Meyler, Aidan

Working Paper

Forecast performance in the ECB SPF: Ability or chance?

ECB Working Paper, No. 2371

Provided in Cooperation with:
European Central Bank (ECB)

Suggested Citation: Meyler, Aidan (2020) : Forecast performance in the ECB SPF: Ability or chance?, ECB Working Paper, No. 2371, ISBN 978-92-899-4014-6, European Central Bank (ECB), Frankfurt a. M.,
https://doi.org/10.2866/20544

This Version is available at:
https://hdl.handle.net/10419/228985

EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Aidan Meyler

Forecast performance in the ECB SPF: ability or chance?

No 2371 / February 2020

Abstract

In this paper, we consider whether differences in the forecast performance of ECB SPF respondents reflect ability or chance. Although differences in performance metrics sometimes appear substantial, it is challenging to determine whether they reflect ex ante skill or other factors impacting ex post sampling variation such as the nature of economic shocks that materialised or simply which rounds participants responded in. We apply and adapt an approach developed by D'Agostino et al. (2012) who used US SPF data. They developed a test of a null hypothesis that all forecasters have equal ability. Their statistic reflects both the absolute and relative performance of each forecaster and they used bootstrap techniques to compare the empirical results with the equivalents obtained under the null hypothesis of equal forecaster ability. Our results, at a first pass, suggest that there would appear to be evidence of good/bad forecasters. However once we control for the autocorrelation that is caused by the overlapping rolling horizons, we find, like D'Agostino et al. (2012), that the best forecasters are not statistically significantly better than others. Unlike D'Agostino et al. (2012), however, we do not find evidence of forecasters that perform very significantly worse than others. Controlling for autocorrelation is a key feature of this paper relative to previous work. Our results hold considering the whole sample period of the ECB SPF (1999-2018) as well as the pre- and post-global financial crisis samples. We also find that when assessed across all variables and horizons, the aggregate (consensus) SPF forecast performs best.

JEL Classification: C53, E27, E37

Keywords: forecasting, performance, bootstrap.

Non-technical summary

This paper considers the forecasting performance of individual forecasters who have participated in the ECB Survey of Professional Forecasters (ECB SPF) for the euro area over the past twenty years since its inception in 1999. While the focus of the published SPF reports has generally been on the aggregated (average) results, it is of interest to policy makers, practitioners and academics to consider whether it is possible to identify individual forecasters who perform better or worse than average. It is, however, challenging in practice to determine whether any apparent differences in forecasting performance across individuals: (a) are statistically and economically meaningful and (b) reflect ability or chance. For example, imagine that two forecasters, A and B, have the same information set but Forecaster A forecasts lower inflation than Forecaster B owing to different beliefs about how the economy works. Then assume, for example, there is an upward oil price shock, not anticipated by either forecaster, in this case, even if the forecast of Forecaster A was better 'ex ante', it may well be that Forecaster B looks better 'ex post'. Of course it could be argued that if the sample is large enough these shocks should average themselves out. However, it has been the case that some shocks appear to have been relatively persistent and therefore could impact ex post rankings relative to hypothetical ex ante forecast quality. The period for which the ECB SPF has been in existence (since 1999) has been characterised by notable structural events/shocks – such as the establishment of the euro area and single currency itself but also the global financial crisis. Furthermore, the SPF is a voluntary survey and the panel is 'unbalanced' (i.e. not every forecaster participates in each round). Therefore, if a panellist participated when the forecast error was relatively low (high) then his/her average performance might look better (worse).

The evidence from the existing, largely US, literature has been mixed. Stekler (1987) and Batchelor (1990) using the US Blue Chip survey reported conflicting results. The former stated "it is possible to identify 'better' forecasters". The latter argued that this finding owed to "an incorrectly defined test statistic" and when a "more appropriate test is conducted" it is hard to argue that any forecasters are better in a statistically meaningful sense. Zarnowitz and Braun (1993) presented some evidence from the US SPF that suggested any superior performance by individual forecasters tended not to persist. Christensen et al. (2008) also based on the US SPF found mixed evidence depending on the forecast variable considered. Interestingly, a more recent paper by D'Agostino et al. (2012), also based on the US SPF, reported that they were unable to identify forecasters that are statistically significantly better than average, but did find evidence of forecasters that perform worse than average.

In this paper, we apply and adapt to the ECB SPF the approach developed by D'Agostino et al. (2012) who used US SPF data. Their baseline (null) hypothesis is that all forecasters have equal ability. They then develop a test statistic, which reflects both the absolute and relative performance of each forecaster, to assess this hypothesis. Using 'bootstrap' (randomly reallocating errors across individual forecasters) and 'Monte-Carlo' (repeating the exercise a large number of times)

techniques, they compare the empirical results with the equivalents obtained under the null hypothesis of equal forecaster ability.

Relative to the D'Agostino et al. (2012) work, although we use largely the same methodology, there are a number of features/variations. First, they examined the US SPF, whereas we assess forecasters in the ECB SPF, which was established in 1999 to survey professional forecasts for the euro area economy. Second, in addition to assessing inflation and growth forecasts, we also assess unemployment forecasts. Third, as the performance metric is 'normalised' for each variable horizon, we apply the methodology jointly assess across all three variables and across horizons. Fourth, as a robustness cross-check we run the tests using the percentile rank statistic. Using the percentile rank rather than the rank statistic gets around the balanced panel limitation referred to in this literature. Fifth, as ten years have now passed since the global financial crisis, we systematically check difference between pre- and post-crisis periods. Lastly, and crucially, motivated by the seeming discrepancy at first glance between the bootstrap results, which suggest some forecasters might have performed statistically better/worse than others, and the lack of correlation in performance across the pre- and post-crisis period, we attempt to control for autocorrelation in the forecast errors. We find that doing so brings the bootstrap results and the cross sub-sample findings more in line with each other.

The main finding is that, while at first glance there appears to be evidence of forecasters who perform better or worse than average, once autocorrelation (which owes to the fact that forecasts are for so-called rolling horizons one and two years ahead and therefore overlap to some extent from one round to the next) is controlled for there is little evidence of forecasters who have performed better or worse in a statistically significant sense. Thus, like D'Agostino et al. (2012), we find no evidence of forecasters that are statistically significantly better than others, but unlike them we do not find evidence of forecasters that perform very significantly worse than others. This may owe to the fact that, since its inception, the ECB SPF has endeavoured to include forecasting institutions (not just individuals) with experience. In the US SPF, which was initiated in 1968, participation declined over time and the survey was close to being discontinued until the Federal Reserve Bank of Philadelphia took over its administration in 1990. Our results hold considering the whole sample period of the ECB SPF (1999-2018) as well as the pre- and post-global financial crisis samples.

Another interesting feature of the results is that, when assessed across all variables and horizons, the aggregate (consensus) SPF forecast performs best. This finding which is consistent with the analysis of Genre et al. (2013), who report that it is hard to find forecast combination methods that beat the simple average, supports the practice of focusing on the aggregate (average) forecasts in the ECB SPF publications.

1 Introduction

Macroeconomic forecasts and expectations play a central role in economic and monetary analyses. Private agents' expectations can affect the economy because they can influence economic decisions in areas such as saving, consumption and investment, as well as wage and price setting. Furthermore they can be useful to policy makers both as a cross-check of their own macroeconomic forecasts but also for assessing the transmission of their policy decisions. In this context, ever since the launch of euro area economic and monetary union in 1999, the ECB has conducted a quarterly survey of macroeconomic forecasters known as the ECB Survey of Professional Forecasters (SPF). In the twenty years since its inception, the ECB SPF has provided a rich set of information on professional forecasters' projections for euro area HICP, real GDP growth and unemployment rate. As the sample size as grown there has been a growing literature examining many different aspects of the ECB SPF.[1] Although some studies, such as Bowles et al. (2010), Genre et al. (2013) and Grothe and Meyler (2018), have considered forecast performance, they have mostly focused on the aggregate performance.

More generally, given that how people formulate expectations of economic variables is such a key conceptual and empirical issue in macroeconomics, it is hardly surprising that there has been a relatively large literature related to other surveys of professional forecasters. Bonham and Cohen (2001) and Keane and Runkle (1992) examine the issue of whether forecasters forecasts are unbiased. The importance of unbiasedness for the debate about rational expectations probably accounts for the fact that most of the literature on the properties of individual-level forecasts has focused on testing for rationality and unbiasedness.

However, like for the ECB SPF, there has been generally little focus on whether the accuracy of individual forecasters varies in a statistically significant sense. This possibly also reflects the challenging nature of the question. Among those studies that have considered the issue (generally focusing on US forecasts) the findings are mixed and inconclusive. Zarnowitz and Braun (1993) presented some evidence from the US SPF that suggested any superior performance by individual forecasters tended not to persist.[2] Stekler (1987) and Batchelor (1990) using the Blue Chip survey reported conflicting results while Christensen et al. (2008) based on the US SPF found mixed evidence. D'Agostino et al. (2012) reported that they were unable

1 For example, Abel et al. (2016), Glas and Hartmann (2016) Łyziak and Paloviita (2017), and Rich and Tracy (2018) consider various aspects of forecast uncertainty; Grishchenko et al. (2017), Dovern and Kenny (2017) and Beechey et al. (2011) examine the anchoring of inflation expectations; and Reitz et al. (2012) and Frenkel et al. (2011) look at expectations formation.

2 This is somewhat similar to the finding by Croushore (2009) that when considering the issue of bias in the US SPF although there may be some evidence for specific sub-samples when looked over longer time spans these tend to disappear.

to identify forecasters that are statistically significantly better than average, but did find evidence of forecasters that perform worse than average. [3]

This paper adopts the methodology of the latter paper applying a bootstrap and Monte Carlo approach to assess the extent to which the observed data on the performance of participants is consistent with a null hypothesis of equal underlying forecasting ability.[4]

In practice there are a number of challenges to assessing whether differences in forecasting performance: (a) reflect ex ante or ex post outcomes or (b) are statistically significant. It is possible that a forecast that looks good ex post might not have been so good ex ante.[5] For example, imagine that Forecaster A and Forecast B have the same information set but Forecaster A forecasts lower inflation than Forecaster B owing to different beliefs about how the economy works. Then assume there is an upward oil price shock, not anticipated by either forecaster, in this case, even if the forecast of Forecaster A was better ex ante, it may well be that Forecaster B looks better ex post. Of course it could be argued that if the sample is large enough these shocks should average themselves out. However, it has been the case that some shocks appear to have been relatively persistent and therefore could impact ex post rankings relative to hypothetical ex ante forecast quality. It is also true that the period for which the ECB SPF has been in existence (since 1999) has been characterised by notable structural events/shocks – the establishment of the euro area and single currency itself but also the global financial crisis. Furthermore, the SPF is a voluntary survey and the panel is unbalanced. Therefore, if a panellist participated when the forecast error was relatively low (high) then his/her average performance might look better (worse).

Another question that is difficult to answer is how to compare across different variables and horizons. For example, imagine that Forecaster A performs better at forecasting HICP inflation and/or the shorter (one-year) ahead horizon, while Forecaster B performs better for another variable (real GDP growth or the unemployment rate) and/or the longer (two-year) ahead horizon. In this situation, it is not obvious how to determine which forecaster performed best, and even if one did so, whether any differences are statistically significant.

Relative to the existing work on this subject, this paper has a number of features/variations. Although we use the same methodology as D'Agostino et al. (2012), there are a number of differences with respect to their paper. First, they examined the US SPF, whereas we assess forecasters in the ECB SPF, which was established in 1999 to survey professional forecasts for the euro area economy. Second, in addition to assessing inflation and growth forecasts, we also assess

---

3 Gamber et al. (2015) using the same methodology examine the Federal Reserve's performance within the cross-sectional distribution of private-sector forecasts.

4 The approach we take is similar to that used in research such as Kosowski et al. (2006), Fama and French (2010), and Cuthbertson, Nitzsche, and O'Sullivan (2008) to assess the relative performance of mutual funds.

5 Diebold et al. (1999) propose an approach using the Probability Integral Transform for making ex post assessments of forecasters' probability distributions. Clements (2014) also examines forecast uncertainty in the US SPF ex ante and ex post.

unemployment forecasts. Third, as the performance metric is ‘normalised’ for each variable horizon, we apply the methodology jointly assess across all three variables and across horizons. Fourth, as a robustness cross-check we run the tests using the percentile rank statistic. Using the percentile rank rather than the rank statistic gets around the balanced panel limitation referred to in this literature. Fifth, as ten years have now passed since the global financial crisis, we systematically check difference between pre- and post-crisis periods. Lastly, and perhaps most crucially, motivated by the seeming discrepancy at first glance between the bootstrap results, which suggest some forecasters might have performed statistically better/worse than others, and the lack of correlation in performance across the pre- and post-crisis period, we attempt to control for autocorrelation. We find that doing so brings the bootstrap results and the cross sub-sample findings more in line with each other.

2 Methodological approach to testing for differences in forecaster performance

This section outlines some of the previous work on assessing the significance of differences in forecaster performance and then describes the methodology used in this paper.[6]

2.1 Previous Work

Stekler (1987) using data between 1977 and 1982 from the monthly Blue Chip survey of economic indicators argues that "it is possible to identify 'better' forecasters". The metric he uses is the rank of each forecaster based on their root mean squared error in each year. To ensure a balanced panel, he excluded forecasters that did not make predictions for every year thereby reducing the panel from thirty-one to twenty four forecasters.

However, Batchelor (1990) argued that Stekler's findings were "based on an incorrectly defined test statistic" and when a "more appropriate test is conducted", the accuracy rankings did not appear to be statistically different from those that might be expected as a result of sampling error in a population of equally accurate forecasters.

Christensen et al. (2008) using the US SPF test for equal forecasting accuracy by extending the forecast comparison test of Diebold and Mariano (1995) to a case in which there are more than two forecasts to be compared. Their approach requires both a balanced panel and a long time series. This resulted in their panel being shrunk to three forecasters. They report mixed results with the tests suggesting equal predictive accuracy for some variables but not others.

In this paper, we adopt the approach of D'Agostino et al. (2012) who address the question of chance versus ability using bootstrapping and Monte Carlo simulation techniques to see whether observed outcomes differ significantly from what would be expected under null of equal ability. Their basic idea is to take the forecast errors for a given variable horizon in each period and randomly reallocate them across the forecasters who provided a forecast in that period for the specific variable horizon (bootstrapping). They repeat this process a large number of times to simulate the distribution of forecast errors under the assumption (null hypothesis) of equal forecasting ability (Monte Carlo simulation). One of the advantages of their approach is that it does not require a fully balanced panel, unlike the other papers cited above. Furthermore, rather than using the rank metric, they use an adjusted mean squared error statistic which penalises outliers and also controls for whether any given

[6] This section draws extensively from D'Agostino et al. (2012).

variable / horizon / period was ‘easy or hard’ to forecast. Overall they conclude there is little evidence of ‘good’ forecasters but some evidence of ‘bad’ forecasters.

2.2 A test of equal forecasting ability

Which metric?

Before describing the bootstrapping process, we first briefly consider which forecast performance metric to utilize. There are a number of possible options each with some possible advantages and possible disadvantages. The forecast error (calculated as the forecast value minus the actual outturn) is simple but would score a forecaster with offsetting errors the same as a forecaster with no errors. Using the absolute error would avoid this problem as offsetting errors are not cancelled out. However this metric could score equivalently two forecasters even though one generally makes small errors whereas the other makes smaller errors most of the time but sometimes makes relatively large errors (i.e. outliers). The squared error statistic avoids the cancelling out issue as well, but also penalises outliers. This is the most commonly used statistic. However, both the absolute error and squared error metrics suffer the potential drawback that they might penalise forecasters who participated when it was relatively hard to forecast. The forecast rank (usually based on either the absolute or squared error) has the advantage that it is relatively simple and does not penalise forecasters who participated when it was relatively hard to forecast (inter-period). However it does not take into account the relative size of errors in a given period (intra-period). Furthermore, it generally requires a balanced panel; otherwise the scale of the metric will vary according to the number of participants in each period. As discussed above, in the existing literature, this has tended to reduce substantially the panel size. A possible solution to this balanced panel issue is to use the percentile rank, which maps the rankings in each period on to the 1-100 scale. However like the rank metric, the percentile rank does not take into account the relative size of errors either intra-period or inter-period.

In principle, the squared or absolute error metrics would appear to be preferred. However, as noted, it may be the case that it is relatively ‘easy or hard’ to forecast given macroeconomic variables at given points in time. Stock and Watson (2002) coined the phrase “the great moderation” to describe the observed reduction in business cycle volatility since the mid-1980s.[7] However the global financial crisis resulted in substantial movements in macroeconomic variables and large forecasting errors. Given that we use an unbalanced panel, a forecaster who did not participate when forecast errors were relatively large could appear better than others even if their errors were similar at other points in time.

[7] Although Stock and Watson (2006, 2005) consider the issue of whether US inflation became harder to forecast and, if so, why

To address this issue, D'Agostino et al. (2012) construct a normalised squared error statistic for each variable for each period for each forecaster. This is defined as:

$$E_{vit} = \frac{e_{vit}^2}{\left(\sum_{i=1}^{N_{vt}} e_{vit}^2\right)\frac{1}{N_{vt}}}$$

where e_{vit} is the realised error for forecaster i for variable v in period t and N_{vt} is the number of forecasters providing a forecast for variable v in period t.[8] An overall score of each forecaster for each variable is calculated by calculating the average of their normalised squared error statistics. Even if a forecaster drops in and out of the survey, the average score can be calculated based on the periods in which they provide a forecast. Furthermore, as errors are normalised for each variable and period, it should not matter if they do not participate when forecast errors are relatively high or low.

A key advantage of this metric is that it controls for differences over time in the 'forecastability' of a given variable, which should affect all forecasters more or less equally, while taking into account the relative magnitude of individual errors for a given variable in a given period. For instance, an E_{vit} of 2 (0.5) would imply that the squared error for individual i for variable v for period t was twice (half) the mean squared error for that variable in that period.

Another advantage of normalising the error statistics for each variable, horizon and period is that it allows us to aggregate across variables and horizons – see below.

However, one potential drawback of this normalised metric is that a forecaster who makes a relatively large error in a period when the average squared error of all forecaster is very small will suffer a large penalty whereas forecaster who makes a relatively small error when the average squared error of all forecaster is very large will not benefit much.

Bootstrapping and Monte Carlo Simulation

The essential idea behind the test of the hypothesis of equal forecaster ability is to randomly reshuffle and reassign individual forecasts for a given variable in a given period.[9] This is repeated many (e.g. 1,000) times and then we test whether the realized historical distribution of forecaster performance is statistically significantly different from those obtained from this random reshuffling. If the actual distribution of forecast performance lies within given confidence bands (for example, 1% and 99%) of the simulated distributions, then we cannot reject the null hypothesis that forecasters have equal ability and that differences in performance are due to chance.

8 A normalised absolute error metric can be calculated by substituting the absolute error instead of the squared error in both the number and dominator.

9 Following D'Agostino et al. (2012), the bootstrap technique is applied in a way that exactly replicates the original unbalanced nature of the panel. Forecast errors for a given variable in a given period are only reassigned among forecasters who provided forecasts for that variable in that period. Errors are not reassigned across periods.

To see how this works in practice, consider the best performing forecaster for a given variable and horizon. We can compare his/her score with the entire distribution of the best scores for each simulation. If the actual best performer lies outside the 1st or 99th percentiles which give us an indication of the range that might be observed in "best performer" scores under random reshuffling and reassignment. If the best performer in the actual data is statistically significantly better than other forecasters, we would expect their score to lie outside the range represented by these bootstrap percentiles (confidence bands).

3 Application to the ECB Survey of Professional Forecasters

The ECB SPF has measured inflation expectations and other macroeconomic expectations since the beginning of monetary union (1999). At the time of its launch, the ECB SPF was the only gauge of private sector macroeconomic expectations for the euro area as a whole.[10] The survey collects information on the expected rates of consumer price inflation, real GDP growth and unemployment in the euro area at several horizons, ranging from the current year to the longer term.[11] In addition, respondents provide expectations for other variables underpinning their forecasts, such as wage growth, the oil price and the exchange rate, and qualitative comments that enrich their quantitative forecasts. Thus the overall survey results provide a comprehensive depiction of experts' aggregate assessment of the macroeconomic outlook.[12]

Expectations are sampled at different horizons for different purposes. In the ECB SPF there are two broad classes of horizon: the so–called 'rolling horizons' and the 'calendar year' horizons. The rolling horizons, which are the focus of this paper, are for one-year and two-years ahead of the latest available data at the time the survey conducted.[13] For example, when the survey was conducted in January 2018 (Q1 2018), HICP inflation data were available up to December 2017, real GDP growth up to Q3 2017 and the unemployment rate for November 2017. Thus the one-year and two-years ahead horizons respectively were December 2018 and December 2019 for HICP inflation, Q3 2018 and Q3 2019 for real GDP growth and November 2018 and November 2019 for the unemployment. Apart from the advantage of representing a fixed length horizon (one-year and two-years), the rolling horizons can be useful measuring how perceptions of risk and uncertainty evolve over time, because these abstract from the natural decline in uncertainty that tends to occur as the forecast horizon shrinks.

The respondents to the survey are expert economists working in either financial or non-financial institutions located mainly in euro area countries but also in some other countries (United Kingdom, Sweden, Denmark and Switzerland). The majority of respondents are from financial institutions, although a significant number of

[10] Since then other surveys including Consensus Economics and the Euro Zone Barometer have added euro area projections to their questionnaires.

[11] Another feature of the ECB SPF is that for HICP inflation, real GDP growth and unemployment expectations at all horizons, including the longer term, are collected not just in the form of point forecasts, but also probability distributions. This allows quantification of forecast uncertainty and of whether forecasters consider the uncertainty to be broadly balanced around their point forecast or skewed towards the upside or the downside. However, in this paper we focus on point forecasts. For an assessment of these probability distributions see Kenny et al. (2015).

[12] For an early and comprehensive introduction to, and overview of, the ECB SPF see Garcia (2003)

[13] The ECB SPF also ask for calendar year forecasts for the current (at the time of the survey) calendar year, the following two calendar years and for four/five years ahead. The longer-term calendar year expectations are four calendar years ahead in the Q1 and Q2 rounds and five calendar years ahead in the Q3 and Q4 rounds.

economic research institutions also contribute. We return to the issue of forecaster sector (financial/non-financial) and location (euro area/non-euro area) later in the paper. On average approximately 60 responses are received each quarter, which is relatively high compared with other expert macroeconomic surveys for the euro area as a whole. The active panel (loosely defined as those who have participated in the past two years) has tended to average about 75.

Participation in the ECB SPF is voluntary and the panel is unbalanced. Table 1 reports how many individual forecasters have provided forecasts for the different variables and horizons. For instance, while 104 forecasters have provided at least one forecast for HICP inflation one-year ahead, a smaller number, 77, have provided at least 20 forecasts for HICP inflation one-year ahead. Not all forecasters forecast all variables or horizons. They tend to forecast HICP inflation more often than real GDP growth (and in turn real GDP growth more often than the unemployment rate). Also they tend to provide one-year ahead forecasts more often than two-year ahead forecasts. For example, although 77 forecasters have provided at least two forecasts for HICP inflation one-year ahead, only 63 have provided at least twenty forecasts for the unemployment rate two-years ahead.

To avoid that our results are impacted by forecasters who only provided a very small number of forecasts, for each variable and horizon we restrict the panel to those who have provided at least twenty forecasts for that variable horizon.[14] Thus, we consider 77 (70) forecasters when assessing the one-year (two-years) ahead inflation forecasts, 73 (68) for one-year (two-years) real GDP growth and 69 (63) for the one-year (two-years) ahead unemployment rate. Our choice of a threshold of at least twenty was fairly arbitrary but was aimed at not shrinking the panel too much while ensuring that those in the panel have provided a reasonable number of forecasts. D'Agostino et al. (2012) set their threshold at ten for their robustness check. Our results did not change qualitatively when we choose other thresholds (such as one, ten, thirty or forty).

Like most macro-economic data for most countries, euro area macroeconomic statistics tend to be revised from preliminary releases. In fact, revisions may be more of an issue for euro area data as they represent an aggregation of data from (currently) 19 countries. To address this issue we have constructed real-time data for each of the three main variables – HICP inflation, real GDP growth and the unemployment rate.[15] For HICP inflation, we use the first full release (not the 'flash estimate' which is based on partial data and subject to revision) and for real GDP growth we use the second estimate (not the preliminary or first estimates, which are more subject to revisions).

[14] For example, if a forecaster only provided one or two forecasts, and these forecasts coincided with a period when forecast errors were relatively low (high), this forecaster good appear better (worse) than others but it would be hard to argue that this was a meaningful difference.

[15] All our results are qualitatively similar when using current vintage data.

A descriptive summary of forecast errors

Forecast errors in the ECB SPF have been substantial. Table 2 shows that across different variables (HICP inflation, real GDP growth and unemployment) and horizons (one-year ahead and two-years ahead), forecast errors have been sometimes sizeable and one-sided on average. For example, the mean absolute error for real GDP growth two-year ahead forecasts was 1.35 p.p. and both the smallest and largest mean errors (0.23 p.p. and 1.75 p.p. respectively) were positive indicating that all individual forecasters had positive errors on average.

There has been considerable heterogeneity in the forecasting performance of ECB SPF forecasters, although many were relatively tightly clustered around the middle. One indicator of the spread of forecast performance is that the range between the minimum and maximum mean absolute error across variables/horizons is 0.77 p.p. (ranging between 0.48 p.p. for UNEM_1Y to 1.22 p.p. for RGDP_2Y), which is almost as large as the average mean absolute error of 0.91 p.p. (ranging between 0.59 p.p. for UNEM_1Y to 1.35 p.p. for RGDP_2Y).[16] However, the average inter-quartile range (the difference between the 25th and 75th percentiles and therefore spanning the central 50% of the panel) is considerably lower at 0.18 p.p. (ranging from 0.12 p.p. for HICP_1Y and UNEM_1Y to 0.29 p.p. for RGDP_2Y). This indicates that, although there have been substantial differences between the average forecast errors for some forecasters, many are bunched together in a relatively tight range.

Although there has been considerable co-movement of errors, the cross-section dispersion of forecast errors has varied substantially across time, variable and horizon. Chart 1 shows, by variable and horizon, the median and individual forecast errors as well as the inverse of the mean absolute and squared errors. The impact of the global financial crisis is evident for all three variables, although most striking for real GDP growth. Another noteworthy feature is that the pattern of dispersion of forecast errors, although uniformly high (inverse low) around the period of the global financial crisis, has varied by variable and, to a lesser extent, by horizon. For HICP inflation, the cross section dispersion was low (inverse high) shortly before the global financial crisis (around Period 30) and toward the end of the sample period (around Period 70). For real GDP growth, the dispersion was lowest near the end of the sample, while, for the unemployment rate is was lowest in the first half of the sample.

[16] 1Y denotes one-year ahead, 2Y denotes two-years ahead. HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate.

4 Results

As noted above, we calculate the performance statistics for individual forecasters for the three variables (euro area HICP inflation, real GDP growth and the unemployment rate) for each horizon (one-year ahead and two-years ahead). In addition, as our metric is normalised for each of the variable horizon combinations, we can also compute statistics aggregating (a) across the three variables (HICP inflation, real GDP growth and the unemployment rate) for the one-year horizon and for the two-years ahead horizons – hereafter VARX1Y and VARX2Y, and (b) across the three variables (HICP inflation, real GDP growth and the unemployment rate) and both horizons (one-years and two-years ahead) – hereafter VARX1Y2Y.

4.1 Results for all forecasters

We provide the results for HICP inflation one-year ahead graphically (as do D'Agostino) in Chart 2, but for space reasons provide a summary of the results for the other variables and horizons in tabular form (Table 3). The left-hand panel of Chart 2 shows the cumulative distribution of forecasters actual forecast performance statistics as well as the 1% and 99% confidence bands surrounding these. For quite a large portion of the distribution, particularly close to the lower and upper ends, the actual performance statistics lie outside the confidence bands. This would suggest that some forecasters perform better and worse than would be expected if forecasting ability was indeed equal. The right-hand panel of Chart 2 shows the corresponding density distribution. From this, it can be seen that the distribution of actual forecast scores is relatively wide compared with the bootstrapped intervals.

Table 3 shows a summary the cumulative distributions of forecasters' actual forecast performance statistics as well as the 1% and 99% confidence bands surrounding these for each of the variable and horizons. To keep the table tractable, data are shown for seven points along the cumulative distribution (i.e. the best forecaster, the 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles and the worst forecaster – these are the same as shown in D'Agostino et al. (2012). The first row of Table 3 summarises the cumulative distribution for the one-year ahead inflation forecasts shown in Chart 2. From this it can be seen that, when averaging over the 76 outcomes (T), the 5^{th} and 25^{th}percentiles of the 77 forecasters (N) are below the 1% confidence band, while the 75^{th} and 95^{th} percentiles are above the 99% confidence band. The fact that some are below and others are above would suggest that some forecaster perform better and others worse than would be expected if forecasting ability was randomly distributed.

The broad pattern of some better and others worse is also evident for the other variables and horizons, including the synthetic aggregate variables (VARX1Y, VARX2Y and VARX1Y2Y). Taken at face value, the results in Table 3 would suggest that some forecasters have better or worse forecasting ability than what might be expected if forecasting ability was equal.

4.2 Aspects of the forecast performance statistics

As indicated above, the results presented in Table 3 are suggestive of differences in forecasting ability. In this section, we take a closer look at the forecast performance statistics particularly over different sub-samples and across variables and horizons.

As noted above, a key development during the period for which the ECB SPF has existed was the global financial crisis. Apart from the large errors around that time period, there were also notable differences in the period before the crisis (pre-crisis) and the period after the crisis (post-crisis). For example, in the pre-crisis period, HICP inflation was, on average, under-forecast (i.e. forecast errors – forecast minus actual – were negative on average), whereas in the post-crisis period HICP inflation was, on average, over-forecast.

Chart 3 shows that, while for both the pre-crisis and post-crisis samples the same broad pattern of some better and others worse was evident as for the whole sample, there appears to be little correlation in the performance rankings across sub-samples. The upper-left-hand panel of Chart 4 shows a scatter plot of the forecast rankings for one-year ahead HICP inflation with the rankings in the pre-crisis sample shown on the horizontal axis and those in the post-crisis sample on the vertical axis.[17] Although there is a very slightly positive fitted regression line its fit is very low. This reflects the fact that forecasters who ranked well in the pre-crisis sample did not necessarily rank well in the post-crisis sample. For example, there is a forecaster who was in the upper quintile in the pre-crisis sample but in the lower decile in the post-crisis sample (and another who was in the lower decile in the pre-crisis sample but in the upper quintile in the post crisis sample).

This pattern is repeated for other variables and horizons. The other panels of Chart 4 show that generally there was very little correlation in forecast performance across the pre-crisis and post-crisis sub-samples. Table 4 shows numerically that, with the possible exception of the two-year ahead HICP inflation forecasts at 0.33[18], there was essentially zero correlation for each of the three variables between the forecast performances in the period before the crisis (1999-2008) and the period after the crisis (2009-2018).

There was more evidence of correlations in forecasting performance across variables most notably for Okun's Law type (GDP and Unemployment) relationships but also for Philips Curve type (Inflation and GDP or unemployment) relationships. In contrast to the lack of correlation across sub-samples, there was positive correlation

[17] The rankings in the pre- and post-crisis periods were calculated for forecasters who provided at least twenty forecasts in each sub-sample. As an additional robustness check, they were calculated for forecasters who provided at least ten forecasts in each sub-sample, the results were qualitatively similar.

[18] Although even here there was a forecaster who was in the lower decile for the pre-crisis sample but in the upper decile for the post-crisis sample.

across variable combinations. These were strongest for real GDP and unemployment, which is in line with an Okun's Law type relationships – see Table 5. The correlation over the whole sample (0.40 for one-year ahead and 0.43 for two-years ahead), was also evident for the pre- and post-crisis sub-samples. There were also positive correlations between HICP inflation and real GDP growth and between HICP inflation and unemployment, but these were relatively low in most cases within the sub-samples.

The correlation across horizons (one-year ahead and two-years ahead) was relatively large (compared to that across sub-samples or across variables), particularly for unemployment and HICP inflation – see Table 6. The correlation between the forecast performance for one-year ahead and two-year head unemployment forecasts was 0.67 over the whole sample (0.54 in the pre-crisis and 0.68 in the post-crisis sub-sample). The correlation between one-year and two-year ahead forecasting performance was slightly lower for HICP inflation at 0.49. It was lowest for real GDP at 0.24 over the whole sample. This may reflect the lower degree of persistence in the real GDP growth time series relative to HICP inflation or the unemployment rate.

The lack of correlation in forecast performance across sub-samples is at odds with the findings from the bootstrap tests. The bootstrap tests suggested that a substantial portion of forecasters performed better or worse than what would be expected under the null hypothesis of equal ability. However, if these forecasters were truly better or worse, one might expect their performance to carry over across sub-samples. One possible explanation for this seeming paradox is that the rolling horizon nature of the SPF forecasts generates significant autocorrelation in the forecast errors.[19] Indeed, the first order autocorrelation coefficient of the forecast errors was positive for each of the variable horizon combinations. This has implications for the bootstrapping. In the next section, we attempt to adjust for the autocorrelation, and once we do so, we show that the bootstrap tests no longer suggested that many forecasters were better or worse than what would have been expected under the null hypothesis of equal ability.

4.3 Adjusting the bootstrap for autocorrelation

The rolling horizon nature of the forecasts automatically imparts serial autocorrelation into the forecast errors. To see this, consider a hypothetical example. Let us start with the one-year ahead HICP inflation forecast made in the Q1 2017 SPF round. As HICP inflation data were available for December 2016, the one-year ahead forecast is for December 2017. The one-year ahead horizons for the following three SPF rounds (i.e. Q2 2017/ Q3 2017/ Q4 2017) are March 2018/ June 2018/

[19] This was less of a problem for D'Agostino et al. (2012) particularly for their one-quarter ahead horizon, where forecast errors are more likely to be serially independent. While it should have been an issue for their one-year ahead horizon results, it does not appear to have been so strong. This may perhaps reflect their much longer sample period 1968-2009 (i.e. 40 years compared with 20 years in this paper).

September 2018. Now imagine there was a shock (for example a strong oil price increase) in October 2017, immediately after the Q4 2017 SPF was completed. This shock will affect the year-on-year inflation rate for the next twelve months (i.e. up to September 2018). Thus in this hypothetical example the forecast errors for the four SPF rounds 2017 are correlated.

This autocorrelation impacts the comparability of the bootstrapped confidence bands and the actual forecast performance statistics. This is because the former when reshuffled and reassigned will lose much of their autocorrelation. To see this consider the forecast errors of Forecaster A who significantly over-forecasts inflation in two consecutive rounds. Under the bootstrapping these will be reassigned to other forecasters but most likely to different forecasters in the two rounds thereby attenuating the effect of autocorrelation on the bootstrapped confidence bands relative to the actual forecast errors.

One way to eliminate or at least substantially reduce this autocorrelation would to only consider SPF rounds four quarters apart. In the hypothetical example above, when the SPF panellists make their inflation forecasts in the Q1 2018 SPF round, they will have observed the oil price shock in October 2017 and incorporated it into their projections. Thus, in principle, the errors from the Q1 2017 and Q1 2018 SPF rounds should not be correlated.[20]

Rows one to four of Table 7 show the results for considering the one-year ahead HICP inflation forecasts in the Q1, Q2, Q3 and Q4 SPF rounds respectively. For example, HICP1YQ4 is based on the forecast errors from the Q4 1999, Q4 2000, …, Q4 2016, Q4 2017 SPF rounds.[21] In this case, the distribution of actual forecast statistics never falls outside the confidence bands. Chart 5 shows the entire cumulative distribution for the Q1 rounds for the one-year ahead forecasts of HICP inflation, real GDP growth and the unemployment rate. In each case, the distribution of actual forecast statistics generally lies within the confidence bands over the entire distribution. This is in stark contrast to the pattern when the forecast errors are not adjusted for autocorrelation (see Chart 2).

In summary, once autocorrelation was controlled for (which impacted on the bootstrapped confidence bands but not the actual forecast statistics), there was a generalised lack of evidence for forecasters that perform significantly better or worse than what would be expected under the null of equal ability. This finding is more congruent with the lack of correlation in forecasting rank across sub-samples and is also more in line with the finding of D'Agostino et al. (2012).

[20] One potential drawback of this is approach is that it could change the sample and get rid of relevant information. However, this is likely to be of second-order magnitude compared with the autocorrelation problem.

[21] For the two-year ahead forecasts, to allow for the longer horizon, HICP2YQ1a is based on the forecast errors from the Q1 1999, Q1 2001, …, Q1 2015, Q1 2017 SPF rounds, whereas HICP2YQ1b is based on the forecast errors from the Q1 2000, Q1 2002, …, Q1 2016, Q1 2018 SPF rounds

4.4 A robustness check

In their paper, D'Agostino et al. (2012) consider an alternative metric of forecast performance – the absolute error, which is less penalising of forecasters with large outlier errors compared with the squared error metric. We also re-ran our calculations using the absolute error and, like them, found that the results were not qualitative changed from above (both with respect to the unadjusted and adjusted for correlation results).

We also considered an alternative metric, the percentile rank. In the earlier literature above, a significant limitation of using the rank as a metric was the need for a balanced panel.[22] However, using the percentile rank largely circumvents this - particularly when there is a relatively large panel (in our case between 77 and 63 depending on the variable and horizon). The percentile rank takes the rank of each forecaster and maps it to the percentile (1-100). Therefore, if there are 80 forecasts in one round but only 60 in another, the minimum-maximum scale of the percentile rank remains the same (i.e. 1-100) unlike the rank (1-80 or 1-60).

When not controlling for autocorrelation, the results using the percentile rank metric are largely similar to those using the squared error metric – see Table 8. For example, for one-year ahead HICP inflation, the actual 5th percentile lies below the confidence band, while the 75th and 95th percentiles lie above. This broad pattern of some performing better and others worse than what would be expected under the null hypothesis of equal ability is also apparent for the other variables and horizons.

However, owing to the rolling horizons, the percentile rank metric, like the squared error metric, will have in-built autocorrelation. To see this, imagine Forecaster A has a relatively high forecast but then there is a downward oil price shock. As was the case for the squared error metric, this shock will also impact the percentile rank of Forecaster A for at least four quarters. Therefore, we again re-run the bootstraps using the percentile rank but adjusting for autocorrelation – see Table 9 (for space reasons we only show the on-year ahead results, but they are qualitatively similar for the two-year ahead forecasts). In this case the results are even clearer cut as there are no instances where the actual distribution lies outside the confidence bands.

In summary, this robust check would support assessment above that, once autocorrelation was controlled for, there was a generalised lack of evidence for forecasters that perform significantly better or worse than what would be expected under the null of equal ability.

[22] In the case of the ECB SPF, although some forecasters have participated intensively, no forecaster has provided forecasts for each variable in every single round. Genre et al. (2013) interpolate missing forecasts in order to generate a reasonably sized balanced panel, but even doing this they report that their sample for GDP growth forecasts shrinks from 94 forecasters in the raw unbalanced to 33 forecasters in the filtered balanced panel.

5 Additional analyses

The main focus of this paper has been on testing whether it was possible to identify forecasters who were statistically significantly different from what might be expected under the null hypothesis of equal ability. Although, the results suggest it was not possible, there was some, albeit limited, correlation of the forecast performance statistics across sub-samples, variables and horizons. With limited information on the forecasters it is difficult to undertake an in-depth analysis but we consider a number of additional dimensions.

5.1 Correlation between forecast rank and number of forecasts provided

One piece of information we have for each forecaster is the number of forecasts that they have provided for each given variable and horizon. There are some arguments for why there might be a relationship between the number of forecasts a forecaster provides and his/her forecast performance. First, if one allows for learning, the more forecasts that are provided may lead to a better performance. Second, participation may be an indicator of enthusiasm and effort.

Table 10 shows that there is generally a negative correlation between the forecast rank of a forecaster and the number of forecasts he/she has provided. This means the more forecasts provided the lower (or better) is the rank. Although the correlation is generally quite small with a mean (median) of -0.16 (-0.17), it is negative in 22 of the 27 cases. The probability of getting 22 negative correlation coefficients out of 27 is very low - less than 0.1% if each is fully independent. This would suggest a positive, albeit limited, relationship between the number of forecasts provided and the forecast rank.

5.2 Can anything beat the simple average?

The results presented in this paper suggest it is difficult, even ex post, to identify ‘good’ or ‘bad’ forecasters. This supports the practice when publishing the SPF results of reporting of the simple average of all SPF respondents for each variable horizon. As additional information, we considered how this aggregate SPF forecast would perform is treated as an extra forecaster. Table 11 shows the rank of the aggregate SPF forecast compared with the distribution of individual forecasts. A striking feature is that it ranks in the upper quartile for nearly all variable and

horizons (the only exception being two-years ahead HICP inflation – HICP2Y – in the pre-crisis sample) and then its percentile rank was 27) and often ranks first for the synthetic combination of variables and horizons (VARX). This suggests that although, for any given variable horizon combination, the aggregate forecast might not be the 'best', its performance is high across all variable and horizon combinations. It is also in line with the finding by Genre et al. (2013) who find it is difficult, ex ante, to come up forecast combinations which beat the simple average.

5.3 Correlation between forecast rank and other forecaster characteristics

A last exercise we undertake is to check whether there is any link between forecast performance and forecasters' characteristics in terms of their country of location (although a majority of the panel are located in the euro area, a number of located in Europe but outside the euro area) or their sector (a majority of the panel come from the financial sector, although a number come from other types of forecasters such as research institutions). Table 12 reports the forecast performance of the aggregate SPF forecast as before, as well as aggregations based on location (euro area or non-euro area) and sector (financial or non-financial).

With respect to location, in all cases, the aggregate of euro area-based forecasters outperforms that of the non-euro area-based forecasters. This most likely reflects the small number of non-euro area-based forecasters. For example, out of the almost 4,000 individual one-year ahead HICP inflation forecasts provided since the inception of the ECB SPF, 12% come from non-euro area-based forecasters, whereas 88% come from euro area-based forecasters.

With respect to sector, the aggregate of financial forecasters generally outperforms the aggregate of non-financial forecasts, although not for one-year ahead HICP inflation and by a smaller margin than was the case for the euro area – non-euro area aggregates. The smaller margin may reflect the relatively larger portion of non-financial forecasters. Again using the example of the almost 4,000 individual one-year ahead HICP inflation forecasts provided since the inception of the ECB SPF, 41% come from non-financial forecasters, whereas 59% come from financial forecasters. In some instances (RGDP1Y, UNEM1Y and UNEM2Y) the financial aggregate performs better than the overall aggregate, although for HICP1Y, HICP2Y and RGDP2Y the overall aggregate performs better.

6 Summary and conclusions

This paper has addressed the question of whether differences in performance between ECB SPF forecasters reflect statistically significant differences in forecasting ability or sampling variation. Ultimately, once we control for the effect on the bootstrapping and Monte Carlo simulations of autocorrelation in the forecast errors, this paper argues that the null hypothesis of equal forecasting ability cannot be rejected. That is, there is no statistically significant evidence that some forecasters in the ECB SPF are better or worse than others. The inability to identify forecasters worse than others contrasts somewhat with findings for the US SPF. This may owe to the fact that, since its inception, the ECB SPF has endeavoured to include forecasting institutions (not just individuals) with experience. In the US SPF, which was initiated in 1968, participation declined over time and the survey was close to being discontinued until the Federal Reserve Bank of Philadelphia took over its administration in 1990.

The finding (once controlling for autocorrelation) of equal forecasting ability is congruent with the very low degree of correlation of forecast ranks across sub-samples. It is also noteworthy that the aggregate SPF forecast generally performs quite well particularly on average. This finding is also congruent with the analysis of Genre et al. (2013) who report that it is hard to find forecast combination methods that beat the simple average.

7 Tables

Table 1

Number of forecasters in the ECB SPF with at least [X] realised forecasts (by variable and horizon) between Q1 1999 and Q4 2018

	HICP1Y	HICP2Y	RGDP1Y	RGDP2Y	UNEM1Y	UNEM2Y
>0	104	101	104	100	101	97
>=10	88	83	87	81	82	79
>=20	77	70	73	68	69	63
>=30	63	46	60	51	54	42
>=40	45	37	45	35	40	31
>=50	34	25	32	25	27	23
>=60	22	17	24	16	19	13
>=70	8	2	8	2	6	2

Note: The maximum number of possible realised one-year (two-year) ahead forecasts is 76 (72). 1Y denotes one-year ahead, 2Y denotes two-years ahead. HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. To interpret the table, note that 77 forecasters have provided at least 20 forecasts for HICP inflation one-year ahead (see row 4, column 2).

Table 2

Summary forecast statistics from ECB SPF by variable and horizon (1999-2018)

(percentage points)

		HICP_1Y	HICP_2Y	RGDP_1Y	RGDP_2Y	UNEM_1Y	UNEM_2Y
mean error	min	-0.74	-0.56	-0.07	0.23	-0.24	-0.93
	25%	-0.33	-0.24	0.19	0.66	-0.02	-0.40
	mean	-0.19	-0.05	0.34	0.87	0.07	-0.22
	median	-0.19	-0.05	0.32	0.86	0.07	-0.22
	75%	-0.05	0.09	0.46	1.05	0.13	-0.08
	max	0.50	1.06	0.84	1.75	0.52	0.51
	range	1.24	1.62	0.91	1.52	0.76	1.44
	std dev	0.19	0.23	0.13	0.18	0.10	0.14
	IQR	0.28	0.32	0.27	0.39	0.15	0.32
absolute error	min	0.48	0.51	0.51	0.70	0.36	0.65
	25%	0.70	0.73	0.83	1.23	0.53	0.99
	mean	0.76	0.80	0.91	1.35	0.59	1.08
	median	0.76	0.79	0.90	1.34	0.59	1.07
	75%	0.82	0.88	1.00	1.52	0.65	1.18
	max	0.97	1.13	1.36	1.92	0.84	1.59
	range	0.49	0.62	0.85	1.22	0.48	0.94
	std dev	0.05	0.07	0.10	0.12	0.06	0.12
	IQR	0.12	0.15	0.17	0.29	0.12	0.20

Note: 1Y denotes one-year ahead, 2Y denotes two-years ahead, HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. IQR denotes inter-quartile (25^{th} to 75^{th} percentile) range.

Table 3

Results based on normalised MSE (mean squared error) metric (1999-2018)

(percentage points)

	Best	5	25	50	75	95	Worst	N	T
HICP1Y	0.61 0.53-0.78	0.69* 0.73-0.84	0.84* 0.86-0.93	0.97 0.96-1.02	1.14* 1.06-1.13	1.52* 1.17-1.34	1.61 1.24-1.76	77	76
HICP2Y	0.69 0.50-0.79	0.73 0.69-0.83	0.79* 0.85-0.93	0.96 0.95-1.02	1.18* 1.05-1.13	1.71* 1.19-1.47	2.70* 1.28-2.25	70	72
RGDP1Y	0.54 0.49-0.77	0.66* 0.68-0.82	0.82* 0.85-0.92	1.00 0.96-1.02	1.18* 1.06-1.14	1.45* 1.20-1.44	2.04* 1.28-2.03	73	76
RGDP2Y	0.64 0.54-0.79	0.73 0.72-0.85	0.86 0.86-0.93	0.97 0.95-1.01	1.15* 1.05-1.13	1.44* 1.18-1.40	2.67* 1.27-2.16	68	72
UNEM1Y	0.61 0.45-0.75	0.66* 0.66-0.81	0.78* 0.83-0.91	0.95* 0.95-1.02	1.25* 1.06-1.15	1.65* 1.20-1.46	2.45* 1.31-2.08	69	76
UNEM2Y	0.60 0.51-0.78	0.66* 0.69-0.83	0.81* 0.85-0.92	0.97 0.96-1.02	1.19* 1.05-1.14	1.72* 1.19-1.43	2.31* 1.27-1.94	63	72
VARX1Y	0.70 0.68-0.86	0.72* 0.80-0.90	0.89* 0.91-0.95	1.00 0.98-1.01	1.18* 1.04-1.08	1.50* 1.11-1.24	1.81* 1.15-1.48	69	228
VARX2Y	0.73 0.70-0.87	0.77* 0.82-0.90	0.86* 0.92-0.96	0.97* 0.98-1.01	1.18* 1.03-1.09	1.42* 1.10-1.23	1.92* 1.14-1.48	63	216
VARX1Y2Y	0.72* 0.77-0.91	0.75* 0.86-0.93	0.89* 0.94-0.97	1.01 0.98-1.01	1.13* 1.02-1.06	1.54* 1.07-1.16	1.97* 1.11-1.32	63	432

Note: * denotes that this lies outside the 1% or 99% confidence bands, which are shown underneath. Columns 2-7 show different percentiles (best, 5th, 25th, 50th, 75th, 95th and worst), N denotes the number of forecasters considered and T denotes the number of outcomes. 1Y denotes one-year ahead, 2Y denotes two-years ahead, HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. VARX denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate). VARX1Y2Y denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate) and the two horizons (one-year and two-years ahead).

Table 4

Correlations of forecast performance across pre- and post-crisis samples (1999-2008 and 2009-2018)

(percentage points)

	One-year ahead	Two-years ahead
HICP Inflation	0.05	0.33
Real GDP growth	-0.09	0.01
Unemployment rate	0.04	0.04
VARX	0.06	0.20
VARX1Y2Y	0.23	

Note: VARX denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate). VARX1Y2Y denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate) and the two horizons (one-year and two-years ahead). Pre- and post-crisis samples refer to 1999Q1 to 2008Q4 and 2009Q1 to 2018Q4 respectively.

Table 5

Correlation of forecast performance across variables_horizons – whole sample, pre- and post-crisis (1999-2018, 1999-2008 and 2009-2018)

(percentage points)

	One-year ahead	Two-years ahead
HICP and RGDP	0.25 (0.02 / 0.49)	0.28 (0.17 / 0.19)
HICP and UNEM	0.19 (0.06 / 0.05)	0.34 (0.16 / 0.11)
RGDP and UNEM	0.40 (0.49 / 0.28)	0.43 (0.20 / 0.26)

Note: HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate.

Table 6

Correlation of forecast performance across horizons by variable

(percentage points)

	Whole sample (1999-2018)	Pre-crisis (1999-2008)	Post-crisis (2009-2018)
HICP1Y + HICP2Y	0.49	0.45	0.48
RGDP1Y + RGDP2Y	0.24	0.14	0.30
UNEM1Y + UNEM2Y	0.67	0.54	0.70
VARX1Y + VARX2Y	0.67	0.54	0.68

Note: HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. 1Y denotes the one-year ahead horizon and 2Y denotes the two-year ahead horizon. VARX denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate). Pre- and post-crisis samples refer to 1999Q1 to 2008Q4 and 2009Q1 to 2018Q4 respectively.

Table 7

Results based on normalised MSE controlling for autocorrelation (1999-2018)

(percentage points)

	Best	5	25	50	75	95	Worst	N	T
HICP1YQ1	0.46 0.24-0.58	0.60 0.50-0.67	0.78 0.71-0.84	0.94 0.91-1.02	1.26 1.11-1.26	1.65 1.36-1.79	2.13 1.55-2.85	80	19
HICP1YQ2	0.38 0.25-0.61	0.62 0.53-0.71	0.79 0.75-0.86	0.97 0.91-1.01	1.23* 1.09-1.22	1.63 1.32-1.74	1.73 1.51-2.91	78	19
HICP1YQ3	0.36 0.27-0.61	0.62 0.50-0.69	0.78 0.73-0.85	0.91 0.89-1.00	1.29* 1.08-1.24	1.85 1.38-1.95	1.94 1.57-3.24	74	19
HICP1YQ4	0.45 0.27-0.65	0.60 0.54-0.73	0.80 0.77-0.87	0.95 0.91-1.01	1.15 1.07-1.21	1.68 1.30-1.69	2.44 1.50-2.85	78	19
HICP2YQ1a	0.11 0.06-0.46	0.36 0.30-0.54	0.66 0.59-0.76	0.86 0.84-1.00	1.31 1.11-1.36	2.34 1.54-2.41	4.27 1.77-5.27	73	9
HICP2YQ1b	0.33 0.07-0.55	0.44 0.36-0.63	0.71 0.67-0.82	0.92 0.87-1.01	1.21 1.09-1.29	1.86 1.45-2.16	1.97 1.72-3.72	72	9
HICP2YQ2a	0.23 0.09-0.52	0.55 0.36-0.62	0.75 0.63-0.79	0.94 0.84-1.01	1.20 1.07-1.31	1.81 1.48-2.40	3.27 1.82-4.85	65	9
HICP2YQ2b	0.21 0.16-0.57	0.61 0.41-0.66	0.81 0.69-0.83	0.98 0.87-1.01	1.13 1.07-1.26	1.75 1.41-2.09	2.70 1.66-4.10	66	9
HICP2YQ3a	0.44 0.14-0.50	0.52 0.37-0.61	0.70 0.61-0.78	0.89 0.82-1.00	1.21 1.08-1.33	1.63 1.48-2.29	3.63 1.81-4.25	60	9
HICP2YQ3b	0.47 0.16-0.61	0.63 0.46-0.68	0.80 0.70-0.83	0.96 0.87-1.01	1.21 1.07-1.27	1.91 1.35-1.97	2.45 1.61-4.11	61	9
HICP2YQ4a	0.35 0.15-0.57	0.51 0.42-0.66	0.76 0.68-0.82	0.88 0.86-0.99	1.16 1.07-1.26	1.93 1.42-2.15	2.63 1.71-4.18	62	9
HICP2YQ4b	0.44 0.22-0.52	0.54 0.39-0.63	0.72 0.64-0.81	0.89 0.85-1.01	1.18 1.08-1.31	1.94 1.45-2.29	3.33 1.73-5.44	63	9

	Best	5	25	50	75	95	Worst	N	T
RGDP1YQ1	0.34 0.17-0.56	0.54 0.45-0.65	0.68* 0.68-0.82	0.96 0.88-1.01	1.22 1.10-1.29	2.06* 1.40-2.01	2.47 1.71-4.04	77	19
RGDP1YQ2	0.51 0.18-0.55	0.63 0.47-0.65	0.77 0.69-0.83	0.96 0.88-1.01	1.20 1.11-1.28	1.62 1.40-1.88	2.54 1.64-3.17	78	19
RGDP1YQ3	0.47 0.23-0.65	0.64 0.52-0.73	0.76* 0.77-0.87	0.96 0.92-1.02	1.23* 1.09-1.22	1.55 1.31-1.73	1.61 1.46-2.78	73	19
RGDP1YQ4	0.42 0.17-0.61	0.56 0.48-0.69	0.76 0.74-0.85	0.91 0.89-1.01	1.15 1.07-1.22	1.99 1.37-2.00	3.02 1.62-4.14	74	19
RGDP2YQ1a	0.18 0.09-0.57	0.56 0.44-0.66	0.76 0.70-0.84	0.92 0.89-1.01	1.13 1.09-1.28	1.62 1.41-1.93	2.44 1.59-3.26	67	9
RGDP2YQ1b	0.39 0.11-0.59	0.64 0.44-0.69	0.81 0.72-0.85	0.96 0.89-1.01	1.18 1.07-1.25	1.99 1.39-2.01	2.34 1.63-3.64	67	9
RGDP2YQ2a	0.33 0.11-0.55	0.53 0.42-0.63	0.75 0.66-0.80	0.95 0.86-0.99	1.22 1.09-1.30	1.73 1.47-2.16	2.29 1.73-4.15	69	9
RGDP2YQ2b	0.17 0.07-0.56	0.49 0.42-0.65	0.76 0.68-0.82	0.93 0.85-0.98	1.06 1.02-1.22	1.82 1.35-2.56	4.86 1.85-9.28	71	9
RGDP2YQ3a	0.47 0.13-0.55	0.59 0.41-0.65	0.72 0.65-0.80	0.92 0.84-0.99	1.13 1.05-1.28	1.98 1.41-2.22	5.51* 1.81-4.83	57	9
RGDP2YQ3b	0.20 0.05-0.49	0.41 0.32-0.60	0.68 0.60-0.77	0.82 0.82-1.00	1.28 1.07-1.35	1.67 1.50-2.57	3.08 1.86-5.63	57	9
RGDP2YQ4a	0.37 0.14-0.59	0.55 0.46-0.67	0.76 0.70-0.83	0.95 0.87-1.00	1.22 1.07-1.28	1.68 1.39-1.98	2.12 1.65-4.68	60	9
RGDP2YQ4b	0.29 0.06-0.52	0.62* 0.37-0.60	0.76 0.64-0.79	0.88 0.83-1.00	1.21 1.06-1.30	1.84 1.47-2.44	4.35 1.82-5.25	63	9

	Best	5	25	50	75	95	Worst	N	T
UNEM1YQ1	0.41 0.15-0.54	0.55 0.43-0.66	0.72 0.68-0.83	0.91 0.88-1.01	1.24 1.07-1.27	1.67 1.39-2.01	3.72 1.69-4.18	75	19
UNEM1YQ2	0.38 0.15-0.57	0.52 0.41-0.65	0.73 0.71-0.84	0.94 0.90-1.02	1.32* 1.10-1.26	1.86 1.40-1.97	2.96 1.62-3.86	72	19
UNEM1YQ3	0.39 0.17-0.57	0.54 0.44-0.66	0.74 0.69-0.83	0.95 0.87-1.00	1.14 1.08-1.27	1.91 1.44-2.09	2.78 1.67-3.88	71	19
UNEM1YQ4	0.25 0.14-0.57	0.49 0.43-0.65	0.76 0.69-0.82	0.96 0.87-1.00	1.16 1.08-1.28	2.10 1.43-2.16	2.77 1.70-4.53	72	19
UNEM2YQ1a	0.23 0.09-0.50	0.35 0.35-0.59	0.70 0.63-0.80	0.92 0.84-1.00	1.24 1.11-1.33	1.77 1.52-2.30	2.88 1.77-4.08	68	9
UNEM2YQ1b	0.39 0.02-0.41	0.43 0.24-0.53	0.64 0.56-0.73	0.85 0.79-0.99	1.30 1.09-1.40	2.63* 1.55-2.60	4.56 1.95-7.28	64	9
UNEM2YQ2a	0.39 0.10-0.54	0.58 0.39-0.62	0.76 0.65-0.80	0.89 0.84-1.01	1.25 1.09-1.32	1.91 1.47-2.13	2.85 1.73-3.98	63	9
UNEM2YQ2b	0.27 0.05-0.51	0.45 0.33-0.60	0.69 0.62-0.79	0.94 0.84-1.00	1.32 1.09-1.34	1.91 1.45-2.23	2.60 1.77-5.04	60	9
UNEM2YQ3a	0.25 0.12-0.57	0.63 0.43-0.65	0.78 0.66-0.80	0.90 0.84-1.00	1.22 1.07-1.32	1.78 1.42-2.19	2.42 1.70-3.54	55	9
UNEM2YQ3b	0.38 0.06-0.48	0.44 0.34-0.59	0.67 0.59-0.77	0.85 0.83-1.01	1.41* 1.11-1.38	2.12 1.49-2.29	3.69 1.76-3.87	56	9
UNEM2YQ4a	0.41 0.09-0.54	0.50 0.39-0.64	0.72 0.63-0.79	0.87 0.81-0.98	1.18 1.05-1.31	2.16 1.44-2.32	3.79 1.87-5.83	56	9
UNEM2YQ4b	0.34 0.08-0.50	0.50 0.35-0.60	0.74 0.61-0.78	0.92 0.82-1.01	1.18 1.09-1.34	1.92 1.45-2.30	3.48 1.75-4.38	56	9

	Best	5	25	50	75	95	Worst	N	T
VARX1YQ1	0.59 0.42-0.72	0.67 0.61-0.78	0.86 0.82-0.91	1.00 0.94-1.02	1.12 1.07-1.16	1.51 1.24-1.56	2.29 1.35-2.29	74	57
VARX1YQ2	0.66 0.45-0.73	0.67 0.64-0.80	0.83 0.83-0.91	1.00 0.95-1.02	1.20* 1.07-1.16	1.55* 1.23-1.51	1.92 1.33-2.16	72	57
VARX1YQ3	0.59 0.48-0.75	0.71 0.65-0.81	0.85 0.84-0.92	0.98 0.95-1.02	1.18* 1.06-1.15	1.43 1.21-1.49	2.11* 1.31-2.11	71	57
VARX1YQ4	0.63 0.46-0.74	0.72 0.65-0.80	0.83 0.83-0.91	0.97 0.94-1.01	1.20* 1.06-1.16	1.56 1.23-1.58	1.96 1.35-2.52	71	57
VARX2YQ1a	0.42 0.34-0.68	0.53* 0.56-0.75	0.78 0.78-0.88	0.94 0.93-1.02	1.17 1.08-1.21	1.79* 1.28-1.65	2.54* 1.41-2.49	65	27
VARX2YQ1b	0.48 0.35-0.69	0.67 0.55-0.76	0.83 0.77-0.88	0.94 0.91-1.01	1.20 1.07-1.21	1.70* 1.28-1.67	2.13 1.45-3.31	64	27
VARX2YQ2a	0.60 0.37-0.68	0.71 0.59-0.76	0.83 0.77-0.88	0.95 0.91-1.02	1.25* 1.07-1.21	1.64* 1.28-1.62	1.72 1.42-2.60	60	27
VARX2YQ2b	0.58 0.36-0.71	0.71 0.60-0.78	0.81 0.79-0.89	0.99 0.92-1.02	1.18 1.06-1.19	1.81* 1.26-1.66	2.30 1.40-2.61	59	27
VARX2YQ3a	0.63 0.32-0.70	0.70 0.55-0.75	0.83 0.77-0.89	0.98 0.91-1.01	1.16 1.07-1.22	1.70 1.29-1.79	2.63 1.40-2.68	54	27
VARX2YQ3b	0.56 0.39-0.69	0.66 0.58-0.76	0.83 0.77-0.89	0.94 0.92-1.02	1.24* 1.07-1.22	1.55 1.26-1.66	2.77* 1.42-2.68	55	27
VARX2YQ4a	0.63 0.42-0.73	0.71 0.60-0.76	0.82 0.79-0.90	0.97 0.91-1.01	1.16 1.05-1.18	1.71 1.27-1.80	2.26 1.37-2.64	54	27
VARX2YQ4b	0.39 0.36-0.68	0.68 0.59-0.76	0.84 0.77-0.89	0.95 0.91-1.02	1.16 1.08-1.22	1.73* 1.26-1.65	2.23 1.40-2.44	56	27

Note: * denotes that this lies outside the 1% or 99% confidence bands, which are shown underneath. Columns 2-7 show different percentiles (best, 5th, 25th, 50th, 75th, 95th and worst), N denotes the number of forecasters considered and T denotes the number of outcomes. HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. VARX denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate). 1Y denotes the one-year ahead horizon and 2Y denotes the two-year ahead horizon. Q1 denotes the Q1 over Q1 comparisons, Q2 the Q2 over Q2, etc. For the two-year ahead forecasts, to allow for the longer horizon, HICP2YQ1a is based on the forecast errors from the Q1 1999, Q1 2001, …, Q1 2015, Q1 2017 SPF rounds, whereas HICP2YQ1b is based on the forecast errors from the Q1 2000, Q1 2002, …, Q1 2016, Q1 2018 SPF rounds.

Table 8

Results based on percentile rank metric (1999-2018)

(percentage points)

	Best	5	25	50	75	95	Worst	N	T
HICP1Y	0.32* 0.32-0.44	0.39* 0.41-0.46	0.47 0.47-0.49	0.51 0.50-0.52	0.56* 0.53-0.55	0.65* 0.56-0.60	0.68 0.58-0.69	77	76
HICP2Y	0.34 0.32-0.43	0.39* 0.40-0.46	0.47 0.47-0.49	0.50 0.50-0.52	0.58* 0.53-0.55	0.68* 0.57-0.62	0.75* 0.59-0.70	70	72
RGDP1Y	0.36 0.33-0.44	0.39* 0.41-0.46	0.47 0.47-0.49	0.51 0.50-0.52	0.55* 0.53-0.55	0.62* 0.56-0.62	0.63 0.58-0.69	73	76
RGDP2Y	0.35 0.33-0.44	0.42 0.40-0.46	0.46* 0.47-0.49	0.51 0.50-0.52	0.57* 0.53-0.56	0.64* 0.57-0.61	0.67 0.59-0.70	68	72
UNEM1Y	0.36 0.33-0.44	0.40* 0.41-0.46	0.47 0.47-0.49	0.52 0.50-0.52	0.56* 0.53-0.55	0.65* 0.56-0.62	0.71* 0.58-0.69	69	76
UNEM2Y	0.28* 0.32-0.44	0.39* 0.41-0.46	0.48 0.47-0.49	0.51 0.50-0.52	0.57* 0.53-0.56	0.67* 0.56-0.62	0.77* 0.59-0.71	63	72
VARX1Y	0.41 0.41-0.47	0.44* 0.45-0.48	0.49 0.49-0.50	0.52* 0.51-0.52	0.53 0.52-0.54	0.62* 0.54-0.57	0.64* 0.55-0.62	69	228
VARX2Y	0.40* 0.40-0.47	0.44* 0.45-0.48	0.48* 0.49-0.50	0.52* 0.51-0.52	0.55* 0.52-0.54	0.61* 0.54-0.58	0.70* 0.56-0.63	63	216
VARX1Y2Y	0.39* 0.44-0.48	0.45* 0.47-0.49	0.50 0.49-0.51	0.52 0.51-0.52	0.54* 0.52-0.53	0.59* 0.54-0.56	0.70* 0.54-0.59	63	432

Note: * denotes that this lies outside the 1% or 99% confidence bands, which are shown underneath. Columns 2-7 show different percentiles (best, 5th, 25th, 50th, 75th, 95th and worst), N denotes the number of forecasters considered and T denotes the number of outcomes. VARX denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate). 1Y denotes one-year ahead, 2Y denotes two-years ahead, HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate.

Table 9

Results based on percentile rank metric adjusting for autocorrelation (1999-2018)

(percentage points)

	Best	5	25	50	75	95	Worst	N	T
HICP1YQ1	0.27 0.17-0.37	0.36 0.32-0.41	0.45 0.42-0.47	0.50 0.49-0.52	0.58 0.55-0.59	0.68 0.61-0.70	0.79 0.65-0.85	80	19
HICP1YQ2	0.16 0.15-0.36	0.35 0.31-0.40	0.45 0.43-0.47	0.52 0.49-0.53	0.59 0.55-0.59	0.67 0.61-0.70	0.77 0.66-0.87	78	19
HICP1YQ3	0.25 0.16-0.36	0.37 0.30-0.40	0.46 0.43-0.47	0.52 0.49-0.53	0.58 0.55-0.59	0.71 0.62-0.72	0.83 0.66-0.87	74	19
HICP1YQ4	0.20 0.15-0.37	0.36 0.31-0.41	0.46 0.43-0.47	0.50 0.49-0.53	0.57 0.55-0.59	0.67 0.61-0.71	0.85 0.66-0.86	78	19
	Best	**5**	**25**	**50**	**75**	**95**	**Worst**	**N**	**T**
RGDP1YQ1	0.25 0.15-0.37	0.35 0.32-0.41	0.45 0.43-0.47	0.52 0.49-0.53	0.57 0.55-0.60	0.72* 0.61-0.70	0.73 0.66-0.86	77	19
RGDP1YQ2	0.33 0.14-0.35	0.39 0.31-0.40	0.47 0.43-0.47	0.51 0.49-0.53	0.57 0.55-0.59	0.65 0.61-0.71	0.75 0.66-0.87	78	19
RGDP1YQ3	0.28 0.16-0.37	0.37 0.30-0.40	0.46 0.42-0.47	0.52 0.50-0.53	0.59 0.55-0.60	0.65 0.62-0.73	0.71 0.66-0.88	73	19
RGDP1YQ4	0.26 0.15-0.36	0.34 0.30-0.41	0.44 0.43-0.47	0.52 0.49-0.53	0.58 0.55-0.59	0.66 0.62-0.72	0.74 0.65-0.88	74	19
	Best	**5**	**25**	**50**	**75**	**95**	**Worst**	**N**	**T**
UNEM1YQ1	0.30 0.14-0.37	0.36 0.31-0.41	0.45 0.43-0.47	0.52 0.49-0.53	0.59 0.55-0.59	0.67 0.61-0.71	0.75 0.66-0.87	75	19
UNEM1YQ2	0.28 0.15-0.37	0.37 0.30-0.40	0.44 0.43-0.47	0.51 0.49-0.53	0.59 0.55-0.60	0.73 0.62-0.73	0.78 0.66-0.87	72	19
UNEM1YQ3	0.27 0.16-0.36	0.37 0.30-0.40	0.46 0.42-0.47	0.52 0.50-0.53	0.59 0.55-0.60	0.65 0.62-0.73	0.76 0.66-0.87	71	19
UNEM1YQ4	0.21 0.15-0.36	0.33 0.30-0.40	0.47 0.42-0.47	0.51 0.49-0.53	0.58 0.55-0.60	0.67 0.62-0.73	0.87 0.66-0.88	72	19

Note: * denotes that this lies outside the 1% or 99% confidence bands, which are shown underneath. Columns 2-7 show different percentiles (best, 5th, 25th, 50th, 75th, 95th and worst), N denotes the number of forecasters considered and T denotes the number of outcomes. HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. 1Y denotes the one-year ahead horizon and 2Y denotes the two-year ahead horizon. Q1 denotes the Q1 over Q1 comparisons, Q2 the Q2 over Q2, etc.

Table 10

Correlation between number of forecasts provided and performance rank (1999-2018)

(percentage points)

	One-year ahead	Two-years ahead
HICP inflation	-0.11 (-0.03 / 0.02)	-0.23 (-0.14 / 0.04)
Real GDP growth	-0.04 (-0.35 / -0.30)	-0.09 (0.04 / -0.08)
Unemployment rate	-0.22 (-0.09 / -0.42)	-0.26 (0.08 / -0.48)
VARX	-0.17 (-0.19 / -0.25)	-0.21 (0.02 / -0.26)
VARX1Y2Y	-0.23 (-0.02 / -0.32)	

Note: VARX denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate). VARX1Y2Y denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate) and the two horizons (one-year and two-years ahead).

Table 11

Rank of SPF aggregate forecast / out of N+1 forecasters

	Whole sample (1999-2018)	Pre-crisis (1999-2008)	Post-crisis (2009-2018)
HICP1Y	7 / 78	8 / 60	7 / 47
RGDP1Y	5 / 74	5 / 58	5 / 45
UNEM1Y	6 / 70	3 / 54	8 / 40
VARX1Y	1 / 70	3 / 54	1 / 39
HICP2Y	4 / 71	13 / 49*	7 / 35
RGDP2Y	5 / 69	9 / 49	1 / 37
UNEM2Y	5 / 64	4 / 47	7 / 35
VARX2Y	1 / 64	4 / 47	1 / 34
VARX1Y2Y	1 / 64	1 / 47	1 / 33

Note: N denotes the number of forecasters considered. 1Y denotes one-year ahead, 2Y denotes two-years ahead, HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. VARX denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate). VARX1Y2Y denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate) and the two horizons (one-year and two-years ahead). Pre- and post-crisis samples refer to 1999Q1 to 2008Q4 and 2009Q1 to 2018Q4 respectively. * For HICP inflation two-years ahead (HICP2Y) in the pre-crisis period, the percentile rank of the aggregate SPF forecast was 27 (marginally outside the upper quartile).

Table 12

Rank of SPF forecast aggregations by sector and location / out of N+5 forecasters (1999-2018)

	HICP1Y	HICP2Y	RGDP1Y	RGDP2Y	UNEM1Y	UNEM2Y
Aggregate	9	5	7	5	8	7
Financial	11	7	5	6	4	4
Non-financial	5	9	10	8	11	13
Euro Area	8	4	6	7	7	6
Non-Euro Area	16	28	14	12	20	28

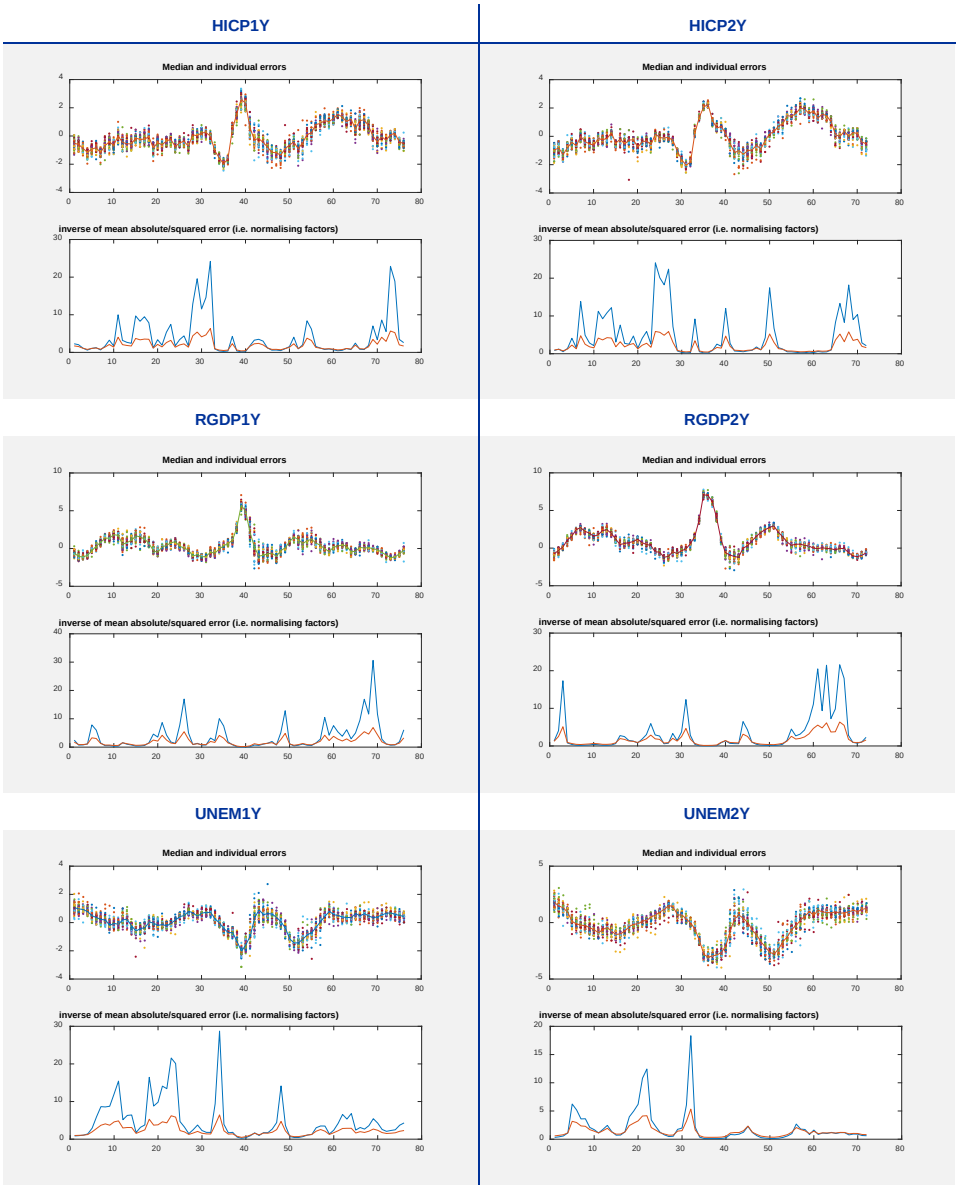
Note: N denotes the number of forecasters considered. 1Y denotes one-year ahead, 2Y denotes two-years ahead, HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate.

8 Charts

Chart 1

By variable and horizon: upper panels - median and individual forecast errors; lower panels – inverse of mean absolute and squared error (1999-2018)

(percent, percentage points)

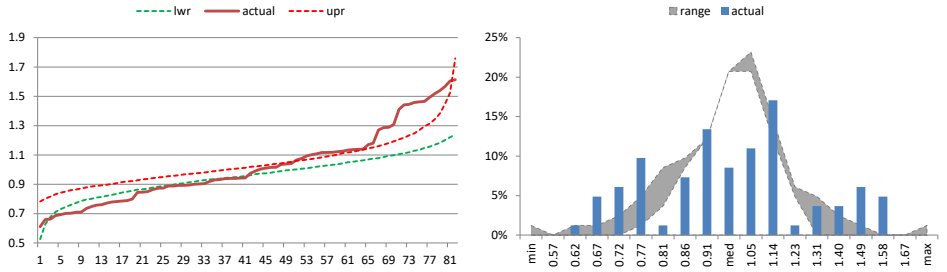


Note: 1Y denotes one-year ahead, 2Y denotes two-years ahead, HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. Horizontal axis denotes SPF rounds (from 1 – Q1 1999 to Q4 2018)

Chart 2

Graphical results of actual and bootstrapped distributions for HICP1Y - unadjusted (1999-2018)

(percent, percentage points)

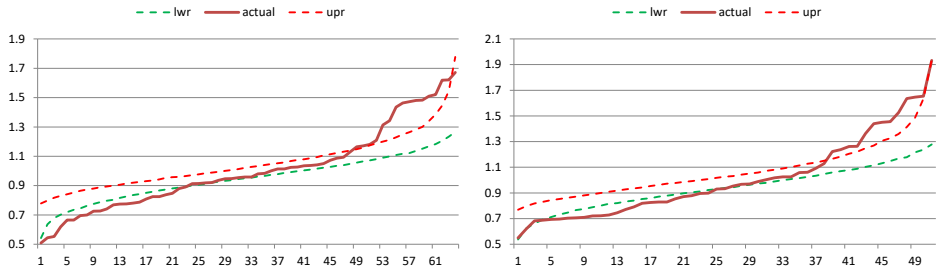


Note: The left-hand panel shows the cumulative distribution of forecasters' actual forecast performance statistics as well as the 1% and 99% confidence bands (derived from bootstrapping) surrounding these. The right-hand panel shows the corresponding density distribution. Unadjusted means not adjusted for autocorrelation – see below.

Chart 3

Graphical results of actual and bootstrapped distributions for HICP1Y – unadjusted – pre- and post-crisis samples (1999-2008 and 2009-2018 lhs and rhs panels respectively)

(percent, percentage points)

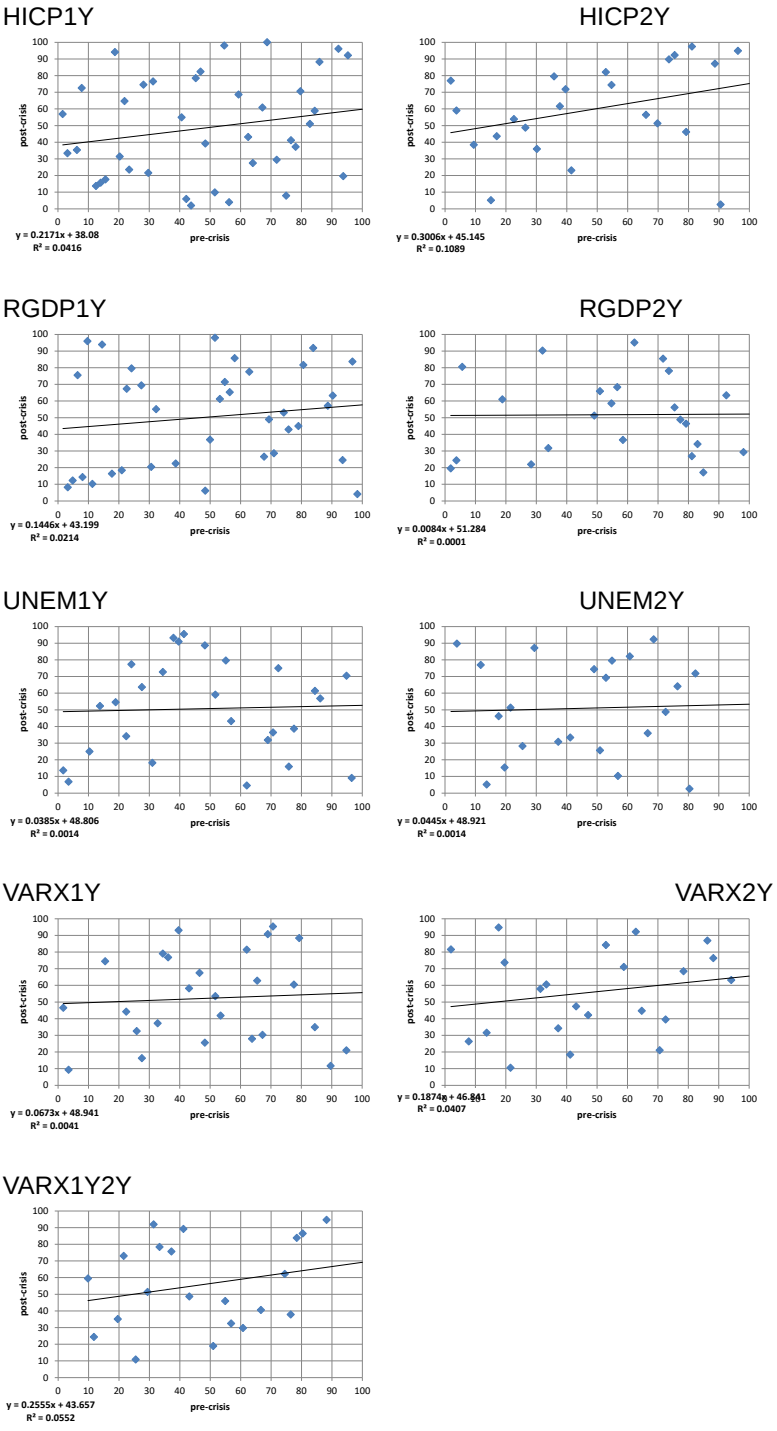


Note: The chart shows the cumulative distribution of forecasters' actual forecast performance statistics as well as the 1% and 99% confidence bands (derived from bootstrapping) surrounding these. Unadjusted means not adjusted for autocorrelation – see below.

Chart 4

Correlation between individual forecast ranks in pre- and post-crisis samples

(percent, percentage points)

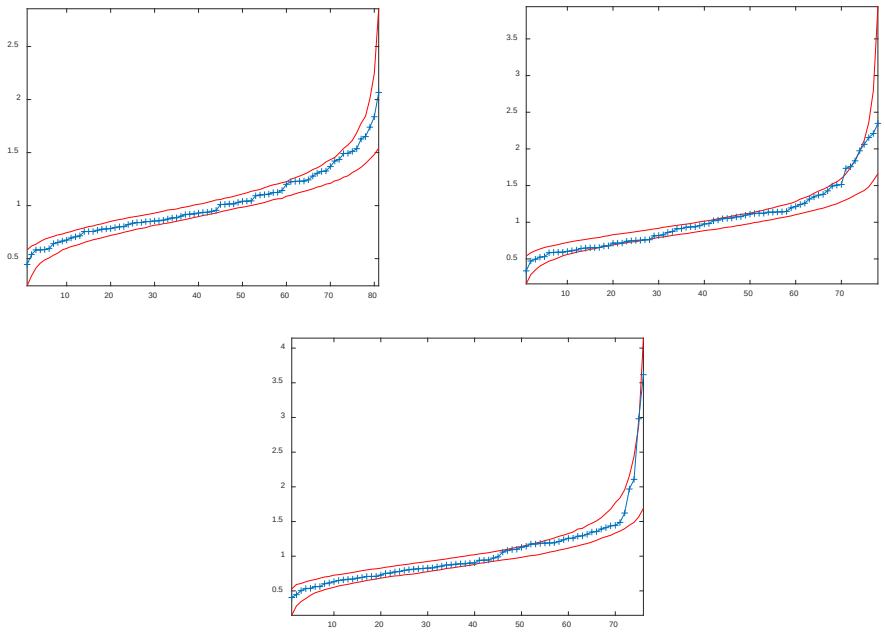


Note: 1Y denotes one-year ahead, 2Y denotes two-years ahead, HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. VARX denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate). VARX1Y2Y denotes the aggregated score across the three variables (HICP inflation, real GDP growth and the unemployment rate) and the two horizons (one-year and two-years ahead). Pre- and post-crisis samples refer to 1999Q1 to 2008Q4 and 2009Q1 to 2018Q4 respectively.

Chart 5

Graphical results of actual and bootstrapped distributions correcting for autocorrelation – results for HICP1YQ1, RGDP1YQ1 and UNEM1YQ1 (1999-2018)

(percent, percentage points)



Note: The chart shows the cumulative distribution of forecasters' actual forecast performance statistics as well as the 1% and 99% confidence bands (derived from bootstrapping) surrounding these. HICP denotes HICP inflation, RGDP denotes real GDP and UNEM the unemployment rate. 1Y denotes the one-year ahead horizon. Q1 denotes the Q1 over Q1 comparisons.

9 References

Abel, J., Rich, R., Song, J. and Tracy, J., “The Measurement and Behaviour of Uncertainty: Evidence from the ECB Survey of Professional Forecasters”, Journal of Applied Econometrics, Vol. 31(3), April/May 2016, pp. 533-550

Batchelor, Roy A. (1990) “All Forecasters Are Equal.” Journal of Business and Economic Statistics, 8, 143–44.

Beechey, M. J., B. K. Johannsen and A. T. Levin (2011) “Are Long-Run Inflation Expectations Anchored More Firmly in the Euro Area than in the United States?”, American Economic Journal: Macroeconomics Vol. 3, No. 2 (April), pp. 104-129.

Bonham, Carl, and Richard Cohen. (2001) “To Aggregate, Pool, or Neither: Testing the Rational Expectations Hypothesis Using Survey Data.” Journal of Business and Economic Statistics, 19, 278–91.

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A. and Rautanen, T., “An Evaluation of the Growth and Unemployment Forecasts in the ECB Survey of Professional Forecasters”, OECD Journal: Journal of Business Cycle Measurement and Analysis, Vol. 2010/2, OECD, December 2010

Christensen, H. Jens, Francis X. Diebold, Georg H. Strasser, and Glenn D. Rudebusch. (2008) “Multivariate Comparison of Predictive Accuracy.” Working Paper available at www.econ.uconn.edu/Seminar%20Series/strasser08.pdf.

Clements, Michael P. (2014) Forecast Uncertainty—Ex Ante and Ex Post: U.S. Inflation and Output Growth, Journal of Business & Economic Statistics, 32, 2, (206).

Croushore, D. (2009) “Philadelphia Fed Forecasting Surveys: Their Value For Research”, Paper presented to the International Workshop on Expectation Formation, Federal Reserve Bank of Philadelphia, 26-27 February.

Cuthbertson, Keith, Dirk Nitzsche, and Niall O’Sullivan. (2008) “UK Mutual Fund Performance: Skill or Luck?” Journal of Empirical Finance, 15, 613–34.

D’Agostino, Antonello, Kieran McQuinn and Karl Whelan. (2012) Are Some Forecasters Really Better Than Others? Journal of Money, Credit and Banking, Vol. 44, No 4 June pp. 715-732.

de Vincent-Humphreys, R., E. Falck, I. Dimitrova and L. Henkel (2019), Results of the third special questionnaire for participants in the ECB SPF, mimeo, February.

Diebold, Francis X., and Mariano Roberto. (1995) “Comparing Predictive Accuracy.” Journal of Business and Economic Statistics, 13, 253–63.

Diebold F. X., A. S. Tay and K. F. Wallis (1999), “Evaluating density forecasts of inflation: the survey of professional forecasters”, in Engle R.F. and H. White (eds.),

Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger, Oxford University Press, 76–90.

Dovern, J. and G. Kenny (2017) “The long-term distribution of expected inflation in the euro area: what has changed since the great recession?” ECB Working Paper Series No 1999, January.

Fama, Eugene F., and Kenneth French. (2010) “Luck versus Skill in the Cross Section of Mutual Fund Returns.” Journal of Finance, 65, 1915–47.

Franses, P. H. and N Maassen (2015) Consensus forecasters: How good are they individually and why? Erasmus School of Economics Econometric Institute Report 2015-21

Frenkel, M., Lis, E.M. and Rülke, J.-C., “Has the economic crisis of 2007-2009 changed the expectation formation process in the Euro area?”, Economic Modelling, Vol. 28(4), July 2011, pp. 1808-1814

Gamber, Edward N., Jeffrey P. Liebner, Julie K. Smith, (2015) The distribution of inflation forecast errors, Journal of Policy Modeling, Vol. 37, No. 1, pp. 47-64.

Garcia, Juan-Angel (2003) An introduction to the ECB's survey of professional forecasters, ECB Occasional Paper Series No 8, September.

Genre, Véronique, Geoff Kenny, Aidan Meyler and Allan Timmermann. (2013) Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting Vol 29, No. 1, Pages 108-121.

Glas, A. and Hartmann, M., “Inflation uncertainty, disagreement and monetary policy: Evidence from the ECB Survey of Professional Forecasters”, Journal of Empirical Finance, Vol. 39(B), December 2016, pp. 215-228

Grishchenko. O. V., S. Mouabbi and J.-P. Renne, 2017. "Measuring Inflation Anchoring and Uncertainty : A US and Euro Area Comparison," Finance and Economics Discussion Series 2017-102, Board of Governors of the Federal Reserve System.

Grothe, M. and A. Meyler, “Inflation Forecasts: Are Market-Based and Survey-Based Measures Informative?”, International Journal of Financial Research, Vol. 9(1), January 2018

Keane, Michael, and David Runkle. (1990) “Testing the Rationality of Price Forecasts: New Evidence from Panel Data.” American Economic Review, 80, 714–35.

Kenny, G., T. Kostka, and F. Masera, (2015) “Density forecast features and density forecast performance: A panel analysis” Empirical Economics, Vol. 48 (3), pp.1203-1231

Kosowski, Robert, Allan Timmerman, Russ Wermers, and Hall White. (2006) “Can Mutual Fund ‘Stars’ Really Pick Stocks? New Evidence from a Bootstrap Analysis.” Journal of Finance, 56, 2551–95.

Łyziak, T. and Paloviita, M., “Anchoring of inflation expectations in the Euro Area: Recent evidence based on survey data”, European Journal of Political Economy, Vol. 46(C), January 2017, pp. 52-73

Meyler, A., A. Melemenidis and M. Karber (2014) Results of the second special questionnaire for participants in the ECB SPF, mimeo, January

Meyler, A. and I. Rubene (2009) Results of a special questionnaire for participants in the ECB SPF, mimeo, April

Reitz, S., Rülke, J.-C. and Stadtmann, G., “Nonlinear expectations in speculative markets – Evidence from the ECB survey of professional forecasters”, Journal of Economic Dynamics and Control, Vol. 36(9), September 2012, pp. 1349-1363

Rich, R. and Tracy, J., “A Closer Look at the Behaviour of Uncertainty and Disagreement: Micro Evidence from the Euro Area”, Working Papers, No 1811, Federal Reserve Bank of Dallas, July 2018

Stekler, Herman. (1987) “Who Forecasts Better?” Journal of Business and Economic Statistics, 5, 155–158.

Stock, James H. and Mark W. Watson (2002). Has the Business Cycle Changed and Why? In M. Gertler and K. Rogoff (Eds.), NBER Macroeconomics Annual 2002. MIT Press.

Stock, James, and Mark Watson. (2005) “Has Inflation Become Harder to Forecast?” ‘Quantitative Evidence on Price Determination, Conference of Board of Governors of the Federal Reserve Board, September 29–30, Washington, DC.

Stock, James, and Mark Watson. (2006) “Why Has U.S. Inflation Become Harder to Forecast?” NBER Working Paper No.12324.

Zarnowitz, Victor, and Philip Bruan. (1993) “Twenty-Two Years of the NBER-ASA Quarterly Economic Outlook Surveys.” In Business Cycle Indicators and Forecasting, edited by James Stock, and Mark Watson. Chicago: University of Chicago Press.

Acknowledgements

The author is grateful to Antonello D'Agostino for sharing the Matlab code used for the D'Agostino et al (2012) paper. The author is grateful for comments from Thomas Westermann, Rupert de Vincent-Humphreys, Elisabeth Falck, Ivelina Dimitrova and Lukas Henkel as well as from an anonymous referee. All remaining errors are under the full responsibility of the author.

Aidan Meyler
European Central Bank, Frankfurt am Main, Germany; email: aidan.meyler@ecb.europa.eu

PDF ISBN 978-92-899-4014-6 ISSN 1725-2806 doi:10.2866/20544 QB-AR-20-023-EN-N