Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts

Michael P. Clements*
Department of Economics,
University of Warwick

October 7, 2003

Abstract

Techniques are proposed for evaluating forecast probabilities of events. The tools are especially useful when, as in the case of the SPF expected probability distributions of inflation, recourse can not be made to the method of construction in the evaluation of the forecasts. The tests of efficiency and conditional efficiency are applied to the forecast probabilities of events of interest derived from the SPF distributions, and supplement a whole-density evaluation of the SPF distibutions based on the probability integral transform approach.

Journal of Economic Literature classification: C53.
Keywords: Density forecasts, event probabilities, encompassing, SPF inflation forecasts.

*Financial support from the U.K. Economic and Social Research Council under grant L138251009 is gratefully acknowledged. Helpful comments were received from David Harvey, Ken Wallis and seminar participants at Monash, Reading and Southampton Universities. Computations were performed using code written in the the Gauss programming language (Aptech Systems) and Givewin 2 and PcGive 10: see Doornik and Hendry (2001).

Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts

Abstract

Techniques are proposed for evaluating forecast probabilities of events. The tools are especially useful when, as in the case of the SPF expected probability distributions of inflation, recourse can not be made to the method of construction in the evaluation of the forecasts. The tests of efficiency and conditional efficiency are applied to the forecast probabilities of events of interest derived from the SPF distributions, and supplement a whole-density evaluation of the SPF distibutions based on the probability integral transform approach.

Journal of Economic Literature classification: C53.
Keywords: Density forecasts, event probabilities, encompassing, SPF inflation forecasts.

1 Introduction

In this paper we propose a number of techniques for evaluating probability event forecasts. These evaluation tools are viewed as complementing existing ways of evaluating probability distributions of macroeconomic variables. Our interest focuses on data sets comprised of a relatively small number of probability distributions, such as the probability distributions of expected inflation from the Survey of Professional Forecasters (SPF). The nature of these data sets affects the ways in which the forecasts can be evaluated. For instance, the SPF density forecasts are produced by averaging the histograms of a number of respondents, so that it appears unlikely that any evaluation of the forecasts will make recourse to the method of construction. By way of contrast, the approach of Li and Tkacz (2001) compares the conditional density function of a particular parametric model to a non-parametric estimate of the conditional density function. The parametric model determines the conditioning variables in the non-parametric estimate of the true (given those conditioning variables) conditional density.

A popular way of evaluating these probability distributions employs the probability integral transform described by Diebold, Gunther and Tay (1998) (but which has a lineage going back at least to Rosenblatt (1952)). As an example, Diebold, Tay and Wallis (1999) apply this approach to the SPF inflation forecasts for the period 1969 to 1995. The probability integral transform evaluates the whole forecast distribution. Diebold *et al.* (1998) and Granger and Pesaran (2000a) establish that a density forecast that coincides with the data generating process will be optimal in terms of minimizing expected loss whatever the loss function of the user, whereas in general rankings between rival forecasts will not be invariant to the user's loss function. This provides a strong case for evaluating the whole forecast density – a forecast density that provides a close match to the true density can be used by all with equanimity, no matter what their individual loss functions. However, a rejection of the forecast densities may not render the forecasts of no value for the purpose at hand. Just as financial risk management tends to focus on a tail quartile of the expected distribution of returns of financial assets (the Value at Risk: VaR), users of inflation forecasts may be primarily interested in whether the probabilities assigned to inflation falling in a certain key range are reasonably accurate.[1] A density forecasts may be well calibrated over the range(s) of interest but be rejected overall.

As well as being of interest in their own right, an assessment of the accuracy of the implied probabilities of certain events (inflation falling in a certain range) may also be informative about the reasons for the 'whole-density' rejection, thereby operating in a constructive manner. Event probability forecasts are intimately related to interval forecasts (commonly also called prediction intervals). Event probability forecasts assign a probability to inflation falling in a pre-specified range. Interval forecasts specify a range for a given probability. This relationship will motivate one of the evaluation tools. A density forecast can be viewed as being comprised of a sequence of intervals forecasts generated by allowing the nominal coverage rate to vary over all values in the unit interval. The evaluation of a sequence of interval forecasts with a specific nominal coverage rate therefore assesses one aspect of the underlying sequence

[1] Since the Spring of 1997, the Monetary Policy Committee in the UK has been charged with delivering an inflation rate of ± 1 percentage points around a target of $2\frac{1}{2}\%$. Whether formalised or not, considerations of this type are an integral part of most Western countries' macroeconomic stabilisation policies.

of forecast densities (such as the VaR in risk management exercises[2]). Similarly, the evaluation of event probabilities assess the forecast densities over particular ranges.

Having accurate forecast probabilities of events of interest may be more important than the forecast density being correctly calibrated throughout its range. To this end, we propose tests of the efficiency of probability forecasts of events derived from the SPF density forecasts. We also propose the use of forecast encompassing tests as a check on the 'conditional efficiency' of the SPF event probability forecasts relative to rival sets of forecasts. Both sets of tests are applicable to event probability forecast evaluation in general.

The plan of the paper is as follows. Section 2 provides a brief review of the probability integral transform approach, and section 3 sets out in detail the event probability forecast evaluation tools we propose. Section 4 describes the nature of the SPF forecasts, and the results of the evaluation using the probability integral transform approach. Section 5 describes the application of the event probability evaluation techniques to the SPF forecasts, and section 6 concludes.

2 Probability distribution forecast evaluation

The key idea of the probability integral transform approach popularised by Diebold *et al.* (1998) is the following. Suppose we have a series of 1-step forecast densities for the value of a random variable $\{Y_t\}$, denoted by $p_{Y,t|t-1}(y)$, where $t = 1, \ldots, n$. The probability integral transforms (pits) of the realizations of the variable with respect to the forecast densities are given by:

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

for $t = 1, \ldots, n$, where $P_{Y,t|t-1}(y_t)$ is the forecast probability of Y_t not exceeding the realized value y_t. When the forecast density equals the true density, $f_{Y,t|t-1}(y)$, it follows that $z_t \sim U(0,1)$. Even though the actual conditional densities may be changing over time, provided the forecast densities match the actual densities at each t, then $z_t \sim U(0,1)$ for each t, and the z_t are independently distributed, such that the time series $\{z_t\}_{t=1}^n$ is independently identically uniform distributed, i.e., $iidU(0,1)$.

Evaluating the forecast densities by assessing whether $\{z_t\}_{t=1}^n$ is $iidU(0,1)$ thus requires testing the joint hypothesis of independence and uniformity. Independence can be assessed by examining correlograms of $\{z_t\}_{t=1}^n$, and of powers of this series (as a check for dependence in higher moments, which would be incompatible with the independence claim), and formal tests of autocorrelation can be performed. Uniformity can also be assessed in a number of ways: whether the empirical cdf of the $\{z_t\}$ is significantly different from the theoretical uniform cdf (a $45°$ line) using e.g., the Kolmogorov Smirnov (KS) test of whether the maximum difference between the two cdfs exceeds some critical value. The effect of a failure of independence on the distribution of the test statistic is unknown, and could be exacerbated by a small sample size. Moreover tests of autocorrelation will be affected by failure of the uniformity assumption. Graphical analyses are often reported as an adjunct to formal tests of the two parts of the joint hypothesis.

[2] See Lopez (1996) for a discussion of the relationship between VaR analysis and interval forecasting.

Other ways of testing probability distributions are given in Thompson (2002), who suggests a frequency domain test of the uncorrelatedness of the $\{z_t\}$ based on the cumulative periodogram approach of Durbin (1969), and the generalized spectral approach of Hong (2001). To aid comparison with the earlier work on the SPF inflation forecasts, and because our primary interest is on event probability forecast evaluation, we will not consider the frequency domain approaches.

3 Event probability forecast evaluation

3.1 Regression-based tests of event probabilities

Let I_t be the event indicator that takes the value unity when the event occurs in period t, and zero otherwise. An event could be inflation being in a certain range, for example. The probabilities p_t attached to the event in each period are calculated by linear interpolation of the the SPF histograms. The evaluation of a sequence of event probability forecasts $\{p_t\}_{t=1}^{n}$ requires an assessment of probabilities which typically vary over time while the ranges defining the events are fixed. Whereas for sequences of interval forecasts the nominal coverage level is fixed and the range defining the 'event' varies over t. The close relationship between the two permits interval evaluation tests to be adapted to evaluate event probability forecasts.

Christoffersen (1998, p. 849–50) and Engle and Manganelli (1999) present regression-based tests of interval forecasts based on:

$$I_t = \alpha + \beta' W_{t-1} + \epsilon_t, \quad t = 1, \ldots, n \tag{2}$$

as a way of testing 'conditional forecast efficiency'. I_t is the 'hit sequence', and equals unity when the t^{th} interval contains the actual, and W_{t-1} is a vector of variables known at $t-1$. W_{t-1} will typically include the lagged values $\{I_{t-1}, I_{t-2,} \ldots\}$. Conditional efficiency is the requirement that a sequence of forecasts has correct conditional coverage, $E\left(I_t | W_{t-1}\right) = p$ for all t. This can in turn be viewed as requiring that the ex post coverage of the set of forecasts equals the nominal coverage rate (correct unconditional coverage, $E\left(I_t\right) = p$) and that hits (alternatively, misses) are not associated with other variables, or combinations of other variables, W_{t-1}. Conditional efficiency requires that £ $\alpha\ \beta'$ ¤ $= [p\ 0']$. A rejection of $\beta = 0$ would signify that the likelihood of a hit varies systematically with information known at the time the interval forecast was made.

A simple way of testing the conditional efficiency of event probability forecasts $\{p_t\}_{t=1}^{n}$ is to include p_t as an explanatory variable in (2), so that now interpreting $\{I_t\}$ as the event indicator we obtain:

$$I_t = \gamma p_t + \beta W'_{t-1} + \epsilon_t, \quad t = 1, \ldots, n \tag{3}$$

The null hypothesis of conditional efficiency is that $E\left(I_t \mid W_{t-1}\right) = p_t$ for all t, which requires that £ $\gamma\ \beta'$ ¤ $= [1\ 0']$. When W_{t-1} simply consists of an intercept, (3) is the realization-forecast regression of Mincer and Zarnowitz (1969). Here the realization $\{I_t\}$ is binary, and the forecast is a probability. Nevertheless, $[\gamma\ \beta] = [1\ 0]$ is a sufficient condition[3] for unbiasedness and also implies that the forecast

[3] See Holden and Peel (1990).

errors $\{I_t - p_t\}$ are uncorrelated with the forecasts.[4] In the context of equation (2) for interval forecasts, Clements and Taylor (2003) note that the binary nature of the dependent variable $\{I_t\}$ suggests fitting a regression model to a logistic transformation of the dependent variable. In terms of the event forecast evaluation regression (3) the logit model is:[5]

$$\Pr\left(I_t = 1\right) = \Lambda(\gamma, \beta; \alpha_t), \quad t = 1, \ldots, n \tag{5}$$

where

$$\Lambda(\gamma, \beta; \alpha_t) = e^{\gamma\alpha_t + \beta}/(1 + e^{\gamma\alpha_t + \beta})$$

and $\alpha_t = \ln(p_t/(1 - p_t))$. The transformation of p_t given by α_t gives $\Pr\left(I_t = 1\right) = p_t$ under the null $[\gamma\ \beta] = [1\ 0]$ as required.

In principle W_{t-1} could include $\{I_{t-1}, I_{t-2}, \ldots p_{t-1}, \ldots\}$ etc., but to aid interpretation we first test with $W_{t-1} = 1$ only in (3) using a logit regression, and then expand the information set to test whether other variables help explain the forecast error $e_t \equiv I_t - p_t$ in regressions such as:

$$I_t - p_t = \beta W_{t-1} + error.$$

Plausible candidate variables in the inflation example are import price inflation to capture the oil price shocks of the seventies, and the unemployment rate as suggested by textbook Phillips Curve models. We will consider instead the information contained in rival sets of probability forecasts, and to that end we consider tests of forecast encompassing in the next section.

3.2 Forecast encompassing tests of probability forecasts

One set of forecasts $\{f_{1t}\}$ is said to forecast encompass another $\{f_{2t}\}$ if the latter contains no useful information not already present in $\{f_{1t}\}$, in the mean-squared error sense that a linear combination of f_{1t} and f_{2t} (with non-zero weight accorded to f_{2t}) has a mean-squared forecast error (MSFE) no smaller than that of f_{1t}. The notion of forecast encompassing is more stringent than the requirement that $\{f_{1t}\}$ has a smaller MSFE than $\{f_{2t}\}$ – $\{f_{1t}\}$ could have a smaller MSFE than $\{f_{2t}\}$ but nevertheless a combination of the two could have a smaller MSFE than $\{f_{1t}\}$.[6]

Forecast encompassing is a natural way to compare rival sets of forecasts. In this section we investigate the validity of tests of forecast encompassing when the forecasts are probability forecasts. We

[4] From (3) with $W_{t-1} = 1$:

$$\left(I_t - p_t\right) = \left(\gamma - 1\right) p_t + \beta + \varepsilon_t$$

and so:

$$E\left[\left(I_t - p_t\right) p_t\right] = \left(\gamma - 1\right) E\ {}^{\text{¡}}p_t^2{}^{\text{¢}} + \beta E\left(p_t\right). \tag{4}$$

[5] When the events cover all possible outcomes the adequacy of the event probability forecasts can be assessed using the multinomial logit model for unordered multi-responses, as described by Patton (2002) working within a similar framework.

[6] There is an extensive literature on forecast encompassing: see Diebold and Lopez (1996) and Newbold and Harvey (2002) for recent surveys, and Chong and Hendry (1986), Clemen (1989), Newbold and Granger (1974) and Stock and Watson (1999) *inter alia*. Forecast encompassing is formally equivalent to the notion of conditional efficiency introduced by Nelson (1972) and Granger and Newbold (1973).

define the forecast errors as $e_{it} = y_t - f_{it}$, $i = 1, 2$, where $\{y_t\}_{t=1}^{n}$ is the sequence of outcomes, and f_{it} is the forecast made of period t at period $t-1$, so the forecasts are 1-step ahead. The forecast error of the combined forecast $f_{ct} = (1-\lambda) f_{1t} + \lambda f_{2t}$ is given by ε_t, so that rearranging implies that the null hypothesis that forecast f_{2t} contains no useful information that is not already present in f_{1t} is given by $\lambda = 0$ in the OLS regression:

$$e_{1t} = \lambda (e_{1t} - e_{2t}) + \varepsilon_t. \tag{6}$$

That is, the expected-squared error of the combined forecast is minimized by $\lambda = 0$, so that no weight is accorded to f_{2t}. The form of this regression implies that the individual forecasts are unbiased ($E(e_{it}) = 0$, $i = 1, 2$) otherwise a constant term would be required in the combination and in (6). The weights on the individual forecasts are restricted to sum to unity, and we restrict λ to $0 \leq \lambda \leq 1$ to rule out forecasts being given negative weights (and weights in excess of one). The alternative hypothesis is therefore one-sided: $\lambda > 0$. The null of $\lambda = 0$ corresponds to $E(e_{1t}, e_{1t} - e_{2t}) = 0$ against the one-sided alternative $E(e_{1t}, e_{1t} - e_{2t}) > 0$.

Under standard assumptions about forecast errors it follows from Diebold and Mariano (1995) that the t-statistic that $\lambda = 0$ in (6) has an asymptotic standard normal distribution. However, Harvey, Leybourne and Newbold (1998) (henceforth HLN) show that when the forecast errors are conditionally heteroscedastic ($E\left[e_{1t}^2 \mid e_{1t} - e_{2t}\right] = g(e_{1t} - e_{2t})$) the standard test will be incorrectly-sized, and they propose the use of heteroscedasticity-robust methods as well as a number of modifications to improve the small-sample performance.

Probability forecasts impose bounds on the ranges of the forecast errors and the disturbance term ε_t in (6). Because y_t is binary and $f_{it} \in (0, 1)$, then $e_{it} \in (-1, 1)$, and under the null $\varepsilon_t \in (-1, 1)$. Typically, probability forecasts will also be characterised by conditional heterocedasticity. To illustrate, consider the data generating process given by:

$$y_t = 1(u_{1t} > v_t), \;\; f_{1t} = u_{1t}, \;\; f_{2t} = u_{2t}, \tag{7}$$

where u_{1t}, u_{2t}, and v_t are independent $U(0, 1)$ random variables. By construction, f_{1t} forecast encompasses f_{2t} because f_{2t} is independent of y_t. This can be seen from $E(e_{1t}, e_{1t} - e_{2t}) = 0$:

$$\begin{aligned} E(e_{1t}, e_{1t} - e_{2t}) &= E\left[(1(u_{1t} > v_t) - u_{1t})(u_{2t} - u_{1t})\right] \\ &= E\left[u_{1t}^2\right] - E\left[u_{1t} 1(u_{1t} > v_t)\right] - E(u_{1t} u_{2t}) + E\left[u_{2t} 1(u_{1t} > v_t)\right] \\ &= 1/3 - 1/3 - 1/4 + 1/4 = 0 \end{aligned}$$

using standard results pertaining to independent $U(0, 1)$ random variables.

The probability forecast errors are also characterised by conditional heteroscedasticity whereby $E\left[e_{1t}^2 \mid e_{1t} - e_{2t}\right]$ depends on $(e_{1t} - e_{2t})^2$. To establish the presence of heteroscedasticity of this form, consider:

$$e_{1t}^2 = \zeta_0 + \zeta_1 (e_{1t} - e_{2t})^2 + \nu_t$$

where ζ_0 is an intercept, and $E\left[\nu_t \mid (e_{1t} - e_{2t})^2\right] = 0$. Heteroscedasticity is present when $\zeta_1 \neq 0$, where:

$$\zeta_1 = \frac{Cov\left[(e_{1t} - e_{2t})^2, e_{1t}^2\right]}{Var\left[(e_{1t} - e_{2t})^2\right]}. \tag{8}$$

Table 1 Monte Carlo estimates of sizes of forecast encompassing tests of probability forecasts for the artificial data generation process.

n	R	R_1	R_2	DM	MDM	$r_{s,1}$	$r_{s,2}$
8	2.76	3.50	0.28	3.14	1.17	2.98	4.01
16	2.95	2.75	1.24	3.10	1.86	2.82	5.18
32	2.85	2.75	2.05	2.98	2.38	2.07	6.34
64	2.90	3.15	2.79	3.22	2.97	1.23	8.17
512	3.24	4.25	4.20	4.26	4.23	0.05	35.77
10,000	3.21	4.83	4.83	4.83	4.83	0.00	100.00

The table records the Monte Carlo rejection frequencies for the true null of forecast encompassing for the data generation process given by (7), based on $40,000$ replications, and for a nominal test size of 5%.

R is the standard t-statistic for $\lambda = 0$; R_1 employs a correction for heteroscedasticity; R_2 an alternative estimator of the denominator of the t-statistic; DM is the Diebold-Mariano test for equal forecast accuracy applied to testing for forecast encompassing; MDM is a modified version of DM; and $r_{s,1}$ is Spearman's rank correlation test against positive correlation, and $r_{s,2}$ is Spearman's rank correlation test implemented as a two-sided tests.

We can replace e_{1t} and e_{2t} by u_{1t}, u_{2t} and v_t, and then substitute numerical values for the resulting moments (and conditional) moments of $U\,(0,1)$ variables. ζ_1 can be calculated rather more simply by simulation, in which case we obtain $\zeta_1 \simeq -0.14$ for the data generation process given by (7).

Table 1 records the results of a small Monte Carlo investigation on the sizes of the standard and HLN-modified forecast encompassing tests for the probability-forecast data generation process given by (7). In brief, the test statistics we consider are: the standard t-statistic for $\lambda = 0$ (R); as R but with heteroscedasticity-consistent-standard-errors (White (1980)) (R_1); as R_1 but with an alternative estimator of the denominator of the t-statistic (R_2); the Diebold and Mariano (1995) test for equal forecast accuracy applied to testing for forecast encompassing (DM); the Harvey, Leybourne and Newbold (1997) modifications to the Diebold-Mariano test for equal forecast accuracy applied to the DM test for forecast encompassing; Spearman's rank correlation test ($r_{s,1}$ and $r_{s,2}$). R and DM are compared to a standard normal, and R_1, R_2 and MDM to a Student t_{n-1}. All are implemented as one-sided tests. Spearman's rank correlation test is a distribution free test that determines whether there is a monotonic relation between two variables, here e_{1t} and $(e_{1t} - e_{2t})$, and is advocated by HLN for 1-step forecasts where it is reasonable to assume that drawings of $\{e_{1t}, (e_{1t} - e_{2t})\}$ are independent. We report a one-sided rank correlation test against the alternative of positive correlation ($r_{s,1}$), and a two-sided test ($r_{s,2}$), for reasons which will become apparent. Details of these tests are provided by HLN.

The table indicates that R_1, R_2, DM and MDM are all correctly-sized asymptotically for tests of probability forecasts, where $n = 10,000$ is taken to correspond to $n = \infty$. All four are under-sized in small-samples, especially R_2 and MDM, although the size distortions to R_1 and DM are not too serious for $n = 32$, which corresponds to the size of the sample of SPF forecasts. Interestingly, MDM is worse than DM. As expected given the conditional heteroscedasticity, the standard test R appears to be undersized even as n gets large, but by a relatively small amount. Perhaps most strikingly, the rank correlation tests are of little use on this type of data. The one-sided test against positive correlation ($r_{s,1}$)

Table 2 Monte Carlo estimates of size and power of forecast encompassing tests of probability forecasts for the ARCH(1) data generation process. .

n	Null hypothesis	R	R_1	R_2	DM	MDM
8	$f_{arch,t}$ FE $f_{r,t}$	3.61	5.16	0.55	3.90	1.73
	$f_{arch,t}$ FE $f_{nc,t}$	11.44	16.27	7.03	13.79	10.33
	$f_{nc,t}$ FE $f_{arch,t}$	17.85	21.80	2.49	17.09	7.92
16	$f_{arch,t}$ FE $f_{r,t}$	3.48	3.54	1.41	3.28	2.11
	$f_{arch,t}$ FE $f_{nc,t}$	7.38	9.79	5.91	8.39	6.92
	$f_{nc,t}$ FE $f_{arch,t}$	34.69	33.76	16.39	29.07	21.66
32	$f_{arch,t}$ FE $f_{r,t}$	3.33	3.27	2.29	3.27	2.67
	$f_{arch,t}$ FE $f_{nc,t}$	6.66	7.28	5.45	6.62	5.93
	$f_{nc,t}$ FE $f_{arch,t}$	60.61	57.66	46.77	53.92	49.80
64	$f_{arch,t}$ FE $f_{r,t}$	3.36	3.41	2.90	3.44	3.13
	$f_{arch,t}$ FE $f_{nc,t}$	6.34	6.12	5.21	5.86	5.45
	$f_{nc,t}$ FE $f_{arch,t}$	88.13	86.78	83.50	85.56	84.30
512	$f_{arch,t}$ FE $f_{r,t}$	3.42	4.20	4.13	4.19	4.15
	$f_{arch,t}$ FE $f_{nc,t}$	6.09	5.11	4.99	5.07	5.02
	$f_{nc,t}$ FE $f_{arch,t}$	100.00	100.00	100.00	100.00	100.00
10,000	$f_{arch,t}$ FE $f_{r,t}$	3.62	4.80	4.79	4.79	4.79
	$f_{arch,t}$ FE $f_{nc,t}$	6.24	5.08	5.07	5.08	5.07
	$f_{nc,t}$ FE $f_{arch,t}$	100.00	100.00	100.00	100.00	100.00

The table records the Monte Carlo rejection frequencies for the null of forecast encompassing for different pairs of forecasts, when the data generation process is given by (9). The calculations are based on $40,000$ replications, and are for a nominal test size of 5%.

The test statistics are explained in the text and in the notes to table 1.

$f_{arch,t}$ FE $f_{r,t}$ is the hypothesis that $f_{arch,t}$ forecast encompasses $f_{r,t}$, etc., where $f_{arch,t}$ are the true forecasts, $f_{r,t}$ contain no useful information by construction. $f_{nc,t}$ is the no-change forecast.

is the most natural as it tests against a positive weight on the rival forecast, but is undersized at $n = 32$, and more pertinently, the size goes to zero as n gets large. The reason is evident from the behaviour of the two-sided rank correlation test $r_{s,2}$, for which the size goes to one, indicating negative correlations in the ranks.

Given that R_1, R_2, DM and MDM have reasonable size properties for tests of probability forecasts, we next carried out a Monte Carlo study to establish the powers of these tests for a data generation process and sets of probability forecasts that match the inflation forecasts in certain respects. The data generation process for inflation (y_t) is an AR(1)-ARCH(1) model,[7] approximately calibrated on the US

[7] See e.g., Engle (1982).

annual data:

$$\begin{aligned} y_t &= 0.5 + 0.8 y_{t-1} + \varepsilon_t \\ \varepsilon_t &= z_t \sqrt{h_t} \\ h_t &= 0.4 + 0.85 \varepsilon_{t-1}^2 \end{aligned} \tag{9}$$

where z_t is a standard normal random variable. We consider probability forecasts of the event that inflation falls within the target range $1\frac{1}{2}$ to $3\frac{1}{2}\%$. The 'true' probability forecasts are denoted by $f_{arch,t}$ and are calculated from the 1-step ahead conditional gaussian densities implied by (9) using the actual values of the coefficients in the conditional mean and variance functions.[8] There is no suggestion that this process is a good representation of the way in which the SPF forecast densities are constructed. All we require is that it captures some of the salient features, such as changing conditional means and variances, so that we can explore the powers of the forecast encompassing tests against rival probability forecasts of the sort used in the empirical study, where these aspects are absent or otherwise incorrectly specified. We calculate 'no-change' $f_{nc,t}$ probability forecasts from conditional gaussian densities with means equal to the realised rates of inflation in the previous periods, and variances equal to the sample variances of the $t-4$ to $t-1$ observations (for a forecast of period t). A rationale for considering forecasts of this type is given in section 5. In addition, we use a rival forecast $f_{r,t} = u_t$, where $u_t \sim U(0,1)$ and $\{u_t\}$ is independent of $\{z_t\}$.

Table 2 reports the results of the Monte Carlo based on the data generation process (9), for $n = \{8, 16, 32, 64, 512, 10000\}$, as before, where again $n = 10,000$ mimics the asymptotic case. For each n, the first test is of whether the true probability forecasts encompass randomly-generated forecasts which are unrelated to y_t. This serves as a check – the rejection frequencies for these rows correspond to sizes, and as expected replicate closely the estimates recorded in table 1. Notice that the rejection frequencies for the null hypotheses that f_{arch} encompasses f_{nc} are in excess of 5% for small n, as the f_{nc} forecast probabilities are correlated with the occurrence of the event. The rejection frequencies of the four modified tests of f_{nc} encompassing f_{arch} are around 50 to 60% for $n = 32$, suggesting that these tests have reasonable power for samples of the size of the SPF forecasts.

4 The SPF probability distributions of expected future inflation

The SPF[9] is a quarterly survey of macroeconomic forecasters of the US economy that began in 1968 as the ASA-NBER survey, administered by the American Statistical Association and the National Bureau of Economic Research, and since June 1990 has been run by the Philadelphia Fed, under its current

[8] West (2001) considers the effects of parameter estimation uncertainty on tests for forecast encompassing: see West (1996), West and McCracken (1998) and West and McCracken (2002) on the impact of parameter estimation uncertainty on other tests of predictive accuracy. Because the SPF probability distributions are not based on a simple model known to the econometrician, it is difficult to see how an allowance could be made for the uncertainty in the respondents' probability assessments, whether this emanates from parameter estimation uncertainty or some other source.

[9] Detailed information on the survey as well as the survey results are available at thURL `http://www.phil.frb.org/econ/spf`. An academic bibliography of articles that either discuss or use data generated by the SPF is also maintained online.

name. The majority of the survey questions ask respondents to report their point forecasts for a number of variables at various forecast horizons, from which median forecasts are calculated, but respondents are also asked to report discrete probability forecasts, or histograms, for output growth and inflation for the current and next year, which are then averaged to produce the density forecasts.

Diebold *et al.* (1999) discuss the survey and the complications that arise in using the inflation forecasts. In order to obtain a non-overlapping series of forecasts – in the sense that the realization of inflation in period t is known before making the next forecast of $t+1$ at period t – they take the density forecasts made in the first quarter of each year of the annual change in that year on the preceding year. This avoids the counterpart of the well-known problem in the point forecast evaluation literature that a sequence of optimal h-step ahead forecasts, where the forecasting interval is one period, will follow a moving-average process of $h-1$. Further complications are that both the base years of the price indices and the indices themselves have changed over time. The change in base years is likely to have had a minor effect on the inflation rate, and we construct a series of realizations of annual inflation that matches the indices for which probability assessments were requested. Thus, for 1969 to 1991 we use the implicit GNP deflator, for 1992 to 1995 the implicit GDP deflator, and for 1996 to 2002 the chain-weighted deflator, correcting for the changes in the definition of the index but not for base-year changes. Moreover, we use the latest available estimates of the realized values.[10]

Finally, as documented by the Philadelphia Fed, the form in which the respondents report their probability assessments has changed over time, with changes in the number of bins and/or their locations and lengths as the perceived likely ranges of the target variables has changed. However, this complication is minor because for the most part we will want to read off probabilities of certain values, and the values that define given probabilities, and both can be achieved by piecewise linear approximation – this approximation 'undoes' the discretization in the histogram.[11]

Figure 1 portrays the inflation density forecasts as Box-Whisker plots along with the realizations. The observations for 1969 to 1996 closely match Diebold *et al.* (1999), and are discussed in detail by those authors. We note that the forecasts and realizations for 1997 and 1998 indicate a continuation of the tendency from the early part of the decade to both over-estimate the uncertainty and level of inflation. The forecast distributions appear too dispersed and the central tendencies consistently indicate higher inflation rates than actually materialise. Inflation is low and falling up until the end of the decade.

Table 3 presents the results of the formal 'whole density' tests of the probability integral transforms $\{z_t\}$. The KS test statistic offers no evidence against the null hypothesis of uniformity. Rather than testing (powers of) the $\{z_t\}$ series for autocorrelation, Berkowitz (2001) suggests taking the inverse

[10] The series were taken from the Federal Reserve Bank of St Louis database (FRED), available at the URL `http://www.stls.frb.org/fred/data/` and have the codes GNPDEF, GDPDEF and GDPCTPI.

[11] As an example, suppose we wish to calculate the forecast probability of observing a value less than $Y = 3.5$. Suppose $\Pr(Y < 2)$ is 0.5, and the bin defined by $[2, 4)$ has a probability of 0.2. Then:

$$\Pr(Y < 3.5) = \Pr(Y < 2) + \frac{1.5}{2}\Pr(Y \in [2,4)) = 0.5 + \frac{1.5}{2}0.2 = 0.65.$$

Linear interpolation follows the assumption implicit in the histogram – that probability mass is uniform within a bin. If a bin is bordered by a high probability bin and a relatively low probability bin, it might be desirable to attach higher probabilities to points near the boundary with the high probability bin.

Table 3 Tests of SPF density forecasts of inflation ((1969–2002)) based on probability integral transforms.

	Test	
Distribution		
	KS test of uniformity	0.16
Independence		
	Bowman-Shenton	0.10
	Doornik-Hansen	0.01
	Berkowitz I	0.00
	Berkowitz II	0.00

The test outcomes are recorded as p-values, except for the KS test, which is the test statistic value. The 5% critical value is 0.23. The Bowman-Shenton test is a two-degree of freedom test with an asymptotic chi-squared distribution, whilst the Doornik-Hansen tests may have better small-sample properties. Berkowitz I is a 1-degree-of-freedom test of no first-order autocorrelation of the transformed probability integral transforms assuming $N(0,1)$. Berkowitz II is a 3-degree-of-freedom test of zero-mean, unit-variance and no first-order autocorealtion of the transformed probability integral transforms assuming normality

normal CDF transformation of the $\{z_t\}_{t=1}^n$ series, to give, say, $\{z_t^*\}_{t=1}^n$, and testing whether the transformed series are $iidN(0,1)$. Berkowitz argues that more powerful tools can be applied to testing a null of $iidN(0,1)$, compared to one of iid uniformity, and proposes a one-degree of freedom test of independence against a first-order autoregressive structure, as well as a three-degree of freedom test of zero-mean, unit variance and independence. In each case the maintained assumption is that of normality, so that standard likelihood ratio tests are constructed using the gaussian likelihoods. We find that both the one and three-degree of freedom tests reject at the 1% level (see table 3). The assumption of normality of $\{z_t^*\}$ is also amenable to testing – the Shenton and Bowman (1977) and Doornik and Hansen (1994) tests of normality return p-values of 0.10 and 0.01 respectively, although these tests of distribution do assume a random sample.

The formal tests of the iid part of the joint hypothesis reject the SPF densities confirming the impression gained from the Box-Whisker plots.

5 Results of the event probability forecast evaluation

From the Box-Whisker plot in figure 1 it appears that the rejection of the SPF probability distributions is due in part at least to the tendency over the recent period to both over-estimate the uncertainty or variability and level of the rate of inflation. It is unclear to what extent these deficiencies detract from their value without a fully-formulated decision-based approach.[12] Whilst such approaches are used in

[12] That is, an evaluation of the forecast densities in terms of the expected ‘economic value’ from basing decisions/actions on those forecasts: see, e.g., Granger and Pesaran (2000a, 2000b) and Pesaran and Skouras (2002, p. 245-7).

Table 4 Inflation event forecasts evaluation.

	Target range event	Directional event
SPF forecasts	0.145	0.547
No change forecasts	0.123	0.306

The table records the p-values of Wald tests of $[\gamma\ \beta] = [1\ 0]$ for the logit regressions. The 'no change' density forecasts of the directional event imply $p_t = 0.5$ for all t, and therefore $\alpha_t = 0.5$, for all t. In this case, the null is that $\beta = 0$ in a regression that omits α_t.

some disciplines such as meteorology, their practical adoption in economics appears some way off. [13] Consequently, we provide two complementary assessments of the quality of the density forecasts, based on how accurately two particular events are forecast, using the techniques of section 3. The first is the event that inflation will be in some target range, here taken to be $1\frac{1}{2}$ to $3\frac{1}{2}\%$. In a number of countries the monetary authorities target a range for the inflation rate. Given the lagged adjustment of the economy to changes in monetary instruments, accurate forecasts of whether the target range is likely to be hit are essential for the efficient conduct of monetary policy. The second event is the direction-of-change of inflation. Accurately forecasting increases versus decreases in rates of growth has obvious appeal in the context of real activity variables, because of the correlation with business cycle phases of contraction and expansion, and Canova (2002) also recommends evaluating inflation forecasts in this way.

Table 4 records the p-values of the tests of $[\gamma\ \beta] = [1\ 0]$ in the logit regression of (3) for the two events. We are unable to reject the nulls that the forecast probabilities of both events are efficient. In addition, event probability forecasts derived from 'no-change' forecast densities are also tested. The latter are the density forecast analog of the widely-reported no-change point forecast (see, e.g., Theil (1966)). No-change forecasts have been viewed as 'naive' predictors, but have recently been shown to be a good forecasting device when there are structurl breaks (see, e.g., Clements and Hendry (1999)). More pertinently, Atkeson and Ohanian (2001) show that a no-change predictor outperforms modern Phillips Curve-based model predictions of inflation in the U.S. over the last fifteen years. We calculate no-change density forecasts by assuming a conditional gaussian density with means equal to the realised rate of inflations in the previous periods, and the variances equal to the sample variances of the $t-4$ to $t-1$ observations (for a forecast of period t). Given the preceding comments it is perhaps not surprising that the no-change forecasts are not rejected for either event. Nevertheless, the no-change forecasts of the directional event simply indicate that there is an evens chance that inflation will be lower (higher) than in the previous period. While the no-change forecasts are uninformative about the directional event by construction, figure 2 shows that the SPF forecasts are also relatively uninformative over the last ten years, at least they vary little until 2002. That said, this is in large part a consequence of the low variability of inflation over the last decade, such that directional changes often correspond to small changes in the rate of inflation, and are therefore hard to predict.

The results of the forecast encompassing test described in section 3.2 are recorded in table 5 for

[13] For example, "Unfortunately the complexity of situations prevents a single definitive cost measure being formulated. Thus a range of simple statistical measures have been developed to measure various aspects of forecast quality" (Encyclopaedia of Statistical Sciences, entry on Forecasting).

Table 5 Forecast encompassing tests.

Null hypothesis	R_1	R_2	DM	MDM
Target range event				
SPF encompasses No-change	0.247	0.246	0.242	0.248
No-change encompasses SPF	0.010	0.029	0.019	0.024
Directional event				
SPF encompasses No-change	0.592	0.592	0.592	0.590
No-change encompasses SPF	0.000	0.002	0.000	0.000

The four variants of forecast encompassing tests are described in section 3.2. The table records the p-values of one-sided forecast encompassing tests.

The range event is that inflation is between 1.5 and 3.5%, the directional event is that inflation is lower than in the previous year.

the SPF and 'no change' forecast probabilities. The four modified tests are all very similar. We do not reject the null that the SPF forecasts encompass the no-change for either the range (inflation between $1\frac{1}{2}$ and $3\frac{1}{2}\%$) or directional events (inflation lower than in previous year). Running the tests in the reverse rejection we find the null hypotheses that the no-change forecasts encompass the SPF forecasts is rejected for both events (at the 1% level for the directional event and at the 5% level for the range event). The no-change forecasts do not contain any useful information not already present in the SPF forecasts, for the two events we consider, although we can comfortably reject the hypotheses that the SPF do not contain any information not already in the no-change forecasts. So whereas the tests of the efficiency of the forecasts in table 4 are unable to reject the SPF or no-change forecasts for either event, the tests of conditional efficiency or forecast encompassing – which include the rival model's forecasts in the information set – do give a definitive outcome.

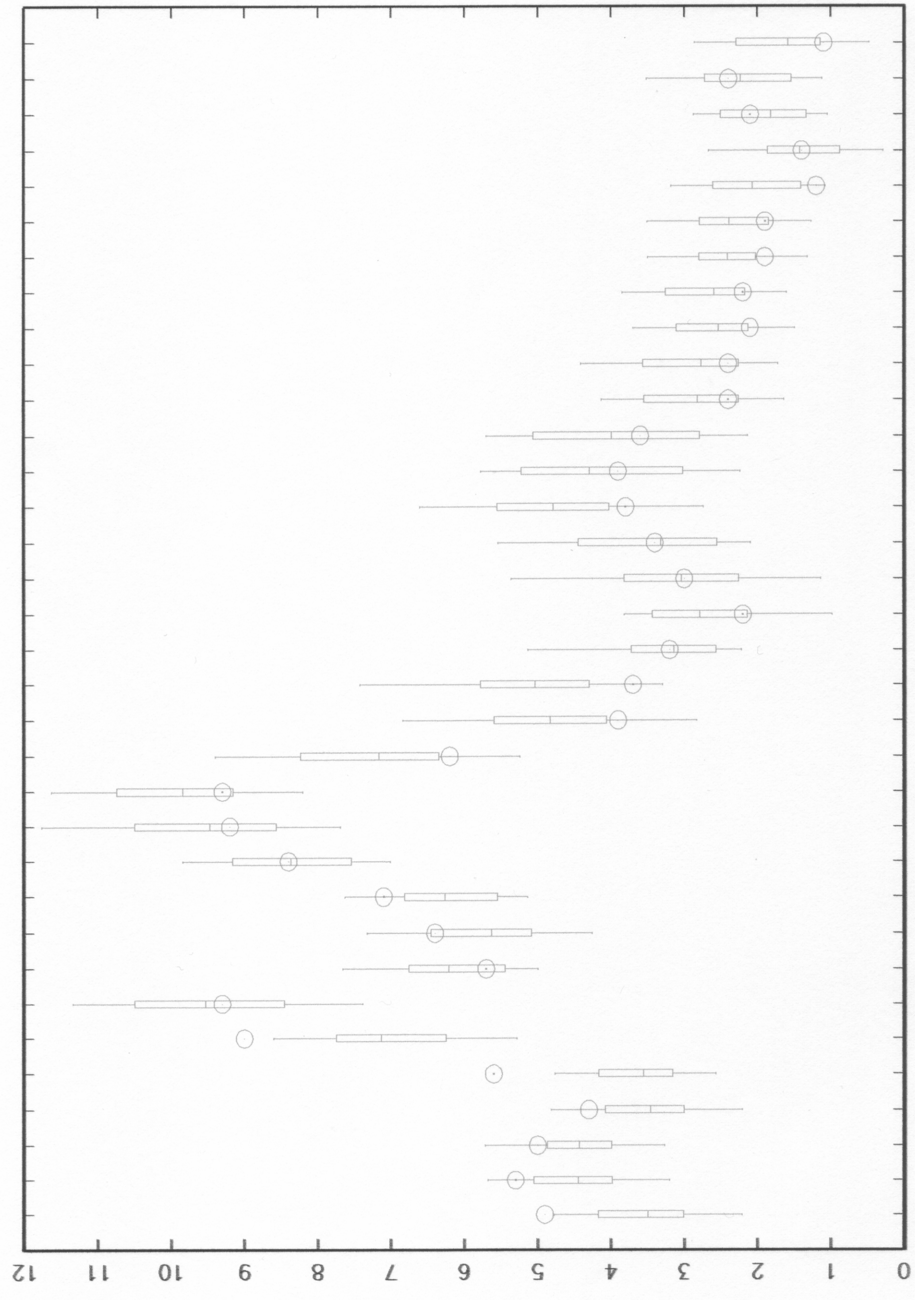
6 Conclusions

We have proposed number of ways of evaluating probability event forecasts to complement the now standard evaluation techniques available for forecast densities. We consider regression-based tests of the efficiency of forecast probabilities paralleling Mincer and Zarnowitz (1969), as well as tests of conditional efficiency or forecast encompassing, to evaluate the forecasts against a wider information set. The tests of forecast encompassing are shown to be affected by conditional heteroscedasticity, but simple modifications to the standard test statistics are shown via Monte Carlo to deliver correctly-sized tests with reasonable power.

These tools are applied in an analysis of the SPF expected probability distributions of inflation. The SPF densities are rejected using the 'whole density' probability integral transform approach, but neither the SPF nor rival no-change forecast probabilities for two events of interest are rejected using a realization-forecast regression related to that of Mincer and Zarnowitz (1969). However, widening the information set to test for conditional efficiency against the information embodied in rival forecasts shows that the SPF forecasts are conditionally efficient against the no-change forecasts (for both events), whilst the conditional efficiency of the no-change forecasts (against the SPF forecasts) is rejected.

The SPF forecasts have been extensively studied by economists. We have suggested new ways in which derived forecast probabilities of particular events of interest may be evaluated, and compared against rival sets of forecast probabilities. The techniques are general, and might be especially useful when, as in the case of the SPF forecasts, the forecasts are not model based (or at least the model is not known to the econometrician). As noted in the introduction, a number of alternative techniques might usefully be brought to bear for model-based forecasts.

Figure 1 Inflation forecast probability distributions shown as Box-Whisker plots and realizations. The boxes represent the inter-quartile, the outer ‘whiskers’ the 10 and 90^{th} percentiles, and the inner line the median. The realizations are circles with dots at the centres.


12
11
10
9
8
7
6
5
4
3
2
1
0
1969
1972
1975
1978
1981
1984
1987
1990
1993
1996
1999
2002

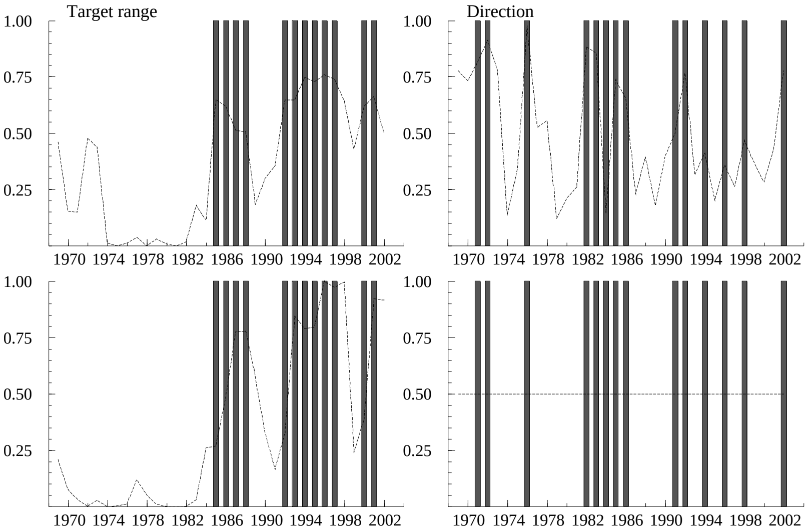


Figure 2 Time series of events ‘inflation in the range 1.5 to 3.5%’ and ‘lower inflation than last year’ and forecast probabilities of these event. Each column refers to one of the two events. The rows relate to the SPF and no-change event forecast probabilities. In each panel, the bars are one-zero event indicators, and the lines the forecast probabilities.

References

Atkeson, A., and Ohanian, L. (2001). Are Phillips curves useful for forecasting inflation?. *Federal Reserve Bank of Minneapolis, Quarterly Review*, **25**, 2–11. (1).

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

Canova, F. (2002). G-7 Inflation forecasts. mimeo, Universitat Pompeu Fabra.

Chong, Y. Y., and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, **53**, 671–690. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, **39**, 841–862.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5**, 559–583.

Clements, M. P., and Hendry, D. F. (1999). *Forecasting Non-Stationary Economic Time Series*. Cambridge, Mass.: MIT Press. The Zeuthen Lectures on Economic Forecasting.

Clements, M. P., and Taylor, N. (2003). Evaluating prediction intervals for high-frequency data. *Journal of Applied Econometrics*, **18**, 445 – 456.

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

Diebold, F. X., and Lopez, J. A. (1996). Forecast evaluation and combination. In Maddala, G. S., and Rao, C. R. (eds.) , *Handbook of Statistics*, Vol. 14, pp. 241–268: Amsterdam: North–Holland.

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

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

Doornik, J. A., and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.

Doornik, J. A., and Hendry, D. F. (2001). *GiveWin: An Interface to Empirical Modelling*. London: Timberlake Consultants Press.

Durbin, J. (1969). Tests for serial correlation in regression analysis based on the periodogram of least-squares residuals. *Biometrika*, **56**, 1–15.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.

Engle, R. F., and Manganelli, S. (1999). CAViaR: Conditional autoregressive Value-at-Risk by regression quantiles. Ucsd discussion paper 99-20, Department of Economics, UCSD.

Granger, C. W. J., and Newbold, P. (1973). Some comments on the evaluation of economic forecasts. *Applied Economics*, **5**, 35–47.

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

Granger, C. W. J., and Pesaran, M. H. (2000b). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, **19**, 537–560.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, **13**, 281–291.

Harvey, D. I., Leybourne, S., and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, **16**, 254–259.

Holden, K., and Peel, D. A. (1990). On testing for unbiasedness and efficiency of forecasts. *Manchester School*, **58**, 120–127.

Hong, Y. (2001). Evaluation of out-of-sample probability density forecasts with applications to stock prices. Manuscript, Department of Economics, Cornell University.

Li, F., and Tkacz, G. (2001). A consistent bootstrap test for conditional density functions with time dependent data. Discussion paper, Dept. of Monetary and Financial Analysis, Bank of Canada.

Lopez, J. (1996). Regulatory evaluation of Value-at-Risk models. Discussion paper 95-6, Federal Reserve Bank of New York.

Mincer, J., and Zarnowitz, V. (1969). The evaluation of economic forecasts. In Mincer, J. (ed.) , *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Nelson, C. R. (1972). The prediction performance of the FRB-MIT-PENN model of the US economy. *American Economic Review*, **62**, 902–917.

Newbold, P., and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society A*, **137**, 131–146.

Newbold, P., and Harvey, D. I. (2002). Forecasting combination and encompassing. In Clements, M. P., and Hendry, D. F. (eds.) , *A Companion to Economic Forecasting*, pp. 268–283: Oxford: Blackwells.

Patton, A. J. (2002). Modelling time-varying exchange rate dependence using the conditional copula. Mimeo, University of California, San Diego.

Pesaran, M. H., and Skouras, S. (2002). Decision-based methods for forecast evaluation. In Clements, M. P., and Hendry, D. F. (eds.) , *A Companion to Economic Forecasting*, pp. 241–267. Oxford: Blackwells.

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

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

Stock, J. H., and Watson, M. W. (1999). A comparison of linear and nonlinear models for forecasting macroeconomic time series. In Engle, R. F., and White, H. (eds.) , *Cointegration, Causality and Forecasting*, pp. 1–44. Oxford: Oxford University Press.

Theil, H. (1966). *Applied Economic Forecasting*. Amsterdam: North-Holland.

Thompson, S. B. (2002). Evaluating the goodness of fit of conditional distributions, with an application to affine term structure models. Manuscript, Harvard University.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, **64**, 1067–1084.

West, K. D. (2001). Tests for forecast encompassing when forecasts depend on estimated regression parameters. *Journal of Business and Economic Statistics*, **19**, 29–33.

West, K. D., and McCracken, M. W. (1998). Regression-based tests of predictive ability. *International Economic Review*, **39**, 817–840.

West, K. D., and McCracken, M. W. (2002). Inference about predictive ability. In Clements, M. P., and Hendry, D. F. (eds.) , *A Companion to Economic Forecasting*, pp. 299–321: Oxford: Blackwells.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838.