{cases} y_{\bullet k} & \text{if } k \in s_d \\ 0 & \text{if } k \in s-s_d.\end{cases}$$

The total mean square error of $\hat{Y}_{\bullet s}(d)$ is defined as

$$V_{\text{TOT}}(d)=E_pE_q(\hat{Y}_{\bullet s}(d)-Y_U(d))^2,$$

where E_p and E_q are the expectation operators with respect to the sampling design p and the response mechanism q, respectively.

The estimator for the case of 100% response is $\hat{Y}_s(d)=(N/n)\sum_s y_k(d)$. Using it, the total error of $\hat{Y}_{\bullet s}(d)$ can be decomposed as

$$\hat{Y}_{\bullet s}(d)-Y_U(d)=\hat{Y}_s(d)-Y_U(d)+\hat{Y}_{\bullet s}(d)-\hat{Y}_s(d)$$

Note that $\hat{Y}_s(d)-Y_U(d)$ is the sampling error and $\hat{Y}_{\bullet s}(d)-\hat{Y}_s(d)$ is the imputation error.

We then obtain

$$V_{\text{TOT}}(d)=V_{\text{SAM}}(d)+V_{\text{IMP}}(d)+2V_{\text{MIX}}(d)$$

where

$$V_{\text{SAM}}(d)=E_p\{\hat{Y}_s(d)-Y_U(d)\}^2$$

$$V_{\text{IMP}}(d)=E_pE_q\{\hat{Y}_{\bullet s}(d)-\hat{Y}_s(d)\}^2$$

$$V_{\text{MIX}}(d)=E_pE_q\{(\hat{Y}_{\bullet s}(d)-\hat{Y}_s(d))(\hat{Y}_s(d)-Y_U(d))\}.$$

Subsets of the population that are used for specific imputation procedures will be termed "imputation groups". In this paper, we assume that imputation groups coincide with strata. Point and variance estimations are discussed in this context. This means that we focus on the estimation problem for a single stratum which plays the role of the population defined at the beginning of this section.

We consider four imputation methods: respondent mean, hot-deck, ratio and nearest neighbour imputation. For ratio and nearest neighbour imputation, we assume that an auxiliary variable, x, such that $x_k>0,\ \forall k\in s$ is available for the whole sample but not necessarily for all units in U. The imputed values $\hat{y}_k$ for the four methods are as follows:

Respondent mean (RM) imputation

$$\hat{y}_k=\bar{y}_r=\frac{1}{m}\sum_r y_k.$$

Hot-Deck (HD) imputation

$$\hat{y}_k=y_{g(k)}$$

where $y_{g(k)}$ is the y-value given by the donor unit, $g(k)\in r$, drawn at random (with replacement) from the m respondent units.

Ratio (RA) imputation

$$\hat{y}_k=\hat{B}_r x_k$$

where $\hat{B}_r=\sum_r y_k/\sum_r x_k$.

Nearest neighbour (NN) imputation

$$\hat{y}_k=y_{g(k)}$$

where $y_{g(k)}$ is the y-value given by the donor unit $g(k)$ such that $\min_{g\in r}|x_g-x_k|$ occurs for $g=g(k)$. If this yields more than one unit, a donor is randomly selected among them.

Both NN and HD require finding a donor among the respondents $g\in r$. However, a fundamental difference exists between the two methods. While the HD method randomly draws donors, the NN method uses an auxiliary variable to select donors by the minimum distance criterion.

3. Methods Based on the Model-Assisted Approach

The survey statistician who uses imputation for missing data appeals (unconsciously, perhaps) to one or more model assumptions. More specifically, the imputed values are assumed, on the average, to be good substitutes for the missing values. That is, in expectation the difference between the true, unobserved value y_k and the imputed value $\hat{y}_k$ is assumed to be zero. The statistician uses these assumptions implicitly or explicitly in the model-assisted approach to obtain variance estimators. These assumptions deal with the response mechanism and/or the relationship between the variable of interest y and the auxiliary variable x used to impute missing y values. We can distinguish three model components:

1. The response mechanism;
2. The regression relationship;
3. The variance structure for the regression.

In the following, the response mechanism is assumed to be uniform, and the model (denoted by ξ) is stated as ξ: $y_k = \beta x_k + \varepsilon_k$; $E_\xi(\varepsilon_k) = 0$; $E_\xi(\varepsilon_k^2) = \sigma^2 x_k$; and $E_\xi(\varepsilon_k \varepsilon_{k'}) = 0$ for $k \neq k'$.

The objective is to obtain an estimator of $V_{\text{TOT}}(d)$ for $\hat{Y}_{\bullet s}(d)$ by constructing estimators of the components $V_{\text{SAM}}(d)$, $V_{\text{IMP}}(d)$, and $V_{\text{MIX}}(d)$. In the model-assisted approach, the variance components $\hat{V}_{\text{SAM}}(d)$, $\hat{V}_{\text{IMP}}(d)$, and $\hat{V}_{\text{MIX}}(d)$ satisfy

$$E_\xi(\hat{V}_{\text{SAM}}(d) - V_{\text{SAM}}(d)) = E_\xi(\hat{V}_{\text{IMP}}(d) - V_{\text{IMP}}(d)) = E_\xi(\hat{V}_{\text{MIX}}(d) - V_{\text{MIX}}(d)) = 0.$$

In particular, $\hat{V}_{\text{SAM}}(d)$ is constructed as the sum of two terms, the first of which is the "ordinary formula", $\hat{V}_{\text{ORD}}(d)$, computed on the completed data set,

$$\hat{V}_{\text{ORD}}(d) = N^2 \frac{1-f}{n} S^2_{y\bullet s}(d)$$

with $S^2_{y\bullet s}(d) = \frac{1}{n-1} \sum_s \{y_{\bullet k}(d) - (\sum_s y_{\bullet k}(d)/n)\}^2$.

The ordinary formula is already available in GES and gives correct results for 100% response. However, when imputed values are present in the data set, $\hat{V}_{\text{ORD}}(d)$ underestimates $V_{\text{SAM}}(d)$ for some of the imputation methods. The second term, a correction term denoted by $\hat{V}_{\text{DIF}}(d)$ becomes necessary. The estimated sampling variance is then

$$\hat{V}_{\text{SAM}}(d) = \hat{V}_{\text{ORD}}(d) + \hat{V}_{\text{DIF}}(d).$$

It follows that the correction term $\hat{V}_{\text{DIF}}$ should be constructed to satisfy

$$E_\xi\{\hat{V}_{\text{DIF}}(d)\} = \frac{N^2(1-f)}{n} E_\xi\{S^2_{ys}(d) - S^2_{y\bullet s}(d)\}.$$

We now present the $\hat{V}_{\text{DIF}}(d)$, $\hat{V}_{\text{IMP}}(d)$, and $\hat{V}_{\text{MIX}}(d)$ components for each of the four types of imputation methods considered.

<u>RM imputation</u>

For RM imputation, the model assumed is model ξ with $x_k = 1$ for all units. The components of the variance estimator are:

$$\hat{V}_{\text{DIF}}(d) = \frac{N^2(1-f)}{n^2} l_d \hat{\sigma}^2$$

$$\hat{V}_{\text{IMP}}(d) = \frac{N^2}{n^2} l_d \left(\frac{l_d}{m} + 1\right) \hat{\sigma}^2$$

$$\hat{V}_{\text{MIX}}(d) = \frac{N^2(1-f)}{n^2} l_d \left(\frac{m_d}{m} - 1\right) \hat{\sigma}^2$$

with $\hat{\sigma}^2 = \frac{m-1}{m} S^2_{yr}$ and $S^2_{yr} = \frac{1}{m-1} \sum_{k \in r} (y_k - \bar{y}_r)^2$.

<u>HD imputation</u>

For HD imputation, the model is the same as for RM imputation, but the components of the variance estimator are different:

$$\hat{V}_{\text{DIF}}(d) = 0$$

$$\hat{V}_{\text{IMP}}(d) = \frac{N^2}{n^2} l_d \left(\frac{l_d}{m} + 2\right) \hat{\sigma}^2$$

$$\hat{V}_{\text{MIX}}(d) = \frac{N^2(1-f)}{n^2} l_d \left(\frac{m_d}{m} - 1\right) \hat{\sigma}^2$$

with $\hat{\sigma}^2 = \frac{m-1}{m} S^2_{yr}$.

<u>RA imputation</u>

For RA imputation, model ξ is assumed. The components of the variance estimator are:

$$\hat{V}_{\text{DIF}}(d) = \frac{N^2(1-f)}{n^2} \left\{\sum_o x_k(d)\right\} \hat{\sigma}^2$$

$$\hat{V}_{\text{IMP}}(d) = \frac{N^2}{n^2} \left\{\sum_o x_k(d)\right\} \left\{\frac{\sum_o x_k(d)}{\sum_r x_k} + 1\right\} \hat{\sigma}^2$$

$$\hat{V}_{\text{MIX}}(d) = \frac{N^2(1-f)}{n^2} \left\{\sum_o x_k(d)\right\} \left\{\frac{\sum_r x_k(d)}{\sum_r x_k} - 1\right\} \hat{\sigma}^2$$

with $\hat{\sigma}^2 = \sum_r e_k^2 / \sum_r x_k$ and $e_k = y_k - \hat{B}_r x_k$.

Note that RM imputation is a special case of RA imputation when $x_k = 1$ for all k.

<u>NN imputation</u>

The same model as for RA imputation is assumed and the components of the variance estimator are:

$$\hat{V}_{\text{DIF}}(d) = 0$$

$$\hat{V}_{\text{IMP}}(d) = \frac{N^2}{n^2} l_d \left\{\left(\frac{l_d}{m} + 1\right) \frac{\sum_r x_k}{m} + \frac{\sum_o x_k(d)}{l_d}\right\} \hat{\sigma}^2$$

$$\hat{V}_{\text{MIX}}(d) = \frac{N^2(1-f)}{n^2} l_d \left\{\frac{\sum_r x_k(d)}{m} - \frac{\sum_o x_k(d)}{l_d}\right\} \hat{\sigma}^2$$

with $\hat{\sigma}^2 = \sum_r e_k^2 / \sum_r x_k$ and $e_k = y_k - \hat{B}_r x_k$.

Note that HD imputation is a special case of NN imputation when $x_k = 1$ for all k.

4. Methods Based on the Jackknife Technique

We first explain the procedure for estimation of the population total Y_U. We then adapt it to estimation of the domain total $\hat{Y}_U(d)$.

The principle of the jackknife technique is to recalculate the estimator after deleting a unit from the sample. The variance between the recomputed estimates is used to obtain an estimate of the variance of the estimator calculated using the complete sample. After the deletion of unit j, the estimator of the whole population total (Y_U) is given by

$$\hat{Y}_{\bullet s}^{(j)} = \frac{N}{n-1} \sum_{k \neq j \in s} y_{\bullet k}$$

where the superscript j denotes that unit j was deleted. This is performed for all units $j \in s$. The jackknife variance estimator is

$$\hat{V} = \frac{n-1}{n} \sum_{j \in s} (\hat{Y}_{\bullet s}^{(j)} - \hat{Y}_{\bullet s})^2.$$

For data sets containing imputed values, this estimator does not take the imputation into account. Rao and Shao (1992) proposed a jackknife variance estimator that corrects the estimator by adjusting the imputed values when the deleted unit is in the response set. For some imputation methods, the adjusted values are the reimputed values based on the reduced response set after deletion of the jth unit. If the jth unit deleted is a nonrespondent, the imputed values are unchanged. The data set after adjusting the imputed values is given by

$$y_{\bullet k}^{(aj)} = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k + a_k^{(j)} & \text{if } k \in o \text{ and } j \in r \\ \hat{y}_k & \text{if } k \in o \text{ and } j \in o \end{cases}$$

where $y_{\bullet k}^{(aj)}$ is the adjusted imputed value and $a_k^{(j)}$ is called the adjustment. The jackknife variance estimator is then given by

$$\hat{V}_{\text{JACK}} = \frac{n-1}{n} \sum_{j \in s} (\hat{Y}_{\bullet s}^{(aj)} - \hat{Y}_{\bullet s}^{(a)})^2$$

where $\hat{Y}_{\bullet s}^{(aj)} = \frac{N}{n-1} \sum_{k \neq j \in s} y_{\bullet k}^{(aj)}$ and $\hat{Y}_{\bullet s}^{(a)} = \frac{1}{n} \sum_{j \in s} \hat{Y}_{\bullet s}^{(aj)}$. This estimator works well when the finite population correction (fpc), $1-f$ with $f = n/N$ is negligible. However, when the fpc is nonnegligible, the jackknife variance estimator overestimates the true variance and its bias can be substantial (Rancourt, Lee and Särndal, 1993). A direct application of the fpc to the jackknife estimator would overcorrect, thereby underestimating the true variance. Only the sampling variance component needs to be corrected. A proper correction can be obtained by first decomposing the total variance into two components, sampling variance and other (mostly imputation) variance. The fpc should be applied only to the sampling variance component. However, the jackknife variance estimator does not readily yields these components. Lee, Rancourt and Särndal (1995) proposed the following fpc-corrected jackknife variance estimator

$$\hat{V}_{\text{JACK}}^* = \hat{V}_{\text{JACK}} - N\hat{S}_{yU}^2$$

where $\hat{S}_{yU}^2$ is an unbiased estimator of the population variance S_{yU}^2. Since the response mechanism is uniform by assumption, we can use $\hat{S}_{yU}^2 = S_{yr}^2$ which is the respondent variance. A simple expression for the fpc-corrected jackknife variance estimator is then given by

$$\hat{V}_{\text{JACK}}^* = \hat{V}_{\text{JACK}} - NS_{yr}^2.$$

This adjustment yields almost unbiased estimates of the variance. (For a more detailed discussion, see Lee, Rancourt and Särndal, 1995).

The adjustment values $a_k^{(j)}$ used in our calculation of $\hat{V}_{\text{JACK}}$ depend on the particular imputation method. Note that adjustments are needed only for $j \in r$, $k \in o$. The adjustment for RM and HD is found in Rao (1991), Rao and Sitter (1992), the one for RA imputation in Rao (1991), Rao and Sitter (1995), and the one for NN imputation (which agrees with the RA adjustment) is found in Kovar and Chen (1994).

RM imputation

The adjustment for RM imputation is given by $a_k^{(j)} = \bar{y}_r^{(j)} - \bar{y}_r$ where

$$\bar{y}_r^{(j)} = \frac{1}{m-1} \sum_{k \neq j \in r} y_{\bullet k}.$$

The adjusted imputed value is then

$$\hat{y}_k + a_k^{(j)} = \bar{y}_r + (\bar{y}_r^{(j)} - \bar{y}_r) = \bar{y}_r^{(j)}.$$

Note that the same value would be obtained by RM imputation based on the reduced response set $r - \{j\}$.

HD imputation

The adjustment for HD imputation is the same as the one used for RM imputation. The adjusted imputed value is given by

$$\hat{y}_k + a_k^{(j)} = y_{g(k)} + (\bar{y}_r^{(j)} - \bar{y}_r).$$

RA imputation

The adjustment for RA imputation is $a_k^{(j)} = \left(\frac{\bar{y}_r^{(j)}}{\bar{x}_r^{(j)}} - \frac{\bar{y}_r}{\bar{x}_r} \right) x_k$. The adjusted imputed value is

$$\hat{y}_k + a_k^{(j)} = \frac{\bar{y}_r}{\bar{x}_r} x_k + \left(\frac{\bar{y}_r^{(j)}}{\bar{x}_r^{(j)}} - \frac{\bar{y}_r}{\bar{x}_r} \right) x_k = \frac{\bar{y}_r^{(j)}}{\bar{x}_r^{(j)}} x_k.$$

Note that the same value would be obtained by RA imputation based on the reduced response set $r - \{j\}$.

NN imputation

As mentioned earlier, the same adjustment as for RA imputation is used here and thus, we have

$$\hat{y}_k + a_k^{(j)} = y_{g(k)} + \left(\frac{\bar{y}_r^{(j)}}{\bar{x}_r^{(j)}} - \frac{\bar{y}_r}{\bar{x}_r} \right) x_k .$$

Now consider jackknife variance estimation for the estimator $\hat{Y}_{\bullet s}(d)$ of the domain total $Y_U(d)$. Define the completed data set after imputation with unit j deleted as:

$$y_{\bullet k}^{(aj)}(d) = \begin{cases} y_{\bullet k}^{(aj)} & \text{if } k \in d \\ 0 & \text{if } k \notin d. \end{cases}$$

The jackknife variance estimator for $\hat{Y}_{\bullet s}(d)$ is then given by

$$\hat{V}_{\text{JACK}}^{*}(d) = \hat{V}_{\text{JACK}}(d) - N\hat{S}_{yU}^2(d)$$

where $\hat{V}_{\text{JACK}}(d)$ is computed using $y_{\bullet k}^{(aj)}(d)$ and $\hat{S}_{yU}^2(d)$ is an unbiased estimator for $S_{yU}^2(d)$.

5. Implementation in GES

The results given in this paper can be implemented within the GES general framework without involving heavy reprogramming. For the model-assisted approach, the terms $\hat{V}_{\text{DIF}}(d)$, $\hat{V}_{\text{IMP}}(d)$ and $\hat{V}_{\text{MIX}}(d)$ have to be added in GES to the ordinary formula currently used. For the jackknife technique, adjusted imputed values have to be calculated when the jth unit is deleted from the respondent set. Note that the adjusted imputed values are to be used only in variance estimation.

If the multiple imputation technique were used, no changes or no additions to GES would be needed. The current GES could be used several times and then, point and variance estimates would then be calculated outside of GES.

For both the model-assisted approach and the jackknife technique, identification of the imputed values and the specification of the imputation method(s) (via imputation flags) are required to identify the proper variance formula specified in this paper.

We have assumed that imputation groups coincide with the strata. Hence there is no need for specifying the imputation groups. However, the imputation groups need not coincide with strata.

It is desirable that the future GES be able to provide separate variance estimates for the sampling and the imputation components. This can be achieved straitforwardly with the model-assisted approach and the multiple imputation technique. For the jackknife technique, one can include the correction given in Section 4 and calculate the two components as follows:

$$\hat{V}_{\text{SAM}}(d) = \frac{N^2(1-f)}{n}\hat{S}_{yU}^2(d)$$

$$\hat{V}_{\text{IMP}}(d) = \hat{V}_{\text{JACK}}(d) - \frac{N^2}{n}\hat{S}_{yU}^2(d).$$

In the above decomposition, the MIX term (which could be nonzero) is included in $\hat{V}_{\text{IMP}}(d)$.

6. Future Work

In this paper, we have described suitable methods for variance estimation in the presence of imputation. It was assumed that imputation groups coincided with strata and the response mechanism was uniform. We have only dealt with the simple expansion estimator. It represents, however, the first step towards the goal of proper variance estimation in estimation packages such as GES.

GES has been built upon the Generalized Regression Estimator (GREG) theory. An important feature of GES is that it can incorporate auxiliary information in estimation. Extension of the variance estimation methods described in this paper to the Generalized Regression Estimator is one of the next priorities for research.

The next hurdle to overcome is the case of more general imputation groups than those considered in this paper. This is important since in reality imputation groups often cut across the stratum boundaries. For the case where some imputation groups are composed of collapsed strata, conservative variance estimates could be obtained using the collapsed strata as design strata.

Only four imputation methods have been considered. Variance formulae will have to be derived for other methods. Further work is also needed for cases when more than one imputation method is used within the same data set.

Violations of assumptions on the model for imputation and/or on the uniform response mechanism will have some effects on the validity of the methods presented in this paper. Lee, Rancourt and Särndal (1994) studied robust aspect of some variance estimation procedures under several scenarios of violated assumptions of the imputation model and the response mechanism. They found that some variance estimation procedures can be seriously affected under some conditions. More research is needed in this area. Also, there has been little done so far to evaluate the robustness of the jackknife variance estimator for data sets with imputed values.

Finally, it is important to realize that domain estimation can be seriously biased, especially when RM or HD imputation is used even under a uniform response mechanism.

7. References

Deville, J.-C., and Särndal, C.E. (1991). Estimation de la variance en presence de donnée imputées. In *Proceedings of Invited Papers for the 48th Session of the International Statistical Institute*, Book 2, Topic 17, 3/17.

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

Kovar, J., and Chen, E. (1994). Jackknife Variance Estimation of Imputed Survey Data. *Survey Methodology*, 20, 45-52.

Lee, H., Rancourt, E., and Särndal, C.E. (1991). Experiment with Variance Estimation from Survey Data with Imputed Values. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 690-695.

Lee, H., Rancourt, E., and Särndal, C.E. (1994). Experiment with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics*, 10, 231-243.

Lee, H., Rancourt, E., and Särndal, C.E. (1995). Jackknife Variance Estimation for Survey Data with Imputed Values. (Presented at the Annual Conference of the Statistical Society of Canada, Montreal, July 9-13, 1995).

Rancourt, E., Lee, H., and Särndal, C.E. (1994). Variance Estimation under More Than one Imputation Method. In *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 374-379.

Rancourt, E., Lee, H., and Särndal, C.E. (1994). Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse. *Survey Methodology*, 20, 137-147.

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

Rao, J.N.K. (1990). Variance Estimation under Imputation for Missing Data. Technical Paper, Statistics Canada, Ottawa.

Rao, J.N.K. (1991). Jackknife Variance Estimation under Imputation for Missing Data. Technical Paper, Statistics Canada, Ottawa.

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

Rao, J.N.K., and Sitter, R.R. (1995). Jackknife Variance Estimation under Imputation for Missing Survey Data. *Biometrika*, 82, 453-460.

Rubin, D.B. (1978). Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 20-34.

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

Särndal, C.E. (1990). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. In *Proceedings of Symposium '90: Measurement and Improvement of Data Quality*, 337-347. Ottawa: Statistics Canada.