SSC Annual Meeting, June 2001
Proceedings of the Survey Methods Section, pp.75-80

VARIANCE ESTIMATION IN THE PRESENCE OF IMPUTATION

Charlie Arcaro and Wesley Yung[1]

ABSTRACT

The problem of missing or incomplete survey data is a common problem faced by statistical agencies. In many surveys conducted by Statistics Canada, missing data is dealt with by replacing (or imputing) them with values derived from current or historical data. While imputation methods can produce unbiased point estimates, conventional variance estimators can lead to underestimation of the true variance if the imputed values are treated as true values. Several methods have been proposed to compute the variance in the presence of imputation. We study two of them; one based on an adjusted jackknife and the other based on a model-assisted approach. We compare the properties of these two methods using theoretical and empirical arguments

KEY WORDS: Single value imputation; Jackknife linearization; Generalized regression estimator; Mean and ratio imputation; Model assisted approach.

RÉSUMÉ

Le problème des données manquantes ou incomplètes est un problème commun rencontré dans les agences statistiques. Pour plusieurs enquêtes menées par Statistique Canada, les données manquantes sont traitées en les remplaçant (ou imputant) par des valeurs dérivées de données historiques ou courantes. Bien que les méthodes d'imputation puissent produire des estimations ponctuelles non-biasées, les estimateurs conventionnels de la variance peuvent sous-estimer cette dernière si les valeurs imputées sont traitées comme vraies. Plusieurs méthodes ont été proposées pour calculer l'estimation de la variance en présence d'imputation. Nous en étudions deux; une basée sur la méthode du jackknife ajusté et l'autre basée sur un modèle assisté. Nous comparons les propriétés de ces deux méthodes en utilisant des concepts théoriques et empiriques.

MOTS CLÉS: Approche basée sur un modèle; estimateur de régression généralisée; Imputation par quotient et par moyenne; imputation simple; jackknife linéarisé.

1. INTRODUCTION

The problem of dealing with missing or incomplete survey data is an issue faced by many statistical agencies. In particular the problem of estimating variances from such data is drawing more and more attention. The use of conventional variance estimation techniques on imputed data can lead to a substantial underestimation of the true variance. This is due to an extra component of variability being added to the dataset in the form the of imputed values. There is now a wide variety of literature available to address this problem. They include the model based approach, (Särndal, 1992), the adjusted jackknife technique (Rao and Shao, 1992), the two-phase approach (Rao and Sitter, 1995), the bootstrap (Shao and Sitter, 1996) and more recently the approach of Shao and Steel (1999). Some of these techniques have been extended since their original conception to estimators that use auxiliary data such as the Generalized Regression (GREG) estimator (Yung and Rao, 2000; Gagnon, Lee, Rancourt, and Särndal, 1996). All of these methods have provided theoretical background for different sampling estimators and designs. In addition to an adjusted jackknife variance estimator for the GREG estimator, Yung and Rao (2000) presented a jackknife linearization variance estimator that is computationally simpler than the normal jackknife variance estimation technique.

In this paper we will compare the jackknife linearization and model based variance estimators for the GREG estimator under weighted mean and weighted ratio imputation for a stratified simple

[1] Charlie Arcaro and Wesley Yung, Statistics Canada, 3rd floor, RH Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, arcacha@statcan.ca.

random sample (SRS) design. Section 2 describes the sample design and introduces the GREG estimator. Sections 3 and 4 present the model based and the jackknife linearization variance estimators respectively and their properties under a weighted ratio imputation assuming a single imputation class. These estimators are then decomposed into components that are based on sample respondent and non-respondent. Section 5 presents some empirical results based on a limited simulation study.

2. NOTATION AND SAMPLE DESIGN

We consider a stratified SRS design with population set U of size N partitioned into H design strata and containing N_h population units in each stratum. A sample of size n_h is drawn from each stratum of the sample set S to give a total sample size of $n = \sum_h n_h$. We define the response indicator variable $\delta_{hk} = 1$ if unit k in stratum h is in the response set and 0 otherwise. Let y_{hk} be the variable of interest for unit hk. We wish to estimate the population total $Y_U = \sum_h \sum_k y_{hk}$. For the non-respondent set, we replace the missing value y_{hk} with the imputed value $\hat{y}_{hk}$, giving the completed dataset $\{y_{\bullet hk} : hk \in s\}$ where $y_{\bullet hk} = y_{hk}\delta_{hk} + \hat{y}_{hk}(1-\delta_{hk})$.

At the stage of estimation, it is common to use auxiliary data to improve the precision of estimators. The GREG estimator is one such estimator. Let $\mathbf{x}_{hk}$ be an auxiliary vector that is observed for every sampled unit with known population total vector $\mathbf{X}$. The imputed GREG estimator is given by $\hat{Y}_{\bullet s} = \sum_s w_{hk}^* y_{\bullet hk}$, where $w_{hk}^* = w_{hk} g_{hk}$ is the GREG adjusted final weight, w_{hk} is the original design weight and the GREG adjustment factor is

$$g_{hk} = 1 + (\mathbf{X} - \hat{\mathbf{X}})'\mathbf{A}^{-1}\mathbf{x}_{hk}$$

with $\mathbf{A} = \sum_s w_{hk}\mathbf{x}_{hk}\mathbf{x}'_{hk}$ and $\hat{X} = \sum_s w_{hk} x_{hk}$.

In this paper, we consider two imputation methods: weighted mean and weighted ratio imputation. The weighted mean method simply imputes the weighted mean of the respondents for missing values. That is,

$$\hat{y}_{hk} = \frac{\sum_s w_{hk}^* y_{hk}\delta_{hk}}{\sum_s w_{hk}^* \delta_{hk}}.$$

The weighted ratio method assumes that an additional auxiliary variable, z_{hk}, is available for all units hk in the sample and is related to the variable of interest. For missing values, the weighted ratio method imputes

$$\begin{aligned} \hat{y}_{hk} &= \frac{\sum_s w_{hk}^* y_{hk}\delta_{hk}}{\sum_s w_{hk}^* z_{hk}\delta_{hk}} z_{hk} \\ &= \hat{\beta} z_{hk}. \end{aligned} \tag{2.1}$$

Noting that the weighted mean imputation is the same as the weighted ratio imputation with $z_{hk} = 1$ for all (hk), in what follows we will develop the weighted ratio method only.

3. MODEL BASED APPROACH

The model based approach assumes there exists an underlying model that explains the relationship between the variable of interest y and the imputation variable z. This model can be written as:

$$y_k = \beta z_k + \varepsilon_k$$

where $E_\xi(\varepsilon_k) = 0$, $V_\xi(\varepsilon_k) = \sigma^2 z_k$, $E_\xi(\varepsilon_k \varepsilon_l) = 0$ for $k \neq l$ and E_ξ, V_ξ denote expectation and variance with respect to the model.

The total variance for $\hat{Y}_{\bullet s}$ can be decomposed as

$$\begin{aligned} V_{TOT} &= E_p E_q E_\xi (\hat{Y}_{\bullet s} - Y_U)^2 \\ &= E_p E_q E_\xi (\hat{Y}_s - Y_U + \hat{Y}_{\bullet s} - \hat{Y}_s)^2 \\ &= V_{SAM} + V_{IMP} + 2V_{MIX} \end{aligned}$$

where E_p and E_q represent expectation with respect to the sampling design p and the response mechanism q. Here we assume an ignorable response mechanism here whereby the probability of response may depend on the sample and auxiliary data but not on y_{hk} (Särndal, 1992). Also $\hat{Y}_s = \sum_s w_{hk}^* y_{hk}$ is the GREG estimator assuming 100% response. Estimators for these components, $\hat{V}_{SAM}$, $\hat{V}_{IMP}$ and $\hat{V}_{MIX}$, are constructed so that they are approximately model unbiased for their respective parameters. The calculation of the three components is discussed in the following subsections:

3.1 Calculation of $\hat{V}_{SAM}$

To estimate V_{SAM}, we can treat the imputed values as true values and apply the usual variance estimator, i.e.,

$$\hat{V}_{SAM} = \sum_h \frac{n_h}{n_h - 1} \sum_k (e_{hk} - \bar{e}_h)^2 \qquad (3.1)$$

where

$e_{hk} = w^*_{hk}(y_{\bullet hk} - \hat{\mathbf{B}}'\mathbf{x}_{hk})$ and $\hat{\mathbf{B}} = \mathbf{A}^{-1}\sum_s w_{hk}\mathbf{x}_{hk}y_{\bullet hk}$

Note that, for simplicity, the finite population correction factor is ignored. This estimator will underestimate V_{SAM} under some deterministic imputation methods unless a correction term is introduced. It was proposed in Gagnon, Lee, Rancourt, and Särndal (1996) that a residual term $d_i = (y_i - \hat{\beta}z_i)\delta_i$ from the respondent set be added to the imputed value $\hat{y}_{hk}$. Here, $\hat{\beta}$ is a parameter estimate from the imputation model as defined in (2.1). With the residual term included, $\hat{V}_{SAM}$ is obtained by using (3.1) but replacing $y_{\bullet hk}$ with $y'_{\bullet hk} = y_{\bullet hk} + d_i(1 - \delta_{hk})$.

3.2 Calculation of $\hat{V}_{IMP}$

We derive an expression for $\hat{V}_{IMP}$ by calculating $E_\xi(\hat{Y}_{\bullet s} - \hat{Y}_s)^2 = E_\xi\left(\sum_s w^*_{hk}(\hat{y}_{hk} - y_{hk})(1-\delta_{hk})\right)^2$. Under a weighted ratio imputation, we have

$$\hat{V}_{IMP} = \left[\frac{\left(\sum_s w^*_{hk} z_{hk}(1-\delta_{hk})\right)^2}{\left(\sum_s w^*_{hk} z_{hk}\delta_{hk}\right)^2}\sum_s w^{*2}_{hk} z_{hk}\delta_{hk} + \sum_s w^{*2}_{hk} z_{hk}(1-\delta_{hk})\right]\hat{\sigma}^2$$

where $\hat{\sigma}^2 = \sum_s d^2_{hk}\delta_{hk} / \sum_s z_{hk}\delta_{hk}$ and the d_{hk} are defined as the residual term, d_i, assigned to the (hk)-th unit. For weighted mean imputation, we obtain $\hat{V}_{IMP}$ by substituting $z_{hk} = 1$ into the above equation.

3.3 Calculation of $\hat{V}_{MIX}$

We derive $\hat{V}_{MIX}$ by calculating

$$E_\xi(\hat{Y}_s - Y_U)(\hat{Y}_{\bullet s} - \hat{Y}_s) = E_\xi\left(\hat{Y}_s - Y_U\right)\left(\sum_s w^*_{hk}(\hat{y}_{hk} - y_{hk})(1-\delta_{hk})\right).$$

For weighted ratio imputation, we have

$$\hat{V}_{MIX} = \left[\frac{\sum_s w^*_{hk} z_{hk}(1-\delta_{hk})}{\sum_s w^*_{hk} z_{hk}\delta_{hk}}\sum_s w^*_{hk}(w^*_{hk} - 1)z_{hk}\delta_{hk} - \sum_s w^*_{hk}(w^*_{hk} - 1)z_{hk}(1-\delta_{hk})\right]\hat{\sigma}^2$$

For weighted mean imputation, we substitute $z_{hk} = 1$. Note that in most cases $\hat{V}_{MIX}$ will be very small compared to $\hat{V}_{TOT}$.

4. JACKKNIFE LINEARIZATION

Rao and Shao (1992) proposed an adjusted jackknife procedure to produce valid variance estimators in the presence of imputation for missing values. Yung and Rao (2000) extended the Rao-Shao adjusted jackknife to estimators that incorporate auxiliary data and presented linearization type variance estimators obtained by linearizing the adjusted jackknife variance estimators. For more on the jackknife linearization variance estimators, see Yung and Rao (1996). The jackknife linearization assumes a uniform response mechanism $q(\bullet|s)$ in that each unit has a constant (usually unobserved) probability of response that is also independent of other units.

Under a weighted ratio imputation we have

$$\hat{Y}_{\bullet s} = \sum_s w^*_{hk} y_{hk}\delta_{hk} + \sum_s w^*_{hk}\hat{y}_{hk}(1-\delta_{hk}) = \hat{\beta}\sum_s w^*_{hk} z_{hk},$$

where $\hat{\beta} = \sum_s w^*_{hk} y_{hk}\delta_{hk} / \sum_s w^*_{hk} z_{hk}\delta_{hk} = \hat{S}/\hat{T}$. To construct a jackknife variance estimator, we first define the jackknife weights when the j-th unit in g-th stratum has been deleted

$$w_{hk(gj)} = \begin{cases} 0 & \text{if } hk = gj \\ \dfrac{n_h}{n_h - 1} w_{hk} & \text{if } h = g,\ i \neq j \\ w_{hk} & \text{otherwise.} \end{cases}$$

The adjusted imputed values are defined as

$$\hat{y}^*_{hk(gj)} = \hat{y}_{hk} + \hat{\beta}_{(gj)} z_{hk} - \hat{\beta} z_{hk}$$

where $\hat{\beta}_{(gj)}$ is obtained from $\hat{\beta}$ by replacing the final weights, w^*_{hk}, by the jackknife adjusted final weights, $w^*_{hk(gj)} = w_{hk(gj)} g_{hk(gj)}$, and $g_{hk(gj)}$ is obtained from g_{hk} by replacing the design weights with the jackknife weights. The adjusted imputed estimator is then

$$\hat{Y}_{\bullet s(gj)} = \sum_s w^*_{hk(gj)} y_{hk} \delta_{hk} + \sum_s w^*_{hk(gj)} \hat{y}^*_{hk(gj)} (1-\delta_{hk}) = \hat{\beta}_{(gj)} \sum_s w^*_{hk(gj)} z_{hk}$$

and the jackknife variance estimator is

$$\hat{V}_J = \sum_g \frac{n_g}{n_g - 1} \sum_j (\hat{Y}_{\bullet s(gj)} - \hat{Y}_{\bullet s})^2. \qquad (4.1)$$

To obtain a jackknife linearization variance estimator, we note that

$$\hat{Y}_{\bullet s(gj)} - \hat{Y}_{\bullet s} = \hat{\beta}_{(gj)} \sum_s w^*_{hk(gj)} z_{hk} - \hat{\beta} \sum_s w^*_{hk} z_{hk} \approx \frac{\hat{Z}}{\hat{T}} \left[(\hat{S}_{(gj)} - \hat{S}) - \hat{\beta}(\hat{T}_{(gj)} - \hat{T}) \right] + \hat{\beta}(\hat{Z}_{(gj)} - \hat{Z})$$

where $\hat{Z} = \sum_s w^*_{hk} z_{hk}$ and $\hat{Z}_{(gj)}, \hat{S}_{(gj)}$ and $\hat{T}_{(gj)}$ are obtained from $\hat{Z}, \hat{S}$ and $\hat{T}$ respectively by replacing the GREG adjusted weights with the jackknife adjusted GREG weights. From the definition of the jackknife weights, one can show that

$$\hat{S}_{(gj)} - \hat{S} \approx \left(\frac{n_g}{n_g - 1} \right) (\bar{e}_{g,1} - e^*_{gj,1})$$

where

$$\bar{e}_{g,1} = \frac{1}{n_g} \sum_j e_{gj,1},$$

$$e^*_{gj,1} = w^*_{gj} (y_{gj} \delta_{gj} - \hat{\mathbf{B}}'_1 \mathbf{x}_{gj})$$

and

$$\hat{\mathbf{B}}_1 = \mathbf{A}^{-1} \sum_s w_{hk} \mathbf{x}_{hk} y_{hk} \delta_{hk}.$$

Similarly, we have

$$\hat{T}_{(gj)} - \hat{T} \approx \left(\frac{n_g}{n_g - 1} \right) (\bar{e}_{g,2} - e^*_{gj,2})$$

$$\hat{Z}_{(gj)} - \hat{Z} \approx \left(\frac{n_g}{n_g - 1} \right) (\bar{e}_{g,3} - e^*_{gj,3})$$

with

$$e^*_{gj,2} = w^*_{gj} (z_{gj} \delta_{gj} - \hat{\mathbf{B}}'_2 \mathbf{x}_{gj})$$

$$e^*_{gj,3} = w^*_{gj} (z_{gj} - \hat{\mathbf{B}}'_3 \mathbf{x}_{gj}),$$

and

$$\hat{\mathbf{B}}_2 = \mathbf{A}^{-1} \sum_s w_{hk} z_{hk} \mathbf{x}_{hk} \delta_{hk}$$

$$\hat{\mathbf{B}}_3 = \mathbf{A}^{-1} \sum_s w_{hk} z_{hk} \mathbf{x}_{hk}.$$

Combining these expressions with (4.1), we obtain the jackknife linearization variance estimator

$$\hat{V}_{LJ} = \sum_g \frac{n_g}{n_g - 1} \sum_j (e_{gj} - \bar{e}_g)^2 \qquad (4.2)$$

where

$$e_{gj} = w^*_{gj} \left[\frac{\hat{Z}}{\hat{T}} \left((y_{gj} - \hat{\beta} z_{gj}) \delta_{gj} - (\hat{\mathbf{B}}_1 - \hat{\mathbf{B}}_2)' \mathbf{x}_{gj} \right) + \hat{\beta}(z_{gj} - \hat{\mathbf{B}}'_3 \mathbf{x}_{gj}) \right].$$

To obtain the jackknife linearization variance estimator for weighted mean imputation, we substitute $z_{hk} = 1$ into (4.2).

In order to compare the components of $\hat{V}_{LJ}$ with that of $\hat{V}_{TOT}$, we decomposed $\hat{V}$ into terms that were similar to $\hat{V}_{SAM}$, $\hat{V}_{IMP}$ and $\hat{V}_{MIX}$ and denoted them as $\tilde{V}_{SAM}$, $\tilde{V}_{IMP}$ and $\tilde{V}_{MIX}$,

$$\hat{V}_{LJ} = \sum_g \frac{n_g}{n_g - 1} \sum_j (e_{gj} - \bar{e}_g)^2$$
$$= \sum_g \frac{n_g}{n_g - 1} \left(\sum_j (e_{gj,1} - \bar{e}_{g,1})^2 + \sum_j (e_{gj,2} - \bar{e}_{g,2})^2 + 2\sum_j (e_{gj,1} - \bar{e}_{g,1})(e_{gj,2} - \bar{e}_{g,2}) \right)$$
$$= \tilde{V}_{SAM} + \tilde{V}_{IMP} + 2\tilde{V}_{MIX}$$

where

$$e_{gj,1} = w^*_{gj} \frac{\hat{Z}}{\hat{T}} \left[(y_{gj} - \hat{\mathbf{B}}'_1 \mathbf{x}_{gj}) \delta_{gj} + (\hat{\beta} z_{gj} - \hat{\mathbf{B}}'_1 \mathbf{x}_{gj})(1 - \delta_{gj}) \right]$$

$$e_{gj,2} = w^*_{gj} \left[\frac{\hat{Z}}{\hat{T}} (\hat{\mathbf{B}}'_2 \mathbf{x}_{gj} - \hat{\beta} z_{gj}) + \hat{\beta}(z_{gj} - \hat{\mathbf{B}}'_3 \mathbf{x}_{gj}) \right].$$

5. EMPIRICAL STUDY AND COMPARISON

5.1 Details

A simulation study was conducted to investigate the finite-sample properties of both the jackknife linearization and the model based variance estimators. We used a fixed finite population of 15,425 units divided into 12 design strata (H=12). We also had 6 post-strata (4 age and 2 sex) that were used by the GREG estimator. The sample was calibrated to the marginal totals of the post-strata variables.

We drew R=10,000 independent simple random samples of size n=1543 (sampling fraction of 10%). An imputation variable z was used in the study and was generated from a $\Gamma(3,16)$ distribution. In addition, there were 3 y-variables generated. The first y-variable y_1 was generated from a $\Gamma(1.5z,4z)$ distribution. The variables y_2 and y_3 were generated as follows:

$$y_2 = z_1^2 + 5000 * r$$
$$y_3 = z_1^2 + 35000 * r$$

Where r is a randomly generated value from a $N(0,1)$ distribution. The purpose of having these 3 y-variables is to compare the performance of the variance estimators when the imputation model does and does not hold. The linear correlations between y_1, y_2, y_3 and z were .95, .56 and .10 respectively. We shall also look at these results for different non-response rates (10%, 20% and 30%). The response for each unit was generated according to a uniform response mechanism

To compare the performance of the jackknife linearization against the model based variance estimator we computed the empirical relative bias (ERB) for each estimator:

$$ERB = \left(\frac{1}{R} \frac{\sum_R \hat{V}_{TOT,r}}{MSE} - 1 \right) * 100\%$$

where $\hat{V}_{TOT,r}$ denotes the value of $\hat{V}_{TOT}$ for the r^th^ simulation and MSE denotes the empirical mean squared error of the estimator of Y,

$$MSE = \frac{1}{R} \sum_R (\hat{Y}_{s,r} - Y)^2$$

Table 1. ERB(%) For Variance Estimation Under Ratio Imputation.

Y Variable	NR Rate	Model Based Determin	Model Based Stochastic	Linear Jackknife
Y_1	0.5	-4	0	-1
	0.3	-3	0	-0.7
	0.1	-0.7	0	0
Y_2	0.5	-26	-5	-6
	0.3	-20	-4	-3
	0.1	-8	0.7	0
Y_3	0.5	-24	0	-5
	0.3	-18	2	-0.8
	0.1	-8	0.5	-0.5

where $\hat{Y}_{s,r}$ denotes the estimate of Y for the r^th^ simulation.

5.2 Comparing Total Variance

For the model based estimator we investigated the effect of using the stochastic term d_i as mentioned in section (3.1). In Table 1, we look at the ERB for the model based estimator using both deterministic imputation (without d_i term) and stochastic imputation (with d_i term).

Here NR Rate refers to non-response rate and 'Determin' and 'Stochastic' refers to deterministic and stochastic imputation respectively. Table 1 shows that, in general, both estimators underestimate the true MSE, however this underestimation is much more prominent when using deterministic imputation for the model based estimator. This highlights the importance of applying a correction term to the imputed value when using the model based estimator, however the effect of the correction term diminishes significantly when the variable of interest and imputation variable have a very strong correlation.

5.3 Comparing Separate Variance Components

The jackknife linearization performs very well without requiring a stochastic term. A possible explanation for this is the way the V_{SAM} term is calculated. For the jackknife linearization component $\tilde{V}_{SAM}$ and model based $\hat{V}_{SAM}$, we note that $\tilde{V}_{SAM} = (\hat{Z}/\hat{T})\hat{V}_{SAM}$ where $\hat{Z}/\hat{T} = \sum_s w_{hk}^* z_{hk} / \sum_s w_{hk}^* z_{hk} \delta_{hk}$ is a weighted "non-response adjustment" term for the imputation variable

Table 2. Estimates of Variance for V_{SAM}, V_{IMP} and V_{MIX} for variable Y_1 (in millions).

NR Rate	Estimator Type	Estimate of V_{SAM}	Estimate of V_{IMP}	Estimate of V_{MIX}
0.1	Model (D)	258	3	0
	Model (S)	261	3	0
	Linear Jack	319	4	-30
0.3	Model (D)	253	12	0
	Model (S)	261	12	0
	Linear Jack	520	45	-146
0.5	Model (D)	248	3	0
	Model (S)	261	29	0
	Linear Jack	1,002	243	-479

z. This appears to adjust the underestimation of V_{SAM} by $\hat{V}_{SAM}$ when a stochastic term is not used. This was the case in Lee, Rancourt, and Särndal (1995) where $\hat{V}_{SAM}$, as in (3.1), was defined as $\hat{V}_{ORD}$ with $\hat{V}_{SAM} = \hat{V}_{ORD} + \hat{V}_{DIF}$ and $\hat{V}_{DIF} = E_{\xi}(\hat{V}^{*}_{ORD} - \hat{V}_{ORD})$ with $\hat{V}^{*}_{ORD}$ being the same as $\hat{V}_{ORD}$ but replacing y_{*hk} with y_{hk}. Note, to calculate $\hat{V}_{DIF}$, the model expectation is needed since y_{hk} is not defined for non-respondents.

When comparing the separate components of variance for the jackknife linearization and the model based estimators we obtain the following results where Model (D) and Model (S) refer to deterministic and stochastic imputation respectively. As already mentioned the estimators for V_{SAM} were different thus we expected differing results for the other components also.

6. CONCLUSIONS

The model based approach, with the stochastic term, performed extremely well when the imputation model was true and performed only slightly poorer when the model was weaker. However the model based approach without the stochastic term performed poorly especially when the model didn't hold. Turning to the jackknife linearization, it performed well for all the models except when the non-response rate was high. Due to the differences in the two approaches, comparison of the components is difficult and requires further investigation to be fully understood.

ACKNOWLEGMENTS

The authors wish to thank to Eric Rancourt, Jean-François Beaumont and Mike Hidiroglou for the many interesting conversations and insight in this area of work.

REFERENCES

Gagnon, F., Lee, H., Rancourt, E. and Särndal, C.-E. (1996), "Estimating the Variance of the Generalized Regression Estimator in the Presence of Imputation for the Generalized Estimation System" *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 151-156.

Lee, H., Rancourt, E. and Särndal, C.-E. (1995), "Variance Estimation in Presence of Imputation for the Generalized Estimation System" *Proceedings of the Section Research Methods, American Statistical Association*, 384-389.

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.

Särndal, C.-E. (1992), "Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used" *Survey Methodology*, 18, 241-252.

Shao, J and Sitter, R.R (1996), "Bootstrap for Imputed Data," *Journal of the American Statistical Association*, 91, 1278-1288.

Shao, J and Steel, P (1999), "Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions," *Journal of the American Statistical Association*, 94, 254-265.

Yung, W. and Rao, J.N.K. (1996), "Jackknife Linearization Estimators Under Stratified Multi-Stage Sampling" *Survey Methodology*, 22 23-31.

Yung, W. and Rao, J.N.K. (2000), "Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information," *Journal of the American Statistical Association*, 95, 903-915.