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

Gian Luigi Mazzi[‡], James Mitchell[+,†,*] and Gaetana Montana[‡]

[‡] Eurostat

[+] Department of Economics, University of Leicester

[†] National Institute of Economic and Social Research

April 20, 2012

Abstract

Combined density nowcasts for quarterly Euro-area GDP growth are produced based on the real-time performance of component models. Components are distinguished by their use of "hard" and "soft", aggregate and disaggregate, indicators. We consider the accuracy of the density nowcasts as within-quarter information on monthly indicators accumulates. We focus on their ability to anticipate the recent recession probabilistically. 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 at least one month of within-quarter "hard" data better calibrated densities are obtained by giving a higher weight in the combination to "hard" indicators.

*Corresponding author. James Mitchell, Department of Economics, University of Leicester, U.K.. E-Mail: jm463@le.ac.uk. We are grateful to Silvia Lui, an anonymous referee and participants at the ISF June 2010 and the Eurostat Colloquium September 2010 for helpful comments.

1 Introduction

Statistical offices publish 'official' GDP data at a lag. Eurostat publishes its so-called Flash estimate of quarterly GDP growth for the Euro-area (EA) about 45 days after the end of the quarter. This meant that Eurostat data did not indicate, for example, that the Euro-area was in "recession" until 14th November 2008. As is common, a "recession" is defined as two successive quarters of negative quarterly growth. Economists and policymakers therefore had to wait 45 days to be told that the economy shrank in 2008q3 - as well as 2008q2. This was despite the fact that published qualitative survey data, as well as other so-called indicator variables, were at the time being interpreted by some as convincing evidence that the Euro-area economy was already in recession. But without a formal means of assessing the utility of these qualitative survey data, and relating them to official Eurostat GDP data, it is impossible to know how much weight to place on them when forming a view about the current state of the economy. An accurate, but timely, impression of the state of the economy is important for policymakers.

There is always a pressure on statistical offices to speed up the delivery of these estimates. Inevitably, with resource constraints impeding the production of earlier/higher frequency official quantitative surveys, this means relying increasingly on forecasting. Or, more accurately, this should be called nowcasting - as within quarter information on indicator variables is exploited. But there is an expected trade-off between the timeliness and accuracy of nowcasts. Nowcasts can always be produced more quickly by exploiting less information; but, we might expect the quality of the nowcasts to deteriorate as a result.

In this paper we suggest a formal but computationally convenient method for establishing what role, if any, these indicator variables should play when constructing nowcasts of current quarter GDP growth. Importantly, the uncertainty associated with the nowcast is acknowledged, and subsequently evaluated, by constructing density nowcasts. Publication of these uncertainty estimates, alongside the central estimate, provides a means of indicating to the user the 'quality' of the nowcast, as measured by the confidence associated with the nowcast. Increasingly, in fact, forecasts are being presented probabilistically, with many central banks now publishing "fan charts". To construct the density nowcasts we take combinations across a large number of competing models. In this approach, model uncertainty, in particular uncertainty about what indicator variables should be used, is explicitly accommodated. The approach is also based on the belief that the candidate component models are all incorrectly specified. But the components allow the modeller

to explore a wide range of uncertainties. The resulting combination reflects the model uncertainty by taking a weighted average across many (simple) component models, with the component models distinguished by what indicator variables they consider. The post-data weights on the components can be time-varying and reflect the relative fit of the individual model forecast densities. The combination becomes very flexible as the number of component models rises; and aims to approximate an unknown but likely complex (non-linear and non-Gaussian) data-generating-process. A related approach has been applied by Jore et al. (2010) to forecast US macroeconomic aggregates. This paper considers both how it can be used, and assesses its efficacy in an application, when nowcasting with real-time mixed-frequency data.

While GDP growth is published at a quarterly frequency, many indicator variables are available at a higher frequency. We consider how nowcasts of quarterly GDP can be constructed as within quarter, monthly, information on these indicator variables accrues. Thereby, our density nowcasts reflect the publication lags of each indicator variable. Following Giannone et al. (2008), we distinguish between quantitative ("hard") and qualitative ("soft") indicator variables, with the soft indicators typically published ahead of hard data. And we consider both EA (aggregate) and country-level (disaggregate) indicators. Examination of country-level indicator data might prove efficacious if, following Hendry & Hubrich (2011), these disaggregates contain information over and above that in aggregate indicators. Moreover, some countries publish their hard data more quickly than others, indeed more rapidly than Eurostat publishes the corresponding aggregate.

We then assess the ability of the combination methods to anticipate the 2008-2009 recession. We assess their ability to predict the probability of a recession and more generally examine their density nowcasts. How well did they do at flagging up the recent contraction to GDP growth, and how far ahead did they successfully call the recession? The latter is important, given the likely trade-off between the timeliness and accuracy of nowcasts. We also seek to identify what, if any, indicator variables were most helpful in anticipating the recession. Thereby, this paper provides timely evidence about the performance of nowcasts over this unusual period. It also extends previous work, such as Giannone et al. (2008), by explicitly constructing density nowcasts. At times of heightened uncertainty it is particularly important to quantify, in real-time, the degree of uncertainty associated with any nowcast. Moreover, in contrast to the majority of applied studies studying the Euro-area, we exploit real-time (aggregate and disaggregate) data made available to us by Eurostat. This means that our out-of-sample simulations are genuinely, rather than 'pseudo', real-time.

The plan of the remainder of this paper is as follows. Section 2 motivates the use of density nowcasts. Section 3 explains the importance of accommodating model uncertainty when nowcasting by considering a set of component models rather than a single model. Section 4 describes the component models used in our nowcasting application. Section 5 explains how our combined density nowcasts are computed and Section 6 provides the empirical results. Section 7 concludes.

2 Density nowcasts

It has become increasingly well understood that it is not a question of this nowcast proving to be 'right' and that nowcast proving to be 'wrong'. Point nowcasts, the traditional focus, are better seen as the central points of ranges of uncertainty. A GDP growth nowcast of, say, 2% must mean that people should not be surprised if actual growth turns out to be a little larger or smaller than that. Moreover, perhaps, at a time of heightened economic uncertainty, they should not be very surprised if it turns out to be much larger or smaller. Consequently, to provide a complete description of the uncertainty associated with the point nowcast many forecasters now publish density nowcasts/forecasts, or more popularly "fan charts".

More formally, density forecasts of GDP growth, say, provide an estimate of the probability distribution of its possible future values. In contrast to interval forecasts, which give the probability that the outcome will fall within a stated interval, such as GDP growth falling within its target range, density forecasts provide a complete description of the uncertainty associated with a forecast. They can thus be seen to provide information on all possible intervals.

What really matters is how nowcasts, or indeed forecasts, affect decisions. The 'better' nowcasts are those that deliver 'better' decisions. On this basis it is argued that the appropriate way of evaluating nowcasts is not to use some arbitrary statistical loss function, but the appropriate economic loss function; e.g. see Granger & Pesaran (2000). Only when the nowcast user has a symmetric, quadratic loss function, and the constraints (if relevant) are linear, is it correct to focus on the point nowcast alone. This is what textbook's call "certainty equivalence". In the more general case, the degree of uncertainty matters. Users are not indifferent to the degree of uncertainty about the point forecast. Uncertainty is expected to attenuate responses to the point nowcast.

The importance of publishing density nowcasts then follows from the fact that we tend,

in reality, not to know users' loss functions. But we should not expect these (unknown to us) functions to be quadratic. For example, we should expect the range of uncertainty to matter. When the user's loss function is asymmetric, such that positive and negative forecasting errors have differing costs, the user's "optimal" forecast need not equal the conditional mean; e.g., see Zellner (1986). By publishing the whole density nowcast for GDP growth the statistics office, the putative producer of the densities in our application, ensures the user is free to extract from the density any feature of concern to them. This feature might be the conditional mean. But interest often focuses in tail events, which require an explicit statement about the uncertainty associated with the nowcast. Users of growth forecasts may be concerned about the probability of recession. These probability event forecasts can readily be extracted from the density forecast.

The trend towards forecasters publishing density forecasts is also explained by the obvious advantages they bring when communicating with the public. It reminds them that the statisticians/forecasters themselves expect the point forecasts to be 'wrong'. It also lets users assess the balance of risks associated with the nowcast.

3 Combination forecasts - and nowcasts

Bayesian Model Averaging (BMA) offers a conceptually elegant means of dealing with 'model uncertainty'. BMA forecasts condition not on a single 'best' model but take a weighted average over a range of candidate models; see Hoeting et al. (1999). This follows from appreciation of the fact that, although one model may be 'better' than the others, we may not select it with probability one. We may not be sure that it is the best forecast. Therefore, if we considered this single forecast alone, we would be overstating its precision.

Similarly in practical macroeconomic forecasting exercises, whether within a Bayesian context or not, it is a stylised fact that combination forecasts are hard to beat. The estimated parameters of a single forecasting model are commonly found to exhibit instabilities and these can be difficult to identify in real-time. In the presence of these so-called 'uncertain instabilities' it can be helpful to combine the evidence from many models. For example, Clark & McCracken (2010) examine the scope for taking linear combinations of point forecasts in real time, motivated by the desire to circumvent the uncertain instabilities in any particular specification. In a series of influential papers, Stock & Watson (2004) have documented the robust performance of point forecast combinations using various types of models for numerous economic and financial variables. Selecting a single

model has little appeal under 'uncertain instabilities' when the single best model suffers from instability. This might happen either if the 'true' model is not within the model space considered by the modeller, or if the model selection process performs poorly on short macroeconomic samples. We may better approximate the truth, and account for the uncertainty in model selection, by combining forecasts.

While methods for combining point forecasts are well established and much exploited, less direct attention has been given in econometrics to the combination of density forecasts.[1] Although, as Wallis (2005) has noted, density forecasts, in fact, have been combined in the ASA-NBER Survey of Professional Forecasters since 1969. The forecasters' densities are combined by taking a linear average, a so-called "linear opinion pool", as in BMA. Mitchell & Hall (2005) and Hall & Mitchell (2007) used this combination rule (but non-Bayesian weights) to combine and then analyse density forecasts from the Bank of England and the National Institute of Economic and Social Research. They considered how, in practice, the densities in the combination might be weighted, considering alternatives to equal weights. Jore et al. (2010) examine linear combinations of densities from VAR models, and Bache et al. (2011) take a linear combination of VAR and DSGE densities. Alternatives include consideration of logarithmic pooling rules; see Kascha & Ravazzolo (2010) and Wallis (2011). Using several examples, and theoretical analysis, Geweke & Amisano (2011) demonstrate the scope for pooled forecast densities to produce superior predictions, even if the set of components to be combined excludes the 'true' model - namely, when there is an 'incomplete model space'.

In this paper we follow in the spirit of this forecasting literature and 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; and we describe this below. We then produce density forecasts from a large set of component models, which differ in terms of the indicators variables, and transformations thereof, they consider. We suggest the use of simple regression based methods. Density forecasts from the component models are then produced analytically. This means that our approach is easy to apply.

[1] Outside the econometrics literature, the benefits of producing density forecasts by combining information across different models have been recognised for some time. For a review see Genest & Zidek (1986).

4 Nowcasting component models

The nowcasts are produced by statistical models. These statistical models by construction, and unlike structural or economic models, are reduced-form. They seek to explain and then nowcast GDP growth by exploiting information on indicator variables. These are variables which are meant to have a close relationship with GDP but are made available more promptly than the data for which they stand as a proxy; moreover, often these indicators are published at a higher frequency (e.g. monthly) than GDP itself, which is typically published on a quarterly basis only.

But there is uncertainty about what indicator variable or variables to use; in practice there is a large number to choose from. Following Giannone et al. (2008), we distinguish between quantitative ("hard") and qualitative ("soft") indicator variables, with the soft indicators typically published ahead of hard data. But the cost of that timeliness is their qualitative nature. These surveys typically ask respondents to provide qualitative categorical answers to a number of questions including what has happened to their output in the recent past and what they expect to happen to their output in the near future. Respondents say whether output has fallen, stayed the same or risen and which of the three they anticipate over some specified future period. The difference between the proportion of those expecting or reporting a rise and those expecting or reporting a fall is then related to GDP growth.

The set of indicator variables increases further when, as possible indicators, we consider variables not directly related to GDP but presumed to have some indirect relationship. For example, interest rates or exchange rates might be considered as they might help predict GDP growth. Our proposed combination approach provides a way of accommodating uncertainty about what indicator(s) to use.

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 a linear regression of quarterly GDP growth on a single indicator variable. We then combine the component density nowcasts using the linear opinion pool.

4.1 Indicator variables: aggregate and disaggregate

To illustrate use of the density forecast combination method we focus on a widely considered set of soft and hard indicator variables, generically denoted $x^m_{soft,t}$ and $x^m_{hard,t}$,

respectively, where $m = 1, 2, 3$ denotes the month in quarter t. As soft indicators, we consider the Economic Sentiment Indicator (ESI) and the spread between short term and 10 year Euro interest rates (available from the ECB). The ESI, published by the European Commission, is a widely used composite indicator. It combines various information from qualitative business tendency surveys, including expectations questions, into a single cyclical confidence indicator. As hard indicators we consider real-time monthly industrial production (IP) data.

As well as considering these data at the Euro-area aggregate, EA(12), level, we examine them at the disaggregate (national) level for each of the twelve Euro-area countries. The EA(12) comprise Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal and Spain. Again real-time data (vintages) are used for these national data. We consider the three monthly releases of national IP, but use only the first release values of national GDP, rather than the three within quarter estimates produced by Eurostat for the EA; this reflects variability across the European countries in terms of their publication of within quarter GDP data. We supplement the national qualitative survey data published by the European Commission (except for Greece and Ireland where data are unavailable) with additional business survey data for Germany, from Ifo, on the business climate, situation and expectations, given it is the largest economy in the EA. Use of these disaggregate data considerably increases the set of indications available; and allows for the possibility that a specific data series from a given country may help explain the aggregate over and above the aggregate information itself. Examination of hard country-level data can also prove efficacious given that some countries, as discussed below, publish their hard data more quickly than others and indeed more quickly than Eurostat publishes the corresponding aggregate. These disaggregate data can therefore be exploited when nowcasting the aggregate; cf. Hendry & Hubrich (2011).

The soft variables are published at the end of the month to which they refer. The hard variables for month m tend to be published, both for the Euro-area aggregate and most countries, around the middle of month $m + 2$ (i.e. at about t+45 days). The Quarterly National Accounts, which include the GDP data, are also updated around the middle of each month.

But there are some differences across countries in terms of how quickly they publish their data, and this must be reflected when producing nowcasts to different timescales. Portugal publishes its monthly industrial production data at the end of month $m + 1$; similarly Belgium publishes its quarterly GDP data at the end of month $m + 1$ (i.e., at t+30 days); therefore, below, we will introduce a nowcast of the Euro-area aggregate

($j = 5$, below) which exploits these data.[2] In fact, given the putative role of globalisation, we also condition this nowcast on the *advance* quarterly GDP data for the US, given that they are published at t+30 days too. Specifically, we use US real-time GDP vintage data from the Federal Reserve Bank of Philadelphia.

In principle, one could add further to our set of indicator variables, especially to the set of soft indicator variables. One could follow Giannone et al. (2008) and extract common factors from a panel data containing information on many indicators and regress GDP on them ('bridging with factors') in an additional component model. This could then be added to our existing set of models. We mention these possible extensions, like others previously, simply to illustrate the generality of our approach.

4.2 The trade-off between the timeliness and accuracy of nowcasts

We focus on producing nowcasts of quarterly GDP growth for the EA(12) to six timescales: t-30, t-15, t+0, t+15, t+30 and t+45 days. t denotes quarter, so that t-30, for example, means that a nowcast for quarter t is produced 30 days before the end of the quarter for which we want a quarterly GDP estimate. In contrast, the nowcast produced at t+45 is produced 45 days after the end of the quarter of interest.

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. Eurostat's Flash GDP estimate for the current quarter is released at about t+45 days. So our nowcast computed at t+45 is produced to the same timescale and therefore provides a benchmark for these official Eurostat estimates. By producing a density nowcast we are also able to characterise the uncertainty at t+45 days.

Since a monthly indicator, by construction, is released three times a quarter, following Kitchen & Monaco (2003), we estimate, in principle when the full quarter's data are available, three component models for each indicator. These involve relating quarterly GDP growth, Δy_t, to $x^m_{k,t}$ ($m = 1, 2, 3$; $t = 1, ..., T$). $x^m_{k,t}$ denotes the k-th indicator variable drawn from the information set Ω^j_t where j ($j = 1, ..., 6$) denotes the first, second, third, fourth, fifth and sixth nowcast formed at t-30, t-15, t+0, t+15, t+30 and t+45 days, respectively. Each successive nowcast exploits an ever larger information set. This

[2]Spain also recently began production of an earlier GDP estimate, at t+30 days. Once a sufficient number of vintages are available one could exploit these data too in out-of-sample simulations of the type undertaken in Section 6 below.

reflects the fact that with the passage of time more and more (aggregate and disaggregate) indicator data become available. Specifically, the component models take the form:

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

where e_t is assumed to be normally distributed. Let $x_{k,t}$ denote the quarterly analogue of $x^m_{k,t}$.

As an alternative to this approach of Kitchen & Monaco (2003), one could relate quarterly GDP growth directly to $x_{k,t}$, with monthly "bridge equations" used to forecast any missing monthly indicator data prior to aggregating to the quarterly estimate $x_{k,t}$, which is then regressed against Δy_t; e.g. see Baffigi et al. (2004). Our combination approach could also be employed if we adopted a related approach and, following Clements & Galvão (2008), used MIDAS regressions, say, to produce the component forecasts.[3] MIDAS regressions also provide a means of running regressions that allow the regressand and regressors (indicators) to be sampled at different frequencies.

4.2.1 The information set as within-quarter data accrues

The information set available at t-30, t-15, t+0, t+15, t+30 and t+45 accumulates as follows, where :

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

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

   $N_1 = no$ of elements of Ω^1_t. p_1 and p_2 denote the number of lags of the quarterly variables $x_{k,t}$ ($k = hard$) and Δy_t. We do not consider lagged quarterly values of the soft data; this seems reasonable, since when the nowcasts are produced we always have in our information set at least two months of within-quarter information on the soft indicators. But, to accommodate dynamics, we do consider previous quarter information about the hard indicators and GDP growth itself, given that these variables are published at a greater lag. In particular, for previous quarter GDP growth, Δy_{t-1}, we consider all three EA national accounts (vintage) estimates ending with $(T-1)$ information; plus we consider lagged values of the GDP vintage

[3] Alternatively, monthly GDP estimates could be estimated using mixed-frequency regression, VAR or factor based methods; these impose an aggregation constraint so that monthly GDP is consistent with the published quarterly values (e.g., see Mitchell et al. (2005), Mariano & Murasawa (2003) and Angelini et al. (2010)).

containing the first release of GDP for data up to quarter T (since this is available at about t+45 days, it has been known for about 15 days at j=1). Simultaneous consideration of multiple vintages means that, implicitly without modelling, the revisions process to GDP is accommodated; our density combination exercise does not simply use only the most recent vintage. In sum, (2) means Ω_t^1 includes two months of within-quarter soft data, as well as previous quarter hard indicator data and lagged GDP data. Ω_t^1 is then related to Δy_t, as measured by the first release of GDP growth, via (1).

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

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

$N_2 = no$ of elements of Ω_t^2. This means Ω_t^2 now includes the first month of within-quarter hard data, as well as Ω_t^1. Ω_t^2 is related to Δy_t, as measured by the second release of GDP growth data.

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

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

$N_3 = no$ of elements of Ω_t^3. Ω_t^3 now includes the final month of within-quarter soft data, as well as Ω_t^2. In practice, given that we use the same GDP release as at j=2 to measure Δy_t, to avoid duplicating component forecasts we do not re-consider indicators already used at j=2. This means the only new component forecast at j=3 involves regressing $x_{soft,t}^3$ on Δy_t.

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

$$\Omega_t^4 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^3, \left\{ x_{hard,t}^m \right\}_{m=1}^2, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right) \tag{5}$$

$N_4 = no$ of elements of Ω_t^4. Ω_t^4 now includes the second month of within-quarter hard data, as well as Ω_t^3. Ω_t^4 is related to Δy_t, as measured by the third release of GDP growth data.

5. j=5. t+30: 30 days after the end of the quarter.

$$\Omega_t^5 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^3, \left\{ x_{hard,t}^m \right\}_{m=1}^2, x_{hard,t}^{3,Por}, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \Delta y_t^{Bel,US}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right) \tag{6}$$

$N_5 = no$ of elements of Ω_t^5. Ω_t^5 now includes full quarter hard data for Belgium (for GDP), Portugal (for IP) and the US (for GDP), as well as Ω_t^4. Ω_t^5 is related to Δy_t, as measured by the third release of GDP growth data.[4]

6. j=6. t+45: 45 days after the end of the quarter.

$$\Omega_t^6 = \left(\left\{ x_{soft,t}^m \right\}_{m=1}^3, \left\{ x_{hard,t}^m \right\}_{m=1}^3, \left\{ x_{hard,t-l} \right\}_{l=1}^{p_1}, \left\{ \Delta y_{t-l} \right\}_{l=1}^{p_2} \right) \tag{7}$$

$N_6 = no$ of elements of Ω_t^6. We ignore the fact that we know, at t+45, Eurostat's Flash estimate for GDP growth in quarter t and now use all three months of within-quarter information on the monthly hard indicator data. At t+45 quarterly GDP data are also available for some of the Euro-area countries, but again we ignore these as a means of isolating the informational content of the monthly indicator data.

In each case ($j = 1, ..., 6$), each element (i.e. indicator) from Ω_t^j is related to Δy_t via (1) for $t = 1, ..., T$. We then use this model, and its estimated coefficients from the sample $t = 1, \ldots, T$, and the quarter $T + 1$ values of the indicator variables, Ω_{T+1}, to nowcast. Recall that the quarter $T + 1$ values of the indicator variables are published ahead of the quarter $T + 1$ values for Δy_t and can therefore be exploited when nowcasting. The nowcasts can be evaluated when Δy_{T+1} is subsequently published.

We set $p_1 = 1$ and $p_2 = 1$. These lag length assumptions rule out processes with lengthy lags in exogenous variables. We believe this is plausible when nowcasting, in particular. It is difficult for a statistics office to defend a situation where GDP is sharply influenced, for example, by movements in some indicator variable more than three months ago. (A professional forecaster, on the other hand, may have no difficulties in explaining the economic transmission mechanism.) Our assumption is designed to comply with the criterion that the models used to produce nowcasts should be credible to policymakers and other non-statisticians. The role of forecasting, as opposed to nowcasting, should be minimised. In practice, empirically we did experiment with the use of longer lags. These did not improve the accuracy of the density nowcasts. Unsurprisingly, within-quarter data is more informative.

[4] We did also consider the use of retail trade data, since they are published at t+30 days. But these data are available over a restricted sample period, which limited the scope of the out-of-sample simulations, and were in any case not found to improve accuracy.

4.2.2 Temporal transformation of indicator variables

When relating the monthly variables $x^m_{hard,t}$ and $x^m_{soft,t}$ to quarterly GDP growth Δy_t, via (1), there is an issue about how these monthly data should be transformed. Rather than *a priori* or *a posteriori* selecting one particular transformation, our approach is to consider simultaneously various transformations of a given indicator variable $x^m_{k,t}$ and then essentially treat these as additional component models (and nowcasts).

The qualitative survey data are considered in both monthly first-differences and quarterly differences. The quarterly transformation of the monthly survey data involves transforming $x^m_{k,t}$ in a manner consistent with the quarterly variable Δy_t (which represents quarterly growth at a quarterly rate). This is achieved, for example, following Mariano & Murasawa (2003). Consider the monthly variable k in levels (rather than log differences) $z^m_{k,t}$, where again the subscript t indicates the particular quarter and m the month within that quarter, $m = 1, 2, 3$, $t = 1, ..., T$. Then, by some simple arithmetic,

$$x_{k,t} = \log z_{k,t} - \log z_{k,t-1} = \frac{1}{3}\Delta \log z^3_{k,t} + \frac{2}{3}\Delta \log z^2_{k,t} + \Delta \log z^1_{k,t} + \frac{2}{3}\Delta \log z^3_{k,t-1} + \frac{1}{3}\Delta \log z^2_{k,t-1} \tag{8}$$

where $\Delta \log z^3_{k,t}$ is monthly growth. The monthly qualitative survey data $x^m_{k,t}$ are also considered in (monthly) levels and quarterly levels, with the quarterly transformation again involving smoothing as in (8) but without application of the logarithmic first difference. Consideration of these four transformations of $x^m_{k,t}$ accommodates uncertainty: (i) about whether the soft data, in levels, are stationary and (ii) whether the informational content of these data is higher when a first or quarterly difference is taken. This means inference does not depend on application of some unit root test. This is attractive, given that unit root tests are well known to suffer from low power in macroeconomic samples. The interest rate spread, which we believe to be stationary, is considered in monthly and quarterly levels only. The hard data, which unlike these survey data we believe to be integrated of order one, are considered in both monthly log first-differences and quarterly differences. Given how we treat $x^m_{hard,t}$, when defining $x_{hard,t-1}$ (e.g. in (2)) we again consider each monthly release separately (for each transformation) rather than aggregate $x^m_{hard,t-1}$, across $m = 1, 2, 3$, to obtain a single lagged quarterly series.

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 $N_6 = 444$.

5 The linear opinion pool

We formalise density combination in a way that extends the commonly-adopted convex mix of point forecasts by utilising the linear opinion pool approach.

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

$$p(\Delta y_\tau) = \sum_{i=1}^{N_j} w_{i,\tau,j}\ g(\Delta y_\tau \mid \Omega_\tau^j), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{9}$$

where N_j $(j = 1, ..., 6)$ where $N_{j+1} > N_j$; $g(\Delta y_{\tau,h} \mid \Omega_\tau^j)$ are the nowcast forecast densities from component model i, $i = 1, \ldots, N_j$ of Δy_τ each conditional on one element (indicator/transformation) from the information set Ω_τ^j. These densities, as we discuss below, are obtained having estimated (1). The non-negative weights, $w_{i,\tau,j}$, in this finite mixture sum to unity.[5] Furthermore, the weights may change with each recursion in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$.

The predictive densities for Δy_τ (with non-informative priors), $g(\Delta y_\tau \mid \Omega_\tau^j)$, allowing for small sample issues, are Student-t; see Zellner (1971). Since each component model, (1), considered produces a forecast density that is t, the combined density defined by equation (9) will be a mixture —accommodating skewness and kurtosis. That is, the combination delivers a more flexible distribution than each of the individual densities from which it was derived. As N_j increases, the combined density becomes more and more flexible, with the potential to approximate non-linear specifications.

We construct the weights $w_{i,\tau,j}$ in two ways.

First, we consider equal weights (EW). The EW strategy attaches equal (prior) weight to each model with no updating of the weights through the recursive analysis: $w_{i,\tau,j} = w_{i,j} = 1/N_j$. We present results for the EW strategy without (prior) truncation of the set of models to be included, although we do experiment below with different groupings of the models. The EW strategy is often recommended when combining point forecasts,

[5]The restriction that each weight is positive could be relaxed; for discussion see Genest & Zidek (1986). Note that in (9) the only unknown parameters to be estimated are the $w_{i,\tau,j}$. The N component densities are taken as given. Somewhat confusingly, in "mixture models" these weights are interpreted on the basis of a latent binary random variable, which is often assumed to have a Markov structure; see Geweke & Amisano (2011) and Mitchell & Wallis (2011). But in these models the parameters of the component models are often estimated simultaneously with $w_{i,\tau,j}$. In so-called BMA for ensemble forecasting models (see Raftery et al. (1995)), the component densities $g(.)$ are centered on the point forecasts from the competing component models, but the variance of the component density forecasts is assumed common across N and estimated simultaneously with $w_{i,\tau,j}$.

although its effectiveness for density forecasts has been questioned (see Jore et al. (2010) and Garratt et al. (2011)).

Secondly, we construct the weights $w_{i,\tau,j}$ based on the fit of the individual model forecast densities: the Recursive Weight (RW) strategy. Following Jore et al. (2010), we use the logarithmic score to measure density fit for each model through the evaluation period. 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.[6] Specifically, the recursive weights for the nowcast densities take the form:

$$w_{i,\tau,j} = \frac{\exp\left[\sum_{\underline{\tau}-8}^{\tau-1} \ln g(\Delta y_\tau \mid \Omega_\tau^j)\right]}{\sum_{i=1}^{N} \exp\left[\sum_{\underline{\tau}-8}^{\tau-1} \ln g(\Delta y_\tau \mid \Omega_\tau^j)\right]}, \qquad \tau = \underline{\tau}, \ldots, \overline{\tau} \tag{10}$$

where the $\underline{\tau}-8$ to $\underline{\tau}$ comprises a two-year training period, since we employ quarterly data, used to initialise the weights. Computation of these weights is feasible even for large N_j. Given the uncertain instabilities problem, the recursive weights should be expected to vary across τ.

From a Bayesian perspective, density combination based on recursive logarithmic score weights has many similarities with an approximate predictive likelihood approach (see Eklund & Karlsson (2007)). Given our definition of density fit, the model densities are combined using Bayes' rule with equal (prior) weight on each model—which a Bayesian would term non-informative priors. Hall & Mitchell (2007) and Geweke & Amisano (2011) consider iterative algorithms to select weights that maximise the logarithmic score. Nevertheless, there are important differences with (predictive) BMA as Geweke & Amisano (2011) explain. When the component models are assumed to constitute an incomplete model space, the conventional Bayesian interpretation of the weights as reflecting the posterior probabilities of the components is inappropriate. We note that instead of looking at fit over the entire density, with a larger out-of-sample window than available in our application the component models could be scored according to their ability to forecast specific probability events of interest.

5.1 Occam's Window: excluding *bad* component models

There is always a question about how one should choose the set of models over which one combines. We start by employing an uninformative prior on all component models and

[6]The logarithmic score of the density forecast, $\ln g(\Delta y_\tau \mid \Omega_\tau^j)$, is the logarithm of the probability density function $g(. \mid \Omega_\tau^j)$, evaluated at the outturn Δy_τ.

use the data (Bayes' rule) to update the weight on each model as evidence accumulates. But we also consider whether there are empirical benefits to excluding some *bad* models, prior to taking the combination.

Madigan & Raftery (1994) propose the use of Occam's Window, under which one averages over a subset of preferred models, treating all the worst fitting models outside this subset as having zero posterior probability. We select the preferred, *better* fitting, models using $w_{i,\tau,j}$ as computed in (10). Specifically, model i is discarded from the combination if it predicts far less well according to the logarithmic score than the best model, i.e. if:

$$\frac{\max\{w_{i,\tau,j}\}_{i=1}^{N}}{w_{i,\tau,j}} > c \tag{11}$$

where c is a constant; following Madigan & Raftery (1994) (and Hoeting et al. (1999)) we set c at 20 in analogy with the 0.05 cutoff used for P-values. In principle, one might think of alternative means of excluding *bad* or uninformative models, e.g. based on subjecting component models to specification tests or dropping those component models where the indicator (on an in-sample basis) is poorly correlated with Δy_t. But here we rely on Occam's window, given both its Bayesian pedigree and its use of an out-of-sample measure of fit (namely $w_{i,\tau,j}$ via (10)), which we might hope offers protection against data mining (or snooping or over-fitting). Having used Occam's window to discard the *bad* models, both equal and recursive weight combinations of the remaining densities are then taken. As an alternative to the combination-based density nowcasts, we also consider the performance of that model which is recursively selected as the best single model, according to $w_{i,\tau,j}$ as estimated in (10).

5.2 Evaluation of nowcast densities

In constructing the combined densities using the linear opinion pool, we evaluate the density forecasts using the logarithmic score at each recursion. These weights provide an indication of whether the support for the component models is similar, or not, based on the score of the individual densities. A finding of similar weights across component models would be consistent with the equal-weight strategy.

A common approach to forecast density evaluation provides statistics suitable for tests of (absolute) forecast accuracy, relative to the "true" but unobserved density. A popular method evaluates using the probability integral transforms (*pits*) of the realisation of

the variable with respect to the forecast densities. See Mitchell & Wallis (2011) for a review. A density forecast can be considered optimal (regardless of the user's loss function) if the model for the density is correctly conditionally calibrated. We gauge calibration by examining whether the *pits* z_τ, where $z_\tau = \int_{-\infty}^{\Delta y_\tau} p(u)du$, are uniform and independently and identically distributed. In practice, therefore, density evaluation with the *pits* requires application of tests for goodness-of-fit and independence at the end of the evaluation period.[7] Mitchell & Wallis (2011) refer to this two component condition as "complete calibration". In the face of alternative goodness-of-fit and independence tests, and in order to build up a robust impression of how well calibrated the densities are, we undertake a battery of tests widely used in the literature.

The eight goodness-of-fit tests employed on the *pits* include, firstly, the Likelihood Ratio (LR) test proposed by Berkowitz (2001). We use a three degrees-of-freedom variant with a test for independence, where under the alternative z_τ follows an AR(1) process. Secondly, and thirdly, we follow Berkowitz (2001) and report a censored LR test which focuses on the 10% top and bottom tails of the forecast densities. Fourthly, we consider the Anderson-Darling (AD) test for uniformity, a modification of the Kolmogorov-Smirnov test, intended to give more weight to the tails. Fifthly, we follow Wallis (2003) and employ a Pearson chi-squared test which divides the range of the z_τ into eight equiprobable classes and tests whether the resulting histogram is uniform. The remaining three tests are for independence of the *pits*; we use a Ljung-Box (LB) test, based on autocorrelation coefficients up to four. To investigate possible higher order dependence we undertake tests in the first, second and third powers of the *pits*.

6 Application nowcasting Euro-Area GDP growth

We compare the accuracy of density nowcasts of Euro-area GDP growth at the six horizons ($j = 1, .., 6$) in recursive out-of-sample experiments using real-time data. The evaluation period is 2003q2-2010q4 (Eurostat published its first Flash estimate for GDP growth for 2003q2). Specifically, we use the real-time data triangles for real GDP and industrial production, for the EA aggregate and the twelve countries, available from Eurostat's real-time (EuroIND) database. The qualitative survey data are not revised (in a significant manner at least). Models are estimated on data vintages back to 2001 with data back to

[7] Given the large number of component densities under consideration, we do not allow for estimation (parameter) uncertainty when evaluating the *pits*. Corradi & Swanson (2006) review *pits* tests computationally feasible for small N_j.

1991q1. Seasonally adjusted data are used. It is important to use real-time data, namely data available at the time rather than the latest release, given data are revised.

The nowcasts are evaluated by defining the 'outturn' as the first (Flash) GDP growth estimate from Eurostat. The exercise could be repeated for different definitions of the outturn, say the second or third QNA release. But as our primary interest is in accelerating delivery of national accounts data, the first estimate does appear to be the natural benchmark.

We break our results into two parts: the RW weights on the soft indicators, the hard indicators and lagged GDP growth derived from the logarithmic score of the component forecast densities; and, the evaluations of the recursive weight, RW, and equal weight, EW, strategies for combination. We also consider strategies that focus only on (equal-weighted) combinations formed using the soft and hard data only.

6.1 Weights on the components

Figure 1 presents the Recursive Weights on the soft indicators (i.e., ESI survey data and the interest rate spread), hard indicators (i.e., IP) and lagged values of GDP growth for the six nowcast horizons, $j = 1, ..., 6$. The interest rate spread, in fact, received little or no weight and henceforth we equate the soft data with the (qualitative) survey data. Note that these weights, on a given type of indicator, say 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; when the aggregate indicators receive a high weight the two lines, for a given indicator, will be close.

We draw out four features from Figure 1. First, comparison of the six panels indicates that the weight on IP increases as j increases; as more hard data become available they get a higher weight in the combination, since the data suggests that their consideration improves (out-of-sample) density fit. In particular, at t-15 days, when the first month of within-quarter IP data become available, the weight on IP increases dramatically, relative to the (soft) survey data. The weight on the IP indicator data further increases, particularly towards the end of the evaluation period when it approaches one, on receipt, at t+15 days, of the second month of within-quarter IP data.

Secondly, the fact that the weights on the component models change over time is consistent with the uncertain instabilities literature referred to above. In particular, during the recession, no doubt thanks to their forward-looking nature, the weight on the soft data dramatically and suddenly increases at t-30, t-15, and t-0 days. Indeed, during the depth of the recession the weight on these soft data became close to unity. In turn, the weight on IP declined rapidly during the recession, but rose as it ended. But when two months of within-quarter IP data are available, i.e. if one is willing to wait until t+15 days, the weight remains high on the IP data even over the recessionary period.

Thirdly, we see that the disaggregate indicators are more informative than the aggregate ones, except at t+45 days when all the weight is given to aggregate indicators. For the three nowcasts produced earlier than t+15 days, prior to receipt of the second month of within quarter IP data, Figure 1 shows that the disaggregate indicators dominate. At t+0 days, Spanish qualitative survey data account for much of the increased weight given to the disaggregate survey data during the recessionary period. But at t+15 and t+30 days, we see that it is the aggregate IP data that account for the majority of the increasing weight given to IP components. From 2009 the utility of the aggregate IP data decreases dramatically in favour of disaggregate IP indicators.

Fourthly, consistent with the stylised fact that it is hard to beat an auto-regressive (AR) model when point forecasting, by deduction from Figure 1 we can infer that the weight on the AR components in the combination is small but non-negligible (since the weights on the soft and hard indicators sum to more than 0.9 but often less than unity). Although, close inspection reveals that the weight on the AR components does decline as within-quarter information accumulates. By t+45 days the AR components receive less than a third of the weight received at t-30 days; on average a weight of about 2.5% rather than 8%. Importantly, we also see the utility of the AR models, as we should expect given that they adjust to change only with a lag, decline during the recessionary period. Once the evaluation period is extended to include the recession the autoregressive nowcasts are clearly beaten by the indicator-based nowcasts, which adapt more quickly to the recessionary "regime". The utility of constructing nowcasts using indicator variables increased over the recessionary period. The weight on that sub-component AR model which constructs the forecast using only the most recent GDP data (the most recent column of data) is quite small, less than 0.01 (or 1%). This indicates that there is informational content in previous, as well as the latest, EA GDP release(s).

There is always an issue about how one should choose the length of the training period to calibrate the weights in (10). There is a trade-off involved. The shorter the length of

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. In (10), and in Figure 1, an increasing window of data is used. Experimentation with rolling windows did not indicate any gains in density fit, as we summarise below. Alternatives, for future research and for datasets where $\overline{\tau}-\underline{\tau}$ spans more than one recession, are to let the weights follow a Markov-switching process such that they vary across regimes (cf. Waggoner & Zha (2012)); or, following the suggestion of a referee, condition the weights on a threshold. For example, when the soft data indicate a high probability of recession (negative GDP growth), the combination weights could be estimated not over all τ, as in (10), but based only on observations from previous recessions.

6.2 Evaluating the nowcast densities

Table 1 summarises the results of the eight *pits* tests, for different j, employed on ten forecasting strategies. These are the EW and RW density combinations, and equal-weighted density averages from those component models which use only survey, IP or lagged GDP (AR) data. EW and RW combinations are also considered when aggregate indicator data only are used, as a means of isolating the relative informational content of the disaggregate indicator data.[8] Finally, AR, Occam's window and model selection densities are examined. To avoid presenting evaluations for each *pits* test separately (see the working paper version of this paper for full details), we follow the pragmatic approach of weighting all eight *pits* tests equally. Table 1 reports how many of these eight tests indicate that the density forecast is correctly calibrated at a 95% significance level—that is, when we cannot reject the null hypothesis that the densities are correctly calibrated on the basis of each individual test.[9] Table 2 presents, again for $j = 1, ..., 6$, the average logarithmic score of each density over the evaluation period.

[8]These EW combinations, taken across all of the component models, effectively give a higher weight to those indicators, in particular the ESI, which are transformed in more ways; e.g. the ESI is considered in quarterly and monthly levels and differences. We found that this implicit weighting is not innocuous. We experimented with EW combinations which group the different indicators (IP, ESI, the interest rate spread and lagged GDP growth) so that, across transformations of a given indicator, the four different indicators have the same weight. Across the board we found that this latter strategy led to less accurate density nowcasts. On average, across j, it led to the average logarithmic score falling by 0.05 units compared with Table 2 below. We eschew formal tests of equal predictive performance given our small-samples.

[9]To control the joint size of the eight evaluation tests, at a 95% significance level, would require 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%. Table 1 can therefore be seen to offer a conservative impression of calibration.

Table 1 shows that, in general across the ten different forecasting strategies, the calibration of the density nowcasts improves as we accumulate within-quarter information. Similarly, Table 2 indicates that the density nowcasts become sharper, and produce a higher log score, as within-quarter information accrues. It is also worth remarking (detailed results are again in the working paper version of this paper) that the story is essentially the same if we look at point forecast accuracy, as measured by root mean squared error. Similarly, we note that experimentation with rolling instead of expanding windows to estimate the weights in RW combinations, via (10), did not deliver clear gains and in some instances led to obvious losses. For example, use of a two-year rolling window led to modest increases in the average log score of "RW Disag" at $j = 1, 2, 3$ (-0.854 to -0.796, -0.804 to -0.796 and -0.791 to 0.789) but more substantial decreases (losses) at $j = 4, 5, 6$ (-0.495 to -0.634, -0.496 to -0.635 and -0.483 to -0.607).

By t+45 days, only one combination strategy (RW) has a correct calibration ratio of 8/8. This is the case whether one considers either the aggregate and disaggregate indicator data jointly or the aggregate data alone. But model selection and the use of Occam's window, even with equal weights, also both deliver calibration ratios of 8/8 at t+45 days and similar log scores to RW (Table 2), precisely because, like RW, they pick up the aggregate IP indicator data, now available for all three months of the quarter. Recall from Figure 1 that at t+45 days these aggregate IP indicators are by far the most informative indicators. Indeed, as a result, by the end of the evaluation period, as the evidence has accumulated, Occam's window in fact eliminates from the combination all but two of the 444 models considered by the RW and EW combinations. This explains the comparable performance of equal and weighted combinations having eliminated *bad* models (indicators/transformations).

In contrast, Table 1 shows that all ten strategies produce poorly calibrated densities according to at least four of the *pits* tests at t-30 days - when no within quarter IP data are available. From t-15 days onwards, when the first month of IP data is published, there is some evidence that calibration improves for the EW and RW combinations which focus exclusively on the aggregate indicators. Indeed the EW combination of aggregate indicators only produces the highest log score at t-15 days. At t+0 days, Table 1 indicates that calibration further improves, with two or three fewer *pits* tests failed for EW and RW combinations using aggregate indicators, relative to Occam's window and selection. Therefore, while later in the quarter the informational content of specific indicators is strong and stable enough for either selection or Occam's window to work, earlier in the quarter one is better off taking a combination, given the additional instabilities. Table

2 shows that selection and Occam's window, relative to the other strategies, do particularly poorly on the basis of the log score until t+15 days. Until t+15 days, the EW combination produces the highest, or amongst the highest, log scores in Table 2. Indeed, it is only marginally beaten by survey-based combinations in one instance at t-30 days. Only with the arrival of the second month of within-quarter IP data is there enough stability (i.e. consensus over the preferred indicator(s); cf. Figure 1) for Occam's window and selection to work. Otherwise, as seen in Figure 1, since the evidence in favour of particular component models changed abruptly over the recessionary period - with the relative informational content of the survey data increasing - one is better off weighting each component model equally. The RW strategy is not as effective as EW, although far preferable to model selection certainly when aggregate indicators only are combined. Nevertheless, the RW combination struggles keep up with the pace of change this early in the quarter. It is only later in the quarter that RW pays off, due to the increased informational content of the IP data.

On receipt of the second month of within-quarter IP data, at t+15 days, we see in Tables 1 and 2 a marked improvement in calibration and density fit, with the RW combination, of aggregate and disaggregate indicators, now having a correct calibration ratio of 8/8. The EW strategy now performs noticeably worse than RW, failing three more pits tests, unless Occam's window is used. This failure is again because EW does not weight highly enough the IP data. The gains associated with the RW combined density at t+15 days are much weaker when only the aggregate indicators are considered, with the average logarithmic score rising from -0.74 to -0.61 rather than -0.79 to -0.50. In fact, looking at the RW weights, it is the latest Spanish IP data that appear particularly helpful in improving density fit. When one considers only the aggregate indicators, the EW combination in fact performs slightly better than the RW combination even at t+15 days. This demonstrates the gains from consideration of the disaggregate IP data.

It does not appear to be worth waiting an additional 15 days for receipt, at t+30 days, of the Portuguese IP data, the Belgian GDP data and the US GDP data. Consideration of these indicators, despite the fact that they are known for the whole quarter, does not improve calibration (Table 1) or density fit (as measured in Table 2). This is consistent with the view that Belgium and Portugal are too small, relative to the EA aggregate, for their data to offer a reliable guide as to likely movements in the aggregate; similarly, while the US economy may influence the European economy at a lag, receipt of the latest US data does not appear to help when nowcasting.

Finally, from Tables 1 and 2 we see that both the survey-based and AR densities do

not improve in accuracy as time passes. Only when produced very early are these densities competitive relative to the other approaches. Even then they fail three or four calibration tests.

6.2.1 Probability of a recession

To evaluate further the accuracy of these density nowcasts we evaluate not the entire density, as above, but the probability forecast of an event of specific interest. We rely on graphical evaluation of these probability event forecasts, rather than formal statistical tests; e.g., see Clements (2004). This is sufficient to illustrate our main findings. Specifically, Figure 2 extracts from the EW and RW combined density nowcasts (using both the aggregate and disaggregate indicators) the implied probability of a (one quarter) recession. In the bottom right panel of the figure we present the outturn, as measured by Eurostat's first (Flash) estimate, for quarterly GDP growth. Given that we have seen that the fit of the densities did not improve at all on receipt, at t+30 days, of the disaggregate hard data for Belgium, Portugal and the US, we do not present the implied probability forecasts at t+30 days; these are identical to those seen at t+15 days in Figure 2.

Figure 2 shows that the EW combination gave about a 10% chance to a recession from 2003 until 2008. The RW combination indicated less than a 10% chance of a recession over this period of sustained economic growth, which seems better. Figure 2 also shows that the RW combination, as we have seen because it increasingly put a high weight on the soft data during the recession, picks up the recession earlier than the EW combination even when the density nowcast is formed at t+0 days or earlier. But the *pits* tests did indicate evidence of calibration failure for these densities as a whole. But from t+15 days, when the RW density does appear to be well-calibrated, we again see the RW combination picking up the recession earlier, and more confidently, than EW.

7 Conclusion

Official GDP data are published with a delay. In order to form a view about the current state of the economy, policymakers therefore rely on a wide range of incomplete data, such as industrial production, which ignore important sectors of the economy, like services, disaggregate (national) rather than aggregate (EA) data, and/or subjective survey data, which tend to be qualitative rather than quantitative. But these data can often point in different directions, particularly at the onset of a recession, and their relative informational

content likely depends on how far ahead of the statistical office's estimate a view about the economy is formed. This paper provides a formal means of assessing the utility of these different indicator variables, and relating them to official Eurostat GDP data. The density combination methods set out make it possible to know how much weight to place on different indicators when forming, at various points in time before and after the end of the quarter of interest as monthly information accumulates, a view about the current state of the economy. The uncertainty associated with these nowcasts is acknowledged, and subsequently evaluated, by constructing density nowcasts.

In a real-time out-of-sample application, density nowcasts for Euro-area quarterly GDP growth are computed as within-quarter monthly "soft" and "hard" indicator data accrues, at both the aggregate and disaggregate (country) level. Alternative temporal transformations of the monthly indicators are considered, delivering a wide range of within-quarter indicators. A linear regression model is then computed for each indicator/transformation and Student-t predictive densities are then combined using the linear opinion pool. Our results suggest that equal-weighted combinations delivered better, but in absolute terms poorly, calibrated densities when limited within-quarter indicator data are available. Equal-weighted combinations were more robust to observed instabilities in terms of the relative importance of "soft" data; the utility of such survey data increased dramatically in the recession, but this was hard to detect in real-time and it therefore pays to weight indicators equally. But, as within-quarter information accumulates, and in particular when the second month of within-quarter industrial production data is published at t+15 days, time-varying weighted combinations are more effective and deliver well-calibrated densities. They do so by giving a higher weight in the combination to the available monthly "hard" data. Equal weighted combinations can perform similarly well at t+15 days onwards, but only if the *bad* models are eliminated prior to taking the combination using Occam's window, so that effectively the available monthly "hard" data are again given a higher weight. 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. Finally, we find that density nowcasts at t+15 days are as accurate as those which involve waiting an additional 15 days for receipt of full-quarter "hard" indicator data from some countries.

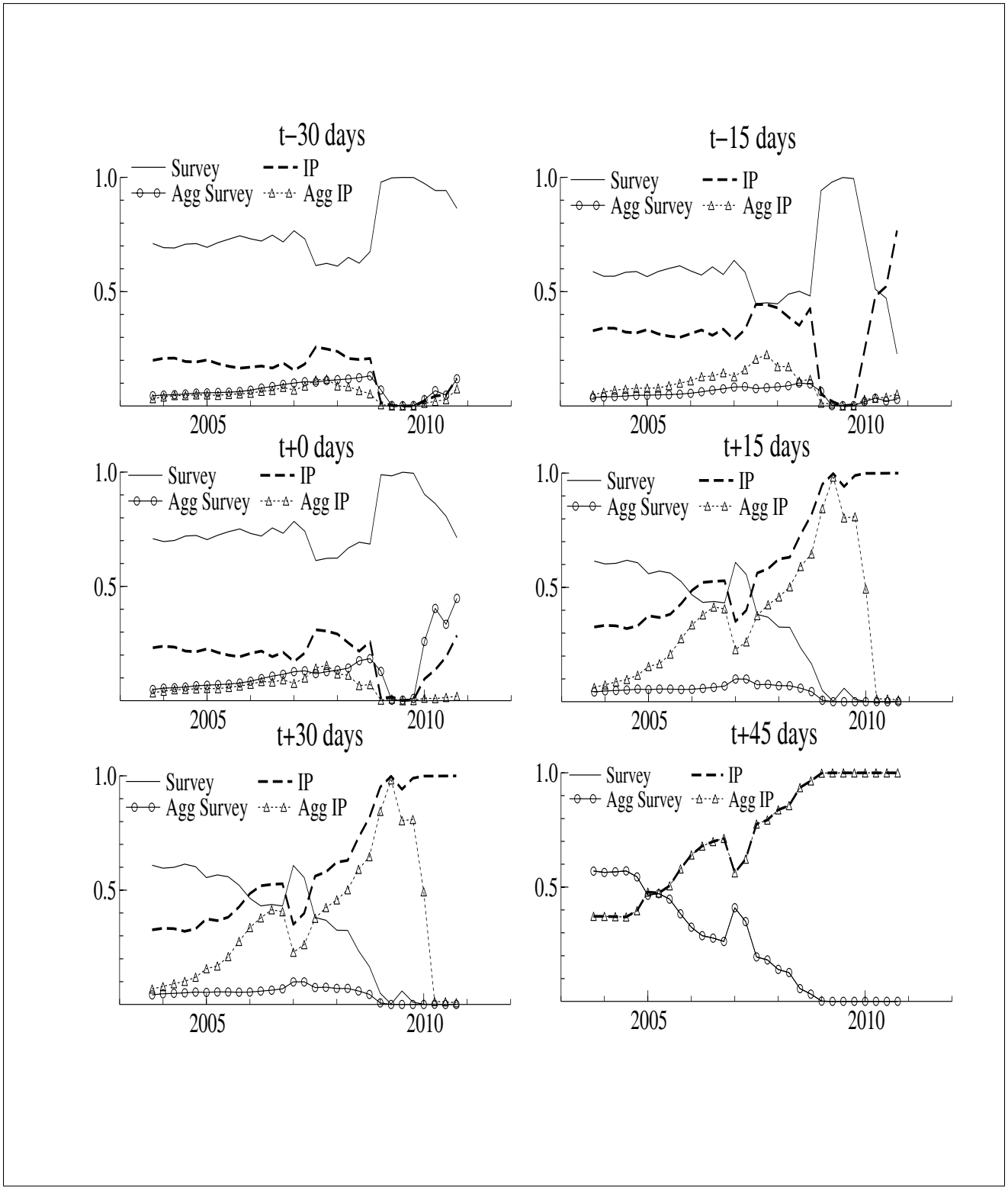

Figure 1: Log-score weights on aggregate/disaggregate soft indicators (survey data) and aggregate/disaggregate hard indicators (IP), and the weights on the aggregate (Agg) indicators only, as within-quarter monthly data accrues ($j = 1, ..., 6$)

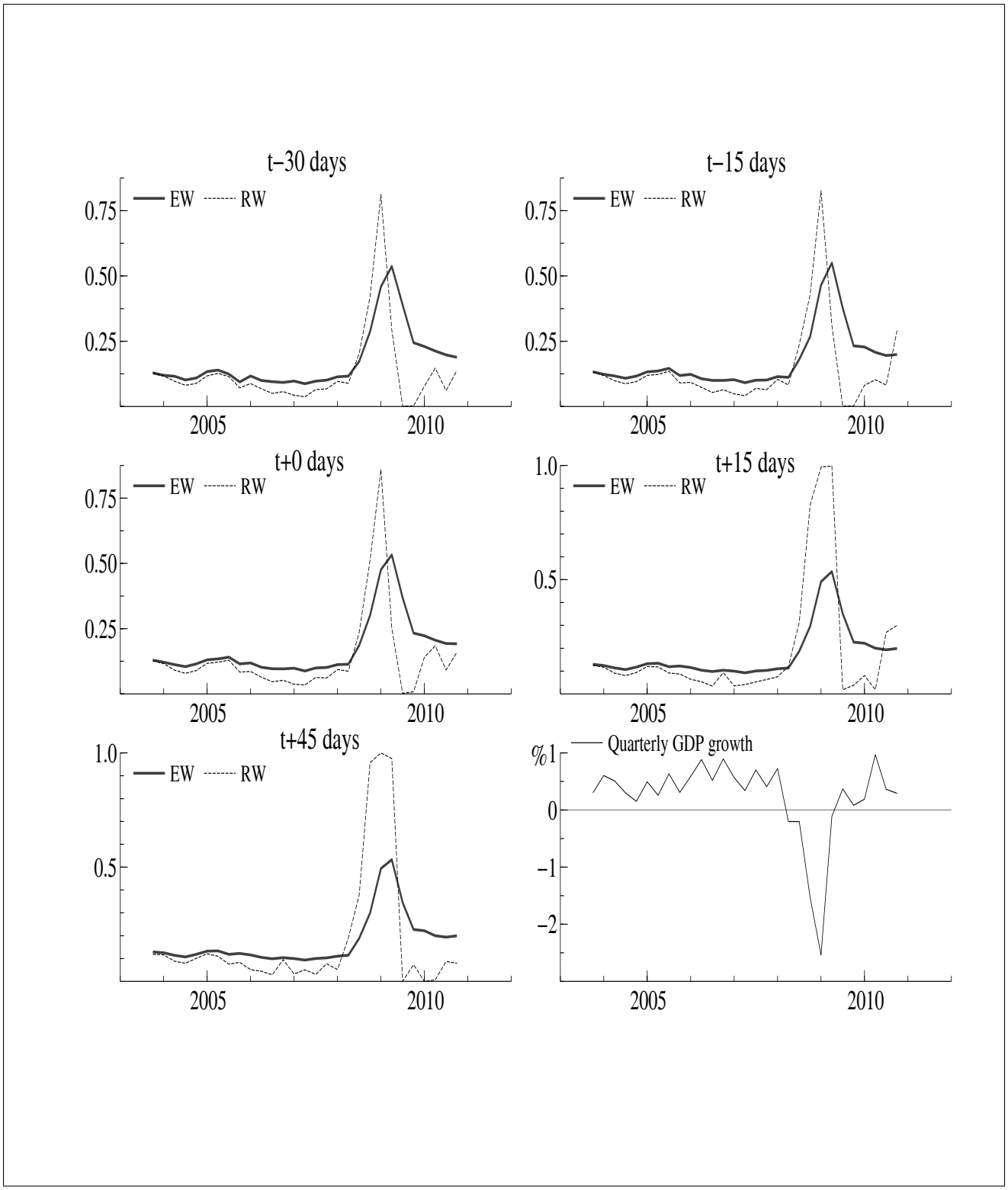

Figure 2: Probability of negative GDP growth according to the EW and RW combination densities

Table 1: 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 $j=5$	t+45 days $j=6$
EW	3	4	3	5	5	4
RW	2	2	3	8	8	8
Survey	4	4	5	5	5	5
IP	4	5	5	5	6	4
EW (Agg)	4	5	6	6	6	6
RW (Agg)	3	6	6	7	7	8
AR	5	4	5	4	5	5
Occam: EW	2	3	2	7	7	7
Occam: RW	2	2	3	8	8	7
Select	3	3	4	8	8	8

Notes: EW is an equal-weighted density combination of all the component models; RW takes a log score weighted average of all of the models; (Agg) denotes combinations of aggregate component models only. Survey is the equal-weighted density combination of those component models that use soft data only; similarly, IP considers the hard indicators only; AR takes equal-weighted density combinations from AR(1) models estimated using all available (multiple vintage) EA GDP growth data. Occam denotes use of Occam's Window. Select is that single model selected according to (10).

Table 2: Average logarithmic score: 2003q2-2010q4

	EW	RW	Survey	IP	EW	RW	AR	Occam		Select
	Disag	Disag			Agg	Agg		EW	RW	
$t-30:j=1$	−0.73	−0.85	−0.70	−0.81	−0.71	−0.82	−0.84	−0.87	−0.90	−1.35
$t-15:j=2$	−0.72	−0.80	−0.71	−0.77	−0.66	−0.82	−0.87	−0.87	−0.84	−1.30
$t+0:j=3$	−0.69	−0.79	−0.70	−0.77	−0.64	−0.74	−0.87	−0.85	−0.86	−0.89
$t+15:j=4$	−0.66	−0.50	−0.70	−0.68	−0.60	−0.61	−0.84	−0.48	−0.46	−0.54
$t+30:j=5$	−0.66	−0.50	−0.70	−0.70	”	”	−0.84	−0.48	−0.46	−0.54
$t+45:j=6$	−0.65	−0.48	−0.70	−0.70	−0.53	−0.46	−0.85	−0.51	−0.46	−0.43

References

Angelini, E., Banbura, M. & Rünstler, G. (2010), 'Estimating and forecasting the euro area monthly national accounts from a dynamic factor model', *OECD Journal: Journal of Business Cycle Measurement and Analysis* **2010**(1).

Bache, I. W., Jore, A. S., Mitchell, J. & Vahey, S. P. (2011), 'Combining VAR and DSGE forecast densities', *Journal of Economic Dynamics and Control* **35**(10), 1659–1670.

Baffigi, A., Golinelli, R. & Parigi, G. (2004), 'Bridge models to forecast the euro area GDP', *International Journal of Forecasting* (20), 447–460.

Berkowitz, J. (2001), 'Testing density forecasts, with applications to risk management', *Journal of Business and Economic Statistics* **19**, 465–474.

Clark, T. E. & McCracken, M. W. (2010), 'Averaging forecasts from VARs with uncertain instabilities', *Journal of Applied Econometrics* **25**, 5–29.

Clements, M. P. (2004), 'Evaluating the Bank of England density forecasts of inflation', *Economic Journal* **114**, 844–866.

Clements, M. P. & Galvão, A. (2008), 'Macroeconomic forecasting with mixed-frequency data', *Journal of Business and Economic Statistics* **26**, 546–554.

Corradi, V. & Swanson, N. R. (2006), Predictive density evaluation, *in* G. Elliott, C. W. J. Granger & A. Timmermann, eds, 'Handbook of Economic Forecasting', North-Holland, North Holland, pp. 197–284.

Eklund, J. & Karlsson, S. (2007), 'Forecast combination and model averaging using predictive measures', *Econometric Reviews* **26**(2-4), 329–363.

Garratt, A., Mitchell, J., Vahey, S. P. & Wakerly, E. C. (2011), 'Real-time inflation forecast densities from ensemble Phillips curves', *North American Journal of Economics and Finance* **22**, 77–87.

Genest, C. & Zidek, J. (1986), 'Combining probability distributions: a critique and an annotated bibliography', *Statistical Science* **1**, 114–135.

Geweke, J. & Amisano, G. (2011), 'Optimal prediction pools', *Journal of Econometrics* **164**, 130–141.

Giannone, D., Reichlin, L. & Small, D. (2008), 'Nowcasting: The real time informational content of macroeconomic data releases', *Journal of Monetary Economics* **55**, 66576.

Granger, C. W. J. & Pesaran, M. H. (2000), 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting* **19**, 537–560.

Hall, S. G. & Mitchell, J. (2007), 'Combining density forecasts', *International Journal of Forecasting* **23**, 1–13.

Hendry, D. F. & Hubrich, K. (2011), 'Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate', *Journal of Business and Economic Statistics* **29**, 216–227.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999), 'Bayesian model averaging: a tutorial', *Statistical Science* **14**, 382–417.

Jore, A. S., Mitchell, J. & Vahey, S. P. (2010), 'Combining forecast densities from VARs with uncertain instabilities', *Journal of Applied Econometrics* **25**, 621–634.

Kascha, C. & Ravazzolo, F. (2010), 'Combining inflation density forecasts', *Journal of Forecasting* **29**(1-2), 231–250.

Kitchen, J. & Monaco, R. (2003), 'Real-time forecasting in practice: The U.S. Treasury staff real-time GDP forecast system', *Business Economics* **38**, 1019.

Madigan, D. M. & Raftery, A. E. (1994), 'Model selection and accounting for model uncertainty in graphical models using Occam's window', *Journal of the American Statistical Association* **89**, 1335–1346.

Mariano, R. S. & Murasawa, Y. (2003), 'A new coincident index of business cycles based on monthly and quarterly series', *Journal of Applied Econometrics* **18**(4), 427–443.

Mitchell, J. & Hall, S. G. (2005), 'Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR "fan" charts of inflation', *Oxford Bulletin of Economics and Statistics* **67**, 995–1033.

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

Mitchell, J. & Wallis, K. F. (2011), 'Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness', *Journal of Applied Econometrics* **26**(6), 1023–1040.

Raftery, A. E., Gneiting, T., Balabdaoui, F. & Polakowski, M. (1995), 'Using Bayesian Model Averaging to Calibrate Forecast Ensembles', *Monthly Weather Review* **133**, 1155–1174.

Stock, J. H. & Watson, M. W. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**, 405–430.

Waggoner, D. F. & Zha, T. (2012), Confronting model misspecification in macroeconomics, NBER Working Papers 17791.

Wallis, K. F. (2003), 'Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts', *International Journal of Forecasting* (19), 165–175.

Wallis, K. F. (2005), 'Combining density and interval forecasts: a modest proposal', *Oxford Bulletin of Economics and Statistics* **67**, 983–994.

Wallis, K. F. (2011), 'Combining forecasts - forty years later', *Applied Financial Economics* **21**, 33–41.

Zellner, A. (1971), *An introduction to Bayesian inference in econometrics*, New York: John Wiley and Sons.

Zellner, A. (1986), 'Biased predictors, rationality and the evaluation of forecasts', *Economics Letters* **21**, 45–48.