EVALUATING PROBABILITY FORECASTS:
Calibration Isn't Everything

by

Kajal Lahiri*
University at Albany – SUNY

and

J. George Wang
College of Staten Island – CUNY

Abstract: Using evaluation methodologies for rare events from meteorology and psychology, we examine the value of probability forecasts of real GDP declines during the current and each of the next four quarters using data from the *Survey of Professional Forecasters*. We study the quality of these probability forecasts in terms of calibration, resolution, the relative operating characteristic (ROC), and alternative variance decompositions. We find that even though QPS and the calibration tests suggest the forecasts for all five horizons to be useful, the other approaches clearly identify the longer-term forecasts (Q3-Q4) having no skill relative to a naïve baseline forecast. For a given hit rate of (say) 90%, the associated high false alarm rates for the longer-term forecasts make these unusable in practice. We find conclusive evidence that the shorter-term forecasts (Q0-Q1) possess significant skill in terms of all measures considered, even though they are characterized by considerable excess variability.

Key Words: Survey of Professional Forecasters, Subjective probability, QPS decomposition, Calibration, Resolution, Skill score, Relative Operating Characteristics, Recession.

JEL Classification: B22; C11; C53

*Corresponding author: K. Lahiri, Department of Economics, University at Albany - SUNY, Albany, NY 12222, USA. Tel: (518) 442 4758; Fax: (518) 442 4736. Email: klahiri@albany.edu

Evaluating Probability Forecasts: Calibration Isn't Everything

1. INTRODUCTION

Forecasting rare business events such as recessions has long been a challenging issue in business and economics. As witnessed during past decades, the track record of large scale structural macro and VAR models, or real GDP forecasts that are obtained from professional surveys (e.g., Blue Chip, Survey of Professional Forecasters, OECD, etc.), in predicting or even timely recognizing postwar recessions has not been very impressive. Even for probability forecasts based on modern time series models, the improvement in forecast performance has been limited at best.[1]

One of the main purposes of this paper is to argue that an evaluation of recorded probability forecasts by professional economists can suggest reasons for forecasting failures, and can help define limits to the current capability of macro economic forecasts. The traditional way of evaluating probability forecasts is the Mean Square Error (MSE) type of measure such as Brier's Quadratic Probability Score (QPS), which evaluates the external correspondence between the probability forecasts and the realization of the event. This approach, however, can fail to identify the ability of a forecasting system to evaluate the odds of the occurrence of an event against its non-occurrence, which is a very important characteristic that the users of forecasts need to assess. Meanwhile, a high performance score can be achieved by totally unskilled forecasts without any information value. Thus, the traditional approach can be inadequate in evaluating the usefulness of probability forecasts, particularly for rare events.

In this paper we will study the usefulness of the subjective probability forecasts that are obtained from the *Survey of Professional Forecasters* (*SPF*) as predictors of GDP downturns using several distinct evaluation methodologies. Even though these forecasts

[1] Forecasting failures have been documented extensively in the literature, see, for instance, Filardo (1999, 2004), Fildes and Stekler (2002), Fintzen and Stekler (1999), Juhn and Loungani (2002), Krane (2003), McNees (1991), and Zarnowitz and Moore (1991). For recent models generating probability forecasts, see Kling (1987), Neftci (1987), Hamilton (1989), Stock and Watson (1991, 1993) and Zellner *et al*. (1991).

are available since 1968, and have drawn media attention[2], very little systematic analysis has been conducted to look into their usefulness as possible business cycle indicators.[3] We will try to pinpoint the importance of alternative evaluation approaches and emphasize the more important characteristics of a set of forecasts from the standpoint of end-users. Our study also has implications for the users of Logit, Probit and other probability models for relatively rare events. Often the conventional goodness-of-fit statistics in these models like the pseudo R-square, fraction correctly predicted, etc. fail to identify the type I and type II errors in predicting the event of interest.

The plan of this paper is as follows: In section 2, we will introduce the data, and explain the set up. In section 3, we will evaluate the probability forecasts using the traditionally popular calibration approach with statistical tests. In section 4, we will explore the multi-dimension nature of the probability forecasts using alternative methodologies. In section 5, we will suggest some effective ways to evaluate the performance of the probability forecasts of rare business events in terms of ROC curve. Finally, concluding remarks will be summarized in section 6.

2. SPF PROBABILITY FORECASTS OF REAL GDP DECLINE

The Survey of Professional Forecasters (SPF) [4] has been collecting subjective probability forecasts of real GDP/GNP declines during the current and four subsequent quarters since its inception in 1968. At the end of the first month of each quarter, the individual forecasters in SPF form their forecasts. The survey collects probability assessments for a decline in real GDP in the current quarter, and in each of the next four quarters conditional on the growth in the current period. The number of respondents has varied between 15 and 60 over the quarters. In this study we use probabilities averaged over

[2] The New York Times columnist David Leonhardt (September 1, 2002) calls the one-quarter-ahead GDP decline probability as the "Anxious Index".

[3] Notable exceptions include Braun and Yaniv (1992), Graham (1996), and Stock and Watson (2003). However, these studies emphasized different aspects of the data.

[4] Formerly the surveys were carried out under the auspices of the American Statistical Association and the National Bureau of Economic Research (ASA-NBER). Since June 1990, the Federal Reserve Bank of Philadelphia has conducted the survey. See Croushore (1993) for an introduction to SPF.

individuals.[5] Using the July revisions, during our sample period from 1968:4 to 2004:2 there were 20 quarters of negative GDP growth -- those beginning 1969:4, 1973:4, 1980:1, 1981:3, 1990:3 and 2001:1 -- which consist of six separate episodes of real GDP declines. The annualized real time real GDP growth issued every July is used as the forecasting target, against which the forecasting performance of the SPF forecasts will be evaluated.[6] The SPF probabilities for real GDP declines during the current and next four quarters are depicted against the real time real GDP growth in Figures 1a -1e. The shaded bars represent the NBER defined recessions.

From Figures 1a–1e, several notable patterns can be observed. First, the mean probabilities generated by the professional forecasters fluctuate over time, varying from as high as 80% to as low as less than 5%. Second, the fluctuations in the probabilities seem to be roughly coincident with those real GDP growth and the NBER defined peaks and troughs. Third, for different forecasting horizons, the probabilities either precede or follow the cyclical movement of the real GDP with different leads or lags. Finally, the high end of the mean probability tends to decrease as the forecasting horizon increases. As shown in the figures, the high-end probability decreases steadily from about 80% for the current quarter to only about 30% for three and four-quarter-ahead forecasts. All these observations suggest that the information content, hence the value, of the SPF probability forecasts may be horizon-dependent.

3. CALIBRATION OF *SPF* PROBABILITY FORECASTS

The traditional way of evaluating probability forecasts for the occurrence of a binary event is to assess the calibration of the forecasts against realizations, that is, to assess the external correspondence between the probability forecasts and the actual occurrence of the event.

[5] In future we would like to consider the optimum combination of the individual probability forecasts. Based on nine selected forecasters, Graham (1996) found that pooling techniques that allow for correlation between forecasts performed better than the simple average of forecasts.

[6] We also conducted our analysis with the 30-day announcements as the real time data. Results were virtually unchanged. All real time data were downloaded from the Federal Reserve Bank of Philadelphia web site.

3.1. Brier's Quadratic Probability Score

A measure-oriented approach simply compares the forecast probability with the realization of a binary event that is represented by a dummy variable taking value 1 or 0 depending upon the occurrence of the event. A most commonly used measure is (half) Brier's Quadratic Probability Score (QPS), a probability analog of mean squared error, i.e.:

$$QPS = 1/T\sum_{t=1}^{T}(f_t - x_t)^2 \qquad (1)$$

where f_t is the forecast probability made at time t, x_t is the realization of the event at time t. T is the total number of the observations or forecasting quarters.

The QPS ranges from 0 to 1 with a score of 0 corresponding to perfect accuracy, and is a function only of the difference between the assessed probabilities and realizations. The calculated QPS for each forecasting horizon from the current quarter (Q0) to the next four quarters (Q1, Q2, Q3, and Q4) are 0.077, 0.098, 0.103, 0.124, and 0.127, respectively. Thus, even though these scores deteriorate as the forecast horizon increases, all seem to suggest good calibration and are close to zero. It may be noted that QPS figures are seldom reported with their associated standard errors. In next section, we will show that these figures, indeed, are not significantly different from their respective expectations under the hypothesis of perfect forecast validity.

3.2 Prequential Test for Calibration

Dawid (1984) and Seillier-Moiseiwitsch and Dawid (1993) (henceforward SM-D) suggested a test for calibration. The SM-D test statistic is constructed by a weighted sum of the difference between the predictive probability and the realization of the event such as

$$Z_j = (r_j - e_j)/\sqrt{w_j} \qquad (2)$$

where e_j and r_j are the predicted probability and realization for probability group j, respectively; w_j is the weight determined by

$$w_j = n_j \pi_j (1 - \pi_j) \qquad \text{(2a)}$$

where π_j is the midpoint of the group *j*, and n_j is the number of quarters in group *j*.

For any given set of sequential probability forecasts, its calibration or accuracy can be tested in each probability interval using the Z_j statistic. If it lies too far out in the tail of the standard normal distribution, it might be regarded as evidence against forecast accuracy. Similarly, the overall performance of the forecast can be evaluated using χ^2 test with *j* degree of freedom, which is constructed as $\sum Z_j^2$. Thus, using SM-D calibration test, the accuracy of probability forecasts can be assessed with explicitly expressed uncertainty as indicated by the confidence level.

The results from the SM-D test to assess the SPF probability forecasts are reported in Table 1 where we find that the forecasts of all forecasting horizons appear to be well calibrated. The χ^2 values for each forecasting horizon fall into the acceptance area with confidence level of 90% and the appropriate degrees of freedom. While the values of χ^2 statistics vary from 3.49 to 12.68, surprisingly the lowest value of χ^2 is obtained for the four-quarter-ahead forecasts (Q4).[7]

Given $QPS = 1/T \sum_{t=1}^{T} (f_t - x_t)^2$, SM-D showed how their calibration test could be converted to a test of whether OPS is significantly different from its expected value $1/T \sum_{t=1}^{T} f_t (1 - f_t)$ under the hypothesis of perfect forecast validity using a standard

[7] Using a Bayesian posterior odds approach, it will be interesting to study the analytical power of the SM-D test against alternatives such as Q3 or Q4 forecasts. We should, however, emphasize that a more powerful calibration test will not minimize the importance of resolution in evaluating probability forecasts.

$N(0,1)$ approximation for the distribution of

$$Y_n = \sum_{i=1}^{n}(1-2f_i)(x_i - f_i)/[\sum_{i=1}^{n}(1-2f_i)^2 f_i(1-f_i)]^{1/2} \tag{3}$$

Then, the observed value of Y_n can be obtained as:

$$Y_n = [\sum_{j=1}^{11}(1-2\pi_j)(r_j - e_j)]/[\sum_{j=1}^{11} n_j \pi_j (1-\pi_j)(1-2\pi_j)^2]^{1/2} \tag{4}$$

The test results are reported in Table 1 as well. We find that, for all forecast horizons, none of the calculated statistics fall in the (one-sided) rejection region at the 5% significance level, which is consistent with the SM-D calibration test results that forecasts for all horizons satisfy the hypothesis of perfect forecast validity.

4. FURTHER DIAGNOSTIC VERIFICATION OF PROBABILITY FORECASTS

Some of the results from the calibration tests in the previous section may seem counter-intuitive. While the probability forecasts for the longer forecasting horizons, especially Q3 and Q4, never exceed 40% even when the event has already occurred, the SM-D test showed that they are well calibrated. This observation leads to a question of whether the calibration is an adequate measure of forecast validity, and why, if it is not. The issue may be analyzed using some alternative approaches.

4.1. The Skill Score

Skill Score (SS) measures the relative accuracy of a forecast compared to a benchmark. We calculated the bellwether skill measure

$$SS(f,x) = 1 - [MSE(f,x)/MSE(\mu_x, x)] \tag{5}$$

where MSE (μ_x,x) is the accuracy associated with the constant base rate or the constant relative frequency forecast (CRFF) which is μ_x= 0.14 in our sample. For each forecasting horizon we found the SS values to be 0.36, 0.19, 0.05, -0.002, -0.05, respectively. Note that the use of this historical average value as the base rate presumes substantial knowledge on part of the forecasters.[8] While the skill score for the shorter run forecasts (Q0-Q1) indicate significant improvement of the SPF forecasts over the benchmark base rate forecast, the longer run forecasts (Q3 and Q4) do not show any clear-cut relative advantage. The SS value for Q2 forecasts is marginal. These results were not discernable using the calibration tests.

Note that the skill score in (5) can be decomposed as (*cf.* Murphy (1988)):

$$SS(f,x) = \rho_{fx}^2 - [\rho_{fx} - (\sigma_f / \sigma_x)]^2 - [(\mu_f - \mu_x)/\sigma_x]^2 \qquad (6)$$

where ρ_{fx} is the correlation coefficient between forecast and the actual binary outcome, σ_f^2 and σ_x^2 are their variances, and (μ_f, μ_x) are the respective sample averages. The decomposition shows that SS is simply the square of the correlation between *f* and *x* adjusted for any miscalibration penalty (second term) and the normalized difference in the sample averages of the actual and the forecast (third term). This decomposition for Q0-Q4 are given in Table 2a where we find that the last two terms of the decomposition are close to zero, and thus, the skills for Q0-Q4 forecasts in effect reflect the correlations between the forecasts and the actual. For Q0, Q1, and Q2, these correlations are 0.393, 0.220, and 0.077 respectively, and are found to be statistically significant using the simple t-test for no correlation.[9] The correlations for Q3-Q4 are very small and statistically insignificant. This decomposition again shows that the long-term probability

[8] Alternatively, one can consider using the last realization as the forecast to form a binary time varying base rate. Thus, for current and next four quarters, the last quarter realization is used. The associated skill scores of SPF forecasts were significantly more than those with μ_x = 0.14 implying that the latter base rate is considerably more informative than the use of the lagged actual. Other base rate alternatives, *e.g.,* eternal optimist (*f*=0), eternal pessimist (*f*=1), or a coin flipper (*f*=0.5), are also considerably less informative than the alternative in (5); *cf.* Zellner *et al.* (1991).

[9] The t-values were obtained from a regression of $(\rho_{fx}/\sigma_x^2)\, x$ on *f*.

forecasts have no skill compared to the benchmark forecast even though they are found to be well calibrated like the near term forecasts.

4.2 The Murphy Decomposition

In addition to calibration, there are several other features that also characterize good probability forecasts. Murphy (1972) decomposed the *MSE* or the half-Brier Score into three components:

$$MSE(f,x) = \sigma_x^2 + E_f(\mu_{x/f} - f)^2 - E_f(\mu_{x/f} - \mu_x)^2 \qquad (7)$$

The first term on the RHS of (7) is the variance of the observations, and can be interpreted as the *MSE* of constant forecasts equal to the base rate. It represents forecast difficulty. The second term on the RHS of (7) measures the calibration or reliability of the forecasts, which measures the difference between the conditional mean of the occurrence on the probability group and the forecast probability. The third term on the RHS of (7) is a measure of the resolution or discrimination that requires some subtleties in interpretation, *cf.* Yates (1994). In general, it is desirable for the relative frequency of occurrence of the event to be larger (smaller) than the unconditional relative frequency of occurrence when f is larger (smaller). Thus, resolution refers to the ability of a set of probability forecasts to sort individual outcomes into probability groups which differ from the long-run relative frequency. Calibration can be interpreted as a labeling skill that expresses uncertainty correctly. Even though calibration is a natural feature to have, it is resolution that makes the forecasts useful in practice.[10]

Thus, the calibration and resolution refer to the two distinct attributes of a forecast. For perfectly calibrated forecasts, $\mu_{x/f=}f$ and $\mu_x = \mu_f$, and the resolution term equals the variance of the forecasts, σ_f^2. Resolution or discrimination (or sharpness) refers to the marginal or predictive distribution of the forecasts $p(f)$. A sample of probability forecasts is said to be completely resolved if the probability only takes values zero and one. Apparently, completely refined forecasts would be miscalibrated due to the inability

[10] See Dawid (1986), and DeGroot and Fienberg (1983).

of the forecasters to predict the future with certainty. Conversely, well-calibrated probability forecasts generally exhibit only a moderate degree of refinement. Thus, possible trade-off between the calibration and resolution exists to minimize *MSE*. Forecasts possess positive *absolute* skill when the resolution reward exceeds the miscalibration penalty.

The distributions of $p(x/f)$ for Q0-Q4 are depicted in Figures 2a-2e. In these figures, $\mu_{x/f}$ is plotted against *f,* and referred to as the attributes diagram. The calculations are explained in Tables 2b and 2c. Figures 2a-2e plot the relationship between $\mu_{x/f}$ and *f* for the relevant sample of forecasts and observations, and also contain several reference or benchmark lines. The straight 45^0 line for which $\mu_{x/f} = f$ represents perfectly calibrated forecasts. The horizontal line represents completely unresolved forecasts, for which $\mu_{x/f} = \mu_x$. The dotted line equidistant between the 45^0 line and the horizontal dashed line represents forecasts of zero skill in terms of SS where the resolution award is equal to the miscalibration penalty. To the right (left) of the vertical auxiliary line at $f = \mu_x$, skill is positive above (below) the zero-skill line and negative below (above) it. This is because, when $\mu_{x/f}$ is on the right (left) of the vertical line and above (below) the zero-skill line, the resolution award will be greater than the miscalibration penalty. Hence, the *MSE* of the SPF would be smaller than that of the base rate, leading to a positive SS. Thus, Figures 2a-2e permit qualitative evaluation of resolution and skill as well as calibration for individual forecasts, see Murphy and Winkler (1992).

An examination of Figures 2a – 2e indicates that, similar to the previous findings, the SPF forecasts with shorter forecasting horizons (Q0-Q2) are generally well calibrated. Most points on the empirical curves fall in regions of positive skill. However, SPF forecasts with longer forecasting horizons (Q3-Q4) reveal less satisfactory performance with negative overall skill scores. In Figures 3a – 3e, the graph is split into two conditional likelihood distributions given $x = 1$ (GDP decline) and $x = 0$ (no GDP decline). For these two conditional distributions, the means were calculated to be (0.56, 0.38, 0.26, 0.19 and 0.18) for $x = 1$ and (0.14, 0.16, 0.17, 0.17 and 0.18) for $x = 0$, respectively. Good discriminatory forecasts will give two largely non-overlapping marginal distributions, and, in general, their vertical differences should be as large as possible. While the shorter run forecasts (Q0-Q2) display better discriminatory power, the longer run forecasts (Q3-Q4) display poor discrimination due to the over-use of low probabilities during both regimes (i.e., $x = 0$ and $x = 1$). So the two distributions overlap. In particular, the mean values for $x = 1$ (GDP decline) and $x = 0$ (no GDP decline) for the 4-quarter ahead forecasts (Q4) are almost identical.[11] Numerical values of the Murphy decomposition are given in Table 3 where we find that *MSE* improves by about 35%, 16% and 6% for the current (Q0), one quarter- (Q1), and 2-quarter–ahead (Q2) forecasts, respectively, over the constant relative frequency forecast (CRFF). The 3-quarter-ahead (Q3) forecasts are even with CREF, and the *MSE* of the 4-quarter-ahead (Q4) forecasts are worse by nearly 4%.

The major contributor for the improvement in *MSE* is resolution, which helps to reduce the baseline *MSE* (CRFF) by about 47%, 25%, 17%, 6%, and 8% for Q0 to Q4, respectively. On the other hand, the miscalibration increases *MSE* of CRFF by 12%, 9%, 11%, 5% and 13%, respectively – they are relatively small for all forecast horizons. The improvement due to resolution is greater than the deterioration due to miscalibration for the up to 2-quarter-ahead forecasts, and the situation is opposite for the 4-quarter-ahead forecasts. In the case of 3-quarter-ahead forecasts, the impact of the resolution and miscalibration pretty much cancel out each other. As indicated by attributes diagrams (Figs.2a – 2e) and the overlapping of the $p(f/x=1)$ distribution with $p(f/x=0)$ (Figs.3a – 3e), the SPF forecasters are conservative in assigning high probability during quarters when recession occurs. This also suggests that distinguishing between occurrences and non-occurrences, and assigning higher probabilities in quarters when recession occurs, can improve the resolution of the forecasts. It may be noted that the assignment of low probability for rare events is not unusual, and is actually quite common in weather

[11] Cramer (1999) suggested the use of this difference in the conditional means as a goodness-of-fit measure in binary choice models with unbalanced samples where one outcome dominates the sample. See also Greene (2003).

forecasting. When the diagnostic information or “cue” is not adequate to make informed forecasts, the tendency for the forecaster is to assign the average base rate probability.[12]

4.3 The Yates Decomposition

Yates (1982), and Yates and Curley (1985) showed that calibration and resolution components in the Murphy decomposition are algebraically confounded with each other, and suggested a covariance decomposition of *MSE* that is more basic and revealing than the Murphy decomposition, see also Björkman (1994). The Yates decomposition is written as:

$$MSE(f,x) = u_x(1-u_x) + \Delta\sigma_f^2 + \sigma_{f,\min}^2 + (u_f - u_x)^2 - 2\sigma_{f,x}^2 \tag{8}$$

where $\sigma_{f,\min}^2 = (u_{f/x=1} - u_{f/x=0})^2 u_x(1-u_x)$, and $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$.

As noted before, the outcome index variance $\sigma_x^2 = u_x(1-u_x)$ provides a benchmark reference for the interpretation of *MSE.* The conditional minimum forecast variance $\sigma_{f,\min}^2$ reflects the double role that the variance of the forecast plays in forecasting performance. On the one hand, minimized σ_f^2 will help reduce the *MSE.* On the other hand, minimized forecast variance can be achieved only when the constant forecast is offered. The constant forecast would lead to zero covariance between the forecast and event, which will, in turn, increase the *MSE*. So the solution is to minimize the forecast variance given the covariance that demonstrates the fundamental forecast ability of the forecasters. The conditional minimum value of forecast variance is achieved when the forecaster has perfect foresight such that he or she can exhibit perfect discrimination of the instances in which the event does and does not occur.

[12] Diebold and Rudebusch (1989, 1991) and Lahiri and Wang (1994) used QPS and its resolution and calibration components to study the value of recession forecasts generated from probability models of Neftci (1984) and Hamilton (1989), respectively. Bessler and Ruffley (2004) have studied probability forecasts from a 3-variable VAR model of stock returns by a bootstrap-type procedure under the normality assumption. They found forecasts to be well calibrated but have very low resolution to be useful.

Since $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$, the term may be considered as the excess variability in the forecasts. If the covariance indicates how responsive the forecaster is to information related to an event's occurrence, $\Delta\sigma_f^2$ might reasonably be taken as a reflection of how responsive the forecaster is to information that is not related to the event's occurrence. Note that $\Delta\sigma_f^2$ can be calculated as $(N_1\sigma_{f/x=1}^2 + N_0\sigma_{f/x=0}^2)/N$, where ($N_i, i = 0,1,$) is the number of periods associated with the occurrence (i =1) and non-occurrence (i = 0), $N_1 + N_0 = N$. So the term is the weighted mean of the conditional forecast variances.

Using the SPF probability forecast, the components of equation (8) were computed and presented in Table 4.[13] For the shorter forecasting horizons up to 2-quarters (Q0-Q2), the overall *MSE* values are less than the constant relative frequency forecast variance, which demonstrate the absolute skillfulness of the SPF probability forecasts. For the longer run forecasting horizons (Q3-Q4), the overall *MSEs* are slightly higher than those of the constant relative frequency forecast. The primary contributor of the performance is the covariance term that helps reduce the forecast variance by almost 84%, 44%, 18% and 5% for up to 3-quarter-ahead forecasts, but makes no contribution for the 4-quarter-ahead forecasts. The covariance reflects the forecaster's ability to make a distinction between individual occasions in which the event might or might not occur. It assesses the sensitivity of the forecaster to specific cues that are indicative of what will happen in the future. It also shows whether the responsiveness to the cue is oriented in the proper direction. This decomposition is another way of reaching some of the conclusions as the decomposition of skill score in Table 2a.

The excess variability of the forecasts, $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$, for each horizon is found to be 0.0330, 0.0212, 0.0113, 0.0046, and 0.0040, respectively. Compared to the overall forecast variances 0.0541, 0.0272, 0.0123, 0.0047, and 0.004, the excess variability's of SPF probability forecasts are 61%, 77%, 91%, 97% and 100% for Q0-Q4 forecasts,

[13] Note that the QPSs and the variances for Q0-Q4 in Tables 3 and 4 are slightly different because Yates decomposition could be done with ungrouped data whereas the Murphy decomposition was done with probabilities grouped as in Table 1.

respectively. Thus, they are very high, and this means that the subjective probabilities are scattered unnecessarily around $\mu_{f/x=1}$ and $\mu_{f/x=0}$. Since the difference in conditional means, $\mu_{f/x=1} - \mu_{f/x=0}$, are very close to zero for Q3-Q4 forecasts, all of their variability is attributed to excess variability. Assigning low probabilities in periods when GDP actually fell seems to be the root cause of the excess variance. In our sample real GDP fell 20 times. However, in10 of these quarters, the assigned probabilities for Q0-Q2 forecasts never exceeded 0.5; for Q3-Q4 forecasts, the assigned probabilities were even below 0.2. In contrast, for Q0-Q2, in more than 90% of the quarters when GDP growth did not decline, the probabilities were assigned correctly below 50% (for Q3-Q4 the probabilities were below 30%). This explains why *Var (f/x=1)* is much larger than *Var (f/x=0)* when the forecasts have any discriminatory power (*cf*. Figs. 3a-3e).

Overall, both the Murphy and Yates decompositions support the usefulness of shorter run SPF probabilities as predictors of negative real GDP growth, and suggest possible ways of improving the forecasts, particularly at the short run horizons. The longer-term forecasts have very little discriminatory power. While the overall accuracy or the calibration of the forecasts are pretty much similar for each forecasting horizon, the usefulness of the shorter run forecasts primarily comes from their better discriminatory power. These probabilities embody effective information related to the occurrence of the event, and the overall average forecast probabilities are close to the relative frequency of the occurrence of the event. However, improvement can be made by further distinguishing factors related to the occurrence of recessions, while keeping the sensitivity of the forecasts to information that are actually related to the occurrence of GDP declines. This would imply a reduction of unnecessary variance of forecasts particularly during GDP declines, thereby increasing resolution further.

5. RELATIVE OPERATING CHNARACTERISTIC (ROC) ANALYSIS

As the Murphy and Yates decompositions indicated, the traditional measure - the overall calibration or accuracy (or calibration at-large) - can be decomposed into two distinct components. The calibration by group (or calibration at-small), as assessed by SM-D validity test, is only one of the characteristics that a forecast possesses. Resolution or discrimination is another, and actually a more important feature to a forecast end-user. However, one vital issue is that, when the resolution is high and the judgment is accurate, the overall calibration will be good or even perfect. But if the resolution is high and the judgment is not accurate, the increased resolution may, instead, cause poor calibration score. Given the inability of any forecaster to forecast perfectly accurately, the frequent trade-off between the calibration and resolution is unavoidable. In this case, if the performance is measured by the traditional calibration and the QPS is used as the metric, it will encourage hedging behavior as is evidenced by the preponderance of assigned probabilities close to the historical base rate. As a result, the end users may frequently receive forecasts with decent calibration scores, but the forecast actually contains no value to them at all.

In contrast, the discrimination characteristic focuses on the fundamental value of a forecast: its ability to capture the occurrence of an event with an underlying high hit rate, while maintaining the false alarm rate to some acceptable level. The Murphy and Yates decompositions analyzed the structure of the total forecast error and the impact or the relative contribution of each component to the total error, but they could not provide a stand-alone single measure for the discrimination ability of a forecast. In evaluating rare event probabilities, it is crucial to minimize the impact of the predominant outcome on the outcome score. More specifically, the impact of correctly identifying the frequent event, which is the primary source of the hedging, should be minimized. So a better approach to forecast performance should concentrate on the hit rate of the infrequent event, instead of the "percentage correctly predicted" that is the very nature of QPS, *cf.* Doswell *et al* (1990) and Murphy (1991).

In addition, one important but often overlooked issue in the evaluation of the probability forecast is the impact of the selection of the threshold. The performance of an ensemble of probability forecasts in terms of discrimination ability is actually the result of the combination of the intrinsic discrimination ability of a forecasting system and the selection of the threshold. In these regards, Relative Operating Characteristic (ROC) is a convenient approach to use, but unfortunately has drawn little attention in econometrics.[14]

Using ROC approach, the weight of evidence in support of the occurrence of an event can be represented by a point (say, *W*) on a scale. Higher values of *W* correspond to a higher probability of occurrence. The decision to issue the occurrence forecast or non-occurrence forecast is then made based on the predetermined threshold (say, *w*) on the weight of evidence scale. The occurrence forecast is announced if $W > w$, the non-occurrence is announced otherwise. It is further assumed that *W* has a probability density $f_0(w)$ before non-occurrences and $f_1(w)$ before occurrences.

Following the signal detection model, the conditional probability of a hit (*H*) is then the probability that the weight of evidence exceeds the threshold *w* if the event occurs. That is $H = \int_w^\infty f_1(w)dw$. Similarly, the conditional false alarm rate (*F*) is the probability that the weight of evidence exceeds *w* when the event does not occur. That is $F = \int_w^\infty f_0(w)dw$. ROC can be represented by a graph of the hit rate against the false alarm rate as *w* varies, with the false alarm rate plotted as the *X*-axis and the hit rate as the *Y*-axis. The location of the entire curve in the unit square is determined by the intrinsic discrimination capacity of the forecasts, and the location of specific points on a curve is determined by the decision threshold *w* that is selected by the user. As the decision threshold *w* varies from low to high, or the ROC curve moves from right to left, *H* and *F* vary together to trace out the ROC curve. Low thresholds lead to both high *H* and *F* towards the upper right hand corner. Conversely, high thresholds make the ROC points move towards the lower left hand corner along the curve. Apparently, perfect discrimination is represented by an ROC that rises from (0,0) along the *Y*-axis to (0,1), then straight right to (1,1). The diagonal $H = F$ represents zero skill, indicating that the forecasts are completely non-discriminatory. ROC points below the diagonal represent the same level of skillful

[14] See Jolliffe and Stephenson (2003), Stephenson (2000), and Swets and Pickett (1982) for additional analysis on the use of ROC.

performance as they would if reflected about the diagonal, but just mislabeled. Forecast of non-occurrence should be taken as occurrence.

In Figures 4a-4e the ROC curves for the current quarter and the next four quarters are displayed. It can be seen that the ROC for the current quarter (Q0) is located maximally away from the diagonal towards the left upper corner demonstrating the highest discrimination ability of the SPF forecasts, followed by the one-quarter-ahead forecasts. For longer-term forecasts ROCs become rapidly flatter as the forecasting horizon increases. For the four-quarter-ahead forecasts, the ROC mildly snakes around the diagonal line, indicating forecasts have basically no skill or discrimination ability for any value of the threshold. Given the relative costs of type I and type II errors in a particular forecasting situation, an end-user can pick a suitable hit rate (or false alarm rate) of choice along the ROC curve to find the corresponding false alarm rate (or hit rate). This will also give an optimal threshold for making decisions.

The hit rates and false alarm rates for each selected threshold are reported in Table 5, where one can find the mix of hit and false alarm rates that are expected to be associated with each horizon-specific forecast. For example, for achieving a hit rate of 90% with Q0 forecasts, one should use 0.25 as the threshold, and the corresponding false alarm rate is expected to be 0.163. Table 5 also shows that at this threshold value, even though the false alarm rates are roughly around 0.15 for forecast of all horizons, the hit rate steadily declines from 90% for Q0 to only 21% for Q4 - clearly documenting the rapid speed of deterioration in forecast capability as the forecast horizon increases. Though not reported in Table 5, for the same hit rate of 90%, the false alarm rates for Q1 through Q4 forecasts are 0.189 (*w=0.237)*, 0.636 (*w=0.13)*, 0.808 (*w=0.115)* and 0.914 (*w=0.10*) respectively. Thus, for the same hit rate, the corresponding false alarm rates for Q2-Q4 forecasts are so large (64%, 80% and 91% respectively) that they can be considered useless for all practical purposes. Interestingly, the relative inferiority of the Q2 forecasts was not clear in our earlier analysis.

In contrast, Q0 and Q1 forecasts seem to have acceptable operating characteristics. Given the relative costs of two types of classification errors, the end-user can choose an appropriate threshold *w* to minimize the total expected cost of misclassification. This type of optimal decision rule cannot be obtained by the Murphy-Yates decompositions of QPS.[15] Moreover, for forecasting relatively rare business events like recessions, ROC analysis is essential for the probability forecasts to have operational value. This is because, in ROC analysis, the success rate in predicting the predominant event is not part of the goodness of fit measure.

6. CONCLUSION

In this paper we have evaluated the subjective probability forecasts for real GDP declines during 1968-2004 using alternative methodologies developed in psychology and meteorology. *The Survey of Professional Forecasters* record probability forecasts for real GDP declines during the current and next four quarters. We decomposed the traditional Brier's QPS associated with these probability forecasts into calibration, resolution, and alternative variance decompositions. We found conclusive evidence that the shorter run forecasts (Q0-Q1) possess significant skill, and are well calibrated. The resolution or the discrimination ability is also reasonable. Q2 forecasts have borderline value. However, the variance of these forecasts, particularly during cyclical downturns, is significantly more than necessary, given their discriminatory power. The analysis of probability forecasts, thus, shows that forecasters respond *also* to cues that are not related to the occurrence of negative GDP growths. This leads to worse resolutions.

[15] Swets and Pickett (1982) suggest the use of area under ROC and the discrimination distance to find optimal values of the threshold. This, however, makes the choice of the threshold *w* independent of the relative costs of type I and type II errors in a specific forecasting context. On the other hand, in a series of papers Zellner (1986) has formulated the problem of forecasting business cycle turning points in a Bayesian decision theoretic framework allowing for asymmetric costs of misclassification, see Garcia-Ferrer *et al.* (1987), Zeller and Hong (1991) and Zellner *et al.* (1991). In this framework, a low threshold value of 0.25 would imply that the relative cost of a false signal is almost four times the cost of missing a downturn. With SPF data, we find that the optimal threshold gets smaller as the horizon gets larger. This is consistent with the supposition that the cost of a false signal is apt to be smaller for longer-horizon forecasts. However, an asymmetric loss structure is unlikely to be the sole explanation for the rapid deterioration of the forecasts over 5 quarterly horizons.

In contrast, longer run forecasts (Q3-Q4) exhibit poor performance as measured by negative skill scores, low resolutions, dismal ROC measures, and insignificant correlations with actual outcomes. Interestingly, the Seillier-Moiseiwitsch and Dawid (1993) test for perfect forecast validity failed to detect any problem with the longer-term forecasts. However, it is clear from our analysis that our professional forecasters do not have adequate information to forecast meaningfully at horizons beyond 2 quarters; they lack relevant discriminatory cues. Since the SPF panel is composed of professional economists and business analysts who forecast on the basis of models and informed heuristics, their failure for the long-term forecasts may indicate that at the present time forecasting real GDP growth beyond two quarters may not be possible with reasonable type I and Type II errors. Since survey probabilities embody important additional information over point forecasts, an analysis of the probability forecasts provided us with a unique opportunity to understand the reasons for forecast failures. As Granger (1996) has pointed out, in some disciplines forecasting beyond certain horizons is known to be not possible; for instance, in weather forecasting the boundary seems to be four or five days. Our analysis of probability forecasts suggests that in macro GDP forecasts, two quarters appears to be the limit at the present time.

We have also emphasized that for forecasting rare events, it is important to examine the ROC curves where the relative odds for the event can be studied at depth. The analysis also helps find an optimum probability threshold for transforming the probability forecasts to a binary decision rule. In many occasions the selection of the threshold is quite arbitrary. In this regard, ROC analysis provides a simple but an objective criterion, incorporating the end user’s loss function for missed signals and false alarms. The ROC analysis in our case revealed that for a pre-assigned hit rate of (say) 90%, the associated false alarm rates for the Q3-Q4 forecasts are so high that they can be considered useless for all practical purposes.

Other interesting implications of this study are as follows: First, decomposition methodologies introduced in this paper have much broader implications for evaluating model fit in Logit, Probit and other limited dependent variable models. These models generate probabilities of discrete events. Again, often in economics, we try to identify events that are relatively rare (e.g., loan defaults, hospital stays, road accidents, crack babies, etc.) in terms of observable predictors. Usually the model fit criteria look excellent, but the estimated model hardly identifies the small population of interest. Using the evaluation methodology of probability forecasts, one can study the true value of the estimated probability models for out-of-sample predictions.

Second, given the multi-dimension nature of the forecasts and the possible trade-offs between the different characteristics of the forecasts such as calibration and resolution, discrimination ability should be taken as an important characteristic with high priority for the end users. As the ROC analysis revealed, a fundamental issue for forecasting a binary event is to distinguish the occurrence of an event from its non-occurrence. A forecast with higher discrimination ability should certainly be considered a better one over others. As revealed by our analysis, a decent external correspondence may not necessarily represent a truly useful forecast. Instead, it could be just the result of the "hedging" behavior on part of the forecasters. Most importantly, a higher accuracy score can be achieved at the expense of lowered discrimination ability.

Third, considering the fact that the chronologies of the NBER recessions are usually determined long after the recession is over, negative GDP growth projections are probably a reasonable way of tracking business cycles in real time. We have found conclusive evidence that the SPF subjective probability forecasts for the near term are useful in this regard, even though these probability forecasts are characterized by excess variability. In principle, the quality of these forecasts can be improved by further distinguishing factors related to the event from those that are not, while keeping the sensitivity of the forecasts to correct information.

One wonders if the SPF forecasters can be trained to do better. In the current situation, forecasting improvement may not be possible for various reasons. In most psychological and Bayesian learning experiments, the outcomes are readily available and are known with certainty; thus prompt feedback for the purpose of improvement is possible. In contrast, the GDP figures are announced with considerable lag, and are then revised repeatedly. Also, as we have mentioned before, correct and dependable cues for predicting recessions a few quarters ahead may not be available to economists. The excess variability of forecasts and the observed lack of discriminating ability may just be a reflection of that hard reality. It may be the same reason why model-based forecasts over business cycle frequencies have not succeeded in the past. Given the loss/cost structure facing the forecasters and lacking useful cues, issuing low probabilities for future recessions may be the optimal predictions for the forecasters under considerable uncertainty, particularly when the course of the cycle can be manipulated by government policies.

ACKNOWLEDGMENTS:
We are grateful to Roy Batchelor, Richard Cohen, Nigel Harvey, Thad Mirer, David Stephenson, Herman Stekler, Kenneth Wallis and Arnold Zellner for many helpful comments and suggestions. However, we are solely responsible for any remaining errors and omissions.

Fig. 1a: Probability of Decline in Real GDP in the Current Quarter

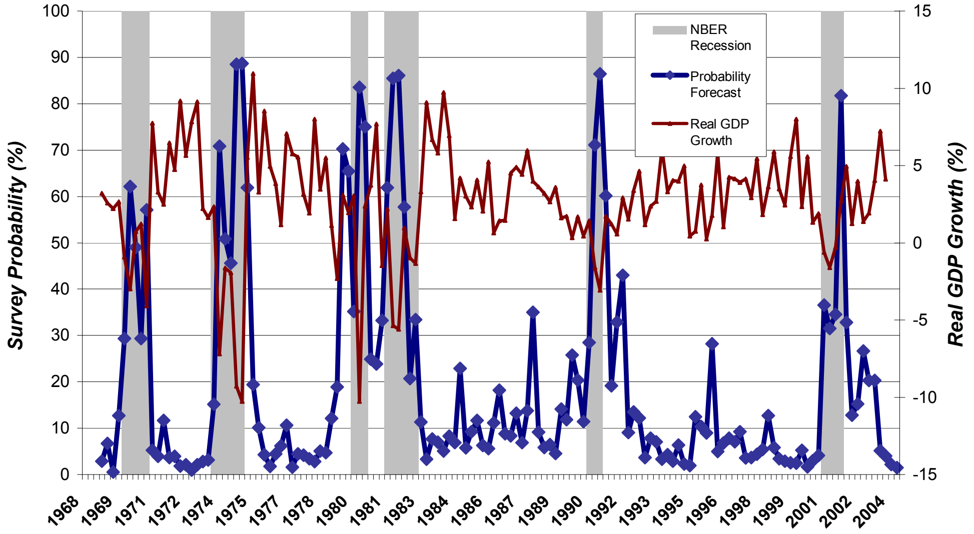


Fig. 1b: Probability of Decline in Real GDP in the Following Quarter

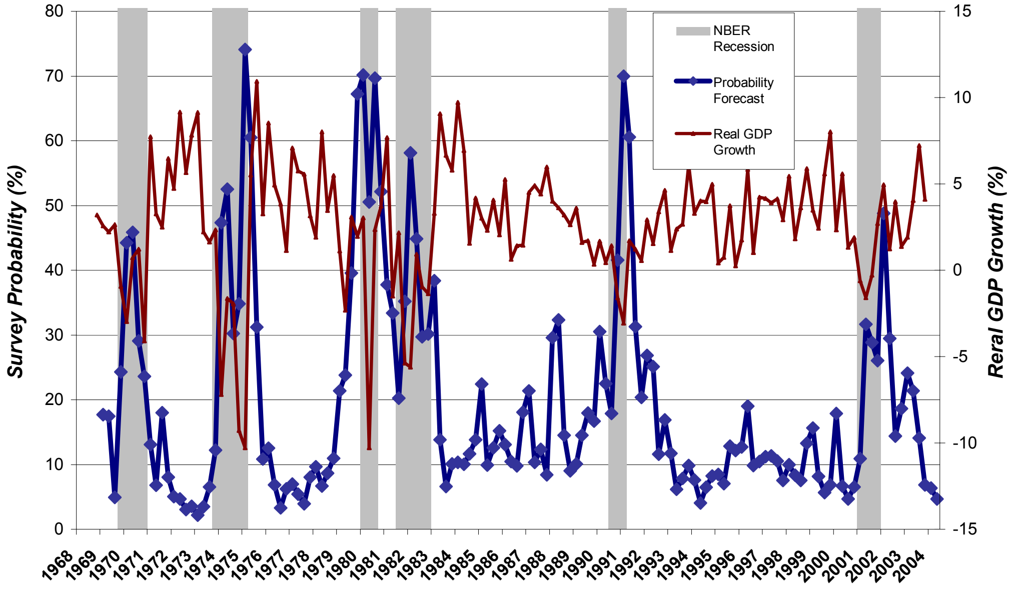


Fig. 1c: Probability of Decline in Real GDP in Second Following Quarter

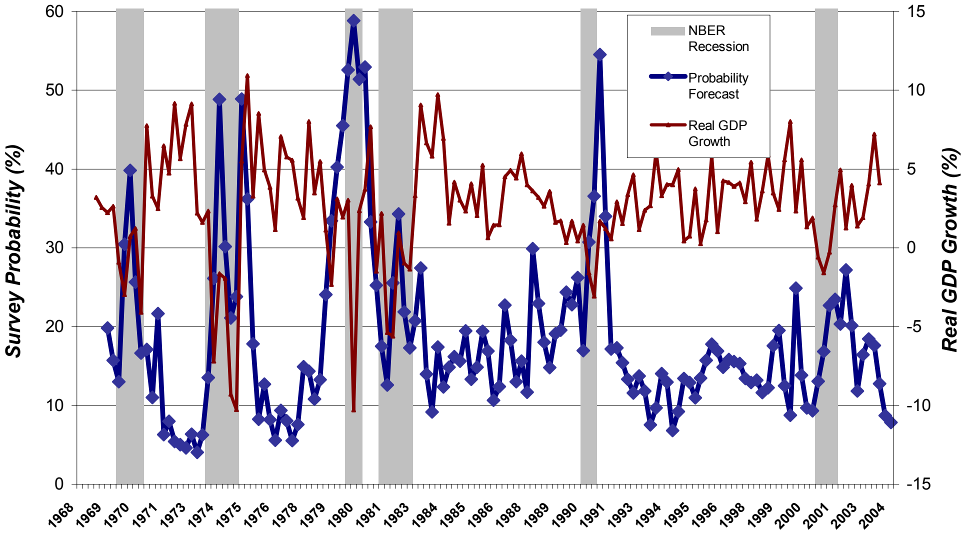


Fig. 1d: Probability of Decline in Real GDP in Third Following Quarter

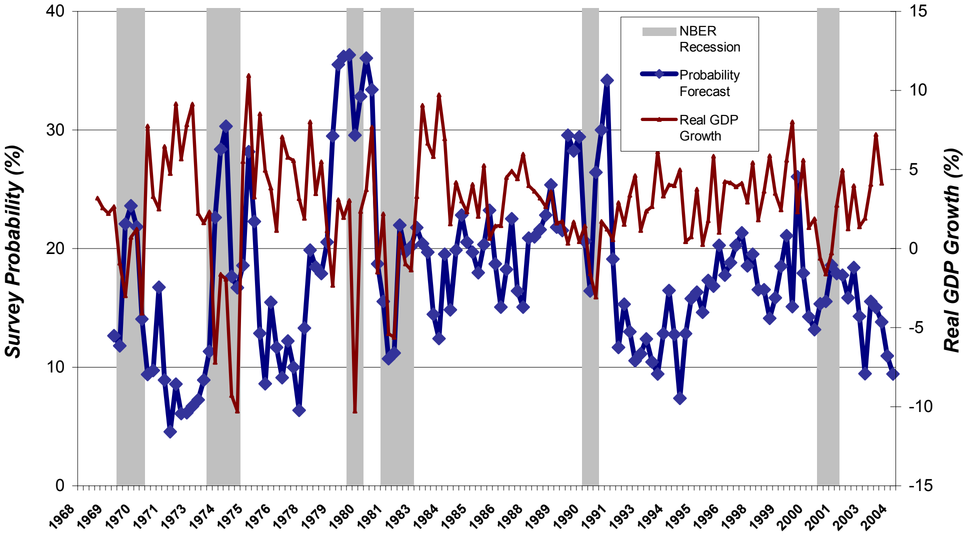


Fig. 1e: Probability of Decline in Real GDP in Fourth Following Quarter

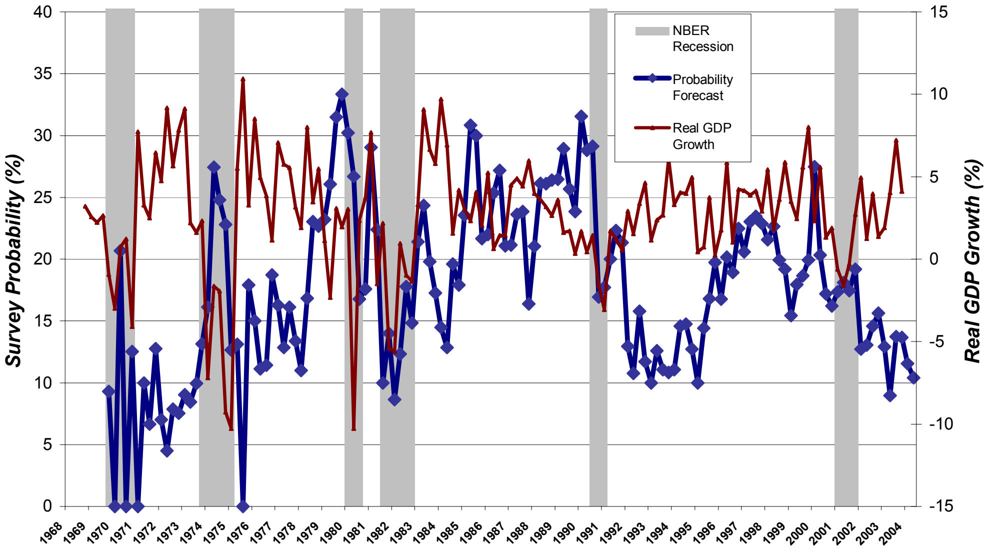


Figure 2a: Attributes Diagram (Q0)

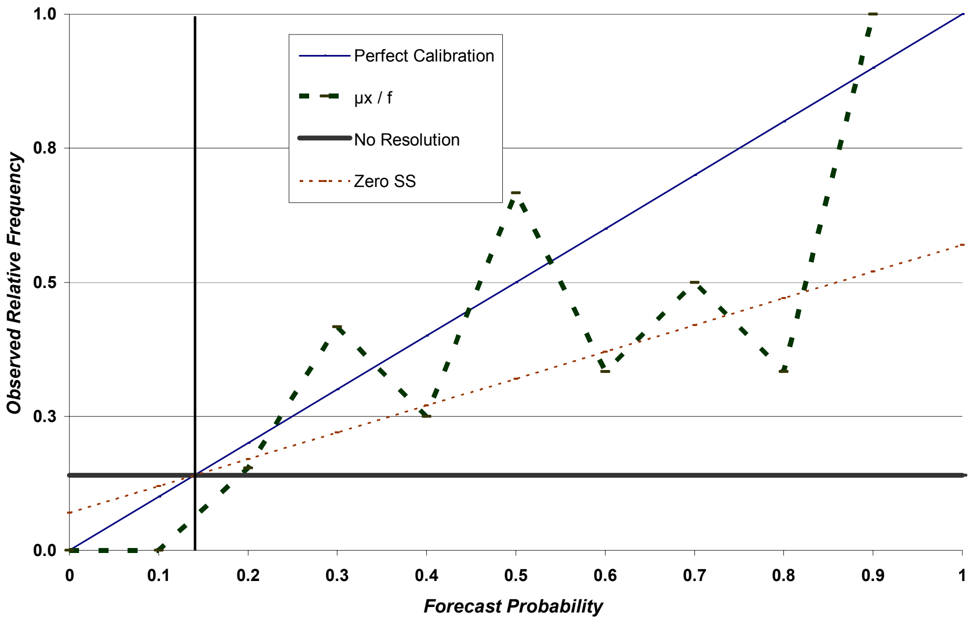


Figure 2b: Attributes Diagram (Q1)

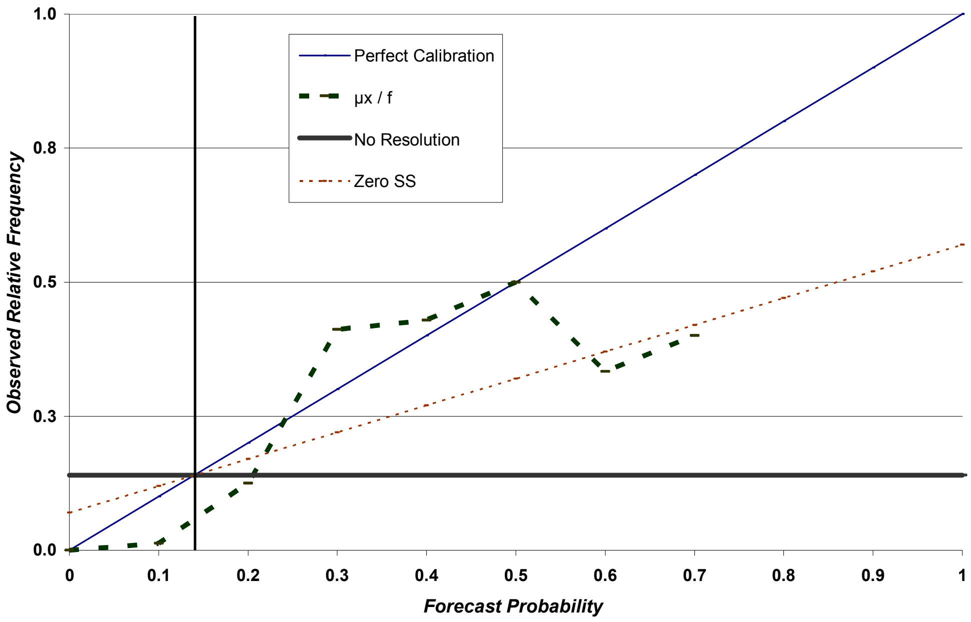


Figure 2c: Attributes Diagram (Q2)

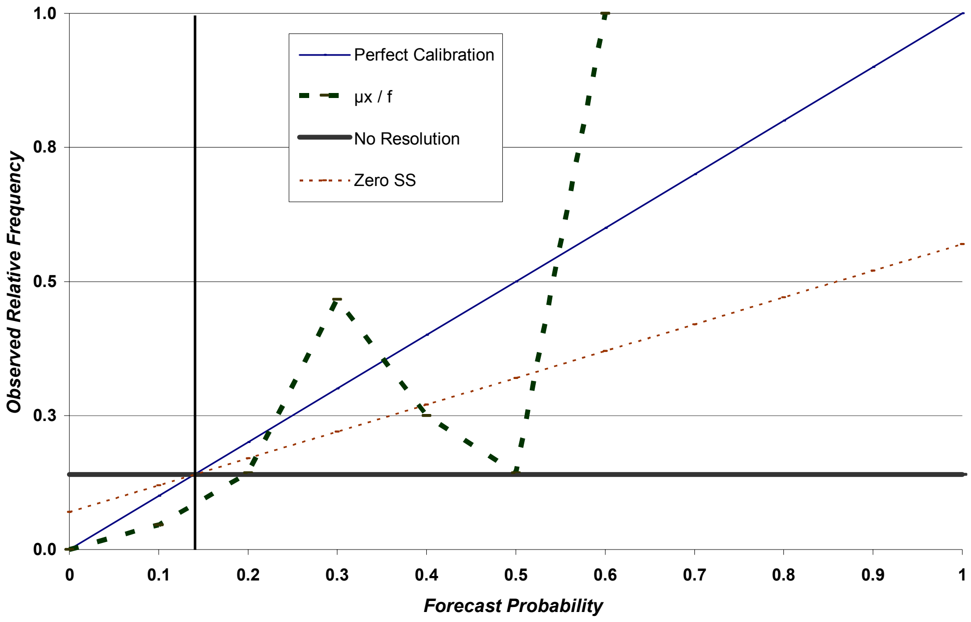


Figure 2d: Attributes Diagram (Q3)

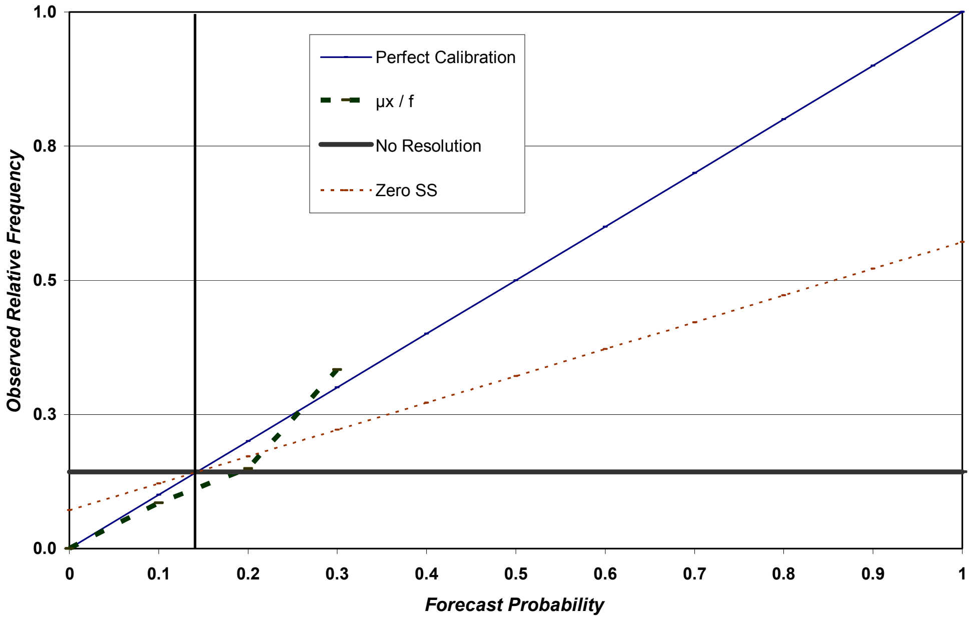


Figure 2e: Attributes Diagram (Q4)

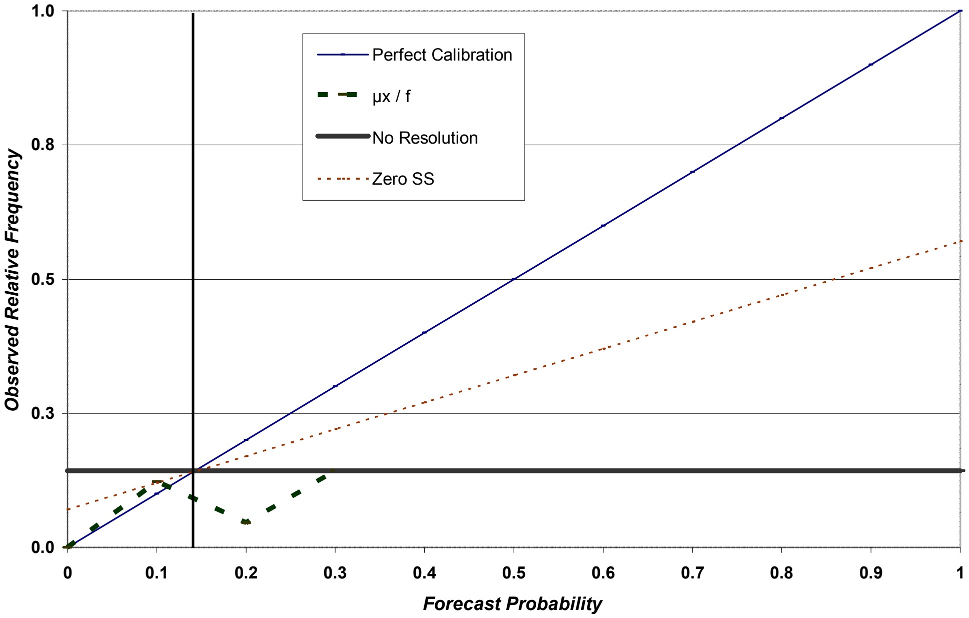


Fig. 3a: Likelihood Diagram (Q0)

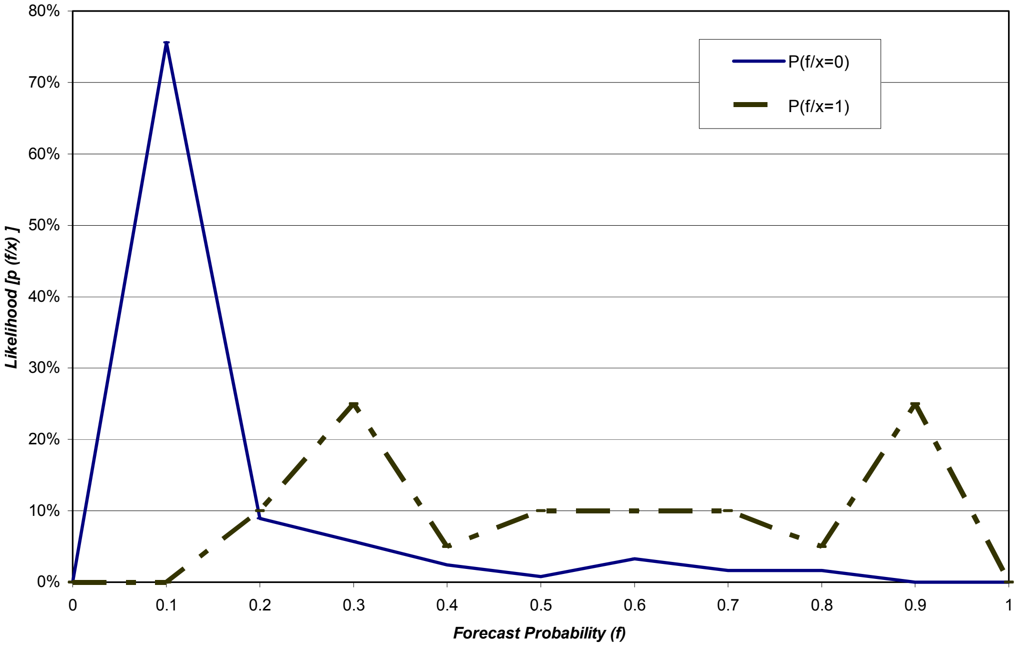


Fig. 3b: Likelihood Diagram (Q1)

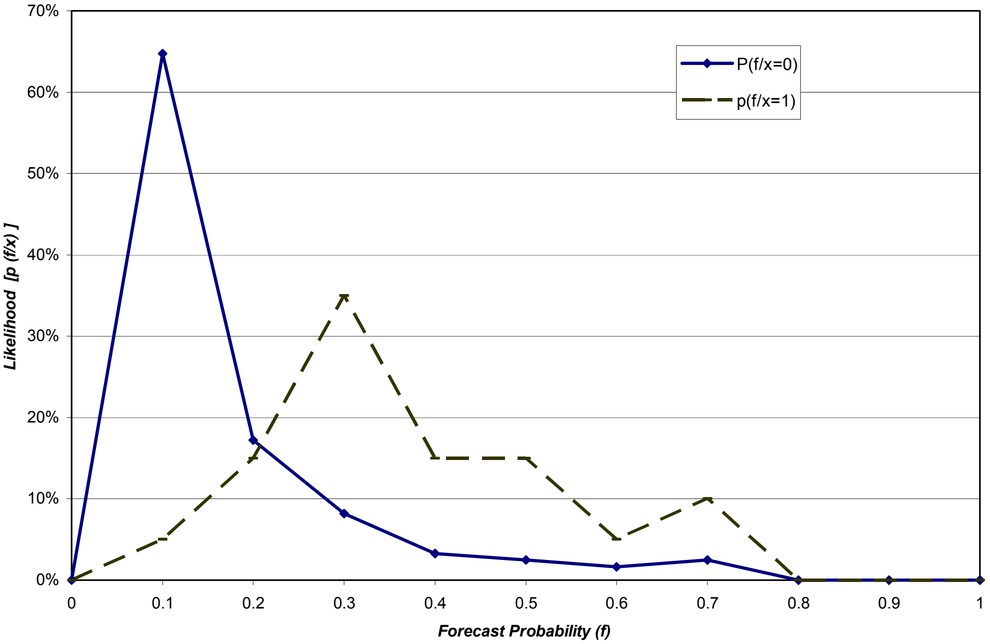


Fig. 3c: Likelihood Diagram (Q2)

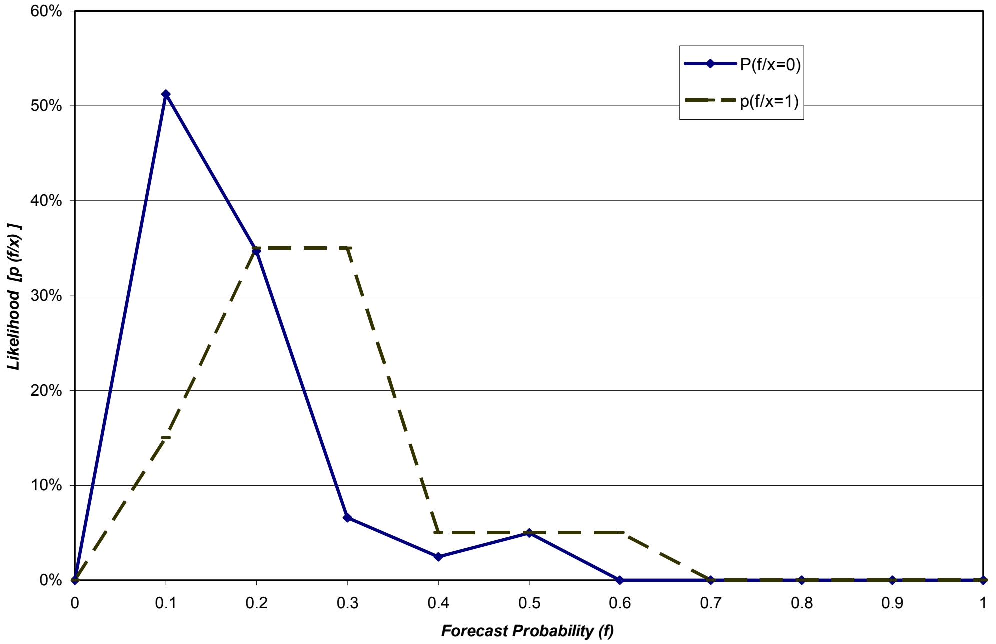


Fig. 3d: Likelihood Diagram (Q3)

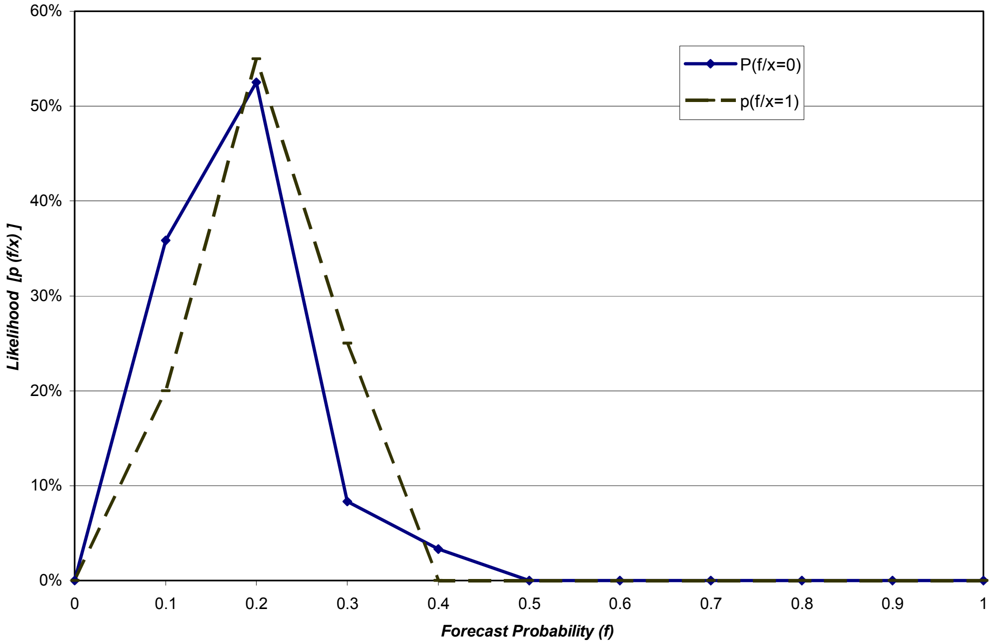


Fig. 3e: Likelihood Diagram (Q4)

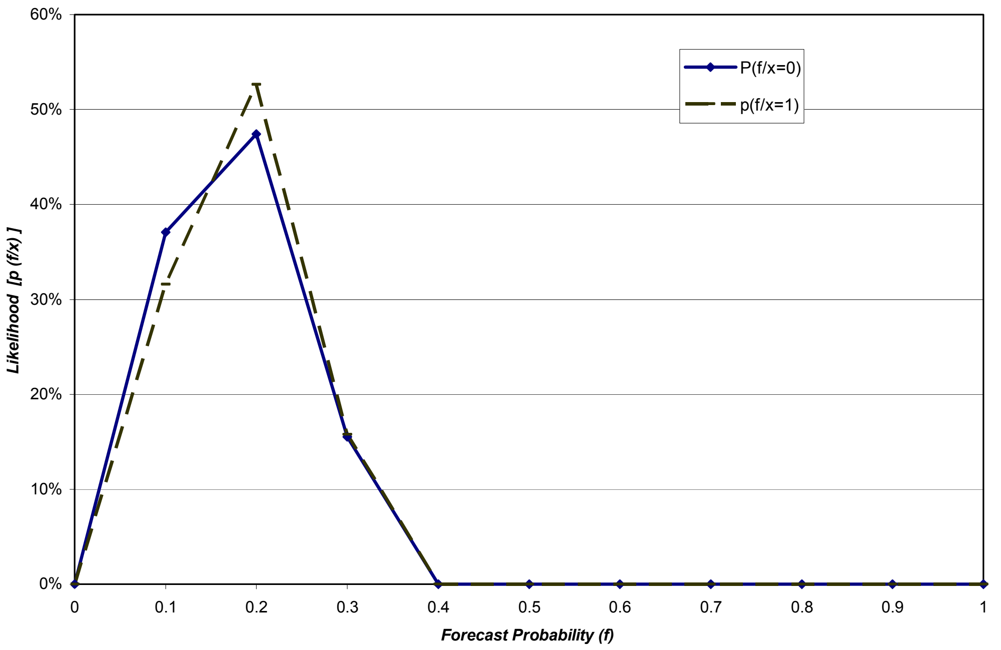
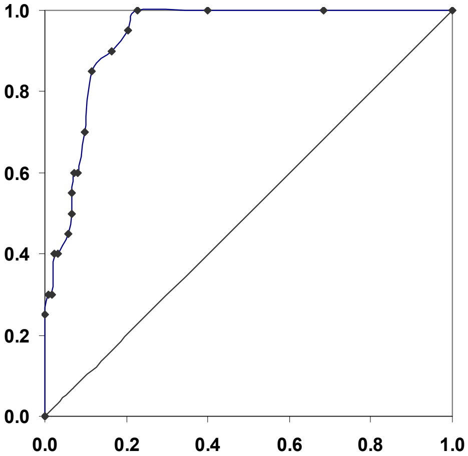
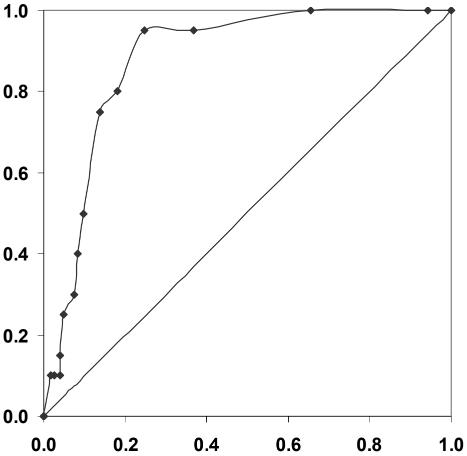


Figure 4a: ROC for Q0



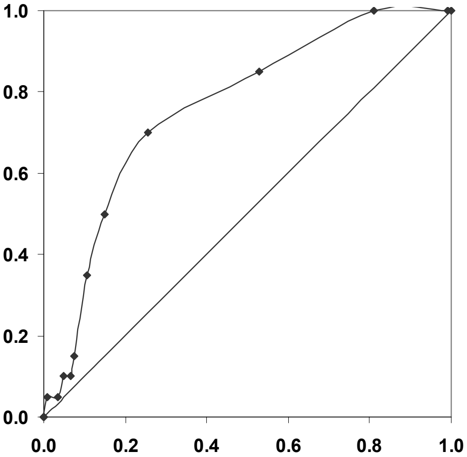
Y-axis: Hit Rate; X-axis: False Alarm Rate

Figure 4b: ROC for Q1



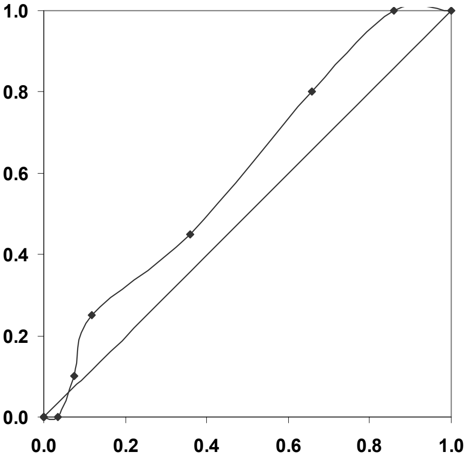
Y-axis: Hit Rate; X-axis: False Alarm Rate

Figure 4c: ROC for Q2



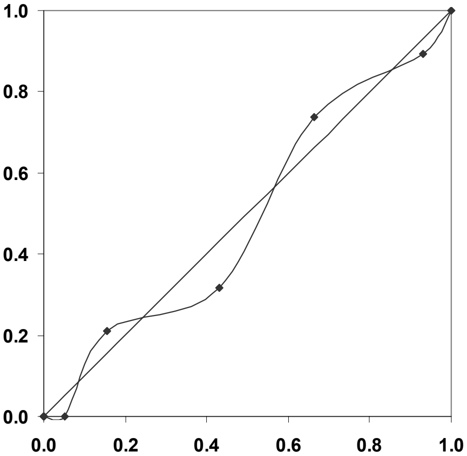
Y-axis: Hit Rate; X-axis: False Alarm Rate

Figure 4d: ROC for Q3



Y-axis: Hit Rate; X-axis: False Alarm Rate

Figure 4e: ROC for Q4



Y-axis: Hit Rate; X-axis: False Alarm Rate

Table 1: Calibration Test

Midpoint	Zj (0)	Zj (1)	Zj (2)	Zj (3)	Zj (4)
0.025	-1.04	-0.55	-0.23	-0.16	-0.16
0.1	-2.38	-2.37	-1.45	-0.43	0.37
0.2	1.00	-0.92	-1.00	-1.10	-0.93
0.3	-0.22	1.01	1.41	0.28	-1.57
0.4	0.94	0.15	-0.61	-1.63	0.00
0.5	-0.82	0.00	-1.89	0.00	0.00
0.6	-0.41	-0.94	0.82	0.00	0.00
0.7	-1.39	-1.46	0.00	0.00	0.00
0.8	1.12	0.00	0.00	0.00	0.00
0.9	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00
χ^2	12.7	10.82	9.74	4.17	3.50
QPS Test (N (0,1))	-1.597	-1.481	-1.393	-1.137	-0.916

Table 2a: Decomposition of Skill Score

Lead Time	SS (Skill Score)	= Association	- Calibration	- Bias
Q0	0.3644	0.3930	0.0017	0.0269
Q1	0.1942	0.2202	0.0000	0.0260
Q2	0.0594	0.0774	0.0015	0.0165
Q3	-0.0019	0.0140	0.0059	0.0100
Q4	-0.0454	0.0000	0.0323	0.0130

Table 2b: Summary Measures of Marginal & Joint Distributions of Forecasts & Observations

Lead Time	Means		Variances		Correlation	Sample
	μ_f	μ_x	Var (f)	Var (x)	Coefficient	Size
Q0	0.1969	0.1399	0.0541	0.1211	0.6269	143
Q1	0.1971	0.1408	0.0272	0.1219	0.4693	142
Q2	0.1868	0.1418	0.0123	0.1226	0.2781	141
Q3	0.1780	0.1429	0.0047	0.1233	0.1184	140
Q4	0.1806	0.1407	0.0040	0.1218	0.0017	135

Table 2c: Summary Measures of Conditional Distribution Given Observations

Lead Time	Means		Variances		Sample	Sample
	$\mu_{f / x = 0}$	$\mu_{f / x = 1}$	Var (f) / x = 0	Var (f) / x = 1	n (x = 0)	n (x = 1)
Q0	0.1383	0.5573	0.0284	0.0632	123	20
Q1	0.1659	0.3875	0.0206	0.0258	122	20
Q2	0.1743	0.2624	0.0110	0.0139	121	20
Q3	0.1747	0.1978	0.0048	0.0039	120	20
Q4	0.1806	0.1809	0.0042	0.0032	116	19

Table 3: Murphy Decomposition

Lead Time	MSE (Accuracy)	= Uncertainty	+ Reliability	- Resolution
Q0	0.0793	0.1211	0.0153	0.0572
Q1	0.1018	0.1219	0.0108	0.0308
Q2	0.1150	0.1226	0.0135	0.0210
Q3	0.1226	0.1233	0.0062	0.0069
Q4	0.1270	0.1218	0.0155	0.0103

Table 4: Yates Decomposition

Lead Time	MSE =	Var (x) +	Δ Var (f) +	Min Var(f) +	$(\mu_f - \mu_x)^2$ -	2*Covar (f,x)
Q0	0.0769	0.1203	0.0330	0.0211	0.0033	0.1008
Q1	0.0977	0.1210	0.0212	0.0059	0.0032	0.0536
Q2	0.1146	0.1217	0.0113	0.0009	0.0020	0.0214
Q3	0.1227	0.1224	0.0046	0.0001	0.0012	0.0057
Q4	0.1265	0.1209	0.0040	0.0000	0.0016	0.0001

Table 5: Hit Rate / False Alarm Rate

		Q0	Q1	Q2	Q3	Q4
0.95	H	0.000	0.000	0.000	0.000	0.000
	F	0.000	0.000	0.000	0.000	0.000
0.85	H	0.250	0.000	0.000	0.000	0.000
	F	0.000	0.000	0.000	0.000	0.000
0.75	H	0.300	0.000	0.000	0.000	0.000
	F	0.016	0.000	0.000	0.000	0.000
0.65	H	0.400	0.100	0.000	0.000	0.000
	F	0.033	0.025	0.000	0.000	0.000
0.55	H	0.500	0.150	0.050	0.000	0.000
	F	0.065	0.041	0.008	0.000	0.000
0.45	H	0.600	0.300	0.100	0.000	0.000
	F	0.073	0.074	0.050	0.000	0.000
0.35	H	0.700	0.500	0.150	0.000	0.000
	F	0.098	0.098	0.074	0.033	0.000
0.25	H	0.900	0.800	0.500	0.250	0.211
	F	0.163	0.180	0.149	0.117	0.155
0.15	H	1.000	0.950	0.850	0.800	0.737
	F	0.228	0.369	0.529	0.658	0.664
0.05	H	1.000	1.000	1.000	1.000	1.000
	F	0.683	0.943	0.992	1.000	1.000

References:

Bessler, D.A. and Ruffley, R. (2004), “Prequential Analysis of Stock Market Returns”, *Applied Economics*, 36, 399-412.

Björkman, M. (1994), “Internal Cue Theory: Calibration and Resolution of Confidence in General Knowledge”, *Organizational Behavior and Human Decision Process*, 58, 386-405.

Braun, P. and Yaniv, I. (1992), “A Case Study of Expert Judgment: Economists’ Probabilities versus Base Rate Model Forecasts”, *Journal of Behavioral Decision Making*, 5, 217-231.

Cramer, J. (1999), “Predictive Performance of the Binary Logit Model in Unbalanced Samples”, *Journal of the Royal Statistical Society*, Series D, (The Statistician), 48, 85-94.

Croushore, D. (1993), “Introducing: The Survey of Professional Forecasters”, *Federal Reserve Bank of Philadelphia Business Review*, November/December, 3-13.

Dawid, A.P. (1984), “Statistical Theory: A Prequential Approach”, *Journal of the Royal Statistical Society*, 147, 279-297.

Dawid, A.P. (1986), “Probability Forecasting,” in S. Kotz, N.L. Johnson and C.B. Mead (eds.), *Encyclopedia of Statistical Sciences* (Vol. 7), New York: Wiley-Interscience, 210-218.

DeGroot, M. H. and Fienberg, S. E. (1983), “A Comparison and Evaluation of Forecasters”, *Journal of the Royal Statistical Society*, Series D, (The Statistician), 32, No. 1/2, 12-22.

Diebold, F. X. and Rudebusch, G. D. (1989), “Forecasting Output with the Composite Leading Index: A Real Time Analysis”, *Journal of the American Statistical Association*, 86, 603-610.

Diebold, F. X. and Rudebusch, G. D. (1991), “Turning Point Prediction with the Composite Leading Index: An *ex ante* Analysis”, in K. Lahiri and G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Approaches*, Cambridge University Press, Cambridge.

Doswell, C. A., Davies-Jones, R. and Keller, D. L. (1990), “On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables”, *Weather and Forecasting*, 5, 576-585.

Filardo, A.J. (1999), “How Reliable Are Recession Prediction Models?” *Federal Reserve Bank of Kansas City Economic Review*, 35-55.

Filardo, A.J. (2004), “The 2001 US Recession: What Did the Recession Prediction Models Tell US?” *Bank of International Settlements, BIS Working Paper* No 148, March.

Fildes, R. and Stekler, H. (2002), “The State of Macroeconomic Forecasts”, *Journal of Macroeconomics,* 24, 435-468.

Fintzen, D. and Stekler, H.O. (1999), “Why Did the Forecasting Fail to Predict the 1990 Recession?” *International Journal of Forecasting*, 15, 309-323.

Garcia-Ferrer, A., Highfield, R. A., Palm, F., and Zellner, A. (1987), “Macroeconomic Forecasting Using pooled International Data,” *Journal of Business and Economic Statistics*, 5, 53-67.

Graham, H. R. (1996), “Is a Group of Forecasters Better Than One? Than None?” *Journal of Business*, 69 (2), 193-232.

Granger, C.W. J. (1996), “Can We Improve the Perceived Quality of Economic Forecasts?” *Journal of Applied Econometrics*, 11, 455-473.

Greene, W. H. (2003), *Econometric Analysis*, 5th ed. Prentice Hall, Upper Saddle River, N.J.

Hamilton, J.D. (1989), “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle”, *Econometrica*, (57), 375-384.

Jolliffe, I.T and Stephenson, D.B. Eds. (2003), *Forecasting Verification: A Practitioner’s Guide in Atmospheric Science*, John Wiley & Son Limited.

Juhn, G. and Loungani, P. (2002), “Further Cross-Country Evidence on the Accuracy of the Private Sector Output Forecasts”*, IMF Staff Papers*, 49, 49-64.

Kling, J.L.(1987), “Predicting the Turning Points of Business and Economic Time Series”, *Journal of Business*, 60, 201-238.

Krane, S. (2003), “An Evaluation of Real GDP forecasts: 1996-2001”, *Economic Perspectives* – Federal Reserve Bank of Chicago, 1Q, 2-21.

Lahiri, K and Wang, J. G. (1994), "Predicting Cyclical Turning Points with Leading Index in a Markov Switching Model", *Journal of Forecasting*, 245-263.

McNees, S.K. (1991), “Forecasting Cyclical Turning Points: The Record in the Past Three Recessions”, in Lahiri, K. and Moore, G.H. (eds), *Leading Economic Indicators: New Approaches and Forecasting Records*, Cambridge: Cambridge University Press, 149-165.

Murphy, A. (1972), “Scalar and Vector Partitions of the Probability Score: Part I. Two-state Situation”, *Journal of Applied Meteorology*. 11, 273-282.

Murphy, A.H., (1988), “Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient”, *Monthly Weather Review*, 116, 2417-2424.

Murphy, A.H. (1991), “Probabilities, Odds, and Forecasters of Rare Events”, *Weather and Forecasting*, 6, 302-306.

Murphy, A.H. and Winkler, R.L. (1992), “Diagnostic Verification of Probability Forecasts”, *International Journal of Forecasting,* 7, 435-435.

Neftci, S.N. (1984), “Are Economic Time Series are Asymmetric over the Business Cycle?” *Journal of Political Economy*, 92, 305-328.

Seillier-Moiseiwitsch, F. and Dawid, A.P. (1993), “On Testing the Validity of Sequential Probability Forecasts”, *Journal of the American Statistical Association*, 88 (421), 355-359.

Stephenson, D.B. (2000), “Use of the ‘Odds Ratio’ for Diagnosing Forecast Skill”, *Weather and Forecasting*, 15, 221-232.

Stock J. H. and Watson, M. W. (1991), “A Probability Model of the Coincident Economic Indicators” in K. Lahiri and G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records*, Cambridge University Press, 63-85.

Stock, J.H. and Watson, M.W. (1993), “A Procedure for Predicting Recessions with Leading Indicators: Econometric Issues and Recent Experience”, in J.H. Stock and M.W. Watson (eds.), *New Research on Business Cycles, Indicators, and Forecasting*, University of Chicago Press, Chicago, 95-153.

Stock, J.H. and Watson, M.W. (2003), “How Did Leading Indicator Forecasts Perform During the 2001 Recession?” *Federal Reserve Bank of Richmond Economic Quarterly*, 89 (3), 71-90.

Swets, J.A. and Pickett, R.M. (1982), *Evaluation of Diagnostic System: Methods from Signal Detection Theory*, Academic Press. New York.

Yates, J.F. (1982), “External Correspondence: Decompositions of the Mean Probability Score”, *Organizational Behavior and Human Performance*, 30,132-156.

Yates, J.F. (1994), “Subjective Probability Accuracy Analysis”, in G. Wright and P. Ayton, (eds.), *Subjective Probability,* John Wiley, Chichester, UK, 381- 410.

Yates, J.F. and Curley, S.P. (1985), “Conditional Distribution Analysis of Probabilistic Forecasts”, *Journal of Forecasting,* 4, 61-73.

Zarnowitz, V. and Moore, G.H. (1991), “Forecasting recessions under the Gramm-Rudmann-Hollings Law”, in K. Lahiri and G.H. Moore (eds.), *Leading Economic*

Indicators: New Approaches and Forecasting Records, Cambridge University Press: Cambridge, 257-273.

Zellner, A. (1986), "Bayesian Estimation and Prediction Using Asymmetric Loss Functions," *Journal of The American Statistical Association*, 81, 446-451.

Zellner, A. and Hong, C. (1991), "Bayesian Methods for Forecasting Turning Points in Economic Time-Series: Sensitivity of Forecasts to Asymmetry of Loss Structures," in K. Lahiri and G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records*, Cambridge University Press: Cambridge, 129-140.

Zellner, A., Hong, C., and Min, C-K. (1991), "Forecasting Turning Points in International Growth Rates Using Bayesian Exponentially Weighted Autoregression, Time Varying Parameter, and Pooling Techniques", *Journal of Econometrics*, 49, 275-304.