Methods for temporal disaggregation of rolling quarterly VAT-based turnover

Paul Labonne [1] Martin Weale [2]

[1]King's College London & ESCoE

[2]King's College London, ESCoE & Centre for Macroeconomics

ESCoE Conference on Economic Measurement 2018

Motivation
Estimating a short-term economic indicator

- The objective is to develop a new output measure of GDP for the UK using VAT returns from HMRC.
- $GDP(O) = \sum^{Industries} GVA$ with GVA = Output - Intermediate Consumption.
- For the most part of the non-financial market economy, output is approximated using business turnover. The latter is currently measured with the Monthly Business Survey (MBS).
- The VAT returns provide an alternative measure of business turnover. All firms with a turnover greater than £85,000 have to submit VAT returns. When doing so they also indicate their turnover for the period in question.
- For the largest businesses (firms in size bands 4 and 5) the MBS is a census, and since it is more timely than the VAT data it is assumed that it will remain the measure of turnover for these businesses.
- However, for smaller businesses (size bands 1 to 3) the MBS is a sample. Since the VAT data are more complete than the MBS they could be used as an alternative measure of turnover for this population.

Data

The nature of the data

- Firms submitting VAT returns can adopt sixteen possible reporting patterns, which we refer as *staggers*. Our aim is to derive monthly estimates of turnover from these mixed-frequency data.

Table: Representation of the VAT staggers

Stag.	Freq.	J	F	M	A	M	J	J	A	S	O	N	D	J
0	M	x	x	x	x	x	x	x	x	x	x	x	x	x
1	Q			x			x			x			x	
2	Q	x			x			x			x			x
3	Q		x			x			x			x		
4	A	x												x
5	A		x											
6	A			x										
7	A				x									
8	A					x								
9	A						x							
10	A							x						
11	A								x					
12	A									x				
13	A										x			
14	A											x		
15	A												x	

x = VAT returns

Data
The nature of the data

- We use VAT-based turnover from 76 industries, which account for approximately 60% of total GVA in the economy. The data range from March 2011 to December 2016.
- The main problem is to disaggregate temporally the VAT-based rolling quarterly turnover for size bands 1 to 3. The MBS for size bands 4 and 5 and the VAT monthly stagger could be used as indicators to inform on the monthly path of the interpolands.

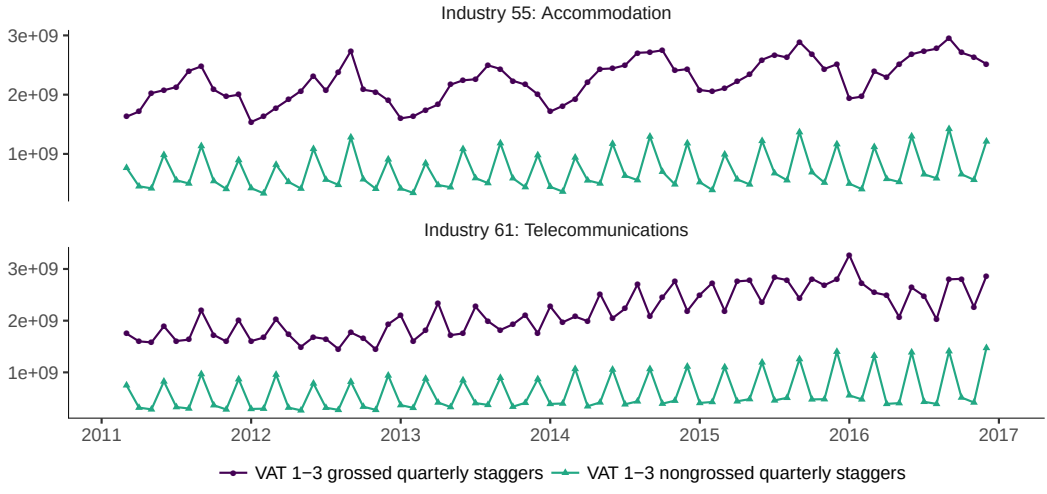
Characteristics of the VAT quarterly data:

- They are on a rolling monthly frequency;
- They are subject to a time-varying stagger bias (i.e. the populations reporting in each stagger are unequal);
- They exhibit significant noise;
- They are not seasonally adjusted.

Data description

An illustration of the quarterly staggers from two industries

Figure: VAT quarterly staggers, Mar 2011 - Dec 2016



We work from the raw data and assume that each stagger represents a distinct population.

- Step 1: We identify outliers, assumed to be reporting errors, and generate dummies for each one.
- Step 2: We estimate seasonally adjusted monthly interpolands for the three quarterly staggers separately at the industry level. The interpolation is carried out with a least-squares model with data in logarithms.
- Step 3: The three estimated monthly series are aggregated together with the figures of the monthly and annual staggers to produce output estimates for each industry. This requires an intermediate step where we seasonally adjust the monthly stagger figures (using X11-ARIMA) and calendarise the annual figures (using Chow-Lin without indicators).
- Step 4: The output estimates for each industry are indexed to March 2011 = 100.
- Step 5: The indexed series are aggregated to the economy level using GVA weights.

- Our approach is an extension of Mitchell *et al.* (2005) model in logs to non-seasonally adjusted data. We assume the following model

$$\log x_t = \beta_i \sum_{i=1}^{4} z_{it} + \log s_t,$$

$$\log s_t = \log s_{t-1} + u_t,$$

with x_t the non-seasonally adjusted interpolands, z_{it} the seasonal dummies, β_i the seasonal effects and s_t the seasonally adjusted interpolands.

- To minimise the sum of the squared errors under the rolling quarterly constraints we set the Lagrangian

$$\mathcal{L}(\log x, \beta, \beta^o, \lambda) = (\log x - Z\beta)' D' D(\log x - Z\beta) - 2\lambda(y - O\beta^o - Cx),$$

with

$$D = \begin{pmatrix} -1 & 1 & 0 & . & 0 & 0 \\ 0 & -1 & 1 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & -1 & 1 \end{pmatrix}; C = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & 0 & . & 1 & 1 & 1 \end{pmatrix},$$

and Z and O the matrices of seasonal and outlier dummies respectively. Since the programme is nonlinear we use a Gauss-Newton algorithm to find a solution.

We want to find the approximate solution $\hat{x}_1 = \hat{x}_0 + \Delta\hat{x}$.

(1) Approximate $\log\hat{x}_1 = \log\hat{x}_0 + \hat{X}_0^{-1}\Delta\hat{x}$;
(2) Solve

$$\begin{aligned}\text{Min} \quad \mathcal{L}(\Delta\hat{x}, \beta, \lambda) = (\log\hat{x}_0 + \hat{X}_0^{-1}\Delta\hat{x} - \mathrm{Z}\beta)'\mathrm{D}'\mathrm{D}(\log\hat{x}_0 + \hat{X}_0^{-1}\Delta\hat{x} - \mathrm{Z}\beta) \\ - 2\lambda(y - O\beta^o - \mathrm{C}(\hat{x}_0 + \Delta\hat{x})),\end{aligned}$$

which solution is given by

$$\begin{pmatrix} \Delta\hat{x} \\ \hat{\beta} \\ \hat{\beta^o} \\ \hat{\lambda} \end{pmatrix} = \begin{pmatrix} \hat{X}_0^{-1}D'D\hat{X}_0^{-1} & -\hat{X}_0^{-1}D'DZ & 0 & -C' \\ -Z'D'D\hat{X}_0^{-1} & Z'D'DZ & 0 & 0 \\ 0 & 0 & 0 & O' \\ C & 0 & O & 0 \end{pmatrix}^{-1} \begin{pmatrix} -\hat{X}_0^{-1}D'D\log\hat{x}_0 \\ Z'D'D\log\hat{x}_0 \\ 0 \\ y - C\hat{x}_0 \end{pmatrix};$$

(3) Set $\hat{x}_0$ to $\hat{x}_1$ and repeat (1) and (2) until $\Delta\hat{x}$ arbitrarily shrinks to zero.

To derive seasonally adjusted figures from these, we use:

$$\hat{s} = \exp(\log\hat{x}_1 - Z\hat{\beta}).$$

We assume that all staggers are equally representative, which implies that each monthly figure is identical in all staggers. This means that all monthly estimates, except for the two outer months, are subject to three quarterly constraints.

- Step 1: The VAT quarterly figures are grossed up to the VAT population. This is done by the ONS.
- Step 2: We identify outliers, assumed to be reporting errors, and replace them with missing values.
- Step 3: We estimate seasonally adjusted monthly estimates of turnover for each industry using a state-space model in logs.
- Step 4: The estimated series are indexed to March 2011 $= 100$.
- Step 5: The indexed series are aggregated to the economy level using annual GVA weights.

- The observation and state equations are

$$
\begin{aligned}
y_t &= s_t + \delta_t + \epsilon_t, && \epsilon_t \sim \mathsf{N}(0, \sigma_\epsilon^2), && t = 1, \\
y_t &= \log(e^{s_t} + e^{s_{t-1}}) + \delta_t + \epsilon_t, && && t = 2, \\
y_t &= \log(e^{s_t} + e^{s_{t-1}} + e^{s_{t-2}}) + \delta_t + \epsilon_t, && && t = 3, ..., T, \\
s_t &= s_{t-1} + \nu_t, && \eta_t \sim \mathsf{N}(0, \sigma_\nu^2), && t = 1, ..., T, \\
\delta_t &= -\sum_{j=1}^{11} \delta_{t-j} + \omega_t, && \omega_t \sim N(0, \sigma_\omega^2), && t = 1, ..., T.,
\end{aligned}
$$

  with y_t the rolling quarterly observations, s_t the unbiased seasonally adjusted interpolands and δ_t the seasonal and bias effects.

- The matrix form is given by (full representation in the appendix)

$$
\begin{aligned}
y_t &= \mathsf{Z}_t(\alpha_t) + \epsilon_t, && \epsilon_t \sim \mathsf{N}(0, \sigma_\epsilon^2), \\
\alpha_{t+1} &= T\alpha_t + \mathsf{R}\zeta_t, && \zeta_t \sim \mathsf{N}(0, Q).
\end{aligned}
$$

- Since $\mathsf{Z}_t(\cdot)$ is a nonlinear transformation of the state vector, we estimate the model using the extended Kalman filter.

- At each step of the Kalman filter recursion we linearise the observation equation using a first-order Taylor approximation at the predicted state:

$$y_t = \mathrm{Z}_t(a_t) + \tilde{Z}_t \cdot (\alpha_t - a_t) + \epsilon_t,$$

with

$$\tilde{Z}_t = \left.\frac{\partial \mathrm{Z}_t(\alpha_t)}{\partial \alpha_t}\right|_{\alpha_t = a_t}.$$

- The Kalman filter yields the state predictions $a_{t+1} = \mathrm{E}(\alpha_t | y_1, ..., y_t)$ and the signal prediction errors $v_t = y_t - \mathrm{Z}(a_t)$ as well as their variance estimates, respectively P_{t+1} and f_t. The recursion is initialised using an approximate diffuse method.

- The variance parameters are estimated by maximising the log-likelihood function derived from the prediction error decomposition (Harvey, 1989):

$$\log \mathrm{L}(y_1, ..., y_T; \sigma_\epsilon^2, Q) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}(\log|f_t| + \frac{v_t^2}{f_t}).$$

- The Kalman smoother makes use of the Kalman filter output when the parameters are set to their MLE estimates and gives the smoothed state vector $\hat{\alpha}_t = \mathrm{E}(\alpha_t | y_1, ..., y_T)$ and an estimation of the noise $\hat{\epsilon}_t$.

- The Kalman filter and smoother algorithms are taken from Durbin and Koopman (2012).

Results

All industries

Figure: State-space and least-squares monthly estimates compared with the MBS, seasonally adjusted figures, Mar 2011 - Oct 2016

- The MBS for bands 1 to 3 has been seasonally adjusted using X11-ARIMA. Similarly to the least-squares and state-space procedures, the series are then indexed to March = 100 and aggregated using GVA weights.

Results

Excluding troublesome industries

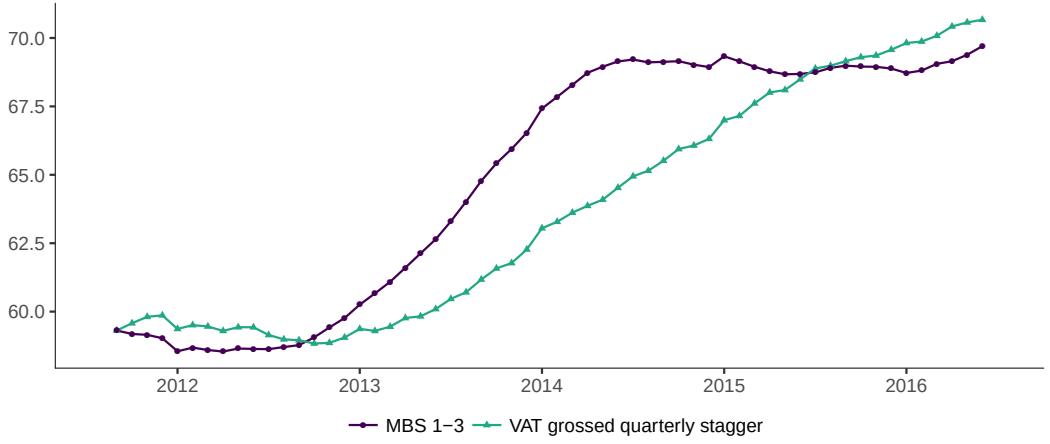
Figure: State-space and least-squares monthly estimates compared with the MBS, seasonally adjusted figures, Mar 2011 - Oct 2016

- We exclude industries for which the MBS population diverges from the VAT population and those for which turnover is not considered as a good proxy for output.

Results

Moving average analysis on all industries

Figure: Thirteen-month moving averages with half-weights on the two outer-most months, Sept 2011 - Jun 2016



- We apply simple thirteen-month moving averages, with the two outer-most months given half-weights, on the raw data. The series are then indexed to September =100 and aggregated using GVA weights. From these, we can verify that the diverging trends in the VAT and MBS series is an intrinsic feature of the data.

Next steps

On the least-squares and state-space models:

- Accounting for Easter effects.
- Adding the MBS 4-5 and VAT monthly stagger as regressors for the log-changes in the interpolands. Preliminary results suggest that the relationships between these variables vary significantly across industries.

On the state-space model:

- Multivariate extension with the MBS 4-5 or VAT monthly stagger as covariates to inform on the monthly volatility of the interpolands.
- Modeling the seasonal effects using a trigonometric form, this could allow to distinguish the seasonal effects from the stagger bias.
- Potentially modeling the quarterly staggers as different observation series to estimate distinct stagger biases and noise components.
- Linearisation via mode estimation (equivalent to a Sequential Linear Constraint optimisation method in the Gaussian model, see Proietti (2006)), which should reduce the approximation error to any chosen tolerance value.

References

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

Harvey, A. C. (1989). *Forecasting, structural time series models and the kalman filter*. Cambridge University Press.

Mitchell, J., Smith, R. J., Weale, M. R., Wright, S., & Salazar, E. L. (2005). An indicator of monthly gdp and an early estimate of quarterly gdp growth. *The Economic Journal*, *115*(501), F108-F129. doi: 10.1111/j.0013-0133.2005.00974.x

Proietti, T. (2006). On the estimation of nonlinearly aggregated mixed models. *Journal of Computational and Graphical Statistics*, *15*(1), 18-38. doi: 10.1198/106186006x100515

Appendix: Full representation of the state-space model

The observation equation is

$$y_t = Z_t(\alpha_t) + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2)$$

with

$$\begin{aligned} Z_t(\alpha_t) &= s_t + \delta_t, && t = 1, \\ Z_t(\alpha_t) &= \log(e^{s_t} + e^{s_{t-1}}) + \delta_t, && t = 2, \\ Z_t(\alpha_t) &= \log(e^{s_t} + e^{s_{t-1}} + e^{s_{t-2}}) + \delta_t, && t = 3, ..., T. \end{aligned}$$

The state equation is

$$\alpha_{t+1} = T\alpha_t + R\zeta_t, \qquad \zeta_t \sim N(0, Q),$$

with $\alpha_t = \begin{pmatrix} s_t \\ s_{t-1} \\ s_{t-2} \\ \delta_t \\ \delta_{t-1} \\ \delta_{t-2} \\ \delta_{t-3} \\ ... \\ \delta_{t-10} \end{pmatrix}$, $T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & ... & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & ... & -1 & -1 \\ 0 & 0 & 0 & 1 & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & ... & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & ... & 0 & 0 \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & ... & 1 & 0 \end{pmatrix}$, $R = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ . & . \\ 0 & 0 \end{pmatrix}$ and $Q = \begin{pmatrix} \sigma_\nu^2 & 0 \\ 0 & \sigma_\omega^2 \end{pmatrix}$.