Proceedings of Statistics Canada Symposium 2002
Modeling Survey Data for Social and Economic Research

THE SYSTEM FOR ESTIMATION OF VARIANCE DUE TO NONRESPONSE AND IMPUTATION (SEVANI)

Jean-François Beaumont and Charles Mitchell[1]

ABSTRACT

The System for Estimation of Variance due to Nonresponse and Imputation (SEVANI) is a SAS-based prototype system with a graphical user interface. Variance estimation is based on the quasi-multi-phase framework. In this framework, a nonresponse model is required and an imputation model can also be used. Two types of nonresponse treatment methods can be dealt with: nonresponse weighting adjustment and imputation. If imputation is chosen, SEVANI requires that one of the following four imputation methods be used: deterministic linear regression imputation, random linear regression imputation, auxiliary value imputation or nearest-neighbour imputation. In this paper, the methodology that SEVANI is based on will be described and an example using real data will be presented to illustrate the theory in practice.

KEY WORDS: Quasi-randomization; Two-phase sampling; Nonresponse model; Imputation model.

1. INTRODUCTION

Most surveys, if not all, have to deal with the problem of nonresponse. Different reasons might explain the presence of nonresponse, such as a refusal to provide the desired information for at least one question or an impossibility to contact a given unit. Nonresponse can also be produced at the editing step of the survey in an attempt to resolve problems of inconsistent or suspect responses. Therefore, nonresponse takes here its general meaning of missing values in the sample.

Nonresponse leads inevitably to an observed sample of smaller size than the sample originally selected. This sample size reduction is most of the time accompanied by an increase in the variance of the estimates, no matter which method is chosen to treat nonresponse. This increase in variance is called the nonresponse variance. The imputation variance is defined here as a component of the nonresponse variance which is due to the use of a random imputation method.

At the beginning of the 80's, most research efforts on nonresponse were focused on the evaluation and the reduction of nonresponse bias. Since that time, a lot of research work has been devoted to estimating the nonresponse variance, especially when imputation is used to replace the missing values (see Lee, Rancourt and Särndal, 2001 for a review). The result of some of this research activity lead to the development of the System for Imputation Variance (SIMPVAR), which was presented at the Statistics Canada Symposium 1997 (Rancourt, Gagnon, Lee, Provost and Särndal 1997). More recently, work has been done to allow more nonresponse treatment methods to be handled in a unified framework for variance estimation, called the quasi-multi-phase (or quasi-randomization in the terminology of Oh and Scheuren, 1983) framework. Särndal and Swensson (1987), Särndal (1992), Rao and Sitter (1995) and Beaumont (2000) are examples of papers that are based on the quasi-two-phase framework. As a result of the substantial modifications in SIMPVAR, the name was changed to the System for Estimation of Variance due to Nonresponse and Imputation (SEVANI). SEVANI is a prototype system developed using SAS v8. Therefore, SEVANI only supports SAS v8 files and libraries. It is a collection of macros that can be run separately or by using the graphical user interface.

Version 1.0 of SEVANI requires that the population parameter to be estimated be a domain total or a domain mean and that the full response estimator be the Horvitz-Thompson estimator or in the family of calibration estimators. Two types of nonresponse treatment methods can be dealt with: nonresponse weighting adjustment and imputation. If imputation is

[1] Statistics Canada, R.H. Coats building, 16th floor, Tunney's Pasture, Canada, K1A 0T6
(jean-francois.beaumont@statcan.ca and charles.mitchell@statcan.ca)

chosen, SEVANI requires that one of the following four imputation methods be used: deterministic linear regression imputation, random linear regression imputation, auxiliary value imputation or nearest-neighbour imputation.

2. WHY SHOULD THE NONRESPONSE VARIANCE BE ESTIMATED?

As noted by Rancourt and al. (1997), there are several reasons for estimating the variance of an estimator whether there is nonresponse or not. When there is nonresponse, the following four main reasons of estimating the nonresponse variance can be emphasized:

i) To obtain valid inferences in the presence of nonresponse;
ii) To measure properly the quality of estimates produced by the survey and to inform users of the data quality;
iii) To better allocate survey resources;
iv) To compare different nonresponse treatment methods and make better decisions.

The third reason means that if the nonresponse variance is large compared to the sampling variance in a given stratum then it might be desirable to put more resources on preventing nonresponse (for example, more follow-ups of nonrespondents) for that stratum. To achieve this objective for a given survey cost, the desired sample size might have to be reduced. This will lead to an increase in the sampling variance but a larger reduction of the nonresponse variance might be anticipated and, thus, the total variance should decrease.

A good modeling effort is always required to minimize the nonresponse bias as much as possible and to find a nonresponse treatment method. If one model is better than all other models, then there is no need to estimate the nonresponse variance in order to choose a method. However, if there are competing models, estimating the nonresponse variance can be used as a criterion to make a decision on the nonresponse treatment method to be chosen as pointed out in the fourth reason above.

3. THE FULL RESPONSE CASE

Let us assume that it is desired to estimate a domain total or a domain mean for a given population *U*, and let us denote this unknown population parameter by θ. In the case of a domain total, $\theta = \sum_{k \in U} d_k y_k$, where y_k is the value of the variable of interest *y* for population unit *k* and d_k is an indicator variable equaling 1 if unit *k* belongs to the domain of interest and 0 otherwise. In the case of a domain mean, $\theta = \sum_{k \in U} d_k y_k / \sum_{k \in U} d_k$.

Since it is usually not possible to survey all population units, a random sample *S* is selected according to some sampling design. The selection probability for unit *k* is denoted by π_k and the probability that a given sample *s* be selected is denoted by $p(s)$. Expectations under the sampling design are denoted by $\mathrm{E}_p(\cdot)$. An estimator $\hat{\theta}$ of θ is usually obtained by computing estimation weights $\tilde{w}_k$, for $k \in S$. In the case of a domain total, $\hat{\theta} = \sum_{k \in S} \tilde{w}_k d_k y_k$, and in the case of a domain mean, $\hat{\theta} = \sum_{k \in S} \tilde{w}_k d_k y_k / \sum_{k \in S} \tilde{w}_k d_k$. If no auxiliary information is available at the population level then the usual Horvitz-Thompson (HT) estimator of θ is obtained by taking $\tilde{w}_k = 1/\pi_k$.

It is not unusual that a vector of auxiliary variables $\mathbf{x_1}$ be available for all sample units and that population totals, $\mathbf{t}_{\mathbf{x}_1} = \sum_{k \in U} \mathbf{x_{1k}}$, be known for these variables. In that case, it is possible to use this auxiliary information in the form of a calibration estimator to improve the usual HT estimator of θ. For example, the Generalized Regression (GREG) estimator can be used with the estimation weights $\tilde{w}_k = w_k g_k$, where

$$g_k = 1 + \frac{\mathbf{x}'_{\mathbf{1k}}}{c_{1k}} \left(\sum_{k \in S} \frac{w_k}{c_{1k}} \mathbf{x}_{\mathbf{1k}} \mathbf{x}'_{\mathbf{1k}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}_1} - \sum_{k \in S} w_k \mathbf{x}_{\mathbf{1k}} \right),$$

$w_k = 1/\pi_k$ and c_{1k} is a known positive function of $\mathbf{x}_{\mathbf{1k}}$ that corresponds to the variance structure of the estimation model underlying the GREG estimator. The GREG estimator is asymptotically *p*-unbiased, that is, $\mathrm{E}_p(\hat{\theta}) \approx \theta$.

The sampling variance of $\hat{\theta}$ is denoted by $\mathrm{V}_{\mathrm{sam}} = \mathrm{V}_p(\hat{\theta})$. It is assumed that some variance estimator exists for $\mathrm{V}_{\mathrm{sam}}$, denoted by $\hat{\mathrm{V}}_{\mathrm{sam}} = \hat{\mathrm{V}}_p(\hat{\theta})$. For example, $\hat{\mathrm{V}}_{\mathrm{sam}}$ could be obtained using techniques such as Taylor linearization, bootstrap or jackknife.

4. THE QUASI-TWO-PHASE FRAMEWORK

In the quasi-two-phase framework, nonresponse is viewed as a second phase of selection and the variable *y* is only observed for part of the sample *s*. The random sample of respondents is denoted by R_1 (the subscript 1 indicates that it is the first phase of nonresponse). The probability that sample unit *k* respond is denoted by $p_{1k} = P(k \in R_1 \mid s)$. The probability that two different sample units *k* and *l* respond is denoted by $p_{1kl} = P(k \in R_1, l \in R_1 \mid s)$. The probability that a given sample of respondents r_1 be observed is denoted by $q_1(r_1 \mid s)$ and expectations under the nonresponse mechanism are denoted by $\mathrm{E}_{q_1}(\cdot \mid s)$. Expectations under the sampling design and the nonresponse mechanism are denoted by $\mathrm{E}_{pq_1}(\cdot) = \mathrm{E}_p \mathrm{E}_{q_1}(\cdot \mid s)$.

In the presence of nonresponse, the GREG or HT estimators $\hat{\theta}$ cannot be calculated since y_k is not observed for $k \in s - r_1$. Nonresponse weighting adjustment or imputation is usually performed to obtain an adjusted estimator $\hat{\theta}^*$. It is assumed that the adjusted estimator is asymptotically q_1-unbiased conditional on the realized sample *s*, that is, $\mathrm{E}_{q_1}(\hat{\theta}^* \mid s) \approx \hat{\theta}$. Consequently, $\hat{\theta}^*$ is approximately unconditionally unbiased ($\mathrm{E}_{pq_1}(\hat{\theta}^*) = \mathrm{E}_p \mathrm{E}_{q_1}(\hat{\theta}^* \mid s) \approx \theta$) and its variance can be approximated by

$$\begin{aligned} \mathrm{V}_{pq_1}(\hat{\theta}^*) &= \mathrm{V}_p \mathrm{E}_{q_1}(\hat{\theta}^* \mid s) + \mathrm{E}_p \mathrm{V}_{q_1}(\hat{\theta}^* \mid s) \\ &\approx \mathrm{V}_p(\hat{\theta}) + \mathrm{E}_p \mathrm{V}_{q_1}(\hat{\theta}^* \mid s) \\ &= \mathrm{V}_{\mathrm{sam}} + \mathrm{V}_{\mathrm{nr1}}, \end{aligned} \tag{4.1}$$

where $\mathrm{V}_{\mathrm{nr1}} = \mathrm{E}_p \mathrm{V}_{q_1}(\hat{\theta}^* \mid s)$ is the first-phase nonresponse variance. Therefore, an approximately unbiased variance estimator for $\mathrm{V}_{\mathrm{nr1}}$ can simply be obtained by finding an approximately unbiased estimator for $\mathrm{V}_{q_1}(\hat{\theta}^* \mid s)$. Since the nonresponse mechanism is unknown in practice, a nonresponse model is needed to approximate it. Therefore, the variance in (4.1) is valid only if the postulated nonresponse model is a good substitute for the true unknown nonresponse mechanism. This is the reason why the framework is termed quasi-two-phase.

When imputation is used to treat nonresponse, more effort is usually devoted to finding a good imputation model (model for the variable *y*, denoted by *m*) than to finding a good nonresponse model. In this case, it might be preferable not to rely too heavily on the nonresponse model and to use the imputation model to obtain a variance estimator. One such approach is described in Särndal (1992). Instead of estimating $\mathrm{V}_{pq_1}(\hat{\theta}^*)$, he proposed to estimate the model expectation of $\mathrm{V}_{pq_1}(\hat{\theta}^*)$, which is given by

$$\begin{aligned} \mathrm{E}_m \mathrm{V}_{pq_1}(\hat{\theta}^*) &\approx \mathrm{E}_m \mathrm{V}_p(\hat{\theta}) + \mathrm{E}_m \mathrm{E}_p \mathrm{V}_{q_1}(\hat{\theta}^* \mid s) \\ &= \mathrm{V}^{\mathrm{m}}_{\mathrm{sam}} + \mathrm{V}^{\mathrm{m}}_{\mathrm{nr1}}, \end{aligned} \tag{4.2}$$

where, here, the sampling variance is $\mathrm{V}^{\mathrm{m}}_{\mathrm{sam}} = \mathrm{E}_m \mathrm{V}_{\mathrm{sam}}$ and the first-phase nonresponse variance is $\mathrm{V}^{\mathrm{m}}_{\mathrm{nr1}} = \mathrm{E}_m \mathrm{V}_{\mathrm{nr1}}$. Under this approach, it is generally assumed that the sampling design and the nonresponse mechanism are ignorable with respect to the imputation model *m*, as defined in Rubin (1976). As a practical matter, this means that all the relevant design information and the information related to the nonresponse mechanism have been included in the imputation model. This assumption makes variance estimation much easier. In particular, provided that $\mathrm{E}_{q_1}(\hat{\theta}^* \mid s) \approx \hat{\theta}$, the first-phase nonresponse variance can be written as

$$\mathrm{V}^{\mathrm{m}}_{\mathrm{nr1}} = \mathrm{E}_p \mathrm{E}_{q_1} \mathrm{E}_m \left\{ (\hat{\theta}^* - \hat{\theta})^2 \right\}.$$

As a result, an approximately unbiased variance estimator for $\mathrm{V}^{\mathrm{m}}_{\mathrm{nr1}}$ can simply be obtained by finding an approximately unbiased estimator for $\mathrm{E}_m \left\{ (\hat{\theta}^* - \hat{\theta})^2 \right\}$, which does not require modeling the unknown nonresponse mechanism, apart from assuming that it is ignorable. This nonresponse model (assumption that the nonresponse mechanism is ignorable) is much weaker than the nonresponse model specification usually required for estimating (4.1). However, this approach requires that the imputation model *m* be valid. The choice between estimating (4.1) or (4.2) should therefore depend on the confidence the survey methodologist has in the nonresponse model or in the imputation model. If the nonresponse model is thought to be of better quality than the imputation model then estimating (4.1) should be preferred over estimating (4.2) and vice-versa.

5. MODELS USED IN SEVANI

As a consequence of the unknown nonresponse mechanism, response probabilities p_{1k}, for $k \in s$, and joint response probabilities p_{1kl}, for $k, l \in s$, are also unknown. In order to estimate the variance in (4.1), these probabilities are generally estimated using a nonresponse model. This is the nonresponse model approach. A commonly used nonresponse model is the uniform-within-class nonresponse model, where all units within a given class are assumed to respond independently of one another with the same response probability.

In practice, joint response probabilities p_{1kl} are never estimated directly. Therefore, it is necessary to assume some form of independence to get rid of joint response probability estimation and to simplify estimation of variance (4.1). In SEVANI, it is assumed that clusters of units *c*, for $c \in s'$, respond independently, where s' is the selected sample of clusters. That is, all sample units within a cluster *c* are observed simultaneously or they are all missing simultaneously. The response probability for all units $k \in s_c$ is denoted by p'_{1c}, where s_c is the sample of units in cluster *c*. The joint response probability of a unit *k* in cluster *c* and a unit *l* in a different cluster c' is equal to $p'_{1c} p'_{1c'}$, while the joint response probability of any two units in the same cluster *c* is equal to p'_{1c}. A typical example where clusters of units can be assumed to respond independently occurs when dwellings (clusters) are selected but the desired information is collected for all people in the selected dwellings. Note that a cluster could also contain only one sample unit and, thus, correspond exactly to a unit of analysis *k*.

Provided that the selected nonresponse model is satisfactory, the estimated response probabilities should be close to the true unknown response probabilities. In SEVANI, the estimated response probabilities are assumed exactly equal to the true response probabilities p_{1k} (or p'_{1c}). This should lead to an underestimation of the nonresponse variance $\mathrm{V}_{\mathrm{nr1}}$ when $\hat{\theta}^*$ depends on these response probabilities, as it is the case when a nonresponse weighting adjustment is performed or, sometimes, when imputation is used. However, some simulation studies (for example, Mantel, Nadon and Yeo, 2000) have shown that the underestimation is often negligible.

In order to find an estimator of the variance (4.2), SEVANI assumes that the sampling design and the nonresponse mechanism are ignorable with respect to the imputation model. Also, SEVANI uses the imputation model *m* such that $\mathrm{E}_m(y_k \mid \mathbf{x_k}) = \boldsymbol{\beta}'\mathbf{x_k}$, $\mathrm{V}_m(y_k \mid \mathbf{x_k})$ is proportional to c_k and observations are independent of one another. The vector $\mathbf{x_k}$ contains auxiliary variables, available for all $k \in s$, used to obtain the imputed estimator $\hat{\theta}^*$, $\boldsymbol{\beta}$ is a vector of unknown model parameters and c_k is a known positive function of $\mathbf{x_k}$ that corresponds to the variance structure of the imputation model *m*. This is the imputation model approach. Sometimes, imputation is performed independently within imputation classes. SEVANI deals with this situation by fitting a different model for each class. It is also possible to deal with imputation classes by including dummy variables into the vector $\mathbf{x_k}$, which indicate to which class the unit *k* belongs.

A good modeling effort is always necessary to reduce the nonresponse bias and the nonresponse variance as much as possible, no matter which nonresponse treatment method is chosen. As part of the modeling effort, the appropriate selection of auxiliary variables, that are used to obtain the adjusted estimator $\hat{\theta}^*$, is the key to reduce the impact of nonresponse.

6. NONRESPONSE TREATMENT METHODS

6.1 Nonresponse Weighting Adjustment

A common remedy to unit nonresponse consists of performing a nonresponse weighting adjustment. When auxiliary variables $\mathbf{x_1}$ are available, the resulting adjusted estimator $\hat{\theta}^*$ of θ is obtained by viewing unit nonresponse as a second phase of selection and by using the GREG estimator under two-phase sampling. In the case of a domain total, $\hat{\theta}^* = \sum_{k\in R_1} \tilde{w}_k^* d_k y_k$, and in the case of a domain mean, $\hat{\theta}^* = \sum_{k\in R_1} \tilde{w}_k^* d_k y_k \big/ \sum_{k\in R_1} \tilde{w}_k^* d_k$, where $\tilde{w}_k^* = w_k^* g_k^*$ is the estimation weight of unit *k*,

$$g_k^* = 1 + \frac{\mathbf{x}'_{\mathbf{1k}}}{c_{1k}} \left(\sum_{k\in R_1} \frac{w_k^*}{c_{1k}} \mathbf{x_{1k}} \mathbf{x}'_{\mathbf{1k}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}_1} - \sum_{k\in R_1} w_k^* \mathbf{x_{1k}} \right) \tag{6.1}$$

and $w_k^* = w_k / p_{1k}$ is the nonresponse adjusted weight of unit *k*. When no auxiliary variable is available, the adjusted estimator $\hat{\theta}^*$ has the same form, with g_k^* in (6.1) replaced by $g_k^* = 1$.

6.2 Imputation

Imputation is generally used to compensate for item nonresponse and sometimes used to deal with unit nonresponse. Imputation consists in replacing the missing value y_k of a nonresponding unit $k \in S - R_1$ by an imputed value y_k^*. The adjusted estimator $\hat{\theta}^*$ of θ is $\hat{\theta}^* = \sum_{k\in S} \tilde{w}_k d_k y_{\bullet k}$ in the case of a domain total and $\hat{\theta}^* = \sum_{k\in S} \tilde{w}_k d_k y_{\bullet k} / \sum_{k\in S} \tilde{w}_k d_k$ in the case of a domain mean, where $y_{\bullet k} = R_{1k} y_k + (1 - R_{1k}) y_k^*$ and R_{1k} is a random variable indicating whether unit *k* responded ($R_{1k} = 1$) or not ($R_{1k} = 0$). Each imputation method leads to a different way of obtaining imputed values y_k^* and a different adjusted estimator $\hat{\theta}^*$.

With Deterministic Linear Regression (DLR) imputation, the imputed value y_k^* is equal to the predicted value $\hat{\boldsymbol{\beta}}'\mathbf{x_k}$ obtained from the imputation model *m*, where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in R_1} \frac{a_k}{c_k} \mathbf{x_k} \mathbf{x_k'} \right)^{-1} \sum_{k \in R_1} \frac{a_k}{c_k} \mathbf{x_k} y_k$$

is the estimator of $\boldsymbol{\beta}$ and a_k is a regression weight. DLR imputation includes many special cases such as, for example, mean imputation or ratio imputation. When there are imputation classes, $\hat{\boldsymbol{\beta}}$ is estimated separately for each class.

To each DLR imputation method, there is a corresponding Random Linear Regression (RLR) imputation method. With RLR imputation, the imputed value y_k^* is equal to $\hat{\boldsymbol{\beta}}' \mathbf{x_k} + e_k$, where e_k is a random component. That is, $e_k = (y_l - \hat{\boldsymbol{\beta}}' \mathbf{x_l}) \sqrt{c_k / c_l}$ and unit *l* is a respondent ($l \in r_1$), in the same imputation class as unit *k*, selected randomly with replacement and with probability proportional to a_k. An example of RLR imputation is random hot-deck imputation, which is the random version of mean imputation. With random hot-deck imputation, a missing value is replaced by the value of a randomly selected respondent in the same imputation class as the nonrespondent.

With Auxiliary Value (AV) imputation, the missing value y_k of a nonresponding unit $k \in s - r_1$ is replaced by an auxiliary value z_k. The imputed value $y_k^* = z_k$ is obtained by using only data that come from the nonresponding unit *k*, such as historical data. Carry-forward imputation is an example of AV imputation, where the auxiliary value z_k is the value of the variable *y* observed at a previous period of the survey. This imputation method can be justified by the imputation model *m* with $\boldsymbol{\beta}' \mathbf{x_k}$ being known. Consequently, there is no unknown vector of parameters to be estimated, as opposed to regression imputation methods.

Finally, with Nearest-Neighbour (NN) imputation, the imputed value y_k^* is equal to y_l, where unit *l* is the closest respondent of unit *k* with respect to the auxiliary variables $\mathbf{x}$ and *l* is in the same imputation class as *k*. To determine the closest respondent of unit *k*, a distance measure is required. In SEVANI, it is assumed that the L_p norm has been used. It is also assumed that the auxiliary variables are numerical and have been standardized. This is done either by subtracting the mean and dividing by the standard deviation each auxiliary variable or by replacing the values of each auxiliary variable by their ranks.

7. VARIANCE ESTIMATION

7.1 Sampling variance estimation

In the presence of nonresponse, usual sampling variance estimators $\hat{\mathrm{V}}_{\mathrm{sam}}$ of $\mathrm{V}_{\mathrm{sam}} = \mathrm{V}_p(\hat{\theta})$ cannot be computed since y_k is not observed for $k \in s - r_1$. However, it is possible to estimate $\mathrm{V}_p(\hat{\theta} \mid r_1)$ by $\hat{\mathrm{V}}_p(\hat{\theta} \mid r_1)$ using standard variance estimation techniques. This can be achieved by treating R_{1k}, for $k \in U$, as being fixed and using existing variance estimation systems valid in the full response case. An approximately unbiased variance estimator $\hat{\mathrm{V}}_{\mathrm{sam}}^* = \hat{\mathrm{V}}_p(\hat{\theta} \mid r_1) + \hat{\mathrm{V}}_{\mathrm{cor}}$ of $\mathrm{V}_{\mathrm{sam}}$ can then be obtained, where $\hat{\mathrm{V}}_{\mathrm{cor}}$ is constructed such that $\mathrm{E}_q(\hat{\mathrm{V}}_{\mathrm{cor}} \mid s) \approx \hat{\mathrm{V}}_{\mathrm{sam}} - \mathrm{E}_q\left\{\hat{\mathrm{V}}_p(\hat{\theta} \mid r_1) \mid s\right\}$. In version 1.0, SEVANI does not estimate the sampling variance. However, in the case of nonresponse weighting adjustment, SEVANI computes $\hat{\mathrm{V}}_{\mathrm{cor}}$, if requested, using the Taylor linearization technique.

When imputation is used, estimating $\mathrm{V}_{\mathrm{sam}}^{\mathrm{m}} = \mathrm{E}_m \mathrm{V}_{\mathrm{sam}}$ instead of $\mathrm{V}_{\mathrm{sam}}$ might be more appropriate, as noted in section 4. In fact, what is really desired is to find an approximately model unbiased predictor for $\hat{\mathrm{V}}_{\mathrm{sam}}$, that is, a predictor $\hat{\mathrm{V}}_{\mathrm{sam}}^*$ such that $\mathrm{E}_m\left\{\left(\hat{\mathrm{V}}_{\mathrm{sam}}^* - \hat{\mathrm{V}}_{\mathrm{sam}}\right) \mid s, r_1\right\} \approx 0$. It is easy to show that such a predictor $\hat{\mathrm{V}}_{\mathrm{sam}}^*$ is *mpq*-unbiased for $\mathrm{V}_{\mathrm{sam}}^{\mathrm{m}}$.

One way to obtain an approximately model unbiased predictor for $\hat{V}_{sam}$ consists of imputing the missing values by drawing values from the estimated distribution of $\{y_k : k \in s - r_1\}$ given $\{y_k : k \in r_1\}$ before using a standard variance estimator valid in the full response case. Essentially, this means that a random component is added to the imputed values in the case of DLR and AV imputation before using a standard variance estimation system. In the case of RLR and NN imputation, imputed values do not need to be modified. Note that these modified imputed values are only used for variance estimation and they have no effect on the adjusted estimator $\hat{\theta}^*$. Note also that this method of predicting $\hat{V}_{sam}$ has not been implemented in the current version of SEVANI.

To obtain a more efficient estimator, more than one set of imputed values could be drawn. The final sampling variance estimate would be obtained by averaging the sampling variance estimates associated to each set of imputed values. This looks like multiple imputation, except that, here, imputation does not need to be proper. In practice, it may be more convenient to have only one set of imputed values. This should lead to reasonable sampling variance estimates unless the nonresponse rate is very high.

7.2 Nonresponse variance estimation

As noted in section 4, when the nonresponse model is thought to be of better quality than the imputation model, using a nonresponse model approach (estimating 4.1) is more natural than using an imputation model approach (estimating 4.2). This is the case when a nonresponse weighting adjustment is performed since no effort is devoted to finding an imputation model. An approximately unbiased variance estimator for V_{nr1} can simply be obtained by finding an approximately unbiased estimator for $V_{q_1}(\hat{\theta}^* \mid s)$. In SEVANI, $V_{q_1}(\hat{\theta}^* \mid s)$ is estimated using the Taylor linearization technique and the assumption that clusters of units respond independently.

When imputation is used, estimating V_{nr1}^m might be more appropriate than estimating V_{nr1}. In SEVANI, estimating V_{nr1}^m is achieved by estimating $E_m\left\{(\hat{\theta}^* - \hat{\theta})^2\right\}$ and by assuming that the sampling design and the nonresponse mechanism are ignorable with respect to the imputation model *m*. Since $\hat{\theta}^*$ and $\hat{\theta}$ are linear estimators, a closed-form expression for $E_m\left\{(\hat{\theta}^* - \hat{\theta})^2\right\}$ is easily obtained.

Whether it is chosen to estimate V_{nr1} or V_{nr1}^m, an additional component of variance must be computed when RLR imputation is used. This additional component of variance takes into account the variability that is due to the random imputation process. With RLR imputation, the nonresponse variance is thus equal to the nonresponse variance of the corresponding DLR imputation method plus the additional random imputation variance. It easy to obtain a closed-form expression for the random imputation variance and it does not depend on the approach chosen (nonresponse model approach or imputation model approach).

NN imputation is a nonparametric imputation method since it does not require specifying the imputation model to justify the form of the adjusted estimator $\hat{\theta}^*$ and to determine its model unbiasedness property. Therefore, using a linear regression imputation model to obtain a variance estimator might not always be appropriate, especially when the linear model does not hold satisfactorily. If a nonresponse model approach is chosen, the adjusted estimator $\hat{\theta}^*$ is not smooth and complications arise when the Taylor linearization technique is used.

To cope with problems associated to the use of NN imputation, NN imputation can be viewed as a random hot-deck imputation method, where the number of imputation classes is equal to the number of distinct values of the variable *y*. This corresponds to an imputation model with no degree of freedom. Therefore, it is not possible to estimate the variability within an imputation class since, by definition of NN imputation, there is no variability. SEVANI deals with this problem by approximating the adjusted estimator using NN imputation by the adjusted estimator using random hot-deck imputation within imputation classes, where the classes are small but contain at least two respondents. To form classes, SEVANI uses the "*k*-means" clustering algorithm implemented in the FASTCLUST procedure of SAS. SEVANI uses the same standardization method and the same distance measure as those used to perform NN imputation.

8. THE QUASI-MULTI-PHASE FRAMEWORK

The quasi-two-phase framework can be easily extended to the quasi-multi-phase framework. In the latter framework, it is possible to estimate the nonresponse variance associated to more than one nonresponse mechanism or, in other words, more than one cause of nonresponse. For example, most surveys suffer from unit and item nonresponse and these two types of nonresponse are likely to be explained by different nonresponse mechanisms. Moreover, they are often not treated in the same way. Unit nonresponse is usually treated by a nonresponse weighting adjustment technique while item nonresponse is usually treated by an imputation technique.

Version 1.0 of SEVANI can be used to estimate up to three components of nonresponse variance; each of them associated to a different nonresponse mechanism. Each additional nonresponse mechanism is viewed as an additional phase of selection (or phase of nonresponse). Therefore, when there is more than one nonresponse mechanism, the approach is termed quasi-multi-phase instead of quasi-two-phase. Although different methods can be used at different phases of nonresponse, SEVANI restricts nonresponse weighting adjustment to the first phase of nonresponse only.

In the case where there are two phases of nonresponse, the variable *y* is only observed for part of r_1. The sample of respondents at the second phase is denoted by r_2. The probability that a given sample of respondents r_2 be observed is denoted by $q_2(r_2 \mid s, r_1)$ and expectations under the second phase nonresponse mechanism are denoted by $\mathrm{E}_{q_2}(\cdot \mid s, r_1)$. Expectations under the sampling design and both nonresponse mechanisms are denoted by $\mathrm{E}_{pq_1q_2}(\cdot) = \mathrm{E}_{pq_1}\mathrm{E}_{q_2}(\cdot \mid s, r_1)$.

When there are two phases of nonresponse, the adjusted estimator $\hat{\theta}^*$ cannot be calculated since y_k is not observed for $k \in r_1 - r_2$. Nonresponse weighting adjustment or imputation is usually performed to obtain a double adjusted estimator $\hat{\theta}^{**}$. It is assumed that $\hat{\theta}^{**}$ is approximately q_2-unbiased conditional on the observed samples *s* and r_1, that is, $\mathrm{E}_{q_2}(\hat{\theta}^{**} \mid s, r_1) \approx \hat{\theta}^*$. Consequently, $\hat{\theta}^{**}$ is approximately unconditionally unbiased and its variance can be approximated by

$$\mathrm{V}_{pq_1q_2}(\hat{\theta}^{**}) \approx \mathrm{V}_{pq_1}(\hat{\theta}^*) + \mathrm{E}_{pq_1}\mathrm{V}_{q_2}(\hat{\theta}^{**} \mid s, r_1),$$

where $\mathrm{V}_{\mathrm{nr2}} = \mathrm{E}_{pq_1}\mathrm{V}_{q_2}(\hat{\theta}^* \mid s, r_1)$ is the second-phase nonresponse variance. If an imputation model approach is chosen then $\mathrm{E}_m\mathrm{V}_{pq_1q_2}(\hat{\theta}^{**})$ is estimated. Estimating the second-phase nonresponse variance is very similar to estimating the first-phase nonresponse variance and the same techniques can be used. Estimation of $\mathrm{V}_{pq_1}(\hat{\theta}^*)$ has already been discussed in the previous sections. Note that SEVANI adds a random component to the imputed values when AV or DLR imputation is used at the second phase. This is to make sure that the first-phase nonresponse variance is properly estimated. This is also necessary for estimating the sampling variance. The extension to a third phase of nonresponse is straightforward.

9. AN EXAMPLE

This example shows how SEVANI can be used to compare different nonresponse treatment strategies using Labour Force Survey (LFS) data. In the LFS, we are concerned with determining whether carry-forward imputation could be replaced with a longitudinal random hot-deck method. One way to measure which method performs better consists of estimating the nonresponse variance for each method. In this example, the population parameter of interest is the total number of employed individuals in the population.

The LFS is a monthly survey whereby 5/6 of the sample is common from month to month. Therefore when there is nonresponse and historical information exists, the current practice is to transfer this information to the current month. This practice is known as carry-forward imputation.

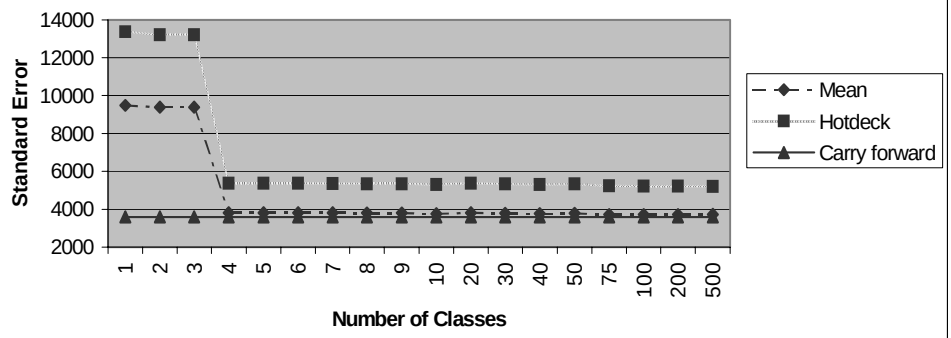
As mentioned in section 5, random hot-deck imputation consists of replacing a missing value by the value of a randomly selected respondent in the same imputation class as the nonrespondent. In longitudinal random hot-deck imputation, classes are constructed using historical auxiliary information. Longitudinal random hot-deck imputation is a random version of longitudinal mean imputation, where a missing value in a given class is replaced by the respondent mean within that class.

The construction of imputation classes is based on the score methodology (for example, Haziza, Charbonnier, Chow and Beaumont, 2001). First, two logistic regression models were used: one model for the probability that an individual is employed and one model for the probability that an individual is unemployed. Stepwise regression was then used in an attempt to find the best historical auxiliary variables for each model. A full year of LFS data was used in the process of choosing the final models. In the end, the labour force status of the previous month was the most important variable included in both models.

Once the models were obtained, historical data was used to get the estimated probability of being employed and unemployed for all units that responded in the previous month. A clustering algorithm (PROC FASTCLUS in SAS) was then used to form classes homogeneous with respect to the two probabilities. Hot-deck and mean imputation was performed within each class.

As seen in figure 1.0, the nonresponse variance is always lowest when carry-forward imputation is used. However, with a larger number of classes, longitudinal hot-deck and longitudinal mean imputation almost give the same results as carry-forward imputation.

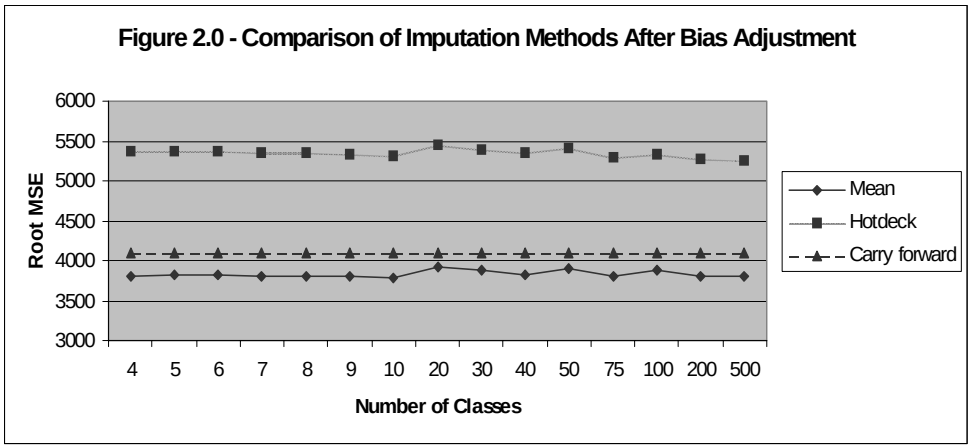
Figure 1.0 - Comparison of Imputation Methods



Note that SEVANI only estimates the nonresponse variance and not the nonresponse bias. Since carry-forward imputation is likely to have a higher nonresponse bias than the longitudinal mean or hot-deck methods, we tried to estimate the mean square error (MSE) instead of the variance. In an attempt to estimate the bias, mean imputation within classes was performed using a very large number of classes (1500). The point estimate was obtained and this was used as the benchmark for estimating the bias and the MSE.

As seen in Figure 2.0, in terms of the MSE, carry-forward imputation is no longer the best method when the number of classes exceeds 4. Longitudional mean imputation now outperforms all three methods. Longitudinal hot-deck imputation still has a higher MSE than carry-forward imputation, but the difference in values is not that large. In fact, perhaps the higher bias involved with carry-forward imputation would make the longitudinal hot-deck method a more attractive alternative.

As has been seen by the above example, SEVANI can be used as part of the process in evaluating nonresponse treatment strategies. In this case, SEVANI showed that carry-forward imputation was performing the best in terms of the nonresponse variance. However, other factors, such as nonresponse bias, also need to be considered.



REFERENCES

Beaumont, J.-F. (2000), "On Regression Imputation in the Presence of Nonignorable Nonresponse", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 580-585.

Haziza, D., Charbonnier, C., Chow, O.S.Y., and Beaumont, J.-F. (2001), "Construction of Imputation Cells for the Canadian Labour Force Survey", *Proceedings of the Statistics Canada Symposium* 2001, *Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.

Lee, H., Rancourt, E., and Särndal, C.-E. (2001), "Variance Estimation from Survey Data under Single Imputation", in Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds), *Survey Nonresponse*, New-York: John Wiley & Sons, Inc., pp. 315-328.

Mantel, H.J., Nadon, S., and Yeo, D. (2000), "Effect of Nonresponse Adjustments on Variance Estimates for the National Population Health Survey", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 221-226.

Oh, H.L., and Scheuren, F.J. (1983), "Weighting Adjustment for Unit Nonresponse", in W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2, New-York: Academic Press, pp. 143-184.

Rancourt, E., Gagnon, F., Lee, H., Provost, M., and Särndal, C.-E. (1997), "Estimation of Variance in Presence of Imputation", *Proceedings of the Statistics Canada Symposium* 1997, *New Directions in Surveys and Censuses*, Statistics Canada, pp. 273-277.

Rao, J.N.K., and Sitter, R.R. (1995), "Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data", *Biometrika*, 82, pp. 453-460.

Rubin, D.B. (1976), "Inference and Missing Data", *Biometrika*, 63, pp. 581-590.

Särndal, C.-E., and Swensson, B. (1987), "A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse", *International Statistical Review*, 55, pp. 279-294.

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