The Classification of the Economic Activity*

Abstract

Which would you prefer, a prediction of business cycle turning points where expansion periods are correctly classified 90% of the time but recessions are misclassified as expansions 40% of the time; or a prediction that correctly classifies recessions 90% of the time but misclassifies expansions 40% of the time? An economist might be quick to answer that it depends on the forecast loss function, specially if it is felt that the costs and benefits of economic policy are asymmetric in each circumstance. If we were to abstract from the loss function momentarily, and since we observe expansion/recession periods 80/20% of the time, we may be tempted to choose the second of the two predictions since it classifies accurately events that are seldom observed. It may surprise the reader to learn that both of these classifications were generated with the same model. This paper discusses methods to classify economic activity appropriately, and to evaluate forecasting models of business cycle turning points.

- *JEL Codes:* E32, E37, C14
- *Keywords:* business cycle turning points, receiver operating characteristics (ROC) curve, relative operating levels (ROL) curve, Business Cycle Dating Committee of the National Bureau of Economic Research.

Travis J. Berge and Òscar Jordà
Department of Economics
U.C. Davis
One Shields Ave.
Davis, CA 95616

E-mail (Berge): tjberge@ucdavis.edu

E-mail (Jordà): ojorda@ucdavis.edu

*We thank John Fernald, James Hamilton, Kevin Lansing, Jeremy Piger, Glenn Rudebusch, Alan Taylor, and John Williams for useful comments and suggestions.

1 Introduction

The Business Cycle Dating Committee (BCDC) of the National Bureau of Economic Research (NBER) was formed in 1978 to establish a historical chronology of business cycle turning points. The NBER itself was founded in 1920 and it published its first business cycle dates in 1929, although records are now available retrospectively starting with the trough of December 1854. Public disclosures about contemporary cyclical turning points are often made with more than a year's delay – the mission of the BCDC is not to serve as an early warning system to policy makers but to be a repository of the classification of economic activity for the historical record. Although other countries now have similar committees (including the Euro Area Business Cycle Dating Committee of the Centre for Economic Policy Research founded in 2002), it is fair to say that the length of historical coverage and the experience of the BCDC have no equal.

This paper investigates several aspects about the BCDC's chronology. First, we ask how successful is the BCDC in classifying business cycle phases given that the true state of the economy is unobservable. Second, we find that two business conditions indices maintained by the Chicago and Philadelphia Federal Reserve Banks are highly accurate, real-time barometers of economic activity with respect to the BCDC's ex-post chronology. However, these indices provide poor advance warning on impending shifts in economic activity. Instead and thirdly, we find that the spread between long and short yields of treasury securities have very reasonable classification ability 12 to 15 months into the future. The contribution of this paper is to provide formal answers about the dating of business cycles generated by the BCDC, the process of which remains largely unknown.

The desire to keep a chronology of turning points – peaks versus troughs of economic activity and hence implicitly the classification of historical economic time series into periods of expansion and recession – reflects the notion that there are fundamental differences between these two phases of the economic cycle. Otherwise, the dating of business cycles would amount to a mindless, mechanical, accounting exercise. Rather, the BCDC's definition of a recession available in its most recent press release of December 11, 2008; states that:

> A recession is a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in production, employment, real income, and other indicators.
>
> *Determination of the December 2007 Peak in Economic Activity, December 11, 2008. Business Cycle Dating Committee of the National Bureau of Economic Research.*

This definition, which harkens back to Burns and Mitchell (1946), makes clear that the BCDC does not simply take, for example, a negative observation of industrial production to indicate that the economy is in recession – that same negative datum for industrial production will sometimes be classified as belonging to an expansion and other times as belonging to a recession. It is this classification of economic activity into expansions and recessions that suggests economic activity can be thought of, for example, as coming from a mixture of two distinct distributions. Analysis and predictions of the binomial state of the economy is an interesting and useful tool for at least two reasons. Information regarding a binomial variable describing aggregate economic activity is simple and thus easily understood by both policy makers and the general public. Indeed, for this reason

policy makers may be more interested in a probabilistic statement regarding recession/expansion states than a point-estimate of aggregate GDP growth. In addition, however, it has long been recognized that a wide variety of economic variables across many sectors of the economy exhibit high levels of conformity. The knowledge of the state of the economy is thus a valuable addition to an econometrician's information set when making predictions of particular variables of interest.

We begin our analysis by: (1) examining the BCDC's chronology itself; (2) comparing the BCDC's chronology to other classification methods, such as the rule requiring two consecutive quarters of negative growth to declare a recession that is so often cited in the press, as well as more sophisticated statistical rules based on hidden Markov models, see e.g. Chauvet and Hamilton (2005) and Chauvet and Piger (2008); and (3) investigating the BCDC's dating in regard to the optimal timing for different macroeconomic indicators that may not by fully coincident with each other. For example, the phase shift between the BCDC's peak-trough dates and payroll/household employment growth suggests that employment growth lags by about seven months the end of a recession.

The methods that we use to accomplish this analysis, although uncommon in economics, have their earliest origin perhaps in Peirce (1884)'s "Numerical Measure of the Success of Predictions." Peirce's definition of the "science of the method" is the precursor to the Youden (1950) index for rating medical diagnostic tests and the receiver operating characteristics (ROC) curve introduced by Peterson & Birdsall (1953) in the field of radar signal detection theory. The ROC curve methodology was then quickly adopted into medicine by Lusted (1960) and is now a common standard of evaluation of medical and psychological tests (see Pepe (2003) for an extensive monograph). The ROC curve approach has also been adopted into the atmospheric sciences (see Mason (1982) for an early refer-

ence), where it has now become part of the World Meteorological Organization's (WMO) Standard Verification System for assessing the quality of weather forecasts (see Stanski, Wilson & Burrows (1989) and Organization (2000)). These methods are now also popular in the machine learning literature (see Spackman (1989) for an early discussion). Recent applications to economics include Jordà and Taylor (2009a, b).

The second question we ask is how informative are business conditions indices about the current state of the business cycle given that the BCDC's classification is available only with a considerable lag. We investigate three such indices: (1) the Chicago Fed National Activity Index (CFNAI), which is a monthly weighted average of 85 indicators of economic activity derived from Stock and Watson (1999); (2) the Aruoba, Diebold and Scotti (ADS) Business Conditions Index at the Federal Reserve Bank of Philadelphia, which is a daily index based on Aruoba, Diebold and Scotti (2009); and (3) the Purchasing Managers Index (PMI), an index released by the Institute for Supply Management that consists of data from surveys of purchasing managers across the country. We find that all three indices are accurate, real-time barometers of economic activity with respect to the BCDC dates available ex-post.

A contemporaneous measurement of the economic climate is useful for policy makers but perhaps not as useful as having an advanced warning on business cycle shifts. In this respect, we find the CFNAI, ADS and PMI indices to be poor predictors, as we will show. We instead show that the spread between long and short maturity yields of treasury securities have very reasonable classification ability 12 to 15 months into the future, echoing similar results reported in the literature, for example in Rudebusch and Williams (2009).

Typical measures of forecasting accuracy for binary outcomes include the *mean absolute error* (MAE); the *root mean square error* (RMSE); and the *log probability score* (LPS), which are all based on specific underlying forecast loss functions. Instead, another novel contribution of our paper is to introduce a set of statistical tools based on ROC analysis that offer several advantages over these traditional measures: (1) the new measures do not depend on the overall prevalence of recessions over the sample examined (i.e., if recessions are observed only 16 percent of the time such as in our sample, then a rule that predicts every period to be an expansion will correctly predict expansions 84 percent of the time but this does not make that rule very useful); (2) because strictly monotone transformations of prediction indices have the same ROC curve, the new evaluation methods automatically encompass a larger class of specifications and are not directly tied to modeling ability but to the information content of the variables considered; and (3) the ROC curve is independent of the forecast loss function considered.

The next section describes these methods in more detail so as to provide a suitable introduction to readers new to this literature and as a launching pad for applications of these methods in other areas of economics. Following this discussion, we then evaluate the BCDC's skill at dating phases of the economic cycle. Once we establish the BCDC's ability, we show that the CFNAI and ADS indices can be used to produce accurate, real-time stand-ins for the BCDC's ex-post classification. However, since these indices are poor at predicting turning points, we then show that yield spreads can be used with considerable success to obtain predictions 12 to 15 months in advance.

2 Statistical Methodology: ROC and ROL curves

This section outlines the statistical methods that we use in subsequent sections. We begin by introducing the basic ingredients. Let $S_t \in \{0, 1\}$ denote the true state of the economy with 0 denoting that t is an expansion period and 1 a recession period instead. For the time being, assume that the BCDC can determine the value of this variable with 100% accuracy, albeit with a considerable delay. Meanwhile, consider the index Y_t and the threshold c which together define the binary prediction *recession* whenever $Y_t \geq c$, and *expansion* whenever $Y_t < c$. Y_t may denote a real-time probability prediction about S_t (such as that obtained with a probit model), a linear index (e.g. such as would be available from a linear probability model), an index from a more complicated statistical model (e.g. a neural network estimator as is often used for classification purposes), or simply an observable variable (e.g. a leading indicator). The distinction is unnecessary for the methods we describe.

Associated to these variables, we can define the following conditional probabilities:

$$\begin{aligned} TP(c) &= P\left[Y_t \geq c | S_t = 1\right] \\ FP(c) &= P\left[Y_t \geq c | S_t = 0\right] \end{aligned}$$

$TP(c)$ is typically referred to as the *true positive rate*, *sensitivity*, *recall rate*, in the more familiar Neyman-Pearson nomenclature *1-Type II error*, or the *power* of the test (the test here being the ability to correctly predict the state of the economy at time t). $FP(c)$ is known as the *false positive rate, (1-specificity), Type I error* or the *size* of the test.

In biostatistics, a common summary of the properties of a binary predictor is the receiver operating

characteristics (ROC) curve. The ROC curve plots the entire set of possible combinations of $TP(c)$ and $FP(c)$ for $c \in (-\infty\infty)$. When $c \to \infty$ then $TP(c) = FP(c) = 0$ and conversely, when $c \to -\infty$ then $TP(c) = FP(c) = 1$. Thus, the ROC curve is a monotone increasing function in [0,1] $\times$ [0,1] space. If Y_t is unrelated to the underlying state of the economy S_t, then Y_t is an uninformative classifier, $TP(c) = FP(c)$ $\forall c$ and the ROC curve would be the 45^0 line – a natural benchmark with which to compare classifiers. On the other hand, if Y_t is a perfect classifier, then the ROC curve will lie along the vertical and upper borders of the positive unit quadrant. Typical applications result in ROC curves between these two extremes with the quality of the classifier being reflected in how far the ROC curve is from the 45^0 line.

As an illustration, Figure 1 displays the ROC curve for a home-made index of business conditions we prepared. This index is based on the number of news items with the word "recession" in the Lexus-Nexus database.[1] This index is similar in spirit to what Google Trends (visit www.google.com) does to track the incidence of influenza throughout the year. Specifically, by tracking search activity on influenza related searches, Google is able to provide a useful two-week ahead prediction of influenza incidence as reported by the Centers for Disease Control. We use our index in raw form – there is no model here, we just want to evaluate how useful is the index to classify the data into recessions and expansions based on the BCDC's chronology.

The top panel of Figure 1 articulates the different trade-offs in accuracy for expansion/recession predictions. For example, the point with $TP(c) = 0.75$; $FP(c) = 0.25$ in the ROC curve displayed says that our index will correctly classify recession periods 75% of the time while incorrectly clas-

[1] The index takes the raw counts of incidences and adjusts for the trend in the number of news outlets included in the Lexus-Nexus database over time and for seasonality. A more detailed description is available in the appendix.

sifying expansion periods as recessionary 25% of the time. Increasing the accuracy of the recession predictions to, say about 90% would naturally increase the rate at which we incorrectly classify expansion periods as recession to about 50%. Conversely, lowering the correct classification of recession periods to about 50% would essentially halve the false positive rate to about 12%. Depending on the costs and benefits associated to policy hits (pursuing expansionary policy to stimulate an economy in recession) and policy mistakes (inflating an economy that is not), the optimal operating point can be determined as the point where the slope of the ROC coincides with the marginal rate of substitution between a policy maker's utility of hits and misses. For completeness, the bottom panel of Figure 1 displays our index and the Google trends index of searches of the word recession, which unfortunately is only available for a much shorter sample.

There are several ways to summarize the visual information contained in a ROC curve with formal statistics. The most widely used summary statistic is the *area under the ROC curve* (AUROC). Let $r \equiv FP(c)$. Since the mapping between c and r is unique, we can define $ROC(r) = TP(c)$ for $r \in (0, 1)$. The ROC curve is a plot of $\{ROC(r), r\}$ for $r \in (0, 1)$, and the AUROC can be defined as

$$AUROC = \int_0^1 ROC(r)dr; \qquad AUROC \in [0.5, 1]. \tag{1}$$

A perfect classifier Y_t has an $AUROC = 1$ whereas a coin-toss classifier has an $AUROC = 0.5$, with most classifiers falling somewhere in-between. Given two classifiers, Y_t and Z_t such that $ROC_y(r) \geq ROC_z(r)\ \forall r \in (0, 1)$ then it is also the case that $AUROC_y \geq AUROC_z$ although the converse is not necessarily true.

One way to compute the AUROC nonparametrically on a given sample where there are no ties (with

continuous distributions, a zero probability event) is (see Mason & Graham (2002)):

$$\widehat{AUROC} = 1 - \frac{1}{n_0 n_1} \sum_{i=1}^{n_1} f_i$$

where f_i is the number of observations with associated $S_t = 0$ but that have a higher value of Y_t than the current observation i for which it is the case that $S_i = 1$, where n_0 is the number of observations for which $S_t = 0$, n_1 is the number of observations when $S_t = 1$, so that $n_0 + n_1 = T$, the total sample size.[2]

Bamber (1975) was the first to point out that the AUROC can be rescaled into a Mann-Whitney (Mann & Whitney 1947) U-statistic as follows

$$U = n_0 n_1 (1 - AUROC),$$

where

$$U = \sum_{i=1}^{n_1} \rho_{1i} - \frac{n_1(n_1 + 1)}{2},$$

where ρ_{1i} is the rank of the i^{th} case of the sample of observations for which $S_t = 1$. The Mann-Whitney statistic ranks the T observations in the sample and then compares the sum of the ranks in the sample for which $S_t = 1$ with the sum of the ranks for which $S_t = 0$. The test entertains the null that the two samples of observations come from the same distribution. The advantage of the Mann-Whitney statistic (and therefore the AUROC) is that it has an asymptotic normal distribution (see, e.g., Sheskin (2000)). The mean of the Gaussian approximation for AUROC is 0.5 and Hanley and McNeil (1982) provide formulas for efficient estimators of the variance.

[2] See Hanley and McNeil (1982) for a more efficient non-parametric procedure. Here the simpler formula makes clear the connection to the Mann-Whitney statistic.

The AUROC is also related to the Wilcoxon statistic (Wilcoxon 1945):

$$W = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_0} I(Y_i, Y_j)}{n_0 n_1},$$

where

$$I(Y_i, Y_j) = \begin{cases} 1 & \text{if } Y_i > Y_j \\ 0 & \text{otherwise} \end{cases}.$$

The statistic W is a non-parametric estimator of $P[Y^1 > Y^0]$, i.e., the probability that the index for an observation for which $S_t = 1$ is higher than that for an observation with $S_t = 0$. It is usually computed to test whether the outcome of a variable in one population tends to be greater than in the second population, without actually assuming how the variable is distributed in the two populations. Pepe (2003) shows that $AUROC = P[Y^1 > Y^0]$, giving the AUROC an intuitive interpretation.

Under general conditions (see Hsieh and Turnbull, 1996 for specifics) then

$$\sqrt{T}(\widehat{AUROC} - P[Y^1 > Y^0]) \rightarrow N(0.5, \sigma^2)$$

with an analytic estimator for σ^2 available, e.g. in Hanley and McNeil (1982) and Obuchowski (1994). See also the discussion in Jordà and Taylor (2009a) for bootstrap methods. Hence, a test of the null $H_0 : AUROC = 0.5$ (i.e., no classification ability) can be simply obtained with

$$t = \frac{\widehat{AUROC} - 0.5}{\sigma^2} \rightarrow N(0, 1).$$

Similarly, tests to compare the ability between two classifiers can be simply derived from the traditional Wald principle (see, e.g., Hanley and McNeil, 1983). For a more detailed discussion on other

inferential procedures, see Jordà and Taylor (2009a). Returning to Figure 1, the AUROC for our home-made index of economic conditions is $\widehat{AUROC} = 0.81$, which is statistically different from 0.5 with a p-value that is essentially zero.

The ROC curve and associated statistics provide a formal and general framework with which to assess binary predictions in which the BCDC's turning points are taken to reflect the true state of the economy at the time. A more intriguing question is to ask how good a job is the BCDC doing. However, since the true state is unobservable, how is it even possible to answer this question?

The approach that we take here is to ask ourselves whether a recession, as classified by the BCDC chronology, corresponds, not just to periods when economic activity is below normal, but to periods when economic activity is *considerably* below normal. In other words, if the BCDC's dating amounted to no more than a coin-toss, one would expect the resulting empirical distributions for expansions and recessions to be very similar. The better the BCDC is at classifying the data, the more *different* these empirical distributions will be. How can we therefore assess this difference? The trick consists in thinking of the S_t provided by the BCDC as forecasts rather than realizations, and to investigate the dual of the ROC curve, which is called the *relative operating levels* (ROL) curve (see Mason and Graham, 2002).

Essentially, the AUROC in Figure 1 is a non-parametric estimate of this separation between empirical distributions and when this value (which we will label *area under the relative operating levels* (AUROL) curve to emphasize that the S_t are predictions rather than realizations) is close to 0.5, we will conclude that the BCDC has low classification skill, whereas an AUROL close to 1 indicates high classification skill. The asymptotic distribution of this Mann-Whitney U-statistic is still Gaussian.

Thus, we have a formal framework to compare classification ability of economic cycles between the BCDC and alternative dating chronologies.

This discussion forms the basis of our subsequent analysis. We begin the next section by providing an overall assessment of the BCDC's implied classification skill to obtain a better understanding of which indicators are best classified by the BCDC. In order to do this, we calculate ROL curves for each of the economic indicators used by the BCDC at a variety of horizons to account for differences in the coherence of different indicators. After having obtained an understanding of the BCDC's recession dates and the BCDC's ability to classify economic activity, we then investigate the performance of a number of popular binary predictors of recessions. Even if the BCDC were to perfectly classify the data, having an understanding of the accuracy of these predictors is important, since they have the benefit of being available to researchers in real-time (or at least with less delay than official BCDC dates).

3 Assessing the Business Cycle Dating Committee

3.1 Four Definitions of Recession

The BCDC produces the definitive series of business cycle turning points for the U.S. economy in the form of the month within which the day of a peak or a trough of economic activity occurs (see the BCDC's release of 2008). Each peak and trough month is therefore a mix of economic expansion and recession. It is generally accepted that trough months should be classified as recessions but there is more ambiguity as to how peak months should be classified. For example, Wright (2006) and Chauvet and Hamilton (2005) define recessions as the period between a peak and a trough,

inclusive of both peak and trough months. We denote these methods BCDC-PI (for *peak included*). Rudebusch and Williams (2009) instead choose to date recessions by excluding peak months. We denote this rule BCDC-PE (for *peak excluded*).

In addition, there are two common rule-of-thumb definitions of recessions used in the literature. A mechanical definition consists in classifying as recessions any periods in which GDP growth is negative. Rudebusch and Williams (2009) call this rule *R1.* Similarly, the press often refers to the unwritten rule that a recession requires at least two consecutive quarters of negative GDP growth. Following Rudebusch and Williams (2009) we will call this method $R2$.

The first order of business is therefore to establish which of the two BCDC rules is preferable and how do they compare to the more mechanical $R1$ and $R2$ rules. Table 1 tabulates the salient features of each classification method. Of the 750 months between January 1947 and June 2009, the series BCDC-PE classifies 122 months as recessionary, so that the US economy is in recession 16 percent of the time. BCDC-PI generates 11 more months of recession, so that the economy is in recession for 18 percent of the time. As can be seen in the fourth column of the table, the average NBER-defined recession lasts about 12 months.

The series $R1$ is very noisy relative to the competing rules, in the sense that the recessionary periods are on average much shorter but occur much more frequently. Although $R1$ identifies a similar number of recessionary months as does the BCDC – 123 – $R1$ produces a total of 82 incorrect signals relative to the BCDC-PE series. Of the incorrect signals, there are 36 false positives (expansion months misclassified as recessionary) and 46 false negatives (recession months misclassified as expansionary). The missed signals often result from upticks in GDP growth within a BCDC-defined

recession, or downticks immediately before or following a BCDC recession. Conversely, $R2$ is much more conservative – only 75 months are identified as recessionary, 10 percent of the sample. Of those 75 months, there are only 10 false positives, but because the rule is so strict it produces 68 false negative signals relative to BCDC-PE. The $R2$ rule completely misses the 2001 NBER-declared recession.

How do we determine which of these four definitions of recession best classifies economic activity? Clearly the choice of definition has important implications down the line when we evaluate business conditions indices and yield curve spreads as possible concurrent and leading indicators of business cycles. Our approach is to examine the ROL curves and statistics for each of the economic indicators cited by the BCDC of the NBER in the 2008 release and compare their respective AUROLs. Because there could be some phase shifts across indicators (for example, it is well-known that employment tends to lag considerably in economic recoveries), we examine a window of ± 24 months around turning points. Hence for each observation t, let $h \in \{-24, ..., 0, ...24\}$ be how the window is constructed.

Specifically, the BCDC bases its decisions on five monthly indicators of economic activity: industrial production (IP); real personal income less transfers (PI); payroll employment (PE); household employment (HE); and real manufacturing and trade sales (MTS). In addition, the BCDC considers two quarterly indicators: real gross domestic product (GDP); and real gross domestic income (GDI). We constructed the indicators as indicated in the appendix of the NBER's most recent declaration (2008) and obtained the data directly from the sources listed there. Details are provided in the appendix of this paper. We constructed interpolated monthly series of GDP and GDI using the same

method indicated to interpolate the GDP deflator used in constructing MTS to allow for direct comparison between the quarterly and monthly indicators. For each variable, we take the 12-month (or 4-quarter) log difference to compute yearly growth rates in a way that produces a smoother series than taking the annualized monthly (quarterly) log differences. This process also has the benefit that we do not have to worry about seasonal adjustments to our data since we consider year-over-year growth rates.

Table 2 reports the AUROL estimates of these experiments for the optimal value of h. It is readily apparent that both BCDC series are superior to the mechanical $R1$ and $R2$ series in all but one case (and there, the difference is not statistically significant). Moreover, the AUROLs for the $R1$ and $R2$ series are in most cases statistically different from the BCDC-PE series, which we take henceforth as our benchmark because it achieves the highest AUROLs overall.

As a complement to Table 2, the top panel of Figure 2 displays the AUROLs associated with the BCDC-PE classification for the monthly interpolation of GDP against h, $h \in \{-24, ..., 0, ...24\}$ along with the lower and upper bounds at the 95% confidence interval associated to the AUROL at each h. The panel reveals that the AUROL is maximized at $h = 3$ (rather than at $h = 0$) with an AUROL that is virtually equal to 1 (perfect classification ability). We are cautious interpreting this three-month phase shift because the series have been interpolated from quarterly data. For this reason, the bottom panel of Figure 2 displays the same exercise for the complete list of economic indicators considered.

Several results deserve comment. As we observed in the top panel of Figure 2, GDP maximizes the AUROL for $h^* = 3$ (where h^* is the notation that denotes the h that maximizes the AUROL) but

this series is interpolated from quarterly data. Similarly GDI has $h^* = 4$ but it is also interpolated. At first we suspected that the interpolation might be causing the apparent phase shift between the BCDC dates and the AUROLs for GDP and GDI. However, we then found that MTS has $h^* = 3$ whereas IP has $h^* = 4$. In addition, HE and PI have $h^* = 6$ and PE has $h^* = 7$. Although the latter results conform well with the observation that the recovery in employment tends to lag the end of recessions, we found the results on production quite intriguing. To clarify, the results on IP, MTS, GDP and GDI suggest an *average* phase shift between the BCDC's dating and when the AUROLs are maximized of about 3 months. From the bottom panel of Figure 2, it is difficult to be overly precise ± 1 month around these numbers but beyond that, the deterioration in AUROLs is quite quick.

We find the results summarized in Figure 2 particularly novel. At first blush, the disparity of horizons at which the AUROLs are maximized may appear to impede the BCDC's effort to come up with a unique chronology of peaks and troughs. Instead, we think it can be advantageous for the BCDC to exploit the timing at which each of these series peaks to come up with more accurate pronouncements on business cycles.

3.2 The BCDC versus Statistical Dating Rules

The BCDC's dating of business cycles is the universally accepted gold standard against which competing methods of turning point prediction are evaluated. Thus, even models in which the state of the economy is a latent process that can be estimated independently of the BCDC's classification (such as the class of hidden Markov models spawned by Hamilton's (1989) seminal work), measure

their success by comparing estimates of the smoothed state probabilities against the BCDC peak-trough dates. This section turns this view point on its head and instead asks how well does the BCDC compare to Chauvet and Hamilton's (2005) and Chauvet and Piger's (2008) expansion/recession dating.

The first algorithmic recession that we explore is due to Chauvet and Hamilton (2005). We consider a interpolated version of their quarter Markov-switching model. This index has the benefits that it is readily available (see: `http://www.econbrowser.com/archives/rec_ind/description.html`), transparent and relatively simple. The quarterly version of the index developed in Chauvet and Hamilton (2005) (henceforth the CH index) is based on the observation that recessions are persistent and that the empirical distributions of GDP growth are informative about the underlying state of the economy. The authors utilize empirical information on GDP growth in order to estimate a two-state Markov chain, where the state represents whether the economy is in expansion or recession. Given a new observation of GDP growth, the CH index produces a probability that the economy is in a recession. In order to translate this probability into a zero-one indicator of the state of the economy at a monthly frequency, we first interpolate the quarterly index into a monthly one using the same methodology as we do for the other quarterly indicators. We then apply the simple rule-of-thumb that any period with a recession probability greater than a particular threshold value is classified as recessionary. In order to find the threshold that maximizes the associated AUROCs, we performed a grid-search over the space 0.5-0.9. We find that the optimal threshold is 0.75, as this value produces the greatest maximized AUROC for the majority of variables analyzed here.

Chauvet and Piger (2008) produce a similar index using a variety of indicators available at a monthly

frequency instead of solely quarterly GDP growth rates. The Chauvet Piger (CP) index is also readily available (see Jeremy Piger's homepage, `http://www.uoregon.edu/~jpiger/`). Chauvet and Piger estimate a dynamic factor model in the vein of Stock and Watson (1989) using data on four coincident variables: nonfarm payroll employment; industrial production; real manufacturing and trade sales; and real personal income less transfer payments. The common factor is assumed to have a switching mean, given by $\mu = \mu_0 + \mu_1 S_t$, where S_t is the variable that indicates the state of the underlying economy. Estimation of the model again produces an index that can be interpreted as the probability that the economy is in recession. To produce a binomial variable, Chauvet and Piger use a two-step process. First, CP wait until the index is greater than or equal to 0.80 for a series of three months. Let the first month of this series be t. The first month of the recession is then identified as the first month prior to month t for which the probability of recession is greater than 0.5.

Table 3 displays the AUROCs associated with each of these recession definitions for each of the indicators previously considered. Columns 1-3 show the contemporaneous AUROCs, while columns 4-6 display the AUROC that is maximized by varying the time horizon h, as well as the associated time horizon. The results in Table 3 suggest that the BDDC does a very good job in classifying economic data, even if it does not mechanically combine these data in any optimal way.

Unsurprisingly, the two algorithmic indices best describe data that are used to produce that index, but describe other data less well. For example, we see that the CH index performs just as well as the BCDC and the CP index when describing GDP growth, but tends to fall short with other indicators where the CH index must rely only on the comovement of these variables across the business cycle.

Interestingly, although the CH index is produced completely independently of the BCDC turning points, the CH index seems to have similar timing as the BCDC dates given that the maximum AUROCs tend to occur at similar horizons.

Similarly, the CP index outperforms the other two indices for three of the four variables used in the estimation of the dynamic factor model. The only time that the BCDC is outperformed for a particular variable in a statistically meaningful sense is by the CP index for Household Employment – the maximum AUROC for household employment for the CP index is larger than the maximum AUROC from the BCDC (both maxima are achieved at $h=7$) at the 90 percent confidence interval.

All in all, this analysis indicates that BCDC's flexible approach allows for the classification of a broad range of variables while sacrificing very little by means of misclassification of any individual series. For only one indicator, household employment, can we be fairly certain that the BCDC is being outperformed by an algorithmic procedure. Moreover, neither the CH or CP index here outperforms the BCDC for all economic variables considered in this analysis.

4 Dating Recessions in Real Time

The previous section demonstrated that the BCDC chronology provides a very useful retrospective classification of economic activity. However, since the BCDC's statements about turning points are released with considerable delay (upwards of a year), it would be useful to have real-time assessments about the business cycle. In this section and in contrast to section 3, we take the BCDC chronology to be the true historical record (rather than a prediction) of the state of the economy and we ask whether there are indices of business conditions that can accurately and in real-time, determine that

which the BCDC only provides with a lag.

We investigate three indicators (plus our home-made indicator from Section 2), that are reasonably representative of popular business conditions indices even if the list is by no means are exhaustive. We include several different types of indices, all of which are freely and publicly available. Two of the indices represent state-of-the-art approaches to measuring aggregate economic activity in real time. First we consider the Chicago Fed National Activity Index (CFNAI), a monthly index constructed as a weighted average of 85 monthly indicators of national activity, drawn from four broad categories: production and income; employment, unemployment and hours; personal consumption and housing; and sales, orders and inventories. The CFNAI corresponds to the index of economic activity in Stock and Watson (1999). More details can be found in their paper and in the Federal Reserve Bank of Chicago's website. The second index we consider is the Aruoba, Diebold and Scotti (ADS) Business Conditions Index maintained by the Federal Reserve Bank of Philadelphia. The ADS index is a new daily index designed to track real business conditions at very high frequencies. It is based on a smaller number of indicators that CFNAI. The details about its construction can be found in Aruoba, Diebold and Scotti (2009) and at the Federal Reserve Bank of Philadelphia's website.

The other two indices that we investigate here represent a different approach to measuring economic activity in real-time. Instead of attempting to measure economic activity directly, these two indices rely on information available to and taken from market participants. The first index is the Purchasing Managers Index (PMI), which has been issued since 1948 by the Institute of Supply Management. The data for the index are collected through a survey of 400 purchasing managers in the manufacturing sector. The PMI is available at a monthly frequency (more details can be

found at the Institute of Supply Management's website). We also include the index that we introduced in Section 2 based on a standardized measure of the counts of news items containing the word "recession" contained in the Lexus-Nexus academic database (more details provided in the appendix). This crude index is meant to provide a benchmark of comparison for the three other indices described above.

The Conference Board produces two other popular indices that we do not consider here: the index of coincident indicators (ICI) and the index of leading indicators (ILI). The reason is that the ICI contains four series (payroll employment, personal income, industrial production and manufacturing and trade sales) that were the focus of computations in Section 3. In the next section, we investigate the predictive power of some of the individual series contained in the ILI.

The evaluation is done with the data vintages most recently available since there is a limit to how far back real-time vintages are available for the indices that we consider. We do not think this is an important limitation - although data revisions can sometimes be dramatic for a single variable, these changes affect the indices to a much smaller degree. Moreover, Chauvet and Piger (2008) show that data revisions do not seem to affect the actual dating of business cycle turning points.

The results of this analysis are reported in Figures 3, 4 and 5, each of which contains two panels. The top panel displays the time-series of the index itself for the period for which it is available, whereas the bottom panel displays the ROC curve based on the BCDC-PE recession dates discussed in section 3. Table 4 provides more detailed results for $h \in \{0, 1, ..., 24\}$ for completeness. Both the CFNAI and ADS indices do very well at $h = 0$ with AUROC values of 0.93 and 0.95 respectively, and whose confidence interval nearly indicates perfect classification. Since the PMI focuses only on

one aspect of the economy – production – it may be unsurprising that the PMI describes BCDC turning points less well, though the contemporaneous AUROC value of 0.9 indicates that the PMI has significant classification ability.

The ADS index includes variables that are evaluated explicitly by the BCDC except for initial jobless claims (we cannot rule out that these data are also used in the BCDC's deliberations, of course). Interestingly, this particular variable is well-classified by the BCDC-PE dates, with an AUROL of 0.95, which is larger than any of the other variables the BCDC claims to be considering. Perhaps for this reason, it is less surprising that ADS scores a very high AUROC. Used in raw form (that is, without the benefit of a sophisticated econometric model), these indices would allow one to classify 95% of the recession periods correctly and generate a false positive rate below 10%. As a benchmark, compare these results with our home-made index displayed in Figure 1, which has an AUROC of 0.78. For a similar success in classifying recessions, our index would have to tolerate a considerably higher false positive rate (possibly as high as 50%).

The same success is absent if one considers the use of these indices for prediction purposes. Within a year all indices are no better than a coin-toss at picking up recessions from expansions, as Table 4 shows. The values of the AUROCs only worsen as one considers longer time horizons. For this reason, we switch our analysis toward classifiers that may be less useful contemporaneously but that may be more helpful over horizons about 12 months into the future.

5 Future Turning Points

The ability to make predictions about economic variables is a valuable skill; accurate forecasts of the future economic environment can be highly profitable to businesses or investors. Policy makers also require forecasts of a variety of economic variables in order to pursue sensible policy. Consequently, a class of models has been developed whose purpose is to predict not any particular economic variable, but instead the underlying state of the aggregate economy. In this section, we evaluate the ability of several different types of these models. We do so by treating the BCDC's business cycle dates as the 'gold standard' against which to compare predictions of recessions. In this section we will focus on the components of the ILI, and in particular the long-short yield spread of treasury securities.

5.1 Forecasting Recessions with the Yield Curve

The yield curve has long been recognized as a leading indicator, where an inversion of the yield curve portends recession in the economy. Ignoring the term premium, the expectations hypothesis states that the difference between the yield on current long-term bonds and current short-term bonds is equal to the difference between current short-term rates and expected future short-term rates. An inverted yield curve thus indicates that the market expects lower short-term rates in the future, implicitly an expectation of tenuous economic performance since low short-term rates are generally due to either quantitative easing from the monetary authority or reduced investment demand.

In this section, we evaluate how accurate are models based on the yield curve; which spread has the most predictive power; and at which horizons. In order to answer these questions, we construct a series of predictions of whether month $(t + h)$ will have been declared recessionary by the

BCDC. Following Rudebusch and Williams (2009), this prediction takes the form of the simple probit model:

$$P(NBER_{t+h}|I_t) = \Phi(\beta_0 + \beta_1 YS_{t-1}), \tag{2}$$

where Φ is the Normal cumulative distribution function, YS_{t-h-1} is the spread between short- and long-term rates, and $h \in (0, 24)$. We note that it is not necessary for our methods to estimate the probit model (since any monotone transformation of the yield spread has the same ROC) but we keep the Rudebusch and Williams (2009) to facilitate comparability. The baseline yield spread is the yield on a 3-month T-Bill less that of a 10-year T-Note. A forecaster therefore re-runs each of 25 probit regressions – one for each forecast horizon – for each month using the data available at each point in time. The information set at time t includes data on yields from Treasury securities (and later, the Federal Funds rate) from all previous periods. Note that this timing convention implies that when a forecaster is making a prediction for whether or not the current month will have been declared recessionary, she does not have information on the current month's yield curve. This is consistent with the fact that data on yields are monthly averages and the thought experiment that a forecaster is making forecasts on the 1st of the month about whether the current month will have been declared recessionary by the NBER.

Data on Treasury yields were collected from Global Financial Data and cover the period 1947:1 - 2009:6. With the exception of forecasts made with 6-month yields (see below), predictions using the yield curve are made out-of-sample beginning in January 1960, with a sample that expands as new data is received. Data for the effective Federal Funds rate are also monthly and come from the H15 release of the Federal Reserve Board.

Figure 6 displays the ROC values for the baseline yield curve model described above at forecast horizons that vary between 0 and 24 months ahead, as well as the corresponding 95% confidence intervals for AUROCs at each h. The model based on the 3-month - 10-year yield spread is most powerful at predicting recessions approximately 12 months into the future. Figure 7 displays the ROC curves associated to $h = 0$ and $h = 12$, respectively. Here we can see the dramatic improvement in the predictive ability of this basic model as we vary the forecast horizon. To understand the difference, notice that for $h = 0$ a 75% of correctly predicted recessions is associates with an astounding 68% false positive rate whereas for $h = 12$, the same 75% rate of correctly classified recessions generates less than a 25% false positive rate.

A natural question from the perspective of a forecaster is whether additional information from the yield curve can add to the predictive ability of the simple model described above. We address this question by re-running the model described by equation 2 above while varying the long rate used in the yield spread. Figure 8 displays AUROC plots over $h \in \{0, ..., 24\}$ from models that use the spread between the 3-month T-Bill and the 6-month T-Bill, as well as 1-, 2-, 5- and 10-year T-Notes.

No matter the definition of the yield spread, all models produce their most accurate forecasts at similar horizons - approximately 12 months ahead. Five- and ten-year spreads have essentially the same predictive power, and produce forecasts that dominate other spreads at all forecast horizons. The spread using the 6-month Treasury note is peculiar. However, data for yields on 6-month Treasuries are relatively sparse, as the six-month bill was introduced in 1982 (data on most of the other yields begin in 1947). Out-of-sample predictions for the model based on the 6-month T-Note

are made beginning in January 1992 (compared to 1960 for all other yield lengths), and the model based on the 6-month long rate includes only three recessions.

5.2 Including the Federal Funds rate

Wright (2006) showed that the inclusion of the Federal Funds Rate into the probit model *slightly* improves the model's ability to forecast recessions out-of-sample (in the sense of a lower RMSE). We therefore augment the Probit model above with the Fed Funds rate. In general, including the Fed Funds rate improves the model's fit, as its inclusion allows the model to distinguish between moves in the yield spread that are due to movement in the short rates versus long rates. This benefit may not be seen in summary statistics such as root mean squared error, since this statistic is an average of particular moments in the data. However, we may expect that the addition of the Fed Funds rate may improve particular *regions* of the ROC curve, even if the overall improvement (in terms of AUROC) is small. The augmented model takes the form:

$$P(NBER_{t+h}|I_t) = \Phi(\beta_0 + \beta_1 YS_{t-1} + \beta_2 FFR_{t-1}),$$

where again the yield spread is defined as the difference between a 3-month yield and either the 6-month, 1-, 2-, 5-, or 10-year yield.

Figure 9 shows the AUROC plots as a function of $h \in \{0, 24\}$ for the models developed above, but including the Federal Funds rate. The addition of the Federal Funds rate improves forecast ability for all models, particularly when forecasting at short forecast horizons. For example, at $h = 0$, the model that uses only the spread between three-month and 10-year yields produces an AUROC of

0.55. By contrast, using the same yield spread and incorporating the Federal Funds achieves an AUROC of 0.65. Figure 10 plots the ROC curves for these two models for $h = 0$. More dramatic, however, is the improvement in the models that use shorter-horizon yield spreads. The addition of the Fed Funds rate improves models that use the 2-year and 5-year yields to the degree that they rival the baseline model that uses the 10-year yield. Likewise, the model using the spread between the 3- and 6-month yields dramatically improves at all time horizons, with an AUROC of close to 0.95 at $h = 7$ but with an AUROC that is still above a respectable 0.85 for horizons $h = 10$ to 15 months out.

5.3 Probit with All Yield Spreads

A natural consequence of the previous discussion is to consider a model that includes all available information about the shape of the yield curve. For example, consider the model

$$\begin{aligned} P(NBER_{t+h}|I_t) &= \Phi(\beta_0 + \beta_1 FFR_{t-1} + \beta_2 YS^{1y}_{t-1} + \beta_3 YS^{2y}_{t-1} \\ &+ \beta_4 YS^{5y}_{t-1} + \beta_5 YS^{10y}_{t-1}), \end{aligned} \tag{3}$$

We did not include the 6-month yield because it would shorten the available sample considerably. Figure 11 displays the AUROC as a function of h over 24 periods along with 95% confidence intervals for individual entries. As is readily apparent from this figure, the model in expression (3) does not seem to outperform the models that include the Fed Funds rate and a longer-term yield spread alone: A model that includes only the Federal Funds rate and the yield spread between the 3-month yield and a long-term T-Note, such as the 5- or 10-year yield. contains as much predictive ability

as the model that includes information about the entire shape of the yield curve.

To summarize, out-of-sample forecasts using yield curve data from 1947-2009 indicate that simple models of the yield curve have significant predictive ability with respect to business cycle turning points and are surprisingly robust. The models based on the yield curve are appealing because of their simplicity, and the baseline model considered here may serve as a transparent baseline for comparison of more complex models that also predict business cycle turning points. At short time horizons, the model based on the difference between the 3- and 6-month yields outperforms other models, achieving AUROCs in the range of 0.70. When forecasting at longer time horizons, however, it is necessary to use longer spreads. Models using 5- or 10-year spreads perform equally well, and are most prescient at time horizons approximately 12 months into the future. The addition of the Federal Funds rate generally improves these models, but only when the models are used to forecast at very short time horizons.

6 Discussion

This paper shows that the BCDC of the NBER produces a chronology of business cycle turning points with very high classification ability, by which we mean that the empirical distributions resulting from the expansion/recession classification, are easily separable in a statistical sense. However, we were surprised to find that classification ability is highest, not for the dates selected by the BCDC, but for dates that shift by approximately 3-months on average for output indicators but by as much as six to seven months for employment and income indicators. Of several definitions of turning points (two based on the BCDC dates, two based on popular, mechanical but misguided rules, and two based on

hidden Markov models) the BCDC's classification scores very highly. Overall these results should simultaneously boost confidence on the BCDC dating and suggests avenues for improvement.

A historical repository of the business cycle chronology is useful for researchers but investors, entrepreneurs, households and policy-makers are concerned about the current and future state of the economic cycle. Our aim here was not to introduce more refined statistical models of prediction but to evaluate the predictive content of several candidate indicators. In real-time, we found that the indices maintained by the Federal Reserve Banks of Chicago and Philadelphia (the CFNAI and ADS indices respectively) provide very accurate, real-time signals about the state of the business cycle but have little information about future turning points. Instead, we found that a relatively crude model based on a probit that includes the federal funds rate and a three month - ten year yield spread on treasury securities, has surprisingly strong classification ability 12 to 15 months out.

All of these results are based on non-parametric techniques that are relatively new in economics. The advantages of our approach are several. First, we provided a formal statistical assessment of the BCDC's classification ability, which to our knowledge is in itself a novel contribution. Second, the techniques we used are independent of loss function and are valid for strictly monotone transformations of the indices that we considered. Thus in one fell swoop, our results are general and encompass a wide class of possible specifications that could be considered in the literature.

We think our paper opens several new and interesting research questions. For example, two classifiers can have similar AUROC but different ROC curves, that is, one classifier can do better in some regions than the other and vice versa. Hence, the outer-envelope ROC curve constructed from these two classifiers can have a higher AUROC that each taken individually. Thus, we speculate that there

may be some improvements that investigators can pursue in trying to predict turning points.

Another interesting result was our finding about the phase shifts between the BCDC's dating and different economic indicators. This suggests that an improved classification could be obtained by optimally combining the differences in timing that different indicators display. Finally, because the ROC curve is explicit about the trade-offs between true positive and false positive rates, and to the extent that we view entering a recession period as undesirable and something to be prevented with activist economic policy, the ROC framework provides for a natural environment in which one can formally evaluate the trade-offs of applying policy when is needed versus when it is not.

7 Appendix

7.1 Data Sources and Calculations

This is a summary of the economic indicators, transformations and data sources provided in the appendix of the December 11, 2008 press release of the Business Cycle Dating Committee of the National Bureau of Economic Analysis and available from their website (www.nber.org).

Indicator	*Sample Available*	*Source and Method*
Industrial Production	1919:1 - 2009:6	FRB index B50001
Real Personal Income less transfers	1959:1 - 2009:5	BEA Table 2.6, line 1 less line 14, both deflated by a monthly interpolation (see below) of BEA Table 1.1.9 line 1
Payroll Employment	1939:1 - 2009:6	BLS Series CES0000000001
Household Employment	1948:1 - 2009:6	BLS Series LNS12000000
Real Manufacturing and Trade Sales	1997:1 - 2009:5	BEA Table 2BU, line 1
Real Gross Domestic Product	1947:I - 2009:II	BEA Table 1.1.6, line 1 (2009:II advanced estimate)
Real Gross Domestic Income	1947:I - 2009:I	BEA Table 1.10, line 1, divided by BEA Table 1.1.9, line 1

Websites:

- Federal Reserve Board industrial production index:
  `www.federalreserve.gov/releases/g17/iphist/iphist\_sa.txt`
- Bureau of Economic Analysis, U.S. Department of Commerce, all but sales:
  `www.bea.gov/national/nipaweb/SelectTable.asp?Selected=N`,
  sales: `www.bea.gov/national/nipaweb/nipa\_underlying/SelectTable.asp`

- BLS payroll survey: `http://data.bls.gov/cgi-bin/surveymost?ce`
- BLS household survey: `http://data.bls.gov/cgi-bin/surveymost?ln`

Interpolation of GDP deflator:

The value of the index in the first month of the quarter is one third of the past quarter's value plus two-thirds of the current quarter's value. In the second month, it is the quarter's value. In the third month, it is two-thirds of the quarter's value plus one third of the next quarter's value.

Treasury Yields and other Indices:

Indicator	*Sample Available*	*Source and Method*
Effective Federal Funds Rate	1954:7 - 2009:6	Federal Reserve Release H15
Six-month T-Note Yield	1982:I - 2009:6	Federal Reserve Release H15
Treasury Yields	1947:I - 2009:6	Global Financial Data
Aruba Diebold Scotti Index	1960:2 - 2009:6	Federal Reserve Bank of Philadelphia
Chicago Fed National Activity Index	1967:3 - 2009:6	Federal Reserve Bank of Chicago
Purchasing Managers Index	1948:1 - 2009:6	Institute for Supply Management
Chauvet-Hamilton Index	1967:11 - 2009:2	Chauvet and Hamilton (2005)
Chauvet-Piger Index	1967:2 - 2009:6	Chauvet and Piger (2008)

Websites:

- Global Financial Data: `http://www.globalfinancialdata.com`
- ADS Index: `http://www.philadelphiafed.org/research-and-data/real-time-center/business-conditions-index/`
- Chicago Fed Index: `http://www.chicagofed.org/economic_research_and_data/cfnai.cfm`
- Purchasing Managers Index: `http://www.ism.ws/`

- Chauvet-Hamilton Index: `http://www.econbrowser.com/archives/rec_ind/description.html`

- Chauvet-Piger Index: `http://www.uoregon.edu/~jpiger/us_recession_probs.htm`

The Lexis-Nexis News Index:

The index is a standardized count of the number of news items that appear in the Lexis-Nexis Academic database (see `http://www.lexisnexis.com/us/lnacademic`). In particular, the count is the number of news articles or news abstracts that Lexis-Nexis retrieves when searching for the word "recession" within "US Newspapers and Wires" source. Our database is at a monthly frequency, beginning in July 1970 and running through June 2009. Each monthly observation is the average daily count for all days within that month, which we then standardize by removing a time trend and adjusting for seasonal variation in the number of counts.

References

Bamber, D. (1975), 'The area above the ordinal dominance graph and the area below the receiver operating characteristic graph', *Journal of Mathematical Psychology* **12**.

Lusted, L. B. (1960), 'Logical analysis in roentgen diagnosis', *Radiology* **74**, 178–93.

Mann, H. & Whitney, D. (1947), 'On a test of whether one of two random variables is stochastically larger than the other', *Annals of Mathematical Statistics* .

Mason, I. B. (1982), *A Model for the Assessment of Weather Forecasts*, Australian Meterological Society.

Mason, S. & Graham, N. (2002), 'Areas beneath the relative operating characteristics and relative operating levels curves: Statistical significance and interpretation', *Quarterly Journal of the Royal Meterological Society* **128**.

Organization, W. M. (2000), *Standardized Verification System for Long-Range Forecasts*, World Metereological Organization.

Peirce, C. S. (1884), 'The numerical measure of the success of predictions', *Science* **4**, 428–441.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford, U.K.

Peterson, W. W. & Birdsall, T. G. (1953), The theory of signal detectability: Part i. the general theory, Technical Report 13, Electronic Defense Group.

Sheskin, D. (2000), *Handbook of Parametric and Non parametric Statistical Procedures*, Chapman Hall: Boca Raton.

Spackman, K. A. (1989), Signal detection theory: Valuable tools for evaluating inductive learning, Technical report, Proceedings of the Sixth International Workshop on Machine Learning.

Stanski, H. R., Wilson, L. J. & Burrows, W. R. (1989), Survey of common verification methods in metereology, Research report no. 89-5, Atmospheric Environment Service, Forecast Research Division.

Wilcoxon, F. (1945), 'Individual comparisons by ranking methods', *Biometrics Bulletin* **1**, 80–83.

Wright, J. H. (2006), The yield curve and predicting recessions, Technical report, Federal Reserve Board.

Youden, W. J. (1950), 'Index for rating diagnostic tests', *Cancer* .

Table 1 - Recession Summary Statistics

	NBER-PI	NBER-PE	R1	R2
Number of Recessions	11	11	26	10
Total Months	133	122	123	75
Average length (months)	12.1	11.1	4.7	7.5

Notes:

Table 2 – ROLs for Recession Indicators

Variable	Indicator					
	BCDC-PI	**BCDC-PE**	**R1**	**R2**	**CP Index**	**CH Index**
GDP	0.9834	0.9848	0.8775**	0.9798	0.9855	0.9801
	(0.0039)	(0.0036)	(0.0203)	(0.0045)	(0.0053)	(0.0063)
	3	3	4	4	4	3
GDI	0.9811	0.9824	0.8623**	0.9655*	0.9790	0.9696
	(0.0045)	(0.0043)	(0.0212)	(0.0071)	(0.0068)	(0.0072)
	4	3	4	4	3	3
PI	0.9595	0.9589	0.8836**	0.9331	0.9394	0.9334
	(0.0079)	(0.0077)	(0.0231)	(0.0135)	(0.0130)	(0.0109)
	6	6	8	7	4	6
IP	0.9665	0.9680	0.8483**	0.9165*	0.9789	0.9475
	(0.0074)	(0.0071)	(0.2189)	(0.0195)	(0.0059)	(0.0110)
	4	4	7	6	5	4
MTS	0.9658	0.9626	0.8834	0.9553	0.9799	0.9398
	(0.0077)	(0.0083)	(0.0234)	(0.0099)	(0.0065)	(0.0117)
	3	2	5	3	3	2
PE	0.9666	0.9665	0.8346**	0.9120**	0.9701	0.9412
	(0.0071)	(0.0074)	(0.0239)	(0.0189)	(0.0090)	(0.0115)
	7	6	8	7	7	7
HE	0.9444	0.9453	0.8368**	0.9511	0.9768	0.9318
	(0.1250)	(0.0111)	(0.0247)	(0.0122)	(0.0061)	(0.0144)
	6	7	8	6	7	7

* Indicates that AUROL is different from BCDC-PE at 90 percent confidence interval

** Indicates that AUROL is different from BCDC-PE at 95 percent confidence interval

Notes:

Table 3 - ROCs for BCDC/CH/CP

Variable	Indicator					
	BCDC-PE	**CP Index**	**CH Index**	**BCDC-PE**	**CP Index**	**CH Index**
GDP	0.9361	0.9300	0.9461	0.9848	0.9855	0.9801
	(00115)	(0.0213)	(0.0142)	(0.0036)	(0.0053)	(0.0063)
	--	--	--	3	4	3
GDI	0.9260	0.9274	0.9259	0.9824	0.9790	0.9696
	(0.0123)	(0.0233)	(0.0162)	(0.0043)	(0.0068)	(0.0072)
	--	--	--	3	3	3
PI	0.8703	0.8821	0.8425	0.9589	0.9394	0.9334
	(0.0169)	(0.0380)	(0.0219)	(0.0077)	(0.0130)	(0.0109)
	--	--	--	6	4	6
IP	0.8937	0.8892	0.8591	0.9680	0.9789	0.9475
	(0.0150)	(0.0240)	(0.0220)	(0.0071)	(0.0059)	(0.0110)
	--	--	--	4	5	4
MTS	0.9426	0.9452	0.9134	0.9626	0.9799	0.9398
	(0.0106)	0.0132)	(0.0142)	(0.0083)	(0.0065)	(0.0117)
	--	--	--	2	3	2
PE	0.8384	0.8541	0.7788	0.9665	0.9701	0.9412
	(0.0172)	(0.0235)	0.0275	(0.0074)	(0.0090)	(0.0115)
	--	--	--	6	7	7
HE	0.8042	0.8587	0.7645	0.9453	0.9768*	0.9318
	(0.0224)	(0.0270)	(0.0317)	(0.0111)	(0.0061)	(0.0144)
	--	--	--	7	7	7

* Indicates that AUROL is different from BCDC-PE at 90 percent confidence interval

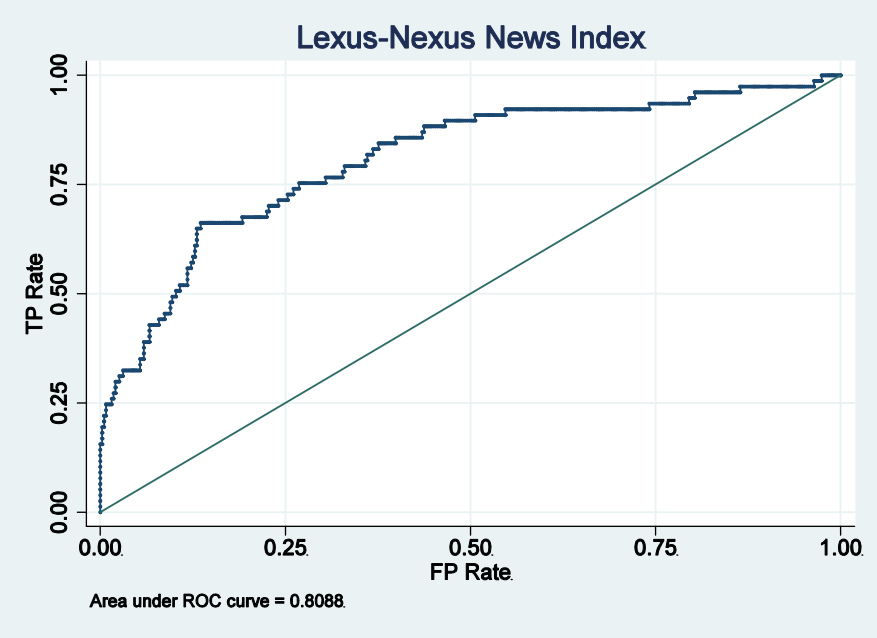
** Indicates that AUROL is different from BCDC-PE at 95 percent confidence interval

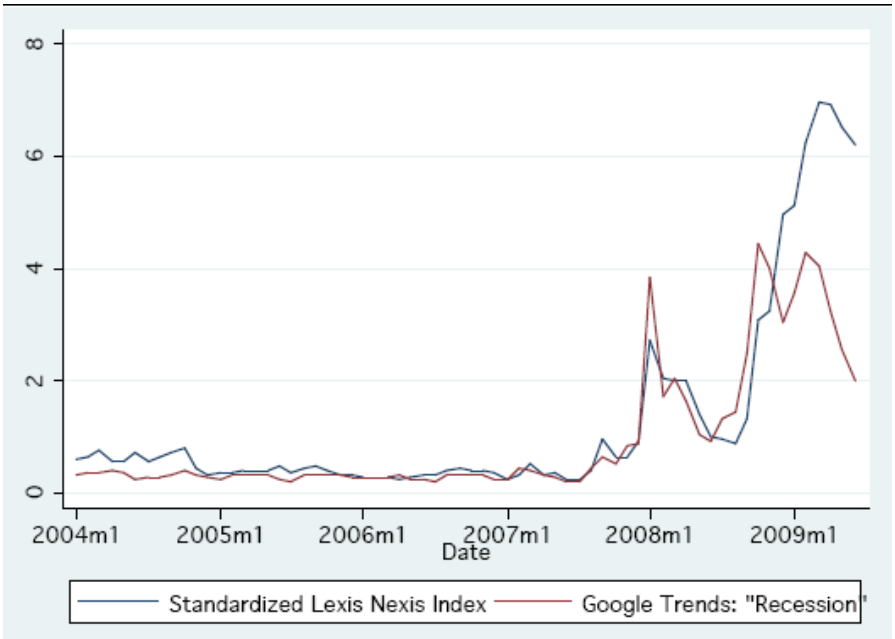
Table 4 – ROCs for CFNAI, ADS, PMI, and News Index

Model	Horizon						
	0	**4**	**8**	**12**	**16**	**20**	**24**
ADS	0.9507	0.8493	0.7310	0.5784	0.5135	0.4584	0.4397
	(0.0131)	(0.0206)	(0.0266)	(0.0318)	(0.0341)	(0.0336)	(0.0344)
NAI (MA-3)	0.9331	0.8171	0.6646	0.5283	0.4580	0.4000	0.3953
	(0.0172)	(0.0252)	(0.0322)	(0.0350)	(0.0368)	(0.0361)	(0.0346)
PMI	0.8875	0.7705	0.6237	0.4858	0.4504	0.4189	0.4407
	(0.0195)	(0.0246)	(0.0267)	(0.0275)	(0.0297)	(0.0304)	(0.0290)
LexisNexis	0.7765	0.6030	0.4785	0.3990	0.3285	0.2839	0.3177
	(0.0302)	(0.0347)	(0.0363)	(0.0372)	(0.0385)	(0.0394)	(0.0394)

Notes:

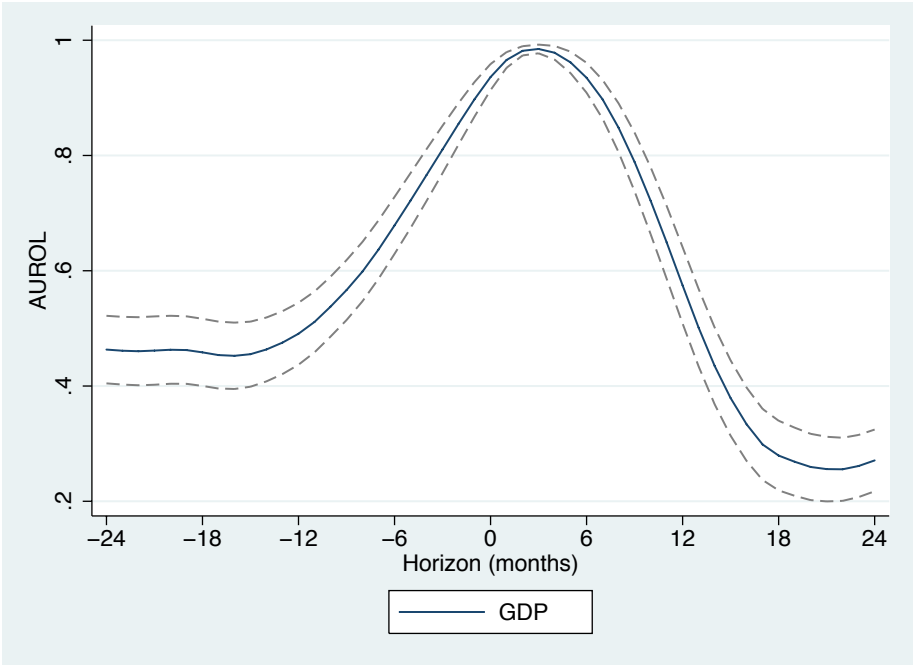
Figure 1 – The ROC curve for news items containing the word "Recession"

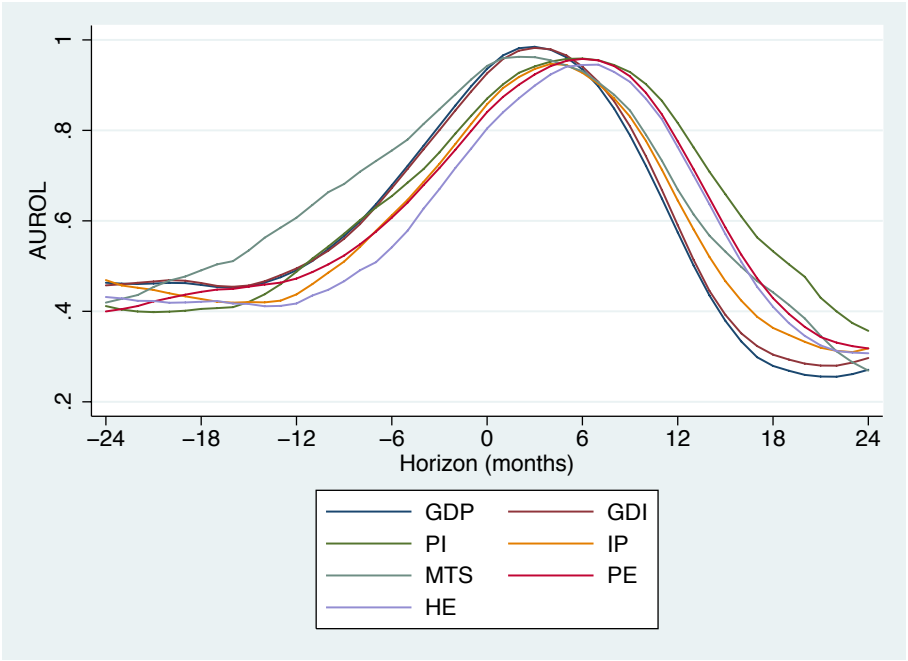




Note: See appendix for description of Lexis Nexis index.

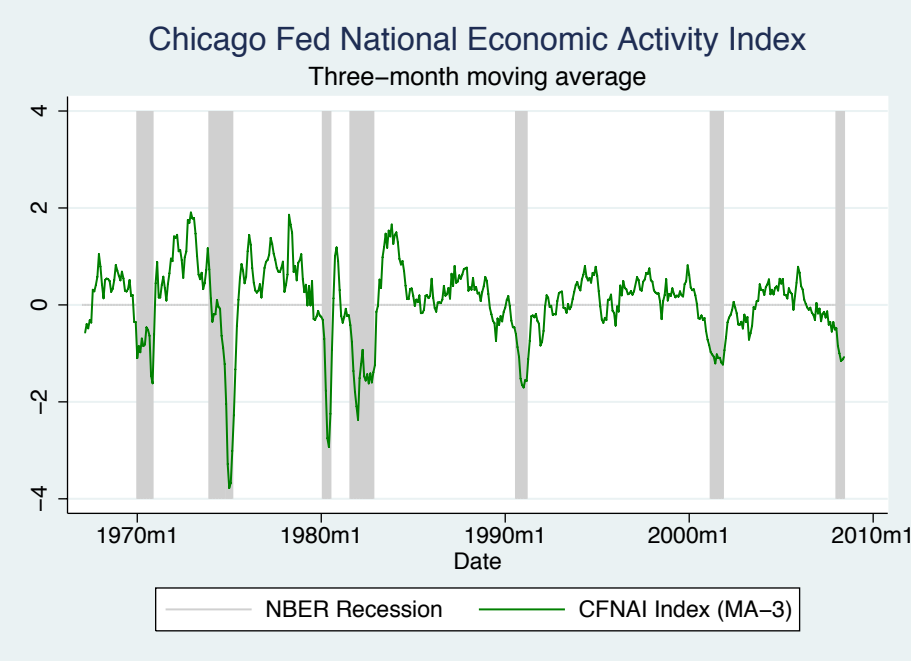
Figure 2 – The AUROC over time for BCDC economic indicators

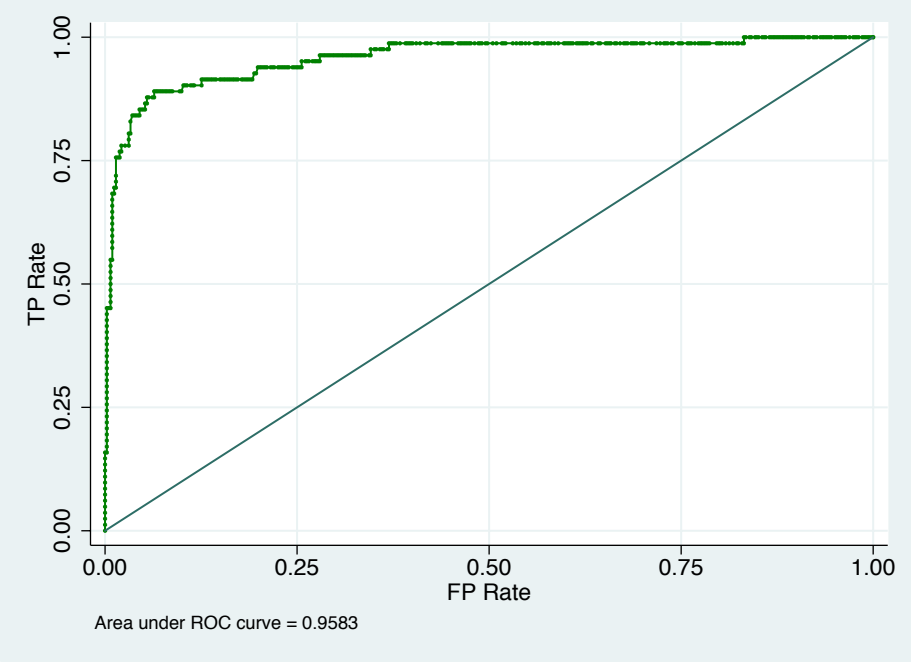




Notes:

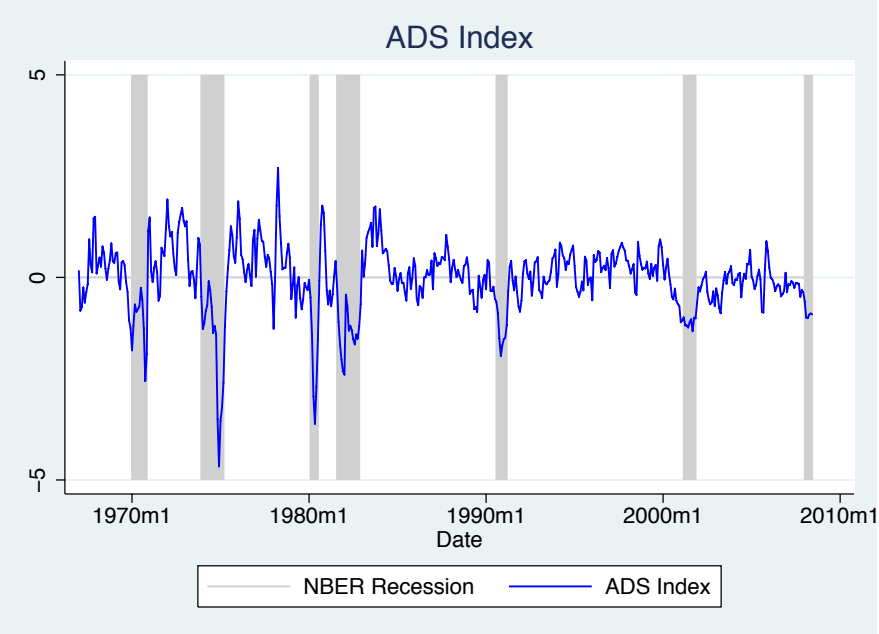
Figure 3 – The Chicago Fed National Activity Index and its ROC

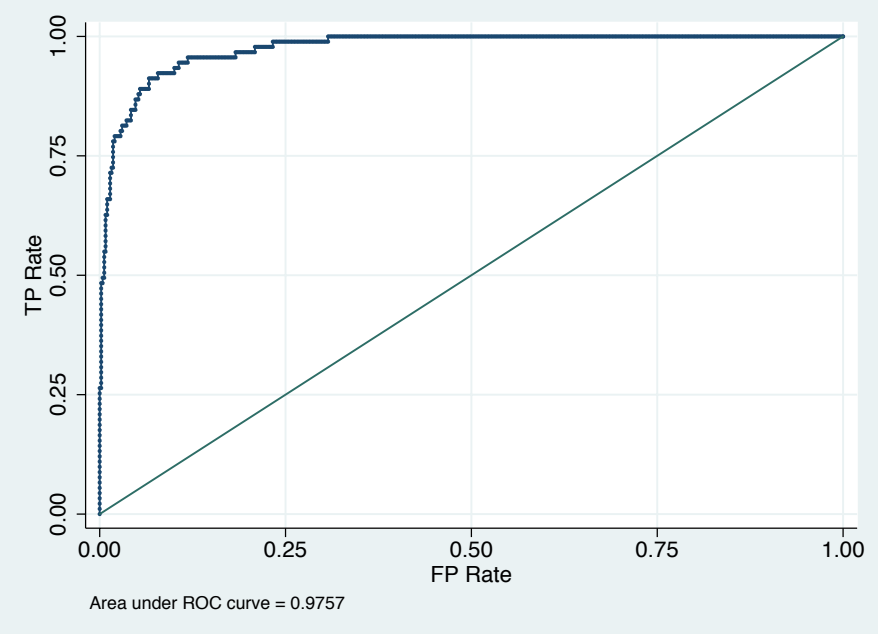




Notes:

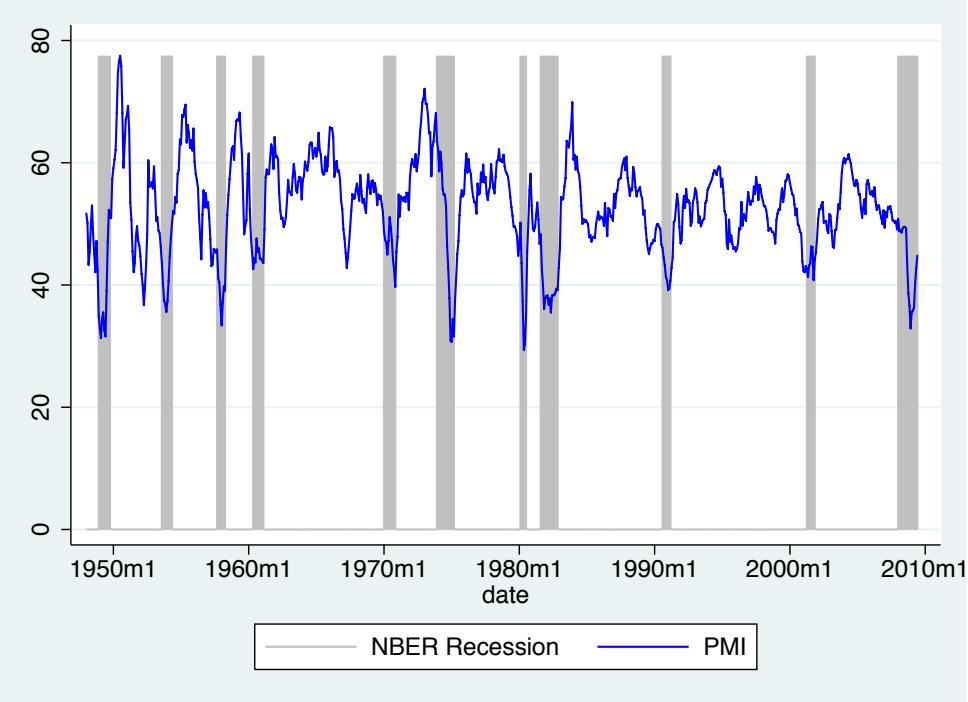
Figure 4 – The Arouba, Diebold and Scotti Index and its ROC





Notes:

Figure 5 – The Producer Managers Index and its ROC



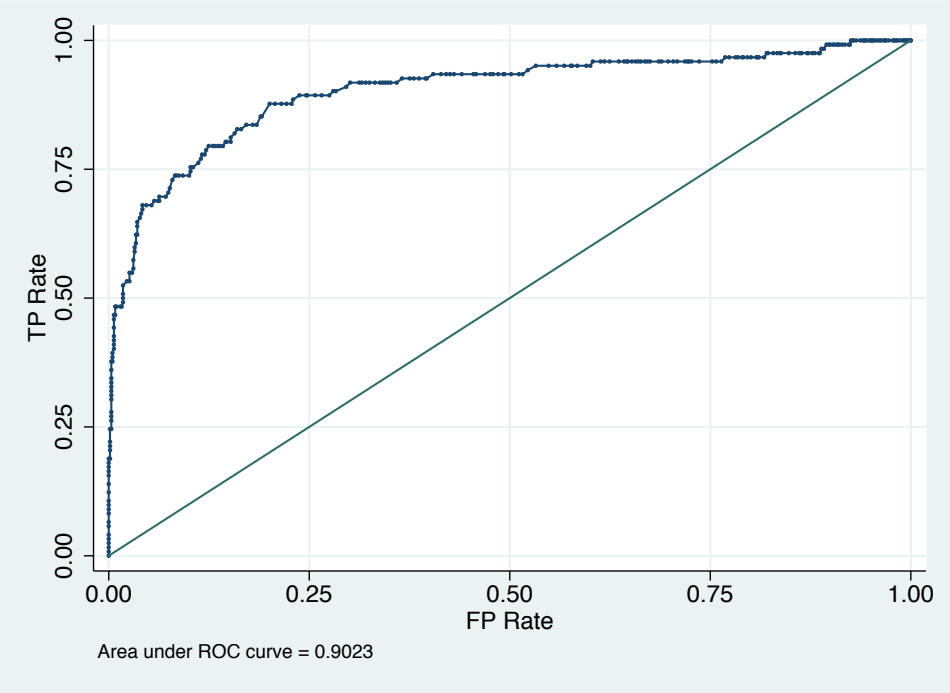
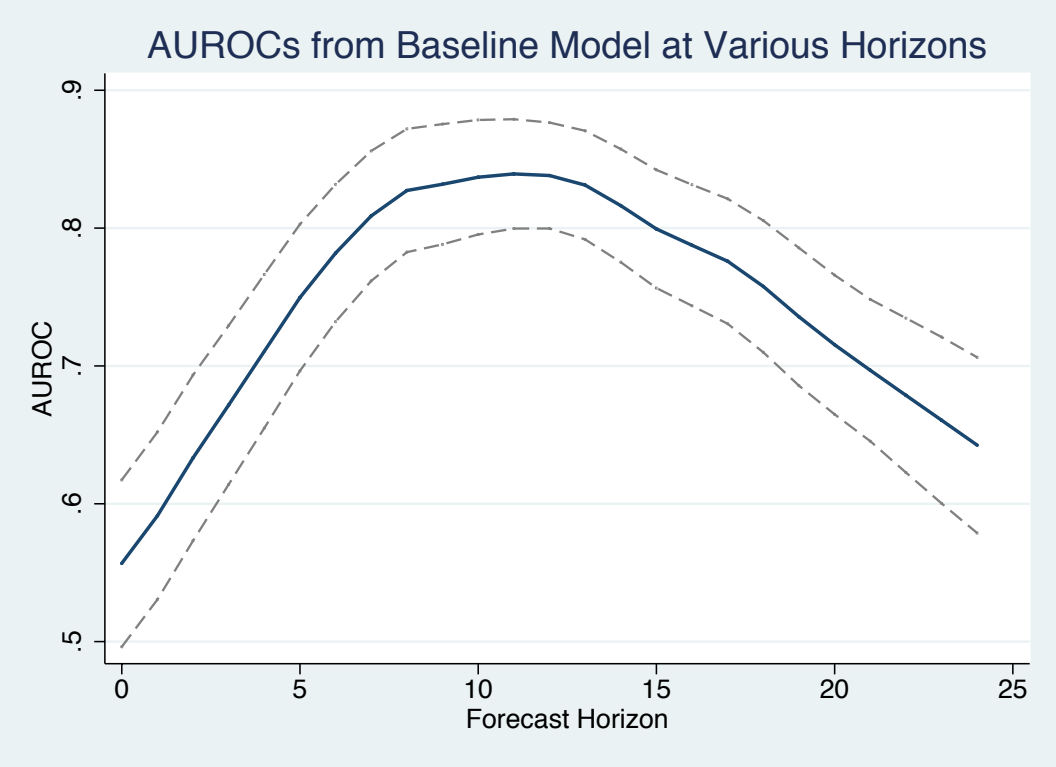
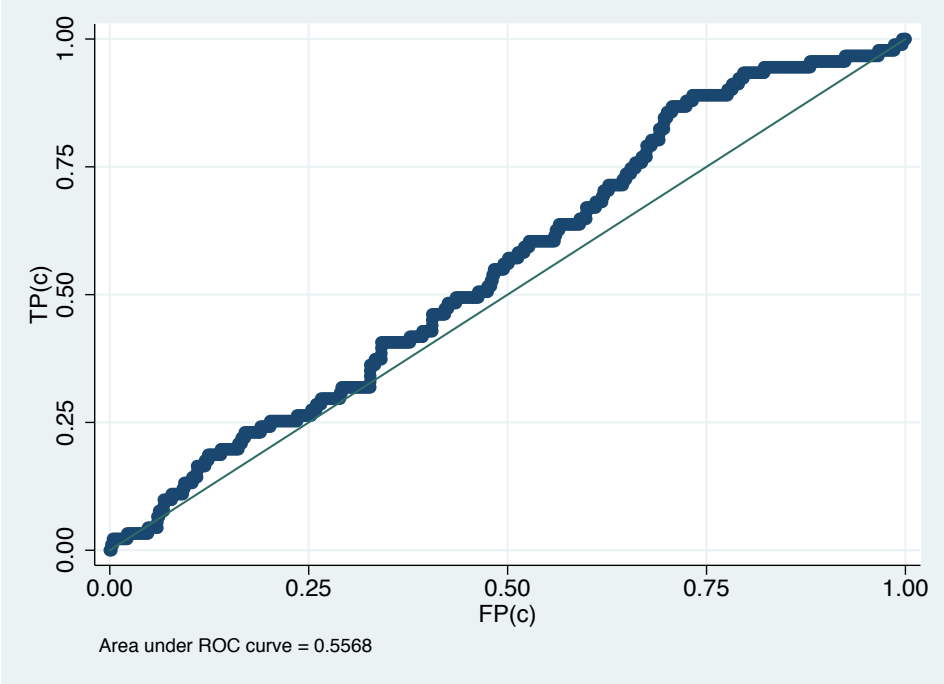


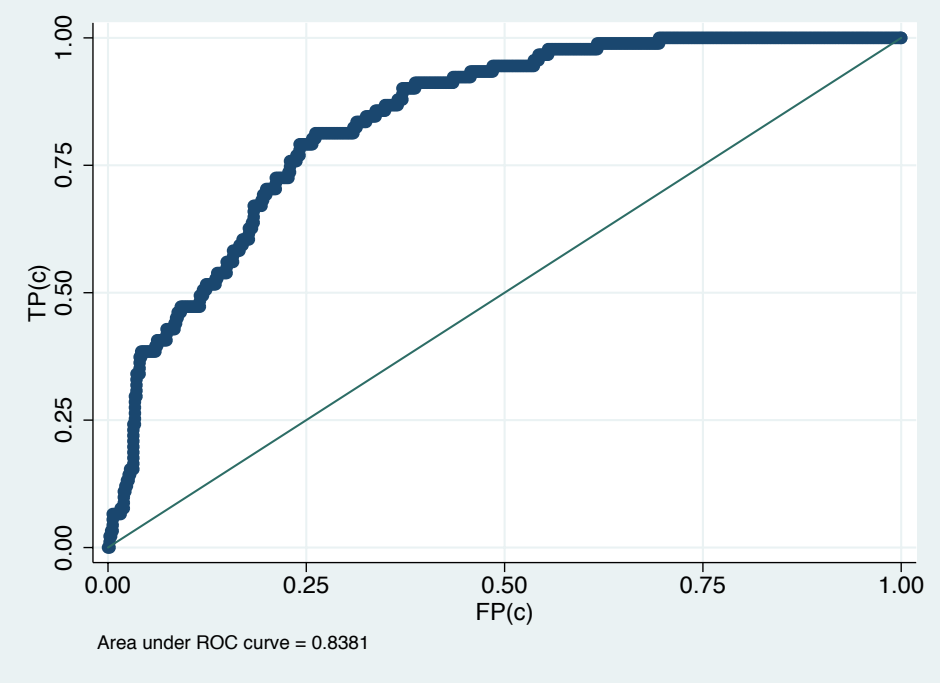
Figure 6 – ROCs for the baseline yield curve model at various horizons



Notes:

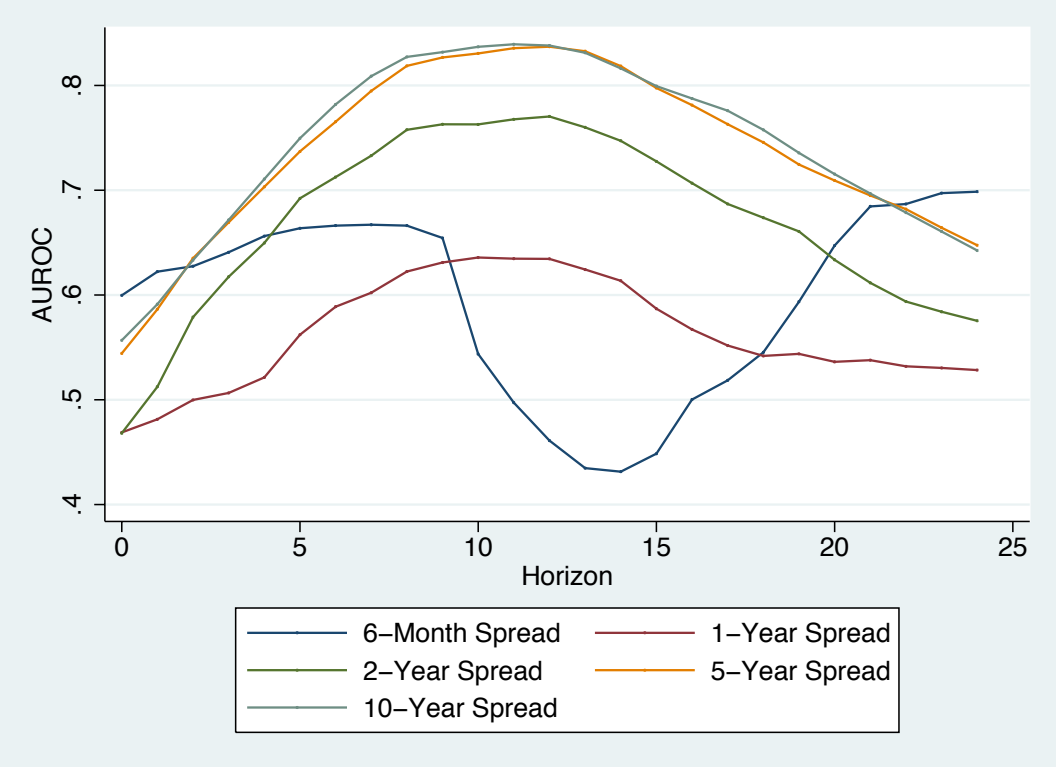
Figure 7 – ROC curves for the baseline yield curve model at horizons *h = 0* and *h = 12*





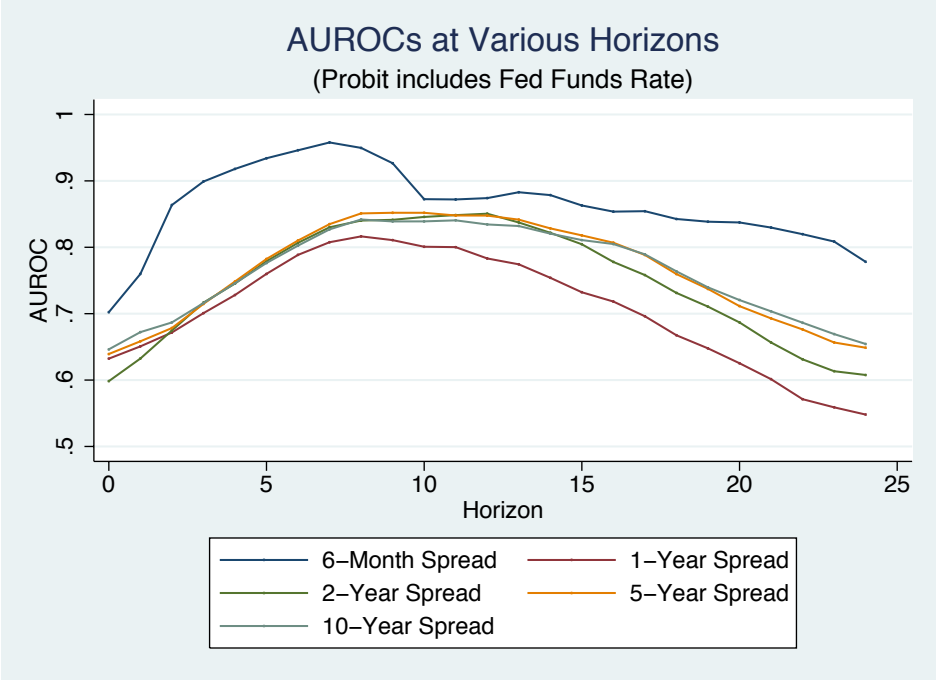
Notes:

Figure 8 – AUROC for different yield spreads



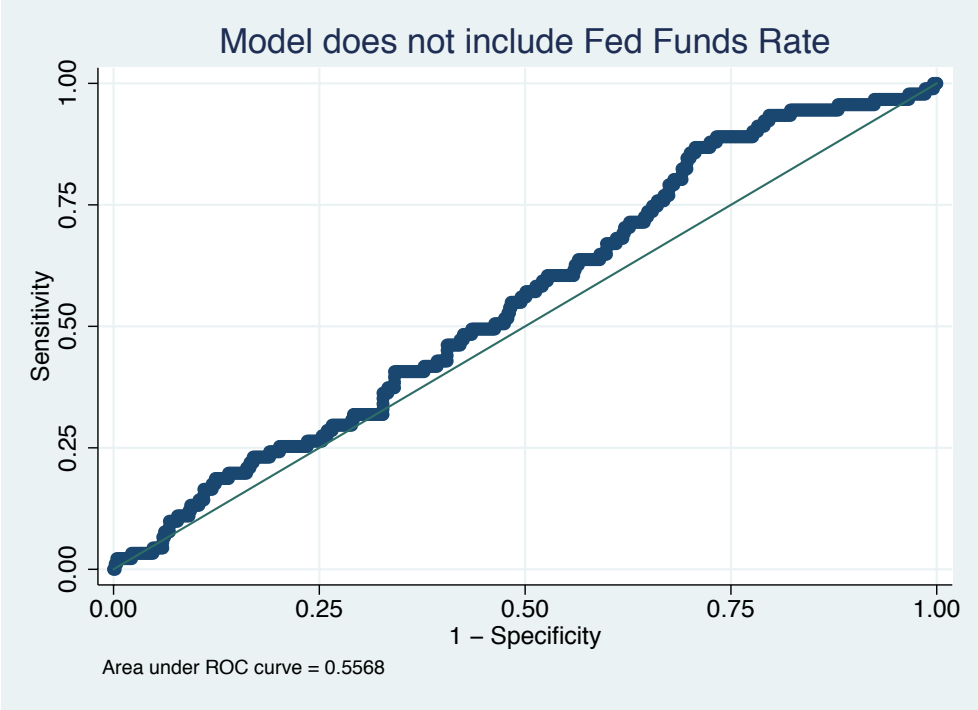
Notes:

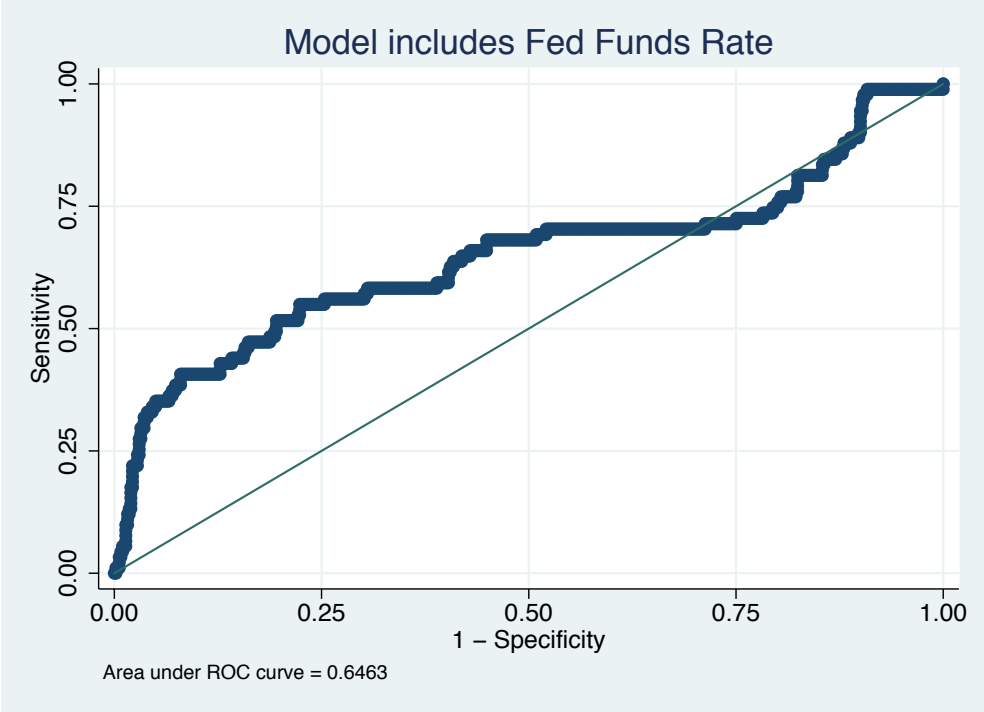
Figure 9 – AUROCs for various yield spread models when the Fed Funds rate is included



Notes:

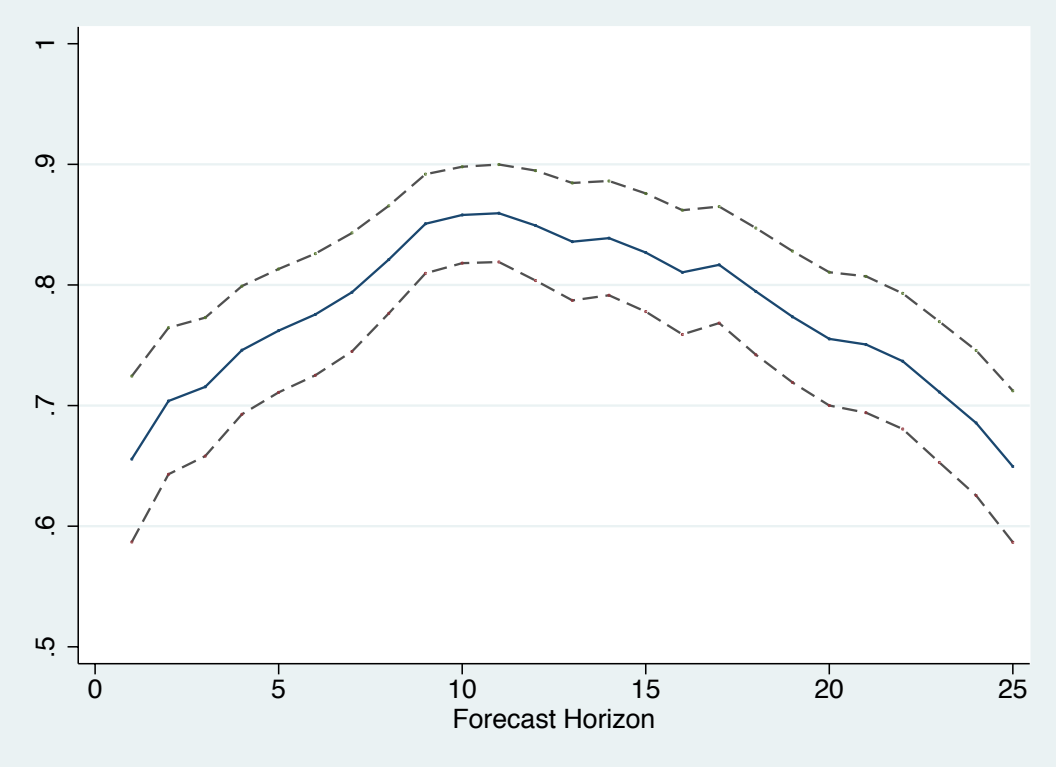
Figure 10 – AUROC for baseline model at *h = 0*: including and excluding the Fed Funds rate





Notes:

Figure 11: The AUROC over time when the Fed Funds rate is included



Notes: