3

Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters

FRANCIS X. DIEBOLD, ANTHONY S. TAY,.
AND KENNETH F. WALLIS

1 Introduction

Economic decision makers routinely rely on forecasts to assist their decisions. Until recently, most forecasts were provided only in the form of point forecasts, although forecasters sometimes attached measures of uncertainty, such as standard errors or mean absolute errors, to their forecasts. Recently, the trend has been to accompany point forecasts with a more complete description of the uncertainty of the forecasts, such as explicit interval or density forecasts. An interval forecast indicates the likely range of outcomes by specifying the probability that the actual outcome will fall within a stated interval. The probability may be fixed, at say 0.95, and the associated interval may then vary over time, or the interval may be fixed, as a closed or open interval, and the forecast probability presented, as in the statement that "our estimate of the probability that inflation next year will be below 2.5 percent is p." A density forecast is stated explicitly as a density or probability distribution. This may be presented analytically, as in "we estimate that next year's inflation rate is normally distributed around an expected value of 2 percent with a standard deviation of 1 percent," or it may be presented numerically, as when a histogram is reported.

Density forecasts were rarely seen until recently but are becoming more common. In finance, practical implementation of recent theoretical developments has dramatically increased the demand for density forecasts; the boom-

• Dean Croushore and Tom Stark provided valuable assistance and advice, and an anonymous referee provided helpful comments. We thank the National Science Foundation, the Sloan Foundation, the University of Pennsylvania Research Foundation, the National University of Singapore and the Economic and Social Research Council for support. Our collaboration was initiated at the 1997 UC San Diego Conference on Financial Econometrics organized by Rob Engle, Clive Granger and Gloria Gonzalez-Rivera.

ing field of financial risk management, for example, is effectively dedicated to providing density forecasts of changes in portfolio value, as revealed by a broad reading of literature such as J.P. Morgan (1996). There is also a growing literature on extracting density forecasts from options prices, which includes Aït-Sahalia and Lo (1998) and Söderlind and Svensson (1997). In macroeconomics, there has also been increased discussion of density forecasts recently, in response to criticism of the lack of transparency of traditional forecasting practice, and to demands for acknowledgment of forecast uncertainty in order to better inform the discussion of economic policy. Macroeconomic density forecasts are the subject of this article.

In the USA the Survey of Professional Forecasters has, since its introduction in 1968, asked respondents to provide density forecasts of inflation and growth. In the early days of the survey these received little attention, with the notable exception of Zarnowitz and Lambros (1987); more recently the distributions, averaged over respondents, have featured in the public release of survey results. In the UK the history is much shorter. In November 1995 the National Institute of Economic and Social Research began to augment its long-established macroeconomic point forecasts with estimates of the probability of the government's inflation target being met and of there being a fall in GDP. This was extended in February 1996 to a complete probability distribution of inflation and growth forecasts. In the same month the Bank of England launched the presentation of an estimated probability distribution of possible outcomes surrounding its conditional projections of inflation. In November 1996 the Treasury's Panel of Independent Forecasters, following repeated suggestions by one of the present authors, reported its individual members' density forecasts for growth and inflation, using the same questions as the US Survey of Professional Forecasters. Our success was short-lived, however, as the new Chancellor of the Exchequer dissolved the panel shortly after taking office in May 1997.

The production and publication of any kind of forecast subsequently requires an evaluation of its quality. For point forecasts, there is a large literature on the *ex post* evaluation of *ex ante* forecasts, and a range of techniques has been developed, recently surveyed by Wallis (1995) and Diebold and Lopez (1996). The evaluation of interval forecasts has a much newer literature (Christoffersen, 1998), as does the evaluation of density forecasts. In this article we use the methods of Diebold, Gunther, and Tay (1998), augmented with resampling procedures, to evaluate the density forecasts of inflation contained in the Survey of Professional Forecasters. Forecasts of inflation are of intrinsic interest, especially in the monetary policy regime of inflation targeting that is common to many OECD economies, and it is also of interest to demonstrate the use of new tools for forecast evaluation and their applicability even in very small samples. As with most of the forecast evaluation literature we pay no attention to the construction of the forecast, and consider only the assessment of its adequacy, after the fact. That is, because little is known about the construction of the density forecasts reported by the survey respondents, we concentrate on the outputs, not the inputs. The density forecast could be based on a formal statistical or econometric model, an ARCH model for a single financial time series or a large-scale macroeconometric model for aggregate macroeconomic variables, for example, or it could be based on more subjective approaches, blending the forecaster's judgment informally with a model-based forecast or using expert elicitation methods.

The remainder of this article is organized as follows. In Section 2 we present a brief description of the Survey of Professional Forecasters, its advantages and disadvantages, leading to our selection of the series of first-quarter current-year mean density forecasts of inflation for evaluation. In Section 3 we develop our evaluation methods, based on the series of probability integral transforms of realized inflation with respect to the forecast densities and the null hypothesis that this is a series of independent uniformly distributed random variables. We present the results in Section 4, and we conclude in Section 5.

2 The Survey of Professional Forecasters

The Survey of Professional Forecasters (SPF) is the oldest quarterly survey of macroeconomic forecasters in the USA. The survey was begun in 1968 as a joint project by the Business and Economic Statistics Section of the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER) and was originally known as the ASA–NBER survey. Zarnowitz (1969) describes the original objectives of the survey, and Zarnowitz and Braun (1993) provide an assessment of its achievements over its first 22 years. In June 1990 the Federal Reserve Bank of Philadelphia, in cooperation with the NBER, assumed responsibility for the survey, at which time it became known as the Survey of Professional Forecasters (see Croushore, 1993).

The survey is mailed four times a year, the day after the first release of the National Income and Product Accounts data for the preceding quarter. Most of the questions ask for point forecasts, for a range of variables and forecast horizons. In addition, however, density forecasts are requested for aggregate output and inflation. The output question was unfortunately switched from nominal to real in the early 1980s, thereby rendering historical evaluation of the output forecasts more difficult, whereas the inflation question has no such defect and provides a more homogeneous sample. Thus we focus on the density forecasts of inflation. Each forecaster is asked to attach a probability to each of a number of intervals, or bins, in which inflation might fall, in the current year and in the next year. The definition of inflation is annual, year over year. The probabilities are averaged over respondents, and for each bin the SPF reports the mean probability that inflation will fall in that bin, in the current year and in the next year. The report on the survey results that was previously published in the *NBER Reporter* and the *American Statistician* did not always refer to the density forecasts, and sometimes combined bins, but means for all the bins in the density forecasts have been included in the Philadelphia Fed's press release since 1990, and the complete results dating from 1968 are currently available on their Web page (http://www.phil.frb.org/econ/spf/spfpage.html). This mean probability distribution is typically viewed as a representative forecaster and is our own focus of attention. The mean forecast was the only one available to analysts and commentators in real time.

There are a number of complications, including:

(a) The number of respondents over which the mean is taken varies over time, with a low of 14 and a high of 65.

(b) The number of bins and their ranges have changed over time. From 1968:4 to 1981:2 there were 15 bins, from 1981:3 to 1991:4 there were 6 bins, and from 1992:1 onward there are 10 bins.

(c) The base year of the price indexes has changed. For surveys on or before 1975:4, the base year is 1958, from 1976:1 to 1985:4 the base year is 1972, and from 1986:1 to 1991:4 the base year is 1982. Beginning in 1992:1, the base year is 1987.

(d) The price index used to define inflation in the survey has changed over time. From 1968:4 to 1991:4 the SPF asked about inflation as assessed via the implicit GNP deflator, and from 1992:1 to 1995:4 it asked about inflation as assessed via the implicit GDP deflator. Presently the SPF asks about inflation as assessed via the chain-weighted GDP price index.

(e) The forecast periods to which the SPF questions refer have changed over time. Prior to 1981:3, the SPF asked about inflation only in the current year, whereas it subsequently asked about inflation in the current year and the following year. Errors occurred in 1985:1, 1986:1 and 1990:1, when the first annual forecast was requested for the previous year and the second forecast for the current year, as opposed to the current and the following year.

Most of the complications (e.g., a, b, c and d) are minor and inconsequential. Complication (e), on the other hand, places very real constraints on what can be done with the data. It is apparent, however, that the series of first-quarter current-year forecasts represents an unbroken sample of annual three-quarter ahead inflation density forecasts, with non-overlapping innovations. (If the information set consists only of data up to the final quarter of the preceding year, then this is a conventional annual series of one-step-ahead forecasts; it is likely, however, that information on the current year available in its first few weeks is also used in constructing forecasts.) The sample runs from 1969 to 1996, for a total of 28 annual observations (densities), which form the basis of our examination of inflation density forecast adequacy.

3 Evaluating Inflation Density Forecasts

We evaluate the forecasts using the methodology proposed by Diebold, Gunther, and Tay (1998), the essence of which is consideration of the series of probability integral transforms of realized inflation $\{y_t\}_{t=1}^{28}$ with respect to the forecast densities $\{p_t(y_t)\}_{t=1}^{28}$. That is, we consider the series:

$$\{z_t\}_{t=1}^{28} = \left\{ \int_{-\infty}^{y_t} p_t(u)du \right\}_{t=1}^{28}.$$

Diebold, Gunther, and Tay (1998) show that if the density forecasts are optimal (in a sense that they make precise), then $\{z_t\}_{t=1}^{28} \overset{\text{iid}}{\sim} \mathrm{U}(0, 1)$. The basic idea is to check whether the realizations y_t come from the forecast densities $p_t(y_t)$ by using the standard statistical result that, for a random sample from a given density, the probability integral transforms of the observations with respect to the density are iid U(0, 1), extended to allow for potentially time-varying densities. In a forecasting context, independence corresponds to the usual notion of the efficient use of an information set, which implies the independence of a sequence of one-step-ahead errors. For our inflation density forecasts, an "error" is an incorrect estimate of the probability that inflation will fall within a given bin; a correct estimate of the tail area probability, for example, implies that we observe the same relative frequency of correspondingly extreme forecast errors, in the usual sense of the discrepancy between point forecast and actual outcome for inflation.

Formal tests of density forecast optimality face the difficulty that the relevant null hypothesis—iid uniformity of z—is a joint hypothesis. For example, the classical test of fit based on Kolmogorov's D_n-statistic, the maximum absolute difference between the empirical cumulative density function (c.d.f.) and the hypothetical (uniform) c.d.f., rests on an assumption of random sampling. The test is usually referred to as the Kolmogorov–Smirnov test, following Smirnov's tabulation of the limiting distribution of D_n and introduction of one-sided statistics, while other authors have provided finite-sample tables (see Stuart and Ord, 1991, §30.37). Little is known, however, about the impact on the distribution of D_n of departures from independence; thus test outcomes in either direction may be unreliable whenever the data are not generated by random sampling. More generally the test is not constructive, in that if rejection occurs, the test itself provides no guidance as to *why*.

More revealing methods of exploratory data analysis are therefore needed to supplement formal tests. To assess unconditional uniformity we use the obvious graphical tools, estimates of the density and c.d.f. We estimate the density with a simple histogram, which allows straightforward imposition of the constraint that z has support on the unit interval, in contrast to more sophisticated procedures such as kernel density estimates with the standard kernel functions. To assess whether z is iid, we again use the obvious graphical tool, the correlogram. Because we are interested not only in linear dependence but also in other forms of nonlinear dependence such as conditional heteroskedasticity, we examine both the correlogram of $(z - \overline{z})$ and the correlogram of $(z - \overline{z})^2$.

It is useful to place confidence intervals on the estimated histogram and correlograms, in order to help guide the assessment. There are several complications, however. In order to separate fully the desired U(0, 1) and iid properties of z, we would like to construct confidence intervals for histogram bin heights that condition on uniformity but that are robust to dependence of unknown form. Similarly, we would like to construct confidence intervals for the autocorrelations that condition on independence but that are robust to non-uniformity. In addition, the SPF sample size is small, so we would like to use methods tailored to the specific sample size.

Unfortunately, we know of no asymptotic, let alone finite-sample, method for constructing serial-correlation-robust confidence intervals for histogram bin heights under the U(0, 1) hypothesis. Thus we compute histogram bin height intervals under the stronger iid U(0, 1) assumption, in which case we can also compute the intervals tailored to the exact SPF sample size, by exploiting the binomial structure. For example, for a 5-bin histogram formed from 28 observations, the number of observations falling in any bin is distributed binomial (28, 5/28) under the iid U(0, 1) hypothesis. (This formulation relates to each individual bin height when the other four bins are combined, and the intervals should not be interpreted jointly.)

To assess the significance of the autocorrelations, we construct finite-sample confidence intervals that condition on independence but that are robust to deviations from uniformity by sampling with replacement from the observed z series and building up the distribution of the sample autocorrelations. The sampling scheme preserves the unconditional distribution of z while destroying any serial correlation that might be present.

Two practical issues arise in the construction of the z series. The first concerns the fact that the forecasts are recorded as discrete probability distributions, not continuous densities, and so we use a piecewise linear approximation to the c.d.f. For example, suppose the forecast probability for $y < 4$ is 0.4 and the forecast probability for $4 \leq y < 5$ is 0.3. If the realization of y is 4.6, then we compute z as $0.4 + 0.6(0.3) = 0.58$. Further, the two end bins are open; they give the probabilities of y falling above or below certain levels. When a realization falls in one of the end bins, to apply the piecewise linear approximation we assume that the end bins have the same width as all the other bins. This occurs for only three observations, and in each case the realized inflation rate is very close to the interior boundary of the end bin.

The second issue is how to measure realized inflation: whether to use real-time or final-revised data, and for which inflation concept. As regards the use of real-time vs. final-revised data, we take the view that forecasters try to forecast the "true" inflation rates, the best estimates of which are the final revised values. Thus we use the most recently revised values as our series for realized inflation. Regarding the inflation concept, we noted earlier that the price index used to define inflation in the survey has changed over time from the implicit GNP deflator to the implicit GDP deflator to the chain-weighted price index. Accordingly, we measure realized inflation as the final revised value of the inflation concept about which the survey respondents were asked. From 1969 to 1991 we use the percentage change in the implicit GNP deflator, from 1992 to 1995 we use the percentage change in the implicit GDP deflator, and for 1996 we use the percentage change in the chain-weighted price index.

Two previous studies of the SPF inflation density forecasts merit discussion. Zarnowitz and Lambros (1987) use the survey results to draw the important distinction between uncertainty, as indicated by the spread of the probability distribution of possible outcomes, and disagreement, as indicated by the dispersion of respondents' (point) forecasts: consensus among forecasters need not imply a high degree of confidence about the commonly predicted outcome. Zarnowitz and Lambros find that the variance of the point forecasts tends to understate uncertainty as measured by the variance of the density forecasts. The former varies much more over time than the latter, although the measures of consensus and certainty (or the lack thereof) are positively correlated. Zarnowitz and Lambros also find that expectations of higher inflation are associated with greater uncertainty. Throughout their paper, however, they summarize the individual density forecasts by their means and standard deviations prior to averaging over respondents; thus they use only part of the information in the density forecasts.

McNees and Fine (1996) evaluate the individual inflation density forecasts of a sample of 34 forecasters who responded to the survey on at least 10 occasions. They proceed by calculating the implied 50 percent and 90 percent prediction intervals, and test whether the actual coverage—the proportion of occasions on which the outcome fell within the interval—corresponds to the claimed coverage, 50 percent or 90 percent as appropriate, using the binomial distribution. Again, only part of the information in the density forecasts is used. Moreover, even in the more limited framework of interval forecast evaluation, the McNees–Fine procedure examines only *unconditional* coverage, whereas in the presence of dynamics it is important to examine *conditional* coverage, as in Christoffersen (1998). Put differently, in the language of density forecast evaluation, McNees and Fine implicitly *assume* that z is iid in order to invoke the binomial distribution; they test only whether z is unconditionally $U(0, 1)$.

4 Results

We show the basic data on realized inflation and "box-and-whisker" plots representing the density forecasts in Figure 1. The bottom and top of the box are the 25 percent and 75 percent points, the interior line is the median, the bottom whisker is the 10 percent point, and the top whisker is the 90 percent point. The box-and-whisker plots point to a number of features of the forecasts and their relationship to the realizations. First, comparing forecasts and realizations, similar patterns to those observed by Zarnowitz and Braun (1993, 30–1) in the distribution of individual point forecasts for the period 1968:4–1990:1 can be seen: "in 1973–74, a period of supply shocks and deepening recession, inflation rose sharply and was greatly underestimated. . . . The same tendency to underpredict also prevailed in 1976–80, although in somewhat weaker form. . . . In between, during the recovery of 1975–76, inflation decreased markedly and was mostly overestimated. Another, much longer disinflation occurred in 1981–85. . . . Here again most forecasters are observed to overpredict inflation. . . . Finally, in 1986–89, inflation . . . was generally well predicted," and this has been maintained up to the end of our present sample, when the errors, although persistently of the same sign, are relatively small. There is also evidence of adaptation: although inflation is *unexpectedly* high when it initially turns high, and *unexpectedly* low when it initially falls, forecasters do eventually catch up.

Second, the data seem to accord with the claim that the level and uncertainty of inflation are positively correlated, as suggested by Friedman (1977). Although this hypothesis has typically been verified by relating the variability of inflation to its actual level, in a forecasting context the relevant hypothesis is that expectations of high inflation are associated with increased uncertainty, and this is verified for a shorter sample of these data by Zarnowitz and Lambros (1987), using different techniques, as noted above. In Figure 1 the forecasts for 1975 and 1980 immediately catch the eye, with two of the largest values of the interdecile range—the distance between the whiskers—corresponding to two of the highest median forecasts. Overall there is a strongly significant positive association between these measures; the coefficient in a regression of the interdecile range on the median forecast has a p-value of 0.0198 (with allowance made for positive residual autocorrelation, discussed below). On the other hand the forecasts for 1986 and 1987 are outliers: these give the two largest values of the interdecile range, at relatively low median forecasts (and yet lower realizations). Perhaps this reflects genuine uncertainty about the impact of the fall in the world price of oil, or simply indicates sampling problems, because the number of survey respondents was falling through the late 1980s, prior to revival of the survey by the Philadelphia Fed.

Third, there has been a gradual tightening of the forecast densities since

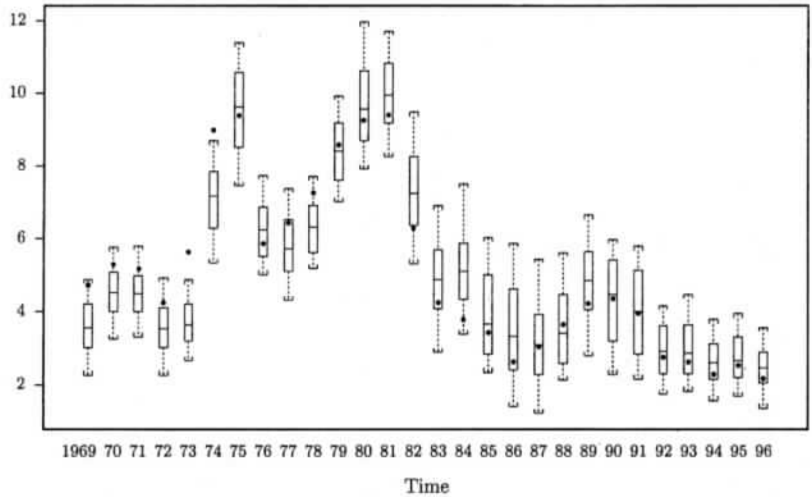

Figure 1. Inflation forecasts and realizations

Notes: The density forecasts are represented by box-and-whisker plots. The boxes represent the inter-quartile range of the forecasts, and the inner line represents the median; the tails represent the 10th and 90th percentiles. We represent inflation realizations with •.

the late 1980s, perhaps due to a reduction of perceived likely supply and demand shocks, an increase in central bank credibility, a reduction in uncertainty associated with the lower level of inflation, or some combination of these. The distributions nevertheless seem to be still too dispersed, because most of the realizations over this period fall squarely in the middle of the forecast densities.

Next, we compute the z series by integrating the forecast densities up to the realized inflation rate, period by period, and we plot the result in Figure 2, in which large values correspond to unexpectedly high values of realized inflation, and conversely. Even at this simple graphical level, deviations of z from iid uniformity are apparent, as z appears serially correlated. In the first half of the sample, for example, z tends to be mostly above its average, whereas in the second half of the sample it appears that the representative forecaster overestimated the uncertainty of inflation, because most of the values of z cluster around 0.4, and they vary little compared to the first half of the sample. This is the counterpart to the observation in Figure 1 that most of the recent realizations are near the middle of the forecast densities, a result that diverges from Chatfield (1993) and the literature he cites, which often finds that forecasters are overconfident, in that their interval forecasts are too *tight*, not too wide.

To proceed more systematically, we examine the distributional and autocorrelation properties of z. We show the histogram and empirical c.d.f. of z in

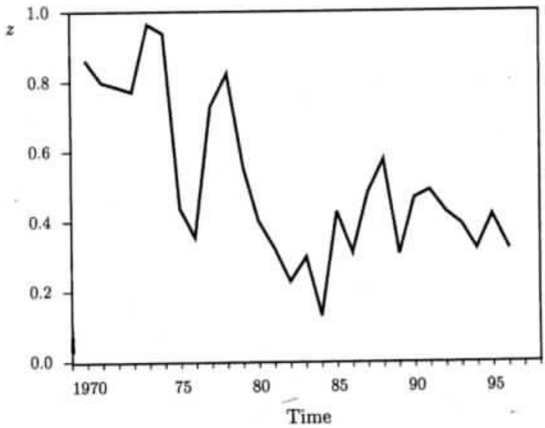

Figure 2. Time series plot of z

Figure 3, together with finite-sample 95 percent confidence intervals calculated by simulation under the assumption of iid uniformity. The unavoidably wide intervals reflect the small sample size.

The empirical c.d.f. lies within the 95 percent confidence interval. Kolmogorov's D_n-statistic has a value of 0.2275, which is less than the 5 percent critical value of 0.24993 given for this sample size by Miller (1956), although little is known about the impact of departures from randomness on the performance of this test, as noted above. In the histogram two bins lie outside their individual 95 percent confidence intervals. The chi-square goodness-of-fit statistic has a value of 10.21, which exceeds the simulated 5 percent critical value for this sample size of 9.14 (the corresponding asymptotic chi-square (4) value is 9.49), although the above caveat again applies.

Two features of the data stand out in both panels of Figure 3. First, too few realizations fall in the left tail of the forecast densities to accord with the probability forecasts, resulting in an empirical c.d.f. that lies substantially below the 45-degree line in the lower part of its range, and a significantly small leftmost histogram bin. This reflects the fact that many of the inflation surprises in the sample came in the 1970s, when inflation tended to be unexpectedly *high*; episodes of unexpectedly low inflation are rarer than the survey respondents think. Second, the middle histogram bin is significantly too high and the empirical c.d.f. lies above the 45-degree line in this range, both indicating too many realizations in the middle of the forecast densities, an already-noted phenomenon driven primarily by the events of the late 1980s and 1990s. The observations from the first half of the sample are shaded in the histogram and are seen to be more uniformly distributed, except again for the lowest

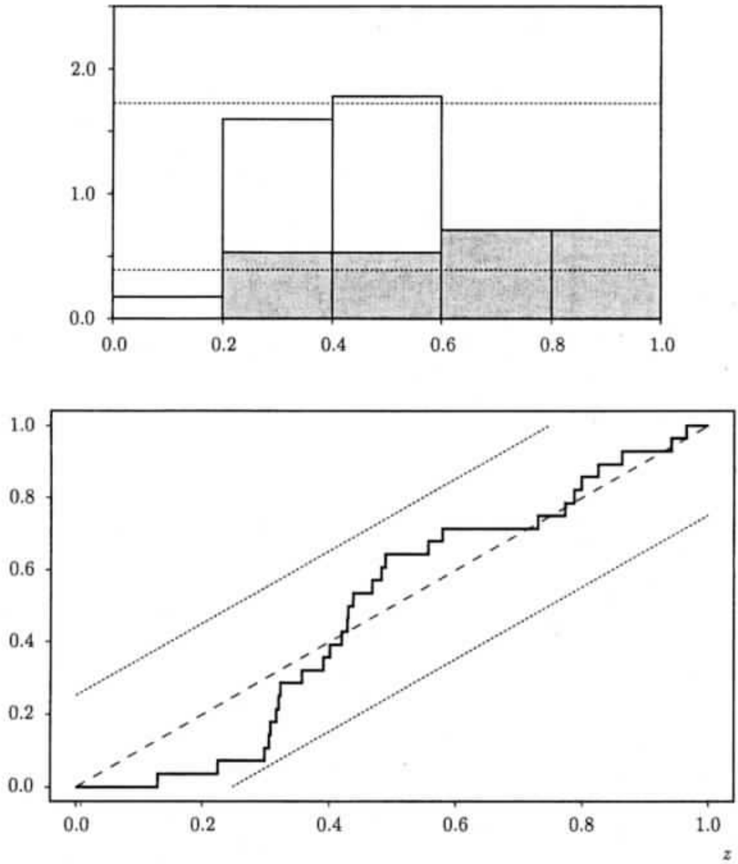

Figure 3. Histogram and empirical cumulative density function of z

Notes: In the top panel, dashed lines represent 95% confidence intervals for individual bin heights under the hypothesis that z is iid U(0, 1). The shaded region corresponds to the first fourteen z observations. In the bottom panel, we superimpose on the empirical c.d.f. a U(0, 1) c.d.f., together with 95% confidence intervals under the hypothesis that z is iid U(0, 1). See text for details.

values, illustrating once more the different characteristics of the two subperiods.

We show the correlograms of $(z-\overline{z})$ and $(z-\overline{z})^2$ in Figure 4, together with finite-sample 95 percent confidence intervals for the autocorrelations computed by simulation under the assumption that z is iid but not necessarily U(0, 1). The first correlogram clearly indicates serial correlation in z itself. The first sample autocorrelation, in particular, is large and highly statistically significant, and most of the remaining sample autocorrelations are positive and significant as well. A Ljung–Box test on the first five sample autocorrelations

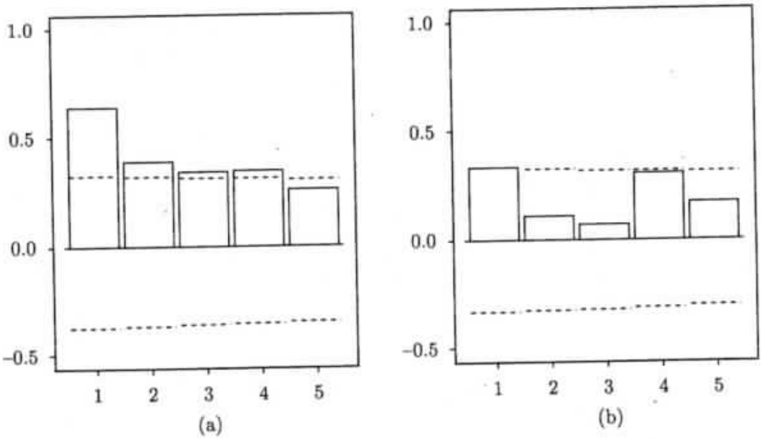

Figure 4. Sample autocorrelation functions of $(z - \bar{z})$ and $(z - \bar{z})^2$

Note: The dashed lines indicate 95% confidence intervals computed under the hypothesis that $\{z_t\}_{t=1}^{m}$ - iid. See text for details.

of $(z - \bar{z})$ rejects the white noise hypothesis at the 1 percent level, using simulated finite-sample critical values computed in the same way as for the correlogram confidence intervals.

Several explanations come to mind, one being the possibility that forecasters are more adaptive than rational, noted above. The inflation series itself is highly persistent, and the forecast densities might not be expected to change rapidly; hence forecasters might use a more-than-optimal amount of extrapolation. Forecast errors are often autocorrelated due to information lags: if a forecast for time $t + 1$ made at time t is based on an information set dated $t - 1$, then it is in effect a two-step-ahead forecast and so, even if optimal, its errors will exhibit an MA(1) correlation structure. The present forecasts are made at the beginning of the year, at which time forecasters have data on the previous year, albeit liable to revision. Because the forecast relates to the current year it is close to a genuine one-step-ahead forecast, and the impact of data revisions is unlikely to be sufficient to cause substantial autocorrelation in forecast errors. An examination of the autocorrelations of z based on preliminary inflation figures supports the later claim; a Ljung–Box test on the first five sample autocorrelations, again using simulated critical values, also rejects the iid hypothesis at the 1 percent level. In any event, the autocorrelations at higher lags in Figure 4 are not suggestive of a moving average structure. It is not clear precisely what kinds of autocorrelation in z might be expected once the density forecasts depart from optimality, but here also there is evidence of too much persistence.

It is also possible that serial correlation in z may be due to the departure or inclusion over time of forecasters who tend to be systematically optimistic or pessimistic. There is no way to check whether this is indeed the case without examining the survey returns of individual respondents, but the problem is likely to be pertinent only if the number of respondents is small. As it turns out, the number of respondents was greater than twenty in all years but four. Furthermore, Figures 2 and 3 suggest that any systematic inclusion of optimistic forecasters would have been in the early years of the sample, but that is the period when the survey enjoyed the greatest number of respondents.

It is interesting to note that although $(z - \bar{z})$ appears serially correlated, there is little evidence of serial correlation in $(z - \bar{z})^2$. Serial correlation in $(z - \bar{z})^2$ would suggest that the inflation density forecasts tend to miss heteroskedasticity in realized inflation. Hence the serial dependence in z appears to be associated with dynamics in the conditional mean of inflation neglected by the density forecasts, not with neglected dynamics in the conditional variance of inflation.

5 Conclusion

Our overall conclusion is that the density forecasts of inflation reported in the Survey of Professional Forecasters are not optimal—the probability integral transforms of the realizations with respect to the forecast densities are non-uniform and autocorrelated. Formal hypothesis tests more clearly support the autocorrelation part of this joint rejection, because here our resampling procedures produce tests that are robust to non-uniformity. The impact of this autocorrelation on the behavior of goodness-of-fit tests is not known, and our rejection of uniformity rests to a greater extent on descriptive methods. In general the density forecasts overestimate the probability that inflation will fall substantially below the point forecast, because there are too few observations in the left tail of the z density: negative inflation surprises occur less often than these forecasters expect. In the more recent data this tendency extends to both tails of the z density, and surprises of either sign occur less often than expected. In the 1990s the forecasters were more uncertain than they should have been, perhaps because they did not recognize, at least to a sufficient degree, that expectations of lower inflation are associated with lower uncertainty. This conclusion was already documented by Zarnowitz and Lambros (1987), and is endorsed here.

We have treated the mean density forecast as a collective forecast, although the sample over which the mean is taken varies in size and composition over time, and so it would be interesting to repeat the analysis for individual forecasters. One of the original aims of the survey was to keep a comprehensive record of forecasts so that forecast evaluation could be conducted on a "broader, more objective and systematic basis" (Zarnowitz, 1969), and we have clearly benefited from the archive that has been accumulated. On the other hand a little scrutiny reveals the difficulties in extending our analysis to individual forecasters, again because the survey's coverage varies, with high turnover of participants; hence only a relatively short series of forecasts is available for most individuals. The number of forecasts might be increased by adding the second, third and fourth quarter forecasts and even, in most of the recent years of the survey, also including forecasts for the following year as well as the current year. However, the pattern of the optimal evolution of density forecasts in such situations is not immediately apparent. For point forecasts, tests of the optimality of a sequence of fixed-event forecasts are based on the independence of successive forecast revisions (Clements, 1997), and the counterpart for density forecasts awaits further research. In the meantime, the evaluation methods for a conventional series of density forecasts employed in the present application are commended for wider use as such series accumulate.

References

Aït-Sahalia, Y. and Lo, A. W. (1998), "Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices," *Journal of Finance*, 53: 499–547.

Chatfield, C. (1993), "Calculating Interval Forecasts," *Journal of Business and Economic Statistics*, 11: 121–35.

Christoffersen, P. F. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39: 841–62.

Clements, M. P. (1997), "Evaluating the Rationality of Fixed-Event Forecasts," *Journal of Forecasting*, 16: 225–39.

Croushore, D. (1993), "The Survey of Professional Forecasters," *Business Review*, Federal Reserve Bank of Philadelphia, November/December: 3–15.

Diebold, F. X. and Lopez, J. A. (1996), "Forecast Evaluation and Combination," in G. S. Maddala and C. R. Rao (eds.), *Handbook of Statistics 14: Statistical Methods in Finance*, pp. 241–68. Amsterdam: North-Holland.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998), "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39: 863–83.

Friedman, M. (1977), "Nobel Lecture: Inflation and Unemployment," *Journal of Political Economy*, 85: 451–72.

J.P. Morgan (1996), "RiskMetrics: Technical Document," 4th edn., New York.

McNees, S. K. and Fine, L. K. (1996), "Forecast Uncertainty: Can It Be Measured?" paper presented at the Conference on Expectations in Economics, Federal Reserve Bank of Philadelphia, October, 1996.

Miller, L. H. (1956), "Table of Percentage Points of Kolmogorov Statistics," *Journal of the American Statistical Association*, 51: 111–21.

Söderlind, P. and Svensson, L. E. O. (1997), "New Techniques to Extract Market Expectations from Financial Instruments," *Journal of Monetary Economics*, 40: 383–429.

Stuart, A. and Ord, J. K. (1991), *Kendall's Advanced Theory of Statistics*, 5th edn., vol. 2. London: Edward Arnold.

Wallis, K. F. (1995), "Large-Scale Macroeconometric Modeling," in M. H. Pesaran and M. R. Wickens (eds.), *Handbook of Applied Econometrics*, pp. 312–55. Oxford: Blackwell.

Zarnowitz, V. (1969), "The New ASA–NBER Survey of Forecasts by Economic Statisticians," *American Statistician*, 23: 12–16.

Zarnowitz, V. and Braun, P. (1993), "Twenty-Two Years of the NBER–ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance," in J. H. Stock and M. W. Watson (eds.), *Business Cycles, Indicators, and Forecasting* (NBER Studies in Business Cycles, vol. 28), pp. 11–84. Chicago: University of Chicago Press.

Zarnowitz, V. and Lambros, L. A. (1987), "Consensus and Uncertainty in Economic Prediction," *Journal of Political Economy*, 95: 591–621.