

Survey-based forecast distributions for Euro Area growth and inflation: ensembles versus histograms

Fabian Krüger¹

Received: 16 February 2016 / Accepted: 3 October 2016 / Published online: 17 January 2017
© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract Ensemble methods can be used to construct a forecast distribution from a collection of point forecasts. They are used extensively in meteorology, but have received little direct attention in economics. In a real-time analysis of the ECB’s Survey of Professional Forecasters, we compare ensemble methods to histogram-based forecast distributions of GDP growth and inflation in the Euro Area. We find that ensembles perform very similarly to histograms, while being simpler to handle in practice. Given the wide availability of surveys that collect point forecasts but not histograms, these results suggest that ensembles deserve further investigation in economics.

Keywords Forecasting · Survey data · Macroeconomics

JEL Classification C52 · C53

1 Motivation

Quantifying the uncertainty around economic forecasts has recently received much attention (e.g., [Jurado et al. 2015](#); [Müller and Watson 2016](#)). The most general information on uncertainty is provided by forecast distributions, such as the ‘fan charts’ of inflation issued by the Bank of England. A key benefit of forecast distributions is that each forecast user—such as a pension fund manager or a private consumer—can extract the specific piece of information that is relevant to them.

✉ Fabian Krüger
fabian.krueger@awi.uni-heidelberg.de

¹ Alfred-Weber-Institute for Economics, Heidelberg University, Bergheimer Straße 58, 69115 Heidelberg, Germany

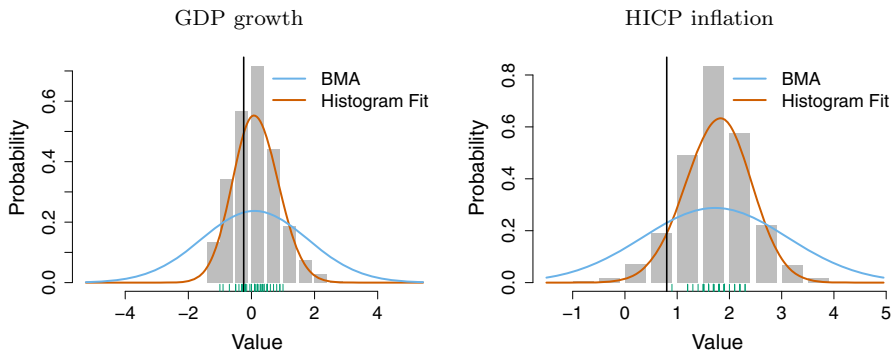


Fig. 1 *Left panel* Forecast distributions for the GDP growth rate in 2013:Q3, made in February 2013. *Right panel* Forecast distributions for the inflation rate of the harmonized index of consumer prices (HICP) in December 2013, made in February 2013. *Both panels* Gray bars represent ECB-SPF histogram forecasts; small vertical green lines at bottom are ECB-SPF point forecasts; vertical black lines are realizing observations

The present paper seeks to construct forecast distributions from the European Central Bank's Survey of Professional Forecasters (SPF). The data contain point forecasts made by several participants, and we consider so-called ensemble methods (e.g., [Raftery et al. 2005](#)) for converting these into a forecast distribution. Furthermore, SPF participants assess the probability that the predictand falls into one of several pre-specified ranges (below zero, between 0 and 0.5%, between 0.5 and 1%, etc). Taken together, these assessments imply a histogram-type forecast distribution.

Comparing ensemble versus histogram methods is relevant for several reasons: First, in a situation where both methods are available, the comparison provides guidance for choosing one of them. Second, given that histogram-type forecasts are unavailable in many data sets, there is a question of whether ensembles are a satisfactory alternative. Third, the results of the comparison might inform the design of future expert surveys, provided that the focus lies on forecast performance. Krüger and Nolte (2016, Section 6) have recently compared ensembles versus histograms for the US SPF data. The present study conducts a similar, but more detailed, comparison for the ECB-SPF data. The ECB-SPF data are ideally suited for that purpose, since it covers point forecasts and histograms for exactly the same target variable. By contrast, the formats of point and probabilistic forecasts are not aligned in the US SPF, such that comparisons require approximations or are restricted to a subset of time periods. Other popular economic surveys, such as the commercial Blue Chip and Consensus data, do not cover probabilistic forecasts.

Figure 1 illustrates our analysis for the 2013:Q2 edition of the ECB-SPF survey: While the gray bars represent histogram probabilities (on average over participants), the blue curve is the forecast distribution implied by an ensemble method (BMA) that derives from the point forecasts, and the orange curve is a continuous distribution fitted to the histogram. We emphasize that the forecast distributions are out-of-sample, that is, they are constructed from data that would have been available in real time. Figure 1 shows that for both GDP growth and inflation, the ensemble distribution is much more dispersed than the histogram. As detailed below, this effect arises because the ensemble

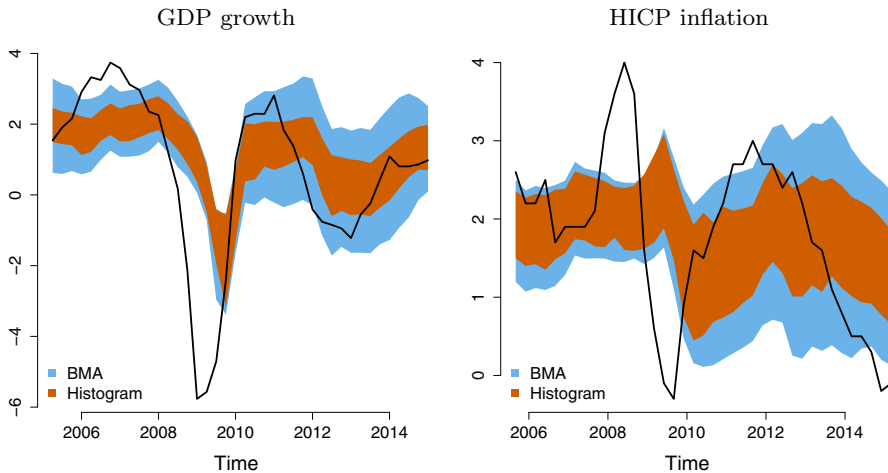


Fig. 2 Prediction intervals and realizations, plotted over time. *Blue* areas correspond to BMA prediction intervals (ranging from the 15th to 85th percentile of the forecast distribution); *orange* areas correspond to histogram prediction intervals. *Both panels* refer to 1-year-ahead forecasts

method incorporates information on past forecast errors, which were large around the great recession. The ensemble variance is an objective measure of forecast uncertainty. This objective measure clearly exceeds survey participants' subjective uncertainty as expressed by the histogram distribution. This relationship extends beyond the 2013:Q2 example. Figure 2 visualizes the 70% central prediction intervals of both methods over time: The BMA prediction intervals are wider than those of the histogram, particularly toward the end of the sample, in the wake of the great recession. The latter caused sharp downward spikes in GDP growth and inflation, as well as large absolute errors of the methods' point forecasts. As a consequence, the spread of the ensemble methods' prediction intervals increases markedly. By contrast, the spread of the histogram prediction intervals increases only marginally after the crisis (see Sect. 4.3).

Our analysis relates to a number of recent empirical studies on the ECB-SPF data (e.g., Genre et al. 2013; Kenny et al. 2014, 2015; Abel et al. 2016; Glas and Hartmann 2016). Compared to these studies, the main innovation of the present paper is our use of ensemble methods for constructing forecast distributions. Ensemble methods are used extensively in meteorology (e.g., Gneiting and Raftery 2005); they have received little attention in economics, with the exception of Gneiting and Thorarinsdottir (2010) and Krüger and Nolte (2016) who analyze US SPF data. Furthermore, ensemble methods may account for disagreement in the point forecasts of individual participants.¹ Starting with Zarnowitz and Lambros (1987), several economic studies have analyzed the relation between disagreement and various notions of forecast uncertainty. As an important contribution to this literature, Lahiri and Sheng (2010) present a factor model which bridges the gap to structural models of expectation formation (e.g., Lahiri and

¹ In the example of Fig. 1, disagreement is reflected in the spread of the small green vertical lines at the bottom of each graph. It constitutes one out of two components in the variance of the BMA ensemble method.

Sheng 2008). By contrast, our results constitute reduced form evidence on the role of forecast disagreement in the ECB-SPF data. Finally, our approach of constructing out-of-sample forecast distributions is different from estimating a historical trajectory of economic uncertainty given all data available today. This conceptual difference explains why the prediction intervals in our Fig. 2 do not cover the large downward spikes of the great recession, whereas uncertainty measures such as the ones reported in Jurado et al. (2015, Figures 1 to 4) tend to peak during recessions.²

The rest of this paper is organized as follows. Section 2 describes the data, highlighting stylized facts that are important for the design of forecasting methods. Section 3 presents the formal framework of our analysis, detailing methods for evaluating and constructing forecasts. Section 4 presents the empirical results, and Sect. 5 concludes.

2 Data on forecasts and realizations

Survey forecasts from the ECB-SPF are available quarterly, starting in 1999:Q1 (see European Central Bank 2016a). We consider forecasts of GDP growth and HICP inflation in the Euro Area; the horizon is one and two years ahead, relative to the most recent available observation. For example, when the 2006:Q4 round of the survey was conducted, the latest available observation on GDP was for 2006:Q2. Hence, the one and two-year-ahead forecasts refer to 2007:Q2 and 2008:Q2, respectively.

For the variables and sample period we consider, the number of participants per survey round ranges from 32 to 64, with a median of 47. However, the composition of this group varies heavily over time. As an example, consider one-year-ahead point forecasts of GDP growth. There are 101 distinct participants who submitted a forecast in at least one of the survey rounds. However, in a typical survey round, only about half of these participants are available. This missing data issue creates a major challenge for forecast combination in the ECB-SPF data. See Genre et al. (2013) for further details, and Capistrán and Timmermann (2009) on missingness in the US SPF data, for which the situation is similar. These data properties motivate a number of simple yet principled combination methods which we describe in Sect. 3.2.

Figure 3 shows time series plots of the two predictands, based on real-time data available from the ECB's statistical data warehouse (European Central Bank 2016b). The relevant series codes are `RTD.Q.S0.S.G_GDPM_TO_C.E` and `RTD.M.S0.N.P_C_OV.A`. GDP is measured at quarterly frequency, whereas HICP inflation is monthly. For both series, updated data vintages are released about once per month.

In our empirical analysis, we adopt a real-time perspective in order to roughly match the information set of SPF participants: When fitting a statistical forecast for a given quarter, we use historical data that would have been available on the first day of the quarter's second month (e.g., February 1, 2013 for 2013:Q1). While the exact schedule of the ECB-SPF survey varies slightly over time (see the documentation by

² The empirical implementation of the latter uncertainty measures is detailed in Jurado et al. (2015, pp. 1189–1191). See also Ozturk and Sheng (2016) who propose a global measure of economic uncertainty.

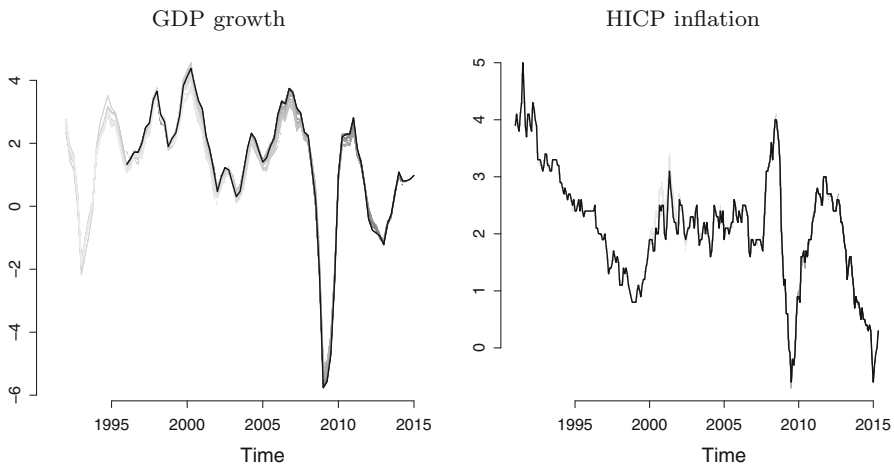


Fig. 3 Time series plots of the GDP growth rate (*left*) and the growth rate of the HICP (*right*). Both series are year-over-year growth rates. In *both panels*, the June 2015 vintage of the data is printed in *black*, and earlier vintages are printed in *grayscale*

European Central Bank 2016a), our procedure provides a simple approximation that seems reasonably accurate. Finally, we compare all forecasts against the 'final' (June 2015) vintage of the data.

3 Forecasting methods

This section describes how we evaluate forecasts, and introduces the forecasting methods we consider.

3.1 Measure of forecast accuracy

We measure the accuracy of distribution forecasts via the Continuous Ranked Probability Score (CRPS; Matheson and Winkler 1976; Gneiting and Raftery 2007). For a cumulative distribution function F (i.e., the forecast distribution) and a realization $y \in \mathbb{R}$, the CRPS is given by

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - 1(z \geq y))^2 dz, \quad (1)$$

where $1(A)$ is the indicator function of the event A . Closed-form expressions of the integral in (1) are available for several types of distributions F , such as the two types we use below: the two-piece normal distribution (Gneiting and Thorarinsdottir 2010) and mixtures of normal distributions (Grimet et al. 2006). Our implementation is based on the R package `scoringRules` (Jordan et al. 2016) which includes both of these variants.

3.2 Ensemble-based methods

Let x_{t+h}^i denote the point forecast of participant i , made at date t for date $t + h$. Furthermore, let \mathcal{S}_{t+h} denote the collection of available participants, and denote by $N_{t+h} = |\mathcal{S}_{t+h}|$ the number of participants. As described above, the set of available participants may change over time and across forecast horizons. Ensemble methods construct a forecast distribution for Y_{t+h} , based on the individual forecasts $\{x_{t+h}^i\}_{i \in \mathcal{S}_{t+h}}$. The ensemble forecast may depend on ‘who says what’, that is, it may vary across permutations of the indexes $i \in \mathcal{S}_{t+h}$. If that is the case, the individual forecasters are said to be non-exchangeable. Alternatively, if the ensemble forecast is invariant to ‘who says what’, the forecasters are said to be exchangeable.

In order to treat forecasters as non-exchangeable, one must be able to compare them. For example, if we knew that Anne is a better forecaster than Bob, we might want to design an ensemble method which puts more weight on Anne’s forecast than on Bob’s. However, relative performance is hard to estimate in the ECB-SPF data set, since the past track records of different forecasters typically refer to different time periods. Similarly, estimating correlation structures among the forecasts, as is required for regression-type combination approaches (e.g., [Timmermann 2006](#)), requires to impute the missing forecasts in some way. Given the difficulties just described, it is perhaps not surprising that [Genre et al. \(2013\)](#) find very little evidence in favor of either performance-based or regression methods for the ECB-SPF data. Motivated by these concerns, we consider two simple ensemble methods which treat the forecasters as exchangeable:

1. The first method (‘BMA’) constructs a forecast distribution as follows:

$$f_{t+h}^{\text{BMA}} = \frac{1}{N_{t+h}} \sum_{i \in \mathcal{S}_{t+h}} \mathcal{N}(x_{t+h}^i, \theta), \quad (2)$$

where $\theta \in \mathbb{R}_+$ is a scalar parameter to be estimated. In words, (2) posits an equally weighted mixture of N_{t+h} forecast distributions, each of which corresponds to an individual forecaster i . Each of these distributions is assumed to have the same variance, θ . The method has been proposed by [Krüger and Nolte \(2016\)](#), who find that it performs well for the US SPF data. The label ‘BMA’ hints at the method’s close conceptual connection to the Bayesian model averaging approach proposed by [Raftery et al. \(2005\)](#) in the meteorological forecasting literature. Denote the mean of the survey forecasts by $\bar{x}_{t+h} = N_{t+h}^{-1} \sum_{i \in \mathcal{S}_{t+h}} x_{t+h}^i$. Then, the variance of the forecast distribution in (2) is given by $D_{t+h} + \theta$, where

$$D_{t+h} = N_{t+h}^{-1} \sum_{i \in \mathcal{S}_{t+h}} \left(x_{t+h}^i - \bar{x}_{t+h} \right)^2$$

is the cross-sectional variance of the survey forecasts (‘disagreement’). The two-component structure of the BMA forecast variance is similar to the model by [Ozturk and Sheng \(2016\)](#) which decomposes forecast uncertainty into a common

component (similar to θ in Eq. 2) and an idiosyncratic component (which they proxy by forecaster disagreement, D_t).

2. The ‘EMOS’ method assumes that

$$f_{t+h}^{\text{EMOS}} = \mathcal{N}(\bar{x}_{t+h}, \gamma), \quad (3)$$

where $\gamma \in \mathbb{R}_+$ is a parameter to be estimated. The method simply fits a normal distribution around the mean of the survey forecasts. Unlike in the BMA method, forecaster disagreement does not enter the variance of the distribution. The label ‘EMOS’ alludes to the method’s similarity to the ensemble model output statistics approach of [Gneiting et al. \(2005\)](#), again proposed in a meteorological context.

Both ensemble methods require only one parameter to be estimated, which can be done via grid search methods. In each case, we fit the parameter to minimize the sample average of the CRPS (see Sect. 3.1), based on a rolling window of 20 observations. Conceptually, the BMA method is based on the idea of fitting a simplistic distribution to each individual forecaster, and then averaging over these distributions. This is why [Krüger and Nolte \(2016\)](#) call it a ‘micro-level’ method. In contrast, the EMOS method fixes a normal distribution and fits the parameters of that distribution via a summary statistic (the mean) from the ensemble of forecasters.

3.3 Survey histograms

In our out-of-sample analysis, we consider the average histogram over all forecasters (for a given date, variable, and forecast horizon). In order to convert the histogram into a complete forecast distribution, we consider a parametric approximation, obtained by fitting a two-piece normal distribution ([Wallis 2004](#), Box A) to the histogram. Specifically, the parameters of the approximating distribution solve the following minimization problem (suppressing time and forecast horizon for ease of notation):

$$\min_{\mu, \sigma_1, \sigma_2} \sum_{j=1}^J [P_j - F_{2\text{PN}}(r_j; \mu, \sigma_1, \sigma_2)]^2, \quad (4)$$

where r_j is the right endpoint of histogram bin $j = 1, \dots, J$; P_j is the cumulative probability of bins 1 to j , and $F_{2\text{PN}}(\cdot; \mu, \sigma_1, \sigma_2)$ is the cumulative distribution function of the two-piece normal distribution with parameters μ , σ_1 and σ_2 .

Parametric approximations to survey histogram forecasts have been proposed by [Engelberg et al. \(2009\)](#), who consider a very flexible generalized beta distribution as an approximation. In our case, replacing the two-piece normal distribution by a conventional Gaussian distribution yielded very similar forecasting results, suggesting that the flexibility of the two-piece normal is well sufficient.

3.4 Simple time series benchmark model

We also compare the ECB-SPF to a simple Gaussian benchmark forecast distribution, with mean equal to the random walk prediction, and variance estimated

Table 1 Parameter estimates for the ensemble methods introduced in Sect. 3.1 (median over rolling windows)

The sample periods are as defined below Table 2

		GDP growth		HICP inflation	
		$h = 1$	$h = 2$	$h = 1$	$h = 2$
BMA	θ	1.45	2.31	0.88	0.81
EMOS	γ	1.59	2.42	0.94	0.86

from a rolling window of 20 observations (in line with the sample used for fitting the ensemble methods). Specifically, denote the series of training sample observations by $\{y_t\}_{t=T-19}^T$. Then, the benchmark distribution has mean y_T and variance $h_q \times \frac{1}{19} \sum_{t=T-18}^T (y_t - y_{t-1})^2$, where h_q is the forecast horizon (in quarters). Our choice of the random walk is motivated by the quarter-on-quarter definition of the predictands, which implies considerable persistence almost by definition (see Fig. 3).³

4 Empirical results

This section summarizes parameter estimates of interest, presents the main results of our out-of-sample analysis, and then provides some further analysis of forecast uncertainty in ensembles and histograms.

4.1 Parameter estimates for ensembles

As described in Sect. 3.2, the two ensemble methods entail different assumptions on the role of forecaster disagreement; Table 1 summarizes the parameter estimates of both methods. Recall that the forecast variance of BMA is given by the sum of θ and the cross-sectional variance of point forecasts (disagreement); the forecast variance of EMOS is given by γ . As shown in Table 1, the estimates of θ are only marginally smaller than the estimates of γ . This implies that disagreement plays a very limited role in predicting forecast uncertainty for the ECB-SPF data. Also, note that the ensemble parameters for inflation are slightly smaller for two-year-ahead forecasts than for one-year-ahead forecasts, which points to possible model misspecification (c.f. Patton and Timmermann 2012).

4.2 Forecast performance

Table 2 contains our main empirical results. In addition to the average CRPS values, the table indicates the results of Diebold and Mariano (1995) type tests for the null hypothesis of equal predictive ability compared to the BMA method. We construct the test statistic as described in Krüger et al. (2015, Section 5). The results can be summarized as follows.

³ Using more sophisticated benchmark models (Bayesian autoregressive models with stochastic volatility) yielded the same qualitative conclusions. Results on these additional benchmarks are omitted for brevity, but are available upon request.

Table 2 Continuous Ranked Probability Score (CRPS), averaged over the evaluation period (2005:Q3 to 2015:Q1 for 1-year-ahead forecasts; 2007:Q3 to 2015:Q1 for 2-year-ahead forecasts)

	GDP growth		HICP inflation	
	$h = 1$	$h = 2$	$h = 1$	$h = 2$
BMA	0.98	1.61	0.64	0.79
EMOS	0.99	1.61	0.64	0.78
Histogram	1.04	1.71	0.67	0.82
Random walk	1.82*	2.43**	0.82	1.13

Smaller values indicate better forecasts. Stars indicate that the null hypothesis of Equal Predictive Ability (compared to BMA) is rejected at the 10% level (*one star*) or 5% level (*two stars*), using a two-sided test

- The ensemble-based methods (BMA and EMOS) perform very similarly for both variables and forecast horizons. This is striking, given that the two specifications feature quite different variance specifications. Together with the parameter estimates reported in Table 1, the result suggests that forecaster disagreement does not have a systematic impact on forecast performance.
- The accuracy of the survey histograms is comparable to the ensemble-based methods. While the histograms attain slightly worse CRPS scores for both variables and forecast horizons, these differences are not statistically significant at the 5% level.
- The random walk time series benchmark attains worse CRPS scores than the (ensemble or histogram based) survey methods in all instances. The difference to BMA is statistically significant at the 10% level for GDP, but not for inflation.

Our finding that disagreement is of limited help for distribution forecasting differs from the results by [Krüger and Nolte \(2016\)](#), who find that disagreement in US SPF forecasts does have predictive power for some variables and forecast horizons. A possible explanation for this discrepancy is that [Krüger and Nolte \(2016\)](#) consider a longer sampling period, with more variation of disagreement over time. This may help to identify effects which are hard to identify in our short sample. This interpretation is also supported by the results of [Boero et al. \(2015\)](#). They consider data from the Bank of England's Survey of External Forecasters (SEF) for the 2006–2012 sample period and find no close relation between disagreement and squared forecast errors. They conclude that “[...] the joint results from the US [SPF] and UK [SEF] surveys suggest the encompassing conclusion that disagreement is a useful proxy for uncertainty when it exhibits large fluctuations, but low-level high-frequency variations are not sufficiently correlated.” ([Boero et al. 2015](#), p. 1044). Our results for the ECB-SPF data are in line with this view. Furthermore, the forecast horizons we consider are fairly long by macroeconomic standards. The limited role of disagreement at long horizons is in line with the factor model by [Lahiri and Sheng \(2010\)](#).

4.3 Forecast uncertainty in ensembles and histograms

As seen from Table 2, the performance of ensemble and histogram methods is not statistically different. However, their similar performance does not necessarily mean

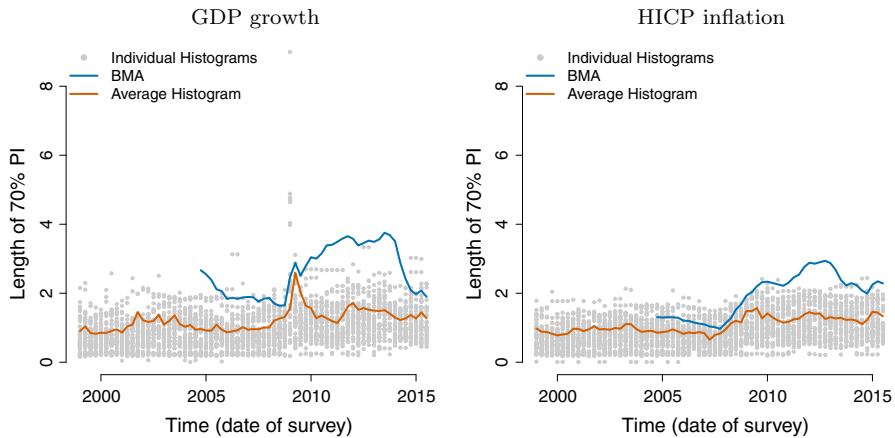


Fig. 4 Length of 70% prediction intervals, for individual and average histograms (approximated via two-piece normal distributions), and for the BMA ensemble method. Forecasts are 1 year ahead

that the methods produce similar forecast distributions. Indeed, as suggested already by Fig. 2, the methods tend to differ substantially in practice.

Comparing the prediction intervals in Fig. 2 to the realizing values (black line), it turns out that the BMA intervals cover the realization in 68% of all cases for GDP, and in 62% of all cases for inflation. These numbers are reasonably close to the desired nominal coverage level of 70%. By contrast, the histogram prediction intervals cover the realizations in only 38% of all cases (for both GDP and inflation), implying that the histogram-based forecast distributions are overconfident.⁴

In principle, our result that the *average* survey histogram is overconfident may conceal important differences in individual-level uncertainty, which are studied by Lahiri and Liu (2006, US SPF data), Boero et al. (2015, UK SEF data), and others. To investigate this possibility, Fig. 4 plots the length of the individual participants' prediction intervals over time. These numbers are computed by fitting a two-piece normal distribution to each individual histogram, as described in Sect. 3.3, and then computing the lengths of the prediction intervals for the resulting two-piece normal distributions.⁵ For most quarters between 2010 and 2014, *all* individual forecast histograms yield shorter prediction intervals than the BMA ensemble, indicating that our result is not driven by a few individuals. Finally, the findings from Fig. 4 are in line with Kenny et al. (2015) who analyze the determinants of individual-level forecast performance in the ECB-SPF. They find that "[...] many experts [...] are underestimating uncertainty

⁴ These results pose no contradiction to the similar CRPS scores of the BMA and histogram methods as reported in Table 2: The CRPS rewards both sharpness (which works in favor of the histogram method) and correct calibration (where BMA performs better); see Gneiting and Katzfuss (2014, Section 3). The balance of these two effects appears to work slightly in favor of BMA, which attains numerically smaller CRPS scores. However, as mentioned earlier the differences are not statistically significant.

⁵ The design of the figure loosely follows Boero et al. (2015, Figure 3). We exclude two individual histograms (one-year-ahead forecasts for GDP growth in 2009:Q1, IDs 52 and 70) where all probability mass was put on the leftmost bin, such that the two-piece normal approximation is not well defined.

and could improve their density performance by simply increasing their variances.” (Kenny et al. 2015, p. 1229)

5 Conclusion

The key result of this paper is that point forecasts are useful for constructing forecast distributions, by means of ensemble methods. Motivated by earlier results for the ECB-SPF (Genre et al. 2013), we focus on very simple variants of these methods, based on a single parameter to be estimated. We find that these methods perform well relative to survey histograms also contained in the ECB-SPF. This suggests that ensemble methods are promising for other survey data sets, which typically comprise point forecasts only.

We further find that ensembles—which estimate the statistical uncertainty from past forecast errors—consistently generate wider prediction intervals than subjective survey histograms (see Figs. 2, 4). This gap between objective versus subjective uncertainty calls for further analysis as the time series dimension of the ECB-SPF data grows larger. Kajal Lahiri’s research on uncertainty, heterogeneity and learning in macroeconomic expert surveys offers important methodological and empirical insights to guide such an analysis.

Acknowledgements This work has been funded by the European Union Seventh Framework Programme under Grant agreement 290976. The author thanks the Klaus Tschira Foundation for infrastructural support at the Heidelberg Institute for Theoretical Studies (HITS), where the author was employed during the writing of the present article. Helpful comments by Tilmann Gneiting, Aidan Meyler, and an anonymous reviewer are gratefully acknowledged.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Abel J, Rich R, Song J, Tracy J (2016) The measurement and behavior of uncertainty: evidence from the ECB survey of professional forecasters. *J Appl Econ* 31:533–550
- Boero G, Smith J, Wallis KF (2015) The measurement and characteristics of professional forecasters’ uncertainty. *J Appl Econ* 30:1029–1046
- Capistrán C, Timmermann A (2009) Forecast combination with entry and exit of experts. *J Bus Econ Stat* 27:428–440
- Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *J Bus Econ Stat* 13:253–263
- Engelberg J, Manski CF, Williams J (2009) Comparing the point predictions and subjective probability distributions of professional forecasters. *J Bus Econ Stat* 27:30–41
- European Central Bank (2016a) ECB survey of professional forecasters. <http://www.ecb.europa.eu/stats/prices/indic/forecast/html/index.en.html>. Accessed 2016-02-10
- European Central Bank (2016b) Statistical data warehouse. <https://sdw.ecb.europa.eu/>. Accessed 2016-02-10
- Genre V, Kenny G, Meyler A, Timmermann A (2013) Combining expert forecasts: Can anything beat the simple average? *Int J Forecast* 29:108–121
- Glas A, Hartmann M (2016) Inflation uncertainty, disagreement and monetary policy: evidence from the ECB survey of professional forecasters. *J Empir Financ* 39:215–228

- Gneiting T, Katzfuss M (2014) Probabilistic forecasting. *Annu Rev Stat Appl* 1:125–151
- Gneiting T, Raftery AE (2005) Weather forecasting with ensemble methods. *Science* 310:248–249
- Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 102:359–378
- Gneiting T, Thorarindottir TL (2010) Predicting inflation: Professional experts versus no-change forecasts. <http://arxiv.org/abs/1010.2318>. Preprint. Accessed 2016-09-21
- Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon Weather Rev* 133:1098–1118
- Grimit E, Gneiting T, Berrocal V, Johnson N (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q J R Meteorol Soc* 132:2925–2942
- Jordan A, Krüger F, Lerch S (2016) R package `scoringRules`: scoring rules for parametric and simulated distribution forecasts. <https://cran.r-project.org/web/packages/scoringRules/>. Accessed 2016-08-30
- Jurado K, Ludvigson SC, Ng S (2015) Measuring uncertainty. *Am Econ Rev* 105:1177–1216
- Kenny G, Kostka T, Masera F (2014) How informative are the subjective density forecasts of macroeconomists? *J Forecast* 33:163–185
- Kenny G, Kostka T, Masera F (2015) Density characteristics and density forecast performance: A panel analysis. *Empir Econ* 48:1203–1231
- Krüger F, Nolte I (2016) Disagreement versus uncertainty: evidence from distribution forecasts. *J Bank Financ* 72:S172–S186
- Krüger F, Clark TE, Ravazzolo F (2015) Using entropic tilting to combine BVAR forecasts with external nowcasts. *J Bus Econ Stat* (forthcoming)
- Lahiri K, Liu F (2006) Modelling multi-period inflation uncertainty using a panel of density forecasts. *J Appl Econom* 21:1199–1219
- Lahiri K, Sheng X (2008) Evolution of forecast disagreement in a Bayesian learning model. *J Econom* 144:325–340
- Lahiri K, Sheng X (2010) Measuring forecast uncertainty by disagreement: the missing link. *J Appl Econom* 25:514–538
- Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. *Manage Sci* 22:1087–1096
- Müller UK, Watson MW (2016) Measuring uncertainty about long-run predictions. *Rev Econ Stud* 83:1711–1740
- Ozturk EO, Sheng X (2016) Measuring global and country-specific uncertainty. Paper presented at the Stanford SITE workshop on ‘The Macroeconomics of Uncertainty and Volatility’. <https://site.stanford.edu/sites/default/files/measuring-global-and-country-specific-uncertainty.pdf>. Accessed 2016-09-25
- Patton AJ, Timmermann A (2012) Forecast rationality tests based on multi-horizon bounds. *J Bus Econ Stat* 30:1–17
- Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mon Weather Rev* 133:1155–1174
- Timmermann A (2006) Forecast combinations. In: Elliott G, Granger C, Timmermann A (eds) *Handbook of economic forecasting*, vol 1. Elsevier, Amsterdam, pp 135–196
- Wallis KF (2004) An assessment of Bank of England and National Institute inflation forecast uncertainties. *Natl Inst Econ Rev* 189:64–71
- Zarnowitz VA, Lambros LA (1987) Consensus and uncertainty in economic prediction. *J Polit Econ* 95:591–621