The closer we get, the better we are?

Nathan Goldstein[*]

Ben-Zion Zilberfarb [**]

MAY 2017

PRELIMINARY – NOT TO BE QUATED WITHOUT PERMISSION

Abstract

Applying a bootstrap test procedure to a unique Israeli dataset of inflation forecasts, we show that forecasts tendency to improve across forecasting horizons, is state dependent. Our findings suggest that the accuracy of updated forecasts is significantly improved only in periods of relatively high inflation rates, while in a low inflation environment the improvement is statistically insignificant.

JEL codes: D83, E27,E37, E47

Keywords: survey forecasts, accuracy, forecasting horizon, bootstrap.

[*] Department of Economics, Bar Ilan University.

[**] Dean, School of Banking and Capital Markets , Netanya Academic College.

1. Introduction

The accuracy of survey forecasts has been extensively explored in the literature. However, most of the studies have concentrated on the relative performance of forecasters (comparing to other forecasters, the consensus forecast, model based forecasts etc.), while few studies considered the accuracy of forecasts across forecasting horizons.

As a simple implication of rational behavior, accuracy of fixed-event predictions - that is, predictions made at several periods in past which refer to the same target period - should improve over time. As the time passes, more relevant information is available and should be incorporated in the more updated forecasts, in order to reduce forecast errors. Accuracy improvement, though, may be less pronounced for longer horizons, as was documented in Fildes and Stekler (2002) and Isiklar and Lahiri (2007).

In this study we examine whether this improvement pattern is state-dependent. Our conjecture is that forecasters may not always invest in reducing their predictions errors. Specifically, significant accuracy improvements in updated forecasts may not be observed in tranquil times as in times of economic instability. To examine this, we utilize a unique Israeli data set of inflation forecasts, taking advantage of the vast changes in inflation levels and dynamics during the sample period, which allows examining accuracy patterns for very distinct inflation regimes.

In addition, most of the evidence on forecasts accuracy, like in the studies above, is presented in a descriptive-statistics form, without formal testing. Since most of the available tests can compare only a pair of forecasts series, they are less useful for multiple series, like in our case of comparing several forecasting horizons. An exception is a study by Kolb and Stekler (1990), which used rank-based tests to present significant evidence for unequal accuracy across forecasting horizons.

Considering the limitations of these tests, we further show that they provide contradicting evidence for our sample. Therefore, we suggest a more novel bootstrap approach in the same line recently used to assess relative performance of mutual funds (eg. Kosowski et al., 2006; Fama and French,2010) and forecasters (D'agostino et al., 2012).

Interestingly, according to the bootstrap results, accuracy improvement is significant only during the high inflation period of our sample, but not when inflation levels became low and more stable. This result may be in accordance with some new theories about information rigidities and inattentiveness (see review by Mankiw and Reis, 2010).

The next section describes the survey data and the inflation dynamics in Israel, during the sample period. In section 3 preliminary results using rank-based tests are presented, followed by a discussion of their drawbacks. Section 4 applies the bootstrap approach for testing the accuracy pattern of the forecasts and discusses the results. Concluding remarks are provided in section 5.

2. Data

The data set is based on a survey conducted since 1980q1 until 2009q1 among economists and business executives from industrial, commercial and financial Israeli firms[1]. On average, there are about 85 quarterly forecasters. Overall, more than 1000 people from more than 400 firms took part in the survey during the above period[2].

Participants were asked to forecast CPI inflation rates for one, two, three and four quarters ahead. Questionnaires were mailed at the 15th of the first month of the quarter, after the inflation rate for the previous month was published. The sample includes all forecasts that were mailed back until the 15th of the next month (before the next publication of monthly inflation).

The Israeli economy went through several very different stages in the survey period. Especially, it had witnessed dramatic changes in the inflation process. A closer look at the data presented in figure 1 reveals three distinct sub-periods.

I. 1980q1-1985q4.

II. 1986q1-1996q4.

III. 1997q1-2009q1.

The first sub-period is characterized by high inflation rates, averaging 30% per quarter, and reaching almost 60% at the peak. A successful stabilization program was implemented in July 1985[3]. The program included freezing of prices, exchange rate and wages, as well as cutting the budget deficit dramatically from roughly 15% of GDP to 1.3% surplus at the end of 1985. The anti-inflation program was extremely successful, bringing quarterly inflation rate down to 6.5% in 1985q4, and a further gradual decline in the second sub-period of 1986-1996. The average inflation rate per quarter during that period was

[1] The survey was arranged by Ungar and Zilberfarb from Bar-Ilan university. This data (in part) with its special features was utilized in few former studies like Ungar and Zilberfarb (1993) and Kandel and Zilberfarb (1999), which addressed other hypotheses about macroeconomic expectations.

[2] These large numbers of participants is another unique feature of our data set, relative to most popular surveys, like the SPF, Livingstone survey or the Economic Consensus.

[3] For more on this program see Bruno (1993).

3.5%. Since 1997, the Bank of Israel adopted an inflation target regime, putting the fight against inflation as its only target. As a result, inflation has declined towards the rates characterizing developed countries. The quarterly average inflation rate since 1997 until 2009 was about 0.65%.

Our analysis examines the forecasts accuracy for the three separated sub-periods, as well as for the whole sample and the 1986-2009 sub-sample, which excludes only the first sub-period of extreme inflation rates.

Table 1 presents descriptive evidence on the average forecasts performance using the RMSE measure. As previously described, there are four consecutive forecasts which refer to the inflation of each quarter in the sample period. Thus, we can compare prediction accuracy across four forecasting horizons. The RMSE values, as reported in table 1, indicate overall improvement in accuracy across forecasting horizons for the whole sample, as well as our defined sub-periods. Notice, though, that for the low inflation sub-period (1997-2009), one quarter ahead forecasts have the least RMSE, but for 2 to 4 quarters ahead predictions the RMSE statistic is virtually the same.

The improvement pattern in the inflation forecasts of the current survey is in line with the patterns observed in other surveys of forecasts (see for example in the review by Fildes and Stekler, 2002) .The next sections provide tests to examine whether the improvement is statistically significant[4].

[4] An alternative approach is a null hypothesis stating that forecasts accuracy does not worsen as we get closer. Patton and Timmerman (2012) have recently proposed a testing procedure for a null of “weak rationality”, relying on this approach. However, this approach is less useful in our study, in light of the observed forecast improvement, as presented in table 1.

3. Rank based tests

Kolb and Stekler (1990) examined the accuracy improvement of USA GNP forecasts across forecasting horizons, using two rank-based tests: Friedman's test (1937) and Kruskal-Wallis test (1952). We apply these two tests to our data set as well, before introducing the bootstrap method.

For the Friedman's test, we rank forecasting horizons in each quarter according to SE (squared errors) performance of the average forecast and calculate the sum of the ranks over all sample quarters to obtain four ranking measures denoted by R_h, where $h = 1, \dots, 4$ indicates the horizon. Then, we compute the test statistic as[5]:

$$F = \frac{12}{TH(H+1)} \sum_{h=1}^{H} \left(R_h - \frac{T(H+1)}{2} \right)^2 \tag{1}$$

where T is the sample total number of quarters and $H = 4$ is the number of forecasting horizons.

For the Kruskal-Wallis test, which extends the Wilcoxin Rank Sum Test, we pool all quarters together and then rank every prediction according to SE criterion, and sum the pooled ranks for each forecasting horizon to obtain four measures denoted by RS_h. The test statistic is:

$$KS = \frac{12}{T^2 H(TH+1)} \sum_{h=1}^{H} \left(RS_h - \frac{T(TH+1)}{2} \right)^2 \tag{2}$$

Both statistics are χ^2_{H-1} asymptotically distributed under the null of equal accuracy.

The results of the two tests, as appear in table 2, are contradicting: While the Friedman's test rejects the equality of accuracy across horizons for every sub-period and

[5] This is a corrected version of the formula used in Kolb and Stekler (1990). See Batchelor (1990).

for the whole period at the 1% significance level, the Kruskal-Wallis test rejects the null only for the whole sample at the 5% significance level, but not for any of the sub-periods. This may reflect the main limitation of the rank-based tests which ignore errors magnitude by attaching the ranks. In particular, while the Friedman's test completely ignores prediction error size by assigning new ranks for every quarter, the Kruskal-Wallis test, which pools the forecasts of all quarters together and then rank them, does give some weight to the relative sizes of the errors across the sample period. As a consequence, although the tendency of accuracy improvement as observed in the data is found to be significant by the Friedman's test, it is possible that this tendency is more frequent in quarters with larger forecast errors, which are given overall low ranks by Kruskal-Wallis rank sum test, and therefore does not turn the test statistic of this test to be significant. This explanation may also highlight a new issue of state-dependency. Specifically, the improving pattern of accuracy of updated forecast may be less pronounced in times of relative economic stability, which are more predictable.

Another disadvantage of the rank-based tests is that they test for overall equal accuracy across the four forecasting horizons, but a rejection does not suggest the particular pattern of improving forecasts that is observed in data, actually holds significantly. In order to overcome the limitations of the rank-based test and to shed more light on the possibility of state-dependency, we suggest the bootstrap method in the next section.

4. The bootstrap test

We build four bootstrap distributions under the null of equal accuracy, as follows: In each draw (1000 in total), we resample forecasts for each quarter across the four forecasting horizons. That is, for every quarter in the sample we have four (average) predictions which refer to it and we resample among them, doing the same for each quarter. As a consequence, differences in accuracy across forecasting horizons in the resampled data are now random, which reflects the null hypothesis.

Next, we calculate four accuracy measures for the four forecasting horizons, order them and assign them to four bootstrap distributions, reflecting four levels of accuracy performance. Thus, the procedure recognizes the possibility of observed differences in accuracy occurring by pure chance. We then compare the four accuracy measures from the real data with their four corresponding bootstrap distributions. The distribution of the best accuracy measures will be assigned to the best accuracy measure among forecasting horizons from the real data. The distribution of the second best accuracy measures will be assigned to the second best accuracy measure among forecasting horizons and so on. Following D'agostino et al. (2012), our accuracy measure is a sum over all quarters of $SE_h/(\sum_h SE_h/H)$, which normalizes the large differences in forecast errors between the sub-periods, as observed in table 1.

The results, presented in table 3 and figure 2, imply a compromise between the previous contradicting results of the rank-based tests: For the high and moderate inflation sub-periods (1980-1986 and 1986-1996, respectively), the most updated forecast is significantly "good", while the earliest forecast is significantly "bad"[6]. The results for the whole sample may also be dominated by these sub-periods. However, for the low inflation

[6] That is, the accuracy measure falls in the 5% left ("good") or right ("bad") tail area of the corresponding bootstrap distribution, .

period the null of equal accuracy is not rejected by the bootstrap[7]. This result is well-illustrated in figure 2, which shows how the observed accuracy measures of 1997-2009 resemble quite close the middle of the bootstrap distributions.

Further, difference between high and low inflation periods was observed, despite the use of an accuracy measure that normalizes differences in RMSE between the periods[8]. Thus, our findings indicate that the common observation in the literature of accuracy improvement across horizons should be handled carefully by proper testing that accounts for the possibility of state dependency.

5. Conclusion

We apply a bootstrap approach to a unique dataset of inflation forecasts, during a period of very distinct inflation regimes, and show that forecasts performance does not always improve across forecasting horizons. In particular, our findings suggest that forecasters do not improve their forecasts significantly in relatively tranquil times with low inflation levels.

A possible explanation is that forecasters do not act as a simple rationality model would imply, because they face information constraints, so they need to allocate resources in order to keep themselves updated. This activity may not be worthwhile in times of

[7] Note that the four quarters ahead forecast performs even slightly better than the one quarter ahead forecast, according to the normalized accuracy measure. However, the performance of both forecasts is not significantly “good”, as the test results point out.

[8] Applying the bootstrap procedure with RMSE as an accuracy measure, yields quite similar results. For the low inflation period, the one quarter-ahead forecast performs significantly “good”, though, but the forecasts for longer horizons are not significantly “bad”, while for the high and moderate inflation periods the accuracy improvement is significant. Hence, differences between inflation regimes are still pronounced.

relative economic stability, as in the period of low inflation in our survey, and consequently improvements in prediction accuracy are not significant.

Recent macroeconomic theories rely on information rigidities models, to explain expectation formation process and macroeconomic dynamics. Two well-known models, for example, are the sticky information model of Mankiw and Reis (2002) and the noisy information model of Sims (2003) and Woodford (2003)[9].Consequently, agents are not fully updated, and the updating process may take a state-dependent form, as modeled by few studies (Gorodnichenko,2008; Branch et al.,2009; Woodford, 2009). Our evidence may support this new line of research.

Literature Cited

Batchelor, Roy A. (1990). "All Forecasters Are Equal." *Journal of Business and Economic Statistics*, 8 (1), 143–44.

Branch, William A., John Carlson, George Evans and Bruce McGough (2009). "Monetary Policy, Endogenous Inattention, and the Volatility Trade-Off." *Economic Journal*, 119, 123-157.

Bruno, Michael (1993). *Crisis, Stabilization and Economic Reform: Therapy by Consensus*. Oxford: Clarendon press.

D'Agostino, Antonello, Kieran McQuinn and Karl Whelan (2012). "Are Some Forecasters Really Better Than Others?" *Journal Money Credit and Banking*, 44 (4), 715-732.

Fama, Eugene F. and Kenneth French (2010). "Luck versus Skill in the Cross Section of Mutual Fund Returns." *Journal of Finance*, 65 (5), 1915-1947.

Fildes, Robert and Herman O. Stekler (2002). "The State of Macroeconomic Forecasting." *Journal of Macroeconomics*, 24 (4), 435-468.

[9] see review by Mankiw and Reis (2010).

Friedman, Milton (1937). "The Use of Ranks to Avoid The Assumption of Normality Implicit in The Analysis of Variance." *Journal of American Statistical Association*, 32 (200), 675-701.

Gorodnichenko, Yuriy (2008). "Endogenous Information, Menu Costs and Inflation Persistence." NBER Working Paper 14184.

Isiklar, Gultekin and Kajal Lahiri (2007). "How Far Ahead Can We Forecast? Evidence From Cross-Country Surveys." *International Journal of Forecasting*, 23 (2), 167-187.

Kandel, Eugene and Ben-Zion Zilberfarb (1999). "Differential Interpretation of Information in Inflation Forecasts." *Review of Economics and Statistics*, 81 (2), 217–226.

Kolb, R. A. and Herman O. Stekler (1990). "The Lead and Accuracy of Macroeconomic Forecasts." *Journal of Macroeconomics*, 12 (1), 111-123.

Kosowski, Robert, Allan Timmerman, Russ Wermers, and Hall White. (2006) "Can Mutual Fund 'Stars' Really Pick Stocks? New Evidence from a Bootstrap Analysis." *Journal of Finance*, 61 (6), 2551–95.

Kruskal, William H. and W. Allen Wallis (1952). "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association*, 47 (260), 583-621.

Mankiw, N. Gregory and Ricardo Reis (2002). "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics*, 117 (4), 1295-1328.

Mankiw, N. Gregory and Ricardo Reis (2010). "Imperfect Information and Aggregate Supply." In: *Handbook of Monetary Economics,* edited by B. Friedman and M. Woodford, Elsevier-North Holland, vol. 3A, chapter 5, 183-230.

Patton, Andrew. J. and Allan Timmermann (2011). "Forecast Rationality Tests Based on Multi-Horizon Bounds," *Journal of Business and Economic Statistics*, 30 (1), 1-17.

Sims, Christopher A. (2003). "Implications of Rational Inattention." *Journal of Monetary Economics*, 50 (3), 665-690.

Ungar, Meyer and Ben-Zion Zilberfarb (1993). "Inflation and Its Unpredictability—Theory and Empirical Evidence." *Journal of Money, Credit, and Banking*, 25 (4), 709–720.

Woodford, Michael (2003). "Imperfect Common Knowledge and the Effects of Monetary Policy." In: *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, edited by Philippe Aghion, Romain Frydman, Joseph Stiglitz and Michael Woodford, Princeton University Press.

Woodford, Michael (2009). “Information-Constrained State-Dependent Pricing.” *Journal of Monetary Economics*, 56 (S), S100–S124.

TABLE 1
Root mean square error (RMSE) of inflation forecasts

Forecast Horizon	1980-2009	1986-2009	1980-1985	1986-1996	1997-2009
1 quarter	2.819	1.299	5.972	1.590	0.964
2 quarters	5.485	1.806	12.202	2.238	1.302
3 quarters	6.873	2.385	15.206	3.192	1.283
4 quarters	8.254	3.822	17.469	5.387	1.289

TABLE 2
Rank-based tests for equal accuracy

	F	KS
	Friedman’s test	Kruskal-Wallis test
1980-2009	30.705**	7.866*
1986-2009	19.877**	6.787
1980-1985	13.971**	6.672
1986-1996	12.136**	5.514
1997-2009	11.302**	2.281

Notes: The table reports test statistics by the Friedman’s and Kruskal-Wallis tests, which are χ_3^2 asymptotically distributed, under the null of equal accuracy across forecasting horizons.
* Denotes rejection of the null at the 5% significance level.
** Denotes rejection of the null at the 1% significance level.

TABLE 3
Bootstrap tests for equal accuracy

Forecast Horizon	1980-2009	1986-2009	1980-1985	1986-1996	1997-2009
1 quarter	**0.757**	**0.829**	**0.436**	**0.666**	0.976
	(0.000)	(0.008)	(0.000)	(0.001)	(0.537)
2 quarters	0.959	0.960	0.951	**0.849**	1.060
	(0.183)	(0.246)	(0.545)	(0.012)	(0.273)
3 quarters	1.008	0.991	1.087	0.970	1.009
	(0.313)	(0.110)	(0.718)	(0.057)	(0.323)
4 quarters	**1.276**	**1.220**	**1.525**	**1.516**	0.954
	(1.000)	(0.995)	(0.987)	(1.000)	(0.850)

Notes: The table reports normalized accuracy measures of forecasts series across the four forecasts horizons in the sample and the corresponding percentiles in the associated bootstrap distributions in parentheses. Bootstrap percentiles are obtained under the null of equal accuracy across forecasts horizons and attached to the respective accuracy measure, according to its relative size.

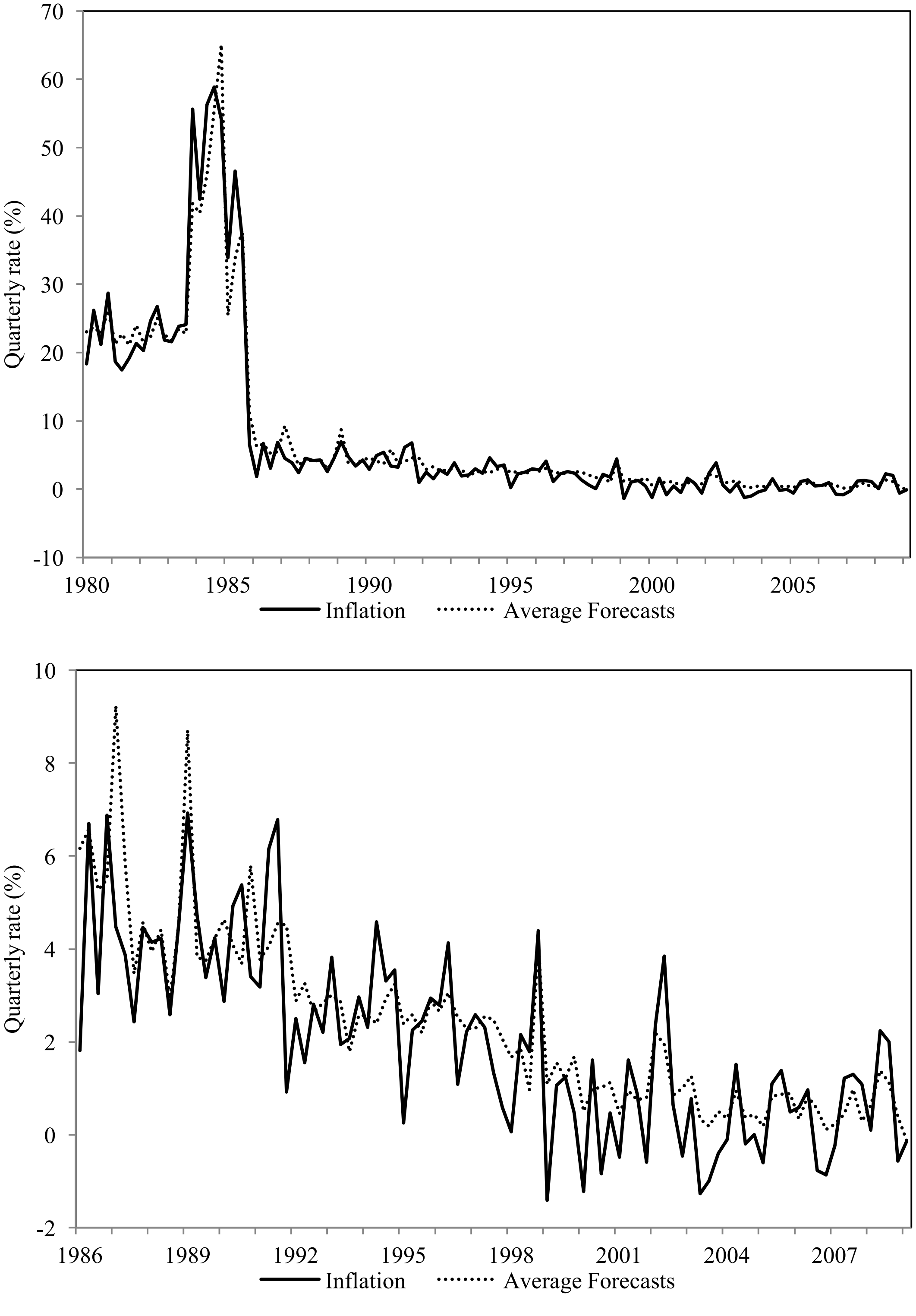

Fig. 1. Quarterly Inflation Rates and One Quarter Ahead Forecasts

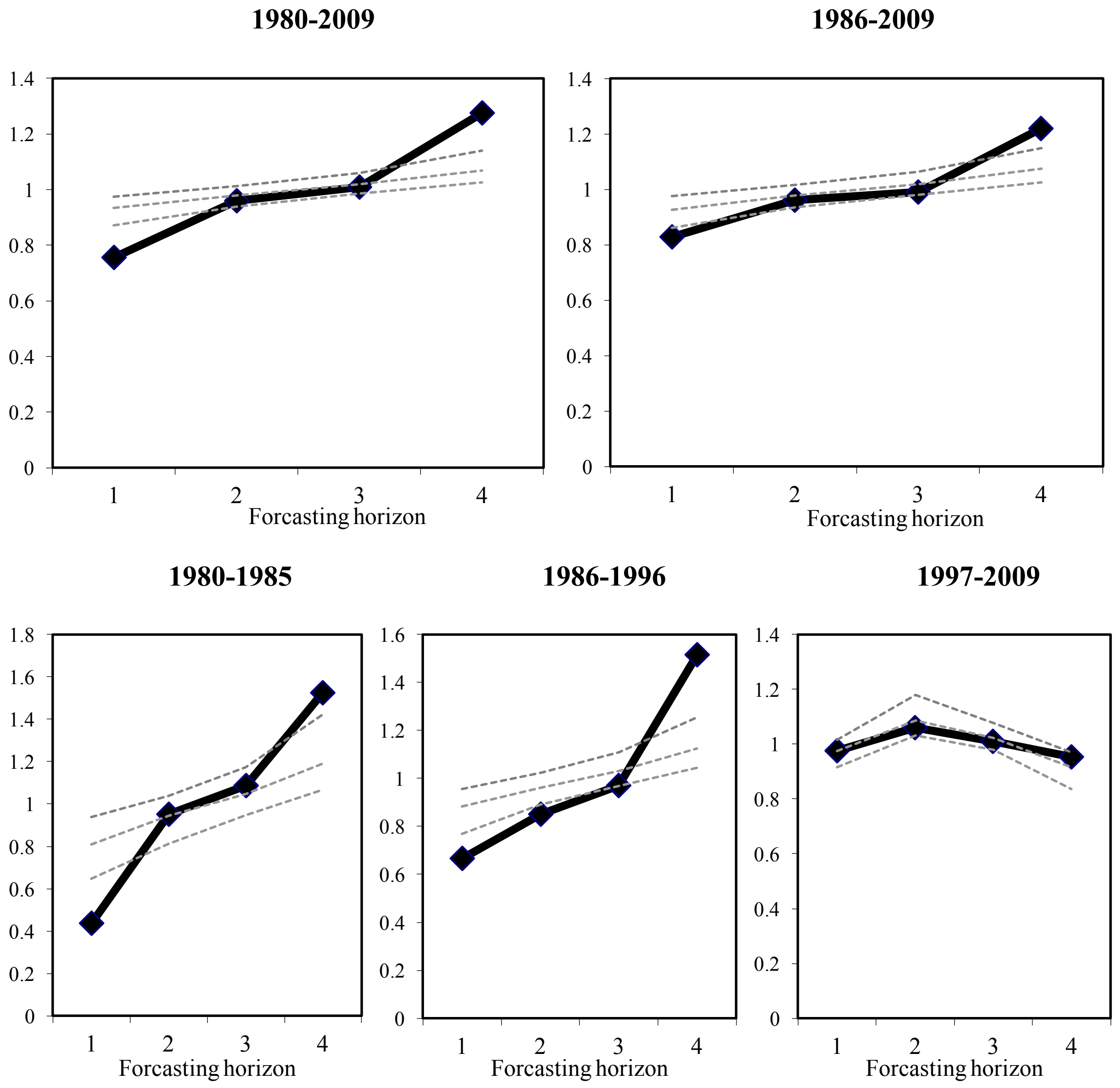

Fig. 2. Accuracy Measures and Bootstrap Distributions

Notes: The figure displays accuracy and bootstrap distributions for various sample periods, according to the results presented in table 3.The solid line denotes accuracy measures, and the dotted lines denotes the 5%, 50% and 95% percentiles in the corresponding bootstrap distributions.