Density nowcasts and model combination: nowcasting Euro-area GDP growth over the 2008-9 recession

Workshop on "Uncertainty and Forecasting in Macroeconomics", Deutsche Bundesbank

Gian Luigi Mazzi[‡], James Mitchell[+,†] and Gaetana Montana[‡]
[‡] Eurostat
[+] Department of Economics, University of Leicester
[†] National Institute of Economic and Social Research, London

1-2 June 2012

Faster production of GDP data

- Statistical offices publish 'official' GDP data at a lag
- Eurostat publishes its *Flash* estimate of quarterly GDP growth for the Euro-area (EA) about 45 days after the end of the quarter
  1. This meant that Eurostat did not identify the EA "recession" (negative quarters in 2008q1 and 2008q2) until 14th November 2008
  2. This was despite the fact that published qualitative survey data, and other *indicators*, were at the time interpreted by some as convincing evidence that the EA was already in recession
  3. But without a formal means of assessing the utility of these incomplete (sectoral, qualitative survey etc.) data, and relating them to official GDP data, we don't know how much weight to place on them when forming a view about the current state of the economy

The generic statistics office

- Is under pressure to speed up delivery of their quarterly GDP estimates
- But resource constraints mean they must rely increasingly on *nowcasting* models, rather than faster official surveys
  - use of within-quarter information on indicator variables
  - as we shall see, there are many possible higher-frequency indicators, "hard" and "soft", aggregate and disaggregate
- Expect a trade-off between the timeliness and accuracy of nowcasts
  - it is therefore important to quantify this
- This paper suggests a formal but computationally convenient method for establishing what role, if any, indicator variables should play when constructing nowcasts of current quarter GDP growth
- The uncertainty associated with the nowcast is acknowledged, and subsequently evaluated, by constructing density nowcasts
  - with the density nowcasts produced at various publication lags

Methodology

1. Density forecast combination, with N large
2. Used in other applications; e.g., density forecasting US inflation (Jore, Mitchell & Vahey 2010 JAE), Norwegian aggregates (Bache et al. 2011 JEDC) and the output gap (Garratt, Mitchell & Vahey 2011)

Methodology

1. Density forecast combination, with N large
2. Used in other applications; e.g., density forecasting US inflation (Jore, Mitchell & Vahey 2010 JAE), Norwegian aggregates (Bache et al. 2011 JEDC) and the output gap (Garratt, Mitchell & Vahey 2011)
3. This paper considers how to implement the methodology, and assesses its peformance including over the recent recession, when nowcasting EA GDP with mixed-frequency data as monthly (within-quarter) data accrue
    - The density nowcasts reflect the publication lags of each indicator
    - To construct density nowcasts for GDP growth we take combinations across a large number of competing *component* models
    - Component models are distinguished by their use of “hard” and “soft”, aggregate and disaggregate, indicators
    - The post-data weights on the components are time-varying and reflect the relative fit of the component model forecast densities

Background on Methodology

- Density combination (a kind of ensembling) a great way to produce more accurate/robust probabilistic forecasts
- Now used at central banks (in particular Norges Bank) when nowcasting & forecasting using a suite of models
- Probabilistic Forecasting Institute (ProFI) has been set up
  - to stimulate and coordinate research into new methods for probabilistic forecasting, evaluation and communication
  - to exchange ideas for operationalising methodologies

More on Methodology

- In the presence of 'uncertain instabilities' it can be helpful to combine the evidence from many models
  - Large extant literature combining point forecasts
  - Equal weights tend to outperform weighted alternatives
  - In density context combination helps, but equal weights can be beaten (JMV JAE 2010)
- Selecting a single model has little appeal when the single best model suffers from instability
  - This might happen either if the 'true' model is not within the model space, or if the model selection process performs poorly on short macro samples
- We use the linear opinion pool to combine density nowcasts
- The design of the model space and the number of components to be considered needs to be specified

Linear opinion pool (LOP)

Given $i = 1, \ldots, N_j$ component models, the combination densities for GDP growth are given by the LOP:

$$p(\Delta y_\tau) = \sum_{i=1}^{N_j} w_{i,\tau,j} \; g(\Delta y_\tau \mid \Omega_\tau^j), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau},$$

where N_j (j denotes the j-th nowcast) is such that $N_{j+1} > N_j$

- $g(\Delta y_{\tau,h} \mid \Omega_\tau^j)$ are the nowcast forecast densities from component model i each conditional on one element from the information set Ω_τ^j
- The non-negative weights, $w_{i,\tau,j}$, sum to unity
- $g(\Delta y_\tau \mid \Omega_\tau^j)$ (with non-informative priors), allowing for small sample issues, are Student-t

Go large

- We exploit large N density combinations
  - As N increases, the combined density becomes more flexible, with the potential to approximate non-linear, non-Gaussian specifications
  - Some similarities with *ensembles* in the meteorology literature
- Contrast with small N combinations
  - Hall & Mitchell (2005/7) combine BoE and NIESR densities
  - Amisano and Geweke (2012) combine DSGE, BVAR and DFM densities
- Component models might all be individually misspecified; but some might work reasonably well at some points in time
  - differ in how they adapt to structural changes (incl. the recession)
  - components can include robust forecasting models
  - we consider a range of AR type models below

Model space

- The nowcasts are produced by statistical models which relate GDP growth to *indicator* variables
- These are variables which are meant to have a close relationship with GDP but are made available more promptly
    - they are often published at a higher frequency (monthly)
- But there is uncertainty about what indicator variable(s) to use; e.g.
    1. *Hard* monthly data on Industrial Production, IP (typically published at t+30 days), retail trade...
    2. *Soft* qualitative survey data (published at t+0 days)
    3. The set of possible indicators increases further when we consider variables not directly related to GDP but presumed to have some indirect relationship (e.g. interest rate spread)

Aggregate and disaggregate indicators

- As well as considering data at the aggregate, EA(12), level we examine them at the disaggregate (national) level too - for each of the 12 EA countries
  - real-time data (vintages) are used for these national data
- Use of disaggregate data in an aggregate model can better approximate the infeasible but (RMSE) efficient multivariate forecast; see Hendry and Hubrich (2011, JBES)

Aggregate and disaggregate indicators

- As well as considering data at the aggregate, EA(12), level we examine them at the disaggregate (national) level too - for each of the 12 EA countries
  - real-time data (vintages) are used for these national data
- Use of disaggregate data in an aggregate model can better approximate the infeasible but (RMSE) efficient multivariate forecast; see Hendry and Hubrich (2011, JBES)
  - Alternatively a global VAR could be used to nowcast an aggregate using disaggregate VARs (Lui & Mitchell, 2012, GVAR Handbook) or a large BVAR
  - Ravazzolo and Vahey (2012) consider disaggregate density forecast combinations
- Disaggregate data can also help as some countries publish their hard data more quickly than others (incl. Eurostat)
  - Portugal publishes monthly IP data at the end of month $m+1$
  - Belgium and Spain currently publish quarterly GDP data at the end of month $m+1$ (i.e., at t+30 days)

Nowcasting models

- Different nowcasting models involve different ways of linking the indicator variables to GDP
- This can be done at a quarterly, monthly or mixed frequency
- It is an empirical question which is most sensible
- Appealing to Occam's razor, we focus on simple component models
    - we estimate, à la Kitchen & Monaco (2003, Business Economics), a linear regression of quarterly GDP growth on a single k-th indicator variable $x^m_{k,t}$ (which might be a lag)

$$\Delta y_t = \beta_0 + \beta_1 x^m_{k,t} + e_t; \;\; (m = 1, 2, 3)$$

      where e_t is assumed normally distributed
    - but combination methodology also appropriate for other models (bridge models, MIDAS, mixed-frequency VAR, dynamic factor models, with temporal aggregation constraint etc.)

Accrual of within-quarter information

- Combine the component density nowcasts across k using the linear opinion pool as within-quarter information accumulates
- We produce to six timescales ($j = 1, ..., 6$)
- At all six timescales we know the value of GDP in the previous quarter
- But this (t-1) estimate may be measured by the first (Flash), second or third release from Eurostat
  - If we know >1 release we consider all known releases (accommodate any predictability in data revisions)

Production of nowcasts as within-quarter data accumulate

1. j=1. t-30: 30 days before the end of the quarter

$$\Omega_t^1 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^2 , \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1} , \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right)$$

Production of nowcasts as within-quarter data accumulate

1. j=1. t-30: 30 days before the end of the quarter

$$\Omega_t^1 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^2, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right)$$

2. j=2. t-15: 15 days before the end of the quarter

$$\Omega_t^2 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^2, x_{hard,t}^1, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right)$$

Production of nowcasts as within-quarter data accumulate

1. j=1. t-30: 30 days before the end of the quarter

$$\Omega_t^1 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^2, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right)$$

2. j=2. t-15: 15 days before the end of the quarter

$$\Omega_t^2 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^2, x_{hard,t}^1, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right)$$

3. j=3. t+0: 0 days after the end of the quarter

$$\Omega_t^3 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^3, x_{hard,t}^1, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right)$$

4. j=4. t+15: 15 days after the end of the quarter

$$\Omega_t^4 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^3, \left\{ x_{hard,t}^m \right\}_{m=1}^2, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right)$$

5. j=5. t+30: Also includes $x_{hard,t}^{3,Por}$ and $\Delta y_t^{Bel,US}$
6. j=6. t+45: Includes $\left\{ x_{hard,t}^m \right\}_{m=1}^3$ too

Component models and data transformations

- Consider different transformations of $x^m_{k,t}$ to accommodate uncertainty about whether the informational content of these data is higher when a first or quarterly difference is taken
- Treat these various transformations of a given indicator variable $x^m_{k,t}$ as additional component models
- The qualitative survey data are considered in levels as well as monthly first-differences and quarterly differences
- The quarterly transformation is

$$x_{k,t} = \frac{1}{3}\Delta \log z^3_{k,t} + \frac{2}{3}\Delta \log z^2_{k,t} + \Delta \log z^1_{k,t} + \frac{2}{3}\Delta \log z^3_{k,t-1} + \frac{1}{3}\Delta \log z^2_{k,t-1}$$

- Given these assumptions, and the availability of the aggregate and disaggregate data, $N_1 = 214$; $N_2 = 293$; $N_3 = 351$; $N_4 = 430$; $N_5 = 438$ and $N_6 = 444$

Constructing the combination weights

- The EW strategy attaches Equal (prior) Weight to each model with no updating of the weights: $w_{i,\tau,j} = w_{i,j} = 1/N_j$
- Recursive Weights

$$w_{i,\tau,j} = \frac{\exp\left[\sum_{\underline{\tau}-8}^{\tau-1} \ln g(\Delta y_\tau \mid \Omega_\tau^j)\right]}{\sum_{i=1}^{N} \exp\left[\sum_{\underline{\tau}-8}^{\tau-1} \ln g(\Delta y_\tau \mid \Omega_\tau^j)\right]}, \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}$$

The logarithmic scoring rule is intuitively appealing as it gives a high score to a density forecast that assigns a high probability to the realised value

  - The model densities are combined using Bayes' rule with equal (prior) weight on each model—which a Bayesian would term non-informative priors
  - Some similarities with an approximate predictive likelihood approach
  - When the model space is incomplete, the conventional Bayesian interpretation of the weights as reflecting the posterior probabilities of the components is inappropriate

Selecting the model space

- Economically or statistically?

Selecting the model space

- Economically or statistically?
- Forecast diversity is important
  - Dependence
- Recall Tobin's advice when picking financial assets:
  - 'don't put your eggs in one basket'
- We consider whether there are empirical benefits to excluding some *bad* models, prior to taking the combination
- À la Madigan and Raftery (1994, JASA) model i is discarded from the combination if it predicts less well according to the logarithmic score than the best model, i.e. if:

$$\frac{\max\{w_{i,\tau,j}\}_{i=1}^{N}}{w_{i,\tau,j}} > c$$

- Variant of Granger's thick modelling
- We also consider the performance of that model which is recursively selected as the best single model, according to $w_{i,\tau,j}$

Evaluation of nowcast densities

1. We test for *complete* calibration (see Mitchell and Wallis, 2011 JAE) by examining whether the *pits* z_τ, where $z_\tau = \int_{-\infty}^{\Delta y_\tau} p(u)du$, are uniform and i.i.d.
2. Undertake a battery of goodness-of-fit and independence tests widely used in the literature
    - To control the joint size of our eight evaluation tests, at a 95% significance level, requires the use of a stricter *p*-value for each individual test than the 5% value we use. The Bonferroni correction indicates a *p*-value threshold, for a 95% significance level, of $(100\% - 95\%)/8 = 0.6\%$ rather than 5%
3. Recent alternatives suggested by Malte Knüppel, and Barbara Rossi and others

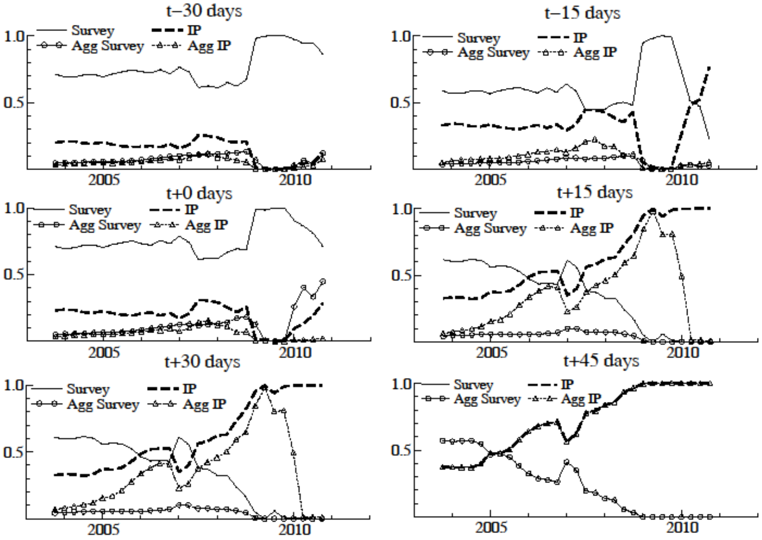
Nowcasting EA GDP growth

- Recursive out-of-sample experiments using real-time data vintages
- Evaluation period is 2003q2-2010q4
  - Eurostat published its first *Flash* estimate for GDP growth for 2003q2
- The nowcasts are evaluated by defining the 'outturn' as the first (Flash) GDP growth estimate from Eurostat
- Break our results into two parts:
  1. the RW weights on the soft indicators, the hard indicators and lagged GDP growth derived from the logarithmic score of the component forecast densities
     - is there support for EW? For an AR? Do the weights change over time (instabilities)? Do disaggregate data help?
  2. the evaluations of the recursive weight, RW, and equal weight, EW, combination strategies

Weights on the components

- The interest rate spread received little or no weight and henceforth we equate the soft data with the (qualitative) survey data
- AR components receive a weight less than 0.1 and weight declines as j increases
- These weights, on a given type of indicator, e.g. the survey data or IP, involve summing the weights on all of the component models estimated using various transformations of the given indicator
- For the hard indicators (i.e., IP and GDP growth) it also involves summation of the weights given to component models which use lagged instead of contemporaneous values
- To identify the relative informational content of the aggregate versus the disaggregate indicators, we also plot the weights when aggregate indicators only are considered

Weights on the components



The length of the training period

- The shorter the training period the more quickly the combined density can adjust to changes over time in the performance of the different models
- But the longer the length of the training period the better the combination weights are estimated
- Here we use an increasing window; but a rolling window did not deliver gains
- Alternatives are to let the weights vary by *regime* (Waggoner & Zha, 2012), or depend on the region of the density of interest (Fawcett, Kapetanios, Mitchell & Price, in progress)

Evaluating the nowcast densities

Table: Number of *pits* tests (out of eight) which indicate correct calibration at 95 percent

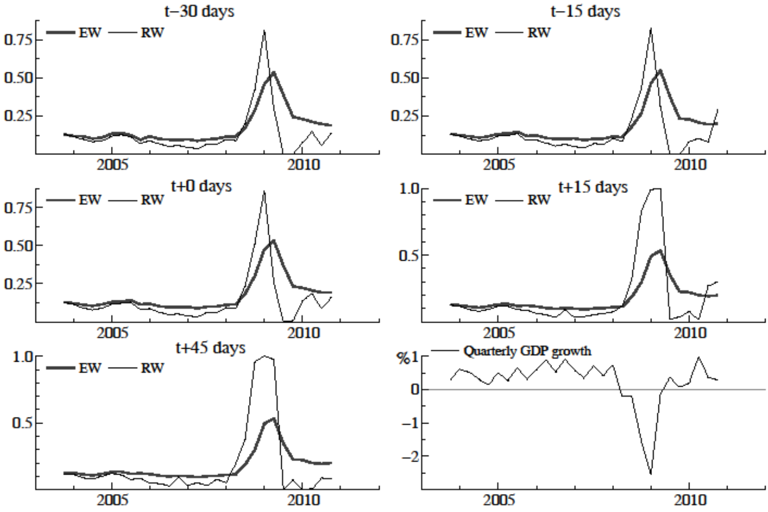
	t-30 days $j = 1$	t-15 days $j = 2$	t+0 days $j = 3$	t+15 days $j = 4$	t+30 days $j = 5$
EW	3	4	3	5	5
RW	2	2	3	8	8
Survey	4	4	5	5	5
IP	4	5	5	5	6
EW (Agg)	4	5	6	6	6
RW (Agg)	3	6	6	7	7
AR	5	4	5	4	5
Occam: EW	2	3	2	7	7
Occam: RW	2	2	3	8	8
Select	3	3	4	8	8

Evaluating the nowcast densities

Table: Negative of the average logarithmic score: 2003q2-2010q4

j	EW Dis	RW Dis	Soft	Hard	EW Agg	RW Agg	AR	Occam EW	Occam RW	Sel
$t-30$	0.73	0.85	0.70	0.81	0.71	0.82	0.84	0.87	0.90	1.35
$t-15$	0.72	0.80	0.71	0.77	0.66	0.82	0.87	0.87	0.84	1.30
$t+0$	0.69	0.79	0.70	0.77	0.64	0.74	0.87	0.85	0.86	0.89
$t+15$	0.66	0.50	0.70	0.68	0.60	0.61	0.84	0.48	0.46	0.54
$t+30$	0.66	0.50	0.70	0.70	"	"	0.84	0.48	0.46	0.54
$t+45$	0.65	0.48	0.70	0.70	0.53	0.46	0.85	0.51	0.46	0.43

Probability of a recession: a region of interest



Conclusions

- Density combination methods make it possible to know how much weight to place on different indicators when forming, at various points in time as monthly information accumulates, a view about the current state of the economy

Conclusions: Nowcasting EA GDP Growth

- We find that the relative utility of "soft" data increased suddenly during the recession
- But as this instability was hard to detect in real-time it helps, when producing nowcasts not knowing any within-quarter "hard" data, to weight the different indicators equally
- On receipt of two months of within-quarter "hard" data (at t+15 days) better calibrated densities are obtained by giving a higher weight in the combination to "hard" indicators unless the *poor* models are eliminated prior to combining
- Similarly, selecting the *best* model is also effective from t+15 days onwards, given there is by then more of a consensus about the preferred indicator(s)
- But earlier in the quarter, given the observed instabilities and uncertainties about the right indicator, selection performs poorly relative to both equal and weighted density combinations

Ongoing work

- Methodological
  - Copula opinion pools to accommodate density forecast dependence
  - Let the weights depend on the region of the density of interest
  - Economic (cost-loss) evaluation of density forecasts incl. density nowcasts (Garratt et al.)
- Communication
  - Presentation of histograms, rather than densities, to emphasise uncertain uncertainty?