Evaluation of the Quality of Imputation of Goods and Services Tax (GST) Revenue for Late Transactions

Joanne Leung
Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON, K1A 0T6, Canada

Abstract
Goods and Services Tax (GST) data are used in many sub-annual business surveys at Statistics Canada. A study was done to evaluate existing imputation methods for late GST transactions by comparing imputed revenues to the reported values. In this paper, we will highlight some results obtained, especially the finding that imputing a new value each month only improves the quality for about half of the units. As a follow-up, we explored the possibility of imputing late values once and keep them at subsequent processing. Results showed that this approach reduces revisions to the data without imposing large differences at the macro level. In light of the findings, this strategy was implemented in the GST Processing System in September 2009. We will show the impact of this change using real data. Finally, we will give some recommendations to further improve the quality of imputation.

Key Words: administrative data, imputation, sub-annual data, data revisions.

1. Introduction

The use of administrative data is of growing importance at Statistics Canada (StatCan). One of the sources of data available is the Goods and Services Tax (GST) database. The GST is a value added tax on most goods and services sold or provided in Canada. Businesses collect the tax and must remit it to the Canada Revenue Agency (CRA) which has been sharing this information with StatCan on a monthly basis. The earliest data available to StatCan is 1997. Once processed, GST data are a source of monthly administrative data that offer a good alternative to the cost and response burden associated with business survey activities. One of the most widely used variables is the revenue for each business.

In 2008, one of the sub-annual surveys team discovered that a large part of revisions to their data was related to the imputation of GST revenue. Therefore, a series of studies comparing imputed revenue and the corresponding reported value was conducted. The objective was to evaluate the quality of imputation of revenue for late remitters, as well as the quality of revised imputed values. In this paper, we will look at some results from the study. The study led to change in the GST imputation process and we will describe the changes briefly. We will also show the impact of this change to various users of the data.

1.1 Overview of GST Data Processing

The CRA collects the GST data from enterprises and provides them to Statistics Canada on a monthly basis. Once the GST data is received at StatCan, the GST Processing

Systems processes the data in three phases: Edit and Imputation (E&I), calendarization and allocation.

The remitting frequency for each business depends on its annual revenue. It can be remitted on a monthly, quarterly or annual basis. Each remittance is called a transaction. Monthly transactions are expected every month, while quarterly transactions are expected once every three months and annual transactions are expected only once a year. For an expected transaction, it may or may not be received on time. This is mainly because monthly and quarterly remitters have a one-month grace period to report to CRA after the end of the reference period. Annual remitters have a grace period of three months. Expected transactions that are not yet received in time of processing at StatCan are considered as late transactions. For a majority of late transactions, they eventually become reported within six months.

Approximately seven weeks after the end of each reference month, StatCan receives GST data files from the CRA. At that point, up to 70% of the expected transactions are available. Each month, StatCan processes all GST transactions in the 19 most recent reference months in the E&I process. Outliers and missing values are detected on the reported transactions. Imputation is done using historical data or using similar units. Late transactions are also imputed. However, transactions that have been late for an extended period will be handled by an inactivation process (Dubreuil, Pierre, Labelle-Blanchet and Liu, 2003), which is not covered in this paper.

After the E&I, all transactions in the database go through the calendarization process (Beaulieu and Quenneville, 2008), where annual and quarterly transactions are split into monthly values. Transactions that are not expected are extrapolated using a similar approach as calendarization. The calendarized revenue for each business is then allocated from the legal entity level to the operating entity lower level. The processed data are available to various users in StatCan. These include several sub-annual business surveys (Brodeur and Pierre, 2003; Dubreuil, Hidiroglou and Pierre, 2003), the Business Register and the System of National Accounts.

1.2 Variable of Interest

Revenue reported on the GST database is the main variable of interest to the sub-annual business surveys at StatCan. However, since the CRA is mostly interested in the tax amounts collected by businesses, the revenue field is actually not mandatory and therefore sometimes missing in the remittance. In such a case, the revenue will require imputation. Imputation is also required when an expected transaction is late, and when the revenue reported is detected as an outlier. Various imputation methods are used to impute the revenue for different reasons. In this paper, we focus on the imputation methods used at the E&I process when an expected transaction is late.

2. Imputation Methods Available for Late Transactions

When a transaction is late, an artificial transaction is created. The revenue of this late transaction has to be imputed. Three different imputation methods are in place in the GST Processing System: the stratum trend method, the carry forward method and the stratum mean method. In this paper, a stratum refers to an imputation class. It is defined as a group of units that have the same remitting frequency, lie within the same annual revenue range and belong to the same industry group according to the North American Industry Classification System (NAICS).

The stratum trend method and the carry forward method are used when a transaction exists for a business exactly one year before the current reference month. When the revenue from the year before is non-zero, the stratum trend method is used to impute the late revenue at the current reference month. The imputed revenue is obtained by multiplying the trend of revenue in the selected stratum by the revenue of the business from the year before. The trend of revenue in a stratum is defined by the ratio between the current mean revenue and the mean revenue from the year before, from all units in the stratum. When the revenue from the year before is zero, the carry forward method is used, in which we carry forward the $0 revenue from the year before to the current transaction. The stratum mean method is used as the last resort. It is used to impute the revenue of a late transaction when there is no information found for the business exactly one year before the current transaction. In that case, the current mean revenue of the selected stratum is taken as the imputed revenue of the late transaction.

Figure 1 shows the distribution of imputation methods being used to impute late transactions in the last quarter of 2007. We can see that the stratum trend method is the most commonly used. The carry forward method is used for about 20% of the late transactions, while the stratum trend method is only used to impute the late revenue for 5% of the transactions.

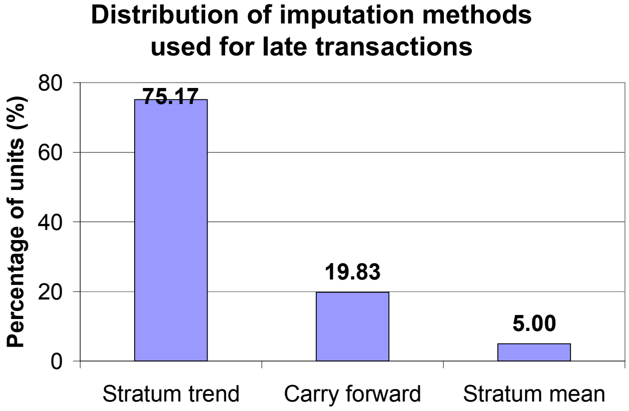

Figure 1: Distribution of imputation methods used for late transactions based on transactions in the last quarter of 2007.

Using late transactions in the last quarter of 2007, we explored the quality of each imputation method by comparing the imputed revenue to the reported one eventually received. Figure 2 shows a breakdown of the quality for each method. For stratum trend and stratum mean methods, there were more units being overestimated than underestimated. For the carry forward method, the imputed revenue, by definition, is always equal to the reported revenue when the businesses report $0 as their revenue. However, the method is inaccurate when the businesses report a non-zero value. We can see from this graph that it occurs in about 20% of the time.

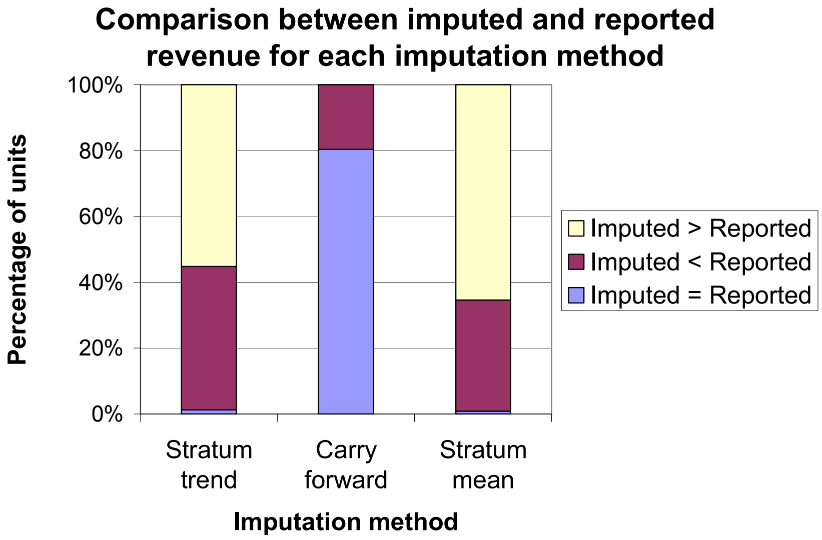

Figure 2: Comparison between imputed and reported revenue for each imputation method based on transactions in the last quarter of 2007.

2.1 Problem in the Imputation Process

In spring 2008, one of the sub-annual surveys team discovered that 65% of the large revisions to their data are related to the imputation of GST revenue. Recall that, each month the 19 most recent reference months of data are being re-processed. If an expected transaction is late, it is imputed. If the same transaction is still late at the following processing, the previous imputed value is erased. At the same time, some late transactions for the same reference months become reported. They become part of the stratum for their corresponding reference month. Revenue of late transactions that are still not received at that point is then imputed again, most likely with a different value. In other words, if the same transaction is late for more than one month, an independent imputation procedure is done at each processing. With new transactions for other businesses coming into the system, the change in stratum composition led to a change in the imputed value. Thus, it created revisions and instability to the imputed data of any given reference month.

3. Study

Since GST data is used by various users in StatCan, we decided to study the overall quality of imputation of revenue for late transactions. The study is based on late transactions between reference months of October 2007 and December 2007.

The study contains three parts. We first checked whether the first imputed revenue was being overestimated or underestimated. Then, we analyzed the quality of each re-imputation done to the late transactions. Finally, we explored the possibility of imputing the late revenues only once and keeping the same value at subsequent processing of the given reference month when the units are still late.

3.1 Over- or under-estimation?

In the first part of the study, we kept transactions that were late but become reported six months after it was first imputed. To determine whether the imputed revenue is over- or under-estimated, we checked to see if the imputed revenue of the late transactions corresponds exactly to, is lower than, or higher than the reported revenue.

Thus, for each late transaction, we computed the ratio

$$\text{ratio} = \frac{\text{imputed revenue}}{\text{reported revenue}}.$$

Figure 3 shows the distribution of transactions having various ratios between imputed and reported revenue. Note that, in GST data, revenue cannot be negative. Therefore, the ratios we computed were all positive. In this graph, we define underestimation as a ratio below 0.8, as well as when the imputed revenue is zero while the reported revenue is non-zero. Overestimation is defined as having a ratio above 1.25, and when the reported revenue is zero while the imputed revenue is non-zero. When the ratio is close to 1 or when both the imputed and reported revenue are zero, it is considered as “no bias”.

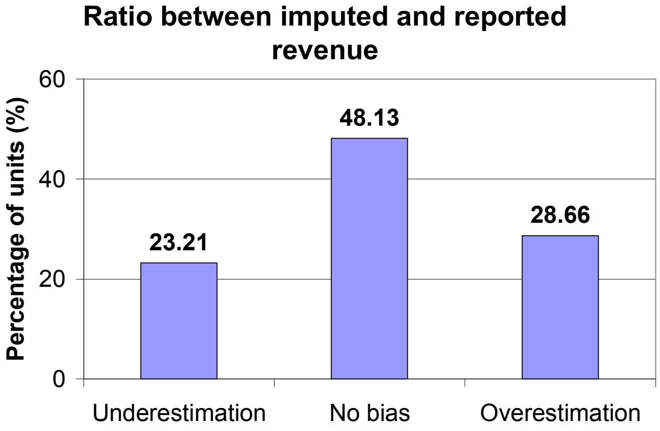

Figure 3: Distribution of ratio between imputed and reported revenue for late transactions in the last quarter of 2007.

We observed from Figure 3 that about half of the units belonged to the “No bias” category, while the other half were overestimated or underestimated. In particular, there were more units with their revenue overestimated than underestimated.

3.2 Quality of Revised Imputation

In Section 2.1, we explained how the imputed revenue can get unstable over time. In the second part of the study, we analyzed the quality of imputation of revenue for late remitters at each processing month. In other words, we looked at how the imputed revenue evolved through time.

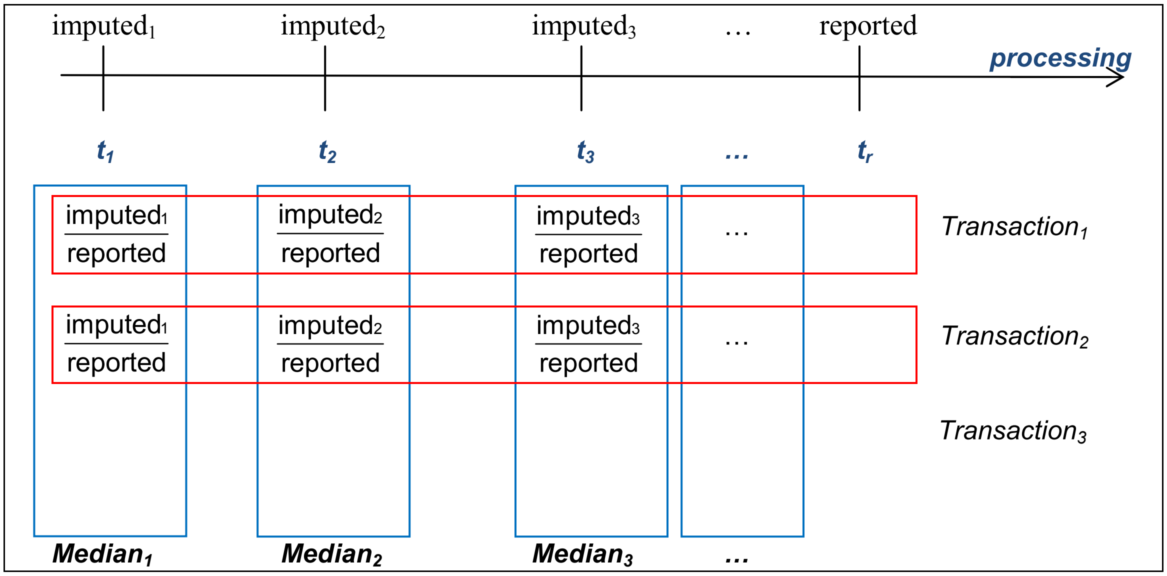

Figure 4: Illustration of how transactions are being re-imputed at subsequent processing periods of a particular reference month until a reported value is received.

Figure 4 illustrates our approach. Let us start with one transaction, *Transaction$_1$*. At time t_1, the transaction is late for the first time and is imputed with a value of *imputed$_1$*. In the next processing at time t_2, the transaction is still late. We then erased the previously imputed value and re-imputed it with the value *imputed$_2$*. At t_3, the transaction is still late and we re-imputed the revenue with a value of *imputed$_3$*. Eventually, the reported value was received at time t_r. Then, we can evaluate the quality of imputation by looking at the ratios between each imputed value and the reported value, i.e., $\frac{imputed_1}{reported}$, $\frac{imputed_2}{reported}$, $\frac{imputed_3}{reported}$, …, $\frac{imputed_{r-1}}{reported}$. We repeat the same process to compute these ratios for the other late transactions *Transaction$_2$*, *Transaction$_3$*, etc.

We determined whether the quality improved over time by computing the medians of the ratios at each processing. For example, we computed *Median$_1$* among all the ratios $\frac{imputed_1}{reported}$ computed for each late transaction *Transaction$_1$*, *Transaction$_2$*, etc., at the first processing at time t_1 and we computed *Median$_2$* among all the ratios computed for each late transaction at the second processing, and so on. If the medians computed converge to 1 then the imputed revenues at each subsequent processing gets closer and closer to the reported value. In other words, if the medians converge to 1, then the revisions at each subsequent processing are good.

Based on all late transactions in the last quarter of 2007 that were imputed more than once, we found that the last imputed values $imputed_{r-1}$ were not always the best ones. That is, the medians of the ratios were not closest to 1 at the last processing before the businesses became reported. Also, subsequent imputations did not always give medians closer to 1 than in the first processing. The medians of the ratio at first and last imputations are shown in Table 1 for a subset of late transactions selected in this study.

Table 1: Median ratios of imputed to reported revenue at the first and last imputations.

Transactions become reported	Median ratio at first imputation (*Median*$_1$)	Median ratio at last imputation (*Median*$_{r-1}$)
At time t_2	1.0041	N/A
At time t_3	1.0022	1.0011
At time t$_4$	1.0193	1.0252
At time t_5	1.0130	1.0420
At time t_6	1.0347	1.0167
At time t_7	1.0621	1.0226

3.2.1 Quality of subsequent imputations vs. quality of first imputation

We took a step further to compare the quality of subsequent imputations against the quality of the first imputation. The absolute value of (ratio – 1), where ratio is defined in Section 3.1, can be rewritten as

$$diff_i = \left| \frac{imputed_i \text{ - } reported}{reported} \right|,$$

where $diff_i$ is the magnitude of the relative difference between the imputed and reported revenue at the ith imputation. We would like to see if the subsequent imputations, namely imputation at time t_2, t_3, …, t_{r-1}, show a smaller relative difference than the one obtained at time t_1. In other words, we would like to see if there is an improvement in the performance of the imputation at subsequent processing periods over the first processing. If, for example, the magnitude of the relative difference at t_1 is larger than that at t_2, i.e., $diff_1$ is larger than $diff_2$, then we say that the imputation at t_2 gives an improvement in quality. Thus, for a transaction that becomes reported at t_4, for example, we are able to tell if the imputation of revenue at t_2 and t_3 give better quality than that at t_1. If, after this part of the study, the quality of previous imputations are deemed to be better than the later ones, or if the subsequent imputations show no better than the first one, then our imputation strategy could be revised such that, once a transaction's revenue is imputed, the value is kept until we receive a reported value, or when the unit is considered inactive according to the inactivation strategy.

We computed each of $diff_2$, $diff_3$, …, $diff_{r-1}$ and compared them against $diff_1$ for each business. If the quality remains unchanged then the imputed value is stable over time. If the quality deteriorates, it indicates that the imputed value is unstable. Figure 5 shows the distribution of the quality improvements and deteriorations found.

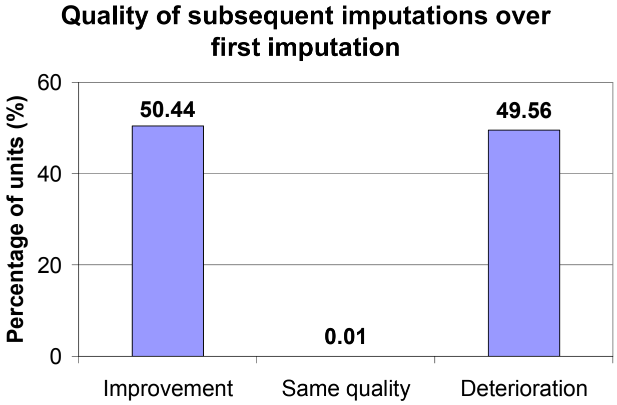

Figure 5: Quality of subsequent imputations over the first imputation for late transactions in the last quarter of 2007.

In our results, we saw improvements in subsequent imputations over the first imputation for about 50% of the units. On the other hand, deterioration in subsequent imputations over the first imputation is observed for the other half of the units. Because of the mixed results, we cannot conclude that subsequent imputations show better quality than the first one. The extent of each improvement seen may not be enough to compensate the quality loss to the other half of the units.

3.3 Testing the Impact of Imputing Late Revenue Once

Since the first imputation is better than subsequent imputations half of the time, we proposed a strategy to impute the late revenue only once. We then applied this imputed revenue to the subsequent processing when the units are still late. That is, we set

$imputed_k = imputed_1$, where k=2, 3, ..., r-1.

In that case, we eliminate the instability produced each time when we re-impute.

Based on all late transactions selected in this study for five different industries found in sub-annual business surveys in StatCan, namely manufacturing, wholesale, retail, food services and business and consumer services industries, Table 2 shows the difference in revenue being imputed once versus revenue being re-imputed at the second processing. Under the proposed strategy, we observed a difference in the data: if we re-impute, there is no gain in the quality; instead, we are imposing a revision at the subsequent processing. For example, for the food services industry, we were imposing a revision of 2.9% to the second processing of the late transactions identified in the last quarter of 2007.

Table 2: Revisions in revenue for late units in specific industries.

Industry	Revenue imputed once (in millions)	Revenue re-imputed (in millions)	Relative difference (%)
Manufacturing	5,580	5,617	-0.7
Wholesale	4,988	5,106	-2.3
Retail	6,847	6,806	0.6
Food services	1,847	1,901	-2.9
Business and consumer services	15,383	15,353	-0.2

3.4 Outcome from the Study

Our study showed that imputed revenue is unstable over time. Also, re-imputing the revenue month after month did not guarantee an improvement in quality. Therefore, as of September 2009, the GST Processing System at StatCan stopped re-imputing late transactions. Revenue is only imputed once when it is first identified as late. The same value is then used in subsequent processing of a given reference month if the unit is still late.

4. Conclusion and Future Work

We conclude that imputation of revenue has a good overall performance. Also, imputing the revenue month over month only improves the quality for about half of the units. By eliminating the effect of re-imputation, we were able to reduce the month-to-month fluctuations induced by updates brought to units in the strata. In this way, the proposed imputed revenue is more stable over time, without altering the quality of imputation.

In the future, we will take a closer look at the carry forward method. In this paper, we showed that the carry forward method has an accuracy of 80%, but it underestimates the revenue when the units later report any non-zero revenue. In other words, 20% of the units reported positive revenue even when their revenue was $0 in the year before. The total amount of underestimation produced accounted for 45% of the overall underestimation observed for all late units.

As part of the future work, we will verify if it is possible to identify units that cause such underestimation in the carry forward method. That is, we will look for characteristics of transactions that have a tendency to report non-zero revenue even if their revenue was $0 one year before. Once we have identified such characteristics, we will identify a different imputation method and apply such method to impute the late revenue for these units.

In addition, we will analyze the impact of eliminating the underestimation at the macro level. To this end, we could use a probabilistic approach. That is, for 20% of the time, we would impute with a non-zero revenue, while for 80% of the time we would continue to carry forward the $0 revenue from the previous year. Identification of units that are more likely to report non-zero values would help to bring the accuracy of the method to the micro level.

References

Beaulieu, M. and Quenneville, B. (2008). “Calendarization of the Goods and Services Tax (GST) data: Issues and solutions”, Proceedings of the Joint Statistical Meetings 2008, Section on Survey Research Methods, Denver, CO.

Brodeur, M. and Pierre, L. (2003). “Use of tax data: An application of Goods and Services Tax (GST) data”, Proceedings of Statistics Canada Symposium 2003, Statistics Canada.

Dubreuil, G., Hidiroglou, M. A. and Pierre, L. (2003). “Use of administrative data in modelling of monthly survey data”, Proceedings of Survey Methods Section, SSC Annual Meeting 2003, Statistical Society of Canada.

Dubreuil, G., Pierre, L. Labelle-Blanchet, S. and Liu, K. (2005). “Analyse et impact des unités présumées inactives sur la base de données de la taxe sur les produits et services”, Proceedings of the Colloque francophone sur les sondages 2005, Société Française de Statistique.