Wohlrabe, Klaus; Carstensen, Kai; Ziegler, Christina

Conference Paper

Predictive Ability of Business Cycle Indicators under Test: A Case Study for the Euro Area Industrial Production

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Forecasting Methods, No. A11-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Wohlrabe, Klaus; Carstensen, Kai; Ziegler, Christina (2010) : Predictive Ability of Business Cycle Indicators under Test: A Case Study for the Euro Area Industrial Production, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Forecasting Methods, No. A11-V3, Verein für Socialpolitik, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/37143

Predictive Ability of Business Cycle Indicators under Test: A Case Study for the Euro Area Industrial Production

Kai Carstensen
Klaus Wohlrabe
Christina Ziegler

CESIFO WORKING PAPER NO. 3158
CATEGORY 12: EMPIRICAL AND THEORETICAL METHODS
AUGUST 2010

CESifo *Working Paper No. 3158*

Predictive Ability of Business Cycle Indicators under Test: A Case Study for the Euro Area Industrial Production

Abstract

In this paper we assess the information content of seven widely cited early indicators for the euro area with respect to forecasting area-wide industrial production. To this end, we use various tests that are designed to compare competing forecast models. In addition to the standard Diebold-Mariano test, we employ tests that account for specific problems typically encountered in forecast exercises. Specifically, we pay attention to nested model structures, we alleviate the problem of data snooping arising from multiple pairwise testing, and we analyze the structural stability in the relative forecast performance of one indicator compared to a benchmark model. Moreover, we consider loss functions that overweight forecast errors in booms and recessions to check whether a specific indicator that appears to be a good choice on average is also preferable in times of economic stress. We find that there is not one best indicator that uniformly dominates all its competitors. The optimal choice rather depends on the specific forecast situation and the loss function of the user. For 1-month forecasts the business climate indicator of the European Commission and the OECD composite leading indicator generally work well, for 6-month forecasts the OECD composite leading indicator performs very good by all criteria, and for 12-month forecasts the FAZ-Euro indicator published by the Frankfurter Allgemeine Zeitung is the only one that can beat the benchmark AR(1) model.

JEL-Code: C32, C53, E32.

Keywords: weighted loss, leading indicators, euro area, forecasting.

Kai Carstensen
Ifo Institute for Economic Research at the University of Munich
Poschingerstrasse 5
Germany – 81679 Munich
carstensen@ifo.de

Klaus Wohlrabe
Ifo Institute for Economic Research at the University of Munich
Poschingerstrasse 5
Germany – 81679 Munich
wohlrabe@ifo.de

Christina Ziegler
Ifo Institute for Economic Research at the University of Munich
Poschingerstrasse 5
Germany – 81679 Munich
ziegler@ifo.de

This version: June 2010 - Preliminary— Please do not quote
We thank the participants of various seminars and the two unknown referees for valuable comments and suggestions.

1 Introduction

The euro area is a rather new subject in the literature on macroeconomic forecasting. However, it is all the more interesting, especially because the European Central Bank conducts its monetary policy explicitly with a view to the euro area as a whole. The forward-looking elements of this policy requires to generate accurate forecasts of inflation and economic activity. In this paper, we consider the latter, concentrating on euro area industrial production which is the most timely "hard indicator" of aggregate output that is available. Specifically, we assess whether several popular "soft indicators" reveal early information that helps to improve the accuracy of industrial production forecasts.

In standard empirical out-of-sample forecasting exercises the performance of leading indicators is often measured by the (root) mean squared error which is derived from a symmetric quadratic loss function. Furthermore, in order uncover significantly forecasting differences between pairs of indicators, typically the popular Diebold-Mariano test is employed.

In line with the recent literature, we challenge this "standard assessment approach" in several ways. First, we allow for a flexible weighting scheme of the forecasting errors in the relevant loss function. This can be more satisfactory in situations where some observations are more important than others, as argued by van Dijk and Franses (2003). The flexible weighting scheme allows to judge the predictive ability of leading indicators during booms or recessions which might be particularly important times for monetary policy decisions and, thus, accurate forecasts, the recent financial and economic crisis being an impressive example. To take these issues into account, we include a weighted loss function into the standard Diebold-Mariano type tests.

Second, we pay attention to the the aspect of nested models in forecast comparisons. Starting with Clark and McCracken (2001) this aspect has been increasingly discussed in the literature. The basic idea is that the comparison of, say, an indicator model with a nested benchmark model (that does not include the indicator) has to take into account the estimation uncertainty associated with estimating the additional parameters for the indicators. Neglecting this uncertainty gives rise to a bias in favor of the benchmark model. For example, in such a situation the Diebold-Mariano test would signal too often that the indicator model is not able to improve upon the benchmark. Specifically, we employ the recently proposed test by Clark and West (2007) to account for this issue.

Third, we note that our forecast comparison—like almost all work in this field—does not literally contrast one model with a single competitor which is the setting the standard pairwise tests such as the one proposed by Diebold and Mariano (1995) are designed for. Instead, we aim at finding the most promising indicators from a possibly large set of can-

didates. In such a situation, a few pairwise tests can signal dominance of one indicator over the other simply by chance, much like repeated draws from, say, the standard normal distribution will yield from time to time values that exceed conventional critical values and lead to the rejection of the mean zero hypothesis. To account for this data snooping problem we apply the test for superior predictive ability (SPA) proposed by Hansen (2005) and based on the seminal paper by White (2000).

Finally, we take a first look at the stability issue of forecast dominance. As argued by Giacomini and Rossi (2008) the relative forecast performance of one indicator to another may change over time, possibly due to structural instabilities, e.g., as the consequence of booms or recessions. A practitioner would of course prefer an indicator that has at least in past shown stable dominance over its competitors. To this end, we implement the fluctuation test proposed by Giacomini and Rossi (2008) which is based on a series of local Diebold-Mariano tests. To the best of our knowledge, we are the first who allow for weighted loss differentials within this framework to assess the forecasting stability also for booms and recessions.

The remainder of the paper is structured as follows. In Section 2, we briefly overview the related literature. In Section 3, we discuss the weighted loss function we use to compare to forecast models before we outline in Section 4 the various forecast accuracy tests we employ. The setup of our out-of-sample forecast exercise is described in Section 5 and the results are presented in Section 6. Section 7 summarizes and concludes.

2 Related Literature

As the euro area is a rather new entity, it is has become only recently a topic in the field of macroeconomic forecasting. Accordingly, there are only few directly related papers available. While we study point forecasts, most of the work done on the euro area focuses on turning point prediction for industrial production, or point forecasts for gross domestic product and inflation. Only the study by Bodo et al. (2000) uses one of the indicators we consider, namely the European Economic Sentiment indicator. Therefore, we are among the first who assess the point forecasting ability of leading indicators for the euro area.

Bodo et al. (2000) provide one of the first studies to forecast euro area industrial production. Besides univariate and vector autoregressive models referring to the four largest euro area countries, the authors employ a two-country vector autoregressive model for the euro area and the US. They study whether the inclusion of survey-based business climate indicator published by the European Commission helps to improve the forecasts. Employing the modified Diebold-Mariano test, they find that the benchmark ARIMA model is

outperformed by the two-country model with the survey indicator.

Marcellino et al. (2003) forecast quarterly euro area macroeconomic time series, among them industrial production, using a dynamic factor model framework with country-specific data. They find, based on a number of different model specifications, that country-specific information matters, albeit without testing for significant differences in predictive ability. Forni et al. (2003) show in a dynamic factor framework that including financial variables does not improve forecast accuracy for euro area industrial production. Marcellino (2008) provides evidence that artificial neural networks perform on average better than simple linear models without indicators.

Ozyildirim et al. (2010) construct a leading economic index (LEI) consisting of eight different time series (including sentiment, interest rate spread and monetary aggregate). In addition the authors propose a coincident economic index consisting (CEI) of industrial production, employment, manufacturing and retail trade. Besides the identification and forecasting of a Euro area business cycle turning points, Ozyildirim et al. (2010) compare the forecasting performance of the LEI for the CEI. Using real-time data it is shown that LEI improves the forecasting accuracy upon an AR benchmark model.

Using different forecast targets, there are quite a few papers that apply the newly developed tests of forecast accuracy discussed above. However, they typically focus on exchange rate and financial forecasting. As an exception, Milas and Rothman (2008) use weighted loss differentials as proposed by van Dijk and Franses (2003) to assess macroeconomic forecasting performance. They use smooth transition vector error-correction models in a simulated out-of-sample forecasting experiment for the unemployment rates in the U.S., the U.K., Canada, and Japan. They find that the forecast performance of the models can differ between booms and recessions. Caggiano et al. (2009) use the test proposed by Clark and West (2007) to account for nested model structures when comparing forecast models for the euro area and other countries. The aspect of data snooping has recently been taken into account by Clark and McCracken (2009) who compare a very large set of forecasting models for U.S. macroeconomic variables. The fluctuation test is used in Fichtner et al. (2009) to assess the stability in the predictive ability of the OECD composite leading indicator for industrial production in 11 OECD countries. It is also used by Rossi and Sekhposyan (2010) to check whether the forecasting performance of various economic models for US output growth and inflation has changed over time. They find that during the Great Moderation many forecasting models became essentially useless.

3 Weighted Loss Functions

The standard period-t loss function used in most of the forecast evaluation literature ist the squared forecast error

$$\mathcal{L}_{i,t} = e_{i,t}^2, \tag{1}$$

where $e_{i,t} = y_t - y_{i,t}^f$ is the forecast error of model i, y_t is the realization of the target variable, $y_{i,t}^f$ is the value predicted by model i. While theoretical results are available for quite general loss functions, see, e.g., Diebold and Mariano (1995), the applied literature concentrates on the quadratic loss function. Comparing the average loss difference of two competing models 1 and 2 then means to compute their mean squared forecast errors (MSFE)

$$\text{MSFE}_i = \frac{1}{P} \sum_{t=T+1}^{T+P} e_{i,t}^2, \quad i = 1,2, \tag{2}$$

over the forecast period $T+1$ to $T+P$ and choose the model with the smaller MSFE. However, one can think of many occasions in which different loss functions can make more sense for the applied forecaster but also for the user of a forecast such as a politician or the CEO of a company. For example, the recent recession demonstrated that a good forecast of a rather extreme event might be of special interest beyond that of minimizing an average squared error: banks could have taken earlier measures to shelter against the turmoil, governments could have started stimulus packages in time, and firms might have circumvented their strong increase in inventories.

As argued by van Dijk and Franses (2003), a weighted squared forecast error can be used to place more weight on unusual events when evaluating forecast models. Specifically, they propose to use the loss function

$$\mathcal{L}_{i,t}^w = w_t e_{i,t}^2, \tag{3}$$

where the weight w_t is specified as

1. $w_{\text{left},t} = 1 - \widehat{F}(y_t)$, where $F(\cdot)$ is the cumulative distribution function (cdf) of y_t, to overweight the left tail of the distribution. This gives rise to a "recession" loss function. In the following, we construct the empirical cdf, $\widehat{F}(y_t)$, as the proportion of observations less than or equal to y_t.

2. $w_{\text{right},t} = \widehat{F}(y_t)$, to overweight the right tail of the distribution. This gives rise to a "boom" loss function.

Obviously, the weighted loss function (3) collapses to the standard loss function (1) when equal weights $w_t = 1$ are imposed. This gives rise to the conventional "uniform" loss function.

Using a weighted loss function complicates things only slightly. To evaluate a forecast model i over a forecast period $T+1$ to $T+P$ simply requires to calculate the weighted mean squared forecast error

$$\text{MSFE}_i = \frac{1}{P}\sum_{t=T+1}^{T+P} w_t e_{i,t}^2. \tag{4}$$

In order to compare, say, model i to a benchmark model 0, one calculates the weighted loss difference

$$d_{i,t} = \mathcal{L}_{0,t}^w - \mathcal{L}_{i,t}^w = w_t e_{0,t}^2 - w_t e_{i,t}^2 \tag{5}$$

and averages over the the forecast period

$$\overline{d}_i = \frac{1}{P}\sum_{t=T+1}^{T+P} d_{i,t} = \frac{1}{P}\sum_{t=T+1}^{T+P} w_t e_{0,t}^2 - \frac{1}{P}\sum_{t=T+1}^{T+P} w_t e_{i,t}^2 \tag{6}$$

In the remainder of this paper, we will use this weighted loss and analyze the forecast accuracy of different models (which in turn are based on different indicators) with respect to the different weighting schemes introduced above.

Figure 1 depicts the empirical cdf of the target variable in our application, namely the growth rate of euro area industrial production. It demonstrates that observations smaller than -0.04 and larger than 0.04 receive a particularly high weight in the analysis of recessions and booms, respectively. The evolution of euro area industrial production and of the weight series is displayed in Figure 2. In the upper panel, the extreme fall in euro area industrial production during the winter of 2008/2009 catches the eye. Hence, this event also dominates the recession weights (lower panel). However, the recession in 2001/2002 receives almost the same weights. Therefore, our results are not solely driven by a single event. On the flip side, the boom weights are particularly high during the rapid expansion in 2000 and in the period of 2006 to 2008 (middle panel).

4 Forecast Accuracy Tests

To analyze whether empirical loss differences between two or more competing models are statistically significant, there is a large number of tests proposed in the literature, among which the pairwise test introduced by Diebold and Mariano (1995) seems to be the most influential and most widely used. Therefore, we also apply it to our setting. We augment our analysis with three further tests which are designed to account for additional important features of the forecast evaluation problem and which have not been used very often in applied work. First, the test proposed by Clark and West (2007) takes into account that our benchmark model—a simple AR(1) model—is nested in all the competing models to

which early indicators are added. Second, the test suggested by Hansen (2005) circumvents the problem of data snooping that arises when a number of pairwise tests are conducted. Finally, the fluctuation test by Giacomini and Rossi (2008) is useful to examine whether the relative forecast performance of one model has changed over time relative to the benchmark. In the following, we briefly introduce these test.

4.1 Modified Diebold-Mariano Test

The standard way to discriminate between the forecasting performances of two competing models is to apply the forecast accuracy test proposed by Diebold and Mariano (1995). In this paper, we apply the modified Diebold-Mariano (MDM) test proposed by Harvey et al. (1997), which corrects for a small sample bias. It evaluates whether the average loss differences between the two models is significantly different from zero. Hence, it is a pairwise test that is designed to compare two models at a time, say, model i with benchmark model 0. Specifically, the null hypothesis of the MDM test is that of equal forecast performance,

$$E\left[d_{i,t}\right] = E\left[\mathcal{L}_{0,t}^{w} - \mathcal{L}_{i,t}^{w}\right] = 0. \tag{7}$$

Following Harvey et al. (1997), we use the MDM test statistic

$$\text{MDM} = \left(\frac{P+1-2h+P^{-1}h(h-1)}{P}\right)^{1/2} \widehat{V}(\overline{d}_i)^{-1/2}\overline{d}_i, \tag{8}$$

where h is the forecast horizon and $\widehat{V}(\overline{d}_i)$ the estimated long-run variance of series $d_{i,t}$. The MDM test statistic is compared with a critical value from the t-distribution with $P-1$ degrees of freedom.

4.2 Forecast accuracy test for nested models

In our setting presented in more detail below, one of the the benchmarks is an AR(1) model against which competing models augmented with more lags and additional indicators are tested. If the AR(1) was the true model, the benchmark model would be nested in the competing models. When testing the null hypothesis of equal forecast accuracy for two nested models, a complication arises as argued by, inter alia, Clark and McCracken (2001) and Clark and West (2007). Consider the typical case in the applied forecast evaluation literature that a simple benchmark model is compared with a rival model which is augmented by additional explanatory variables such as further lags or indicators. Under the null, the additional variables are useless and their coefficients are zero. Estimating these coefficients

introduces noise in the derived forecasts of the rival model. Hence, under the null, the forecast accuracy of the parsimonious benchmark model is higher than (and not equal to) that of the larger rival model. Neglecting this fact leads to undersized tests with poor power, see Clark and McCracken (2001) and Clark and West (2005). In this sense, conventional tests favor the parsimonious model too often. Therefore, Clark and West (2007) propose an adjusted test that takes the nested model structure into account.

Specifically, for a test in the spirit of Diebold and Mariano (1995), Clark and West (2007) define the adjustment term

$$\bar{a}_i = \frac{1}{P}\sum_{t=1}^{P} w_t \left(y_{0,t}^f - y_{i,t}^f\right)^2, \tag{9}$$

where $y_{0,t}^f$ is the forecast of the parsimonious benchmark model and $y_{i,t}^f$ is the forecast of the augmented rival model. As they consider an unweighted loss functions, they set $w_t = 1$. The test statistic is defined as

$$\text{CW} = \widehat{V}(\overline{d}_i - \bar{a}_i)^{-1/2}\left(\overline{d}_i - \bar{a}_i\right), \tag{10}$$

where $\widehat{V}(\overline{d}_i - \bar{a}_i)$ is the estimated long-run variance of the adjusted loss difference $\overline{d}_i - \bar{a}_i$. Note that it is essential that the forecasts be computed from a rolling regression. As demonstrated in a simulation study by Clark and West (2007), using forecasts computed from a rolling regression scheme and applying the normal distribution leads to a fairly good but somewhat undersized test. For example a test with 10 percent nominal size will typically have a true size between 5 and 10 percent.[1] For our purpose, this should be a good approximation.

Note that is sensible to report the results both for the MDM test, which assumes non-nested models, and the CW test, which assumes nested models because it is unknown which of the two cases holds in reality. For example, if the AR(1) benchmark was not correct, the difference between the forecast errors of this model and an augmented indicator model need not converge to zero. This would make the CW test invalid. On the other hand, a comparison between different indicator forecasts is not necessarily non-nested if both indicators are uninformative so that the difference of the forecast errors converges to zero. This would make the MDM test invalid.

[1] A referee pointed out that this need not hold in our application. Therefore, we performed a simulation analysis tailored to our data and forecasting procedure presented below. It turned out that the size performance of the CW test is as described by Clark and West (2007).

4.3 Testing rationality with the Mincer-Zarnowitz regression

The rationality test proposed by Mincer and Zarnowitz (1969) and Zarnovitz (1985) is based on the idea that the error of an efficient forecast has to be unbiasend and uncorrelated with the forecast itself which gives rise to the standard Mincer-Zarnowitz regression

$$y_{t+1} = \alpha + \beta y^f_{t+1} + u_{t+1}. \tag{11}$$

For general loss functions, Elliott et al. (2008) show that efficient forecasts require that the generalized forecast error $\mathcal{L}'_{i,t}$ be orthogonal to the information set available to the forecaster. This can be implemented in a generalized Mincer-Zarnowitz regression by regressing $\mathcal{L}'_{i,t}$ on a set of variables v_t which are part of that information set. Using the weighted loss function (3) yields $\mathcal{L}'_{i,t} = 2w_t e_{i,t}$ and constraining v_t to a constant gives rise to

$$w_{t+1} e_{i,t+1} = \alpha + u_{t+1}, \tag{12}$$

or, in a slightly more unrestricted form,

$$w_{t+1} y_{t+1} = \alpha + \beta w_{t+1} y^f_{i,t+1} + u_{t+1}, \tag{13}$$

which collapses to the standard Mincer-Zarnowitz regression under uniform loss. The null hypothesis of efficiency can be formulated as $H_0 : \alpha = 0, \beta = 1$. A Wald test and the F distribution are used to test this null. As we allow u_{t+1} to exhibit serial dependency, we use the estimated long-run variance to construct the test statistic.

4.4 Superior Predictive Ability Test

Conventional econometric techniques for forecast evaluation focus on the comparison of two models at a time. Applying such pairwise tests sequentially to a number of models gives rise to the problems related to multiple testing procedures, particularly invalidating standard critical values. Effectively, comparing several different models to a benchmark model may result in spuriously identifying a superior model just by chance. To account for this data snooping problem we apply the test for superior predictive ability (SPA) proposed by Hansen (2005) which is based on the seminal paper by White (2000). The idea of this test is basically to compare a benchmark forecast model simultaneously to the whole set of m rival forecast models with the null hypothesis being that the benchmark is not inferior to any of the rivals. The null is formulated as the multiple hypothesis

$$H_0 : E(d_{i,t}) \leq 0 \quad \forall i = 1, \ldots, m. \tag{14}$$

and is rejected when at least one of the rival models yields significantly more accurate forecasts—and thus a smaller expected loss—than the benchmark model.

Of course, the expectation of $d_{i,t}$ is unknown, but it can be consistently estimated with the sample mean $\overline{d}_i$, $i = 1, ..., m$. White (2000) proposes the reality check test statistic

$$RC = \max_k P^{1/2}\overline{d}_i \,. \tag{15}$$

Note that the limiting distribution of *RC* is not unique under the null hypothesis. Therefore, the stationary bootstrap method of Politis and Romano (1994) is utilized.

As a major drawback, the *RC* test depends heavily on the set of competing models. If this set contains poor or irrelevant models delivering bad forecasts then the test is conservative in the sense that the critical value, which the *RC* statistic has to exceed in order to reject the null, increases with the number of included alternatives. Hence, adding enough irrelevant models could, in principle, lead to accepting the null hypothesis no matter how good a single competing model might be. As a solution to this problem, Hansen (2005) proposes the studentized test statistic

$$SPA = \max\left[\max_k \widehat{V}(\overline{d}_i)^{-1/2}\overline{d}_i, 0\right], \tag{16}$$

where $\widehat{V}(\overline{d}_i)$ denotes the consistently estimated long-run variance of $\overline{d}_i$. Assuming that irrelevant models deliver high forecast errors, the studentization downweights such models. Thereby, the size of the *SPA* test should be stable even if irrelevant models are added. Since the limiting distribution of the test statistic is not unique under the null hypothesis, a stationary bootstrap is used. Moreover, the distribution theory requires the use of a rolling estimation window in contrast to the recursive scheme used for the other tests.

4.5 Fluctuation Test

To analyze the stability of the forecasting performance over time, we implement the fluctuation test proposed by Giacomini and Rossi (2008). The test is based on the idea that due to potential structural instabilities—in our context possibly as the consequence of booms or recessions—the relative forecast performance of two competing models may change. Therefore, the authors propose to assess the development of a local loss difference over time in contrast to concentrating on the average (global) loss difference as in conventional tests. This may supply important information for a forecaster. In particular, indicator models that deliver accurate forecast only in specific situations or only at the beginning of the historical out-of-sample experiment might be downweighted.

To implement the fluctuation test, Giacomini and Rossi (2008) calculate the centered local loss differences of the Diebold-Mariano type,

$$\overline{d}_{i,t}^{\,\text{local}} = \frac{1}{Q}\sum_{\tau=t-Q/2}^{t+Q/2-1} \widehat{V}(\overline{d}_i)^{-1/2} d_{i,\tau}, \quad t = T+Q/2+1, \ldots, T+P-Q/2+1, \tag{17}$$

where Q is the length of the sub-sample. They check whether this sequence crosses the appropriate critical values which can be derived from a non-standard limiting distribution and are provided by the authors. If it does, then an instability is detected. Note that in our application below we calculate the forecasts from a rolling regression scheme and set the sub-sample length to $Q = 48$ months.

When interpreting the results of the fluctuation test in comparison to a conventional Diebold-Mariano test, one should keep in mind that the null hypothesis of equal forecast accuracy is tested against slightly different alternatives. In the conventional approach, the alternative hypothesis is that one of the two models delivers a smaller expected loss than the other *on average* over a fixed evaluation period. Hence, the approach presupposes structural invariance. In contrast, the fluctuation test uses the alternative hypothesis that one of the two models delivers a smaller expected loss at *some point* in the evaluation period. As this point is unknown, to prevent the test from spuriously detect instability, the absolute critical values are larger than in the conventional approach. This result is well known from standard structural break tests, such as the "sup" tests discussed by Andrews (1993). Therefore, in finite samples it might well be the case that the null hypothesis of equal forecast accuracy is rejected on average (assuming structural stability) but not locally (dropping the assumption of structural stability).

5 Empirical Setup

5.1 Database

We consider seven different business cycle indicators that are often used for the prediction of economic growth in the euro area. These indicators are constructed and published by different institutions such as the European Commission, the OECD, the ZEW, the DZ-Bank, and the CEPR. Table 1 contains a list of the indicators and their components. Our target series is the the year-over-year (yoy) growth rate of the industrial production index for the euro area as published by Eurostat. Although industrial production accounts only for around 20 percent of total GDP, it is regarded as a well-suited and quickly available business cycle indicator as argued, inter alia, by Breitung and Jagodzinski (2001). Our sample spans from $1992M02$ to $2009M6$. Unit root tests indicate that the FAZ indicator must be transformed into differences to be stationary.[2]

[2]For all other variables, GLS-based unit root tests reject the null hypothesis of non-stationarity. Detailed results are available upon request.

5.2 Forecast model

In our forecast exercise we consider the standard autoregressive distributed lag (ADL) model for generating forecasts. The h-step-ahead model is given by

$$y_{t+h} = \alpha + \sum_{i=1}^{p} \phi_i y_{t+1-i} + \sum_{j=1}^{r} \theta_j x_{t+1-j} + \varepsilon_t \tag{18}$$

where y_t is the year-on-year growth rate of euro area industrial production and x_t denotes one of the aforementioned leading indicators which are taken as exogenous. Hence, we refrain from modeling feedback effects. We allow for a maximum of 12 lags both for the endogenous and the exogenous variable. The lag length is chosen via the AIC criterion. We employ a rolling forecasting scheme as required for the Hansen test. The initial estimation period ranges from 1992:02 to 2000:1 ($T = 96$) which is moved forward through up to 2009:05. At each point in time equation (18) is re-specified and the forecasts are calculated. The initial forecast date is 2000:01 plus the forecast horizon and the final forecast date is 2009:06. We generate short-term ($h = 1$), medium-term ($h = 6$) and long-term forecasts ($h = 12$). The number of calculated forecasts ranges from $P = 113$ for $h = 1$ to $P = 102$ for $h = 12$. We employ two benchmark models, an AR(1) model which is always nested in (18) and an AR(p) model.

6 Results

In a first step, we report the uniform, boom and recession weighted MSFE for all indicator models and the autoregressive benchmark models (Table 2). As a general result, the average forecast errors based on the uniform weighting scheme are strongly driven by the forecast errors made during recessions which are substantially higher than during booms. This holds for all models and forecast horizons. It implies that improvements in terms of indicator construction and model building should aim at better predictions of recession periods.

Comparing the indicators, we find that their ranking in some—but by far not in all—cases differs considerably between boom and recession periods. For the short-term forecasts ($h = 1$), we observe that the EJ indicator ranks as number 1 or 2 in all weighting schemes. Also, the EC and ZEW indicators always rank as number 3 and 8, respectively. Hence the relative performance of these indicators is largely unaffected by the specific economic situation. On the other hand, the relative performance of the ESI and OECD indicators depend on whether a boom or a recession has to be predicted. While the ESI is particularly useful in recessions, the OECD indicator has its strengths in booms. Over-

all, it is reassuring that all indicator models outperform the AR(1) benchmark model. The differences are more pronounced for recession forecasts.

Forecasting six months ahead leads to a somewhat different picture. Now the OECD indicator uniformly outperforms its competitors by a noticeable amount. The FAZ indicator follows closely behind for boom forecasts but is much less suited for recession forecasts. In contrast, the ZEW indicator works well for recession forecasts but ranks only as number 7 for boom forecasts. The EJ indicator which performed well for the short horizon cannot be recommended for the 6-month horizon.

Looking at the 12-month forecasts, the AR(1) model becomes number 2 on average. It is only outperformed by the FAZ indicator. All the other indicators do not seem to add useful information to the simple autoregressive benchmark. This result is not very surprising because it is conventional wisdom that early indicators have little to say about the developments beyond a horizon of 6 months or so.

In practice, the choice of an appropriate indicator should depend on both the forecast horizon and on the specific loss function. Forecasters who particularly dislike forecast errors during recessions should use a slightly different set of indicators than forecasters who are more interested in correct boom prediction. For example, at the 1-month horizon the top three models for booms are (in this order) based on the OECD, EJ, and EC indicators while the top three models for recessions are based on the EJ, ESI, and EC indicators.

In a second step, the modified Diebold-Mariano test is used to check the significance of the above findings, see Tables 3 to 4. At the horizon of one month, only the EJ indicator significantly outperforms the AR(1) model for all weighting schemes. It is also significantly better than some of its competitors, particularly for uniform loss. At the horizon of six months, both the FAZ and the EC indicator can significantly outperform three of their competitors under uniform loss while the OECD indicator, which yields the smallest forecast errors, shows only one significant result. This picture changes considerably under boom loss, where both the OECD and FAZ indicators outperform the other indicators. For recession forecasts, the EC, ZEW, FAZ, and OECD indicators are indistinguishable by the Diebold-Mariano test, even though the differences in MSFE between, e.g., the OECD and the FAZ indicator are considerable. At the horizon of 12 months, no indicator is able to significantly dominate the benchmark models, not even the FAZ indicator that has smaller MSFE. Under boom weights, the FAZ indicator at least outperforms five of its competitors. As a negative result, for all horizons and weighting schemes, it is difficult to find indicators that significantly dominate all or most of their rivals. The CFI and the ZEW exhibit a particularly poor performance. However, we are careful with these results because, as argued before, there a several caveats to take into account. Therefore, we supplement the

Diebold-Mariano test and possibly qualify its results in the following.

The Clark-West test is computed to reassess the performance of the indicator models in comparison to the AR(1) benchmark. Since the modified Diebold-Mariano test is biased in favor of the nested AR(1) model, the results should be more in favor of the indicator models. In fact, for $h = 1$ we obtain the result that all indicator models significantly outperform the benchmark, see Table 5. For $h = 6$, again all indicator models dominate the benchmark (with one borderline case). This is in strong contrast to the results of the Diebold-Mariano test which are much more pessimistic with regard to the additional information content of the indicators. At the 12-months horizon the Diebold-Mariano test does not find a single indicator model that outperforms the benchmark, while the Clark-West test identifies the FAZ indicator as being significantly better for all weighting schemes.

The SPA test of Hansen takes into account that we are ultimately interested in comparing each of the models simultaneously to all its competitors. Pairwise significance as attested by the Diebold-Mariano test might be spurious in some cases. In Table 6, we test for each model the null hypothesis that it has equal predictive ability as all its competitors against the alternative that at least one other model yields more accurate predictions. For the AR(1) model the null is rejected at forecast horizons of 1 and 6 months which corresponds to the finding that it is difficult to beat only at the 12-month horizon. The general autoregressive model is even dominated at the 12-month horizon. Similarly, the ESI and CFI indicators are almost always outperformed by at leat one competitor and may therefore be safely disregarded in forecasting exercises. For the EJ indicator the previous result that it performs excellent in terms of MSFE at the 1-month horizon is confirmed as the null is only rejected at this very short horizon. The ZEW indicator is a borderline case with p-values between 0.07 and 0.13. The remaining three indicators (EC, FAZ, OECD) are—with one exception—not significantly dominated by any competitor, irrespective of the forecast horizon or the weighting scheme.

The results of the Mincer-Zarnowitz tests for forecast efficiency and unbiasedness yield mixed results, see Table 7. For the 1-month horizon, the null generally cannot be rejected for uniform and boom loss, while it is rejected for recession loss. This result is robust to excluding the most recent recession from the sample which implies that the simple linear one-indicator models leave some information unused for recession forecasts. For the 6-month horizon, this result is alleviated somewhat as there are some indicators that seem to yield efficient and unbiased forecasts even under recession loss. At the 12-month horizon, however, this null is again rejected in most cases. In general, the indicators that yield the best forecasts for a given horizon and weighting scheme—as identified by the MSFE—are not necessarily also efficient in terms of the Mincer-Zarnowitz regression suggesting

that adding further information to the forecast equations could lead to improved forecasts, especially for the 12-month horizon. To further investigate this, we augmented the Mincer-Zarnowitz regressions with seasonal dummies to account for potential seasonality and with lagged forecasts to account for neglected dynamics. However, they all turned out to be insignificant. We conclude that the leading indicators used in this paper are not sufficiently informative for rather long-term business cycle forecasts, which is also reflected by the finding that at the 12-month horizon it is extremely difficult to beat the simple AR(1) model.

Finally, we use the fluctuation tests to check the structural stability of the modified Diebold-Mariano (MDM) test results. In Figure 3, the MDM based fluctuation statistics for the horizon of $h = 1$ are displayed over the period from the beginning of 2003 to the middle of 2007 (remember that the statistics are centered so that the last 24 months of the sample cannot be considered). Each statistic refers to a pairwise test of the respective indicator model against the AR model. A value above the upper critical value indicates that the indicator is significantly more accurate than the benchmark while a value below the lower critical value indicates the opposite case. However, significant sub-samples seem rare. This is mainly due to the fact that to signal local significance of the MDM test, a higher critical value has to be crossed than to signal average significance as reported by the standard MDM test above. Nevertheless, within the interval of insignificance, we still observe some variance of the statistics. For example, under all loss functions the usefulness of the indicators seems to deteriorate in the sub-samples with midpoints in 2005 and the beginning of 2006. Thereafter, for sub-samples including the recent recession, the quality of the indicators improves again, especially under recession loss. This indicates that the simple benchmark might be better suited for rather tranquil times while the strength of the indicators is to contain early information on changes in the business cycle. This impression is, however, only weakly supported by the medium-term and long-term fluctuation tests, where the fluctuation tests do not detect much instability, see Figures 4 and 5.

7 Summary

In this paper we assessed the predictive abilities of seven widely recognized leading indicators for euro area industrial production. We went beyond standard forecast evaluation approaches in several respects, taking up recent methodological developments. We allowed for departures from the uniform symmetric quadratic loss function typically used in forecast evaluation exercises. Specifically, we overweighed forecast errors during periods of high or low growth rates to check how the indicators perform during booms and recessions,

i.e., in times of particularly high demand for good forecasts. It turned out that some indicators are well-suited for booms or recessions only while others are largely unaffected by the business cycle situation.

We also took the issue of nested models into account when comparing indicator models with a simple autoregressive benchmark. Unlike the standard Diebold-Mariano test, the test proposed by Clark and West (2007) identified all indicators as significantly outperforming the benchmark at short to medium-term forecast horizons. This result confirms the usefulness of the seven early indicators for euro area industrial production.

In order to prevent the problem of data snooping when searching for the best of the seven indicators by performing multiple pairwise tests, we implemented the test for superior predictive ability proposed by Hansen (2005). The results pointed to the existence of a group of three top indicators (EC, FAZ, OECD) that are generally not dominated by others. However, it is not possible to significantly discriminate between these three. For short-term forecasts, also the Business Climate Indicator (EJ) published by the European Commission performed excellent.

Mincer-Zarnowitz regressions revealed that most indicators yield efficient forecasts for short horizons under uniform and boom loss while they have particular problems for 1-year forecasts. This indicates that, at least in principle, further improvements are possible. This, however, is left for future research.

Finally, we implemented the fluctuation test introduced by Giacomini and Rossi (2008) to assess the forecasting stability of each model both on average and during booms and recessions. It indicated that the simple autoregressive benchmark model might be difficult to beat in rather tranquil times while the strength of the indicators is to contain early information on booms and recessions.

All in all, the results indicate that there is not one best indicator that uniformly dominates all its competitors. The optimal choice rather depends on the specific forecast situation and the loss function of the user. For 1-month forecasts the EJ and OECD indicators work well, for 6-month forecasts the OECD indicator performs very good by all criteria, and for 12-month forecasts the FAZ indicator is the only one that can beat the AR(1) model. Still, however, the Mincer-Zarnowitz regressions suggest that none of the indicator uses all available information.

References

Andrews, Donald W. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821–856.

Bodo, G., R. Golinelli, and G. Parigi (2000), "Forecasting industrial production in the euro area," *Empirical economics*, 25(4), 541–561.

Breitung, Jörg, and Doris Jagodzinski (2001), "Prognoseeigenschaften alternativer Indikatoren für die Konjunkturentwicklung in Deutschland," *Konjunkturpolitik*, 47, 292–314.

Caggiano, G., G. Kapetanios, and V. Labhard (2009), "Are more data always better for factor analysis? Results for the euro area, the six largest euro area countries and the UK," Working paper series, European Central Bank.

Clark, T.E., and M.W. McCracken (2009), "Averaging forecasts from VARs with uncertain instabilities," *Journal of Applied Econometrics*, 25(1), 5–21.

Clark, Todd E., and Michael W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85–110.

Clark, Todd E., and Kenneth D. West (2005), "Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis," *Journal of Econometrics*, 135, 155–186.

Clark, Todd E., and Kenneth D. West (2007), "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 138, 291–311.

Diebold, Francis X., and Roberto S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13(3), 253–263.

Elliott, G., I. Komunjer, and A. Timmermann (2008), "Biases in macroeconomic forecasts: irrationality or asymmetric loss?" *Journal of the European Economic Association*, 6(1), 122–157.

Fichtner, F., R. Ruffer, and B. Schnatz (2009), "Leading indicators in a globalised world," Working paper series, European Central Bank.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2003), "Do financial variables help forecasting inflation and real activity in the euro area?" *Journal of Monetary Economics*, 50(6), 1243–1255.

Giacomini, Raffaella, and Barbara Rossi (2008), "Forecasting Comparisons in Unstable Environments," Working Paper 08–04, Duke University, Department of Economics.

Hansen, Peter Reinhard (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, 23(4), 365–380.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold (1997), "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 13, 281–291.

Marcellino, Massimiliano (2008), "A linear benchmark for forecasting GDP growth and inflation?" *Journal of Forecasting*, 27(4), 305–340.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2003), "Macroeconomic forecasting in the Euro area: Country specific versus area-wide information," *European Economic Review*, 47(1), 1–18.

Milas, Costas, and Philip Rothman (2008), "Out-of-sample forecasting of unemployment rates with pooled STVECM forecasts," *International Journal of Forecasting*, 24(1), 101–121.

Mincer, J., and V. Zarnowitz (1969), *The Evaluation of Economic Forecasts in J. Mincer (ed.), Economic Forecasts and Expectations*, National Bureau of Economic Research, New York.

Ozyildirim, A., B. Schaitkin, and V. Zarnowitz (2010), "Business cycles in the euro area defined with coincident economic indicators and predicted with leading economic indicators," *Journal of Forecasting*, 29(1-2), 6–28.

Politis, Dimitris N., and Joseph P. Romano (1994), "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89(428), 1303–1313.

Rossi, B., and T. Sekhposyan (2010), "Have economic models' forecasting performance for US output growth and inflation changed over time, and when?" *International Journal of Forecasting*, In press.

van Dijk, Dick, and Philip Hans Franses (2003), "Selecting a Nonlinear Time Series Model using Weighted Tests of Equal Forecast Accuracy," *Oxford Bulletin of Economics and Statistics*, 65, 727–744.

White, Halbert (2000), "A reality check for data snooping," *Econometrica*, 68(5), 1097–1126.

Zarnovitz, Victor (1985), "Rational expectations and macroeconomic forecasts," *Journal of Business and Economic Statistics*, 3, 293–311.

Table 1: Overview over the euro area indicators

Indicator	**Components**	**Source**
European Sentiment Indicator (ESI)	Industry Confidence Indicator, Services Confindence Indicator Consumer Confidence Indicator (CFI) Construction Confidence Indicator Retail Trade Confidence Indicator	European Commission
Consumer Confidence Indicator (CFI)	Consumer surveys	European Commission
Business Climate Indicator (EJ)	Industry survey about: production trends in recent months, order books export order books, stocks and production expectations	European Commission
FAZ-Euro-Indicator (FAZ)	New job vacancies, order entries, Reuter purchasing manager´s index (PMI), building and planning permissions, production, interest rate spread, cosumer confidence, Morgan-Stanley- Capital-International Index, real money (M3)	DZ-Bank
OECD Composite Leading Indicator (OECD)	Composite by individual OECD indicators for EU-12: variables for surveys by national institutes, new job vacancies, orders inflow/demand, spread of interest rates, production, finished goods stocks, passenger car registration, other national indicators	Organisation for Economic Co-operation and Development (OECD)
ZEW Indicator of Economic Sentiment (ZEW)	Medium-term expectations for development of the macroeconomic trend, inflation rate, short-term and long-term interest rates, stockmarket, exchange rates, profit situation of different German industries (only financial experts)	Centre for European Economic Research (ZEW)
EuroCoin (EC)	Data from 11 categories: industrial production, producer prices, monetary aggregates, interest rates, financial variables, exchange rates, surveys by the European Commission, surveys by national institutes, external trade, labour market	Centre for Economic Policy Research (CEPR)

Table 2: Root Mean Squared Forecast Errors

	Uniform		Boom		Recession	
	MSE	Rank	MSE	Rank	MSE	Rank
			$h = 1$			
AR(1)	0.016	9	0.013	9	0.019	9
AR	0.015	6	0.012	4	0.018	6
ESI	0.014	4	0.012	5	0.017	2
EJ	0.013	1	0.011	2	0.015	1
CFI	0.015	5	0.012	7	0.017	5
EC	0.014	3	0.012	3	0.017	3
ZEW	0.016	8	0.013	8	0.018	8
FAZ	0.015	7	0.012	6	0.018	7
OECD	0.014	2	0.011	1	0.017	4
			h=6			
AR(1)	0.051	8	0.022	4	0.069	8
AR	0.048	6	0.022	6	0.063	6
ESI	0.050	7	0.025	8	0.065	7
EJ	0.047	5	0.022	5	0.062	5
CFI	0.054	9	0.027	9	0.071	9
EC	0.041	2	0.020	3	0.055	3
ZEW	0.042	3	0.025	7	0.055	2
FAZ	0.044	4	0.017	2	0.060	4
OECD	0.034	1	0.016	1	0.045	1
			$h = 12$			
AR(1)	0.063	2	0.031	4	0.084	2
AR	0.067	7	0.034	6	0.089	7
ESI	0.071	8	0.037	7	0.093	9
EJ	0.067	6	0.038	8	0.087	5
CFI	0.072	9	0.042	9	0.093	8
EC	0.064	3	0.030	2	0.086	3
ZEW	0.066	4	0.034	5	0.086	4
FAZ	0.061	1	0.028	1	0.081	1
OECD	0.066	5	0.031	3	0.088	6

Notes: This Table reports the root MSFEs and the corresponding ranking for each forecasting horizon and weighting scheme.

Table 3: Modified Diebold-Mariano test for uniform weights ($w_t = 1$)

$h=1$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)										0	1
AR	1.285 (0.201)									0	1
ESI	1.473 (0.143)	0.812 (0.419)								0	1
EJ	2.655 (0.009)	2.171 (0.032)	1.918 (0.058)							6	0
CFI	1.302 (0.196)	0.542 (0.589)	−0.653 (0.515)	−2.368 (0.020)						0	1
EC	1.652 (0.101)	0.852 (0.396)	0.143 (0.886)	−1.120 (0.265)	0.603 (0.548)					0	0
ZEW	0.588 (0.557)	−0.487 (0.627)	−0.898 (0.371)	−1.939 (0.055)	−0.629 (0.531)	−0.998 (0.320)				0	1
FAZ	0.999 (0.320)	−0.202 (0.840)	−0.783 (0.435)	−1.749 (0.083)	−0.416 (0.678)	−1.212 (0.228)	0.331 (0.741)			0	1
OECD	1.631 (0.106)	0.845 (0.400)	0.184 (0.854)	−1.046 (0.298)	0.639 (0.524)	0.083 (0.934)	0.979 (0.330)	0.988 (0.325)		0	0
$h=6$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)										0	1
AR	0.961 (0.339)									1	1
ESI	0.469 (0.640)	−1.315 (0.191)								1	2
EJ	1.122 (0.264)	0.993 (0.323)	1.377 (0.171)							1	0
CFI	−0.860 (0.392)	−1.739 (0.085)	−1.684 (0.095)	−1.706 (0.091)						0	6
EC	1.500 (0.137)	1.695 (0.093)	1.858 (0.066)	1.636 (0.105)	1.884 (0.062)					3	0
ZEW	1.161 (0.248)	1.193 (0.235)	1.417 (0.159)	1.070 (0.287)	1.547 (0.125)	−0.618 (0.538)				0	0
FAZ	1.711 (0.090)	1.151 (0.252)	2.042 (0.044)	0.814 (0.418)	2.493 (0.014)	−0.745 (0.458)	−0.344 (0.732)			3	0
OECD	1.456 (0.148)	1.499 (0.137)	1.631 (0.106)	1.444 (0.152)	1.715 (0.089)	1.273 (0.206)	1.590 (0.115)	1.165 (0.247)		1	0
$h=12$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)										0	0
AR	−1.066 (0.289)									0	0
ESI	−1.261 (0.210)	−1.403 (0.164)								0	1
EJ	−1.419 (0.159)	0.033 (0.974)	0.848 (0.399)							0	1
CFI	−1.613 (0.110)	−1.431 (0.156)	−0.372 (0.711)	−1.054 (0.294)						0	3
EC	−0.418 (0.677)	1.340 (0.183)	1.552 (0.124)	1.426 (0.157)	1.944 (0.055)					1	0
ZEW	−0.675 (0.501)	1.049 (0.296)	1.511 (0.134)	0.858 (0.393)	1.741 (0.085)	−0.492 (0.624)				1	1
FAZ	0.789 (0.432)	1.651 (0.102)	1.705 (0.091)	1.922 (0.057)	2.089 (0.039)	0.980 (0.329)	1.800 (0.075)			4	0
OECD	−0.617 (0.539)	0.607 (0.545)	1.127 (0.262)	0.493 (0.623)	1.292 (0.199)	−0.415 (0.679)	−0.048 (0.962)	−1.091 (0.278)		0	0

Notes: For each pair of models the modified DM test statistic is reported together with the two-sided p-value in brackets below. A negative sign indicates that the MSFE of column model is smaller than that of the row model and vice versa. The last two columns count the number of times the row model significantly (to the level of 10%) outperforms its competitors (column "+") and is outperformed by its competitors (column "-").

Table 4: Modified Diebold-Mariano test for boom and recession weights

										recession		boom	
$h=1$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-	+	-
AR(1)		−2.095 (0.038)	−1.493 (0.138)	−2.437 (0.016)	−1.500 (0.136)	−1.885 (0.062)	−1.022 (0.309)	−1.455 (0.149)	−2.342 (0.021)	0	1	0	4
AR	0.421 (0.675)		0.271 (0.787)	−1.111 (0.269)	0.982 (0.328)	−0.171 (0.865)	1.348 (0.180)	0.315 (0.753)	−1.295 (0.198)	0	1	1	0
ESI	1.089 (0.279)	1.058 (0.292)		−1.537 (0.127)	0.349 (0.727)	−0.699 (0.486)	0.746 (0.457)	0.056 (0.956)	−1.375 (0.172)	0	0	0	0
EJ	2.003 (0.048)	2.084 (0.039)	1.654 (0.101)		1.789 (0.076)	0.839 (0.403)	1.777 (0.078)	1.469 (0.145)	−0.147 (0.883)	3	0	3	0
CFI	0.873 (0.384)	1.059 (0.292)	−0.640 (0.523)	−2.050 (0.043)		−0.789 (0.432)	0.461 (0.646)	−0.275 (0.784)	−1.751 (0.083)	0	1	0	2
EC	1.141 (0.256)	0.923 (0.358)	−0.019 (0.985)	−0.974 (0.332)	0.416 (0.678)		1.152 (0.252)	0.684 (0.496)	−0.865 (0.389)	0	0	1	0
ZEW	0.140 (0.889)	−0.180 (0.857)	−0.784 (0.434)	−1.559 (0.122)	−0.566 (0.573)	−0.783 (0.435)		−0.679 (0.499)	−1.791 (0.076)	0	0	0	2
FAZ	0.322 (0.748)	−0.106 (0.916)	−0.842 (0.402)	−1.546 (0.125)	−0.570 (0.570)	−1.155 (0.251)	0.109 (0.913)		−1.365 (0.175)	0	0	0	0
OECD	0.843 (0.401)	0.591 (0.556)	−0.314 (0.754)	−1.243 (0.216)	0.119 (0.905)	−0.505 (0.615)	0.564 (0.574)	0.686 (0.494)		0	0	3	0

										recession		boom	
$h=6$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-	+	-
AR(1)		0.156 (0.876)	0.920 (0.360)	0.143 (0.886)	1.413 (0.161)	−0.870 (0.386)	0.829 (0.409)	−1.866 (0.065)	−2.376 (0.019)	0	0	0	2
AR	1.067 (0.288)		1.694 (0.093)	−0.069 (0.945)	2.062 (0.042)	−1.068 (0.288)	1.006 (0.317)	−3.595 (0.000)	−2.566 (0.012)	0	0	2	2
ESI	0.789 (0.432)	−1.037 (0.302)		−1.312 (0.192)	1.531 (0.129)	−1.728 (0.087)	−0.106 (0.916)	−3.136 (0.002)	−2.605 (0.010)	0	1	0	4
EJ	1.205 (0.231)	1.222 (0.224)	1.231 (0.221)		1.765 (0.080)	−1.314 (0.192)	0.992 (0.323)	−2.557 (0.012)	−2.681 (0.008)	0	0	1	2
CFI	−0.563 (0.575)	−1.467 (0.145)	−1.506 (0.135)	−1.492 (0.139)		−2.118 (0.036)	−0.757 (0.451)	−3.248 (0.002)	−2.788 (0.006)	0	2	0	5
EC	1.492 (0.139)	1.651 (0.102)	1.704 (0.091)	1.562 (0.121)	1.692 (0.094)		1.671 (0.098)	−1.668 (0.098)	−2.363 (0.020)	2	0	3	2
ZEW	1.315 (0.191)	1.415 (0.160)	1.481 (0.141)	1.316 (0.191)	1.515 (0.133)	0.034 (0.973)		−2.332 (0.022)	−2.635 (0.010)	0	0	0	3
FAZ	1.566 (0.120)	0.843 (0.401)	1.555 (0.123)	0.504 (0.615)	2.015 (0.046)	−0.962 (0.338)	−0.747 (0.457)		−0.538 (0.592)	1	0	7	0
OECD	1.380 (0.171)	1.376 (0.172)	1.446 (0.151)	1.313 (0.192)	1.516 (0.132)	1.125 (0.263)	1.198 (0.234)	1.168 (0.246)		0	0	7	0

										recession		boom	
$h=12$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-	+	-
AR(1)		0.931 (0.354)	1.341 (0.183)	1.469 (0.145)	1.447 (0.151)	−0.141 (0.888)	0.842 (0.402)	−0.657 (0.512)	−0.064 (0.949)	0	0	0	0
AR	−1.034 (0.303)		1.640 (0.104)	0.907 (0.367)	1.119 (0.266)	−1.261 (0.210)	−0.118 (0.906)	−2.384 (0.019)	−1.150 (0.253)	0	0	0	1
ESI	−1.165 (0.247)	−1.245 (0.216)		0.246 (0.806)	0.765 (0.446)	−1.731 (0.086)	−1.383 (0.170)	−2.651 (0.009)	−1.833 (0.070)	0	0	0	3
EJ	−1.011 (0.314)	0.819 (0.414)	1.188 (0.238)		0.387 (0.699)	−1.595 (0.114)	−1.050 (0.296)	−1.732 (0.086)	−1.512 (0.134)	0	0	0	1
CFI	−1.405 (0.163)	−1.507 (0.135)	0.106 (0.916)	−1.458 (0.148)		−1.578 (0.118)	−1.107 (0.271)	−1.885 (0.062)	−1.501 (0.136)	0	2	0	1
EC	−0.527 (0.599)	1.264 (0.209)	1.399 (0.165)	0.881 (0.380)	1.790 (0.077)		1.104 (0.272)	−0.963 (0.338)	0.041 (0.968)	1	0	1	0
ZEW	−0.597 (0.552)	1.117 (0.267)	1.374 (0.172)	0.464 (0.643)	1.751 (0.083)	−0.217 (0.829)		−1.846 (0.068)	−0.997 (0.321)	1	0	0	1
FAZ	0.731 (0.466)	1.330 (0.186)	1.393 (0.167)	1.297 (0.197)	1.635 (0.105)	0.883 (0.379)	1.359 (0.177)		1.077 (0.284)	0	0	5	0
OECD	−0.721 (0.473)	0.362 (0.718)	0.879 (0.382)	−0.127 (0.899)	0.927 (0.356)	−0.454 (0.651)	−0.307 (0.760)	−1.006 (0.317)		0	0	1	0

Notes: See notes in Table 3. The lower triangular reports the results for recession weights (w_{left}) and the upper for the boom weights (w_{right}).

Table 5: Results of the Clark-West Test

	Uniform			Boom			Recession		
	$h=1$	$h=6$	$h=12$	$h=1$	$h=6$	$h=12$	$h=1$	$h=6$	$h=12$
AR	0.000	0.028	0.788	0.000	0.035	0.834	0.000	0.024	0.727
ESI	0.000	0.013	0.776	0.000	0.025	0.828	0.000	0.009	0.692
EJ	0.000	0.024	0.750	0.000	0.023	0.729	0.000	0.026	0.759
CFI	0.000	0.055	0.743	0.000	0.114	0.678	0.000	0.056	0.795
EC	0.000	0.018	0.368	0.000	0.015	0.409	0.000	0.022	0.331
ZEW	0.000	0.039	0.481	0.001	0.037	0.546	0.000	0.041	0.411
FAZ	0.000	0.002	0.035	0.000	0.002	0.043	0.002	0.004	0.032
OECD	0.000	0.040	0.442	0.000	0.032	0.517	0.000	0.046	0.365

Notes: Table reports p-values for the one-sided modified Clark-West test. A small p-value indicates that the row indicator has a significantly smaller MSE than the nested AR(1) benchmark model.

Table 6: Results of the Hansen Test for Superior Predictive Ability

	Uniform			Boom			Recession		
	$h=1$	$h=6$	$h=12$	$h=1$	$h=6$	$h=12$	$h=1$	$h=6$	$h=12$
AR1	0.04	0.04	0.04	0.04	0.04	0.04	0.15	0.28	0.39
AR	0.08	0.09	0.11	0.06	0.06	0.06	0.07	0.07	0.07
ESI	0.07	0.10	0.11	0.04	0.05	0.05	0.04	0.04	0.04
EJ	0.69	0.69	0.98	0.05	0.05	0.07	0.05	0.05	0.05
CFI	0.01	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02
EC	0.30	0.33	0.50	0.08	0.11	0.18	0.19	0.25	0.39
ZEW	0.11	0.13	0.13	0.07	0.07	0.09	0.09	0.09	0.13
FAZ	0.19	0.21	0.23	0.12	0.12	0.23	0.72	0.72	0.97
OECD	0.28	0.33	0.42	0.64	0.64	0.97	0.20	0.22	0.26

Notes: Reported are p-values of SPA tests with the null hypothesis that the row model has equal predictive ability as all its competitor models against the alternative that at least one competitor yields more accurate predictions.

Table 7: Results of the Mincer-Zarnowitz Regression Test

	Uniform			Boom			Recession		
	$h=1$	$h=6$	$h=12$	$h=1$	$h=6$	$h=12$	$h=1$	$h=6$	$h=12$
AR1	0.417	0.272	0.000	0.157	0.122	0.913	0.023	0.220	0.015
AR	0.217	0.218	0.000	0.060	0.626	0.370	0.003	0.157	0.061
ESI	0.485	0.214	0.000	0.669	0.188	0.007	0.132	0.155	0.026
EJ	0.024	0.214	0.000	0.722	0.826	0.002	0.023	0.120	0.040
CFI	0.060	0.090	0.000	0.821	0.007	0.000	0.006	0.103	0.007
EC	0.112	0.094	0.000	0.954	0.397	0.623	0.027	0.001	0.071
ZEW	0.380	0.313	0.000	0.040	0.560	0.111	0.003	0.034	0.004
FAZ	0.139	0.036	0.043	0.727	0.041	0.945	0.018	0.010	0.000
OECD	0.018	0.001	0.000	0.419	0.174	0.701	0.009	0.000	0.020

Notes: Table reports p-values for the Mincer-Zarnowitz regression test. A small p-value indicates that for the row indicator the null hypothesis of rationality is rejected.

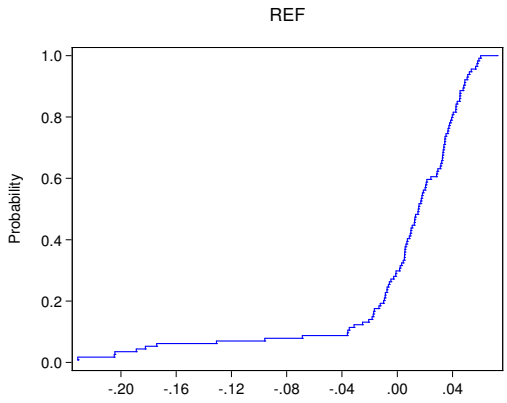

Figure 1: Empirical Cumulative Distribution Function $\widehat{F}(y_t)$

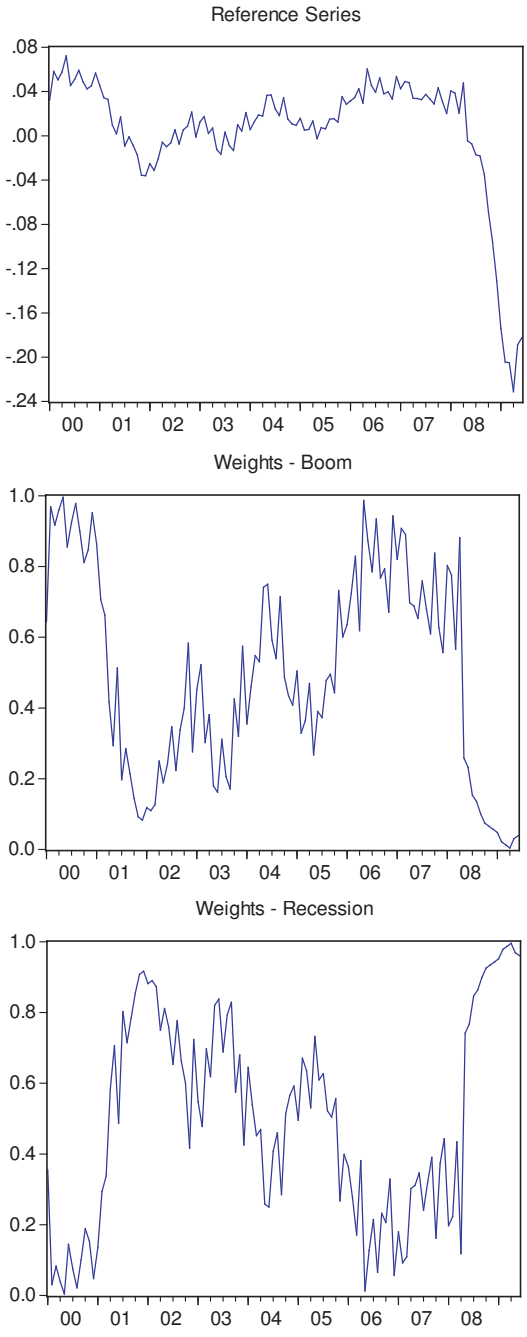

Figure 2: Reference Series and Weights

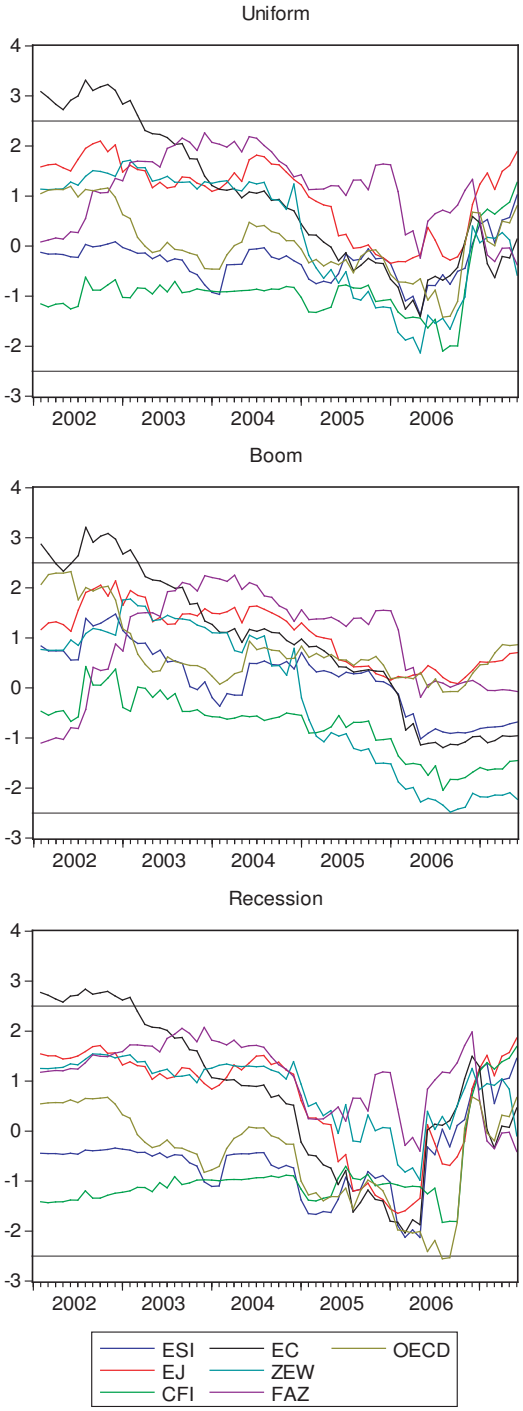

Figure 3: Fluctuation MDM test for $h = 1$ against the AR benchmark

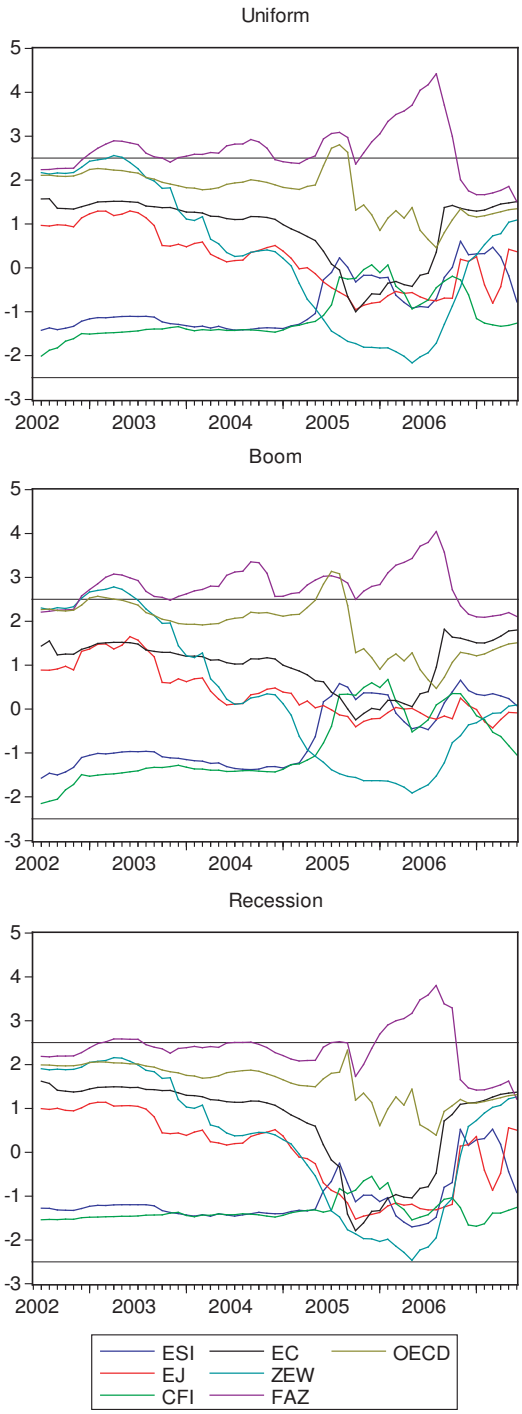

Figure 4: Fluctuation MDM test for $h = 6$ against the AR benchmark

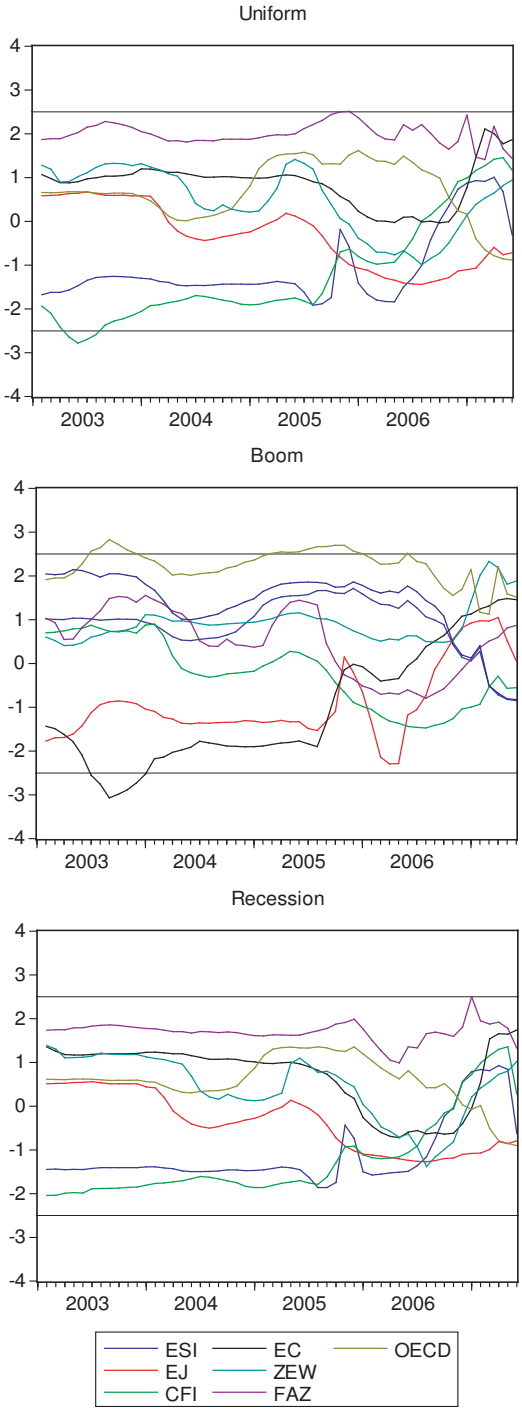

Figure 5: Fluctuation MDM test for $h = 12$ against the AR benchmark