

Catalogue no. 11-522-XIE

**Statistics Canada International
Symposium Series - Proceedings**

**Symposium 2005 :
Methodological Challenges for
Future Information needs**



2005



**Statistics
Canada**

**Statistique
Canada**

Canada

Combining Administrative and Respondent Data by the Monthly Survey of Manufacturing

Steve Thomas, Krista Cook ¹

ABSTRACT

The use of administrative data sources by statistical agencies has become very popular since the use of these data sources can significantly reduce respondent burden and monetary costs. A recent development at Statistics Canada is the availability of monthly calendarized revenue data, available through an agreement with Canada Revenue Agency (CRA). This information has been shown to have a strong relationship with the revenue information collected by Statistics Canada's Monthly Survey of Manufacturing (MSM). This presentation will give a brief overview of the MSM and the information available from the CRA and will concentrate on how the available data were integrated into the survey process.

KEY WORDS: Administrative Data; Calendarization; Respondent Burden.

1. INTRODUCTION

The use of administrative data sources by statistical agencies has become very popular as respondent burden and the monetary costs associated with collection become more of an issue. The use of administrative data sources can significantly reduce monetary costs and reduce respondent burden while, at the same time, improve data quality if the data source is dependable. Collection costs are usually the largest costs associated with surveys. Respondent burden has always been an issue, but with the use of administrative data, the respondent will only have the burden of continuing to supply data to the administrative source. In addition to the cost savings and reduced respondent burden, administrative data can be of better quality than survey data and, if used properly, can actually improve survey estimates. For more on the issues related to the use of administrative data, refer to the paper by Brackstone (1987).

The use of administrative data may improve data quality if the data source is dependable. However, the advantages of using administrative data may come at a cost to data quality if the data do not correspond to that collected by the survey. The unit that the administrative data represents may not correspond to the statistical unit of interest; the period of time covered by the administrative data may not correspond to the period of interest for the survey; the variable of interest may not correspond to the variable from the administrative source; or the timing of the data may not correspond to the needs of the survey. A further problem with administrative data is that they can be very difficult to validate. Data often come from the administrative source "as is" and it is generally up to the statistical agency to edit and impute the data.

After careful examination of the administrative data and its source, it is up to the statistical agency to create methodologies that use this data to its full potential while considering its possible drawbacks. This was the challenge that Statistics Canada's Monthly Survey of Manufacturing (MSM) faced with the use of Goods and Services Tax (GST) data from the Canada Revenue Agency (CRA). In the spring of 2003, the MSM was chosen as a potential candidate to use tax data to replace survey data. After extensive testing, development, and implementation, the MSM became the first survey to publish monthly estimates using GST data as an administrative data source in the fall of 2004. This paper will describe the survey, the Goods and Services Tax, and some of the challenges that were faced in implementing procedures to use GST data.

¹Steve Thomas, Statistics Canada, R.H. Coats Building, 16th Floor, Tunney's Pasture, Ottawa, Ontario, K1A 0T6; Krista Cook, Statistics Canada, R.H. Coats Building, 11th Floor, Tunney's Pasture, Ottawa, Ontario, K1A 0T6

2. THE GOODS AND SERVICES TAX

The Goods and Services Tax (GST) was introduced in 1991 as a tax on most of the goods and services sold in Canada. In general, the GST is a tax of 7% on sales of goods and services, however, there are exceptions. For three Atlantic provinces (New Brunswick, Nova Scotia, and Newfoundland and Labrador), the tax is harmonized with the provincial tax to create an overall tax rate of 15%. As well, there are some goods and services, such as parts of the food industry, where no tax is applied. The tax is collected by the Canadian Revenue Agency (CRA) except in the province of Quebec, where the tax is collected by the provincial government and then forwarded to the CRA.

Every business with annual revenues over \$30,000 must report GST to the government. The reporting frequency is based on the annual revenue of the business. Businesses with between \$30,000 and \$500,000 of revenue file annually; businesses with revenue between \$500,000 and \$6 million report quarterly; and businesses with revenue greater than \$6 million are required to file monthly. Annual remitters have up to 3 months after the reporting period end date to report while quarterly and monthly reporters must report within the 30 days following the period end date.

All transactions are handled through a 15-digit Business Number (BN) assigned to the business by the CRA. The first 9-digits of this number identify a business, a part of a business, or a group of businesses depending on how the business is structured and prefers to report, while the remaining 6-digits represent the multiple GST accounts that the business may have. With each remittance, or transaction, the company provides total sales and other revenue, input tax credits, and the GST collected under the 15-digit BN for the taxable period. In the calendar year 2004, there were approximately 8.4 million transactions from 2.5 million businesses received by CRA. In terms of counts, most of these transactions were quarterly but in terms of sales, most were monthly (see Table 1). For more information on the GST and the CRA, visit the CRA website at www.cra-arc.gc.ca.

Table 1: GST Transactions – 2004

Frequency	Business Counts	Transaction Counts	Revenue
Monthly	8.1%	24.4%	80.5%
Quarterly	59.4%	66.0%	16.7%
Annually	32.5%	9.6%	2.8%

3. TREATMENT OF THE GST DATA AT STATISTICS CANADA

The GST data is remitted to the CRA, who compiles an administrative data file that includes the period covered (start and end dates), sales and other revenue, along with other information. The information on this file may not correspond to what is required by Statistics Canada. First, the GST data may not yet be available for a considerable number of units (late filers, deaths, etc.) Second, the GST data may not be reported on a consistent basis: they can be reported on a monthly, quarterly, or on an annual basis. Third, the reporting period may differ within the monthly, quarterly, and annual reporters, and finally, the reporting period may change over time for each business that remits GST to the CRA. Regardless of the reporting period, the reporting frequency or any missing data, the available data are provided “as is” by the CRA to Statistics Canada within seven weeks after the end of each reference month. This implies that all transactions still need to be edited, imputed and calendarized.

Statistics Canada’s Tax Data Division (TDD) is responsible for the processing of the GST file each month. This processing includes editing in the form of range edits and outlier detection; imputation for late returns, inconsistent values and critical outliers; as well as calendarization of the revenue data. Calendarization transforms all the GST data, including non-monthly data (quarterly, annual, 13-period), to correspond to the calendar month data required by business surveys. For more information on calendarization, refer to the work by Quenneville, Cholette, and Hidioglou in 2003. Once all of the processing is complete, a file is made available to the surveys within 4 days after the CRA file has been received.

4. THE MONTHLY SURVEY OF MANUFACTURING

The Monthly Survey of Manufacturing (MSM) is a sample survey that has been conducted at Statistics Canada for more than 50 years. The survey collects information on shipments, inventories, and orders from manufacturing establishments in Canada. This information is used in the calculation of the Gross Domestic Product (GDP), an important economic indicator. The System of National Accounts is the principal user of the collected data. Industry Canada, the Department of Finance, and the manufacturing companies themselves also use the estimates.

4.1 Background on the MSM

4.1.1 Overview

The MSM collects data from businesses at the statistical establishment level, where the characteristics of shipments (Goods of Own Manufacturing – GOM), inventories (Raw Materials – RM, Goods in Process – GIP, Finished Products – FP), and orders (Unfilled Orders – UO) are collected on a monthly basis. For the MSM, a manufacturing establishment is any establishment that is coded to manufacturing through the use of the North American Industrial Classification System (NAICS).

4.1.2 Frame

The MSM uses Statistics Canada's Business Register (BR) as a list frame. The BR contains approximately 100,000 manufacturing establishments and is updated monthly with births, deaths, and any other changes to the establishments. In order to lessen response burden and lower the collection cost, it is necessary to create take-none boundaries, where the smallest units that constitute the bottom 2% of total size (based on revenue from shipments) for each province are excluded from the sampling population. This results in the sampling frame being reduced to just 35,000 establishments.

4.1.3 Sampling

A sample of about 12,000 units is taken from the sampling population. The sampling population is stratified based on industry, province, and size. The size strata are constructed using the Lavallée-Hidiroglou procedure (Lavallée and Hidiroglou, 1988) where the size boundaries are created in such a way to achieve a target coefficient of variation (CV) of 3.5% for the size variable at the industry (NAICS 4, 5, or 6-digit) by province level. A power allocation with $p=0.5$ is used to allocate the sample. The sample is the same from month to month, except for births, which are sampled with the same probability as units in the original population. Approximately 7,000 take-alls are included in the sample.

4.1.4 Imputation

Missing values, which can be the result of total or partial non-response, are imputed using historical imputation where a trend factor is applied to previous month values. This trend factor is determined from a group of responding establishments within the same NAICS code and province. In situations where previous month values are not available for imputation (i.e. births), ratio imputation is used, which is based on Gross Business Income (GBI) available from the BR.

4.1.5 Estimation

Estimates are calculated using a Horvitz-Thompson estimator where each observation's response is multiplied by the sampling weight and values are summed over domains. The domains for the MSM are the province and industry, which is based on a grouping of the 6-digit NAICS code to the 3, 4, 5, or 6-digit level. Prior to publication, all estimates are benchmarked to the most recent Annual Survey of Manufacturing (ASM) values. The benchmark factor is basically the ASM value divided by the sum of the months of the MSM for the same reference year. After the benchmark factor is applied the level of the estimate is dictated by the ASM, while the trend of the estimate is taken from the MSM.

5. MSM AND THE GST

5.1 Overview

The goal for the MSM is to use GST data to reduce response burden and reduce data collection costs with minimal impact on data quality. There are several obstacles that have to be overcome with respect to GST data before it can be used effectively by the MSM. First, the GST data does not correspond to the unit of interest for the survey. The GST is collected via a Business Number (BN), which is usually equivalent to the statistical enterprise, while the MSM collects data at the statistical establishment level. Second, the GST data are not always available when needed by the MSM. Finally, the GST data only contain information on sales or other revenue and no information on stocks such as inventories or orders.

5.2 Establishment Number vs. Business Number

At Statistics Canada, a statistical structure is used on the Business Register (BR) to represent businesses in Canada. The hierarchical structure has four levels with statistical enterprises at the top followed by companies, establishments, and then locations. As stated earlier, the statistical establishment is used by the MSM since it is at this level that the business can supply the required financial information. The statistical structure can be simple, with one enterprise having one company with one establishment and one location, or complex where many companies, establishments or locations are part of an enterprise. The GST data is at the 15-digit BN level, with the first 9-digits of the BN roughly corresponding to the statistical enterprise. The remaining 6-digits correspond to the multiple accounts that the business or enterprise may have. This means that the business may report several times for the same enterprise, and given the structure of the BR, we may have many establishments reporting under the same 9-digit BN.

5.2.1 Simple Establishment Definition

Given the fact that the 9-digit BN is connected to the enterprise and the MSM is collecting at the establishment level, it makes sense that the survey only replaces those establishments that are part of simple statistical structures. To attempt to replace establishments with complex structures would mean that enterprise data would have to be allocated among its establishments. For the MSM, a simple establishment is defined as any unit that is part of the Non-Integrated Portion (NIP), as defined by the business register, as well as any unit that is part of the Integrated Portion (IP) that does not have any of the multi-flags associated with the enterprise. The multi-flags consist of the following:

1. Multi-Location
2. Multi-Establishment
3. Multi-Province
4. Multi-Activity
5. Joint Ventures
6. Partnerships
7. Combined Proprietorships

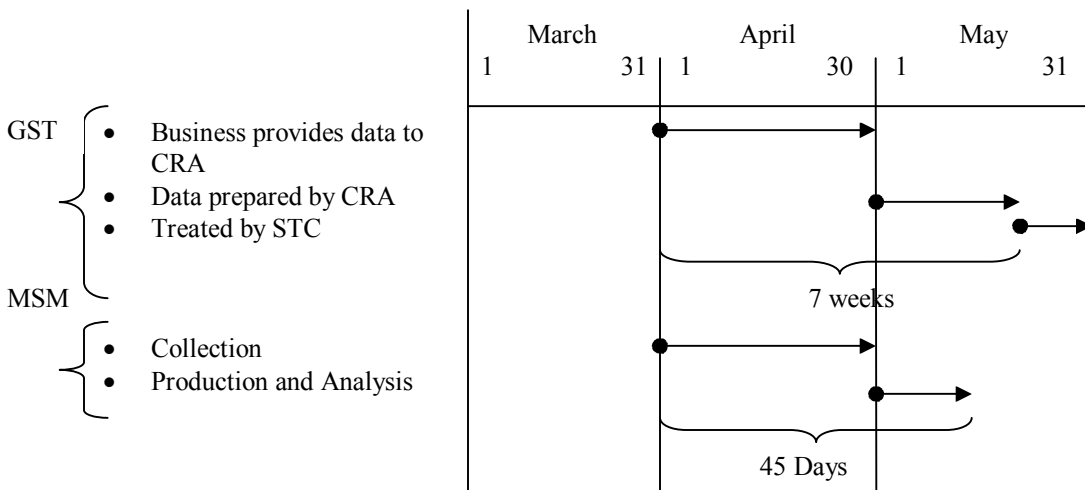
There are exceptions to this definition. If a structure consists of a single establishment with multiple locations and all of the locations are manufacturers in the same province then it should be safe to use the GST values for the establishment since the establishment will represent all of the locations. If a unit is defined as simple by the MSM but reports its GST using multiple accounts (i.e. multiple 16-digit accounts for the same 9-digit BN) then the unit will not be available for replacement. Aggregations of the multiple accounts were studied and it appeared that the correlation between the aggregated 9-digit BN information and shipments was not as good as was observed for the single BNs. Therefore, no attempt was made to use aggregated information in the estimation process.

5.3 Timeliness

Figure 1 illustrates the timeliness of MSM and GST data. The MSM produces its first set of estimates one month after the end of the reference month and the final set of estimates is usually published 45 days after the reference

month (month m). Comparing this schedule to the GST schedule resulted in the MSM deciding to use the GST data from two months prior to the reference month ($m-2$). It was also thought that data from month $m-2$ would be of better quality than using data from month $m-1$ since there should be less imputation. As well, imputation models for the GST should be stronger given the fact that there is more respondent data to base these models on. TDD now receives the GST data file from CRA seven weeks after the reference month and takes up to an additional four days to process the data. This means that surveys wishing to use GST as a data source can now use previous month ($m-1$) data instead and this may be a future goal of the MSM.

Figure 1: Current time-line of GST and MSM data availability - March Reference Month



5.4 Data availability

The GST data that are supplied to the survey areas by TDD contain sales and other revenue, collected GST, and input tax credits, which are all variables related to sales. No information is available for inventories or orders and the correlations between the inventories / orders characteristics and the other information available was found to be weak. A second study looked at the correlations of the inventories with the Input Tax Credit with the idea that the credit is applied on goods used to manufacture other goods. Unfortunately, the correlations were similar to those available from GST revenue. Therefore, it was decided that modeling inventory/orders characteristics using GST auxiliary variables was not appropriate.

As a last resort, a possibility of modeling inventory/orders variables using the current imputation process was investigated. In this situation, current month values were modeled based on previous month values and a trend. The trend was calculated based on response values from the imputation group and this trend was applied to the previous months reported or imputed value at the micro data level. This approach was supported by the fact that the trends for inventories and orders were basically flat and the fact that extra steps were taken to ensure that the replacement units did not represent significant proportions of the overall estimate.

6. SURVEY PROCESS

6.1 Initial Identification of the GST Sample

The first step in the survey process was to identify the units available for data replacement where we begin by identifying simple establishments. The same definition as described in section 5.2.1 was used. Note that this definition requires a link between the survey and the GST data. A timing issue was identified in section 5.3 where it was mentioned that GST data from month $m-2$ would have to be used in estimation. However, because of the timing issue, for the identification of the units to replace during sampling, we had to link with even older data ($m-3$) and hope that those units identified would remain simple at the time of estimation.

6.2 Size Thresholds

It was originally thought that simple units were always the smaller establishments that did not contribute much to the overall total of the estimates. However, it was found that some larger establishments do have a simple statistical structure. To ensure that the largest establishments were not replaced, a 75/50/50 rule was put in place. If the establishment was amongst the largest establishments that represented 75% of shipments, 50% of total inventories or 50% of Unfilled Orders, then it was not available for replacement. This ensured that the units available for replacement represented less than 25% of shipments, 50% of Total Inventories, and 50% of Unfilled Orders. One year of responses were used to prevent any monthly or seasonal fluctuations from having an effect. The months of $m-2$ to $m-14$ were used since data from current and previous month were not available. Given that units that were in sample for less than 1 year were also eligible for replacement, an adjustment for these units was made to have a 'projected' year total for comparison purposes. The projected year total is calculated as:

$$\text{Projected Year Total} = (\text{year total}) * 12 / (\# \text{months in sample})$$

For our purposes, a unit was considered to be a large contributor if it was below the 75/50/50 thresholds at the NAICS group (i.e. 4-digit NAICS for transportation and 3-digit NAICS for the rest of manufacturing) by province level for each of the three main characteristics. The 75/50/50 size cut-offs were used to create a threshold file, which could be used to create the boundaries for introducing new units to the replacement portion (S2) in the future.

After determining which units were simple and applying the thresholds, certain additional units were excluded to finalize the list of units available for replacement. Subject matter officers provided a list of important units that should not be replaced; units that were part of a combined report were removed from the eligibility list since these units will continue to be part of the collection entity; and, as part of the initial selection, we also removed units that were dead/out-of-scope on the MSM death management table but were still live on the BR. These were units that were dead according to the MSM but the survey feedback was not processed by the BR. The removed units were not considered eligible for replacement or for the model.

6.3 Replacement Sampling

The next step was to select a 50% sample from the units that had been identified as eligible. This was just a simple random sample of the eligible units at the NAICS 3 and 4-digit by province level. The units selected plus all simple refusals became the replacement sample, which was identified as S2. For simple refusals it was thought that replacing these units using GST information, rather than existing imputation techniques, would give more reliable data. The simple units that were not selected remain in sample and continue to be collected on a monthly basis. These units were identified as S1 and are used in the modeling of the S2 units. The sample split of S1 and S2 remains constant from month to month except in situations where the unit is no longer simple or loses the link with the GST data.

6.4 Monthly Estimation

Once the necessary survey data have been collected and the appropriate GST files are available, estimates can be calculated. It should be noted that in the description below there is a portion of the replacement sample (S2) that does not have a link to GST which is identified as $S2_{\text{GST}}$. This is the portion that has been identified for replacement

during sampling but at the time of estimation (up to 2 months later) no link can be made. If we now let y represent the MSM shipment revenues for month m , while x represents the most up to date GST sales (for month $m-2$) then the estimate totals can be created as follows.

$$\tilde{Y} = \sum_s w_k y_k^*$$

$$\text{where } y_k^* = \begin{cases} y_k & \text{if } k \notin S2 \\ \left(\frac{\hat{Y}_{S1_{GST}}}{\hat{X}_{S1_{GST}}} \right) \times x_k & \text{if } k \in S2_{GST} \text{ and live} \\ \text{Historical Imputation} & \text{if } k \in S2_{\overline{GST}} \text{ and live} \\ 0 & \text{if } k \in S2 \text{ and dead.} \end{cases}$$

To ensure that a proper model is created, outliers are removed from the model group (S1 with GST) using the same multivariate outlier detection routine that is used in the regular MSM production process (Franklin and Brodeur, 1997). These are observations where the GST revenue is not in agreement with the shipment values that are reported. Observations with large shipments and zero revenue or large revenue and zero shipment values could be excluded from the model. However, in general, such observations should be included to account for those unknown in the replacement portion.

Once the model is created, it is applied to the replacement units (S2). Outliers are removed from this population as well. In this situation, a univariate outlier detection method is used since only the GST revenue variable is available for these units. For the outliers where the model is not applied, the regular historical imputation method is used. For this process, current month values are imputed by multiplying historical values by a trend which is calculated from the responding units in the imputation group. One of the main advantages of this approach is that once the modeled micro values have been imputed, the existing estimation programs can be used.

7. RESULTS

For the June 2005 reference month, 4,928 units were identified as simple using the definition in section 5.2.1. From those units, 2,226 were identified for replacement and placed in S2, while the remaining units were placed in S1. Of the 2,226 units in S2 only 2,123 were actually replaced. 103 units were not replaced because they were outliers or did not link to the GST data file. These units were imputed with their previous months values multiplied by a trend.

7.1 Shipments

In order to measure the impact of using GST data for replacement, a measure of the proportion of the estimate coming from each of the different sources (respondent data, GST data, and imputation) was created for June 2005. For the national all-manufacturing estimate, 8% of the estimate was from the units replaced with GST data. For the 21 major-group industry estimates, the replacement portion represents 11% on average. This portion represents 9% on average for the provincial estimates, and 20% on average for the detailed industry (NAICS 4,5, and 6-digit level) by province estimates. Note that the percentage is high at the most detailed level because an entire domain could be replaced using GST data at this level.

Another recent development is the availability of variances that take into consideration the use of GST data. This work is in its early stages but initial results suggest that the impact on the coefficients of variation is minimal. For more information, refer to Hurtubise (2004).

7.2 Inventories and Orders

For the inventories and orders characteristics, the impact was more significant than was seen for shipments. This is not surprising given that a greater proportion of the estimate was allowed to be replaced using GST data. At the national all-manufacturing level, 16.7%, 16.1%, 15.8%, and 18.3% of the estimates for RM, GIP, FP, and UO

respectively were represented by values modeled using GST data. No coefficient of variation analysis is available at this time.

8. FUTURE CONSIDERATIONS

At this point, the inclusion of GST data in the survey design of the MSM is working quite well. In the future, with time and resources permitting, several ideas should be investigated. First, the current design does not allow for any rotation of the units that are modeled using GST data. One option is to rotate units into and out of the S2 portion on a monthly basis. The other option is to do a complete rotation every year or during a re-stratification, which is every three years. With rotation, basic profiling information, such as NAICS classification, size, and a characteristic profile can be kept up to date.

The second future consideration is to increase the number of units being replaced using GST data. The current production process has shown that the replacement of 50% of those establishments eligible is working quite well with little impact on the estimates. Can this proportion be increased to 60% or 70% to further reduce response burden and increase the cost savings?

Should the MSM be calibrated to the GST revenue totals? It is a fact that using GST data will have an impact on quality that can not be measured until the variance is calculated that incorporates the variance due to imputation modeling. Calibration should help increase the quality as long as we are using the correct population, correct variable, calibrating to the correct totals, and using the correct model groups. These methods should be considered during any future redesign of the survey.

REFERENCES

- Brackstone, G. J. (1987), "Issues in the Use of Administrative Records for Statistical Purposes", *Survey Methodology*, 13, pp. 29-43.
- Dubreuil, G., Hidiroglou, M.A., Pierre, L. (2003), "Use of Administrative Data in Modeling of the Monthly Survey Data", *Proceedings of the Survey Methods Section, Statistical Society of Canada*.
- Franklin, S. and Brodeur, M. (1997), "A Practical Application of a Robust Multivariate Outlier Detection Method", *1997 Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 186.
- Hurtubise D. (2004), "Variance due to ratio model imputation", *The Imputation Bulletin, Vol. 4 no 2, Methodology Branch, Statistics Canada*.
- Lavallée, P. and Hidiroglou, M. (1988), "On the Stratification of Skewed Populations", *Survey Methodology*, 14, pp. 33-43.
- 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*.
- Yung, W., Cook, K., Thomas, S. (2004), "Use of GST Data by the Monthly Survey of Manufacturing", *Proceedings of the Survey Research Methods Section, American Statistical Association, 2004, à paraître*.