

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UN/ECE Work Session on Statistical Data Editing**  
(Cardiff, United Kingdom, 18-20 October 2000)

**COMMENTS ON “EVALUATING EFFICIENCY OF STATISTICAL  
DATA EDITING: GENERAL FRAMEWORK”**

Submitted by Statistics Canada<sup>1</sup>

**Contributed paper**

**I. INTRODUCTION**

1. At the 1999 plenary session of the Conference of European Statisticians, a general framework on statistical data editing was discussed and resulted in the current version of a document prepared by Professor Svein Nordbotten (2000). The goal of the present paper is to describe how the framework described by Nordbotten fits into the Canadian context of data editing. This covers some experiments already done at Statistics Canada, with some views and recommendations.

**II. THE MARKET PERSPECTIVE**

2. The framework described by Nordbotten gives details on a market perspective that statistical agencies may adopt. Among other things, it mentions that the targeted level of quality should optimize the gain for the agency, i.e., the difference between the value and the production cost of a statistical product. Although this approach makes sense, it is difficult to apply in the monopoly context of Statistics Canada (STC). Like at most national statistical agencies, STC statistical products are often developed for one main user, usually another government department that needs information to monitor its programs. Then, specific requirements have to be satisfied, and the resulting products are offered to other users only after the fact.

3. Nevertheless, the three indicators mentioned in the paper, relevance, timeliness and accuracy, are used by STC as quality components. Three other indicators are also taken into account when assessing quality even if not directly related to the editing process: accessibility, interpretability and coherence (STC, 1998a). Technically, most of these components can be evaluated by the users, with the exception of accuracy which is only measured by the statistical agency. STC realizes that consistent quality across products contributes in maintaining its reputation and this makes its products more desirable.

4. For many reasons, no real effort has been made to identify a value function based on the six components, nor to linearize the six components into one unique indicator. Since users have different priorities, it is very difficult to identify a unique model for the value function. Indeed, scoring the quality components would depend on individual users.

**III. STATISTICS CANADA INITIATIVES**

5. Many initiatives were undertaken at STC to address quality issues. The quality guidelines (STC, 1998a) and a quality assurance framework (STC, 1997) were developed for that purpose. The

---

<sup>1</sup> Prepared by Claude Poirier.

policy for reporting quality (STC, 1992) and the guidelines for reporting non-response (STC, 1993) are other examples. In the latter case, the recommendation is not to produce rates but a series of counts like total number of units, out-of-scopes, deaths, misclassifications, total refusals, partial refusals, etc. This allows the user to derive rates specific for their own requirements, with no misunderstandings or conceptual discrepancies.

6. A quality secretariat was recently put in place at STC to deal with various issues. It does not exclusively target the editing process but will obviously impact on it. Its mandate is to promote sound quality management practices, to design studies, to assist programs and to support performance reports on quality. Since timeliness was identified as one of STC's weaker points by the Economists, it is one of the priorities for the secretariat. Timeliness issues are: its reporting, its comparison among similar surveys and across countries, its improvement over the years, and the identification of best practices. Meanwhile, resources are being put into the evaluation of selective editing to improve timeliness.

7. The STC committee on imputation practices is another internal initiative to improve quality. Its mandate is to provide a forum for exchange on current practices through presentations and documentation. In identifying the best practices, listing references, encouraging presentations and maintaining an internal web site, it contributes to help survey developers in choosing strategies based on colleagues' experiences with similar surveys. This directly impacts on the timeliness and the accuracy of the resulting product, and helps to reduce the survey costs.

8. The coherence of concepts and definitions for social surveys (STC, 1998b) is another aspect of quality on which resources are spent. Again, this is closely related to the editing process since cross-validation of survey estimates is desirable.

#### **IV. WHAT ARE THE MOTIVATIONS FOR STATISTICAL EDITING?**

9. As mentioned by Nordbotten, statistical editing should not only be used to identify and fix errors in data but also to evaluate other survey processes. At STC, editing results are used to review the design of questionnaires, the frame content, the sampling design, etc. Subject matter officers, with their survey management responsibilities, make the bridge between members of multidisciplinary teams. They regularly require changes in various processes based on editing diagnostics.

#### **V. MEASURING STATISTICAL QUALITY**

10. The paper makes distinctions between *output oriented* and *input oriented* approaches. It describes the output approach as the deviation of the resulting product size from the target size. To quantify such a difference, full and rigorous editing on subsets of units is proposed, assuming this would lead to a set of real data. In practice, we think that edited data, as good as the editing process may be, can be close to but not exactly equal to the real values. Then, the quality predictor derived from the regular editing and the full editing processes gives a good indication of the accuracy, but it can be negatively biased.

11. As for the input oriented approaches, the various editing and imputation indicators listed in the paper are regularly used at STC as diagnostic tools. To be meaningful, they have to be used together rather than individually. We noticed that comparisons with their historical values help in detecting problems in the process.

12. For STC, an important aspect of quality measurement that is not described in the output nor the input approaches is the uncertainty of the measurement process, the editing and the imputation actions. To address this issue, several studies on measurement errors were done, or are being done. We know that the timing of the interview, not only within the season but also within the week and the day affects the

responses. On top of that, concepts are sometimes difficult to apply. For instance, an agriculture survey which distinguishes seeding intentions and seeding reality has problems when respondents seed their land before being asked for intentions. This problem is observed in Canada, a wide country with various climatic conditions.

13. An approach that was considered in order to reduce the variability of the editing and imputation process was to involve respondents in the training of editing staff or directly in the production phases.

14. Finally, the measurement of the overall variability is important at STC. A lot of resources are being spent on development of the theory for calculating the variance due to imputation. The resulting system, SIMPVAR, will process four main imputation methods by the Summer 2001 (ratio, mean, hotdeck and nearest neighbor imputation). Target users include major surveys like the Labour Force Survey and the Unified Enterprises Survey. Gagnon et al. (1997) present some aspects of the methodology.

## **VI. ANALYSIS**

15. Section 5 of Nordbotten's document presents a causal model where the Demands, the Resources and the Architecture drive the Performance, the Costs and the Quality. In the STC context, such a model is difficult to apply because users' demands include a compromise of cost and quality requirements. This would correspond to a cyclical version of Nordbotten's causal model.

## **VII. FURTHER RESEARCH**

16. As illustrated by Nordbotten, it is understood that research on editing has to take into account some related components: Production systems, users' systems, development, applications, other processes. Some developments on the description of editing processes are observed through an evolving Information Meta-Database. Currently at STC, as mentioned earlier in this paper, a technical challenge is to make sure over-editing is minimized, maybe with selective editing. This would improve timeliness and costs. On the evaluation aspect, some interest exists in the functional analysis of edit and imputation systems. On the other hand, internal resources are regularly assigned in the generation of synthetic data in order to evaluate methods. Like Nordbotten, we realize that synthetic data provide more control in the generation and evaluation of errors. Finally, some interest in the development of an Internet knowledge base has been noted. This would encourage users to share expertise through an international database.

17. As for research projects in a market perspective, the STC monopoly situation unfortunately does not provide the motivation to investigate such aspects.

## **References**

- Rancourt, E., Gagnon, F., Lee, H., Provost, M. and Särndal, C.E. (1997). "Estimation of Variance in Presence of Imputation", Proceedings of the 1997 Symposium, Statistics Canada, 273-277.
- Nordbotten, S. (2000). "Evaluating Efficiency of Statistical Data Editing: General Framework", Methodological material for the Conference of European Statisticians, UN/ECE.
- Statistics Canada (1992). "Policy on Informing Users of Data Quality and Methodology", Policy Manual, 2.3
- Statistics Canada (1993). "Standards and Guidelines for Reporting of Nonresponse Rates", Technical report, Statistics Canada.
- Statistics Canada (1997). "An Outline of Statistics Canada's Quality Assurance Framework", Technical report, Methods and Standards Committee, Statistics Canada.

Statistics Canada (1998a). "Quality Guidelines", Third edition, Ottawa.

Statistics Canada (1998b). "Report on Progress on the Harmonization of Social Statistics", Presented at the Work Session on Statistical Metadata, Conference of European Statisticians, Statistical Commission and Economic Commission for Europe, February 1998.