# Estimation of monthly business sector output from Value Added Tax data in the UK

NTTS 2019
Temporal disaggregation for higher frequencies data

Paul Labonne [1] Martin Weale [2]

[1]King's College London & ESCoE

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

# Motivation: A temporal disaggregation problem

- The independent review of UK economic statistics (Bean, 2016) advocates for increasing the use of administrative data in the National Accounts.
- At present ONS have access to Value Added Tax returns collected by HMRC. These returns indicate business turnover in addition to any tax due.
- Business turnover data form the core of the output estimate of GDP, which is the earliest estimate in the UK. Currently these data are collected through the Monthly Business Survey (MBS).
- Since the sample of the VAT data is many time larger than the MBS's, they could replace the MBS and improve the output estimate of GDP.
- While ONS publish monthly output figures, the VAT data take the form of overlapping quarterly aggregates. Therefore, it is necessary to disaggregate temporally the VAT data into monthly estimates (reffered to as interpolands) to make full use of them.

# A closer look at the Monthly Business Survey data

- The MBS collects monthly turnover data from a sample of UK businesses.
- The survey's sample is based on five strata : strata 1 to 3 refer to small and medium size businesses; strata 4 and 5 refer to the largest businesses.
- The MBS 4 & 5 is a census and is more timely than the VAT data. Consequently, it is assumed to remain the measure of turnover for the largest businesses. We want to explore the use of VAT data to replace the MBS 1-3.
- Since it will remain, the MBS 4 & 5 can be used as an indicator series while estimating the interpolands from the VAT data.
- We use the MBS 1-3 as a comparison to the VAT-based interpolated series, as both series should represent the same population.
- Additionally we use also the MBS 1-3 as a synthetic series to test the robustness of our method on 'clean' data. For this we aggregate the MBS 1-3 monthly figures into rolling quarterly aggregates, which we subsequently disaggregate temporally using our method. The resulting interpolated series can thus be compared to the original figures.

# The characteristics of the Value Added Tax data

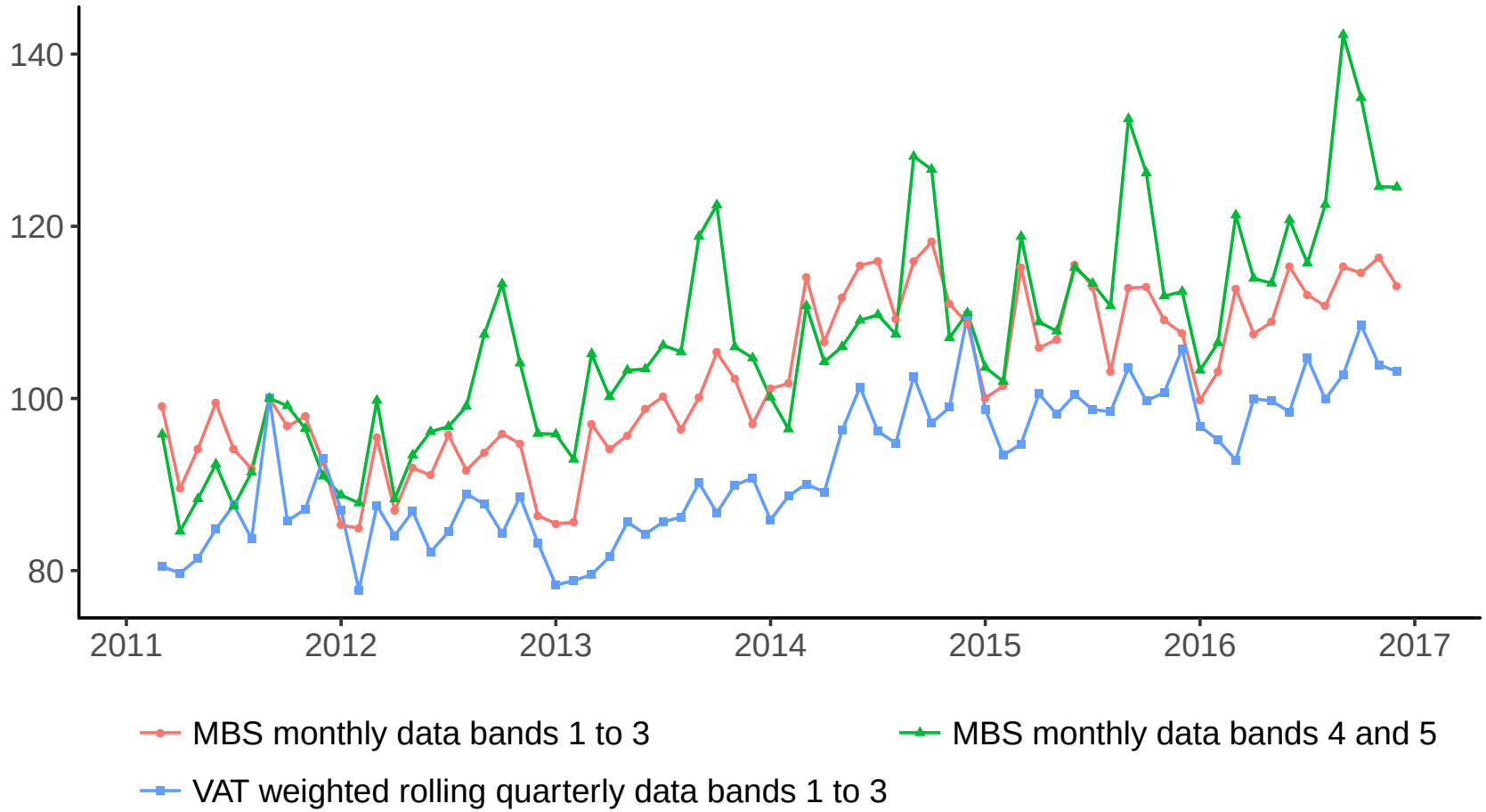
- Most firms submit quarterly returns, but they can start reporting in any month, generating three possible quarterly reporting patterns (called staggers). There are also a small number of firms reporting monthly and annual returns, but each quarterly stagger is weighted to represent all VAT-reporting firms.
- These weighted rolling quarterly VAT data show three notable characteristics:
  - They are overlapping, as represented in table 1.
  - They exhibit dynamic seasonality. Staggers are not equally populated which creates a stagger bias. If this bias is changing over time it will appear in the seasonal effects.
  - They are noisy. One reason for this could be apportionment errors when the VAT-based turnover figures are apportioned from VAT registration units to ONS statistical units.

Table: 1: Representation of the quarterly staggers; x = quarterly turnover total

| | Month | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | J | F | M | A | M | J | J | A | S | O | N | D | J |
| Stagger 1 | | | x | | | x | | | x | | | x | |
| Stagger 2 | x | | | x | | | x | | | x | | | x |
| Stagger 3 | | x | | | x | | | x | | | x | | |

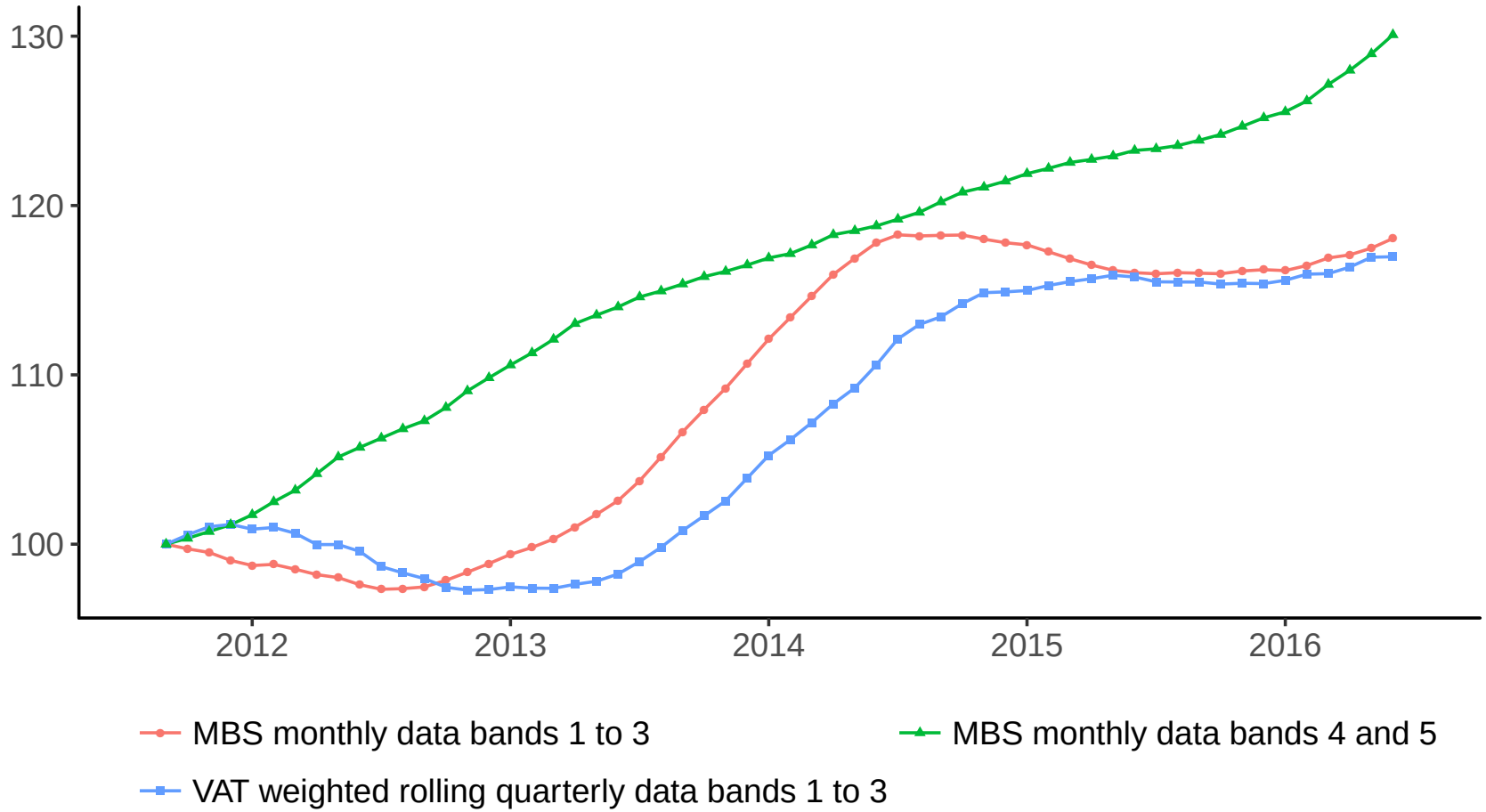
# A look at the VAT data: Aggregated data from 75 industries

Figure: Levels of the aggregated raw series, index September 2011 = 100



# A look at the VAT data: Aggregated data from 75 industries

Figure: Thirteen-month moving averages of the aggregated series, index September 2011 = 100



# State space models as temporal disaggregation method

- The characteristics of the VAT data imply that popular least-squares techniques such as Chow and Lin (1971), Fernandez (1981) and Litterman (1983) are not a viable option for interpolation of the VAT figures.
- Constraining the interpolands to noisy rolling quarterly totals produce erratic estimates.
- It is necessary to estimate the noise in the aggregate figures and allow for dynamic model components such as time-varying trends. This can be achieved using state space methods.
- Specifically we use a Seemingly Unrelated Time Series Equations (SUTSE) state space framework. Hence we can make use of an indicator series whilst making relatively weak assumptions on the form that the relationship between the interpolands and the covariates can take – we simply assume that both series (or some of their unobserved components) are affected by a common environment.

# A structural model

- We use a local linear trend model for the monthly seasonally adjusted estimates:

$$\begin{aligned} x_t &= \mu_t + e_t, \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim \mathrm{N}(0, \sigma_\xi^2), \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim \mathrm{N}(0, \sigma_\zeta^2), \end{aligned} \tag{1}$$

  where $x_t$ is the log monthly seasonally adjusted estimate, $\mu_t$ the time-varying trend and $\nu_t$ the time-varying slope. The irregular components are assumed to follow an auto-regressive process:

$$\Phi(B) e_{t+1} = \kappa_t, \qquad \kappa_t \sim \mathrm{N}(0, \sigma_\kappa^2). \tag{2}$$

- We capture the seasonality with a dummy seasonal model:

$$\gamma_{t+1} = -\sum_{j=1}^{11} \gamma_{t+1-j} + \omega_t, \qquad \omega_t \sim N(0, \sigma_\omega^2). \tag{3}$$

- We account for the Easter effect:

$$E_t = \beta(h_t - \sum_{t=1}^{s} h_t/s) = \beta h_t^a. \tag{4}$$

# An nonlinear overlapping temporal aggregation method

- We use an overlapping temporal aggregation function to link the disaggregated estimates to the aggregated figures:

$$y_t = \log(e^{x_t} + e^{x_{t-1}} + e^{x_{t-2}}) + \gamma_t + \beta_t h_t^a, \quad t = 1, ..., N. \tag{5}$$

- This kind of overlapping function is common in the now-casting literature (see for instance Aruoba et al. (2009)).
- This approach is different from the method of Harvey and Pierse (1984) (see also Harvey (1989)) which relies on the use of a cumulator variable. Using a cumulator variable is useful to limit the size of the state vector, but this is not an issue in our case.
- Equation (5) exhibits nonlinearities which arise from aggregating interpolands in logs. The approximation for nonlinear aggregation constraints of Mitchell et al. (2005) is not a viable option in our case.

# A multivariate nonlinear structural model

The SUTSE framework is:

$$
\begin{aligned}
y_{1,t} &= \log(e^{x_{1,t}} + e^{x_{1,t-1}} + e^{x_{1,t-2}}) + \gamma_{1,t} + \beta_{1,t} h^a_{1,t}, \\
y_{2,t} &= x_{2,t} + \gamma_{2,t} + \beta_{2,t} h^a_{2,t}, \\
\boldsymbol{x}_t &= \boldsymbol{\mu}_t + \boldsymbol{e}_t, \\
\Phi(L)\boldsymbol{e}_{t+1} &= \boldsymbol{\kappa}_t, \qquad & \boldsymbol{\kappa}_t \sim \mathrm{N}(0, \Sigma_\kappa), \\
\boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t + \boldsymbol{\nu}_t + \boldsymbol{\xi}_t, & \boldsymbol{\xi}_t \sim \mathrm{N}(0, \Sigma_\xi), \\
\boldsymbol{\nu}_{t+1} &= \boldsymbol{\nu}_t + \boldsymbol{\zeta}_t, & \boldsymbol{\zeta}_t \sim \mathrm{N}(0, \Sigma_\zeta), \\
\boldsymbol{\gamma}_{t+1} &= -\sum_{j=1}^{11} \boldsymbol{\gamma}_{t-j} + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t \sim N(0, \Sigma_\omega), \\
\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t.
\end{aligned}
\tag{6}
$$

We define the covariance matrix $\Sigma_h$, with $h = \kappa, \xi, \zeta$, as

$$
\Sigma_h = \begin{pmatrix} \sigma^2_{1,h} & \rho_h \sigma_{1,h} \sigma_{2,h} \\ \rho_h \sigma_{1,h} \sigma_{2,h} & \sigma^2_{2,h} \end{pmatrix}
$$

with $\sigma^2_{1,h}$ the variance of the interpoland's $h$ component and the $\sigma^2_{2,h}$ the variance of the covariate's $h$ component. $\rho_h \neq 0$ if $h = \kappa$ and zero otherwise.

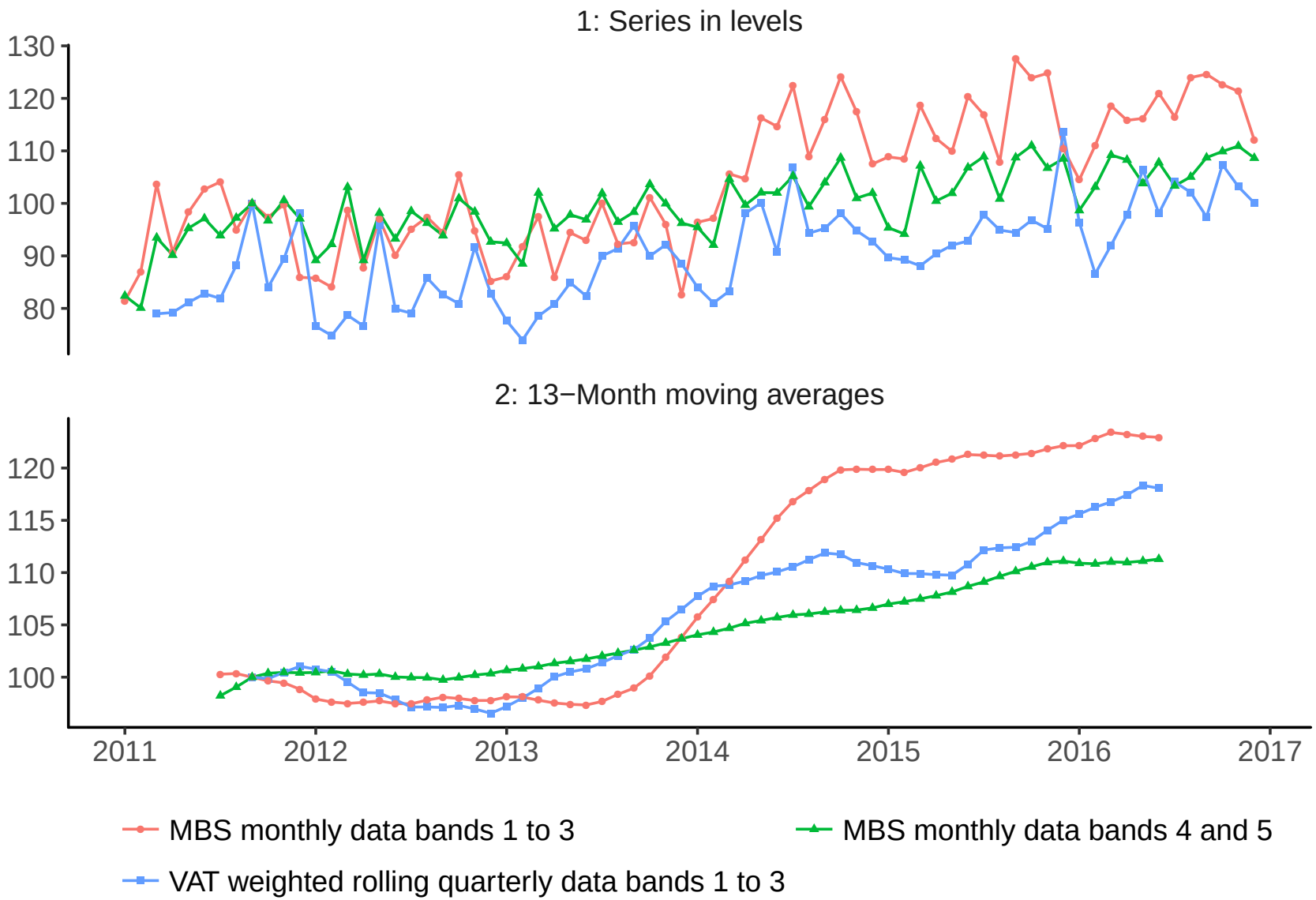
# State space form, linearisation and estimation

- The model may be written in state space form as

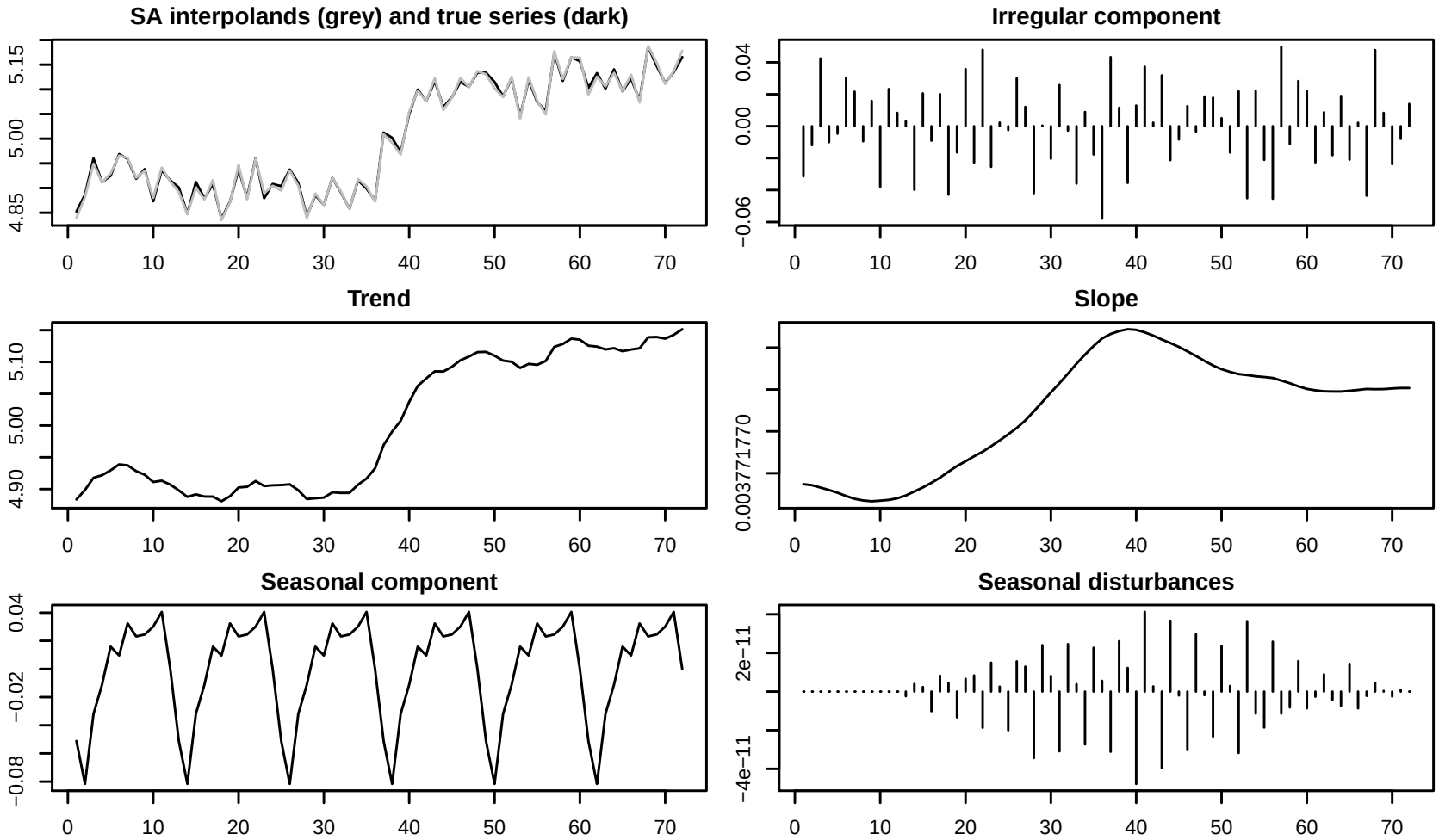
$$\begin{aligned} \boldsymbol{y}_t &= Z_t(\boldsymbol{\alpha}_t), \\ \boldsymbol{\alpha}_{t+1} &= T\boldsymbol{\alpha}_t + R\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathrm{N}(0, Q), \\ \boldsymbol{\alpha}_1 &\sim \mathrm{N}(\boldsymbol{a}_1, P_1). \end{aligned} \tag{7}$$

- Model (7) relies on a nonlinear observation function and cannot be estimated with standard methods because the Kalman filter does not apply.
- We follow Proietti and Moauro (2006) and use a *sequential linear constrained* (SLC) method to linearise the model.
- We estimate the approximate linear model once the SLC algorithm has converged. We maximise the log-likelihood derived from the prediction error decomposition of the Kalman filter's output. The interpolands are subsequently derived from the Kalman smoother.
- The Kalman filter is initialised with a diffuse initialisation (see Durbin and Koopman (2012)) and we use the computer codes from the R package of Helske (2017).

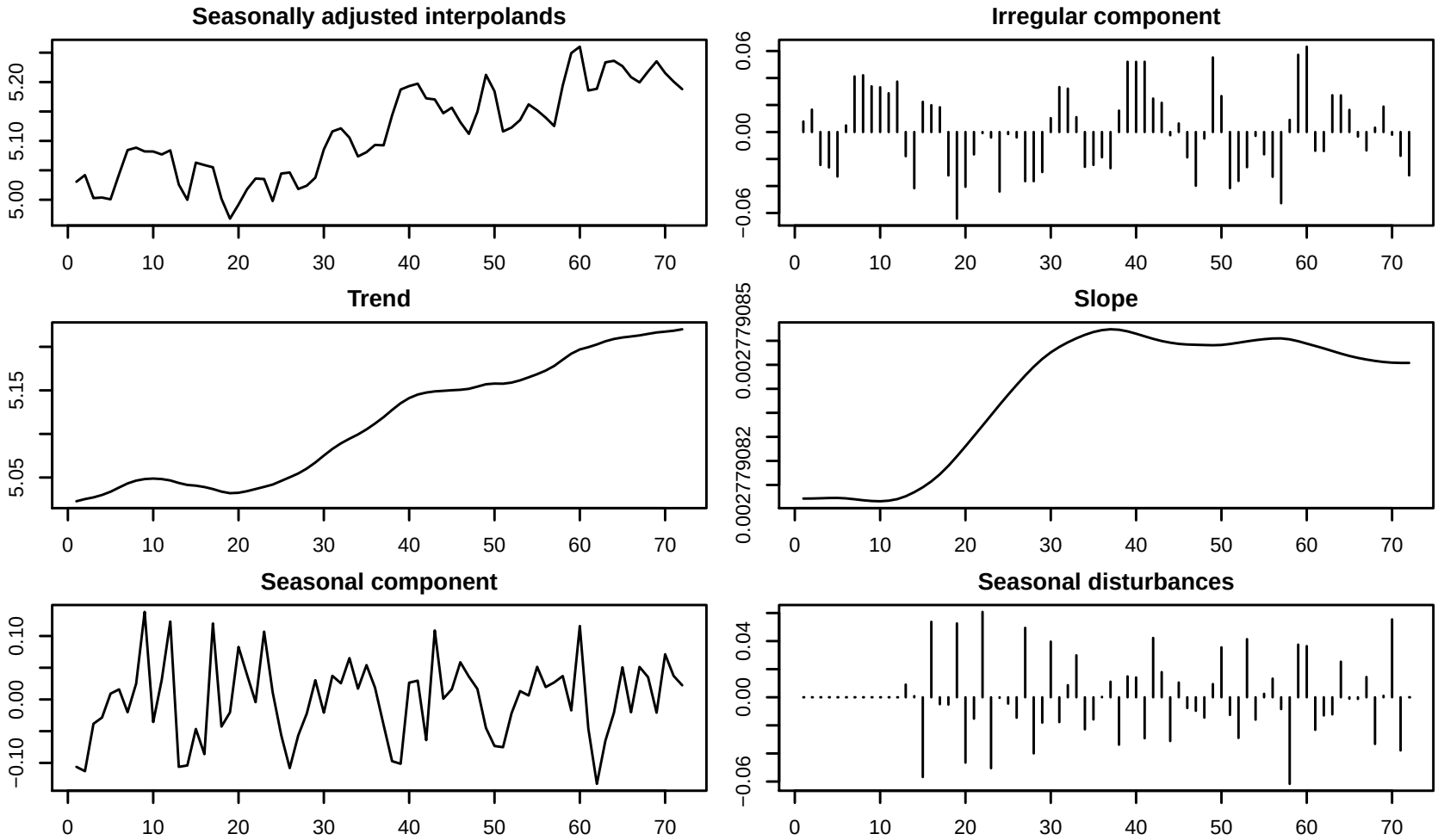
# Case study: Raw data from industry 435T495



# Case study: Results from the synthetic series with the univariate model

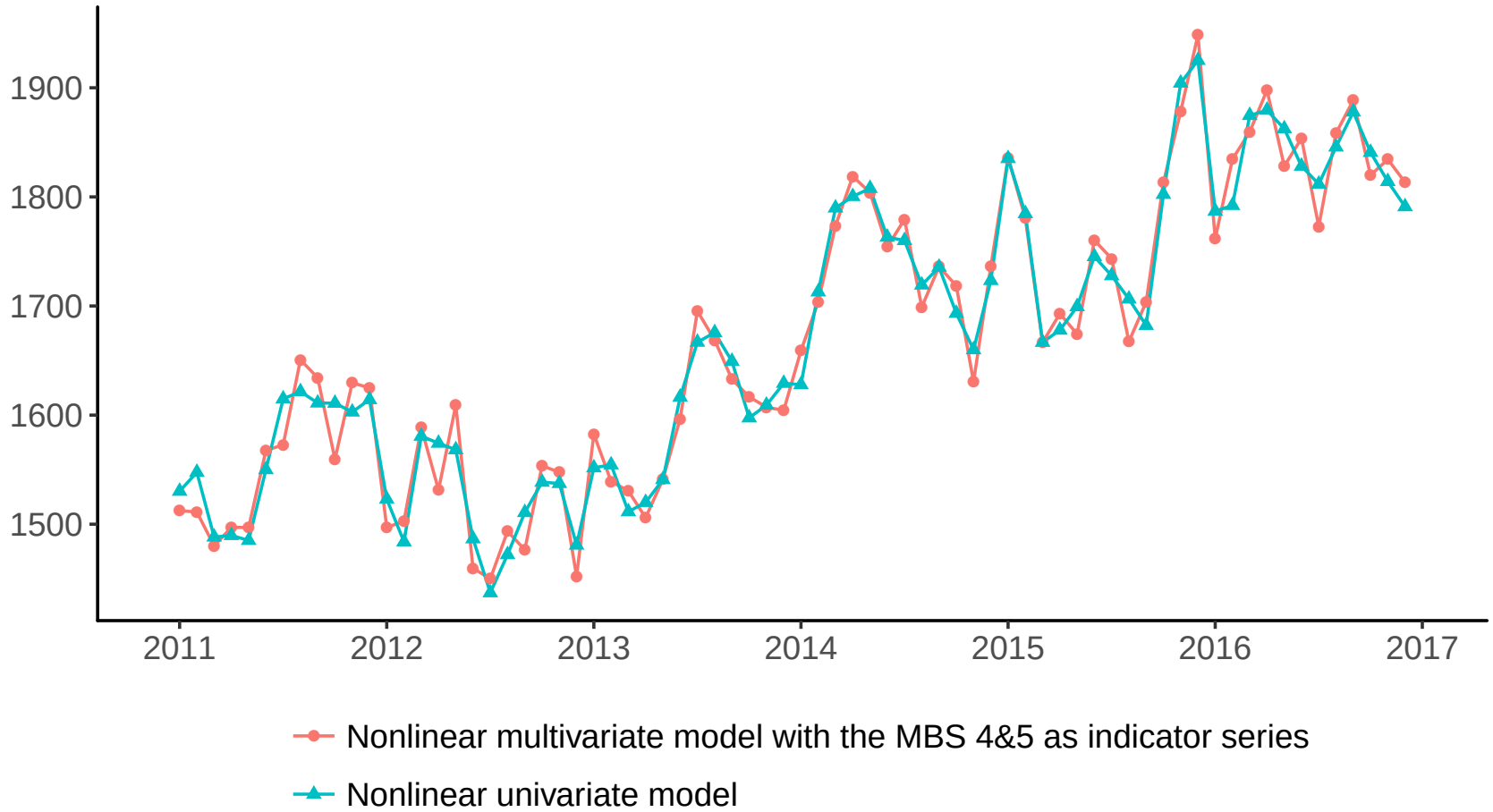


# Case study: Results from the VAT series with the univariate model



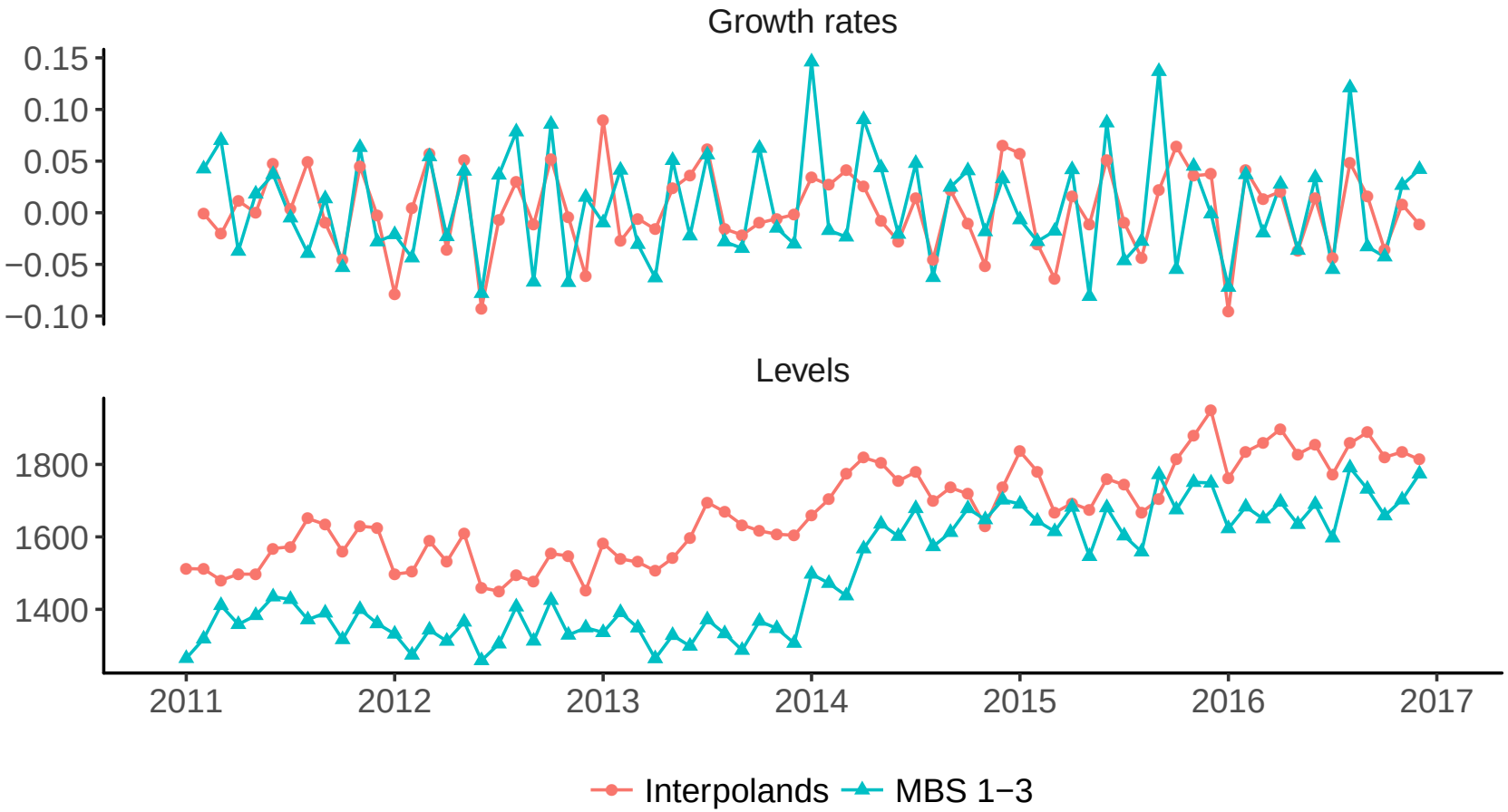
# Case study: Results from the VAT series with the multivariate model

Figure: Seasonally adjusted interpolands, in £million

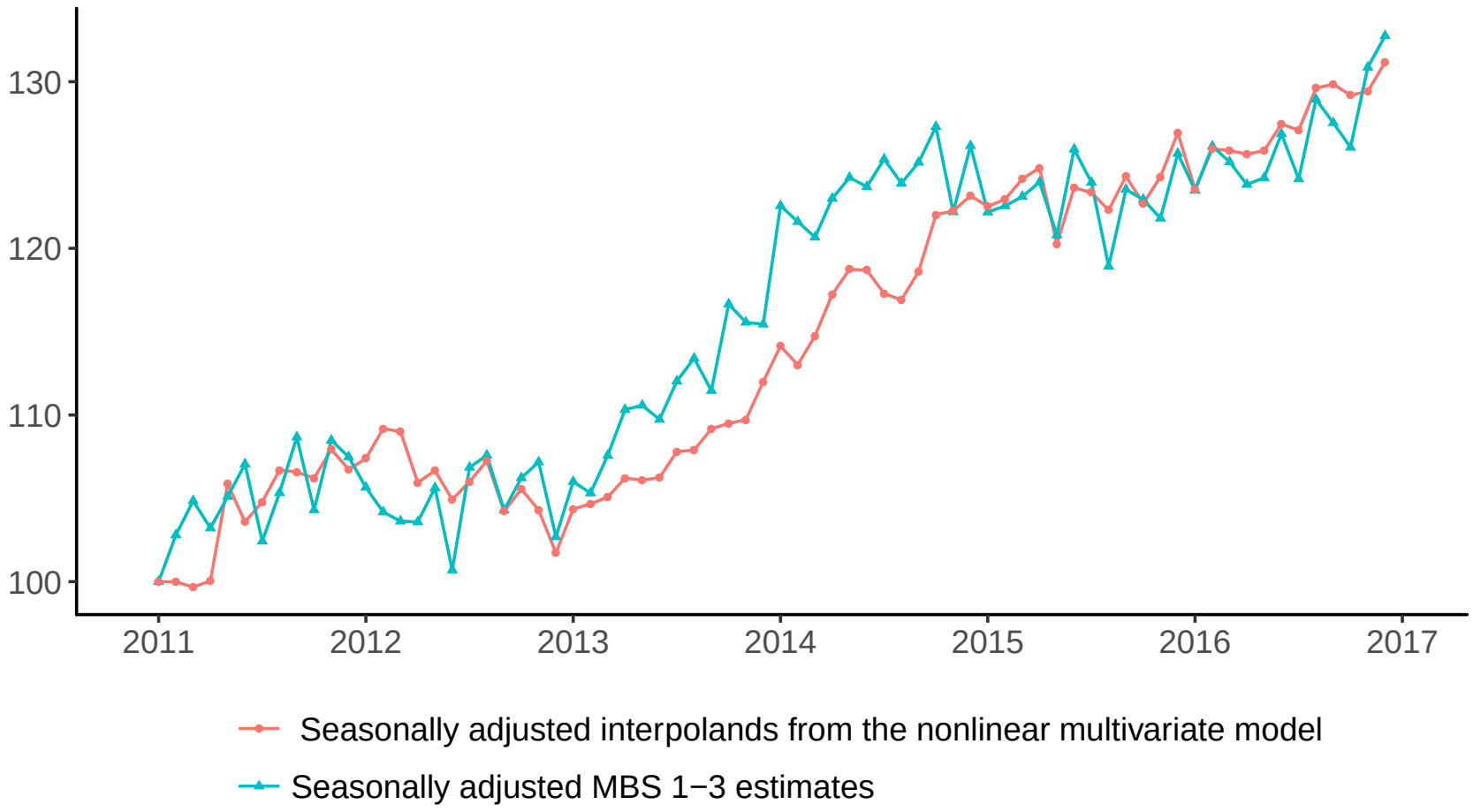


# Case study: Comparison with the MBS

Figure: Seasonally adjusted interpolands and MBS 1-3 figures, levels in £million



# Results for 75 industries accounting for a quarter of the UK economy



# Concluding remarks

- Multivariate structural state space models provide a flexible framework for interpolation when the data are noisy and exhibit dynamic unobserved components.
- We have shown that the VAT figures yield monthly estimates less volatile than the MBS figures and show a different time profile. Replacement of survey data by administrative data may, thus, lead to some rewritting of economic history.
- The seasonal disturbances, which capture the noise in the data, show non-gaussian features. Treating the outlying observations with standard methods does not seem to be efficient.
- Score-driven models (GAS/DCS) could be used to clean the raw figures from important outliers in a first stage.
- To produce a timely series it is necessary to forecast the late returns. This now-casting exercise can be carried out in the state space framework by augmenting the observation vector with the different vintages of the data.

# References


Aruoba, S. B., F. X. Diebold, and C. Scotti (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics 27*(4), 417–427.

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

Chow, G. C. and A.-L. Lin (1971). Best linear unbiased estimation of missing observations in an economic time series. *Review of Economics and Statistics 53*, 372–5.

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

Fernandez, R. B. (1981). A methodological note on the estimation of time series. *The Review of Economics and Statistics 63*(3), 471.

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

Harvey, A. C. and R. G. Pierse (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association 79*(385), 125–131.

Helske, J. (2017). Kfas: Exponential family state space models in r. *Journal of Statistical Software 78*(10).

Litterman, R. B. (1983). A random walk, markov model for the distribution of time series. *Journal of Business & Economic Statistics 1*(2), 169.

Mitchell, J., R. J. Smith, M. R. Weale, S. Wright, and E. L. Salazar (2005). An indicator of monthly gdp and an early estimate of quarterly gdp growth. *The Economic Journal 115*(501), F108–F129.

Proietti, T. and F. Moauro (2006). Dynamic factor analysis with non-linear temporal aggregation constraints. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 55*(2), 281–300.