STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 29
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Madrid, 20-23 October 2003)

STATISTICS CANADA'S NEW SOFTWARE TO BETTER UNDERSTAND AND MEASURE THE IMPACT OF NONRESPONSE AND IMPUTATION

Supporting paper

Submitted by Statistics Canada[1]

Abstract: In the survey-taking context, treatment of nonresponse - most often imputation - is always part of the data processing steps. Over the years, Statistics Canada has invested in research not only to constantly improve imputation methods and good practices, but also to be able to measure and understand the impact of imputation. Recently, the imputation research activities have expanded from more academic-type research into a broader set of activities such as what is usually found in a resource centre. These activities consist in research papers, a committee on practices in imputation (COPI), an imputation bulletin, courses, consultation, participation in workshops and now software. Two new systems, both in SAS have been developed. The first, GENESIS, is a generalized system for imputation simulations. It allows users to carry out simulation studies under a wide variety of conditions (nonresponse mechanisms, imputation classes, imputation methods, sampling designs, variances estimators). It then provides Monte-Carlo measures such as bias, variance and MSE. Results are stored in tables and can easily be accessed. The second system is SEVANI, a system to estimate the variance in presence of nonresponse and imputation. It is designed to provide users with the portion of variance that is due to the nonresponse adjustments (whether it is compensated for by re-weighting, by imputation or both). This paper highlights the two systems and the context in which they were created.

I. INTRODUCTION

1. Measuring a phenomenon or a population value often has a highly important value. However, for inference purposes or to appropriately inform users, it is imperative to provide a precision measure for any statistics produced. At Statistics Canada, there is a Policy on informing users of data quality and methodology (Statistics Canada, 2001) that serves as a framework to determine which quality measures to use. Further, there are Quality Guidelines (Statistics Canada, 1998) that provide directions on how to measure quality at each step of surveys.

[1] Prepared by Eric Rancourt with Jean-François Beaumont, David Haziza and Charles Mitchell, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. eric.rancourt@statcan.ca.

2. When there is nonresponse, an extra source of error is present and it calls for new measures or adapted ones. One such measure is the variance, but it needs to be adapted to situations where imputation is used (or in the context of editing as in Rancourt, 2002).

3. In the last two decades, there has been a large amount of research on estimating the variance in presence 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 Rancourt, Lee and Särndal (2000, 2002) and in Shao (2002).

4. In the context of 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.

5. Recently, 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 measure of the variance that is due to nonresponse and/or imputation.

6. This paper is divided as follows: Section 2 describes Statistics Canada's context in which software and systems such as GENESIS and SEVANI are designed and developed. Section 3 and 4 respectively present GENESIS and SEVANI. The conclusion follows in Section 5.

II. TOOLS, SOFTWARE AND SYSTEMS AT STATISTICS CANADA

7. Since the 80's, Statistics Canada has invested in the development of generic software for various survey steps. For example, systems such as GEIS, the Generalized Edit and Imputation System (Cotton, 1991) or as GES, the Generalized Estimation System (Estevao and Hidiroglou, 1995) are now software regularly used at Statistics Canada and in a few other agencies. Such systems are usually built in three phases. The first phase is composed of research, development and preparation of specifications within the Methodology Branch. The second one is composed of system analysis and programming in the Informatics Branch. Finally, the systems go through extensive testing within the Informatics Branch and in the Methodology Branch. There are also programs and software that are built specifically for surveys. For example, the Canadian Census has its own E&I system, CANCEIS, the Canadian Census Edit and Imputation System, based on the Nearest-neighbour Imputation Methodology (NIM) (Bankier, Lachance and Poirier, 2000).

8. As computers are more and more powerful, and as the employees and incoming new staff are more and more at ease with computer programming, there is a new type of systems that has started coming to life. These systems (often in the form of routines or SAS macros) are built by survey methodologists for a specific purpose in the context of one survey. However, given his/her knowledge of computer programming and other surveys, a methodologist will occasionally step back from the context of the survey at hand, think about a number of extensions and make his/her program more general so that several surveys may then use it. At some point, these systems may eventually become part of Statistics Canada's official software series. Among those are programs such as IMPUDON (Methodology Branch, 2001) for donor

imputation, used for several business surveys and BOOTVAR (Statistics Canada, 2002) for variance estimation based on the bootstrap technique and used by data analysts.

9. Nonresponse and imputation have been a regular methodology research and development field since the mid 1980's. In the 1990's, research started to focus on trying to compute the variance due to imputation and at the end of the 1990's the emphasis shifted to providing tools and support to methodologists.

10. There is now a series of activities surrounding imputation much in the form of an imputation resource centre. This group is involved in several activities. They are:

11. Research: Each year, a research budget of about one and a half person-year is allocated among a number of methodologists who are involved in theoretical work on imputation. These activities usually lead to published papers or presentations at conferences.

12. The Committee on Practices in Imputation (COPI): The committee has been formed at the beginning of 2000 and meets regularly. Each time, the imputation strategy of a survey or a research/development topic is presented and discussed. These presentations serve as an input to the group for potential research areas and the presenters receive advices on the issues raised.

13. The Imputation Bulletin: The Bulletin is produced twice a year and is aimed at establishing a bridge between imputation theory and practice. It is a key source of dissemination of information on imputation. Each Bulletin covers three to four imputation topics and presents a unique application. The Bulletin also informs methodologists of new papers released in various journals and upcoming conferences and courses.

14. Participation in workshops and seminars: Members of the group present the results of their findings and developments in conferences and participate in workshop in other statistical agencies or universities.

15. Teaching courses: A number of nonresponse and or imputation courses are offered on a regular basis by members of the group at Statistics Canada. Courses have also been developed to be presented as part of one to three day workshops during conferences.

16. Consultation: Various methodologists often consult members of the imputation group on issues related to nonresponse, imputation and estimation in their survey. Some of these consultations may give birth to research projects.

17. Development of software: 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)

18. 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 simple to use and a 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.

19. **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.

20. 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.

21. **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).

22. GENESIS then generates nonresponse 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.

23. 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.

24. 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).

25. 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).

26. 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.

27. 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.

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

29. 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.

30. 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.

31. 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.

32. 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 NONRESPONSE AND IMPUTATION (SEVANI)

33. SEVANI v1.0 (Beaumont and Mitchell, 2002) is a SAS-based prototype system that can be used to estimate the nonresponse 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.

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

35. Variance estimation is based on the quasi-multi-phase framework (Beaumont and Mitchell, 2002), where nonresponse is viewed as additional phases of selection. Since the survey methodologist does not control the nonresponse mechanisms, a nonresponse model is required. When imputation is used to treat nonresponse, strength can be gained by using an imputation model. In SEVANI, 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.

36. Nonresponse 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 nonresponse. This increase in variance is called the nonresponse variance. The imputation variance is defined in SEVANI as a component of the nonresponse variance, which is due to the use of a random imputation method.

37. SEVANI can deal with situations where nonresponse has been treated either by a nonresponse 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.

38. 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.

39. As noted by Rancourt, Lee and Särndal (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:

- To obtain valid inferences in the presence of nonresponse;
- To properly measure the quality of estimates and to inform users of the data quality;
- To better allocate survey resources between sample size and nonresponse related activities;
- To compare different nonresponse treatment strategies in order to make better decisions.

40. 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.

41. 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.

V. CONCLUSION

42. Both GENESIS and SEVANI have recently been launched at Statistics Canada for production in surveys. They are still in the form of prototypes, but several methodologists working on various surveys have started using them or are considering their use. These systems should help quantify the impact of nonresponse and imputation.

References

BANKIER M., LACHANCE M. and POIRIER P. – 2001 Canadian Census Minimum Change Donor Imputation Methodology, Proceedings *of the Workshop on Data Editing, UN-ECE*, United Kingdom, Cardiff, 2000.

BEAUMONT J.-F., MITCHELL C. – 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,* 2002.

COTTON C. – Functional description of the Generalized Edit and Imputation System, *Technical Document*, Business Survey Methods Division, Statistics Canada, 1991.

ESTEVAO, V., HIDIROGLOU M.A., SÄRNDAL C.-E. – Methodological Principles for a Generalized Estimation System at Statistics Canada, *Journal of Official Statistics,* 11, 181-204, 1995.

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 Nonresponse*, Groves, R. et al eds., J. Wiley and Sons, New York, 315-328, 2002.

METHODOLOGY BRANCH. – *The Imputation Bulletin*, Vol. 1, No. 1, Statistics Canada, 2001.

RANCOURT E. – Using Variance Components to Measure and Evaluate the Quality of Editing Practices. *Working paper No. 11*. Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing, Helsinki, 2002.

RANCOURT E., GAGNON F., LEE H., PROVOST M., and SÂRNDAL C.-E. – Estimation of Variance in Presence of Imputation, *Proceedings of the Statistics Canada Symposium* 1997, *New* Directions in Surveys and Censuses, Statistics Canada, 273-277, 1997.

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 Nonresponse in Surveys*. New York, John Wiley, 1987.

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 Nonresponse*, Groves, R. et al eds., J. Wiley and Sons, New York, 303-314, 2002.

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.

STATISTICS CANADA. – Estimation of the Variance Using Bootstrap Weights. *User's Guide for BOOTVARE_V20.SAS program*. Health Division. 2002

STATISTICS CANADA. – *Policy on Informing Users of Data Quality and Methodology*, Statistics Canada Policy Manual, 2001.

STATISTICS CANADA. – *Quality Guidelines*. Catalogue No. 12-539-XIE. Third Edition, October 1998.