Quadratic Scoring Rules and Density Forecast Histograms

Gianna Boero, Jeremy Smith and Kenneth F. Wallis*

Department of Economics
University of Warwick
Coventry CV4 7AL, UK

JBES submission, 30 March 2009

Abstract This paper provides a practical evaluation of some leading density forecast scoring rules in the context of forecast surveys. We analyse the forecasts of UK inflation obtained from the Bank of England's Survey of External Forecasters, considering both the survey average forecasts published in the quarterly *Inflation Report*, and the individual survey responses recently made available by the Bank. The density forecasts are collected in histogram format, and the ranked probability score (RPS) is seen to have clear advantages. Missing observations are a feature of forecast surveys, and we propose an adjustment to the RPS, based on the Yates decomposition, to improve its comparative measurement of forecaster performance in the face of differential non-response. As an alternative combined forecast, a trimmed mean scores better than the published survey average.

Keywords Density forecast evaluation; Brier (quadratic probability) score; Epstein (ranked probability) score; Bank of England Survey of External Forecasters; missing data; forecast comparison; forecast combination

Acknowledgments We are grateful to David Bessler and Gabriel Casillas-Olvera for supplying their code, Valentina Corradi and Norman Swanson for helpful suggestions, and Bank of England staff for assembling and helping to clean the survey dataset. Readers wishing to gain access to the data should write to the Publications Editor, Inflation Report and Bulletin Division, Bank of England, Threadneedle Street, London EC2R 8AH, UK.

***Corresponding author** email K.F.Wallis@warwick.ac.uk; tel +44 24 7652 3026; fax +44 24 7652 3032

1. Introduction

Probability forecasting is the process of attaching a numerical probability to an uncertain future event, and scoring rules measure the quality of probability forecasts by a numerical score based on the forecast and the eventual outcome. The earliest example of such a rule, introduced by Brier (1950) and subsequently bearing his name, concerns the situation in which, on each of a series of occasions, an event can occur in only one of a small number of mutually exclusive and exhaustive categories, and a forecast consists of a set of probabilities, one for each category, that the event will occur in that category. The Brier score is then given as the mean over occasions of the squared differences between the forecast probabilities and an indicator variable that takes the value 1 in the category in which the event occurred and 0 in all other categories. Much of the theoretical work underpinning probability forecast construction and evaluation originally appeared in the meteorological literature, and the example in Brier's article concerned the verification of probability forecasts of rain or no-rain in given periods, which has only two categories and is sometimes called an event probability forecasting problem. The mathematical formulation adopted by Brier has also resulted in the use of the name "quadratic probability score" (QPS), which is used below, although it is potentially misleading, because a family of quadratic scoring rules exists, of which the Brier score is just one member (Stael von Holstein and Murphy, 1978).

In many forecasting applications the focus of attention is the future value of a continuous random variable, and the presentation of a density forecast or predictive distribution – an estimate of the probability distribution of the possible future values of the variable – is becoming increasingly common. Tay and Wallis (2000) survey applications in macroeconomics and finance, and more than half of the inflation targeting central banks, worldwide, now present density forecasts of inflation in the form of a fan chart. The best-known series of density forecasts in macroeconomics dates from 1968, when the American Statistical Association and the National Bureau of Economic Research jointly initiated a quarterly survey of macroeconomic forecasters in the United States, known as the ASA-NBER survey; Zarnowitz (1969) describes its original objectives. In 1990 the Federal Reserve Bank of Philadelphia assumed responsibility for the survey and changed its name to the Survey of Professional Forecasters (SPF). Survey respondents are asked not only to report their point forecasts of several variables but also to attach a probability to each of a number of pre-assigned intervals, or bins, into which output growth and inflation, this year and next year, might fall. In this way, respondents provide density forecasts of these two variables, in the form of histograms. The probabilities are then averaged over respondents to obtain survey average density forecasts, again in the form of histograms, which are published. More recently the Bank of England (since 1996) and the European Central Bank (since 1999) have conducted similar surveys with similar density forecast questions, and they also share the practice of the SPF in making the individual responses to the survey, suitably anonymised, available for research purposes. The empirical applications in the present paper extend the explorations of the Bank of England Survey of External Forecasters (SEF) dataset by Boero, Smith and Wallis (2008a,b).

The Brier score and its covariance decomposition (Yates, 1982, 1988), discussed below, are used by Casillas-Olvera and Bessler (2006) in a comparative evaluation of the published survey average density forecasts from the SEF and the density forecasts of the Bank of England's Monetary Policy Committee (MPC). However, the Brier score's set-up in terms of distinct classes or categories, in relation to a set of histogram bins, neglects the ranking or ordering of the bins in terms of the values of the underlying continuous variable. For four-bin histograms where the outcome falls in the bin that has been assigned probability 0.3 and the other bins have probability 0.5, 0.1 and 0.1, for example, the Brier score is indifferent to the location of these last three probabilities, but forecasts that placed 0.5 in a bin adjacent to the bin in which the outcome fell would generally be regarded as better forecasts than those that did not. The Ranked Probability Score introduced by Epstein (1969), a further member of the class of quadratic scoring rules, takes account of the ordering of the categories, but appears not to have been previously used in the evaluation of density forecasts expressed as histograms. On the other hand, its extension to continuous distributions, the continuous ranked probability score (CRPS), has recently attracted attention in the meteorological literature (Gneiting and Raftery, 2007).

Gneiting and Raftery's (2007) review of scoring rules, their characterisations and properties, includes a leading alternative to the quadratic scores, namely the logarithmic score. Originally proposed by Good (1952), this is defined as

$$\log \mathrm{S}\left(f, x_t\right) = \log f\left(x_t\right)$$

for density forecast *f* of the random variable X_t evaluated at the outcome x_t. The logarithmic score has many attractive features, but it is inappropriate for our present dataset of density forecasts reported as histograms. From time to time in the individual survey responses it happens that the outcome falls in a histogram bin to which the respondent has assigned zero probability, so that the log score is undefined. To assign an arbitrarily large value to the score on such occasions is an unsatisfactory solution, since the ranking of competing forecasts is sensitive to the chosen value. On the other hand zero-probability forecast outcomes are readily accommodated by the quadratic scores.

In this paper we compare and contrast the Brier and Epstein rules, or QPS and RPS, in applications to survey density forecasts of UK inflation. Section 2 contains a formal presentation of the rules and discussion of the decompositions that have been proposed. Section 3 extends the study of Casillas-Olvera and Bessler (2006) by considering both of the rules and a longer sample period. Section 4 turns to the individual respondents to the Bank of England's survey and again uses both rules to evaluate their forecast performance: it is seen that the RPS is preferred. Incomplete data are a feature of this survey, like all forecast surveys, and an adjusted score, RPS*, is proposed, to standardise comparisons in the face of missing observations caused by differential non-response. An alternative combined forecast to the published survey average density forecast is also considered. Section 5 concludes.

2. Quadratic Scoring Rules

2.1. The Brier and Epstein rules

We consider a categorical variable whose sample space consists of a finite number *K* of mutually exclusive events, and for which a probability forecast of the outcome at time *t* is a vector of probabilities $\left(p_{1t},...,p_{Kt}\right)$. We have in mind applications in which the categories are the *K* bins of a histogram of a continuous random variable *X*, and we define indicator variables d_{kt}, $k=1,...,K$, which take the value 1 if the outcome x_t falls in bin *k*, otherwise $d_{kt}=0$. Also in mind are time series forecasting applications, in which each forecast of the outcome at times $t=1,...,T$ is formed at some previous time. For a sample of forecasts and realisations of the categorical variable, the Brier score is given as

$$\text{QPS}=\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(p_{kt}-d_{kt}\right)^{2}. \tag{1}$$

The range is usually stated as $0 \le \text{QPS} \le 2$, although the extreme values are obtained in extreme circumstances in which, in every period, all the probability is assigned to a single bin and the outcome either does or does not fall into it. More generally, there is a non-zero lower bound that corresponds to a best fit. If the bin probabilities are constant over time, $p_{kt} = p_{ks} = p_k$, say, $t \ne s$, $k = 1,\ldots,K$, this is obtained for any forecast sequence in which the relative bin frequencies

$$\bar{d}_k = \frac{1}{T}\sum_{t=1}^{T} d_{kt},\; k = 1,\ldots,K$$

match the probabilities p_k, whereupon the score achieves its minimum value

$$\text{QPS}_{\min} = 1 - \sum_{k=1}^{K} p_k^2 \;\; .$$

Note that this is indifferent to the ordering of the time series of observations.

The Brier score is also indifferent to the fact that, in the histogram context, there is a natural ordering of the categories, or an implicit measure of the distance between them, which should be taken into account, as noted above. To do this, Epstein's (1969) proposal replaces the density functions implicit in the Brier score with their corresponding distribution functions (Murphy, 1971). Defining these as

$$P_{kt} = \sum_{j=1}^{k} p_{jt}, \quad D_{kt} = \sum_{j=1}^{k} d_{jt}, \quad k = 1,\ldots,K \,,$$

with $P_{Kt} = D_{Kt} = 1$, the ranked probability score is

$$\text{RPS} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(P_{kt} - D_{kt}\right)^2 . \qquad (2)$$

The RPS penalises forecasts less severely when their probabilities are close to the actual outcome, and more severely when their probabilities are further from the actual outcome. Like the Brier score, its minimum value is 0, occurring in the same extreme circumstance of the outcomes falling in bins whose forecast probability is 1. Similarly, the maximum value of the RPS occurs when some $p_{kt} = 1$ and the outcome falls in a different bin, but the actual value depends on how far from the *k*th bin that is. In extremis, with the outcomes and the unit-probability bins being located at opposite ends of the range, this value is $K - 1$.

2.2. Decompositions of the scores

Several decompositions or partitions of the Brier score and, by extension, the Epstein score have been proposed, with the aim of obtaining information about different aspects of forecast performance. Early contributions focused on the event probability forecasting problem and used a simplified version of the Brier score given in equation (1), which we denote QPSE, namely

$$\text{QPSE} = \frac{1}{T}\sum_{t=1}^{T}\left(p_t - d_t\right)^2 . \tag{3}$$

Here p_t is the forecast probability, and $d_t = 1$ if the event occurs or zero if it does not. The QPSE score is equal to half of the value obtained from equation (1) with $K = 2$, since it neglects the complementary non-occurrence of the event, whose forecast probability is $1 - p_t$.

Sanders (1963) requires that all probabilities be expressed in tenths and partitions the *T* forecasts into eleven subsets of size T_j, say, in which the forecast probability is $p_j = j/10,\ j = 0,\dots,10$. To consider QPSE subset-by-subset we rearrange the summation in equation (3) as

$$\text{QPSE} = \frac{1}{T}\sum_{j=0}^{10}\sum_{t\in T_j}\left(p_j - d_{jt}\right)^2 .$$

Expanding the terms in the inner summation gives

$$\sum_{t\in T_j}\left(p_j - d_{jt}\right)^2 = T_j\left(p_j - \bar{d}_j\right)^2 + \sum_{t\in T_j}\left(d_{jt} - \bar{d}_j\right)^2 = T_j\left[\left(p_j - \bar{d}_j\right)^2 + \bar{d}_j\left(1 - \bar{d}_j\right)\right],$$

where $\bar{d}_j$ is the relative frequency of occurrence of the event over the T_j occasions on which the forecast probability is p_j. Summing the first term on the right-hand side over *j* and dividing by *T* gives the component of QPSE that measures what is variously called validity, reliability or calibration. A plot of $\bar{d}_j$ against p_j is called a reliability diagram or calibration curve: for a "well-calibrated" forecaster this is close to a diagonal line. The sum over *j* of the second term on the right-hand side, divided by *T*, involves only the outcome indicators but nevertheless reflects forecaster behaviour, because the indicators are sorted into classes according to the forecaster's probabilities. Sanders (1963) refers to this term as a measure of the "sharpness" of the forecasts, using a term introduced by Bross (1953, Ch.3); "resolution" and “refinement” are also in use. Its maximum value is obtained when each $\bar{d}_j$ is 0.5, that is, the forecaster’s probabilities have not succeeded in discriminating high-probability and low-probability occurrences of the event, and sharpness is lacking.

The second term in Sanders’ decomposition can be further partitioned as

$$\frac{1}{T}\sum_{j=0}^{10} T_j \bar{d}_j\left(1-\bar{d}_j\right)=\bar{d}\left(1-\bar{d}\right)-\frac{1}{T}\sum_{j=0}^{10} T_j\left(\bar{d}_j-\bar{d}\right)^2,$$

where $\bar{d}$ is the overall rate of occurrence of the event (Murphy, 1973). This separates out the variance or uncertainty of the indicator variable, $\bar{d}\left(1-\bar{d}\right)$, which depends only on nature’s determination of the occurrence or otherwise of the event. Murphy argues that the remainder can then more appropriately be called resolution, since it measures the degree to which the relative frequencies for the 11 subcollections of forecasts differ from the overall relative frequency of occurrence of the event: high resolution improves (lowers) the QPS.

This three-component decomposition is used in a study of the Bank of England Monetary Policy Committee’s density forecasts of inflation and growth by Galbraith and van Norden (2008). An event probability forecast is derived from a published density forecast by calculating the forecast probability that the variable in question exceeds a given threshold. The resulting probabilities take continuous values, rather than the discrete values assumed in the preceding derivations, and one could simply group the probabilities into bins. Instead, Galbraith and van Norden use a kernel estimator to obtain a smoothed calibration curve.

Calculating above-threshold and below-threshold probabilities from a density forecast in effect reduces the MPC’s density forecast, which has the two-piece normal functional form, to a two-bin histogram. The Bank’s forecast survey questionnaire most often specifies a six-bin histogram, and generalisations of these decompositions of the QPS for $K > 2$ are available in the literature. However, they depend on similar discretisation and grouping of the forecasts to that used in the above derivations, although with six categories and probabilities stated in tenths (or similarly rounded) the number of possible forecasts is 3003, from Murphy’s (1972) equation (1). Many of these possible configurations are of little practical relevance to the SEF individual dataset, where the forecast histograms are almost invariably unimodal, although the tail probabilities in the first and/or last open-ended bins are sometimes sufficiently large to give the impression of an additional local peak. Nevertheless the number of distinct configurations observed in the SEF histograms analysed in Section 4 is typically close to the time series sample size, and a decomposition of the individual scores into reasonable estimates of forecast reliability and resolution is not practicable.

A decomposition of the QPS which does not require such a grouping of forecasts into distinct subcollections is the covariance decomposition due to Yates (1982, 1988), obtained as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(p_{kt}-d_{kt}\right)^2=\sum_{k=1}^{K}\frac{1}{T}\sum_{t=1}^{T}\left[(p_{kt}-\overline{p}_k)-(d_{kt}-\overline{d}_k)+(\overline{p}_k-\overline{d}_k)\right]^2$$

$$=\sum_{k=1}^{K}\left[\operatorname{var}\left(p_k\right)+\operatorname{var}\left(d_k\right)+\left(\overline{p}_k-\overline{d}_k\right)^2-2\operatorname{cov}\left(p_k,d_k\right)\right]. \qquad (4)$$

Yates (1988) notes that the second term in this last expression, the sum of the outcome indicator sample variances $\operatorname{var}(d_k)=\overline{d}_k(1-\overline{d}_k)$, is outside the forecaster's influence, while the third term, the sum of squared "biases", indicates the miscalibration of the forecasts. He offers further algebraic rearrangement of the first and fourth terms, as in the initial event-probability derivation with $K=2$ (Yates, 1982), although their interpretations do not readily generalise to the case $K>2$.

The Yates decomposition is reported by Casillas-Olvera and Bessler (2006) in their comparative study of the MPC and SEF survey average density forecasts, noted above and extended in the next section. The contribution of the variance of d to the total QPS varies over subperiods, but is the same for the two forecasts under consideration, as indicated by the above derivation. When working with the forecasts supplied by individual respondents to the survey, however, we face the familiar problem of individual non-response, which differs across individuals, so that the data have the form of an unbalanced panel. Thus the individual scores are calculated over different subsamples of the maximum possible T observations, and it is no longer the case that the contribution of the variance of d is the same for all individual forecasters. Since this term remains outside the forecasters' influence, to evaluate comparative forecast performance we propose an adjusted score, denoted QPS*, obtained by replacing the individual subsample outcome variance component of the QPS by the full-sample outcome variance.

We are not aware of a comparable covariance decomposition of the RPS, although the (mostly meteorological) literature contains much discussion of extensions of the previous reliability-resolution-uncertainty decomposition to the RPS and its continuous generalisation. Nevertheless it is clear that the derivation in equation (4) applies equally well to the RPS given in equation (2), on replacing lower-case p and d by upper-case P and D. As a result, a similar variance of D term can be identified that is a function of the outcomes alone. For comparing forecast performance in the face of differential non-response we again propose an adjusted score RPS*, obtained by replacing the individual-specific measure of outcome variance in the RPS by its full-sample equivalent.

3. The SEF average and MPC density forecasts of inflation: QPS and RPS

In this section we extend Casillas-Olvera and Bessler's (2006) comparative evaluation of the average density forecasts of inflation, two years ahead, from the Survey of External Forecasters, and the Monetary Policy Committee's fan chart forecasts of inflation for the same horizon. We increase their time series sample size, of 14 quarterly forecasts, to 36, and we consider the RPS as well as the QPS. Both forecasts are published in the Bank of England's quarterly *Inflation Report*, although to obtain numerical values of the parameters of the two-piece normal distribution on which the MPC's fan charts are based, it is necessary to consult the Bank's spreadsheets.

The Bank of England's quarterly Survey of External Forecasters began in 1996. The institutions covered in the survey include City firms, academic institutions and private consultancies, and are predominantly based in London. The sample changes from time to time as old respondents leave or new survey members are included, and not every institution responds every quarter, nor answers every question. Although there is no record of the response rate, the publication of summary results in the *Inflation Report* always includes the number of responses on which each reported statistic is based; typically this is in the low twenties.

Initially the SEF questionnaire asked for forecasts of inflation in the last quarter of the current and following years. Such questions eventually deliver sequences of fixed-event forecasts, but not quarterly series of fixed-horizon forecasts. However, in 1998 a third question was added, asking for forecasts two years ahead, and this marks the start of the series analysed below. (At this time a second variable, GDP growth, was also added.) In May 2006 all three questions were switched to a fixed-horizon format, focusing on the corresponding quarter one, two and three years ahead. In the UK's inflation targeting policy regime, the Government chooses the targeted measure of inflation and its target value, and the SEF has sought forecasts of the same variable, namely the Retail Prices Index excluding mortgage interest payments (RPIX) until the end of 2003, then the Consumer Prices Index (CPI). Thus forecasts collected in the eight quarters through 2002-3 have to be evaluated against outcomes in 2004-5 for the previous target variable, not the then-current target variable. At the time of writing, outcome data are available to the end of 2008, hence we use the surveys from 1998Q1 to 2006Q4, a total of 36. The histograms in the first five of these surveys have four bins (<1.5, 1.5-2.5, 2.5-3.5, >3.5), then the two interior bins were further divided and from 1999Q2 there are six bins (<1.5, 1.5-2, 2-2.5, 2.5-3, 3-3.5, >3.5); in 2004Q1 the whole grid was shifted downwards by 0.5, following the change in the target from 2.5% RPIX inflation to 2% CPI inflation. For comparative purposes we convert the MPC's fan chart forecasts at the two-year horizon to sets of probabilities for the same bins, using the MPC's parameterisation of the two-piece normal distribution (Wallis, 2004, Box A).

Table 1. Scores of SEF average and MPC density forecasts of inflation

	SEF	MPC
QPS	0.711	0.759
RPS	0.566	0.596

Note: $T = 36$; forecasts of inflation two years ahead published 1998Q1-2006Q4.

The scores of the two forecasts are shown in Table 1. It is seen that the survey average forecast has a smaller QPS than the MPC forecast, strengthening Casillas-Olvera and Bessler's finding for the first 14 of these 36 quarterly observations. The RPS gives the same ranking of the two forecasts, although the two RPS values are slightly closer together than the two forecasts' QPS values. The RPS values are smaller than the QPS values, since the forecast densities are unimodal and, most of the time, the outcomes fell towards the centre of these distributions: the positioning of relatively high probabilities close to the bin in which the outcome fell is acknowledged by the RPS, but not by the QPS.

To study the comparative behaviour of the two scores in greater detail we turn to Figure 1, which illustrates observation-by-observation the components of the calculation, namely the histogram probabilities and the location of the inflation outcome, for the SEF average forecast and MPC forecast in the upper and lower panels respectively. The coloured segments of the vertical columns show, with reference to the left-hand scale, the allocation of forecast percentage probabilities to the histogram bins. For most of the period there are six bins, and the colours follow a rainbow array. The key to the figure records the RPIX inflation range for each bin; from 2004Q1 all these numbers should be reduced by 0.5, following the switch to CPI inflation. For the first five observations there are four bins, with the two interior bins combining, pairwise, the four interior bins of the six-bin grid, as described above: their colours are intermediate, in the same spectral sense, between the separate colours of their corresponding pairs. The large black dots show in which bin the inflation outcome, two years later, fell. There is no inflation scale in Figure 1, and the dots are simply placed in the centre of the probability range of the appropriate bin; this is the same bin for both forecasts, since we have calculated the MPC's probabilities as if the MPC was answering the SEF questionnaire, as noted above. Readers wishing to see a plot of actual inflation outcomes should consult Figure 2. The QPS and RPS for each observation are shown with reference to the right-hand scale; these points are joined by solid and dashed lines respectively, and their mean values over the 36 observations are the content of Table 1.

For most of the period, the inflation outcomes fell in one of the two central bins of the histograms, and the RPS is smaller than the QPS because it correctly acknowledges the appropriate unimodal shape of the densities, for both forecasts. The SEF scores are generally smaller than the MPC scores in these circumstances, because the SEF densities have smaller dispersion. However the last three forecasts provide an interesting contrast. The outcomes, with CPI inflation in excess of 3%, fell in the upper open-ended bin, and the MPC's greater tail probabilities result in its lower scores. The difference with the SEF is more marked in the case of the RPS, where the MPC correctly benefits from greater probabilities not only in the upper bin, but also in the adjoining bin. However these three observations are not sufficient to offset the overall lower scores of the SEF average forecasts, as indicated by the sample means in Table 1. Nevertheless these different episodes illustrate the advantage of the RPS in better reflecting probability forecast performance in categorical problems which have a natural ordering, such as these density forecast histograms, and its continued use is recommended.

The inclusion of Figure 2 for the benefit of readers who are unfamiliar with the UK's inflationary experience over this period also allows us to relate a further comparison between the SEF average forecasts and the MPC's forecasts. Figure 2 shows the inflation outcomes, 2000Q1-2008Q4, together with point forecasts made two years earlier, namely the MPC density forecast means as published on the Bank's spreadsheets and the corresponding means calculated from the SEF average histograms. The general tendency of the external forecasts to stay close to the inflation target irrespective of the inflation experience at the time the forecasts were made is often taken to be an indication of the credibility of the MPC and the inflation targeting policy regime. Viewed simply as forecasts, however, as in the analysis of the MPC's forecasts by Groen, Kapetanios and Price (2009), we find that their respective forecast RMSEs are 0.65 (MPC) and 0.61 (SEF), which matches the ranking of these forecasts given in Table 1 by the scoring rules.

4. Scoring the individual SEF respondents

4.1. QPS and RPS for regular respondents

The dataset of individual SEF responses made available by the Bank of England gives each respondent an identification number, so that their individual responses, including non-response, can be tracked over time, and their answers to different questions can be matched. The total number of respondents appearing in the dataset is 48, but there has been frequent entry and exit, as in other forecast surveys, and no-one has answered every question since the beginning. To avoid complications caused by long gaps in the data, and to maintain degrees of freedom at a reasonable level, we follow the practice of US SPF researchers and conduct our analyses of individual forecasters on a subsample of regular respondents. For the present purpose we define "regular" as "more than two-thirds of the time", which gives us a subsample of 16 individual respondents, who each provided between 25 (two individuals) and 36 (one individual) of the 36 possible two-year-ahead density forecasts of inflation over the 1998Q1-2006Q4 surveys.

We first extend the QPS-RPS comparison of the previous section to the individual level. Each regular respondent's scores are calculated from their available forecasts and outcomes, thus T in equations (1) and (2) is between 25 and 36. A scatter diagram of the results is presented in Figure 3, which also includes the SEF average density forecast as a point of reference, plotted at the values given in Table 1. Bearing in mind the difference in scales, it is seen that all 16 points lie below the "45°" line, thus Section 3's finding for the SEF average forecast that the RPS is less than the QPS extends to these individual forecasts, for the same general reasons discussed above. The scatter of points is positively sloped, and the correlation between the QPS and RPS of the regular respondents is 0.87. Nevertheless there are some small reversals of rankings: whenever the line joining two points has a negative slope, the QPS and RPS disagree about the ranking of the corresponding individuals.

For detailed individual scrutiny we first pick out individual 26, who is the only ever-present regular respondent, and is highly ranked (3rd) on both scores, and is an outlier in one further respect. Whereas almost three-quarters of all the individual forecasts in the sample (357 out of 485) utilise all available histogram bins, there are 21 forecasts which have non-zero entries in only two bins, and 17 of these are individual 26's forecasts. The upper panel of Figure 4 shows the observation-by-observation components of the score calculations for individual 26 as in Figure 1; on the five occasions when inflation fell in outer bins with zero forecast probabilities, the large black dots are placed on the boundary of the grids. These include two quarters with inflation below 2% (the 2000Q2,Q3 forecasts) and two with inflation above 3% (the 2006Q2,Q4 forecasts). For each of these four observations the QPS takes approximately the same value, in the range 1.50-1.58, suggesting that the four forecasts are of approximately equal quality. On the other hand the RPS gives a well-separated ranking of these forecasts: 2006Q2 is clearly worst, followed by 2006Q4, whereas 2000Q2,Q3 are rather better. Given the location of the various probabilities forming the histograms, this latter view is correct, and the QPS's indifference to this question again emphasises its inadequacy as an indicator of the quality of these density forecasts.

4.2. Missing data

For comparison we include in the lower panel of Figure 4 the corresponding data for individual 25, who has the best RPS result, as shown in Figure 3. Although the first seven forecasts do not score as well as those of individual 26, the local peaks in the latter's RPS at the zero-probability outcomes have much diminished counterparts in individual 25's scores.

Also very noticeable, however, is that individual 25's last two forecasts are missing, whereas these observations make relatively high contributions to individual 26's overall RPS.

To place such comparisons on an equal basis, one might consider calculating the scores over the subsample of observations common to both forecasters, thus in the above case simply using the first 34 datapoints for both individuals. However this neglects available information on the forecast performance of the individual who has responded more often. Moreover to make multiple comparisons among our 16 regular respondents this is not a practical solution. Although none of these respondents is missing more than 11 of the 36 possible forecasts, the more-or-less random occurrence of the missing forecasts means that there are only three occasions when all 16 individual forecasts are available. Overall, 91 of the possible $16 \times 36 = 576$ forecasts are missing, comprising 77 cases of complete non-response to the questionnaire, and 14 cases of an incomplete questionnaire being returned, known as *item non-response* to survey practitioners. There is no evidence that the process leading to missing forecasts depends on either forecasts or inflation outcomes, and the missing data can be called *missing at random* and the observed data *observed at random* using terms introduced by Rubin (see Little and Rubin, 2002). Neither imputation-based methods nor model-based methods for handling incomplete data, as discussed by Little and Rubin, appear applicable in the present context, although we note an interesting application to the construction of combined point forecasts in the face of missing data in the US SPF by Capistran and Timmermann (2007).

Instead, as discussed at the end of Section 2, we focus on the components of the score that reflect forecaster performance, by correcting the score for variation in the outcome variance term identified in the Yates decomposition (equation (4), or its generalisation to the RPS). To retain comparability with the uncorrected score, we replace the outcome variance calculated over an individual's subsample by the full-sample outcome variance. Thus the score for individual 26, who has no missing observations, does not change. (For the purpose of this calculation we assume six histogram bins throughout, and there is no difficulty in assigning the first five outcomes accordingly.)

The results are shown in Figure 5, as a scatter diagram of RPS and adjusted RPS (denoted RPS*) values. Points lying above the $45°$ line represent individuals whose score has increased as a result of the adjustment, and their previous lower score might be considered to be the result of having missed some hard-to-forecast occasions. This adjective certainly applies to the last three inflation outcomes in our sample, and individuals 2, 25 and 27 did not respond on two of these occasions, while individual 8 missed all three. The adjustment corrects for the smaller outcome variance in their respective subsamples and increases their scores, resulting in a more accurate picture of their relative forecast performance. In particular, the adjustment moves individual 25 from 1st to 4th position in the ranking, and individual 8 from 8th to 14th.

As a final illustration at the individual level the data for the two respondents whose scores are decreased most as a result of the adjustment are shown in Figure 6. Individual 9, in the upper panel, has the same number of missing observations – ten – as individual 8, but these correspond to outcomes that fell in the central bins of the histograms. Thus the subsample outcome variance is greater than the full-sample variance and the adjustment reduces the score. Nevertheless individual 9 remains ranked in last place, as a result of the excessive dispersion of the forecast histograms, in particular the high probabilities attached to forecast outcomes in the lowest, open-ended bin, which did not materialise. On the other hand for individual 31, in the lower panel of Figure 6, who has eleven missing observations similarly distributed, the adjustment changes the ranking, from 6th on RPS to the top ranked position on RPS*. The scores for the four forecasts made between 2002Q3 and 2003Q2 are unusually small, as a result of placing rather high probabilities in the bins into which inflation duly fell, and zeroes in the outer bins. Throughout, unlike individual 9, individual 31 placed small, or zero, probabilities in the lower open-ended bin, and the latter's relative scores benefited from this choice, except in 2000Q2,Q3.

The overall effect of these adjustments for differential non-response is to reduce the dispersion of the individual scores. The $\operatorname{var}\left(D_k\right)$ terms in the Yates decomposition are outside the forecasters' influence, and assuming that these are independent of the factors that result in individual non-response from time to time, the adjusted score RPS* that corrects for the differential impact of these terms gives a better comparative summary of individual forecast performance. There remains considerable dispersion in the RPS* scores, however, and this heterogeneity in individual density forecasting performance mirrors the finding of considerable heterogeneity in point forecasting performance in this survey by Boero, Smith and Wallis (2008b).

4.3. Forecast combination

The idea that combining different forecasts of the same event might be worthwhile has gained wide acceptance since the seminal article of Bates and Granger (1969), although the resulting literature has been mostly concerned with point forecasts, as was that article. The relative neglect, until recently, of density forecast combination is surprising, given that the first article on the US survey of forecasters (Zarnowitz, 1969) included a survey average density forecast constructed as a simple average of respondents' forecast histograms: this appeared in the same year as Bates and Granger's article. The practice of reporting survey average forecast histograms continues to this day, as seen above. Recent contributions on density forecast combination more generally include Wallis (2005), Hall and Mitchell (2007) and Geweke and Amisano (2008).

Density forecast combination as yet lacks an optimality result comparable to that of Bates and Granger (1969) for point forecast combination. They showed that a linear combination of two competing point forecasts using the optimal (variance minimising) weight in general has a smaller expected squared forecast error than either of the two competing forecasts. The only case in which no improvement in forecast performance in this sense is possible is that in which one forecast is already the optimal forecast (with minimum expected squared error), whereupon the optimal weights are 1 and 0. On the other hand, for two density forecast histograms we can show that the expected QPS of any linear combination of two forecasts exceeds that of the better forecast, thus combining a reasonably well-performing forecast with an inferior forecast makes matters worse. This might accord with one's intuition, but for point forecasts this intuition was overturned by Bates and Granger's result and many subsequent empirical studies.

In the light of the continued use of simple averages of survey forecast histograms and the location of its score in the middle of the cluster of individual scores seen in Figure 3, we investigate the applicability of a different result from the empirical point forecast combination literature, namely that trimmed means often outperform other combinations: for a recent survey see Timmermann (2006). The heterogeneity of individual performance among the 16 regular forecasters considered above suggests that improvement in the scores of an average forecast might be obtained if some of the individuals with higher scores were excluded from the average.

To mimic a real-time setting for the exercise we divide our sample period into "training" and "evaluation" subperiods. The first subperiod is used to identify the eight individuals with lowest RPS* scores. Remembering that in order to calculate the score for a given forecast we have to wait two years for the outcome to materialise we set the total length of this period at five years, thus selection into the trimmed sample is based on the RPS* scores over the first 12 forecasts. A trimmed average forecast is then constructed from the forecast histograms of these individuals, and its observation-by-observation data are shown in Figure 7. For the first five years this is an ex-post construction which could not have been done at the time, whereas over the second subperiod, the final four years, we have an ex-ante forecast which can be evaluated by comparing its scores to those of the published SEF average shown in Figure 1. Over these 16 quarters, the mean RPS is 0.683 for the SEF average and 0.616 for the trimmed average, thus some gains are available from this device.

We end this section with an observation on the respective empirical literatures on point and density forecast comparison and combination. Smith and Wallis (2009) note that there are two strands in the point forecast comparison literature, one concerned with comparisons across individual participants in forecast surveys and the other with comparisons of competing forecast models and methods constructed by interested researchers. Greater heterogeneity in individual forecast performance is often found in analyses of surveys of economic forecasters than in comparisons of forecasts from competing statistical models. The corresponding density forecast comparison literature is as yet sparse, but we already notice a similar outcome. As a summary measure of the dispersion of individual RPS* scores shown in Figure 5 we obtain a coefficient of variation of 0.143. Geweke and Amisano (2008) construct six competing forecasting models for the daily S&P 500 returns from the ARCH, stochastic volatility and Markov mixture families. The fitted models yield continuous densities whose logarithmic scores are reported in their Table 1: their coefficient of variation is 0.051. Whether the forecasting models that researchers construct for comparative purposes continue to exhibit less diversity than respondents to forecast surveys remains to be seen.

5. Conclusion

This paper provides a practical evaluation of some leading density forecast scoring rules in the context of forecast surveys. We analyse the forecasts of UK inflation obtained from the

Bank of England's Survey of External Forecasters, considering both the survey average forecasts published in the quarterly *Inflation Report*, and the individual survey responses recently made available by the Bank. The density forecasts are collected as a set of probabilities that future inflation will fall in one of a small number of preassigned ranges, and thus are examples of categorical forecasts in which the categories have a natural ordering. The ranked probability score was initially proposed as an alternative to the quadratic probability score for precisely these circumstances, and our exercise makes its advantages clear.

Missing observations are endemic in surveys, and we have two answers to this problem in the present context. First, in common with much other research on forecast surveys, our study of individual forecast performance is conducted on a subsample of regular respondents. In our case these are the 16 respondents who are each missing less than one-third of the possible two-year-ahead forecasts collected between 1998Q1 and 2006Q4. Their forecast scores have considerable dispersion, part of which is due to differences in the inflation outcomes over the different subperiods for which individuals provided their forecasts. Accordingly, and secondly, we propose an adjustment to the score, based on the Yates decomposition, which corrects for the differential impact of the component of the score that depends only on the outcome and not on the forecast, and hence gives a clearer measure of forecaster performance. We recommend the adjusted ranked probability score, denoted RPS*, to other analysts of forecast surveys facing the familiar problems of non-response.

Density forecast combination is receiving increasing attention, which parallels several themes in the well-established point forecast combination literature. One such theme concerns the sensitivity of a simple mean of several forecasts to extreme forecasts, and the use of a trimmed mean, obtained by removing the worst-performing forecasters from the set before the mean is calculated, as an alternative. In a similar exercise, we identify the eight better-performing forecasters over the first three years of our surveys, remembering that we have to wait a further two years for the outcome data, and then find that the average forecast of these eight outperforms the published survey average forecast over the remaining four years of our sample period. The Survey of External Forecasters provides an input to the Monetary Policy Committee's quarterly forecasting round, and the use of trimmed means for this purpose is seen to have advantages. On the other hand, the survey is also used as an indicator of current sentiment about future macroeconomic developments and the credibility of monetary policy, and it might be argued that, for this purpose, no-one's view should be discarded.

Previous work analysing the point forecasts of inflation and GDP growth from the SEF (Boero, Smith and Wallis, 2008b) found considerable heterogeneity among individual respondents, shown by the failure of standard tests of equality of idiosyncratic error variances and evidence of different degrees of asymmetry in forecasters' loss functions. Similar dispersion of forecast scores from their density forecasts of inflation again indicates that some respondents are better at forecasting than others. This leads us to close this paper with the same final thought as that article, that "several of our findings prompt questions about the individual forecasters' methods and objectives, the exploration of which would be worthwhile".

References

Bates, J.M. and Granger, C.W.J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20, 451-468.

Boero, G., Smith, J. and Wallis, K.F. (2008a). Uncertainty and disagreement in economic prediction: the Bank of England Survey of External Forecasters. *Economic Journal*, 118, 1107-1127.

_____ (2008b). Evaluating a three-dimensional panel of point forecasts: the Bank of England Survey of External Forecasters. *International Journal of Forecasting*, 24, 354-367.

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

Bross, I.D.J. (1953). *Design for Decision*. New York:Macmillan.

Capistran, C. and Timmermann, A. (2007). Forecast combination with entry and exit of experts. *Journal of Business and Economic Statistics*, forthcoming.

Casillas-Olvera, G. and Bessler, D.A. (2006). Probability forecasting and central bank accountability. *Journal of Policy Modeling*, 28, 223-234.

Epstein, E.S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8, 985-987.

Galbraith, J.W. and van Norden, S. (2008). Calibration and resolution diagnostics for Bank of England density forecasts. Presented at the 'Nowcasting with Model Combination' Workshop, Reserve Bank of New Zealand, 11-12 December 2008.

Geweke, J. and Amisano, G. (2008). Optimal prediction pools. http://www.biz.uiowa.edu/faculty/jgeweke/papers/paperC/paper.pdf

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359-378.

Good, I.J. (1952). Rational decisions. *Journal of the Royal Statistical Society* B, 14, 107-114.

Groen, J.J.J., Kapetanios, G. and Price, S. (2009). A real time evaluation of Bank of England forecasts of inflation and growth. *International Journal of Forecasting*, 25, 74-80.

Hall, S.G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23, 1-13.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd edn). Hoboken, NJ: Wiley-Interscience.

Murphy, A.H. (1971). A note on the ranked probability score. *Journal of Applied Meteorology*, 10, 155-156.

_____ (1972). Scalar and vector partitions of the probability score: Part II. *N*-state situation. *Journal of Applied Meteorology*, 11, 1183-1192.

_____ (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12, 595-600.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, 2, 191-201.

Smith, J. and Wallis, K.F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, in press.

Stael von Holstein, C.S. and Murphy, A.H. (1978). The family of quadratic scoring rules. *Monthly Weather Review*, 106, 917-924.

Tay, A.S. and Wallis, K.F. (2000). Density forecasting: a survey. *Journal of Forecasting*, 19, 235-254. Reprinted in *A Companion to Economic Forecasting* (M.P. Clements and D.F. Hendry, eds), pp.45-68. Oxford: Blackwell, 2002.

Timmermann, A. (2006). Forecast combinations. In *Handbook of Economic Forecasting* (G. Elliott, C.W.J. Granger and A. Timmermann, eds), pp.135-196. Amsterdam: North-Holland.

Wallis, K.F. (2004). An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review*, No.189, 64-71.

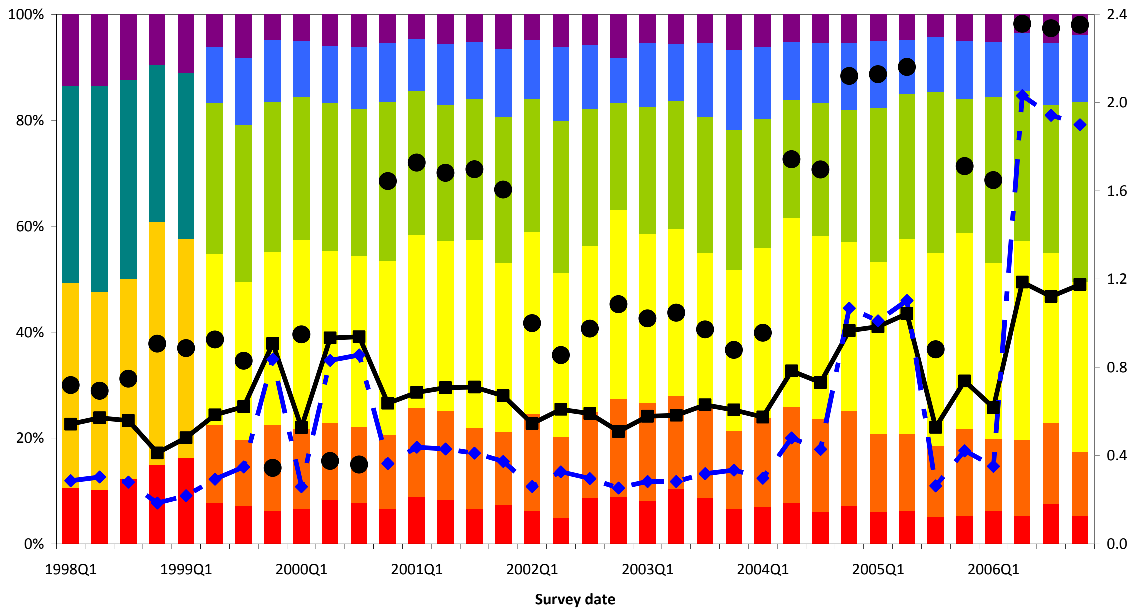
_____ (2005). Combining density and interval forecasts: a modest proposal. *Oxford Bulletin of Economics and Statistics*, 67, 983-994.

Yates, J.F. (1982). Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30, 132-156.

_____ (1988). Analyzing the accuracy of probability judgments for multiple events: an extension of the covariance decomposition. *Organizational Behavior and Human Decision Processes*, 41, 281-299.

Zarnowitz, V. (1969). The new ASA-NBER survey of forecasts by economic statisticians. *American Statistician*, 23(1), 12-16.

Figure 1. Forecast probabilities two years ahead, inflation indicators, QPS and RPS
Upper panel: SEF average forecast; lower panel: MPC forecast



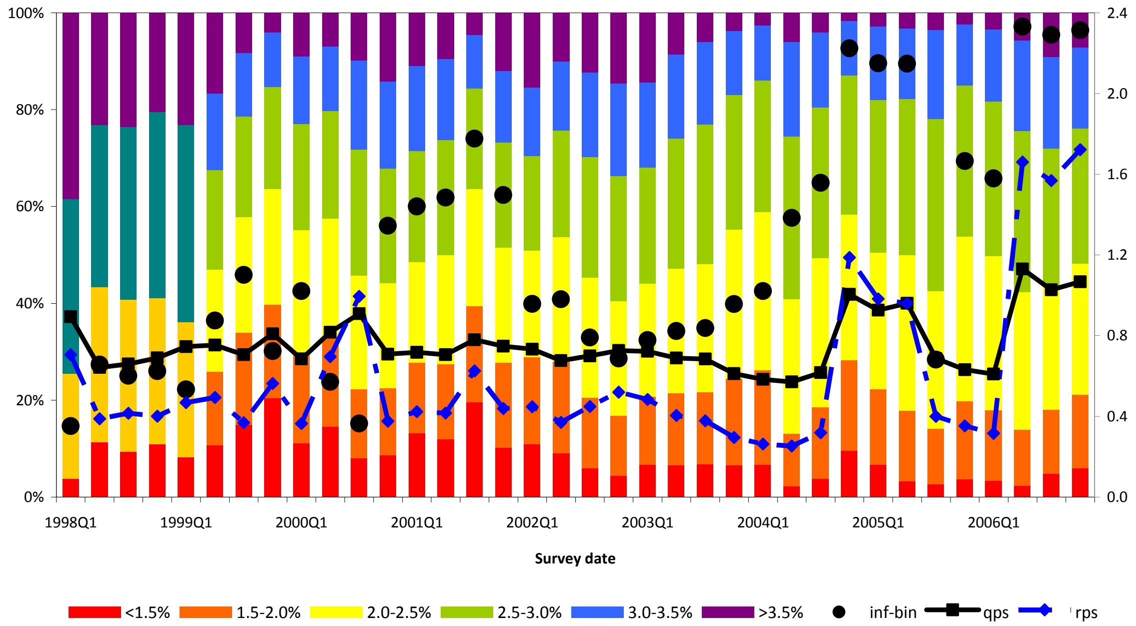


Figure 2. Inflation, 2000Q1-2008Q4, and mean forecasts made two years earlier

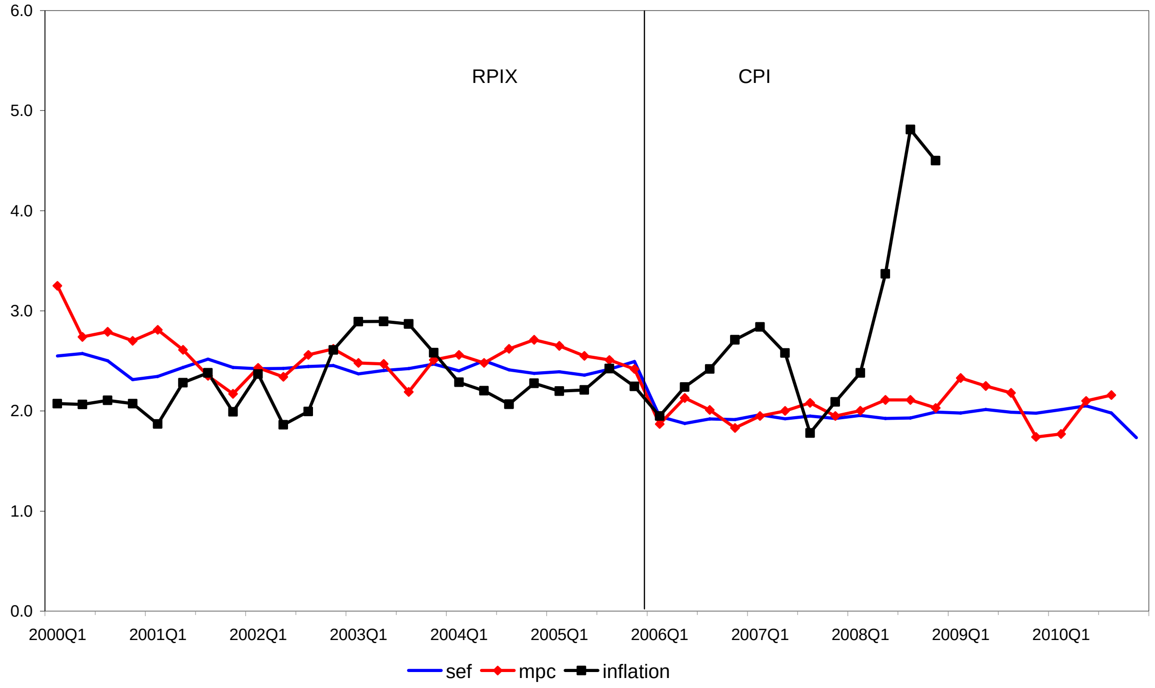


Figure 3. QPS and RPS for 16 regular respondents and the SEF average (filled square)

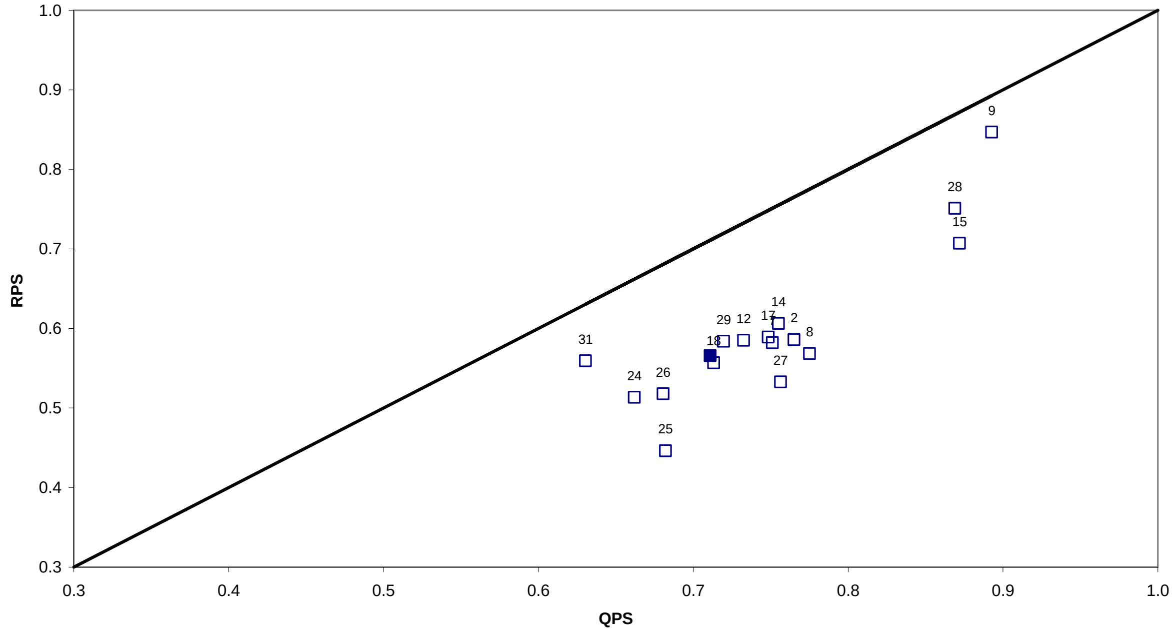
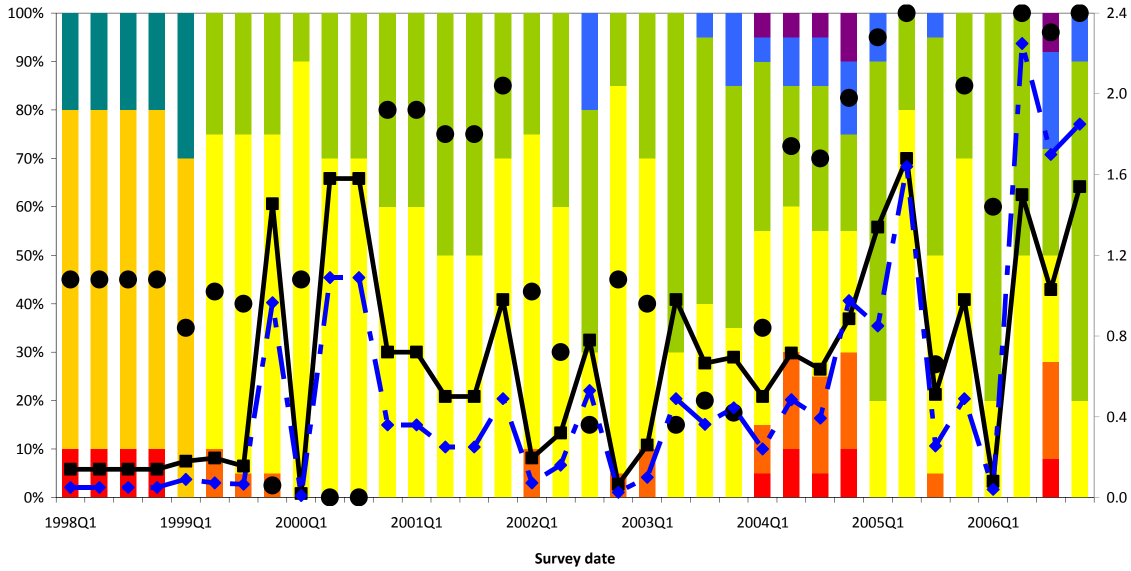


Figure 4. Forecast probabilities two years ahead, inflation indicators, QPS and RPS
Upper panel: individual 26; lower panel: individual 25



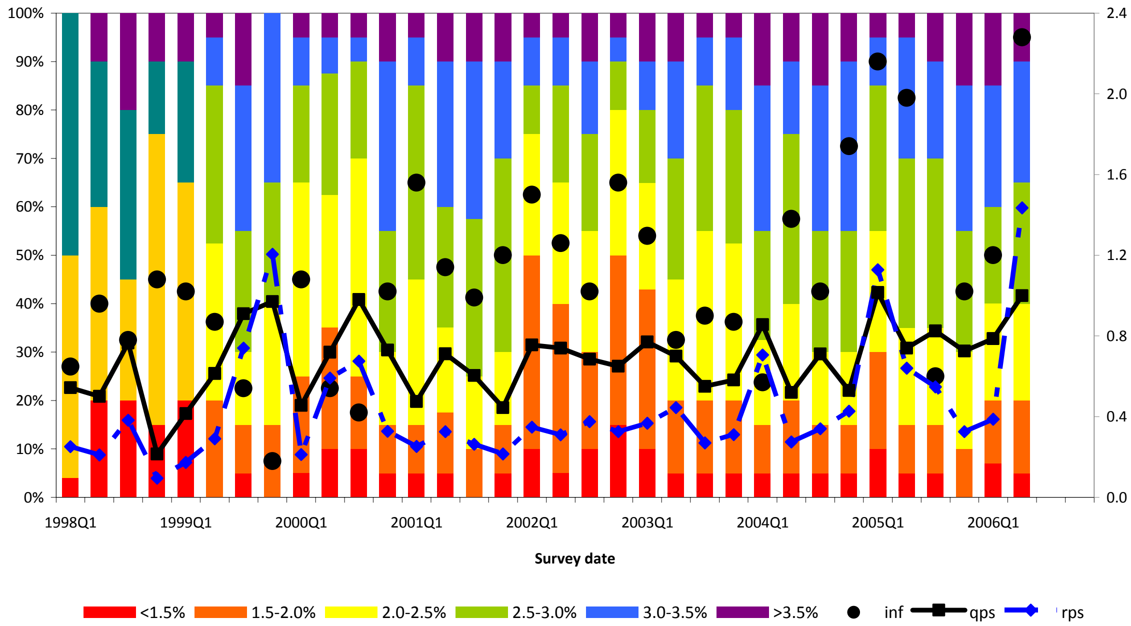


Figure 5. RPS and RPS* for 16 regular respondents

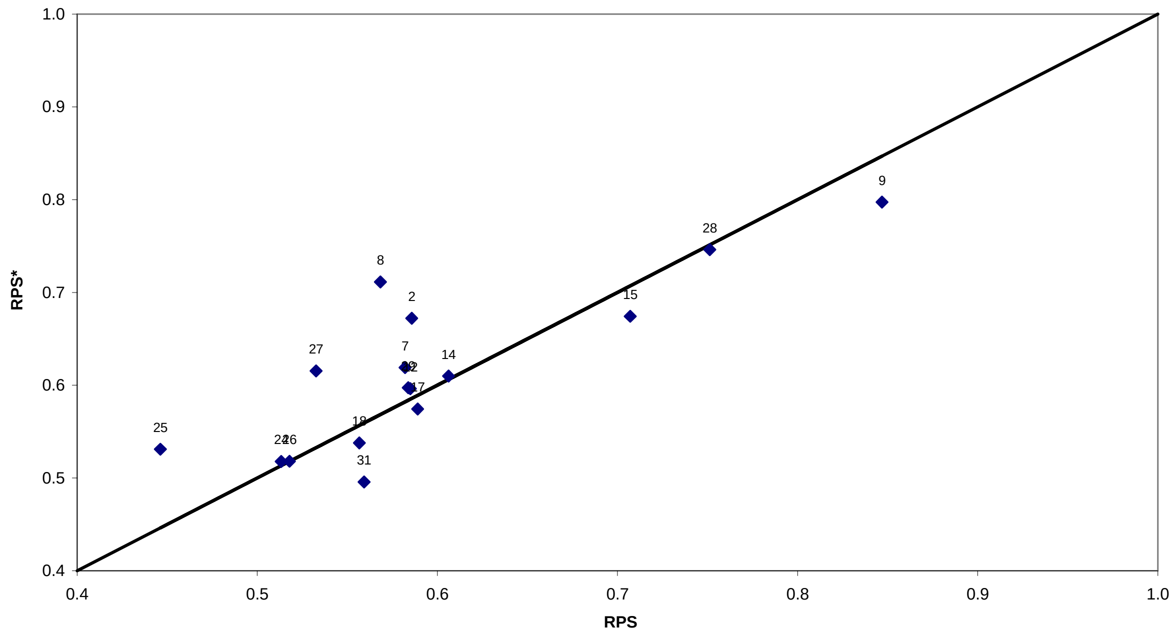
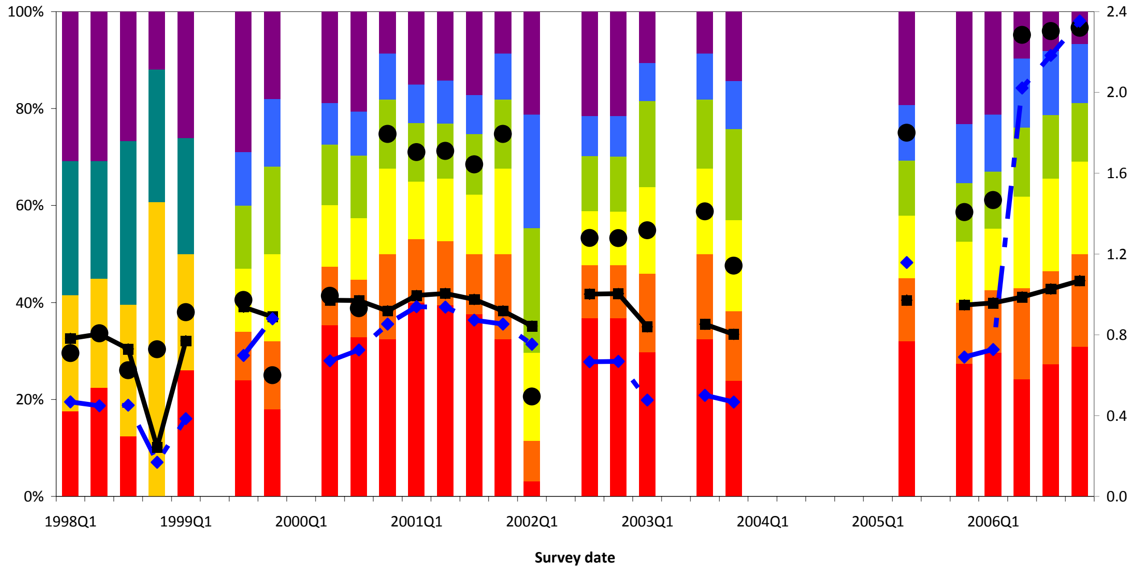


Figure 6. Forecast probabilities two years ahead, inflation indicators, QPS and RPS
Upper panel: individual 9; lower panel: individual 31



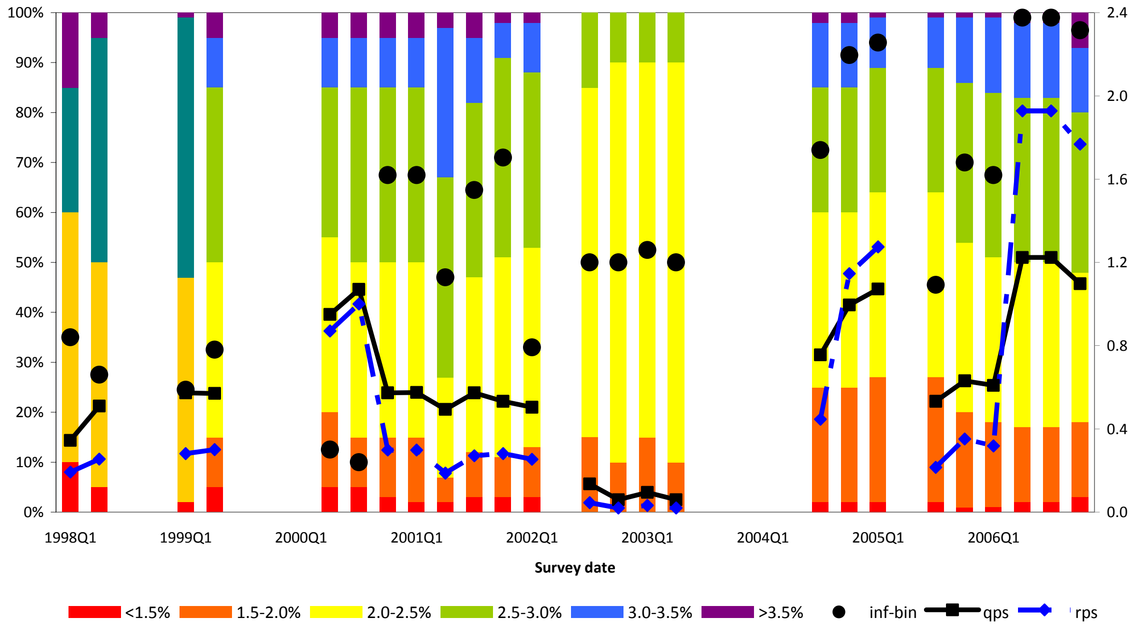


Figure 7. Forecast probabilities two years ahead, inflation indicators, QPS and RPS
SEF trimmed average forecast