June 2012

Combining expert forecasts: Can anything beat the simple average?[1]

V. Genre[a], G. Kenny[a,*], A. Meyler[a] and A. Timmermann[b]

Abstract

This paper explores the gains from combining surveyed expert forecasts using the ECB Survey of Professional Forecasters (SPF). The analysis encompasses combinations based on principal components and trimmed means, performance-based weighting, least squares estimates of optimal weights as well as Bayesian shrinkage. For GDP growth and the unemployment rate, only few of the individual forecast combination schemes outperform the simple equally weighted average forecast in a pseudo-out-of-sample analysis, while for the inflation rate there is stronger evidence of improvement over this benchmark. Nonetheless, when we account for the effect of multiple model comparisons through White's reality check, the results caution against any assumption that the identified improvements would persist in the future.

Keywords: forecast combination, forecast evaluation, multiple model comparisons, real-time data, Survey of Professional Forecasters

JEL: C22, C53

[a] European Central Bank, Kaiserstrasse 29, D-60311 Frankfurt, Germany

[b] Rady School of Management and Department of Economics, University of California, San Diego, USA.

[*] Corresponding author: DG Research, European Central Bank, Kaiserstrasse 29, D-60311 Frankfurt, Germany. Tel: +49 69 1344 6416. Fax: +49 69 1344 6575. Email: geoff.kenny@ecb.europa.eu

[1] The authors would like to thank the editor, Michael Clements, along with two anonymous referees for many helpful comments. We also thank Tommy Kostka for providing excellent research support with the SPF data. In addition the paper has benefited from comments and discussions with participants in seminars at the ECB, the Central Bank of Ireland, the Eurosystem's Working Group on Forecasting and Working Group on Econometric Modelling, the 31st International Symposium on Forecasting in Prague, June 2011, as well as from an anonymous referee from the ECB Working Paper Series. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the ECB or the Eurosystem. Any errors are the sole responsibility of the authors.

1. Introduction

Coinciding with the launch of the euro currency in January 1999, the European Central Bank (ECB) started a Survey of Professional Forecasters (SPF) as part of its information gathering and analysis of the euro area macroeconomic outlook. Since its inception, the forecast data collected in the SPF has normally been summarised by way of a simple average of the surveyed forecasts. Although a large literature exists on optimal forecast combination – see Timmermann (2006), Newbold and Harvey (2002) and Clemen (1989) – such an approach was reasonable given the lack of any available track record among SPF panel members at that time. Moreover, empirical studies have shown that such a simple equally weighted pooling of forecasts performs relatively well in practice compared with other approaches that rely on estimated combination weights – a phenomenon dubbed the "forecast combination puzzle".

This paper explores different combinations of the SPF forecasts with a view to optimising the quality of SPF information that is made available to decision makers and the public. Forecast combination seeks to reduce the information in a vector of forecasts to a single summary or combined forecast using weights chosen to minimize the expected loss. Our analysis encompasses a variety of methods that have been proposed in the literature including statistical combinations based on principal components analysis and trimmed means, performance-based weighting, optimal weighting as well as Bayesian shrinkage. We also employ statistical techniques and construct sub-groups of forecasters to deal with the relatively large cross sectional dimension of the SPF dataset and reduce the effect of estimation error. Given that we test a large number of combination methods on a single historical dataset, we employ the White (2000) "reality check" to deal with the multiple comparison problem. Also, given the significant revisions to euro area macroeconomic variables over our sample period, we examine the sensitivity of our results to the chosen vintage of the outcome for the forecast target variable. Finally, we consider the sensitivity of the results to the period of exceptional macroeconomic volatility following the financial crisis of 2008-2009.

Over the sample period analysed, we find that the equal-weighted combination sets a high benchmark that performs well relative to other forecasts. Notwithstanding the relatively good performance of the SPF benchmarks, a number of different combination strategies achieve gains relative to this benchmark. Looking across variables, the scope for improvements from more sophisticated combination strategies appears the most significant for inflation with smaller gains achievable for the unemployment rate and GDP growth. Overall, however, we would not conclude that there exists a strong case to consider combinations other than equal weighting as a means of better summarising the information collected as part of the regular quarterly rounds of the ECB SPF. The variation in the best performing specification through time, across target variables and across horizons together with the likely role of chance in explaining the success of some models in our sample highlights the inherent difficulty to successfully pick out in real time a preferred or best combination method.

The remainder of the paper is organised as follows. Section 2 reviews the different combination methods and explains how they are applied to the ECB SPF. Section 3 provides background information on the SPF and key features of the associated dataset, focusing on the cross-sectional information that is available and some practical issues (such as the frequent non-responses of some individual forecasters to the survey) that need to be overcome when implementing several combination methods. Section 4 presents performance evaluation measures and out-of-sample evaluation results for inflation, GDP growth and the unemployment rate. Finally, Section 5 concludes with a summary of our main findings.

2. Forecast Combination Methods

This section briefly introduces the main approaches we apply for the estimation of combination weights and the various benchmarks against which they are evaluated. Let $\hat{y}_{i,t+h}$ be the *i*'th survey participant's forecast of the outcome in period *t+h*, based on the forecaster's information at time *t*. Forecast combination aims to reduce the information in a vector of N forecasts ($\hat{y}_{i,t+h}$, *i* = 1, . . . , N) to a single summary or combined forecast $\hat{y}^{c}_{t+h}(\hat{y}_{i,t+h}, w)$ where *w* represents the vector of combination weights, $w_{i,t+h}$, i = 1, …, N. The optimal combination chooses $w_{i,t+h}$, such that the conditional expected loss of the errors of the combined forecast is minimised. If the forecast is linear in the combination weights and the loss function is of the Mean Squared Error (MSE) type, combination weights will depend on the first two moments of the joint distribution of the vector of forecasts and the outcome and can be estimated by linear projection of the individual forecasts on the target variable (Granger and Ramanathan (1984)). In this set-up, the equally weighted combination is optimal in population only under strongly restrictive assumptions. If these fail to hold, there are potentially significant opportunities to better exploit the information content of the SPF by estimating weights in order to combine the individual forecasts (see, e.g., Aiolfi et al. (2011), Stock and Watson (2004), Hendry and Clements (2004) and Diebold and Pauly (1987) and Clemen and Winkler (1986)).

In practice, however, there are important limits to the gains from attempting to combine forecasts optimally. Smith and Wallis (2009), highlight estimation error as a plausible explanation for the "forecast combination puzzle" by which more simple combination schemes – such as the equally weighted combination – often perform best in practice. Such error may be particularly important in situations where the number of individual forecasts is large relative to the number of time series observations. This is clearly the case for the ECB SPF, which is a quarterly survey launched only in 1999 with approximately 90 participants from across the EU. Such estimation error reflects the dependence of the optimal weights on the full conditional covariance matrix of forecasts which – when the number of forecasts is high – entails a large number of unknown parameters.

The remainder of this section outlines different approaches to combining forecasts and how we have adapted them for use with the ECB SPF. We restrict ourselves to the class of linear combinations and focus on those methods which emphasise parsimony with a view to minimising as much as possible estimation error.

Trimming and other statistical combinations: These include the median and other trimmed mean measures which remove extreme values from the cross-section of forecasts, assigning zero weight to some forecasts and equal weights to all others.

Performance-based combinations: These assign higher weights to forecasts with a relatively good forecasting track record and lower weights to forecasts with a poor performance (see Bates and Granger (1969)). Stock and Watson (2004) propose a weighting scheme based on the individual forecast's past performance:

$$w_{it} = \frac{m_{it}^{-1}}{\sum_{j=1}^{N} m_{jt}^{-1}} \quad \text{, where } m_{it} = \sum_{s=T_0}^{t-h} \delta^{t-h-s} (y_{s+h} - \hat{y}_{i,s+h})^2 \qquad (1)$$

Here δ is a discount factor and m_{it} is the cumulative sum of past discounted forecast errors computed since the start of the sample (T_0). Values of δ below unity assign higher weight to more recent forecast errors in the calculation of the combination weights and we set δ = 1, 0.95, and 0.85. We also consider rolling performance with no discounting applied based on a window of length v to account for possible time variation in relative forecast performance. Performance is assessed for v = 1, 4 and 8 quarters. As a special case of this, the recent best method assigns all weight to the individual forecaster with the lowest MSE over a window of v = 1 and v = 4 quarters.

Principal components combination: Following Stock and Watson (2004), these combinations use principal components analysis to estimate the static common factors from the panel of forecasts and regress a subset of these on the target variable. We consider up to three principal components, (labelled p = 1, 2 and 3) and use a rolling window with 20 quarterly observations for estimation, controlling for publication lags in the regression to preserve the "real time" character of the resulting combination.[2]

Least squares (optimal) combination weights: Following Granger and Ramanathan (1984) these use least squares regression to estimate the optimal combination weights:

$$y^c_{t+h} = w_{0,h} + \sum_{i=1}^{N} w_{i,h}\hat{y}_{i,t+h} + \varepsilon_{t+h} \qquad (2)$$

We consider various restricted versions of (2) that either omit the constant, constrain the estimated weights to sum to unity or impose a convexity constraint to rule out negative weights and weights greater than unity, i.e. $0 \leq w_{i,h} \leq 1.0 \quad \forall\, i = 1,\ldots,N$.[3] In practice, it is not feasible to apply (2) to combine the individual SPF forecasts given the large cross-sectional dimension and the relatively small time series available. We therefore follow the k-mean clustering approach of Aiolfi and Timmermann (2006) and replace the N individual forecasts in equation (2) with the mean forecast computed for a small number of clusters. We consider two or three clusters (c= 2 or c = 3 and base the clustering on the most recently observed squared forecast error, thus yielding groups of either High/Low or High/Medium/Low performing forecasters. Using these forecasts, the weights are estimated from a rolling window with 20 quarterly observations.

Projection on the mean: Following Capistrán and Timmermann (2009), this method uses linear projection of the target variable on the equally weighted forecast $\bar{y}_{t+h}$:

$$y^c_{t+h} = w_{0,h} + \overline{w}\,\bar{y}_{t+h} + \varepsilon_{t+h} \qquad (3)$$

[2] For a survey conducted in month (quarter) t, an estimate of the outcome for inflation and the unemployment rate is available for month t-1 and t-2 respectively while the outcome for GDP growth is available only for quarter t-2.

[3] The convexity constraints are implemented using non-linear least squares. In principle, such a non-negativity constraint may be sub-optimal. However, in practice, such constraints may help improve forecast performance by limiting the impact of estimation error on the combined forecast.

Here $\overline{w}$ is the estimated slope parameter in the combination regression. Equation (3) can again be estimated either with or without the bias adjustment parameter ($w_{0,h}$), using the rolling pseudo real time estimation procedure described above.

Shrinkage (Bayesian) Combinations: These shrink the least squares weights toward a prior of equal weights, giving the resulting combination a Bayesian interpretation (Diebold and Pauly (1990)). As implemented in Stock and Watson (2004), our shrinkage weights take the form

$$w_{i,h} = \psi\, \hat{w}_{i,h} + (1-\psi)\,(1/N) \tag{4}$$

where $\psi = \max\,(0\,, 1-\kappa\, N/\,(T-h-N-1)$. The parameter κ governs the amount of shrinkage and T denotes the number of observations used in the least squares regressions. We estimate the shrinkage combination using the same clusters employed in the least squares combinations and vary the intensity of the shrinkage parameter between κ =4 and κ =6.[4]

3. The ECB SPF

This section provides an overview of the ECB SPF, focusing on the key features of the survey and its associated dataset. We highlight the extent of missing observations in the panel and, given that a number of combination methods require a panel without missing observations, we also present a simple approach to create a fully balanced panel.

3.1 Key features

The SPF forecasts are described in detail in previous studies such as Bowles et al. (2007) and Garcia (2003). The complete dataset can be downloaded directly at http://www.ecb.europa.eu/stats/prices/indic/forecast/html/index.en.html. The survey was launched in the first quarter of 1999 and has since been conducted on a quarterly basis with the main results communicated to policy makers (the ECB Governing Council) and published regularly in the ECB Monthly Bulletin. The aim of the survey is to provide forward looking information on inflation expectations as measured by the expected change in the Harmonised Index of Consumer Prices (HICP) for the euro area, though forecasts for GDP growth and the unemployment rate have also been collected. The underlying panel is comprised of macroeconomic experts with a strong forecasting

[4] The prior mean of the bias adjustment parameter is set equal to zero. The choice of shrinkage parameter allows the weight on the prior mean to vary between 25% and 75% depending on the number of observations and clusters used.

record. Although the survey questionnaire related to macroeconomic aggregates in the euro area as a whole, the forecasters were drawn from across the European Union (EU).[5]

An important feature of the ECB SPF is the definition and transformations of the predicted variables. For both the one-year and two-year horizons, these refer to the annual changes in the level of GDP in quarter *t+h* compared with quarter *t+h-4* and in the level of the HICP in month *t+h* compared with month *t+h-12* and to the *level* of the unemployment expressed as a percentage of the euro area labour force in month *t+h*. Another issue worth highlighting in the context of the SPF is that forecasters are asked to forecast each variable one-year and two-years ahead of the latest available outcome. Given that the forecast rounds have also been scheduled to coincide with the previous month's HICP release (i.e. Q1 rounds take place mid-Jan., Q2 in mid-Apr., Q3 in mid-Jul. and Q4 in mid-Oct.), this means that the 'one-year ahead' forecast is actually around six-to-eight months ahead for GDP growth, eleven months ahead for the unemployment rate and twelve months ahead for HICP inflation.[6] We denote these rolling horizons as H = 1 and H = 2 years, respectively.

Figure 1 plots the equally weighted mean SPF forecasts for the three variables and two horizons. The first vintage outcomes are shown for each variable together with the forecast errors for the equal-weighted forecast calculated using the first vintage. The Figure highlights the relatively sizeable and often persistent forecast errors from the SPF; the errors are particularly sizeable for the quarters starting in 2008Q3 reflecting the impact of the 2008-2009 financial crisis. The inflation and GDP forecast errors show a one-side pattern while those for the unemployment rate are more two-sided.[7] This graphical presentation also highlights some difference between the one-year ahead and the two-year ahead forecasts, in the sense that the latter have been much smoother and, hence, less correlated with the outcome.

Table 1 provides further summary information on the forecast performance of the SPF over the period since it was launched, reporting the mean errors and Root MSE

[5] Bowles et al. (2007) examine the forecasting performance of forecasters headquartered inside and outside the euro area and report no noticeable difference.

[6] For example, for the one-year ahead forecast in the Q1 round, which takes place in mid-January, forecasters are asked to forecast annual HICP inflation in December, year-on-year GDP growth in Q3 and the unemployment rate in November in that year. As this paper is a forecast evaluation paper we generally refer to the target period rather than the round when the survey is actually conducted.

[7] For a more comprehensive analysis of the properties of the SPF forecasts for growth and the unemployment rate see Bowles et al. (2010).

(RMSE) over different samples. With the exception of the inflation forecasts, performance as measured by the RMSE deteriorates with an increase in the length of the horizon. The recent financial crisis also impacted forecast accuracy negatively. For example, the average RMSE on the one-year ahead SPF forecasts for GDP rose from 0.9 when calculated over the period prior to the crisis to 1.4 when calculated over the full sample. A similar deterioration in forecast accuracy is observed for the unemployment rate, while the crisis impacted less the accuracy of the inflation forecasts.

To provide some information on the heterogeneity embedded in the SPF panel, Table 1 also reports the minimum and maximum values for the above summary statistics taken from the individual level data. The minimum and maximum ranges highlight significant heterogeneity in forecasting performance. For example, over the full sample, the best performing GDP forecaster has a RMSE 0.4 percentage points below the equivalent measure for the worst performing forecaster. Similar heterogeneity in forecast performance at the individual level is observed for unemployment rate and inflation forecasts. It is precisely this heterogeneity in forecast performance that the combination methods we employ seek to exploit.

Another feature of the data evident from Figure 1 and Table 1 is the presence of possible bias in SPF forecasts. In the case of GDP, the bias has tended to be negative (i.e. as defined here the forecasted level has tended to be above the actual outcome). In the case of the one-year and two-year ahead unemployment rate forecasts the average bias has tended to be quite small and also alternating in sign depending on whether one-year or two-year ahead forecasts are considered. In the case of inflation, there is also evidence of positive bias, i.e. on average the forecasted level for inflation has tended to be below the outcome. Overall the heterogeneity across forecasters and apparent bias suggest that, for some variables and some horizons, combination methods, particularly those which allow for bias adjustment, could yield superior out-of-sample performance relative to the equally weighted combination (which does not adjust for bias).[8]

3.2 Filtering and balancing the panel

A major practical challenge that arises in forecast surveys is the frequent and extensive gaps in the panel of forecasts reflecting the non-responses by some participants, the

[8] Most combinations we apply incorporate some form of bias adjustment via the constant in the combination regressions. A relatively large bias in individual forecasts will also tend to result in a relatively low weight when using the performance based weighting schemes discussed in Section 2.

introduction of a new panel member or the dropping out of an existing participant. Figure 2 provides an illustration of the extent of the unbalanced nature of the panel for the case of GDP growth (H=1). In the raw ECB SPF data, as illustrated in Figure 2, the gaps in the dataset are considerable. Most of the 'entry' and 'exit' actually reflects non-responses and are as such determined by the respondents rather than by those conducting the survey.[9] Such a clearly unbalanced panel prohibits any meaningful comparison of unfiltered individual forecast performance. For example, some forecasters could perform poorly (or reasonably well) simply because they contributed to the survey during a period when the target variable exhibited above (below) average volatility. In the case of GDP, the cohort of forecasters who entered the panel only in 2007 and 2008 performed particularly poorly reflecting the exceptionally high volatility in the macroeconomic environment around this time.

To reduce any sampling distortions associated with frequent non-responses and missing observations in the raw SPF dataset, we have filtered the data so as to include only those forecasters who have been contributing relatively frequently. The filter is such that forecasters with more than four consecutive missing observations are excluded from the panel. For example, those forecasters entering in 2007 and 2008 are not included in the filtered panel given that they would not satisfy the requirement of no more than four consecutive missing values over the sample period. For GDP growth (H=1), this filtering reduces the number of forecasters from 94 in the unfiltered dataset to 31 in the filtered panel. Although this reduction may imply some loss of information, it has the advantage that the surviving forecasters can then be compared on a reasonably consistent basis given the much smaller incidence of missing data. The impact of the filtering on the mean forecast is generally trivial.

Even after filtering our irregular respondents, the data involve several missing values and gaps. To fill out these gaps, we use a simple approach that focuses on the dynamics of relative individual forecasts using a panel regression of the form:

$$\hat{y}_{i,t+h} - \bar{y}_{t+h} = \beta_i(\hat{y}_{i,t+h-1} - \bar{y}_{t+h-1}) + \varepsilon_{i,t+h} \qquad (5)$$

Equation (5) posits a simple AR(1) process, whereby the relative deviation of each forecaster from the simple average in period *t* is linked to its relative deviation in period *t-1.* If $\beta_\text{i} = \beta = 1.0$, missing observations for individual forecasts are simply set to the

[9] For instance, the participation rate tends to be lower in the Q3 (July) round, presumably reflecting holiday absences. However, Bowles et al. (2010, p. 86) report "no noteworthy impact of participation rates, either in terms of an "improvement or worsening" in aggregate forecast performance.

previously reported individual forecast updated with the change in the average of those forecasters who do respond. For $0 \leq \beta \leq 1.0$, the missing values for forecaster i in period t are replaced with the period t average forecast plus a fraction of the previously observed deviation from the average forecast. β can be estimated recursively over the sample period to ensure that the method used to balance the panel preserves the pseudo real time nature of the resulting dataset.

Figure 1 reports the equal-weighted average SPF forecast for each variable and horizon, using both the balanced and the unbalanced panel. Although the two series are virtually identical over the sample period, we use the headline SPF indicator, based on the unbalanced and unfiltered panel as the primary benchmark in our subsequent evaluation. Importantly, this equal-weighted combination remains the headline SPF indicator that is reported and published in the ECB monthly bulletin. At the same time, the methods we have proposed may be of analytical use, e.g. to construct a constant sample update of the SPF in order to check for compositional effects and also to estimate combinations which require a balanced panel.

3.3. Real-time data issues

A key practical complication that arises in forecast combination and evaluation relates to the impact of data revisions. Survey forecasts are by definition "real time" in the sense that they cannot use information that was unavailable at the time the survey was carried out and combinations of such forecasts also possess a corresponding real-time dimension. However, data revisions alter the estimate of the outcome for the forecast target variable and the evaluation of different combinations may be sensitive to the choice of outcome vintage. For our baseline results, we use the first estimate for each variable in deriving forecast performance statistics. However, to the extent that measurement error (or "noise") partly accounts for subsequent data revisions, they may have a predictable component which would suggest a possible preference to focus on the revised vintages of data in the evaluation. Given this alternative hypothesis, we report our forecast evaluation results also using the current vintage of estimates and check the sensitivity of the performance of different combinations to this choice.

To get a sense of the role of such revisions for the evaluation of SPF combinations, Figure 3 plots the difference between the first estimate provided by Eurostat (the European Statistical Agency) for each of the three SPF variables and the corresponding "current" estimates available in December 2011.[10] Substantial revisions in euro area data are apparent for both GDP growth and the unemployment rate. Compared with initial published results, the more recent estimates of euro area GDP growth have been revised upward substantially over most of the period since 1999.[11] A notable exception was the recessionary period in 2008 and 2009 where the first estimate was revised down. Considerable revisions are also evident for the unemployment rate (with downward revisions in the first half of the sample being followed by significant upward revisions). In the case of inflation, there have been more limited revisions and mainly in the early years of the sample.

4. Measuring Forecast Performance and Results

Our evaluation is presented in the form of the MSE of the different SPF combinations ($\hat{y}_{c,t+h}$) *relative* to the benchmark equal-weighted combination, $\bar{y}_{t+h}$. Assuming our "holdout" sample (used for the out-of sample evaluation) runs from period T$_1$ to period T, our performance evaluation measure is given by

$$\text{(Relative) MSE} = \sum_{t=T_1}^{T}[y_{t+h} - \hat{y}^{c}_{t+h}]^2 \Big/ \sum_{t=T_1}^{T}[y_{t+h} - \bar{y}_{t+h}]^2 \qquad (6)$$

This is less than unity whenever a given combination performs better than the simple equal-weighted combination. To help gauge the overall performance of the equal-weighted benchmark, we also report the Relative MSE for three simple time series models. In particular, we consider a naïve forecast which sets the projected level of the target variable equal to its current level as known at the time of the survey and allowing for publication lags.[12] We also estimate a random walk with drift for the seasonally adjusted *level* of GDP and the *level* of the consumer price index (HICP).[13] Lastly, to capture persistence in the dynamics of the three variables, we also estimate an AR(1) process for the log change in GDP and HICP and for the level of the unemployment

[10] Our real time data is fully consistent with the estimates in the real time database of the Euro Area Business Cycle Network as described in Giannone et al. (2010a).

[11] The importance of such data revisions for the euro area is similar to the evidence for the US which is reviewed in Croushore and Stark (2003).

[12] Publication lags imply that the known current level for each forecast variable is approximately lagging the survey month by 1 month in the case of inflation, by 2 months in the case of the unemployment rate and by 2 quarters in the case of the annual GDP growth. This information on the level of each forecast variable that is known at the time of the survey is provided to ECB SPF participants when they receive the survey questionnaire.

[13] Given that it is not a clearly trending variable like GDP and the HICP, a random walk without drift would seem more appropriate for the level of the unemployment rate. This is, however, equivalent to the naïve forecast for the level of the unemployment rate.

rate. While these time series benchmarks are quite simple, they have proven to be quite difficult to beat in practice - particularly at horizons beyond one-quarter.

Given that we are evaluating a large set of combinations repeatedly using the same historical dataset, chance alone may be able to explain a statistically significant result for any given combination. As a result, inference based on the use of multiple pair-wise comparisons of predictive ability is not appropriate in this context. We therefore report the White (2000) reality check. The reality check tests the null hypothesis that the expected performance of the best performing model is no better than that of the benchmark model. The test provides useful information as to whether or not the identification of some improvement compared with the equal-weighted combination is genuine and thus likely to persist over time. Denoting f_j as a measure of the out-of-sample forecasting performance of the j^{th} combination (j = 1, . . . , J) relative to the benchmark, e.g. $\mathrm{MSE_j} - \mathrm{MSE_0}$, where the benchmark (represented by model zero) is the equal-weighted combination, the White Reality Check (RC) test can be applied to the performance statistic:

$$\text{White RC} = \underset{j}{Min}\left\{T^{\frac{1}{2}}\bar{f}_j\right\}, \text{ where } \bar{f}_j = T^{-1}\sum_{t=1}^{T} f_{t,j} \qquad (7)$$

Here $\bar{f}_j$ is the sample mean of the forecasting performance of model j measured relative to that of the benchmark, While a closed form solution for the distribution of the minimum in (7) is not available, it can be approximated using a bootstrap sampling procedure and the relevant P-values can then be reported for the null hypothesis that the expected performance of the best performing combination is no better than that of the equal-weighted combination.

A limitation of the reality check procedure is that it assesses the performance of the best combination scheme jointly with the performance of a large cross-section of competing specifications which may reduce its power. To deal with this issue, we propose an alternative approach which attempts to mimic the situation confronting a decision maker seeking to choose *in real time* among different combinations. Subject to having a minimum initial track record, at each point in time we choose that combination strategy which over the most recent four quarters generated the smallest MSE. The identity of this model may change through time so we are effectively referring to the forecasting performance of the combination selection or ‘search’ rule. Such a recent best method is not subject to the multiple hypothesis testing criticism highlighted above since it effectively only uses one combination at each point in time.

4.1 Comparison of SPF with other simple benchmarks

Table 2 reports the out-of-sample Relative MSEs for GDP growth, HICP inflation and the unemployment rate. The MSEs are calculated over "normal" business cycle conditions and excluding the very large macroeconomic shocks associated with the period since the end of 2008. To provide graphical insight into the performance of the estimated combinations and their evolution over time, Figure 4 plots for each variable and horizon, the best performing combination (which changes depending on whether we include the crisis or not) as well as a shaded region representing the range of forecast values encompassed by all the estimated combinations.[14]

One notable feature of the results is the relatively good performance of the (equally weighted) SPF forecasts for the real variables (GDP and unemployment) relative to simple time series models (Random Walk, Naïve or AR(1)). However, naïve time series predictors for inflation tend to outperform the SPF average at both horizons. These results are consistent with previous studies of the SPF (e.g. Bowles et al. (2007)) and euro area inflation forecasting (e.g. Benalal et al. (2004) and Giannone et al. (2010b)). In the case of inflation, the relatively poor performance of the equal-weighted SPF forecast certainly motivates the case for examining the extent to which other combinations might yield some predictive gain. There are no notable gains from either trimming or focusing on the median SPF forecast for inflation, suggesting little evidence of "noisy" inflation forecasters. However, for GDP growth and particularly for the unemployment rate at short horizons (H=1), some trimming of forecasters results in a noticeable improvement in relative performance.

4.2 Relative performance of different combination methods

To get a sense of the relative performance of the various combination methods, panel (a) of Table 3 summarises the detailed results presented in Table 2 by displaying the relative MSEs for the *best* performing specification within seven main combination categories for all variables and both horizons. Looking first at the results for GDP growth (columns 2 and 3), a few combination methods outperform moderately the

[14] More detailed results are available from a discussion paper version of our paper.

equal-weighted combination. At the one-year horizon, the gains are strongest for least squares combinations and to a lesser extent combinations based on rolling performance. At the two-year horizon, the scope for improving on the equal-weighted combination appears smaller. However, strategies such as the recent best as well as rolling performance weighting demonstrate improvements close to 10%. Another notable feature of the GDP results is that principal components combination as well as the projection on the mean perform very poorly.

Table 3a (columns 4 and 5) depicts the relative performance of the different combination strategies for inflation where several methods are identified which outperform the equal-weighted combination. The gains are quantitatively larger for regression based combinations such as the projection on the mean, principal components as well as least squares and shrinkage based combinations. However, the performance-based strategies also generally outperform the benchmark, most noticeably the strategy of using the recent best forecaster. At the one- and two-year horizons the best methods for inflation are, respectively, projection on the mean and least squares – both without any explicit bias adjustment. Compared with the equal-weighted combination, the best performing models deliver a reduction in the MSE that is between 24% and 37%. The relatively strong performance for inflation may link to the persistent downward bias in the equal-weighted combination over the sample period analysed. It may be that more flexible combination strategies better “correct” for bias. In this context, Genre et al. (2010) have highlighted stronger persistence in relative forecast performance at the individual level for the SPF inflation forecasts. Hence, a possible explanation for the results for inflation is that compared with equal weighting the other combination methods better handle such features (persistent bias). However, compared with equal weighting, while more sophisticated combination strategies may better capture the level of inflation, they do not fare much better in capturing cyclical dynamics or turning points (see also Figure 4).

Finally, the corresponding results for the unemployment rate in Table 3a (columns 6 and 7) indicate some scope for achieving improvement in forecast performance relative to the equal-weighted combination. The best performing methods for the sample excluding the financial crisis, yield gains close to 20%. However, the results are very variable across the different methods. As can be seen from Table 2, this good performance is particularly sensitive to the chosen specification.

4.3 Sensitivity to data vintage and the financial crisis

Table 3b reports the relative MSEs computed using the more recent December 2011 vintage of outcome variables. Compared to the results using the first vintage (Table 3a), the performance of the different combinations does not appear to change excessively, although performance is generally slightly worse for GDP growth. At the same time, the best performing methods generally continue to perform best when evaluated against the current vintage, i.e., the ranking of different combinations appears quite insensitive to the vintage. The scope for improving on the equal-weighted combination remains most evident for the case of inflation; indeed the performance of the various inflation combinations is broadly unaffected given that inflation was hardly revised during the evaluation period.

Table 3c summarises the results for the SPF sample which includes the observations since the end of 2008 and which are therefore strongly influenced by the recession following the financial crisis. Compared with Table 3a, it is clear that the results are sensitive to the crisis period. For example, for GDP growth one-year ahead, the relatively good performance of the least squares combination during the period prior to the crisis is lost once the sample is extended. In the extended sample, the best performing method at the one-year horizon is based on rolling performance, although the gain of close to 5% is for practical purposes very small. Most methods cannot, improve on the equal-weighted benchmark at both one-year and two-year horizons. A similar picture emerges for inflation and unemployment rate forecasts where, once again, relative performance of the various combination methods deteriorates in the extended sample.

These subsample comparisons suggest that models which perform well during normal times may not be best suited to periods of exceptional macroeconomic volatility. This can be seen in Table 4 which reports the best performing combinations for each of the two samples (see also Figure 4 which plots each of these combinations). For no variable or horizon is it the case that the best performing specification is unchanged when the sample period is extended to include the crisis. In the extended sample, it is noteworthy that performance-based combination strategies often perform best. Another feature associated with the financial crisis seen in Figure 4 is a clear increase in the range of forecasts implied by the various combination strategies employed. For nearly all forecasts, this range increased substantially following the great recession in 2009. Figure 4 also highlights the lagging nature of the combination strategies we employ when it comes to predicting turning points. For example in the second half of 2009, many combination strategies continued to point to a large drop in GDP growth which had by that time already started to recover. All of these features, point to the important challenges that would be faced by a forecast user trying to employ such combination strategies in real time during such periods.

4.4 Empirical results for the "reality check" and recent best combination

Table 4 reports the empirical results of the "reality check" procedure in the form of White P-values which provide an estimate of the likelihood that the best performing model does *not* outperform the equal-weighted combination. For GDP the White test indicates (at both horizons) that it is more likely than not that the best performing models do not outperform the equal-weighted combination. For inflation, the White P-values are considerably lower. However, only for the one-year ahead inflation forecasts, is the reality check indicating a significant improvement of the best performing combination scheme relative to the benchmark at the 10% significance level. In the case of the unemployment rate, the reality check suggests no significant improvements. The above relatively strong result for inflation is only valid during "normal times", however. When the sample period is extended to include the financial crisis, the improvements identified for all combinations appear to "fail" the reality check at standard levels of significance. These findings are therefore more in line with the forecast combination puzzle. They also caution against any tendency to take a relatively good past performance of a given combination strategies as a strong indication of a likely good performance in the future.

Table 5 also reports results from the evaluation of the recent best combination procedure, i.e., the decision rule to select the combination with the best performance based on the most recently observed four forecast errors. For the period excluding the crisis, quantitative improvements are identified for inflation two years ahead and, to a lesser extent, for the unemployment rate one-year ahead. For the other variables and horizons, this decision rule fails to outperform the equal-weighted benchmark. According to the P-values from a Giacomini-White (2006) test – which is applicable here since we are evaluating the performance of a single approach - the improvement for inflation at the two-year horizon is also statistically significant. When the sample is extended to include the crisis, the recent best combination procedure performs worse than the benchmark for all variables and all horizons. These results highlight the overall robustness of the equal-weighted combination for a real time user of surveyed forecasts and suggest that past performance of a given combination provides little information about its likely performance during the crisis period.

5. Concluding remarks

We reviewed the potential for forecast performance improvements through the application of forecast combination methods to surveyed expert forecasts from the ECB Survey of Professional Forecasters. Over the sample period analysed, although the equal-weighted combination sets a high benchmark, a number of more sophisticated combination strategies are shown to achieve quantitative gains relative to this benchmark. Looking across variables, the scope for improvements appears strongest for inflation with smaller gains achievable for the unemployment rate and GDP. However, our results do not identify any single combination approach which appears to dominate either across variables or at different horizons. Nor do they in general suggest statistically significant improvements for the single best performing combination. Instead, the best performing combination methods vary depending on the horizon and the variable considered.

Overall, we would conclude from this study that there exists only a very modest case to consider combinations other than equal weighting as a means of better summarising the information collected as part of the regular quarterly rounds of the ECB SPF. The variation in the best performing specification through time, across target variables and across horizons together with the likely role of chance in explaining the success of some models in our sample would caution strongly against any temptation to try and pick out a preferred or best combination method. Given these findings, we would certainly see no case to replace equal weighting as the headline indicator to summarise the forecasts in the ECB SPF. At most, our results would argue in favour of reporting a suite of different combinations which forecast users could draw on taking into account their historical track record and the prevailing economic context.

References

Aiolfi, M. and Timmerman, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, 135 (1-2), 31-53.

Aiolfi, M., Capistrán, C. and Timmermann, A. (2011). Forecast combination, Ch. 12. In M.P. Clements and D.F. Hendry (Eds.) The Oxford Handbook of Economic Forecasting, Oxford University Press.

Bates, J.M. and Granger, C.W.J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20 (4), 451-468, December.

Benalal, N., Diaz del Hoyo, J.L., Landau, B., Roma, M. and Skudelny F. (2004). To aggregate or not to aggregate? Euro Area Inflation Forecasting. *ECB Working Paper series* No. 374, July.

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A. and Rautanen, T. (2010). An evaluation of the growth and unemployment rate forecasts in the ECB Survey of Professional Forecasters. *Journal of Business Cycle Measurement and Analysis*, Issue 2, pp. 63-90.

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A. and Rautanen, T. (2007). The ECB Survey of Professional Forecasters: A review after eight years' experience. *ECB Occasional Paper* No. 59, April.

Capistrán, C. and Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business and Economic Statistics*, 27 (4), 428-440.

Clemen, R.T (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5 (4), 559-583.

Clemen, R.T. and R.L. Winkler (1986). Combining economic forecasts. *Journal of Business and Economic Statistics*, 4 (1), 39-46.

Croushore, D. and T. Stark (2003), A Real-Time Data Set for Macroeconomists: Does the data vintage matter? *The Review of Economics and Statistics*, 85(3), 605-617

Diebold, F.X. and Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, 6 (4), 503-508.

Diebold F.X. and Pauly, P. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, 6 (1), 21-40.

Garcia, J.A. (2003). An introduction to the ECB Survey of Professional Forecasters. *ECB Occasional Paper* No. 8, September.

Genre, V., Kenny, G., Meyler A. and Timmermann A. (2010), "Combining the forecasts in the ECB SPF: Can anything beat the simple average?", *ECB Working Paper* No. 1277, December.

Giacomini, R. and White H. (2006), Tests of conditional predictive ability, *Econometrica*, 74 (6), 1545-1578

Giannone, D., Henry, J., Lalik, M. and Modugno, M. (2010a). An area-wide real time database for the euro area. *ECB Working Paper* No 1145, January.

Giannone, D., Lenza, M., Momferatou, D., and Onorante, L. (2010b). Short-term inflation projections: A Bayesian Vector Autoregressive approach. *CEPR Discussion Paper*, No. 7746, March.

Granger, C.W.J. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3 (2), 197-204.

Hendry, D.F. and Clements, M.P. (2004). Pooling of forecasts. *The Econometrics Journal*, 7 (1), 1-31.

Newbold, P. and D. I Harvey (2002), Forecast combination and encompassing, in Clemenets, M. P. and Hendry, D. F. (eds.) A companion to economic forecasting, Oxford, Blackwells.

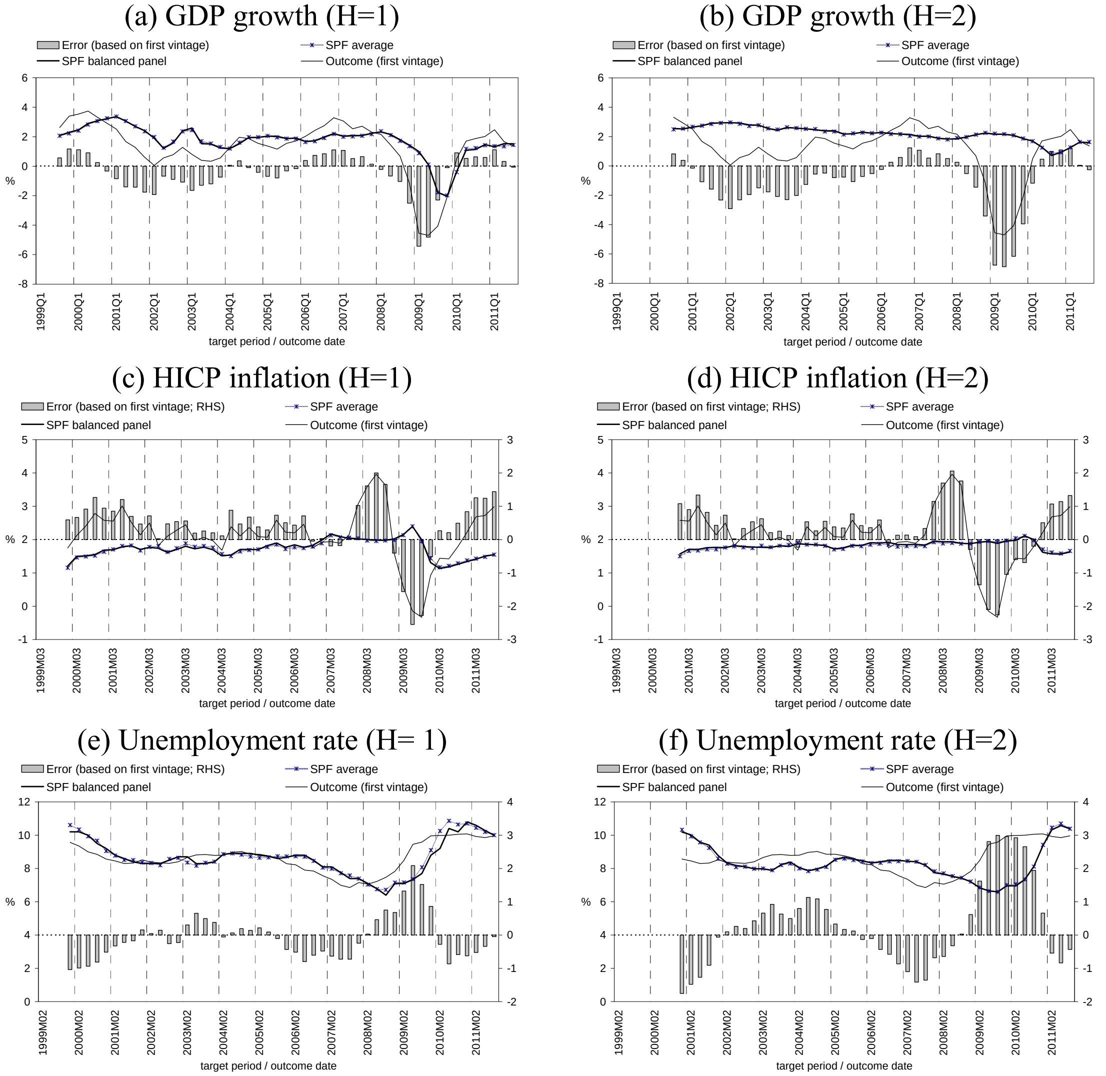
Smith, J. and Wallis, K.F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71 (3), 331-355.

Stock, J.H. and Watson, M.W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23 (6), 405-430.

Timmermann, A. (2006). Forecast combinations, Ch. 4. In G. Elliott, C.W.J. Granger and A. Timmermann (Eds.) Vol. 1, Handbook of Economic Forecasting, North-Holland.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68 (5), 1097-1126.

Figure 1: ECB SPF – Equally weighted forecasts, outcomes and forecast errors



(a) GDP growth (H=1)

(b) GDP growth (H=2)

(c) HICP inflation (H=1)

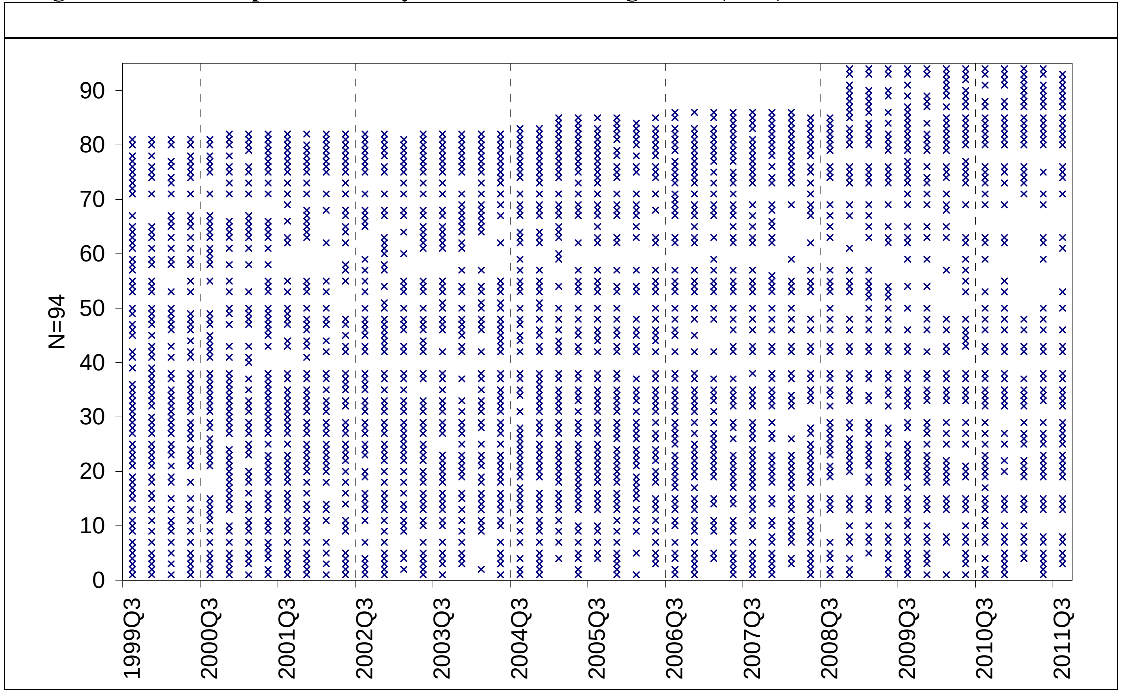
(d) HICP inflation (H=2)

(e) Unemployment rate (H= 1)

(f) Unemployment rate (H=2)

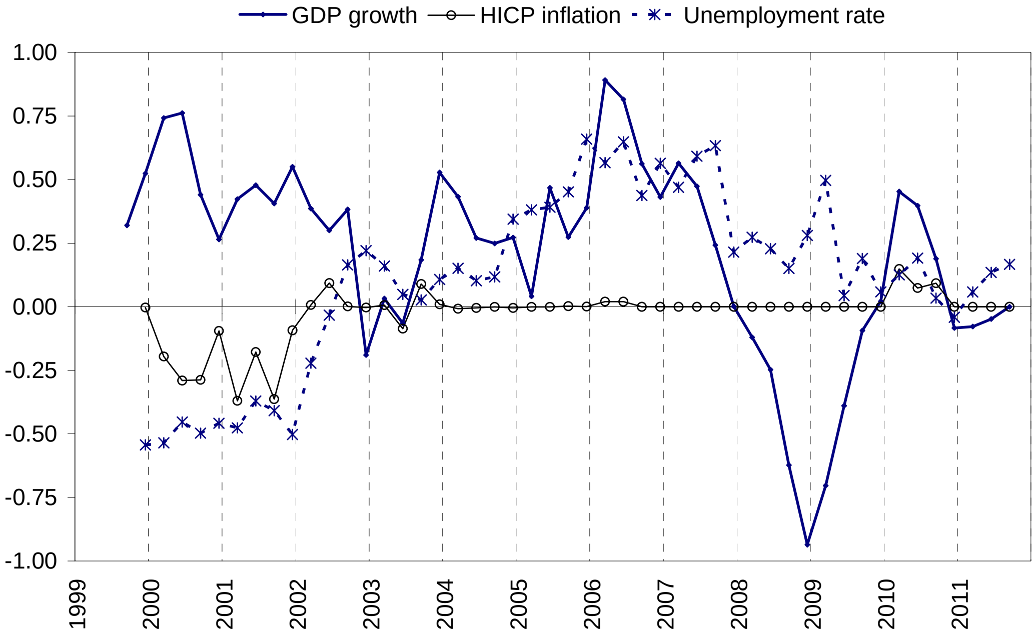
Notes: H = 1 refers to the one-year ahead forecasts, H = 2 refers to the two-year ahead forecasts. SPF average refers to the official (equally weighted) average of ECB SPF forecasts. SPF balanced panel refers to the (equally weighted) average of a balanced panel of the ECB SPF (see text for more detail). It should be noted that the two SPF measures are virtually indistinguishable in the Figures, which serves to highlight that the filtering and balancing of the panel does not impact the mean forecast in any meaningful way. First vintage outcome refers to the first official value for each variable published (by Eurostat).

Figure 2: The SPF panel – 'entry' and 'exit': GDP growth (H=1)



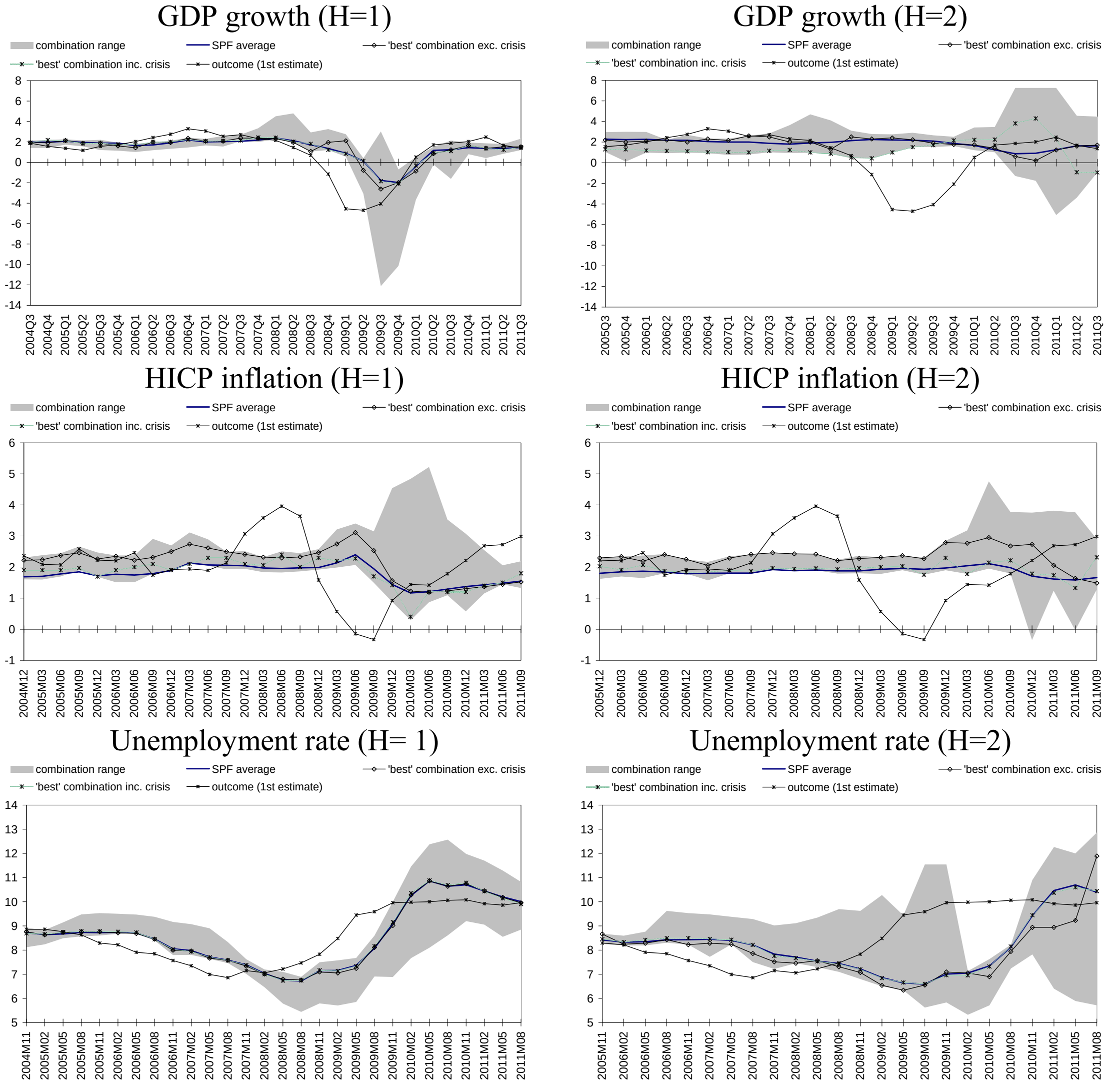
Note: *The Figures indicate for the example of GDP growth forecasts at horizon (H=1), whether a particular individual (tracked along the Y axis) submitted a survey response (indicated by an* ***x****). A blank space thus indicates that no response was submitted for a given survey round. The survey dates are reported on the X-axis. N denotes the total number of forecasters who have submitted at least one response over the full sample.*

Figure 3: Difference between first and current (December 2011) vintages of outcomes for SPF variables *(percentage points)*



Note: *The Figure reports the December 2011 vintage for the outcome variables minus the first vintage. Hence a positive value indicates an upward revision.*

Figure 4: ECB SPF forecast combination with 'best' performing combinations



GDP growth (H=1)

GDP growth (H=2)

HICP inflation (H=1)

HICP inflation (H=2)

Unemployment rate (H= 1)

Unemployment rate (H=2)

Notes: H = 1 refers to the one-year ahead forecasts, H = 2 refers to the two-year ahead forecasts. Combination range refers range from the minimum to maximum forecasts arising from the 31 combination methods considered. SPF average refers to the official (equally weighted) average of ECB SPF forecasts. Best combination excluding (including) crisis refers to the forecast combination with the lowest relative mean squared error for the evaluation period excluding (including) the crisis period - see text for more details. First estimate refers to the first official estimated value for each variable published (by Eurostat).

Table 1: Forecast performance statistics for the ECB SPF: Different samples

	Sample exc. crisis	Sample inc. crisis	Sample 1st half	Sample 2nd half
	GDP growth one-year ahead			
	1999Q3-2008Q3	**1999Q3-2011Q3**	**1999Q3-2005Q3**	**2005Q4-2011Q3**
Mean forecast value	2.1	1.7	2.2	1.3
Mean error	-0.3	-0.4	-0.5	-0.4
	(-0.4 ; -0.1)	*(-0.6 ; -0.2)*	*(-0.7 ; -0.4)*	*(-0.6 ; 0.0)*
RMSE	0.9	1.4	1.0	1.8
	(0.9 ; 1.1)	*(1.3 ; 1.7)*	*(0.9 ; 1.2)*	*(1.6 ; 2.2)*
	GDP growth two-years ahead			
	2000Q3-2008Q3	**2000Q3-2011Q3**	**2000Q3-2006Q1**	**2006Q2-2011Q3**
Mean forecast value	2.4	2.2	2.6	1.8
Mean error	-0.7	-1.1	-1.2	-1.0
	(-0.9 ; -0.5)	*(-1.2 ; -0.9)*	*(-1.4 ; -0.9)*	*(-1.1 ; -0.7)*
RMSE	1.3	2.2	1.5	2.8
	(1.1 ; 1.5)	*(2.0 ; 2.4)*	*(1.3 ; 1.7)*	*(2.6 ; 3.0)*
	HICP inflation one-year ahead			
	1999M12-2008M12	**1999M12-2011M09**	**1999M12-2005M09**	**2005M12-2011M09**
Mean forecast value	1.8	1.7	1.7	1.8
Mean error	0.6	0.4	0.6	0.2
	(0.3 ; 0.8)	*(0.1 ; 0.6)*	*(0.3 ; 0.9)*	*(0.0 ; 0.4)*
RMSE	0.8	0.9	0.7	1.2
	(0.7 ; 1.0)	*(0.9 ; 1.1)*	*(0.5 ; 0.9)*	*(1.1 ; 1.3)*
	HICP inflation two-years ahead			
	2000M12-2008M12	**2000M12-2011M09**	**2000M12-2006M03**	**2006M06-2011M09**
Mean forecast value	1.8	1.8	1.8	1.9
Mean error	0.5	0.3	0.5	0.1
	(0.3 ; 0.7)	*(0.0 ; 0.5)*	*(0.2 ; 0.7)*	*(-0.1 ; 0.4)*
RMSE	0.8	0.9	0.6	1.2
	(0.6 ; 0.9)	*(0.8 ; 1.1)*	*(0.4 ; 0.8)*	*(1.1 ; 1.4)*
	Unemployment rate one-year ahead			
	1999M11-2008M11	**1999M11-2011M08**	**1999M11-2005M08**	**2005M11-2011M08**
Mean forecast value	8.5	8.7	8.8	8.5
Mean error	-0.2	0.0	-0.1	0.1
	(-0.4 ; -0.1)	*(-0.4 ; 0.1)*	*(-0.4 ; 0.0)*	*(-0.4 ; 0.2)*
RMSE	0.5	0.7	0.4	0.9
	(0.4 ; 0.7)	*(0.6 ; 0.9)*	*(0.4 ; 0.8)*	*(0.7 ; 1.0)*
	Unemployment rate two-years ahead			
	2000M11-2008M11	**2000M11-2011M08**	**2000M11-2006M02**	**2006M05-2011M08**
Mean forecast value	8.4	8.3	8.5	8.1
Mean error	-0.2	0.3	0.1	0.5
	(-0.5 ; 0.2)	*(-0.2 ; 0.6)*	*(-0.3 ; 0.4)*	*(-0.2 ; 0.9)*
RMSE	0.8	1.3	0.8	1.6
	(0.7 ; 1.0)	*(1.2 ; 1.5)*	*(0.6 ; 1.1)*	*(1.5 ; 1.9)*

Notes: Figures in brackets refer to minimum and maximum values across the balanced panel of SPF forecasters

Table 2: Relative mean squared errors for different forecast combinations*

	GDP growth		HICP inflation		Unemp. rate	
	H=1	**H=2**	**H=1**	**H=2**	**H=1**	**H=2**
Benchmark						
Equal Weighted SPF (balanced panel)	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Simple Times Series Models						
Naïve	2.65	2.73	0.71	0.74	4.62	3.84
Random Walk (with drift)	1.90	0.96	0.80	0.80	-	-
AR(1)	1.99	5.89	0.81	1.13	4.16	4.52
Recent Best						
v = 1 quarter	1.15	0.88	0.84	0.87	0.89	1.00
v = 4 quarters	1.06	0.97	0.89	0.85	0.96	0.92
Trimmed Means						
Symmetric Trim (5%)	0.94	1.01	1.01	0.99	0.82	1.09
Median (50%)	1.03	1.02	1.01	1.01	0.80	1.01
Recursive Performance						
$\delta = 1.0$	1.01	0.99	0.98	0.95	0.81	1.00
$\delta = 0.95$	1.01	0.99	0.98	0.95	0.82	1.01
$\delta = 0.85$	1.00	1.00	0.98	0.95	0.83	1.03
Rolling Performance						
v = 1 quarter	0.92	0.88	1.05	0.94	0.81	1.03
v = 4 quarters	1.02	0.97	0.97	0.94	0.84	1.08
v = 8 quarters	1.00	1.00	0.97	0.94	0.84	1.04
Projection on Mean (PM)						
PM	1.97	2.34	0.88	0.75	2.39	3.79
PM ($w_{0,h} = 0$)	1.80	3.30	0.76	0.67	1.54	2.83
Principal Components (PC)						
PC (p = 1)	3.28	3.07	0.85	0.72	5.54	4.40
PC (p = 2)	2.97	3.32	0.86	0.70	5.56	4.51
PC (p = 3)	3.49	3.11	0.83	0.75	5.49	4.56
Least Squares (LS)						
LS (c= 2, $w_{0,h} = 0$)	1.93	3.29	0.90	0.63	1.76	2.70
LS (c= 2)	2.02	3.37	0.93	0.83	3.11	3.94
LS (c= 2, $w_{0,h} = 0, \sum w_{i,h} =1.0$)	1.15	1.56	0.98	1.00	0.97	0.71
LS (c= 2, $w_{0,h} = 0, 0 \leq w_{i,h} \leq 1.0$)	1.72	3.36	0.91	0.63	1.70	2.70
LS (c= 3, $w_{0,h} = 0$)	1.51	3.02	0.81	0.68	1.61	2.73
LS (c= 3)	1.74	3.40	0.93	0.76	2.61	3.97
LS (c= 3, $w_{0,h} = 0, \sum w_{i,h} =1.0$)	0.71	1.15	1.09	0.95	0.84	1.05
LS (c= 3, $w_{0,h} = 0, 0 \leq w_{i,h} \leq 1.0$)	1.66	3.36	0.82	0.66	1.50	2.92
Shrinkage Weights (SW)						
SW ($w_{0,h} = 0$, c=2, $\boldsymbol{\kappa} = 4$)	1.61	1.73	0.83	0.76	1.32	1.41
SW ($w_{0,h} = 0$, c=3, $\boldsymbol{\kappa} = 4$)	1.16	1.20	0.88	0.90	1.04	1.12
SW ($w_{0,h} = 0$, c=2, $\boldsymbol{\kappa} = 6$)	1.50	1.31	0.88	0.86	1.15	0.94
SW ($w_{0,h} = 0$, c=3, $\boldsymbol{\kappa} = 6$)	1.07	0.99	0.95	1.01	0.87	0.93
SW (c=2, $\boldsymbol{\kappa} = 4$)	1.67	1.82	0.93	0.92	1.98	1.81
SW (c=3, $\boldsymbol{\kappa} = 4$)	1.26	1.35	0.95	0.93	1.35	1.28
SW (c=2, $\boldsymbol{\kappa} = 6$)	1.54	1.35	0.98	0.96	1.61	1.11
SW (c=3, $\boldsymbol{\kappa} = 6$)	1.12	1.09	0.99	1.01	0.99	0.90

Notes: Calculations are made using first vintage estimates of each variable. The periods used for the one- and two-year ahead forecasts respectively are: GDP growth 2004Q3-2008Q3 and 2005Q3-2008Q3; HICP inflation 2004M12-2008M12 and 2005M12-2008M12; Unemployment rate 2004M11-2008M11 and 2005M11-2008M11.

Table 3 Out-of-sample comparison of forecast performance
(Relative MSE for best performing specification of different combination strategies)

	GDP	GDP	HICP	HICP	Unemp.	Unemp.
	H=1	H=2	H=1	H=2	H=1	H=2
	(a) 1st vintage data / Sample exc. crisis					
Recent Best Forecaster	1.06	0.88	0.84	0.85	0.89	0.92
Recursive Performance	1.00	0.99	0.98	0.95	0.81	1.00
Rolling Performance	0.92	0.88	0.97	0.94	0.81	1.03
Projection on the Mean	1.80	2.34	0.76	0.67	1.54	2.83
Principal Components	2.97	3.07	0.83	0.70	5.49	4.40
Least Squares	0.71	1.15	0.81	0.63	0.84	0.71
Shrinkage Weights	1.07	0.99	0.83	0.76	0.87	0.90
	(b) Latest vintage data / Sample exc. crisis					
Recent Best Forecaster	1.08	0.93	0.84	0.85	0.78	0.80
Recursive Performance	1.00	1.00	0.98	0.95	0.84	1.01
Rolling Performance	0.94	0.93	0.97	0.94	0.70	1.01
Projection on the Mean	1.65	2.01	0.76	0.67	1.06	3.59
Principal Components	2.03	2.43	0.83	0.70	5.29	6.58
Least Squares	0.83	0.97	0.81	0.63	1.00	0.75
Shrinkage Weights	1.05	0.98	0.83	0.76	0.84	0.84
	(c) 1st vintage data / Sample inc. crisis					
Recent Best Forecaster	1.05	1.02	0.91	0.97	0.94	1.32
Recursive Performance	0.99	1.00	0.98	0.99	0.85	1.01
Rolling Performance	0.95	1.00	0.98	0.98	0.84	1.01
Projection on the Mean	1.12	1.31	1.20	1.14	1.56	1.29
Principal Components	1.43	1.00	1.59	1.13	3.16	1.88
Least Squares	1.06	1.07	1.00	0.97	1.50	1.01
Shrinkage Weights	1.06	1.08	1.12	1.01	1.10	1.08

Note: The out-of-sample periods excluding the crisis are as reported in Table 2. First vintage data refer to the first estimate of each variable. Latest vintage data refer to the vintage at the cut-off (i.e. December 2011). The out of sample period including the financial crisis extends the sample to 2011Q3, 2011M09, and 2011M08 for GDP growth, HICP inflation and the unemployment rate forecasts respectively.

Table 4 White "Reality Check"

Sample excluding the financial crisis			
Variable (horizon)	**Best model**	**Rel. MSE**	**White P-Value**
GDP growth (H=1)	Least Squares (c= 3, $w_{0,h} = 0$, $\sum$ $w_{i,h}$ =1.0)	**0.71**	0.69
GDP growth (H=2)	Recent Best (v = 1 quarter)	**0.88**	0.90
HICP inflation (H=1)	Projection on the Mean ($w_{0,h} = 0$)	**0.76**	**0.06**
HICP inflation (H=2)	Least Squares (c= 2, $w_{0,h} = 0$)	**0.63**	0.11
Unemp. rate (H=1)	Median	**0.80**	0.78
Unemp. rate (H=2)	Least squares (c= 2, $w_{0,h} = 0$, $\sum$ $w_{i,h}$ =1.0)	**0.71**	0.23
Sample including the financial crisis			
Variable (horizon)	**Best model**	**Rel. MSE**	**White P-Value**
GDP growth (H=1)	Rolling Performance (v = 1 quarter)	**0.95**	0.95
GDP growth (H=2)	Principal Component (p = 1)	1.00	0.99
HICP inflation (H=1)	Recent Best (v = 1 quarter)	**0.91**	0.68
HICP inflation (H=2)	Least Squares (c= 3, $w_{0,h} = 0$, $\sum$ $w_{i,h}$ =1.0)	**0.97**	0.72
Unemp. rate (H=1)	Rolling Performance (v = 4 quarters)	**0.84**	0.91
Unemp. rate (H=2)	Recursive Performance (δ = 0.85)	1.01	0.98

White refers to the White (2000) 'reality check' test.

Table 5 Evaluation based on recent best combination

	exc. crisis		**inc. crisis**	
Variable (horizon)	**Rel. MSE**	**GW P-Value**	**Rel. MSE**	**GW P-Value**
GDP growth (H=1)	1.35	0.77	1.12	0.78
GDP growth (H=2)	1.10	0.68	1.50	0.89
HICP inflation (H=1)	1.13	0.86	1.38	0.93
HICP inflation (H=2)	**0.70**	**0.05**	1.13	0.72
Unemployment rate (H=1)	**0.94**	0.35	1.27	0.81
Unemployment rate (H=2)	1.52	1.00	1.37	0.98

GW refers to the Giacomini and White (2006) test.