ICMA Centre

Henley Business School
UNIVERSITY OF READING

Discussion Paper

Are Macroeconomic Density Forecasts Informative?

April 2016

Michael P Clements
ICMA Centre, Henley Business School, University of Reading

Discussion Paper Number: ICM-2016-02
ICM-2016-02

The aim of this discussion paper series is to disseminate new research of academic distinction. Papers are preliminary drafts, circulated to stimulate discussion and critical comment. Henley Business School is triple accredited and home to over 100 academic faculty, who undertake research in a wide range of fields from ethics and finance to international business and marketing.

admin@icmacentre.ac.uk

www.icmacentre.ac.uk

© Clements, April 2016

Are Macroeconomic Density Forecasts Informative?

Michael P. Clements
ICMA Centre,
Henley Business School,
University of Reading,
Reading RG6 6BA

April 7, 2016

Abstract

We consider whether survey density forecasts (such as the inflation and output growth histograms of the US Survey of Professional Forecasters) are superior to unconditional density forecasts. The unconditional forecasts assume that the average level of uncertainty experienced in the past will prevail in the future, whereas the SPF projections ought to be adapted to current conditions and the outlook at each forecast origin. The SPF forecasts might be expected to outperform the unconditional densities at the shortest horizons, but this does not transpire to be the case, for either the aggregate or individual respondents' forecasts.

Keywords: probability distribution forecasts, aggregation, Kullback-Leibler information criterion.
JEL classification: C53.

1 Introduction

There has been much interest in survey forecasts in recent years, driven in part by the opportunities they offer to test theories of expectations-generation (see, e.g., Pesaran and Weale (2006), Coibion and Gorodnichenko (2012, 2015) and Andrade and Le Bihan (2013), amongst many others) and the opportunities they offer for improved forecasting accuracy, either as direct forecasts themselves (see, e.g., Ang, Bekaert and Wei (2007), Clements (2015)) or as an adjunct to other forecasting models (see, e.g., Wright (2013)). As well as collecting point predictions, some surveys elicit respondents' subjective probability distributions, in the form of histograms, offering the promise of 'direct' measures of forecast uncertainty (see, e.g., Giordani and Söderlind (2003)), Rich and Tracy (2010) and Clements (2014a)) to supplement less theoretically satisfactory measures such as forecaster disagreement, as given by some measure of the cross-sectional dispersion of the point predictions (Zarnowitz and Lambros (1987)).

Where both are available, point predictions and histograms have been compared for variables such as real GDP growth and inflation, in terms of what they have to say about central tendencies, or the most likely outcomes. Naturally, the two generally match, but some systematic differences have been documented and explored in the literature (Engelberg, Manski and Williams (2009), Clements (2010, 2014b)). There is also a literature on evaluating density forecasts (see the next section), but it remains an open question as to whether survey respondents are able to form probability assessments about the future values of key macro-variables (such as output growth and inflation) which are more accurate than 'unconditional' benchmark densities. Little is known about how the information content of the subjective probability assessments varies with the forecast horizon: one might surmise that survey forecasters would outperform the benchmarks at short horizons, but any advantage would dissipate as the horizon increases, but (to the best of my knowledge) there is little evidence as to whether this is this case. Also of interest is the performance of the aggregate distributions (i.e., averaging across individual respondents), and of the individual forecasters' assessments, i.e., we consider the role combination (or aggregation) plays.

We consider the US Survey of Professional Forecasters (SPF). We regard the SPF densities as adding value if they are more accurate than the benchmarks, at least at short horizons. Describing the benchmark forecasts as 'unconditional' is a misnomer in the sense that the forecast densities are centred on the median point predictions (of the SPF respondents), and hence draw on forecast origin information, but serves to underline that the dispersion of the forecast density is based on the historical variance of the forecast errors. One would expect the SPF forecasts to outperform the unconditional densities at the shortest horizons (just as a conditional mean or variance forecast would outperform an unconditional mean or variance forecast) but that the relative improvements would diminish as the forecast horizon lengthens, as the role of current conditions in predicting future developments lessens. Our results suggest the opposite: the aggregate and individual histograms are rejected in favour of the unconditional histograms at the shorter horizons, reflecting the under-confidence of the survey respondents at within-year horizons, as documented by Clements (2014a). That is, the survey respondents tend to over-estimate the degree of uncertainty they face when forecasting at the shorter horizons. We show that this is true at the aggregate level and also holds for individuals. Moreover, at least at the level of the aggregate histograms the mis-specification is found to be systematic, and 'future' densities can be successfully corrected based on the performance of an in-sample or training set.

Our empirical investigation considers both whether the SPF densities and the benchmark densities are correctly specified, and provides a comparison of the two, not requiring that either set closely approximates

the truth. We are careful to check that our findings are not dependent on the way the histogram forecasts are elicited, and how this has changed over time, or on any mismatch between point predictions and histogram means (e.g., Engelberg *et al.* (2009)). We regard the value of survey macro-forecasts as established in the case of first-moment prediction, but as unproven in terms of the probability assessments implied by the histogram forecasts.

The remainder of the paper is as follows. Section 2 reviews the literature on forecast density evaluation, and on comparisons between forecast densities. Section 3 describes the survey data. Section 4 describes the construction of the benchmark densities, used to gauge the value of conditioning on forecast origin information. The results are given in section 5. Section 6 assesses the robustness of our main findings to the assumptions which have been made. Section 7 considers a simple correction based on the past performance of the SPF densities, which delivers more accurate densities over the out-of-sample period. Section 8 concludes.

2 The Evaluation of Survey Density Forecasts, and Comparisons to Benchmarks

A popular way of evaluating survey density forecasts is based on the probability integral transform, dating back at least to Rosenblatt (1952), with recent contributions by Shephard (1994), Kim, Shephard and Chib (1998) and Diebold, Gunther and Tay (1998). Diebold *et al.* (1998) and Granger and Pesaran (2000) show that a density forecast that coincides with the 'true' forecast density will be optimal in terms of minimizing expected loss irrespective of the form of the (generally unknown) user's loss function. In some applications only a portion of the forecast density may be relevant, as for example in financial risk management, where the tail quartile of the expected distribution of returns plays a prominent role (in Value at Risk calculations), or in macro inflation forecasting, where the focus is on the probability that inflation will exceed a target value. Tools have been developed for the study of quartiles and for events derived from density forecasts (see, e.g., Engle and Manganelli (2004), Clements (2004)) but our focus will be on the whole density.

There are two parts to our empirical investigation. In the first part, we assess how well the SPF forecast densities approximate the true, unknown densities, using the probability integral transform, and in particular, the extension due to Berkowitz (2001). In the second, we compare the SPF densities to unconditional benchmark densities. Hence the SPF densities may be mis-specified - of interest is whether they are nonetheless superior to the benchmarks. These comparisons are motivated by Lee, Bao and Saltoglu (2007) and Mitchell and Hall (2005), and the recognition that the Kullback-Leibler Information Criterion (Kullback and Leibler (1951), KLIC) measure of the divergence between a forecast density and the true density can be adapted to compare two or more densities, without making the assumption that any of the densities is correctly-specified. The KLIC will be used to compare the SPF densities against the benchmark densities in the form of a Diebold and Mariano (1995) test of equal predictive ability. The benchmark densities are such that the rejection of the null of equal density accuracy in favour of the SPF densities would imply that the conditioning on the forecast origin information implicit in the survey forecasts results in the superior accuracy.[1] One would expect the forecast origin information would become less valuable for determining

[1] In a similar vein, for point forecasts Diebold and Kilian (2001) suggest measuring predictability by comparing the expected loss of a short-horizon forecast to a long-horizon forecast. We use the unconditional density as the 'long-horizon' forecast, against which the SPF density forecasts are compared as the 'short-horizon' forecasts. Note that our benchmark forecasts are

the scale of the forecast density as the forecast horizon increases. By considering density forecasts with horizons from approximately one quarter to one year ahead, it might be possible to determine the horizon at which the survey forecasts become 'uninformative', in the sense that, evaluated by KLIC, the conditioning information yields no improvement in accuracy.[2]

2.1 Density evaluation

Suppose we have a series of 1-step forecast densities for the value of a random variable $\{Y_t\}$, denoted by $p_{Y,t-1}(y)$, where $t = 1, \ldots, n$. The probability integral transforms (p.i.t.'s) of the realizations of the variable with respect to the forecast densities are given by:

$$z_t = \int_{-\infty}^{y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}\left(y_t\right) \tag{1}$$

for $t = 1, \ldots, n$, where $P_{Y,t-1}\left(y_t\right)$ is the forecast probability of Y_t not exceeding the realized value y_t. In terms of the random variables $\{Y_t\}$, rather than their realized values $\{y_t\}$, we obtain random variables denoted by $\{Z_t\}$:

$$Z_t = \int_{-\infty}^{Y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}\left(Y_t\right).$$

When the forecast density equals the true density, $f_{Y,t-1}\left(y\right)$, it follows by a simple change-of-variables argument that $Z_t \sim U\left(0,1\right)$, where $U\left(0,1\right)$ is the uniform distribution over $(0,1)$. Even though the actual conditional densities may be changing over time, provided the forecast densities match the actual densities at each t, then $Z_t \sim U\left(0,1\right)$ for each t, and the Z_t are independently distributed of each other, such that the realized time series $\{z_t\}_{t=1}^n$ is an iid sample from a $U\left(0,1\right)$ distribution.

This suggests we can evaluate whether the conditional forecast densities match the true conditional densities by testing whether $\{z_t\}_{t=1}^n$ is iid$U\left(0,1\right)$. Berkowitz (2001) suggested taking the inverse normal CDF transformation of the $\{z_t\}_{t=1}^n$ series, to give, say, $\{z_t^*\}_{t=1}^n$, on the grounds that more powerful tools can be applied to testing the null that the $\{z_t^*\}_{t=1}^n$ are iid$N(0,1)$ (for $h = 1$) compared to one of iid uniformity of the original $\{z_t\}_{t=1}^n$ series. He proposes a one-degree of freedom test of independence against a first-order autoregressive structure, as well as a three-degree of freedom test of zero-mean, unit variance and independence. In each case the maintained assumption is that of normality, so that standard likelihood ratio tests are constructed using the gaussian likelihoods. [3]

The Berkowitz (2001) three degree of freedom test is given by:

$$LR_B = -2(L\left(0,1,0\right) - L\left(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}\right)) \tag{2}$$

where $L\left(0,1,0\right) = \sum_{t=1}^n \left[\ln \phi\left(z_{t|t-1}^*\right)\right]$ is the value of the Gaussian log-likelihood for an independently

unconditional in terms of the dispersion about the mean, but the mean is conditioned on forecast origin information.

[2] The way in which this discussion is framed assumes that the one-quarter ahead survey forecasts will outperform, and that 1-year forecasts will be no better, but this does not turn out to be the case.

[3] The assumption of normality of $\{z_t^*\}_{t=1}^n$ is also amenable to testing, for example, using the Shenton and Bowman (1977) two-degree of freedom asymptotic chi-squared test or the test recommended by Doornik and Hansen (2008).

distributed standard normal random variable ($\phi(.)$ is the $N(0,1)$ pdf), and:

$$L\left(\hat{\mu},\hat{\sigma}^2,\hat{\rho}\right)=\sum_{t=1}^{n}\left[\ln\left(\phi\left(\left(z_{t|t-1}^{*}-\hat{\mu}-\hat{\rho}z_{t|t-1}^{*}\right)/\hat{\sigma}\right)/\hat{\sigma}\right)\right]$$

is the maximized log-likelihood for an AR(1) with Gaussian errors ($\hat{}$'s on parameters denote maximum likelihood estimates). As noted by Lee *et al.* (2007), the assumptions of a first-order process, and that it is Gaussian, can be generalized: they allow instead a higher-order autoregression, with iid disturbances that follow a semi-non-parametric density function. However, for quarterly macro data relatively short-sample sizes perhaps warrant the simpler assumptions.

2.2 Density comparison

Lee *et al.* (2007) show that the Berkowitz test can be interpreted as a particular form of KLIC-based evaluation of a forecast density compared to the true density, and that the KLIC can also be used to compare (two or more) mis-specified identities.

Firstly, comparing a forecast density to the true density. The KLIC is defined as:

$$KLIC_{t|t-h}=E\left[\ln\left(f_{Y,t-h}\left(y_t\right)\right)-\ln\left(p_{Y,t-h}\left(y_t\right)\right)\right]$$

where the expectation is with respect to the true density, so that:

$$KLIC_{t|t-h}=\int f_{Y,t-h}\left(y_t\right)\ln\left\{\frac{\left(f_{Y,t-h}\left(y_t\right)\right)}{\left(p_{Y,t-h}\left(y_t\right)\right)}\right\}\partial y_t.$$

Berkowitz (2001, Proposition 2, p.467) shows that $\ln\left(f_{Y,t-h}\left(y_t\right)\right)-\ln\left(p_{Y,t-h}\left(y_t\right)\right)=\ln q_{Z^*,t-h}\left(z_t^*\right)-\ln\phi\left(z_t^*\right)$, where $q_{Z^*,t-h}$ is the density of z_t^* and ϕ is the standard normal. The KLIC is estimated as the sample average of $d_{t|t-h}\equiv\ln q_{Z^*,t-h}\left(z_t^*\right)-\ln\phi\left(z_t^*\right)$ (over t, for a given h), and if we allow that z_t^* is a Gaussian AR(1), we obtain:

$$\begin{aligned}\overline{KLIC}_h &= \frac{1}{n}\sum_{t=1}^{n}d_{t|t-h}=\frac{1}{n}\sum_{t=1}^{n}\left[\ln\left(\phi\left(\left(z_{t|t-h}^{*}-\hat{\mu}-\hat{\rho}z_{t|t-h}^{*}\right)/\hat{\sigma}\right)/\hat{\sigma}\right)\right]-\frac{1}{n}\sum_{t=1}^{n}\left[\ln\phi\left(z_{t|t-h}^{*}\right)\right]\\ &= \left(2n\right)^{-1}LR_B,\end{aligned}\tag{3}$$

where LR_B is given in (2).Hence the KLIC and Berkowitz test are directly related. The assumption that the $z_{t|t-h}^{*}$ are independent in our framework for optimal density forecasts is valid for our forecasts.

The KLIC can also be used as the loss function in a *comparison* of two density forecasts, using the approach to testing equal predictive accuracy of Diebold and Mariano (1995). Letting the loss differential between the KLICs of the two densities be $d_{t|t-h}$, then

$$\begin{aligned}d_{t|t-h} &= \left(\ln\left(f_{Y,t-h}\left(y_t\right)\right)-\ln\left(p_{Y,t-h}^{2}\left(y_t\right)\right)\right)-\left(\ln\left(f_{Y,t-h}\left(y_t\right)\right)-\ln\left(p_{Y,t-h}^{1}\left(y_t\right)\right)\right)\\ &= \ln\left(p_{Y,t-h}^{1}\left(y_t\right)\right)-\ln\left(p_{Y,t-h}^{2}\left(y_t\right)\right),\end{aligned}$$

where $p^1()$ and $p^2()$ denote rivals sets of forecast densities.

The average loss differential is:

$$\overline{d}_h = \frac{1}{n}\sum_{t=1}^{n} d_{t|t-h}$$

with limiting distribution:

$$\sqrt{n}\left(\overline{d}_h - E\left(d_{t|t-h}\right)\right) \xrightarrow{d} N\left(0, \xi^2\right) \tag{4}$$

where ξ^2 is the limiting variance.

We will use (4) to compare the SPF forecasts (say, $p^1_{Y,t-h}$) against the benchmark forecasts ($p^2_{Y,t-h}$). The SPF forecast densities are *given* and so the issue of parameter uncertainty does not arise.[4] In which case, under the null of equal accuracy, $E\left(d_{t|t-h}\right) = 0$, and $\xi^2 = \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j$, with $\gamma_j = E\left(d_{t|t-h}d_{t-j|t-j-h}\right)$, and (4) is simply:

$$\frac{\sqrt{n}\overline{d}_h}{\xi} \xrightarrow{d} N\left(0,1\right). \tag{5}$$

Our forecasts are non-overlapping, in the sense that, say, the realization for the previous year's Q1 survey forecast is known before this year's Q1 survey forecast is made. Hence γ_i can be assumed to be zero for $i > 0$.

3 Data description and distributional assumptions

We use the quarterly US Survey of Professional Forecasters (SPF) respondents' probability distributions for inflation and output growth. The SPF began as the NBER-ASA survey in 1968:4 and runs to the present day: see Croushore (1993). It has been extensively used for academic research into the nature of expectations formation: see https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/academic-bibliography.

We use the 130 quarterly surveys from 1981:Q3 to 2013:Q4, inclusive.[5] Over this period, the survey provides respondents' histograms for output growth and inflation, in terms of the percentage change in the survey year relative to the previous year. Hence we have a sequence of (approximately) year-ahead histogram forecasts from the first quarter surveys of each year, down to a sequence of 1-quarter ahead forecasts from the Q4 surveys of each year. The evaluation of the histograms require the actual values, i.e., the realizations of the random variables being forecast. We use early-vintage actual values, specifically, the 'advance estimates' of output growth and inflation released in Q1 of the year following the year being forecast. [6]

The survey also provides fixed-event forecasts of the current calendar-year inflation and output growth rates, as well as rolling-event forecasts of the quarterly values (at annual rates) of these variables at horizons up to 4-quarters ahead. These are used to construct the benchmark densities, as described in section 4.

The reported histograms provide an incomplete estimate of the densities. We fit normal distributions to

[4] Much of the literature supposes the densities are based on models defined up to an unknown vector, $\boldsymbol{\theta}$, so that the test may be affected by the need to estimate the unknown parameter vector.

[5] The data were downloaded in December 2015, from http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/

[6] These are taken from the quarterly Real Time Data Set for Macroeconomists (RTDSM) maintained by the Federal Reserve Bank of Philadelphia: see Croushore and Stark (2001). This consists of a data set for each quarter that contains only those data that would have been available at a given reference date: subsequent revisions, base-year and other definitional changes that occurred after the reference date are omitted.

the histograms (see, e.g., Giordani and Söderlind (2003, p. 1044) and Boero, Smith and Wallis (2015)) when there are 3 or more bins with non-zero probability attached, and otherwise we fit triangular distributions, in precisely the same way as explained in Engelberg *et al.* (2009, see p.37-8). From these distributions we obtain estimates of z and log scores.

4 The benchmark density forecasts.

Our benchmark density forecasts (BMs) are constructed from the historical distributions of past median SPF point prediction forecast errors, centred on the median point predictions of the annual year-on-year growth rates. The rolling-event forecasts of the quarterly values (at annual rates) at horizons up to 4-quarters ahead are used. These forecasts are used to construct *ex ante*, real-time distributions of forecast errors that are comparable to the errors in forecasting the calendar-year annual growth rates. We require: (i) that the forecast horizons match, and (ii) that the targets match. We explain how this is achieved by way of an example. Consider the first survey 1981:Q3. The histogram forecasts made in response to this survey are 2-quarters ahead (in the sense that the latest data values for output growth or inflation are for 1981:Q2). Hence we need 2-step ahead forecast errors to construct the BMs. They must also be of the annual growth rate relative to the previous year. We calculate 50 2-step ahead forecasts errors, using the 1968:Q4 survey to the 1981:1Q1 survey, inclusive, where the target is the year-on-year growth rate. That is, the first forecast error is the actual growth rate in the four quarters up to and including 1969:Q1, relative to the four quarters through 1968:Q1, all taken from the 1969:Q2 vintage, compared to the forecast of the four quarters through 1969:Q1 from the 1968:Q4 survey, relative to the four quarters through 1968:Q1. The forecast from the 1968:Q4 survey consists of forecasts of 1969:Q1 and 1968:Q4, and all the other quarterly values used to construct the forecast of the year-on-year growth rate are data taken from the 1968:Q4 data vintage.

The last of the 50 forecast errors is the forecast of the four quarters through 1981:Q2 relative to the previous four quarters, from the 1981:Q1 survey (so that we use forecasts of 1981:Q2 and 1981:Q1, and vintage 1981:Q1 data values of the other quarters), and the actual value is the same quantity taken from the 1981:Q3 data vintage. Hence, all these 50 forecast errors are available at the time of the 1981:Q3 survey, they are all of length 2, all forecast the change in one year relative to the previous year, and in each case the actual uses the real-time vintage (the advance estimate for the most recent quarter being forecast).

For the 1981:Q4 survey, we again use the latest available set of 50 forecast errors, but recognizing that the forecast horizon should now be 1 quarter. Hence the first forecast error is constructed using the forecasts of the four quarters through 1969:Q2 relative to the previous four periods, from the 1969:Q2 survey (so that 1969:Q2 is a forecast, and the other quarters are actual data from the 1969:Q2 vintage), and the actual growth rate is constructed as the same quantity taken from the 1969:Q3 vintage. The last of the 50 forecast errors is the forecast of the four quarters through 1981:Q3 relative to the previous four quarters, from the 1981:Q3 survey (so, a forecast of 1981:3, and 1981:3 data vintage values of the other quarters), compared to the growth in the four quarters to 1981:Q3 relative to the previous four quarters, all from the 1981:Q4 vintage. This set of 50 forecasts are all of length one, but otherwise are the same as those calculated for the 1981:Q3 survey: they are of year-on-year growth rates, and are real-time. We continue through the sample up to 2013:4 in this way, ensuring the forecasts are of the same length as the histogram forecasts (i.e., of length 4, 3, 2 or 1 depending on whether the survey is in Q1, Q2, Q3 or Q4, and refer to year-on-year quantities.

Our choice of benchmark forecasts are motivated by Rossi and Sekhposyan (2015). They use the empirical distribution of SPF forecast errors to construct an uncertainty index, based on the percentile of the historical distribution which corresponds to the realization: realizations in the tails are deemed more uncertain than those away from the tails.[7] Our focus is different, but nevertheless past SPF forecast errors provide distributions against which the SPF conditional distributions can be compared. We fit normal distributions to the past distributions. For example, the 1981:Q3 benchmark distribution (BM) is assumed to be normal with mean given by the median SPF point prediction of the rate of growth of 1981 over 1980 made in response to the 1981:Q3 survey, and variance given by the sample variance of the past forecast errors. This facilitates the calculation of z^* for evaluation of the BMs as well as the log score for comparison to the SPF histogram-based distributions.[8]

5 Empirical Results

5.1 The Aggregate Distributions

Aggregate distributions calculated by equal weighting of the individual respondents' forecast distributions are often the object of the analysis. Equal weighting is known in the literature as the linear opinion pool (see Genest and Zidek (1986)), and such forecasts are reported by the US SPF with the individual histograms, although there are other ways of combining density forecasts (Hall and Mitchell (2009) provide a review). Denoting individual i's density forecast for Y_t made at time $t-h$ by $p_{Y,i,t-h}(y_t)$, with mean and variance $\mu_{i,t|t-h} = \int_{-\infty}^{\infty} y_t p_{Y,i,t-h}(y_t)\,\partial y_t$ and $\sigma^2_{i,t|t-h} = \int_{-\infty}^{\infty} (y_t - \mu_{it})^2 p_{Y,i,t-h}(y_t)\,\partial y_t$, for $i = 1, \ldots, N$ the aggregate density is: $p_{Y,t-h}(y_t) = \frac{1}{N}\sum_{i=1}^{N} p_{Y,i,t-h}(y_t)$ with mean and variance given by:

$$\begin{aligned} \mu_{t|t-h} &= \tfrac{1}{N}\sum_{i=1}^{N} \mu_{i,t|t-h} \\ \sigma^2_{t|t-h} &= \tfrac{1}{N}\sum_{i=1}^{N} \sigma^2_{i,t|t-h} + \tfrac{1}{N}\sum_{i=1}^{N} \left(\mu_{i,t|t-h} - \mu_{t|t-h}\right)^2. \end{aligned}$$

Hence the mean of the aggregate distribution is the simple average of the means of the individual distributions, whereas the variance is the average of the individual variances plus the second term which measures disagreement between forecasters, and serves to increase the aggregate variance relative to the cross-section average: see, e.g., Lahiri, Teigland and Zaporowski (1988) and Wallis (2005), *inter alia*.

Use of the aggregate histograms allows unbroken sequences of forecasts across the entire sample, as the average is taken across all respondents (so 'N' varies over t) and the changing composition is ignored. For many individuals there are many non-response surveys, generally due to late joining or leaving the survey, which is obviously exacerbated by the the survey's long historical duration (relative to similar surveys, such as those run by the ECB and Bank of England, for example). The aggregate histograms are often regarded as

[7] In principle, at least, their measure conflates 'predictable' and 'unpredictable' uncertainty. In the sense that when (conditional) uncertainty is high, a realization in the tail of the historical (unconditional) distribution is far less surprising than a tail realization when conditional uncertainty is low. In practice, obtaining reliable estimates of conditional uncertainty may prove difficult. As we show in this paper, survey histogram forecasts - a prima facie candidate measure - perform relatively poorly.

[8] We do not need to assume normality to calculate z (and hence z^*): we could simply look at the proportion of the historical errors which are less than the realization. However, when this is 0 or 1, the calculation of z^* is problematic as the inverse normal cdf is not defined.

a summary measure of the information in the survey, in much the same way as the median point prediction is often taken as *the* survey point prediction, and used in comparisons with model forecasts. But whereas the average point prediction is a 'good' summary measure (see, e.g., Clemen (1989), Aiolfi, Capistrán and Timmermann (2011) and Manski (2011) as examples of a very large literature), the same may not be true of density combination because of the inflation of the variance of the aggregate histogram (relative to a randomly-selected individual histogram). We address this issue by evaluating the aggregate and individual histograms.

The aggregate histograms always allocate non-zero probability to more than 2 bins, so that the normal distribution is fit to all the histograms, and is used to calculate z^* and the log scores.[9] We find that the SPF output growth forecasts are not rejected at the longest horizon forecasts (Quarter 1 surveys, corresponding to $h = 4$), but all the shorter-horizon forecasts are rejected: see table 1. In stark contrast, only the shortest horizon (quarter four) BM forecasts are rejected, on the grounds that the corresponding z^*'s are not zero-mean and unit-variance. The results for the BMs suggest that unconditional forecasts adequately approximate the true densities (in a statistical sense) for all but the 1-quarter ahead horizon. However, the SPF forecasts are mis-specified at all but the longest horizon.

Table 2 suggests that the SPF forecasts are rejected at all four horizons for inflation, whereas the BM forecasts are well-specified for the longer two horizons, but not at 1 and 2 quarters ahead. Hence the results for the BMs match those for output, and are intuitively plausible - failing to condition on forecast origin information leads to rejection at the shorter horizons (1 quarter ahead, for output growth, and up to 2 quarters ahead for inflation).

A possible explanation of the rejection of the aggregate SPF forecasts is that disagreement inflates the variances of the aggregate histograms, leading to the histograms being 'too dispersed' given the realizations, and the z^* variables showing too little dispersion: this is consistent with the estimates of the variances of the AR(1) model fit to $\{z_t^*\}$ reported in tables 1 and 2, which are less than unity at all h for both variables. Clements (2014a) reports estimates for the variances of the aggregate histograms' standard deviations and the averages of individual standard histograms, averaged over the years 1983–2010. At $h = 4$ the aggregate histogram standard deviations are around 20% higher for both variables, and at $h = 1$ the aggregate histogram measure is 46% higher for output growth, and 22% higher for inflation. Consequently, the effect of aggregating the individual histograms could explain the rejection of the inflation histograms.

That said, some recent work on forecast uncertainty suggests that appropriate measures of the uncertainty of the consensus forecast should include uncertainty arising from the heterogeneity of individual forecasters (see Lahiri and Sheng (2010) and Lahiri, Peng and Sheng (2015)). According to this line of reasoning, the effect of aggregation on the variance of the distribution ought not be a reason for rejection.

Finally, the final columns of tables 1 and 2 report the comparisons of SPF and BM forecasts based on KLIC (or log scores) for the two variables (equation (5)). The entries in the tables are constructed such that a value less than 0.05 leads to a rejection of equal accuracy in favour of the BM at the 5% level (one-sided test), and an entry greater than 0.95 rejects in favour of the SPF being more accurate (at the 5% level in a one-sided test). The aggregate SPF are less accurate at all but the longer two horizons for output growth (and the test outcome is marginal for the 3-quarter horizon), while for inflation the SPF are rejected at all

[9] When, as here, z is calculated after fitting a Gaussian distribution to the histogram, taking the inverse standard normal cdf of z to give z^* results in $z^* = (y - \hat{\mu})/\hat{\sigma}$, where y is the realization, and $\hat{\mu}$ and $\hat{\sigma}^2$ are the mean and variance of the fitted distribution.

but the longest horizon.

5.2 The Individual Distributions

We report results for all individuals who filed more than 15 forecasts of a given horizon. We fit normal distributions when histograms have 3 or more non-zero bins, and (isosceles) triangular distributions when either one bin has all the probability mass, or it is distributed across two (adjacent) bins. Triangular distributions result in values of z of 0 or 1 when the realization falls outside of the support: these are arbitrarily set to 0.01 and 0.99, respectively. (Hence z^* is well defined, and an interpretation is that no realized values are viewed as 'impossible' by the SPF respondents). Moreover, log scores are set to $\ln(0.01)$ in both these circumstances.

Tables 3 and 4 report results for output growth and inflation. We report a two-degree of freedom test of zero-mean and unit-variance, but do not allow an AR(1) as the unrestricted model because there are fewer observations in most individual cases and because there are missing values, which would complicate the fitting of an autoregressive model. (Moreover, it seems likely that power comes from testing the mean and variance of z^*, not from testing for autocorrelation). For each individual and horizon, then, we report the number of forecast observations, the p-value of the Berkowitz two-degree of freedom test, the estimates of the mean and variance, and the Diebold-Mariano test of the log scores. The results for individuals are sorted by p-value of the Berkowitz test within each survey quarter (equivalently, forecast horizon).

Consider table 3 for output growth. Except at the longest horizon, the statistical adequacy of the SPF densities is rejected for fewer than one third of respondents at the 5% level. (Specifically, for Q2, 2 of 12; for Q3, 2 of 9; and for Q4, 4 of 13). This suggests the scales of the individual respondents' histograms may be better calibrated than the scales of the aggregate distributions, which would be consistent with the concern over 'variance-inflation' from aggregating the individual forecasts. However, the fewer rejections at the individual level may also reflect lower power due to the smaller sample sizes. Tellingly, the BM forecasts are statistically more accurate on log score than the forecasts of each individual SPF respondent at the shortest horizon (i.e., for Q4 surveys). (The entries in the table are constructed such that a value greater than 0.95 suggests the BM are more accurate at the 5% level (one-sided test), and an entry less than 0.05 rejects in favour of the SPF being more accurate (at the 5% level in a one-sided test). At longer horizons there are fewer rejections (e.g., for Q1 forecasts there are 3 out of 12, and for Q2 6 out of 12) but it is never the case that we reject in favour of an SPF respondent at any horizon. The comparisons against the BM forecasts suggests that the individual SPF densities are poor, especially at the shortest horizon, and the failure to reject on tests of correct specification is likely due to small sample sizes.

For inflation (table 4) the rejection of the SPF forecasts is less equivocal, in that the SPF forecasts are rejected on the Berkowitz test for 7 out of 8 and 9 out of 12 respondents for Q3 and Q4 forecasts, respectively, and in addition, the forecasts of all these individuals are found to be statistically less accurate than the BM forecasts.

6 Robustness of the results to the assumptions

We consider whether the findings discussed in section 5 are unduly dependent on the assumptions we have made in order to construct forecast densities from the histograms.

6.1 The wider histogram bin widths before 1992:Q1.

Prior to 1992, respondents assigned operabilities to intervals of width 2 percentage points. From 1992Q1 onwards a finer gradation was adopted with intervals being halved. The use of wide intervals may give a misleading picture when uncertainty is low, and all the probability is assigned to one interval. In such circumstances, a symmetric triangular distribution (with support on the full interval) results in a variance of 0.0416 when the interval width is one, but (four times as large) at 0.1666 when the interval width is two. Individuals are more likely to assign all the probability to one interval in response to Q4 surveys, because perceived uncertainty will be smaller at the shortest horizon. Hence our approach will place a floor on the variance (and will affect the z and log score calculations) for earlier-period one-bin histograms which may be particularly distortionary prior to 1992.

We tackle this in two ways.

Firstly, we approximate the one-bin histograms prior to 1992 by a symmetric triangular distribution with a base of one (rather than two), located centrally within the interval. There is no way of knowing whether this provides a more accurate representation of the underlying subjective distribution. We simply wish to assess whether the way we treat the wide single-bin histograms is driving the results. This has no effect on the aggregate results for output growth, because all the aggregate histograms allot probability to 3 or more bins. The results by individual are qualitatively unaffected - for the shortest horizon we still reject for 4 of the 13 respondents, and for all respondents reject in favour of the BMs on log score comparisons. (These results are available in an Appendix). The same is true for inflation.

Secondly, we set the beginning of the forecast sample to 1992:Q1. This has the drawback of discarding around a third of the observations, and reducing the number of individual respondents we can separately analyze, but counters the potentially more insidious effects of the wider bins not picked up by the first strategy. The aggregate histogram results for output growth and inflation are broadly unchanged, apart from the SPF output growth forecast no longer being rejected at $h = 3$. Hence we continue to reject the SPF inflation forecasts being well specified at all horizons, and the SPF output growth forecasts are rejected at the two shorter horizons. The pattern of results by individual is also largely unchanged. At $h = 1$ we find evidence against the forecasts being well specified for 5 of 11 and 7 of 10 respondents for output growth and inflation, respectively, while for all these respondents we reject in favour of the BMs being more accurate.

6.2 Centreing the SPF densities on the point predictions.

Engelberg *et al.* (2009) find inconsistencies between the central tendencies and the point predictions for some SPF respondents, and Clements (2010) finds evidence that the point predictions tend to be more accurate in terms of traditional forecast evaluation criteria such as squared error loss. This raises the suspicion that the relatively good performance of the BMs may result from their being centred on the (cross-sectional) median *point* prediction. This turns out not to be the case. Centreing the SPF aggregate histograms on the median point predictions, and the individual histograms on the individuals' point predictions,[10] does not result in a marked improvement in the SPF forecasts.

[10] For the individual respondent histograms with probability assigned to 3 bins or more, a normal density is fit to the histogram, as in the standard approach, and the estimated mean is then replaced by the individual's point prediction. For the one and two bin histograms for which we assume triangular distributions no use is made of the point prediction.

7 Correcting the SPF aggregate density forecasts

Given that our findings appear relatively robust, we consider whether the SPF forecasts can be improved with simple mechanical corrections. Such corrections are commonplace in the point prediction forecasting literature, and are sometimes viewed as a way of 'fixing' a model's forecasts for mis-specification resulting from structural change (see, e.g., Castle, Clements and Hendry (2015)). It is possible to (re-)calibrate future forecast densities for the apparent mis-specification of past densities (for which realizations are available): see, e.g., Dawid (1984), Kling and Bessler (1989) and Diebold, Hahn and Tay (1999). However, given the relatively small number of forecast densities of a given horizon, we consider whether a simple scaling of the SPF aggregate densities would improve their accuracy. Based on the first 15 forecast densities (that is, the densities of 1982 to 1996 for the Q1 and Q2 surveys, or 1981 to 1995 for the Q3 and Q4 surveys), we calculate a horizon-specific scale factor that maximizes log score over this in-sample period,[11] and then apply these factors to the variances of the remaining, out-of-sample density forecasts. These corrections are calculated for, and applied to, the aggregate histograms centred on the median point predictions.

Success would require that the variances of the SPF densities systematically over (or under) estimate the uncertainty over the in-sample period, and that the same remains true of the out-of-sample period. Table 5 records the results for the out-of-sample forecast densities 1996(1997) to 2013.

The first three rows of each panel show that the results for our reduced sample period match the whole sample: the DM tests reject the null hypotheses that the SPF densities are as accurate as the BM densities for all but the longest-horizon Q1 surveys forecasts. Notwithstanding the smaller sample of forecasts, the differences in the log scores recorded in the table are sufficiently large to reject the null that the SPF and DM densities are of equal accuracy in population.

The table shows marked improvements on log score from scaling-down the variances by the fixed in-sample factors, such that the SPF output growth densities are not rejected at any horizon, and furthermore, the inflation forecasts are now more accurate than the benchmark forecasts at the shorter two horizons. That such corrections yield significant improvements in accuracy indicates that the SPF forecast densities are systematically mis-specified at the shorter horizons, in that the dispersions are too wide. Corrections based on past performance up to the mid 1990s yield improvements on average over the remainder of the 1990's and the period up to 2013.

8 Conclusions

Other papers have considered the aggregate US SPF histograms, see, e.g., Diebold, Tay and Wallis (1999). The novelty of the current contribution is the investigation of the term structure of the aggregate densities

[11] Given a set of normal forecast densities defined by $\{\mu_t, \sigma_t^2\}_{t=1}^N$ and realizations $\{x_t\}_{t=1}^N$, the log score is defined by:

$$\sum_{t=1}^N \ln p\left(x_t; \mu_t, \sigma_t^2\right) = -\sum_{t=1}^N \ln \sigma_t - \frac{1}{2}\sum_{t=1}^N \ln 2\pi - \sum_{t=1}^N \frac{(x_t - \mu_t)^2}{2\sigma_t^2}$$

Choosing λ to maximize the log score over 1 to N, where $\hat{\sigma}_t^2 = \lambda \sigma_t^2$:

$$\sum_{t=1}^N \ln p\left(x_t; \mu_t, \hat{\sigma}_t^2\right) = -\frac{1}{2} N \ln \lambda - \sum_{t=1}^N \ln \sigma_t - \frac{1}{2} N \ln 2\pi - \frac{1}{2\lambda}\sum_{t=1}^N \frac{(x_t - \mu_t)^2}{\sigma_t^2}$$

results in $\hat{\lambda} = \frac{1}{N}\sum_{t=1}^N \left((x_t - \mu_t)/\sigma_t\right)^2$.

and those of individual survey respondents, and the comparison to the benchmark forecasts. The latter are designed such that the SPF densities would be superior were the respondents able to gauge the degree of uncertainty characterizing the macro-outlook at the time the forecasts are made.

Our findings suggest that the aggregate and individual forecast densities tend to be too dispersed at the shorter forecast horizons, such as one quarter and two quarters ahead, consistent with the findings of Clements (2014a) in his study of *ex ante* and *ex post* forecast uncertainty. Because of the excess dispersion of the survey densities at short horizons, the expected dominance of the SPF densities over the unconditional-variance benchmark distributions at short horizons does not materialize. The benchmark densities are rejected at short horizons on tests of correct specification, as expected, given that the dispersions of these distributions are not conditioned on developments at the time the forecasts are made. However, the excess dispersion on the aggregate SPF densities renders these densities less accurate than the benchmarks on log score comparisons.

Of course the forecasters' subjective probability assessments at short horizons may well be driven by motives other than maximizing accuracy as judged by log score. The respondents' loss functions may be such that the respondents are incentivized to ensure that realized outcomes fall well within the likely range of outcomes implied by their probability assessments, for example. Be that as it may, the excess dispersion of the aggregate histograms at short horizons is sufficiently large and persistent that the application of correction factors (based on a training sample of forecasts and realizations) to out-of-sample probability assessments leads to marked improvements in their (log score) accuracy.

Although we have considered a single macro-survey, it is the longest and probably most used in terms of the probability assessments it provides. Our findings question the reliability of the short-horizon survey densities.

References

Aiolfi, M., Capistrán, C., and Timmermann, A. (2011). Forecast combinations, chapter 11. In Clements, M. P., and Hendry, D. F. (eds.), *The Oxford Handbook of Economic Forecasting*, pp. 355–388: Oxford University Press.

Andrade, P., and Le Bihan, H. (2013). Inattentive professional forecasters. *Journal of Monetary Economics*, *60*(8), 967–982.

Ang, A., Bekaert, G., and Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation better?. *Journal of Monetary Economics*, **54(4)**, 1163–1212.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, **19(4)**, 465–474.

Boero, G., Smith, J., and Wallis, K. F. (2015). The measurement and characteristics of professional forecasters' uncertainty. *Journal of Applied Econometrics*. Forthcoming.

Castle, J. L., Clements, M. P., and Hendry, D. F. (2015). Robust approaches to forecasting. *International Journal of Forecasting*, **31**, 99–112.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5**, 559–583. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

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

Clements, M. P. (2010). Explanations of the Inconsistencies in Survey Respondents Forecasts. *European Economic Review*, **54(4)**, 536–549.

Clements, M. P. (2014a). Forecast Uncertainty - Ex Ante and Ex Post: US Inflation and Output Growth. *Journal of Business & Economic Statistics*, **32(2)**, 206–216. DOI: 10.1080/07350015.2013.859618.

Clements, M. P. (2014b). Probability distributions or point predictions? Survey forecasts of US output growth and inflation. *International Journal of Forecasting*, **30(1)**, 99–117. DOI: 10.1016/j.ijforecast.2013.07.010.

Clements, M. P. (2015). Are professional macroeconomic forecasters able to do better than forecasting trends?. *Journal of Money, Credit and Banking*, **47,2-3**, 349–381. DOI: 10.1111/jmcb.12179.

Coibion, O., and Gorodnichenko, Y. (2012). What can survey forecasts tell us about information rigidities?. *Journal of Political Economy*, *120*(1), 116 – 159.

Coibion, O., and Gorodnichenko, Y. (2015). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, *105*(8), 2644–78.

Croushore, D. (1993). Introducing: The Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Business Review*, **"November"**, 3–15.

Croushore, D., and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, **105(1)**, 111–130.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of The Royal Statistical Society, ser. A*, **147**, 278–292.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts: With applications to financial risk management. *International Economic Review*, **39**, 863–883.

Diebold, F. X., Hahn, J. Y., and Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: High frequency returns on foreign exchange. *Review of Economics and Statistics*, **81**, 661–673.

Diebold, F. X., and Kilian, L. (2001). Measuring predictability: Theory and macroeconomic applications. *Journal of Applied Econometrics*, **16**, 657–669.

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253–263.

Diebold, F. X., Tay, A. S., and Wallis, K. F. (1999). Evaluating density forecasts of inflation: The Survey of Professional Forecasters. In Engle, R. F., and White, H. (eds.), *Festschrift in Honor of C. W. J. Granger*, pp. 76–90: Oxford: Oxford University Press.

Doornik, J. A., and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, **70**, 927–939.

Engelberg, J., Manski, C. F., and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics*, **27(1)**, 30–41.

Engle, R. F., and Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression

Quantiles. *Journal of Business & Economic Statistics*, **22**, 367–381.

Genest, C., and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, **1**, 114–148.

Giordani, P., and Söderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, **47(6)**, 1037–1059.

Granger, C. W. J., and Pesaran, M. H. (2000). A decision-based approach to forecast evaluation. In Chan, W. S., Li, W. K., and Tong, H. (eds.), *Statistics and Finance: An Interface*, pp. 261–278: London: Imperial College Press.

Hall, S. G., and Mitchell, J. (2009). Recent developments in density forecasting. In Mills, T. C., and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*, pp. 199–239: Palgrave MacMillan.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility : likelihood inference and comparison with ARCH models. *Review of Economic Studies*, **81**, 361–393.

Kling, J. L., and Bessler, D. A. (1989). Calibration-based predictive distributions: An application of prequential analysis to interest rates, money, prices and output. *Journal of Business*, **62**, 477–499.

Kullback, L., and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Sciences*, **22**, 79–86.

Lahiri, K., and Sheng, X. (2010). Measuring forecast uncertainty by disagreement:the missing link. *Journal of Applied Econometrics*, **25**, 514–538.

Lahiri, K., Teigland, C., and Zaporowski, M. (1988). Interest rates and the subjective probability distribution of inflation forecasts. *Journal of Money, Credit and Banking*, **20(2)**, 233–248.

Lahiri, K., Peng, H., and Sheng, X. (2015). Measuring Uncertainty of a Combined Forecast and Some Tests for Forecaster Heterogeneity. Cesifo working paper series 5468, CESifo Group Munich.

Lee, T.-H., Bao, Y., and Saltoglu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, *26*(3), 203–225.

Manski, C. F. (2011). Interpreting and combining heterogeneous survey forecasts. In Clements, M. P., and Hendry, D. F. (eds.), *Oxford Handbook of Economic Forecasting, Chapter 16*, pp. 457–472. Oxford: Oxford University Press.

Mitchell, J., and Hall, S. G. (2005). Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR 'Fan' Charts of Inflation. *Oxford Bulletin of Economics and Statistics*, *67*(s1), 995–1033.

Pesaran, M. H., and Weale, M. (2006). Survey expectations. In Elliott, G., Granger, C., and Timmermann, A. (eds.), *Handbook of Economic Forecasting, Volume 1. Handbook of Economics 24*, pp. 715–776: Elsevier, Horth-Holland.

Rich, R., and Tracy, J. (2010). The relationships among expected inflation, disagreement, and uncertainty: Evidence from matched point and density forecasts. *Review of Economics and Statistics*, **92(1)**, 200–207.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, **23**, 470–472.

Rossi, B., and Sekhposyan, T. (2015). Macroeconomic Uncertainty Indices Based on Nowcast and Forecast Error Distributions. *American Economic Review*, *105*(5), 650–55.

Shenton, L. R., and Bowman, K. O. (1977). A bivariate model for the distribution of $\sqrt{b_1}$ and b_2. *Journal of the American Statistical Association*, **72**, 206–211.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika*, **81**, 115–131.

Wallis, K. F. (2005). Combining Density and Interval forecasts: A Modest Proposal. *Oxford Bulletin of Economics and Statistics*, **67(s1)**, 983–994.

Wright, J. H. (2013). Evaluating Real-Time VAR forecasts with an informative democratic prior. *Journal of Applied Econometrics*, **28**, 762–776. DOI: 10.1002/jae.2268.

Zarnowitz, V., and Lambros, L. A. (1987). Consensus and uncertainty in economic prediction. *Journal of Political Economy*, **95(3)**, 591–621.

Table 1: Aggregate Histograms of Output Growth: SPF & Benchmark

		Evaluation based on z^*						Comparison (Eqn. 5)
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	
1	SPF	0.14	0.30	0.46	0.17	0.29	0.76	0.46
	BM	0.27	0.19	0.17	0.24	0.21	0.64	
2	SPF	0.17	0.01	0.01	0.03	0.26	0.38	0.04
	BM	0.80	0.08	0.04	-0.01	0.05	0.46	
3	SPF	0.08	0.00	0.00	0.13	0.32	0.24	0.00
	BM	0.48	0.20	0.12	0.13	0.13	0.58	
4	SPF	0.90	0.00	0.00	0.12	0.02	0.25	0.00
	BM	0.88	0.01	0.00	0.14	0.03	0.35	

The table records the results of evaluating the densities using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (5).
The first column denotes the survey quarter, whereby '1' indicates a first quarter of the year survey, and a forecast horizon of 4 quarters, and '4' a fourth quarter survey (and a horizon of 1 quarter). The column headed 'Ind.' is the p-value of a test for independence: in terms of Eqb. (2) the test is based on $-2(L\left(\tilde{\mu},\tilde{\sigma}^2,0\right) - L\left(\hat{\mu},\hat{\sigma}^2,\hat{\rho}\right))$, where $\tilde{\mu},\tilde{\sigma}^2$ denote MLEs with $\rho = 0$ imposed.
The next column is the three-degree of freedom test in Eqn. (2), and the column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence. The next 3 columns report the estimates of the unrestricted AR(1).
The final column is the p-value of the test of SPF and Benchmark densities using Eqn. (5).

Table 2: Aggregate Histograms of Inflation: SPF & Benchmark

		Evaluation based on z^*						Comparison (Eqn. 5)
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	
1	SPF	0.04	0.00	0.00	-0.13	0.40	0.27	0.05
	BM	0.05	0.25	0.82	0.07	0.37	0.90	
2	SPF	0.03	0.00	0.00	-0.22	0.40	0.17	0.00
	BM	0.27	0.09	0.08	-0.19	0.20	0.56	
3	SPF	0.67	0.00	0.00	-0.37	-0.08	0.12	0.00
	BM	0.12	0.01	0.01	-0.40	-0.27	0.42	
4	SPF	0.74	0.00	0.00	-0.34	-0.06	0.14	0.00
	BM	0.19	0.01	0.01	-0.28	0.24	0.49	

See notes to table 1.

Table 3: Evaluation of Individual Respondents' Output Histograms

			Evaluation based on z^*			Comparison (Eqn. 5)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
421	1	20	0.66	0.11	1.26	0.85
431	1	18	0.24	0.35	1.31	0.95
446	1	18	0.16	0.21	1.68	0.86
99	1	16	0.04	-0.34	1.96	0.62
428	1	16	0.02	0.60	1.63	0.97
411	1	17	0.01	0.33	2.17	0.94
426	1	21	0.01	0.37	1.95	0.94
484	1	17	0.00	0.79	0.89	0.99
20	1	17	0.00	0.38	2.54	0.94
433	1	17	0.00	0.53	2.58	0.96
420	1	16	0.00	0.01	3.11	0.89
407	1	16	0.00	0.92	3.47	0.94
421	2	21	0.99	-0.01	0.95	1.00
446	2	20	0.81	0.13	0.92	0.99
411	2	16	0.72	-0.16	0.84	1.00
431	2	18	0.46	0.12	0.67	0.93
99	2	16	0.41	-0.09	0.61	0.56
426	2	19	0.24	0.18	0.59	0.98
463	2	17	0.22	-0.24	0.58	0.99
433	2	18	0.21	0.38	1.26	0.82
484	2	16	0.18	0.29	0.57	1.00
20	2	23	0.15	-0.11	1.66	0.94
407	2	16	0.01	0.08	2.39	0.84
65	2	18	0.01	0.13	2.30	0.93
84	3	19	0.51	0.21	1.25	0.64
20	3	21	0.50	-0.26	1.03	0.95
421	3	17	0.43	0.31	0.99	1.00
420	3	20	0.23	0.16	0.58	0.89
407	3	19	0.19	0.38	1.28	0.69
433	3	19	0.11	0.42	0.70	0.28
426	3	21	0.05	0.27	0.48	1.00
446	3	18	0.02	0.37	0.41	1.00
65	3	17	0.01	0.35	2.14	0.99
84	4	27	0.70	0.15	1.10	1.00
421	4	21	0.60	0.14	0.78	1.00
446	4	17	0.43	0.31	0.99	1.00
411	4	18	0.31	0.24	0.66	1.00
433	4	18	0.25	0.39	1.04	1.00
20	4	20	0.16	-0.18	1.66	1.00
407	4	19	0.08	0.52	1.02	1.00
472	4	16	0.06	0.29	0.44	1.00
431	4	17	0.06	0.57	1.12	1.00
99	4	18	0.02	0.06	2.24	1.00
426	4	20	0.02	0.19	0.36	1.00
420	4	20	0.00	0.36	0.32	1.00
463	4	17	0.00	0.09	0.15	1.00

The table records the results of evaluating the densities of individual respondents using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (5), to compare each individual's forecasts against the Benchmark.
'N' is the number of forecast densities by the individual 'id' made in response to 'Qtr' surveys. The column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence, and the next 2 columns report the estimates of mean and variance.
The final column is the p-values of the test of an SPF indivdual against the Benchmark densities using Eqn. (5).
We consider all respondents who made more than 15 forecasts of a given horizon.

Table 4: Evaluation of Individual Respondents' Inflation Histograms

			Evaluation based on z^*			Comparison (Eqn. 5)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
411	1	18	0.44	0.23	0.75	0.88
446	1	18	0.23	-0.16	0.56	0.89
433	1	18	0.12	0.05	1.84	0.98
420	1	16	0.12	-0.20	1.83	0.98
484	1	17	0.11	-0.19	0.47	0.95
426	1	21	0.09	-0.34	1.58	0.92
99	1	16	0.03	-0.56	0.60	0.91
421	1	18	0.02	-0.34	0.39	0.99
431	1	19	0.00	-0.08	0.22	0.69
20	1	16	0.00	-0.28	3.52	0.97
407	1	16	0.00	-0.88	2.50	1.00
411	2	16	0.54	-0.21	1.27	1.00
431	2	17	0.36	-0.35	0.95	0.98
99	2	17	0.29	-0.35	1.25	1.00
433	2	18	0.20	-0.31	0.65	0.97
65	2	17	0.09	-0.44	1.50	0.98
446	2	20	0.06	-0.33	0.52	1.00
426	2	19	0.04	-0.56	0.77	1.00
463	2	18	0.02	-0.38	0.41	1.00
407	2	16	0.00	-0.87	1.81	0.99
421	2	19	0.00	-1.06	1.73	0.98
20	2	23	0.00	-1.08	3.17	1.00
84	3	20	0.07	-0.43	0.64	1.00
426	3	21	0.02	-0.59	0.71	1.00
420	3	20	0.00	-0.44	0.35	1.00
446	3	18	0.00	-0.36	0.24	1.00
407	3	18	0.00	-0.85	0.55	1.00
433	3	19	0.00	-0.31	0.21	1.00
65	3	17	0.00	-0.49	2.76	1.00
20	3	21	0.00	-1.12	1.44	1.00
84	4	28	0.34	-0.19	1.32	1.00
431	4	17	0.22	-0.42	1.03	1.00
411	4	18	0.18	-0.11	0.52	1.00
99	4	18	0.04	-0.53	1.46	1.00
463	4	17	0.02	-0.28	0.36	1.00
20	4	19	0.02	-0.65	1.16	1.00
433	4	17	0.01	-0.66	1.50	1.00
426	4	19	0.01	-0.47	0.40	1.00
446	4	18	0.00	-0.47	0.34	1.00
421	4	20	0.00	-0.74	0.77	1.00
407	4	18	0.00	-0.72	1.63	1.00
420	4	20	0.00	-0.57	0.25	1.00

See notes to table 3.

Table 5: Effects of in-sample adjustment of SPF aggregate density variances out-of-sample

Survey quarter	Q1	Q2	Q3	Q4
Output Growth				
SPF	-1.18	-0.78	-0.57	-0.17
BM	-1.21	-0.70	-0.36	0.47
DM:SPF versus BM	0.78	0.03	0.00	0.00
SPF cal.	-1.19	-0.70	-0.40	0.27
DB:SPF cal. versus BM	0.55	0.51	0.36	0.19
Inflation				
SPF	-0.79	-0.57	-0.37	-0.14
BM	-0.68	-0.10	0.47	1.05
DM:SPF versus BM	0.27	0.00	0.00	0.00
SPF cal.	-0.70	-0.08	0.58	1.17
DM:SPF cal. versus BM	0.34	0.57	0.99	0.97

Except for the rows prefixed by 'DM:', the entries are average log scores for the years 1997 - 2013 for the Q1 and Q2 surveys, and for 1996 - 2013 for the Q3 and Q4 surveys. The rows labelled 'SPF cal.' show average log scores when the SPF variances are scaled to optimise log score for the years 1982 - 1996 (Q1 and Q2 surveys) or 1981 - 1995 (Q3 and Q4 surveys). The rows prefixed by 'DM:' record the p-values of Diebold-Mariano tests of equal predictive ability on log score, computed such that values close to 1 reject in favour of the SPF forecasts, and values close to zero reject in favour of the benchmark forecasts.

9 Not For Publication Appendix

The tables in the appendix provide the detailed results which are summarized in section 6.

For output growth and inflation, tables 6 and 7 provide the results for individual respondents when the calculations for single-bin histograms assume the bin is of width 1.

Tables 8 and 9 report results for the aggregate histograms when the forecast sample begins in 1992:Q1, and tables 10 and 11 report the results for the individual respondents.

Finally, tables 12 and 13 report results for the aggregate SPF histograms centred on the median point predictions.

Table 6: Evaluation of Individual Respondents' Output Histograms: Single-bin histogram widths set to 1

			Evaluation based on z^*			Comparison (Eqn. 5)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
421	1	20	0.66	0.11	1.26	0.85
431	1	18	0.24	0.35	1.31	0.95
446	1	18	0.16	0.21	1.68	0.86
99	1	16	0.04	-0.34	1.96	0.62
428	1	16	0.02	0.60	1.63	0.97
411	1	17	0.01	0.33	2.17	0.94
426	1	21	0.01	0.37	1.95	0.94
484	1	17	0.00	0.79	0.89	0.99
20	1	17	0.00	0.38	2.54	0.94
433	1	17	0.00	0.53	2.58	0.96
420	1	16	0.00	0.01	3.11	0.89
407	1	16	0.00	0.92	3.47	0.94
421	2	21	0.99	-0.01	0.95	1.00
446	2	20	0.81	0.13	0.92	0.99
411	2	16	0.72	-0.16	0.84	1.00
431	2	18	0.46	0.12	0.67	0.93
99	2	16	0.41	-0.09	0.61	0.56
426	2	19	0.24	0.18	0.59	0.98
463	2	17	0.22	-0.24	0.58	0.99
433	2	18	0.21	0.38	1.26	0.82
484	2	16	0.18	0.29	0.57	1.00
20	2	23	0.15	-0.11	1.66	0.94
407	2	16	0.01	0.08	2.39	0.84
65	2	18	0.01	0.07	2.49	0.96
84	3	19	0.52	0.20	1.25	0.57
20	3	21	0.50	-0.26	1.03	0.95
421	3	17	0.43	0.31	0.99	1.00
420	3	20	0.23	0.16	0.58	0.89
407	3	19	0.19	0.38	1.28	0.69
433	3	19	0.11	0.42	0.70	0.28
426	3	21	0.05	0.27	0.48	1.00
446	3	18	0.02	0.37	0.41	1.00
65	3	17	0.00	0.23	2.55	0.99
421	4	21	0.62	0.14	0.79	1.00
446	4	17	0.43	0.31	0.99	1.00
84	4	27	0.41	0.15	1.33	1.00
411	4	18	0.31	0.24	0.66	1.00
433	4	18	0.25	0.39	1.04	1.00
20	4	20	0.16	-0.18	1.66	1.00
407	4	19	0.08	0.51	1.03	1.00
472	4	16	0.06	0.29	0.44	1.00
431	4	17	0.06	0.57	1.12	1.00
426	4	20	0.02	0.19	0.36	1.00
99	4	18	0.01	0.00	2.45	1.00
420	4	20	0.00	0.36	0.32	1.00
463	4	17	0.00	0.09	0.15	1.00

See notes to table 3.

Table 7: Evaluation of Individual Respondents' Inflation Histograms: Single-bin histogram widths set to 1

			Evaluation based on z^*			Comparison (Eqn. 5)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
411	1	18	0.44	0.23	0.75	0.88
446	1	18	0.23	-0.16	0.56	0.89
433	1	18	0.12	0.05	1.84	0.98
420	1	16	0.12	-0.20	1.83	0.98
484	1	17	0.11	-0.19	0.47	0.95
426	1	21	0.09	-0.34	1.58	0.92
99	1	16	0.03	-0.56	0.60	0.91
421	1	18	0.02	-0.34	0.39	0.99
431	1	19	0.00	-0.08	0.22	0.69
20	1	16	0.00	-0.28	3.52	0.97
407	1	16	0.00	-0.88	2.50	1.00
411	2	16	0.54	-0.21	1.27	1.00
431	2	17	0.36	-0.35	0.95	0.98
99	2	17	0.29	-0.35	1.25	1.00
433	2	18	0.20	-0.31	0.65	0.97
65	2	17	0.16	-0.35	1.48	0.97
446	2	20	0.06	-0.33	0.52	1.00
426	2	19	0.04	-0.56	0.77	1.00
463	2	18	0.02	-0.38	0.41	1.00
407	2	16	0.00	-0.87	1.81	0.99
421	2	19	0.00	-1.06	1.73	0.98
20	2	23	0.00	-1.08	3.17	1.00
84	3	20	0.21	-0.37	0.83	1.00
426	3	21	0.02	-0.59	0.71	1.00
420	3	20	0.00	-0.44	0.35	1.00
446	3	18	0.00	-0.36	0.24	1.00
407	3	18	0.00	-0.85	0.55	1.00
433	3	19	0.00	-0.31	0.21	1.00
65	3	17	0.00	-0.51	2.71	1.00
20	3	21	0.00	-1.12	1.44	1.00
431	4	17	0.22	-0.42	1.03	1.00
411	4	18	0.18	-0.11	0.52	1.00
84	4	28	0.08	-0.11	1.70	1.00
99	4	18	0.03	-0.48	1.71	1.00
463	4	17	0.02	-0.28	0.36	1.00
20	4	19	0.02	-0.65	1.16	1.00
433	4	17	0.01	-0.66	1.50	1.00
426	4	19	0.01	-0.47	0.40	1.00
446	4	18	0.00	-0.47	0.34	1.00
421	4	20	0.00	-0.74	0.77	1.00
407	4	18	0.00	-0.72	1.63	1.00
420	4	20	0.00	-0.57	0.25	1.00

See notes to table 3.

Table 8: Aggregate Histograms of Output Growth: SPF & Benchmark, Forecast Sample begins in 1992:Q1

		Evaluation based on z^*						Comparison (Eqn. 5)
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	
1	SPF	0.16	0.53	0.88	0.06	0.31	0.83	0.48
	BM	0.25	0.44	0.50	0.09	0.26	0.66	
2	SPF	0.15	0.07	0.08	0.01	0.34	0.39	0.03
	BM	0.80	0.30	0.16	-0.01	0.06	0.50	
3	SPF	0.06	0.00	0.00	0.17	0.43	0.23	0.03
	BM	0.53	0.34	0.23	0.26	0.15	0.67	
4	SPF	0.30	0.00	0.00	0.16	0.23	0.21	0.00
	BM	0.79	0.02	0.01	0.33	-0.07	0.35	

See notes to table 1.

Table 9: Aggregate Histograms of Inflation: SPF & Benchmark, Forecast Sample begins in 1992:Q1

		Evaluation based on z^*						Comparison (Eqn. 5)
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	
1	SPF	0.14	0.02	0.02	-0.08	0.35	0.32	0.12
	BM	0.11	0.25	0.45	0.14	0.36	1.08	
2	SPF	0.02	0.00	0.00	-0.14	0.50	0.14	0.00
	BM	0.27	0.48	0.54	-0.08	0.25	0.66	
3	SPF	0.75	0.00	0.00	-0.34	-0.08	0.11	0.00
	BM	0.94	0.03	0.01	-0.18	-0.02	0.33	
4	SPF	0.99	0.00	0.00	-0.36	0.00	0.11	0.00
	BM	0.39	0.18	0.13	-0.23	0.20	0.55	

See notes to table 1.

Table 10: Evaluation of Individual Respondents' Output Growth Histograms: Forecast Sample begins in 1992:Q1

			Evaluation based on z^*			Comparison (Eqn. 5)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
421	1	19	0.57	0.14	1.31	0.81
431	1	17	0.20	0.36	1.38	0.95
446	1	18	0.16	0.21	1.68	0.86
426	1	20	0.01	0.49	1.73	0.91
411	1	16	0.01	0.36	2.28	0.94
484	1	17	0.00	0.79	0.89	0.99
420	1	16	0.00	0.01	3.11	0.89
433	1	16	0.00	0.62	2.62	0.95
421	2	20	1.00	-0.01	1.00	1.00
446	2	20	0.81	0.13	0.92	0.99
431	2	17	0.53	0.14	0.70	0.90
426	2	18	0.25	0.21	0.60	0.96
463	2	17	0.22	-0.24	0.58	0.99
484	2	16	0.18	0.29	0.57	1.00
433	2	17	0.14	0.45	1.26	0.81
421	3	16	0.39	0.34	1.03	0.99
420	3	20	0.23	0.16	0.58	0.89
407	3	17	0.10	0.48	1.32	0.53
433	3	18	0.06	0.48	0.67	0.25
426	3	20	0.06	0.28	0.50	0.99
446	3	18	0.02	0.37	0.41	1.00
421	4	19	0.70	0.16	0.86	1.00
411	4	16	0.46	0.24	0.74	1.00
446	4	17	0.43	0.31	0.99	1.00
84	4	17	0.27	0.29	1.41	1.00
433	4	17	0.22	0.42	1.08	1.00
472	4	16	0.06	0.29	0.44	1.00
431	4	16	0.04	0.62	1.14	1.00
407	4	17	0.04	0.62	1.05	1.00
426	4	19	0.02	0.20	0.38	1.00
420	4	18	0.00	0.44	0.24	1.00
463	4	17	0.00	0.09	0.15	1.00

See notes to table 3.

Table 11: Evaluation of Individual Respondents' Inflation Histograms: Forecast Sample begins in 1992:Q1

			Evaluation based on z^*			Comparison (Eqn. 5)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
411	1	17	0.52	0.23	0.80	0.83
446	1	18	0.23	-0.16	0.56	0.89
426	1	20	0.15	-0.27	1.57	0.90
420	1	16	0.12	-0.20	1.83	0.98
433	1	17	0.12	0.13	1.85	0.96
484	1	17	0.11	-0.19	0.47	0.95
421	1	17	0.02	-0.36	0.41	0.98
431	1	18	0.00	-0.10	0.22	0.63
431	2	16	0.37	-0.35	1.00	0.97
433	2	17	0.25	-0.32	0.69	0.97
446	2	20	0.06	-0.33	0.52	1.00
426	2	18	0.02	-0.63	0.74	1.00
463	2	18	0.02	-0.38	0.41	1.00
421	2	18	0.00	-1.11	1.79	0.98
426	3	20	0.01	-0.65	0.68	1.00
420	3	20	0.00	-0.44	0.35	1.00
407	3	16	0.00	-0.86	0.60	1.00
446	3	18	0.00	-0.36	0.24	1.00
433	3	18	0.00	-0.30	0.22	1.00
84	4	17	0.31	-0.37	1.07	1.00
411	4	16	0.30	-0.13	0.57	1.00
431	4	16	0.23	-0.42	1.09	1.00
463	4	17	0.02	-0.28	0.36	1.00
426	4	18	0.01	-0.44	0.40	1.00
433	4	16	0.01	-0.69	1.58	1.00
446	4	18	0.00	-0.47	0.34	1.00
421	4	18	0.00	-0.81	0.81	1.00
407	4	16	0.00	-0.75	1.77	1.00
420	4	18	0.00	-0.51	0.23	1.00

See notes to table 3.

Table 12: Aggregate Histograms of Output Growth: SPF & Benchmark, SPF Centred on Median Point Prediction

				Evaluation based on z^*				Comparison (Eqn. 5)
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	
1	SPF	0.24	0.27	0.29	0.23	0.24	0.72	0.87
	BM	0.27	0.19	0.17	0.24	0.21	0.64	
2	SPF	0.71	0.01	0.01	0.01	0.07	0.37	0.03
	BM	0.80	0.08	0.04	-0.01	0.05	0.46	
3	SPF	0.35	0.00	0.00	0.10	0.18	0.31	0.00
	BM	0.48	0.20	0.12	0.13	0.13	0.58	
4	SPF	0.51	0.00	0.00	0.08	0.12	0.08	0.00
	BM	0.88	0.01	0.00	0.14	0.03	0.35	

See notes to table 1.

Table 13: Aggregate Histograms of Inflation:SPF & Benchmark, SPF Centred on Median Point Prediction

			Evaluation based on z^*					Comparison (Eqn. 5)
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	
1	SPF	0.03	0.00	0.00	0.03	0.40	0.25	0.05
	BM	0.05	0.25	0.82	0.07	0.37	0.90	
2	SPF	0.13	0.00	0.00	-0.09	0.27	0.11	0.00
	BM	0.27	0.09	0.08	-0.19	0.20	0.56	
3	SPF	0.23	0.00	0.00	-0.13	-0.20	0.05	0.00
	BM	0.12	0.01	0.01	-0.40	-0.27	0.42	
4	SPF	0.15	0.00	0.00	-0.06	0.26	0.02	0.00
	BM	0.19	0.01	0.01	-0.28	0.24	0.49	

See notes to table 1.