WP. 10
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (ii): Implementing editing strategies and links to other parts of processing

ASSESSING AND DEALING WITH THE IMPACT OF IMPUTATION THROUGH VARIANCE ESTIMATION

Invited Paper

Submitted by Statistics Canada[1]

Abstract: Imputation is a well-known approach to treat non-response in surveys. However, it can have a number of impacts on data and other processes, but more importantly, on estimates produced from these data. In recent years, imputation research has lead to the development of a number of methods and approaches as well as software to assess the quality of imputation by means of estimation of the variance under imputation. These methods are either based on a model, or on re-sampling techniques. In this field, we are now at a point where the main questions are about which method(s)/tool(s) should be used to perform variance estimation under imputation; about which quality measures should be developed based on these and about how to interpret results coming out of them. This paper attempts to respond to such questions by presenting the various aspects of the problems, by giving an overview of the existing methods and tools and by proposing potential applications and research avenues.

I. INTRODUCTION

1. Surveys are faced with the problem of missing and non-responding data. The missing information may be at the unit level (total non-response) or at the item level (partial non-response). Also, the information may be available but inconsistent or unusable. In this case, editing rules usually identify such records. In this paper, we shall consider all of these, but concentrate on non-response.

2. In the presence of non-response extra care must be taken to deal with this potential source of error. First, it must be detected from the data after collection. Then, it needs to be treated and finally, the resulting quality of estimates produced should be assessed.

3. In the context of surveys, non-response (especially partial non-response) is very often treated by imputation. To perform imputation, there are numerous methods that are all the resulting expression of a modelling exercise between the variable of interest and other variables present on the file. The degree of sophistication of the model will depend on the auxiliary information available, on the needs for simplicity and the analysts' needs.

4. Imputation is very attractive, because it leads to a complete data file that can be used by regular complete data software. However, this too often brings a false sense of security about the quality of data and estimates produced. Imputation has impacts that should and can be measured. These may be direct

[1] Prepared by Eric Rancourt, Business Survey Methods Division, Statistics Canada, (eric.rancourt@statcan.ca).

(on the process, on estimates) or indirect (on the next design, on secondary analysis). It is thus clear that quality measures are essential.

5. Statisticians have always been uncomfortable with inference from samples containing imputed data. In the last few decades, a significant amount of research has been carried out to provide a theoretical framework to imputation that would enable statisticians to measure the quality of imputed data and of estimates based on them. The first instance is in Rubin (1977) where the multiple imputation approach was introduced. Equipped with this tool, producers of micro data files can provide external users with properly imputed files from which correct inference can be made. In the 90's, then followed a number of approaches to estimating the variance under imputation. For instance, Lee, Rancourt and Särndal (2000; 2002) and Shao (2002) present a detailed account of such methods along with extensive comparisons, both qualitative and quantitative. Rao (1996) also presents a review of the topic.

6. There have been recent developments in the field of quality assessment of imputation, particularly with respect to tools, software and their application. Some methods can now be used to assess the quality of manual imputation (Rancourt, 1997) or the quality of editing (Rancourt, 2002). Further, a number of methods are now available in software such as SEVANI (Beaumont and Mitchell, 2002) or GENESIS (Haziza, 2003).

7. This paper is divided as follows: Section II briefly describes and discusses variance estimation methods under imputation. Then Sections III and IV respectively present Statistics Canada's systems: GENESIS and SEVANI. Section V outlines potential applications of imputation variance estimation and Section VI briefly describes some research avenues.

II. VARIANCE ESTIMATION METHODS

8. As described in Rancourt (2001), having at hand a method to evaluate the impact (variance) of imputation allows one to construct a better imputation strategy. Estimation of the total variance (including imputation) can be very useful. For example, it allows for:

- More precise estimation of the total variance, and as a result, a better knowledge of the data quality by the statistical agency;
- The possibility to make correct inferences;
- Better quality reporting to users, who can sometimes be misled by knowing only the sampling variance;
- Improved evaluation and choice of imputation methods by assessing the variations in precision (imputation variance) resulting from various approaches;
- Production of separate estimates of sampling and imputation variances to plan or adjust budgets between the sample size and the imputation/follow-up effort.

9. Variance estimation methods can be grouped into four categories: the two-phase approach; the reversed approach; re-sampling and multiple imputation.

10. Two-phase approach. The two-phase approach is based on the idea that the response mechanism is the second phase of a two-phase selection mechanism. It is similar to two-phase or double sampling but different in that the second phase process in NOT controlled by the survey statistician. In this approach, a model is necessary. One may use a model for the response mechanism, or a model for the data.

11. The response model approach was presented in Rao and Sitter (1995). The simplifying assumption that non-response is uniform is often made, but the approach is general enough to allow one to model non-response in any fashion. In the end, it is as if one tries to restore the second phase inclusion

probabilities by estimating response probabilities. The resulting variance estimator conveniently splits into two terms, one that represents sampling and one that represents non-response (imputation)

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{NRP}}.$$

12. The data model approach (Särndal, 1992) consists of using the relationships that exist between the variable of interest and other variables among the respondents to evaluate the expected model error caused by imputed nonrespondents. Under imputation, the total error of an estimator can be decomposed as the sampling error, the non-response (imputation) error and a mixed term. The estimator for the total variance is made of a sampling component which is the usual sampling variance estimator[2] under imputation and an imputation component. There is also a mix term, but in many situations it is exactly or approximately zero. The expression for the variance estimator is

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + \hat{V}_{\text{MIX}}.$$

13. Reversed approach. In the reversed approach, being a respondent or a non-respondent is assumed to be a characteristic of every unit in the population. As a result, the *selection* of respondents can be viewed as taking place before selection of the sample. Therefore, we have a population of respondents selected from the population and a sample of respondents selected from the population of respondents. This approach was introduced by Fay (1991) and Shao and Steel (1999). The corresponding estimator still has two components, but their meaning is slightly different. In the case of negligible sampling fractions, the second component $(\hat{V}_2)$ becomes negligible and we obtain a simplified one-term variance estimator. In this case,

$$\hat{V}_{\text{TOT}} = \hat{V}_1 + \hat{V}_2.$$

14. Re-sampling techniques. The usual re-sampling techniques have now been adapted to the context of imputed data for estimation of V_1 above. The main ones are the Jackknife technique (Rao and Shao, 1992) and the Bootstrap technique (Shao and Sitter, 1996). They are reviewed in Shao (2002).

15. The jackknife technique (Rao and Shao, 1992) was developed to estimate the total variance. It is justified in the context of the reversed approach and actually measures its first term $(\hat{V}_1)$. The principle is that whenever a unit is deleted, a correction has to be made to all imputed values that were originally influenced by this value. Once the corrections are made, the usual jack-knife variance estimator can be used:

$$\hat{V}_{\text{JKNF}} = \sum \frac{n-1}{n} \sum \left(\hat{q}^{j} - \hat{q}\right)^2.$$

16. The Bootstrap technique is a natural extension of the idea of simulating non-response and imputation to measure their impact. In the bootstrap (Shao and Sitter, 1996), the principle is to repeat the non-response and imputation processes within each of the bootstrap samples, thereby allowing the ordinary formula to be appropriately used. However, it can be implemented without replicating non-response by using adjustments to imputed codes. For the bootstrap, the variance estimation formula is the usual one:

[2] The usual sampling variance estimator can be used directly provided that errors are added to the imputed values for variance estimation purposes. For instance, see Gagnon, Lee, Rancourt and Särndal (1996). Otherwise, an adjustment component is required.

$$\hat{V}_{\text{BOOT}} = \frac{1}{L}\sum_{l=1}^{L}\left(\hat{q}^{j} - \bar{\hat{q}}\right)^{2}.$$

17. Multiple imputation. Multiple imputation was developed by (Rubin 1977, 1987), and Rubin (1996). It was develop to answer the needs for quality measure of estimates by external (or secondary) users of statistics. Multiple imputation is based on the approximate Bayesian bootstrap where more than one value is imputed for missing values so that a variance between sets of imputed values may be obtained. To get the final estimator, it is combined with the average variability within sets of imputed data. The estimator has the form:

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{BET}} + \left(\frac{M+1}{M}\right)\hat{V}_{\text{WITH}}.$$

18. There are also other approaches such as the balanced repeated replication method that was adapted for the case of imputed data, but they are not treated here.

19. In practice, the main issue with estimation of the variance in presence of imputation is not about how to proceed, but about which method to use. Each has its own merits and the choice should be made based on a number of considerations, including

- The approach implemented in production under full response;
- The need for separate components for sampling and imputation variance;
- Whether users are internal or external to the agency or group producing imputation;
- Simplicity;
- Complexity of the design and estimator.

20 Since the 1980s, Statistics Canada has invested in the development of generic software for various survey steps. Recently, two gaps have been identified with respect to imputation software: the ability to perform repeated simulation studies with missing data and imputation without having to write a new program each time and taking imputation into account while estimating the variance of statistics in surveys. The systems presented here (GENESIS and SEVANI) are major steps trying to fill the gaps.

III. THE GENERALISED SIMULATION SYSTEM (GENESIS)

21. GENESIS v1.1 (Haziza, 2003) is a menu driven system based on SAS Release 8. It contains SAS macros linked to menus using SAS/AF. The system was developed to address the fact that several methodologists at Statistics Canada regularly conduct simulation studies in the presence of imputation. It therefore seemed appropriate to create a tool that would enable users to conduct such simulation studies without having to write a program each time. GENESIS is a simple to use and relatively efficient system in terms of execution time. The system assumes that a population data file is provided in SAS format. This population file is used as the starting point for 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.

22. **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.

23. For several designs, GENESIS computes the Horvitz-Thompson, ratio and regression estimators. It displays several useful Monte Carlo results such as the relative bias of point and variance estimators, the mean squared error and coverage of the confidence interval. GENESIS also displays several useful graphics that facilitates the comparison between estimators.

24. **In the imputation module**, simulation studies can be carried out to test the performance of imputed estimators (and, in some cases, variance estimators) under different scenarios. From the population provided, GENESIS draws simple random samples without replacement of size *n* (specified by the user).

25. GENESIS then generates non-response to the variable of interest according to one of the following three response mechanisms:

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

26. 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 non-response so that the probability of response increases or decreases with a function of the auxiliary variables or with the variable of interest.

27. In terms of imputation methods, the user may select one of the following:

- Previous value (or historical) imputation;
- Mean imputation;
- Ratio imputation;
- Regression imputation;
- Random hot deck imputation;
- Nearest neighbour imputation (for which the user may specify the choice of distance).

28. For some imputation methods, GENESIS estimates the variance of the estimators by the following methods:

- The two-phase approach under the MCAR mechanism (Rao and Sitter, 1995);
- The two-phase approach based on a model (Särndal, 1992);
- The reverse approach under the MCAR mechanism (Shao and Steel, 1999);
- The reverse approach based on a model (Shao and Steel, 1999).

29. Steps (1) to (4) are repeated *R* times where *R* is the number of iterations specified by the user. 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), the relative bias of the variance estimators, etc.

30. GENESIS stores important results tables (SAS tables) in a database that gives the user more processing flexibility. For example, the user can easily calculate Monte Carlo measures other than those offered by GENESIS.

31. **In the Imputation/Reweighting classes module,** GENESIS allows the user to test the performance of methods for constructing imputation classes (method by cross-classification and score method).

32. GENESIS provides a means of examining the behaviour of two methods of forming imputation classes: the method by cross-classification and the score method. Within the classes, the user can choose to impute by mean or by random hot deck.

33. Cross-classifying method**:** This method involves forming imputation classes by cross-classifying auxiliary categorical variables specified by the user. He or she may also specify a number of constraints such as a minimum number of respondents per class or that the number of respondents 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.

34. Scores method**:** The first step in this method is to predict the variable of interest or the probability of response using the respondent units, leading to two "scores": $\hat{y}$ et $\hat{p}$. 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 the equal quantiles method, which forms imputation classes of approximately equal size or using the classification method based on an algorithm that makes it possible to create homogeneous classes with respect to the selected score.

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

IV. THE SYSTEM FOR ESTIMATION OF THE VARIANCE DUE TO NON-RESPONSE AND IMPUTATION (SEVANI)

36. SEVANI v1.0 (Beaumont and Mitchell, 2002) is a SAS-based prototype system that can be used to estimate the non-response and imputation variance portions in a survey context when a domain total or mean is estimated. SEVANI is designed to function in a SAS v8 environment either directly using the macros or through the graphical user interface.

37. To be able to provide estimated variances, the system requires the sample data file, final survey weights and sampling variance estimates (before taking non-response/imputation into account). Then SEVANI will provide in a SAS file, the portion of the variance that is due to non-response, to imputation, their proportion to total variance as well as the total variance (total of sampling, non-response and/or imputation)

38. Variance estimation is based on the quasi-multi-phase framework (Beaumont and Mitchell, 2002), where non-response is viewed as additional phases of selection. Since the survey methodologist does not control the non-response mechanisms, a non-response model is required. When imputation is used to treat non-response, strength can be gained by using an imputation model. In SEVANI, it is possible to estimate the non-response variance associated to more than one non-response mechanism or, in other words, more than one cause of non-response. For example, most surveys suffer from unit and item non-response and these two types of non-response are likely to be explained by different non-response mechanisms. Moreover, they are often not treated in the same way. Unit non-response is usually treated by a non-response weighting adjustment technique while item non-response is usually treated by an imputation technique.

39. Non-response inevitably leads to an observed sample of smaller size than the sample originally selected. This sample size reduction is usually accompanied by an increase in the variance of the estimates, no matter which method is chosen to treat non-response. This increase in variance is called the non-response variance. The imputation variance is defined in SEVANI as a component of the non-response variance, which is due to the use of a random imputation method.

40. SEVANI can deal with situations where non-response has been treated either by a non-response weighting adjustment or by Imputation. If imputation is chosen, SEVANI requires that one of the following four imputation methods be used (within imputation classes or not):

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

41. 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.

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

V. POTENTIAL APPLICATIONS OF IMPUTATION VARIANCE ESTIMATION

43. Estimating the variance in the presence of imputation is only one dimension of measuring the overall quality of survey estimates. There is a large potential to extending the methods developed for estimation of the variance in presence of imputation. We outline the cases of manual imputation (or adjustments) and editing.

46. Manual imputation. This includes any type of manual intervention to modify data in the course of editing or analyzing the data. As in Rancourt (1997), manual imputation can be linked to previous value imputation. If the process can be replicated in a controlled setting (e.g. analysts making changes to values that are known to be correct), then the associated error can be evaluated using the data model method within the two-phase approach in paragraph 12.

47. Editing. It can be considered as part of the non-response mechanism. This is often the implicit assumption when dealing with the problem of missing data. However, the editing process is completely distinct from the actual response mechanism and should be studied accordingly. Since an editing model cannot be used, a *data* model (ratio) such as

$$x: y_k = b\, x_k + e_k, \quad \mathrm{E}_x(e_k) = 0, \ \mathrm{E}_x(e_k e_{k'}) = 0, \quad \mathrm{E}_x(e_k^2) = s^2 x_k$$

may be considered as in Rancourt (2003).

48. Under a data model, the editing process can be fixed, since the variables of interest are random. The total variance (sampling, non-response and editing) can be worked out to be

$$V_{\mathrm{TOT}} = V_{\mathrm{Editing}} + V_{\mathrm{Nonresponse}} + V_{\mathrm{Sampling}} + \text{ Small mix terms.}$$

Then with measures such as $\hat{V}_{\mathrm{Editing}}$ and $\hat{V}_{\mathrm{Nonresponse}}$, one can study various impacts of editing to monitor the editing process.

VI. RESEARCH AVENUES

49. The context of survey taking is in constant evolution and has increased in complexity from the original (theoretical) idea of univariate inference from a simply drawn sample to a population. Areas of research in the context on imputation could include

- The context of combining survey and administrative data through direct replacement of survey data whether they are adjusted or not;
- Rolling surveys and censuses, where data are missing by design for some areas or groups at any given point in time;
- Multivariate aspects of imputation and relationships between variables;
- Multi-level aspects of imputation when imputation classes are sequentially used in a hierarchical setting.

References

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

FAY R.E. – A Design-Based Perspective on Missing Data Variance. Proceedings of the Annual Research Conference, US Bureau of the Census, 429-440, 1991.

GAGNON F., LEE H., RANCOURT E. and SÄRNDAL C.-E. – Estimation the Variance of the Generalized Regression Estimator in the Presence of Imputation for the Generalized Estimation System, *Proceedings of the Survey Methods Section*, 151-156, 1996.

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

LEE H., RANCOURT E., SÄRNDAL C.-E. – 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, 1997.

LEE H., RANCOURT E., SÄRNDAL C.-E. – Variance Estimation from Survey Data under Single Value Imputation, *Working Paper HSMD – 2000 – 006E*, Methodology Branch, Statistics Canada, 2000.

LEE H., RANCOURT E., SÄRNDAL C.-E. – Variance Estimation from Survey Data under Single Value Imputation, in *Survey Non-response*, Groves, R. et al eds., J. Wiley and Sons, New York, 315-328, 2002.

RANCOURT E. – Estimation de la Variance en Présence d'Imputation par Valeur Précédente. Colloque Francophone sur les Sondages, Rennes, 1997

RANCOURT E. – Edit and Imputation: From Suspicious to Scientific Techniques. *Proceedings*, International Association of Survey Statisticians, 604-633, 2001.

RANCOURT E. – Using Variance Components to Measure and Evaluate the Quality of Editing Practices, *Working paper No. 10*, Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing, Helsinki, 2002.

RANCOURT E. – Statistics Canada's New Software to Better Understand and Measure the Impact of Non-response and Imputation, *Working paper No. 10*, Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing, Madrid, 2003.

RAO J.N.K., SHAO J. – Jackknife Variance Estimation with Survey Data under Hot-deck Imputation, Biometrika, 79, 811-822, 1992.

RAO J.N.K. – On Variance Estimation with Imputed Survey Data. *Journal of the American Statistical Association*, 91, 499-506, 1996.

RAO J.N.K., SITTER R.R. – Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data, *Biometrika*, 82, 453-460, 1995.

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

RUBIN D.B. – *Multiple Imputation for Non-response in Surveys*. New York, John Wiley, 1987.

RUBIN D.B. – Multiple Imputation after 18 + Years. Journal of the American Statistical Association, 91, 473-489, 1996.

SÄRNDAL C.-E. – Method for Estimating the Precision of Survey Estimates when Imputation Has Been Used, *Survey Methodology*, 241-252, 1992.

SHAO J. – Replication Methods for Variance Estimation in Complex Surveys with Imputed Data, in *Survey Non-response*, Groves, R. et al eds., J. Wiley and Sons, New York, 303-314, 2002.

SHAO J., SITTER, R.R. – Bootstrap for Imputed Survey Data. Journal of the American Statistical Association, 91, 1278-1288, 1996.

SHAO J., STEEL P. – Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions, *Journal of the American Statistical Association*, 94, 254-265, 1999.
