New Tools at Statistics Canada to Measure and Evaluate the Impact of Nonresponse and Imputation

J.-F. Beaumont, D. Haziza, C. Mitchell and E. Rancourt

Statistics Canada, R.H. Coats Building, Ottawa (ON), Canada K1A 0T6.
jean-francois.beaumont@statcan.ca; david.haziza@statcan.ca; charles.mitchell@statcan.ca; eric.rancourt@statcan.ca.

Abstract

Since the 80's, Statistics Canada has invested in the development of generic software for editing and imputation. From the outset, it has been known by practicians and researchers that treating nonresponse, in particular by imputing for missing values, has impacts on estimates. In the 90's, resources were devoted to research how these impacts could be measured. These studies provided a number of answers that have led to the development of two systems (or sets of routines). One is a flexible tool that is used to carry out simulation studies in presence of imputation and under full response. It is called the GENEralised SImulation System (GENESIS). The second one is a System for Estimation of VAriance due to Nonresponse and Imputation (SEVANI). It is designed to provide users with one or more non-sampling variance component due to total or partial nonresponse when re-weighting or imputation is used to compensate for missing data. Both systems are SAS-based. The paper describes the features of the systems and presents an overview of their methodology.

1. Introduction

Statistical surveys such as those carried out at Statistics Canada are designed to collect and provide a wealth of information on and to the Canadian Society. This information takes the form of numerous point estimates. However, for inference purposes or to appropriately inform users, it is imperative to provide precision measures for any statistics produced. At Statistics Canada, the *Policy on Informing Users of Data Quality and Methodology* (Statistics Canada, 2001) provides a framework for quality measures. Further, there are *Quality Guidelines* (Statistics Canada, 2003) that guide survey statisticians on how to measure quality at each step of surveys.

When there is nonresponse, an extra source of error is present and it calls for new measures or adapted ones. Among others, these may take the form of simulation studies to evaluate the impact of a nonresponse treatment strategy or specific techniques such as variance estimation that appropriately take nonresponse and/or imputation into account.

In recent years, there has been a large amount of research on evaluating the impact of nonresponse especially when it is treated by means of imputation. Starting with multiple imputation (Rubin 1977,1987), there is now a fairly large range of methods that have been developed. A detailed review can be found in Lee, Rancourt and Särndal (2000, 2002) and in Shao (2002). Also, the context of nonresponse is better understood (see for instance the book Groves *et al*. eds, 2002) and imputation is now considered a statistical field in its own right (Rancourt, 2001).

With sample surveys and official statistics where single imputation is the norm, there has been (despite the fact that there are new variance estimation techniques) a lack of available computer tools specifically designed to help methodologists choosing, evaluating and measuring the impact of nonresponse and imputation. In this context, Statistics Canada has undertaken the development of two SAS-based systems that try to answer those needs. The first system, GENESIS, is a simulation system designed to help survey methodologists select their nonresponse/imputation strategy and quantify the relative performance of imputation methods through simulation studies. The second one, SEVANI, is a system designed to provide survey statisticians with a means to measure the variance that is due to nonresponse and/or imputation in production.

In the next sections, this document presents a brief overview of the two systems.

2. The Generalized Simulation System (GENESIS)

GENESIS 1.1 (Haziza, 2003) is a menu driven system based on SAS Release 8. The system contains SAS macros linked to menus using SAS/AF. GENESIS is a simple-to-use and relatively efficient system in terms of execution time. The user must first provide a data file in SAS format. This file represents the population used as the starting point for the simulations. The user then chooses a variable of interest and auxiliary variables. GENESIS contains three main modules:

(1) Full response module;
(2) Imputation module;
(3) Imputation/Reweighting classes module.

GENESIS is a two-fold tool. First, it may be used in a survey sampling course by the instructor in order to illustrate theoretical concepts in survey sampling such as the choice of sampling design, choice of an estimator or choice of allocation in the case of stratified random sampling. Also, GENESIS may be used in survey applications to help decide on an imputation strategy for a given survey. For example, GENESIS may be useful to answer such questions as: which imputation method should be used; how imputation classes should be formed; or how many classes there should.

Full Response Module

In the full response module, several sampling designs are available: simple random sampling, proportional-to-size sampling with and without replacement, stratified random sampling, Poisson sampling, one-stage and two-stage cluster sampling, two-phase sampling and the Rao-Hartley-Cochran method. For several designs, GENESIS draws R samples of size n (both user-specified) and can compute the Horvitz-Thompson, ratio and regression estimators for population totals and means and their estimated variances. Then, GENESIS displays several useful Monte Carlo results such as the relative bias of point and variance estimators, the mean squared error and coverage probability. GENESIS also displays several useful graphics that facilitate comparisons between estimators.

Imputation Module

In the imputation module, it is possible to carry out simulation studies to test the performance of imputed estimators (and, in some cases, variance estimators) under different scenarios.

(i) From the population, GENESIS draws simple random samples without replacement of size n (specified by the user).

(ii) GENESIS then generates nonresponse to the variable of interest according to one of the following three response mechanisms:

(1) MCAR (Missing Completely At Random): the probability of response is constant
(2) MAR (Missing At Random): the probability of response depends on one or more auxiliary variables
(3) NMAR(Not Missing At Random): the probability of response depends on the variable of interest

The user must specify the desired response rate. In the case of the MAR and NMAR mechanisms, the user can also choose to generate the nonresponse so that the probability of response increases or decreases with a function of the auxiliary variables or with the variable of interest.

(iii) The user then selects an imputation method among the following: previous value (or historical) imputation, mean imputation, ratio imputation, regression imputation, random hot deck imputation or nearest neighbour imputation (the user may specify the choice of distance). Finally, GENESIS computes the imputed estimator of the population mean.

(iv) For some imputation methods, GENESIS can also estimate the variance of the imputed estimator using the following methods:

(1) The two-phase approach under the MCAR mechanism (Rao, 1990);
(2) The two-phase approach based on a model (Särndal, 1992);
(3) The reverse approach under the MCAR mechanism (Shao and Steel, 1999);
(4) The reverse approach based on a model (Shao and Steel, 1999).

(v) Steps (1) to (4) are repeated R (user-specified) times.

Lastly, a number of Monte Carlo measures are proposed, such as the relative bias of the imputed estimators, their root mean squared error, the estimators of variance (when the estimation of variance option is selected) and the relative bias of the variance estimators. GENESIS stores important result tables (SAS tables) in a database that gives the user more flexibility in processing the results. For example, the user can easily calculate Monte Carlo measures other than those offered by GENESIS.

Imputation/Reweighting Classes Module:

In practice, it is customary to first form classes and then impute/reweight within each class. The primary objective of forming classes is to reduce nonresponse bias. Instead of forming classes, one could impute values directly using a regression model. However, there are at least two reasons for using classes: firstly, it is more practical when it is a matter of imputing a large number of variables at once, and secondly, classes provide a degree of robustness compared to the use of regression imputation.

To estimate the population mean, a random sample without replacement of size n is drawn. Suppose that the units respond to item y independently of one another such that the response probability for unit i is $p_i,\ i=1,...,n.$ An imputed estimator for $\overline{Y}$, denoted as $\overline{y}_I$, is defined by

$$\overline{y}_I = \frac{1}{\sum_{i\in s} w_i}\left[\sum_{i\in s_r} w_i y_i + \sum_{i\in s_m} w_i y_i^*\right] \qquad (1)$$

where $w_i = 1/\pi_i$ is the survey weight attached to unit i and $\pi_i = P(i \in s)$ is its first-order inclusion probability, s_r is the set of r units that responded to item y, s_m is the set of m units that did not respond to item y ($r+m=n$), and y_i^* is the imputed value created in order to "fill the hole" for the missing value y_i . Under mean imputation, the imputed estimator (1) is biased and the bias is given by

$$Bias(\overline{y}_I) = E(\overline{y}_I) - \overline{Y} \approx \frac{1}{N\overline{P}}\sum_{i=1}^{N}(p_i - \overline{P})(y_i - \overline{Y}) \qquad (2)$$

where $\overline{P} = \frac{1}{N}\sum_{i=1}^{N} p_i$. Expression (2) of the bias is equal to 0 if the covariance in the population between variables p and y is zero, which is the case, for example, if all units in the population have the same probability of responding (uniform response mechanism) and/or if the value of the variable of interest is the same for all units in the population. Obviously, these two requirements are very rarely met in practice. To reduce the nonresponse bias, it is common practice to divide the population into disjoint imputation classes. Imputation is then performed independently within each class, leading to an adjusted imputed estimator similar to (1), but based on C classes.

In the case of mean imputation within classes, the bias for the adjusted estimator becomes by

$$Bias(\overline{y}_{I,c}) \approx \frac{1}{N}\sum_{c=1}^{C}\overline{P}_c^{-1}\sum_{i\in U_c}(p_i - \overline{P}_c)(y_i - \overline{Y}_c), \qquad (3)$$

where $\overline{P}_c = \frac{1}{N_c}\sum_{i\in U_c} p_i$ and $\overline{Y}_c = \frac{1}{N_c}\sum_{i\in U_c} y_i$. The bias in (3) is equal to zero if the covariance between the variables p and y is zero not overall as in (2), but only within each class. This is more practical as it is usually possible to meet this requirement by forming imputation classes that are homogeneous with respect to the response probabilities p_i's and/or to the variable of interest y.

Various methods are used in practice to form imputation classes. GENESIS allows the user to test the behavior of two methods for constructing imputation classes: the cross-classification method and the score method.

The *Cross-classifying method* involves forming imputation classes by cross-classifying the auxiliary categorical variables that the user specifies. As is often the case in practice, the user may specify a number of constraints such as the minimum number of respondents per class or the fact that the number of respondents must be greater than the number of non-respondents in the classes. If the constraints are not met, GENESIS will eliminate one of the auxiliary variables and the remaining variables will be cross-classified.

In the *Scores method*, the first step in this method is to predict the variable of interest or the probability of response using the respondent units, which will produce two "scores": $\hat{y}$ and $\hat{p}$. The user must specify the desired number of classes C. After selecting one of the two scores (or both), the imputation classes are then formed using one of the following two methods, which the user will specify: the equal quantiles method, which forms imputation classes of approximately equal size or the classification method based on a classification algorithm that makes it possible to create homogeneous classes with respect to the selected score.

For both methods, GENESIS provides Monte Carlo measures, such as the relative bias of the imputed estimator or the relative root mean square error (RRMSE). For the scores method, GENESIS also provides graphics showing the behavior of the relative bias and the RRMSE when 1, 2,..., and C imputation classes are used.

3. The System for Estimation of the Variance Due to Nonresponse and Imputation (SEVANI)

SEVANI v1.0 (Beaumont and Mitchell, 2002) has recently been launched at Statistics Canada for production in surveys. It has already been considered by several surveys and has been tested for the Unified Enterprise Survey (UES). It is a SAS-based prototype system that can be used to estimate the nonresponse and imputation variance in a survey context when domain totals or means are estimated. SEVANI is designed to function in a SAS Release 8 environment either through the graphical user interface or directly with the macros.

To be able to provide estimated variances, the system requires the sample data file, final survey weights, sampling variance estimates and additional specifications about the estimator and the nonresponse treatment method used. Then, SEVANI will provide in a SAS file quantities such as the portion of the variance that is due to nonresponse/imputation, its proportion to total variance as well as the total variance (total of sampling and nonresponse/imputation variance).

SEVANI can deal with situations where nonresponse has been treated either by a nonresponse weighting adjustment or by imputation. For imputation, SEVANI handles the following four methods (within imputation classes or not):

1. Deterministic Linear Regression (such as mean or ratio imputation);
2. Random Linear Regression (such as random hot-deck imputation);
3. Auxiliary Value (such as carry-forward imputation);
4. Nearest Neighbour.

Note that auxiliary value imputation covers all methods for which the imputed value for a given unit k is obtained by using auxiliary data that come from this unit k only. Therefore, no information from the respondents is used to compute imputed values.

Variance estimation is based on the quasi-multi-phase framework (Beaumont and Mitchell, 2002), where nonresponse is viewed as additional phases of selection. In SEVANI, it is thus 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.

The Quasi-Two-Phase Framework

In the quasi-two-phase framework (or quasi-randomization in the terminology of Oh and Scheuren, 1983), nonresponse is viewed as a second phase of selection and the variable of interest y is only observed for part of the sample s. When there is nonresponse, the usual calibration or Horvitz-Thompson estimators, say $\hat{\theta}$, of the population parameter θ cannot be calculated since variable y is not observed for nonrespondents. 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-unbiased conditional on the realized sample s, that is, $\mathrm{E}_q(\hat{\theta}^* \mid s) \approx \hat{\theta}$, where q indicates that the expectation is evaluated with respect to the nonresponse mechanism. Consequently, the variance of $\hat{\theta}^*$ can be approximated by

$$\mathrm{V}_{pq}(\hat{\theta}^*) = \mathrm{V}_{\mathrm{sam}} + \mathrm{V}_{\mathrm{nr}} \;, \tag{4}$$

where $\mathrm{V}_{\mathrm{sam}} = \mathrm{V}_p(\hat{\theta})$ is the sampling variance, $\mathrm{V}_{\mathrm{nr}} = \mathrm{E}_p \mathrm{V}_q(\hat{\theta}^* \mid s)$ is the nonresponse variance and the subscript p refers to the sampling design. An approximately unbiased variance estimator for V_{nr} can simply be obtained by finding an approximately unbiased estimator for $\mathrm{V}_q(\hat{\theta}^* \mid s)$. In SEVANI, the Taylor linearization approach is used for this purpose.

Since in practice the nonresponse mechanism is unknown, a nonresponse model is needed to approximate it. Therefore, the variance in (4) 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. Note that the extension to the quasi-multi-phase framework is straightforward and can be found in Beaumont and Mitchell (2002).

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 use the imputation model to obtain a variance estimator as in Särndal (1992). Instead of estimating $\mathrm{V}_{pq}(\hat{\theta}^*)$, he proposed to estimate the model expectation of $\mathrm{V}_{pq}(\hat{\theta}^*)$, which is given by

$$\begin{aligned} \mathrm{E}_m\mathrm{V}_{pq}(\hat{\theta}^*) &\approx \mathrm{E}_m\mathrm{V}_p(\hat{\theta}) + \mathrm{E}_m\mathrm{E}_p\mathrm{V}_q(\hat{\theta}^* \mid s) \\ &= \mathrm{V}_{\mathrm{sam}}^{\mathrm{m}} + \mathrm{V}_{\mathrm{nr}}^{\mathrm{m}}\ , \end{aligned} \tag{5}$$

where, here, the sampling variance is $\mathrm{V}_{\mathrm{sam}}^{\mathrm{m}} = \mathrm{E}_m\mathrm{V}_{\mathrm{sam}}$ and the nonresponse variance is $\mathrm{V}_{\mathrm{nr}}^{\mathrm{m}} = \mathrm{E}_m\mathrm{V}_{\mathrm{nr}}$. 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 simplifies variance estimation. In particular, provided that $\mathrm{E}_q(\hat{\theta}^* \mid s) \approx \hat{\theta}$, the nonresponse variance can be written as

$$\mathrm{V}_{\mathrm{nr}}^{\mathrm{m}} = \mathrm{E}_p\mathrm{E}_q\mathrm{E}_m\left\{(\hat{\theta}^* - \hat{\theta})^2\right\}. \tag{6}$$

As a result, an approximately unbiased variance estimator for $\mathrm{V}_{\mathrm{nr}}^{\mathrm{m}}$ 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, but only assuming that it is ignorable. This assumption is much weaker than the nonresponse model usually required for estimating (4). However, this approach requires that the imputation model *m* be valid. The choice between estimating (4) or (5) 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) should be preferred over estimating (5) and vice-versa.

Models Used in SEVANI

In practice, joint response probabilities are never estimated directly. Therefore, it is necessary to assume some form of independence to avoid estimating them and to simplify variance estimation. In SEVANI, it is assumed that clusters of units respond independently. That is, all sample units within a given cluster are observed simultaneously or they are all missing simultaneously. A typical example occurs when dwellings (clusters) are selected but the desired information is collected for all people in the selected dwellings.

In SEVANI, the estimated response probabilities are assumed to be equal to the true response probabilities. This should lead to an underestimation of the nonresponse variance V_{nr} 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, 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 (5), SEVANI uses a multiple linear regression imputation model *m* in which observations are assumed to be independent of one another. This is the imputation model approach. Sometimes, imputation is performed independently within imputation classes. SEVANI does it by fitting a different model for each class.

Imputation Variance

Whether it is chosen to estimate V_{nr} or $\mathrm{V}_{\mathrm{nr}}^{\mathrm{m}}$, an additional component of variance must be computed when Random Linear Regression (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 Deterministic Linear Regression 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).

Nearest-Neighbour (NN) imputation is a nonparametric imputation method since it does not require specifying the imputation model either to justify the form of the adjusted estimator $\hat{\theta}^*$ or to determine its model unbiasedness property. Therefore, using a linear regression imputation model to obtain a variance estimator may not always be appropriate, especially when the linear model does not hold satisfactorily. If the 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 this definition of NN imputation, there is no variability. SEVANI deals with this problem by approximating 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. The user may select the same standardization method and the same distance measure as those used to perform NN imputation.

4. References

Beaumont, J.-F. and Mitchell, C. (2002). The System for Estimation of Variance Due to Nonresponse and Imputation (SEVANI), *Proceedings of Statistics Canada Symposium 2002: Modeling Survey Data for Social and Economic Research.*

Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (eds.) (2002) *Survey Nonresponse*. John Wiley and Sons.

Haziza, D. (2003). The Generalized Simulation System (GENESIS). *Proceedings of the Section on Survey Research Methods*, American Statistical Association, to appear.

Lee, H., Rancourt, E., Särndal, C.-E. (2000). Variance Estimation from Survey Data under Single Value Imputation, *Working Paper HSMD - 2000 - 006E*, Methodology Branch, Statistics Canada.

Lee, H., Rancourt, E., Särndal, C.-E. (2002). Variance Estimation from Survey Data under Single Value Imputation, in Groves, R., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds.), *Survey Nonresponse*, John Wiley & Sons, Inc., New York.

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*, 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, 143-184.

Rancourt, E. (2001). Edit and Imputation: From Suspicious to Scientific Techniques. *Proceedings of the International Association of Survey Statisticians*, 634-655.

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

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

Rubin, D.B. (1977). Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, 72, 538-543.

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

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

Shao, J. (2002). Replication Methods for Variance Estimation in Complex Surveys with Imputed Data, in Groves, R., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds.), *Survey Nonresponse*, John Wiley & Sons, Inc., New York.

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.

Statistics Canada. (2001). - *Policy on Informing Users of Data Quality and Methodology*, Statistics Canada Policy Manual.

Statistics Canada. (2003). *Quality Guidelines*. Catalogue No. 12-539-XIE. Third Edition, October 2003.