Survey density forecast comparison in small samples

Laura Coroneo[*]
University of York

Fabrizio Iacone
Università degli Studi di Milano
University of York

Fabio Profumo
University of York

20th December 2023

Abstract

We apply fixed-b and fixed-m asymptotics to tests of equal predictive accuracy and of encompassing for survey density forecasts. We verify in an original Monte Carlo design that fixed-smoothing asymptotics delivers correctly sized tests in this framework, even when only a small number of out of sample observations is available. We use the proposed density forecast comparison tests with fixed-smoothing asymptotics to assess the predictive ability of density forecasts from the European Central Bank's Survey of Professional Forecasters (ECB SPF). We find an improvement in the relative predictive ability of the ECB SPF since 2010, suggesting a change in the forecasting practice after the financial crisis.

Keywords: survey density forecast comparison, ECB SPF, Diebold-Mariano test, forecast encompassing, fixed-smoothing asymptotics

JEL Classification: C12, C22, E17

[*]Corresponding author: Laura Coroneo, Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD, UK. The support of the ESRC grant ES/J500215/1 is gratefully acknowledged. We thank the editor George Kapetanios, the associate editor, an anonymous referee, Liudas Giraitis, Alastair Hall, and participants to the Sheffield Advances in Econometrics Workshop, the 42nd International Symposium on Forecasting (ISF 2022 - University of Oxford), the 10th Italian Congress of Econometrics and Empirical Economics (ICEEE - University of Cagliari) and the 12th ECB Conference on Forecasting Techniques (ECB) for useful comments.

1 Introduction

Expectations play a key role in economic decision-making and largely determine policy outcomes. This is particularly true for monetary policy, as its effects heavily depend on expectations. For this reason, central banks around the world regularly run surveys of professional forecasters to gather information about private agents' expectations.

Survey respondents are asked to report their point forecasts for a set of macroeconomic fundamentals and, increasingly, to provide a density forecast that describes the predicted probability distribution of the variables of interest. Compared to the more popular point forecasts, density forecasts provide a wider understanding of the uncertainty associated with the prediction, see Fair (1980) and Dawid (1984) for some early references, and Tay and Wallis (2000) for a more recent detailed discussion.

Well-known examples of survey density forecasts include the Survey of Professional Forecasters (SPF) currently managed by the Federal Reserve Bank of Philadelphia, the Survey of External Forecasters managed by the Bank of England and the European Central Bank's Survey of Professional Forecasters (ECB SPF). A large amount of work has been devoted to analysing the density forecasts provided by the US SPF, see among others Diebold, Tay and Wallis (1999) and Clements (2014), and the Bank of England's Survey of External Forecasters, see among others Boero, Smith and Wallis (2008) and Mitchell and Hall (2005). The literature dedicated to density forecasts provided by the ECB SPF is more limited, see de Vincent-Humphreys, Dimitrova, Falck and Henkel (2019) for a survey, possibly because the ECB SPF started only recently, in 1999.

A challenge in forecast comparison studies for survey data is that traditional inference methods suffer from relevant small sample size distortions, which can lead to spurious results, as well documented by Clark (1999) for the Diebold and Mariano (1995) equal predictive accuracy test. This shortcoming is of course especially relevant when the analysis is performed on subsamples, as for example when only using the post-great financial crisis sample.

In this paper, we apply fixed-b and fixed-m asymptotics to address the small sample bias of density forecast comparison tests. We compare alternative density forecasts by testing two null hypotheses. The first hypothesis is the null of equal predictive accuracy of two forecasts, this is the Diebold and Mariano (1995) equal predictive ability test. The second one is the null of forecast encompassing in Harvey et al. (1998), which involves testing whether one forecast is encompassed by the other. To accommodate forecasts reported as probabilities for intervals, or bins, as typical for survey forecasts, we use two loss functions: the Quadratic Probability Score by Brier (1950) and the Ranked Probability Score by Epstein (1969). With these loss functions, we show that both tests can be performed in the framework of semiparametric inference on the mean of a process. In the case of the test of equal predictive accuracy, this coincides with the framework in Diebold and Mariano (1995), so we will loosely refer to it as the DM framework in the remainder of the paper, even when we apply it to the forecast encompassing test.

The DM framework is particularly appealing as it is simple and the test statistic is easy to compute. To overcome the small sample bias of the DM framework, we use an alternative approach based on fixed-smoothing asymptotics. In particular, we consider fixed-b asymptotics by Kiefer and Vogelsang (2005) and fixed-m asymptotics by Hualde and Iacone (2017). This approach proved capable of eliminating size distortion in the equal predictive accuracy test for comparing point forecasts, see Coroneo and Iacone (2020). In an original Monte Carlo exercise, we first document that standard asymptotics deliver unreliable density forecast comparison tests in small samples, and we then verify that fixed-b and fixed-m asymptotics can be used with success to perform tests of equal predictive accuracy and encompassing for density forecasts.

We apply the proposed density forecast comparison tests to assess the accuracy of the ECB SPF density forecasts for three key macroeconomic variables (real GDP growth, inflation and the unemployment rate). We are interested in establishing whether ECB SPF density forecasts can beat and/or encompass simple benchmarks, such as a uniform, an unconditional Gaussian, a Gaussian random walk, a Gaussian random walk with drift and a naive forecast taken from the previous round of ECB SPF forecasts. All benchmark forecasts are produced in real-time, by using the same information available to professional forecasters at each survey deadline.

Results indicate that ECB SPF density forecasts for unemployment and real GDP growth outperformed and sometimes encompassed the benchmarks, especially at one-year ahead and in the second subsample. On the other hand, survey forecasts for inflation do not easily outperform nor encompass the benchmarks. For all the variables, however, the application of fixed-smoothing asymptotics strengthens the evidence of an improvement in relative predictive ability since 2010, suggesting a change in the forecasting practice after the financial crisis. We also find that the ECB SPF easily outperforms and encompasses the naive benchmark, indicating that professional forecasters update their information set when making their predictions and that previous round forecasts are uninformative.

This paper contributes to the literature on forecast evaluation by introducing fixed-smoothing asymptotics to density forecast comparison tests. This type of asymptotics is becoming popular for point forecast comparison, see Choi and Kiefer (2010), Harvey et al. (2017), Li and Patton (2018), Coroneo and Iacone (2020), Coroneo et al. (2022), but, to the best of our knowledge, their properties for density forecast comparison tests have not been analysed. Our novel Monte Carlo exercise confirms the small sample bias of standard density forecast comparison tests, and indicates that fixed-smoothing asymptotics successfully addresses this issue. We also contribute to the literature on forecast encompassing by showing how the forecast encompassing test for density forecasts can be implemented in the DM framework: Clements and Harvey (2010) introduce it for dichotomic variables but we extend it to continuously distributed variables.

The remainder of the paper is organised as follows. In Section 2 we describe how to perform tests of equal predictive accuracy and encompassing for survey density forecasts.

In Section 3, we show how to apply fixed-smoothing asymptotics to these tests. We investigate the properties of the tests in Section 4, where we present a Monte Carlo exercise and provide recommendations for the bandwidths. In Section 5 we carry out the empirical study, and in Section 6 we conclude.

2 Density forecast comparison

We compare two h-step ahead density forecasts made at time $t-h$ for the variable of interest y_t using loss functions. The h-step ahead survey density forecast i specifies the probability that the variable of interest y_t falls in bin k given the information available at time $t-h$, $\mathbf{f}_{t,i} = [f_{t,i}^1, \ldots, f_{t,i}^k, \ldots, f_{t,i}^K]'$, where $f_{t,i}^k = P_{t-h,i}(y_t \in k)$ for $k = 1, \ldots, K$.

The vector of realisations is $\mathbf{y}_t = [y_t^1, \ldots, y_t^k, \ldots, y_t^K]'$, where the indicator variable $y_t^k = I(y_t \in k)$ takes the value of 1 if the outcome at time t falls in bin k and zero otherwise, so that $K-1$ elements of $\mathbf{y}_t$ are set to 0 and one takes value 1. The forecast error is then $\mathbf{e}_{t,i} = \mathbf{y}_t - \mathbf{f}_{t,i}$.

The cumulative distribution function of the density forecast is $\mathbf{F}_{t,i} = [F_{t,i}^1, \ldots, F_{t,i}^k, \ldots, F_{t,i}^K]'$, where $F_{t,i}^k = \sum_{l=1}^k f_{t,i}^l$, and the cumulative outcome variable is $\mathbf{Y}_t = [Y_t^1, \ldots, Y_t^k, \ldots, Y_t^K]'$, where $Y_t^k = \sum_{l=1}^k y_t^l$. Finally, the cumulative forecast error is given by $\mathbf{E}_{t,i} = \mathbf{Y}_t - \mathbf{F}_{i,t}$.

We consider two loss functions that naturally accommodate forecasts reported as histograms: the Quadratic Probability Score by Brier (1950) and the Ranked Probability Score by Epstein (1969). The Quadratic Probability Score (QPS) associated with each forecast is given by

$$QPS_{t,i} = \sum_{k=1}^K (y_t^k - f_{t,i}^k)^2 = \mathbf{e}_{t,i}'\mathbf{e}_{t,i}. \quad (1)$$

This loss function penalizes equally any probability assigned to events that do not occur. As a consequence, forecasts that assign a large probability in a neighbourhood of the realised outcome are treated in the same way as forecasts that assign a small probability to that same neighbourhood and put more probability on very distant outcomes. This may be appropriate in some situations; in many cases, however, it is desirable to consider the forecast clustering more probability in the intervals near the realised outcome as more precise. For this reason, we also consider the Ranked Probability Score (RPS) associated with each forecast, given by

$$RPS_{t,i} = \sum_{k=1}^{K} (Y_t^k - F_{t,i}^k)^2 = \mathbf{E}'_{t,i}\mathbf{E}_{t,i}. \qquad (2)$$

This loss function has the advantage of considering the overall tendency of the forecast probability density function, as it penalizes less severely density forecasts assigning relatively larger probabilities to outcomes that are close to the true outcome. Therefore, the RPS has the desirable property of being proper in the sense that encourages the forecasters to reveal their true beliefs, see Gneiting and Raftery (2007).

Another appealing property of the QPS and the RPS is that they are always defined, even when the realisation falls in a histogram bin to which the survey forecast has assigned a zero probability. On the contrary, the more popular logarithmic score would be undefined in this case.

We use two approaches to compare the performance of two density forecasts. The first involves testing the null hypothesis of equal predictive accuracy of the two forecasts according to the QPS or the RPS loss function. This can be implemented with the test for equal predictive accuracy proposed by Diebold and Mariano (1995). The second approach involves testing for whether one density forecast is encompassed by the other one, in the sense that the predictive accuracy (according to the QPS or the RPS loss function) of the encompassing density forecast cannot be improved by a linear combination with the encompassed forecast. This is the forecast encompassing test and, for point forecasts, Harvey et al. (1998) show that, by redefining the loss differential, it is possible to implement it using the DM framework.

In what follows, we first show how the DM framework can be used also to perform the forecast encompassing test for density forecasts. We then discuss the limitations of standard asymptotics when applied to the DM framework, and apply fixed-smoothing asymptotics to the DM framework for density forecast comparison.

2.1 Equal predictive accuracy

A test of equal predictive accuracy allows testing the null hypothesis that two alternative forecasts have equal forecasting accuracy according to a user-chosen loss function, which in the case of density forecasts can be the QPS or the RPS loss function.

Denote by L^i the loss function for $i = 1, 2$, so that $L_t^i = QPS_{t,i}$ if the QPS loss is used or $L_t^i = RPS_{t,i}$ if the RPS loss is used, and the loss differential by

$$d_t = L_t^1 - L_t^2, \tag{3}$$

the null hypothesis of equal forecasting ability is

$$H_0 : \{E(d_t) = 0\}. \tag{4}$$

2.2 Forecast encompassing

A forecast encompassing test involves testing whether one set of forecasts encompasses another one, in the sense that the accuracy of one set of (encompassing) forecasts $\mathbf{f}_{t,1}$ cannot be improved through a linear combination with a second set of (encompassed) forecasts $\mathbf{f}_{t,2}$. To this end, we consider the density forecast combination

$$\mathbf{f}_{t,c}(\lambda) = (1 - \lambda)\mathbf{f}_{t,1} + \lambda\mathbf{f}_{t,2} \tag{5}$$

where λ $(0 \leq \lambda \leq 1)$ is a scalar and denotes the weight associated with forecasts $\mathbf{f}_{t,2}$. In this context, $\mathbf{f}_{t,1}$ encompasses $\mathbf{f}_{t,2}$ if the optimal weight in the QPS (or RPS) sense is equal to zero.

In Appendix A, we show that if we define d_t as

$$d_t = \begin{cases} \mathbf{e}'_{t,1}(\mathbf{e}_{t,1} - \mathbf{e}_{t,2}), \text{for } QPS, \\ \mathbf{E}'_{t,1}(\mathbf{E}_{t,1} - \mathbf{E}_{t,2}), \text{for } RPS, \end{cases} \tag{6}$$

then the null of density forecast encompassing can be expressed as

$$H_0 : \{E(d_t) = 0\}$$

and the density forecast encompassing test can be conducted against the one-sided alternative $E(d_t) > 0$ (i.e., $\lambda > 0$), given the assumption of a non-negative combination weight.

2.3 Diebold-Mariano framework

In sections 2.1-2.2, we showed how both the equal predictive accuracy and the forecast encompassing tests can be performed in the framework of inference on the mean of the process d_t also in the context of density forecast evaluation. The difference between the two tests lies in how the process d_t is defined. For the test for equal predictive accuracy, d_t is defined as in (3), while for the test for density forecast encompassing d_t is defined as in (6).

Denoting the sample average as $\overline{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ and the long run variance as $\sigma_T^2 = var(\sqrt{T}\,\overline{d})$, then the test statistic is $\sqrt{T}\,\overline{d}/\sigma_T$. Under regularity conditions as for example in Giacomini and White (2006) and under H_0,

$$\sqrt{T}\frac{\overline{d}}{\sigma_T} \rightarrow_d N(0,1). \tag{7}$$

The test statistic in (7) is unfeasible as σ_T is unknown, but this may be replaced by an estimate, say $\widehat{\sigma}$. If the latter is consistent, in the sense $\widehat{\sigma} - \sigma_T = o_p(1)$, the feasible statistic obtained in this way retains the standard normal limiting distribution.

One estimate that, under regularity conditions, fits this purpose, is the Weighted Covariance Estimate (WCE)

$$\widehat{\sigma}^2_{WCE} = \widehat{\gamma}_0 + 2\sum_{j=1}^{T-1} k(j/M)\widehat{\gamma}_j$$

where $\widehat{\gamma}_j$ is the sample autocovariance at lag j, $k(.)$ is a kernel function and M is a bandwidth parameter. A popular kernel is the triangular (Bartlett) kernel, yielding the estimate

$$\widehat{\sigma}^2_{WCE-B} = \widehat{\gamma}_0 + 2\sum_{j=1}^{M} \left(\frac{M-j}{M}\right)\widehat{\gamma}_j.$$

Regularity conditions to ensure consistency include $M \to \infty$ and $M/T \to 0$ as $T \to \infty$.

A second class of estimates of the long run variance is the Weighted Periodogram Estimate (WPE)

$$\widehat{\sigma}^2_{WPE} = 2\pi\sum_{j=1}^{T/2} K_M(\lambda_j)I(\lambda_j) \tag{8}$$

where $K_M(\lambda_j)$ is a symmetric kernel function, and $I(\lambda_j)$ is the periodogram of d_t computed at the Fourier frequency λ_j. A popular kernel in this case is the Daniell kernel, as this estimate of the long run variance has a very simple formula in the frequency domain,

$$\widehat{\sigma}^2_{WPE-D} = 2\pi\frac{1}{m}\sum_{j=1}^{m} I(\lambda_j). \tag{9}$$

where m is a user-chosen parameter that, with a slight abuse of notation, is referred to as bandwidth too. When $m \to \infty$ and $m/T \to 0$ as $T \to \infty$, the estimate is consistent.

Unfortunately, the DM framework is subject to severe size distortion in small and medium-sized samples, as documented, for example, in Clark (1999). Obviously, finite sample size distortion is not a problem affecting only the DM framework, it is common to any test that makes inference on the mean (or on a regression parameter) using a heteroskedasticity autocorrelation consistent estimate of the long run variance and maintaining the limit normality assumption for the standardised statistic, see for example Newey and West (1994). In fact, in any finite sample, the ratio M/T is still non-zero, and in a moderate size sample this ratio may be non-negligible. Thus, this size distortion may be more severe in the context of forecast comparisons, as in many cases the sample size is relatively small, when compared to other macro and financial applications.

3 Fixed-smoothing asymptotics

Neave (1970) shows that treating the ratio M/T as constant can provide a better measure of the variance of the weighted covariance estimate of a spectral estimate. Kiefer and Vogelsang (2002a,b, 2005) apply the same intuition to the problem of testing hypothesis about the mean for a weakly dependent process, deriving the distribution of the feasible test statistic when $M/T \to b \in (0,1]$ as $T \to \infty$. Under this assumption $\widehat{\sigma}^2$ is not consistent, and the test statistic has a non-standard limit distribution that depends both on b and on the kernel choice. Because of the dependence on b of the limit distribution, this approach is often referred to as "fixed-b".

In the context of the DM framework, for the Bartlett kernel the results of Kiefer and Vogelsang (2005) imply that, under H_0 and regularity conditions, when $M/T \to b \in (0,1]$ as $T \to \infty$,

$$\sqrt{T}\frac{\overline{d}}{\widehat{\sigma}_{WCE-B}} \to_d \Phi^B(b) \tag{10}$$

$\Phi^B(b)$ is characterised in Kiefer and Vogelsang (2005) and a cubic equation is provided for critical values.

In the frequency domain, fixed-b corresponds to keeping m constant when the Daniell kernel is used. This naturally leads to considering asymptotics for fixed m. Under H_0 and regularity conditions, Hualde and Iacone (2017) consider m constant as $T \to \infty$, in this case we have

$$\sqrt{T}\frac{\overline{d}}{\widehat{\sigma}_{WPE-D}} \to_d t_{2m}. \tag{11}$$

We refer to Appendix B for a sufficient assumption to establish the limits (10) and (11). Fixed-b and fixed-m asymptotics can be heuristically understood as undersmoothing in the context of estimating the spectral density at frequency zero. For this reason, many references, for example Sun (2013), refers to them collectively as fixed-smoothing. Monte Carlo simulations in Kiefer and Vogelsang (2005) suggest that critical values obtained using fixed-b asymptotics result in better empirical size for tests. This was later justified theoretically by Sun (2014), that shows that fixed-b asymptotics provides a higher order refinement. Moreover, fixed-smoothing asymptotics gives a justification (and suitable critical values) even for bandwidths that researchers would not consider when using standard asymptotics: it is even possible to choose $M = T$ when using the weighted covariance Bartlett estimate, or to choose $m = 1$ when using the weighted periodogram Daniell estimate. This allows a further correction in the empirical size, as in Monte Carlo simulations larger bandwidths M (smaller m) are associated to better empirical size. For example, Monte Carlo simulations in Coroneo and Iacone (2020) indicate that it is possible to completely eliminate the size distortion documented by Clark (1999).

4 Monte Carlo study of size and power

We analyse the empirical size and power of the tests of equal predictive accuracy and of encompassing for density forecast by means of a Monte Carlo experiment. Since Kiefer and Vogelsang (2005), simulation studies have by now covered a fairly wide range of situations, including inference in regression models, in non-linear models, and others. We refer to Lazarus, Lewis, Stock and Watson (2018) for a recent, comprehensive study. In point forecasting, studies include Coroneo and Iacone (2020) on forecast evaluation in small samples, Harvey, Leybourne and Whitehouse (2017) on forecast encompassing, and Li and Patton (2018) on forecast evaluation in large samples.

We already noticed that simulation studies find that fixed-smoothing asymptotics yield better approximation of the empirical size, and that this improvement is stronger the larger is the bandwidth M (the smaller is m). These works also find that the finite sample power is decreasing with the bandwidth, therefore documenting the existence of a trade-off between correct size and power. Lazarus, Lewis, Stock and Watson (2018), drawing on their extensive simulation study, recommend $M = \lfloor 1.3T^{1/2} \rfloor$ and $m = \lfloor 0.2T^{2/3} \rfloor$.

In this section, we check whether the size improvements for the equal predictive ability and forecast encompassing tests still hold in the case of density comparisons. We use a rather small sample that replicates the dimension of the sample of our dataset. We also examine the issue of bandwidth selection, and compare our results with Lazarus, Lewis, Stock and Watson (2018) and Coroneo and Iacone (2020).

In our Monte Carlo study, for simplicity we only consider the QPS loss function. We consider a sample of T observations, and we assume that the probability that the variable of interest y_t falls in bin k, for $k = 1, 2, 3$, is given by $\mathbf{y}'_t = (0, 1, 0)$. We also assume that we have two density forecasts that assign the probability that y_t falls in bin k as follows

$$\mathbf{f}'_{1,t} = (A_t, 1 - A_t, 0);$$
$$\mathbf{f}'_{2,t} = (0, 1 - B_t, B_t);$$

where

$$A_t = a_t + a_{t-1} + ... + a_{t-Q};$$
$$B_t = b_t + b_{t-1} + ... + b_{t-Q};$$

and a_t, ..., a_{t-Q} are realisations from a uniform distribution in $[0, \alpha/(Q+1)]$, b_t, ..., b_{t-Q} are realisations from a uniform distribution in $[0, \beta/(Q+1)]$, and a_t, ..., a_{t-Q}, b_t, ..., b_{t-Q} are all independently distributed.

The forecast errors are then given by

$$\mathbf{e}'_{1,t} = (-A_t, A_t, 0);$$
$$\mathbf{e}'_{2,t} = (0, B_t, -B_t).$$

In this setting, $E(\mathbf{e}'_{1,t}\mathbf{e}_{1,t}) = E(2A_t^2)$, $E(\mathbf{e}'_{2,t}\mathbf{e}_{2,t}) = E(2B_t^2)$. This means that the null hypothesis of the equal predictive ability test, $E(d_t) = E(\mathbf{e}'_{1,t}\mathbf{e}_{1,t}) - E(\mathbf{e}'_{2,t}\mathbf{e}_{2,t}) = 0$, follows from setting $\alpha = \beta$. For the forecast encompassing test, $E(\mathbf{e}'_{1,t}(\mathbf{e}_{1,t} - \mathbf{e}_{2,t})) = 2E(A_t^2) - E(A_t)E(B_t)$, so to obtain the null hypothesis $E(d_t) = E(\mathbf{e}'_{1,t}(\mathbf{e}_{1,t} - \mathbf{e}_{2,t})) = 0$, we set $\beta = 8\frac{4+3Q}{12(Q+1)}\alpha$. We can investigate the power of the equal predictive ability test setting $\beta = \sqrt{\alpha^2 - 3/2 \times c/\sqrt{T}}$ as we increase the value of c.

In our experiment we set $\alpha = \beta = 1$ for the equal predictive accuracy test and $\alpha = 3/8$ for the forecast encompassing test, and Q up to 6, with sample size set at $T = 40$, 80, and we repeat the experiment for 10,000 replications. Our sample size is much smaller than the sample size of Lazarus, Lewis, Stock and Watson (2018), and it matches the dimension of the sample available for our empirical study. Indeed, checking the empirical performance in such small samples is one reason of interest in this experiment.

In Tables 1 and 2, we report the empirical size of the test with one-sided alternative $H_1 : \{E(d_t) > 0\}$ when 5% critical values from both standard asymptotics and fixed-smoothing asymptotics are used. In columns WCE, the long run variance estimate is computed using a Bartlett kernel with bandwidths $M = \lfloor T^{1/3} \rfloor$, $M = \lfloor T^{1/2} \rfloor$ and $M = T$. In columns WPE, we use the Daniell kernel with bandwidths $m = \lfloor T^{1/4} \rfloor$, $m = \lfloor T^{1/3} \rfloor$, $m = \lfloor T^{1/2} \rfloor$ and $m = \lfloor T^{2/3} \rfloor$. Consistently with results from other simulation studies, standard asymptotics are associated with size distortions. The performance deteriorates as the dependence increases with Q, especially when the bandwidth m is too long (or, to a lesser extent, when M is too short), reflecting the fact that the dependence causes a curvature in the spectral density at larger frequencies, and thus a bias in the estimation of the spectral density in zero. The second source of distortion is due to the approximation

Table 1: Empirical size of the test of equal predictive ability

Panel A: standard asymptotics

		WCE			WPE			
T	Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
40	0	0.063	0.078	0.201	0.085	0.076	0.062	0.058
	2	0.102	0.098	0.218	0.084	0.078	0.082	0.117
	4	0.158	0.129	0.234	0.101	0.096	0.130	0.192
	6	0.197	0.152	0.245	0.109	0.117	0.176	0.232
80	0	0.061	0.071	0.208	0.090	0.070	0.061	0.055
	2	0.082	0.080	0.206	0.086	0.067	0.066	0.092
	4	0.117	0.098	0.220	0.090	0.077	0.087	0.160
	6	0.148	0.111	0.221	0.092	0.083	0.112	0.194

Panel B: fixed-smoothing asymptotics

		WCE			WPE			
T	Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
40	0	0.048	0.049	0.049	0.048	0.051	0.050	0.049
	2	0.083	0.064	0.058	0.047	0.051	0.067	0.108
	4	0.136	0.089	0.072	0.058	0.069	0.113	0.183
	6	0.173	0.112	0.081	0.066	0.082	0.157	0.222
80	0	0.051	0.048	0.049	0.049	0.050	0.050	0.051
	2	0.070	0.055	0.053	0.050	0.049	0.057	0.085
	4	0.104	0.074	0.059	0.054	0.058	0.077	0.152
	6	0.134	0.084	0.062	0.053	0.062	0.100	0.188

Note: empirical size of the equal predictive ability test with standard asymptotics (panel A) and fixed-smoothing asymptotics (panel B). The theoretical size is 5%, for a one-sided alternative hypothesis. Q indicates the dependence in the process. WCE refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance; WPE refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance.

Table 2: Empirical size of the forecast encompassing test

Panel A: standard asymptotics

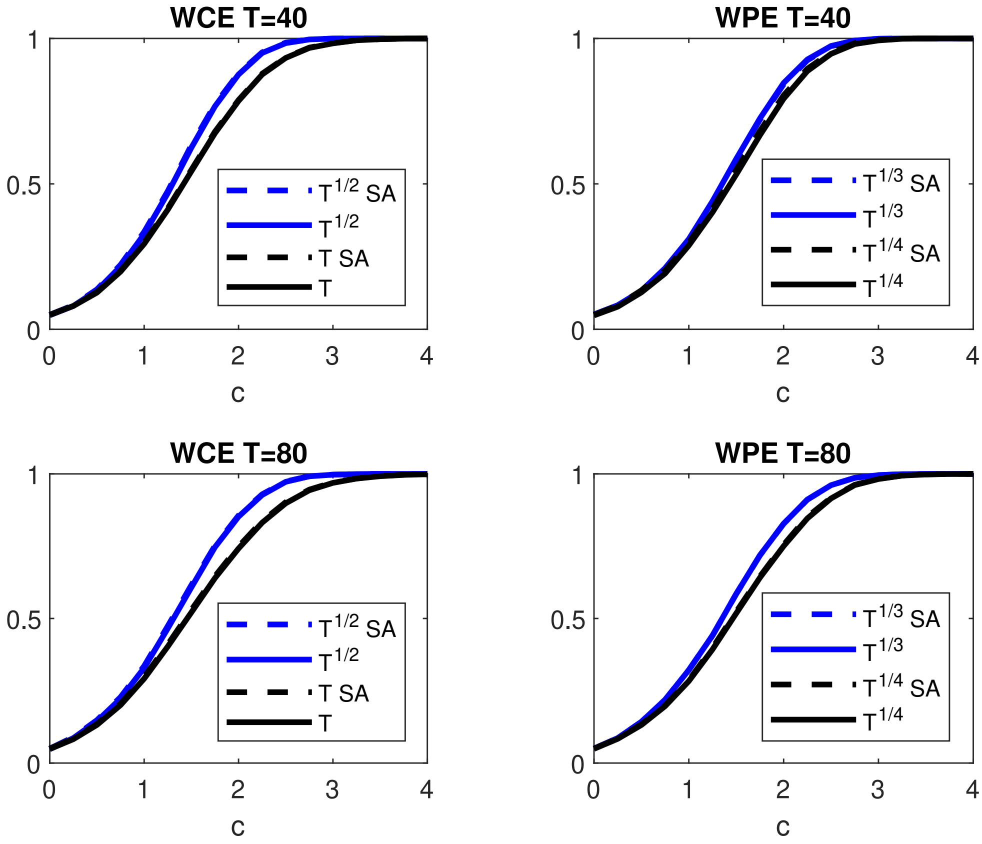
		WCE			WPE			
T	Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
40	0	0.049	0.061	0.187	0.072	0.060	0.047	0.044
	2	0.073	0.069	0.188	0.058	0.053	0.054	0.090
	4	0.122	0.091	0.203	0.072	0.067	0.096	0.159
	6	0.162	0.118	0.211	0.079	0.084	0.142	0.199
80	0	0.048	0.057	0.198	0.079	0.059	0.049	0.045
	2	0.064	0.059	0.186	0.069	0.051	0.047	0.071
	4	0.094	0.075	0.200	0.072	0.055	0.064	0.136
	6	0.124	0.086	0.194	0.069	0.059	0.090	0.167

Panel B: fixed-smoothing asymptotics

		WCE			WPE			
T	Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
40	0	0.035	0.036	0.040	0.040	0.038	0.037	0.039
	2	0.055	0.038	0.039	0.029	0.030	0.041	0.079
	4	0.100	0.060	0.049	0.038	0.042	0.080	0.149
	6	0.140	0.079	0.054	0.041	0.056	0.123	0.191
80	0	0.037	0.038	0.040	0.041	0.040	0.039	0.040
	2	0.050	0.039	0.039	0.035	0.033	0.038	0.065
	4	0.079	0.052	0.044	0.037	0.038	0.054	0.129
	6	0.110	0.061	0.046	0.035	0.041	0.077	0.161

Note: empirical size of the forecast encompassing test with standard asymptotics (panel A) and fixed-smoothing asymptotics (panel B). The theoretical size is 5%, for a one-sided alternative hypothesis. Q indicates the dependence in the process. WCE refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance; WPE refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance.

Figure 1: Finite sample power.



Note: power performances of the equal predictive ability test in a samples of size $T = 40$ and $T = 80$, for the theoretical size 5% and for a one-sided alternative hypothesis. The dashed lines refer to power performances using size-adjusted critical values while solid lines use fixed-smoothing asymptotics. The parameter c indicates the distance from the null hypothesis. WCE refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance; WPE refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance.

of the average periodogram as its probability limit, and this is more evident when m is too short ($m = \lfloor T^{1/4} \rfloor$), and when the bandwidth M is too long ($M = T$). Using fixed-smoothing asymptotics always improves the empirical size. As usual, the best performance is for for $M = T$ or the smallest m (as the size distortion due to the curvature of the spectral density is least, in this case), but on balance we observe correctly sized tests with WCE already with bandwidth $M = \lfloor T^{1/2} \rfloor$, likewise, we observe correct size with WPE already with bandwidth $m = \lfloor T^{1/3} \rfloor$.

For the power study we only consider the equal predictive ability test. We set $Q = 0$ and increasing values of c up to 4. In this case, we only consider bandwidths that are associated to good empirical size properties, namely WCE with $M = \lfloor T^{1/2} \rfloor$ and $M = T$, and WPE with $m = \lfloor T^{1/4} \rfloor$ and $m = \lfloor T^{1/3} \rfloor$, in all cases only for fixed-smoothing asymptotics. For the purpose of comparison only, we also plot the size adjusted power. Power performances are reported in Figure 1. In all cases the empirical power is a very good approximation of the size adjusted power, again offering support to the assumption that fixed-smoothing asymptotic is a valuable instrument for inference. We also find that, as a general rule, larger bandwidths M (smaller m) are associated to lower power, consistently with other similar simulation studies. Overall we suggest $M = \lfloor T^{1/2} \rfloor$ and $m = \lfloor T^{1/3} \rfloor$. Given our sample size, these bandwidth rules seem in line with the recommendation in Lazarus, Lewis, Stock and Watson (2018).

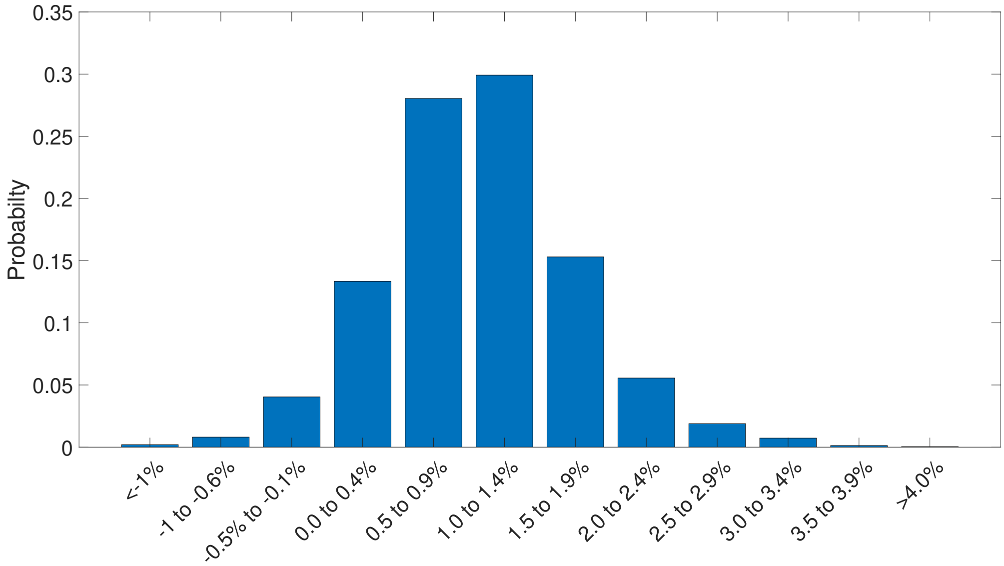
5 Application

We use the proposed density forecast evaluation tests with fixed-smoothing asymptotics to evaluate the predictive ability of density forecasts from the European Central Bank's Survey of Professional Forecasters (ECB SPF) for HICP inflation, the unemployment rate and real GDP growth against five simple benchmark density forecasts: a uniform, an unconditional Gaussian, a Gaussian random walk, a Gaussian random walk with drift, and naive benchmark that uses the previous survey value for the same forecast horizon. Of course, one can use more sophisticated benchmarks, but here our objective is to assess whether the ECB SPF survey forecasts can at least beat simple benchmarks.

5.1 The ECB Survey of Professional Forecasters

We use aggregate ECB SPF density forecasts at one and two years ahead for inflation (year-on-year percentage change of the Harmonised Index of Consumer Prices, HICP), real GDP growth (year-on-year percentage change of real GDP) and the unemployment rate (as percentage of the labour force).

Figure 2: ECB SPF density forecast for HICP one-year ahead, December 2016



Note: histogram of the one-year ahead aggregate density forecast for HICP from the 2016.Q1 survey round. Participants are asked to report a probability for the realisation in December 2016 to fall in each bin.

The ECB SPF is administered quarterly to a panel of forecasters (about 80 institutions with an average of 60 responses each round). Participants are experts affiliated with financial or non-financial institutions based within the European Union, and form an heterogeneous group to guarantee the representativeness and independence of the expectations collected.

Participants are asked to provide a forecast for the current calendar year, the following calendar year, the calendar year after that, a long term horizon, a rolling horizon one year ahead of the latest available data and a rolling horizon two years ahead of the latest available data. For more information on the ECB SPF see Garcia (2003) and Bowles, Friz, Genre, Kenny, Meyler and Rautanen (2007).

To report their density forecasts, participants are given a set of specific ranges and are asked to predict the probability that the target variable will fall in each specific range, or bin, with the first and the last being open intervals. The number of ranges given in every survey round can change but their width is fixed. The ECB SPF reports both the anonymised individual density forecasts and the aggregate density forecast, constructed by summing up the individual probabilities reported in the SPF and dividing by the number of respondents. For example, in Figure 2 we present the one year-ahead density forecast for the December 2016 HICP produced in the 2016.Q1 survey round.

5.2 Benchmark Density Forecasts

We compare ECB SPF density forecasts against five simple benchmark density forecasts: a uniform, an unconditional Gaussian, a Gaussian random walk, a Gaussian random walk with drift, and a naive forecast based on the lagged ECB SPF density forecast that, as such, incorporates all the information available at the previous survey round.

As forecasters operate using data as available at the time the forecasts are made, we construct the benchmark density forecasts using only the real-time information available to professional forecasters up to the deadline for responding to each survey round by using the historical vintages from the Euro Area Real-Time Database available on the European Central Bank Statistical Data Warehouse.

For the uniform benchmark, we use a uniform distribution with constant probability between the maximum and minimum of the target variable historical values as available at the survey deadline. For the unconditional Gaussian benchmark, we use a Gaussian distribution with mean and variance obtained from the historical observations of the target variable as available at each survey deadline. For the Gaussian random walk benchmark, we use a normal distribution with conditional expectation equal to the last observation available at the survey deadline and variance calculated using all historical observations as available at each survey deadline. For the Gaussian random walk with drift benchmark, we use a normal distribution with conditional mean estimated using a random walk with drift and variance calculated using all historical observations as available at each survey deadline. From these predictive distributions, we compute the probability that the realization of the target variable falls inside each bin.

For the naive benchmark, we simply use the last available ECB SPF density forecast for the same horizon, i.e. $f_t^{k,Naive} = f_{t-1}^{k,SPF}$. In the case of different bins available from a survey round to the following, the forecasts are adjusted to accommodate the new bin structure. If in the new survey round there are more bins than in the previous, the probability of the last bin is equally split across the additional bins available; if there are less bins in the current survey round than the previous round, the probabilities of extreme bins are added up and placed in the only available bin. For additional discussion about the changing bin structure see D'Amico et al. (2008) and Manzan (2021).

5.3 Empirical Results

We analyse the ECB SPF aggregate density forecasts at the rolling horizons of one and two years for the unemployment rate, real GDP growth and HICP inflation for the surveys between 2000.Q1 and 2019.Q4, corresponding to a total of 80 quarterly observations. We also split the sample into two equally sized subsamples: 2000.Q1-2009.Q4 and 2010.Q1-2019.Q4, of 40 observations each. As shown in Section 4, with such small sample sizes the DM framework with standard asymptotics suffers from large size distortions, but fixed-smoothing asymptotics can still provide reliable inference.

The first vintage of the Euro Area Real-Time Database is for January 2001, while our sample starts in 2000.Q1. As a result, for the first year, we construct the benchmarks in real-time using the first vintage available (January 2001), and we mimic the real-time publication delays. Additionally, we construct all forecast errors using the first release as the actual value.

Density forecast evaluation test results are reported in Table 3 for the full sample, and in Tables 4 and 5 for the two subsamples. In Panel A, we report the equal predictive accuracy test, and in Panel B we report the forecast encompassing test. A negative value of the equal predictive accuracy test indicates that the benchmark is performing better than the ECB SPF forecast, while a negative value for the forecast encompassing test

Table 3: Forecast evaluation tests. Full sample Q1.2000 - Q4.2019, $T = 80$.

Panel A: Equal Predictive Ability Test

		1 year ahead					2 years ahead				
Variable	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	4.76*	5.01*	3.91*	3.74*	5.42*	1.75	2.12*	1.83	2.85*	3.74*
	WPE	3.82*	4.07*	3.55*	3.33*	6.58*	1.44	1.79	1.67	2.48*	4.99*
GDP	WCE	2.43*	2.24*	2.08*	2.14*	2.70*	-1.11	-1.39	1.72	2.03*	1.47
	WPE	2.10*	2.04*	1.76	1.81	2.44*	-0.90	-1.12	1.56	1.83	1.57
HICP	WCE	0.90	0.58	0.94	1.15	2.19*	-0.23	0.46	1.96*	2.07*	1.14
	WPE	0.89	0.50	0.87	1.06	1.74	-0.21	0.40	1.92*	1.89*	0.99

Panel B: Forecast Encompassing Test

		1 year ahead					2 years ahead				
Variable	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	0.34	0.38	-0.29	-0.46	-4.18	1.36	1.36	0.02	-0.94	-2.83
	WPE	0.27	0.30	-0.27	-0.44	-4.24	1.09	1.09	0.02	-0.80	-3.53
GDP	WCE	0.77	0.57	0.06	0.10	-2.63	2.75*	2.69*	1.48	1.26	-0.52
	WPE	0.64	0.50	0.05	0.09	-2.46	2.20*	2.16*	1.33	1.13	-0.60
HICP	WCE	3.48*	1.41	1.55	1.19	-1.47	3.74*	1.33	0.10	-0.02	-0.57
	WPE	3.54*	1.27	1.40	1.07	-1.18	3.38*	1.17	0.10	-0.02	-0.50

Note: Equal predictive ability test statistics and the forecast encompassing test statistics for one-year and two-year ahead ECB SPF density forecasts against the uniform, unconditional Gaussian, the Gaussian random walk, the Gaussian random walk with drift, and the naive benchmark forecasts on the full sample Q1.2000 - Q4.2019 ($T = 80$) using the RPS loss. A negative equal predictive ability test statistic sign implies that benchmark performs better than the ECB SPF, and a negative value for the forecast encompassing test indicates that the estimated unrestricted weight on the benchmark is negative. Long run variances are estimated using WCE with Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$. One-sided significance at the 5% level is indicated with a grey shaded cell using standard asymptotics and with * using fixed-smoothing asymptotics.

indicates that the unrestricted weight on the benchmark is negative, as it does not have any additional information with respect to the ECB SPF forecast. One-sided rejections at the 5% level from standard asymptotics critical values are indicated shading the appropriate cell and rejections using fixed-smoothing asymptotics critical values are reported using *. We only present results for the RPS loss, because as noted in Section 2, it is a proper loss function in the sense that it encourages the forecasters to reveal their true beliefs; results for the QPS metric are available in Appendix D. The long run variance is estimated using WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$.

For the unemployment rate, the equal predictive ability test indicates that the one-year ahead ECB SPF forecasts outperform all the benchmarks in all the samples. At two-year ahead, the ECB SPF forecasts still outperform the Gaussian random walk with drift and the naive benchmarks on the full sample. The subsample analysis reveals an improvement of the relative performance in the second subsample with evidence of superior predictive ability against the uniform, the unconditional Gaussian and the naive benchmarks. Moreover, the ECB SPF encompasses all the benchmarks at both horizons. Overall, these results indicate that for unemployment the ECB SPF provides more accurate predictions than the benchmarks at the one year-ahead horizon and, in the second period, also at the two-year horizon.

Results for real GDP growth are less favourable for the ECB SPF. At the one-year horizon, the ECB SPF outperforms the benchmarks on the full sample (but the results are less clear cut if the WPE estimate is used). This is mostly due to the superior performance of the ECB SPF in the second subsample. At the two-year horizon, the ECB SPF does not consistently outperform any benchmark on the full sample. In fact, the forecast encompassing test suggests that at the two-year horizon the ECB SPF can be improved by combinations with the uniform and the unconditional Gaussian benchmarks in the whole sample and in the first subsample. Overall, these results indicate that for

Table 4: Forecast evaluation tests. Subsample Q1.2000 - Q4.2009, $T = 40$.

Panel A: Equal Predictive Ability Test

		1 year ahead					2 years ahead				
	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	3.08*	3.67*	3.18*	2.93*	2.80*	0.37	1.02	1.47	2.81*	2.27*
	WPE	2.52*	3.05*	2.64*	2.44*	2.37*	0.31	0.87	1.32	2.56*	1.87
GDP	WCE	1.31	1.20	1.55	1.64	2.20*	-1.49	-2.23	1.28	1.56	1.04
	WPE	1.08	0.99	1.45	1.53	2.11*	-1.40	-2.01	1.16	1.43	1.31
HICP	WCE	-0.65	-1.05	1.32	1.64	0.90	-0.57	-1.06	1.29	1.55	-0.27
	WPE	-0.58	-0.94	1.19	1.48	0.83	-0.46	-0.82	1.13	1.32	-0.26

Panel B: Forecast Encompassing Test

		1 year ahead					2 years ahead				
	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	0.50	0.39	0.10	-0.20	-1.74	1.47	1.41	0.97	-0.73	-1.70
	WPE	0.42	0.33	0.11	-0.20	-1.93	1.23	1.20	1.16	-0.88	-1.41
GDP	WCE	0.95	0.94	0.32	0.36	-1.99	2.75*	2.92*	0.90	0.77	-0.13
	WPE	0.81	0.78	0.29	0.32	-2.05	2.52*	2.64*	0.81	0.69	-0.17
HICP	WCE	3.27*	3.61*	0.49	-0.13	-0.30	3.15*	3.22*	0.13	-0.50	0.69
	WPE	2.87*	3.14*	0.43	-0.11	-0.29	2.47*	2.38*	0.11	-0.40	0.67

Note: Equal predictive ability test statistics and the forecast encompassing test statistics for one-year and two-year ahead ECB SPF density forecasts against the uniform, unconditional Gaussian, the Gaussian random walk, the Gaussian random walk with drift, and the naive benchmark forecasts on the subsample Q1.2000 - Q4.2009 ($T = 40$) using the RPS loss. A negative equal predictive ability test statistic sign implies that benchmark performs better than the ECB SPF, and a negative value for the forecast encompassing test indicates that the estimated unrestricted weight on the benchmark is negative. Long run variances are estimated using WCE with Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$. One-sided significance at the 5% level is indicated with [grey shading] using standard asymptotics and with * using fixed-smoothing asymptotics.

Table 5: Forecast evaluation tests. Subsample Q1.2010 - Q4.2019, $T = 40$.

Panel A: Equal Predictive Ability Test

		1 year ahead					2 years ahead				
Variable	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	4.25*	3.79*	2.21*	2.39*	5.10*	2.99*	2.72*	1.36	1.62	2.34*
	WPE	3.52*	3.17*	1.98*	2.16*	4.82*	2.65*	2.39*	1.26	1.49	2.10*
GDP	WCE	3.89*	3.19*	2.82*	2.82*	2.67*	-0.11	0.14	1.80	2.08*	0.86
	WPE	3.35*	2.69*	3.35*	3.33*	2.50*	-0.11	0.13	1.97*	2.25*	0.73
HICP	WCE	2.50*	1.24	-0.01	0.07	2.32*	0.12	1.34	1.61	1.55	1.98
	WPE	1.97*	1.10	-0.01	0.06	1.90	0.10	1.15	1.25	1.22	1.97*

Panel B: Forecast Encompassing Test

		1 year ahead					2 years ahead				
Variable	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	-0.05	0.15	-0.51	-0.54	-4.80	0.36	0.47	-0.51	-0.81	-1.84
	WPE	-0.04	0.12	-0.46	-0.48	-4.79	0.32	0.42	-0.47	-0.74	-1.67
GDP	WCE	-0.09	-0.50	-0.71	-0.71	-2.25	1.58	1.35	1.54	1.47	-0.53
	WPE	-0.08	-0.42	-1.02	-1.01	-2.12	1.50	1.28	1.25	1.16	-0.44
HICP	WCE	1.70	-0.08	1.73	1.65	-1.76	2.25*	-0.31	0.04	0.27	-1.63
	WPE	1.30	-0.07	1.40	1.37	-1.41	1.87	-0.26	0.03	0.21	-1.65

Note: Equal predictive ability test statistics and the forecast encompassing test statistics for one-year and two-year ahead ECB SPF density forecasts against the uniform, unconditional Gaussian, the Gaussian random walk, the Gaussian random walk with drift, and the naive benchmark forecasts on the subsample Q1.2010 - Q4.2019 ($T = 40$) using the RPS loss. A negative equal predictive ability test statistic sign implies that benchmark performs better than the ECB SPF, and a negative value for the forecast encompassing test indicates that the estimated unrestricted weight on the benchmark is negative. Long run variances are estimated using WCE with Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$. One-sided significance at the 5% level is indicated with ▒ using standard asymptotics and with * using fixed-smoothing asymptotics.

real GDP growth, ECB SPF can outperform simple benchmarks at least at one-year horizon, especially in the second subsample.

In the case of inflation, the ECB SPF does not statistically outperform the benchmarks, except, in the whole sample, the Gaussian Random walks at the two-year ahead horizon, and, in the second subsample, the uniform (at the one-year horizon) and possibly the naive. In fact, the forecast encompassing tests indicate that in the first subsample the uniform and the unconditional Gaussian are not encompassed by the ECB SPF, suggesting that one can improve the ECB SPF density forecasts by combining them with these benchmarks.

As for the benchmarks, the ECB SPF easily outperforms and encompasses the naive benchmark, indicating that professional forecasters update their information set when making their predictions and that previous round forecasts are uninformative. On the other hand, the uniform, and to a lesser extent, the unconditional Gaussian benchmarks seem the most difficult to outperform and encompass, especially for two-year ahead forecasts.

Comparing the application of standard asymptotics with fixed-smoothing asymptotics, we reject the null hypotheses of equal predictive ability or of no encompassing more frequently for the tests with standard asymptotics. Spurious rejections occur in the whole sample especially at the two-year horizon, due to a higher level of dependence in d_t, which exacerbates the size distortions induced by standard asymptotics, see Section 4. For example, in Table 3 standard asymptotics indicate that, for the unemployment rate, the ECB SPF outperforms all the benchmarks at the two-year horizon, but this evidence is partially spurious and therefore not fully confirmed when using fixed-smoothing asymptotics.

The application of fixed-smoothing asymptotics also allows us to perform reliable inference even on the two equally-sized subsamples of only 40 observations, when the size distortions of standard asymptotics can be quite large, see Section 4. The forecast encompassing test results for the second subsample in Table 5 indicate that for inflation the ECB SFP encompasses all the benchmarks, whereas standard asymptotics would have often lead us to the opposite conclusion, demonstrating the risks of using standard asymptotics in such a small sample. Therefore, fixed-smoothing asymptotics reinforces one of our key findings, namely that the performance of the ECB SPF improves relative to the benchmarks in the second subsample.

6 Conclusions

In this paper, we apply fixed-b and fixed-m asymptotics to tests of equal predictive accuracy and encompassing for survey density forecasts. In an original Monte Carlo design, we verify that fixed-smoothing asymptotics delivers correctly sized tests in this framework, even when only a small number of out of sample observations is available.

We apply the density forecast evaluation tests with fixed-smoothing asymptotics to evaluate the predictive ability of density forecasts from the European Central Bank's Survey of Professional Forecasters (ECB SPF) over the period 2001.Q1-2019.Q4, taking as benchmarks simple density forecasts generated from a uniform distribution, an unconditional Gaussian distribution, a Gaussian random walk distribution, a Gaussian random walk with drift distribution, and the previous survey round forecast.

Our results indicate that ECB SPF density forecasts for unemployment and real GDP growth outperformed and sometimes encompassed the benchmarks, especially at one year ahead and in the second subsample. Survey forecasts for inflation do not easily outperform nor encompass the benchmarks, but again the relative performance improved in the second subsample. For all the variables, the application of fixed smoothing asymptotics strengthens the evidence of an improvement in relative predictive ability since 2010, indicating a change in the forecasting practice after the financial crisis.

A Density forecast encompassing test

In this Appendix, we show the null hypothesis of density forecast encompassing can be tested using the DM framework by defining d_t as in (6).

If we denote the forecast errors associated with $\mathbf{f}_{t,c}(\lambda)$ in (5) as $\mathbf{e}_{t,c}(\lambda) = \mathbf{y}_t - \mathbf{f}_{t,c}(\lambda)$, then, the optimal weight in the minimum QPS sense has

$$\widehat{\lambda} = \arg\min \sum_{t=1}^{T} (\mathbf{e}_{t,c}(\lambda)'\mathbf{e}_{t,c}(\lambda)). \tag{12}$$

The derivative is

$$\frac{\partial}{\partial\lambda}\sum_{t=1}^{T}(\mathbf{e}_{t,c}(\lambda)'\mathbf{e}_{t,c}(\lambda)) = \sum_{t=1}^{T} 2(\mathbf{y}_t - \mathbf{f}_{t,c}(\lambda))'\frac{\partial}{\partial\lambda}(-\mathbf{f}_{t,c}(\lambda)) = \sum_{t=1}^{T} 2(\mathbf{y}_t - \mathbf{f}_{t,c}(\lambda))'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2})$$

and the first order condition therefore gives

$$\sum_{t=1}^{T} 2(\mathbf{y}_t - \mathbf{f}_{t,c}(\lambda))'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2}) = 0.$$

which is met for $\lambda = \widehat{\lambda}$ (i.e., $\widehat{\lambda}$ is defined in this way).

Let

$$d_t(\lambda) = -(\mathbf{y}_t - \mathbf{f}_{t,c}(\lambda))'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2})$$

If $\mathbf{y}_t$, $\mathbf{f}_{1,t}$ and $\mathbf{f}_{2,t}$ are jointly mixing with a sufficient rate, then so is $d_t(\lambda)$.

Denoting $\sigma_T^2(\lambda) = Var(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} d_t(\lambda))$ as the long run variance, assuming that $d_t(\lambda)$ is mixing with sufficient rate and $\sigma_T(\lambda) > 0$ then we have a CLT for standardised sum of $d_t(\lambda)$. This suggests a LM type test for forecast encompassing. Mimicking the first order condition, denote λ_0 as the value of λ that gives $E(d_t(\lambda))|_{\lambda=\lambda_0} = 0$, then

$$\sqrt{T}\frac{1/T\sum_{t=1}^{T} d_t(\lambda_0)}{\sigma_T(\lambda_0)} \to_d N(0,1)$$

and, under $H_0 : \{\lambda_0 = 0\}$ then

$$\sqrt{T}\frac{1/T\sum_{t=1}^{T} d_t}{\sigma_T} \to_d N(0,1)$$

where we used d_t and σ_T in place of $d_t(0)$ and $\sigma_T(0)$ to shorten the notation.

Rewriting

$$d_t(\lambda) = -(\mathbf{y}_t - \mathbf{f}_{t,1} - \lambda(\mathbf{f}_{t,1} - \mathbf{f}_{t,2}))'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2})$$

then

$$d_t = \mathbf{e}'_{t,1}(\mathbf{e}_{t,1} - \mathbf{e}_{t,2})$$

These facts suggest for $H_0 : \{\lambda_0 = 0\}$ the test statistic

$$\sqrt{T}\frac{1/T\sum_{t=1}^{T} d_t}{\widehat{\sigma}} = \sqrt{T}\frac{1/T\sum_{t=1}^{T} \mathbf{e}'_{t,1}(\mathbf{e}_{t,1} - \mathbf{e}_{t,2})}{\widehat{\sigma}}$$

for an appropriate estimate of the long rung variance $\widehat{\sigma}$.

To complete the specification of the test and to check the power, we rewrite

$$\begin{aligned}
\frac{\sqrt{T}}{T}\sum_{t=1}^{T} d_t =& \frac{\sqrt{T}}{T}\sum_{t=1}^{T}(d_t - d_t(\lambda_0) + d_t(\lambda_0)) \\
=& \frac{\sqrt{T}}{T}\sum_{t=1}^{T}(d_t - d_t(\lambda_0))) + \frac{\sqrt{T}}{T}\sum_{t=1}^{T}(d_t(\lambda_0)) \\
=& \frac{\sqrt{T}}{T}\sum_{t=1}^{T}(d_t - d_t(\lambda_0)) + O_p(1)
\end{aligned}$$

and notice that

$$\begin{aligned}
d_t - d_t(\lambda_0) =& -(y_t - \mathbf{f}_{t,c}(0))'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2}) + (\mathbf{y}_t - \mathbf{f}_{t,c}(\lambda_0))'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2}) \\
=& (\mathbf{f}_{t,c}(0) - \mathbf{f}_{t,c}(\lambda_0))'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2}) \\
=& (\mathbf{f}_{t,1} - (1-\lambda_0)\mathbf{f}_{t,1} - \lambda_0\mathbf{f}_{t,2})'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2}) \\
=& \lambda_0(\mathbf{f}_{t,1} - \mathbf{f}_{t,2})'(\mathbf{f}_{t,1} - \mathbf{f}_{t,2}) = \lambda_0(\mathbf{e}_{t,1} - \mathbf{e}_{t,2})'(\mathbf{e}_{t,1} - \mathbf{e}_{t,2})
\end{aligned}$$

Thus, for the alternative $H_A : \{\lambda_0 > 0\}$ the null hypothesis is rejected if the test statistic takes a value larger than the critical value.

Notice that the value that solves $E(d_t(\lambda_0)) = 0$ is

$$\lambda_0 = \frac{E(\mathbf{e}'_{t,1}(\mathbf{e}_{t,1} - \mathbf{e}_{t,2}))}{E(\mathbf{e}_{t,1} - \mathbf{e}_{t,2})'(\mathbf{e}_{t,1} - \mathbf{e}_{t,2})}$$

so if, for example, $\mathbf{e}_{t,1}$ and $\mathbf{e}_{t,2}$ are vectors of independent, identically distributed sequencies, independent from each other, then $\lambda_0 = 1/2$. On the other hand, if $E(\mathbf{e}'_{t,1}(\mathbf{e}_{t,1} - \mathbf{e}_{t,2})) = 0$ then $\lambda_0 = 0$.

The Ranked Probability Score (RPS) may be treated in the same way, using the cumulative distribution functions of each density forecast $\mathbf{F}_{t,i}$ and of the individual realisation $\mathbf{Y}_t$.

B Assumption for fixed-smoothing limits

Assumption 1 *Partial sums of d_t are such that the functional central limit theorem (FCLT) holds*

$$\frac{\sqrt{T}}{T}\frac{1}{\sigma_T}\sum_{t=1}^{\lfloor rT \rfloor} d_t \Rightarrow W(r)$$

where $\lfloor . \rfloor$ denotes the integer part of a number, $r \in [0,1]$ and $W(r)$ is a standard Brownian motion.

Assumption 1 is sufficient to establish the fixed-smoothing limits (10) and (11). This assumption is not primitive, but it is convenient because it may be established under a range of conditions. For example, Phillips and Solo (1992) consider linear processes of independent, identically distributed innovations of martingale difference sequences. On the other hand, Wooldridge and White (1988) consider mixing processes, thus allowing for forms of heteroskedasticity that may also induce non-stationarity, under the additional assumption that $Var\left(\sqrt{T}/T \; 1/\sigma_T \sum_{t=1}^{\lfloor rT \rfloor} d_t\right) \to r$. In view of the non-linearity in the loss function, establishing a linear representation for d_t from primitive assumptions on $f_t^{k,i}$ and y_t^k may be very challenging, whereas establishing mixing properties may be easier, especially when $f_t^{k,i}$ and y_t^k are limited to being M-dependent processes. However, as the two classes may overlap but are not included in each other, see discussion in Phillips and Solo (1992) and Andrews (1984), we prefer the more general assumption given here, that encompasses them both.

C Average scores and summary statistics

Sample averages for the RPS scores are reported in Table 6. The ECB SPF always achieves one of lowest score across variables and forecasting horizons, although situations in which the ECB SPF is outperformed by a benchmark also occurred (most notably, in the 2000-2009 subsample for GDP at two years with the uniform and for HICP with the unconditional Gaussian). Averages scores increase in the second subsample for the unemployement rate and, at the two-year horizon, for HICP, suggesting a decrease in predictability since 2010 for these variables. However, in these cases the deterioration of the average RPS scores is usually less marked for the ECB SPF than for the benchmarks.

Summary statistics of the process d_t for unemployment, GDP and HICP forecasts against the benchmarks are in Tables 7-9. We report the full sample mean, the standard deviation, and autocorrelations up to the fourth lag and the eighth lag. The latter may be worth looking at because in the presence of large autocorrelation of d_t (relative to the sample size), the DM test may have low power, see Coroneo and Iacone (2021).

Results for the equal predictive ability case are in the top panel of each table. A positive entry for the sample mean indicates that the forecasters are more accurate than the benchmark, which is often the case. The autocorrelation profile seems usually more relevant when the unconditional Gaussian benchmark is used. Fortunately, even in these cases we can see that the autocorrelation coefficients at lag eight are usually quite small, indicating a quick decay of the dependence.

Results for the forecast encompassing case are in the bottom panel of each table. A negative entry for the sample mean is associated with an unrestricted negative estimate of the weight in the forecast combination (which is not feasible given that the weight needs to be between 0 and 1). We interpret this result as evidence that the SPF forecast cannot be improved by a linear combination with the benchmark. This most often happens for the naive benchmark. In Table 10, we report the estimated forecast combination weights for the full sample (from 2000.Q1 to 2019.Q4) and the two subsamples (2000.Q1

Table 6: Average RPS Scores.

Full sample Q1.2000 - Q4.2019, $T = 80$

		SPF	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	1YR	0.72	1.81	2.04	0.96	0.99	0.97
	2YR	1.41	1.99	2.20	1.65	1.79	1.62
GDP	1YR	1.05	1.56	1.55	1.79	1.85	1.33
	2YR	1.85	1.59	1.57	2.45	2.64	1.88
HICP	1YR	1.10	1.18	1.21	1.26	1.31	1.15
	2YR	1.21	1.19	1.27	1.64	1.75	1.23

Subsample Q1.2000 - Q4.2009, $T = 40$.

		SPF	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	1YR	0.66	1.50	1.94	0.96	1.03	0.90
	2YR	1.39	1.58	2.02	1.58	1.82	1.61
GDP	1YR	1.15	1.65	1.64	1.97	2.06	1.57
	2YR	1.95	1.47	1.37	2.73	3.03	1.99
HICP	1YR	1.11	1.03	0.97	1.43	1.51	1.14
	2YR	1.09	1.03	0.97	1.49	1.62	1.08

Subsample Q1.2010 - Q4.2019, T = 40.

		SPF	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	1YR	0.77	2.13	2.14	0.96	0.96	1.04
	2YR	1.43	2.40	2.38	1.72	1.75	1.63
GDP	1YR	0.95	1.47	1.46	1.62	1.64	1.10
	2YR	1.74	1.71	1.77	2.16	2.25	1.77
HICP	1YR	1.09	1.34	1.44	1.08	1.10	1.16
	2YR	1.34	1.36	1.57	1.79	1.87	1.37

Note: Sample averages of the RPS scores for one-year and two-year ahead forecast, ECB SPF and benchmarks. The top panel refers to the full sample (Q1.2000 - Q4.2019, $T = 80$), the middle panel to the first half-sample (Q1.2000 - Q4.2009, $T = 40$) and the bottom panel to the second-half sample (Q1.2010 - Q4.2019, $T = 40$).

to 2009.Q4 and 2010.Q1 to 2019.Q4). Comparing across the two subsamples, we can see that the estimated forecast combination weights are in general smaller in the second period, suggesting that overall the benchmarks contained less additional information in the second part of the sample.

Table 7: Summary Statistics of d_t for unemployment rate.

Panel A: Equal Predictive Ability Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	1.10	1.33	0.25	0.28	0.25	0.58	0.79	0.24	0.37	0.21
STD	1.05	1.17	0.37	0.41	0.40	1.48	1.62	0.64	0.65	0.48
AC1	0.82	0.81	0.63	0.66	0.30	0.86	0.86	0.74	0.71	0.35
AC2	0.59	0.61	0.23	0.28	-0.06	0.63	0.64	0.38	0.32	0.02
AC3	0.36	0.40	0.04	0.10	-0.06	0.41	0.43	0.22	0.18	-0.04
AC4	0.20	0.25	0.01	0.04	-0.06	0.22	0.26	0.15	0.19	-0.02
AC8	-0.22	-0.23	-0.01	-0.11	0.01	-0.23	-0.19	-0.04	-0.19	-0.07

Panel B: Forecast Encompassing Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.03	0.04	-0.01	-0.01	-0.07	0.25	0.27	0.00	-0.05	-0.07
STD	0.43	0.49	0.16	0.17	0.16	0.79	0.88	0.28	0.25	0.19
AC1	0.72	0.71	0.50	0.50	0.28	0.87	0.85	0.69	0.64	0.39
AC2	0.45	0.43	0.05	0.05	-0.14	0.64	0.62	0.27	0.19	0.08
AC3	0.27	0.28	0.11	0.11	-0.05	0.44	0.43	0.15	0.10	-0.05
AC4	0.18	0.20	0.21	0.17	-0.10	0.25	0.27	0.16	0.18	-0.05
AC8	-0.14	-0.11	0.31	0.15	0.23	-0.23	-0.24	0.26	0.12	-0.01

Note: sample mean, standard deviation (STD), and autocorrelation coefficients up to order 4, and order 8 (AC1, AC2, AC3, AC4, AC8) of d_t for the unemployment rate using the RPS loss function. The top panel refers to d_t as defined for the equal predictive ability test in (3), and the bottom panel refers to d_t as defined for the forecast encompassing test in (6).

Table 8: Summary Statistics of d_t for GDP growth.

Panel A: Equal Predictive Ability Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.51	0.50	0.74	0.80	0.28	-0.25	-0.28	0.60	0.80	0.03
STD	1.07	1.21	1.71	1.83	0.60	1.00	0.84	1.95	2.22	0.25
AC1	0.70	0.66	0.72	0.70	0.50	0.81	0.82	0.75	0.74	0.15
AC2	0.40	0.33	0.44	0.41	0.18	0.55	0.58	0.33	0.31	-0.19
AC3	0.18	0.11	0.27	0.25	0.18	0.37	0.39	0.02	0.00	-0.19
AC4	0.06	0.01	0.13	0.11	0.14	0.27	0.32	-0.11	-0.12	-0.12
AC8	-0.05	-0.04	0.05	0.06	0.11	0.03	0.24	-0.04	-0.02	0.09

Panel B: Forecast Encompassing Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.05	0.03	0.00	0.01	-0.08	0.34	0.29	0.15	0.15	-0.01
STD	0.35	0.35	0.41	0.43	0.21	0.52	0.47	0.49	0.55	0.13
AC1	0.67	0.52	0.71	0.72	0.45	0.83	0.81	0.80	0.80	0.12
AC2	0.38	0.26	0.41	0.43	-0.01	0.60	0.55	0.49	0.50	-0.20
AC3	0.17	0.05	0.15	0.16	-0.03	0.41	0.35	0.25	0.25	-0.20
AC4	0.14	0.02	0.04	0.03	0.04	0.31	0.27	0.08	0.09	-0.13
AC8	0.05	0.08	0.15	0.15	0.18	0.10	0.25	0.06	0.07	0.09

Note: sample mean, standard deviation (STD), and autocorrelation coefficients up to order 4, and order 8 (AC1, AC2, AC3, AC4, AC8) of d_t for GDP growth using the RPS loss function. The top panel refers to d_t as defined for the equal predictive ability test in (3), and the bottom panel refers to d_t as defined for the forecast encompassing test in (6).

Table 9: Summary Statistics of d_t for HICP.

Panel A: Equal Predictive Ability Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.08	0.11	0.16	0.21	0.05	-0.02	0.06	0.43	0.53	0.01
STD	0.53	0.79	0.88	0.92	0.23	0.52	0.58	1.08	1.24	0.11
AC1	0.51	0.72	0.61	0.63	0.15	0.52	0.69	0.71	0.73	-0.02
AC2	0.33	0.61	0.38	0.40	-0.29	0.27	0.51	0.43	0.46	-0.02
AC3	0.20	0.51	0.18	0.22	-0.03	0.12	0.36	0.26	0.28	0.03
AC4	0.06	0.32	0.08	0.16	0.13	-0.04	0.19	0.15	0.18	0.19
AC8	-0.04	0.04	-0.28	-0.26	0.09	-0.21	-0.02	-0.34	-0.34	-0.08

Panel B: Forecast Encompassing Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.14	0.11	0.10	0.09	-0.02	0.16	0.08	0.01	0.00	0.00
STD	0.25	0.36	0.35	0.36	0.11	0.26	0.29	0.38	0.43	0.05
AC1	0.42	0.64	0.57	0.60	0.12	0.51	0.68	0.68	0.72	-0.02
AC2	0.26	0.53	0.45	0.49	-0.30	0.27	0.48	0.45	0.49	-0.02
AC3	0.13	0.42	0.21	0.26	-0.02	0.12	0.34	0.28	0.32	0.05
AC4	-0.01	0.25	0.09	0.16	0.13	-0.04	0.15	0.06	0.11	0.18
AC8	-0.09	0.04	-0.25	-0.20	0.07	-0.23	-0.08	-0.27	-0.29	-0.10

Note: sample mean, standard deviation (STD), and autocorrelation coefficients up to order 4, and order 8 (AC1, AC2, AC3, AC4, AC8) of d_t for HICP using the RPS loss function. The top panel refers to d_t as defined for the equal predictive ability test in (3), and the bottom panel refers to d_t as defined for the forecast encompassing test in (6).

Table 10: Estimated forecast combination weights

Full sample Q1.2000 - Q4.2019, $T = 80$

	1 year ahead					**2 years ahead**				
Variable	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	0.03	0.03	0.00	0.00	0.00	0.23	0.21	0.01	0.00	0.00
GDP	0.09	0.06	0.01	0.01	0.00	0.80	0.95	0.17	0.14	0.00
HICP	0.38	0.34	0.28	0.23	0.00	0.53	0.37	0.02	0.00	0.00

Subsample Q1.2000 - Q4.2009, $T = 40$.

	1 year ahead					**2 years ahead**				
Variable	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	0.07	0.04	0.01	0.00	0.00	0.41	0.30	0.16	0.00	0.00
GDP	0.15	0.14	0.05	0.05	0.00	1.00	1.00	0.15	0.11	0.00
HICP	0.64	0.72	0.09	0.00	0.00	0.61	0.73	0.03	0.00	0.76

Subsample Q1.2010 - Q4.2019, $T = 40$.

	1 year ahead					**2 years ahead**				
Variable	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	0.00	0.01	0.00	0.00	0.00	0.05	0.07	0.00	0.00	0.00
GDP	0.00	0.00	0.00	0.00	0.00	0.54	0.45	0.20	0.17	0.00
HICP	0.19	0.00	0.50	0.48	0.00	0.47	0.00	0.01	0.06	0.00

Note: LS estimate of λ in (5) using the RPS loss function. The top panel refers to the full sample (Q1.2000 - Q4.2019, $T = 80$), the middle panel to the first half-sample (Q1.2000 - Q4.2009, $T = 40$) and the bottom panel to the second-half sample (Q1.2010 - Q4.2019, $T = 40$).

D QPS results

In this Appendix, we report sample averages for the QPS scores, summary statistics of the process d_t, the equal predictive ability tests and the forecast encompassing tests using the Quadratic Probability Score (QPS) loss function.

Sample averages for the QPS scores are reported in Table 11. Results are similar to the ones with the RPS loss function, and the ECB SPF still achieves some of the lowest scores across benchmarks and horizons. However, with this loss function, the superior performance of the ECB SPF against the benchmarks and the change in average scores for the second sub-sample are less noticeable.

Summary statistics of the process d_t for unemployment, GDP and HICP forecasts are in Tables 12-14. We report the full sample mean, the standard deviation, and autocorrelations up to the fourth lag and the eighth lag for the QPS loss function. Results for the equal predictive ability case are in the top panel of each table. A positive entry for the sample mean indicates that the forecasters are more accurate than the benchmark. Results for the forecast encompassing case are in the bottom panel of each table. A negative entry for the sample mean is associated with an unrestricted negative estimate of the weight in the forecast combination.

In Table 15, we report the estimated forecast combination weights for the full sample (from 2000.Q1 to 2019.Q4) and the two subsamples (2000.Q1 to 2009.Q4 and 2010.Q1 to 2019.Q4). Weights have a pattern similar to the one observed with RPS, but these are often higher.

Density forecast evaluation test results are reported in Table 16, for the full sample, and in Tables 17 and 18, for the two subsamples. In Panel A, we report the equal predictive accuracy test, and in Panel B we report the forecast encompassing test. A negative value of the equal predictive accuracy test indicates that the benchmark is performing better than the ECB SPF forecast, while a negative value for the forecast encompassing test indicates that the unrestricted weight on the benchmark is negative, as it does not have any additional information with respect to the ECB SPF forecast. One-sided rejections at the 5% level from standard asymptotics critical values are indicated shading the appropriate cell and rejections using fixed-smoothing asymptotics critical values are reported using *. The long run variance is estimated using WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$.

Overall, using the QPS in the equal predictive ability test yields qualitatively the same outcome as using the RPS, although significant results are less frequent when the QPS is used, see for example the case of the real GDP growth. For the forecast encompassing test, we find that the null of no encompassing is rejected more often with the QPS. The comparison of the outcomes of the tests with the two different loss functions indicates that the ECB SPF place more probability in the neighbourhood of the effective outcome, often near-missing the true realisation. As for the performance of the individual benchmarks, we again find that the forecasts from naive benchmark are outperformed and encompassed statistically more often, confirming the conclusion that this evidence supports the conjecture that the participants in the survey successfully update their predictions.

Table 11: Average QPS Scores.

Full sample Q1.2000 - Q4.2019, $T = 80$

		SPF	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	1YR	0.79	0.92	0.97	0.84	0.84	0.88
	2YR	0.91	0.94	0.98	0.91	0.92	0.96
GDP	1YR	0.87	0.92	0.90	0.96	0.97	0.91
	2YR	0.97	0.92	0.90	1.00	1.02	0.97
HICP	1YR	0.91	0.89	0.87	0.88	0.89	0.91
	2YR	0.89	0.89	0.88	0.93	0.94	0.90

Subsample Q1.2000 - Q4.2009, $T = 40$.

		SPF	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	1YR	0.75	0.90	0.97	0.83	0.84	0.84
	2YR	0.88	0.92	0.97	0.89	0.91	0.95
GDP	1YR	0.91	0.92	0.90	0.95	0.97	0.96
	2YR	1.04	0.92	0.89	0.99	1.03	1.04
HICP	1YR	0.90	0.87	0.83	0.89	0.90	0.90
	2YR	0.86	0.88	0.83	0.90	0.91	0.86

Subsample Q1.2010 - Q4.2019, $T = 40$.

		SPF	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	1YR	0.83	0.94	0.97	0.85	0.85	0.93
	2YR	0.94	0.95	0.99	0.93	0.93	0.97
GDP	1YR	0.84	0.92	0.91	0.96	0.97	0.86
	2YR	0.91	0.92	0.91	1.00	1.01	0.90
HICP	1YR	0.92	0.91	0.90	0.87	0.87	0.93
	2YR	0.93	0.91	0.92	0.95	0.96	0.93

Note: Sample averages of the QPS scores for one-year and two-year ahead forecast, ECB SPF and benchmarks. The top panel refers to the full sample (Q1.2000 - Q4.2019, $T = 80$), the middle panel to the first half-sample (Q1.2000 - Q4.2009, $T = 40$) and the bottom panel to the second-half sample (Q1.2010 - Q4.2019, $T = 40$).

Table 12: Summary Statistics of d_t for unemployment rate, QPS.

Panel A: Equal Predictive Ability

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.13	0.18	0.05	0.06	0.10	0.03	0.07	0.00	0.02	0.05
STD	0.23	0.25	0.20	0.20	0.18	0.22	0.25	0.15	0.16	0.14
AC1	0.47	0.56	0.48	0.49	0.27	0.67	0.73	0.59	0.60	0.10
AC2	0.25	0.35	0.29	0.30	-0.06	0.35	0.43	0.17	0.23	-0.06
AC3	0.04	0.12	0.08	0.09	0.05	0.13	0.22	0.03	0.11	-0.02
AC4	-0.04	0.07	0.05	0.05	0.04	-0.02	0.12	0.00	0.07	0.12
AC8	-0.09	0.01	-0.06	-0.08	0.12	-0.17	-0.04	-0.18	-0.16	0.03

Panel B: Forecast Encompassing Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.02	0.02	0.02	0.02	-0.03	0.05	0.05	0.03	0.02	-0.01
STD	0.11	0.12	0.10	0.10	0.07	0.11	0.12	0.07	0.07	0.05
AC1	0.45	0.48	0.44	0.44	0.29	0.70	0.74	0.57	0.60	0.11
AC2	0.26	0.26	0.26	0.27	-0.10	0.36	0.40	0.13	0.20	-0.08
AC3	0.04	0.04	0.06	0.06	0.09	0.15	0.18	-0.02	0.06	-0.07
AC4	-0.02	0.03	0.04	0.04	0.03	-0.01	0.09	-0.01	0.03	0.07
AC8	-0.06	0.05	-0.07	-0.09	0.19	-0.20	-0.11	-0.17	-0.15	0.08

Note: sample mean, standard deviation (STD), and autocorrelation coefficients up to order 4, and order 8 (AC1, AC2, AC3, AC4, AC8) of d_t for the unemployment rate using the QPS loss function. The top panel refers to d_t as defined for the equal predictive ability test in (3), and the bottom panel refers to d_t as defined for the forecast encompassing test in (6).

Table 13: Summary Statistics of d_t for GDP growth, QPS.

Panel A: Equal Predictive Ability

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.05	0.03	0.08	0.10	0.04	-0.06	-0.07	0.02	0.05	0.00
STD	0.28	0.26	0.31	0.31	0.17	0.23	0.21	0.26	0.28	0.05
AC1	0.50	0.48	0.47	0.47	0.17	0.76	0.75	0.72	0.73	0.16
AC2	0.05	0.02	0.09	0.10	-0.04	0.46	0.45	0.37	0.36	-0.29
AC3	-0.05	-0.07	0.02	0.01	0.05	0.28	0.32	0.20	0.16	-0.22
AC4	-0.11	-0.08	0.04	0.01	-0.05	0.18	0.26	0.07	0.04	-0.01
AC8	0.08	0.09	-0.02	-0.05	0.03	0.08	0.19	-0.09	-0.15	0.08

Panel B: Forecast Encompassing Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.05	0.05	0.05	0.05	0.00	0.09	0.08	0.07	0.07	0.00
STD	0.13	0.12	0.13	0.13	0.07	0.12	0.11	0.10	0.10	0.03
AC1	0.53	0.49	0.47	0.46	0.10	0.77	0.75	0.74	0.75	0.13
AC2	0.08	0.04	0.04	0.04	-0.12	0.49	0.46	0.45	0.45	-0.24
AC3	-0.02	-0.04	-0.06	-0.07	0.03	0.31	0.32	0.30	0.30	-0.15
AC4	-0.08	-0.06	-0.09	-0.09	-0.10	0.21	0.26	0.20	0.20	-0.04
AC8	0.08	0.11	0.11	0.11	0.06	0.10	0.20	0.10	0.10	0.06

Note: sample mean, standard deviation (STD), and autocorrelation coefficients up to order 4, and order 8 (AC1, AC2, AC3, AC4, AC8) of d_t for GDP growth using the QPS loss function. The top panel refers to d_t as defined for the equal predictive ability test in (3), and the bottom panel refers to d_t as defined for the forecast encompassing test in (6).

Table 14: Summary Statistics of d_t for HICP, QPS.

Panel A: Equal Predictive Ability

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	-0.02	-0.05	-0.03	-0.03	0.00	0.00	-0.02	0.03	0.04	0.00
STD	0.25	0.22	0.24	0.24	0.09	0.25	0.19	0.23	0.23	0.06
AC1	0.30	0.23	0.30	0.33	-0.14	0.38	0.30	0.44	0.49	-0.25
AC2	0.23	0.22	0.20	0.22	-0.20	0.29	0.15	0.31	0.37	-0.12
AC3	0.12	0.14	0.10	0.11	0.03	0.15	0.03	0.10	0.17	0.13
AC4	-0.01	-0.03	-0.04	0.00	-0.04	-0.09	-0.22	-0.15	-0.09	0.00
AC8	-0.16	-0.13	-0.21	-0.21	0.01	-0.07	-0.06	-0.33	-0.34	-0.11

Panel B: Forecast Encompassing Statistic

	1 year ahead					2 years ahead				
	Uni	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
Mean	0.08	0.07	0.07	0.07	0.00	0.07	0.05	0.04	0.04	0.00
STD	0.13	0.11	0.11	0.11	0.04	0.13	0.10	0.10	0.11	0.03
AC1	0.33	0.28	0.28	0.31	-0.17	0.39	0.31	0.35	0.41	-0.26
AC2	0.24	0.24	0.21	0.22	-0.17	0.29	0.12	0.21	0.27	-0.12
AC3	0.12	0.15	0.10	0.11	0.02	0.16	0.03	0.04	0.10	0.16
AC4	-0.02	-0.01	0.03	0.05	-0.07	-0.09	-0.22	-0.24	-0.17	0.00
AC8	-0.18	-0.12	-0.23	-0.23	0.00	-0.07	-0.08	-0.10	-0.15	-0.11

Note: sample mean, standard deviation (STD), and autocorrelation coefficients up to order 4, and order 8 (AC1, AC2, AC3, AC4, AC8) of d_t for HICP using the QPS loss function. The top panel refers to d_t as defined for the equal predictive ability test in (3), and the bottom panel refers to d_t as defined for the forecast encompassing test in (6).

Table 15: Estimated forecast combination weights, QPS.

Full sample Q1.2000 - Q4.2019, $T = 80$

	1 year ahead					**2 years ahead**				
Variable	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	0.14	0.10	0.21	0.19	0.00	0.40	0.30	0.47	0.38	0.00
GDP	0.34	0.37	0.26	0.25	0.00	0.72	0.89	0.44	0.37	0.50
HICP	0.57	0.74	0.64	0.62	0.35	0.51	0.63	0.35	0.31	0.29

Subsample Q1.2000 - Q4.2009, $T = 40$.

	1 year ahead					**2 years ahead**				
Variable	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	0.12	0.09	0.15	0.12	0.00	0.39	0.30	0.42	0.32	0.00
GDP	0.45	0.52	0.38	0.35	0.08	0.93	1.00	0.64	0.51	0.29
HICP	0.59	0.81	0.54	0.51	0.79	0.46	0.66	0.36	0.31	0.93

Subsample Q1.2010 - Q4.2019, $T = 40$.

	1 year ahead					**2 years ahead**				
Variable	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**	**Uni**	**UG**	**GRW**	**GRWD**	**Naive**
UN	0.15	0.12	0.33	0.34	0.00	0.42	0.30	0.58	0.53	0.00
GDP	0.20	0.20	0.15	0.14	0.00	0.46	0.48	0.24	0.22	0.95
HICP	0.54	0.64	0.81	0.77	0.00	0.57	0.58	0.34	0.33	0.00

Note: LS estimate of λ in (5) using the QPS loss function. The top panel refers to the full sample (Q1.2000 - Q4.2019, $T = 80$), the middle panel to the first half-sample (Q1.2000 - Q4.2009, $T = 40$) and the bottom panel to the second-half sample (Q1.2010 - Q4.2019, $T = 40$).

Table 16: Forecast evaluation tests. Full sample Q1.2000 - Q4.2019, $T = 80$, QPS.

Panel A: Equal Predictive Ability Test

		1 year ahead					2 years ahead				
Variable	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	3.61*	3.81*	1.49	1.58	4.22*	0.71	1.37	0.12	0.57	2.97*
	WPE	3.69*	3.48*	1.35	1.43	4.13*	0.62	1.18	0.11	0.51	3.46*
GDP	WCE	1.16	0.83	1.76	1.98*	1.91*	-1.10	-1.57	0.43	0.96	0.02
	WPE	0.92	0.68	1.40	1.57	1.72	-0.90	-1.30	0.37	0.83	0.02
HICP	WCE	-0.56	-1.48	-0.89	-0.71	0.40	-0.05	-0.76	0.91	1.11	0.31
	WPE	-0.48	-1.41	-0.77	-0.61	0.37	-0.04	-0.77	0.92	1.06	0.27

Panel B: Forecast Encompassing Test

		1 year ahead					2 years ahead				
Variable	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	1.37	1.11	1.17	1.10	-2.68	2.78*	2.18*	2.52*	2.08*	-1.74
	WPE	1.36	1.03	1.07	1.02	-2.40	2.41*	1.84	2.30*	1.87*	-1.89
GDP	WCE	2.44*	2.45*	2.50*	2.52*	-0.04	3.46*	3.49*	3.29*	3.26*	1.47
	WPE	1.96*	2.02*	2.01*	2.03*	-0.03	2.84*	2.87*	2.83*	2.79*	1.54
HICP	WCE	4.34*	4.17*	4.22*	4.02*	0.95	3.22*	3.66*	2.74*	2.31*	0.43
	WPE	3.73*	4.04*	3.73*	3.46*	0.92	2.71*	3.98*	2.62*	2.12*	0.37

Note: Equal predictive ability test statistics and the forecast encompassing test statistics for one-year and two-year ahead ECB SPF density forecasts against the uniform, unconditional Gaussian, the Gaussian random walk, the Gaussian random walk with drift, and the naive benchmark forecasts on the full sample Q1.2000 - Q4.2019 ($T = 80$) using the QPS loss. A negative equal predictive ability test statistic sign implies that benchmark performs better than the ECB SPF, and a negative value for the forecast encompassing test indicates that the estimated unrestricted weight on the benchmark is negative. Long run variances are estimated using WCE with Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$. One-sided significance at the 5% level is indicated with [grey shading] using standard asymptotics and with * using fixed-smoothing asymptotics.

Table 17: Forecast evaluation tests. Subsample Q1.2000 - Q4.2009, $T = 40$, QPS.

Panel A: Equal Predictive Ability Test

		1 year ahead					2 years ahead				
	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	2.13*	2.78*	1.38	1.51	2.15*	0.53	1.07	0.27	0.65	2.31*
	WPE	1.70	2.31*	1.12	1.21	1.93	0.42	0.92	0.22	0.53	2.15*
GDP	WCE	0.26	-0.08	0.70	0.97	1.36	-2.39	-3.95	-1.15	-0.10	0.63
	WPE	0.22	-0.07	0.61	0.85	1.35	-2.29	-3.96	-1.10	-0.10	0.62
HICP	WCE	-0.50	-1.51	-0.26	-0.07	-0.54	0.21	-0.79	0.84	1.20	-0.49
	WPE	-0.45	-1.37	-0.26	-0.07	-0.49	0.18	-0.62	0.62	0.90	-0.48

Panel B: Forecast Encompassing Test

		1 year ahead					2 years ahead				
	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	0.71	0.62	0.59	0.52	-0.89	1.74	1.68	1.55	1.28	-1.23
	WPE	0.58	0.52	0.48	0.42	-0.86	1.38	1.41	1.27	1.04	-1.15
GDP	WCE	2.63*	2.70*	2.60*	2.63*	0.35	5.04*	5.60*	5.12*	4.97*	0.75
	WPE	2.31*	2.44*	2.27*	2.30*	0.36	4.79*	5.56*	4.84*	4.64*	0.77
HICP	WCE	3.00*	3.54*	3.28*	3.20*	1.58	2.18*	3.45*	2.81*	2.28*	1.00
	WPE	2.69*	3.22*	3.14*	3.07*	1.48	1.82	2.57*	2.19*	1.80	1.01

Note: Equal predictive ability test statistics and the forecast encompassing test statistics for one-year and two-year ahead ECB SPF density forecasts against the uniform, unconditional Gaussian, the Gaussian random walk, the Gaussian random walk with drift, and the naive benchmark forecasts on the subsample Q1.2000 - Q4.2009 ($T = 40$) using the QPS loss. A negative equal predictive ability test statistic sign implies that benchmark performs better than the ECB SPF, and a negative value for the forecast encompassing test indicates that the estimated unrestricted weight on the benchmark is negative. Long run variances are estimated using WCE with Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$. One-sided significance at the 5% level is indicated with [grey shading] using standard asymptotics and with * using fixed-smoothing asymptotics.

Table 18: Forecast evaluation tests. Subsample Q1.2010 - Q4.2019, $T = 40$, QPS.

Panel A: Equal Predictive Ability Test

		1 year ahead					2 years ahead				
Variable	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	4.76*	3.31*	0.72	0.72	5.29*	0.48	0.94	-0.30	-0.14	2.40*
	WPE	4.28*	2.78*	0.65	0.65	5.46*	0.40	0.80	-0.28	-0.12	2.18*
GDP	WCE	1.72	1.58	2.01*	2.04*	1.91	0.14	0.07	1.16	1.30	-0.61
	WPE	1.65	1.49	2.10*	2.14*	1.53	0.13	0.06	1.20	1.35	-0.58
HICP	WCE	-0.24	-0.48	-0.93	-0.84	1.80	-0.29	-0.25	0.46	0.55	1.40
	WPE	-0.19	-0.41	-0.75	-0.69	2.03*	-0.24	-0.22	0.37	0.44	1.54

Panel B: Forecast Encompassing Test

		1 year ahead					2 years ahead				
Variable	LRV	Uni	UG	GRW	GRWD	Naive	Uni	UG	GRW	GRWD	Naive
UN	WCE	2.37*	1.62	1.67	1.73	-4.18	2.84*	1.64	2.57*	2.36*	-1.41
	WPE	2.23*	1.43	1.53	1.59	-4.32	2.46*	1.44	2.34*	2.05*	-1.31
GDP	WCE	1.08	1.04	1.15	1.16	-0.91	1.46	1.39	1.35	1.33	1.33
	WPE	1.02	0.97	1.14	1.15	-0.73	1.37	1.30	1.24	1.22	1.26
HICP	WCE	2.89*	2.42*	2.60*	2.53*	-0.66	2.41*	2.01*	1.23	1.23	-0.82
	WPE	2.27*	1.97*	2.11*	2.08*	-0.72	2.06*	1.79	1.00	1.00	-0.91

Note: Equal predictive ability test statistics and the forecast encompassing test statistics for one-year and two-year ahead ECB SPF density forecasts against the uniform, unconditional Gaussian, the Gaussian random walk, the Gaussian random walk with drift, and the naive benchmark forecasts on the subsample Q1.2010 - Q4.2019 ($T = 40$) using the QPS loss. A negative equal predictive ability test statistic sign implies that benchmark performs better than the ECB SPF, and a negative value for the forecast encompassing test indicates that the estimated unrestricted weight on the benchmark is negative. Long run variances are estimated using WCE with Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$. One-sided significance at the 5% level is indicated with [grey shading] using standard asymptotics and with * using fixed-smoothing asymptotics.

References

Andrews, Donald W. K. (1984) 'Non-strong mixing autoregressive processes.' *Journal of Applied Probability* 21(4), 930–934

Boero, Gianna, Jeremy Smith, and Kenneth F Wallis (2008) 'Uncertainty and disagreement in economic prediction: the Bank of England Survey of External Forecasters.' *The Economic Journal* 118(530), 1107–1127

Bowles, Carlos, Roberta Friz, Veronique Genre, Geoff Kenny, Aidan Meyler, and Tuomas Rautanen (2007) 'The ECB Survey of Professional Forecasters (SPF)-a review after eight years' experience.' *ECB Occasional Paper*

Brier, Glenn W (1950) 'Verification of forecasts expressed in terms of probability.' *Monthly Weather Review* 78(1), 1–3

Choi, Hwan-sik, and Nicholas M Kiefer (2010) 'Improving robust model selection tests for dynamic models.' *The Econometrics Journal* 13(2), 177–204

Clark, Todd E (1999) 'Finite-sample properties of tests for equal forecast accuracy.' *Journal of Forecasting* 18(7), 489–504

Clements, Michael P (2014) 'Forecast uncertainty—ex ante and ex post: US inflation and output growth.' *Journal of Business and Economic Statistics* 32(2), 206–216

Clements, Michael P, and David I Harvey (2010) 'Forecast encompassing tests and probability forecasts.' *Journal of Applied Econometrics* 25(6), 1028–1062

Coroneo, Laura, and Fabrizio Iacone (2020) 'Comparing predictive accuracy in small samples using fixed-smoothing asymptotics.' *Journal of Applied Econometrics* 35(4), 391–409

_ (2021) 'Testing for equal predictive accuracy with strong dependence.' Technical Report, Department of Economics, University of York

Coroneo, Laura, Fabrizio Iacone, Alessia Paccagnini, and Paulo Santos Monteiro (2022) 'Testing the predictive accuracy of covid-19 forecasts.' *International Journal of Forecasting*

D'Amico, Stefania, Athanasios Orphanides et al. (2008) 'Uncertainty and disagreement in economic forecasting.' *Federal Reserve Board Finance and Economics Discussion Series*

Dawid, A Philip (1984) 'Present position and potential developments: Some personal views statistical theory the prequential approach.' *Journal of the Royal Statistical Society: Series A (General)* 147(2), 278–290

de Vincent-Humphreys, Rupert, Ivelina Dimitrova, Elisabeth Falck, and Lukas Henkel (2019) 'Twenty years of the ECB survey of professional forecasters.' *Economic Bulletin Articles*

Diebold, Francis X, and Robert S Mariano (1995) ‘Comparing predictive accuracy.’ *Journal of Business and Economic Statistics* 13(3), 253–262

Diebold, Francis X, Anthony S Tay, and Kenneth F Wallis (1999) ‘Evaluating density forecasts of inflation: the survey of professional forecasters.’ *in R. Engle and H. White (eds.), A Festschrift in Honour of Clive WJ Granger* pp. 76–90

Epstein, Edward S (1969) ‘A scoring system for probability forecasts of ranked categories.’ *Journal of Applied Meteorology* 8(6), 985–987

Fair, Ray C (1980) ‘Estimating the expected predictive accuracy of econometric models.’ *International Economic Review* 21(2), 355–378

Garcia, Juan A (2003) ‘An introduction to the ECB’s Survey of Professional Forecasters.’ *ECB Occasional Paper*

Giacomini, Raffaella, and Halbert White (2006) ‘Tests of conditional predictive ability.’ *Econometrica* 74(6), 1545–1578

Gneiting, Tilmann, and Adrian E Raftery (2007) ‘Strictly proper scoring rules, prediction, and estimation.’ *Journal of the American Statistical Association* 102(477), 359–378

Harvey, David I., Stephen J. Leybourne, and Emily J. Whitehouse (2017) ‘Forecast evaluation tests and negative long-run variance estimates in small samples.’ *International Journal of Forecasting* 33(4), 833 – 847

Harvey, David I, Stephen J Leybourne, and Paul Newbold (1998) ‘Tests for forecast encompassing.’ *Journal of Business and Economic Statistics* 16(2), 254–259

Hualde, Javier, and Fabrizio Iacone (2017) ‘Fixed bandwidth asymptotics for the studentized mean of fractionally integrated processes.’ *Economics Letters* 150, 39–43

Kiefer, Nicholas M, and Timothy J Vogelsang (2002a) ‘Heteroskedasticity–autocorrelation robust standard errors using the bartlett kernel without truncation.’ *Econometrica* 70(5), 2093–2095

_ (2002b) ‘Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size.’ *Econometric Theory* 18(6), 1350–1366

_ (2005) ‘A new asymptotic theory for heteroskedasticity-autocorrelation robust tests.’ *Econometric Theory* 21(6), 1130–1164

Lazarus, Eben, Daniel J. Lewis, James H. Stock, and Mark W. Watson (2018) ‘HAR inference: Recommendations for practice.’ *Journal of Business and Economic Statistics* 36(4), 541–559

Li, Jia, and Andrew J. Patton (2018) ‘Asymptotic inference about predictive accuracy using high frequency data.’ *Journal of Econometrics* 203(2), 223 – 240

Manzan, Sebastiano (2021) 'Are professional forecasters bayesian?' *Journal of Economic Dynamics and Control* 123, 104045

Mitchell, James, and Stephen G Hall (2005) 'Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and niesr fan charts of inflation.' *Oxford Bulletin of Economics and Statistics* 67, 995–1033

Neave, Henry R (1970) 'An improved formula for the asymptotic variance of spectrum estimates.' *The Annals of Mathematical Statistics* 41(1), 70–77

Newey, Whitney K., and Kenneth D. West (1994) 'Automatic lag selection in covariance matrix estimation.' *The Review of Economic Studies* 61(4), 631–653

Phillips, Peter C. B., and Victor Solo (1992) 'Asymptotics for linear processes.' *Annals of Statististics* 20(2), 971–1001

Sun, Yixiao (2013) 'A heteroskedasticity and autocorrelation robust F test using an orthonormal series variance estimator.' *The Econometrics Journal* 16(1), 1–26

Sun, Yixiao (2014) 'Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference.' *Journal of Econometrics* 178, 659 – 677

Tay, Anthony S, and Kenneth F Wallis (2000) 'Density forecasting: a survey.' *Journal of Forecasting* 19(4), 235–254

Wooldridge, Jeffrey M., and Halbert White (1988) 'Some invariance principles and central limit theorems for dependent heterogeneous processes.' *Econometric Theory* 4(2), 210–230