Seventh ECB Workshop on Forecasting Techniques

The Measurement and Characteristics of Professional Forecasters' Uncertainty

Kenneth F. Wallis

Emeritus Professor of Econometrics, University of Warwick
http://go.warwick.ac.uk/kfwallis

Joint work with Gianna Boero and Jeremy Smith

Good morning!

- This paper is about the construction and interpretation of measures of uncertainty from forecast surveys that ask for density forecasts as well as point forecasts.
- We consider several statistical issues that arise, with application to the Bank of England Survey of External Forecasters.
- Using our preferred measure, mostly based on fitted normal distributions, we find
    - substantial heterogeneity of individual forecast uncertainty, and
    - strong persistence in individuals' relative level of uncertainty.
- This latter finding is new. It is similar to the individual optimism and pessimism already established in other point forecast surveys, and which we replicate in the present dataset.
- The main findings are replicated with data from the European Central Bank's Survey of Professional Forecasters.

Contents

The Bank of England Survey of External Forecasters

- the survey covers a sample of City firms, academic institutions and private consultancies, predominantly based in London. Their identities are not known to us.

- a summary of aggregate survey results is published in each Bank of England *Inflation Report* (Feb, May, Aug, Nov). Each chart or table notes the number of responses on which it is based; this is typically in the low twenties, and varies over time, by variable, and by forecast horizon, due to both ***complete*** and ***item non-response***.

- we consider the forecasts of CPI inflation and GDP growth one-, two- and three-years-ahead (h=5, 9 and 13 quarters) provided by individual respondents over 23 surveys from May 2006.

- May 2006 saw two previous fixed-target questions (last-quarter this year, last-quarter next year) replaced by fixed-horizon questions (one-year-ahead, three-years-ahead). The recent period has also seen more action in the data, compared with the preceding ***n***on-***i***nflationary ***c***onsistently ***e***xpansionary – ***'nice'*** – period studied in our previous articles.

- as in other survey research, our statistical analysis of individual forecaster performance is based on a subsample of ***regular respondents***. In the present exercise these are the 17 respondents whose item response rate over this period exceeds two-thirds.

- a recent CPI inflation question, and the published survey mean point forecasts, follow.

Figure 1. Bank of England questionnaire, November 2010 survey, inflation question

PROBABILITY DISTRIBUTION OF 12-MONTH CPI INFLATION OVER THE MEDIUM TERM

Please indicate the percentage probabilities you would attach to the various possible outcomes in 2011 Q4, 2012 Q4 and 2013 Q4. The probabilities of these alternative forecasts should of course add up to 100, as indicated.

PROBABILITY OF 12-MONTH CPI INFLATION FALLING IN THE FOLLOWING RANGES			
	2011 Q4	2012 Q4	2013 Q4
<0%			
0.0% to 1.0%			
1.0% to 1.5%			
1.5% to 2.0%			
2.0% to 2.5%			
2.5% to 3.0%			
> 3.0%			
TOTAL	100	100	100

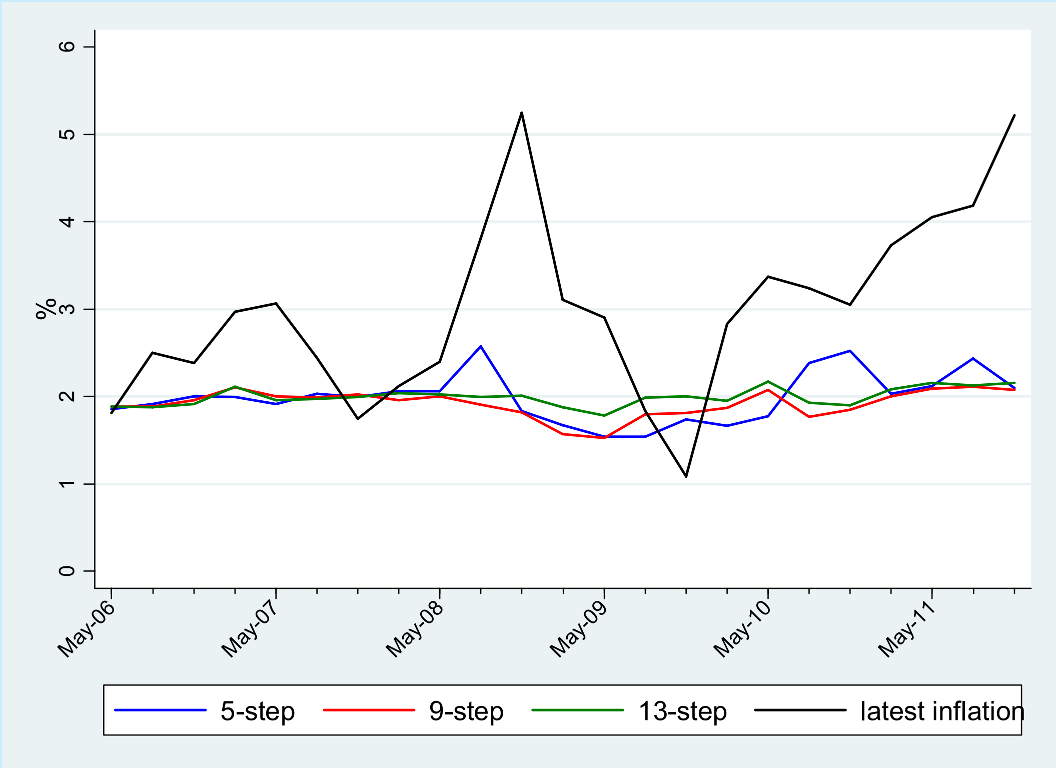
CENTRAL PROJECTION FOR 12-MONTH CPI INFLATION		
2011 Q4	2012 Q4	2013 Q4

Four half-percentage-point bins cover the range 1-3%.

The previous lower open bin (<1%) was divided as shown in February 2009.

Figure 2. Mean point forecasts, and latest (monthly) data available to forecasters

Upper panel: CPI inflation; (lower panel: GDP growth)

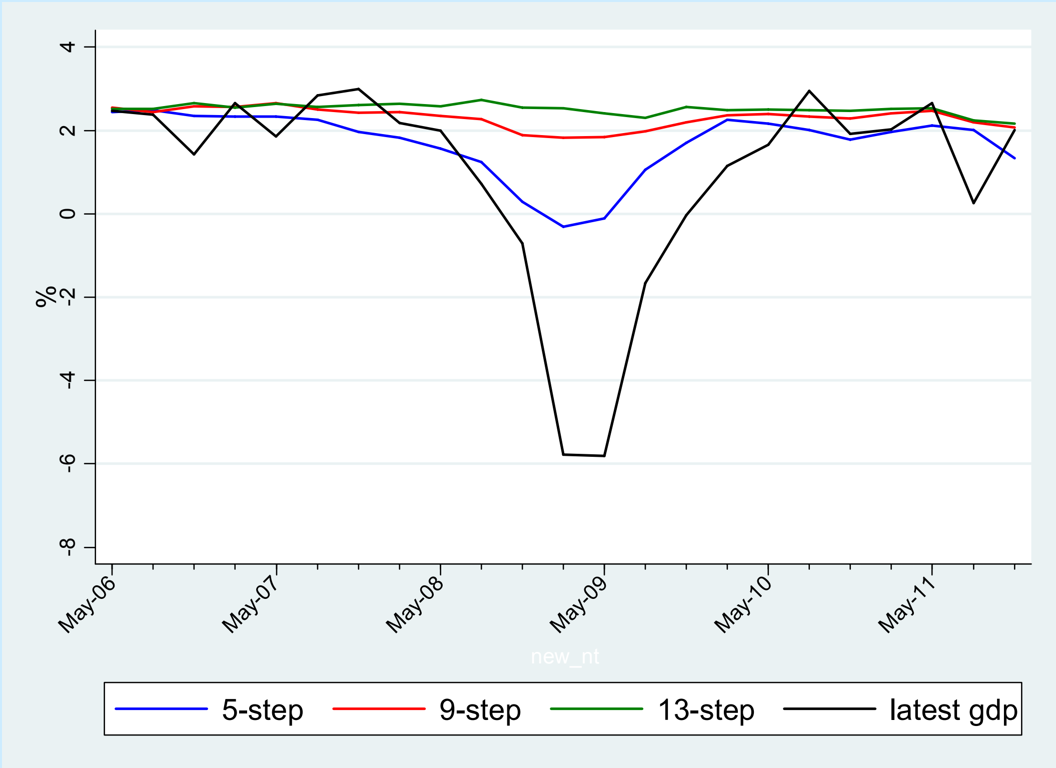


The survey mean point forecasts show little reaction to current conditions.

At longer horizons, the mean forecast stays close to the official CPI inflation target of 2%.

Figure 2. Mean point forecasts, and latest (monthly) data available to forecasters

(Upper panel: CPI inflation;) **lower panel: GDP growth**



At longer horizons, the mean forecast stays close to the trend growth rate of 2.5%.

The mean forecasts conceal much disagreement between individual point forecasts: see below.

2. Measuring uncertainty: a framework for analysis

2.1. Elicitation and probability distributions

There are some parallels between the process of elicitation of an expert's probability distribution about an uncertain quantity, via a facilitator, and the fitting of probability distributions to forecast survey responses. The following observations are relevant to both activities.

- Any probability distribution chosen to represent expert beliefs or to summarise a histogram density forecast expresses uncertainty about the variable in more detail than has been provided, and should not be interpreted as a perfect representation of expert or forecaster uncertainty.

- Reported probabilities are imprecise – there is uncertain uncertainty – but 'there cannot be a fully probabilistic solution to the problem of imprecision in probability assessments, as the notion of an imprecise probability itself is in violation of an axiom of subjective probability' (O'Hagan *et al*, 2006, p.160).

- The elicitation process and the reporting of a histogram density forecast are not processes of sampling from a population, and do not support the use of classical hypothesis tests of the goodness-of-fit of the chosen distribution.

2.2. The use of probability distribution functions

We use fitted distributions to facilitate the estimation of moments, as an alternative to the traditional method of estimating moments of distributions specified in histogram form.
This applies standard formulae for discrete distributions, using the midpoints of the histogram bins (treating the open bins as closed bins with the same width as the interior intervals).
Thus the distribution is treated as if all probability mass is located at the interval midpoints.

However the underlying variable is continuous, and typical macroeconomic density forecasts are unimodal, suggesting that more of the probability in each bin is located closer to the centre of the distribution, which motivates estimation of moments by fitting

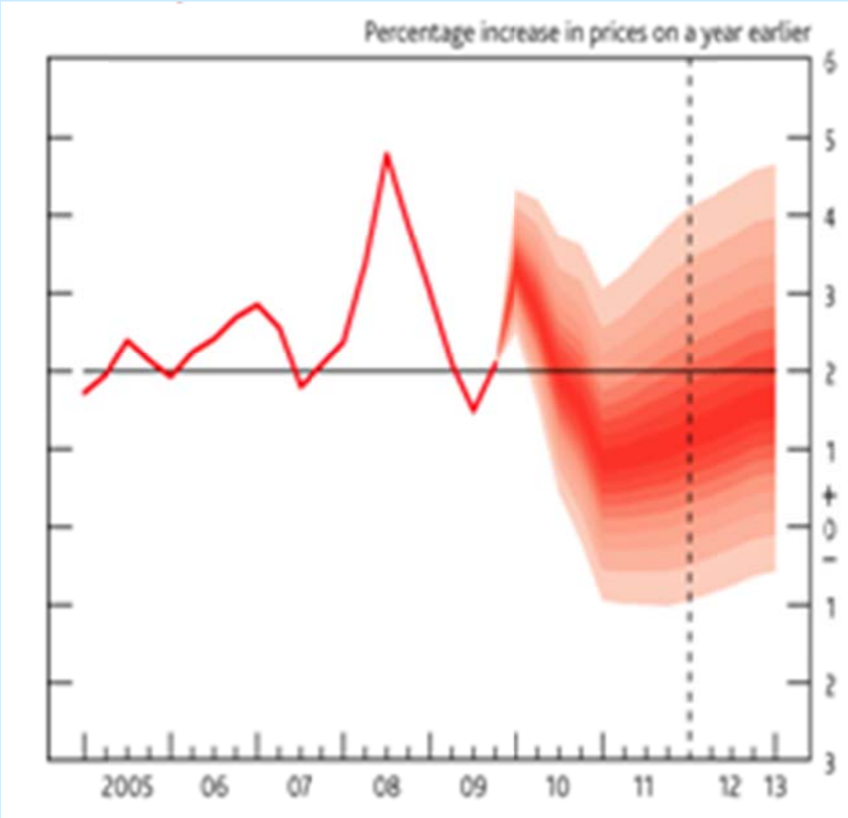
- normal distributions (Giordani and Soderlind, 2003; Lahiri and Liu, 2006, …)
- generalised beta distributions with $p>1$, $q>1$ (Engelberg *et al*., 2009; Clements, 2012)

To fit these distributions, non-zero probabilities in at least three bins are needed; in two-bin cases Engelberg *et al*. fit symmetric triangular distributions.

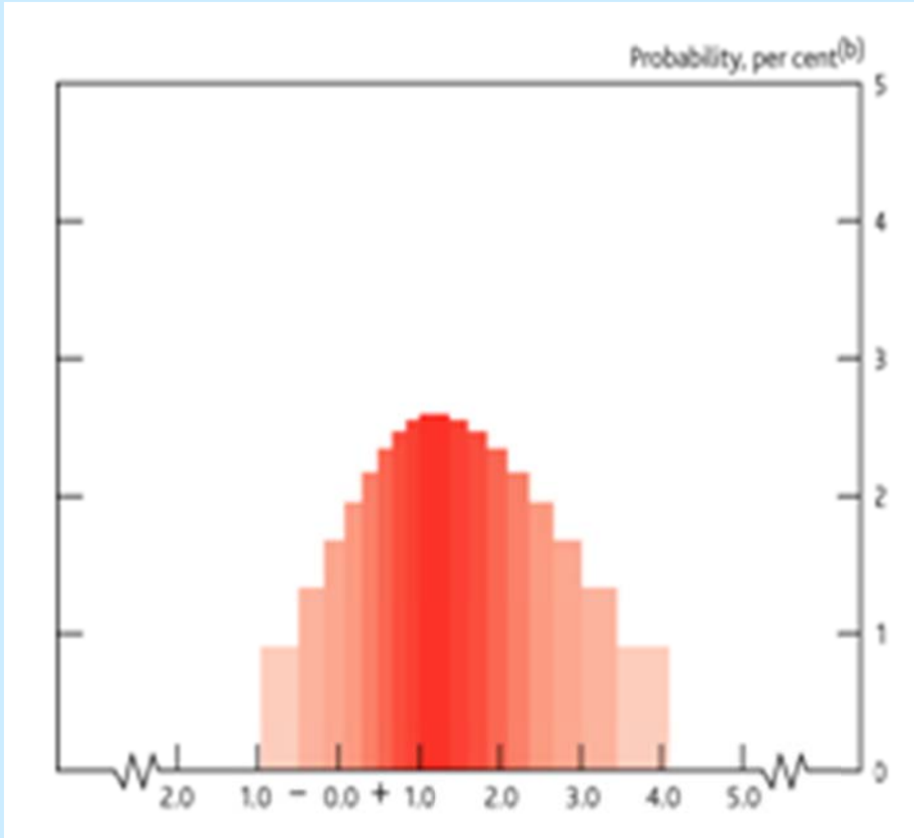
Asymmetric distributions based on the normal distribution are also in use; see central bank ‘fan chart’ forecasts, for example.

Figures 3, 4. Bank of England MPC inflation forecast, February 2010

The fan chart



Two-years-ahead cross-section



The 'facilitators' use the two-piece normal distribution to calculate the histogram intervals, but the histogram is all that is agreed by the MPC (the 'experts'):

'*the distribution … of tail events is not explicitly specified, as to do so would require a spurious degree of precision on the part of the MPC*' (King, 2010).

2.3. Uncertain uncertainty

Beware interpreting any elicited distribution as a perfect representation of expert uncertainty:

- we have insufficient probabilities to specify a unique distribution

  (but maybe the quantity of interest, here the variance, is robust to different specifications)
- it is difficult for experts to give precise numerical values for their probabilities.

It is sometimes assumed that reported probability = 'true' probability + random error

(such an assumption might support goodness-of-fit tests).

But this does not match survey data, in which uncertainty about subjective probabilities is demonstrated by widespread use of round numbers, varying across individuals, who have different patterns of rounding, to different extents.

Table 1. The numerical character of (almost 15,000) reported probabilities

Reported percent probability	Percentage of cases
Multiple of 10	35.2
Otherwise multiple of 5	30.5
Other integer	30.7
Non-integer	3.6

3. The properties of individual uncertainty measures

3.1. Choosing a measure

We compare the standard deviation estimates from normal and beta distributions fitted to the 1381 density forecasts of inflation that use three or more bins. In the great majority of cases the two estimates are very close.

The beta distribution has difficulties when histograms plotted in the usual way are U or J-shaped. This happens because the range of the central half-percentage-point bins is not wide enough. The beta distribution can match these shapes, with one or both parameter estimates <1, but our prior is a single interior mode. There are insufficient probabilities to estimate the support (L, U).

The normal distribution does not require closure of the open bins, and can handle these cases:

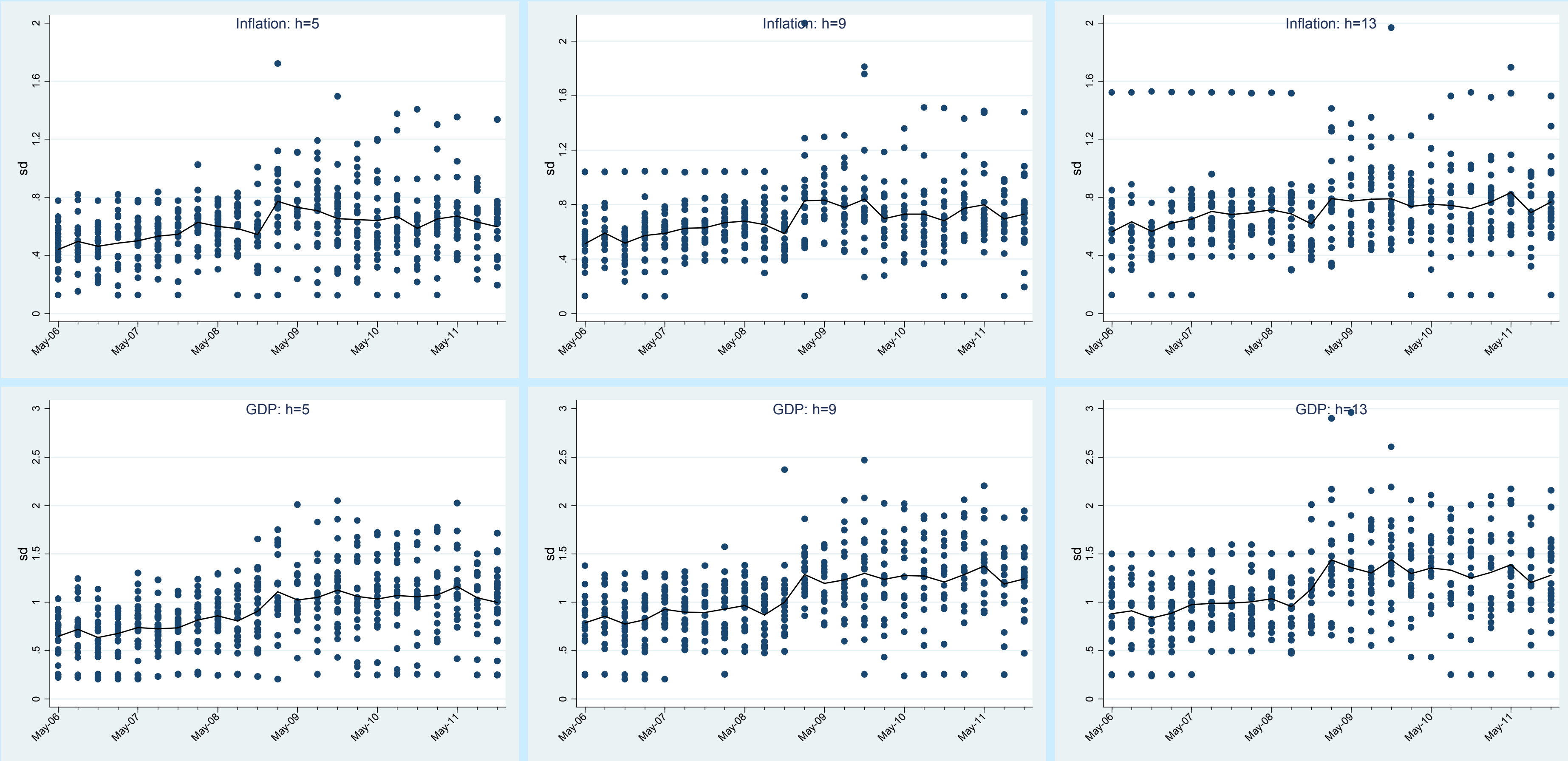
- extreme example: probabilities 5%, 5%, 90% in bins 2-2.5%, 2.5-3%, >3%.

GDP growth forecasts have one-percentage-point bins and more problem cases for the beta distribution (also more two-bin forecasts and seven with 100% in a single bin).

Conclusion: our preferred uncertainty measure is the sd of the fitted normal distribution in all cases with non-zero probability in at least three bins; otherwise we use symmetric triangular distributions as in Engelberg, Manski and Williams (2009). The results are shown in Figure 5.

Figure 5. Spread of individual uncertainty measures (sd), and their median

Upper panels: CPI inflation; lower panels: GDP growth; *h*=5, 9, 13



The median has a local peak in February 2009, in some cases signalling a shift in level; the overall dispersion also increases. Uncertainty increased as the crisis spread, and the MPC cut bank rate by 3 percentage points since the previous survey.

Lahiri and Liu (2006) find similar time-varying heterogeneity of uncertainty in the US SPF.

3.2. The persistence of individual relative uncertainty

Systematic study of the forecasts across individuals and over time is handicapped by missing data, so we now turn to the subsample of 17 regular respondents, who are missing no more than one-third of the total possible responses. The overall subsample item response rate is 87%.

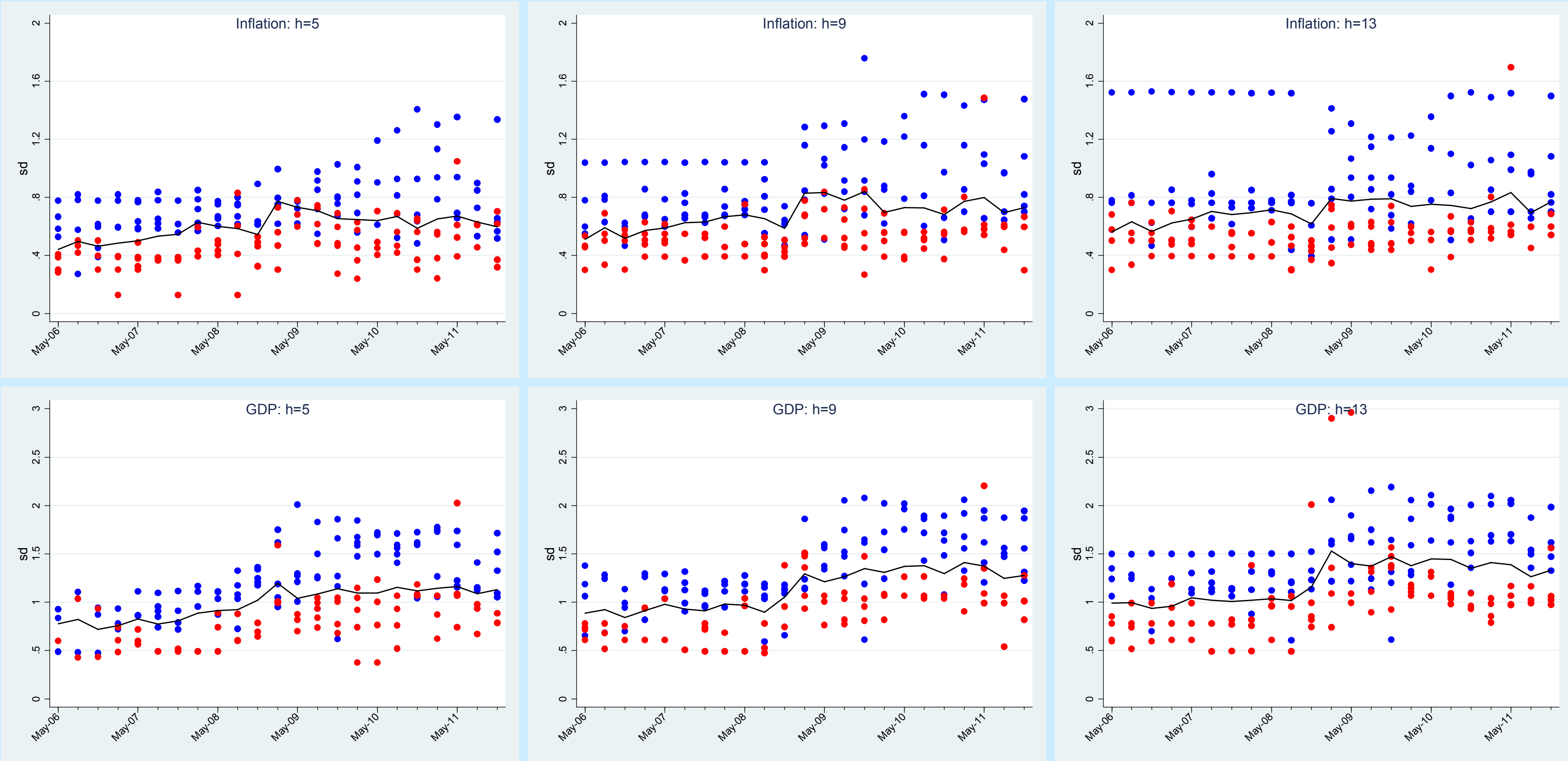
To study possible persistence in the relative position of individual forecasters' uncertainty, we rank the 17 (or fewer) individuals from the highest to the lowest uncertainty for each variable (2), horizon (3), and time period (23), that is, for each column of dots in Figure 5.

For each variable and horizon, that is, each panel of Figure 5, we calculate each forecaster's average rank over the (23 or fewer) surveys for which they are present.

As an illustration of possible persistence in Figure 5, in each panel of Figure 6 we pick out the uncertainty measures of the five individuals with the highest average rank and the five individuals with the lowest average rank in each case.

(Note that most of them have some forecasts missing in each panel of the figure, also that these are not exactly the same individuals across panels, a question we return to below.)

Figure 6. Uncertainty measures of the five highest-ranked (blue) and lowest-ranked (red) regular respondents in each panel



There is a very clear indication of persistence in relative forecast uncertainty, with blue dots tending to stay high and red dots tending to stay low.

*The Kendall coefficient of concordance**

As a statistical measure of persistence among the rankings of forecasters, we use the Kendall coefficient of concordance, denoted W, and defined as the ratio of the sum of squared mean deviations of the observed average ranks to its maximum possible value. Thus $0 \leq W \leq 1$.

With 17 regular respondents and no missing data, perfect agreement of the rankings across all 23 surveys would give average ranks 1, 2,…,17 in some order, with sum of squared mean deviations equal to 408, which is the maximum possible value in this case. At the other extreme the average ranks all tend to equal 9 when rankings of individuals are purely random over time.

With missing observations, however, the maximum rank is less than 17. For each survey we rank the forecasts ignoring non-respondents, and individuals' average ranks are calculated over the occasions on which they responded. We calculate a revised maximum possible sum of squared mean deviations of average ranks conditional on the observed pattern of missing data in each case, and with the observed average ranks we obtain the results in the following table.

With random rankings and no missing data, the scaled statistic has an approximate chi-squared distribution: for 23 rankings of 17 items, the 99th percentile of W is 0.09.

*****Kendall, M.G. and Gibbons, J.D. (1990). *Rank Correlation Methods* (5th edn), Ch.6.

Table 2. Measures of agreement over time between forecasters' rankings with respect to their uncertainty measures: Kendall coefficients of concordance

	h=5	*h=9*	*h=13*
CPI inflation	0.40	0.50	0.47
GDP growth	0.44	0.40	0.47

These coefficients indicate considerable stability over time in the relative level of individual forecasters' uncertainty, to a similar extent for both variables and all three forecast horizons.

In order to pool these cases, we calculate the concordance between the six rankings given by the time-averaged scores in each case. The Kendall coefficient is 0.92; under the above null its 99^{th} percentile for 6 rankings of 17 individuals is 0.33. The rankings of individual forecasters by their uncertainty levels are almost identical across the two variables and three forecast horizons.

This strong evidence of persistence in individual forecasters' relative levels of uncertainty, as expressed in their subjective probabilities, is a new finding. We are not aware of models of forecaster behaviour that incorporate this feature.

3.3. *Ex ante and ex post measures of inflation uncertainty*

Published density forecasts of inflation in the UK date from February 1996, which saw the first appearances of the Bank of England's 'fan chart' and the National Institute's tabled histogram. The National Institute also published a density forecast of real GDP growth at this time, whereas a growth fan chart did not appear in the Bank's *Inflation Report* until November 1997.

In both institutions the variance of the density forecast was calibrated with reference to past point forecast errors, with judgmental adjustment. This has remained the case, although the reduction in inflation variance during the 1990s was not recognised quickly enough (Wallis, 2004; Mitchell, 2005; Bank of England, 2005). This prompts the question whether survey respondents behave in the same way or, at arm's length, whether our ex ante density forecast uncertainty measure is related, at the individual level, to past point forecast errors.

We measure ex post uncertainty by individual point forecast RMSE over the preceding four quarters. Since the quarterly series of one- and three-years-ahead forecasts began only in May 2006, the need to wait for four forecast outcomes makes the available time series samples too small. However the two-years-ahead forecast questions were in the earlier questionnaire, so we have sufficient past data to utilise the full sample, with T=23, subject to missing observations.

We consider only the inflation forecasts, since data revisions mess up GDP forecast evaluation.

Ex ante and ex post measures of forecast uncertainty, results

Joint estimation of a regression of density forecast standard deviation on point forecast RMSE for 17 regular respondents, allowing intercept and slope to vary across individuals, yields strong rejections of the equality of intercepts and of the equality of slope coefficients.

Inequality of intercepts is in agreement with the persistence in relative uncertainty seen above. Of the individual slope coefficients, eight are positive and significantly different from zero, and eight are positive but not significantly different from zero. In the final case there is a significant negative coefficient, which is clearly due to the presence of two distinct subsamples in the data.

Again there is evidence of different forecasters following different practices: some do appear to calibrate their density forecasts with reference to past point forecast errors; some do not, at least with the measure we have chosen.

If we neglect this heterogeneity and impose equality of the individual coefficients on point forecast RMSE, then the resulting common coefficient is significantly different from zero, in contrast to the results of Lahiri and Liu (2006) with US SPF data, also those of Rich and Tracy (2003) that they cite. However Rich and Tracy were working with survey average data, while Lahiri and Liu used the level and absolute value of the immediate past forecast error rather than a forecast RMSE measure.

4. Persistence in the relative level of individual point forecasts

In view of the findings of persistent individual biases towards optimism or pessimism in *Consensus Economics* point forecasts by Batchelor (2007), see also Patton and Timmermann (2010), we apply the ranking methods developed above to the point forecasts in the Bank of England survey. The complete data are shown in Figure 7.