November 2009 v.4

The Classification of Economic Activity *

Abstract

The Business Cycle Dating Committee (BCDC) of the National Bureau of Economic Research provides a historical chronology of business cycle turning points. This paper investigates three central aspects about this chronology: (1) How skillful is the BCDC in classifying economic activity into expansions and recessions? (2) Which indices of business conditions best capture the current but unobservable state of the business cycle? And (3) Which indicators predict future turning points best and at what horizons? We answer each of these questions in detail with methods novel to economics designed to assess classification ability. In the process we clarify several important features of business cycle phenomena.

- *JEL Codes:* E32, E37, C14
- *Keywords:* business cycle turning points, receiver operating characteristics (ROC) 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 Richard Dennis, John Fernald, Pierre Olivier Gourinchas, James Hamilton, Kevin Lansing, Jeremy Piger, Glenn Rudebusch, Alan Taylor, Hal Varian, and John Williams for useful comments and suggestions. Jordà is grateful for the hospitality of the Federal Reserve Bank of San Francisco during the preparation of this manuscript. Partial financial support was provided by the Spanish Ministerio de Ciencia e Innovación, grant SEJ-2007-63098.

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 asks three important questions about cyclical economic activity: (1) How accurate is the taxonomy of expansions and recessions implied by the peak and trough dates recorded by the BCDC? (2) Because the BCDC releases are retrospective, Which indicators best signal the current stage of the business cycle? And (3) Which indicators provide advance warning on future turning points? Using methods novel to economics (but common in many other sciences) —the receiver operating characteristics (ROC) curve— we find that economic activity is best classified by shifting the start and end dates of recessions by three months relative to the peak and trough dates reported by the BCDC. Employment indicators ought to be lagged an additional three to four months. Second, we find that the Aruoba, Diebold and Scotti (ADS) index of business conditions, maintained by the Federal Reserve Bank of

Philadelphia, and the Chicago Fed National Activity Index (CFNAI) provide accurate signals regarding the current state of the business cycle. Finally, optimal predictive classification varies across the components of the Conference Board's Index of Leading Indicators (ILI). Each index's classification ability depends on the forecast horizon in non-monotonic manner, an important finding that suggests that parsimonious affine specifications lack sufficient texture to take full advantage of the predictive information contained in the ILI. We provide out-of-sample evidence about direct predictive-classification ability up to 24 months into the future.

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 about when GDP growth is observed to be negative. The BCDC's definition of a recession[1] 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 1, 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

[1] `www.nber.org/cycles/`

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 as coming from a mixture of two distinct distributions, a feature that we take advantage of in our analysis. Moreover, information regarding a binomial variable describing aggregate economic activity is simple and easily understood by both policy-makers and the general public: policy-makers may prefer to craft policy responses with a probabilistic statement regarding the recession/expansion state of the economy than with a more uncertain point-estimate of aggregate growth in GDP.

The methods we use in this paper are new to economics, although their earliest origin perhaps traces back to Peirce's (1884) "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, as well as the receiver operating characteristic (ROC) curve introduced by Peterson and Birdsall (1953) in the field of radar signal detection theory. The ROC curve methodology was 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 been adopted into fields as diverse as the atmospheric sciences (see Mason, 1982 for an early reference, as well as Stanski, Wilson and Burrows, 1989; and the World Meteorological Organization, 2000) and machine learning (see Spackman, 1989 for an early discussion). Recent applications to economics include Jordà and Taylor (2009a, b).

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), all of which are rely on the specification of an underlying forecast loss function. A major contribution of our paper is to introduce a set of statistical tools based on ROC analysis that offer several advantages over these traditional measures. The ROC curve is independent of any forecast loss function, providing a non-parametric method for judging different potential classification indices. Strictly monotone transformations of the same prediction index have the same ROC curve: these new evaluation methods are not directly tied to modeling ability but to the information content of the indices themselves and automatically encompass a larger class of specifications – the main focus of this paper. Lastly, the new measures do not depend on the overall prevalence of recessions over the sample examined – this is important since recessions are observed only about 16 percent of the time. A rule that predicts every period to be an expansion will correctly predict expansions 84 percent of the time, a seemingly good number but such a rule is clearly useless to policy-makers trying to head-off recessions since the rule has a 100% error rate (it misses all the recessions). Our methods are set-up to explicitly recognize the policy trade-offs of these two error rates.

2 Classification Ability: The ROC Curve

The methods that we use in this paper will be unfamiliar to most economists. The economics tradition is that one proposes a statistical model from which to generate predictions about the state of the economy (expansion, recession). One then evaluates an indicator's performance using appropriate inferential procedures. The loss functions associated with this predictive evaluation may vary, but if the specification of the model is a correct representation of the

data generating process (DGP), one obtains unbiased estimates of the true model. However, when the statistical model is only an approximation, different loss functions result in different models and parameter estimates, and therefore possibly different conclusions about the usefulness of a particular economic indicator (see Hand and Vinciotti, 2003). The methods that we use here do not require that we construct specific models and hence, the decision problem is independent of the loss function one may consider. We now explain our approach in detail by discussing first how to evaluate indicators taking the BCDC's dating to be the true classification of business cycles before discussing the more nuanced question of how one can evaluate the BCDC's dating itself.

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, as we know). Meanwhile, consider the index Y_t, which we require only to be real-valued and ordinal. Y_t may denote a real-time probability prediction about S_t, a linear index, an index from a more complicated statistical model (e.g. a neural network estimator), or simply an observable variable (e.g. a leading indicator). The distinction is unnecessary for the methods we describe. Y_t, together with a the threshold c, define a binary prediction *recession* whenever $Y_t \geq c$, and *expansion* whenever $Y_t < c$.

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*, or *recall rate;* and $FP(c)$

is known as the *false positive rate,* or *(1-specificity).*

The ROC curve plots the entire set of possible combinations of $TP(c)$ and $FP(c)$ for $c \in (-\infty, \infty)$. As $c \to \infty$, $TP(c) = FP(c) = 0$. Conversely, when $c \to -\infty$, $TP(c) = FP(c) = 1$, so that 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 and is an entirely 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 hug the north-west border of the positive unit quadrant. Most applications generate ROC curves between these two extremes. Thus, since the abscissa is $FP(c)$ and c uniquely determines $TP(c)$, it is customary to represent the ROC curve with the Cartesian convention $\{ROC(r), r\}_{r=0}^{1}$ where $ROC(r) = TP(c)$ and $r = FP(c)$.

As an illustration, Figure 1 displays the ROC curve for an index of business conditions that we constructed. The index is based on the number of news items with the word "recession" appearing in the LexisNexis database every month.[2] The ROC curve displayed in the top panel of Figure 1 articulates the relative trade-offs in predicting recessions and expansions accurately. For example, correctly classifying 90% of all recessions results in a high rate of false positives (expansions incorrectly coded as recessions): 50%. By predicting recessions slightly less accurately (say 75%), the false positive rate would be cut in half to 25%. For completeness, the bottom panel of Figure 1 displays our index and the Google

[2] The index takes the raw counts of incidences per month, and adjusts for the trend in the number of news outlets included in the LexisNexis database over time and for seasonality. This index is similar in spirit to what Google Trends (visit www.google.com/trends) does to track the incidence of, e.g., influenza throughout the year. By tracking search activity on influenza related word 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. We provide a more detailed description in the appendix.

Trends index for the word recession over the longest sample available for Google Trends.

In general, there may be different benefits and costs associated with making accurate predictions and errors and hence the overall utility of the classification can be expressed as (see Baker and Kramer, 2007):

$$\begin{aligned} U &= U_{11}ROC(r)\pi + U_{01}(1 - ROC(r))\pi + \\ &\quad U_{10}r(1-\pi) + U_{00}(1-r)(1-\pi) \end{aligned} \tag{1}$$

where U_{ij} is the utility (or disutility) associated with the prediction i given that the true state is j, $i, j \in \{0, 1\}$ and π is the unconditional probability of observing a recession in the sample. It is easy to see that utility is maximized when

$$U_{11}\frac{\partial ROC}{\partial r}\pi - U_{01}\frac{\partial ROC}{\partial r}\pi + U_{10}(1-\pi) - U_{00}(1-\pi) = 0$$

or rearranging

$$\frac{\partial ROC}{\partial r} = \frac{U_{00} - U_{10}}{U_{11} - U_{01}}\frac{(1-\pi)}{\pi} \tag{2}$$

that is, that point where the slope of the ROC curve equals the expected marginal rate of substitution between the net utility of accurate expansion and recession prediction.

Underlying the classification problem is the view that the observations of Y_t reflect a mixture of two distributions. Specifically, let Z_t denote the observations of Y_t for which $S_t = 1$, with probability density function (*pdf*) given by f, and cumulative probability distribution (*cdf*) given by F. Similarly, let X_t denote the observations of Y_t for which $S_t = 0$ and with *pdf* given by g and *cdf* given by G. Then, the ROC curve can also be seen as a plot of $ROC(r) = 1 - G(F^{-1}(1-r))$ versus r, $r \in [0, 1]$, so that the slope of the ROC curve in

(1) is

$$\frac{\partial ROC}{\partial r} = \frac{g(F^{-1}(1-r))}{f(F^{-1}(1-r))}$$

that is, the slope of the ROC curve is the likelihood ratio between f and g. Hence, expression (1) relates the likelihood ratio between the expansion and recession distributions and the expected marginal relative utility from correct classification.

Given U_{ij}, $i, j \in \{0, 1\}$, one can therefore determine the *optimal operating point* as the threshold c^* that meets the equilibrium condition (2). Under the assumption $U_{ii} = 1$ and $U_{ij} = -1$ and $\pi = 0.5$, the optimal operating point maximizes the distance between $TP(c)$ and $FP(c)$, which is the well-known Kolmogorov-Smirnov statistic (Kolmogorov, 1933; Smirnov, 1939). Clearly the assumption $\pi = 0.5$ is violated for our data and we do not know the values of U_{ij} that the BCDC uses.

A summary of all the trade-offs contained in the ROC curve and a commonly used measure of overall classification ability is the area under the ROC curve ($AUROC$) :

$$AUROC = \int_0^1 ROC(r)dr; \qquad AUROC \in [0.5, 1], \tag{3}$$

where it is clear that a perfect classifier has $AUROC = 1$ whereas a coin-toss classifier has $AUROC = 0.5$.

The $AUROC$ has several other convenient statistical interpretations. Green and Swets (1966) show that $AUROC = P[Z > X]$. Therefore, a simple, non-parametric estimate of (3) is:

$$\widehat{AUROC} = \frac{1}{n_0 n_1} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \left\{ I(Z_i > X_j) + \frac{1}{2} I\left(Z_i = X_j\right) \right\} \tag{4}$$

where $I(A)$ is the indicator function and is equal to 1 when A is true, 0 otherwise, and $n_k, k = 0, 1$ indicates the number of observations for the kth state. The last term in (4)

is a tie-breaking rule typically rarely needed in large samples. Bamber (1975) and Hanley and McNeil (1982) show that $\widehat{AUROC}$ is a two-sample, rank-sum statistic that can be reconfigured and reinterpreted as a Wilcoxon-Mann-Whitney U-statistic (Mann and Whitney, 1947 and Wilcoxon, 1945). Using the theory of U-statistics, Hsieh and Turnbull (1996) show that under mild regularity conditions (described in detail in their paper):

$$\sqrt{n_1}\left(\widehat{AUROC} - P[Z > X]\right) \xrightarrow{d} N(0, \sigma^2)$$

$$\sigma^2 = \frac{1}{n_0 n_1} AUROC(1 - AUROC) + (n_1 - 1)(Q_1 - AUROC^2) + (n_0 - 1)(Q_2 - AUROC^2)$$

$$Q_1 = \frac{AUROC}{2 - AUROC}; Q_2 = \frac{2AUROC^2}{1 + AUROC}.$$

For more details on the formulas for the variance see, e.g. Hanley and McNeil (1982), Obuchowski (1994), and Greiner, Pfeiffer and Smith (2000). The asymptotic normality result is very convenient because many hypothesis tests can be articulated using the familiar Wald principle (e.g. see Pepe, 2003). Bootstrap procedures are also available (see. e.g. Obuchowski and Lieber, 1998) although large sample approximations have been found to do well even in relatively small samples (again, see Pepe, 2003).

ROC curve methods provide formal assessment of classification ability: given the classifier Y_t, how well can it separate the classes associated with the true underlying states $S_t \in \{0, 1\}$. A non-parametric estimate of the $AUROC$ is easy to compute and its asymptotic distribution is Gaussian under general conditions so that inference against the null of no classification ability ($H_0 : AUROC = 0.5$) or comparisons of classification ability across classifiers, are straightforward (see Jordà and Taylor, 2009b for a detailed survey on other ROC-based testing procedures). The $AUROC$ is a two-sample, rank-sum statistic that compares the f and g densities implicit in the mixture distribution of Y generated by S for the basic problem

of evaluating

$$P[Y_t \geq c | S_t = 1].$$

In the next section we consider a different evaluation problem: if Y_t is generated by an unobserved mixture process, we want to know whether the BCDC dates properly classify the data into each component of the mixture. What makes this evaluation problem difficult is that the true state of the business cycle is not directly observable.

2.1 Evaluating the BCDC's Dating

The BCDC dating has been taken by the profession and the public as the final word on the historical chronology of cyclical turning points. However, since the BCDC does not provide a mathematical or statistical algorithmic procedure that can be directly and formally evaluated, it is difficult to form a judgment about its quality.[3] Here we propose a possible solution to this problem.

We begin by taking the view that economic activity can be approximately represented by a mixture model so that an observation Y_t of e.g. GDP, could have come from a density f that characterizes recessions, or a density g that characterizes expansions. Moreover, it is natural to expect that the more extreme an observation (say an observation of 10% GDP growth) the more likely it is that it belongs to one or the other distribution (i.e. 10% GDP growth is more likely to belong to the expansion distribution g than say, 2% GDP growth is). For illustrative purposes, Figure 2 displays a kernel density estimate of the two distributions implied by the BCDC dating. For reasons that will become clear momentarily, we note that

[3] Specifically, the latest public release of December 1, 2008 states that "Although the indicators described above are the most important measures considered by the NBER in developing its business cycle chronology, there is no fixed rule about which other measures may contribute information to the process in any particular episode."

we shift the BCDC definition one quarter forward. The mean of the recession distribution is located at -0.3% annual GDP growth whereas the expansion distribution is centered at 4.2% (although both have a similar standard deviation of 1.8-1.9%). However, even though there is considerable overlap of the two distributions, we will show that BCDC classification has very high skill.

Therefore, think of the BCDC's dating as a filtered probability prediction $\widehat{S}_t$ of the unobservable, underlying class marker S_t. If $\widehat{S}_t$ were generated by a fair coin-toss, the resulting f and g densities of the mixture for Y would be identical to a null model in which Y is assumed to come from a non-mixture process. The $AUROC$ for this coin-toss classifier would be 0.5, the typical null. Instead, the more skill in the construction of $\widehat{S}_t$, the clearer the distinction between the implied mixture distributions and in fact, perfect classification will generate $AUROC = 1$. Because a considerable portion of the paper consists in evaluating potential classifiers of the true state of the economy based on the BCDC dates, we begin our empirical analysis by first assessing the skill of the BCDC itself against the coin-toss null but also against alternative dating schemes based on two specifications of Hamilton's (1989) well known hidden Markov mixture model.

3 Assessing the Business Cycle Dating Committee

We begin this section by investigating four different definitions of recession, and hence the optimal dating of recessions and expansions from the trough and peak dates provided by the BCDC. In the analysis we focus directly on the yearly growth rates of the indicators of economic activity the BCDC publicly acknowledges to use. Our focus on yearly growth rates implicitly assumes a constant growth trend so it is reasonable to consider the effect

of different detrending methods on our analysis. Therefore, next we compare cyclical GDP detrended with three alternative methods. We conclude the section by comparing the BCDC directly against the classification generated from two popular specifications of Hamilton's (1989) hidden Markov mixture model.

3.1 Four Definitions of Recession

The BCDC produces a series of business cycle turning points for the U.S. economy that contains the month within which the day of a peak or a trough of economic activity occurs (see the BCDC's release of December 1, 2008). Each peak and trough month is therefore some 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. The BCDC itself (see `www.nber.org/cycle/`), Chauvet and Hamilton (2005) and Wright (2006) define BCDC-defined recessions as the period between a peak and a trough, including both the peak and trough months. We denote the series produced by this method *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, we consider two alternative and popular "rule-of-thumb" definitions of recessions. One mechanical definition classifies recessions as any period in which GDP growth is negative. The other, quite popular with the media, calls a recession when there are at least two consecutive quarters of negative GDP growth. Following Rudebusch and Williams (2009) we will call the two series resulting from these rules *R1* and *R2*, respectively.

Table 1 tabulates the salient features of each of these four classification methods. 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 (since there are 11 recessions in our sample), so that the economy is in recession 18 percent of the time. The average NBER-defined recession lasts about 12 months. The series $R1$ is very noisy relative to the other rules since recessionary periods are on average much shorter but occur 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 (momentarily entertaining that the BCDC is truth). Of the incorrect signals, there are 36 false positives (that is, expansion months misclassified as recessionary) and 46 false negatives. 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 (or 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.

Next, we use the ROC curve analysis to compare the classification skill of each of these four definitions relative to each of the coincident indicators mentioned in the BCDC's release of December 1, 2008. In addition and because there could be phase shifts across indicators (for example, it is well-known that employment tends to lag considerably in economic recoveries), we examine a window h of ± 24 months around turning points. At each horizon h, we calculate the corresponding AUROC and denote the horizon at which the maximum AUROC occurs as h^*.

The BCDC claims to base 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). It also considers two quarterly indicators: real gross domestic product (GDP) and real gross domestic income (GDI). Consequently, we constructed these indicators with data obtained directly from the sources listed by the BCDC (more details are provided in the appendix). To allow for direct comparison between the quarterly and monthly indicators, we construct monthly interpolated series of GDP and GDI using the linear interpolation method described by the BCDC. We then take the 12-month log differences to compute year-over-year growth rates, which are less noisy than the annualized monthly growth rates. The next section compares the implicit constant growth assumption of this procedure with other detrending methods commonly used in macroeconomics. However, focusing on growth rates here is clearly least controversial.

Table 2 reports the AUROC estimates for h^*. Both BCDC series are clearly superior to the mechanical $R1$ and $R2$ series, with the exception of household employment, where R2 outperforms the two BCDC-based datings, although the difference is not statistically significant. For most variables, the AUROCs for the $R1$ and $R2$ series are statistically inferior to the BCDC-PE series, which we take henceforth as our benchmark since it achieves the highest AUROCs overall.

Figure 3 displays the AUROCs associated with the BCDC-PE classification for the monthly indicators. The top panel of Figure 3 displays the AUROC of GDP against h, $h \in \{-24, ..., 0, ...24\}$, along with the lower and upper bounds of the 95% confidence interval associated to the AUROC at each h as an illustration, whereas the bottom panel of Figure 3 displays the AUROCs associated with each indicator against h, although in the interest of

legibility we suppress the 95% confidence bands.

Both Table 2 and Figure 3 reveal that the AUROC for GDP is maximized at $h = 3$ (rather than at $h = 0$), with an AUROC that is virtually equal to 1 (for GDI $h^* = 4$). We do not think this three-month shift is due to the interpolation of the quarterly data into monthly since the AUROC is maximized at $h = 3$ for MTS and $h = 4$ for IP (both monthly). Moreover, the AUROC is maximized at $h = 6$ for HE and PI and at $h = 7$ for PE. These results conform well with the well-worn observation that the recovery in employment tends to lag the end of recessions.

We find the phase shift results intriguing and novel, and upon closer inspection of Figure 2, indeed sensible. The explanation is that during the first few months after a trough of economic activity, the first few observations thereafter are better classified as still coming from the recession distribution (a similar effect happens a few months after a peak but with the expansion distribution instead). The effect is naturally more pronounced for employment.

3.2 Trends and Cycles

In a growing economy, classification of economic activity into expansions and recessions refers to its cyclical component – broadly speaking, the behavior of the economy around its secular trend. In a stable economy like the U.S., it seems uncontroversial to examine the yearly growth rates of the set of coincident economic indicators used by the BCDC. This method implicitly assumes a constant growth path and does not require specific modelling of the trend process. However in macroeconomics it is common to investigate business cycle phenomena by applying some filtering method to the raw data in the levels. We find this problematic for several reasons: (1) there is no consensus about the appropriate trend-cycle

decomposition; (2) filtered trend estimates are sensitive to the sample used and may vary as the sample grows over time; (3) trends across indicators are likely to differ; and (4) common filtering methods often introduce additional and unwanted dynamic elements into the cyclical component.

However and for the sake of completeness, in this section we examine the classification skill of the BCDC-PE rule for output deviations from a Hodrick and Prescott (1997) trend (HP); a Baxter and King (1999) trend (BK) where the cycle is defined over frequencies between 6 to 32 quarters; and from estimates of potential output reported by the Congressional Budget Office (CBO). Figure 4 displays the growth rates of these trends to get a sense of their variation over the business cycle.

Several results deserve comment. First notice that the maintained constant growth assumption implicit in our earlier calculations would be displayed as a flat line in Figure 4 and is therefore omitted for clarity. Second, there is very strong conformity across the three trends, with some slight differences between CBO on one side, and HP and BK on the other: for example, in the 2001 recession, HP and BK trend output are declining whereas CBO potential output is increasing. Third, some of the time trend output grows during recessions and some of the time it declines although recessions tend to coincide with periods in which trend output is low.

Table 3 reports AUROC estimates for the deviations of GDP from these three trends and using the BCDC-PE rule and for $h = 0$ and h^* by allowing h to vary between ± 24. For $h = 0$, AUROC values decline considerably and in a statistically significant manner with respect to results reported in Table 2. Interestingly, the results for the optimal h are rather

better and all three detrending methods coincide in choosing $h^* = 6$ as the optimal horizon. This apparently surprising result, however, is consistent with what we see in Figure 4. When the trend is allowed to vary over time, trough dates tend to coincide with simultaneous improvements of trend and cycle. Therefore, the deviations of output from trend tend to persist a while longer and get classified as belonging to the recession distribution for a few periods more than when the trend is fixed, resulting in a shift of h^* from 3 to 6 months.

3.3 The BCDC versus Statistical Dating Rules

The BCDC's dating of business cycles is held as the universally accepted gold standard against which competing methods of turning point prediction are evaluated. Even models in which the underlying state of the economy can be estimated independently of the BCDC's classification (such as the class of hidden Markov mixture models spawned by Hamilton's 1989 seminal work) evaluate their success when estimates of the smoothed state probabilities line up against the BCDC's peak-trough dates. This section turns this view point on its head and instead asks how well the BCDC dates compare to smoothed state probabilities available in Chauvet and Hamilton (2005) and Chauvet and Piger (2008).

We first consider an interpolated version of Chauvet and Hamilton's (2005) quarterly Markov-switching smoothed transition state probability index (henceforth the CH index), which is readily available (see: `www.econbrowser.com/archives/rec_ind/description.html`) and transparent. The dependent variable is GDP growth and the two-state Markov chain specified in the model captures the underlying unobserved state of whether the economy is in expansion or recession. In order to translate the transition state probabilities into monthly zero-one indicator about the state of the economy, we interpolate the quarterly index lin-

early and then apply the simple rule-of-thumb that any period with a recession probability greater than a given threshold value is classified as a recession. In order to find the optimal threshold, we performed a grid-search over the space 0.5-0.9 and found 0.75 to maximize the AUROC for the majority of indicators analyzed here.

Chauvet and Piger (2008) produce a similar index (available from Jeremy Piger's homepage at `www.uoregon.edu/~jpiger/` and which we will denote CP). However, CP first 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 μ is assumed to follow the process $\mu = \mu_0 + \mu_1 S_t$, where S_t is an unobserved latent variable about the state of the economy. Estimation of the model produces an estimate of the probability that the economy is in recession. CP use a two-step process to then translate this probability into a binomial variable. First, CP record when does the estimated probability become greater than or equal to 0.80 for three consecutive months. These dates are classified as recession. Let the first month of this series be month t. Then the beginning of the recession is dated as the first month prior to month t for which the probability of recession is greater than 0.5.

Table 4 displays the AUROCs associated with each of these three recession indicators and for each of the coincident indicators examined by the BCDC. Columns 1-3 show the contemporaneous AUROCs, while columns 4-6 display the maximum AUROC values when the horizon h is allowed to vary over ± 24. Broadly speaking, the results in Table 4 suggest that the BCDC-PE rule does a very good job in classifying economic data relative to the

two statistical procedures examined here, which are meant to search for the optimal mixture in the sample. The AUROCs associated with the BCDC-PE dates, both contemporaneously and at h^*, are statistically indistinguishable from the AUROCs that result from CP and CH.

The results are interesting, specially since the CP index combines information from several of the coincident indicators used by the BCDC and optimally allocates the data into the two distributions in the mixture. However, the BCDC's dating process appears to classify a broad range of variables while sacrificing very little by means of misclassification for any individual series. Except for household employment, there is no statistical basis to suggest that the two statistical procedures outperform the BCDC for any of the coincident indicators considered.

4 Indices of Business Conditions

In contrast to section 3, in this section we take the BCDC chronology to be the historical record of the true 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: the current state of the business cycle.

We investigate three indices of aggregate activity plus the news-based indicator introduced in Section 2 to act as a benchmark. These are indices commonly used in the profession and are freely and publicly available. Two of the indices represent state-of-the-art approaches to measuring aggregate economic activity in real time. The Chicago Fed National Activity Index (CFNAI) is 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 introduced

in Stock and Watson (1999). More details can be found in their paper and in the Federal Reserve Bank of Chicago's website.[4] The second index included is the Aruoba, Diebold and Scotti (ADS) Business Conditions Index maintained by the Federal Reserve Bank of Philadelphia. The ADS index is a new 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.[5]

The other two indices we investigate rely on information from market participants instead of attempting to measure economic activity directly. 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[6]). 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" in the LexisNexis 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.

We evaluate these indices with the most recently available data vintage since real-time vintages are not available for a long enough period. We do not think this is an important limitation - although data revisions can sometimes be considerable for a single variable (such

[4] www.chicagofed.org/economic_research_and_data/cfnai.cfm

[5] www.philadelphiafed.org/research-and-data/real-time-center/business-conditions-index/

[6] www.ism.ws/ISMreport/content.cfm?ItemNumber=10752&navItemNumber=12961

as GDP), 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 5 (for CFNAI), 6 (for ADS) and 7 (for PMI), each of which contains two panels. The top panel displays the ROC curve (using the BCDC-PE recession dates discussed in Section 3) and the bottom panel the time series for the index. Table 5 provides more detailed results for various values of h. 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 intervals include near-perfect classification ability. The PMI index has an $\widehat{AUROC} = 0.9$, which is somewhat lower but PMI is a narrow indicator for production rather than a broad based measure such as CFNAI and ADS. As a benchmark, our LexisNexis index has an $\widehat{AUROC} = 0.81$, which is statistically inferior to any of the three indices considered.

A more detailed investigation into the indices themselves revealed that out of the variables included to construct the ADS index, initial jobless claims alone has an $\widehat{AUROC} = 0.95$, which is considerably higher than any of the other variables and approximately the value attained by the ADS index itself. One way to interpret an $\widehat{AUROC} = 0.95$ is to notice that this implies, for example, 95% correct classification of recessions with a false positive rate below 10% (or in Neyman-Pearson nomenclature, the classifier is as effective as a test with 0.95 power at a 90% confidence level). For reference value, the LexisNexis index would generate a 50% false positive rate for the same 95% correct classification rate of recessions.

We conclude this section with two observations. First, the classification ability of all the indices considered deteriorates very rapidly when used to predict turning points into the

future: within a year, they are no better than a coin-toss at distinguishing recessions from expansions. As we will see in the next section, the components of the ILI go a long way to remedy this situation. Second, we considered those threshold values that would maximize the utility of the classification so as to check the values recommended by the different agencies that publish these data. Specifically, assume that the benefits of hits equals the costs of misses in magnitude, then the optimal threshold can be determined from expression (1) as:

$$\max_{c} \left(2\widehat{\pi}\widehat{TP}(c) - \widehat{\pi}\right) - \left(2\left(1-\widehat{\pi}\right)\widehat{FP}(c) - \left(1-\widehat{\pi}\right)\right)$$

The resulting estimates of the optimal thresholds are for CFNAI, $c^* = -0.82$; ADS $c^* = -0.80$; and $PMI = 44.74$, which are somewhat lower than the values commonly used as rules of thumb, specifically for CFNAI and ADS, $c^{\bigstar} = 0$; and for PMI, $c^{\bigstar} = 50$. Of course, these estimates would vary under different assumptions about the relative utility of classification hits and misses.

5 Future Turning Points

The last of the three main questions we set out to investigate in this paper considers the ability to predict future business cycle turning points. In this section we focus on the components of the Conference Board's Index of Leading Indicators (ILI), a complete description of which is provided in the appendix. Throughout this section we maintain the working convention that the BCDC's chronology is the "gold standard" that these predictions should try to properly classify. Within this section, we accomplish two tasks. First we use ROC analysis to determine the relative classification ability of each individual component of the ILI over horizons ranging from 0 to 24 months in advance. Interestingly, we find considerable

variation in classification ability across predictors depending on the forecast horizon considered and more importantly, we find that at some horizons, positive values of the predictor are associated with higher likelihood of recession, whereas at other horizons the association is with the negative values of the predictors instead. This non-monotonicity is revealing because it suggests that parsimonious affine models will often lack sufficient texture to generate accurate predictions of the economic cycle, even a few periods into the future. Thus, the second task we carry out is a direct prediction-classification exercise and out-of-sample evaluation over several horizons to determine the best methods of business cycle turning point prediction.

5.1 The Conference Board Index of Leading Indicators: ROC Analysis

The Conference Board's Index of Leading Indicators includes ten individual components (see appendix for data sources and description). Several of these variables are meant to capture market or consumer expectations about future economic activity—for example, the S&P 500 and the Treasury debt yield spread both speak to market expectations, while the University of Michigan consumer survey directly measures household expectations. The remaining variables—building permits for new housing units, average weekly hours in manufacturing, manufacturers' new orders, initial claims for unemployment insurance, and the index of supplies deliveries — are more direct measures or precursors of future economic activity.

Figure 8 displays the AUROCs across horizons $h \in (0, 36)$ for all ten leading indicators used by the Conference Board. In the interest of readability, we group the indicators into two panels and suppress confidence intervals. Many indicators achieve AUROC maxima at horizons very close to $h = 0$. Interestingly, however, these indicators then achieve minima

at horizons between 12 and 18 months into the future. We pause here to clarify that an $AUROC < 0.5$ means that it is the negative of the classifier considered that has classification ability, consequently many of the indicators appear to have valuable information to forecast recessions at distant horizons as long as one flips the sign of the index. Several indicators—the S&P 500, permits for new housing units, consumer expectations, new orders for consumer goods, and initial claims—achieve their highest AUROCs contemporaneously, although the inverse of these series generally have modest but detectable predictive abilities at longer time horizons.

Table 6 presents the AUROCs corresponding to different forecast horizons. The second column presents the value $(1 - AUROC)$ so as to present classifier ability in terms of the familiar values above 0.5, its standard error, and horizon where the indicator's AUROC achieves its minimum (or the maximum AUROC for the negative of the indicator). The third column corresponds to the contemporaneous AUROC value for that indicator, while forth column corresponds to the maximum value of the AUROC, as well as the associated forecast horizon.

As an example, consider the results for the indicator initial claims for unemployment insurance. Initial unemployment claims attains the highest AUROC (0.96) of any of the Conference Board's leading indicators and at a horizon of $h^* = 1$ (visible in the bottom panel of Figure 8). The indicator's AUROC dips below 0.5 at $h > 12$ months, and at $h = 24$, initial claims achieves its minimum value of 0.28—or changing the sign on the index, a value of 0.72—indicating that although the index is clearly not as powerful at $h = 24$ as it is contemporaneously, initial claims contains significant predictive ability at

horizons approximately two years into the future. A number of the components of the ILI display a similar behavior suggesting that parsimonious affine specifications may be insufficient to capture the non-monotonic classification behavior and therefore that no single linear combination of the components of the ILI is adequate to forecast at all horizons. The next section uses direct classification methods and an out-of-sample evaluation with ROC analysis to investigate the predictive ability of the ILI

5.2 Forecasting Business Cycle Turning Points

Let w_t denote the vector of components of the ILI and let $S_t \in \{0, 1\}$ denote the state variables implied by the BCDC-PE dates. In this section we are interested in modeling the posterior probabilities $P[S_{t+h} = s|w_t]$ for $h \in \{0, ..., 24\}$ (we include $h = 0$ as a nowcast) More specifically, we assume the log-odds ratio at time h is a linear function of w_t, so that

$$\log \frac{P[S_{t+h} = 0|w_t]}{P[S_{t+h} = 1|w_t]} = \beta_{h0} + \beta_h' w_t; \qquad h \in \{0, ..., 24\}$$

which results in the well-known logistic model. The parameters of this model can be easily maximized with standard techniques by maximum likelihood or iterated least squares. Moreover, this is a popular model for classification in biostatistics. In fact, linear discriminant analysis (LDA), a standard classification algorithm, consists of the logistic regression we propose and a marginal model for w_t. Hastie, Tibshirani and Friedman (2009) however argue that the logistic model may be a safer choice than LDA. Since most economists are familiar with logistic regression but not necessarily with LDA, we prefer to take the safer route.

The prediction problem over more than one horizon into the future can be done in one of two ways: by specifying the one period ahead model and iterating forward as needed, or by

estimating a specific model for each forecast horizon. We prefer to take the latter approach for several reasons. First, the iterative approach would require us to specify a model for w_t that we could use to iterate as well. Second, the specification of the conditional model would have to be sufficiently parametrically intensive to capture the non-monotonicities that we uncovered in the previous section. Third, the nonlinearity of the logistic model would require simulation techniques to construct forecasts beyond one period ahead. This would complicate the out-of-sample computations we are about to describe.

The classification-prediction exercise uses a rolling window of fixed width that is used as a training sample. The first window begins January, 1968 and ends December, 1977 (120 observations). With this training sample we generate a set of forecasts for $h = 0, ..., 24$ and then roll the training sample by one month and repeat. We use the collection of out-of-sample classification-predictions to calculate the per-horizon AUROCs that are displayed in Figure 9. The figure shows that the ILI begins with nearly perfect classification ability at $h = 0$ (which is not surprising since in section 3 we discovered that initial claims of unemployment can generate an AUROC of about 0.96) which gradually deteriorates as the forecast horizon increases. However, over the first year, classification ability remains very high, with AUROCs around 0.9. A more steady decline occurs after month 10 or 11 so that two years out we still do better than a coin-toss but not by a very large margin.

6 Discussion

Cyclical fluctuations of economic activity have long been categorized into expansions and recessions in implicit recognition that the economy evolves differently in each state. Policy-makers may not be as concerned with momentary lapses into economic weakness as they

may be with full transitions into the recessionary state. This paper offers fresh views on the problem of classifying economic activity into expansions and recessions.

This paper makes several contributions. To our knowledge, we are the first to directly provide a measure of the quality of the chronology of business cycles provided by the BCDC. This is important because we are able to provide researchers and the public with some assurance that the chronology has considerable classification value, even when compared to statistical models tailored to optimize how the data should be categorized. Furthermore, our analysis yields further insight into the timing of transitions: maximum classification accuracy of economic activity could be achieved by shifting the beginning and end of recessions by three-to-four months, with employment shifted by an additional three-to-four months.

In order to design an effective policy response one must determine what is the current state of the economy and when are future transitions expected to occur. Business conditions indices maintained by the Federal Reserve Bank's of Chicago and Philadelphia provide very accurate signals in real time. Prediction-classification up to horizons of one year is also fairly accurate (with an AUROC close to 0.9 throughout) but quickly tapering off thereafter, although significantly better than a coin toss even two years out. Here a novel observation is that no single linear combination of the components of the ILI is likely to work well since we have uncovered strong variation across horizons and in the manner in which the components help classify future turning points.

Understanding the difference between classification ability and model fit is important. As an illustration, think of least squares: model fit improves when the Euclidean distance between an observation and the regression line is made small, regardless of the sign of the

regression error. Therefore, extreme observations tend to drive the slope of the regression line. However, in a classification scenario the sign of the regression error is much more important – extreme events are easily assigned to the correct class but it is much more difficult to assign observations in the neighbourhood of the regression line. Tilting of the regression line can therefore result in better fit and worse classification. Statistical methods tailored for classification, such as linear of quadratic discriminant analysis, neural networks, support vector machines and boosting algorithms will undoubtedly become more commonplace in economics, ours being a modest contribution in this new direction.

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 third 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.

Indices

Indicator	*Sample Available*	*Source and Method*
Chauvet-Hamilton Index	1967:11 - 2009:2	Chauvet and Hamilton (2005)
Chauvet-Piger Index	1967:2 - 2009:6	Chauvet and Piger (2008)
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

Websites:

- Chauvet-Hamilton Index: `www.econbrowser.com/archives/rec_ind/description.html`
- Chauvet-Piger Index: `www.uoregon.edu/~jpiger/us_recession_probs.htm`
- ADS Index: 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: `www.ism.ws`

Conference Board Index of Leading Indicators

Indicator	*Sample Available*
Average weekly hours, manufacturing	1939:1 - 2009:6
Average weekly initial claims for unemployment insurance	1967:1 - 2009:6
Building permits, new private housing units	1960:1 - 2009:6
Index of supplier deliveries—vendor performance	1948:1 - 2009:6
Interest rate spread, 10-year Treasury bonds less federal funds rate	1954:8 - 2009:6
Manufacturer's new orders, consumer goods and materials	1959:1 - 2009:6
Manufacturer's new orders, nondefense capital goods	1959:1 - 2009:6
Money supply, M2	1959:1 - 2009:6
Stock prices, S&P 500	1921:1 - 2009:6
University of Michigan index of consumer expectations	1959:11 - 2009:6

The LexisNexis News Index:

The index is a standardized count of the number of news items that appear in the LexisNexis Academic database (see `www.lexisnexis.com/us/lnacademic`). In particular, the count is the number of news articles or news abstracts that LexisNexis 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

Aruoba, S. Borağan, Francis X. Diebold and Chiara Scotti (2009) "Real-Time Measurement of Business Conditions," *Journal of Business and Economic Statistics,* 27(4): 417-427.

Baker, Stuart G. and Barnett S. Kramer (2007) "Peirce, Youden, and Receiver Operating Characteristic Curves," *The American Statistician,* 61(4): 343-346.

Bamber, D. (1975) "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characterisitc Graph," *Journal of Mathematical Psychology,* 12: 387-415.

Baxter, Marianne and Robert G. King (1999) "Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series," *Review of Ecomomics and Statistics,* 81: 575-593.

Burns, Arthur F. and Wesley C. Mitchell (1946) **Measuring Business Cycles.** NBER Book Series Studies in Business Cycles N. 2. New York: NBER.

Business Cycle Dating Committee of the National Bureau of Economic Research, December 1, 2008 press release available at: http://www.nber.org/cycles/dec2008.html.

Chauvet, Marcelle and James D. Hamilton (2005) "Dating Business Cycle Turning Points," *National Bureau of Economic Research,* working paper 11422.

Green, David M. and John A. Swets (1966) **Signal Detection Theory and Psychophysics.** Peninsula Publishing: Los Altos, California, USA.

Greiner, Matthias, Dirk Pfeiffer and Ronald D. Smith (2000) "Principles and Practical Application of the Receiver Operating Characteristic Analysis for Diagnostic Tests," *Preventive Veterinary Medicine,* 45: 23-41.

Hamilton, James D. (1989) "A New Approach to the Economic Analysis of Nonstationary Time Series Subject to Changes in Regime," *Econometrica,* 57(2): 357-384.

Hand, David J. and Veronica Vinciotti (2003) "Local versus Global Models for Classification Problems: Fitting Models Where It Matters," *The American Statistician*, 57(2): 124-131.

Hanley, James A. and Barbara J. McNeil (1982) "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology,* 143: 29-36.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009) **The Elements of Statistical Learning.** Springer: New York, New York, USA.

Hodrick, Robert and Edward C. Prescott (1997) "Postwar U.S. Business Cycles: An Empirical Investigation," *Journal of Money, Credit and Banking, 29*: 1-16.

Hsieh, Fushing and Bruce W. Turnbull (1996) "Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve," *Annals of Statistics*, 24: 25-40.

Jordà, Òscar and Alan M. Taylor (2009a) "The Carry Trade and Fundamentals: Nothing to Fear but FEER Itself," U.C. Davis, *mimeo.*

Jordà, Òscar and Alan M. Taylor (2009b) "Investment Performance of Directional Trading Strategies," U.C. Davis, *mimeo.*

Kolmogorov, Andrey N. (1933) "Sulla Determinazione Empirica di una Legge di Distribuzione," *Giornale dell'Istituto Italiano degli Attuari 4: 83-91.*

Lusted, Lee B. (1960) "Logical Analysis in Roentgen Diagnosis," *Radiology*, 74: 178-93.

Mann, H. B. and D. R. Whitney (1947) "On a Test of Whether One of Two Radom Variables is Stochastically Larger than the Other," *Annals of Mathematical Statistics*, 18: 50-60.

Mason, Ian B. (1982) "A Model for the Assessment of Weather Forecasts," *Australian Meterological Society*, 30: 291-303.

Obuchowski, Nancy A. (1994) "Computing Sample Size for Receiver Operating Characteristic Curve Studies," *Investigative Radiology,* 29(2): 238-243.

Obuchowski, Nancy A. and Michael L. Lieber (1998) "Confidence Intervals for the Receiver Operating Characteristic Area in Studies with Small Samples," *Academic Radiology,* 5(8): 561-571.

Peirce, Charles S. (1884) "The Numerical Measure of the Success of Predictions," *Science*, 4, 428-441.

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

Peterson W. Wesley and Theodore G. Birdsall (1953) "The Theory of Signal Detectability: Part I. The General Theory," Electronic Defense Group, Technical Report 13, June 1953. Available from EECS Systems Office, University of Michigan.

Rudebusch, Glenn D. and John C. Williams (2009) "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve," *Journal of Business and Economic Statistics,* 27(4): 492-503.

Smirnov, Nikolai V. (1939) "Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples (in Russian)," *Bulletin Moscow University,* 2: 3-16.

Spackman, Kent A. (1989) "Signal Detetection Theory: Valuable Tools for Evaluating Inductive Learning," Proceedings of the Sixth International Workshop on Machine Learning. 160-163.

Stanski, Henry R., Laurence J. Wilson and William R. Burrows (1989) "Survey of Common Verification Methods in Metereology," Research Report No. 89-5, Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin Street, Downsview, Ontario, Canada.

Stock, James H. and Mark W. Watson (1989) "New Indexes of Coincident and Leading Indicators," in **NBER Macroeconomics Annual**, 4: 351-393.

Wilcoxon, Frank (1945) "Individual Comparisons by Ranking Methods," *Biometrics Bulletin,* 1, 80-83.

World Metereological Organization (2000) **Standardized Verification System (SVS) for Long-Range Forecasts (LRF).** World Meterological Organization, Geneva, Switzerland.

Wright, Jonathan (2006) "The Yield Curve and Predicting Recessions," Finance and Economics Discussion Series 2006-07*, Federal Reserve Board.*

Youden, William J. (1950) "Index for Rating Diagnostic Tests," *Cancer*, 3: 32-35.

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 of recession (in months)	12.1	11.1	4.7	7.5

Notes: NBER-PI refers to NBER recessions defined when the peak and trough month are included. NBER-PE are NBER recessions where the peak date is excluded. R1 refers to the rule that classifies a recession as any observation where GDP growth is negative. R2 is the rule that instead requires two consecutive quarters of negative growth.

Table 2 – AUROCs for Four Recession Indicators

Variable	Indicator			
	BCDC-PI	**BCDC-PE**	**R1**	**R2**
GDP	0.9834	0.9848	0.8775**	0.9798
	(0.0039)	(0.0036)	(0.0203)	(0.0045)
	3	3	4	4
GDI	0.9811	0.9824	0.8623**	0.9655*
	(0.0045)	(0.0043)	(0.0212)	(0.0071)
	4	3	4	4
PI	0.9595	0.9589	0.8836**	0.9331
	(0.0079)	(0.0077)	(0.0231)	(0.0135)
	6	6	8	7
IP	0.9665	0.9680	0.8483**	0.9165*
	(0.0074)	(0.0071)	(0.2189)	(0.0195)
	4	4	7	6
MTS	0.9658	0.9626	0.8834	0.9553
	(0.0077)	(0.0083)	(0.0234)	(0.0099)
	3	2	5	3
PE	0.9666	0.9665	0.8346**	0.9120**
	(0.0071)	(0.0074)	(0.0239)	(0.0189)
	7	6	8	7
HE	0.9444	0.9453	0.8368**	0.9511
	(0.1250)	(0.0111)	(0.0247)	(0.0122)
	6	7	8	6

* Indicates that AUROC is different from BCDC-PE AUROC at 90 percent confidence interval
** Indicates that AUROC is different from BCDC-PE AUROC at 95 percent confidence interval

Notes: NBER-PI refers to NBER recessions defined when the peak and trough month are included. NBER-PE are NBER recessions where the peak date is excluded. R1 refers to the rule that classifies a recession as any observation where GDP growth is negative. R2 is the rule that instead requires two consecutive quarters of negative growth.

Table 3 - AUROCs for Cyclical GDP Using Three Alternative Detrending Methods

	HP Filter	BK Filter	Potential GDP (CBO)	HP Filter	BK Filter	Potential GDP (CBO)
n	748	676	724	748	676	724
AUROC	0.7032	0.7066	0.7518	0.8982	0.9048	0.8660
Std. err.	(00279)	(0.0269)	(0.0260)	(0.0157)	(0.0138)	(0.0211)
h*	--	--	--	6	6	6

Notes: GDP data filtered at quarterly frequency, then interpolated to monthly observations. $\lambda = 1600$ for HP filter. Baxter and King filter set to select frequencies between 6 and 32 quarters. Left hand panel refers to AUROCs for contemporaneous BCDC-PE dating. Right-hand panel reports the AUROCs at the optimal horizon *h*.

Table 4 - AUROCs for BCDC versus Statistical Dating Rules: Chauvet and Hamilton (2005) and Chauvet and Piger (2008)

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

Notes: Columns 1-3 correspond to AUROCs for $h = 0$, while columns 4-6 give the AUROCs, along with the h that produces the highest AUROC (reported as the last line in each entry). Standard errors reported in parentheses.

Table 5 – AUROCs for ADS, CFNAI, PMI, and LexisNexis News Index

Model	Horizon						
	0	**4**	**8**	**12**	**16**	**20**	**24**
ADS	0.9773	0.8759	0.7663	0.6022	0.5121	0.4803	0.4321
	(0.0059)	(0.0188)	(0.0245)	(0.0317)	(0.0341)	(0.0344)	(0.0348)
NAI (MA-3)	0.9548	0.8432	0.7006	0.5478	0.4644	0.4125	0.3947
	(0.0134)	(0.0233)	(0.0303)	(0.0345)	(0.0376)	(0.0373)	(0.0356)
PMI	0.9023	0.8000	0.6616	0.5047	0.4555	0.4268	0.4222
	(0.0181)	(0.0239)	(0.0263)	(0.0272)	(0.0294)	(0.0303)	(0.0292)
LexisNexis	0.8056	0.6451	0.5034	0.4121	0.3400	0.2942	0.3070
	(0.0293)	(0.0333)	(0.0359)	(0.0369)	(0.0386)	(0.0383)	(0.0390)

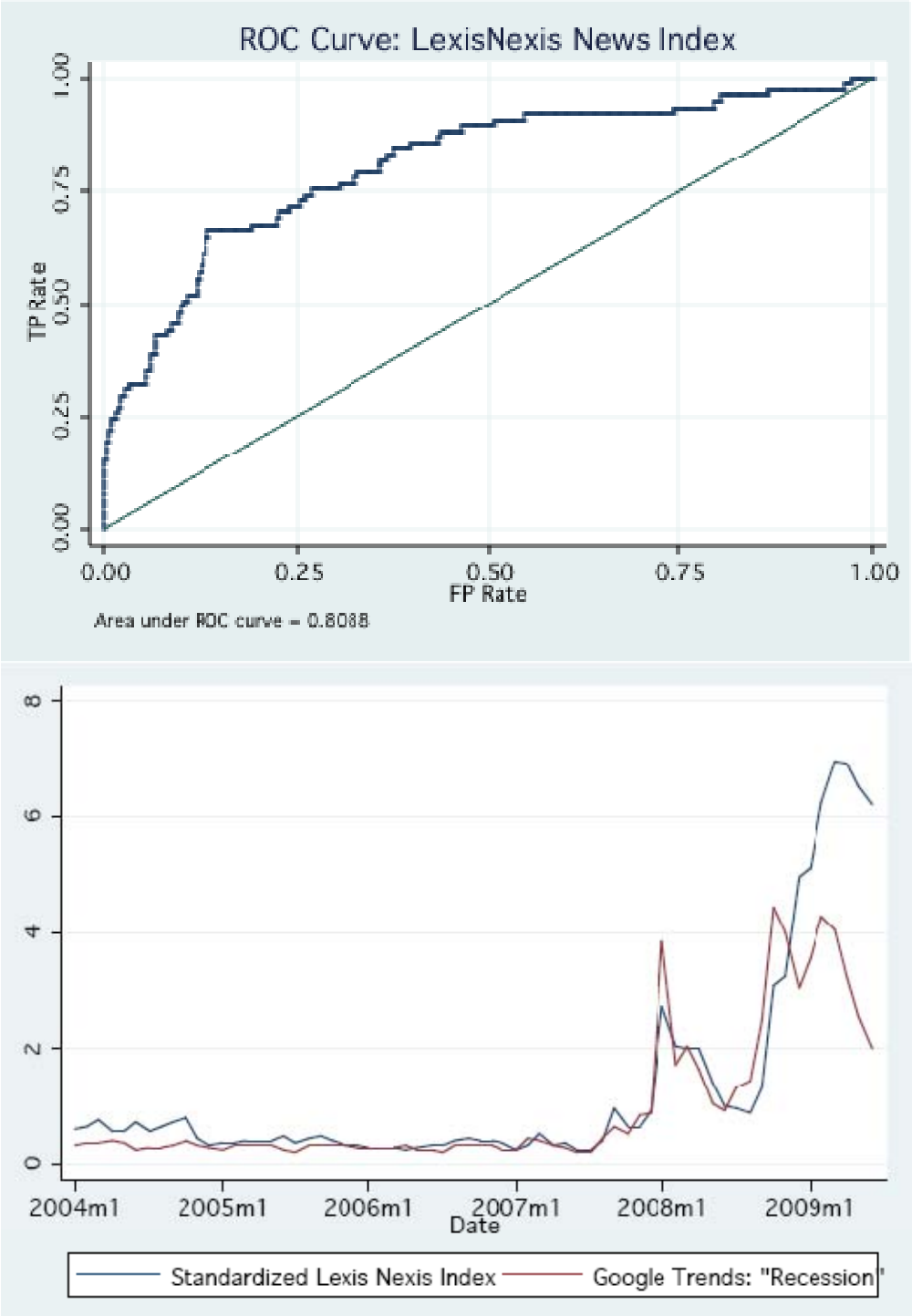
Notes: AUROC values at various horizons. Standard errors included in parentheses.

Table 6 - AUROCs and horizon of maximal for Conference Board's leading index

	AUROC at h=0	Positive $AUROC^{max}$	Negative $AUROC^{max}$
S&P 500	0.8279	0.8279	0.6204
	(0.0166)	(0.0166)	(0.0260)
	--	0	14
Vendor Performance	0.7531	0.7531	0.7827
	(0.0236)	(0.0236)	(0.0277)
	--	0	16
Average weekly hours, Manufacturing	0.9000	0.9235	0.7666
	(0.0138)	(0.0120)	(0.0272)
	--	3	17
New private housing units	0.8396	0.8396	0.8026
	(0.0295)	(0.0295)	(0.0321)
	--	0	15
Michigan survey consumer Expectations	0.8256	0.8256	0.8013
	(0.0275)	(0.0275)	(0.0261)
	--	0	14
Manufacturers' new orders: consumer goods	0.9337	0.9418	0.7259
	(0.0124)	(0.0108)	(0.0355)
	--	1	21
Manufacturers' new orders: capital goods	0.8291	0.9003	0.6075
	(0.0255)	(0.0173)	(0.0307)
	--	4	35
M2	0.6793	0.6793	0.7297
	(0.0341)	(0.0341)	(0.0321)
	--	0	22
Initial Claims	0.9563	0.9687	0.7849
	(0.0127)	(0.0093)	(0.0384)
	--	1	24
10-year T-bill less FFR	0.4122	0.7334	0.5825
	(0.0309)	(0.0238)	(0.0310)
	--	18	0

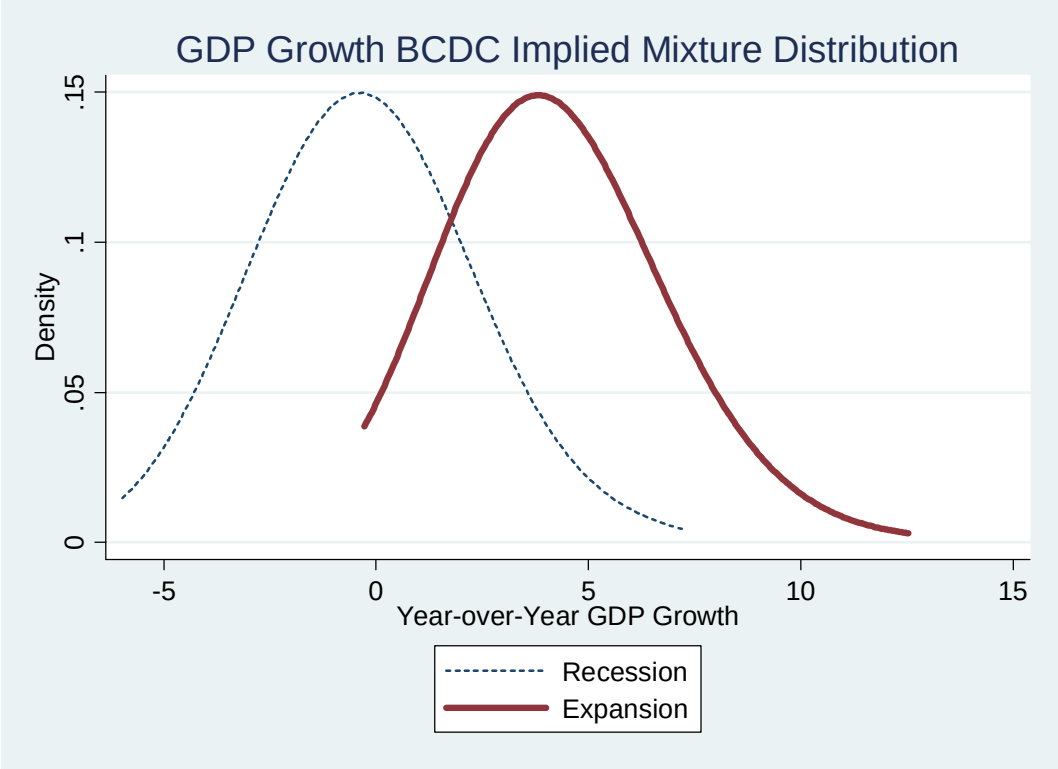
Notes: Standard errors in parentheses. The first column is the AUROC associated with $h = 0$. The second column is the maximum AUROC for the index at h^*. The third column is the maximum AUROC for the negative of the index at h^*. For each indicator, we show the AUROC, with standard errors in parentheses. The third number displays the forecast horizon associated with the maximum value of the AUROC.

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



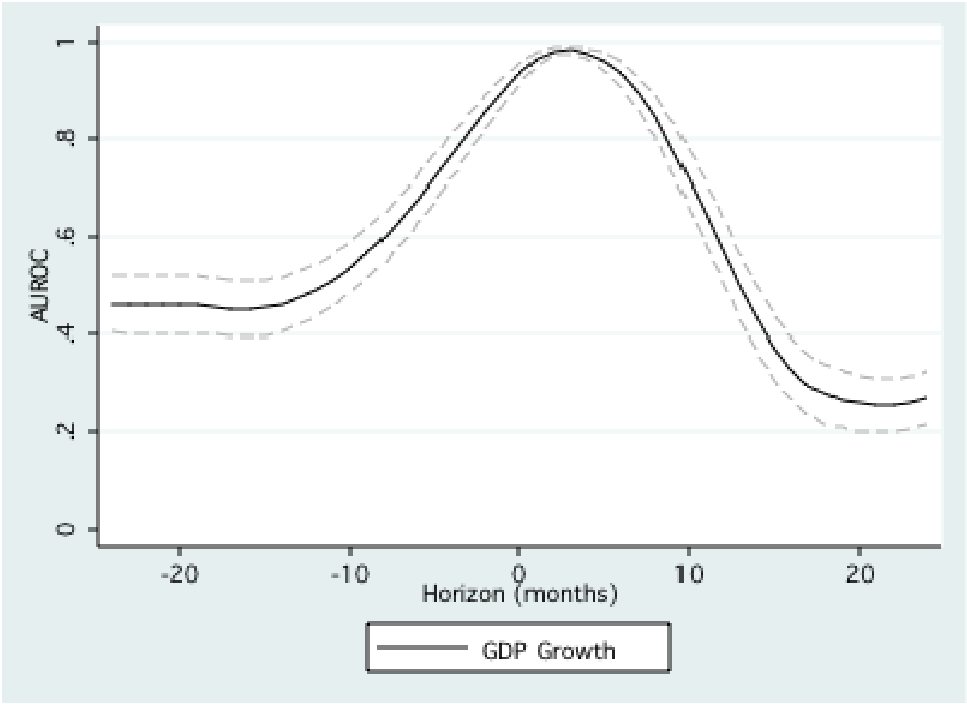
Note: See appendix for description of LexisNexis index.

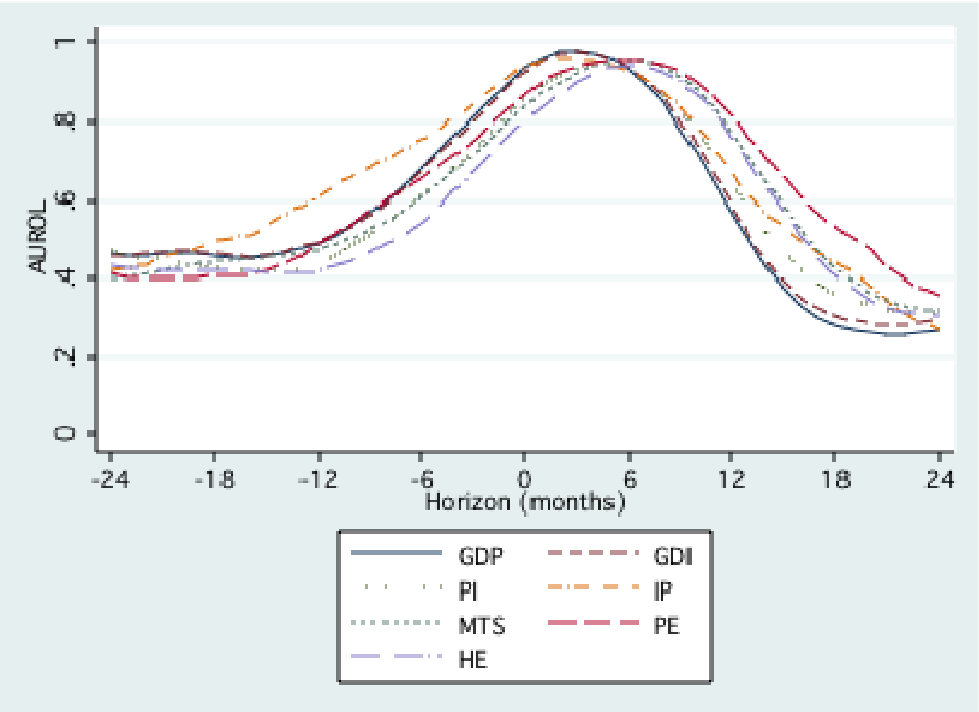
Figure 2 – Mixture Distribution for Yearly GDP Growth Implied by the BCDC Dating. Sample 1947:I – 2009:II



Notes: Kernel density estimates of year-over-year growth rate in GDP (not annualized quarterly rates) based on a Gaussian kernel with window width 2. The mean of the recession distribution is approximately -0.28%, with standard deviation 1.82%. The minimum is -3.98% and the maximum is 5.25%. The mean of the expansion distribution is 4.17% with a standard deviation of 1.92%. The minimum is -0.27% and the maximum is 12.54%.

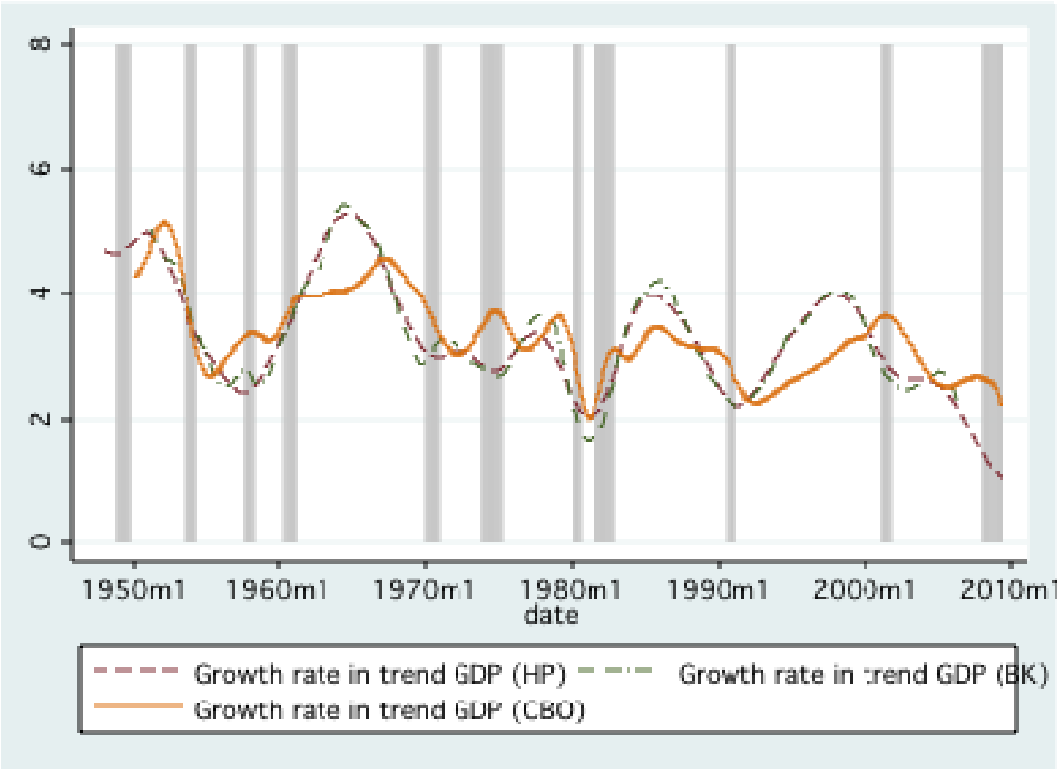
Figure 3 – The AUROC over time for BCDC economic indicators





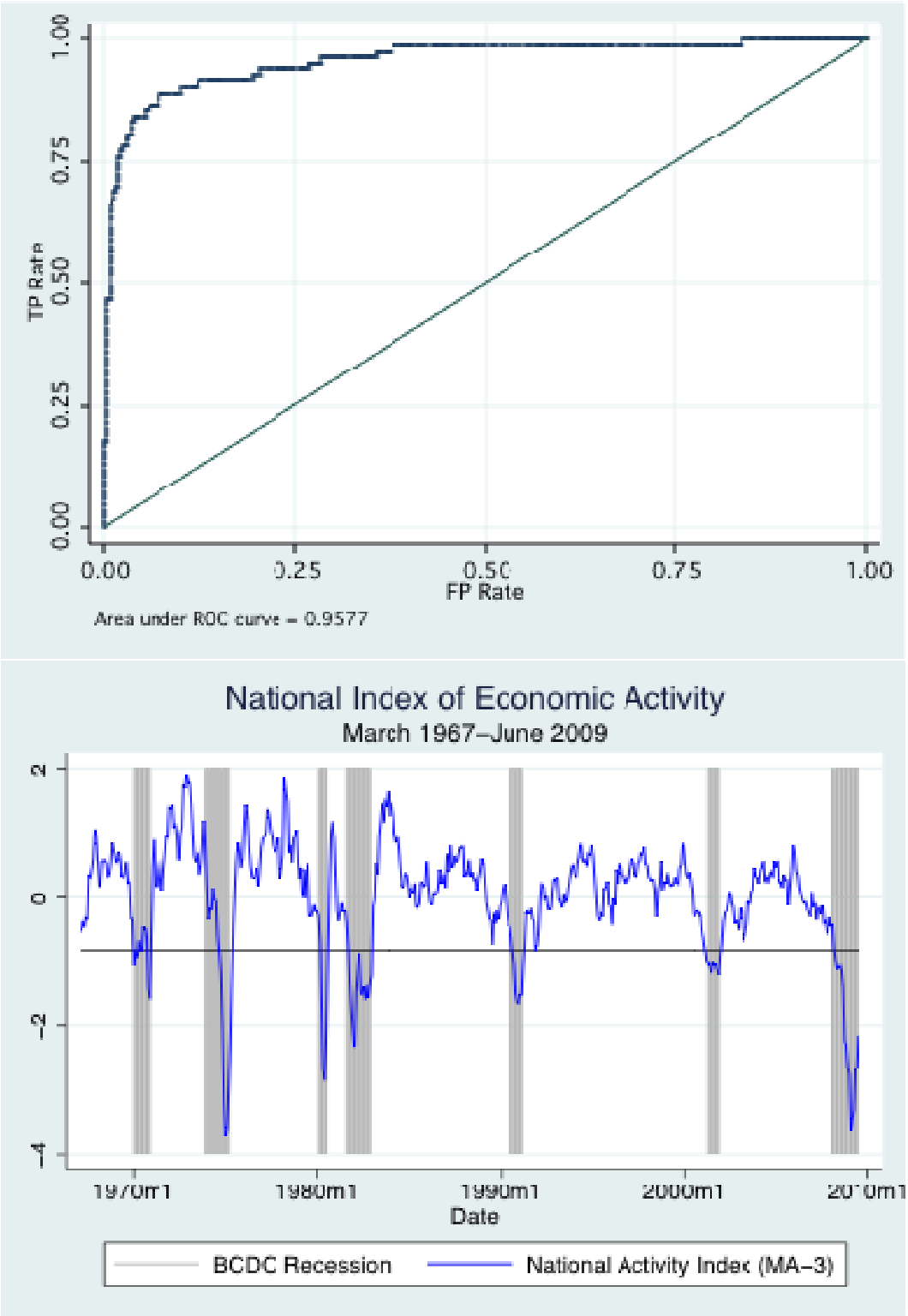
Notes: The top panel displays AUROCs and 95% confidence bands for classification ability of GDP. The bottom panel only displays the AUROCs for each of the coincident indicators the BCDC claims to use to determine peak and trough dates according to the December 1, 2008 release.

Figure 4 – Growth Rates of HP, BK and CBO trends for GDP over the Business Cycle



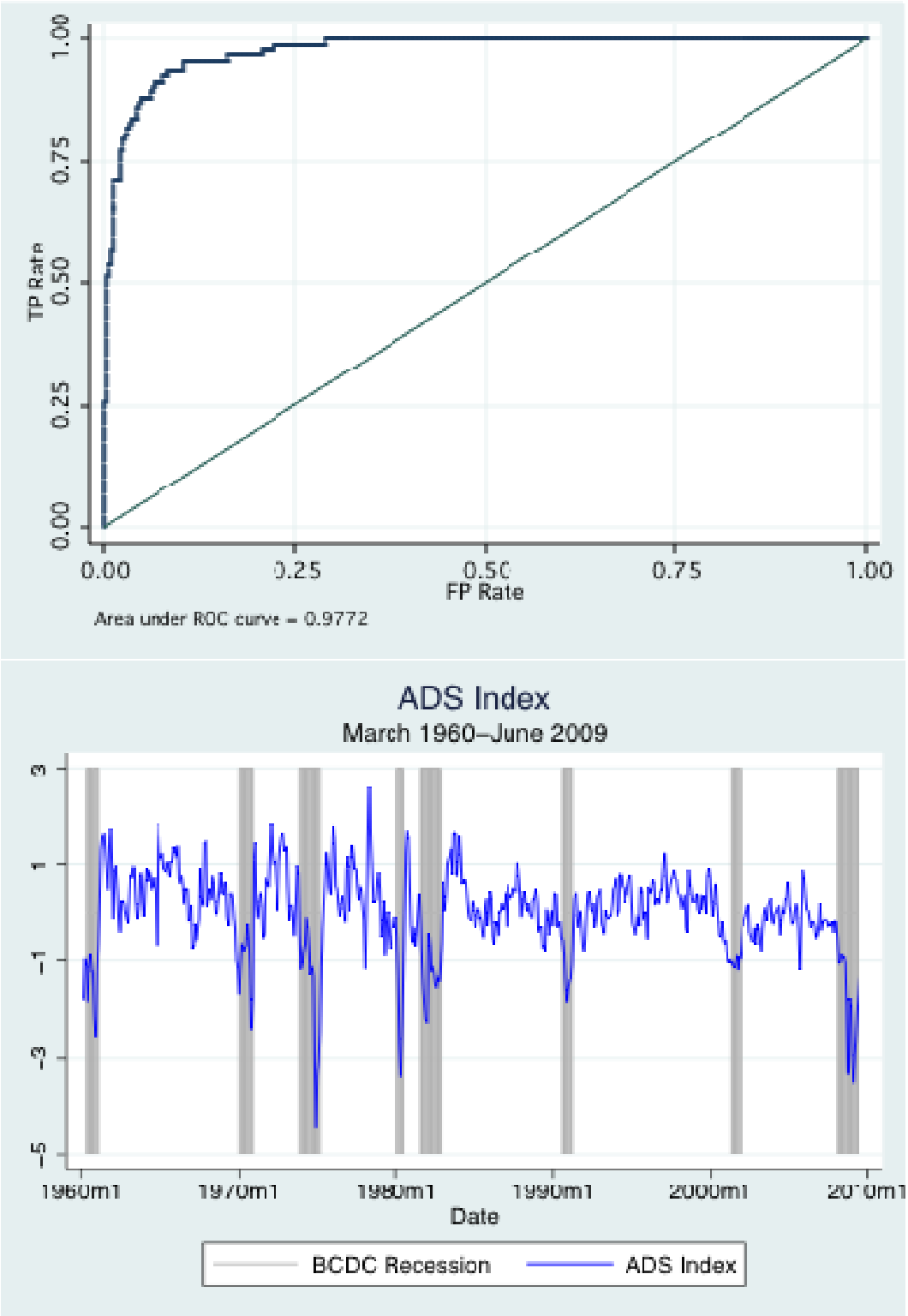
Notes: Shaded areas are NBER recessions. HP GDP trend refers to the Hodrick-Prescott trend output; BK GDP trend refers to the Baxter and King trend for frequencies above 32 quarters; CBO Potential Output is reported directly by the Congressional Budget Office. The graph depicts the yearly growth rates (in percentages) for each trend.

Figure 5 – The Chicago Fed National Activity Index and its ROC



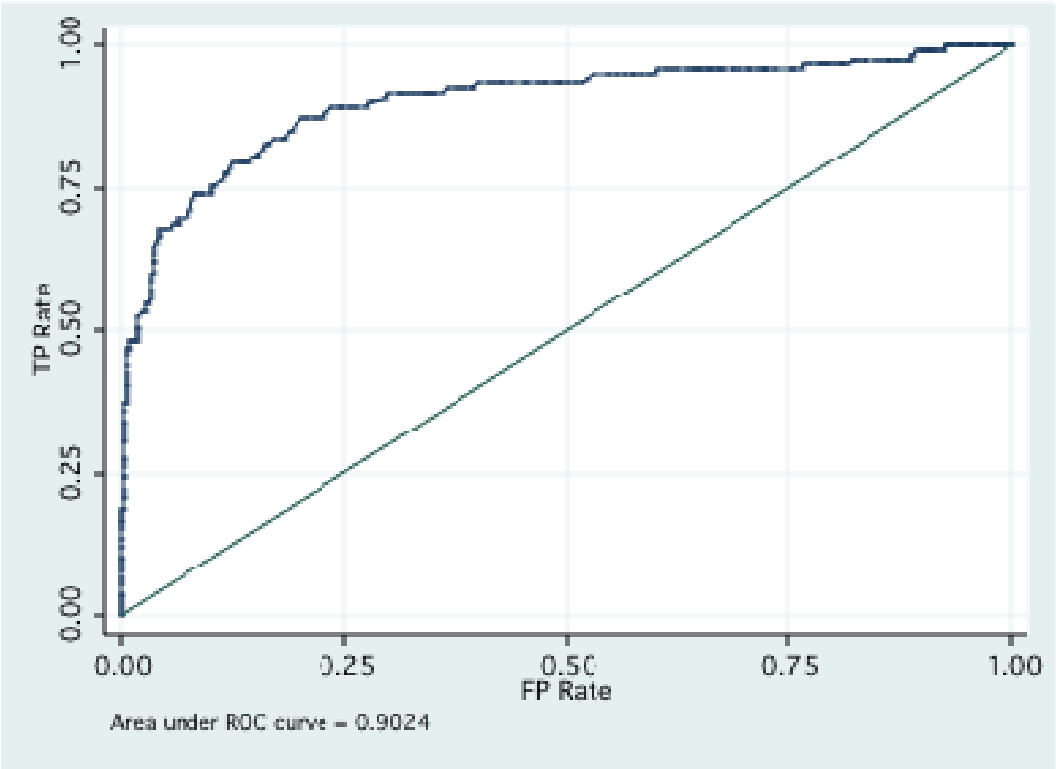
Notes: The top panel displays the contemporaneous ROC curve. The horizontal line corresponds to the value of the index that maximizes the utility of the method, assuming equally weighted benefits and costs.

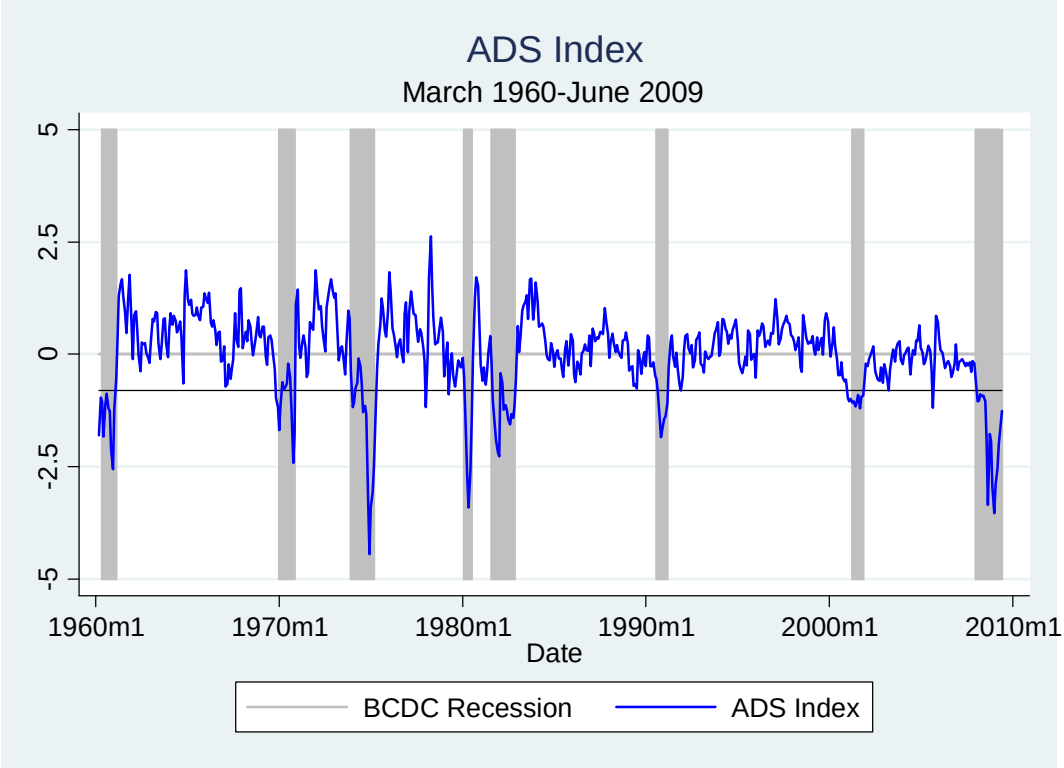
Figure 6 – The Arouba, Diebold and Scotti Index and its ROC



Notes: The top panel displays the contemporaneous ROC curve. The horizontal line corresponds to the value of the index that maximizes the utility of the method, assuming equally weighted benefits and costs.

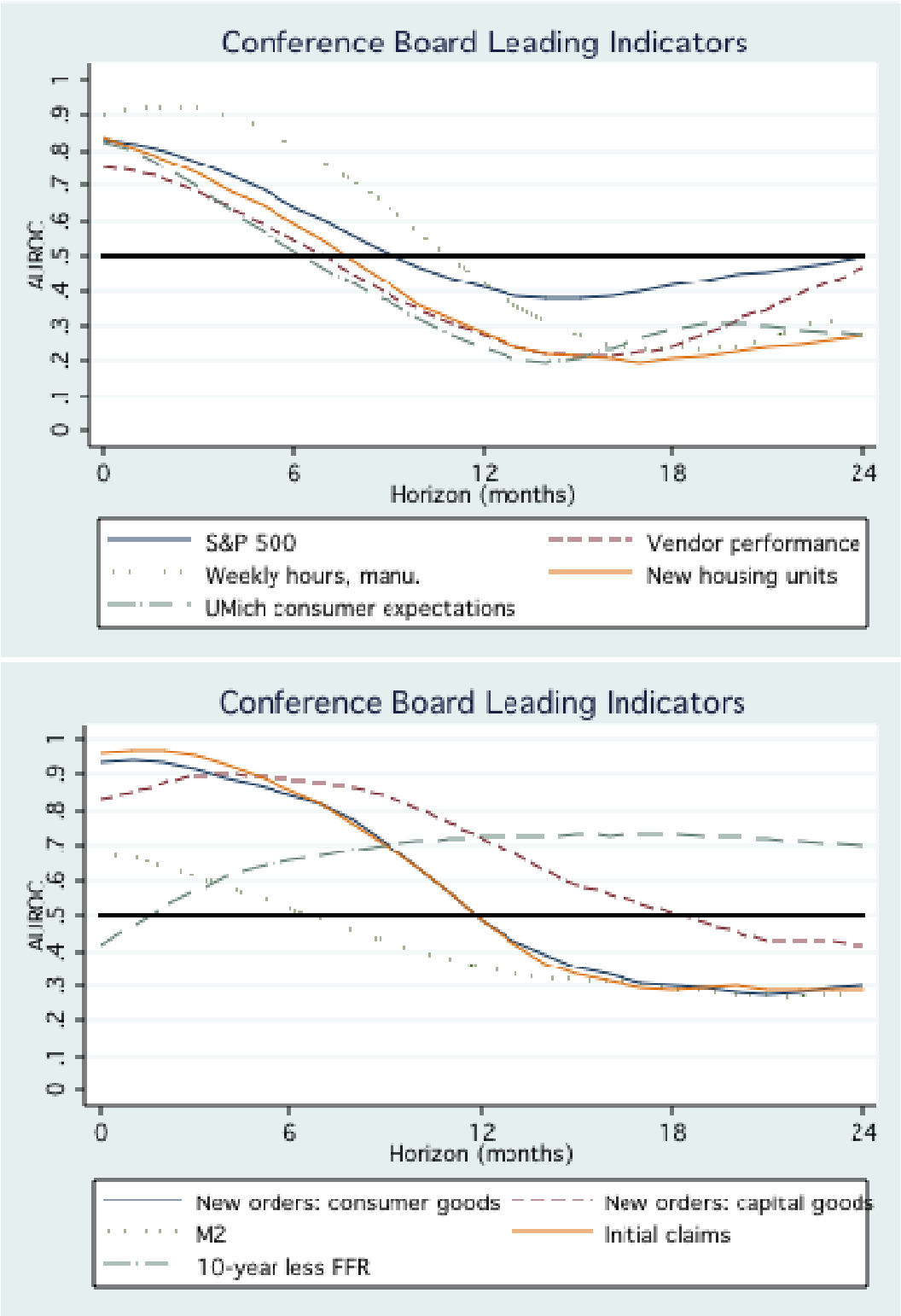
Figure 7 – The Producer Managers Index and its ROC





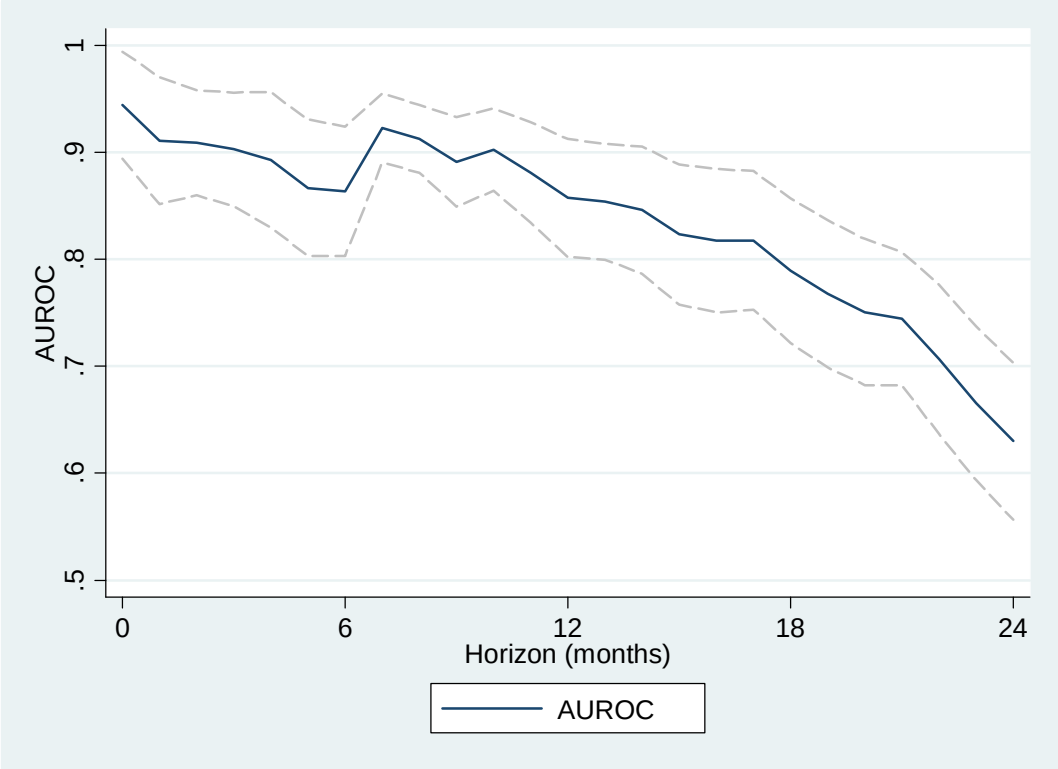
Notes: The top panel displays the contemporaneous ROC curve. The horizontal line corresponds to the value of the index that maximizes the utility of the method, assuming equally weighted benefits and costs.

Figure 8 – AUROCs 0 to 24 months into the future for the components of the Index of Leading Indicators published by the Conference Board



Notes: Confidence intervals suppressed for in the interest of readability. See Table 6 for detailed results.

Figure 9 – Out of sample AUROC values over 0 to 24 periods into the future from direct logistic regression of the Index of Leading Indicators from the Conference Board



Notes: Predictive model is a logistic regression containing all elements of the ILI. Predictions at all horizons made out-of-sample with a rolling (fixed-window) regression. Initial predictions made in January 1978 (initial sample of January 1968 to December 1977).