Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

USE OF TAX DATA: AN APPLICATION OF GOODS AND SERVICES TAX (GST) DATA

Marie Brodeur and Louis Pierre[1]

ABSTRACT

For several years, Statistics Canada has been using tax data as direct replacement in imputation or estimation or as a data certification tool. A new project has been launched to use data from the Goods and Services Tax (GST) more extensively. The project's goal is to reduce the response burden and collection costs of sub-annual surveys at Statistics Canada. This article provides an overview of the work to date using GST data at Statistics Canada. We will provide a general description of the GST project and more specifically, of the development of the database, design of the editing and imputation system, calendarization and the estimation model. We will also discuss the key challenges in designing this type of project.

KEYWORDS: Calendarization; Estimation; GST; Imputation.

1. INTRODUCTION TO THE GST PROJECT

For several years, Statistics Canada (SC) has been using tax data in its social and economic statistical programs. The tax data are used as direct replacement in imputation or estimation or as a data certification tool. Statistics Canada recently took a large step forward with the Strategic Streamlining Initiative launched in 2002. The main objective 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. A further objective is to develop new annual and sub-annual indicators not currently possible using survey data.

The Tax Data Division is involved in two key development projects linked to the Strategic Streamlining Initiative. The first is the T1/T2 project to use data from T1 (individuals) and T2 (corporations) tax returns to replace simple unit survey data from annual surveys associated with the Unified Enterprise Survey. The second project 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 to begin implementation and replacement of single establishments in sub-annual surveys. Statistics Canada has signed an agreement with the Canada Customs and Revenue Agency (CCRA) 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*.

Use of GST-generated data has always been of significant interest to SC but it has not had the necessary resources to carry out the analysis. In 2000, negotiations began with CCRA 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 relatively large, quarterly if it is of average size and annually for small companies. Quarterly and annual remitters account for 90% of businesses but for only 22% of total revenue in the economy. The objective of the GST project is to replace 50% of the single 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 to combine survey data and GST data. The sections below deal with each of these aspects.

[1]Marie Brodeur and Louis Pierre, Statistics Canada, Ottawa, Ontario, Canada, K1A OT6, marie.brodeur@statcan.ca, louis.pierre@statcan.ca

2. DESCRIPTION OF THE DATABASE

Once a month, CCRA 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 gross data file. Gross 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 "statistics" 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 enterprise 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.

CCRA 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 distribution 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 CCRA 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, CCRA 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 for 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. EDITING AND IMPUTATION

CCRA and SC do not use GST data for the same purposes. The main variable of interest for CCRA, for example, is the amount of GST, while for SC, it is the amount of sales. CCRA 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 lengths. 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 average, 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 limits 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 Hidiroglou-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 gross transactions with tax rates calculated around “0%”. 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. In contrast, suspect and key values will be removed from the calculation of strata averages for imputation purposes.

3.4 Definition of transactions to impute

The final process is to determine which transactions to impute. We impute key outliers, missing revenue and expected, but late, transactions. In the latter case, we will first determine if a unit is late or deceased 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 average 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

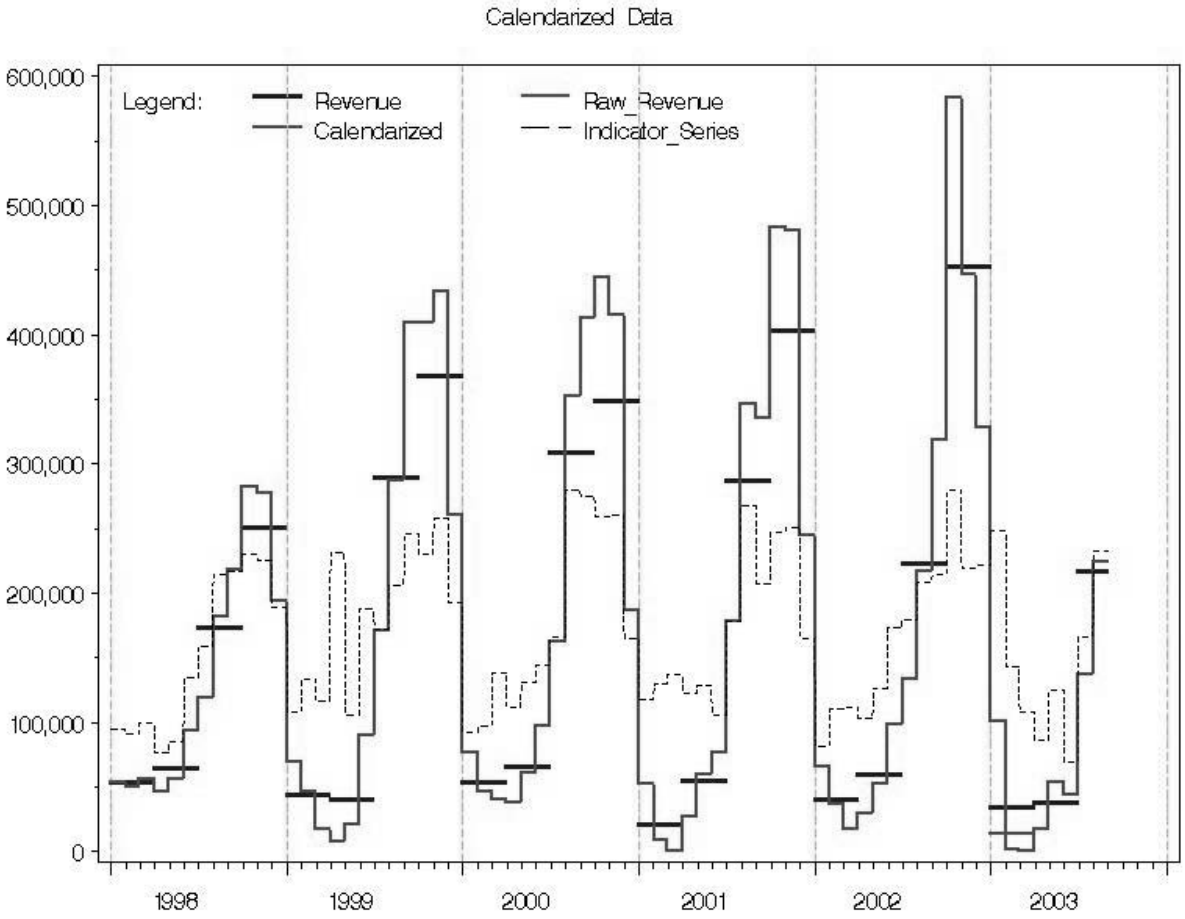
As mentioned in the introduction, calendarization of the data is very important to the successful use of the GST data because the goal is to replace monthly survey data. It is therefore important to design a database to meet this objective. Transactions generated by the E&I program can have reference 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 re-staged monthly indicator series for 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 Hidiroglou (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. Further details on the models are available in Dubreuil, Hidiroglou and Pierre (2003).

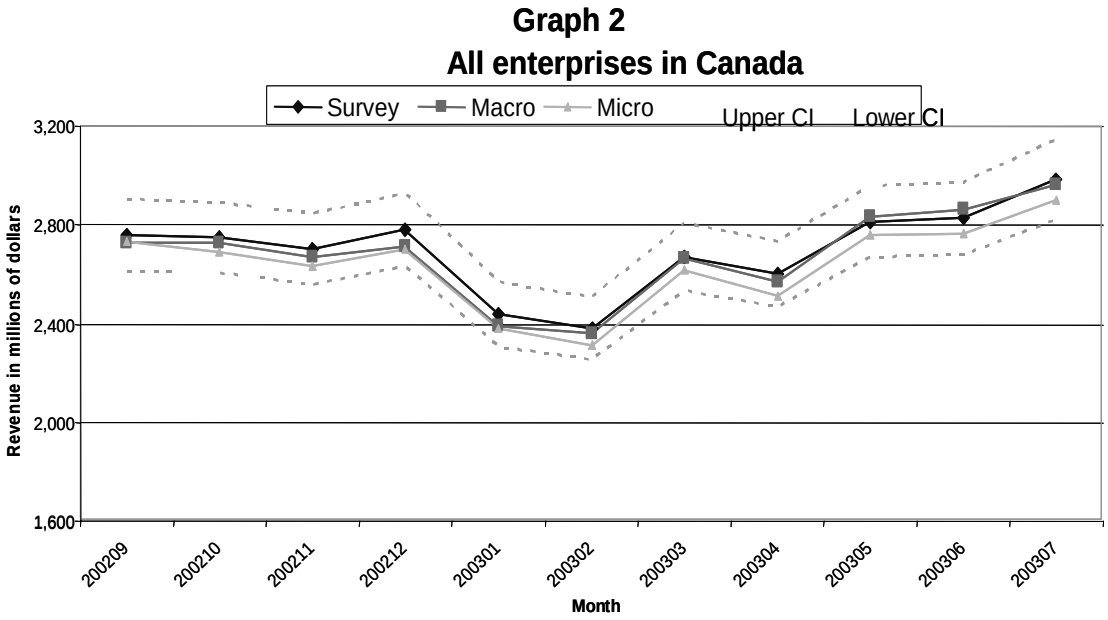
To date, the GST data have been used for two surveys -- the Monthly Restaurant, Caterers and Taverns Survey (MRCTS) and the Monthly Survey of Manufacturing (MSM). For each survey, historical simulations covering several months were carried out to properly measure the impact of seasonal variations. The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

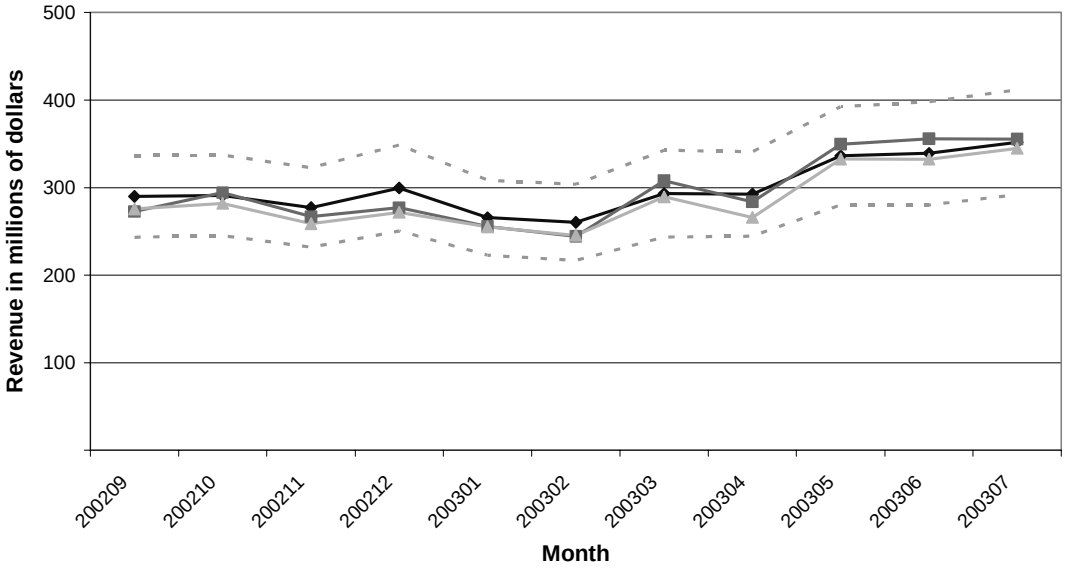
Approximately 70% of the estimate of sales from the MRCTS come from single establishments. This illustrates the potential importance of replacing 50% of these establishments with GST data. The survey's response rate is about 70% and there is a great deal of total and partial imputation. The survey was 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 services (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 excellence 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 gross 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.

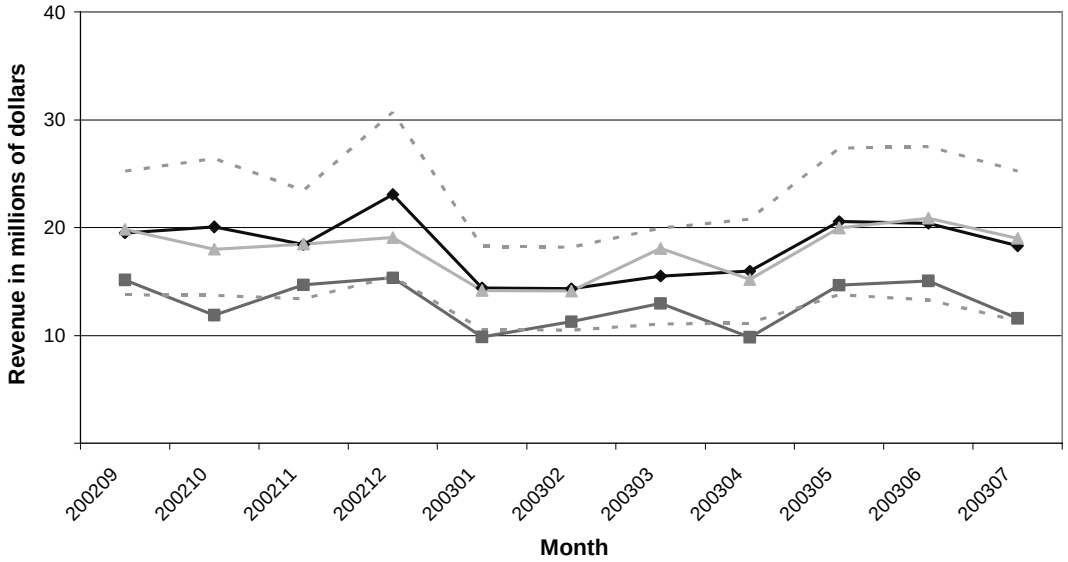
Results of the MRCTS simulations from September 2002 to July 2003



Graph 3 - Quebec
Full service restaurants



Graph 4 - Quebec
Caterers, Taverns and Mobile Canteens



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.

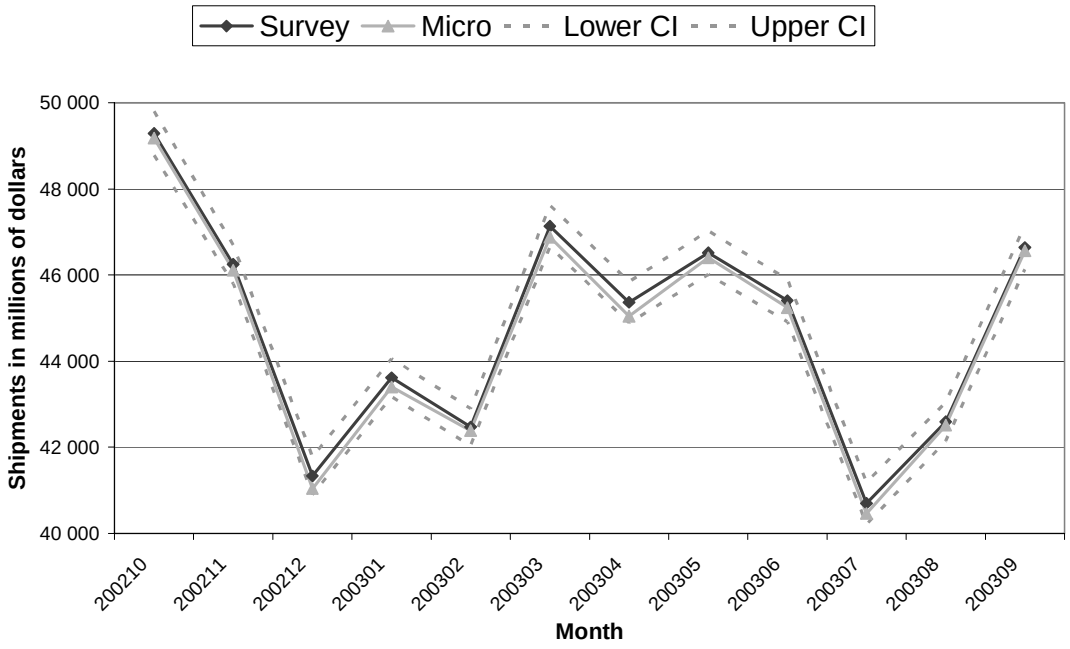
The MRCTS therefore opted to proceed and to replace 34% of its sampled single establishments with the MICRO model. A parallel test is planned for between October 2003 and February 2004, with implementation in April 2004 for the survey's February reference month.

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. The survey does not present methodology problems like the MRCTS. However, we always 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 an additional 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. Of course, these are national level estimates for all industry codes covered by the survey. Similar results are observed for all industry codes and all provinces taken separately.

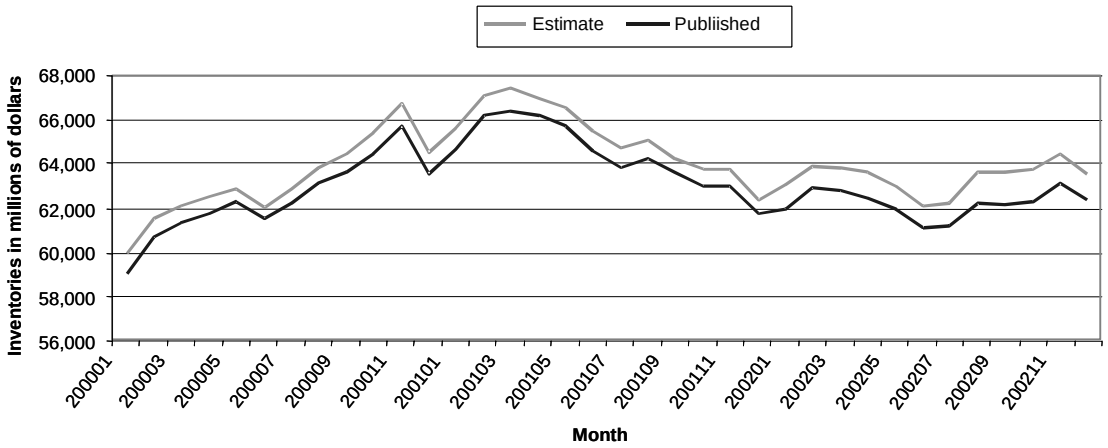
Graph 5 - Shipments - All manufactures



In terms of inventories, a number of solutions were considered. First, an effort was made to model using the GST data but the correlation was too weak. We next tried to use the enterprises' annual tax data to develop a model. This was complicated and did not produce satisfactory results. Lastly, we chose to use the survey's imputation system. Simulations were done over a period of three years and we found that the same trend was preserved (see Graph 6). Consequently, we decided to go with this option.

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 will take place between May and August 2004 and everything should be in place for September 2004 (July reference month).

Figure 6 - Inventories – All manufactures



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 replace 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

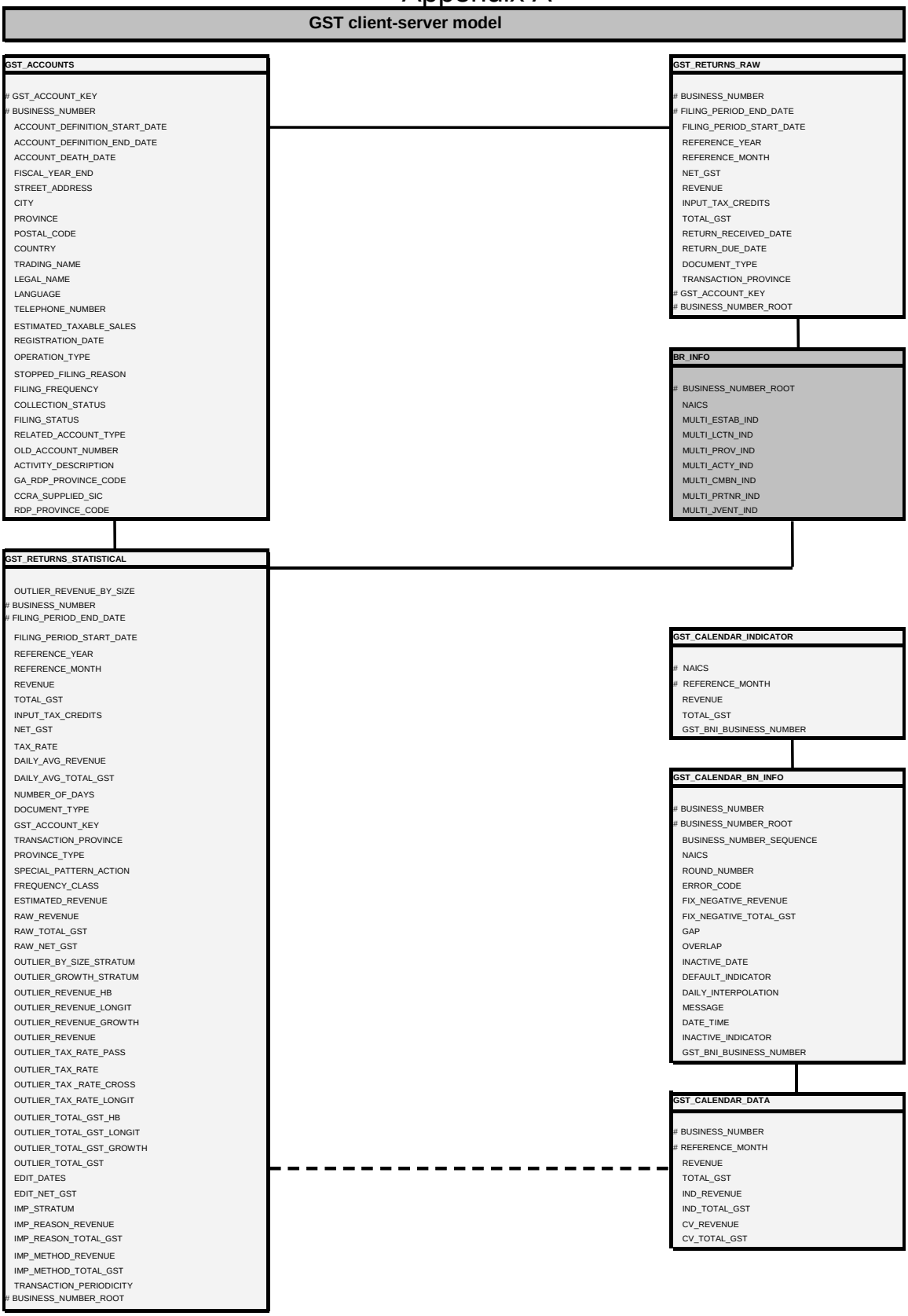
Dubreuil, G., Hidiroglou, M.A. and Pierre L. (2003), ''Use of Administrative Data in Modeling of the Monthly Survey Data'', Proceedings of the Survey Methods Section, Statistical Society of Canada.

Hamel, N. et Lothian, J. (2002), ''L'utilisation du Jackknife pour l'imputation des données administratives, conférence, 3ième Colloque francophone sur les sondages.

Hidiroglou, 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 Hidiroglou, 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


GST client-server model
GST_ACCOUNTS
GST_ACCOUNT_KEY
BUSINESS_NUMBER
ACCOUNT_DEFINITION_START_DATE
ACCOUNT_DEFINITION_END_DATE
ACCOUNT_DEATH_DATE
FISCAL_YEAR_END
STREET_ADDRESS
CITY
PROVINCE
POSTAL_CODE
COUNTRY
TRADING_NAME
LEGAL_NAME
LANGUAGE
TELEPHONE_NUMBER
ESTIMATED_TAXABLE_SALES
REGISTRATION_DATE
OPERATION_TYPE
STOPPED_FILING_REASON
FILING_FREQUENCY
COLLECTION_STATUS
FILING_STATUS
RELATED_ACCOUNT_TYPE
OLD_ACCOUNT_NUMBER
ACTIVITY_DESCRIPTION
GA_RDP_PROVINCE_CODE
CCRA_SUPPLIED_SIC
RDP_PROVINCE_CODE
GST_RETURNS_RAW
BUSINESS_NUMBER
FILING_PERIOD_END_DATE
FILING_PERIOD_START_DATE
REFERENCE_YEAR
REFERENCE_MONTH
NET_GST
REVENUE
INPUT_TAX_CREDITS
TOTAL_GST
RETURN_RECEIVED_DATE
RETURN_DUE_DATE
DOCUMENT_TYPE
TRANSACTION_PROVINCE
GST_ACCOUNT_KEY
BUSINESS_NUMBER_ROOT
BR_INFO
BUSINESS_NUMBER_ROOT
NAICS
MULTI_ESTAB_IND
MULTI_LCTN_IND
MULTI_PROV_IND
MULTI_ACTY_IND
MULTI_CMBN_IND
MULTI_PRTNR_IND
MULTI_JVENT_IND
GST_RETURNS_STATISTICAL
OUTLIER_REVENUE_BY_SIZE
BUSINESS_NUMBER
FILING_PERIOD_END_DATE
FILING_PERIOD_START_DATE
REFERENCE_YEAR
REFERENCE_MONTH
REVENUE
TOTAL_GST
INPUT_TAX_CREDITS
NET_GST
TAX_RATE
DAILY_AVG_REVENUE
DAILY_AVG_TOTAL_GST
NUMBER_OF_DAYS
DOCUMENT_TYPE
GST_ACCOUNT_KEY
TRANSACTION_PROVINCE
PROVINCE_TYPE
SPECIAL_PATTERN_ACTION
FREQUENCY_CLASS
ESTIMATED_REVENUE
RAW_REVENUE
RAW_TOTAL_GST
RAW_NET_GST
OUTLIER_BY_SIZE_STRATUM
OUTLIER_GROWTH_STRATUM
OUTLIER_REVENUE_HB
OUTLIER_REVENUE_LONGIT
OUTLIER_REVENUE_GROWTH
OUTLIER_REVENUE
OUTLIER_TAX_RATE_PASS
OUTLIER_TAX_RATE
OUTLIER_TAX _RATE_CROSS
OUTLIER_TAX_RATE_LONGIT
OUTLIER_TOTAL_GST_HB
OUTLIER_TOTAL_GST_LONGIT
OUTLIER_TOTAL_GST_GROWTH
OUTLIER_TOTAL_GST
EDIT_DATES
EDIT_NET_GST
IMP_STRATUM
IMP_REASON_REVENUE
IMP_REASON_TOTAL_GST
IMP_METHOD_REVENUE
IMP_METHOD_TOTAL_GST
TRANSACTION_PERIODICITY
BUSINESS_NUMBER_ROOT
GST_CALENDAR_INDICATOR
NAICS
REFERENCE_MONTH
REVENUE
TOTAL_GST
GST_BNI_BUSINESS_NUMBER
GST_CALENDAR_BN_INFO
BUSINESS_NUMBER
BUSINESS_NUMBER_ROOT
BUSINESS_NUMBER_SEQUENCE
NAICS
ROUND_NUMBER
ERROR_CODE
FIX_NEGATIVE_REVENUE
FIX_NEGATIVE_TOTAL_GST
GAP
OVERLAP
INACTIVE_DATE
DEFAULT_INDICATOR
DAILY_INTERPOLATION
MESSAGE
DATE_TIME
INACTIVE_INDICATOR
GST_BNI_BUSINESS_NUMBER
GST_CALENDAR_DATA
BUSINESS_NUMBER
REFERENCE_MONTH
REVENUE
TOTAL_GST
IND_REVENUE
IND_TOTAL_GST
CV_REVENUE
CV_TOTAL_GST