STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 2
English only

CONFERENCE OF EUROPEAN STATISTICIANS

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

Topic (i): Planning and management of statistical data editing

ANALYSIS OF DATA SLICES AND METADATA TO IMPROVE SURVEY PROCESSING

Invited Paper

Submitted by Statistics Canada[1]

Abstract: Survey processing, especially components that are highly labour intensive can be expensive. It is therefore important to find ways to gain efficiencies. Managers of Statistics Canada's largest multi-sector business survey have access to versions of data and to additional processing metadata that describe how the data were transformed from collection through post-imputation correction. We are able to use this information to detect inefficient or inappropriate methods and to replace them in favour of more efficient and appropriate methods. The paper presents the findings of a study involving a few years of survey data.

I. INTRODUCTION

1. The Unified Enterprise Survey (UES), initiated at Statistics Canada (STC) for 1997 with seven industries, now integrates just under 50 annual business surveys into one centralized survey system. Businesses of all sizes are in scope for the UES. Large firms are always in sample; the smaller businesses are randomly selected each year.

2. For the first few years of the UES, most industries have employed two questionnaires – a long questionnaire that asks for all the variables of interest to the industry and in an attempt to ease response burden a shorter questionnaire directed to smaller firms. The intention was that the details for these smaller firms could be derived from their tax data.

3. Editing and/or imputation is carried out in each of the first four phases of the survey process: (1) Data collection, (2) Post-collection review and correction, (3) Automated imputation and (4) Post-imputation review and correction.

4. The data and metadata resulting from each of these phases are housed in a central data repository such that four versions of data with accompanying metadata (one for each of the four processes above) are available for analysis. In addition, for mail-back units, the raw captured data are available for research, but reside outside the central repository.

5. Only recently have we finally had the opportunity to assess the data available for the years 1997 to 2000. With this data, we wanted to:

---

[1]Prepared by Colleen Martin and Claude Poirier (poircla@statcan.ca). Acknowledgements to Jackey Mayda, Pamela Ramage-Morin and Rob Williams of STC for their research activities.

- quantify the effect that manual intervention has on the data since the costs, in terms of people and time, are high;
- determine whether these costs could be decreased by directing interventions more efficiently;
- determine whether pre-specified automated edit and imputation procedures are the most appropriate or whether they are in fact leading to a need for more manual intervention.

II. THE STUDY

A. Background to the Collection Process

6. Survey data for the UES are collected via mail-back questionnaire or by telephone. All initial contact is made via regular mail questionnaire. Follow-up for total non-response and for certain edit failure is conducted by telephone.

7. Mail-back questionnaires are captured using a Quick Data Entry system with virtually no editing. Captured data are then batch edited and categorized according to the severity of their edit failures.

8. Two slightly different follow-up strategies are applied for mail-back units that fail capture edits. For the non-manufacturing sector, questionnaires categorized as having severe edit failures are flagged for follow-up. For the manufacturing sector, "critical" units categorized as having severe edit failures are flagged for follow-up. Mail-back questionnaires having only non-severe edit failures and manufacturing "non-critical" units are not flagged for follow-up.

9. All total non-responses, i.e. units that do not mail back their questionnaire, are flagged for follow-up.

10. The main concern during this phase of processing is the follow-up cost. It is estimated that a telephone follow-up takes on average 15 minutes. This estimated time does not account for the unsuccessful follow-up attempts that often precede a final contact.

B. Findings Related to the Collection Process

Edit failure and Follow-up Rates

11. Our study concentrated on reference year 2000 and on units that mailed back their questionnaires. The path followed by mail-back units provides us with a fuller set of data and metadata. For these units, we have two versions of data - raw captured data and the data resulting from follow-up. In addition, we have the edit flags resulting from the batch edit and flags resulting from the follow-up process. Less information is recorded for units whose data are collected entirely by telephone.

12. For one specific industry, we identified the ten edits with the highest percentage of failing cases. All ten edits were severe, the category requiring follow-up. One edit was failing 27% of the time. The smallest of the ten was failing 13% of the time. Of these edits, we noted that six were query edits – not highlighting mathematical impossibilities, but rather potential unusual relationships.

13. Across all industries, we detected that:

- Less than 3% of units passed all edits;
- Units which failed any edits tended to fail severe edits;
- All units failing severe edits in the non-manufacturing industries were flagged for follow-up and all critical units failing severe edits in the manufacturing industries were flagged for follow-up;
- The rate of follow-up was lower for manufacturing industries than for non-manufacturing since manufacturing did not follow-up non-critical units;
- In some industries, 100% of units were flagged for follow-up;

- Short questionnaires were flagged for follow-up at a lower rate than long questionnaires, since they had fewer variables and therefore fewer edits.

14. As mentioned earlier, we had in addition a large percentage of the sample that did not mail back their questionnaires and who therefore required telephone follow-up also.

15. While complete follow-up was possible in the early days of UES when we did not yet have 50 industries involved, we do not have the resources to follow-up at this rate with the current volume.

16. These observations pointed us to several areas where we should immediately direct our attention, in an effort to decrease costs:

- *We needed to find a way to encourage respondents to use our mail-back questionnaire in order to minimize the cost associated with telephone data collection;*
- *We needed to re-visit our edits, paying stricter attention to what should constitute an edit follow-up, so that those units responding by mail would not be contacted by telephone simply to have their reported data confirmed;*
- *We needed to find a way to prioritize individual units for follow-up in an even stricter fashion than was currently employed for the manufacturing sector.*

Impact of Follow-up

17. By comparing the raw data with the post-follow-up data, we split changes resulting from follow-up into 3 categories to determine what is the effect of each on the data:

- Missing data becomes present through follow-up - This brought about an increase of up to 20% for the weighted variable in its industry. On average, it produced a 2% increase.
- Value (properly captured) changes through follow-up - This brought about a change of up to 4% for the weighted variable in its industry. On average, it produced a 1-% change. A high percentage of queried responses were confirmed by the respondent to be correct.
- Value (improperly captured) changes through follow-up - This brought about a change of up to 62% for the weighted variable in its industry. On average, it produced a 10% change. A very small number of changed records produced a very large actual change.

18. As expected, missing data becoming present had a large impact. However, it is the normal job of automated imputation (a post-collection process) to correct this specific situation, and as long as we could feel confident that we have good imputation methods in place, there would not be any need to concentrate on these units. On the other hand we were quite concerned about one broad area:

- *We needed to direct our attention to the problem of edit failure caused by improper capture.*

C. Background to the Post Collection Processes

19. There are 3 main steps in the post-collection process - post-collection review and correction by survey data analysts, automated imputation and post-imputation review and correction.

20. Automated imputation for UES comprises techniques for solving key variables and techniques for solving details. Three or four key variables, generally totals, are identified for each industry. Key variables are equivalent to, or a subset of, mandatory variables from the collection phase.

21. For missing key variables, the automated system applies in order:

- Derivation rules involving other reported data for the individual record;
- Previous year's ratios amongst key variables for the individual record;

- Current year ratios amongst key variables within a group of like records;
- Year over year trend for each key variable within a group of like records.

22. This set of processes ensures that we have key variables for all units that report at least one key variable, and for all units that were in the survey the previous period.

23. For units that have their key variables solved, the distribution of details for the short questionnaires was taken from each unit's tax data for reference years 1997 to 1999. For distribution of missing details for long questionnaires, donor imputation was used on a section-by-section basis using key variables to find the nearest neighbour.

24. When key variables for new-in-sample-units can not be solved using the above techniques, the record undergoes mass imputation using the data of a single donor for all sections of the questionnaire and for all variables, both key variables and details. During the 1997 to 1999 period, tax data were used to find the nearest neighbour and the ratios observed between the donor vs recipient tax data were applied to the donor's reported data to get a more tailored result for the recipient, rather than simply copying the donor's data.

25. Throughout the automated imputation process, "reported" data are very rarely changed. It must be noted however, that the automated system cannot differentiate between properly captured and improperly captured data; both are treated as "reported".

26. The primary concern during post-collection processing is the cost of review and correction conducted by the survey analysts. We learned early in the study that most analysts do little manual correction before automated edit and imputation, so we concentrated our efforts on the changes made after automated imputation. At this stage of processing we have the data out of automated imputation together with the meta data that identify where each data value came from – reported, imputed by method-A, imputed by method-B etc., and we have the data after manual intervention, with each changed value easily identified.

D. Findings Related to the Post-Collection Process

Manual Imputation Rates

27. So far, our study has concentrated on tracking the Total Operating Revenue and the Total Operating Expense, which are key variables common to all UES industries. We looked at each of the variables across the entire UES for 1998 and 1999 to determine how often and under what condition the variable had been manually changed. We wanted to see whether data imputed using one method were changed more often than data imputed by other methods. The 1998 results are shown below. Reference year 1999 findings were quite similar.

- The Total Operating Revenue variable was manually changed over 15% of the time.
  - 67% of these changes were for units that had been imputed through mass imputation;
  - 24% of the changes were changes to "reported" data;
  - 4% of the changes provided data for units that could not be imputed through the automated process and so still had missing data.

- The Total Operating Expense variable was manually changed over 21% of the time.
  - 55% of these changes were for units that had been imputed through mass imputation;
  - 35% of the changes were changes to "reported" data;
  - 5% provided data for units that could not be imputed through the automated process and so still had missing data.

These findings strongly indicated that we should be concerned with mass imputation and changes to reported data.

28. We investigated the methods surrounding automated mass imputation. As stated earlier, the imputation method depended on there being a consistent relationship between tax data and reported data for units in sample. These tax data had undergone considerable in-house processing to bring them to the level of the collection entity, which is also the level at which we edit and impute. When we looked more closely at the results, we found that there was not always a consistent relationship between tax data and reported data. For the majority of units, tax data and reported survey data were quite similar. For other cases there were huge disparities. It was in fact the inconsistency across units that was causing us problems. When a donor with great disparity was chosen for a recipient whose tax data in fact was very close to "reality", we ended up applying outlandish ratios resulting in outlandish imputed values.

29. We conducted a rather subjective survey amongst subject matter analysts to try and determine why they felt the need to change "reported" data. Two answers emerged - data that had been badly captured and the respondent had misunderstood the question.

30. The findings suggested several areas of concern:

- *We needed to find a substitute for tax data or improve the processes leading to our version of tax data so that there would be a consistent correlation between auxiliary data and survey data;*
- *We needed to address the issue of badly captured data, so that analysts could feel confident that "reported" data were truly reported;*
- *We needed to revisit the content/wording of our questionnaires, so that respondents would not misunderstand.*

Impact of Manual Imputation

31. We next proceeded to look at the changes themselves, to see if we could determine whether the change in the actual estimate was proportionate to the effort being put into micro correction. By comparing net change to absolute change we discovered quite a disparity, although it was clear that the net change had a considerable effect on the estimate:

- For the Total Operating Revenue variable in the 1998 and 1999 reference years
- Net change amounted to –4.39% and –5.29% for the overall industry;
- Absolute change was 52% and 37%.

- For the Total Operating Expense variable in the 1998 and 1999 reference years
- Net change amounted to –39.1% and –23.2% for the overall industry;
- Absolute change was 69% and 42%.

32. We ranked individual units that had been changed and made some further discoveries:

- While 15% to 21% of these variables had been changed, when we looked at what would have happened had we made only the top 50 changes, these results were within a very small margin of the final estimates;
- Estimates resulting from the top 10 changes were also quite similar, especially in some industries.

33. With this information we felt that:

- *We needed to find a way to identify fewer, large impact units that would yield the greatest improvement in the estimates.*

III. CHANGES TO OUR PROCESSES RESULTING FROM THE FINDINGS

E. Changes Currently in Effect

34. For reference year 2001, for the non-manufacturing sectors, we now have only one questionnaire per industry. The length of each new questionnaire has been greatly reduced, compared to the equivalent long questionnaire of the reference years 1997 to 2000. The questions have been reworded in an attempt to avoid misinterpretation by respondents.

- We expect to have a more successful mail-back response rate now that the respondent is faced with a much shorter questionnaire;
- The number of collection edits has been greatly reduced as a result of the shorter questionnaire;
- Follow-up rates should be lowered, with fewer edits involved.

35. For the reference year 2000, for mass imputation, we no longer use tax data to find a nearest neighbour, but have turned instead to using historical response for units previously in sample and the size measure that resides on the Business Register for new units in sample.

- The imputed values will improve, assuming we now have a tighter correlation between these auxiliary data and the reported data of the donors.

F. Further On-going or Planned Studies

36. For the manufacturing industries, identifying critical versus non-critical units succeeds in reducing the follow-up rate. A technique to prioritize all UES units for follow-up must be found, such that unimportant, easily imputed units are not given the same attention as large, difficult to impute units. The method must recognize the variation across industries and geographic region – a small unit for one industry/province is not necessarily a small unit for another. Additionally, a large unit is not always difficult to impute, especially as we introduce more appropriate imputation techniques. A method that was devised for the pilot year was never fully implemented. That method and others will be evaluated.

37. The capture edits themselves must be revisited. Query edits in particular must be reviewed and perhaps dropped altogether. As was evident in the study, most of the queries were later confirmed by the respondents. A study is now underway in the area of Data Collection to determine how best to minimize the number and type of edits that trigger follow-up. The results should be ready to be implemented for reference year 2002.

38. The data capture errors must be addressed. Originally, the capture of mail-back questionnaires was done using the same system as telephone capture, with interactive editing. For seasoned keyers this proved to be a very time-consuming process and thus the process was changed. More thought needs to be given to finding the happy medium that assists keyers in finding and correcting their own keying errors, without frustrating them.

39. We have to find a way to help subject matter officers to direct their attention to those units where change will have the most impact. While it was simple to find the top 50 units after the fact, it is difficult to find a way to identify these units before the change is actually made.

40. For post-collection imputation and manual correction, we will investigate at a more detailed level which imputation methods are resulting in values with which analysts are dissatisfied, with the intention of adapting our methods, and if necessary our systems, to bring better results.

41. There continues to be at STC a culture that insists on micro data correction beyond the point where the influence on the estimates is worthwhile. We feel that since we now have the capacity to obtain good information concerning the effect of corrections, we will be able to adapt to more appropriate procedures.
