

Seminar on Registers in Statistics - methodology and quality 21 - 23 May, 2007 Helsinki

Administrative Data in Statistics Canada's Business Surveys: The Present and the Future

*Wesley Yung, Eric Rancourt and Mike Hidioglou
Statistics Canada*

Wesley.Yung@statcan.ca, Eric.Rancourt@statcan.ca and Mike.Hidioglou@statcan.ca

Chapter 1.0 Introduction

In recent years, the use of tax data in Statistics Canada's annual and sub-annual business survey programs has increased to a great extent. 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 for smaller enterprises and a reduction of collection costs for the survey programs. Currently, the annual survey programs at Statistics Canada are using the annual T1/T2 tax data from the Canadian Revenue Agency (CRA), while the sub-annuals are using monthly Goods and Services Tax (GST) and Payroll Deduction (PD7 and PAYDAC) data from CRA.

In the majority of the business survey programs, tax data have been incorporated into existing sample designs and processing systems where possible. In order not to adversely affect the current production environment, the use of tax data has more or less been limited to treating them as data coming from another source and to treat the data as if they were reported by the respondent. This has allowed the business survey programs to continue using their existing systems and procedures with minimal changes. However, from a methodological point of view, this framework may not result in the most efficient use of the available tax data.

In this paper, we present the current framework of using tax data at Statistics Canada and the consequences of using this framework. We then present a new framework that could lead to a more efficient use of tax data. For both frameworks, we will present strengths and weaknesses and issues that remain unresolved. In the following section, we present some background information on the tax data currently being used. In Chapter 3, we present the current framework, followed by a proposed new framework in Chapter 4. In Chapter 5, we present some ideas on how the new framework could be implemented. Finally, some summary comments are given in Chapter 6.

Chapter 2.0 Tax Data

The principal sources of tax data can be broken down into annual data (T1/T2) and monthly data (GST, PD7 and PAYDAC). This distinction is helpful as the two sources are used differently by the business survey programs at Statistics Canada.

Chapter 2.1 Annual Tax Data (T1/T2)

Annual tax data are themselves divided into two sources, both available from CRA. Unincorporated businesses (sole proprietors and partners) file a T1 tax form, while incorporated (corporations) file a T2 tax form. The T1 tax forms are filed by individuals to report their income to CRA for the calendar year. They have until April 30th

of year y+1 to report to CRA. Minimal checks are performed at CRA and the file is then passed on to Statistics Canada where more detailed editing, imputation and outlier detection takes place. Currently Statistics Canada receives only the portion of the T1 population that files electronically (approximately 50%) but does receive a file called the Assessed Record File (ARF) containing two variables (gross business income and net income) for the entire population. For estimates of variables other than gross business and net income, model based methods are used to combine the information from the electronic filers and the ARF. Estimates are provided to the surveys in September of year y+1.

The other source of annual tax data, T2, covers incorporated businesses. Corporations are allowed to choose their own fiscal year but are expected to file their T2 tax information with CRA within six months of the end of their fiscal year. Because of this arrangement, CRA receives T2 tax data throughout the year and provides Statistics Canada monthly files containing T2 data for businesses that have filed in the previous month. Although CRA does some minor editing before sending the data to Statistics Canada, the majority of the processing is performed by Statistics Canada. Information from the business' financial information statements (income statement and balance sheet) are provided in the General Index of Financial Information (GIFI) format. This format is a coding system that allows corporations to prepare their financial statements and report them to CRA in a standardized manner. 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. Once a year, usually in September, a file is produced for the period covering April 1 of the previous year to March 31 of the current year. Prior to this release, units flagged for imputation are imputed through historical or donor imputation. Historically, the imputation rate has been 40 to 50%. Although this rate appears to be quite high, analyses have shown that the imputed data are very close to the reported data, once received. Once this processing has been completed, the tax data are ready to be used by the survey programs.

Chapter 2.2 Monthly Tax Data

Chapter 2.2.1 Goods and Service Tax (GST)

The GST is a 6% tax levied on all goods and services provided in Canada, with some exceptions. For instance there still exist some industries, such as renting, where the tax rate is 0%. In the provinces of New Brunswick, Nova Scotia and Newfoundland and Labrador, the GST is replaced by a harmonized sales tax of 14% that combines the GST and the provincial sales tax. With the exception of the province of Quebec, the GST is collected by the CRA.

All businesses with annual revenues greater than \$30,000 must register for a GST account and are required to file GST remittances. The frequency of remittance depends on their annual revenue. Businesses with annual revenue greater than \$6M file monthly and businesses with annual revenue between \$500K and \$6M file quarterly. Businesses with annual revenues between \$30K and \$500K 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.

Each remittance, or transaction, consists of the business' activity code, Business Number (BN), GST number, the expected filing frequency (monthly, quarterly or annually), period covered (start date and end date), sales and other revenue, the input tax credit and collected GST. Each year, Statistics Canada receives approximately 7.2M transactions, covering approximately 2.2M businesses, from the CRA. In terms of counts, most of these transactions are quarterly but in terms of sales, most are monthly (see Table 1).

Table 1. GST Transactions

	Business Counts	Transaction Counts	Sales
Monthly	6.4%	19.5%	77.1%
Quarterly	65.8%	71.4%	20.2%
Annually	27.8%	9.1%	2.8%

When using administrative data, one issue of concern is the timeliness of the data. This is of particular concern for monthly surveys. The GST data are provided to Statistics Canada by CRA six weeks after the end of the reference month, at which time approximately 30% of the expected transactions have been received. Thus, approximately 70% of the expected transactions need to be imputed or extrapolated using calendarization¹. The following month, CRA provides data for the next reference month as well as for all previous months. Ten weeks after the end of a given reference month, approximately 85% of the expected transactions have been received.

Chapter 2.2.2 Payroll Deduction Data (PD7 and PAYDAC)

In Canada, businesses are required to remit to CRA deductions withheld from employees. Examples of these deductions are deductions for the Canada Pension Plan, Employment Insurance premiums and income tax. Each remittance consists of the gross payroll of the business, the number of employees in the last pay period, the end of the remitting period and the amount paid, as well as information to identify the business. Depending on the size of the business' payroll, it is expected to remit quarterly, monthly or several times per month. Businesses with large payrolls remit one to five times per month depending on the type of payrolls used (i.e. monthly, weekly, biweekly, etc.). CRA provides two monthly files to Statistics Canada; the PD7 file contains the payroll deduction information described above and the PAYDAC file contains information on each of the businesses in the PD7 file. The information on the PAYDAC file includes whether a business is active or not, the type of payroll used in the business and frequency of response. As CRA does minimal processing of the PD7 data, Statistics Canada performs editing and imputation and converts the data into a form that is usable for its statistical programs.

Chapter 3.0 Use of Tax Data at Statistics Canada

Both annual and monthly tax data are being used at many stages of the business survey process at Statistics Canada. In this section, we go through the various stages of the survey process where tax data are being used and describe how they are being used within each of these stages.

Chapter 3.1 Sampling Frames

For the vast majority of business surveys at Statistics Canada, the Business Register (BR) forms the basis of the sampling frame. The BR is a register containing approximately 8.5 million businesses and covers all businesses in Canada. Of these 8.5 million businesses, approximately 2.3 million are known to be active. When a new business is started, a request for a Business Number (BN) is received by CRA which then assigns one to that business. The new BNs are sent to Statistics Canada along with a description of the proposed business activity. Using this description, Statistics Canada assigns a North American Industrial Classification System (NAICS) code and "births" the business on the BR. However, the request for a BN is simply a request and does not necessarily indicate that the business is active. Thus, variables on the BR such as Gross Business Income (GBI),

¹ Reference 1: Estimating Calendar Month Values from Data with Various Reporting Frequencies by Quenneville, B., Cholette, P. and Hidioglou, M.. *Proceedings of the Business and Economic Section*, American Statistical Association, 2003.

number of employees and salary and wages could be zero or missing and will fall into the approximately 6.2 million records that are not known to be active. Those businesses are typically not eligible to be selected in any of the business surveys. Confirmation of business activity could come from administrative sources such as GST, T1 or T2.

Chapter 3.2 Sample Design

For all active units on the BR, a Gross Business Income (GBI) is provided. For a very small portion of the units (corresponding to the larger units), the GBI value comes from profiling activities. For the remaining units, GBI is obtained either from GST data or a value is modeled from employment data (PD7). In the past, the majority of business surveys used GBI as a measure of size during stratification, but recently business surveys have been moving to using data from different sources such as reported survey data from a previous time period or data obtained from the annual tax program (T1/T2). If multiple sources of data are available, typically the largest value is used as the size measure.

Chapter 3.3 Collection

While tax data are not used directly in the collection process, the use of tax data does affect collection by reducing the number of respondents contacted. Currently at Statistics Canada, businesses that are called simple singles are identified as being eligible to have their survey data replaced by tax data. A simple single business is one that has a simple structure and has business activities in only one industry and one province. By a simple structure, we mean that there is a one-to-one link between the business and the tax data and that the tax data represent the business activity in only one industry and one province. Restricting ourselves to the simple singles for tax replacement allows us to ensure that the financial activity reported from tax data is in the correct industry and province.

Units identified as being eligible for tax replacement are treated differently depending on whether the survey is annual or monthly. In annual surveys, approximately 50% of the eligible units in the sample are not sent to the field for collection, but their data are obtained directly from tax data. Equivalence between the tax and survey concepts is ensured by a link between the survey questions and the GIFI formatted data through the Chart of Accounts (COA). In recent years, significant efforts have been made at Statistics Canada to establish and refine the COA. It is now at a level of quality that makes it fully usable by surveys. Once the information has been obtained from tax data, it is treated as a survey response for the remainder of the survey process.

For monthly surveys a couple of issues exist. First of all, the COA does not apply to the GST data and the GST data are not always available on time. Although equivalence of concepts is not guaranteed through the COA, it is felt that the GST data are very similar to revenue data that are typically collected by the monthly surveys. To account for the timeliness issue, a ratio model is used to link GST data for a previous month to the current reference month. That is, for a given reference month m , GST data available for month $m-1$ are used as a proxy for month m data but are adjusted through a ratio model to account for the different reference month. In order to calculate this ratio, survey data from a sample of simple singles for month m are needed, as well as \hat{X}_{m-1} , the weighted total of the GST data for month $m-1$ for the same units. This ratio could be calculated at different levels, for example the industry level, to take into account different behaviours across industries or groups. Responses are modelled for the simple singles identified for tax replacement by using this ratio and the corresponding GST value from the previous reference month. Currently, the typical monthly business survey using GST data is replacing approximately 50% of the eligible units. The remaining 50% are used to build the ratio model. As with the annual business surveys, once data are obtained via the GST data, the responses are treated as if they were obtained from the survey respondents.

Chapter 3.4 Edit and Imputation

For many years, tax data have been used at Statistics Canada for data confrontation purposes. That is, tax data have been used to validate survey responses and estimates. More recently, they have been used for editing and imputation purposes. If a survey response is not received, some units are being treated as described in the previous sub-section. That is, their response is being imputed either directly from tax data or obtained through modelling.

Chapter 3.5 Estimation

At the estimation stage, tax data are typically used to estimate the contribution from the Take-None (TN) portion of the population. In the late 1990's, in an effort to reduce response burden on smaller businesses, it has been decided not to contact any units that fell below a given threshold. These units were small businesses that did not contribute significantly to the overall estimate and were labelled as Take-None units. In order to estimate the contribution of these units, estimates are produced using tax data.

Payroll deduction data are used extensively in the estimation stage of the Survey of Employment, Payroll and Hours (SEPH) through the use of a model. Each month, a sample of approximately 11,000 units is surveyed to obtain the necessary variables to build models that can be applied to the PD7 data to produce a set of values for the entire population. One such variable is the Average Weekly Earnings (AWE) that is collected in the monthly survey but is not available on the PD7 file. The AWE variable is highly correlated to the Average Monthly Earnings (AME) that is available on the PD7 file. The monthly survey data are used to estimate regression parameters, which are then used to predict AWE for all observations on the PD7 file using the report AME. Estimates for various domains are then obtained from this census.

Chapter 3.6 Summary

As one can see, tax data are used extensively throughout the business survey process. However, they have mostly been incorporated into existing survey processes and thus their use has been somewhat limited in scope. For example, tax data that have been used to replace survey data (through design or because of non-response) have been treated as survey data to avoid changes to estimation and data validation systems. This strategy of replacement, and treating tax simply as another source of data, has been implemented this way partially because this allows the use of existing systems. Other uses could have been possible, one being using tax as an auxiliary variable for calibration purposes. One needs to question if tax data are being used to their full capacity in the present framework. In the following section, we assume that we are not restricted by this framework and consider how tax data could be used to their fullest extent.

Chapter 4.0 A Tax Based Framework

One way to use tax data to their fullest is to use the universe of tax data as the target population. Estimates could then be produced from this census of records. Under this approach, businesses would no longer be contacted as their financial information would be obtained from tax data. While this approach sounds easy to implement and should produce high quality estimates (assuming that the quality of tax data is high), there are many issues that need to be addressed before Statistics Canada would be prepared to change from a survey based framework to a tax based framework. Some of these issues are discussed in this section.

For the majority of business surveys at Statistics Canada, the BR has been used to define the survey population. If we are to shift to a tax based framework, there could be some differences between the survey population defined by the BR and the one defined by tax data. For example, for the complex enterprises, there is often not a one-to-one link between the legal structure (tax data) and the operational structure on the BR. These differences would need to be resolved so that any breaks in series due to a change in sampling frames or to population

coverage could be minimized. The best approach would be to still use the BR, but attach tax data to the simple singles.

Currently, the BR is kept up-to-date in terms of NAICS coding, status, structure and, to some extent, GBI through feedback from surveys. If a tax based framework were adopted, feedback from surveys covering financial data only would no longer be available since businesses would no longer be contacted to collect this data. In this case, another method of obtaining up-to-date information would be needed. One possibility would be to run a 'Nature of Business Report' (NBR) survey where a rotating sample of units on the BR would be contacted each year and information on their status, nature of business and structure of the business would be obtained. During this contact, information that would be available through tax data would usually not be asked. For large complex units, the contact could be more often to ensure that the most up-to-date information is available. This could be similar to the profiling exercise that is currently done for the larger units.

Contact of the large units would be important as tax information may not provide enough details. For example, tax information may not be available at the establishment level but only at a higher level. For some businesses that are active in different industries or different provinces, this might mean that the business activity for each of the industries might not be available from tax data. Thus, correct estimates at the industry and/or province level will not be possible. Reliable methods of allocating consolidated data from the higher level to a lower level will need to be developed. Some allocation methods based on employment data exist, but the quality of the allocation is questionable. If the tax based framework is to be adopted, information necessary to perform high quality allocations could be collected during the NBR survey.

By shifting to a tax based framework, one would minimize sampling error (in fact eliminate it) but may increase non-sampling error due to the imputation and the processing of the tax data. Currently, the amount of imputation differs greatly from one variable to another. For the main tax variables, imputation rates are very low, but for some other variables, the imputation rate can reach 50%, especially for variables where a generic-to-detail allocation needs to be performed. In order to report correct quality measures, the non-sampling errors within the tax data need to be quantified.

One of the biggest problems with shifting to a tax based framework is that characteristics data, such as commodities produced, would not be available. While some business surveys produce estimates of financial variables only, some produce estimates of characteristics also. Some work on modeling characteristics from available tax data has been done, but with little success. A survey would need to be conducted to obtain characteristics data, as well as information on structure/nature of business and, in some cases, additional financial information as well.

Under a purely tax based framework, Statistics Canada would be totally dependent on administrative (tax) data. While this can be envisioned in a perfectly designed reception-processing system, it nonetheless would constitute a vulnerability in that the control over changes to tax data would be external. If independence and control is to be retained, the ability to conduct surveys (for both financial and non-financial information) must be maintained so that in case of delays or disruption of the delivery of tax files, the statistical program can be continued.

Even if a new framework were to be based on tax data, the need to maintain an accurate and complete list of units and information to contact them would remain at the core of the economic statistics program. That is why the BR's role would remain central even if the framework were shifted towards tax data. Tax data may provide a greater coverage, but the BR is the sampling frame for surveying businesses regardless of whether its coverage matches the tax data coverage.

An important challenge in the efforts of increasing the use of tax data is the acceptance by users, analysts and those involved in data processing. Attaining a level of comfort by all involved will be achieved by getting people

to know more about tax data and by gradually increasing their time spent on working with tax data. In fact, with the introduction of the tax replacement program, Statistics Canada has forced a large number of people to learn much more about tax data, thereby bringing us to the current point where the conditions are met to increase the use of tax data in a much more fundamental way than by simply increasing the percentage of replacement of units in the samples.

Defining a target strategy is relatively simple given the set of tax data available. However, the key to reaching this goal will be how the transition is handled. As one can see, there are some non-trivial issues that need to be resolved before moving to a purely tax based framework.

Finally, one point that has not been mentioned but needs to be made is that a switch to a tax based framework would almost certainly mean that breaks would be observed in published series. Some of these breaks could be substantial and there must be a willingness by business survey programs and clients to accept these disruptions as a price to improving quality and capacity.

In the meantime, it may be possible to have a framework that combines the survey and tax based frameworks. One possible methodology to achieve this is given in the next section.

Chapter 5.0 Combined Framework

In this section, we outline a possible methodology that would combine the survey and tax based frameworks, and that could be implemented without major changes to existing systems. This methodology could be thought of as a transition between the existing survey based framework and a tax based framework while the issues discussed in Chapter 4 are resolved. Given that annual and monthly surveys use different tax sources and different methodologies to incorporate the tax data, they will be considered separately.

Chapter 5.1 Annual Surveys

In Chapter 3.3, the concept of a simple single was discussed. That is, a simple single is a business for which one can reliably obtain all financial information from tax data. Currently only a portion of the simple singles is being replaced with tax data. The rest are either being contacted by the survey or are being accounted for through survey weights. One way to increase the use of tax data would be to use tax data for all simple singles in the tax universe. In essence, this would imply stratifying the sampling frame first based on whether or not the unit is a simple single. Note that this assumes that the simple singles on the sampling frame coincide with the simple singles in the tax universe. For convenience, we will call complex units the ones that are not simple singles, although some of these complex units may have a fairly simple structure. Estimation of the complex universe could be achieved through the existing survey framework. That is, the complex units could be further stratified, if desired, and a sample of units could be used to estimate for this population. Tax data would be used in some stages of the survey process, as is currently done, and methods to further incorporate tax data, such as calibration and better allocation methods, should also continue to be investigated.

Turning back to the simple singles, it might be necessary to further divide them into incorporated and non-incorporated businesses because of the difference in the availability of T1 and T2 tax data. T2 data are available for all incorporated businesses in Canada, while T1 are available for only a portion of the non-incorporated businesses. Since 2006, only the T1 businesses that report to CRA in electronic form are available to Statistics Canada. Due to this change, the estimation system was modified slightly since the non-random set of paper filers needed to be appropriately accounted for by the electronic filers sample.

As previously mentioned, moving to a tax based framework would mean that information on characteristics, status, industry and structure of the business may not be up-to-date. Because of this, some contact of the simple

singles would be necessary. What we envision is a rotating sample from which the status, the structure and the nature of business of the business would be obtained. For those surveys needing characteristic data, a sample of simple singles could be selected and characteristic data could be collected at the same time. Some co-ordination would be required as the simple single population would cut across different surveys. In addition, if the same contact were to be used to collect characteristic data, then the timing of the contact would need to be co-ordinated to meet the deadlines of the affected surveys.

Chapter 5.2 Monthly Surveys

As with the annual surveys, until some of the issues discussed in Chapter 4 are resolved, increasing the use of tax data for complex units is somewhat limited. Thus, as mentioned above, work should continue on improving the allocation of tax data and calibration methods using tax as an auxiliary variable.

For the simple singles, tax data cannot be used directly as in the annual surveys due to timeliness and conceptual issues as discussed in Chapter 3.3. While the percentage of eligible simple singles that are tax replaced could be increased from the current 50%, there will always be a need to have some simple singles in the sample in order to estimate the ratio that is used to account for the time and conceptual differences. If there is to be an increase in the percentage of units that are tax replaced, two points needed to be kept in mind: an adequate sample of simple singles is needed to reliably estimate the necessary ratios and that by increasing the percentage of tax replaced units, one is trading sampling variability for non-sampling variability. This trade-off needs to be carefully evaluated.

Thus, for the monthly surveys it would not be possible to use tax data for all simple singles in the tax universe but the portion could be larger than today's use. The problem of status, industry, structure and characteristics experienced by annual surveys also occurs for the monthly surveys. However, for the monthly surveys the solution proposed in Chapter 5.1 could be simplified by the fact that some units will have to be contacted for calculating the ratio adjustment. During this contact, up-to-date information on status, values of business industrial activity and structure should also be obtained. Some method of rotation should be implemented so that eventually the entire simple single universe would be covered after a period of time.

Chapter 6.0 Summary

The use of tax data in Statistics Canada's business survey programs has increased significantly in the recent past and there is a willingness to use the administrative sources even more in the future. However, since tax data have been introduced into existing business survey programs and systems, their use has been somewhat constrained. In this paper, we have considered how tax data could be used if these constraints were removed, but we have pointed out that there are still some issues that need to be resolved before Statistics Canada can move to a purely tax based framework

A possible methodology, that can be viewed as a transition between the existing survey based framework and a tax based framework, has been proposed. This proposed methodology would increase the use of tax data and could be implemented while work continues on resolving the issues that are not permitting the use of the tax based framework. Once these issues are resolved, the proposed methodology could easily be adapted to the tax based framework.

References

Quenneville, B., Cholette, P. and Hidirolou, M. (2003). Estimating Calendar Month Values from Data with Various Reporting Frequencies. *Proceedings of the Business and Economic Section*, American Statistical Association, 2003.