WP.11
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (ii): Editing data from multiple sources

RE-EDITING OF DATA IN CANADIAN BUSINESS SURVEYS

Supporting paper

Prepared by Eric Rancourt, Statistics Canada

Abstract: In a business surveys program, some statistics are solely based on a survey or on administrative data. On the other hand, more and more statistical programs are based on both survey and administrative data. In each case, an editing and imputation process has been developed and implemented. However, when the two sources are confronted or used together, discrepancies have lead people with a desire to increase the quality of the resulting data by adding more edits and interventions. This raises questions such as: Is there too much editing in the process? Is this justified to increase consistency? What is the interplay between accuracy of data in a given source and consistency with another source? This paper explores those questions in light of Canadian examples.

I. INTRODUCTION

1. In a business survey program, there is a wide variety of concepts to measure in order to provide relevant information for analysis of the situation. Collection of such data can be performed directly by contacting selected enterprises or indirectly by means of various administrative sources. Such sources can be income tax forms, tax deduction of perception data, payroll information, etc. At the end, a specific industry program may be based on survey data only, on administrative data only, or a combination of the two.

2. No matter which data source is considered, there will invariably be some form of non-response due to a variety of reasons. As well, amongst the available data there will likely be some inconsistencies due to logical or expected relationships that are not satisfied by the data. For these reasons, editing and imputation techniques are usually used to improve the data treatment and help remedy some of the data problems.

3. The editing and imputation processes can be manual or automated (though usually automated). As well, depending on the objectives of the survey (e.g. producing micro data versus aggregates) the process may have various degrees of sophistication. Typically, business data are edited for their internal consistency, consistency with previously reported information and then consistency with other records.

4. In recent years, the amount of administrative data as well as their quality has significantly increased. At the same time, there has been a need to rationalize the survey process to make it more efficient and, ideally, improve its quality. Thus, at the beginning of the years 2000, Statistics Canada began to incorporate a more direct and aggressive use of tax data in its business survey programs.

5. Using both survey and administrative data within the same program provides an opportunity to confront data from different sources and therefore make comparisons and further validate the data. In a sense, this is a third step of editing that can be referred to as re-editing. Indeed, survey data are edited and imputed; administrative data are independently edited and imputed; and finally the two data sets are (at least partially) joined together (see Section IV for examples). Therefore, it potentially creates inconsistencies that need to be dealt with.

6. The question then comes as to whether and how to treat such inconsistencies resulting from a combined use of survey and tax data, but further, whether these data should first be combined altogether before attempting any kind of internal editing and imputation activity. Several issues can be raised, such as the need for a centralized set of clean administrative data, such as the varying timeframes of data availability versus the attempt to find a statistically optimal solution. The following sections discuss such issues.

7. Section II briefly outlines the historical context of survey taking and discusses the editing and imputation phases of surveys while Section III presents the new context of surveys. Then, examples are considered in Section IV and Section V contains a discussion of the issues. The conclusion follows in Section VI.

II. COMMON EDITING AND IMPUTATION PRACTICES

8. Business statistics programs are mainly based on surveys covering one or more specific industrial sectors such as retail, wholesale, mining, manufacturing or services. In general, the nature of information collected is of two different types: First, financial information; and second, information on characteristics of the products/services. In a given survey, there is often one or a few key concepts to measure and these are used to define which records are respondents and which are not.

9. Typically, editing rules are applied to the key variables at collection time, thereby defining a critical path, to identify total non-responses. For other records, a more complete set of edits is applied by groups of variables involved in common edit rules, constituting edit sets. Edit sets are then applied to groups of data to identify values to accept or reject.

10. When carried out, editing follows a number of principles. First, the usual motivating factor is to enhance consistency within and between records. To do so, editing is used as a preliminary step to imputation, in order to identify which records are to be imputed to create a complete "rectangular" data file. Ideally, the process is automated to that it can easily be reproduced. Of course, simplicity is a central feature of any editing and imputation strategy. For instance, guidelines have been defined at Statistics Canada (Statistics Canada, 2003) using such principles.

11. Though consistency is very desirable, the main ***objectives of editing*** are not to "correct" survey data. It is rather, as defined by Grandquist and Kovar, (1997) to:

- Provide the means to better understand the survey process and plan for the future;
- Allow survey practitioners to know and inform users about the quality of the data;
- Tidy up the data so that flagrant data problems can be overcome.

12. Sometimes, and not necessarily consciously, the first two objectives are left adrift and editing becomes too much of a tool being used in an attempt to correct every imaginable error. This tends to be concentrated in manual editing. With computers and software, editing and imputation systems require compromises to be made (from the long list of manual checks) but result in higher overall quality data because of the greater consistency, increased simplicity and improved timeliness.

III. THE NEW SURVEY CONTEXT

13. Since after the end of the Second World War, the history of economic programs in statistical agencies can be categorized as being increasingly survey-centric. That is, whenever new questions have been asked about the economy, the tendency has been to design a new survey to respond to those needs. This was true until the 1990's. At that point, the collection and processing experience as well as the knowledge of administrative data reached a new level where the potential benefits of its active use became evident and warranted. Further, for efficiency reasons, the rationalization of programs became required and the increased use of administrative naturally appeared as part of the solution.

14. In Canada (and in many other countries), direct use of administrative data was implemented relatively recently. As will be seen in Section IV, the annual and monthly surveys started using tax data in lieu of survey data for the financial part of the questionnaires. Tax data (and administrative data in general) are processed separately from survey data. This allows for increased consistency, as only a single version of tax data is processed and produced rather than one for each program. On the other hand, once tax data are joined with survey data, new inconsistencies may arise.

15. There are many ways of defining consistency, and each can be tackled by introducing some form of editing. However, whenever data are forced to follow some rules, there is an increasing risk of introducing bias. In a sense, increasing consistency may come at the expense of a reduced accuracy. Three levels of consistency can be considered:

16. ***Record consistency*** is achieved when data for a given unit pass the set of edit rules.

17. ***Consistency of sources*** is achieved when data for a given unit obtained from more than one source are in agreement according to some rules. For example, a survey objective may be to measure the total operating expenses. For some records, instead of being surveyed, it is extracted from administrative files, but it is nonetheless consistent with other variable obtained from the survey.

18. ***Overall consistency*** is achieved when aggregates from different sources are in agreement. For example, this is the case in the System of National Accounts (SNA) when wages and salaries from tax files (T4) agree with those of surveys as obtain from the Survey of Employment, Payrolls and Hours and/or the Annual Survey of Manufactures (ASM).

19. The potential inconsistencies create a situation where the temptation is high to perform increased editing. It can be implemented formally by following edit rules (see Section IV) or in an ad hoc fashion. This is particularly the case at the analysis time when micro data comprising both survey and administrative data are available for the analysts. In this case, manual interventions can take place and if data are changed, the overall impact is often not assessed.

20. There is nothing wrong with further editing the data after survey and administrative data have been combined. On the other hand, the three objectives of editing in Section II still apply and the temptation to slip into the third one must be guarded against.

21. At Statistics Canada, with the increasing use of tax data in the economic surveys program, a Data Integration Project (DIP) (Trépanier, 2006) was launched in 2005 to study the optimal approaches to combining survey and administrative data. The project is looking at three aspects of the problem : A) Conceptual definition of variables present in both sources; B) concepts and definitions of nonresponse; and C) optimal sample design (mainly allocation of the sample between survey and tax data as well as the choice of estimator). While the DIP is focusing on sample design, the optimized approaches that will stem from this work should guide methodologists in their implementation of re-editing, if it is needed.

IV. EXAMPLES

22. ***The Unified Enterprise Survey*** (Pelletier, 2004; Nadeau, 2005) is a centralized survey vehicle for annual business surveys covering more than 45 sectors of the Canadian economy. The survey collects financial information as well as characteristics of the output produced. In 2002, the direct use of tax data was introduced to reduce the response burden of enterprises as well as the on-going costs of the survey. To do so, the sample is first selected, and a random sub-sample of businesses with a simple structure is randomly selected. For this sub-sample, income tax data from the Canada Revenue Agency (CRA) are used instead of carrying out the survey. The tax files are centrally processed in the Tax Data Division (TDD) of Statistics Canada. What is called tax replacement is performed for variables that have been tested and found to have matching concepts between the two sources. About 55% of simple-structured businesses are "tax-replaced".

23. After tax data have been added to the survey portion, data processing continues with editing and imputation. In the automated process itself, little re-editing of tax data is performed, though in a few cases tax data may be rejected and re-imputed using survey variables. Once a complete data file has been produced, manual review of records takes place and then re-editing (of either survey or tax data) takes place.

24. The main economic ***monthly surveys*** are the Monthly Survey of Manufacturing (Yung, Cook and Thomas, 2004), the Monthly Wholesale and Retail Trade Survey (Trépanier, 2004) and the Monthly Restaurant, Caterers and Taverns Survey (Statistics Canada, 2004). In 2004, these programs started using Goods and Services Tax (GST) data related to sales to perform tax replacement of some of the financial values. The approach is the same as for annual surveys except that replacement is not performed directly but rather through a simple regression model (mostly ratio). As for income tax data, GST data are centrally processed in the TDD.

25. After tax replacement, some re-editing is performed. Month-to-month changes in the GST revenue is looked at as well as the difference between the revenue based on the GST model and that which would be obtained from the survey imputation process. If some thresholds are exceeded, then the GST modeled value is rejected in favor of the survey-based imputed value.

V. DISCUSSION

26. The combined use of survey and administrative data opens up a wide range of possibilities for economic survey programs, but it raises some issues discussed in the following.

27. There has been a lot of progress in the development, implementation and understanding of editing. Nowadays, multiple tools exist to perform editing and tailor-made programs can be developed relatively quickly to allow users access to data and analysis of changes. With this wide and easy access, there is a risk of unduly increasing the amount of editing. Further, the desire to "make" data consistent should be balanced with the potential loss of accuracy.

28. Should data sources be processed independently or jointly? Separate processing allows for breakdown of the processing steps and centralized administrative data collection, processing and management. On the other hand, joint processing provides more flexibility, leads to a more optimal solution to the problem at hand but requires re-processing of the administrative file for each application. With separate administrative data processing, the consistency of sources is maximized, but the overall consistency between programs would suffer.

29. The consistency of sources appears to be a simple issue when only two sources are used. If we imagine this process in a context of multiple sources, then the issue is not so clearly resolved. Centralized processing will prevent the proliferation of multiple versions of the same file and greatly improve overall consistency. However, with centrally produced administrative files, the implementation approach of

using administrative data may have an impact on quality/consistency. If the survey information is first used, and then administrative file 1 is combined with it, re-editing could take place to produce an intermediate file. Then if administrative file 2 is added and a new re-editing process is added, there is a risk of slipping into a sequential editing process that could lead to a dead end. Most likely, the survey information as well as the administrative information should be processed together, but this leads back to the original question. At this time, there does not seem to exist established measures to serve as a guide in the matter of compromising between accuracy and consistency. Such tools could become useful.

30. The compromise between consistency and accuracy can also be paralleled to the compromise between adding noise or not in confidentiality procedures. In this context, publication without adding noise leads to greater accuracy and can be related to editing survey and administrative data together in a specific application. Then adding noise to protect confidentiality can be related to re-editing once survey and administrative data have been processed separately. In both cases, more rules (confidentiality or editing) can be applied but at the expenses of accuracy.

31. Combining survey and administrative data can further be viewed as imputation. When tax data are used instead of surveying businesses, the replacement process is in fact the application of a model to fill in holes, which is nothing else than imputation. In the context of imputation, it is important to properly set up flags to identify which records have been imputed. Similarly, in tax replacement, one should keep track of the source of records (survey or tax). When tax replacement is viewed as imputation, then the framework established to measure the variance due to imputation can be used. Since it can be extended to variance due to editing (Rancourt, 2002), it could perhaps be used to derive a means by which one could balance accuracy (editing) against consistency (re-editing).

32. If the use of administrative data is viewed as a form of imputation, the question of whether to use imputed administrative data to impute missing survey data comes up. In re-editing, should the priority (likelihood to change) be given to values already imputed or not? This issue is akin to the context of longitudinal surveys. Indeed, it would be ideal to wait for all panels of a survey to globally edit and impute but the process must be implemented sequentially one panel at a time (as soon as data from a panel become available). In this case, priority is often given implicitly to what was already imputed.

33. No matter how editing and re-editing of survey data and administrative data are viewed, the main question is always coming back to the concepts targeted. If the concepts are similar, then the need for re-editing should be minimal. On the other hand, if they differ slightly, re-editing may become more important. Finally if the concepts are very different, then the variables should be treated as two completely different variables.

34. At Statistics Canada, major efforts have been made to understand, describe and link administrative (tax) data concepts and survey concepts. To do this, a Chart of Accounts (Vinette, 2003) has been developed and it has greatly facilitated the mapping of variables as well as the modeling approach for tax replacement.

35. When looking at micro data, the need for re-editing may appear more important than actually required. The analysis of combined survey and administrative data seems to be a good candidate for macro editing and possibly selective editing. Before re-editing many individual records, it would perhaps be wiser to consider aggregated data and determine the high impact records and perform selective editing. This may reduce the need for re-editing and contribute to preserving the accuracy while enhancing consistency.

VI. CONCLUSION

36. Quality is not one-dimensional. In this paper, we have explored a link between accuracy and consistency in the context of combined use of survey and administrative data. The statistician may be comfortable with estimates of parts that do not add to a total, for example, but for the user it is almost

intolerable. That is why re-editing of the resulting survey-administrative data files takes place. However one must be careful not to fall into the trap of the third editing objective (attempting to correct data) and always keep in mind the first two (learning about the process; informing users of quality).

37. We have explored the parallel between re-editing and other aspects of surveys such as longitudinal surveys, imputation, protection of confidentiality and macro editing. These offer some starting points to develop a method/approach aimed at controlling accuracy and consistency at the same time. Until a complete approach is found, survey practitioners should keep their focus on concepts measured as well as objectives to guide their re-editing decisions.

References

Granquist L. and Kovar, J. – Editing of Survey Data: How Much is Enough? *Survey Measurement and Process Quality*, Lyberg, L et al eds., J. Wiley and Sons, New York, 1997.

Nadeau C. – Challenges Associated with the Increased Use of Fiscal Data for the Unified Enterprise Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 2005.

Pelletier E. – L'utilisation accrue des données fiscales dans le cadre de l'Enquête unifiée auprès des entreprises. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 2004.

Rancourt E. – Using Variance Components to Measure and Evaluate the Quality of Editing Practices. UN/ECE Work Session on Statistical Data Editing, Helsinki, May 27-29, 2002.

Statistics Canada. - Quality Guidelines. Catalogue No. 12-539-XIE. Fourth Edition, October 2003.

Statistics Canada – Restratification et changements méthodologiques touchant l'Enquête mensuelle sur les restaurants, traiteurs et tavernes (EMRTT) : Répercussions et effet sur les données et les séries chronologiques. Internal Document, Statistics Canada, 2004.

Trépanier J. – Data Integration Project (DIP). Internal document, Statistics Canada, 2006.

Trépanier J. – The Redesigned Canadian Monthly Wholesale and Retail Trade Survey: A Post-mortem of the Implementation. Proceedings of the Survey Research Methods Section, American Statistical Association, 2004.

Vinette L. – The Chart of Accounts (COA) … Its Beginnings, Evolvement and Implementation. Internal document, Statistics Canada, 2003.
