Interpolating and standardizing time series data covering various fiscal intervals using splines

Jack Lothian – Statistics New Zealand

Statistics New Zealand is in the midst of implementing its 2010-2020 Strategic Plan that will transform how the agency functions. The "administrative data first" philosophy is a critical component in the transformation process and the Goods and Sales Tax (GST) data is a key dataset for transforming business surveys. GST can provide data which can be used in sub-annual surveys to replace directly surveyed units or to improve in editing or in calibration. These processes could potentially reduce response burden and collection costs plus improve quality. However, the use of administrative data poses major challenges. Due to late filings, GST data are not all available on time during the production cycle and the data covers a melange of varying and overlapping time intervals. We propose a calendarization method based on interpolating the cumulated flows with splines that provides data with standardization time intervals and short-term forecasts. The methodology improves the timeliness and quality of the GST data and increases the willingness of the survey programs to embrace tax data.

Key words: Calendarization, Administrative data, Cubic splines, Interpolation, Significance editing, Forecasting, Standardization

1. Introduction

In its Statistics 2010-20 Strategic Plan, Statistics New Zealand(Stats NZ) has committed to using “administrative data first” whenever that is feasible. Administrative data will be supplemented by direct collection where necessary. To achieve this objective, the administrative databases must be designed to support regular business survey production cycles. To support ongoing production cycles the data must be standardized, of a reasonable quality and accessible in a timely manner. The amount and quality of the administrative data input into the production cycle cannot change significantly from cycle to cycle. In addition, the databases must permit business surveys to control information gaps in or overlap of coverage across industries and sectors. In my experience this implies that administrative data must strive to provide unit level estimates of key variables for all the in-scope units input into the national accounts by the business surveys. Dozens of ongoing regular business surveys plus numerous ad-hoc or occasional surveys must all be able to extract current/clean/consistent/non-overlapping unit responses for the administrative portion of their survey. In addition, the estimates must be time stamped and be on a consistent calendar basis. In summary to achieve a system that maximises the use of administrative data, one must maximise the consistency, quality and coverage of the unit administrative data. For a more detailed discussion see (Seyb, McKenzie, and Skerrett 2013).

Within a business survey environment, administrative data typically has the following usages:

1. Frame or Business Register maintenance;
2. Improving/enhancing aggregate business survey estimates through:
    I. Calibration of aggregate estimates;
    II. Editing aggregate estimates (macro edits);
3. Improving/enhancing unit responses through:
    I. Replacement of direct survey units;
    II. Editing direct survey responses (micro edits);
    III. Imputation for field and total unit nonresponse.

While Stats NZ's strategic plan focuses on usage 3.I, all of these usages will be required at various points in the production cycle and thus all these usages need to be potentially supported. Key standardization issues for all the uses are calendarization and imputation for data gaps. The data cannot be a melange of time stamps and reporting time intervals with randomly appearing information gaps. The steps that are required to clean and standardize the data are:

1. Calendarization
2. Outlier detection
3. Imputation for unit and item non-response and error correction
4. Forecasting delayed responses

Most countries that process sub-annual administrative data implement these steps in varying orders. This paper will focus on one particular administrative data source: Generalized Sales Tax (GST) data. Section 2 of the paper presents the data and its challenges. Section 3 outlines the proposed standardization methodology and section 4 gives the conclusions.

2. The data challenges

All NZ businesses with sales in excess of $60,000 NZ must file GST reports with the Inland Revenue Department (IRD). GST net revenue is a pseudo value-added tax (VAT) which is calculated by subtracting GST eligible expenses from GST eligible sales. As is typical in most GST/VAT systems larger businesses are expected to file more frequently than small businesses. Thus businesses with over $24 million in sales must file monthly, businesses over ½ million in sales must file bi-monthly and business over $60,000 must file every 6 months. Firms filing bi-monthly can choose to file on either odd or even months and firms filing 6-monthly can start filing on any of the first 6 months in the year. Thus the GST transactions contain 9 different time interval patterns. There are about 440,000 firms reporting GST transactions who generate about 2.5 million transactions per year distributed as shown in Table 1.

Table 1: Share of GST transactions and sales by filing frequency: 1997-2013

Filing frequency	Transactions	Sales
Monthly	12.2%	59.8%
Bi-monthly	75.2%	37.3%
6-monthly	12.6%	2.9%

Additionally, about 1 in 100 firms change their reporting frequency each month and over a firm's lifespan, approximately 1 in 6 firms change their reporting period. Moreover, 1 in 14 transactions have no valid reporting frequency coded on the transaction. This frequently occurs during a firm's start-up period or when the reporting period changes or after a period of non-reporting. The melange of 9 different time-beats plus the uncertainty of properly identifying the reporting period makes using the GST data in a production environment a challenge.

To further complicate the issue, IRD appears to permit firms to file (and to not file) null transactions for an individual reporting period and then catch-up in subsequent transactions. Table 2 summarizes the occurrence of null or no activity transactions by filing frequency.

Table 2: Percentage share of null GST transactions by filing frequency: 1997-2013

Filing frequency	Percent of null transaction present
Monthly	34.0%
Bi-monthly	22.1%
6-monthly	25.8%
Unknown	24.6%
Overall	24.1%

Approximately, 1 in 4 of the transactions show no activity reported and unexpectedly the largest firms are most prone to report no activity in a given month. Large firms are showing no sales activity in a third of the reporting periods. As shown in Table 3, the majority of the nulls are occurring in a time period between two other transactions displaying positive activity.

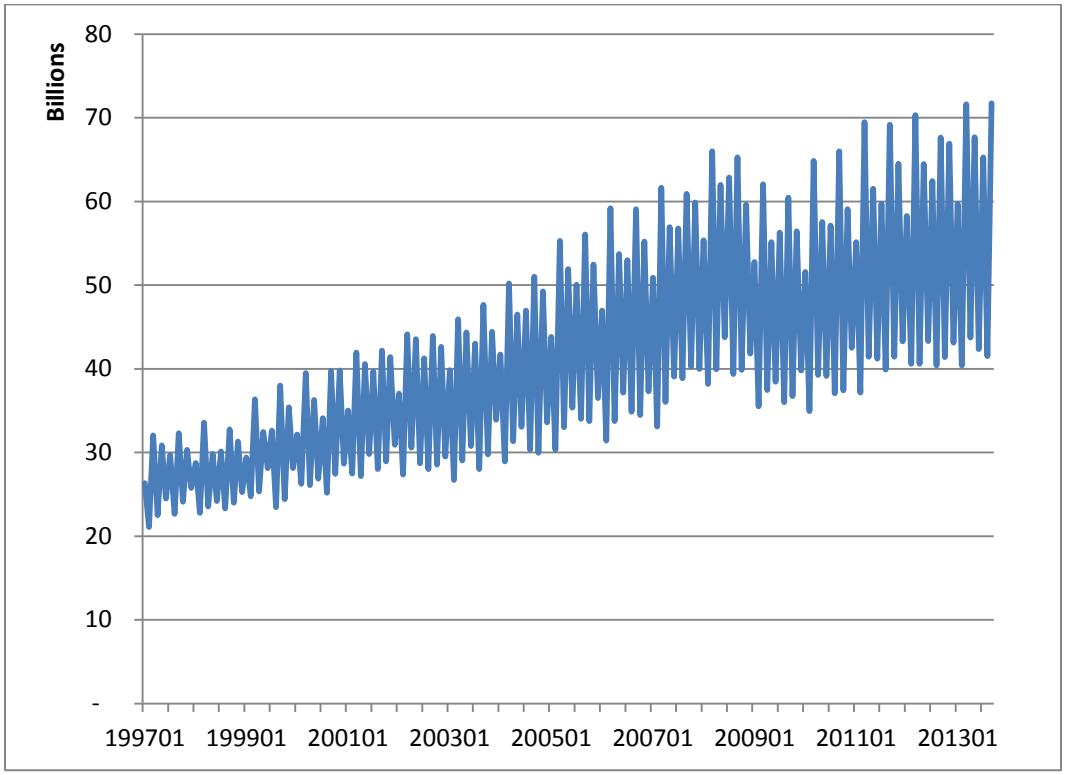
Table 3: Distribution of percentage share of the nulls by time location: 1997-2013

Time location of the nulls	Share points
Nulls in time series that never show any activity	0.8%
Nulls occurring at the beginning or end of the time series	10.9%
Nulls occurring in the middle of an active period in the time series	12.4%
Overall	24.1%

Note that the GST database at Stats NZ does not cover the full reporting period; the database starts in 1996 while GST started in 1986. In addition, months subsequent to year 2013 are not in the database. Thus some of the nulls at the beginning and end of the time series in the database are actually in the middle of otherwise active periods in the full

extended time series. Evidently, most of the observed nulls are concentrated in the middle of otherwise positive activity. The nulls appear to be creating an effect that I call "heaping". Heaping implies that firms file a null transaction followed by a catch-up transaction in the next filing period. The mixture of complex beat-frequencies plus the appearance of random zeroes makes identifying outliers and developing an imputation strategy for non-response problematic. The number of apparent outliers generated by these null responses is orders of magnitude greater than the number of statistical outliers expected. Figure 1 shows the aggregate sales time series for the raw IRD transaction.

Figure 1: Aggregate raw IRD GST sales, monthly time series: 1996-2013



The time inconsistencies confound many of the standard methodologies that are used for cleaning the data and imputing late responses. Without standardizing the time reporting, the successive standardization routines become exceedingly complex black boxes that fundamentally modify a majority of the raw information. Legally, firms must report all their GST eligible revenues and thus the yearly total of a firm's transactions should equal their actual GST eligible revenues during the year. Thus in theory, the yearly totals of a firm's transactions should be preserved. Applying this constraint is difficult.

3. The proposed standardization methodology

Standardizing time

The challenges of standardizing Stats NZ's GST data are so great it seems almost impossible to achieve yet the calendarization strategy recently outlined in (Quenneville, Picard, and Fortier 2013) resolves many of these difficulties. Following the methodology in this paper, we propose a calendarization method based on interpolating the cumulated flows with splines. The output is data with standardization time intervals. The splines standardize the reporting frequencies, eliminate the changes in reporting frequencies and backfill the null transactions. Seasonal factors can be propagated downwards from the observed monthly series to the bi-monthly series and then to the 6-monthly series by transforming the time axis.(Beaulieu and Quenneville 2008) This imposes the observed monthly seasonal factors downwards onto the bi-monthly transactions. The interpolated bi-monthly time series are then seasonally adjusted and the seasonal factors from the bi-monthly are imposed downwards onto the interpolation of the 6-monthly series. The first step in this process is removing all the null transactions from the time series and replacing them with missing value indicators. Next we must transform the GST sales flow (s) into a cumulate or stock (S) by defining:

$$S^T = \sum_{t=t_1}^{t_1+T} s^t$$

Then re-define time (t) as τ :

$$\tau^T = \sum_{t=t_1}^{t_1+T} \left(SF\right)^t$$

where SF^t are the imposed external multiplicative seasonal factors. Note that τ^T will be defined at intermediate time points periods (months) where S^T may not be observed. The missing S^T will be the interpolation points that we desire. Then we fit an interpolating spline through the knots $\left(S^T, \tau^T\right)$. Next, we read off on our curve the interpolated S^T values at all the defined τ^T including the points where S^T was unobserved. The GST flow is then derived:

$$s^{\tau^T} = \Delta S^{\tau^T} = S^{\tau^T} - S^{\tau^{T-1}}$$

The untransformed time variable is simply the index variable $\left(T\right)$ from τ^T . This process injects the pre-defined seasonal factors into the spline fit. The process is akin to standard time series benchmarking techniques. The interesting point is this procedure preserves the raw cumulant values. No sales value is added or subtracted to the time series. For example, if the time series is from a bi-monthly reporter: then re-collapsing the observed time period back to a bi-monthly time series will reproduce the observed raw bi-monthly

data exactly. The spline just drags sales values backward to fill the time gap under a seasonal constraint. If one assumes that all GST revenue is eventually reported then this procedure should be a reasonable assumption. One of our desired objectives was minimizing changes to the actual observed data and this procedure leaves the original observations untouched. We believe that modifying the data as little as possible while standardizing is a strong and positive trait for this procedure. The procedure has the added strength of being easily explainable to non-technical persons. Figure 2 shows the result of applying this spline procedure to the aggregate GST sales data.

Figure 2: Creating the aggregate time standardized GST sales data: 1997-2013

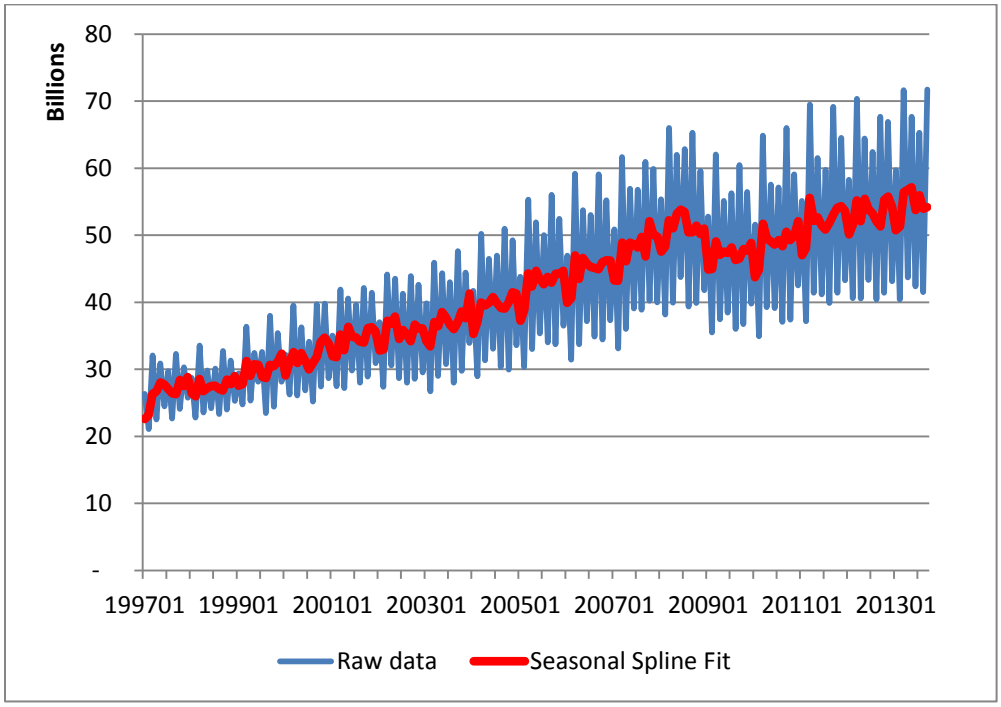
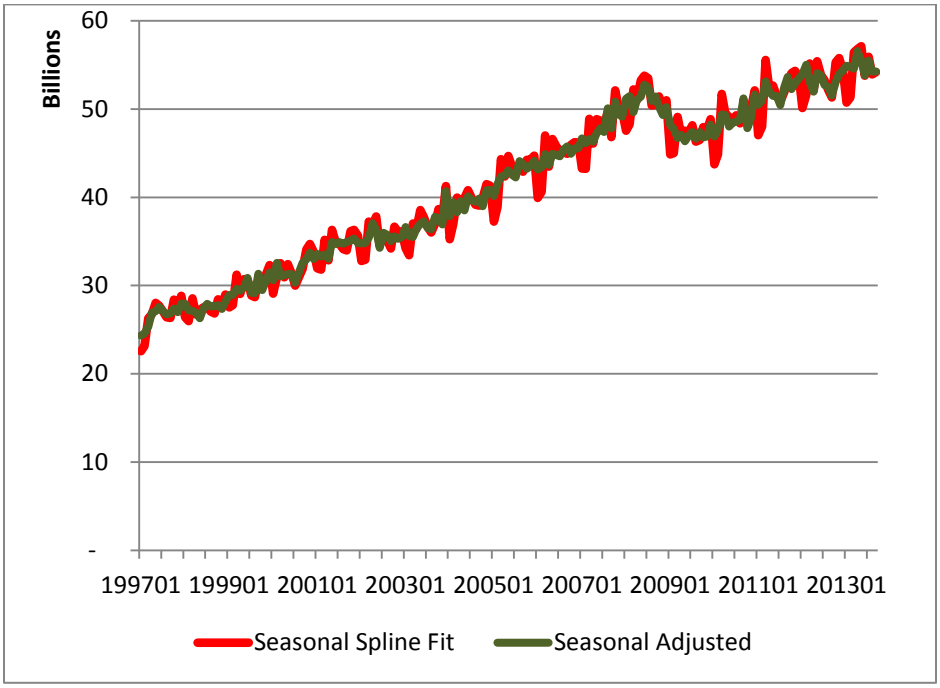


Figure 3 shows the spine-generated aggregate GST sales and the time series with the seasonality removed. The trends are clearly discernable now but the month-to-month changes are still somewhat volatile.

Figure 3: The aggregate seasonally adjusted GST sales: 1997-2013



So in summary, the procedure standardizes reporting periods and eliminates the "heaping" effect, while not changing any of the original observed raw data cumulants. As an added benefit it preserved, as much as possible, the movements of seasonality observed in the large monthly reporters. As an added bonus, the spline interpolation uses the SAS procedure PROC EXPAND and it is trivial to implement and processing is quick even for large time series bases.

Outlier detection

After completing the spline calendarization, the GST data set is standardized to the point where selective editing (also called significance editing or macro editing) can be applied to identify extremes and serious coding issues (de Waal 2013). The idea of selective editing is that edits will be applied based upon an individual records effect on the aggregate or stratum estimates. Small firms with volatile sales but who have minimal impact on the estimated aggregates for the industry will be ignored.

The first step is the macro-level flagging of significant changes in time of stratum estimates. These absolute changes are skewed, have kurtosis and contain trends and seasonal effects. Various transformations of the data can eliminate or mute these effects. To begin, convert the stratum totals into a year-over-year growth rate time series R_h^t; this eliminates seasonal and linear trend effects.

$$R_h^t = X_h^t \Big/ X_h^{t-12} = \sum x_{h,i}^t \Big/ \sum x_{h,i}^{t-12}$$

Unfortunately, the resulting transformed distribution still has significant skewness and kurtosis. If we then take a logarithmic transformation we tend to eliminate skewness but the distribution still may have heavy tails or kurtosis. To address this issue we can use non-parametric estimators for the location (μ) and scale (σ) parameters. Thus the transformed macro-growth variable is:

$$\mathrm{LR}_h^t = \log\left(R_h^t\right)$$

Then parametize the distribution by estimating the median and inter-quartile range of the T values of LR_h^t. The median becomes the estimator for μ_h^R, while $(IQR/1.349)$ becomes the estimator for σ_h^R. A significant (at the 1% level) macro change might then be identified by (hopefully after these transformations are applied, we can appeal to a normal approximation):

$$\frac{abs\left(LR_h^t - \mu_h^R\right)}{\sigma_h^R} > 3$$

Significance editing states that outliers should only exist in stratums that fail this test. (We will relax this constrain eventually.) Alternately, the growth in the stratum total X_h^t can be written as:

$$\Delta X_h^t = R_h^t - 1 = \sum \frac{\left(x_{h,i}^t - x_{h,i}^{t-12}\right)}{x_{h,i}^{t-12}} x_{h,i}^{t-12} = \sum \left(\frac{x_{h,i}^t}{x_{h,i}^{t-12}} - 1\right) x_{h,i}^{t-12} = \sum \Delta r_{h,i}^t x_{h,i}^{t-12}$$

If we assume $r_{h,i}^t$ and $x_{h,i}^{t-12}$ are independent then aggregate change arises from two multiplicative factors or effects: a size $\omega_{h,i}^t = \left(x_{h,i}^t / X_h^t\right)$ effect and a unit or micro-change effect from $r_{h,i}^t$. Then we can go through the same procedure we used for R_h^t with $r_{h,i}^t$ and define our transformed micro-level variable as:

$$\mathrm{lr}_{h,i}^t = \log\left(r_{h,i}^t\right) = \log\left(x_{h,i}^t \Big/ x_{h,i}^{t-12}\right)$$

A significant (at the 1% level) micro changes would then be identified by:

$$\frac{abs\left(lr_{h,i}^t - \mu_h^r\right)}{\sigma_h^r} > 3$$

Significance editing says that an outlier must fail both the macro and micro level tests and $\omega_{h,i}^{t}$ must be sufficiently large to impact the stratum estimates. We could then define a significance edit score that combines these three factors (the size effect, the macro-change, and the micro-change) into one test.

$$score = \frac{\omega_{h,i}^{t}}{k_h} * \frac{LR_{h,i}^{t} - \mu_h^R}{3\sigma_h^R} * \frac{lr_{h,i}^{t} - \mu_h^r}{3\sigma_h^r} > 1$$

The parameter k_h is a tuning constant. Notice, the absolute values were removed and the test is one-sided. Macro and micro changes that move in opposite directions cannot contribute significantly to the stratum change. In addition, only changes that have a gross effect on the stratum total will be detected. If the growth within the stratum is spread across many units, then the chance of detecting an outlier diminishes. When an outlier is detected, by examining the three effects it is relatively easy to explain to a non-technical person why the point was declared or not declared an outlier. Again, the basic principles behind the methodology are minimal change to the raw data and simplicity of the explanations.

Imputation/forecasting

Finally, with clean time standardized data available, simple ARIMA models could be used to forecast current non-responses that have not been received due to late responses or edit failures. See for example, (Dagum 2010). This step has not been finalized at Stats NZ. There are three options being considered: 1) using ARIMA models as done by the U.K. Office of National Statistics (ONS) with their VAT data; 2) using deterministic linear models as done by Statistics Canada with their GST data; or 3) us the interpolating spline to do the forecast by setting the boundary conditions on the last knot so all projections are linear with an imposed seasonal as suggested by (Quenneville, Picard, and Fortier 2013). Each has its merits. The deterministic and spline projections are conservative, simple and easy to explain but sometimes the forecasts must extend forward for almost two years. One wonders if the simple spline linear projection might not be too flexible for such extended forecasts. Under these conditions, the ARIMA or a combined cross-sectional/time series deterministic model might be more successful. Of course, these procedures are harder to explain to non-technical users and present greater technical challenges to implement.

4. Conclusions

These methodologies should improve the timeliness and quality of the GST data and increases the willingness of the survey programs to embrace tax data. The key issue is standardizing the data so that business surveys can use it in ongoing production cycles. This is achieved by calendarizing the data, cleaning it, imputing for non-response, and ensuring the data is released in a timely manner for the business surveys to use on an ongoing basis.

References

Beaulieu, Martin, and Benoit Quenneville. 2008. Caldendardization of the Goods and Services Tax (GST) Data: Issues and Solutions. Paper read at Proceedings of the Section on Survey Research Methods, Joint Statistical Meeting, at Denver, Colorado.

Dagum, Estela Bee. 2010. "Time series modeling and decomposition." *Statistica* no. 70 (4):433-457.

de Waal, Ton. 2013. "Selective Editing: A Quest for Efficiency and Data Quality." *Journal of Official Statistics* no. 29 (4):473-488.

Quenneville, Benoit, Frédéric Picard, and Susie Fortier. 2013. "Calendarization with interpolating splines and state space models." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* no. 62 (3):371–399.

Seyb, Allyson, Ron McKenzie, and Andrew Skerrett. 2013. "Innovative Production Systems at Statistics New Zealand: Overcoming the Design and Build Bottleneck." *Journal of Official Statistics* no. 29 (1):73–97.