Research Program on Forecasting

Evaluating a Leading Indicator: An Application: the Term Spread

Herman O. Stekler
The George Washington University
hstekler@gwu.edu

Tianyu Ye
The George Washington University

RPF Working Paper No. 2016-004
http://www.gwu.edu/~forcpgm/2016-004.pdf

March 15, 2016

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www.gwu.edu/~forcpgm

Evaluating a Leading Indicator: An Application: the Term Spread

Herman Stekler, Tianyu Ye

Abstract

This paper analyzes the procedures that have previously been used to evaluate indicators. These methods determine whether the indicator correctly classifies periods when there was (not) a recession. These approaches do not show whether or not an indicator signaled a turn or failed to predict it. This paper then presents a new approach and applies it to the term spread series. The results are mixed because the indicator predicts every recession but also generates a large number of false signals. This result may explain why economists do not always place great weight on this series.

Key words: leading series, ROC curve; yield spread puzzle

Evaluating a Leading Indicator: An Application: the Term Spread[1]

The process of detecting turning points is different from making quantitative predictions which are generated using some form of a formal model. In focusing primarily on forecasting turning points, the profession has relied on a different methodology which examines the behavior of data or series that customarily signal the onset of a recession. To be useful as an indicator, the particular series should predict most if not all recessions and should not generate many false signals. Consequently, the methodology for evaluating an indicator also differs from the procedures used to judge the accuracy of quantitative forecasts. It is most appropriate that a paper that focuses on these methodologies be in a Festschrift honoring Kajal Lahiri. He, along with his students, has made significant contributions to our understanding of the issues involved in forecasting with indicators.

This paper discusses the way that the indicators have previously been evaluated, presents a new way for judging their accuracy, and then applies this new procedure to the term (interest rate) spread. This series is analyzed for two reasons. First, there has been considerable discussion about the interest rate spread's ability to predict the onset of a recession, and the existing literature suggests that the slope of the interest rate spread can be used as a leading indicator. On the other hand, there is evidence that the information in the yield spread is not consistently

[1] I greatly appreciate the comments made by Olga Bespalova , Bryan Boulier, and Tara Sinclair on earlier versions of this paper.l

utilized by professional forecasters. This has led economists to speculate why this information is over looked and has been called "the yield spread puzzle". (Rudebusch and Williams, 2009; Lahiri et al., 2013.) Our analysis can shed some light on this puzzle.

The next section discusses the methodologies that have been used to evaluate indicators. Subsequent sections review the results of previous studies that have evaluated the yield spread and then present our results and conclusions.

1. Overview of Methodologies

There is little argument about the qualities that a good leading indicator should possess. It should accurately predict turning points in advance and should not generate too many false signals. However, not all methodologies that have been used to evaluate leading indicators focus on these criteria.[2] We illustrate that result by dividing these methodologies into a number of categories and focusing on the respective procedures.

One group of procedures consists of ad hoc rules that identify turns and calculate the tradeoff between the length of the forecasting lead and the number of false turns that are generated. This method can determine (1) whether turning points were predicted, (2) the forecasting lead and (3) the number of false signals.

A more systematic approach for evaluating indicators is based on probit models that are used to calculate the probability that an indicator has signaled a turning point. The indicator is then evaluated using quadratic probability scoring rules, some of its attributes, or Receiver

[2] This clearly is the case for methods which evaluate this series on the basis of the mean squared errors of the probability forecasts for a particular quarter. It is also true, but not so obvious, when ROC curves (that are explained below) are used to analyze the tradeoffs between true hits and false signals.

Operating Characteristics (ROC) curves. These methods show whether the indicator correctly classified the periods during which a recession did (or did not) occur. This evaluation approach does not show whether there was (not) a recession after the indicator generated a signal, or whether there was a recession that had not been predicted. We argue that these are the characteristics that should be evaluated.

> A weather forecasting analogy will illustrate this difference. A motorist going on a long trip would like to know whether it will snow anytime within the next eight hours. The driver is not interested in knowing whether the forecast was correct in classifying that there was (not) snow in each of those eight hours.

The question we ask is whether a particular series, in this case the yield spread, can or cannot predict the onset of a recession. We examine this question in the context of ad hoc rules, probability models, and two forms of the ROC curve.

1.1 Ad hoc Rules

Alexander (1958) and Alexander and Stekler (1959) recognized that in order to evaluate an indicator, it was necessary to develop rules for determining what constitutes a signal from the predictor. These rules were ad hoc and were designed to be implemented in real or pseudo-real time. The original rule was based on the number of months that the indicator was below (above) a peak (trough). (Alexander and Stekler, 1959). If every decline from a peak is counted as a signal, it was shown that there would be a very large number of false predictions. Consequently, the use of an "n or more months up or down" rule was suggested. This rule is an implicit smoothing device that generates the tradeoff between the number of false turns and the average forecasting lead. Vaccara and Zarnowitz (1977) suggested an alternative rule for identifying a

predictive signal from an indicator. There must be three consecutive declines in the indicator before saying that a signal has occurred.[3] The subsequent evaluations did not suggest other ad hoc rules but rather focused on methodologies involving turning point probabilities.

1.2 Probit Models and Probabilities

Probit models have been the preferred statistical models for calculating the probability that an indicator is signaling a recession. The probit is specified as (1):

$$Prob\ (Y_{t+k} = 1) = \int \phi\left(a_0 + a_1\ (X_t)\right) d\,t, \qquad (1)$$

where Y_{t+k} =1 is a dummy that takes on the value 1 if the period is within a recessionary quarter as defined by the National Bureau of Economic Research and is 0 otherwise. X_t represents the value of the indicator that is being evaluated. In the case of the term spread it is defined as the difference between the three-month and ten-year average yields on Treasury securities in quarter t and ϕ represents the standard normal cumulative distribution function.[4]

The probabilities that were generated from these probit regressions have been evaluated using a number of methodologies.[5] Most frequently, the quadratic probability score (QPS), known as the Brier Score, was used to measure the overall accuracy. This measure (eq. 2) is

[3] There is a distinction between the "three month down" and the three month consecutive decline rules. The indicator may be below peak for three months but may not have declined in every one of those three months. Stekler (1991) compared these two ad hoc rules. He found in favor of the "three consecutive month" rule because it significantly reduced the number of false turns without substantially affecting the forecasting lead.

[4] Dynamic specifications of the basic probit were examined by Duecker, 2005; Chauvet and Potter, 2005; and Kauppi and Saikkowen , 2008 among others. Wright also included the Federal Funds Rate in the regressions and Berge (2015) averaged the predictions over several models. In addition, several papers have examined whether the relationships may have changed over time. See Giacomini and Rossi (2005) and Estrella et al. (2003).

[5] See Lahiri and Wang (2013) for a comprehensive overview of the methodologies that can be used to evaluate probability forecasts.

analogous to the mean square error of the typical quantitative forecast and compares the estimated probabilities with the binary (0 or 1) outcomes.

$$QPS = 1/T \sum(F_t - X_t)^2 \text{ , where} \qquad (2)$$

F_t represents the calculated probabilities; X_t is a dummy variable which is either 0 or 1 depending on the state of the world; T is the number of observations. While this approach is appropriate for measuring the association between forecasts and actuals for binary variables, it does not provide definitive information about the accuracy of the indicator in predicting turning points.[6]

Lahiri and Wang (2006) and Lahiri and Yang (2015) discussed the problems involved in using the QPS as an evaluation criterion. Lahiri and Wang then suggested that the odds ratio should be used to determine whether a particular indicator accurately predicted turning points without generating a large number of false positives.[7] They defined the odds ratio as:

$$OR = FO/BO,$$

where FO are the odds in favor of an event, f/(1-f), and BO are the base rate odds, u/(1-u). In this ratio, f is the probability that the event will occur and u is the average percent that the event has occurred during the sample period. If this statistic is used, then different values of this measure could be considered as alternative thresholds for calling a turn. Different values of OR would yield a different number of both the true and the false turns that were predicted.

1.3 Conventional ROC Curves

[6] A variant of the Brier Score can also be used to determine whether the forecaster has skill.

[7] Lahiri and Wang note that Murphy (1991), in the context of weather forecasting, had suggested providing an odds ratio in addition to the raw probabilities.

There is another strand of the literature that also examines the hit and false alarm rates. This method combines these data into a graph known as the Receiver Operating Characteristic curve (ROC).[8] In this literature the hit rate is defined as the percentage of times that a turn that occurred had been accurately predicted; similarly the false alarm rate is calculated from the times that a turn did not occur but had been predicted. Basically the ROC is a classifier. For a given probability threshold, it relates the number of periods that were classified correctly by the indicator and the number of false signals. However, it does not determine whether the turning points were actually predicted or the number of times that there were signals but no turn occurred.

A typical ROC is illustrated in Figure 1 and is related to the contingency table of Table 1. The Y-axis of the ROC is the hit rate; the X-axis is the false alarm rate. Using the notation of the contingency table, the hit rate is defined as a/(a+c), and the false alarm rate is b/(b+d). These rates would be calculated for each threshold, i.e. the predicted probability obtained from the probit. The closer that the ROC is to the upper left corner of the diagram, the greater is the discriminatory power of the classifier. The 45 degree line represents the results that would be obtained when the indicator had zero skill in classifying the observations. Thus the area under the ROC, called the AUC, measures how well the indicator discriminates between recessionary and non-recessionary months. The optimal threshold that would be used in any decision problem is the one that had the maximum vertical distance between the ROC and the no skill diagonal. (See Youden, 1950).

While this curve is now being used more prominently, there are two associated problems. Similar to the problem that Lahiri and Wang (2006) had noted with respect to the

[8] The odds ratio and the ROC curve are related. Johnson (2004) provides a formal analysis.

simple probabilities: how does one interpret the results to evaluate the performance of the indicator in predicting turning points.[9] Second, if the distribution of the events that occurred and the events that did not occur are not similar, the ROC may give too much weight to the correct forecast of not predicting a turn. In the weather forecasting literature the concern is with the forecasts of rare events such as tornadoes. The correct forecast of no event, i.e. no tornado, would dominate the contingency table. (See Doswell et al, 1990). In our case the prediction that there would be no recession when in fact none occurred would be the dominant feature. In the next section we modify the ROC to account for this issue.

1.4 Modified ROC = PR Curve

We have noted that there might be an issue if there is a skewed distribution involving non-events that are correctly not signaled. In that situation, the false-alarm axis of the conventional ROC might emphasize a relationship that most forecasters would not consider important. (Laurette, 2003, p.3054). They then should not be given credit for the large number of no-no observations. If this were the case, the d cell of the 2x2 contingency table would be completely ignored and the false- alarm rate would be calculated as b/(a+b). It should be noted that, with these definitions, a high hit rate can only be achieved with high false alarm rates.

A ROC type curve drawn from these calculations would continue to plot the hit rate, a/(a+c), on the y-axis, but now the false alarm rate on the x-axis would be b/(a+b). This version of the graph has been used in the science, meteorology, machine learning, etc., literature and has been called the Precision-Recall (PR) curve. In both cases, the hit rate is measured on the y-axis; the difference lies in the x-axis. In the ROC case, true negatives are included in the data of the x-

[9] The ROC contains the number of time periods that were classified correctly, but not whether the actual turning points were forecast in advance.

axis, but the PR excludes true negatives and only considers the ratio of false to total signals. See Davis and Goodrich (2006) for a discussion of the relationship between the ROC and PR curves.[10]

In addition to excluding the true negatives in our subsequent analysis, we further modify the PR curve to only focus on whether the event itself was predicted or missed not whether the time periods were correctly classified. *We think that this is the appropriate way to measure the performance of a leading indicator and will use it in our evaluation of the accuracy of the term spread. As noted before, the distinction in the methods is between examining the forecasts for every month of a sample period versus considering only the signals for the particular events (recessions) that occurred in this time period.*

2. Previous Yield Spread Evaluations

On one hand, the existing literature suggests that the slope of the interest rate spread can be used as a leading indicator. The difference between the rate of the 10-year Treasury bond and the 3-month T-bill is the particular version of the spread that has been used in these analyses of this variable as a predictor. The seminal paper by Estrella and Hardouvelis (1991) showed that this interest spread had historically exhibited a negative correlation with the likelihood of a recession. Moreover, other research has shown that this time series is associated with future economic activity. (Among others, also see Estrella and Mishkin, 1998; Lahiri and Wang, 1996; Estrella, Rodrigues and Schich, 2003; Berge, 2015).[11]

[10] In their article, Davis and Goodrich reverse the axes of the PR curve. The PR analysis calls the x axis, precision.

[11] However, some of the newer research has extended the investigations of the relationship between financial and real variables by examining the association between the *entire* yield curve and economic activity. (See Ang, Piazzesi and Wei, 2006; Giacomini and Rossi, 2006; Wright, 2006, Rudebusch and Williams, 2009). Nevertheless, we focus on the term spread because it is that series has been the focus of empirical research.

The existing conclusions about the value of the yield spread in predicting real economic activity generally have been derived from probit regressions which yield probabilities that the economy will be in a recession some period in the future. The QPS was then frequently used to evaluate those predcitions. This was the approach that was taken by Estrella and Hardouvelis (1991), Estrella and Mishkin, 1998; Lahiri and Wang, 1996; Estrella, Rodrigues and Schich, 2003; Wright, 2006; Rudebusch and Williams, 2009; Lahiri et al., 2013; Berge, 2015. These studies all showed that there was information in the yield spread about the direction of economic activity in particular quarters but had not specifically evaluated its ability to forecast recessions.

On the other hand, Boulier and Stekler (2001) focused on the spread's performance in predicting turning points and obtained more equivocal results.[12] The spread predicted all but one of the peaks between 1957 and 1990 but also made a large number of false predictions. They also showed that there were a number of occasions when there were false reversals of true predictions, i.e. showing that a recession was not imminent, just before it occurred.[13] We now turn to the results of our own evaluation of the performance of the term spread.

3. Evaluating the Term Spread: New Results

Our analysis covers the period January 1960- August 2013. We used the NBER dates of recessions and found that there were 101 recessionary months and 543 non-recessionary months in this period. We estimated probits recursively using Equation 1. This yielded pseudo-real time probabilities that that there would be a recession in the next n months. Although the previous studies had shown that a 12-month lead yielded the best results, we also calculated these probabilities for 6-month and 18-month forecast horizons. From these probabilities and the

[12] Boulier and Stekler used the "3-months up or down" ad hoc rule in their analysis.
[13] Friedman and Kuttner (1998) showed that this had occurred prior to the 1990 recession.

actual occurrence (non-occurrence) of a recession, it was possible to tabulate the true signal and false alarm rates for both the ROC that classified each of the months and the PR which focused on the true and false signals of the actual event (recessions). In all cases, these rates depend upon the level of the probability threshold.

3.1 Classifying Recessionary Months-ROC curve

We start by determining how well the indicator classified recessionary and non-recessionary months. Table 2 presents the results for a selected number of thresholds. Even a very low probability that a recession will occur correctly identifies only 50-60% of the recessionary months and generates a large number of false signals. As an example, Table 3 presents the contingency table associated with the 25% probability. At that threshold, only 58 out of the 101 recessionary months are correctly predicted while there are 88 months which are falsely predicted to be recessionary.

The customary ROC curves derived from the data associated with Table 2 are presented in Figure 2. Our analysis confirms that the 12-month horizon is the best. The ROC curve associated with that horizon lies closest to the axes and has the largest AUC. The high value of the AUC is attributable to the large number of true negatives.[14]

3.2 Predicting Turning Points

[14] In the case of the 25% threshold there were 455 true negatives. (Table 3). Although we do not present the results, we generated the PR curve that was generated from the same data that produced Figure 2. This curve shows how well the indicator classified the months between recessionary and non-recessionary when the true negative data are excluded. This curve is further from the left hand corner of the graph and closer to the 45 degree line.

The previous analyses using either the ROC or PR curves showed how well the term spread indicator classified all the months between 1960 and 2013. These analyses do not indicate how well the term spread indicator actually predicted that a recession would occur. For this we again use the probabilities obtained from the probits but do not classify the number of months in which there was a signal. Rather, we determine whether or not a recession occurred within 12 months after that level of the signal threshold was observed. By varying the level of the threshold we can then calculate the tradeoff between the number of recessions that are predicted and the number of false signals that are observed.

Our analysis shows that the indicator can predict every recession if the probability threshold, obtained from the probit, is set very low, i.e. 0.2-0.3. However, at these probability levels, there are as many false predictions as there are correct forecasts. (See Table 4). At higher thresholds, both the number of actual recessions that are predicted and the number of false signals decline. Figure 3 presents the PR curve that displays this result.

3.3 Interpretation of Results

Although we used a different methodology, these mixed results are similar to those of Boulier and Stekler (2001). The term spread does predict every recession but only if one accepts both the low threshold probabilities of the probits and the high number of false signals that are generated even at those low thresholds.

These mixed results may provide an additional explanation for the "yield spread puzzle" that was posed by Rudebusch and Williams, 2009. Lahiri et al. (2013) demonstrated that forecasters, especially the better ones, gave more weight to information other than the yield spread. Based on those results, our findings suggest that less weight is placed on this particular

indicator because individuals may consider that the indicator's "signals" are not accurate enough to be useful.

4.0 Conclusions

This paper has examined a number of methodologies that have been used to evaluate the ability of the leading indicators to forecast recessions. A new technique, that is a modified version of a ROC curve and can be used to evaluate any indicator, was developed. In this paper, the new technique was applied to the term spread index. The results were mixed because the spread predicted every recession but also generated a large number of false signals. This result may provide an additional explanation for the failure of economists to place great weight on this series.

Table 1

Contingency Table Showing Relationship between True and False Predictions of Recessions

		Forecast	
		Recession	No Recession
Recession	**Actual**	a (true positive)	b (false positive)
No		c (false negative)	d (true negative)

Recession		

Table 2: Predicted and actual monthly recession classifications under different thresholds 1960.1-2013.8

Threshold	True Signal Rate: Percentage of true positives out of 101 actual recessionary months			False Alarm Rate: Percentage of false positives out of 543 actual non-recessionary months		
	12-month	6-month	18-month	12-month	6-month	18-month
0.25	58%	43%	30%	16%	18%	10%
0.30	51%	41%	20%	13%	12%	3%
0.40	37%	32%	9%	8%	4%	0.4%
0.50	20%	24%	2%	5%	2%	0
0.60	14%	17%	0	2%	0.7%	0
0.70	9%	10%	0	2%	0.7%	0

Table 3

Classifications of True and False Signals, Associated with a Recession Probability of .25 from Probit

		Forecast	
		Recession	No Recession
Actual	Recession	58	43
	No Recession	88	455

Table 4: Probability of detection and false alarm ratio under different thresholds, 1960.1-2013.8

Threshold	No. true signal (a)	No. false signal (b)	Probability of Detection (a/8)	False Alarm Ratio (b/(a+b))
0.2	8	10	100%	55.56%
0.3	8	7	100%	46.67%
0.4	6	4	75%	40.00%
0.5	3	2	38%	40.00%
0.6	2	2	25%	50.00%
0.7	2	1	25%	33.33%
0.8	2	1	25%	33.33%
0.9	1	0	13%	0

Figure 1
Examples of ROC Curves

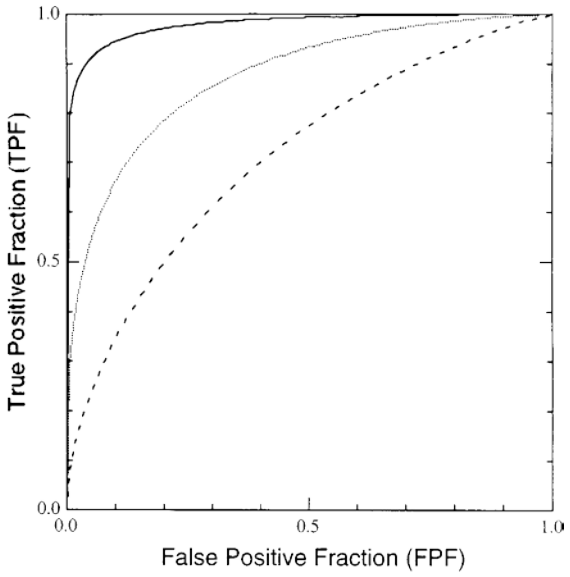


Figure 2: Comparison of ROC Curves,

6-month, 12-month, 18-month recursive predictions

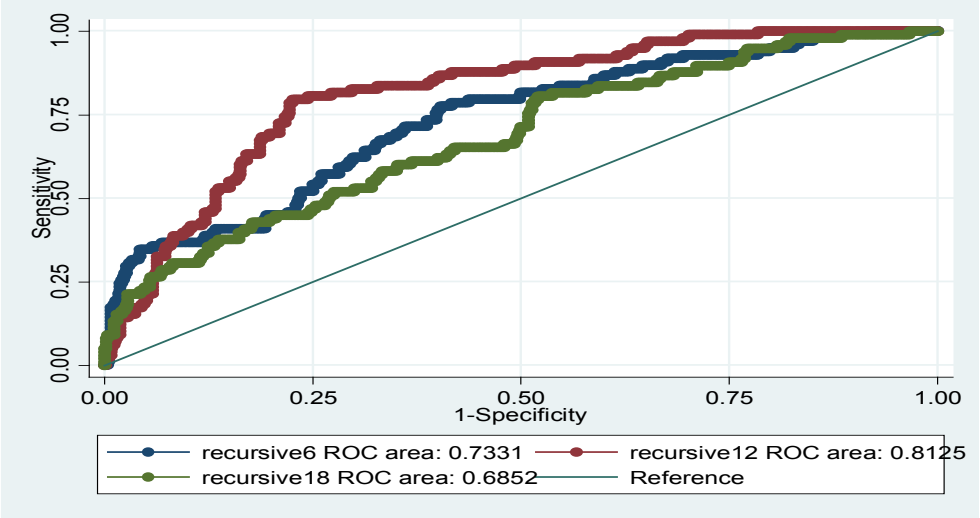
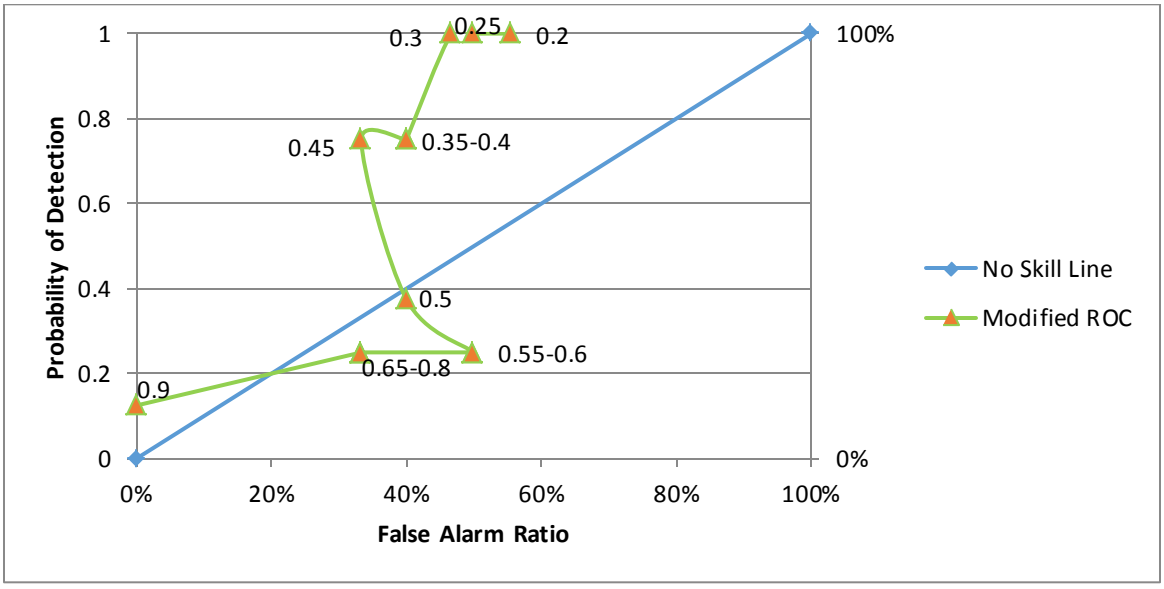


Figure 3: PR Curve,

12-month event-based recursive predictions



Bibliography

Alexander, S. (1958). Rate of change approaches to forecasting: diffusion indexes and first differences, *The Economic Journal*, 68, 288-301.

Alexander, S., Stekler, H. O. (1959). Forecasting industrial production- leading series versus autoregression, *Journal of Political Economy*, 67, 402-409.

ANG PIAZZESI WEI

Berge, T. (2015). Predicting recessions with leading indicators; Model averaging and selection over the business cycle, *Journal of Forecasting*, 34, 455-71.

Boulier, B., Stekler, H. O. (2001). The term spread as a cyclical indicator: A forecasting evaluation, *Applied Financial Economics*, 11, 403-409.

Chauvet, M., Potter, S. (2005). Forecasting recessions using the yield curve, *Journal of Forecasting*, 24, 77-103.

Davis, J., Goodrich, M. (2006). The relationship between Precision-Recall and ROC curves, *Proceedings of the 23rd Conference on Machine Learning*.

Doswell C. A., Davies-Jones, R., Keller, D. L. (1990). On summary measures of skill in rare event forecasting based on contingency tables, *Weather and Forecasting*, 5, 576-585.

Dueker, M. (2005).Dynamic forecasts of qualitative variables: A qual VAR model of U. S. recessions, *Journal of Business and Economic Statistics*, 23, 96-104.

Estrella, A., Hardouvelis, G. A, (1991). The term structure as a predictor of real economic activity, Journal of Finance, 46, 55-576.

Estrella, A., Mishkin, F. S. (1998). Predicting U.S. recessions: Financial Variables as leading indicators, *Review of Economic Statistics*, 80, 45-61.

Estrella, A., Rodrigues, A., Schich, S. (2003). How stable is the predictive power of the yield curve? Evidence from Germany and the United States, 85, 629-644.

Friedman, B. J.,Kuttner, K. N. (1998). Indicator properties of the paper-bill spread; Lessons from recent experience, *Review of Economic Statistics*, 80, 34-43.

Giacomini, R., Rossi, B. (2006). How stable is the forecasting performance of the yield curve for output? *Oxford Bulletin of Economics and Statistics*, 68(s1), 783-795.

Johnson, N. P. (2004). Advantages to transforming the receiver operating characteristic (ROC) curve into likelihood ratio co-ordinates, Statistics in Medicine, 23, 2257-226.

Kauppi, H., Saikkonen, P. (2008). Predicting U.S. recessions with dynamic binary resonse models, *Review of Economic Statistics*, 90, 777-791.

Lahiri, K., Monkroussos, G., Zhao, Y. (2013). The yield spread puzzle and the information content of SPF forecasts, *Economics Letters*, 118, 219-221.

Lahiri, K., Wang, J. G. (2006). Subjective probability forecasts for recessions: Guidelines for use, *Business Economics*, 41, 26-37.

Lahiri, K., Wang, J. G. (2013). Evaluating probability for GDP declines using alternative methodologies, *International Journal of Forecasting*, 29,175-190.

Lahiri, K. Yang L. (2015). A further analysis of the Conference Board's new Leading Economic Index, *International Journal of Forecasting*, 31, 46-453.

Laurette, F. (2003). Early detection of abnormal weather conditions using a probabilistic extreme forecast index, *Quarterly Journal of the Royal Meteorological Society*, 129, 3037-3057.

Murphy, A. H. (1991). Probabilities, odds, and forecasers of rare events, *Weather and Forecasting*, 6, 302-306.

Rudebusch, G. D., Williams, J. C. (2009). Forecasting recessions; The puzzle of the enduring power of the yield curve, *Journal of Business and Economic Statistics*, 27, 492-503.

Stekler, H. O. (1991). Turning point predictions, errors and forecasting procedures, K. Lahiri and G. H. Moore ed., *Leading Economic Indicators* (Cambridge University Press), 169-181.

Vaccara ,B. N., Zarnowitz, V. (1977). How good are the leading indicators?, Proceedings of the Business and Economics Statistics Sections American Statistical Association, 41-50.

Wright, J. H. (2006). The yield curve and predicting recessions, Board of Governors of the Federal Reserve System *Finance and Economics Discussion Series,* No. 2006-7.

Youden, W. J. (1950). Index for rating diagnostic tests, *Cancer*, 3, 32-35.