Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

International Journal of Forecasting 27 (2011) 379–393

international journal of forecasting

www.elsevier.com/locate/ijforecast

Scoring rules and survey density forecasts

Gianna Boero, Jeremy Smith, Kenneth F. Wallis*

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

Abstract

This article provides a practical evaluation of some leading density forecast scoring rules in the context of forecast surveys. We analyse the density forecasts of UK inflation obtained from the Bank of England's Survey of External Forecasters, considering both the survey average forecasts published in the Bank's quarterly *Inflation Report*, and the individual survey responses recently made available to researchers by the Bank. The density forecasts are collected in histogram format, and the ranked probability score (RPS) is shown to have clear advantages over other scoring rules. Missing observations are a feature of forecast surveys, and we introduce 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. The new measure, denoted RPS*, is recommended to analysts of forecast surveys.

Keywords: Density forecast evaluation; Brier (quadratic probability) score; Epstein (ranked probability) score; Logarithmic score; Bank of England Survey of External Forecasters; Missing data; Forecast comparison

1. Introduction

In many forecasting applications, attention is focused on 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 early 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 for 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,

* Corresponding author.
E-mail address: K.F.Wallis@warwick.ac.uk (K.F. Wallis).

0169-2070/$ - see front matter © 2010 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2010.04.003

this year and next year, might fall. In this way, the 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 follow the practice of the SPF in making the individual responses to the survey, made suitably anonymous, available for research purposes. This article considers methods for the comparative assessment of the quality of such forecasts, with the Bank of England Survey of External Forecasters (SEF) as a practical example. Other aspects of the SEF dataset are explored by Boero, Smith, and Wallis (2008a,b,c).

A scoring rule measures the quality of a probability forecast using a numerical score based on the forecast and the eventual outcome, and can be used to rank competing forecasts. The earliest example of such a rule, introduced by Brier (1950) and subsequently bearing his name, involves the situation in which 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 sum 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. The example in Brier's article concerned the verification of probability forecasts of rain or no-rain in given periods: this 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 & Murphy, 1978).

When evaluating survey density forecasts, the distinct classes or categories of the Brier score's set-up are taken to be the set of histogram bins. However, the ranking or ordering of the bins in terms of the values of the underlying continuous variable is neglected. For a four-bin histogram where the outcome falls in the bin that has been assigned a probability of 0.3 and the other bins have probabilities of 0.5, 0.1 and 0.1, for example, the Brier score is invariant to the way in which these last three probabilities are assigned to the bins in which the outcome does not fall. However, forecasts that have placed 0.5 in a bin adjacent to the bin in which the outcome falls would generally be regarded as better forecasts than those that have not. The Ranked Probability Score introduced by Epstein (1969), a second member of the class of quadratic scoring rules, takes the ordering of the categories into account. It does not appear to have previously been used in the present area of research, although its extension to continuous distributions, the continuous ranked probability score (CRPS), has recently attracted attention in the meteorological literature (Gneiting & Raftery, 2007).

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

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

for a density forecast f of the random variable X_t evaluated at the outcome x_t. The logarithmic score has many attractive features, and appears in the literature in many guises. To a Bayesian, the logarithmic score is the log predictive likelihood, and if two forecasts are being compared, the log Bayes factor is the difference between their logarithmic scores. The definition in terms of a continuous density can readily be adapted to discrete distributions and discretised continuous distributions, as in the present context, although there is then a potential difficulty: as can be seen below, from time to time in the individual survey responses the outcome falls in a histogram bin to which the respondent has assigned a zero probability, which means that the log score is undefined. To assign an arbitrary value to the score on such occasions is not a satisfactory solution, since the ranking of competing forecasts will be sensitive to the value chosen. On the other hand, zero-probability forecast outcomes are readily accommodated by the quadratic scores.

In this article we compare and contrast the Brier and Epstein rules, or QPS and RPS, and the logarithmic score, in applications to survey density forecasts of UK inflation. Section 2 contains the technical background to our study, comprising a formal presentation of the rules, a consideration of the relevance of the various decompositions that have been proposed, and a discussion of the statistical tests of predictive ability that we employ. The empirical analysis begins in Section 3 with a comparison of the published survey average density forecasts from the SEF and the density forecasts of the Bank of England's Monetary Policy Committee (MPC). Section 4 turns to the individual SEF respondents and uses the two quadratic scoring rules to evaluate their forecast performances, and it is seen that the RPS is preferred. Incomplete data are a feature of this survey, like all forecast surveys, and our adjusted score, RPS*, is found to provide more reliable rankings of forecasters in the face of missing observations caused by differential non-response. Section 5 concludes.

2. Scoring rules and their applications

2.1. The Brier, Epstein and logarithmic 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 $(p_{1t}, \ldots, p_{Kt})$. 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 $d_{kt}, k = 1, \ldots, K$, which take the value 1 if the outcome x_t falls in bin k, and $d_{kt} = 0$ otherwise. Also in mind are time series forecasting applications, in which each forecast of the outcome at times $t = 1, \ldots, T$ is formed at some previous time. For a sample of forecasts and realisations of the categorical variable, the sample mean Brier score is given as

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

It has negative orientation — smaller scores are better. The range is usually stated as $0 \leq \mathrm{QPS} \leq 2$, although the extreme values are only obtained in extreme circumstances in which, in every period, all of the probability is assigned to a single bin and the outcome either does or does not fall into it.

The Brier score is also invariant to the ordering of the $K - 1$ bins which have $d_{kt} = 0$ at each time t, as noted above. To take this into account, Epstein's (1969) proposal replaces the density functions implicit in the Brier score with their corresponding distribution functions (Murphy, 1971). Defining these as

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

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

$$\mathrm{RPS} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}(P_{kt} - D_{kt})^2 . \tag{2}$$

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, and occurs 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 that is from the kth bin. In extremis, with the outcome and the unit-probability bin located at opposite ends of the range, this value is $K - 1$.

Adapting the definition of the logarithmic score given above to the histogram context gives

$$\log \mathrm{S} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} d_{kt} \log (p_{kt}) .$$

This has a positive orientation — larger scores are better, and since log S typically takes negative values, scores with smaller absolute values are typically better.

2.2. Decompositions of the quadratic 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 Eq. (1), which we denote by QPSE, namely

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

Here p_t is the forecast probability, and $d_t = 1$ if the event occurs or zero if it does not. The QPSE is equal to half of the value obtained from Eq. (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 $p_j = j/10$, $j = 0, \ldots, 10$. The QPSE can then be calculated subset-by-subset by rearranging the summation in Eq. (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. Thus, we have a two-component decomposition of the QPSE as

$$\text{QPSE} = \frac{1}{T}\sum_{j} T_j\left(p_j - \bar{d}_j\right)^2 + \frac{1}{T}\sum_{j} T_j\bar{d}_j\left(1 - \bar{d}_j\right).$$

The first component 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 it is close to a diagonal line. The second component involves only the outcome indicators, but nevertheless reflects the forecaster's 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, Chapter 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 between 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 of the 11 subcollections of forecasts differ from the overall relative frequency of occurrence of the event: a 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, but one could simply round the probabilities to the nearest tenth. 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 a 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. However, they depend on a similar discretisation and grouping of the forecasts to that used in the above derivations, although using 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:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}(p_{kt}-d_{kt})^2 &= \sum_{k=1}^{K}\frac{1}{T}\sum_{t=1}^{T}[(p_{kt}-\bar{p}_k) \\ &\quad -(d_{kt}-\bar{d}_k)+(\bar{p}_k-\bar{d}_k)]^2 \\ &= \sum_{k=1}^{K}[\operatorname{var}(p_k)+\operatorname{var}(d_k) \\ &\quad +(\bar{p}_k-\bar{d}_k)^2-2\operatorname{cov}(p_k,d_k)]. \qquad (4)\end{aligned}$$

Yates (1988) notes that the second term in this last expression, the sum of the outcome indicator sample variances $\operatorname{var}(d_k)=\bar{d}_k(1-\bar{d}_k)$, is outside the forecaster's influence, while the third term, the sum of squared "biases", indicates the miscalibration of the forecasts. He also offers a 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, which we extend 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 is 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, in order to focus on their forecast performance in a comparable manner we standardise the contribution of the variance of d to their individual scores. In order to ensure that the score of an individual with no missing observations is unaltered, we do this by replacing the individual subsample outcome variance component of the QPS with the full-sample outcome variance.

We are not aware of a comparable covariance decomposition of the RPS given in Eq. (2), although the (mostly meteorological) literature contains a considerable amount of discussion on extensions of the previous reliability-resolution-uncertainty decomposition of the RPS and its continuous generalisation (see Candille & Talagrand, 2005, for example). Nevertheless, it is clear that an equivalent covariance decomposition of the RPS can be obtained by replacing the lower-case p and d with the upper-case P and D throughout Eq. (4). As a result, a similar variance-of-D term can be identified that is a function of the outcomes alone. For comparing the forecast performances in the face of differential non-response, we calculate an adjusted score, denoted by RPS*, which is obtained by replacing the individual-specific measure of the outcome variance in the RPS with its full-sample equivalent.

2.3. *Testing predictive ability*

To construct a formal test of the equal predictive ability of two competing forecasts, we follow Giacomini and White (2006). Their framework encompasses point, interval and density forecasts and a wide range of loss functions, and can readily be adapted to the present context, although it is an asymptotic test and our forecast sample size is small. Their focus on what they call the *forecasting method* as the object of evaluation, which encompasses a number of choices made by the forecaster concerning models, data and estimation procedure, is appropriate in our present context, where little information is available about the actual choices made by survey respondents. Adapting and simplifying the notation of their Eqs. (4) and (6), the Wald-type test statistic, given T out-of-sample forecasts, is

$$W = T\left(T^{-1}\sum_{t=1}^{T}h_t\Delta L_t\right)'\Omega^{-1}\left(T^{-1}\sum_{t=1}^{T}h_t\Delta L_t\right),$$

where h_t is a $q \times 1$ vector of test functions, ΔL_t is the difference in the losses or scores of the two forecasts in period t, and Ω is an appropriate heteroscedasticity and autocorrelation consistent (HAC) estimate of the asymptotic covariance matrix. The null hypothesis is

$$E\left(h_t \Delta L_t\right) = 0, \quad t = 1, \ldots, T,$$

under which the distribution of W tends to χ_q^2 as $T \to \infty$. The simplest "unconditional" test of equal forecast performances has $q = 1$ and $h_t = 1$, $t = 1, \ldots, T$; in this case, when using the logarithmic score, the test is equivalent to (the unweighted version of) the likelihood ratio test of Amisano and Giacomini (2007). Possible practical choices of h_t for a "conditional" test are discussed below.

3. The SEF average and MPC density forecasts of inflation

Our empirical study of the scoring rules begins with a 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. Both forecasts are published in the Bank of England's quarterly *Inflation Report*, although it is necessary to consult the Bank's spreadsheets in order to obtain numerical values for the parameters of the two-piece normal distribution on which the MPC's fan charts are based.

The Bank of England's quarterly Survey of External Forecasters began in 1996. The institutions covered in the survey include commercial banks and other financial institutions, 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; this is typically in the low twenties.

The SEF questionnaire initially asked for forecasts of inflation in the last quarter of the current and following years. Such questions eventually deliver sequences of fixed-event forecasts, which were analysed by Boero et al. (2008c), but not quarterly series of fixed-horizon forecasts, which are required for the present exercise. However, in 1998 a third question was added, asking for forecasts two years ahead, and this marks the start of the series analysed here. (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, and thus our series is continued via the second question of the new format. 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, the forecasts collected in the eight quarters over 2002–3 have to be evaluated against the outcomes in 2004–5 for the previous target variable, rather than 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 surveys. 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, so that there are six bins from 1999Q2 onward (<1.5, 1.5–2, 2–2.5, 2.5–3, 3–3.5, >3.5); finally, 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. The survey average forecast, as published in the *Inflation Report* and considered in this section, is then obtained as a simple average over all respondents of their reported probabilities in each bin. For purposes of comparison, 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).

The scores of the two forecasts are shown in Table 1. It is clear that the survey average forecast has a smaller QPS than the MPC forecast, which matches Casillas-Olvera and Bessler's (2006) finding based on the first 14 of these 36 quarterly observations. The RPS and the log score give the same ranking of the two forecasts. The RPS values are smaller than the QPS values, because 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

Table 1
SEF average and MPC density forecasts of inflation: scores and test results.

	SEF	MPC	p-value, $h_t = 1$	p-value, $h'_t = (1, \pi_{t-1})$
QPS	0.711	0.759	0.171	0.118
RPS	0.566	0.596	0.519	0.395
Log score	−1.465	−1.535	0.451	0.336

Note: $T = 36$ (two-year-ahead forecasts published 1998Q1–2006Q4).

the outcome fell is acknowledged by the RPS, but not by the QPS.

The last two columns of Table 1 show the asymptotic p-values of the Giacomini-White test using two test functions. The first is an intercept, giving the "unconditional" test noted above. For a "conditional" test, Giacomini and White (2006, p. 1555) suggest that h_t might be chosen to include such variables as lagged loss differences and business cycle indicators. The former suggestion causes degrees of freedom problems in the present case, since we are working with what are in effect 9-quarter-ahead forecasts and a sample of 36 quarterly observations. As a simple example of the latter possibility we choose the most recent inflation observation at the time when the forecast was made, denoted by π_{t-1}. We use a Newey-West estimate of Ω, allowing for a moving average of order eight in the forecast errors. The results show that, although the three scores agree on the ranking of the two forecasts, in no case is the difference in the scores great enough to reject the hypothesis of equal forecast performances. The p-value for each score is reduced when lagged inflation is added to the test function, providing weak evidence of a differential use of this information by the two forecasts, but not sufficient to overturn the general conclusion.

To study the comparative behaviour of the quadratic scores in greater detail, we turn to Fig. 1, which illustrates, observation-by-observation, the components of the calculation of QPS and RPS, namely the histogram probabilities and the location of the inflation outcome, for the SEF average forecast and the MPC forecast in the upper and lower panels respectively. The segments of the vertical columns show, with reference to the scale on the left, the allocation of forecast percentage probabilities to the histogram bins: each column stacks the bars of a standard histogram diagram, one on top of another. For most of the period there are six bins, and the colours (shown in the online version of the article) follow a rainbow array. The key to the figure records the RPIX inflation range for each bin; from 2004Q1 onward all of 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 Fig. 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 were answering the SEF questionnaire, as noted above. (Readers wishing to see a plot of actual inflation outcomes should consult Fig. 2.) The QPS and RPS for each observation are shown with reference to the scale on the right; these points are joined by solid and dashed lines respectively, and their mean values over the 36 observations are reported in 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 dispersions. 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 lead to 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 the probability forecast performance in categorical problems which have

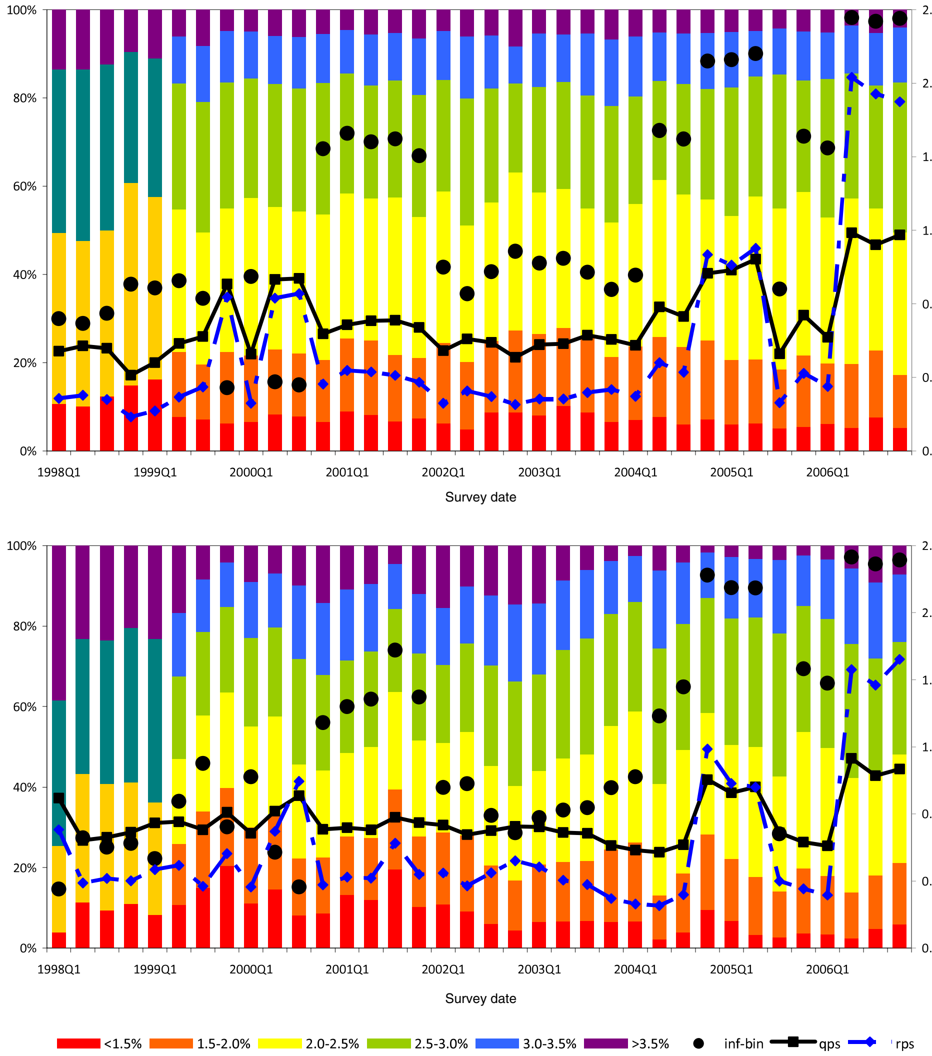

Fig. 1. Forecast probabilities two years ahead, inflation indicators, and the forecast scores (QPS and RPS). Upper panel: SEF average forecast; lower panel: MPC forecast. *Note:* A coloured version of this figure is available in the online version of the article.

a natural ordering, such as these density forecast histograms, and its continued use is recommended.

Combining some of the bins of a six-bin histogram to obtain a four-bin histogram generally improves the score: the probability associated with the bin in which the outcome falls may increase, and, for the quadratic scores, the sum of squares of probabilities is reduced. In principle, the first five scores in Fig. 1 are therefore not directly comparable with the remainder, although in practice they do not appear to be substantially smaller than subsequent scores, for both QPS and RPS, and both the SEF average and MPC forecasts. Such heterogeneity could be avoided either by working with a shorter sample period, or by reducing the 31 six-bin histograms to four bins. However, either course of action would result in a loss of information which we prefer not to incur, and we return to this issue in the context of the individual survey responses below.

The inclusion of Fig. 2 for the benefit of readers who are not familiar 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. Fig. 2 shows the inflation outcomes,

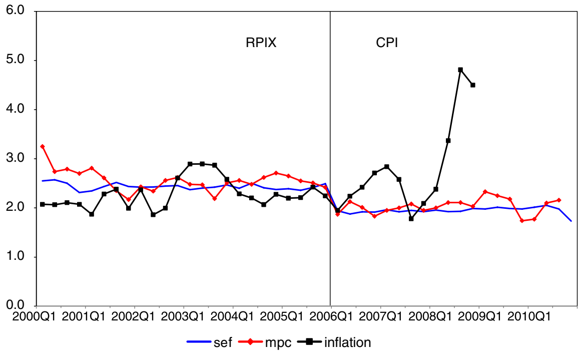

Fig. 2. Inflation over the period 2000Q1–2008Q4, with the mean forecasts made two years earlier.

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. (We apply the standard formula, assuming that the open-ended bins have twice the width of the interior bins; it makes no difference whether the probabilities are assumed to be concentrated at the mid-points of the respective bins, or spread uniformly across each bin.) 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 by the scoring rules, as given in Table 1.

4. Scoring the individual SEF respondents

4.1. QPS and RPS for regular respondents

The dataset of individual SEF responses which has been 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 the complications caused by long gaps in the data, and to maintain the 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 respondents, who each provided between 25 (two respondents) and 36 (one respondent) of the 36 possible two-year-ahead density forecasts of inflation over the 1998Q1–2006Q4 surveys.

Although the survey average forecasts always have non-zero probabilities in every bin, as can be seen in Fig. 1, many individual forecasters use fewer of the available bins. Moreover, for 12 individual forecasts, made by four respondents, inflation falls in a bin which has a forecast probability of zero; hence, a logarithmic score cannot be calculated for these four respondents. It is sometimes suggested that an imputed non-zero value be assigned in such cases, with reference to the degree of rounding of the reported probabilities: see, for example, the robustness check in this connection in Engelberg, Manski, and Williams (2009) study of US SPF forecasts. In the present exercise, the reported probabilities in nine of the forecasts under consideration are multiples of 0.10; in the remaining three forecasts multiples of 0.05 also appear. In addition, when the forecaster uses only the two central bins, it is possible for the outcome to fall in a non-adjacent bin, which is indeed observed in our data. Thus, there is considerable freedom available to the researcher who wishes to second-guess the forecaster, which makes the resulting score unreliable for ranking competing forecasts. A possible solution is to exclude these four respondents from further consideration; however, we prefer to set the logarithmic score aside, and accordingly consider only the quadratic scores in this section.

We first extend the QPS–RPS comparison of the previous section to the individual forecasters. For each regular respondent, both scores are calculated from their available forecasts and outcomes; thus, T in Eqs. (1) and (2) takes values between 25 and 36 for the different respondents. A scatter diagram of the results is presented in Fig. 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 can be seen that all 16 points lie

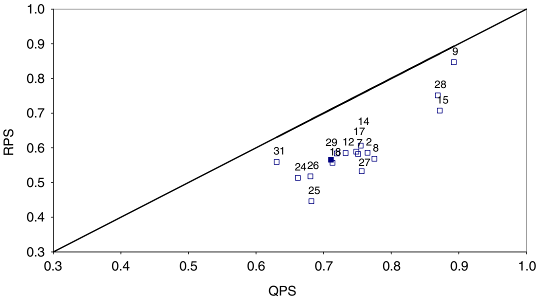

Fig. 3. QPS and RPS for the 16 regular respondents and the SEF average (filled square).

below the "45°" line; thus, the finding in Section 3 for the SEF average forecast, that the RPS is less than the QPS, extends to these individual forecasters as well, for the same general reasons discussed above. The scatter of points is positively sloped, although there is less than perfect agreement between the rankings: the rank correlation coefficient between the QPS and RPS of the regular respondents is 0.76. There are several ambiguous pairwise comparisons: whenever the line joining two points has a negative slope, the QPS and RPS disagree about the relative performances of the corresponding forecasters.

For detailed individual scrutiny, we first pick out individual 26, who is the only ever-present regular respondent, is highly ranked (3rd) on both scores, and is an outlier in one further respect. Whereas almost three-quarters of 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 Fig. 4 shows the observation-by-observation components of the score calculations for individual 26 as in Fig. 1; on the five occasions when inflation fell in the 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). These are of especial interest, because the QPS takes approximately the same value for each of these four observations, 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 the worst, followed by 2006Q4, whereas 2000Q2, Q3 are rather better. A study of the location of the various probabilities forming the histograms shows that this alternative view of the comparative quality of these forecasts is correct, and the QPS's indifference to this question again emphasises its inadequacy as an indicator of the quality of these density forecasts. In the following section we in turn set the QPS aside.

4.2. *Missing data*

For comparison, we include in the lower panel of Fig. 4 the corresponding data for individual 25, who has the best RPS result, as shown in Fig. 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 large 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 one would simply use the first 34 datapoints for both respondents. However, this neglects available information on the performance of the forecaster who has responded more often. Moreover, this is not a practical solution for making multiple comparisons among our 16 regular respondents. Although none of these respondents is missing more than 11 of the 36 possible forecasts, the incidence of missing forecasts shown in Fig. 5 is such 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 to survey practitioners as *item non-response*. Also shown in Fig. 5 are the latest inflation data available at the time the forecasts were prepared (π_{t-1} in Section 3), and there is no evidence of any relationship between the process leading to missing forecasts and this variable, nor any other variable we have considered. Thus, forecasters do not appear to respond differently when the forecasting problem might seem to be more difficult, and no systematic patterns, such as periodicity, are apparent in Fig. 5. Accordingly, we treat the

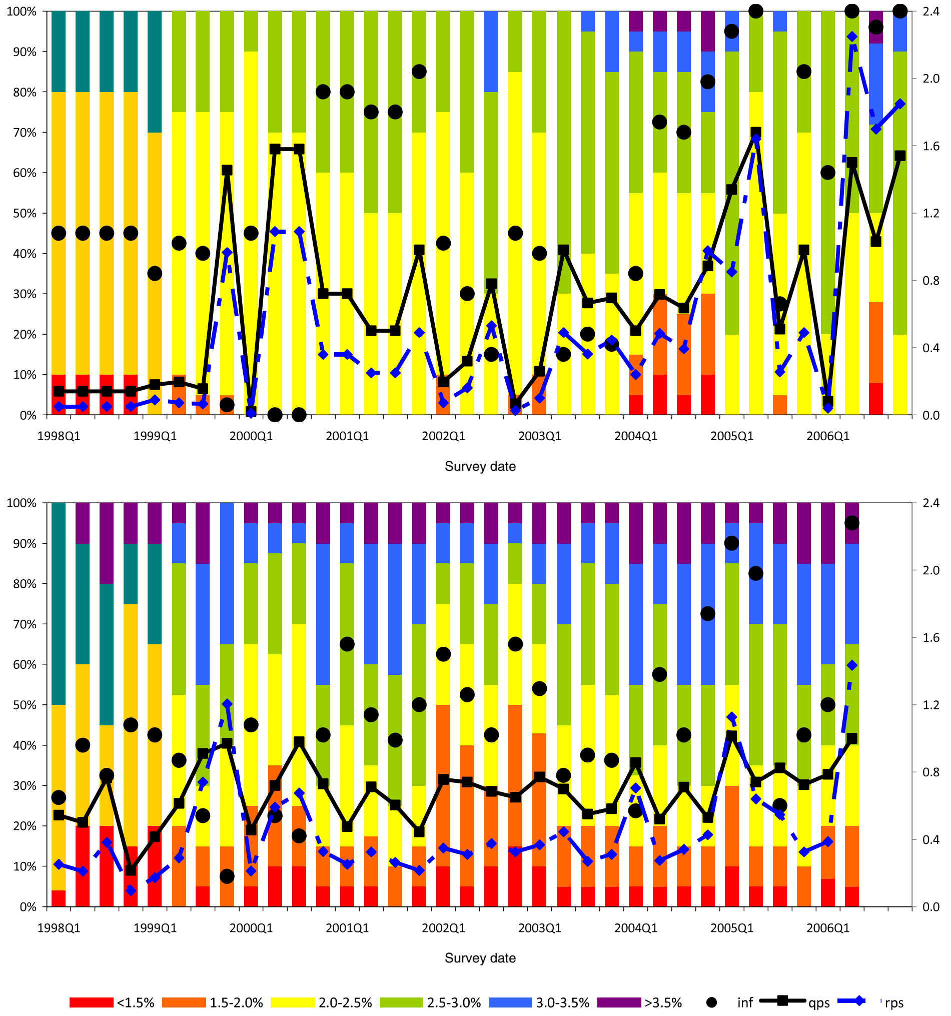

Fig. 4. Forecast probabilities two years ahead, inflation indicators, and the forecast scores (QPS and RPS). Upper panel: individual 26; lower panel: individual 25. *Note:* A coloured version of this figure is available in the online version of the article.

missing data as *missing at random* and the observed data as *observed at random*, using terms introduced by Rubin (see Little & Rubin, 2002). Neither imputation-based methods nor model-based methods for handling incomplete data, as discussed by Little and Rubin, appear to be relevant to the present context of forecast comparison, although we note an interesting application to the construction of a combined point forecast in the face of missing data in the US SPF by Capistran and Timmermann (2009).

Instead, as was discussed at the end of Section 2, we focus on the components of the score that reflect the forecaster performance, by correcting the score for variation in the outcome variance term identified in the Yates decomposition (Eq. (4), generalised 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. (To calculate the variance of *d* or *D* over the full sample, we assume six histogram bins throughout and assign the first five inflation outcomes accordingly, even though those forecasts had only four bins.)

The results are shown in Fig. 6, as a scatter diagram of RPS and adjusted RPS (denoted RPS*) values.

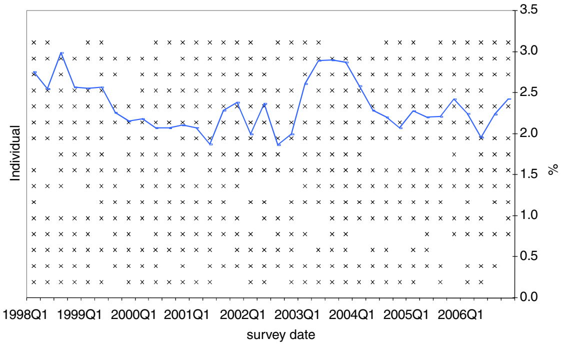

Fig. 5. Incidence of missing two-year-ahead forecasts (blanks) from 16 regular respondents, together with the latest inflation data.

As in Fig. 3, the two scores give different rankings of the forecasters, with a rank correlation coefficient of 0.72. 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 description 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 variances in their respective subsamples and increases their scores, resulting in a more accurate picture of their relative forecast performances. In particular, the adjustment moves individual 25 from the 1st to 4th position in the ranking, and individual 8 from 8th to 14th.

The potential heterogeneity due to the move from four to six bins in the 1999Q2 survey noted above may have a greater impact on the individual comparisons, since Fig. 5 shows that some individuals are missing more of the four-bin observations than others. As a robustness check, we repeat the calculations (a) assuming four bins throughout, or (b) deleting the first five observations. In the first case, all of the scores are reduced, as expected, whereas the second case gives corresponding slight increases in all scores. Overall, however, the results appear to be robust: the cloud of points has the same shape in all cases, and the few changes in rankings that are observed are confined to the centre of the distribution, where the individual scores are close together.

For a final illustration at the individual level, in Fig. 7 we present the data for the two respondents whose scores are decreased most as a result of the adjustment. Individual 9, in the upper panel, has the same number of missing observations — ten — as individual 8, but these correspond to outcomes which 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. However, individual 9 is still ranked last, as a result of the excessive dispersion of the forecast histograms, and 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 Fig. 7), who has eleven missing observations similarly distributed, the adjustment changes the ranking considerably, 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

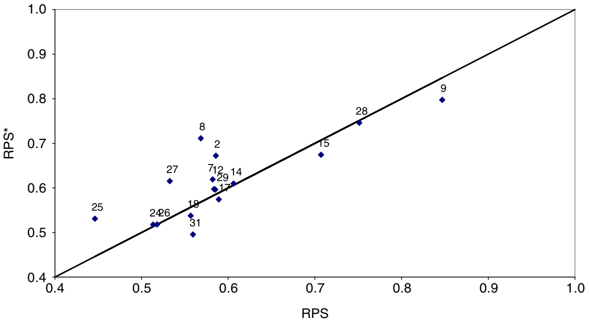

Fig. 6. RPS and RPS* for the 16 regular respondents.

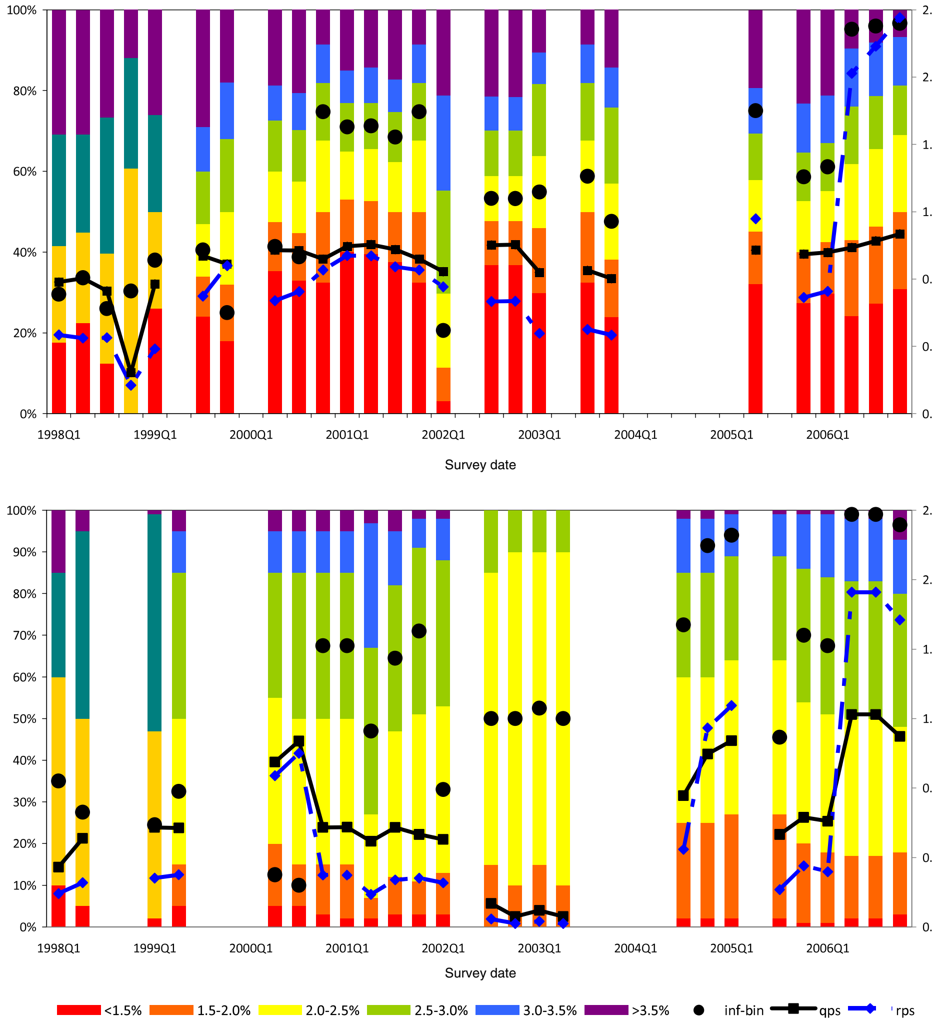

Fig. 7. Forecast probabilities two years ahead, inflation indicators, and the forecast scores (QPS and RPS). Upper panel: individual 9; lower panel: individual 31. *Note:* A coloured version of this figure is available in the online version of the article.

in the lower open-ended bin, and the latter’s relative scores benefited from this choice, except in 2000Q2 and 2000Q3.

The overall effect of these adjustments for missing data is to reduce the dispersion of the individual scores. Part of the dispersion in the unadjusted scores is seen to be the result of differential non-response, over and above differences in forecasting performances. The var(D_k) component of the individual RPS given by the Yates decomposition is outside the forecaster’s influence, and, assuming that this is 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 this component gives a more reliable ranking of the individual forecast performances. There remains a considerable dispersion in the RPS* scores, however, and this heterogeneity in individual density forecasting performances mirrors the finding of a considerable degree of heterogeneity in point forecasting performances in the SEF by Boero et al. (2008b).

5. Conclusion

This article 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, as a set of probabilities that the 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. Epstein's ranked probability score was initially proposed as an alternative to Brier's quadratic probability score for precisely these circumstances, and our exercise makes its advantages clear. The logarithmic score is the leading alternative to the quadratic scoring rules, but, unlike them, is not defined whenever inflation falls in a histogram bin to which the forecaster has assigned a zero probability. Such situations occurred in our sample of individual forecasters, and thus exclude the logarithmic score from consideration in this context.

Missing observations are endemic in surveys, and our answer to this problem comes in two parts. First, in common with much other research on forecast surveys, our study of individual forecast performances 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 a considerable amount of dispersion, part of which is due to differences in the inflation outcomes over the different subperiods for which these respondents provided their forecasts. Accordingly, and secondly, we introduce 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, not on the forecast, and hence gives a clearer ranking of forecaster performance. We recommend the adjusted ranked probability score, denoted RPS*, to other analysts of forecast surveys, in different countries and at different forecast horizons, who nevertheless face the familiar problem of non-response.

Attention in Section 4 of this article is restricted to descriptive comparisons and rankings of competing forecasts, without formal testing. Extensions of the pairwise test used in Section 3 to multiple comparisons, using Bonferroni intervals or other methods (see Miller, 2006), keeping in mind the small-sample context, await future research.

The analysis of the point forecasts of inflation and GDP growth from the SEF in our earlier article in this journal (Boero et al., 2008b) finds a considerable degree of heterogeneity among individual respondents, as shown by the failure of standard tests of the equality of idiosyncratic error variances and by evidence of different degrees of asymmetry in forecasters' loss functions. The 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 article with the same final thought as our earlier article, that the findings "prompt questions about the individual forecasters' methods and objectives, the exploration of which would be worthwhile".

Acknowledgements

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

References

Amisano, G., & Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, *25*, 177–190.

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

Boero, G., Smith, J., & Wallis, K. F. (2008b). Evaluating a three-dimensional panel of point forecasts: the Bank of England Survey of External Forecasters. *International Journal of Forecasting*, *24*, 354–367.

Boero, G., Smith, J., & Wallis, K. F. (2008c). Here is the news: forecast revisions in the Bank of England Survey of External Forecasters. *National Institute Economic Review*, *203*, 68–77.

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.

Candille, G., & Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, *131*, 2131–2150.

Capistran, C., & Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business and Economic Statistics*, *27*, 428–440.

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

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

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

Galbraith, J. W., & 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.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*, 1545–1578.

Gneiting, T., & 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, Series B*, *14*, 107–114.

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

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Miller, R. (2006). Multiple comparisons. In N. Balakrishnan, C. B. Read, & B. Vidakovic (Eds.), *Encyclopedia of statistical sciences* (2nd ed.) (pp. 5055–5065). Hoboken, NJ: Wiley-Interscience.

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

Murphy, A. H. (1972). Scalar and vector partitions of the probability score: part II. N-state situation. *Journal of Applied Meteorology*, *11*, 1183–1192.

Murphy, A. H. (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.

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

Tay, A. S., & Wallis, K. F. (2000). Density forecasting: a survey. *Journal of Forecasting*, *19*, 235–254.
Reprinted in Clements, M. P., & Hendry, D. F. (Eds.) (2002). *A companion to economic forecasting* (pp. 45–68). Oxford: Blackwell.

Wallis, K. F. (2004). An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review*, *189*, 64–71.

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

Yates, J. F. (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.