WP. 46
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (viii): Selective and macro editing

SELECTIVE EDITING FOR BUSINESS SURVEYS AT STATISTICS CANADA

Supporting Paper

Prepared by Lucie Cloutier, Statistics Canada

I. INTRODUCTION

1. The Unified Enterprise Survey (UES) was initiated in 1997 as part of the Project to Improve Provincial Economic Statistics (PIPES). The UES adopted an integrated framework for conducting annual business survey programs. Key to the success of the project was the application of a set of design principles (i.e. use of a single frame, common sampling and processing methodologies) to all UES surveys. These broad-based principles have been essential to ensure that the data produced from the program meet the needs of the Canadian System of National Accounts (CSNA). However, while these principles have been applied to individual survey programs as they entered the UES, their application has been slightly different. The surveys measuring distributive trades (i.e. wholesale and retail) have a collection vehicle that minimizes the data collected, increasing the modelling of other variables required by the CSNA. The approach of the Annual Survey of Manufactures is to use one generic questionnaire, with commodity sections customized for each industry, Finally, the Service Industries survey model is to adopt a modular approach where one generic module and other specific characteristic modules can be "turned on and off" for different survey cycles. Also, there are variants of the above with differing ways to handle commodities, collection of location level data and the use of tax data.

2. After 12 years, it is now necessary to redesign the UES in order to achieve greater harmonization and to simplify processes. Going beyond simply a systems renewal, the UES program will take this redesign as an occasion to reconsider changes to the following: conceptual issues, sample design and use of tax data; alternate data collection strategies; and the UES data model and its processing systems. It is believed that one way to improve efficiency in the redesign of the UES is to incorporate a rational and systematic procedure to edit collected data. The Common Editing Strategy (CES) is used in many countries. Statistics Canada analyzed the Australian model and believes that it can be adapted to our environment.

II. THE COMMON EDITING STRATEGY FOR UES

A. Current practices

3. In the current environment, subject-matter officers edit micro-level records, prior to estimation in order to identify and resolve errors in captured data records before any aggregation of data is done. In the UES, the three main user program areas (manufacturing, service industries and distributive trades) use different automated methods to run data through a set of edits which identify data considered to be either erroneous or questionable. Some subject matter officers then attempt to correct most errors. This strategy

is very costly and time-consuming. Many studies have found that there is a tendency to do too much micro editing[1]. However, it is very hard for subject matter officers to let go of this process since they have a sense of ownership and responsibility regarding the estimates produced with the micro data. It is the common belief that micro data without any errors will produce excellent macro estimates.

4. In March 2007, Yeung (2007) looked at the amount of manual intervention being performed in the UES. This analysis looked at five surveys: Food Services, Primary Metals, Retail Chains, Retail Stores and Wholesale. The table below shows the prevalence of manual imputation (for RY 2004).

Survey	Total No. of units	Units that were changed at least once	Units changed at post-collection but before E&I	Units changed after E&I
Food Services	3440	1383 (40%)	1091 (32%)	416 (12%)
Primary Metals	2924	732 (25%)	514 (18%)	331 (11%)
Retail Chains	1337	1020 (76%)	674 (50%)	664 (50%)
Retail Stores	8126	2438 (30%)	1141 (14%)	1551 (19%)
Wholesale	6112	1538 (25%)	780 (13%)	858 (14%)

5. This table indicated that there is a high degree of manual intervention, particularly in the retail chain survey. In addition, a good number of these units are changed more than once.

Survey	Total No. of units	Units changed both before and after E&I	Units changed both before and after E&I by the same analyst
Food Services	3440	124	76
Primary Metals	2924	113	44
Retail Chains	1337	318	108
Retail Stores	8126	254	108
Wholesale	6112	100	45

6. These tables indicate that there is quite a bit of manual intervention being undertaken by subject matter officers. The adoption of a common strategy and process requires a shift in thinking, from one where the objective is to produce the most accurate data possible, to one where information is determined to be fit-for-use.

B. How the CES can improve the process

7. To improve the timeliness and to harmonize to processes, we evaluated the Australian model for the Annual Integrated Collection (AIC) core editing strategy. They have taken a “big-picture” approach to managing annual business surveys by having a coordinated management and planning phase and by integrating processes.

[1] See References for titles

8. For the UES redesign, Statistics Canada would like to propose some variation of this model to better suit our survey environment. A core financial module will be developed and used by all annual business surveys that will harmonize concepts and standards with the sub-annual surveys. A modular approach to questionnaire design will be developed to be able to turn on / off modules depending on the demand for data, resources and priorities. The core financial variables will be prioritized based on input from the CSNA and external users of our data (other departments, universities, and associations). Once this is determined, a standardized approach can be developed with subject matter divisions to harmonize data processing from collection to dissemination as well as use of tax data. The next section will detail the proposed prioritization of core financial variables and the standardized approach to data processing.

C. Prioritization

9. Requirements from the CSNA have not changed since the inception of PIPES in 1997. The CSNA needs to produce provincial and territorial estimates for 300 industries and 727 commodities. Priorities have been set:

- Priority 1 statistics are **essential** for Input and Output (IO) tables of CSNA. The elements identified as Priority 1 statistics are the principal statistics required to estimate Gross Domestic Products and the tax bases required for Harmonized Sale Tax revenue allocation. This data is required by province within 15 months of the reference period for surveyed industries as well as industries not currently covered by a survey program.
- Priority 2 statistics are sub-aggregates of the principal statistics. These elements are not assigned as high a priority as the principal statistics but are nevertheless important for IO since they are used to determine the output and use of commodities by industry. This data is required by province.
- Priority 3 statistics are the detailed revenue and expense items on Services Industry Division surveys. With the exception of the detailed items identified as principal statistics, CSNA is prepared to accept data in the form of percentages of the appropriate sub-totals. CSNA is also prepared to accept survey details on a less frequent basis (every two years for example) depending on the relative importance particular industries within the HST revenue allocation framework. The level of detail required by CSNA should be reviewed on a survey by survey basis.

Table 1: CSNA Requirements and Priorities

Survey variable	IOD Priority and Requirements		
	Priority 1	Priority 2	Priority 3
	Essential Data	Required but IO willing to compromise on level of detail and frequency	Required but IO willing to compromise on level of detail and frequency
Total Revenues	X		
Total Operating Revenues	X		
Sales of Goods Purchased for Resale	X		
Revenues by Type			X
Other Operating Revenues			X
Total Non-Operating Revenues	X		
Total Expenses	X		
Total Operating Expenses	X		
Depreciation and Amortization	X		
Total Wages, Salaries and Benefits	X		
Wages and Salaries			X
Employer Portion of Employee Benefits			X

Total Energy and Utility Expenses		X	
Energy and Utility Expenses by Type			X
Total Purchased Services Expenses		X	
Purchased Services Expenses by Type			X
Total Material and Supplies Expenses		X	
Materials and Supplies by Type			X
Total Non-Operating Expenses	X		
Total Opening Inventories	X		
Total Opening Inventories of Goods Purchased for Resale	X		
Total Opening Inventories of Raw Materials			X
Total Opening Inventories of Goods in Process and Finished Goods			X
Total Closing Inventories	X		
Total Closing Inventories of Goods Purchased for Resale	X		
Total Closing of Raw Materials			X
Total Closing Inventories of Goods in Process and Finished Goods			X
Sales or Operating Revenues by Type of Client			X
Distribution of Sales or Operating Revenues by Client			X

10. The CSNA will also be asked to prioritize key variables according to these criteria:

A – High quality estimates (estimates fully justified and defendable)
B – Good quality estimates (major movements explained)
C – Basic quality estimates (minimal checking)
D – Unknown – possibly poor quality (estimates may be unchecked

11. This ranking will be done by industrial classification, geography and any other aggregates they would need.

12. Once the needs of the CSNA are identified and ranked, negotiation with partner departments will take place. It is essential to the success of the UES redesign that partner departments, as well as CSNA understand the cost and the response burden associated with each additional variable on a questionnaire. This is where a high-level management body will have to play an important role in deciding the relevance of questionnaire content. At the same time, the partner departments will also be asked to rank their requests for data the same way as the CSNA did with the above criteria of quality.

13. Although the degree of quality is important to reduce the response burden and increase relevance, it is also important to determine the periodicity of variables. CSNA is ready to compromise on frequency of collection for variables, especially commodity variables, some of which may stay relatively constant over time.

14. As a result of this triage, modular questionnaires will be produced. The first essential module will contain core financial variables identified as priority 1. Other modules will be dedicated to commodities, to industry specific financial details, and special topics that are relevant at the time of the survey.

D. The CES standardized approach for data processing

15. In order to have the subject matter officers focus their attention on specific variables, generic tools and guidelines need to be developed. All subject matter officers will be able to view data from collection to dissemination, historical data, tax data, sub-annual related data and any other data files contained in an integrated business database. This element is essential to the UES redesign.

16. The standardized approach will describe the process of data from collection to dissemination as well as the process and tools that need to be used for priority 1, 2 and 3 and for quality A to D variables.

17. The planning phase will gather requirements from subject-matter divisions on developing generalized reports that will be needed to analyze the data. It will be done at the beginning of the redesign and subject matter officers will have to agree on the requirements. Problems are not expected for core financial variables. However, the remaining modules will have to be negotiated with each subject matter involved in the process and get agreement on the best way to analysis the data.

18. The first phase will be the pre-grooming after collection is closed. A set of automated ratios and edits will be developed to detect outliers and failed records that subject matter officers will need to correct before the automated edit and imputation. These ratios and edit will be different depending on quality norms identified previously. This step is essential to ensure that a maximum number records without any errors or missing cells enter the automated edit and imputation from which to impute data for non-response and partial response questionnaires as well as units that will be imputed with tax records. The quality of the data going into the automated edit and imputation is important at this stage since each record could be a potential donor. An error in one donor could be replicated in many records during the imputation process. However, it is not necessary to correct all data in every detail. The main products of statistical offices are aggregated data often based on sample. Thus, a certain level of record in error is acceptable as long as the most influential errors are detected and corrected.

19. The next step in the UES processing is the automated edit and imputation, using Statistics Canada's BANFF generalized system. A set of algorithms will be available to impute the missing variables. Analysts will be consulted on the relationships that variables may have with other variables, if donor or historical imputation is an option, what kind of tax treatment is recommended, etc.

20. After the automated edit and imputation, analysts will then have to go back to treat erroneous records that could not be imputed or didn't satisfy all edits. There should not be more then a handful of records that are in this step and analysts should be able to resolve them very rapidly. Ideally, all records should pass through this step without errors.

21. Estimation will then take place and produce macro estimates. At this phase, the analysts will do analysis using pre-determined tables of ratios, outlier detection, and any other diagnostic tests that have already been established at the planning phase. A drill-down tool will be provided to analysts to correct micro records for quality A and B maybe C variables. This iterative process will have to be completed when the data will be judge as fit for use, depending on a certain level of quality predetermined by methodology.

22. Then the data will be ready to go through the confidentiality process. Ideally, this process will be performed not by the same subject matter analysts but by specialists trained specifically on confidentiality rules and programs. Presently, most subject matter divisions are using Statistics Canada's CONFID1 to perform confidentiality. It becomes very difficult to get the right people to perform confidentiality and subject matter divisions have to invest in training. It would be more efficient to have a centre for all area that would specialize in confidentiality and dissemination using the new Statistics Canada's CONFID2 program. Although specialists would perform this task, the analysts will have to clear the data and work with these specialists to provide specification of the industry.

23. Once the data is cleared for confidentiality, the dissemination steps can proceed. All tables and graphs will be prepared in the same way and will be sent using the generalized channel to Statistics Canada's CANSIM database. Subject matter officers will then be ready to release the data.

24. Processes and tools to be used by all of the subject matter officers in the redesigned UES will need to be harmonized, both for macro, as well as micro editing of data. In the first step of the UES redesign, specification for designing these processes and tools will be gathered from different subject matter areas. Then, an analysis will be performed to come up with generalized common editing processes and tools that could satisfy most of the requests. The focus will be on the generation of globally optimal solutions rather than the locally optimal solutions that have been developed to date.

III. CONCLUSION

25. It will be challenging to get an agreement on the priority variables and on the degree of scrutiny that these variables will undergo. Multiple consultations will have to be scheduled with the CSNA but also with external clients of Statistics Canada to get a consensus on the priority variables.

26. It will also be a challenge for some analysts to change paradigms. Many are convinced that micro editing of all variables is essential to produce high quality estimates. They feel responsible for the final results and they are the ones that will be asked to defend these sets of aggregates. They have been doing the job for many years and are confident in their method of analysis. However, it is important that they adhere to this concept of CES since they will have to release the data much sooner then in the past. Also, they need to understand that all data that will be released don't have to be of equal quality and a certain degree of uncertainty is acceptable for some variables that are only indicators. A good training program and a rigorous strategy of CES will have to be maintained for some years before the change is really adopted by all analysts.

REFERENCES

McCoull, Erica (2007a). "AIC Priority Setting Processes" unpublished - Australian Bureau of Statistics - internal document.

McCoull, Erica (2007b). "Principles and criteria for accepting and prioritizing topics for the AIC program" unpublished Australian Bureau of Statistics - internal document.

McDonald, Andrew (2006a). "Business Case for the Annual Integrated Collection" unpublished - Australian Bureau of Statistics - internal document.

McDonald, Andrew (2006b). "Model for the Annual Integrated Collection" unpublished – Australian Bureau of Statistics - internal document.

Peterson, Greg (2007). "Briefing Note: Australian Bureau of Statistics – AIC core editing strategy" unpublished – Statistics Canada - internal document.

Yeung, Chi Wai (2006), "Analysis on Manual Imputation" unpublished – Statistics Canada - internal document.

Brodeur, Marie; Koumanakos, Peter; Leduc, Jean; Rancourt, Éric; and Wilson, Karen (2006). "The Integrated Approach to Economic Surveys in Canada", Ottawa, Enterprise Statistics Division, Statistics Canada

Granquist, L. (1984), "Data Editing and its impact of the further processing of Statistical data" Workshop on Statistical Computing, Budapest.

Granqist, L. (1997). "The New View on Editing", International Statistical Review

Granquist, L. and Kovar J. (1997), "Editing of Survey Data: How Much is Enough?" John Wiley & Sons, Inc., New-York