STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 11
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (ii): Measuring and evaluating data editing quality

USING VARIANCE COMPONENTS TO MEASURE AND EVALUATE THE QUALITY OF EDITING PRACTICES

Invited paper

Submitted by Statistics Canada[1]

Abstract: Sampling errors have usually attracted the attention of survey statisticians because of their direct link with the well-developed sampling theory. However, it is generally acknowledged that non-sampling errors may account for the greater proportion of the total error. This is particularly true for editing. Although there is no general methodology theory (as opposed to sampling theory) we can build on existing methods to develop tools appropriate to the editing context. One such method is the calculation of nonresponse and/or imputation variance, which can be extended to editing. Its use could shed light on the relative precision of approaches. In this paper, we explore ways in which the components of the total variance can be used to evaluate the quality of editing practices. We also consider how to use these variance components to help make the selection of a best editing strategy among many.

I. INTRODUCTION

1. Survey-taking is an exercise that comprises a large number of different activities. Since all of these activities require human intervention (or have been developed by humans), they are subject to errors. This is true regardless of whether the survey is a sample survey or a census. Therefore, there is a need not only to prevent errors, but also to measure their impact on the quality of estimates and to try to improve the situation. Even in a perfect scenario with all possible and imaginable efforts being made to avoid errors, some would still remain.

2. The errors are usually categorised into two types. The first type is the sampling error. It stems from the fact that only a sample of the population is investigated instead of all units. The theory and methods for this type of error have been studied at length in the past 50 years. In the context of a census, there is no sampling error, and for large surveys, it may not be the dominant source of error.

3. The second type is called non-sampling error. This includes all sources of errors that may cause an estimate not to be in perfect agreement with the parameter it is supposed to estimate. Non-sampling errors can be of various types and are often classified into:

[1] Prepared by Eric Rancourt, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. eric.rancourt@statcan.ca.

(i) Coverage error, caused by a survey frame not matching exactly the target population;
(ii) Response (measurement) error, caused by respondents failing to provide “true” values to answers;
(iii) Nonresponse error, caused by a wide variety of reasons that lead to missing data;
(iv) Processing error, caused by any manipulation that introduces error into estimates.

4. The interest of this paper is in a specific cause for nonresponse, namely editing. Indeed, editing can be viewed as one of the causes that lead to missing data, especially when there is a large number of inconsistencies found. With editing, it is never known to the survey manager or the survey statistician whether enough or too much is performed. Of course, it is desirable to try to find the faulty data, but beyond a certain point, good data are being erased. It is therefore important to measure the impact of editing, especially because the amount of editing can sometimes be fairly substantial or even too large.

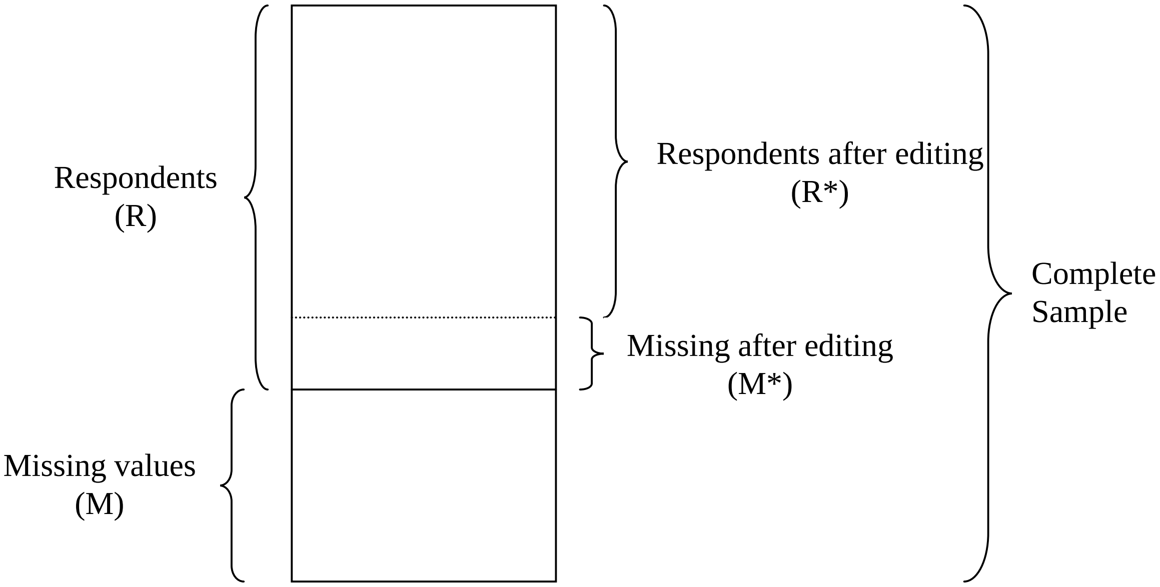
5. A number of methods have been proposed to measure the impact of editing in surveys. Using estimates computed on edited and non-edited data, Madsen and Solheim (2000), defined editing bias measures. Also interested in bias, Barcaroli and D’Aurizio (1997) and Manzari and Della Rocca (1999) used simulations to produce “true” values that could then be used to compute a number of indexes they defined. In Nordbotten (1999), ratios are defined to contrast frequencies of rejected and imputed data to their associated costs.

6. In this paper, we are interested in using the theory of estimation of the variance under imputation (Särndal, 1992; Lee, Rancourt and Särndal, 2002) to produce an editing variance component. A similar idea was used in the metrics proposed by Nordbotten (1997) to evaluate the quality of editing and imputation by contrasting estimates based on edited and unedited values. The next section presents and defines editing as a statistical process. It is followed by Section III with a description of how to measure the impact of editing using variance components. Then, some recent developments at Statistics Canada in this field are highlighted in Section IV and the conclusion follows.

II. EDITING AS A STATISTICAL PROCESS

7. There are many causes for nonresponse. From an estimation perspective, data that are lost, deleted, misreported or corrupted may all have an impact. Therefore, editing can be seen as a process that causes missing data and needs to be followed up and monitored like any other cause whether it is manual or automated.

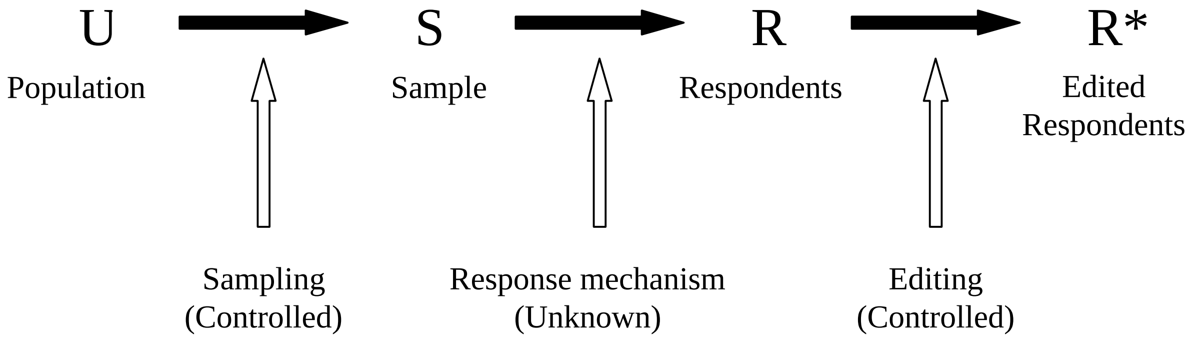
8. Conceptually, the set of nonrespondents in a sample can be divided into those that have missing data as a result of editing, and those that have missing data as a result of other causes. Obviously, this breakdown is not unique and is not specific to editing. Any reason for missing data could be isolated in the same way but here the attention is on editing. In the following, the discussion is assuming total nonresponse, but it could also apply to partial nonresponse. Graphically, the situation can be pictured as follows:



9. The problem with editing is to characterize it as a *statistical* process. With manual editing, the argument can be made that when reproduced, it may lead to different outcomes. In this case the errors could be assumed to behave like measurement errors and the associated theory could be used. However, with automated editing (or editing in general), the process is not necessarily stochastic but rather fixed and so, errors cannot be treated like measurement errors. In fact, it could be assumed that there are potential measurement errors and in this case, even with a fixed set of edit rules the editing process would not really be controlled. But in this paper, it is assumed that measurement errors are absent and that editing is responsible for missing data.

10. Editing can be considered as part of the nonresponse mechanism. This is often the implicit assumption when dealing with the problem of missing data. For example, the re-weighting or the imputation process is most of the time performed on all missing data without regard to the cause of missingness. However, the editing process is completely distinct from the actual response mechanism and should be studied accordingly. Then, the problem remains one of representing the editing process. Since it is fixed, it cannot really be represented by a model.

11. Editing does have an impact on the data. It cannot be modelled, but it affects which part of the data distribution can be observed. In surveys, the sequence of events can be represented as follows:



12. As can be seen, unlike the response mechanism, editing is a process under the survey statistician’s control. For instance, with query edits, the rules are not absolute and the editor may

be causing data to be missing because of edits rules that are much too tight. It is then clear that editing is not a correction tool, as the more editing is performed, the less responses are available. In the case of a census with a 100% response rate, editing becomes the only cause of missing data. One must therefore strive not to overuse editing for correction, but rather to use it to learn about the data in order to minimise other errors.

III. MEASURING THE IMPACT OF EDITING

13. As concluded in Granquist and Kovar (1997): "Editing will have to serve a greater function than just a data correction tool". Its goal is to improve the overall quality. The quality of individual values that have been treated for nonresponse is only secondary. It is with this view in mind that this section presents a tool to measure the impact of nonresponse.

14. Just like a quality control strategy in a production line, attempting to measure the quality of a process such as editing has two aspects. Firstly, the quality of the level (of estimates) must be assessed. This is why the literature has been abundant on the subject. For example, it is necessary to count how many records fail a given edit or a given set of edits. Secondly, the stability of the process must also be assessed. In this case, the literature on the subject is rather limited. There are measures of editing but not in the sense of variability of the estimates. To obtain a complete picture of the quality of editing, both types of measures (level and stability) must be used together.

15. Measuring the variability of processes such as nonresponse and editing cannot be achieved directly: A treatment must be applied. This treatment may be re-weighting and/or imputation. Therefore, the variability due to nonresponse (or editing) needs to be measured using the variability due to imputation or due to re-weighting. In this paper, mainly imputation will be considered.

16. Since an editing model cannot be used, a *data* model (ratio) such as

$$\xi : y_k = \beta\, x_k + \varepsilon_k ,\quad \mathrm{E}_\xi(\varepsilon_k) = 0,\ \mathrm{E}_\xi(\varepsilon_k \varepsilon_{k'}) = 0,\quad \mathrm{E}_\xi(\varepsilon_k^2) = \sigma^2 x_k$$

may be considered. More general models may be used, but this simplified version is sufficient for the current discussion. Under a data model, the editing process can be fixed, since the variables of interest are random. The model selected is the one that is found to best represent the data. It is important to note that the selection of a model requires the usual assessment of the fit and is not obtained as the result of a simple choice.

17. When imputation is used, models of the form of ξ are useful to represent the data. In Särndal (1992), a model is used to develop a framework to evaluate the precision of estimates under imputation. In this case, the variance of the estimator is represented as the sum of the sampling and the nonresponse (imputation) variance. In the context of editing, the variance could be further divided into components due to sampling, nonresponse and editing.

18. In the following, the imputation variance calculation approach is applied to editing. To avoid the need for evaluation of the mean squared error, estimates are assumed to be unbiased. This assumption lies on the premise that everything possible would be attempted by the survey statistician to avoid errors that may cause some bias. Of course, some bias may remain but its measure is not the focus of this paper.

19. Assume that the parameter of interest is the population total, $Y_U = \sum_U y_k$, where y_k is the variable of interest for unit k. The theoretical estimator that would be used under 100% response is called $\hat{Y}_{\text{FULL}}$. The estimator under no editing is $\hat{Y}_{\text{NO-EDIT}}$ and the actual one with editing is $\hat{Y}_{\text{EDIT}}$. Respectively, they are:

$$\hat{Y}_{\text{FULL}} = \sum_S w_k y_k,$$

$$\hat{Y}_{\text{NO-EDIT}} = \sum_R w_k y_k + \sum_M w_k \hat{y}_k,$$

$$\hat{Y}_{\text{EDIT}} = \sum_{R^*} w_k y_k + \sum_{M^*} w_k \hat{y}_k + \sum_M w_k \hat{y}_k,$$

where w_k is the final weight and $\hat{y}_k$ is a value imputed by a given method. Note that in this case, it is assumed that the same imputation model is used for values missing as a result of nonresponse or as a result of editing. As well, this expression does note make a distinction between records that fail an edit as a result of measurement error and those that fail strictly as a result of editing. The total error will reflect both causes.

20. The total error of $\hat{Y}_{\text{EDIT}}$ is

$$\hat{Y}_{\text{EDIT}} - Y_U = (\hat{Y}_{\text{EDIT}} - \hat{Y}_{\text{NO-EDIT}}) + (\hat{Y}_{\text{NO-EDIT}} - \hat{Y}_{\text{FULL}}) + (\hat{Y}_{\text{FULL}} - Y_U).$$

Turning to variance (and assuming no bias),

$$V_\xi(\hat{Y}_{\text{EDIT}}) = E_\xi(\hat{Y}_{\text{EDIT}} - Y_U)^2 = V_{\text{Editing}} + V_{\text{Nonresponse}} + V_{\text{Sampling}} + \text{ Small mix terms.}$$

21. For example, with ratio imputation and assuming that only units from R* are used to impute in both M^* and M, estimation is as follows:

$\hat{V}_{\text{Sampling}}$: As usual; or preferably computed on values imputed with residuals,

$$\hat{V}_{\text{Nonresponse}} = \left[\frac{(\sum_M w_k x_k)^2}{\sum_{R^*} x_k} + \sum_M w_k^2 x_k\right]\hat{\sigma}^2,$$

$$\hat{V}_{\text{Editing}} = \left[\frac{(\sum_{M^*} w_k x_k)^2}{\sum_{R^*} x_k} + \sum_{M^*} w_k^2 x_k\right]\hat{\sigma}^2,$$

where x_k is the auxiliary variable, $\hat{\sigma}^2 = \dfrac{\sum_{R^*}(y_k - \hat{B}x_k)^2}{\sum_{R^*} x_k}$ and $\hat{B} = \dfrac{\sum_{R^*} y_k}{\sum_{R^*} x_k}$.

These formulae are in fact the same ones that apply to estimation of the variance due to nonresponse when there is imputation. For example, forms for many imputation methods can be found in Lee, Rancourt and Särndal (1995) and translated into this editing framework.

22. With tools such as $\hat{V}_{\text{Editing}}$ and $\hat{V}_{\text{Nonresponse}}$, one can study various impacts of editing. For a given set of respondents, a number of editing strategies can be evaluated. When developing editing, it would be possible to choose among the approaches believed to be unbiased, the one with the smallest estimated variance. Another application is in repeated surveys where monitoring of the editing process could be enhanced by adding $\hat{V}_{\text{Editing}}$ to the set of edit counts and rates that are usually looked at.

23. There are several methods to take imputation into account. The model approach is favoured for extension to editing because it allows for production of estimates of each of the components of the total variance. Other methods such as the jackknife and the bootstrap exist for estimating the variance due to nonresponse and they could be used to estimate the total variance. However, the possibility of having variance components ($\hat{V}_{\text{Editing}}$, $\hat{V}_{\text{Nonresponse}}$) is not always an option for these approaches.

IV. RECENT DEVELOPMENTS AT STATISTICS CANADA

24. Statistics Canada has produced quality guidelines (Statistics Canada, 1998) for all steps and processes of surveys. Therein, the multiple goals of editing are clearly stated. For instance, the guidelines say: “the editing process is often very complex. When editing is under the agency’s control, make available detailed and up to date procedures with appropriate training to staff involved, and monitor the work itself. Consider using formal quality control procedures”. The importance of measuring the impact of editing is therefore recognised. Further, it evokes the idea that editing is a process that can be submitted to quality control methods.

25. There has also had a Policy on informing users of data quality and methodology since the late seventies (Statistics Canada, 2001). To improve the information provided to users, large efforts have been made since 1990. A large part of it has been on estimation of the variance under imputation. As presented in Lee, Rancourt and Särndal (2000), there is now a wide array of methods to correctly account for (single) imputation variance in estimation of the total variance.

26. As was seen in Section III, the methods designed to estimate the variance due to imputation can also be used to estimate the variance due to editing. Moreover, recent developments have been made in the creation of tools and software to evaluate the impact of nonresponse and imputation and the quality of estimates.

27. The first tool is a generalised simulation imputation system called GENESIS. This system allows for evaluation of imputation techniques under a simulation framework. Various imputation methods can be compared for a given population in a controlled simulation environment. Currently, a beta version can be used but it is not supported. While it is designed for imputation, GENESIS could also be used to investigate editing. Its simulation-based nature makes it a tool for investigation of the bias. At this time it is designed for evaluating the variance due to imputation, but it will eventually be extended to editing.

28. The second tool is a system designed to estimate the variance under nonresponse and imputation. The system, called SEVANI, can produce nonresponse and/or imputation variance components. It can produce up to three estimated variance components for processes such as re-weighting for nonresponse, imputation or editing. It will therefore allow for the production of variance estimates that more accurately represent the sampling and nonresponse variance. It should be possible to use such a system to produce measures such as $\hat{V}_{\text{Editing}}$ presented in Section III. A beta version of the system is currently being evaluated.

29. Editing is an activity that is an integral part of the process in many surveys at Statistics Canada. This is acknowledged in a survey on editing practices by Gagnon, Gough and Yeo (1994). Like re-weighting and imputation, it does have an effect on estimation. Thanks to the increased awareness of this impact on final results, survey methodologists have started to consider the use of variance components to evaluate the quality of estimates. On top of the sampling variance, survey methodologists have started estimating the imputation variance and there are plans to evaluate the variance due to editing and imputation in a number of business, social and household surveys.

30. The Canadian Labour Force Survey is at the start of a redesign and treatment of missing and inconsistent data is at the top of the agenda. Activities that are considered in the redesign are the treatment of total nonresponse through re-weighting and treatment of partial nonresponse through imputation. As well, cross-sectional and longitudinal editing are to be re-assessed and further developed. To select the approach that will be used in production, GENESIS and SEVANI will be used to produce comparative measures of performance of the editing and imputation as well as for re-weighting for total nonresponse. Then, when the redesigned survey is in place, SEVANI will be used on a continuous basis to monitor the impact ($\hat{V}_{\text{Editing}}$) of editing on a monthly basis.

V. CONCLUSIONS

31. Editing is an important process in surveys. It can be viewed as one of the causes of nonresponse but unlike other response mechanisms, it is fixed. It does nonetheless have an impact on final estimates, which should be measured.

32. This paper has used the imputation variance estimation theory to apply it to editing. By assuming a model for the data, it is possible to use a model approach to obtain an estimate an editing variance component. As well, this information should help in better understanding the editing process. Further, adding the editing variance to the estimated variance allows for analysts to make more precise inference and for data users to be better informed of data quality.

33. The survey manager who has at his disposal measures of sampling, imputation and editing variance is well equipped to make decisions on how to allocate or re-direct funds in the survey. For instance, the following variance estimates may be obtained (as a percent of the total estimated variance).

Variance component	**Scenario 1**	**Scenario 2**	**Scenario 3**
Sampling	80 %	40 %	40 %
Nonresponse	10 %	50 %	20 %
Editing	10 %	10 %	40 %

34. Under Scenario 1, most of the variance is due to sampling. This may indicate that a larger sample could be needed. Conversely, nonresponse and editing appear to have a small impact on the variance. If editing and treatment of nonresponse require large amounts of funds and efforts, then these may be reduced. Under Scenarios 2 and 3 sampling is no longer the dominant factor contributing the overall variance. In Scenario 2, editing still has a minimal impact while it is significant in Scenario 3. In this case, the large figure (40%) may indicate that there is over-editing, that there are many errors in the data or that collection procedures need refinements. Also, since 60% of the total variance is due to editing and nonresponse, perhaps the sample size can be reduced to re-direct funds towards improvement of follow-up and treatment of nonresponse procedures.

35. Of course, bias is also a component of the error that cannot be neglected. The present paper does not present an approach aimed at replacing other methods to evaluate editing. It aims at providing a complementary measure in order to help in fully understanding the editing process. Since editing can be viewed as a process, its stability must be evaluated and monitored. Using the $\hat{V}_{\text{Editing}}$ component presented in Section III appears to be a tool of choice for this purpose.

36. Hopefully, the approach in this paper will help survey statisticians developing new editing strategies or more precisely monitor those in place.

References

BARCAROLI G., D'AURIZIO L. – Evaluating Editing Procedures: The Simulation Approach, *Working Paper No. 17,* Conference of European Statisticians, UN\ECE Work Session on Statistical Data Editing, Prague, 1997.

GAGNON F., GOUGH H. and YEO D. - Survey of Editing Practices in Statistics Canada. Statistics Canada Technical Report, 1994.

GRANQUIST L. and KOVAR, J. – Editing of Survey Data: How Much is Enough?. *Survey Measurement and Process Quality,* Lyberg, L et al eds., J. Wiley and Sons, New York, 1997.

LEE H., RANCOURT E., SÄRNDAL C.-E. – Variance Estimation in the Presence of Imputed Data for the Generalised Estimation System, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389, 1995.

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, 2002.

MADSEN B., SOLHEIM L. – How to Measure the Effect of Data Editing, *Working Paper No. 2*, Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing, Cardiff, 2000.

MANZARI A., DELLA ROCCA G. – A Generalised System Based on a Simulation Approach to Test the Quality of Editing and Imputation Procedures, *Working Paper No. 4,* Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing, Rome, 1999.

NORDBOTTEN S. – Metrics for Predicting the Quality of Editing and Imputation, *Working Paper No. 20,* Conference of European Statisticians, UN\ECE Work Session on Statistical Data Editing, Prague, 1997.

NORDBOTTEN S. – Strategies for Improving Statistical Quality, *Working Paper No. 4*, Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing, Rome, 1999.

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

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.