

Use of GST data¹ in Statistics Canada sub-annual surveys

Louis Pierre²

Introduction

The Canada Revenue Agency (CRA) collects information pertaining to the Goods and Services Tax (GST) from incorporated and unincorporated businesses. Since 1997, CRA has been sharing this information with Statistics Canada (STC) on a monthly basis. After processing, the GST data become a source of monthly administrative data that provides a good way of reducing the cost and response burden associated with the economic survey activities.

After establishing a solid infrastructure, the Tax Data Division (TDD), in conjunction with the Business Surveys Methods Division (BSMD), is proposing a number of methods for incorporating the GST data into the survey process. During Phase I, the objective was to replace half of the simple sampling units with GST data for three monthly surveys.³ For Phase II, the goal is to develop new quarterly surveys that rely on maximum use of the GST data. These surveys are intended to fill some major gaps in the data, especially in the service industries sector. The approaches used for Phases I and II are different since the constraints and opportunities are different.

Issues

The use of GST data appears to present four major issues for which we will need to come up with solutions.

First, the GST data are administrative data. Each month, we receive millions of records of which we must verify the quality without being able to return to the source. Furthermore, a substantial percentage of the expected transactions are late. We have thus developed a sophisticated edit and imputation (E&I) system that processes the information within 24 to 48 hours.

Secondly, taxfilers are required to file their returns according to different frequencies – either monthly, quarterly or annually – depending on their annual income.⁴ Transactions can be of varying durations within the same frequency. We have therefore developed a calendarization methodology that produces estimations of the transactions on a calendar month basis. The calendarization also carries out interpolations and extrapolations.⁵

Thirdly, because of the time granted to taxfilers, CRA does not start capturing transactions until a month after the reference month. Three weeks later, STC asks for the files for the transactions that have been captured and takes a few days to process this information. The clean, imputed and calendarized data are thus available to clients approximately 8 weeks after the reference month. Since STC endeavours to publish the results of the monthly surveys approximately 7 weeks after the reference month, the GST data are always too late for applications for Phase I. This situation has led us to develop and propose two

¹ Goods and Services Tax

² Louis Pierre, Statistics Canada, 120 Parkdale Avenue, Ottawa ON, Canada, K1A 0T6,
louis.pierre@statcan.ca

³ Monthly Restaurants, Caterers and Taverns Survey (MRCTS), Monthly Survey of Manufacturing (MSM), and Monthly Retail and Wholesale Trade Survey (MRWTS)

⁴ Revenue of more than \$6M: monthly; Revenue of between \$0.5M and \$6M: quarterly; Revenue of less than \$0.5M: annually. Monthly and quarterly taxfilers have a month to submit their returns, while annual taxfilers have three months.

⁵ For example, an annual transaction expected in December will be extrapolated for the following 11 months.

integration models that combine the survey data for the current month with the GST data for the previous month.

Lastly, the GST data are available for the legal entity (i.e. the business number (BN)), whereas the survey data follow the statistical structure, generally at the level of the establishment.⁶ In Phase I, we resolved the problem by stipulating that the integration model applied only to simple units for which there is an unequivocal relationship between the BN and the establishment. For Phase II, we propose to define a pseudo simple concept at various levels, i.e. a substantial number of legal entities arising from the GST data could be matched to statistical enterprises.

Edit and imputation

In general, the edit and imputation system seeks to detect outliers and determine the transactions to be imputed. The detection of outliers takes place through a combination of cross-sectional and longitudinal tests. Tests are carried out on tax rates, size and growth rates for the variables of interest. In addition to critical outliers, late transactions and transactions with missing revenue are imputed. A strategy is used to determine whether a transaction is late or whether a unit has become inactive, thereby avoiding overestimation.⁷ There are 7 imputation methods that are selected from on the basis of two decision tables. The methods are based on trends or on the history, and methods related to the auxiliary variable in cases of partial imputation are used in priority.

Calendarization

The objective of calendarization is to generate an estimation for the two variables of interest that corresponds precisely with the calendar months. If any time segments are not covered by transactions, these periods are interpolated or extrapolated by the calendarization program. Calendarization consists essentially in benchmarking the GST data on a monthly indicator series rescaled to the level of a given business.⁸ Indicator series are currently produced at the national level for each industry based on the North American Industry Classification System (NAICS) at 6 digits using monthly or quasi-monthly transactions. We plan to improve these indicator series and to produce others that would address specific needs related to the projects on the data gaps for Phase II. For example, complex businesses could be removed from the existing series (Phase I) and quarterly indicator series could be produced for the industries identified under NAICS to 4-digits, thereby increasing the robustness and homogeneity of the series (Phase II: see below).

Integration models (Phase I)

Two models based on ratios have been developed.⁹ The first is referred to as ‘Macro’ and consists of a calibration on the ratio for the survey and GST data at the population level. The second model, referred to as ‘Micro’, consists in ratio imputation. This involves the same ratio as in the previous model, but calculated at the sample level and applied to the micro data. For operational reasons, it was the ‘Micro’ model that was adopted for the three surveys for Phase I. This is in fact a double ratio, the first being the ratio between the survey data and the GST data at time ‘m-1’ and the second being the ratio between the

⁶ STC has established a framework, a 4-level statistical structure that hierarchically includes the business, company, establishment and location. This structure is different from the business’s legal and operational structure. It is because of this difference that we are seeking to identify reconciliation approaches in this article.

⁷ See Dubreuil, G. *et al.* (2005), “Analyse et impact des unités présumées inactives sur la base de données de la taxe sur les produits et services”, Colloque francophone sur les sondages 2005.

⁸ See Quenneville, Cholette and Hidioglou (2003), “Estimating Calendar Month Values from Data with Various Reporting Frequencies”, Proceedings of the Business and Economic Section of the American Statistical Association.

⁹ See Pierre, Brodeur (2004), “Statistical Use of Goods and Services Tax Data in Statistics Canada’s Monthly Economic Surveys”, Joint Statistical Meetings 2004.

survey data at time 'm' in relation to time 'm-1'. It should be recalled that this model applies only to simple businesses.

Creation of pseudo simple businesses (Phase II)

As part of the initiative of the Services Division (SD) to create Quarterly Services Indicators (QSI), it has been agreed that maximum use of the GST data would be the cornerstone of the project, constituting a completely innovative approach at STC. It is important to note that, in contrast with goods-producing industries and in general, simple businesses make a substantial contribution to the total revenue of a number of service industries. Various constraints related to Phase I have disappeared in this initiative and have given way to new opportunities. For example, the plan is to publish quarterly estimations, which would allow using GST data directly rather than the use of a model that combines data from different periods. Second, there are plans to use a model that produces estimations at the industry level aggregated according to the NAICS to 4 -digits. This represents an opportunity since businesses are being redefined as pseudo simple businesses eligible for direct replacement. Another category of pseudo simple businesses is identified when the predominant activity of the establishments that comprise a business exceed a specified minimum threshold (*eg.* 85%). One last possibility for pseudo simple entities would consist in determining an algorithm that would allow certain BN accounts¹⁰ to be matched with establishments defined in the statistical structure.

Estimation methodology (Phase II)

For some service industries aggregated according to the NAICS to 4-digits, it is believed that the GST data for simple and pseudo simple businesses could often represent the majority of the industry's revenue, sometimes even as much as 85-95%. Another advantage is that these revenues would be those of the population identified and not those of a representative sample, as is the case with the traditional survey approach. For the residual portion of truly complex businesses, the survey managers would select only the largest, for which a quarterly survey would be carried out. It should be recalled that the purpose of such a survey would be to collect for only one variable, that of revenue. The remaining small complex businesses would be estimated using a model, but they are expected to represent only a very small portion of the total estimation.

One criticism that could be made regarding the use of administrative data is that they may not accurately represent the population, the concepts or the variables that a survey is endeavouring to cover. If this were the case, it would be addressed by the fact that the Services Division would not use the revenue levels from the GST, but only the quarterly trends, to be benchmarked to the existing annual surveys. Previous studies have in fact shown that revenues based on the GST are strongly correlated with revenues obtained from the surveys.

Conclusion

The development of Quarterly Services Indicators presents a challenge and an extraordinary opportunity not only to fill data gaps in a steadily growing sector with behaviours that are quite different from goods-producing industries, but to innovate in the manner in which Statistics Canada fulfills its mandate, while controlling the response burden for businesses and the infrastructure costs associated with carrying out surveys.

¹⁰ Business Numbers (BN) include 15 positions, of which the last 4 figures identify 'RT' accounts identified by businesses.