Estimation of monthly business turnover using administrative data in the UK

Keywords: *Temporal disaggregation, State space models, Structural time series models, Administrative data, Monthly GDP.*

1 Introduction

The United Kingdom has traditionally relied on surveys to provide the data needed to compile the national accounts. Following the Bean Review [1], however, there has been an increased focus on the use of administrative data sources. A particular focus has been the use of Value Added Tax (VAT) returns to provide data on business turnover which are currently collected by the Monthly Business Survey (MBS). These data form the core of the output estimate of GDP, which has been published monthly since July 2018. Since the MBS for the largest businesses is fully enumerated and is more timely than the VAT data, it is assumed to remain the measure of turnover for these businesses. This paper explores the role that VAT returns can play in delivering monthly estimates of turnover for small and medium size businesses.

The most prominent issue to be addressed in the use of VAT data arises from the nature of the data. Approximately 10% of respondents make monthly returns, and there is also a small number of annual returns. Most respondents, however, make quarterly returns covering a period of three months, but within this, some of these cover the calendar quarter, some the three months ending in the first month of the calendar quarter and some the three months ending in the second month of the calendar quarter. In order to make use of these data it is necessary to generate monthly estimates (which we refer to as interpolands) from these rolling quarterly aggregates. In this aim we develop a state space approach for filtering, cleaning and disaggregating temporally the VAT figures, which are noisy and exhibit dynamic unobserved components. We notably derive multivariate and nonlinear methods to make use of an indicator series (the MBS for the largest businesses) and data in logarithms, respectively. After illustrating our temporal disaggregation method and estimation strategy using an example industry, we estimate monthly seasonally adjusted figures for the seventy-five industries for which the data are available. We thus produce an aggregate series representing approximately 60% of gross value added in the economy. We compare our estimates with those derived from the Monthly Business Survey and find that the VAT-based estimates show a different time profile and are smoother.

In addition to this empirical work, our paper contributes to the literature on temporal disaggregation in two respects. First, we provide a discussion of the effect that noise in aggregate figures has on the estimation of disaggregated model components. Secondly, we adopt a new temporal aggregation strategy different from the method of Harvey and Pierse [2]. The method we use is easier to model and can easily be generalised to non-rolling data.

2 Methods

Several characteristics of the VAT data such as their dynamic seasonality, rolling nature and the noise they exhibit render their temporal disaggregation problematic.

Applying least-squares techniques such as Chow and Lin [3], Fernandez [4] and Litterman [5] to the rolling quarterly figures produces highly erratic estimates. This stems from the difficulty in estimating observation errors and the impossibility of capturing stochastic model components. Consequently, we choose to develop state space methods, which do not suffer from either of these issues. We are able to clean, filter and disaggregate temporally the quarterly VAT figures in a single framework.

Our main model is a nonlinear multivariate structural time series model given by the following system of equations:

$$\begin{aligned}
y_{1,t} &= \log(e^{x_{1,t}} + e^{x_{1,t-1}} + e^{x_{1,t-2}}) + \gamma_{1,t} + \beta_{1,t} h^a_{1,t}, && \\
y_{2,t} &= x_{2,t} + \gamma_{2,t} + \beta_{2,t} h^a_{2,t}, && \\
\boldsymbol{x}_t &= \boldsymbol{\mu}_t + \boldsymbol{e}_t, && \\
\Phi(L)\boldsymbol{e}_{t+1} &= \boldsymbol{\kappa}_t, && \boldsymbol{\kappa}_t \sim \mathrm{N}(0, \Sigma_\kappa), \\
\boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t + \boldsymbol{\nu}_t + \boldsymbol{\xi}_t, && \boldsymbol{\xi}_t \sim \mathrm{N}(0, \Sigma_\xi), \\
\boldsymbol{\nu}_{t+1} &= \boldsymbol{\nu}_t + \boldsymbol{\zeta}_t, && \boldsymbol{\zeta}_t \sim \mathrm{N}(0, \Sigma_\zeta), \\
\boldsymbol{\gamma}_{t+1} &= -\sum_{j=1}^{11} \boldsymbol{\gamma}_{t-j} + \boldsymbol{\omega}_t, && \boldsymbol{\omega}_t \sim N(0, \Sigma_\omega), \\
\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t.
\end{aligned} \tag{1}$$

where y_{1t} are the aggregate VAT figures for small and medium size businesses and y_{2t} are the covariates (the MBS for the largest businesses). The state components $\boldsymbol{\mu}_t$, $\boldsymbol{\nu}_t$, $\boldsymbol{\gamma}_t$, $\boldsymbol{e}_t$ and $\boldsymbol{\beta}_t$ and the state disturbances $\boldsymbol{\xi}_t$, $\boldsymbol{\omega}_t$, $\boldsymbol{\zeta}_t$ and $\boldsymbol{\kappa}_t$ are vectors of length two with the VAT's component as first term and the MBS's component as second term.

The $\boldsymbol{x}_t$ are the logs of the seasonally adjusted monthly turnover estimates, which we assume follow a local linear trend model. The seasonally adjusted figures are thus composed of a time-varying trend but can also be subject to irregular variations in their level, captured by $\boldsymbol{\mu}_t$ and $\boldsymbol{e}_t$ respectively. We let the slope in the trend slowly vary over time by modelling it as a random walk. Turnover series at industries level can exhibit rapidly changing trends, which this model is thus suited to capture efficiently. We model the irregular variations as an autoregressive process that can be of order zero, one or two.

The seasonality is captured by the vector of seasonal dummies $\boldsymbol{\gamma}_t$. For the interpolands those are 3-month seasonal effects, but they also capture the noise in the data. This means that the seasonal effects are highly dynamic and the seasonal dummies model seems adequate in this case. The Easter effect is captured by $\boldsymbol{\beta}_t$ and $h^a_{i,t}$, $i = 1, 2$, are the Easter dummies.

The first two equations are the observation functions. They relate the state components just described to the figures that we observe. We notably use a nonlinear aggregation function to accommodate VAT data in logarithms; this mitigates the risk of heteroscedasticity while recognising that the aggregation constraints are linear.

Finally, we assume that the disturbances are uncorrelated across time and across state components, but there can be a contemporaneous correlation across the VAT and MBS series. Hence the latter can be used to identify the monthly changes in the interpolands.

Model (1) is easily expressed in state space form since the aggregation function for the VAT series does not rely on a cumulator variable. Models in state space forms can be estimated by maximum likelihood estimation using the Kalman filter outputs and the prediction error decomposition. After the estimation running the Kalman smoother yields the optimal estimates of the state vector (which includes the interpoland) and its variance matrix based on all observations.

Since the model is nonlinear, we first need to derive an approximating linear model. The best strategy when the data are subject to strong seasonal movements is to follow Proietti and Moauro [6]. They present a *sequential linear constrained* (SLC) method to estimate a model with nonlinear aggregation constraints. This algorithm allows us to reduce the approximation error to zero.

3 Results

3.1 An application: Land transport and transport service via pipelines

We illustrate our state space approach to disaggregate temporally the VAT data using the 'Land transport and transport services via pipelines' industry. The time series include seventy data points, starting in March 2011 up to December 2016. We begin by testing the univariate version of model (1) (which is derived by omitting the covariate series) on the MBS data for small and medium size businesses. We have artificially aggregated this series - we can thus compare the interpolands with the true underlying monthly figures. The results of the estimation suggest that the univariate state space model and estimation procedure is satisfactory for disaggregating temporally rolling quarterly turnover figures subject to seasonal movements. However, unlike the VAT data, the synthetic data are not noisy.

When applied to the the VAT data the results from the univariate model display two important differences from the estimation of the synthetic series. First, the estimated interpolands of the VAT figures are smoother. The second important distinction, which is linked to the first, is the volatility in the estimates of the seasonal effects. The noise in the data is captured in the seasonal disturbances, but it seems that the latter also capture some of the volatility in the interpolands.

Overall the estimation of the VAT rolling quarterly series suggests that the univariate model is not efficient in separating the noise in the data from the underlying monthly changes in the interpolands. To remedy this issue we proceed with the estimation of the multivariate model (1), where the MBS for the largest businesses is used as an indicator series for the month-on-month changes in the seasonally adjusted interpolands.

The estimated correlation coefficient between the irregular disturbances in the seasonally adjusted interpolands and covariates is positive, and, as a result, the interpolands' variance increases. We improve the estimation further by detecting outlying observations in the covariate estimates using the auxiliary residuals, and we account for these with dummy variables. We conclude that the covariates are helpful in identifying month-on-month changes in the seasonally adjusted interpolands and that the multivariate model provides satisfactory results.

3.2 Application to all industries

Having illustrated our method with one industry as a case study, we proceed by producing aggregate estimates from the whole set of seventy-five industries. We estimate seasonally adjusted monthly series of turnover for each industry using model (1), which we then aggregate together. We compare this series with the seasonally adjusted aggregate figures from the MBS covering small and medium size businesses. The VAT estimates exhibit a lower volatility than the MBS estimates. The state space estimation will typically generate estimates with a relatively lower volatility, but the noise in the VAT data exacerbates this feature. In addition, the MBS and

VAT aggregated estimates follow different trends. Notably, the VAT data point to a slower recovery after the euro area sovereign debt crisis.

The correlation between the series in growth rates is 0.51. This indicates some similarities between the two series, but, since they should represent the population of businesses, we would expect the correlation to be higher. To understand this divergence it would be helpful to analyse the data at a firm level.

4 Conclusions

The results suggest that our state space framework yields satisfactory monthly seasonally adjusted estimates from the VAT data. So far our analysis has been carried out using historical data from Jannuary 2011 to December 2016. Having access to vintages of the data and a longer time series covering the 2000's and 1990's would allow us to develop our research in two directions.

First, the VAT data are incomplete for the latest months, and, in order to produce timely estimates of turnover from these, we would need to forecast the missing data. VAT returns accrue gradually; it takes around six months for the data to be almost complete. Having access to the vintages of the VAT data would allow us to model and forecast the revisions. This can be achieved by extending our state space framework by modelling the revisions as latent states and augmenting the observation vector with estimates of the quarterly data at different maturities.

Secondly, analysing a longer time series would allow us to test the stability and robustness of our method. Notably, using data covering the Global Financial Crisis would be helpful in examining the ability of our model to capture major turning points.

References

[1] C Bean. Independent review of uk economic statistics. *https://www.gov.uk/government/publications/independent-review-of-uk-economicstatistics-final-report*, 2016.

[2] A. C. Harvey and R. G. Pierse. Estimating missing observations in economic time series. *Journal of the American Statistical Association*, 79(385):125–131, 1984. DOI: 10.1080/01621459.1984.10477074.

[3] Gregory C. Chow and An-Loh Lin. Best linear unbiased estimation of missing observations in an economic time series. *Review of Economics and Statistics*, 53: 372–5, 1971.

[4] Roque B. Fernandez. A methodological note on the estimation of time series. *The Review of Economics and Statistics*, 63(3):471, 1981. DOI: 10.2307/1924371.

[5] Robert B. Litterman. A random walk, markov model for the distribution of time series. *Journal of Business & Economic Statistics*, 1(2):169, 1983. DOI: 10.2307/1391858.

[6] Tommaso Proietti and Filippo Moauro. Dynamic factor analysis with nonlinear temporal aggregation constraints. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(2):281–300, 2006. DOI: 10.1111/j.1467-9876.2006.00536.x.