WP.7
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (i): Editing of administrative and Census data

Processing Methodology of Tax Data at Statistics Canada

Invited Paper

Prepared by F.Brisebois, R.Laroche and R.Manriquez, Statistics Canada, Canada

I. Introduction

1. The use of tax data in Statistics Canada's annual and sub-annual business survey programs has increased to a great extent in recent years. The increase in use has been due in part to improved timeliness and quality of the tax data available and also due to the willingness of the survey programs to embrace tax data. A result of this increased use is a reduction of response burden and a reduction of collection costs for the survey programs.

2. Statistics Canada has a signed agreement with the Canada Revenue Agency (CRA) to have access to all tax microdata. This agreement falls under the jurisdiction of three acts – the *Statistics Act*, the *Income Tax Act* and the *Excise Tax Act*. Information is captured by the CRA which in turn, provides data to Statistics Canada. The Tax Data Division acts as Statistics Canada's single point of contact with the CRA and has the mandate to develop the concepts and methodologies that ensure tax data are consistent with and usable by all Statistics Canada programs.

3. This paper presents an overview of the processing methodology and current challenges for three important sources of tax data: the Goods and Services Tax (GST), T1 and T2. The GST provides a monthly source of tax data which is used in sub-annual business surveys, while the T1 and T2 sources are used by annual business survey programs. They are used for various purposes, but most importantly as direct (or indirect) replacements of collection units. For the majority of business surveys, the approach involves using tax data for small, simple units in the take-none stratum, for some selected sample units in the take-some stratum, and in some cases for units that did not return a survey questionnaire. Readers can consult Yung, Rancourt and Hidiroglou (2007) for a more detailed description of the use of tax data in business surveys at Statistics Canada.

II. Unincorporated businesses: T1 tax data

A. Overview of the processing methodology

4. The T1 Form is filled out annually by all Canadians having earned an income on which income taxes must be paid to the government. Canadians must report their income to the CRA for the calendar year *y* by April 30th of year *y+1*. Individuals recognized as self-employed workers or unincorporated businesses must fill additional schedules that provide details about their business income and related expenses. Their business income must be categorized in one (or more) of the following sources: Farming, Fishing, Rental, Professional, Commission, and Business. Data associated with these income sources are of particular interest to Statistics Canada business surveys.

5. Two datafiles are essential for preparing the ultimate T1 data used by surveys. The first is the Assessed Record File (ARF T1). It contains limited information for the entire population of interest, consisting of four million individuals having reported positive income for at least one of the six sources mentioned above. The main variables available on the file gross business income and net income broken down by income source.

6. The second datafile contains the T1 data. Unlike the ARF, this datafile contains more detailed information, but only for a subset of the population. Up until reference year 2006, Statistics Canada received data only for the portion of the T1 population that had filed electronically, that is, having produced their report using any pre-approved software and sent it electronically to the CRA. Approximately 50% of the population of interest filed electronically. Since reference year 2007, reports filled electronically but submitted by mail on a paper copy have been added to the data sent to Statistics Canada. These extra cases are referred to as the "barcoders" since the printed copy of these income returns include a barcode that contains all the T1 information entered using the software program, which can then be scanned and transformed into electronic data at the CRA. Combining these barcoders cases with the electronic filers, 85% of the T1 population is covered, the remaining part consisting of cases where the T1 form is filled and returned on paper and therefore not available electronically. In order for estimates produced from that file to be representative of the full population, imputation is performed to cover the missing 15%. Estimates for gross business income and net income are produced using the census ARF T1 file, while estimates for other variables are obtained from the created T1 complete file.

7. Minimal checks are performed on these two data files at the CRA before being passed on to Statistics Canada where more detailed processing takes place. The processing strategy involves the following steps: identification of partnerships, editing with deterministic rules, and imputation.

8. One of the first steps consists of identifying partnerships. A partnership exists when two or more individuals are partners in the same unincorporated business. For tax purposes, each partner must report the partnership figures and not the figures for his or her share of the partnership. For example, if a couple owns a dwelling that brings in $5,000 a year in rental income, each partner must report $5,000 in gross income on his or her tax return. Each partner's share of the business is reflected only in his or her net income. To avoid overestimating the revenue and expenditures of unincorporated businesses, it is important to accurately identify such partnerships. In most cases, partnerships can be quickly identified using Social Insurance Numbers of partners which are provided on the form. Otherwise, there is no unique key that identifies these partnerships. For these cases, a number of variables on the ARF (last names, industrial codes, addresses, etc.) and from other data sources are used to identify, by deduction, partnerships.

9. Deterministic rules are applied afterwards in order to balance results within sections (for example, the total for expenses must be equal to the sum of all its components). Records that fail these edits are later handled through systematic or manual corrections.

10. The vast majority of Statistics Canada's annual surveys create their frames from the central business register and therefore require that tax data are available for each unit selected from that frame. There are however four possible reasons for the absence of tax data for some units. The first is the presence of inactive units; these are cases that are present on the business register but for which tax data has not been received for the current and previous years; these units are imputed deterministically by setting all tax data to 0. The second reason is that at the time of estimation, the expected tax data have not all been received. This problem is solved by means of historical imputation with a trend applied to previous year's data to reflect changes in the economy. The third reason for the absence of tax data is the existence of "paper" respondents (where the income tax report is filled and submitted on paper to the CRA), for whom only two variables are available (from the ARF T1 file). Donor imputation is used to resolve this problem. The fourth reason is that data of poor quality are sometimes received for some electronic respondents; when such data cannot be systematically corrected, they are deleted and replaced by donor imputed data.

11. Finally, to achieve consistency across the provided tax data, collected survey data and National Accounts data, a chart of accounts was developed at Statistics Canada. The chart of accounts consists of standardized financial statements, which make it easier to link the three data sources. Therefore, the last processing step consists of aggregating the T1 financial data on the basis of the variables in the chart of accounts.

B. Current challenges

12. Since the use of T1 tax data is growing and could represent an important portion of estimates published for some industries, it is important to inform users of the quality of these tax data. A detailed quality report is being developed and will include various quality indicators, including variance due to imputation estimates. The main challenge for the estimation of this variance is to fully reflect all complexities of the imputation strategy in the calculations. The SEVANI system developed at Statistics Canada (Beaumont and Bissonnette, 2010) is used for that purpose.

13. Although several tax data sources are currently used directly by surveys, or indirectly in the processing of main tax data files, there are still some data sources that could help improving the overall processing strategy. For example, some schedules filled specifically by farm operators might contain data that could be used to help edit or balance fields on the main tax form or in other schedules. In fact, there are at the moment two processing systems in place at Statistics Canada for the processing of T1 and T2 data; the main one (described in this paper) is done by Statistics Canada's Tax Data Division and feeds most annual surveys, while a second was developed in the Agriculture Division to serve the more specific needs of agricultural surveys. There is currently an initiative to integrate both these systems into a single system that would serve the needs of all surveys.

III. Corporations: T2 tax data

A. Overview of the processing methodology

14. Canadian corporations file T2 returns to the CRA for each taxation year. The return provides information regarding corporation's Financial Statements and Tax Returns. Corporations are allowed to choose their own fiscal year but are expected to file their T2 tax information with the CRA within six months of the end of their fiscal year. Because of this arrangement, the CRA receives T2 tax data throughout the year and provides Statistics Canada with monthly files containing T2 data for businesses that have filed in the previous month. As for the T1 tax data, only minimal editing is performed at the CRA. Once processed at Statistics Canada, the T2 database is created for a reference year and constitutes a census of about 1.5 million legal entities.

15. Each month, Statistics Canada receives files containing the T2 tax data and financial information (Income Statement and Balance Sheet) from the CRA. From these files, assessment and reassessment data are extracted, reformatted, and then loaded into the database. It is important to mention that although the T2 tax form contains several hundreds fields, only eight of them are mandatory; many optional fields are nevertheless often of great interest to survey programs. This explains why, even though some editing is done at the CRA, a complete and thorough processing system is a necessity at Statistics Canada.

16. When received by Statistics Canada, the information is passed through a series of edits to balance the data and to identify errors or outliers for correction. Any errors that cannot be corrected are flagged for imputation. The first series of edits consists in flagging obvious erroneous large values; these are typically generated by field concatenation, which occurs when a missed delimiter causes two values to be captured in one cell. Then, deterministic edits ensure that the results in each section of the form balance. The edits are performed at two levels: the first level verifies that the totals in the Balance Sheet and the Income Statement match, while the second one ensures that the various sections of the Balance Sheet and the Income Statement also balance.

17. A series of many more edits are subsequently applied to identify various potential invalid or inconsistent values: overlap between fiscal periods, reallocation of negative values, consistency between financial and tax data, and balancing edits. Errors are corrected in a systematic fashion whenever

possible; manual corrections are applied otherwise. Deterministic imputation is used for three variables to ensure consistency between the different tax data sources available. Refer to Andrews, Hamel, Martineau and Rondeau (2007) to obtain further details on the edit and imputation strategy.

18. One important processing step that is unique to T2 data is the generic-to-detail allocation process. Corporations can report their data in generic fields, in detail or both. As mentioned earlier, only eight fields are compulsory for the CRA. The other fields are not checked. Statistics Canada, on the other hand, needs the details since T2 data are used to replace economic survey data. To meet users' requirements, the financial data are put through a generic-to-detail allocation process.

19. Businesses are assigned to imputation classes using models with a variety of variables where the classes correspond to the distribution of block totals over detail variables. Once imputation classes have been defined for businesses reporting details only, businesses for which imputation is needed (i.e. some amount is reported in the generic field) are divided into imputation classes using discriminant models. These discriminant models can be parametric or non-parametric. In the first case, a model based on the values of reported variables assigns each business to an imputation class. In the second case, the imputation class is based on the 15 closest neighbours in terms of the explanatory variables. Once imputation classes have been generated, ratios of all details are estimated for businesses with known distributions. Generic amounts are then assigned to details using the ratio distributions and the current assignment rules. Huang and Ladiray (2005) provide a good overview of the challenge faced in imputing the details, while Andrews, Brisebois and Hamel (2007) present options evaluated while recently implementing the new allocation methodology.

20. As for the T1 database, even if all tax data have not been received, the T2 database must ensure tax data are available for each unit on annual business surveys frames. When information is not received in time for the creation of the database, imputation is performed to provide values for the units with total nonresponse. These imputed values get overwritten with reported data when they are eventually received by the Agency.

21. The last processing step is the same as for the T1 tax data and consists of deriving the chart of accounts variables; some T2 fields are aggregated into derived variables used by surveys, on the basis of the variables in the chart of accounts.

B. Current challenges

22. T2 and T1 data face many of the same challenges, for instance, there are many other data sources available (associated to various schedules for example) which could be exploited to improve tax data processing. Efforts are also put into improving the coverage of the population, by investigating discrepancies between the coverage obtained from T2 tax data sources and Statistics Canada Business Register. As for T1 data there are several reasons why some units are either on the register but not in the tax source, and the opposite.

IV. Goods and Services Tax (GST)

A. Overview of the processing methodology

23. Instituted in 1991, the GST is a Canadian tax levied on the final consumption of products and services. Businesses collect the tax and remit it periodically to the CRA in the form of transactions. In some provinces, the GST is replaced by a harmonized sales tax that combines the GST and the provincial sales tax.

24. All businesses with annual revenues greater than $30,000 must register for a GST account and are required to file GST remittances. Enterprises only dealing with tax-exempt goods or services are exonerated. The frequency of remittance depends on the annual revenue of the business; the frequency can be either monthly, quarterly or annually. Businesses with annual revenue greater than $6M file monthly and businesses with annual revenue between $1,5M and $6M file quarterly. Businesses with

annual revenues between $30K and $1,5M are required to file annually. Quarterly and monthly filers are required to remit within 30 days of the period end, while annual filers must report within three months.

25. Each remittance, or transaction, consists mainly of the Business Number (BN), GST account number, period covered (start date and end date), sales and other revenue, the input tax credit and collected GST. Information related to the transactions is captured by the CRA which in turn, provides the data to Statistics Canada. Each year, Statistics Canada receives from the CRA approximately 9.3M transactions, covering approximately 2.8M businesses.

26. For a given reference month, the GST data are provided to Statistics Canada by the CRA seven weeks after the end of the reference month, at which time approximately 75% of the expected transactions have been received.

27. The processing steps for the GST consist of editing and imputation, calendarization of non-monthly transactions, and allocation to detailed lower-level business entities. Consult Brodeur and Pierre (2003) for a more detailed introduction to the GST database, as well as for an overview of the processing steps. The following paragraphs highlight the main steps.

28. The CRA and Statistics Canada do not use GST data for the same purposes. The CRA is interested in the amount of GST collected and remitted by businesses, while Statistics Canada is interested in the amount of sales (for statistical purposes). Since minimal editing is applied to sales at the CRA, it is essential to have in place an elaborate editing and imputation process. The process includes the verification of inconsistencies (in transaction dates for example), the identification of outlier or suspect values which could be the result of capture errors, and the imputation of the missing, inconsistent or outlier values. It is also during this first step that business industry codes are assigned, and that some businesses are declared as being inactive. Units are declared inactive as a result of information provided by the CRA or due to a prolonged period with non report. The edited and imputed data are then calendarized.

29. Reporting periods differ from one business to another. As described by Dubreuil, Hidiroglou and Pierre (2003), the objective of calendarization is to generate monthly estimates of sales for each business reporting GST on a non-monthly basis and taking into account industry specific seasonal patterns. Essentially, it involves benchmarking the GST data to the estimated seasonal pattern of the industry. Monthly estimates for period for which a transaction is not due, according to the unit's reporting frequency, are also generated in the calendarization process. These predicted monthly values are referred to as extrapolations. More details on the calendarization process are available in Quenneville, Cholette and Hidiroglou (2003), and most recent issues addressed are discussed in Beaulieu and Quenneville (2008). Once this process is completed, a database of perfect monthly transactions is made available to survey programs. One more process is essential if one is going to use GST data for complex structured businesses. This last step is referred to as allocation.

30. Businesses with complex structures can remit their GST transactions at an aggregated level. The reported amounts may include revenues from several industrial sectors and several provinces. These aggregated data need to be broken down into detailed levels in order to be usable by business surveys, in order to produce detailed indicators at the industrial and provincial levels. This process is referred to as allocation. In short, the allocation process redistributes the high level information to a finer industry-provincial level for complex structured businesses. Factors used in this allocation are based on information present in Statistic Canada's central business register. This information may in some cases be outdated or volatile and lead to an inaccurate allocation, which makes the use of these allocated data less appealing for the moment. The overall quality of this process has yet to be fully assessed.

31. One special aspect of the GST processing methodology is that the monthly process is not only run on new incoming data, but on historical data as well. This is done to incorporate all of the information received up to the current processing month. Although a large amount of transactions are received on time for a given month, some come in late. Reprocessing historical data using the most recent volume of transactions received greatly improves the accuracy of the data. In more practical terms, this means that a previously imputed data value can be revised with an updated imputed value, or with the actual

transaction value once received. When focusing on a given reference month, this translates into having dynamic data points, or more directly said, revisions. These revisions represent a clear source of variability between preliminary estimates typically published 30 days after the reference period for monthly surveys, and the revised ones published 30 days later.

B. Current challenges

32. Several improvements have been implemented in the GST processing strategy over recent years, based on feedback received from survey programs using these data. The improvements targeted the main path of data processing. Although they have increased the already high level of quality of the product, there are still some areas where fine-tuning is possible. Most of them deal with exceptional cases that can in some instances have an important impact on industry detailed estimates. Also, although the monthly updates of the GST database with new transactions received from CRA increase the quality of the data, these frequent revisions can cause problem to the users of the data. For example, the current strategy makes the assumption that the reported tax amount is always of better quality than the sales amount since it is the variable of interest for the CRA. If there were going to be incoherent or abnormal values, they would be more likely to be corrected for that field. However, there could be some situations where an inconsistency between reported tax and sales amounts could be the results of an erroneous tax amount. This is to be investigated. Relaxing the assumption that the tax amount is always of better quality than sales amount could lead to better overall quality in the data.

33. Another improvement currently investigated is for cases labelled as “gaps”. These are cases where reported transactions do not entirely cover a certain period of time, creating a gap. The current strategy simply nullifies all periods for which no information has been received from the CRA. These gaps were typically considered as cases of temporary period of inactivity for a business. However, recent information indicates that there are legitimate reasons to these other than inactivity for these gaps. For example, gaps can be created when a business is being audited by the CRA. In such a scenario, the business transactions for the period under investigations are not transmitted to Statistics Canada until the situation is resolved. Information that a business is being audited is not provided to Statistics Canada, and therefore, these gaps can not be pre-identified and dealt with consequently. The plan is develop a strategy where the decision to either impute a positive value or nullify a gap will be based, among others, on the status of the business at the time of the gap (active, pending closure, etc.) and the duration of the gap.

References

Andrews, J., Brisebois, F. and Hamel, N. (2007). “Methodology of Allocating Generic Field to its Details”, ICES-III, The Third International Conference on Establishment Surveys, 2007, Montreal, Canada. American Statistical Association, Alexandria (Virginia).

Andrews, J., Hamel, N., Martineau, P. and Rondeau, C. (2007). “Methodology for the Processing and Imputation of the Incorporated Enterprises (T2).” Internal document, Statistics Canada

Beaulieu, M. and Quenneville, B. (2008). “Calendarization of the Goods and Services Tax (GST) Data: Issues and Solutions”, 2008, Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, JSM 2008, Denver, CO.

Beaumont, J.-F., and Bissonnette, J. (2010). Variance Estimation under Composite Imputation: The Methodology Behind SEVANI. Survey Methodology (to appear).

Brodeur M. and Pierre L. (2003). “Use of Tax Data: An Application of Goods and Services Tax (GST) Data", Proceedings of Statistics Canada Symposium 2003, Statistics Canada.

Dubreuil, Hidiroglou and Pierre (2003). ‘’Use of Administrative Data in Modeling of the Monthly Survey Data’’, Proceedings of the Survey Methods Section, Statistical Society of Canada.

Huang, R. and Ladiray, D. (2005). Imputing Distributions in Administrative Tax Data. Proceedings of the 2005 Statistics Canada Symposium, Statistics Canada, Catalogue no. 11-522-XIE.

Quenneville, B., Cholette P., Hidiroglou, M. (2003). “Estimating Calendar Month Values from Data with Various Reporting Frequencies”, 2003 Proceedings of the Joint Statistical Meetings, Business and Economic Section, JSM 2003, San Francisco, CA.

Yung, W., Rancourt, E. and Hidirouglou, M. (2007). *Administrative Data in Statistics Canada’s Business Surveys : The Present and the Future*. Seminar on Registers in Statistics – Methodology and Quality. Helsinki, May 21-23, 2007.