

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 3
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 3 of the provisional agenda

CANADA: NATIONAL REPORT

Submitted by Statistics Canada¹

¹ Prepared by John G. Kovar.

I. INTRODUCTION

1. Activities related to data editing and imputation are progressing on several fronts at Statistics Canada. First, in order to further automate the editing efforts, extensive use of the agency's Generalized Edit and Imputation System (GEIS) is being made in many areas. Of particular note is its implementation in the complex settings of the Family Expenditure Survey (FAMEX) on the one hand, and surveys related to the services industry on the other. While in the case of FAMEX the size and complexity of the survey are the primary obstacles, in the case of services surveys, it is the large number of applications and the relatively tight time frame that pose the greatest challenge. Secondly, the use of graphical techniques to detect values which contribute to the change of estimates is gaining in prominence. An interesting and innovative application for the Monthly Survey of Manufactures is described below. Thirdly, with respect to imputation, a full report on the use of the New Imputation Methodology (NIM) as applied to the Canadian Census of Population is provided under item 5 of the agenda for this session.

II. EXPERIENCES IN THE IMPLEMENTATION OF THE GENERALIZED EDIT AND IMPUTATION SYSTEM FOR THE FAMILY EXPENDITURE SURVEY

2. FAMEX provides comprehensive information on the income and expenditures of families across Canada. It is used primarily in the creation of the fixed basket of goods for the Consumer Price Index. Over 16,000 households are included in the sample resulting in approximately 13,000 "usable" questionnaires. Information is collected at both the family and individual level for different sections of the questionnaire. The average interview lasts approximately 3 to 4 hours. Over 1600 variables, with a potential of 200 additional variables for multiple occurrences, are collected. The result is a record just short of 15,000 bytes per household.

3. Editing and imputation had previously been accomplished through manual intervention of subject matter specialists. In order to achieve a systematic and reproducible approach to edit and imputation, a decision was made to implement GEIS. Further benefits of GEIS include reduced development time, as well as easy adjustment of edit and imputation procedures at later stages of the project.

4. Although manual editing has been reduced, an increase of pre-and post-processing of the data is necessary in order to use GEIS. Unusual or influential records are pre-selected for manual edit and imputation, while the automated process is directed towards the cleaning of typical errors. Use of the error localization module is being made for sections with more complex edits, but this time consuming module can be avoided for sections with only simple positivity edits which essentially only identify missing values. In all cases, the done imputation module is used exclusively and effectively.

5. One of the most problematic sections of the questionnaire is the mortgage section where a strategy of derived variables was implemented to eliminate excessive, as well as redundant, information. Furthermore, the hierarchical nature of the questionnaire posed further problems which had to be resolved.

6. The analysis tool of choice at Statistics Canada is SAS, used either on a PC or on the mainframe. However, to use GEIS, in addition to learning to use the product itself, it is necessary to know some SQL and be able to navigate in UNIX - relatively scarce skills at Statistics Canada, outside of the informatics areas. Clearly a modular approach implemented in SAS would have been preferable.

7. None the less, the development of the edit and imputation strategy and implementation has started in December, 1996, with the first GEIS production runs scheduled for June, 1997. It is clear that a customized system could have never been developed and tested in such a short period of time.

III. DEVELOPMENT OF A GENERIC EDIT AND IMPUTATION SYSTEM FOR SURVEYS IN THE SERVICES INDUSTRY

8. Services Division is currently redesigning its survey program in order to produce more reliable and detailed provincial economic statistics as part of Statistics Canada's efforts in this endeavour. The survey program for Services Division is now made up of over 20 annual surveys (many of them new) each focusing on a specific area of the industry. The survey sizes vary from 500 to 3000 units and an average of 250 variables are collected.

9. Methodologically sound methods for sampling, imputation and estimation are being introduced to replace judgemental non-probability samples and manual edit and imputation procedures. Because of the large number of applications to be developed in a very short period of time, Services Division has opted for an edit and imputation system that will feature common but distinct processes for different sectors. It is believed that this will avoid the complexity and manageability problems of a single large system and allow for survey specific requirements to be met when required. Also, because of time constraints, it appears more efficient to develop several small applications with similar underlying structures.

10. GEIS was chosen as the central engine for the edit and imputation system for many reasons. It provides methodologically sound and reproducible methods for edit and imputation, reduces development time and costs over time, ensures consistency among similar surveys, provides process statistics for each step and is maintained and improved by a dedicated group. On the down side, concerns with respect to the system requirements and the limitation in the type of edits that can be specified had to be addressed.

11. A functional prototype is currently being developed for one of the many Services surveys. The plan is to develop a model for future edit and imputation applications to be developed in large part by the client division. The application will be developed in a generic and modular fashion and each step will be fully documented in order to facilitate implementation for other Services Division surveys. The edit and imputation will be complemented by a comprehensive data validation process that will put emphasis on selective and aggregate macro editing. Final results and assessment of the prototype are expected in late July.

IV. DATA VISUALIZATION IN A SURVEY TAKING ENVIRONMENT

12. Three dimensional visualization is one of the most exciting technologies to emerge in the last few years. Statistics Canada has recently added a state-of-the-art, 3-D visualization tool to its redesigned Monthly Survey of Manufactures - a key indicator of the business cycle. Code named CAVEAT (Computer Assisted Validation, Editing and Analysis Tool), it is to be the centrepiece of a system designed to ensure that published estimates meet certain minimum quality standards.

13. The Monthly Survey of Manufactures is currently undergoing a redesign of systems and methodology. The survey collects data on shipments, inventories, and orders on a monthly basis from a sample of 12,000 manufacturers. These data are published by standard industrial classification and province, and are seasonally adjusted. One of the desired capabilities of the new survey system is the ability to perform accurate and timely analysis. To this end, both current and historical data should be readily accessible, using a concise graphical representation of the data's inherent complexities. A graphical representation would allow the subject matter specialist to quickly and easily determine the causes behind unusual changes in estimates by an immediate visual recognition of anomalous or outlying data values. The redesigned survey requires a more effective quality assurance system: users need to put more emphasis on critical elements of the sample, and need a tool that would emphasize the link between sampled responses and the estimates derived from them.

14. CAVEAT meets these needs by integrating survey estimates, sample responses, and other critical indicators into a fully interactive 3-D landscape. CAVEAT gives its users a clearer understanding of their data, greater success in identifying erroneous responses, and an improved ability to interpret estimates. The micro data is to be drawn from an existing database (MS Access), the estimates from existing SAS libraries, and other indicators possibly from the agency's CANSIM data base. Changes to micro data can be made through the same interface by authorized officers. Drilling down capabilities are supported at the standard industrial classification by province level for all variables on both the raw and the seasonally adjusted series. The data of the top ten contributors for the past number of months are displayed to start with, but the possibility of moving deeper in these two dimensions is to be incorporated. The prototype system exists today; a fully functional production system is to be in place by mid 1998.