Use of GST Data by the Monthly Survey of Manufacturing

Wesley Yung, Krista Cook and Steven Thomas
Wesley Yung, Statistics Canada, 11-F R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6

Key Words: Administrative Data; Goods and Services Tax; Monthly Survey of Manufacturing

1.0 INTRODUCTION

Each month Statistics Canada's Monthly Survey of Manufacturing (MSM) collects information on shipments, inventories and unfilled orders from approximately 10,000 manufacturing establishments across Canada. Estimates from this survey are major economic indicators and serve as key inputs to the monthly Gross Domestic Product (GDP) published by Statistics Canada. In an effort to reduce respondent burden and to reduce collection costs, Statistics Canada has been investigating the use of administrative sources as an alternative data source. One such source is the Goods and Services Tax (GST) that was introduced in 1991 as a tax on all goods and services provided in Canada. This tax is collected by the Canada Revenue Agency (CRA), who in turn makes the data available to Statistics Canada.

This paper will present the challenges faced and methodology implemented in integrating the use of the GST data into the MSM. Background on the MSM and the GST will be presented in the next section. Section 3 will present the methods evaluated for incorporating GST data into the MSM. Results of a parallel run are reported in section 4 and, finally, a summary is given in section 5.

2.0 BACKGROUND

2.1 MSM

The Monthly Survey of Manufacturing (MSM) is a sample survey that provides information on shipments, inventories, and orders for all manufacturing establishments in Canada. Each month the characteristics of shipments or goods of own manufacture, inventories (raw materials, goods in process, finished products), and unfilled orders are collected. The System of National Accounts is the principal user of the collected data and uses this information in the calculation of the Gross Domestic Product (GDP), an important economic indicator. Other users of the data include Industry Canada, the Department of Finance, and the manufacturing companies themselves.

The sampling unit of the MSM is businesses at the statistical establishment level as defined by Statistics Canada's Business Register (BR). The BR contains approximately 100,000 manufacturing establishments and is updated monthly with births, deaths, and any other changes to the establishments. A manufacturing establishment is a unit that has a North American Industry Classification System (NAICS) code within the scope of manufacturing (NAICS 311-339).

In order to lessen response burden and lower the collection cost it is necessary to create take-none boundaries. These boundaries are defined such that the smallest units that constitute the bottom 2% of total size (based on shipments if available or gross business income from the BR) for each province are excluded from the population. This results in the sampling frame being reduced to just 35,000 establishments.

The sample is stratified based on industry (as defined by NAICS), 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 as to achieve a target coefficient of variation (CV) of 3.5% at the NAICS industry by province level. A power allocation is used to allocate the sample of about 10,000 units. 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. Deaths remain in the sample to represent other deaths in the population.

Estimates are created using a Horvitz-Thompson estimator. The domains of interest for the MSM are the province and industry (4, 5, or 6-digit NAICS). Each year, all estimates are benchmarked to Statistics Canada's Annual Survey of Manufactures (ASM) using the most recent benchmark correction factor. This factor is the ASM value divided by the sum of the months of the MSM for the same reference year. New ASM values are usually available a year and a half after the reference year and new benchmark correction factors are usually available a few months after the new ASM estimates are available. Even though the benchmark correction factor is based on data that can be up to two years old, it is thought that this prediction gives a more reliable estimate than the unadjusted MSM values. Therefore, the level of the estimate is given by the ASM, while the trend of the estimate is taken from the MSM.

2.2 GST

The GST is a 7% tax levied on all goods and services provided in Canada with some exceptions. In the provinces of New Brunswick, Nova Scotia and Newfoundland and Labrador, the GST is replaced by a harmonized sales tax of 15% that combines the GST and the provincial sales tax. In addition, there still exists some 0% taxed industries. With the exception of the province of Quebec, the GST is collected by the CRA.

All businesses with annual revenues greater than $30,000 must register for a GST account and are required to file GST remittances. The frequency of remittance depends on their annual revenue with businesses with annual revenue greater than $6M filing monthly and businesses with annual revenue between $500K and $6M filing quarterly. Businesses with annual revenues between $30K and $500K are required to file annually. Quarterly and monthly filers are required to remit within 30 days of the period end, while annual filers must report within three months.

Each remittance, or transaction, consists of the business' activity code, Business Number (BN), GST number, filing frequency (monthly, quarterly or annually), period covered (start date and end date), sales and other revenue, the input tax credit and collected GST. Each year, Statistics Canada receives approximately 7.2M transactions, covering approximately 2.2M businesses, from the CRA. In terms of counts, most of these transactions are quarterly but in terms of sales, most are monthly (see Table 1).

Table 1. GST Transactions

	Counts	Transaction Counts	Sales
Monthly	6.4%	19.5%	77.1%
Quarterly	65.8%	71.4%	20.2%
Annually	27.8%	9.1%	2.8%

When using administrative data, one issue of concern is the timeliness of the data. This is of particular concern for a monthly survey such as the MSM. The GST data is provided to Statistics Canada by CRA six weeks after the end of the reference month, at which time approximately 30% of the expected transactions have been received. Thus approximately 70% of the expected transactions need to be imputed or extrapolated using calendarization. For more on calendarization see Quenneville, Cholette and Hidiroglou (2003). The following month, CRA provides data for the next reference month as well as for all previous months. Ten weeks after the end of a given reference month, approximately 85% of the expected transactions have been received. Given this delay in the reporting of the GST data, for reference month *m* for the MSM, it would be reasonable to use the GST data from month *m-1*. However, due to the production schedule of the MSM, it was decided to use *m-2* GST data to ensure that publication deadlines would be respected.

3.0 GST AND THE MSM

Before deciding on how to use the GST data to replace survey data, units to be replaced have to be identified. In order for a unit to be eligible for replacement, first of all, it has to be a simple establishment. For the MSM, a simple establishment is an establishment that has a simple structure. That is, according to Statistics Canada's BR, it is not multi-location, multi-establishment, multi-province, multi-activity, a partnership, an amalgamation or a combined proprietorship establishment. All units that are not simple are classified as complex. Complex units are not eligible for replacement due to the difficulty in allocating the GST data to the possibly different establishments or locations. In addition, if a simple establishment has duplicate BNs on the GST file, then it is no longer eligible for replacement.

Note that both large and small establishments can fall under the definition of a simple establishment. In order to minimize the effect of GST replacement on the MSM, only simple units that are not included in the top contributors in terms of shipments are eligible for replacement. How to define the top contributors will be discussed shortly. This ensures that large establishments are not replaced in the sample. In addition, simple units identified by subject matter can be placed on a 'do not replace' list.

Once units eligible for GST replacement were defined, correlation analyses were performed to identify which variables could be replaced. Using data from the March 2003 MSM, correlations between GST revenue and MSM variables for simple units were produced and are presented in Table 2.

As one can see from Table 2, the GST revenue is highly correlated with MSM shipments but is poorly correlated with the inventory variables (raw materials, goods in process, finished products and total inventory) and unfilled orders. Based on these correlation figures, it appears that GST data could possibly be used for the shipments variable but not necessarily for inventories or unfilled orders. In the next section, we address the use of GST data for shipments through two approaches.

We return to the problem of inventories and unfilled orders in section 3.2.

Table 2. Correlation between GST Revenue and MSM Variables

Variable	Correlation with GST Revenue
Shipments	91%
Raw materials	32%
Goods in process	4%
Finished products	8%
Total inventories	35%
Unfilled orders	19%

3.1 Modeling Shipments

Two main methods of modeling the shipment values using GST data were evaluated for use by the MSM. These two methods were originally developed for use with another monthly business survey at Statistics Canada, the Monthly Restaurants, Caterers and Taverns Survey (MRCTS). The first method, termed the micro approach, consists of modeling shipments at the micro data level and then creating the estimates in the same way the current estimates are created. This method can also be seen as an imputation method. The second approach, termed the macro approach, consists of dropping a part of the sample and re-weighting the affected strata. To help counteract the effect on quality that reducing the sample size can have on the estimates, the GST revenue variable is used to create a calibrated estimate. For those units that do not have GST, an optional additional step is to calibrate on counts. Both these methods consist of replacing or eliminating only simple establishments. It was felt that the best way to evaluate the two methods was to perform a simulation study using MSM data in order to evaluate their ability to preserve both the trends and levels of the existing series. The simulation study used data from the MSM for the months from October 2002 to July 2003.

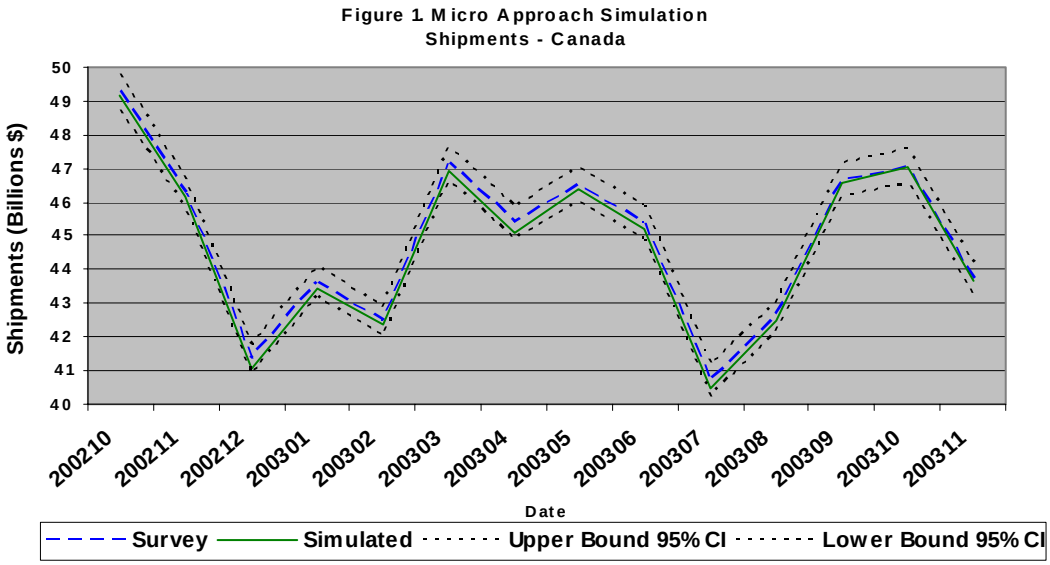
3.1.1 The Micro Approach

The first step in the process of testing the micro approach was to identify which establishments were available for replacement. For the purpose of the simulation, these were the units that were defined as simple, alive, and not in the top 80% of the domain of interest. In the case of this simulation the domain was the 3-digit NAICS by province with the exception of NAICS 336 (transportation) where the domain was the 4-digit NAICS by province. Simple units that were known to be important were also excluded from the replacement eligibility list. However, all chronic non-respondents were included. Ultimately, this resulted in 4,453 establishments being considered available for replacement for the first month of the simulation (October 2002). A simple random sample of approximately half of these establishments was selected which resulted in 2,241 units being selected for replacement.

The next step in the process was to model the shipment values for each of the units selected for replacement. The model was created from the ratio of the shipment values for the current month and the GST revenue values from two months before the reference month. The model was based on those units that were eligible for replacement but remained in the surveyed portion and were still alive according to the BR. At this time, outliers were removed from the model group using the existing MSM outlier detection system. An observation was considered an outlier when the GST revenue value was not in line with the shipment value.

Once the model was created, it was applied to the GST revenue value of the replacement units to create a modeled shipment value. Outliers were also removed from the replacement sample. However, for the replacement sample, we used univariate outlier detection on the GST revenue variable as it was the only variable available. The GST model was not applied to the outliers, so instead the regular imputation method was performed by using the historical values and a trend calculated from the responding units in the imputation group.

As Figure 1 illustrates, the results from the simulation of the micro approach were promising. There was little difference between the estimates with GST modeled data and the original estimates at the total manufacturing level. A 95% confidence interval suggests that the new line is well within the bounds of the original estimate. This was the case for many of the higher level estimates. However, with some of the more detailed domains, the modeling is not necessarily as dependable and outliers can have more of an effect.



One major advantage of using the micro approach is that it is consistent with the existing sampling and estimation methodology which means that from an analysis point of view, there is little change for the analysts to deal with and will be easier for them to remedy errors. Little change to the existing methodology also means that there is minimal impact the survey processing systems. The micro approach also allows us to replace more units than the macro approach since we are able to include take-alls and chronic refusals in the replacement sample. We also have a better control on the impact these processes can have on the estimates by introducing the 80% coverage rule.

However, there were also some disadvantages of using the micro approach which included the fact that it requires analysis of the entire sample which means that there was no improvement in the workload for the analysts. Another disadvantage is that it is difficult to determine the true variance due to the imputation. Current methods do not take into account the imputation with estimating the variance. There are also problems with how to resolve possible dead establishments. This can be seen as a possible problem with the macro approach as well. We have to determine how we will decide if a unit is dead if its data is being modeled from the GST as there is a time lag between the death of a unit and GST reported values of zero.

3.1.2 The Macro Approach

As with the micro approach, a simulation study using the macro approach was run in order to investigate the effect of using GST data on trend as well as its effect on the level of the estimate. It used the same months of MSM data as the simulation study for the micro approach.

The first step in this process was to identify which establishments were available for replacement by GST data. These were the units that were defined as simple and were in a take-some stratum. Since the macro approach involves dropping part of the sample and re-weighting the remaining units in the stratum, take-all strata were not considered for replacement to avoid having take-all strata with an adjusted weight (after re-weighting to take into account the units to be replaced with GST data) greater than one. Also, for the macro approach, the simple establishments in take-some strata with a link to GST were identified in the population. These values were used to calibrate the GST units in sample to the GST revenue total for the population. Since GST data for the population is used for the macro approach, care had to be taken to avoid bias when selecting the sample that was going to be dropped. Consequently, for the macro approach the exclusion of outliers from the sample, the 80% coverage rule and the exclusion of important units was not performed as was done in the micro approach.

This resulted in 3,177 establishments being selected as available for modeling by GST data for the first month of the simulation. A simple random sample of approximately half of these establishments by stratum was selected resulting in 1,579 of the GST eligible units being dropped from the sample. The sample was chosen at the stratum level to avoid the possibility of having empty strata when it came time to adjust the design weights at estimation. Also, it is important to note that in the simulation, there were no restrictions on the

minimum stratum sample counts needed for a unit to be considered GST eligible. In reality, however, we would only want to drop units in strata with many units in sample which would result in far less than 1,579 units being dropped from sample.

After dropping 50% of the take-some simple units with GST from the sample, the design weights were then adjusted using the remaining sample. For the take-some units in sample that could be linked to a GST record, the shipment values were calibrated to the GST revenue of all take-some simple manufacturing units (this excluded the take-none portion) at month *m-2* (i.e. two months before the current MSM reference month). For the units in sample that could not be linked to a GST record, the shipments were estimated using the adjusted weights. The take-all portion of the estimate remained the same as before. It can be noted here that this was a variation of the macro approach proposed for the MRCTS. For the take-some non-GST units, it was decided not to calibrate to the unlinked frame counts.

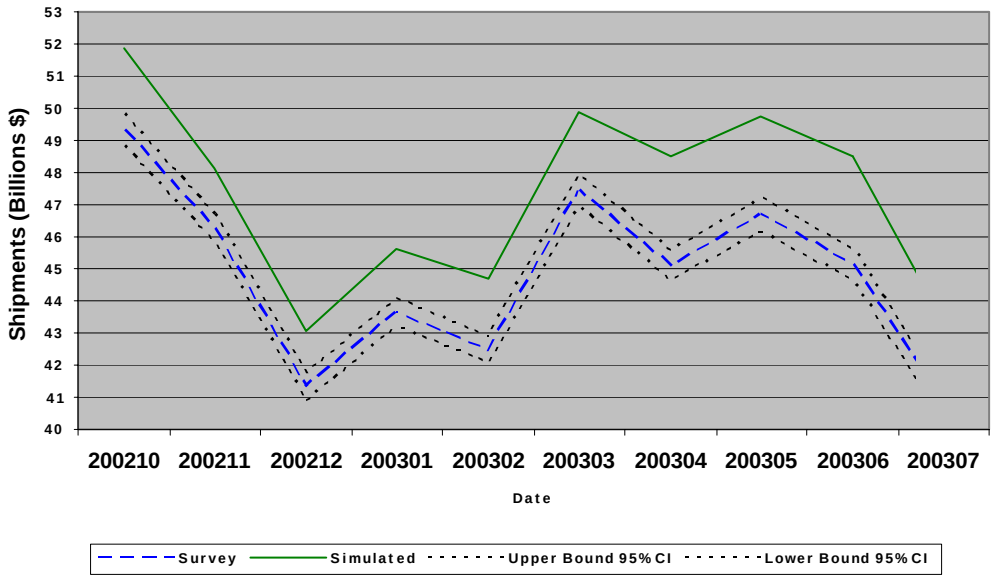
For the GST units, several ways of creating the calibration groups were tried for the simulation including NAICS 3-digit, NAICS 3-digit by province, the most detailed NAICS used for estimation (4,5, or 6-digit NAICS), domain NAICS by province and also by stratum. The results discussed here used the NAICS 3-digit level for calibration.

Figure 2 shows the results of the macro approach simulation for total manufacturing shipments. The estimates for the macro approach were quite different than the original series. In the majority of cases, the macro approach produced estimates that were higher than the original series. Even at the national level, the shipments estimates calibrated to the GST were not within the 95% confidence interval of the published MSM estimate. As estimates get more detailed, outliers in small domains can have a large affect on the trend as well as the level.

It was thought that the macro approach would have some advantages over the micro approach for several reasons. One advantage of the macro approach is that the estimates are supposed to be of better quality in terms of the coefficient of variation. It was also thought that there would be savings in terms of analysis since there would be fewer units to analyze each month.

However, there were a number of disadvantages to using the macro approach. The first problem was the number of units that were dropped from sample. As the macro model excludes units from take-all strata from being dropped, the target number of units to be dropped from the sample cannot be reached. Furthermore, if we restrict the size (in terms of the number of units) of the take-some strata that are eligible for GST replacement then the number of units that can be dropped diminishes to a few hundred units. The re-weighting step in the macro approach also causes some problems. The current MSM design allows a maximum design weight of 30 but when some of these strata are re-weighted after some of the sample is dropped, the design weights can become quite large – up to 60 or 70 in some cases. As mentioned earlier, one of the main disadvantages of the macro approach is the larger change seen in the level of the estimates as compared to the micro

Figure 2. Macro Approach Simulation
Shipments - Canada



approach. Even at the total manufacturing level, the new estimate with the GST is outside the 95% confidence limits of the published estimate. With smaller domains, not only is the level of the estimate affected, but the trend is greatly affected as well. It was also thought that the macro approach would have a higher learning curve with the analysts, as they would have new methodology to deal with and since we would be using GST information for the entire population rather than just the sample, the analysts would find it difficult to diagnose and remedy errors in GST values. Also from a systems point of view, it would not be as easy to adapt the existing MSM systems to work with the macro approach and would require more resources to implement.

Given the disadvantages of the macro approach as well as the better estimates from the micro approach during the simulation, it was decided to implement the micro approach in the MSM.

3.2 Inventories and Unfilled Orders

As previously mentioned, it was decided not to use GST data for inventories and unfilled orders and thus alternative methods were investigated. The first of these methods was to use the inventory to shipment ratio, as the shipment variable would be available for the majority of the establishments in the MSM sample (either from survey responses or the GST data). Although this relationship may have some basis from an economic theory standpoint, data from the MSM showed very little relationship between inventories and shipments. A number of regression studies were performed and they showed that the inventory to shipment ratio was highly correlated to the shipments' movement. This suggested that the inventories had a smooth gradual trend while the shipments drove the change in the ratio.

Another approach that was investigated was the use of the Input Tax Credit (ITC) that is collected by CRA. The ITC is the amount of GST paid on inputs that manufacturers can receive credit for from CRA. This variable is included on the GST files that Statistics Canada receives from CRA, but is not edited, imputed nor calendarized by Statistics Canada. Despite the unknown quality of the ITC data, the relationship between ITC and MSM variables was investigated due to the fact that previous studies have shown a strong correlation between the ITC and the value of inventories. Results from the March 2003 data are given in Table 3.

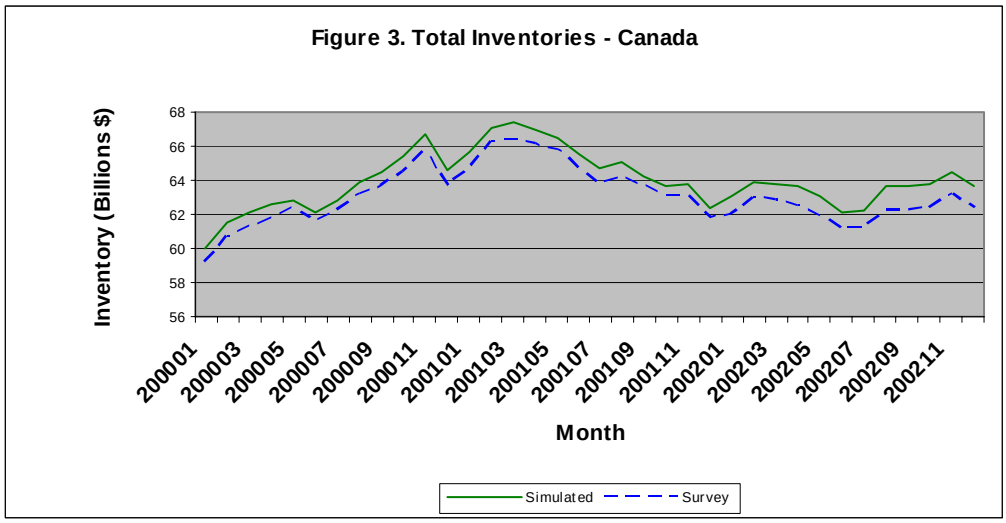
Table 3. Correlation between the ITC and MSM Variables

Variable	Correlation with ITC
Shipments	85%
Raw materials	55%
Goods in process	40%
Finished products	44%
Total inventories	59%
Unfilled orders	26%

As one can see from Table 3, the relationship between the ITC and the value of inventories is not very strong (less than 60%). Because of this, it was decided not to pursue the use of the ITC.

The final approach investigated was the use of the existing imputation system. Presently, missing inventories and unfilled orders are imputed using historical data with a trend based on responding units within the same NAICS/Province group. For those units without historical data, a proxy for shipments is obtained from administrative sources and the inventory to shipments ratio (as weak as it may be) is used. Fortunately, this situation does not arise very often. In order to evaluate the existing imputation system, three years of imputation were simulated for approximately 2,200 units. The period of three years was chosen simply because the MSM sample is renewed every three years by design and it was felt that units that were identified by replacement would have a chance for rotation at that time. Figure 3 shows the total inventory estimates based on the simulation and the survey values at the Canada level for the three years of simulated data.

As one can see, there appears to be a small difference in the level estimates, but the trend of the simulated series appears to follow the survey series. As expected, towards the end of the three year period we see that the two series are beginning to diverge. This difference was expected, given that the historical imputation was performed over a three year time period, but is actually smaller than expected. The results demonstrated here are typical of the other inventory variables as well as for unfilled orders. Based on these encouraging results, it was decided that the existing imputation system would be used for inventories and unfilled orders. The difference in the level would be handled through a macro adjustment or benchmarking.



4.0 PARALLEL RUN

Before implementing the new GST sample, it was decided to run the new GST system in parallel with the existing MSM system for a period of four months. This enabled us to monitor the new system and work out any problems without jeopardizing the monthly production of estimates. The parallel run ran from the April through July 2004 reference months.

4.1 Sample Selection

The initial GST sample was selected in much the same manner as it was for the simulation. There were some slight modifications in determining which units were eligible for GST replacement. In order to ensure that enough units were replaced while still ensuring certain coverage by the surveyed sample, it was decided to reduce the 80% coverage of shipments by 3-digit NAICS and province to 75% coverage of shipments by 3-digit NAICS and province. In addition, conditions of 50% coverage of inventories and unfilled orders were added. For the initial month, there were 4,129 units meeting all the requirements to be considered eligible for replacement and 2,252 of them were selected into sample – this included 383 simple chronic refusals.

4.2 Estimation

When calculating the model for the first month of the parallel run, there were 384 units identified as outliers and removed from the model. In addition, 71 out of the 2,252 units in the GST replacement group were identified as outliers based on their revenue values and were imputed by the existing MSM imputation system.

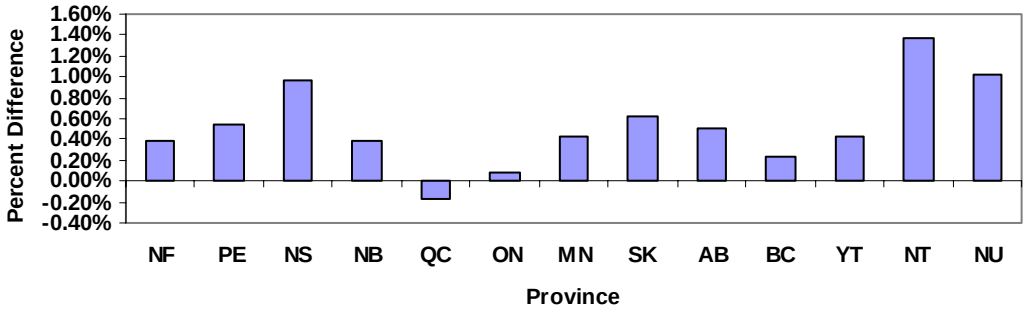
When published and GST estimates for shipments at the total manufacturing level were compared, the differences were small over the course of the parallel run. For all four months, the GST estimates of Total Manufacturing in Canada were well within the 95% confidence interval of the published estimates. The percentage difference for all four months is summarized in Table 4.

Table 4. Percentage difference in GST Estimate from Published Estimate during Parallel Run

Reference Month	Percentage difference from Published Estimate
April 2004	0.11%
May 2004	0.37%
June 2004	0.25%
July 2004	0.20%

The results were good when we looked at more detailed estimates that the MSM publishes at as well. Figure 4 shows the percentage difference between the GST estimates and the published shipments for each of the ten Canadian provinces and the three Canadian territories for April 2004. While it was generally the smaller provinces and territories that showed the largest percentage difference, the overall differences were considered negligible.

Figure 4. GST vs. Published Estimates
Provincial Level (April 2004)



When comparing the parallel run estimates with the published estimates at the NAICS 3-digit level, there was little to no difference in most of the estimates. However, some of the larger percentage differences seen were in the smaller industries where a small change in magnitude can have a larger impact on the percentage change. In some cases, it was found that the simple chronic refusals now modeled by GST were responsible for some of the larger changes and it was thought that the GST values were more accurate than the previously imputed shipment value.

5.0 SUMMARY

For statistical agencies, such as Statistics Canada, the use of administrative data is an attractive alternative for reducing respondent burden and reducing collection costs. As illustrated in this paper, the integration of GST data into Statistics Canada's monthly surveys is not without challenges. However, the MSM has been able to successfully integrate the use of GST data for estimating shipments and use its existing systems of imputation for inventories and unfilled orders. As shown by the simulation work and a parallel run, it is expected that the introduction of GST will have only a minor impact on the estimates.

Despite the encouraging results to date, many challenges still exist for the GST/MSM project. For instance, should units currently being replaced by GST be rotated? As previously noted, units that go out of business are very hard to identify in a timely manner through GST data. CRA and Statistics Canada are reluctant to death a unit unless a pre-determined time has passed with no remittance. Until that time has passed, imputation is used to impute a non-zero GST value and thus appears to still be in business to the MSM. In the same manner, new businesses are hard to identify due to the time lag between the birth and the first GST remittance. Another challenge is the units that remain in-business but change their business activity (ex. no longer a manufacturer). According to the GST data they are still alive but may be out of scope for the MSM. Suggestions of using a Nature of Business Report (NBR) have been made. This report would be a very short contact of units that are being replaced by GST data to verify if the business was still in business and if so, what was the nature of business. Finally, the long term effect of using historical imputation for inventories and unfilled orders still needs to be fully investigated.

References

Lavallée, P. and Hidiroglou, M. (1988). On the Stratification of Skewed Populations, *Survey Methodology*, 14, 33-43.

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*, American Statistical Association, 2003, to appear.