

# Statistical Use of Goods and Services Tax Data in Statistics Canada's Monthly Economic Surveys

Louis Pierre<sup>1</sup>, Marie Brodeur<sup>2</sup>

## ABSTRACT

In a rapidly changing world, policy analysts and decision-makers on the economic scene need quality information if they are to understand economic activity and react quickly. The monthly economic statistics program of Statistics Canada (STC) helps provide this information in a timely fashion. However, the collection of monthly data imposes a significant response burden on Canadian businesses.

With the development of new methodological approaches and processing capabilities based on information technology, STC is increasingly using existing sources of administrative data such as the GST (goods and services tax) files to produce quality economic estimates while reducing the response burden. In Canada, the GST is a tax levied on the consumption of goods and services and is accounted for on the value added at each stage of production. However, the use of administrative data poses major challenges. For example, sophisticated edit and imputation programs must be developed in order to detect and correct various problems that might exist in the raw data. These programs must be very efficient to process millions of records in a timely fashion. Also, administrative data are not all available in time on a monthly basis. It has therefore been necessary to develop methodologies for calendarization and techniques that combine survey data with GST data. This paper will give an overview of these methodologies, focusing on recent, real-life applications.

## 1. INTRODUCTION TO THE GST PROJECT

For several years, Statistics Canada (SC) has been using tax data in its statistical programs, mainly for its annual surveys. In 2002, Statistics Canada launched the Strategic Streamlining Initiative. One of the main objectives of this initiative is to promote expanded use and better integration of tax data in economic statistical programs. More specifically, the goal is to reduce response burden and data collection costs and to obtain new, higher quality statistical data.

Under the Strategic Streamlining Initiative one of the projects undertaken by the Tax Data Division involves use of Goods and Services Tax (GST) data. This article will deal with the GST project. GST-generated data are now accessible to Statistics Canada in the form of a database and the objective is the replacement of simple establishments in sub-annual surveys. Statistics Canada has signed an 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.*

### What is GST?

The Goods and Services Tax is a tax levied on the consumption of goods and services (supplies) in Canada. For instance, goods and services exported outside Canada are exempted while those imported to Canada are taxed. The tax rate on supplies is 7% in non-harmonized provinces and 15% in three harmonized provinces (Newfoundland, Nova Scotia and New Brunswick). There are also zero-rated supplies. GST is a multi-stage tax, which means that the tax is accounted for on the value added of goods and services.

Use of GST-generated data has always been of significant interest to SC. In 2000, negotiations began with CRA to obtain data on a monthly basis and a database was created. An editing and imputation system also had to be designed. One of the major challenges with using GST data is that remitters must file monthly if the company is

<sup>1</sup> Louis Pierre, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6, [louis.pierre@statcan.ca](mailto:louis.pierre@statcan.ca)

<sup>2</sup> Marie Brodeur, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6, [marie.brodeur@statcan.ca](mailto:marie.brodeur@statcan.ca)

relatively large, quarterly if it is of average size and annually for small companies. Quarterly and annual remitters account for 93% of businesses but for only 23% of total revenue in the economy. The objective of the GST project is to replace 50% of the simple establishments in monthly survey samples by the GST. Simple enterprises consist of a single establishment and therefore we will use the term “establishment” in this article. The majority of these establishments are quarterly GST remitters. The major challenge in using GST-generated data is therefore to create monthly data from quarterly data. A calendarization process is used to this end. This is followed by an estimation system that combines survey data and GST data. The sections below deal with each of these aspects.

## 2. DESCRIPTION OF THE DATABASE

Once a month, CRA provides SC with two files, the first containing the reports from each company (transactions) and the second containing the characteristics of the GST accounts.

The monthly transaction file includes transaction updates (additions, changes) for a period covering three to four years. SC considers this file a raw data file. Raw data for the current month, the preceding month and data going back six months go through a complex editing and imputation program (E&I) from which an historical “edited” file is created containing edited data and all relevant information on the detection of outlier values and on imputation. This file then becomes the input to the data calendarization program. The Tax Data Division (TDD) gives itself five working days to process these millions of transactions each month and to make them available to client divisions. The final file essentially contains an estimate of the revenue of each enterprise (at the business number level) on the basis of a calendar month. Appendix A contains a diagram of the structure of the GST database that operates on an Oracle platform. The database includes data from 1998 forward. Current data are available two months after the reference month.

CRA updates its GST files once a week (weekend processing). SC had to determine the optimal date for receiving these files, one that was a compromise between obtaining the maximum number of transactions and the need to receive them as soon as possible to prepare the statistics for our monthly

surveys. At Statistics Canada, monthly survey release dates are set in advance and generally correspond to six weeks after the end of the reference month. Since companies normally submit their reports to CCRA one month after the end of the reference month, it quickly became apparent that the transactions for the reference month would never be available on time. This meant that we needed to develop a model that could use the data from the previous month, what is referred to as “m-1”. As we will see later, the proposed model uses GST data in the form of a ratio. It is therefore perfectly valid to use the “m-1” month, knowing that there is sufficiently strong correlation between the survey data and the GST data.

We therefore asked CRA to provide us with its files seven weeks after the reference month. After seven weeks, the files include about 90% of the expected transactions, in terms of sales, whereas if we had decided on six weeks, this percentage would have varied between 45% and 85%. Unexpected transactions are imputed (extrapolated) by the calendarization module. This is the case with annual and quarterly transactions. For example, CRA does not expect to receive a return from a company whose fiscal year ends in December for the following eleven months of the year. However, we need an estimate for each of these months in our database. The same holds true for a company that submits transactions quarterly. For example, we have to produce an estimate for the months of April and May for a company that files for the quarter from January to March. The total rate of estimation (imputation and extrapolation) for a given month varies between 20% and 30% after seven weeks. This rate can also vary from industry to industry. The eight-week option was rejected because it did not fit with the production time lines of the client divisions.

## 3. EDIT AND IMPUTATION

CRA and SC do not use GST data for the same purposes. The main variable of interest for CRA, for example, is the amount of GST, while for SC, it is the amount of sales. CRA does not have a program to produce data that can be used directly for statistical purposes. It was therefore essential to put in place an editing, outlier detection and imputation program. In early 2000, a GST data processing system was developed. When we began using the data, we found a number of shortcomings in the editing, outlier detection and imputation functions. As an example, when detecting outlier values, we were comparing

transactions that could have different durations. In addition, the choice of imputation methods was based on a criterion of least variance. This strategy meant that a majority of imputations was done using the stratum mean, which is not necessarily appropriate. Consequently, the decision was made to change the imputation strategy and take advantage of this change to review other processing methods. More detail on the former strategy can be found by consulting Hamel and Lothian (2002). The sections below provide an overview of the new program.

### **3.1 Pre-processing**

The first objective of this module is to resolve inconsistencies in transaction dates and multiple transactions. The data also needs to be converted on a daily basis for analysis and comparison purposes. This latter operation is a major improvement over the previous version of the editing and imputation modules. The reporting frequency of each transaction is determined and the estimate of annual revenue is updated on a continuous basis. The reporting frequency class and the estimate of annual revenue are used to build the strata needed to process the GST data. Lastly, the medians for revenue, GST and tax rate are calculated for each unit.

### **3.2 Definition of bounds for outlier detection**

This module determines the categories of estimated annual revenue. The categories, redefined in the second generation of specifications, are linked to industry groups and classes of reporting frequency to build the strata. The strata are in turn regrouped into classes. For each stratum, we calculate the median and the quartiles that are then used to determine the limits for outlier detection. A special methodology was developed for the tests associated with verifying acceptable growth rates. To find out more about this methodology, consult Hidioglou-Berthelot (1986). All of the parameters used in processing the data, such as an acceptable maximum tax rate, are presented in this module.

### **3.3 Outlier detection**

Outlier detection is the result of a combination of cross-sectional and longitudinal tests on standardized data (daily averages). Tests based on tax rates predominate, except for raw transactions with “near 0%” tax rates. There are also a few complementary tests based on levels and on the growth in the two

variables of interest. Comparisons are made with data from month “m-12”. The final step is a complex combination of all of these tests to determine the outliers that are defined as suspect or critical. Critical values will be imputed, while suspect variables will not be. However, suspect and critical values will be removed from the calculation of strata means for imputation purposes.

### **3.4 Definition of transactions to impute**

The final process is to determine which transactions to impute. We impute critical outliers, missing revenue and expected, but late, transactions. In the latter case, we will first determine if a unit is late or dead based on a strategy to avoid overestimation. Unexpected transactions will be extrapolated in the calendarization module.

### **3.5 Imputation strategy**

In the first version of the editing and imputation modules, we selected from five imputation methods the one that had the least variance. The method using the stratum mean was the one selected most often. The new imputation strategy consists of selecting the imputation method from two decision tables, one for revenue and the other for the GST. There are seven imputation models, three of which are based on the other available variable (revenue or GST). This has the advantage of preserving a consistent tax rate (80% of imputations are based on one of these three methods). However, these three methods are not used for units or industries that have tax rates around 0% because there is no relation between the two variables. In cases where both variables must be imputed, revenue is imputed first.

## **4. CALENDARIZATION**

As mentioned in the introduction, calendarization of the data is very important to the successful use of the GST data because the goal is to replace monthly survey data. It is therefore important to design a database to meet this objective. Transactions generated by the E&I program can have reference periods of different lengths even if they are assigned a reporting frequency class (monthly, quarterly, annual). The purpose of calendarization is to generate an estimate of the two variables of interest that exactly matches the tax months and to do so for a specified number of months. If segments of time are not covered by transactions, these periods will be

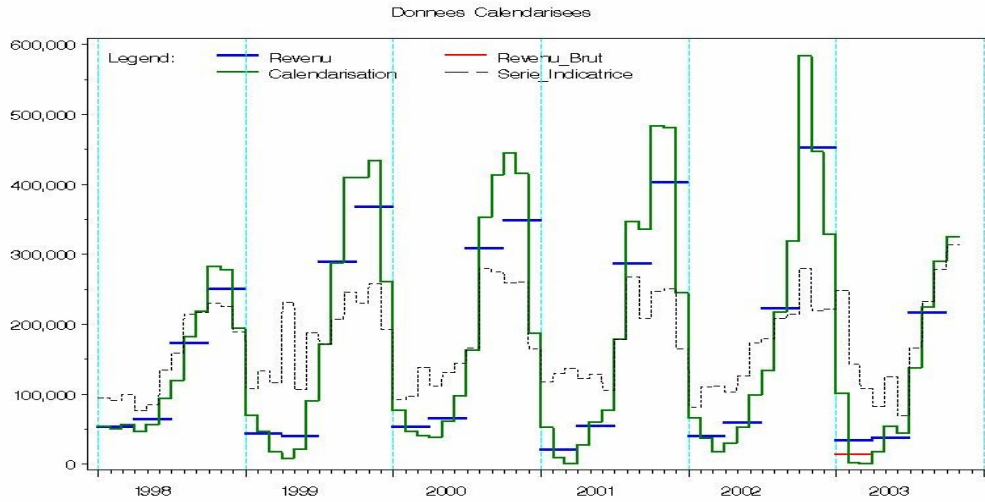
interpolated and extrapolated by the calendarization program, unless these periods were voluntarily nullified in advance (temporarily inactive or deceased units).

Calendarization uses a proportional method developed by Denton and adapted to transactions with variable lengths. Essentially, it involves benchmarking the GST data to a monthly indicator series re-scaled to a given enterprise. Indicator series

are currently produced at the national level for each industry based on the six-digit North American Industry Classification System (NAICS) using monthly or quasi-monthly transactions.

Graph 1 shows the relations between the indicator series, GST data and calendarized data for a unit that reports quarterly. Persons interested in more detail on calendarization can consult the article by Quenneville, Cholette and Hidirolou (2003).

**Graph 1**



## 5. ESTIMATION

The purpose of the GST project is to replace data from sampled single establishments in sub-annual surveys. As mentioned in Section 2, TDD receives data seven weeks after the end of the reference period. At that point, the SC monthly surveys are about to release their data. As a result, direct data replacement is not a viable option. It was necessary to design a model that combines the survey data from the current month with the GST data of the previous month. Two ratio-based models were developed. The first model is called the ‘Macro’ model and involves a calibration on the ratio of survey data and GST data at the population level. The second model is the ‘Micro’ model and involves imputation by ratio. It is the same ratio as in the previous model but calculated at the sample level and applied to the microdata.

The estimate for the Micro approach uses the following formula:

$$\hat{Y} = \sum_{S_{N-E}} w_k y_k + \sum_{S_1} w_k y_k + \sum_{S_2} w_k y_k^*$$

For those units in  $S_2$

$$y_k^* = (\hat{Y}_{S_1} / \hat{X}_{S_1}) x_k$$

where

$$\hat{Y}_{S_1} = \sum_{S_1} w_k y_k,$$

and

$$\hat{X}_{S_1} = \sum_{S_1} w_k x_k$$

The estimate for the Macro approach uses the following formula:

$$\hat{Y} = \sum_{S_{N-E}} w_k y_k + \hat{Y}_{S_1,NGST} \frac{N_{NGST}}{\hat{N}_{S_1,NGST}} + \hat{Y}_{S_1,GST} \frac{X_{GST}}{\hat{X}_{S_1,GST}}$$

Further details on the models are available in Dubreuil, Hidirolou and Pierre (2003).

To date, the GST data have been used for two surveys, the Monthly Restaurant, Caterers and Taverns Survey (MRCTS) and the Monthly Survey of Manufacturing (MSM). For each survey, historical simulations covering several months were carried out to properly measure the impact of seasonal variations. The simulations reflected real conditions for the production of GST data as they are received

after seven weeks, including at the data processing and ratio calculation levels. The next two subsections deal with the results obtained.

### 5.1 Monthly Restaurant, Caterers and Taverns Survey

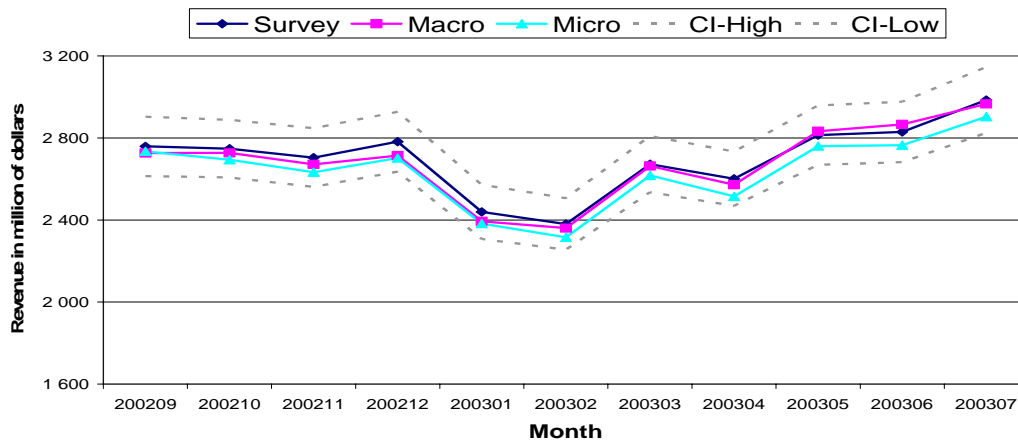
Approximately 70% of the estimate of sales from the MRCTS comes from single establishments. This illustrates the potential importance of replacing 50% of these establishments with GST data. The survey's response rate is about 70% and there is a great deal of total and partial imputation. The survey was last redesigned in 1994 and the sample design is composed of several strata. However, 55% of the estimate comes from four strata that represent restaurants in Ontario and Quebec. This means that there are several small strata, especially in the drinking places (alcoholic beverages) industry, by

province. There are also many units that are not classified in the proper stratum, which results in several strata with quite high coefficients of variation. For these reasons, the decision was made to replace only 34% of single establishments in this survey. However, the impact on the survey's total estimates is 42%.

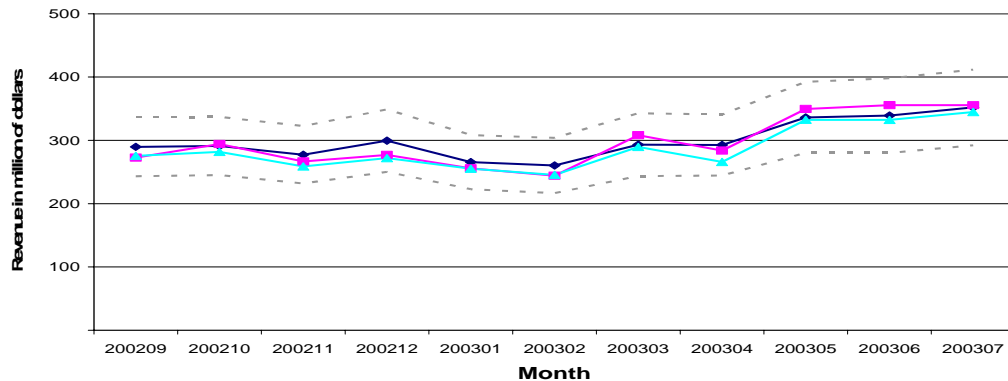
Graph 2 shows that the estimates are excellent at a national level for both the Macro and Micro models. The dotted lines represent the confidence interval (CI) at 95% of the estimates published by the MRCTS. The estimates are also excellent for the big strata such as the one in Graph 3. Graph 4, however, gives an example of a small stratum. We can see that the Macro estimator does not perform as well and that, in particular, the trend from one month to another is sometimes very different.

#### Results of the MRCTS simulations from September 2002 to July 2003

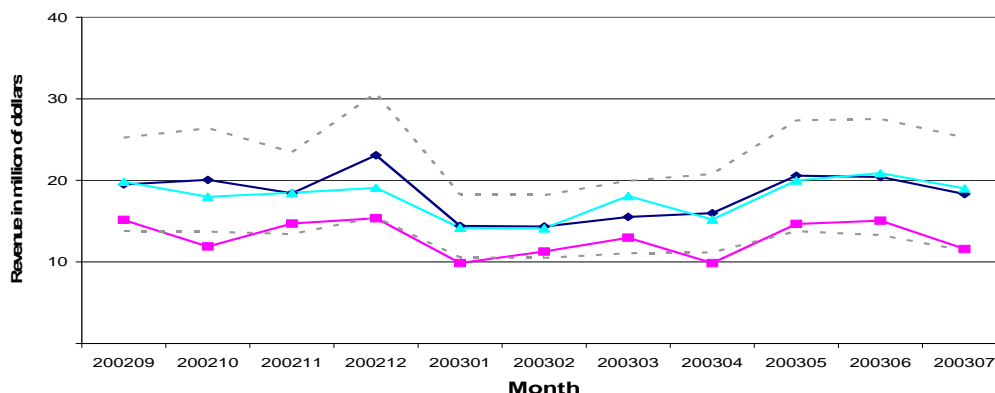
**Graph 2  
All enterprises in Canada**



**Graph 3 - Quebec  
Full-services restaurants**



**Graph 4 - Quebec  
Caterers and mobiles food services**



Since sub-annual surveys are used in particular as indicators of trends, it is important to select a model that preserves the latter. We therefore decided to use the Micro model. Another advantage of the Micro model is that it is very easy to put in place because it involves adding information to the survey microdata file. This means that it is possible to use existing computer programs and simply to include a few additional analytical tools.

The MRCTS therefore opted to proceed and to replace 34% of its sampled single establishments with the Micro model. A parallel test was done between October 2003 and February 2004, with implementation in July 2004 for the survey's May reference month.

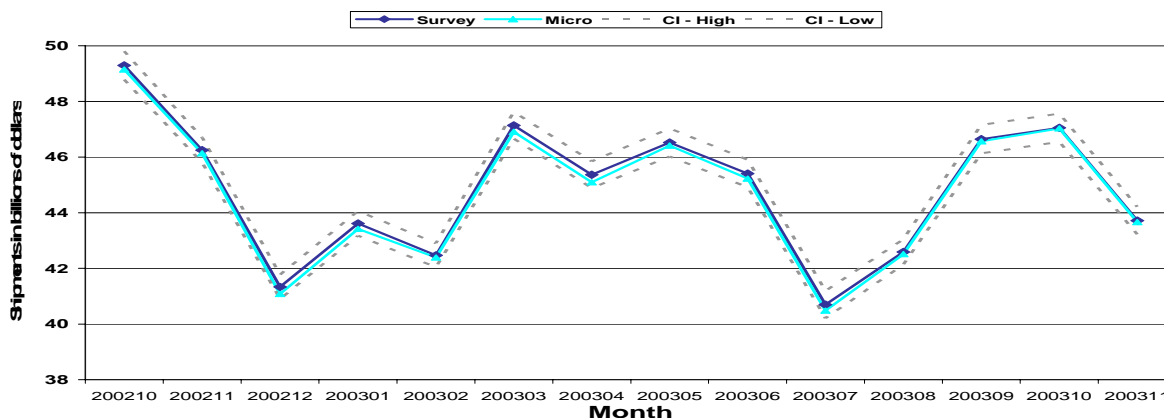
## 5.2 Monthly Survey of Manufacturing

The challenge with this survey is slightly different. First, the aim was to replace the sales from 50% of single establishments but all of the single establishments combined represent only 20% of the

final estimate. The impact is therefore less significant in the context of this survey. However, the MSM also collects data on inventories and that information is not available in the GST files. The challenge was to find a solution for those establishments that would be replaced by the GST and for which no further data would be collected.

The MSM was last redesigned in 1999 and the sample was re-stratified in 2003. However, MSM is benchmark with the Annual Survey of Manufactures. When estimating the MSM, the weight of the survey must be considered with the weight of the benchmarking. Using the Macro model would require adding a third weight. For this reason, it was decided to test only the Micro model. The correlation between GST sales and the survey's shipments is 80%. Graph 5 presents the estimates as published by the MSM along with those of the Micro model. It is obvious that the two lines follow the same trend and the estimates are very close. Similar results are observed for all industry codes and all provinces taken separately.

**Graph 5 - Shipments - All Manufactures**



The MSM therefore decided to proceed and to replace 50% of its sampled single establishments. Since the results were promising, implementation was advanced. A parallel test has taken place between May and August 2004 and everything should be in place for September 2004 (July reference month).

## 6. ISSUES

The GST project met its planned objectives earlier than expected in the case of the MSM. The response burden of small establishments will be reduced along with the survey collection costs. There are still a number of lessons to learn from this experience, which is far from over. First, before using tax data, processing is required. Editing and outlier detection rules are needed. It is also important to have an imputation strategy adapted to use requirements. The GST data will be used on a longitudinal basis and historical imputation methods respond well to this challenge. Calendarization is definitely a key element that made use of the data on a monthly basis possible. Although we are very satisfied with this process, it will be important to measure the impact on the data over time.

Second, it is not always easy to replace units in sample designs where the units are improperly classified as is the case with the MRCTS. The ideal would be for the use of administrative data to be taken into consideration in future when redesigning a survey or designing a new survey.

Third, the Micro model gives very satisfactory results. However, its biggest advantage is its great simplicity since all users can understand it and explain it easily. It is also very easy to implement operationally. However, we will have to re-evaluate use of the Micro model over time. The Macro model is not as sensitive to classification problems and it also allows for use of the full power of the population data and to reduce the coefficient of variation.

Finally, in 2004, the GST project will undertake the replacement of 50% of the single establishments in monthly wholesale and retail trade surveys. These two surveys were just redesigned in 2003. However, single establishments account for 60% of total sales in the retail sector and 45% of total sales in the wholesale sector. Consequently, this will be an extensive replacement of data by the GST model.

## ACKNOWLEDGEMENTS

The authors thank François Maranda and Jocelyn Tourigny for their excellent comments. They also thank Roxane Payeur and Lucy Chung for the preparation of the graphs.

## REFERENCES

- Dubreuil, G., Hidioglou, M.A. and Pierre L. (2003), "Use of Administrative Data in Modeling of the Monthly Survey Data", Proceedings of the Survey Methods Section, Statistical Society of Canada.
- Hamel, N. et Lothian, J. (2002), "L'utilisation du Jackknife pour l'imputation des données administratives", conférence, 3<sup>ième</sup> Colloque francophone sur les sondages.
- Hidioglou, M.A. and Berthelot, J.-M. (1986), "Statistical Editing and Imputing for Periodic Business Surveys", Techniques d'enquêtes, Juin 1986, Vol. 12, No.1, pp. 73-83, Statistique Canada.
- Quenneville, B., Cholette, P. and Hidioglou, M. (2003), "Estimating Calendar Month Values from Data with Various Reporting Frequencies", Proceedings of the Business and Economic Section of the American Statistical Association.

# Appendix A

