Standardizing time in payroll data 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 Business Payroll Tax (PAYE) data is an important input dataset for business and national accounts surveys. Standardized PAYE data can be used in sub-annual surveys to enhance survey data, improve editing and implement calibration. These processes could potentially reduce response burden and collection costs plus improve quality. However, the use of administrative data poses major challenges. 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 PAYE data and increases the willingness of the survey programs to embrace tax data.

Key words: Benchmarking, Calendarization, Administrative data, Cubic splines, Payroll data

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: the monthly payroll taxation for employees (PAYE) 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.

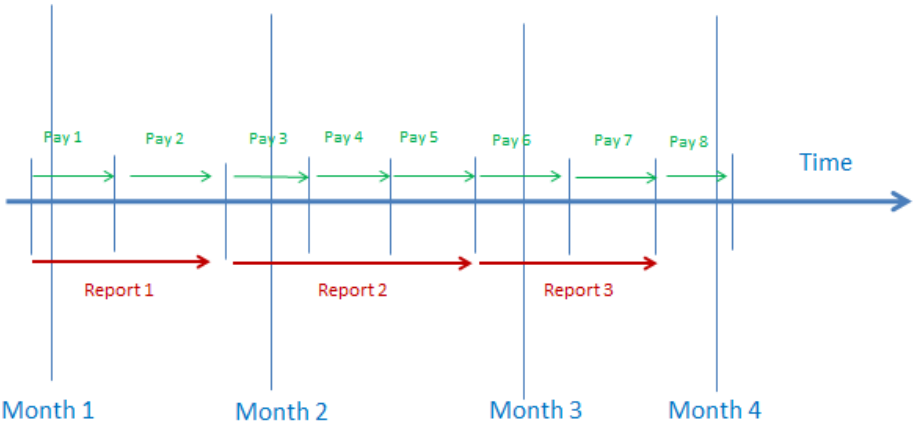
2. The data challenges

All NZ businesses with salaried employees must file monthly PAYE reports with the Inland Revenue Department (IRD). This is a rich data source covering the monthly salaried income of approximately 2.2 million NZ residents (about ½ of the population of NZ). Individual transactions for every employee are filed each month and the transactions automatically cross-link firms' IRD numbers with individuals' IRD numbers.

Superficially, one would think that the time periods for the PAYE should be already standardized since the firms file their PAYE reports monthly. The difficulty is the majority of NZ firms pay their employees weekly, bi-weekly or 4-weekly rather than monthly. IRD does not require firms to go through the costly exercise of recompiling their payroll accounts into a second monthly calendarization. Instead, they allow firms to file based upon their own unique payroll cycle. Thus firms are required to cumulate by employee all the salary paid during the filing month. The filing covers actual salary paid rather than earned and the number of pay periods can change from month to month.

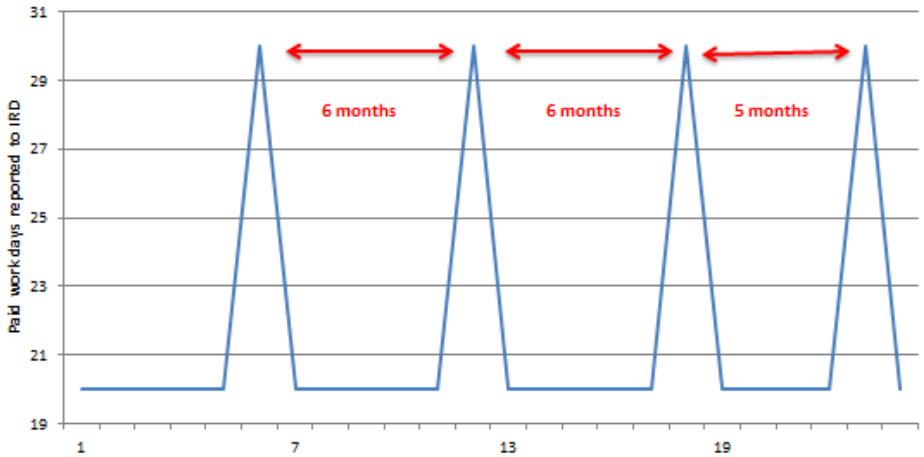
Figure 1 shows the effect of this reporting regulation for a firm with a bi-weekly payroll.

Figure 1: How payroll data gets reported to IRD by a firm with a bi-weekly payroll



The hypothetical bi-weekly PAYE firm will cumulate two payroll periods in Report 1 and 3 payrolls periods in Report 2. Figure 2 shows the impact of the hypothetical bi-weekly payroll on the number of work-days paid over a 2 year period.

Figure 2: Reported paid days for hypothetical bi-weekly payroll



The number of reported paid days fluctuates between 20 and 30 days over the various months. This implies an occasional increase of 50% from one month to another and the effect is purely a calendar effect unrelated to the economy. This effect is so large that it will swamp seasonal, business cycle and trend related changes in the data. Another troubling aspect of Figure 2 is the time intervals between these peaks are not fixed. While there is a periodicity to these interval lengths, the length of the cycle is such that effectively the length of intervals appears to randomly fluctuate between 5 and 6 months.

Firms report salary paid during the month to IRD and not salary earned. Most firms pay employees with a time lag. Thus if employees are paid on a Friday, they will not be receiving their pay for the current week but rather for a previous week. Unfortunately, the payment lag is not fixed for all firms. Some firms pay employees on a Wednesday for the previous week's earned income while other firms may go as long as 4 weeks before they pay salary earned. For bi-weekly payrolls most firms will have a payment lag between 3 and 14 day.

Figure 3: Reported bi-weekly payroll with payment to employee delayed (lagged)

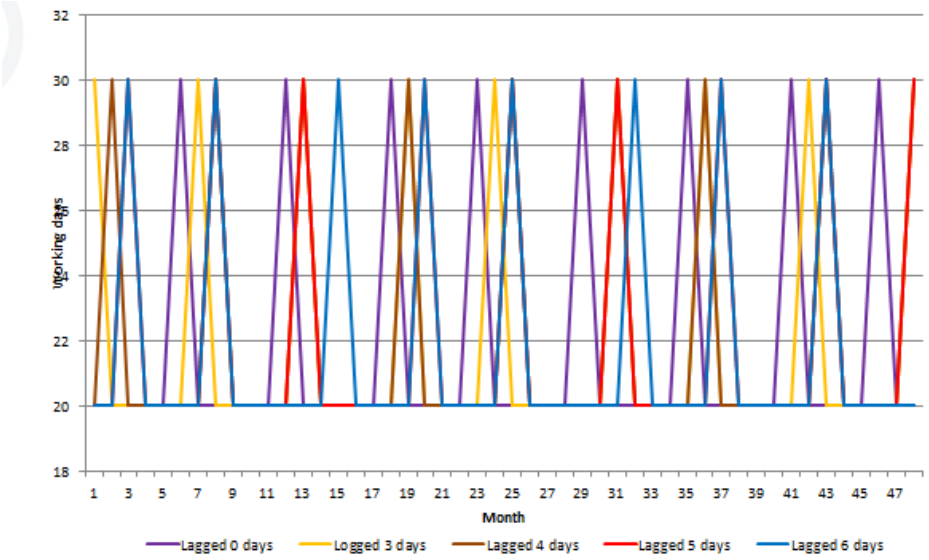


Figure 3 shows the effect of different pay lags on the pay peaks. Note that some colours seem to disappear. This occurs because they are being occulted by other lines. The beat frequencies for the peaks significantly changes based upon the time lag of the payments. For this hypothetical bi-weekly payroll, based upon the payment lag we can get 10 fundamental different beat patterns. In addition, bi-weekly firms can follow even or odd week payroll cycles so there are approximately 20 possible patterns (templates) for bi-weekly payrolls.

Table 1: Possible payroll cycles with size of peak and number of templates

Payroll cycle	Size of peak	Number of templates
Weekly	25%	10
Bi-weekly	50%	20
4-weekly	100%	40
Monthly	None	1

Table 1 summarizes the possible payroll cycles and the number of templates and the peak sizes. Thus all payroll cycles except monthly payrolls will have very large random peaks

in the reported earnings. In addition, there will be up to 71 different peak patterns (or templates) in the reported data.

One might think that aggregating the monthly reports into quarterly reports would significantly reduce or possibly eliminate the peaks but this is not the case. For quarterly reports, weekly payrolls will show 8% peaks, biweekly 17% peaks and 4-weekly 33% peaks. Even an 8% peak will swamp the underlying business cycle in quarterly data. It is probable that most firms pay bi-weekly which implies that quarterly reports will exhibit substantial random peaks.

To complicate matters, a variety of different random effects can influence the data. Unpaid holidays, overtime, inflation, production variations, promotions and seasonal effects can introduce random fluctuations into the data but generally one would suspect that these fluctuations would be less significant at the firm level than the calendar effects. Transaction heaping may occur as well. In this case a firm may not report for a month and then file a cumulative report for 2 months.

The PAYE data is a melange of approximately 71 peak patterns and each firm's payment peak will swamp any business cycle information concerning the month-to-month (or quarter-to-quarter) movements in the firm's payroll. Effectively, it makes the raw PAYE data only useable at the yearly reporting level. As currently constituted PAYE data has limited usefulness in a sub-annual survey.

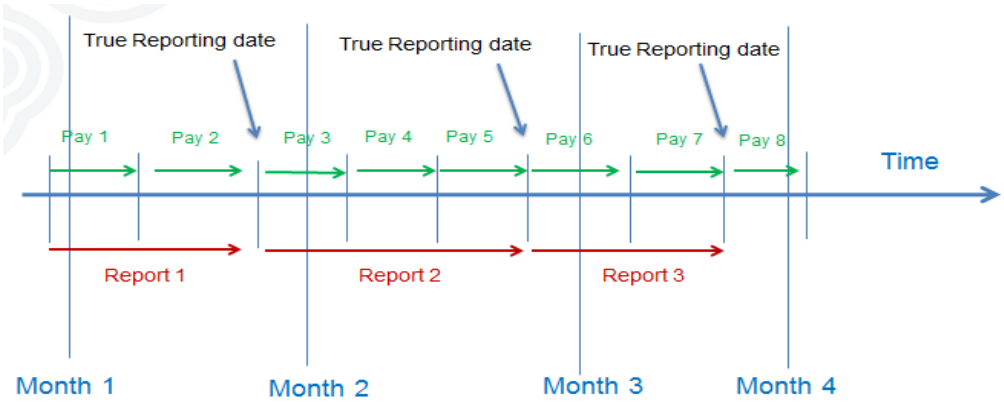
In the past at Stats NZ there have been several attempts at standardizing the PAYE data for use in sub-annual surveys. Most of these studies have used naïve and simple strategies for correcting for the calendar effects and for this reason they have failed. The manner in which the calendar effect exhibits itself in the PAYE data is quite complex and subtle and simple strategies will not work. In the next section, a strategy for calendarizing the data will be proposed.

3. The proposed standardization methodology

Standardizing time

The key to standardizing time and eliminating the calendar effect is discovering the true payment date of the last payroll in the month. Figure 4 shows the true date we are seeking.

Figure 4: Establishing the true reporting date of the last payroll

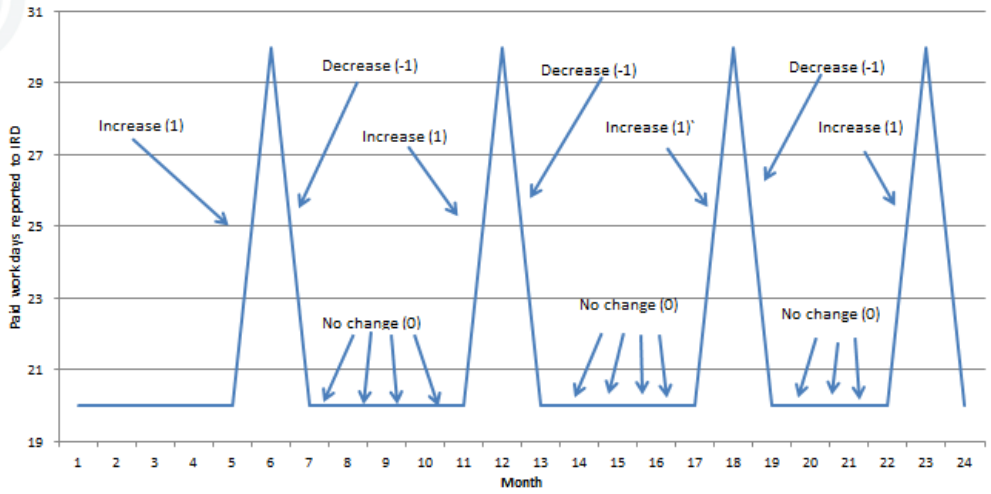


Given the true date, we can establish how many days were paid in each period and we can develop strategies to remove the calendar effects. One possible strategy for establishing the true date is to generate the 71 possible template patterns and test which one fits the times series reported by the firm. Unfortunately, other effects, especially inflationary and seasonal effects will tend to mask the calendar effect and make it difficult to identify the true underlying pattern.

We could try curve matching statistics like the Gini coefficient to identify the best fitted template but the random non-calendar effects seem to make this approach complicated and unreliable. So perhaps we could mitigate these non-calendar effects by simplifying the concept of the calendar effect. Let us assume the calendar effect exhibits itself as 3 change states; increase or decrease or no change (1,-1,0). Instead of trying to see if the peak height is 25% or 50%, we identify whether or not there is an increase or decrease. Thus we would transform our time series into a series of 1, -1, 0 values.

As illustrated in Figure 5, our 71 templates would also be transformed into a series of 1, -1, 0 values. We could then match the observed pattern to the templates and choose the best fit.

Figure 5: Transformation of templates in (1,-1,0) values



Defining the change-state time series for the templates is straight forward because they are integer values but for the observed time series the changes will be continuous values. A transformation is required to make the observed time series categorical. Let us define the continuous change variable $\hat{\Delta}_i^t$ as

$$\hat{\Delta}_i^t = \frac{p_i^t}{p_i^{t-1}}$$

Where p_i^t is the pay at time t and p_i^{t-1} is the pay from 1 month previous. Then we define our categorical indicator as:

$$\tilde{\Delta}_i^t = \left\{ \begin{array}{ccc} \text{missing} & \hat{\Delta}_i^t & \text{undefined or } =0 \\ \text{missing} & \hat{\Delta}_i^t & >2.05 \text{ or } <0.45 \\ 0 & \hat{\Delta}_i^t & >0.85 \text{ and } <1.20 \\ 1 & \hat{\Delta}_i^t & >1.20 \text{ and } <2.05 \\ -1 & \hat{\Delta}_i^t & >0.45 \text{ and } <0.85 \end{array} \right\}$$

The above ranges are the default ranges but in general they will be empirical values tuned to the data. In addition, if the seasonal effects are very strong, a methodology may be required to eliminate the seasonal effect. Once we have $\tilde{\Delta}_i^t$ defined we match it to templates to identify the appropriate pay frequency and payment lag. Then knowing pay frequency and lag, the reporting date t_R can be estimated. Short time series can match to multiple templates so a tie breaking formula is required.

As mentioned, the calendar effects in the PAYE data are significant, complex and subtle. Stats NZ is experimenting with a number of strategies for identifying the true underlying template and the above strategy may change.

Establishing the true date t_R is the first step in fully calendarizing the PAYE data. Next we need to properly redistribute the values in time. To do that we use a calendarization strategy recently outlined in (Quenneville, Picard, and Fortier 2013). Following the methodology in this paper, we propose a calendarization method based on interpolating the cumulated flows with splines.

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 pay which is a flow (p) into a cumulate or stock (P) by defining:

$$P^T = \sum_{t=t_1}^{t_1+T} p^t$$

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 P^T may not be observed. The missing P^T will be the interpolation points that we desire. Then we fit an interpolating spline through the knots $\left(P^T, \tau^T\right)$. Next, we read off on our curve the interpolated P^T values at all the defined τ^T including the points where P^T was unobserved. The PAYE flow is then derived:

$$p^{\tau^T} = \Delta P^{\tau^T} = P^{\tau^T} - P^{\tau^{T-1}}$$

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 pay value is added or subtracted to the time series. The spline

drags pay values backward to fill the time gap under a seasonal constraint. If one assumes that all PAYE 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

So in summary, the procedure standardizes reporting periods while not changing any of the original observed raw data cumulants. 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 PAYE 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 payrolls 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:

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

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).

4. Conclusions

These methodologies should improve the timeliness and quality of the PAYE 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.