Calendarization of the Goods and Services Tax (GST) Data: Issues and Solutions

Martin Beaulieu, Benoît Quenneville
Business Survey Methods Division, Statistics Canada,
100 Tunney's Pasture Driveway, RHC11-Q, Ottawa, ON, Canada, K1A 0T6
(martinj.beaulieu@statcan.gc.ca, benoit.quenneville@statcan.gc.ca)

Abstract

Statistics Canada uses the Goods and Services Tax (GST) data as auxiliary data in several sub-annual surveys. Given that GST remittance can be done on an annual, quarterly, monthly or on a more frequent basis, the GST data must be calendarized. Calendarization means that the data have to be standardized so that they refer to a common reporting period. The calendarization system, which was built and implemented in 2002 (Brodeur and Pierre, 2003), is now being revised in order to improve the quality of the data by solving some of the issues that were raised by the users. This paper will discuss these issues and the solutions to be implemented.

Keywords: calendarization, interpolation, indicator series, administrative data.

1. Introduction

In order to reduce response burden and data collection costs, Statistics Canada (StatCan) business survey programs use more and more administrative data such as the Goods and Services Tax (GST). For example, sub-annual surveys use GST data for imputation by using GST revenue as an auxiliary variable, or for estimation by directly replacing survey data by GST data for small units. StatCan's Business Register also uses GST revenue as a size measure for all establishments.

The GST is a value added tax levied on the final consumption of products and services. Once a month, StatCan receives two files from the Canada Revenue Agency (CRA), one containing the reports from each company and one containing the characteristics of the accounts. Reports from companies will be referred to as transactions. Upon reception of CRA data by StatCan, an edit & imputation (E&I) system including an outlier detection module is executed, followed by the calendarization process (Quenneville, Cholette and Hidiroglou, 2003). For further details on the GST data processing, see Brodeur and Pierre (2003).

Businesses have to report their GST data to the CRA at a given frequency depending on their annual revenue. Businesses with annual revenues greater than 6M$ must report each month, those with revenue between 1.5M$ and 6M$ must report on a quarterly basis and those with revenues smaller than 1.5M$ must report annually. Those boundaries set how often businesses have to report, but businesses can elect to report more frequently than what they are requested. Furthermore, reporting periods do not have to follow calendar months. For example, a monthly remitter can decide to report its revenue from the 15th day of month to the 14th of the following month or report on a "quasi-monthly" frequency, *i.e* periods of four weeks with occasional transactions covering five weeks.

Since most StatCan GST data users are monthly or quarterly surveys, the GST data needs to be calendarized to provide users with monthly data. The goal of the calendarization process is to produce monthly estimates of revenue that are in agreement with the reported revenue on the transaction. For example, for a quarterly remitter, the three monthly estimates produced by the calendarization process must add up to the quarterly revenue reported on the transaction.

Section 2 gives an overview of the calendarization process currently in place. This process is under review in order to improve the quality of the data. Sections 3 to 6 discuss in detail four main issues and their solutions. The issues are related to the selection of the businesses contributing to the estimation of the indicator series; the estimation (and the seasonality) of the indicator series; the daily interpolations process; and the revision strategy. The paper concludes with a short summary.

2. Overview of the Calendarization Process

As mentioned above, calendarization of the GST data is needed due to different reporting frequencies of the businesses, reporting period of different lengths and the StatCan GST data users' need for calendar months estimates. The basic idea of calendarization is to use a monthly indicator for the temporal distribution of the unit's revenue value. This is a benchmarking application where the series to be benchmarked is a monthly indictor series, and the benchmarks are the unit's revenue values on their transactions. The level of the calendarized revenue is thus driven by the revenue reported (or imputed) on the transaction. The monthly movement in the calendarized value is driven by the monthly movement in the indicator series. More information on benchmarking and temporal distribution can be found in Dagum and Cholette (2006). The indicator series are constructed by industry, using the 6-digit North American Industry Classification System (NAICS) as the industry identifier. In total, there are 1009 indicator series in the GST calendarization system

The indicator series is made of the monthly averages of the revenue over the units that are selected to be part of the pool of contributors to the indicator series. A unit is included in the pool of contributors if at least 80% of its transactions are monthly or quasi-monthly. There are two steps to the estimation of the indicator series; first, we obtain strict monthly estimates for the contributors to the indicator series, *i.e.* for the quasi-monthly remitters. Those strict monthly estimates are obtained via a daily interpolation process. Once daily revenues are obtained, they are added over the days of a month to get the corresponding calendar month revenue. The second step is the actual estimation of the indicator, which is the average revenue of the monthly and quasi-monthly remitters using their strict monthly revenues. For more details, see Quenneville, Cholette and Hidiroglou (2003).

The current calendarization process was built in 2002. At the time, some assumptions were made due to a lack of historical data. Six years later, ten years of data are now available to test these assumptions and users have made comments and requests. The calendarization process is currently under review to improve quality of the data. An in-depth study of the process was done and four issues were identified as priorities. These issues are: (i) the criteria for the selection of contributors to the estimation of the indicator series; (ii) the estimation and in particular the seasonality of the indicator series; (iii) the complexity of the daily interpolation process and its processing time; and finally, (iv) the revision strategy. Each of these issues will be discussed in details in the following sections, as well as the solutions to be implemented.

3. Contributors to the Indicator Series

To obtain a reliable monthly movement, we must have an indicator series that reflects the accurate seasonal pattern for a specific industry. In recent years, some unstable indicator series were observed. Two elements of the quality of the indicators series were looked at in the revision of the calendarization process: the pool of contributors and the seasonality of the indicators. The latter will be discussed in section 4.

The only selection criterion for a unit to contribute to the indicator series was that at least 80% of its transactions had to be reported on a monthly or quasi-monthly basis. The absence of more refined criteria paved the way to various problems. First, even with an outlier detection module in the edit and imputation system, there is still a possibility for a few outliers to make it through the system and be part of the pool of contributors. An outlier in the pool of contributors can lead to an indicator series with outlier values, which may generate outliers in the monthly data of all quarterly and annual remitters in that industry.

Figure 3.1 shows an example of a significant outlier in the indicator series of the Automotive Repair and Maintenance industry, in May 2003. A closer look at the industry's contributors indicated that this outlier was due to only one of the units in the pool of contributors. With such an outlier in the indicator series, all annual remitters in that industry have most of their 2003 annual revenue assigned to the month of May. The dotted series shows that the indicator series is more stable with 2003 annual revenue distributed more evenly throughout the year.

Another issue identified was the presence of complex units, *i.e.*, in this context, units active in more than one field of activity. Since the goal of the indicator series is to provide an industry specific monthly movement, the use of multi-activity companies can produce an unrepresentative movement. As such, a multi-activity business may have only 30% of its revenue coming from the industry it is contributing to. This means that some indicator series could be based on businesses coming from other industries. For example, Figure 3.2 shows the impact of adding a few multi-activity units had in 2002 in the Sporting Goods Stores. The indicator series, which used to be very stable, saw a level shift and a change in the monthly movement. As shown by the dotted line, the removal of those units allowed maintaining a stable series and a stable monthly movement.

Finally, another issue was the use of quasi-monthly units in the pool of contributors. This did not lead to issues related with the quality of the indicator series, but issues with the performance of the system. As their name says, the quasi-monthly units do not cover perfect calendar months. In order to use them in the indicator series, a daily interpolation step was necessary. Once daily revenue values are obtained they are summed over all days of the calendar month to obtain a monthly value. This daily interpolation process was time consuming and complex. Tests have shown that the removal of the quasi-monthly units from the pool of contributors has no significant impact on the monthly movement of the indicator series. Thus, only perfect monthly remitters will be used in the pool of contributors. This allows to improve the systems efficiency without significant impact on the monthly movement.

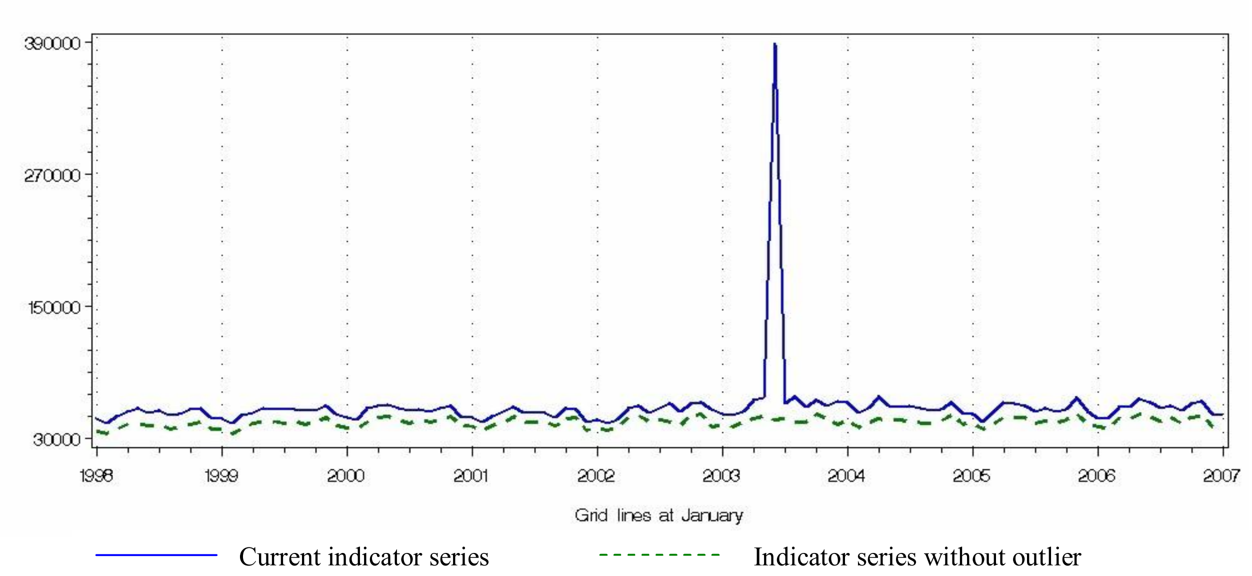

Figure 3.1: Impact of outliers in the pool of contributors to the indicator series

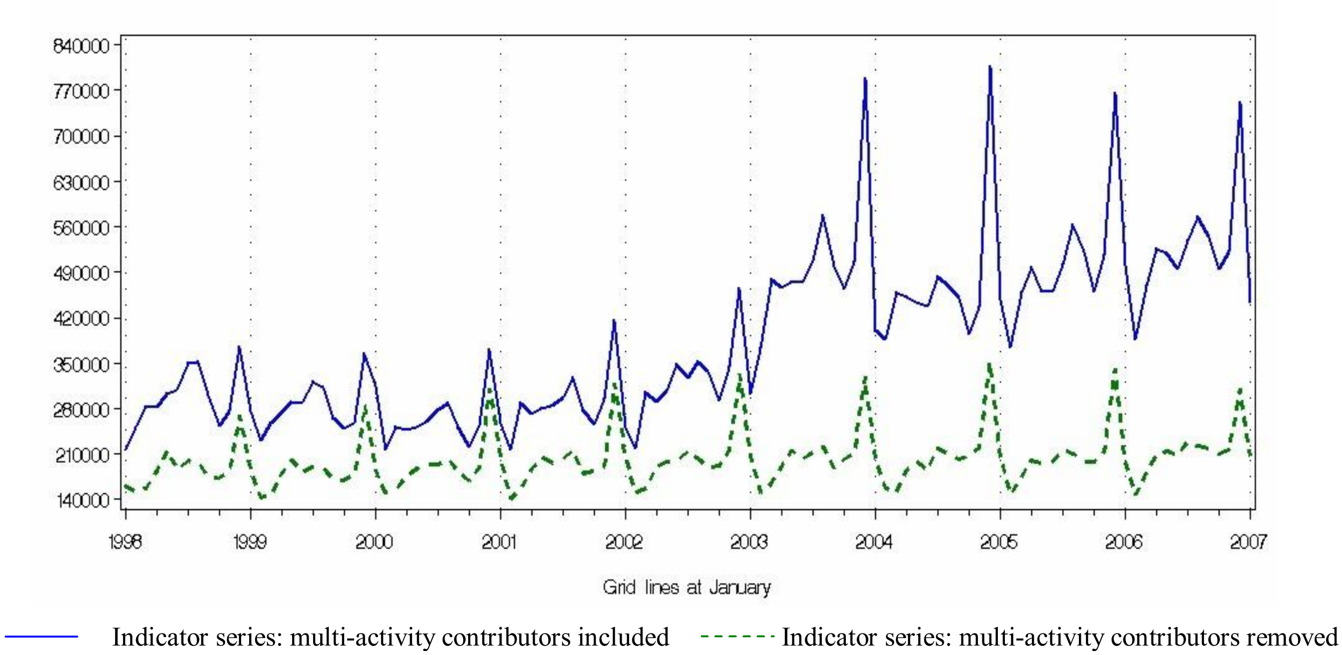

Figure 3.2: Impact of multi-activity contributors to the indicator series

The following summarizes changes made to the selection criteria used to create the pool of contributors to estimate the indicator series: (i) suspect outliers identified in the E&I process are no longer included; (ii) multi-activity units are no longer included; and (iii) only perfect monthly remitters are selected. These changes improved the efficiency of the system and the stability of the monthly movement. The level of the indicator is lowered by these changes, but since the level of the calendarized value is driven by the reported revenue on the transaction, these changes have no impact on the level of the calendarized monthly revenue.

4. Seasonality of the Indicator Series

The improvements made to the pool of contributors have lead to more stable monthly movements. The next aspect considered is the seasonality of the indicator series. The current system systematically calculates an indicator series for every industries and applies the monthly movement of the indicator series to all quarterly and annual remitters within each industry. One could wonder if all industries have a significant seasonal pattern. If it is not the case, then the current process would apply an erratic movement to some industries.

In order to verify the seasonality of the indicator series, U.S. Census Bureau's X-12 ARIMA was run on all 1009 indicator series. X-12 ARIMA is a seasonal adjustment program which decomposes series in different components such as trend-cycle, seasonality and calendar effects (Findley et al., 1998). This has allowed the identification of indicator series with significant seasonal patterns, outliers, trading-day effect and Easter effect. Table 4.1 summarizes the results.

Table 4.1: Significant components of the 1009 indicator series

All Series		%
Outlier(s)	623	61.74%
Seasonal effect	676	67.00%
Trading day	532	52.73%
Easter effect	300	29.73%
Total	1009	100%

Table 4.1 shows that two thirds of the series had a significant seasonal effect (67%), close to 62% had at least one outlier, a little more than half of the series had a trading-day effect (53%) and about 30% had an Easter effect. This shows again that the indicator series were sensitive to outliers. It also shows that, for a third of the series, an erratic movement is applied to calendarize while the series have no significant seasonal effect.

The solutions to correct these issues were to: (i) use constant indicator series for the non-seasonal industries and (ii) use the combined seasonal and calendar adjustment factors (table D16 of the X12-ARIMA output) as the indicator series for the seasonal industries. For the seasonal industries, the use of the combined seasonal and calendar factors permits a more stable temporal distribution over the months by removing the contaminating and unnecessary effects of the trend-cycle, outliers and irregular components on the temporal distribution of the business revenue. For the non-seasonal industries, the use of a constant indicator series produces a smooth temporal distribution of the business revenue over the months without spurious jump between the last month of a transaction and the first month of the next transaction, the so-called step problem that would be obtained by allocating the average monthly revenue to the months pertaining to a given transaction. Figure 4.1 shows the impact of using a constant series for a non-seasonal industry on the aggregated calendarized data. This graph shows that the average levels are similar, but with less irregular variations on the aggregated industry total when the constant indicator series is used.

Another advantage of this improvement will be that the indicator series (seasonal or not) will be stored in a database and revised only once a year. The calculation of the indicator series will not be done every month as it is currently the case, and thus, processing time will be reduced accordingly.

5. Daily interpolations

The daily interpolations are computed for 'non-perfect' remitters. These are remitters that do not report all their transactions on perfect calendar months, *i.e.* there is at least one transaction not starting on the first day of a month and/or ending on the last day of a month. As mentioned above, the daily interpolation process is time consuming and is by far the most complex algorithm of the calendarization system. The goal was to find a way to accelerate and simplify the daily interpolation portion. Part of the solution has already been discussed in section 3 with the removal of the quasi-monthly units from the pool of contributors. However, daily interpolation is still needed for the calendarization of the "non-perfect" remitters.

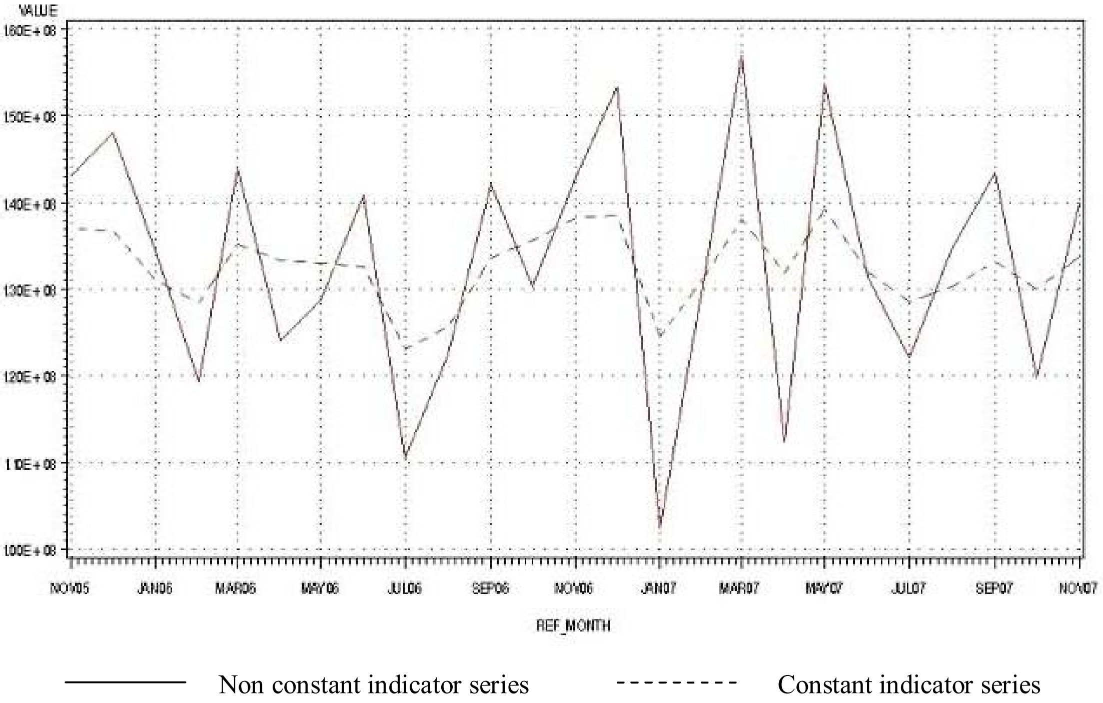

Figure 4.1: Impact of constant indicator on calendarized revenue aggregates

The already existing SAS® procedure Proc Expand can be used to calculate daily interpolation using natural splines. The daily interpolations are calculated on the cumulated revenue against a cumulated daily weight. This daily weight is the product of the estimated X-12-ARIMA's trading-day weight with the appropriate monthly seasonal and holiday factor. As a result of the spline interpolations, we obtain daily values for the cumulated revenue. We get a final monthly value by de-cumulation, *i.e.* by subtracting from the cumulated revenue of the last day of the current month the cumulated revenue of the previous month. Figure 5.1 is a graphic representation of an example. Each dot represents the cumulated revenues at the ending date of each transaction and the curve represents the spline interpolations obtained via SAS® Proc Expand.

In this example, a business reports its first transaction from December 30th to January 31st with $419,454 of revenue. A second transaction covers February 1st to February 23rd with revenues of $104,121. Hence, the cumulated revenue on February 23rd is the sum of the two revenues, which is $523,575. A third transaction starts on February 24th and ends on March 22nd with revenues of $371,530. The cumulated revenue on March 22nd is $895,105, and so on. We obtain as a result of the spline interpolations a value of $585,481 on February 29th. Using that value, the February calendarized revenue is derived by as following $585,481 - $419,454 = 166,027$.

This method has the advantage of being easy to implement and using an already existing procedure. After the daily interpolation step, all the transactions of all the remitters are on "perfect" calendar months. That is, all the businesses have edited transactions with a starting date on the first day of a month and an ending date on the last day of a month. The next step is to use Statistics Canada's Proc Benchmarking (Latendresse et al., 2007) to obtain the temporal distribution of the business' revenue by benchmarking its industry indictor series to its edited perfect calendar month transactions. Once this benchmarking application is completed, the resulting calendarized values are stored in a database. The use of this database will be explained in the next section.

6. Revision Strategy

Another problem with the calendarization process was its revision strategy. With the current process, calendarization is executed every month, and, in the process, data as far back as 1998 is updated every month. This leads to two major issues. First, in terms of systems, space requirements and execution time keep increasing every month. Currently each production cycle overwrites the production of the previous month. Second, from a methodological point of view, this process tends to deteriorate older data.

This deterioration has two causes. The current system has a static NAICS classification, meaning that if a business changes its main activity, the history of that business will be changed to be in line with today's NAICS classification. This will be fixed by using a dynamic NAICS classification. Under the new approach, if the business NAICS changes on a given month, the change will be applied from this month and forward only. The history will not be changed.

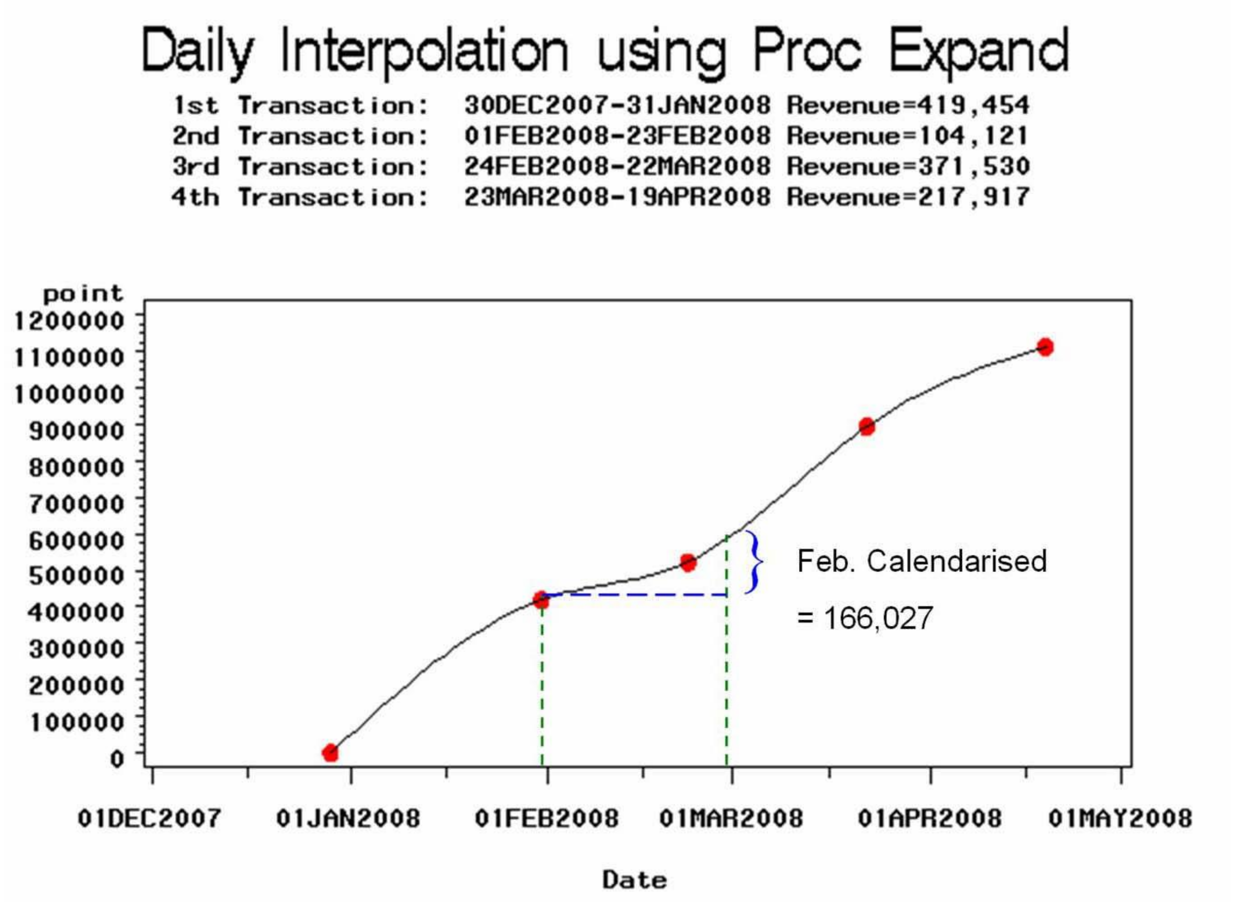

Figure 5.1: Example of calendarisation using SAS® Proc Expand's spline interpolations

The second cause of deterioration is due to the fact that the indicator is recalculated every month with a pool of contributors also updated every month. This means that 1999 data can be calendarized based on the movement of businesses that did not exist then. A part of the solution to that problem comes from the use of the seasonal factors, which will be calculated only once a year, and only the most recent years could see their seasonal factor revised.

The other part of the solution is to use Link-Benchmarking. With the calendarized data kept in a database, this technique will allow the revision of only the recent months using a sliding interval. The following example illustrates the approach. Let us take a perfect quarterly remitter reporting in March, June, September and December for which we want to obtain the calendarized value for October. The idea is to go back three transactions (September, June and March) and use the previously calendarized value for February as a link-benchmark. Calendarized values prior to February are not revised. The calendarized value for February is used as if it were a transaction along with the March, June and September transactions. The indicator series from February to October is benchmarked to those transactions to obtain the calendarized values for February to October. Obviously, the February calendarized value remains unmodified. For the November processing month, February would remain the link-benchmark as no new transaction has been reported. In December, a new transaction is reported. Going back three transactions (December, September and June), the new link-benchmark becomes the May calendarized value. The calendarized values prior to and including May are now frozen in the database.

Link-benchmarking will also be used in the daily interpolation process where the link-benchmark will be the interpolated cumulated daily revenue on an given day in the recent past. This approach is not yet finalised.

7. Conclusion

The revision of the GST calendarization process is a three-year project with two objectives. The most important objective is to improve the quality of the data. After an in-depth analysis of the process, the ways identified to achieve that goal were: to refine the selection criteria for the indicator series' pool of contributors; to assess the seasonality of the indicators series used; to speed-up and simplify the daily interpolation; and to implement a revision strategy. As a refinement to the pool of contributors, only businesses who report for perfect calendar months, who are active in only one industry and who do not have suspect outliers are included. Constant indicator series are used for around 400 industries for which no significant combined seasonal/calendar pattern was found. For the remaining industries, the combined seasonal-calendar and holiday factors will be used as the indicator. X-12-ARIMA will be used for the purpose of updating the indicator series. Daily interpolation will be simplified by using spline interpolations on the cumulated revenues, using SAS® Proc Expand. Finally, the use of Link-Benchmarking with StatCan's Proc Benchmarking will be used to calendarize revenue and to reduce the amount of revisions executed every month.

The other important objective is to reduce the space required by the calendarization system, to speed-up the execution time and to simplify the overall methodology of the system. It will be achieved by removing the quasi-monthly remitters from the pool of contributors. This eliminates the need for daily interpolations in the estimation of the indicator series. With the use of the combined seasonal and calendar factors as the indicator, the indicator series itself will be updated only once a year and the monthly production will find the indicators in a database of indicator series. The use of SAS® Proc Expand and StatCan's Proc Benchmarking will also simplify the system, facilitate future modifications and facilitate the knowledge transfer due to the excellent documentation that comes with those procedures.. Finally, the revamped revision strategy will reduce and stabilize greatly the amount of transactions processed in the monthly calendarization of the GST revenue.

Some of these changes, such as the criteria for the selection of contributors to the indicator series and the constant indicator series for non-seasonal industries, have already been implemented. So far, the users from the sub-annual surveys are satisfied with the results as it already stabilized the calendarized data. The rest of the changes should be implemented early in 2009.

Acknowledgments

The authors would like to thank Pierre Cholette and Estella Bee Dagum for their advice during that revision process, as well as Frédéric Picard and Marie-Ève Mainville for their contribution to the project.

Bibliography

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.

Quenneville, B., Cholette P., Hidiroglou, M. (2003). “Estimating Calendar Month Values from Data with Various Reporting Frequencies”, 2003 Proceedings of the Joint Statistical Meetings, Business and Economic Section, JSM 2003, San Francisco, CA.

Bee Dagum E., Cholette, P. (2006). “Benchmarking, Temporal Distribution and Reconciliation Methods of Time Series”. Springer-Verlag, New York, Lecture notes in Statistics, #186.

Findley, D., Monsell, B., Bell W., Otto M. and Chen B. (1998): “New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program,” Journal of Business and Economic Statistics, 16, 127–177.

Latendresse, E., Djona, M., Fortier, S. (2007): “Benchmarking Sub Abnnual Series to Annual Totals – From Concepts to SAS ® Procedures and SAS® Enterprise Guide Custom Task,” 2007 SAS® Global Forum Proceedings