

USE OF ADMINISTRATIVE DATA IN MODELING OF THE MONTHLY SURVEY DATA

G. Dubreuil, M.A. Hidirolou and L. Pierre¹

ABSTRACT

The cost and response burden associated to the survey activities have always been a major concern to Statistics Canada. These issues are especially important for monthly surveys. One option is the use of administrative data. Statistics Canada has been accessing the Canadian Goods and Services Tax (GST) data since 1997. These data offer the potential of replacing monthly economic survey data. In this paper, we discuss the processing of the monthly GST files, namely, the editing and imputation of outliers and missing data, and the calendarization. Then, we focus on the application of a model that uses clean calendarized GST data to replace monthly survey data while respecting the timeliness issue. Finally, an application relevant to the Monthly Restaurants, Caterers and Taverns Survey is presented.

KEY WORDS: Ratio estimator; Post-stratified estimator; Calibration.

RÉSUMÉ

Le coût et le fardeau de réponse associés aux activités d'enquête ont toujours été une préoccupation majeure à Statistique Canada. Ces questions sont particulièrement importantes pour les enquêtes mensuelles. Une option est l'utilisation de données administratives. Statistique Canada a accès aux données de la taxe sur la consommation de biens et de services au Canada (TPS) depuis 1997. Ces données offrent la possibilité de remplacer des données d'enquêtes économiques mensuelles. Dans cet article, nous discutons du traitement des fichiers de TPS mensuels, à savoir, la vérification et l'imputation des données aberrantes et des données manquantes et la calendarisation. Ensuite, nous nous concentrons sur l'application d'un modèle qui utilise les données de TPS calendarisées et épurées pour remplacer les données d'une enquête mensuelle tout en respectant les échéanciers. Finalement, une application faisant appel à l'Enquête mensuelle sur les restaurants, traiteurs et tavernes est présentée.

MOTS CLÉS : Estimateur par quotient; Estimateur post-stratifié; Calage sur marges.

1. INTRODUCTION

1.1 Context

The Goods and Services Tax (GST) is a Canadian tax on final consumption. Businesses have been remitting GST to the Canadian Customs and Revenue Agency (CCRA) since 1991. Information related to the transactions is captured by CCRA who, in turn, provides data to Statistics Canada (STC) in the form of raw transactions on a monthly basis. These transactions represent monthly, quarterly or annual taxable periods. This information can be used to increase the accuracy of monthly or quarterly sales data and reduce response burden and data collection costs. These advantages are hampered by a number of constraints that include timing, quality and data compatibility issues. In terms of timing, GST data may not be available for a considerable number of units at the time that they are expected. This raises questions as to what action is appropriate when an expected transaction is not received. In addition, transactions may contain errors, or they may simply be missing (partially or totally), thereby raising quality issues. Finally, the taxable period that they represent may not coincide with those required by sub-annual business surveys or even be fully compatible in terms of definition. This implies that the GST data need to go through a number of processes. These processes include (i) edit and imputation and (ii) calendarization. We describe in this paper what constitutes each process. The Monthly Restaurants, Caterers, and Taverns Survey (MRCTS) is used to illustrate how GST data can be used for sub-annual business surveys.

The paper is structured as follows. In Section 2, we provide an overview of GST. The proposed estimation procedure used for the modeling of monthly survey data is given in Section 3. Finally, conclusions are provided in Section 4.

¹ Guylaine Dubreuil, Mike Hidirolou and Louis Pierre, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6, guylaine.dubreuil@statcan.ca.

2. OVERVIEW OF GST

2.1 Description

Every business in Canada earning \$30,000 or more in annual sales is required to register for a GST account with CCRA. The reporting frequency is function of the size of the business. The businesses having annual sales over \$6 million are required to report monthly; the ones with annual sales between \$500,000 and \$6 million have to report at least quarterly. Finally, the businesses having less than \$500,000 can report annually. Monthly and quarterly reporters must remit within 30 days. Annual reporters must remit within 3 months. In general, the GST rate is 7 %, except for the provinces of New Brunswick, Nova Scotia, and Newfoundland where the Harmonized Sales Tax of 15 % has been applied since 1997. Most of the businesses remit quarterly (about 70 %).

Currently, STC has access to the GST remittance files that are made up of transactions. The transactions include the Business Number (BN) and the GST number, the filing frequency (monthly, quarterly, annually), the period covered (start and end dates), the sales (x), and the collected GST (z). STC processes the GST files each month. This processing includes editing (i.e.: range edits, outliers), imputation (late returns, inconsistent values, critical outliers, partial and total non-response) as well as calendarization of the sales data to account for various filing frequencies. For a given reference month, GST transactions are not all provided at once by CCRA to STC. There are two reasons for this: (i) the ending date associated with a transaction may be later than the reference month; or (ii) the transaction may not be captured yet by CCRA. Although transactions are eventually received from CCRA for a reference month of interest, a significant number of transactions need to be imputed the first time that a given reference month is processed.

For a given reference month m , the GST file is obtained from CCRA for the first time seven weeks after the end of that period and STC needs about a week to process it. Up to 70 % of the expected transactions are available at that time. Also, quarterly and annual remitters that are not expected for the reference month m are extrapolated using calendarization. The timing and availability of GST data needs to be taken into account when they are used as auxiliary information for sub-annual business surveys. A monthly business survey typically processes data and publishes results 6-7 weeks after the reference month m . At the releasing time, the GST data are not processed yet for that same reference month m . Consequently, it is reasonable to use the calendarized GST data of the previous occasion $m-1$ as auxiliary data for the current survey occasion data m .

2.2 Editing and imputation

First of all, a pre-processing step examines the validity of each field on each transaction. Contradictions in starting and ending dates such as negative period length, one-day gap and time-period overlaps are resolved. Multiple transactions for a particular processing month are resolved in order to get one record by BN. Data are standardized to a daily average for editing and imputation. An annual revenue is estimated for each business and will be used to determine its size. The median of the unit over time for the raw revenue per day, the raw collected GST per day and the raw tax rate are computed: these medians are used for editing and imputation purposes.

Editing examines the validity of each field on each standardized transaction both in terms of inter-field edits on the same unit, and inter-field edits across similar standardized transactions. Transactions are grouped according to the cross-classification of filing frequencies, industry groups (defined in terms of the North American Industrial Classification System, NAICS), and sizes. Both types of inter-field edits compare sales (x), and collected GST (z). The GST rate $r = z/x$ for a given transaction is also tested in terms of its change in relation to the same unit for previous reference periods. The inter-field edits within each transaction are bounded by predetermined values computed from previous GST data files. The inter-field edits across transactions are statistical and use the Hidioglou-Berthelot (1986) method to determine suspicious data. Inter-field edits compare data both on cross-sectional and longitudinal bases. A set of these tests determines if a record is classified as acceptable, suspect or critical. Suspect outliers are flagged and are not included in the imputation process. Suspect outliers are included as "is" in the estimation. Critical outliers are imputed, replaced with a more credible value, and the resulting imputed values are included in the estimation.

More specifically, active GST transactions will be imputed if they are identified as critical outliers, or have missing values because of partial or total non-response (late transactions). GST transactions are not imputed if a unit is considered as

inactive. A unit is declared as inactive if it has not reported transactions for several contiguous months. The number of months required to declare the inactive status depends on the unit's reporting frequency (monthly, quarterly or annual). GST remitters are expected to remit revenues to CCRA depending on their reporting frequency, and on time. Late transactions are considered as non-respondents, and their data need to be entirely imputed.

For partial non-response, the preferred imputation is to use auxiliary information available for that unit (micro-level imputation). When this is not possible, relations obtained from other similar transactions are used (macro-level imputation). In the case of total non-response, both the macro-level imputation and micro-level imputations are used. Sales (x) are first imputed using the macro-level imputation. Total GST (z) is then imputed using the micro-level imputation. For the micro-level imputation of total GST, different GST rates are used because they differ between provinces. The tax rates can also vary for an enterprise, depending on the proportion of sales of goods and services that have a 0 % GST rate. There are seven possible imputation procedures to impute data for total and partial non-response. The choice of which procedure to use depends on the non-response status (total or partial) of the transaction, as well as the existence of data for previous transactions.

2.3 Calendarization

Each transaction contains GST data for a set of consecutive days referred to as a *reporting period*. A reporting period is defined by a starting date and an ending date. Reporting periods can be approximately represented as monthly periods, multi-weekly periods, fiscal and calendar quarters, and fiscal and calendar years. Data associated with these transactions are referred to as fiscal data. Reporting periods differ from unit to unit, and may even vary within a unit if that unit changes its reporting period (e.g. from quarterly to monthly). The objective of calendarization is to generate estimates on a calendar month basis taking into account seasonal patterns. Calendarized monthly estimates are referred to as extrapolations when they are predicted. These extrapolations represent monthly estimates for transactions not yet received because the transaction is not due according to a previous reporting frequency. More details of the calendarization process are available in Quenneville, Cholette and Hidirolou (2003).

3. MODELING OF MONTHLY SURVEY DATA

3.1 Structural Considerations

A BN identifies each GST transaction and its associated calendarized records. The BN is an identifier assigned by CCRA to each business in Canada and it can be linked to an enterprise (ENT) on the Statistics Canada's Business Register (BR). This linkage can be one BN to one ENT (one-to-one), many BN's to one ENT, and many BN's to several ENT's. There are also ENT's for which we can not find an associated BN. There are four basic statistical units on the BR that are nested within one another. They are, in order of importance, the enterprise, the company, the establishment, and the location. The linkage to the next-lower level may or may not be one-to-one at each level of this nesting. Each BN can be classified as either being *single-linked* or *multi-linked*. A BN is *single-linked* if it satisfies the following two conditions: (i) it has a one-to-one BN-ENT linkage; and (ii) the associated enterprise is linked in a one-to-one fashion all the way down to a single location that covers exactly one NAICS6 in one province. The corresponding entities below the BN level will then also be labeled as single-linked. All other BN's, and any entities linked to such BN's, will be defined as *multi-linked*. The calendarized GST sales may not exactly correspond to the data obtained by direct surveying. For this reason, the modeling of the survey sales using calendarized GST sales will occur within the take-some strata, as for most of the business surveys, the majority of the units in these strata are expected to be single-linked. Multi-linked units and large units are initially sampled as take-all. Such units are surveyed directly regardless of their linkage to the BN's.

3.2 Estimations procedures

In this section, we provide an overview of the proposed methodology to incorporate the calendarized GST sales, as an auxiliary variable, for MRCTS. The objective is to reduce cost and response burden by keeping the same quality of the survey and minimizing changes to the existing design. Stratification for MRCTS was last updated in 1999. Estimates of total sales (and number of locations) are obtained on a monthly basis. Domain estimation reflects changes in industry or geography classification. This is the first monthly business survey to use the GST as auxiliary data, resulting in up to 50% sample size reductions in some of its strata.

GST sales should not be used as auxiliary data in the smaller cells or those where the correlation is weak. Correlations improve as outliers and dead units are excluded from the computations. Robust estimators should be used to remove the impact of outliers (outliers should be removed from the computations). An estimation procedure should be devised to minimize the weakening of the correlation caused by dead surveyed units declared live by GST.

For a given survey occasion m , total sales will be estimated using a mixture of the combined post-stratified ratio estimator (based on x) and the combined post-stratified count estimator (based on counts). The combined post-stratified count estimator is used for sample subsets that can not use the GST data (e.g. non-match) whereas the combined post-stratified ratio estimator is used for sample subsets that can use the GST data. The previous month GST sales, $x(m-1)$, are used as auxiliary data for the post-stratified combined ratio estimator.

The MRCTS sample s consists of a number of stratified samples s_h , $h = 1, 2, \dots, L$, where $\bigcup_{h=1}^L s_h = s$ for a given occasion m . The strata can be take-all (all units are selected with certainty) or take-some (units are selected with a given inclusion probability). The set of strata h are restricted to the take-some strata. Let U_h denote the stratified universe of size N_h . Subsets of s_h where GST data can be linked to statistical units are denoted as $s_{h,GST}$. Subsets of s_h where GST data cannot be linked to statistical units are denoted as $s_{h,\overline{GST}}$.

Take-some samples s_h are randomly split into two groups s_{1h} and s_{2h} provided that: (i) there is a minimum number of units (10) within s_h and (ii) the correlation between survey data and the GST data is good in terms of correlation: $\rho > 0.5(c.v.(x)/c.v.(y))$ where $c.v.(y)$ is the coefficient of variation of y , and similarly for x . This correlation condition holds if the estimator of population total is of the form $\hat{Y}_{RAT} = (\hat{Y}/\hat{X})X$ (Cochran, 1977). Samples $s_{h,GST}$ and $s_{h,\overline{GST}}$ are further split into $s_{1h,GST}$ and $s_{1h,\overline{GST}}$, and $s_{2h,GST}$ and $s_{2h,\overline{GST}}$, respectively. Since dead units have an impact on the correlation and the resulting ratio, they are initially split as equally as possible between $s_{1h,GST}$ and $s_{1h,\overline{GST}}$, and between $s_{2h,GST}$ and $s_{2h,\overline{GST}}$. For s_{1h} , we have sales (y) for the current occasion, and the corresponding GST revenues (x) up to the previous survey occasion. In addition, we only have GST sales (x) up to the previous occasion for the sample $s_{2h,GST}$. The sampling set-up is summarized in Figure 1.

Given the auxiliary GST data, there are two main options for estimating the population total $Y = \sum_{h=1}^L \sum_{U_h} y_k$. In the first option (*micro-record approach*), survey data are continuously collected for sample s_{1h} . Survey data and classification data are collected once for births occurring within s_{2h} . Synthetic data are thereafter predicted for usable survey records within $s_{2h,GST}$ using the ratio estimator. The mean of units in s_{1h} is used to create synthetic data for sampled units belonging to $s_{2h,\overline{GST}}$. Dead units within $s_{2h,\overline{GST}}$ retain their zero value as long as they remain in the sample. In the second option (*macro-record approach*), survey data are continuously collected for the sampled units in s_{1h} . The resulting estimates are calibrated to population totals within $U_{h,GST}$ for population units linked to GST, or $U_{h,\overline{GST}}$ for population units not linked to GST.

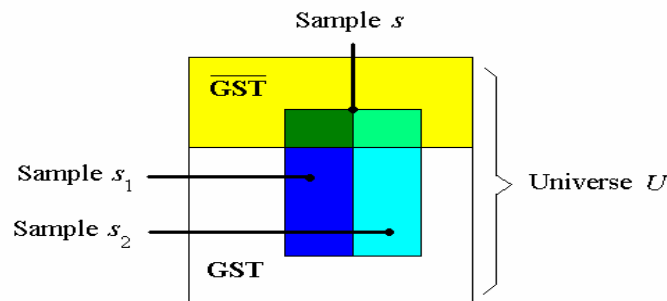


Figure 1: Sampling Process

These two approaches are next summarized. The survey weight for units belonging to the original sample s_h is denoted as $w_k = N_h / n_h$, and $w_{1k} = N_h / n_{1h}$ for units in s_{1h} for estimators that are strictly based on s_{1h} . The various required auxiliary totals are: $\tilde{N}_{s_{1h,GST}} = \sum_{s_{1h,GST}} w_{1k}$, $\tilde{X}_{s_{1h,GST}} = \sum_{s_{1h,GST}} w_{1k} x_k$, $\hat{N}_{s_{1h,GST}} = \sum_{s_{1h,GST}} w_k$, $\hat{X}_{s_{1h,GST}} = \sum_{s_{1h,GST}} w_k x_k$, $\hat{N}_{s_{2h,GST}} = \sum_{s_{2h,GST}} w_k$, $\hat{X}_{s_{2h,GST}} = \sum_{s_{2h,GST}} w_k x_k$, $\hat{N}_{s_{h,GST}} = \sum_{s_{h,GST}} w_k$, and $\hat{X}_{s_{h,GST}} = \sum_{s_{h,GST}} w_k x_k$. The objective is to estimate the population total for a given domain U_d , $Y(d) = \sum_{h=1}^L Y_h(d)$.

Micro-record approach

The estimator of $Y(d)$ is $\hat{Y}^{(MICRO)}(d) = \sum_{h=1}^L \hat{Y}_h(d)$ with $\hat{Y}_h(d) = \sum_{s_h} w_k y_k^*(d)$, where $y_k^*(d)$ is y_k^* , if $k \in U_d$ and zero otherwise. The value for y_k^* is defined as follows. The y_k^* values are equal to the original y_k values if they belong to s_{1h} . Otherwise, they need to be computed depending whether they belong to $s_{2h,GST}$ or to $s_{2h,\overline{GST}}$. That is, each active unit that belongs to $s_{2h,\overline{GST}}$ is assigned the value $\bar{y}_{1h,\overline{GST}} = \sum_{s_{1h,\overline{GST}}} y_k / n_{1h,\overline{GST}}$ where $n_{1h,\overline{GST}}$ is the number of units in $s_{1h,\overline{GST}}$. Each inactive (known dead) unit in $s_{2h,GST}$ gets a 0 value. If a unit belongs to $s_{2h,GST}$, it is predicted with the aid of the GST data. For each unit in $s_{2h,GST}$ the predicted value will be $y_k^* = x_k r$. Here $r = \left(\sum_{h \in G_c} \sum_{s_{1h,GST}} w_{1k} y_k \right) / \left(\sum_{h \in G_c} \sum_{s_{1h,GST}} w_{1k} x_k \right)$ is the ratio of the sum of the survey data y for occasion m to the sum of the GST data x for occasion $m-1$, as some level of regrouping of the original sampling strata. This regrouping (collapsing) is denoted as strata as G_1, \dots, G_c . The set of strata h that belong to group c is denoted as $h \in G_c$. For each unit in $s_{2h,\overline{GST}}$ the predicted value will be $y_k^* = \bar{y}_1$ where $\bar{y}_1 = \left(\sum_{h \in G_c} \sum_{s_{1h,GST}} w_{1k} y_k \right) / \left(\sum_{h \in G_c} \sum_{s_{1h,GST}} w_{1k} \right)$.

If there is no collapsing of strata, the y_k^* values are:

$$y_k^* = \begin{cases} y_k & \text{if } k \in s_{1h} \\ \left(\tilde{Y}_{s_{1h,GST}} / \tilde{X}_{s_{1h,GST}} \right) \times x_k & \text{if } k \in s_{2h,GST} \text{ and live} \\ \left(\tilde{Y}_{s_{1h,\overline{GST}}} / \tilde{N}_{s_{1h,\overline{GST}}} \right) & \text{if } k \in s_{2h,\overline{GST}} \text{ and live} \\ 0 & \text{if } k \in s_{2h} \text{ and dead} \end{cases}$$

where $\tilde{Y}_{s_{1h,\overline{GST}}} = \sum_{s_{1h,\overline{GST}}} w_{1k} y_k$.

If we had strictly kept the sample s_{1h} , and used the auxiliary data from s_h , the estimator $\hat{Y}_h(d)$ would have been a two-phase estimator. Sample counts $n_{1h,\overline{GST}}$ in $s_{1h,\overline{GST}}$ and the GST data (x), in $s_{1h,GST}$ would have been used. That is, $\hat{Y}_h(d)$ would have been

$$\hat{Y}_h(d) = N_{h,\overline{GST}} \bar{y}_{s_{1h,\overline{GST}}}(d) + N_{h,GST} \frac{\bar{y}_{s_{1h,GST}}(d)}{\bar{x}_{s_{1h,GST}}}$$

Macro-record approach

The estimator of the population total $Y(d)$ is

$$\hat{Y}^{(MACRO)}(d) = \sum_{c=1}^C \left[X_{c,GST} \frac{\sum_{h \in G_c} \sum_{s_{1h,GST}} w_{1k} y_k(d)}{\sum_{h \in G_c} \sum_{s_{1h,GST}} w_{1k} x_k} + N_{c,\overline{GST}} \frac{\sum_{h \in G_c} \sum_{s_{1h,\overline{GST}}} w_{1k} y_k(d)}{\sum_{h \in G_c} \sum_{s_{1h,\overline{GST}}} w_{1k}} \right]$$

If there is no collapsing of strata, the estimator of total reduces to $\tilde{Y}^{(MACRO)}(d) = \sum_{h=1}^L \tilde{Y}_h(d)$ where

$$\tilde{Y}_h(d) = \frac{\tilde{Y}_{s_{1h,GST}}(d)}{\tilde{N}_{s_{1h,GST}}} N_{h,GST} + \frac{\tilde{Y}_{s_{1h,GST}}(d)}{\tilde{X}_{s_{1h,GST}}} X_{h,GST}, \text{ with } \tilde{Y}_{s_{1h,GST}}(d) = \sum_{s_{1h,GST}} w_{1k} y_k(d) \text{ and } \tilde{Y}_{s_{1h,GST}}(d) = \sum_{s_{1h,GST}} w_{1k} y_k(d).$$

The micro-record approach uses the least GST data. It can almost be viewed as having originated from a two-phase sample design. It is referred to as the micro model because it results in a full set of micro-data for the sample. It is easier to treat errors in the GST data as we are dealing with the sample only. It is, however, less efficient than the current design. The macro-record approach uses all the available GST data, and the resulting sample is a one-phase sample design. It is referred to as the macro model because it uses population level data. It is not as easy to treat errors in GST data as we are dealing with the whole universe. Its advantage is that it is more efficient than the current design.

Differences between the estimates of totals for the weighted approach currently used in MRCTS and the proposed approaches are illustrated for the December 2002 reference month in Figure 2. They show comparable results for aggregated domains such as sub-industrial level. The associated coefficients of variation also show comparable results.

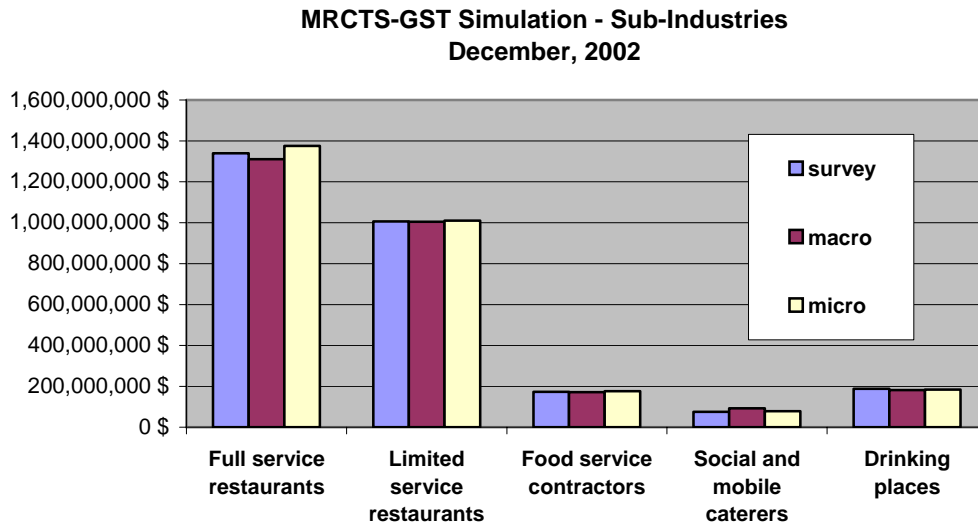


Figure 2

4. CONCLUSION

We have analyzed the relationship between MRCTS collected sales data for a given survey occasion and the corresponding GST data for the previous survey occasion. The combined ratio estimator seems to be a reasonable choice for predicting survey data using calendarized GST data. It is anticipated that this auxiliary data can only be used for subsets of the MRCTS strata that have a sufficient number of units and where the correlation is reasonable. It should be noted that the presence of dead units in the sample implies that the MRCTS data cannot be fully substituted with GST data. We need to further study smaller domains and evaluate the stability of the estimates over time.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques*. 3rd edition, New York: John Wiley and Sons.
Hidiroglou, M.A., Berthelot, J. -M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology*, Vol. 12, 73-83.
Quenneville, B., Cholette P., Hidiroglou M.A. (2003). Estimating calendar month values from data with various reporting frequencies. Contributed paper presented at the Annual Meeting of the Joint Statistical Meeting held in San-Francisco, California.