Nowcasting UK GDP Using Automatic Model Selection

Oleg I. Kitov

Department of Economics and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford

December 3, 2013

Summary

- Use model (variable) selection within bridge equations framework
- Detect contemporaneous breaks in leading indicators
- Exploit non-stationarity and co-breaking of leading indicators
- Employ robust forecasting to correct for contemporaneous breaks

Overview

- Motivation: poor data quality and timeliness
- Solution: nowcast using model selection within bridge equations framework(Castle et al., 2013)
- Automatic model selection: Autometics algorithm (Doornik, 2009)
- Detect breaks: Impulse-indicator saturation (Hendry et al., 2008)
- Use information about breaks: robustifying nowcasts (Hendry, 2006)
- Empirical application: nowcasting UK output growth (this paper)

Motivation: GDP Growth Estimates Data Content

Variable	Estimate	Release lag	Real data
$\Delta y_{t_q}^{v_1}$	Preliminary, first estimate	3.5 weeks	44%
$\Delta y_{t_q}^{v_2}$	Output, Income and Expenditure	8 weeks	67%
$\Delta y_{t_q}^{v_3}$	UK National Accounts, first final estimate	12 weeks	80%
$\Delta y_{t_q}^{v_f}$	Final estimate 3 years later	3 years	
Δy_{t_q}	Latest available, most accurate estimate		

Table : GDP growth estimates, release dates and data completeness

Motivation: UK GDP Growth Revision Series

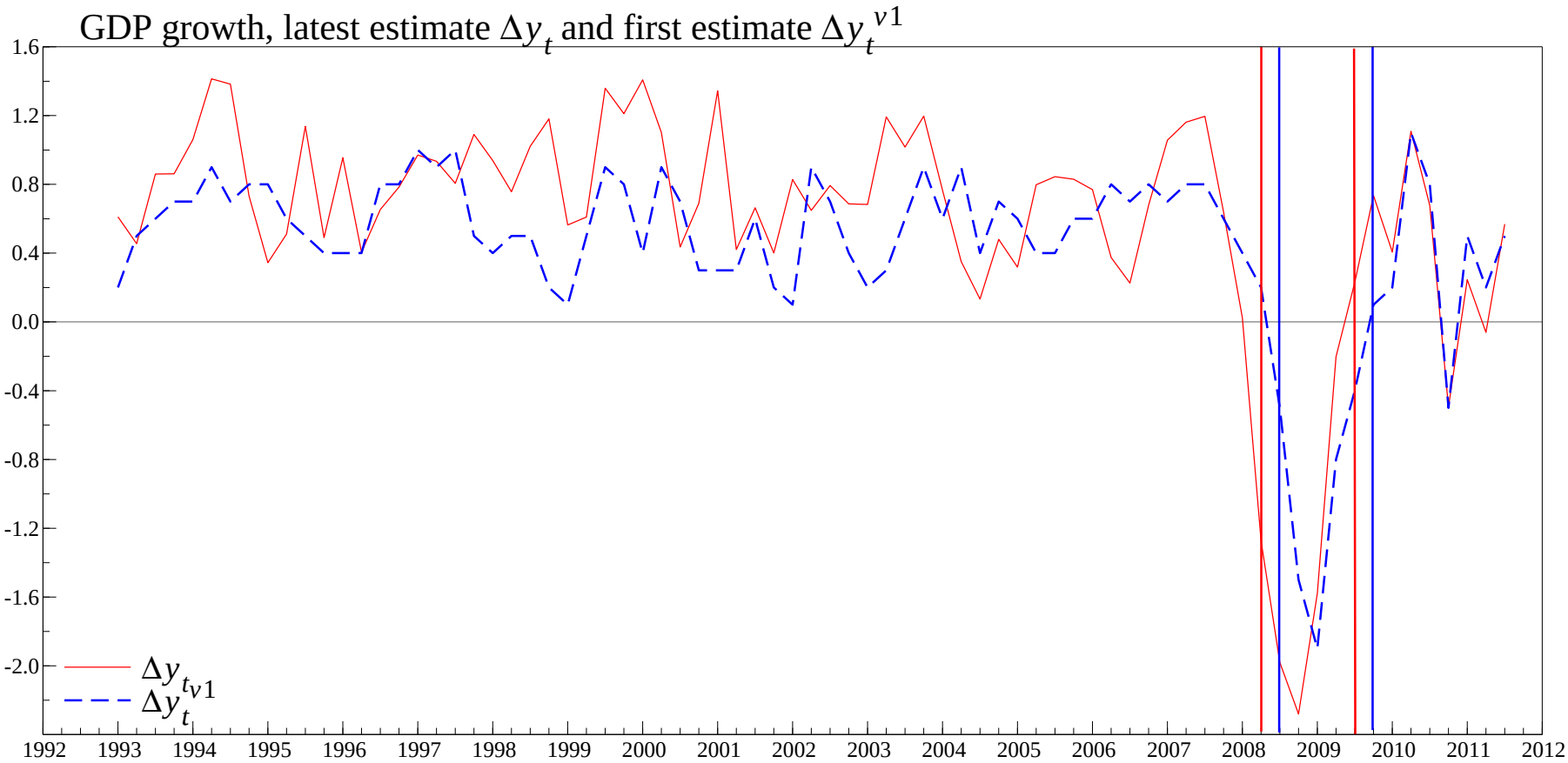

Figure : GDP growth latest estimate Δy_t and flash estimate $\Delta y_t^{v_1}$: ONS missed recession of 2008Q2 by 6 months

Motivation: Multiple Location Shifts

Most nowcasting and forecasting approaches treat data as a-priory stationary or force stationarity: may be inappropriate in face of location shifts.

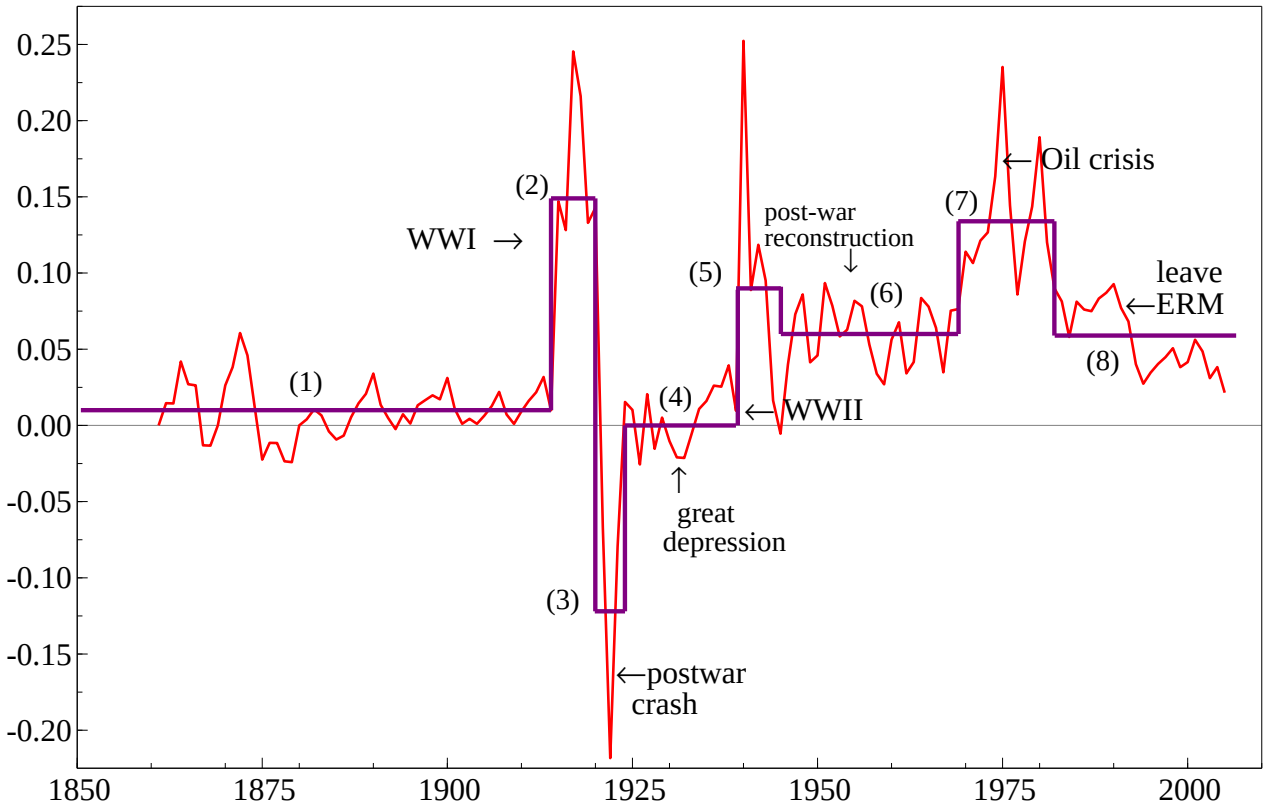

Figure : UK wage inflation

Motivation: Data Stucture and Model Selection

- Mixed frequency data
- Inconsistent release dates of indicators
- More variables than observations
- Unclear which leading indicators are relevant
- Functional form of the relationship is unknown
- Dynamic structure of the relationship is unknown
- Structural breaks

Methodolody: Bridge Equations Framework

Construct a direct "bridge" between aggregate measures and a set of explanatory variables. This approach involves specifying a model for GDP growth, Δy_{t_q}, at a quarterly frequency, s.t.:

$$\Delta \hat{y}_{t_q} = c + \sum_{i=1}^{p} \alpha_i \Delta y_{t_q - i} + \sum_{j=0}^{p} \sum_{i=1}^{k} \beta_{ij} z_{i,t_q - j} + u_{t_q}, \quad \text{for } t = 1, \ldots, T \quad (1)$$

where k leading indicators, z_{i,t_q}, are taken at a quarterly level and transformed to stationarity, p is the lag length and u_{t_q} are $i.i.d.$ residuals.

Methodology: Forecasting Indicators

- The quarterly indicators, z_{i,t_q}, in (1) are the transformed versions of the observed monthly indicators, z_{i,t_m}, a subset of which is unobserved in-sample due to the ragged-edge problem.
- The individual equations that forecast the missing values of the monthly indicators, e.g. AR(p):

$$\hat{z}_{i,t_m|t_m-1} = \sum_{i=1}^{p} \beta_i z_{i,t_m-i} + e_{i,t_m} \tag{2}$$

where $\hat{z}_{i,t_m|t_m-1}$ is the conditional forecast for the monthly series.

- Can use general-to-specific modeling and variable selection (as alternative to State-Space, Factor Models or MIDAS) to correct for ragged edges and produce parsimonious nowcasting models.

General-to-Specific (Gets)

- Gets is based on model discovery and theory of reduction (Hendry, 1987, 1995): find an appropriate local representation of the DGP by reducing a general model (GUM) to a specific one.
- Starting from the GUM that nests all the candidate variables, their lags, functional form transformations and possible breaks, look for a parsimonious model that nests the Local Data Generating Process (LDGP):
  - Step 1: Define a set of N candidate variables - the General Unrestricted Model (GUM)
  - Step 2: Reduce the complexity of GUM by removing insignificant variables, while checking that at each reduction the validity of the model is preserved

Model Selection: Autometrics

- GUM: The general unrestricted model (GUM) is the starting point of the search. The GUM is specified based on broad theoretical considerations to nest the LGDP
- Pre-Search: prior to specific selection, a pre-search lag reduction is implemented to remove insignificant lags
- Search Paths: Autometrics uses a tree search to explore paths. Starting from the GUM, Autometrics removes the least significant variable as determined by the lowest absolute t-ratio. Each removal constitutes one branch, which is back-tested against the initial GUM using an F-test. Branches are followed until no further variable can be removed - arrive at a terminal model
- Diagnostic Testing: each terminal model is subjected to a range of diagnostic tests

Break Detection: Impulse-indicator Saturation (IIS)

- Consider a regression for y_t, where $t = 1, \ldots, T$ is saturated by T impulse indicators: for each time period an indicator $1_{t_j=t}$ is defined taking value of unity when $t_j = t$ and zero otherwise
- Stage 1: Add first $T/2$ impulses as explanatory variables as well as a constant term, select over those:

$$y_t = \mu + \sum_{j=1}^{T/2} \delta_j 1_{t_j=t} + \epsilon_t \tag{3}$$

- Stage 2: Add the other half of the impulses so that $1_{t_j=t}$ for $j = T/2 + 1, \ldots, T$ are added and the model is re-selected.
- Stage 3: The selected impulses from both stages are combined and re-selected.

IIS Properties

- Adding exogenous and lagged dependent variables does not affect the analysis.
- Under the null of no outliers or shifts, the above specification with T indicators results in no efficiency loss for $\alpha \leq 1/T$.
- Mean and variance estimators are unbiased when *Autometrics* is used to select a parsimonious model, where only relevant indicators are retained if $\left| t_{1,\hat{\delta}_t} \right| \geq c_\alpha$ for a given significance critical value c_α.

Advantages of Autometrics

- Autometrics can effectively handle more variables than observations and work with collinear data
- Can control size (selection of irrelevant variable) with nominal significance level
- Deals with structural breaks using IIS
- Joint variable selection and break detection produce positive results: Monte Carlo simulations show that the method works well when there are multiple breaks and doesn't find breaks that do not exist
- Automatic model selection applied to nowcasting when unanticipated structural breaks occur particularly is relevant during major economic changes, e.g. recession
- Knowledge about breaks is successfully applied to adjust and robustify nowcasts

Forecasting During Breaks

- When unanticipated location shifts occur, conditional expectations of a future variable need not be unbiased even based on the in-sample DGP and given all available information
- A failure to locate a break in either the target variable or the leading indicators may result in misspecification of the post-break model and consequently biased nowcasts / forecasts
- Robust forecasting devices may forecast better than any structural model in such shifting processes, as measured by root mean-square forecast errors (RMSFEs)

Forecasting Bias During Breaks

Consider an AR(1) series y_t with a non-zero intercept and a random zero-mean measurement error $\epsilon_t \sim \text{IN}\left[0, \sigma_\epsilon^2\right]$:

$$y_t = \gamma_0 + \gamma_1 y_{t-1} + \epsilon_t \tag{4}$$

In-sample conditional forecast

$$\hat{y}_{t|t-1} = \hat{\gamma}_0 + \hat{\gamma}_1 y_{t-1}$$

will be unbiased if the series is measured correctly and no break occurs for all t.

Location Shifts

Now, assume a location shift occurs at the forecast origin T - the intercept changes from γ_0 to γ_0^*:

$$y_t = \gamma_0^* + \gamma_1 y_{t-1} \tag{5}$$

for $t = T+1, \ldots$. Since the break is not known at the forecast origin, the conditional forecast takes the following form:

$$\hat{y}_{T+1|T} = \hat{\gamma}_0 + \hat{\gamma}_1 y_T \tag{6}$$

and consequently results in a forecasting error of:

$$e_{T+1|T} = y_{T+1} - \hat{y}_{T+1|T} = \gamma_0^* + \gamma_1 y_T - \hat{\gamma}_0 - \hat{\gamma}_1 y_T \tag{7}$$

As $\hat{\gamma}_1$ is unbiased estimator of γ_1, the conditional expectation of the error at T is:

$$\mathrm{E}\left[e_{T+1|T}\right] = \gamma_0^* - \hat{\gamma}_0 \tag{8}$$

Robust Forecast

- In-sample observations do not produce an unbiased estimate of $\hat{\gamma}_0^*$, as the only observation containing information about the break, y_t, becomes available after the forecast origin - at $T+1$.
- Unless the break is predicted *ex-ante*, $e_{T+1|T}$ will be different from zero.
- However, the first observation of the shifted process can be used to forecast the next realization of y_t.: $\hat{y}_{T+2|T+1} = y_{T+1}$, with the expected conditional error of:

$$\begin{aligned} \mathsf{E}\left[e_{T+2|T+1}\right] &= \mathsf{E}\left[\gamma_0^* + \gamma_1 y_{T+1} - \gamma_0^* - \gamma_1 y_T\right] \\ &= \mathsf{E}\left[\gamma_1 \left(y_{T+1} - y_T\right)\right] \end{aligned} \tag{9}$$

- This the smallest error in the mean square sense, since the shifted term cancels out. Consequently, the single observation produces the best possible forecast one period after the break has occurred.

Exploiting Ragged Edges

- Two monthly leading indicators, $z_{1,t}$ and $z_{2,t}$, highly correlated and/or break simultaneously (co-breaking).
- $z_{1,t}$ is released at t, $z_{2,t}$ is available with a one period delay, so only $z_{1,t-1}$ is known at t.
- Allow for a break, of any type, to be realized for $z_{1,t}$. Furthermore, assume that the break is detected by IIS after model selection using *Autometrics*.
- Since $z_{1,t}$ and $z_{2,t}$ are highly correlated or co-breaking, it is likely that $z_{2,t}$ will also experience a break at t.
- This will not explicitly observed until $t+1$, when $z_{2,t}$ becomes available.

Robustifying Forecasts Contemporaneously

- If the break is common, the conditional model selected by *Autometrics* for $z_{2,t}$ will produce a biased forecast $\hat{z}_{2,t|t-1}$.
- Systematic bias can be avoided by using a robust device for $z_{2,t}$ that takes the previous observed value of the variable and is denoted by $\tilde{z}_{2,T+1|T}$:

$$\tilde{z}_{2,T+1|T} = z_{2,T} \tag{10}$$

- This will cancel out the break, if it does occur, and will result in a small error if it doesn't.

Dealing with Mixed Frequency

For every $\mathbf{z}_{t_m} = (z_{t_m}, \ldots, z_0)'$, denote the latest quarter t_q for which an observation $\mathbf{z}_{t_q}$ is available by τ, such that $\tau = t_q \leq t_m$, then:

$$\begin{aligned} \text{first month: } \mathbf{z}^1_{t_q} &= z_{\tau-2}, z_{\tau-5}, \ldots \\ \text{second month: } \mathbf{z}^2_{t_q} &= z_{\tau-1}, z_{\tau-4}, \ldots \\ \text{third month: } \mathbf{z}^3_{t_q} &= z_{\tau}, z_{\tau-3}, \ldots \end{aligned} \tag{11}$$

where $\mathbf{z}^1_{t_q}$, $\mathbf{z}^2_{t_q}$ and $\mathbf{z}^3_{t_q}$ are vectors containing observations for the first, second and third months of the quarter respectively.

Dealing with Ragged Edges

- For a monthly variable z_{k,t_m}, available with a lag l_k, such that the latest observation is z_{k,t_m-l_k}, the following GUM fully saturated by impulse indicators is formed:

$$z_{k,t_m} = \sum_{i=1}^{b_k} \sum_{j=1}^{12} \beta_{i,j} z_{i,t_m-l_i-j} + \sum_{t=1}^{t_m} \zeta_{k,t} 1_{k,t} \tag{12}$$

  where b_k is the total number of variables in the GUM and 12 is the largest lag entering the GUM.
- One, two or three periods ahead forecasts are computed depending on the corresponding release lag $l_i \in \{1, 2, 3\}$ to correct for the ragged edges.

Data: Leading Indicators

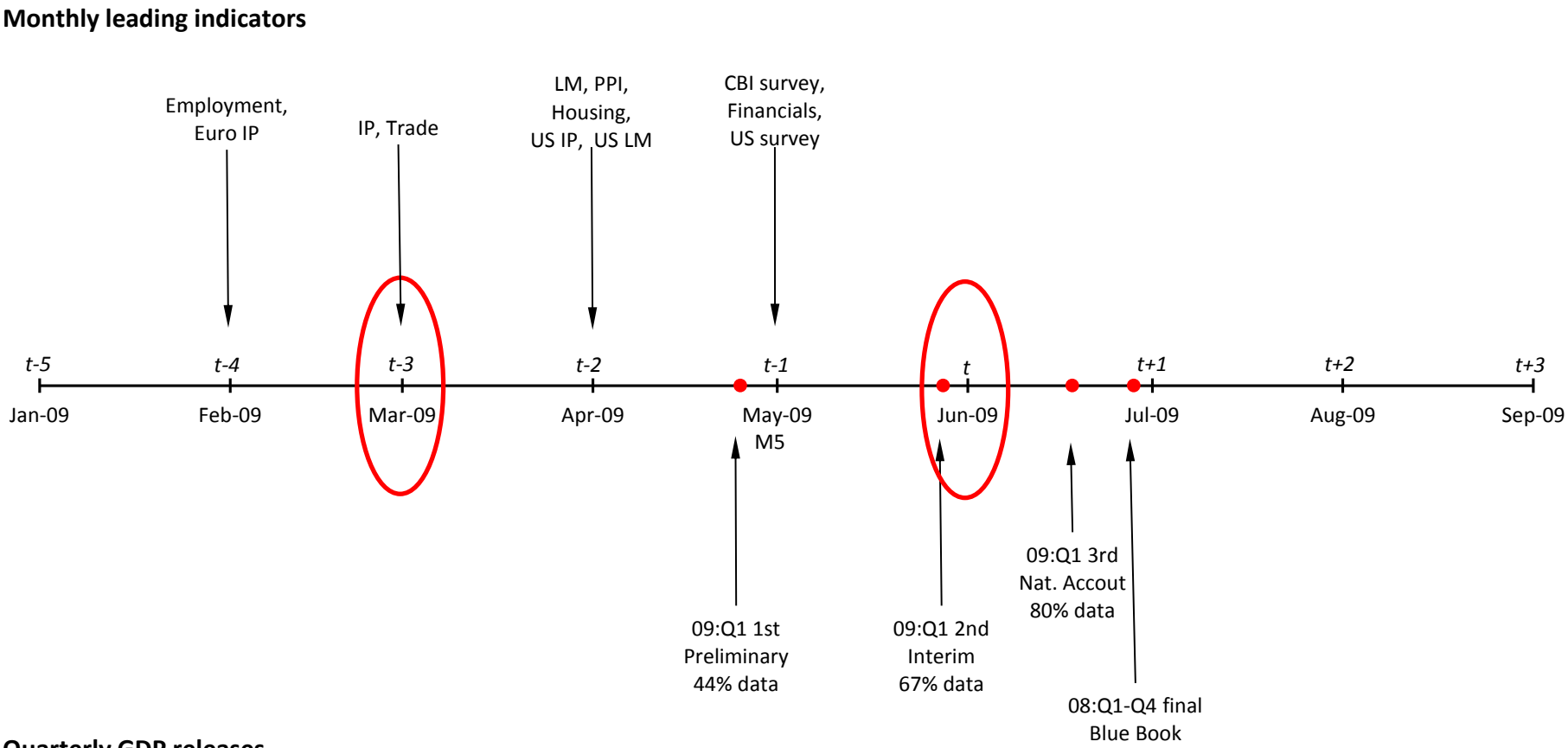

Figure : Pseudo release timing and structure of data vintages for the quarterly GDP measurements and 60 monthly leading indicators for the UK

Nowcast Horizons

	t_{m_1}	t_{m_2}	t_{m_3}	t_{m_4}	t_{m_5}	t_{m_6}	t_{m_7}	t_{m_8}	t_{m_9}	$t_{m_{10}}$	$t_{m_{11}}$	$t_{m_{12}}$
t_{q_1}		$h^1_{t_{q_1}}$	$h^2_{t_{q_1}}$	$h^3_{t_{q_1}}$ $y^{v_1}_{t_{q_1}}$	$y^{v_2}_{t_{q_1}}$	$y^{v_3}_{t_{q_1}}$						
t_{q_2}					$h^1_{t_{q_2}}$	$h^2_{t_{q_2}}$	$h^3_{t_{q_2}}$ $y^{v_1}_{t_{q_2}}$	$y^{v_2}_{t_{q_2}}$	$y^{v_3}_{t_{q_2}}$			
t_{q_3}								$h^1_{t_{q_3}}$	$h^2_{t_{q_3}}$	$h^3_{t_{q_3}}$ $y^{v_1}_{t_{q_3}}$	$y^{v_2}_{t_{q_3}}$	$y^{v_3}_{t_{q_3}}$
t_{q_4}	$h^3_{t_{q_4}}$ $y^{v_1}_{t_{q_4}}$	$y^{v_2}_{t_{q_4}}$	$y^{v_3}_{t_{q_4}}$								$h^1_{t_{q_4}}$	$h^2_{t_{q_4}}$

Table : Nowcast horizons and GDP releases in real time. Each entry corresponds to a nowcast/release of the quarterly GDP growth in a quarter t_{q_i} available in month t_{m_j}

Benchmark: Vintages Models

- A benchmark univariate model that includes the quarterly data on real time GDP growth vintages. To utilize all information available in real time, GUM differs for three nowcasting horizons:

$$
\begin{aligned}
h^1_{t_q} : \Delta\hat{y}_{t_q} &= f\left(\Delta y^{v_1}_{t_q-1},\ldots,\Delta y^{v_1}_{t_q-4};\Delta y^{v_2}_{t_q-1},\ldots,\Delta y^{v_2}_{t_q-4};\Delta y^{v_3}_{t_q-2},\ldots,\Delta y^{v_3}_{t_q-4}\right)\\
h^2_{t_q} : \Delta\hat{y}_{t_q} &= f\left(\Delta y^{v_1}_{t_q-1},\ldots,\Delta y^{v_1}_{t_q-4};\Delta y^{v_2}_{t_q-1},\ldots,\Delta y^{v_2}_{t_q-4};\Delta y^{v_3}_{t_q-1},\ldots,\Delta y^{v_3}_{t_q-4}\right)\\
h^3_{t_q} : \Delta\hat{y}_{t_q} &= f\left(\Delta y^{v_1}_{t_q},\ldots,\Delta y^{v_1}_{t_q-4};\Delta y^{v_2}_{t_q-1},\ldots,\Delta y^{v_2}_{t_q-4};\Delta y^{v_3}_{t_q-1},\ldots,\Delta y^{v_3}_{t_q-4}\right)
\end{aligned}
$$

- For h^1_t , the most recent estimate of $\Delta y^{v_3}_{t_q-1}$ is still not available in real time and is therefore excluded from the GUM.
- The last horizon h^3_t overlaps with the release period for the flash estimate and thus the first contemporaneous vintage of GDP is included in the GUM.

Single-indicator Models

- Growth is projected onto the models each augmented with one single monthly indicator z^j_{i,t_q}, resulting in a total of 60 models for each evaluation period.
- Forecasts for the missing observation is omitted, utilizing in-sample information only.
- The maximum lag for the indicators is 12 and the minimum lag corresponds the earliest observed realization.

Augmented Models with In-sample Indictors

- Augment vintages models with the full set of leading indicators.
- Only in-sample information enters the GUM so that the ragged-edge problem is not corrected for.
- The nowcasting model from the selected specification then has the following form:

$$\Delta \hat{y}_{t_q}^{h_q^k} = \hat{\boldsymbol{\alpha}}' \Delta \hat{\mathbf{y}}_{t_q}^{v} + \hat{\boldsymbol{\beta}}' \hat{\mathbf{z}}_{t_q} + \hat{\boldsymbol{\zeta}}' \hat{\mathbf{d}}_{t_q} \tag{13}$$

  - $\Delta \hat{\mathbf{y}}_{t_q}^{v}$ is a vector containing the selected lags of the growth estimates,
  - $\hat{\mathbf{z}}_{t_q}$ is a vector of lags of the relevant in-sample leading indicators
  - $\hat{\mathbf{d}}_{t_q}$ are significant impulse dummies.

Augmented Models with Forecasted Indicators

- This model utilizes block separation and corrects the ragged edge problem by using forecasts for the missing monthly observations.
- A nowcasting model then has the following form:

$$\Delta \hat{y}_{t_q}^{h_q^k} = \hat{\boldsymbol{\alpha}}' \Delta \hat{\mathbf{y}}_{t_q}^{v} + \hat{\boldsymbol{\beta}}' \hat{\mathbf{z}}_{t_q} + \hat{\boldsymbol{\gamma}}' \tilde{\mathbf{z}}_{t_q} + \hat{\boldsymbol{\zeta}}' \hat{\mathbf{d}}_{t_q} \tag{14}$$

- where all variables are the same as in (13) and $\tilde{\mathbf{z}}_{t_q}$ are the forecasted values for the selected relevant leading indicators.

Augmented Models with Robust Forecasts

- Instead of the conditional monthly forecasts from (12), a robust device is used for individual indicators forecasts that is robust to breaks, such that:

$$\tilde{z}_{i,t_q} = z_{i,t_q - l_i} \tag{15}$$

where l_i is the release lag for variable i.

Results: Vintages Models

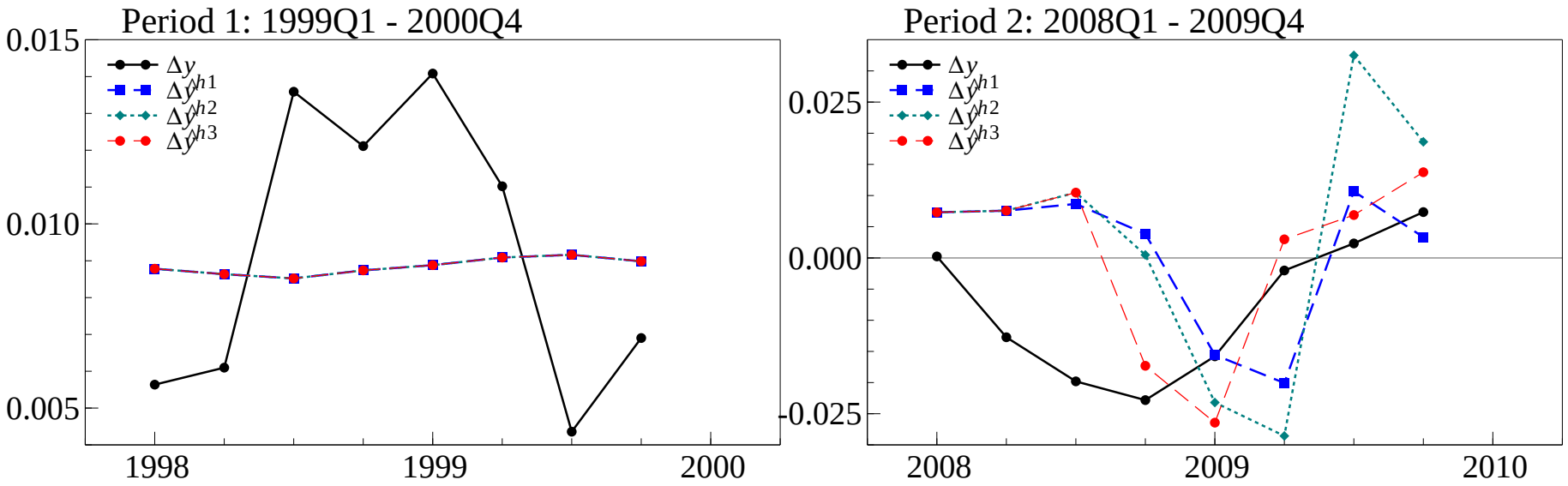

Figure : Nowcasts from vintages model for three horizons and two subsamples: 1999Q1 - 2000Q4 (left panel) and 2008Q1 - 2009Q4 (right panel)

Results: Single Indicator Models

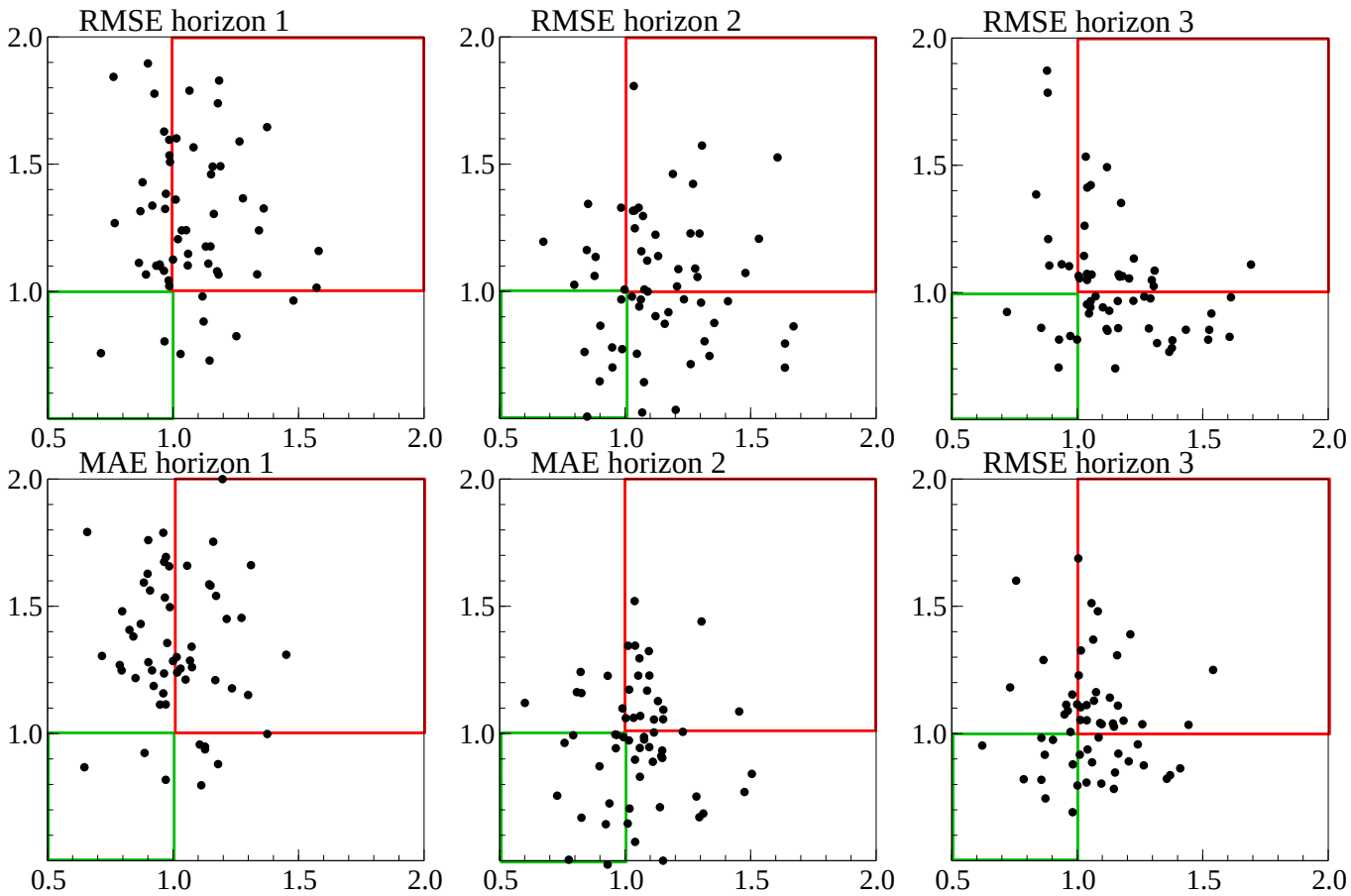

Figure : Ratio of RMSE for single model to vintages model for the first period - 1992-2000 on x-axis and for the second period - 2001-2009 on y-axis. Red area - region of robust underperformance, green - region of robust overperformance

Results: Augmented Models

	RMSE					
	1993 - 2000			2001 - 2009		
Model	$h^1_{t_q}$	$h^2_{t_q}$	$h^3_{t_q}$	$h^1_{t_q}$	$h^2_{t_q}$	$h^3_{t_q}$
Vintages	*0.004*	*0.004*	*0.004*	*0.017*	*0.022*	*0.014*
Augmented in-sample*	1.178	1.868	1.306	0.791	0.556	0.664
Augmented forecasts*	1.442	1.029	1.142	0.784	0.438	0.689
Augmented robust*	1.476	1.208	1.195	0.691	0.438	0.672

*RMSE ratio to RMSE of the corresponding vintages model

Table : RMSE of augmented models as ratio to RMSE of the vintages model

Conclusions

- Use model selection (Autometrics) to nowcast UK real output growth
- Exploit non-stationarity to improve forecasting accuracy
- Detect breaks using impulse-indicator saturation
- Robustify leading indicators forecasts if there is evidence of contemporaneous breaks

References

Castle, J. L., D. F. Hendry, and O. I. Kitov (2013). Forecasting and nowcasting macroeconomic variables: A methodological overview. Economics Series Working Paper 674, University of Oxford, Department of Economics.

Doornik, J. A. (2009). Autometrics. In *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*. Oxford University Press.

Hendry, D. F. (2006). Robustifying forecasts from equilibrium-correction systems. *Journal of Econometrics 135*(1-2), 399–426.

Hendry, D. F., S. Johansen, and C. Santos (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics 23*(2), 337–339.