WP.51

ENGLISH ONLY

UNITED NATION STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (v): Quality indicators and quality reporting

AN UPDATE OF ANALYSIS OF DATA SLICES AND METADATA TO IMPROVE SURVEY PROCESSING

Supporting Paper

Submitted by Statistics Canada[1]

Abstract: In May 2002, in Helsinki, Finland, a paper submitted by Statistics Canada on the Unified Enterprise Survey (UES) was presented at the UN/ECE Work session on Statistical Data Editing. The title of the paper was "Analysis of Data Slices and Metadata to Improve Survey Processing". The paper provided an overview of the UES Collection and Post Collection processes. It also presented the findings of a study evaluating those two processes. Conclusions and actions to be taken were discussed in the paper. Three years later, we provide an updated version of the paper. This paper gives an overview of the collection and post collection processes. It puts the emphasis on the major changes that have taken place in the last few years. It explains how those changes have modified and improved the UES.

I. INTRODUCTION

1. The Unified Enterprise Survey (UES), initiated at Statistics Canada (STC) in 1997 with seven industries, now integrates close to 50 annual business surveys into one centralised survey system. In this paper, we will use the term industry to refer to a specific economic sector. Businesses of all sizes are in scope for the UES. Larger firms are always selected in the sample while the smaller businesses are randomly selected each year.

2. Many things have changed since the beginning of the UES in 1997. In the first few years of the UES, most industries employed two questionnaires – a long version and a short version. In subsequent years this process was changed to simplify the processing and to lower response burden for those that were filling a long form. Now, only one version of the questionnaire is used. This new version is much shorter than the initial long version.

3. Also, a graphical tool was developed to evaluate and monitor the impact of all the changes performed during a complete survey cycle. More details about this tool are given by Hazelton (2003).

4. The most important change that affected the collection and post collection processes is the increased use of fiscal data, also referred to as tax data. In the recent years, this source of data improved drastically in terms of accessibility, accuracy and timeliness facilitating its use in the post collection process. For example, for reference year 2003, tax data accounted for close to 50% of the final estimate for some variables in certain industries.

[1] Prepared by Jean-Sébastien Provençal (provjea@statcan.ca)

5. In this paper, we provide a review of the collection and post collection processes. We summarize the collection process prior to 2002 that was documented by Martin and Poirier (2002). We review the findings related to collection that were pointed out in a study conducted in 2002. We summarize the updates that have been made to the collection process since 2002, and finally we discuss how various issues were addressed by the recent changes that were implemented. We go through the same steps to describe the post collection process.

II. REVIEW OF THE COLLECTION AND POST COLLECTION PROCESSES

A. Background of the collection process and review of the findings resulting from the 2002 study

A.1. Overview of the collection process prior to 2002

6. Survey data for the UES were collected via mail-back questionnaire or by telephone. All initial contacts were made via regular mail questionnaire. Follow-up activities in the presence of total non-response and for certain edit failures were conducted by telephone.

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

8. Two slightly different follow-up strategies were applied to mail-back units that failed capture edits. For the non-manufacturing sector, questionnaires categorized as having severe edit failures were flagged for follow-up. For the manufacturing sector, only "critical" units categorized as having severe edit failures were flagged for follow-up. Critical units were identified through a score function. Mail-back questionnaires having only non-severe edit failures and manufacturing "non-critical" units were not flagged for follow-up.

9. All total non-responses, i.e. units that do not mail back their questionnaire, were flagged for follow-up. A major concern was the cost associated with the follow-up strategy. In 2002, it was estimated that a telephone follow-up takes on average 15 minutes. This estimated time did not account for the unsuccessful follow-up attempts that often precede a final contact.

A.2. Review of the findings on collection resulting from the 2002 study

10. In this sub-section, we provide a summary of the findings and highlights that were pointed out by the study carried out using results from reference year 1997 to reference year 2000. The study made the following recommendations.

Edit failure and Follow-up Rates

- *The need to find a way to encourage respondents to use the mail-back questionnaire in order to minimize the cost associated with telephone data collection;*
- *The need to re-visit the 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;*
- *The need to find a way to prioritise individual units for follow-up in an even stricter fashion than was currently employed for the manufacturing sector.*

Impact of Follow-up

- *The need to direct the attention to the problem of edit failure caused by improper capture.*

11. To conclude, the use of a shorter questionnaire, starting in 2002 for reference year 2001, was expected to improve the response rate and reduce the edit failure rate.

B. Updates made to the collection process since 2002 and their impact

B.1. Updates made to the collection process since 2002

12. Through the last few years, several new methodologies and processes were put in place in order to improve the collection process. One of these was the implementation of a Quality Control (QC) process for QDE starting in 2003 (collection for reference year 2002). Overall, QC results showed that QDE was efficient. It was estimated that less than 1% of the records were found initially in error and less than 0.5% were still in error at the end of the process. Similar results were observed for reference year 2003.

13. Another major change made to the collection process is related to the follow-up strategy. Starting in 2003 (collection for reference year 2002), a score function was implemented for the non-manufacturing sector. The score function provides a day-to-day operational plan for collection. It identifies the units that are the largest potential contributors to the final estimates in each domain (e.g. industry by geography) and follow up is performed only for these most important units. The score function is a tool that helps to better manage the collection effort and reduce the collection cost. It also serves to meet the collection goal of obtaining a high weighted response rate. More details about the score function are given by Poirier, Phillips and Pursey (2003).

14. The follow-up strategies used for manufacturing and non-manufacturing sectors are now more similar. Both sectors are now using a score function to identify critical (followed-up) and non-critical (not followed-up) units. However, each sector uses a different methodology to derive their score function.

15. Finally, more pre-contact activities took place. This was done to account for the increasing contribution of fiscal data to the final estimates. The main purpose of the pre-contact is to validate the geographical and industrial classification of units for which it was intended to use fiscal data instead of survey data.

B.2. Impact of the updates on the collection process

16. The QC done on QDE confirmed that data obtained from the capture process were of good quality. Analysis done on reference years 2002 and 2003 showed similar results.

17. The implementation of a score function was certainly a major change to the collection process. The score function allows for the prioritizing of units selected for follow-up activities in order to reach a maximum coverage for a given domain.

18. On the other hand, the introduction of a score function raises a few issues. The use of the score function led to a minimal collection effort for a large portion of the sample. Units with a low priority are not followed-up and consequently we observe a very low response rate for these units. One could wonder why having an optimal sampling design for a given sample size if we cannot afford a process with a follow–up strategy for each surveyed unit? Should we consider reducing the sample size in order to be able to put a greater collection effort for each sampled unit? There are also some concerns regarding the respondent perception if no effort is done to contact them if they do not respond.

19. These issues are currently being studied. We might want to come up with a strategy where a minimal follow-up effort is carried on for each unit. For example, the score function could be used to control the effort's intensity.

20. Also, since reference year 2002, we reduced resources allocated for collection, and increased the use of tax data. The use of the score function may also be revisited given the increased use of tax data. Since non-financial data can only be obtained from respondents, there is more pressure to obtain even higher response rate from respondents.

C. Background to the post collection process and review of the findings resulting from the 2002 study

C.1. Overview of the post-collection process prior to 2002

21. There were three main steps in the post-collection process: 1) post collection review and correction by survey data analysts, 2) automated imputation, and 3) post-imputation review and correction. This section focuses mainly on automated imputation.

22. The automated imputation included three distinct parts. First, we imputed missing key variables, (e.g. three or four variables, generally totals), identified for each industry, for all partially filled questionnaires or empty questionnaires with historical data available. Second, we imputed missing details (e.g. totals breakdowns or characteristics like number of employees), for all partially filled questionnaires or empty questionnaires with historical data available. Finally, we imputed all empty questionnaires.

Imputation of missing key variables

23. For missing key variables, the automated system imputed a value by using different methods. The choice of the method depends on the data available for the record. The methods are applied in the following order:

- Derivation of 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 similar records
- Year over year trend for each key variable within a group of similar records

24. These processes ensured that we have key variables for all units that reported at least one key variable, and for all units that were in the survey in the previous period.

Imputation of details

25. For units that have a value for their key variables, (either reported or imputed) the distribution of details for the short questionnaires was taken from each unit's tax data report. The distribution of missing details, for the long questionnaires, was obtained through donor imputation. The imputation is done independently for each section of the questionnaire. Key variables are used to find the nearest neighbour for each section.

Imputation for total non-response

26. Empty records underwent 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. Ratios observed between the donor and the 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).

C.2. Review of the findings on post-collection resulting from the 2002 study

27. In this sub-section, we provide a summary of the findings and highlights that were pointed out by the study carried out using results from reference year 1997 to 2000.

Manual Imputation Rates

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

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

Impact of Manual Imputation

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

D. Updates made to the post-collection process since 2002 and their impact

D.1. Updates made to the post-collection process since 2002

29. One major update to the post collection process is the increasing use of tax data as a proxy for some pre-designated variables upfront in the process. Another very important update, that had an impact on the post collection process, is the replacement of the long and short form questionnaires by one redesigned single-form questionnaire.

30. Since reference year 2001, we have been using tax data as a proxy value for specific variables in case of total non-response. Since reference year 2002, we extended this approach in using tax data in lieu of survey data for a portion of the sample where no questionnaire was sent. For reference year 2002, we implemented this process for a limited number of records across a few industries. For reference year 2003, we extended it. For example, for the non-manufacturing sectors, we used this process for less than 6,000 sampling units for reference year 2002 and for over 13,000 sampling units (out of 44,000) across 7 industries for reference year 2003.

31. This process allows us to reduce response burden and the collection cost. However, the use of tax data has its limitation: tax data are easily usable only for simple business structures and not all the traditionally surveyed variables can be obtained through tax data. Variables for which tax data can be used vary from one industry to another. For reference year 2003, the number of financial variables used ranged from 7 to 25 depending on the industry.

32. Records for which we use tax data are treated like records with a partially filled questionnaire and are dealt with in the first two parts of the automated imputation process. At these stages, we process survey data and tax data together. Ratios and trends used to impute key variables are calculated using survey and tax information. Generally, we used survey filled questionnaires as donors to impute the missing details from questionnaires that were partially filled using tax information. Several studies had shown enough similarities between these two sources to justify this approach.

33. We are now using only one form of questionnaire for all sampled units rather than the two forms (a short and a long) that were previously used. The new form is shorter than the old long form. The processing is now simpler. As well, partially completed questionnaires are now more likely to be imputed by one donor.

34. Finally, with a simplified process, we were able to expand the set of techniques used to impute missing details on each partially filled questionnaire. The techniques described below represent the result of the improved imputation strategy.

- Derivation of rules involving other reported data for the individual record
- Carry forward of last year information for the individual record
- Previous year's ratios amongst variables for the individual record
- Current year ratios amongst variables within a group of similar records
- Donor imputation using the nearest neighbour method within a group of similar records

35. Since different imputation situations may call for different techniques, the expansion of the number of techniques helped to better tailor the imputation strategy for each industry.

D.2. Impact of the updates on the post collection process

36. The main factor that had the most impact on the UES Post Collection Process is linked to the increased use of tax data. Tax data have continued to improve over the years in terms of availability, timeliness and consistency, which gave us the opportunity to increase their use.

37. The use of tax data significantly reduces the number of questionnaires with total non-response. Hence, it reduces the use of massive imputation that often leads to major sources of manual revisions by the data analysts. For reference year 1998, 67% of the manual revisions were for units that had been imputed through mass imputation. For example, for the retail trade store industry, the percentage of cases for which we used mass imputation decreased significantly. It went from 25% for reference year 2000 to 7% for reference year 2003 which represent a decrease of few thousand records that needed imputation.

38. Recently, the UES questionnaires have been redesigned to better represent accounting principles that are more easily understood by respondents and that are also used in the tax forms. This initiative should also help reduce partial non-response as the concepts measured should be more readily understood by the respondents. Some surveys have already gone through this process, and others will undertake this task very soon. This project will be finalised in the upcoming years. This initiative will increase the number of variables for which we can use tax data. Hence, it will reduce the imputation rate for these variables.

39. Finally, despite these many improvements, there is still a need to develop a method in order to better identify units that should be reviewed by the data analysts. An improvement in this area would certainly optimize the use of resources, and would contribute to improve quality of the final product, especially its timeliness.

III. CONCLUSION

40. The main purpose of this paper was to give an update of the recent progress regarding UES Collection and Post Collection processes issues that have been reported by Martin and Poirier (2002). Both processes have considerably evolved.

41. On the data collection side, we made changes in order to use the available resources more efficiently and optimize the follow-up strategy. Work is still on going on this issue as the UES context is changing with the increasing use of tax data. We will have to develop new surveys focussing on characteristics for some industries. Optimal use of the resources is necessary in order to achieve the targeted objectives of timeliness, data accuracy and cost reduction. The collection process will need to adapt in order to face all these challenges.

42. On the post collection side, major improvements have been made. Fiscal data is a reliable source of data that helped to improve the process. This source was used to reduce significantly the major cause of manual reviews e.g. records that were imputed through mass imputation. On the other hand, we still need to develop a method to target more efficiently which records should be reviewed manually.

43. Our goals remain to reduce respondent burden, reduce collection cost and make better use of tax data. Our efforts to harmonise our questionnaires with the concepts used by accountants (our respondents) and the tax forms may lead us in different directions. For example, how can we best integrate survey data and tax data?

44. The integrated use of survey data and tax data will require the development of new quality indicators in order to inform the users about the data quality and accuracy of the final products.

Acknowledgements

Acknowledgements to Hélène Bérard, Sylvie DeBlois, Anthony Dupuis, Judy Lee and Claude Poirier of Statistics Canada for their contribution.

References

- Hazelton, F. (2003), Impact of Data Processing on Unified Enterprise Survey Micro Data, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Spain (Madrid). (http://www.unece.org/stats/documents/2003.10.sde.htm)

- Martin, C. and Poirier, C. (2002), Analysis of Data Slices and Metadata to Improve Survey Processing, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Finland (Helsinki). (http://www.unece.org/stats/documents/2003.10.sde.htm)

- Poirier, C., Phillips, R. and Pursey, S. (2003), The Use of a Score Function in a Data Collection Context, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Spain (Madrid). (http://www.unece.org/stats/documents/2003.10.sde.htm)
