

## ESTIMATING THE VARIANCE OF THE GENERALIZED REGRESSION ESTIMATOR IN THE PRESENCE OF IMPUTATION FOR THE GENERALIZED ESTIMATION SYSTEM

F. Gagnon, H. Lee, E. Rancourt and C.-E. Särndal<sup>1</sup>

### ABSTRACT

It is well known that when imputation is used, application of the usual variance estimators on the completed data set leads to underestimation of the true variance. A number of methods currently exist for dealing with this problem and it would be desirable to incorporate such methods into a system like Statistics Canada's Generalized Estimation System (GES), which is based on the Generalized Regression Estimator (GREG). However, the existing methods are designed primarily for estimation without auxiliary data. The purpose of this paper is to discuss variance estimation for the GREG estimator in the presence of imputation for GES.

**KEY WORDS:** Model-assisted approach; Single imputation; Completed data; Imputation variance; Ratio imputation; Nearest neighbour imputation.

### RÉSUMÉ

Il est bien connu que lorsqu'on utilise l'imputation, l'application des estimateurs habituels de variance appliqués sur l'ensemble de données complété mène à une sous-estimation de la véritable variance. Plusieurs méthodes existent présentement pour traiter de ce problème. Il serait souhaitable d'incorporer de telles méthodes dans un système comme le Système généralisé d'estimation (SGE) de Statistique Canada, qui est basé sur l'estimateur par la régression généralisé (GREG). Cependant, les méthodes existantes sont d'abord conçues pour l'estimation en l'absence de données auxiliaires. Le but de cet article est de discuter de l'estimation de la variance pour l'estimateur GREG en présence d'imputation.

**MOTS CLÉS:** L'approche assisté d'un modèle; imputation simple; données complétées; variance due à l'imputation; imputation par quotient; imputation par le plus proche voisin.

### 1. INTRODUCTION

During the last two decades, developments in statistical theory and advances in programming techniques and computers have given birth to software designed to process survey data. Statistics Canada has developed its own survey software, one part of which is the Generalized Estimation System (GES), whose methodology description can be found in Estevao, Hidiroglou and Särndal (1995). The GES is a microcomputer package that has been developed to produce estimates from survey data as well as the precision of those estimates. The system produces domain estimates using various estimators such as Horvitz-Thompson, post-stratified, raking ratio, ratio and regression, for totals, means and ratios.

The GES is designed to accommodate several designs and sampling plans and it has been developed for use with complete data files. However, there are usually missing data in all surveys, resulting in incomplete data files. To circumvent this problem, imputation is often used to complete these files. As a consequence, the estimation system can be used as usual on the *completed data set*.

If the imputation is unbiased, using the GES or a similar system will produce unbiased point estimates. On the other hand, when variance estimates are produced by the system on completed data sets, the true variances are underestimated, often severely. This is caused mainly by the added variability due to imputation which is completely missed by the ordinary variance formula applied to the completed data set, as

---

<sup>1</sup> F. Gagnon, H. Lee and E. Rancourt, Statistics Canada, 11<sup>th</sup> floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6; and C.-E. Särndal, Département de mathématiques et statistique, Université de Montréal, C.P. 6128, succursale A, Montréal, Québec, Canada, H3C 3J7.

will be seen in Section 2.

To address this problem, a number of methods have recently been developed. The first one is multiple imputation, proposed by Rubin (1978, 1987). This method requires imputing two or more values for each missing datum.

In this paper, we consider rather the situation where a data set has been completed using *single imputation*. Single imputation is a common practice in many statistical agencies, where many surveys are carried out and where data are collected on a very large number of variables. There is, therefore, a need for producing proper variance estimates in the presence of single imputation.

To this end, a number of variance estimation methods have been developed. These are the model-assisted method, developed by Särndal (1992), the two-phase approach by Rao and Sitter (1995) and the Jackknife technique by Rao and Shao (1992).

We are interested in providing variance estimates for all situations currently handled by the GES. For the case of domain estimation, the theory has been developed in Lee, Rancourt and Särndal (1995) for the Horvitz-Thompson estimator using the model-assisted approach. The purpose of this paper is to provide variance estimators for the Generalized REGression estimator (GREG) for a number of imputation methods. In particular, ratio and nearest neighbour imputation are studied in this paper.

In Section 2, the model-assisted approach is developed for the GREG estimator in the presence of imputation. The resulting variance estimators have been tested through simulations, the results of which are described and discussed in Section 3. Finally, some concluding remarks are presented in Section 4.

## 2. THEORETICAL DEVELOPMENT

### 2.1 Background

Let  $U = \{1, \dots, k, \dots, N\}$  be the index set of the population. A probability sample  $s$ , of size  $n$ , is drawn from  $U$  by a sampling design  $p$ . Let also  $r$  of size  $m$ , and  $o$  of size  $n-m$  be, respectively, the set of respondents and the set of nonrespondents. Therefore,  $s = r \cup o$ . Let  $y_k$  be the value of the variable of interest,  $y$ , for unit  $k$ . We assume that  $y_k > 0$  for all  $k \in U$ . The parameter to estimate is the population total of  $y$ ,  $Y_U = \sum_U y_k$ .

The respondent set  $r$  is realized by a response mechanism  $q(\cdot|s)$ . That is,  $q(r|s)$  is the (unknown) conditional probability of realizing  $r$ , given  $s$ . We assume that  $q(\cdot|s)$  is a uniform response mechanism. It is also assumed that the relation between the  $y$ -variable and the auxiliary column vector,  $\underline{x}$  is expressed by the linear regression model  $\zeta$  given by

$$\zeta: y_k = \gamma_k' \underline{x}_k + \epsilon_k, \quad (2.1)$$

$E_\zeta(\epsilon_k) = 0$ ,  $E_\zeta(\epsilon_k^2) = c_k \sigma^2$ ,  $E_\zeta(\epsilon_k, \epsilon_l) = 0$ ,  $k \neq l$ , for some constant  $c_k$  assumed to be known.

Values that are missing due to nonresponse are imputed using a single-value imputation method. The *completed data set* is given by  $\{y_{\cdot k} : k \in s\}$  where

$$y_{\cdot k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k & \text{if } k \in o \end{cases} \quad (2.2)$$

and where  $\hat{y}_k$  is the imputed value.

We will focus on two imputation methods: ratio (RA) imputation and nearest neighbour (NN) imputation, where an auxiliary variable  $z$  for imputation is available. For RA imputation, the imputed value for unit  $k$  is given by

$$\hat{y}_k = \hat{B} z_k,$$

where  $\hat{B} = \sum_r y_i / \sum_r z_i$ . The underlying ratio imputation model, denoted  $\xi$ , is given by  $\xi: y_k = \beta z_k + \delta_k$ , where  $E_\xi(\delta_k) = 0$ ,  $E_\xi(\delta_k^2) = z_k \sigma^2$  and  $E_\xi(\delta_k, \delta_l) = 0$  for all  $k \neq l$ . For NN imputation, we have

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

where  $l(k)$  is the donor unit determined in such a way that the minimum of the distance  $|z_l - z_k|$  is obtained, among all  $l \in r$ , when  $l = l(k)$ .

### 2.2 Variance Estimation for the Imputed GREG Estimator

The GREG estimator used in the GES is defined by:

$$\hat{Y}_s = \sum_s w_k y_k$$

with  $w_k = a_k g_k$ , where  $a_k = 1/\pi_k$  is the sampling weight and where the "g-factor",

$$g_k = 1 + \left( \sum_U \underline{x}_i - \sum_s a_i \underline{x}_i \right)' \left( \sum_s a_i \underline{x}_i \underline{x}_i' / c_i \right)^{-1} \frac{\underline{x}_k}{c_k}$$

modifies the sampling weight by incorporating the auxiliary information contained in the known total  $\sum_U \underline{x}_i$ . The estimator  $\hat{Y}_s$  is designed for the 100% response case. In the presence of imputation, the *imputed GREG estimator* is computed instead. It is given by:

$$\hat{Y}_{\cdot s} = \sum_s w_k y_{\cdot k}. \quad (2.3)$$

The total error of the imputed estimator is given by

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

where  $\hat{Y}_s - Y_U$  represents the *sampling error* and  $\hat{Y}_{*s} - \hat{Y}_s$ , the *imputation error*. Then, the total variance is given by

$$V_{\text{TOT}} = E_p E_q (\hat{Y}_{*s} - Y_U)^2 = V_{\text{SAM}} + V_{\text{IMP}} + 2V_{\text{MIX}} \quad (2.4)$$

where

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

is the *sampling variance*,

$$V_{\text{IMP}} = E_p \left\{ E_q [(\hat{Y}_{*s} - \hat{Y}_s)^2 | s] \right\}$$

is the *imputation variance* and

$$V_{\text{MIX}} = E_p \left\{ (\hat{Y}_s - Y_U) E_q [(\hat{Y}_{*s} - \hat{Y}_s) | s] \right\}$$

is the mixed term.

The estimators of the three components of (2.4) are denoted  $\hat{V}_{\text{SAM}}$ ,  $\hat{V}_{\text{IMP}}$  and  $\hat{V}_{\text{MIX}}$ . They are computed as follows.

Calculation of  $\hat{V}_{\text{SAM}}$ : Start with the ordinary variance estimator used in the GES for the case of 100% response,

$$\hat{V}_{\text{ORD}} = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) (g_k e_k) (g_l e_l)$$

where  $e_k = y_k - \hat{G}' x_k$  with

$$\hat{G} = \left( \sum_s a_k x_k x_k' / c_k \right)^{-1} \sum_s a_k x_k y_k / c_k.$$

Compute  $\hat{V}_{\text{ORD}}$  replacing  $\{y_k : k \in s\}$  by  $\{y_k^* : k \in s\}$ , where the  $y_k^*$  are suitably defined "pseudo-values" used for variance calculation only. The result of this operation is  $\hat{V}_{\text{SAM}}$ . The set  $\{y_k^* : k \in s\}$  could be the same as the completed data set  $\{y_k : k \in s\}$  or different. For RA imputation, we suggest to use

$$y_k^* = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k + d_k^* & \text{if } k \in o \end{cases} \quad (2.5)$$

where the value  $d_k^*$  is randomly selected from the set of respondents' imputation residuals,  $\{d_k = y_k - \hat{B}z_k : k \in r\}$ . For NN imputation, however, the same completed data set is used for both point and variance estimation, so that  $y_k^* = y_k$  for all  $k \in s$ .

Calculation of  $\hat{V}_{\text{IMP}}$ : Derive an expression for  $\hat{V}_{\text{IMP}}$  appealing to the ratio imputation model  $\xi$  and following the model-assisted approach. For RA imputation, use

$$\hat{V}_{\text{IMP}} = \left\{ \frac{\left( \sum_o w_k z_k \right)^2}{\sum_r z_k} + \sum_o w_k^2 z_k \right\} \hat{\sigma}^2 \quad (2.6)$$

where  $\hat{\sigma}^2$  is given by:

$$\hat{\sigma}^2 = \sum_r d_k^2 / \sum_r z_k. \quad (2.7)$$

For NN imputation, use

$$\hat{V}_{\text{IMP}} = \left\{ 2 \sum_o w_k^2 z_k + \left[ \left( \sum_o w_k \right)^2 - \sum_o w_k^2 \right] \frac{\bar{z}_r}{m} \right\} \hat{\sigma}^2 \quad (2.8)$$

with  $\bar{z}_r = \sum_r z_k / m$ .

Computation of  $\hat{V}_{\text{MIX}}$ : For RA imputation, use

$$\hat{V}_{\text{MIX}} = \left\{ \frac{\sum_o w_k z_k \sum_r (w_k - 1) z_k}{\sum_r z_k} - \sum_o (w_k - 1) w_k z_k \right\} \hat{\sigma}^2 \quad (2.9)$$

where  $\hat{\sigma}^2$  is given by (2.7). In many situations,  $\hat{V}_{\text{MIX}}$  is small compared to  $\hat{V}_{\text{IMP}}$ . If  $w_k$  is constant for all  $k$ , we have  $\hat{V}_{\text{MIX}} = 0$ . For NN imputation, use  $\hat{V}_{\text{MIX}} = 0$ .

The estimators  $\hat{V}_{\text{IMP}}$  and  $\hat{V}_{\text{MIX}}$  were constructed in such a way that they respectively satisfy  $E_{\xi} \{ E_p E_q (\hat{V}_{\text{IMP}}) - V_{\text{IMP}} \} \approx 0$  and  $E_{\xi} \{ E_p E_q (\hat{V}_{\text{MIX}}) - V_{\text{MIX}} \} \approx 0$  (see Rancourt, Särndal and Lee, 1994).

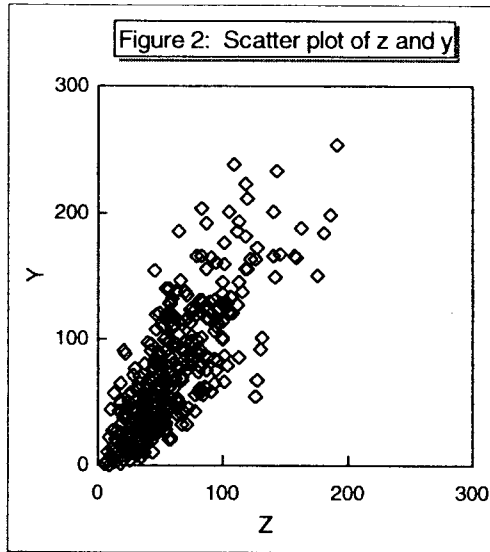
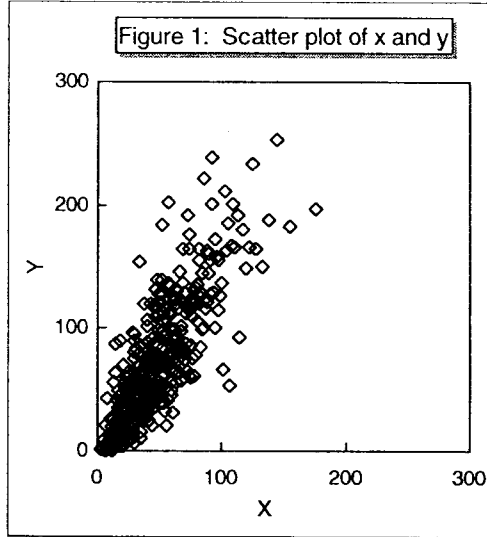
### 3. SIMULATION RESULTS

#### 3.1 Description of the Simulation Study

A Monte Carlo study was conducted in order to examine the basic statistical properties of the proposed estimators for  $V_{\text{TOT}}$  of the ratio estimator when ratio or nearest neighbour imputation is used.

We drew  $R=2000$  repeated simple random samples  $s$  of size  $n=100$  from a population of size  $N=400$ . This artificial population was generated from a ratio model. Two auxiliary variables are attached to the realized population: one for imputation ( $z$ ) and the other for estimation ( $x$ ). These two auxiliary variables have the same level of correlation coefficient (0.8) with the  $y$ -variable. The  $x$ -values were generated from a  $\Gamma(3,16)$  distribution. Then the  $y$ -values were also generated from a gamma distribution with mean  $1.5 x_k$  and variance  $d^2 x_k$ . The constant  $d$  was determined to

obtain the desired (0.8) correlation coefficient. The variable  $z$  was generated similarly to obtain the same correlation with  $y$ . For more details on the generation of the population, see Lee, Rancourt and Särndal (1994). Figure 1 shows the scatter plot of  $x$  and  $y$  and Figure 2 shows the plot of  $z$  and  $y$ .



The ratio estimator for the population total  $Y_U$  is obtained from (2.3) with  $x_k = x_k = c_k$ . For each sample  $s$ , a set of respondents  $r$  was realized assuming a uniform response mechanism with  $E_q(m) = 70$ . This means that the average response rate is 70%. For each realized response set  $r$ , two completed data sets were created: one by RA imputation and the other one by NN imputation. Then the (imputed) ratio estimator was calculated for each completed data set and the variance components  $\hat{V}_{SAM}$  and  $\hat{V}_{IMP}$  were also calculated, as

described in Section 2. Note that for the ratio estimator, the ordinary formula for the variance estimator (that is, the formula for 100% response) is

$$\hat{V}_{ORD} = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \frac{\bar{x}_U}{\bar{x}_s} \right)^2 \frac{1}{n-1} \sum_s \left( y_k - \frac{\bar{y}_s}{\bar{x}_s} x_k \right)^2.$$

Now  $\hat{V}_{SAM}$  is the result obtained when this  $\hat{V}_{ORD}$  is computed on  $\{y_k^* : k \in s\}$ , instead of on the data  $\{y_k : k \in s\}$ . Note that for RA imputation,  $y_k^*$  is given by (2.5), and that for NN imputation,  $y_k^* = y_k$  for all  $k \in s$ .

For the case considered in this simulation (simple random sampling and the ratio estimator), it is appropriate, for both RA and NN imputation, to assume  $\hat{V}_{MIX} = 0$ . We thus compute  $\hat{V}_{TOT} = \hat{V}_{SAM} + \hat{V}_{IMP}$  as our estimate of the total variance of  $\hat{Y}_{s^*}$ .

Tables 1 and 2 show the following summary measures (Monte Carlo Average, MCA, and Monte Carlo Variance, MCV) computed from the simulation:

$$MCA(\hat{V}_{SAM}) = \frac{1}{R} \sum_{j=1}^R \hat{V}_{SAM_j}$$

$$MCA(\hat{V}_{IMP}) = \frac{1}{R} \sum_{j=1}^R \hat{V}_{IMP_j}$$

$$MCA(\hat{V}_{TOT}) = MCA(\hat{V}_{SAM}) + MCA(\hat{V}_{IMP}).$$

We compare  $MCA(\hat{V}_{TOT})$  to the total variance of  $\hat{Y}_{s^*} = \bar{x}_U \bar{y}_{s^*} / \bar{x}_s$ , as estimated from the Monte Carlo study by

$$MCV(\hat{Y}_{s^*}) = \frac{1}{R-1} \sum_{j=1}^R \left\{ \hat{Y}_{s^*_j} - \frac{1}{R} \sum_{j=1}^R \hat{Y}_{s^*_j} \right\}^2.$$

As well, a 95% confidence interval was constructed for each iteration to investigate its coverage rate of the true population value  $Y_U$ .

### 3.2 Results for Ratio Imputation

As defined in Subsection 2.1, the ratio imputed value  $\hat{y}_k$  is given by:

$$\hat{y}_k = \frac{\sum_r y_i}{\sum_r z_i} z_k.$$

In order to estimate the sampling variance component correctly,  $\hat{V}_{SAM}$  is calculated with the pseudo-values  $y_k^*$ , as defined by (2.5). Note that the  $y_k^*$  are used for sampling variance calculation only. The imputation variance  $V_{IMP}$  is estimated by (2.6).

The Monte Carlo Average of the variance estimator  $\hat{V}_{TOT} = \hat{V}_{SAM} + \hat{V}_{IMP}$  equals 1.75, which is close to the Monte Carlo Variance (=1.79). Recall that we assume that  $\hat{V}_{MIX} = 0$ . Using just  $\hat{V}_{SAM}$  as the variance estimator would have grossly underestimated the true variance, because the Monte Carlo Average of  $\hat{V}_{SAM}$  equals 1.10 only.

The true total  $Y_U$  is contained in 94.6% of the confidence intervals built using  $\hat{V}_{TOT}$ . This coverage rate is very close to the nominal value of 95%. By comparison, confidence intervals constructed with variance estimates that ignore the imputation variance component yield much lower coverage rates, because the variance is severely underestimated. For example, we found that the coverage rate was only 88% when the confidence interval was computed using only  $\hat{V}_{SAM}$  as the estimator of total variance (thus effectively setting  $\hat{V}_{IMP}$  to zero). Even worse, the coverage rate was 80% when the confidence interval was based on the ordinary formula  $\hat{V}_{ORD}$  computed on the data  $\{y_k : k \in s\}$  instead of on  $\{y_k^* : k \in s\}$ .

**Table 1. Relative Bias and Coverage Rate of the Proposed Variance Estimator of the Ratio Estimator under Ratio Imputation**

MCA( $\hat{V}_{SAM}$ )	MCA( $\hat{V}_{IMP}$ )	MCA( $\hat{V}_{TOT}$ )	MCV( $\hat{Y}_{..}$ )	Relat. Bias	Cov. Rate
1.10	0.65	1.75	1.79	-2.2%	94.6 %

(numbers are in millions)

### 3.3 Results for the Nearest Neighbour Imputation

When using the NN imputation method, the imputed value  $\hat{y}_k$  is given by:

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

The variance components  $V_{SAM}$  and  $V_{IMP}$  are estimated using  $\hat{V}_{ORD}$  applied to  $\{y_k^* : k \in s\}$ , which is the same as  $\{y_k : k \in s\}$ , and formula (2.8), respectively.

Table 2 shows that the Monte Carlo Average of the estimator  $\hat{V}_{TOT} = \hat{V}_{SAM} + \hat{V}_{IMP}$  (= 2.20) is close to the Monte Carlo Variance (=2.23). It is clear that the use of  $\hat{V}_{SAM}$ (=1.11) only would have grossly underestimated the total variance.

Confidence intervals built using  $\hat{V}_{TOT}$  covered the true value  $Y_U$ , 93.8% of the time. In comparison, the coverage rate was only 84% when the confidence intervals were computed by using only the component  $\hat{V}_{SAM}$  to estimate the total variance.

**Table 2. Relative Bias and Coverage Rate of the Proposed Variance Estimation of the Ratio Estimator under Nearest Neighbour Imputation**

MCA( $\hat{V}_{SAM}$ )	MCA( $\hat{V}_{IMP}$ )	MCA( $\hat{V}_{TOT}$ )	MCV( $\hat{Y}_{..}$ )	Relat. Bias	Cov. Rate
1.11	1.09	2.20	2.23	-1.7%	93.8 %

(numbers are in millions)

## 4. CONCLUSION

The Generalized Estimation System is routinely used at Statistics Canada. However, it does not currently have a facility to calculate the proper variance when imputation is used to complete the data set. It is important to have such a facility in the GES, since almost all data sets include imputed values, often in a high proportion. Such a facility would enable survey practitioners not only to calculate proper variance estimates but also to understand better the total variability. Then, resources can be better allocated so as to maximize their return in terms of data quality. For instance, if an estimate of the imputation variance is a large component of the total variance, we may want to allocate more resources to the data collection, capture and follow-up processes so that imputation is minimized. Another possibility would be to use a more efficient imputation method. On the other hand, if the sampling variance is a large component of the total variance, we may allocate more resources to increase the sample size or to make the sample design more efficient.

It is desired to provide an essentially unbiased variance estimator for the GREG estimator used in the GES in the presence of the single-value imputation methods that are most frequently used. These methods are: ratio imputation, nearest neighbour imputation, respondent mean imputation and hot-deck imputation. At Statistics Canada, the use of single imputation (as opposed to multiple imputation) is likely to continue. That is why proper variance estimates are needed. To meet this need, we have proposed two variance estimators for the ratio estimator: one for the RA imputation method, and one for the NN imputation method, as a step towards the ultimate goal.

We showed in the paper that the proposed variance estimators work very well under the simulated situation. However, whether they work in a more general context remains to be verified. Furthermore, the problem of variance estimation for imputation methods such as hot-deck and mean imputation has yet to be addressed for the GREG estimator. Also, the

work reported here concentrates on the situation where the domain of estimation is the whole population, the imputation group coincides with the whole population, the response mechanism is uniform, and the model assumptions are well satisfied. Obviously, there is more work to be done to obtain more general results. This includes an extension to the hot-deck and respondent mean imputation methods, to estimation for arbitrary domains, to general imputation groups and to nonuniform response mechanisms.

## REFERENCES

- Estevao, V., Hidioglou, M.A., and Särndal, C.-E. (1995). "Methodological principle for a generalized estimation system at Statistics Canada", *Journal of Official Statistics*, 11, 181-204.
- 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). "Variance estimation in the presence of imputed data for the Generalized Estimation System", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 384-389.
- Rancourt, E., Särndal, C.-E., and Lee, H. (1994). "Estimation of the variance in presence of nearest neighbour imputation", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 888-893.
- 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). "Variance estimation under two-phase sampling with application to imputation for missing data", *Biometrika*, 82, 453-460.
- Rubin, D.B. (1978). "Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse", *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. (1992). "Methods for estimating the precision of survey estimates when imputation has been used", *Survey Methodology*, 18, 241-252.