STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 6
English only

CONFERENCE OF EUROPEAN STATISTICIANS

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

Topic I: Management and evaluation of editing and imputation procedures

METADATA – AN AID TO MANAGING THE EDIT AND IMPUTATION PROCESS

Submitted by Statistics Canada[1]

Contributed paper

Abstract

Edit and Imputation is an important and often expensive part of the survey process. Managers and others involved in the process are encouraged to find efficiencies to minimize the cost in terms of persons and time spent. In order to support decisions that must be made in this regard, it is necessary that the decision-makers have at their disposal all relevant information about the process. The method explored here combines the notion of making available versions of data from each important step in the process as well as providing additional processing metadata that describe how the data were transformed.

I. INTRODUCTION

1. The Unified Enterprise Survey (UES) is a relatively new Statistics Canada initiative. Among its many objectives, the survey aims to integrate most annual business surveys into one centralized survey process, using one common methodology and one generalized computer processing system. Each major processing step of the UES has a manager whose role is to work with many diverse subject matter areas to gather their requirements, to oversee the development and implementation of computer systems that meet those requirements and to manage the actual process in production. The Edit and Imputation (E&I) manager is responsible for the automated E&I process as well as the manual review and correction processes that are conducted by the subject matter analysts. The UES E&I manager assumes as well partial responsibility for defining the edits and follow-up procedures carried out in Data Collection.

2. The large volume of data that passes through the E&I system and the fact that the process is centralized demands that strict attention be paid to the cost in terms of people and time as well as to the process in terms of reliability and suitability. A primary goal is to measure the effect of manual intervention in relation to the resources spent and to have available enough detailed information that scarce resources of time and personnel can be directed to the areas that have the most impact. A second goal is to determine whether pre-specified automated procedures are in fact the most appropriate. The overall objective is one of continuous improvement.

3. The systems and the methods that have been put in place must support management's requirements for information about the processes and about the data transformations that take place

[1] Prepared by Colleen Martin.

through those processes. Only then can appropriate decisions be taken regarding changes to the automated system and to the manual interventions, for future survey cycles.

II. AIMS AND OBJECTIVES

4. There are several factors that influence the conduct of a survey processing system, specifically those components related to edit and imputation. A major objective during the development of the UES processing system was to ensure that enough processing metadata resulting from the process itself were retained to allow all interested parties to understand the impact of each of those factors.

5. We wanted to learn about the role and appropriateness of data collection edit and follow-up, the role and appropriateness of systematic edit and imputation, the role of manual review and correction and the influence of questionnaire design. Manual review and correction procedures were of special significance in light of the person-hours and the elapsed time traditionally spent carrying out such manual processes at Statistics Canada.

6. We wanted to measure the rate and impact of the data collection edit failures and the ensuing follow-up. The cost is considerable and the response burden can be onerous.

7. We wanted to measure the impact of systematic imputation on the estimate. For UES, this is largely a measure of the effect of imputing values to replace missing values. The basic methodology of UES rarely changes reported values.

8. We wanted to measure the impact of manual imputation on the estimate. This implies a measure for those instances where missing values are manually imputed, where reported values are overridden and where systematically imputed values are overridden.

9. Poor response is arguably the single factor most affecting E&I. Apart from complete disinterest and refusal to cooperate, the most influential factor leading to poor response is the design of the questionnaire. A questionnaire that is too long, a questionnaire that is imprecise or unclear, a questionnaire that is repetitive, or asks for obscure, rarely available data discourages respondents from 'giving it their best shot'.

10. The design of the questionnaire also directly affects the edits and imputations that must be performed to ensure a coherent record. Details adding to totals are the prime example of arithmetic relationships that must be maintained. There may on the surface appear to be a redundancy when both are asked on a questionnaire, but other factors play a role and conventional wisdom dictates that it is a useful redundancy. On the other hand, asking more than one question to obtain a single (duplicate) answer leads to complications for the E&I processes, and is likely to annoy respondents, at least to some extent.

11. The E&I manager, survey methodologists and subject matter analysts are all interested in learning more about the efficiency and suitability of the entire process.

III. EDIT AND IMPUTATION PROCESSES AND PROCESSING METADATA

12. The definition of the term **metadata** is generally given as data that describe the data of concern… or briefly data about data. **Processing metadata**, for the purposes of this document, refers to the metadata that are retained by the system processes and which allow us to understand the data transformations that have taken place during those processes.

13. Edit and/or imputation are carried out in each of the first four phases of the overall UES survey process. The data and metadata resulting from each of these phases are housed in a central data repository so that at the end of the process, four slices of data with accompanying metadata are available for analysis.

III.1 Data collection phase

14. Survey data for the UES are collected via mail-back questionnaire or by telephone. Follow-up for edit failure is conducted by telephone. Mail-back questionnaires are captured using the same capture and editing system as those collected by telephone.

15. Data Collection edits fall into 3 main categories:
- *Edits that detect missing mandatory variables,*
- Edits that detect arithmetic inconsistency between related variables … e.g. the sum of details and the reported totals, equivalencies,
- Edits that query the relationship amongst related variables, where those relationships are not strictly mathematical but are somewhat more fluid… e.g. a specific expense item should not be more than X% of the total expense.

16. For units that are mailed back, if only query errors are detected on the questionnaire no attempt is made to follow-up, otherwise follow-up is attempted (although it is not always successful). For units collected by telephone, all detected errors are reviewed with the respondent during the interview.

17. For each questionnaire passing through the system, processing meta data are recorded:

- For each edit, meta data provide the current status of the edit:
- Edit is passing - no follow-up has been done,
- *Edit is failing - no follow-up has been done,*
- Edit is passing - (resulting from) follow-up which led to correction,
- Edit is passing - (resulting from) follow-up which confirms the current value(s) (applies to query edits only),
- Edit is failing - follow-up has not confirmed current value(s).

- For each variable, metadata provide the current status of the variable:
- Variable missing - no explanation,
- Variable missing - respondent refused,
- Variable missing - respondent unable to provide,
- Variable present – reported.

III.2 Post collection review and correction phase

18. This is an interactive process that provides subject matter analysts with the first opportunity to review the raw data and to correct errors. The intention in this phase of processing is to transform the data in such a way that the batch edit and imputation process to follow is enhanced. This process is not meant to take the place of systematic batch edit and imputation.

19. For each questionnaire that is manually altered through the system, processing metadata are recorded:

- For each variable that is changed, metadata provide the current status of the variable:
- Variable present – manual imputation.

III.3 Batch edit and imputation phase

20. The E&I system consists of a series of algorithms, designed to detect which variables require imputation, to detect outliers and to impute values where required. A variety of imputation methods are employed.

21. For each questionnaire passing through the system, processing meta data are recorded:

- At the record level, meta data in the form of simple flags provide easy categorization:
- A flag to indicate that one or more of the key variables has been imputed,
- A flag to indicate that the record has been identified as an outlier for purposes of donor imputation,
- A flag to indicate that the record has been identified as an outlier for purposes of trend calculation,
- A flag to indicate that the record is a critical unit contributing more than X% of the estimate for its estimation group,
- A flag to indicate that the record could not be completely imputed using batch E&I algorithms and that it therefore requires manual attention.

- For each variable that is changed, metadata provide the current status of the variable:
- Variable present - imputation method X,
- Variable present - imputation method Y,
- Etc.

- For each set of variables (edit group) imputed by donor method:
- identification of the questionnaire, the edit-group and the donor.

- For each variable causing a record to be an outlier for purposes of donor imputation
- identification of the questionnaire and the outlying variable.

III.4 Post E&I review and correction phase

22. The physical system used in this phase is the same system that is used to review and correct in the post collection phase. There are two main differences in the mode of operation ... the categorization flags are an aid to sub-setting the data and the data cannot leave this phase until all records are error free.

23. For each questionnaire that is manually altered through the system, processing meta data are recorded:

24. For each variable that is changed, metadata provide the current status of the variable:
- Variable present – manual imputation

IV. EVALUATING THE PROCESS

25. From its inception, the UES processing system has been designed with a view to providing processing managers, methodologists and data analysts with more and better information about the processes and data transformations that take place than has generally been available through earlier Statistics Canada survey processing systems. The aim is to study and analyze the process events to find the weaknesses and to make changes where necessary to minimize, perhaps even eliminate those weaknesses.

26. There are two aspects to the processing system that allow such analysis. First, actual data out of each major processing step are stored and kept, such that four versions of survey data result. Second, processing metadata specific to each processing step are an integral part of the output of each process.

The availability of these versions of data and the accompanying metadata support the analysis of data through the four processes.

27. Analysis of the edit status variables retained in Data Collection will allow us to measure edit failure rates and, for query edits, edit confirmation rates.

28. For arithmetic edits, attention will be given to those that have a higher failure rate to try and determine the cause. The overall design of the questionnaires and the clarity of the question, with emphasis on the specific questions related to the edit will be explored.

29. For query edits that have a high incidence of being confirmed, the edit will either be removed or the parameters will be relaxed.

30. For specific questions that are rarely answered or for which respondents make it clear through the item status variable that they do not have available such answers, we will consider dropping the question allowing the data to be included in a more generic way.

31. For the automated E&I procedures, we will review the parameters, variable groupings and record groupings used to see whether they are appropriate in the long term.

32. We will review the results of the various imputation methods employed to see if they are the best choice for the reality of the survey response as opposed to the theory of survey methodology. The item status variables will provide the details required and the versions of data will allow us to substantiate the findings.

33. As stated before, we want to pay special attention to the costly manual processes. The UES processing system provides two opportunities for subject matter specialists to confront directly the data of their questionnaires. Many resources of persons and time are invested at both stages.

34. The long-term view for UES E&I is that we can eventually move away from much of the manual intervention and rely more on automated methods of imputation. Certain questions must be answered for this to happen. What is the effect of manual imputation on the overall estimates and at what cost? Are all manual interventions created equal? What causes subject matter experts to override automated imputations?

35. The majority of time spent in manual review and correction is during that phase which follows automated E&I. By comparing the version of data after E&I with the version of data after manual correction, we will be able to see the actual impact of the changes made. Some studies done in the past would seem to indicate that once a small number of extremely large units are corrected, there are rapidly diminishing returns for the time spent and that overall the changes made often have minimal effect on the final estimates.

36. Where results of analysis indicate that the questionnaire itself is flawed, we will be prepared to make changes as indicated.

V. THE FUTURE DIRECTION

37. Evaluating the conduct of the edit and imputation modules is an on-going process. The E&I manager, methodologists and the subject matter analysts must question always the suitability and the cost of the data transformations that have taken place.

38. The UES processing system is still in its infancy. New processes will be developed and new processing metadata will be added as the system evolves. Some of these changes will be the direct result of analysis conducted on the data of earlier survey cycles.

39. Until now, not a lot of time has been given to analysis of the available information. Recently a position was created for a person whose job it will be to coordinate and indeed carry out much of the analysis. Only then can the worth of the current metadata and version techniques themselves be evaluated. The first report will be available by early next year.

40. Much metadata exist that are not yet useable in a systematic way for purposes of analysis. For example, interviewers go to great lengths to include comments about their respondent contacts. These are unfortunately just text and thus not easily incorporated in automated studies. As well, many of the imputation modules employed for UES are borrowed from Statistics Canada's Generalized Edit and Imputation System (GEIS). The system produces a multitude of logs, but these too are not well formatted to be included in an automated way. Both sources, however, provide interesting and relevant information. A longer-term goal would be to have these bits of information recorded in a more useable way.