Forecasting with a model of data revisions

Jana Eklund,[1]
George Kapetanios[1,2]
& Simon Price[1,3]

[1]Bank of England

[2]Queen Mary University of London

[3]City University (London)

Modelling and Forecasting Economic and Financial Time Series
with State Space Models
Sveriges Riksbank, 17 October 2008

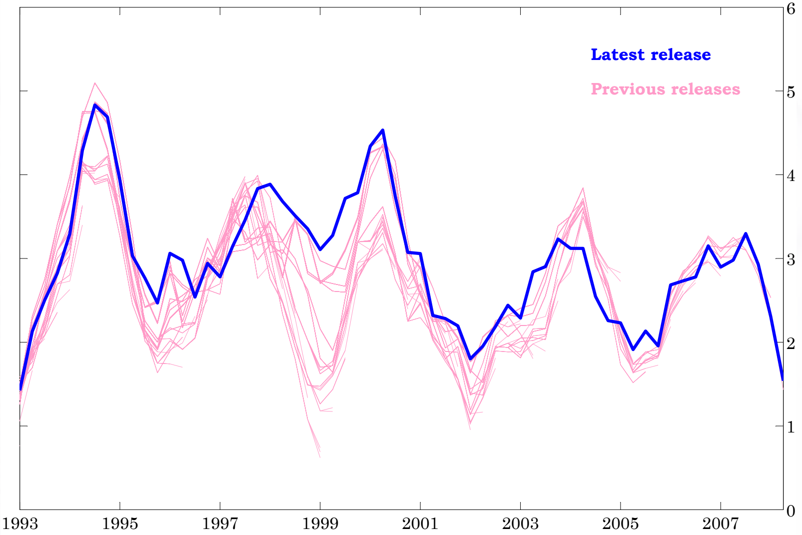
Data uncertainty

- Published data are estimates rather than perfect measures
- Measurement errors due to incomplete samples or proxies
- Statistical agencies revise their estimates - larger samples or better proxies

- State space modelling - usual approach to extracting signals
- Model the cumulative impact of revisions

United Kingdom real GDP

Real-time published estimates from 1993



Revisions in the United Kingdom

Across a range of macroeconomic variables revisions have tended to be

- Large relative to the variance in published data
- Occur several years after the first release ▸ Chart
- Initial estimates tend to be revised upwards
- Revisions to quarterly growth rates tend to be partially offsetting from one quarter to the next (negative serial correlation)

Extract From the Real-time Database

Quarterly Growth of Whole Economy Investment

		Release date				
		2003 Q1	2003 Q2	...	2006 Q3	2006 Q4
Reference date	2002 Q4	-0.15	0.16	...	3.51	3.51
	2003 Q1		-1.13	...	-3.18	-3.18
	⋮			⋱	⋮	⋮
	2006 Q2				1.31	1.21
	2006 Q3					1.32

Stylised Real-time Database - Maturity of Observations

Reference date		Release date				
		2003 Q1	2003 Q2	...	2006 Q3	2006 Q4
	2002 Q4	1	2	...	15	16
	2003 Q1		1	...	14	15
	⋮			⋱	⋮	⋮
	2006 Q2				1	2
	2006 Q3					1

Policy implications

- Policymakers need to know what the state of the economy in order to set policy appropriately: understanding revisions process may help this
- Policy often seen as a forward looking exercise
- Forecasts also summarise dynamic impact of shocks and policy
- Possible - likely? - that better nowcasts would help forecast process

Modelling assumptions

- Official data improve with maturity
- Latest release subsumes earlier vintages
  - Does not forecast specific ONS releases

The model of the published data

$$y_t^{t+n} = y_t + c^n + v_t^{t+n}$$

- y_t^{t+n} - an estimate of y_t published at time $t+n$, $n = 1, \ldots, T-t$
- y_t - the true data
- c^n - bias at maturity n
- v_t^{t+n} - measurement error associated with the published estimate

The model for the true data y_t

$$y_t = \mu + \sum_{i=1}^{q} \alpha_i y_{t-i} + \epsilon_t,$$

Assumptions:

- Stationarity of y_t - suitable for differenced or detrended data
- Linear functional form

Modelling choices

1. Bias

$$c^n = c^1(1+\lambda)^{n-1}$$

- c^1 - initial bias in published data
- $-1 \leq \lambda \leq 0$ - rate at which bias decays as data become more mature

2. Serial correlation with respect to time

$$v_t^{t+n} = \sum_{i=1}^{p} \beta_i v_{t-i}^{t+n} + \varepsilon_t^{t+n}$$

- is a finite AR process with maturity invariant parameters
- $\mathrm{E}\left(\varepsilon_t^{t+n}\right)^2 = \sigma^2_{\varepsilon^n}$

Modelling choices

1. Bias

$$c^n = c^1(1+\lambda)^{n-1}$$

- c^1 - initial bias in published data
- $-1 \leq \lambda \leq 0$ - rate at which bias decays as data become more mature

2. Serial correlation with respect to time

$$v_t^{t+n} = \sum_{i=1}^{p} \beta_i v_{t-i}^{t+n} + \varepsilon_t^{t+n}$$

- is a finite AR process with maturity invariant parameters
- $\mathrm{E}\left(\varepsilon_t^{t+n}\right)^2 = \sigma^2_{\varepsilon^n}$

Modelling choices

3. Heteroscedasticity with respect to maturity

$$\sigma^2_{\varepsilon^n} = \sigma^2_{\varepsilon^1} (1+\delta)^{n-1}$$

- $\sigma^2_{\varepsilon^1}$ - initial variance in published data
- $-1 \leq \delta \leq 0$ - rate at which variance decays as data become more mature

4. Correlation between errors

$$\text{cov}\left(\epsilon_t, \varepsilon^n_t\right) = \rho_{\epsilon\varepsilon}\sigma_{\epsilon}\sigma_{\varepsilon^n}$$

Modelling choices

3. Heteroscedasticity with respect to maturity

$$\sigma^2_{\varepsilon^n} = \sigma^2_{\varepsilon^1} (1+\delta)^{n-1}$$

- $\sigma^2_{\varepsilon^1}$ - initial variance in published data
- $-1 \leq \delta \leq 0$ - rate at which variance decays as data become more mature

4. Correlation between errors

$$\text{cov}\left(\epsilon_t, \varepsilon^n_t\right) = \rho_{\epsilon\varepsilon}\sigma_\epsilon\sigma_{\varepsilon^n}$$

The model

$$\begin{aligned} y_t^T &= c^n + y_t + v_t^T \\ y_t &= \mu + \sum_{i=1}^{q} a_i y_{t-i} + \epsilon_t \\ v_t^T &= \sum_{i=1}^{p} b_i v_{t-i}^T + \varepsilon_t^T \end{aligned}$$

Revisions

$$w_t^{j,n} = y_t^{t+n+j} - y_t^{t+n}$$

- Can obtain matrix **W** of revisions
  - rows contain revisions of a specific maturity
  - columns contain revisions within a single release

Two-step approach

Using only the latest release - does not mean that past releases are uninformative

1. Using revisions to estimate
    - bias (λ, c^1)
    - heteroscedasticity $(\delta, \sigma^2_{\varepsilon^1})$
    - serial correlation (b_i)
    - correlation with economic activity $(\rho_{\epsilon\varepsilon})$
2. Estimate remaining parameters (a_i) using Kalman filter

Reasons for two-step approach

- For $N \to \infty$
  - GMM estimates are $\sqrt{NT}$ consistent
  - ML estimates $\sqrt{T}$ consistent
- More data in the first step - lower variability in the estimates
- In practice: variability not taken into account in the second step

Growth for five National Accounts variables

- GDP
- Household consumption
- Whole economy investment
- Economic exports
- Economic imports

Evaluation setup

- Estimation over 1993Q2 - 2003Q1 releases
- Evaluation period 1998Q2 - 2003Q1
- Two subsamples 1998Q2 - 2000Q3 and 2000Q4 - 2003Q1
- Excluding revisions: 1998Q3 for all variables (ESA 1995)

Revisions

- Comparing forecast from the state space model with published data 16 periods later
- Comparing simple AR forecast based on contemporaneous data with release 16 periods later

Evaluation setup

- Estimation over 1993Q2 - 2003Q1 releases
- Evaluation period 1998Q2 - 2003Q1
- Two subsamples 1998Q2 - 2000Q3 and 2000Q4 - 2003Q1
- Excluding revisions: 1998Q3 for all variables (ESA 1995)
  Revisions
- Comparing forecast from the state space model with published data 16 periods later
- Comparing simple AR forecast based on contemporaneous data with release 16 periods later

Ratio of RMSFE 1998Q2 to 2003Q1

Minimum in bold,* indicates a significant DM statistic

(a) Fixed four lags for y_t

h	GDP	Consumption	Investment	Exports	Imports
1	0.9186	0.9738	**0.8918**	0.9083*	**0.8745***
2	0.9425	**0.9710**	0.9444	0.9257*	0.8813
3	**0.8899**	0.9848	1.0104	**0.8766**	0.9363
4	0.9400	0.9829	1.0009	0.9551*	0.9685

(b) Lag order optimally selected by HQIC

h	GDP	Consumption	Investment	Exports	Imports
1	0.9186	0.9738	**0.8878**	0.9682	1.0270
2	0.9425	**0.9710**	0.9351	0.9313	1.0534
3	**0.8899**	0.9848	1.0472	**0.9302**	1.0116
4	0.9400	0.9829	0.9547	0.9860	**0.9982**

Ratio of RMSFE 1998Q2 to 2000Q3

Minimum in bold,* indicates a significant DM statistic

(a) Fixed four lags for y_t

h	GDP	Consumption	Investment	Exports	Imports
1	**0.8183**	0.9765	**0.9433**	0.8551	0.9058
2	0.8898	**0.9621**	0.9581	0.8586*	**0.8271**
3	0.8210	0.9680	0.9868	**0.6865**	0.9461
4	0.8917	0.9712	0.9609	0.9086	1.0371

(b) Lag order optimally selected by HQIC

h	GDP	Consumption	Investment	Exports	Imports
1	**0.8183**	0.9765	**0.9208**	0.9575	0.9481
2	0.8898	**0.9621**	0.9458	0.8593	0.9707
3	0.8210	0.9680	1.0197	**0.7962**	**0.8898**
4	0.8917	0.9712	0.9147	0.9690	0.9185

Ratio of RMSFE 2000Q4 to 2003Q1

Minimum in bold,* indicates a significant DM statistic

(a) Fixed four lags for y_t

h	GDP	Consumption	Investment	Exports	Imports
1	1.0653	**0.9667**	**0.8698**	**0.9323**	**0.8663**
2	1.0641	1.0134	0.9343	0.9522	0.8979
3	**1.0360**	1.0870	1.0330	1.0039	0.9326*
4	1.0488	1.0420	1.0308	0.9744	0.9328

(b) Lag order optimally selected by HQIC

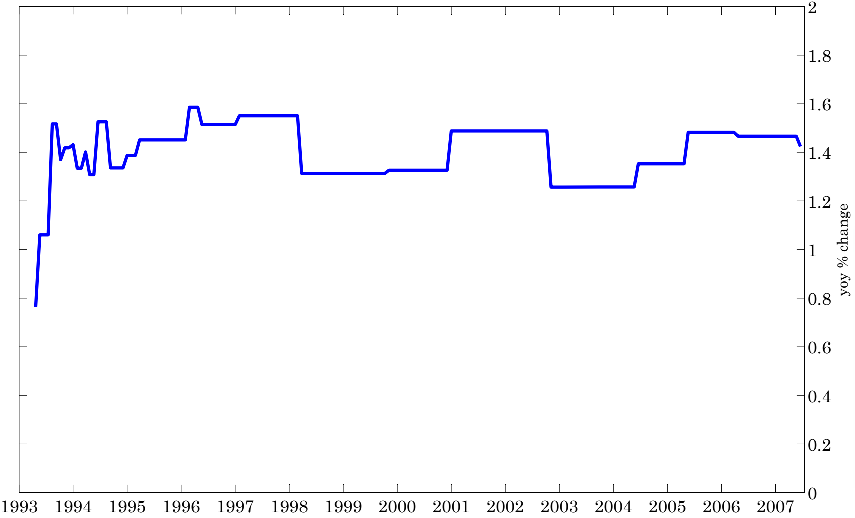
h	GDP	Consumption	Investment	Exports	Imports
1	1.0653	**0.9667**	**0.8731**	**0.9726**	**1.0529**
2	1.0641	1.0134	0.9272	0.9600	1.0797
3	**1.0360**	1.0870	1.0736	1.0017	1.0732
4	1.0488	1.0420	0.9848	0.9925	1.0585

Conclusions

- Using state-space approach to obtain better estimates of the ‘true’ value
- Practical and parsimonious way of producing backcast series
- Only have short periods for estimation and evaluation
- In the majority of cases forecast performance is improved

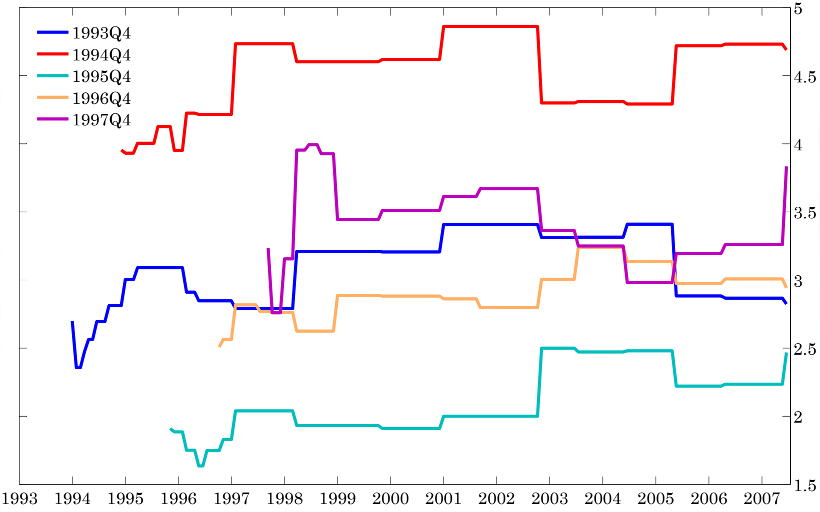
Successive estimates of GDP growth in 1993

◂ Back



United Kingdom real GDP

Back



Ratio of RMSFE for simulated data

$T = 500, a = 0.6, b = 0.6, r = 100, \delta = 0.5$

(a) Fixed 1 lag for y_t

h	Latest	Backcast	Data +16Q
1	0.8983	1.0124	1.2219
2	0.9606	1.0052	0.9708
3	0.9509	0.9933	0.9486
4	0.9730	0.9951	0.9886

(b) Lag order optimally selected by HQIC

h	Latest	Backcast	Data +16Q
1	0.9049	1.0172	1.2155
2	0.9669	1.0080	0.9747
3	0.9564	0.9935	0.9535
4	0.9669	0.9944	0.9934

Ratio of RMSFE for simulated data

$T = 120, a = 0.6, b = 0.6, r = 100, \delta = 0.5$

(a) Fixed 1 lag for y_t

h	Latest	Backcast	Data +16Q
1	0.9043	1.0037	1.2352
2	0.9461	1.0032	0.9643
3	0.9481	0.9900	0.9508
4	0.9662	0.9908	0.9860

(b) Lag order optimally selected by HQIC

h	Latest	Backcast	Data +16Q
1	0.8946	0.9996	1.2290
2	0.9409	1.0112	0.9546
3	0.9412	0.9890	0.9449
4	0.9664	0.9948	0.9926