THE PREDICTIVE POWER OF THE SPF PROBABILITY FORECASTS

Jiazhuo G. Wang
College of Staten Island
City University of New York

ABSTRACT

The probability forecast of decline in real GDP from the Survey of Professional Forecasters (SPF) has long been used by the various users in the public and private sectors as the predictor of the cyclical movement of the economy. However, its predictive power and forecasting performance have not yet been properly evaluated in literature. In this paper, the PT Predictive Power Test and the Kuipers Score are applied to assess the usefulness of the SPF's probability of decline in real GDP as the indicator of the future path of the economy.

INTRODUCTION

Cyclical movement in real GDP has long been the focus of business cycle researchers and business practitioners. As witnessed for the past decades, cyclical movement of real GDP tremendously impacted the economy, and early detection of the phase change could provide enormous values for corporations, individuals and government policy makers. For this purpose, professional forecasts from a variety of forecasting entities using different techniques were conducted with the intention to supply government and business decision makers, as well as general public, with some reliable and timely guidance for the future path of the economy. Among them, the probability forecast of decline in real GDP from the Survey of Professional Forecasters (SPF) is a one that was being constantly monitored and frequently used by various end users in both public and private sectors.

However, given the critical role of the prediction of the future movement of real GDP in influencing business decisions, one important issue in business cycle research is the quality of the forecasts. While high quality forecast can provide its end users with a useful "leading indicator" for their business references, poor performed forecast could be a "misleading indicator" for its users in terms of direction, timing, and magnitude of the future changes in the economy. Therefore, any professional forecasts, including the probability forecasts of decline in real GDP, without associated evaluations should be considered a mission incomplete, and used with extra caution.

The purpose of this paper is intended to evaluate the SPF by assessing the forecasting performance of the probability of decline in real GDP with different forecasting horizons, which have been widely used by users in public and private sectors, but have not yet been analyzed and evaluated in literature with the appropriate evaluation methodologies. It is hoped that the evaluation for the forecasting performance of the probability of decline in GDP in this paper will provide its current and potential users with a needed assessment for its usefulness as a predictor of the cyclical movement of real GDP.

The structure of the paper is as follows. Section II presents the descriptions and empirical data of the SPF probability forecasts. Section III assesses the predictive power of the SPF forecasts. Section IV analyzes the forecasts in terms of balance of the missing signals and the false alarms. Finally, Section V ends the article with some concluding remarks.

THE SPF PROBABILITY FORECAST ON DECLINE IN REAL GDP

As one of the oldest business surveys in the US, since 1968, the American Statistical Association and the National Bureau of Economic Research (ASA/NBER) routinely conducted the quarterly surveys by mailing some questionnaires to professional forecasters and collecting their forecasts for the future economy. The questionnaires are mailed out when the forecasters typically review and update their predictions, and the responses are received by the middle of the second month of the quarter. The number of the responses to the questionnaires usually arrange from about 20 to 150. The survey was commonly referred to as the ASA-NBER Survey in previous literature, and the name was changed to the Survey of Professional Forecasters (SPF) when the Federal Reserve Bank of Philadelphia took over the responsibility for the survey in June 1990.

The SPF generally covers the current quarter, the subsequent four quarters, as well as the current year and next year. The variables to be predicted include GDP related measures, the unemployment rate, the probability of a decline in GDP, and other important macro-economic variables that are closely watched by government and business decision makers as well as the general public.

The graphs of the mean probability of decline in real GDP in the current and the following quarters from Quarter 4 1968 to Quarter 2 2004 are depicted in the Chart 1-5. The lines in each chart display the probability of decline in real GDP in different quarters as the professional forecasters made the predictions over time, and the real time real GDP growth rate, respectively.

The real time real GDP growth rate could be calculated in two different ways. First, calculating the real time real GDP growth rate uses each quarterly release of the real GDP at real time since Quarter 4 of 1968 when the SPF was first time being conducted. At the beginning of each quarter, the Bureau of Economic Analysis of US Department of Commerce issues the preliminary real GDP data for the previous quarter and revises the data for all other previous quarters, if needed. The changes in the revision for the previous quarterly data could be caused by incompletion of the previous data, statistic error, structure changes, measure changes (from GNP to GDP, for example), and any other issues that may call for the revisions. As the consequence, different quarterly releases for the same quarter real GDP could be quite different. It is interesting to note that the real time real GDP growth calculated in such way only shows one quarter negative growth for the most recent recession in 2001.

Second, calculating real time real GDP growth rate uses the July release of each year. It is well known that the July version of the real GDP release is a relatively complete one in each year. It may not be as "real time" as the one calculated using the first approach, but it caught all major consecutively negative growths of the real GDP, which is certainly a better reflection of the cyclical movement of real GDP in reality. But in either case, the real time real GDP and revised real GDP could differ significantly either in terms of magnitude or even in the directions of the changes.

Considering that SPF forecasts were based on the real GDP data available to the SPF respondents at real time, the real time real GDP is used as the target to evaluate the performance of SPF forecasts, and make the comparison on an apple to apple basis. More specifically, we use the real time real GDP growth rate calculated by the July version of each year. As mentioned above, it could be different from the most recently revised data, such as Q3 2004 release for the data up to Q2 2004, but it is probably much closer to what was available to SPF forecasters when they made their predictions in real time.

From the charts 1-5, several notable phenomenon can be observed. First, the mean probabilities generated by the professional forecasters fluctuate over time in a certain pattern. The value of mean probability varies from as high as in 80% range to as low as less than 5%. Second, the fluctuation of ups and downs in mean probability seems coincident with the fluctuations in real GDP growth. That is, around the time with the negative growth rate of real GDP or the recessionary periods, the probabilities suddenly rise up; and in the time associated with the positive growth rate or the expansionary periods, they remain relatively low. Third, for different forecasting horizons, sudden increases or decreases in the probability either precede or follow the cyclical movement of real GDP with different time leads or lags. Finally, the high end of the mean probability tends to decrease as the forecasting horizon increases. As shown in the charts, the high end probability decreases from an 80% range for current quarter to a 70% range for the one quarter ahead, to a 50% range for two quarters ahead, and to a 30% range for three and four quarters ahead.

All these observations indicate that the probability forecasts for the decline in real GDP contain tremendous information about the phase changes of the real GDP and business cycles. Consequently, some proper evaluations need to be conducted for these series for their forecasting ability as the predictors.

PREDICTIVE POWER OF SPF FORECASTS

The predictive power of the SPF probability forecasts is first being examined in this section. Given the binary nature of the event, the Pesaran and Timmermann (PT) test (Pesaran and Timmermann, 1992) is used to assess the predictive power of the SPF probability forecasts in their abilities to predict the future path of the real GDP.

The PT test was designed to testing the prediction of the directional changes such as an occurrence or a non-occurrence of an event (for example, the decline in real GDP or non-decline in real GDP for the current quarter or for the future quarters). However,

the SPF forecasts are the probabilities that real GDP will decline in different forecasting horizons. They are not directly event variables that take value of 1 or 0. Therefore, applying the PT test needs to first translate the SPF probability into an event variable.

One possible way of translation is the traditional naïve approach. That is, a value 1 will be assigned if the forecasting probability for the occurrence of the event is above 50%; a value 0 will be assigned, otherwise. Another possible way of translation is to select the probability that is associated with the critical value that makes difference in the PT test results. For example, given the confidence level of 90% for normal distribution, the SPF probability that is associated with the critical value of 1.645 is 5%; then, use the 5% as the threshold to translate the probability into an event variable. In this way, not only can we test the prediction performance, but also we are able to know how low (or high) the threshold needs to be set up to identify the predictive power of forecasts on the occurrence of the event.

For the PT test, the forecast evaluation is conducted by calculating the difference between the portion of the times that the event is predicted correctly and the mean of the underlying binomial distribution (theoretical portion of the times of the occurrence of the event) under the null hypothesis of independence between the forecast and the occurrence of the event for the 2x2 case. The test statistic is as follows:

$$S_n = (\hat{P} - \hat{P}_*) / (Var(\hat{P}) - Var(\hat{P}_*))^{1/2} \quad (1)$$

where $\hat{P}$ is the portion of the times that the event is predicted correctly,

$\hat{P}_* = P_y P_x + (1 - P_y)(1 - P_x)$, the mean of the binomial distribution under the null hypothesis,

P_y and P_x are the probability of the occurrence of the event and the forecasted probability of the occurrence of the event, respectively. When the theoretic P_y and P_x are unknown, P_y and P_x can be efficiently estimated by

$$\hat{P}_y = \sum_{t=1}^{N} Y_t / N$$

$$\hat{P}_x = \sum_{t-1}^{N} X_t / N$$

respectively, under the null; where Y_t is the occurrence of the event, and X_t is the prediction of the occurrence of the event.

$Var(\hat{P}) = \hat{P}(1 - \hat{P}_*) / N$, is the variance of $\hat{P}$,

$$Var(\hat{P}_*) = (2\hat{P}_y - 1)^2 \hat{P}_x (1 - \hat{P}_x) / N + (2\hat{P}_x - 1)^2 \hat{P}_y (1 - \hat{P}_y) / N + 4\hat{P}_y \hat{P}_x (1 - \hat{P}_y)(1 - \hat{P}_x) / N^2$$

is the variance of $\hat{P}_*$.

Under the null hypothesis of independence between the SPF probability forecasts and the occurrence of the event, the difference between the percentage of the correct forecasts made by SPF ($\hat{P}$) and the percentage of the occurrence of the event ($\hat{P}_*$) should be insignificant as measured by the test statistic S_n in (1) which follows $N(0,1)$ under the null. Reversely, if test fails at any acceptable level of significance, a dependent relationship between the two series will be considered. Then, the existence of the predictive power of the forecast on the occurrence of the event will be supported by the test.

Given the directional nature of the PT test and the density function nature of the SPF forecast, as discussed above, we use two thresholds as the directional indicator for the SPF forecast series: (a) using the naïve approach with 50% as the threshold. The SPF probability of decline in real GDP above 50% is considered a direction change (down), otherwise, it is considered a non-down prediction; (b) using the probability that is associated with the critical value of the normal distribution with the confidence level of 90% as the thresholds. In either case, the PT test was performed using real time real GDP with all forecasting horizons. The test results are displayed in Table 1.

As it turns out, for the naïve approach, all test results for current and next quarter forecast (2 quarter and above forecasts are not applicable, because no forecast probability is above 50%) uniformly reject the independence hypothesis with any commonly used acceptable level, indicating strongly the

existence of the predictable relationship between the SPF forecasts and the actual decline in real GDP. In other words, the SPF forecast is not a groundless predictor for the occurrence of the decline in real GDP in the current and the following quarters. Instead, the SPF contains useful information about the target being predicted, and, thus, is an important source for watching the phase changes of the real GDP cyclical movement.

Similar results were revealed for the critical value approach. Given the way of selecting the threshold, the test statistic with the probability above the critical value is expected to reject the null. The interesting result is that the threshold probability, the one that is needed to reject the null hypothesis, is very low, between 1.5% and 8.5%, depending upon the forecasting horizons. The result, once again but in a different way, strongly supports the existence of the predictive power of the SPF forecast on the occurrence of the decline in real GDP.

BALANCING MISSING SIGNALS AND FALSE ALARMS

The second measure of evaluating the SPF probability forecasts is the balance of the missing signals and the false alarms. As in any probability forecast area, the trade-off between the missing signals and the false alarms always exists. In general, a decision rule based on the forecast probability that tends to decrease the missing signals or increase the "hit rate" will tend to increase the false alarms. So the balance of these "type I" and "type II" error is an important measure for the evaluation of the forecasts. In this regard, the SPF forecasts can be assessed using Kuipers Performance Index or Kuipers Score (Granger and Paseran, 2000) with the contingency matrix calculated using the event variable translated from the SPF with the naïve and the threshold methodologies as discussed above. Kuipers Score (KS) was originally proposed by Pierce (1884), and used widely in evaluating the forecasting performance in metrology. KS is defined as the difference between the "hit rate" (H) and the "false alarm rate" (F) as follows:

$$KS = H - F \qquad (2)$$

Where $H = T_{by} / (T_{by} + T_{bn})$

$F = T_{gy} / (T_{gy} + T_{gn})$

T_{by} is the number of times that the event occurred (the subscript "*b*" stands for bad thing (event) happened) and the forecaster predicted it correctly ("*y*" for "yes" answer given by forecaster). T_{bn} is the number of times that the event occurred but the forecaster failed to predict it (with "no" answer). So the ratio *H,* then, is the "hit ratio" that measures the portion of the times that the forecasters predict correctly when the event indeed occurred.

Similarly, T_{gy} is the number of times that the event didn't occur, but the forecasters mistakenly predicted it. T_{gn} is the number of times that the event didn't occur and the forecasters correctly said no. So the ratio *F* is the "false alarm ratio" that measures the portion of times that the forecasters generated false signals for the occurrence of the event when it actually didn't happen. By definition of the *T's,* the total number of the observations (*T)* is the sum of all *T's*. That is:

$$T = T_{by} + T_{bn} + T_{gy} + T_{gn}$$

Naturally, higher the value of KS with a positive score, better the forecasting performance, with the balance of the missed events and the false alarms. Higher the KS, higher the hit rate with the relatively low false alarm rate. Reversely, the lower KS, or even the negative KS, indicates the higher false alarm rate relative to the hit rate. Obviously, if a forecaster always predicts the occurrence of the event systematically, then KS will be equal to 0 with 100% hit rate and 100% false alarm rate. Similarly, if a forecaster always predicts the non-occurrence of the event systematically, then, the KS will also be equal to 0, but with 0% false alarm rate and 0% hit rate as well.

The Contingency Matrix and the KS for both the naïve approach and the threshold approach with different forecasting horizons are displayed in Table 2. The numbers in the first, second, third and fourth quadrants correspond to $T_{by}, T_{gy}, T_{gn}, T_{bn}$, for each forecasting horizon, respectively.

As the Table 2 shows, compared with two approaches, the threshold approach, with its low threshold (<10%), generated high hit rate (100%), but with high false alarm rate (>90%) as well. In contrast, the naïve approach, with its relatively higher dividing line (50%), generated low false alarm rate (< 5%),

but also relatively low hit rate (60% and < 60%, compared to above 90%). It should be noted that, in the threshold case, the KS appears extremely low, the KS ranges only from 0.024 to 0.065. In contrast, the KS for the naïve approach is much higher, especially for the current quarter and the next quarter, with the KS equal to 0.6012 and 0.2593, respectively. The results indicate that the SPF forecast using the naïve approach can generate much balanced predictions with the considerations of both the hit rate and the false alarm rate. On the other hand, the SPF using the threshold approach can only help determine the critical values that help identify the predictive power of the SPF, but can't be used as the actual threshold to translate the probability into an event variable. Given the low threshold, it systematically generates high hit rates, but meanwhile, it systematically generates high false alarm rates as well. As the results, it produces a low Kuipers Score.

CONCLUDING REMARKS

In this paper, I used the PT predictive power test and the Kuipers Score to evaluate the forecast performance of the SPF probability forecasts for the decline in real GDP with the different forecasting horizons for their usefulness as the guidance of the future path of the economy. In summary, I would like to conclude the paper as follows:

First, the SPF probability forecasts for the decline in real GDP in the current quarter and near future contain tremendous amounts of information about the regime switching of the cyclical movement in real GDP. They are indubitably the important sources for exploring and identifying the possible signals of the forthcoming of the cyclical phase changes.

Second, the PT predictive power test reveals strong evidence for the existence of the dependent relationship between the forecasts and the event being forecasted in various forecasting horizons, as indicated by the rejection of the null hypothesis of independence. That means: the SPF probability forecasts are the predictions for the future real GDP movement with the needed predictive power.

Third, similar to the issue of balancing "Type I" and "Type II" error, a high hit rate needs to be achieved with the balancing of a low false alarm rate. Applying the Kuipers Score to the SPF probability forecasts with different direction translation approaches shows much balanced performance using the naïve way of translation.

Finally, it should be mentioned that the SPF probability forecasts for the decline in real GDP, especially for the longer forecasting horizons, seem conservative, as observed by the relatively low probabilities assigned by the forecasters to the possibility of decline in real GDP, even when the real GDP decline was almost on the corner or already occurred. Therefore, the SPF probability for the decline in real GDP does contain correct direction information for the real GDP, but need to be used with some "corrections" or amplifications.

REFERENCES

Diebold, F. X., J. Hahn and A. S. Tay. Real-Time Multivariate Density Forecast Evaluation and Calibration: Monitoring the Risk of High-Frequency Returns on Foreign Exchange. www.ssc.upenn.edu/~diebold.

Diebold, F. X., T. A. Gunther and A. S. Tay. 1998. Evaluating Density Forecasting with Application to Financial Risk Management. International Economic Review, Vol. 39, No. 4, November: 863-883.

Diebold, F. and R. Mariano. 1995. Comparing Forecast Accuracy. Journal of Business Economics & Statistics, 13: 253-265.

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

Diebold, F. X., A. S. Tay, and K. F. Wallis. 1998. Evaluating Density Forecasting of Inflation: The Survey of Professional Forecasts. Working Paper, September 11: www.ssc.upenn.edu/~diebold.

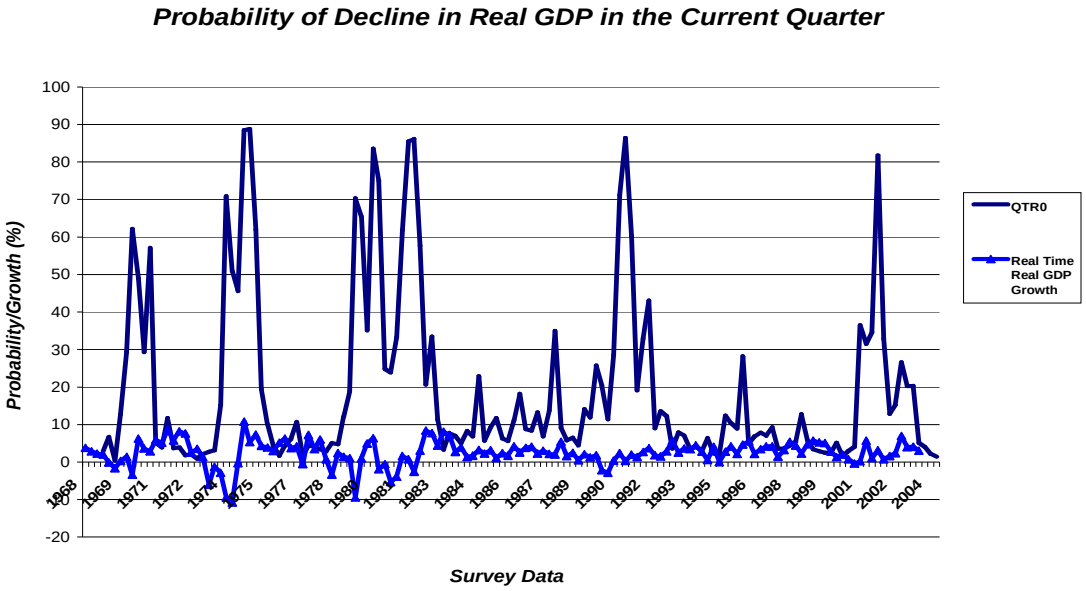
Granger, C. W. J. and M. H. Pesaran. 2000. Economic and Statistical Measures of Forecast Accuracy. Journal of Forecasting, Vol. 19: 537-560.

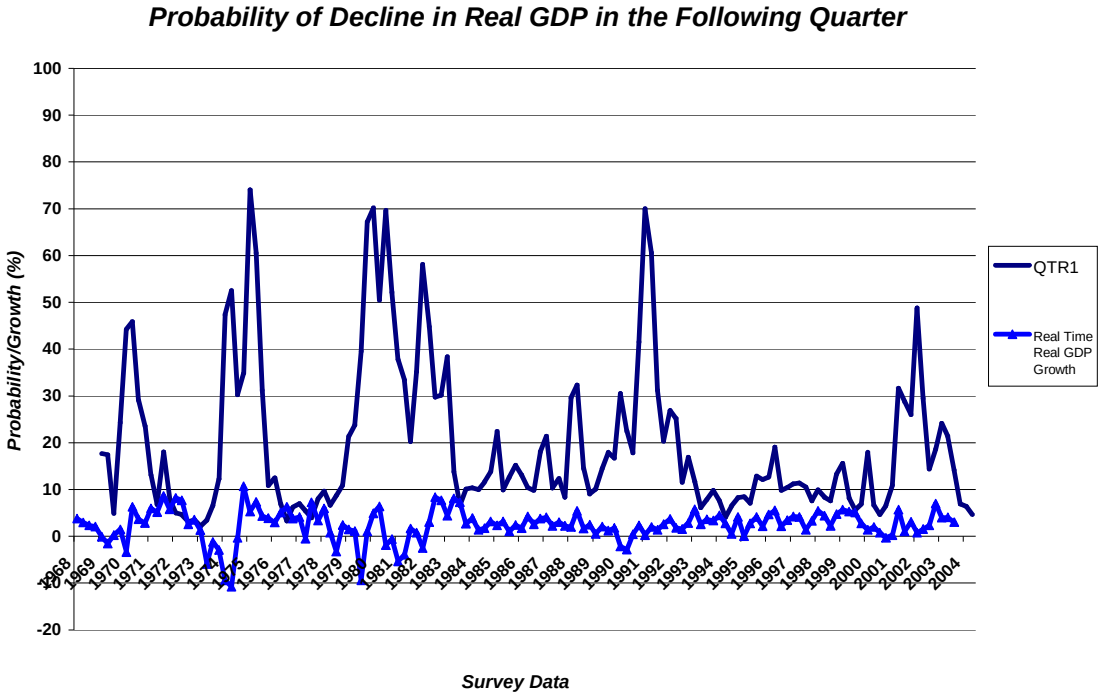
Lahiri, K. and J. G. Wang. 1994. Predicting Cyclical Turning Points with Leading Index in a Markov Switching Model. Journal of Forecasting, Vol. 13, Number 3, May: 245-264.

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, 1991.

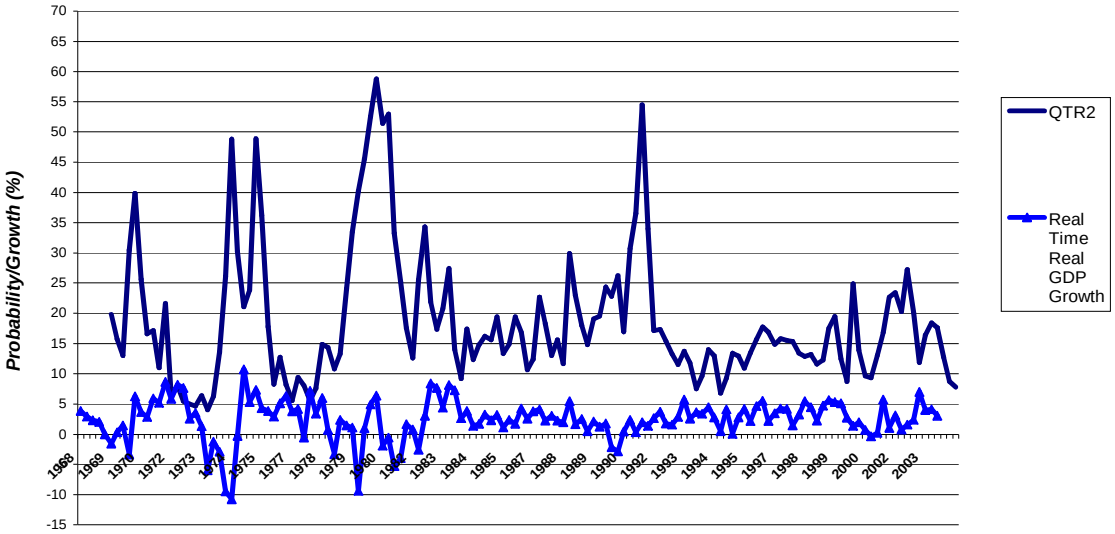
Peirce, C.S. 1884. The Numerical Measure of the Success of Prediction. Science 4: 453-454.

Pesaran, M. H. and A. Timmerman. 1992. A Simple Nonparametric Test of Predictive Performance. Journal of Business & Economic Statistics, October, Vol. 10, No. 4: 461-465.

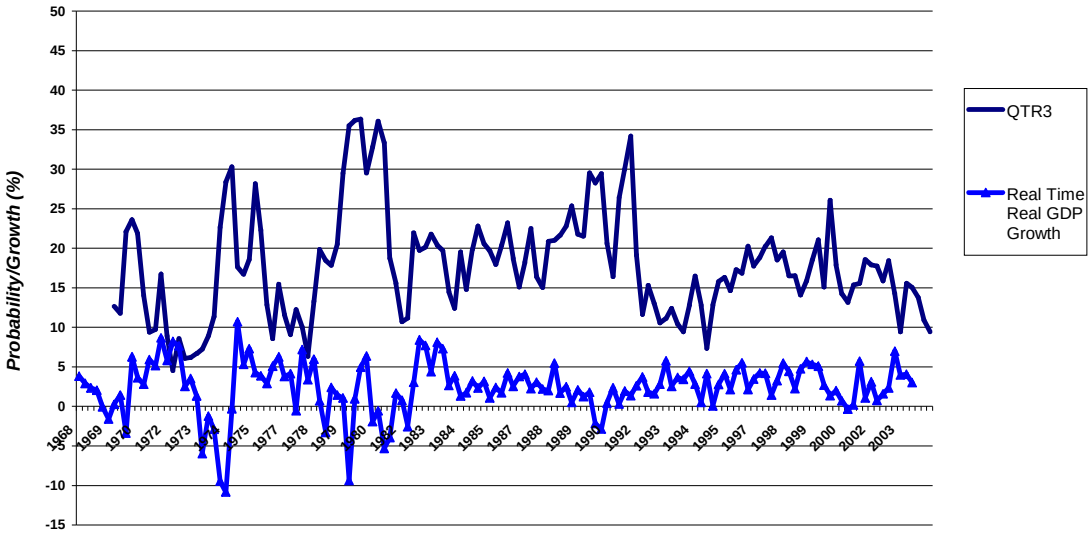

Probability of Decline in Real GDP in the Current Quarter
Probability/Growth (%)
100
90
80
70
60
50
40
30
20
10
0
-20
1968
1969
1971
1972
1974
1975
1977
1978
1980
1981
1983
1984
1986
1987
1989
1990
1992
1993
1995
1996
1998
1999
2001
2002
2004
QTR0
Real Time Real GDP Growth
Survey Data


Probability of Decline in Real GDP in the Following Quarter
Probability/Growth (%)
100
90
80
70
60
50
40
30
20
10
0
-20
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
QTR1
Real Time Real GDP Growth
Survey Data

Probability of Decline in Real GDP in Second Following Quarter



Probability of Decline in Real GDP in Third Following Quarter



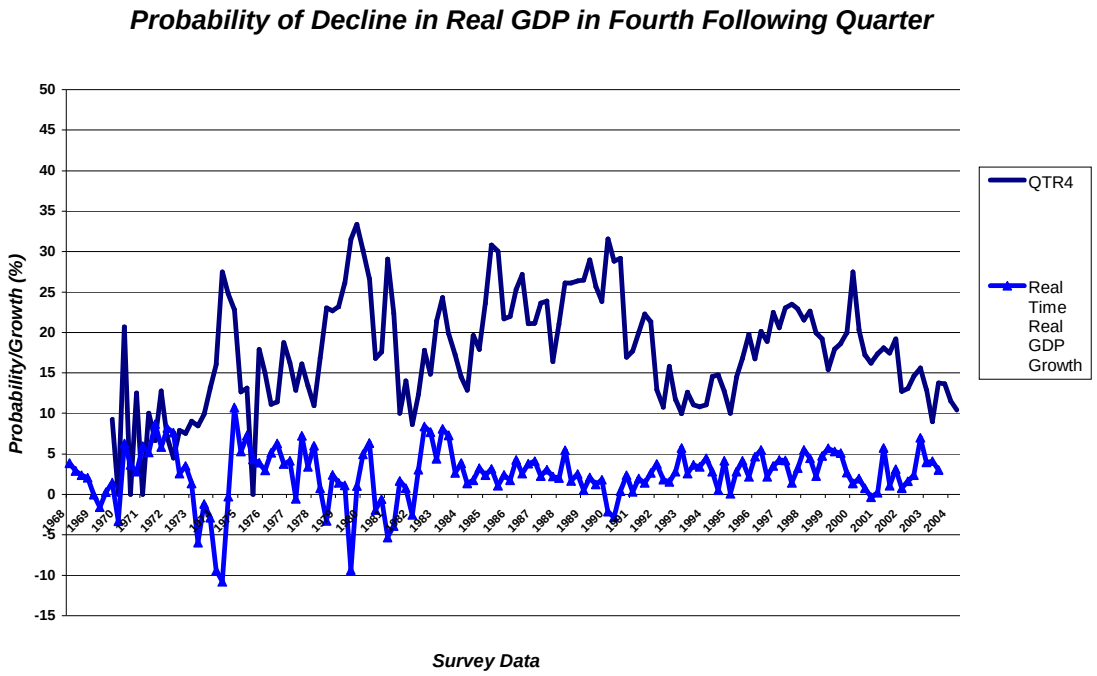


Table 1: PT Predictive Power Test

Horizon	Threshold Probability	Threshold Value	Naïve
Q0	0.015	-1.82	-34.33
Q1	0.035	-1.82	-49.52
Q2	0.050	-3.03	N/A
Q3	0.083	-1.86	N/A
Q4	0.070	-1.86	N/A

Note: (1) Threshold probability indicates the probability that makes difference for the PT test results.
(2) Threshold value is the value of the PT test that is calculated using (1) as the threshold.
(3) Naïve is the value of the PT test that is calculated using naïve translation approach.

Table 2: Contingency Matrix (Threshold)

		Realization (Zt)	
Horizon	**Forecasts/Actions**	**Bad (Zt = 1)**	**Good (Zt = 0)**
Q0	**Yes**	20	120
Q1		20	120
Q2		20	120
Q3		20	115
Q4		19	117
Q0	**No**	0	3
Q1		0	3
Q2		0	3
Q3		0	8
Q4		0	3

Note: (1) Bad (Zt = 1) represents the occurrence of the event
(2) Good (Zt = 0) represents the non-occurrence of the event
(3) Total observations are 143 quarters from Q4 1968 to Q2 2004.

Contingency Matrix (Naive)

		Realization (Zt)	
Horizon	**Forecasts/Actions**	**Bad (Zt = 1)**	**Good (Zt = 0)**
Q0	**Yes**	13	6
Q1		6	5
Q2		1	4
Q3		0	0
Q4		0	0
Q0	**No**	7	117
Q1		14	118
Q2		19	119
Q3		20	123
Q4		19	120

Kuipers Score

	Q0	Q1	Q2	Q3	Q4
Threshold	0.0244	0.0244	0.0244	0.0650	0.0250
Naïve	0.6012	0.2593	0.0175	0.0000	0.0000