ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Hutter, Christian; Weber, Enzo

Article — Accepted Manuscript (Postprint)

Constructing a new leading indicator for unemployment from a survey among German employment agencies

Applied Economics

Provided in Cooperation with:
Institute for Employment Research (IAB)

Suggested Citation: Hutter, Christian; Weber, Enzo (2015) : Constructing a new leading indicator for unemployment from a survey among German employment agencies, Applied Economics, ISSN 1466-4283, Taylor & Francis, London, Vol. 47, Iss. 33, pp. 3540-3558, https://doi.org/10.1080/00036846.2015.1018672 , https://www.tandfonline.com/doi/full/10.1080/00036846.2015.1018672

This Version is available at:
https://hdl.handle.net/10419/263252

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.


http://creativecommons.org/licenses/by-nc/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

This is an Accepted Manuscript version of the following article, accepted for publication in Applied Economics. Christian Hutter & Enzo Weber (2015) Constructing a new leading indicator for unemployment from a survey among German employment agencies, Applied Economics, 47:33, 3540-3558, DOI: 10.1080/00036846.2015.1018672.
It is deposited under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Constructing a New Leading Indicator for Unemployment from a Survey among German Employment Agencies

December 15, 2014

Abstract

The paper investigates the predictive power of a new survey implemented by the Federal Employment Agency (FEA) for forecasting German unemployment in the short run. Every month, the CEOs of the FEA's regional agencies are asked about their expectations of future labor market developments. We generate an aggregate unemployment leading indicator that exploits serial correlation in response behavior through identifying and adjusting temporarily unreliable predictions. We use out-of-sample tests suitable in nested model environments to compare forecasting performance of models including the new indicator to that of purely autoregressive benchmarks. For all investigated forecast horizons (1, 2, 3 and 6 months), test results show that models enhanced by the new leading indicator significantly outperform their benchmark counterparts. To compare our indicator to potential competitors we employ the model confidence set. Results reveal that models including the new indicator perform very well at the 10 percent level.

JEL classification: C22, C52, C53, E24

Keywords: survey data, forecast evaluation, nested models, model confidence set, unemployment

1 Introduction

Since 2005, the German labor market has been attracting growing attention, especially after proving its capability of resistance during the Great Recession. Contrary to many other industrialized countries, unemployment figures were only slightly affected. Continued astonishment about this development showed that there was - and still is - a certain lack of leading indicators from which labor market experts could reliably deduce Germany's unemployment development. All in all, the German labor market is a rather neglected field with respect to leading indicators for aggregate unemployment. Although some surveys, among them the ifo employment barometer published by the ifo institute for economic research in Munich, have the potential for an indirect leading indicator since the original target variable correlates to some extent with unemployment, only few resources seem to be invested in searching and finding a leading indicator that *directly* aims at signaling unemployment changes. As a consequence, there is only little literature on forecasting German unemployment. [1]

Our main contribution for improving the current situation is to construct the first leading indicator in Germany that explicitly aims at forecasting German unemployment in the short run. For this purpose, we exploit a new survey conducted by the German Federal Employment Agency (FEA) among the CEOs or appropriate persons in consultation with the upper management of the regional employment agencies. The survey is conducted every month and forms an innovative data set where concentrated local labor market expertise is collected in a way that consistently assures a 100 percent response rate. We generate an aggregate unemployment indicator going beyond a simple weighted average of all agencies: particularly, we employ an appropriate distinction of temporarily reliable and temporarily unreliable agencies in order to effectively exploit information about serial correlation in the response behavior. This might be a promising direction also for other forecast indices constructed from disaggregate data.

In addition, we investigate the predictive power of the resulting unemployment leading indicator, focussing on out-of-sample tests for equal predictive accuracy in population. Since the investigated autoregressive benchmark models are nested in the larger models including the novel unemployment leading indicator, we follow Clark and West (2007) and use out-of-sample tests that control for the nested model environment. The results reveal that models enhanced by the new indicator significantly outperform purely autoregressive benchmark models with respect to their out-of-sample performance.

Furthermore, the new unemployment indicator is compared to other established leading indicators such as the ifo employment barometer, order inflow and registered vacancies. We employ an approach recently established by Hansen et al. (2011) to compare predictive accuracy of a large number of models: the model confidence set (MCS). For all four investigated forecast horizons (1-, 2-,

[1] For instance, Schanne et al. (2010) use spatial GVAR models to forecast unemployment for all 176 German labor market districts. Askitas and Zimmermann (2009) suggest using data on Internet activity for forecasting German unemployment.

3 and 6-months ahead), we find that specifications which include the novel unemployment leading indicator not only survive the selection procedure but also dominate the model confidence sets.

The remainder of the paper is structured as follows: The new FEA survey is introduced in section 2. Section 3 describes our method to construct the novel unemployment leading indicator. The first part of section 4 compares the forecasting performance of the new indicator to that of pure benchmark models using nested model out-of-sample tests. The second subsection briefly describes alternative labor market indicators and investigates the new indicator's *relative* forecasting power with the help of the MCS. Robustness checks are presented in section 5. The last section concludes.

2 The FEA survey

In fall 2008, the Federal Employment Agency (FEA) started a survey in which the CEOs or appropriate persons of the top-level staff in the local employment agencies report - at a monthly frequency - their expectations of future labor market developments. In total, there are 176 of these local offices that are responsible for implementing the tasks of the FEA on the regional level. The original aim of this survey was to have an early warning system for the labor market during economic turmoil such as the Great Recession of 2008/2009. That is why some of the questions were explicitly designed for economic crises and adapted later on when the economy was recovering again. The most promising question with respect to our goal of constructing a novel leading indicator for German unemployment is the following:

What is your overall expectation for the development of unemployment in your district within the next three months - beyond the usual seasonal pattern?

The question has been continuously available in an unchanged format since the beginning of the survey. Five possible answers are provided (decline strongly, decline, stay constant, increase, increase strongly). In order to move from this ordinal Likert scale to a metric indicator the answers are translated into integers between -2 and 2.[2] The questionnaires can be answered within a time period of several days around mid-month when unemployment and other FEA statistics are counted. Furthermore, there is no option for abstaining or responding indecisively. As a consequence, all respondents have to provide an analyzable answer so that the response rate is consistently 100 percent. Therefore, biases due to panel mortality or a changing response rate can be ruled out. Seasonally adjusted unemployment figures needed to construct the indicator (see section 3) and to evaluate its forecasting performance (see section 4) are provided by the FEA statistics. Data for our study go from November 2008 to June 2012, resulting in 44 observations for all 176 local agencies.

There are several advantages over other macroeconomic surveys. The answering CEOs or appropriate persons in consultation with the upper manage-

[2]We abstain from other ways of conversion in which the distances between the answer options are not treated equally.

ment of a local agency can be considered experts in assessing the particular labor market structure and unemployment development in a certain district. Their specific knowledge has the potential of a leading indicator. For example, employees who expect losing their job are legally obliged to notify the respective regional agency at least three months in advance. Similarly, the local office should be among the first to know when unemployed persons find a job again. In addition, it has insight in meaningful district-level data that can provide relevant signals such as consulting requests and applications for transition companies or short-time work. The answering CEOs should be able to evaluate the district-specific relevance of these signaling variables for unemployment development. Another important source of information stems from periodical meetings with the upper management of neighboring local agencies on announced or expectable labor market developments in the regions. This is especially relevant since a company's foreclosure or opening could be an early available (un)employment signal even if it occurs in an adjacent district. The regional agencies not only stay in touch with neighboring agencies but also with local companies, alliances and chambers (of commerce) all of which might provide useful information. In addition to these local information sources, the answers are most likely influenced as well by information available at the aggregate level since these data are relevant for the region's development, too, as well as easy to collect and permanently present in daily news.

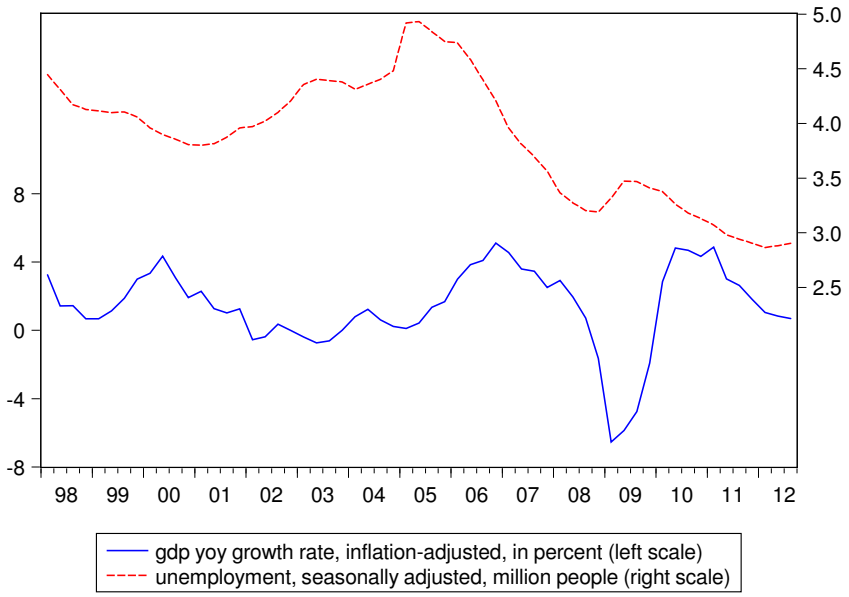
For instance, the unprecedented fall of the GDP during the Great Recession (see solid line in figure 1) in combination with a record level of applications and use of short-time work might have provoked extremely negative expectations among labor market forecasters. However, the deep slump of GDP was not followed by a huge increase of unemployed figures as most experts concluded from previous recessions. Contrariwise, sharply decreasing productivity and working time per capita lessened the impact on unemployment figures and contributed substantially to the German job miracle[3] - as we know today with the benefit of hindsight (see figure 1). Since the first part of our sample is highly dominated by an unprecedented recession with unexpectedly mild consequences for unemployment, it is important not to over-emphasize absolute forecasting performance but to keep an eye on how the respondents of the FEA survey perform in comparison with other established leading indicators during the same period (which is done in subsection 4.2). Plausible information sources, the respondents' expertise in proceeding early signals in a way specific to and suitable for the respective local agency as well as the consistently high response rate are good reasons to expect that an indicator using these survey data could have promising leading indicator properties.

3 The novel unemployment leading indicator

Exploiting survey data collected at a disaggregated level can work by adding up regional forecasts or forecasting on an aggregate level with the help of a single

[3]For a deeper investigation, see Möller (2010)

Figure 1: German GDP year-on-year growth rate and unemployment



indicator that efficiently combines all regional information. Since the underlying paper is not interested in modeling regions (as it is done in Schanne et al. (2010), for instance) but rather in constructing a new unemployment leading indicator for Germany, our focus is on the aggregate approach. As a consequence, the natural way to condense the answers from the local agencies is to average over all cross-section units using some sort of weights in order to account for the different sizes of the local agencies. As weights we tested seasonally adjusted unemployment at the district level and the so-called reference group, i.e. the denominator of the local unemployment rate, a somewhat broader figure that approximately captures all employed and unemployed persons in the respective regional district. Both weights vary over time although the latter is usually adapted only once a year.

A typical feature of many survey-based indicators (for instance, the prominent business climate index published by the ifo institute for economic research in Munich) is that, although the questions explicitly exclude seasonal effects, some seasonal pattern is left in the answers. We found that this is the case for our data, too. We use the standard X.12 ARIMA procedure to adjust for seasonality.

We define a "conventional" approach that uses local unemployment expectations collected at time point t from all agencies, resulting in an unemployment

indicator denoted I_t^{all}. Codified integer assessments ϕ_{it} ranging from -2 to 2 from all 176 agencies are averaged using time varying weights ω_{it}^j where $j = 1, 2$ denotes the two investigated weighting variables: either seasonally adjusted unemployment u_{it} $(j = 1)$ or the reference group r_{it} $(j = 2)$.

$$I_t^{all,j} = \sum_{i=1}^{176} \omega_{it}^j \cdot \phi_{it} \tag{1}$$

Figure 2 shows the resulting unemployment leading indicators that have been created according to equation (1) using unemployment (dashed line) or the reference group (solid line) as weights. Evidently, the choice of weights makes almost no difference. Due to aggregation of integer values, the resulting indicators can take virtually any value between -2 and 2 and have a natural line of zero which means no change in aggregate unemployment is expected within the next three months. Values above (below) zero indicate negative (positive) labor market expectations, hence rising (falling) unemployment. When compared to the aggregate unemployment level during that period (see solid line in figure 3), one can see that at first glance the indicators seem to lead unemployment by three to four months which is also supported by an analysis of the respective cross-correlogram. Since the survey explicitly asks for expected unemployment *changes*, we depict year-on-year and quarterly changes of aggregate seasonally adjusted unemployment, too (see dotted and dashed lines in figure 3). A second look reveals that the surveyed agencies had been too pessimistic on average, especially so in 2009 after the financial crisis.[4] Of course, put in context of the exceptional circumstances prevailing at that time, pessimistic assessments were dominant among other labor market professionals as well.

The following paragraphs focus on finding and exploiting certain non-erratic patterns in response behavior in order to improve leading indicator properties of the new unemployment barometer. A major advantage of our survey data is a consistent response rate of 100 percent which allows in-depth investigations such as monthly reliability checks of the agencies' labor market assessments. A first tendency check pools all cases in which the agencies report *increase* and *increase strongly* (*decline* and *decline strongly*) and treats it as correct prediction each time seasonally adjusted unemployment in the respective districts has increased (declined) after three months. The report *stay constant* is considered as correct prediction if the unemployment change in the respective district has not exceeded 2.5 percent in absolute value.[5]

We then investigate the number of correct and incorrect predictions covering all 7040 observations[6], conditional on the accuracy of the previous month's predictions. In 3273 cases, a correct forecast is followed by another correct forecast, compared to only 971 cases in which the agencies' expectations turned out to be false given they had been correct the month before. Similarly, there

[4] The finding of too pessimistic reports is also supported by Schanne (2012).

[5] Schanne (2012) uses the same critical threshold of 2.5 percent.

[6] 176 cross section units, 44 observations over time, minus 3 months needed to wait for the first evaluation, minus 1 month due to conditioning on the previous month

Figure 2: Monthly unemployment leading indicator, conventional weighting approach

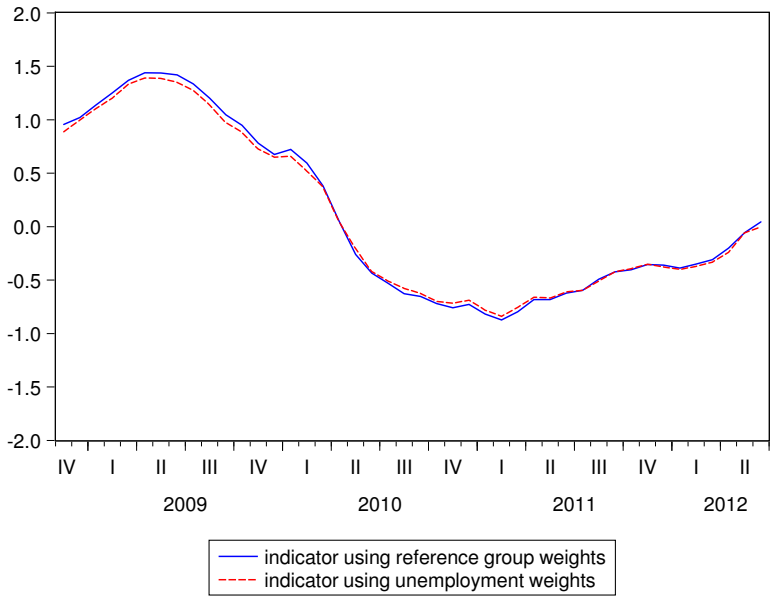
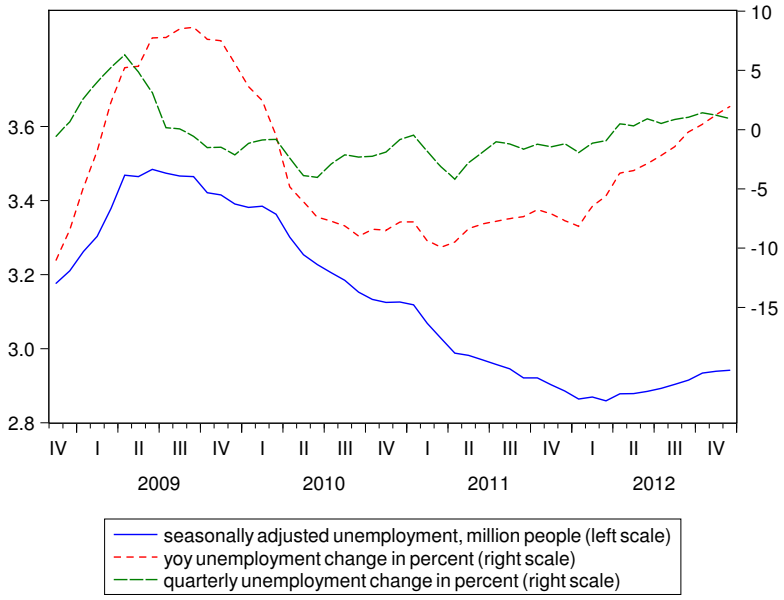


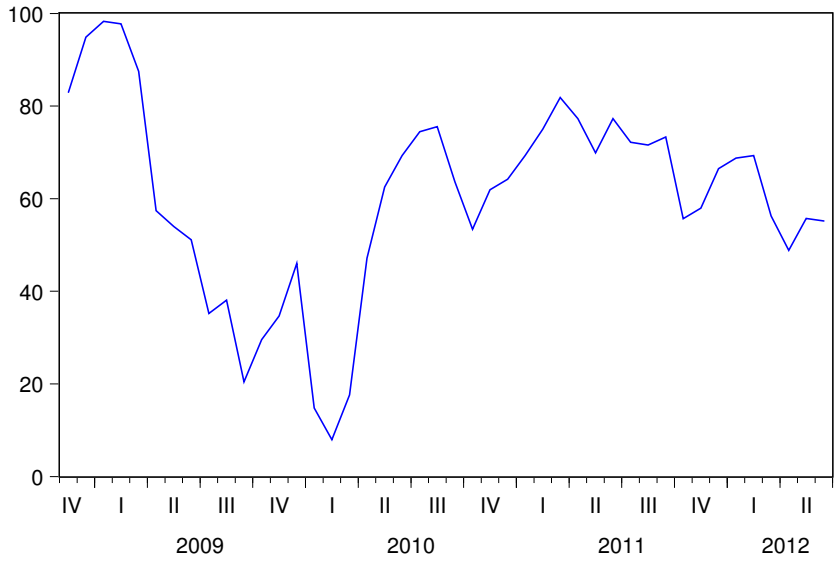
Figure 3: Unemployment, quarterly and year-on-year unemployment changes



are only 924 cases in which a wrong expectation remains a singularity, whereas in 1872 cases it is followed by another wrong prediction. In other words: It is more than twice as likely to get a wrong prediction instead of a correct one given the respective agency has reported a wrong tendency the month before, and 3.5 times as likely to get a correct expectation compared to a false one conditional on a correct previous-month report.

In addition to investigating the aggregate response pattern, we consider its development over time. Figure 4 displays the share of agencies that correctly reported the tendency of unemployment development. Note that the line begins with a delay of three months since one has to wait until February 2009 until the predictions made in November 2008 can be evaluated. On average, the hit count is more than 60 percent. However, there is considerable variation over time. Figure 4 clearly shows that regional agencies had trouble predicting the development of unemployment correctly in the aftermath of the Great Recession. As argued in section 2, regional labor market experts obviously overestimated the impact of the slump in GDP on German unemployment figures. However, the share of correctly reporting agencies seems to be considerably higher during non-recession times. In summary, correct or wrong assessments tend to accumulate at certain times, switching only gradually between periods of collective reliability and unreliability.

Figure 4: Share of agencies reporting correct unemployment tendencies (in percent)



Our findings clearly support the presence of serial correlation in response accuracy. The remainder of this section discusses the question of how to effectively use information on reliability of the agencies' forecasts in order to generate an aggregate leading indicator. We implement an appropriate distinction of *temporarily reliable* and *temporarily unreliable* agencies instead of using a conventional weighted average of *all* agencies. Consequently, if correct or wrong assessments tend to accumulate at certain times, any sorting-out procedure should be flexible enough to select a *varying* share of reliable agencies instead of - say - the best X percent. Hence, an efficient method that temporarily adjusts the respective agencies' reports until their expectations prove to be correct again is expected to improve quality of the novel unemployment leading indicator. However, the horizon of expectation of the survey requires waiting three months to take the reliability decision, which probably reduces expected efficiency gains because serial correlation in response behavior is less pronounced after three months. As a consequence, the benefit of any adjustment procedure needs to be investigated empirically which is done in section 4.

While developing an adjustment procedure, we make use of another major advantage of our survey: There are five (instead of three as in most comparable surveys) answer options that allow a more gracefully built assessment of unemployment changes. This is why we fully exploit all answer categories and decide about reliability of expectations with the help of given scopes. For our preferred version of the new unemployment leading indicator we consider a prediction of *increase strongly* as hit if unemployment rises more than 12.5 percent over the next three months. Reports of *increase* are treated as correct if unemployment growth lies between 2.5 and 12.5 percent. Reports of *decline* and *decline strongly* are treated analogously. *Stay constant* is considered as hit if quarterly unemployment changes are smaller than 2.5 percent in absolute value. This stronger selection obviously lowers the percentage of predictions that are considered as accurate. However, a check of alternative limits revealed that the above-mentioned selection criteria are among the most conservative with respect to the exclusion of agencies, which supports our intention not to deliberately exclude too many agencies. The chosen limits also match actual response behavior: approximately ten percent of all assessments fall into extreme categories (*increase strongly, decline strongly*), corresponding well to roughly ten percent of quarterly unemployment changes that are greater than 12.5 percent in absolute value. As a consequence of this stricter sorting-out procedure, we assure that the remaining agencies not only have been temporarily right about *tendency* but also about *magnitude* of unemployment changes.

Since we also know about the sign of the forecasting errors of the unreliable agencies, we can distinguish between too optimistic and too pessimistic agencies and adjust current predictions by the latest observable bias in response behaviour. In order to quantify this bias both the survey answers and the subsequent district-specific unemployment changes are transformed into integers between -2 and 2 according to the limits described above. This leads to categorized forecasting errors ranging between -4 and +4. Among all temporarily unreliable agencies, we distinguish between the *bias type* and the *variance type*

of respondents. The bias type tends to over- or underestimate unemployment changes, whereas the forecasting errors of the variance type typically differ in sign. Through this distinction we can change the assessments of unreliable agencies in a way that precisely matches the respective types. This distinction requires taking into account at least two successional observations. We do not recommend accounting for more than two observations, since any agency could learn from its mistakes. Therefore, it is more convincing to implement a method that is flexible enough to adapt to changes in response behavior as fast as possible. Hence, we allow an agency's forecasting type to change over time.

Consequently, an agency's assessment always remains unchanged unless there have been two consecutive months of wrong predictions. This way, less than one third of all reports in our sample are considered unreliable and thus require adjustment. If the signs of two consecutive forecasting errors coincide, the agency belongs to the bias type. We exploit serial correlation in response behavior through adjusting its prediction by the current extent of categorized over- or underestimation.[7]

If the signs of two consecutive forecasting errors do not coincide, the agency belongs to the variance type. In this case there is no bias one could adjust for. Therefore, we recommend replacing its prediction. Otherwise it could happen that the agencies remaining in the index at a certain point in time are not representative for all of Germany. As substitute we use the respective agencies' current monthly unemployment changes in percent, translated into integers in accordance with the classification procedure presented above. Monthly instead of quarterly unemployment growth rates are taken in order to improve the lead time of the new unemployment indicator. Consequently, the thresholds for converting metric unemployment changes into integer values are transformed using cubic roots.

Hence, our final version of the new unemployment leading indicator is defined as follows:

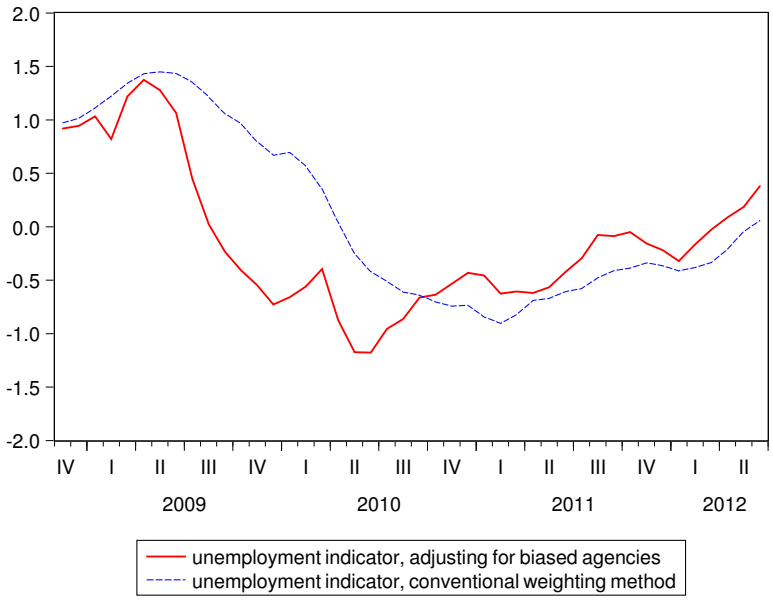
$$I_t^{adj,j} = \sum_{i=1}^{176} \omega_{it}^j \cdot [\gamma_{it} \cdot \phi_{it} + (1 - \gamma_{it}) \cdot (\delta_{it} \cdot \phi_{it}^{adj} + (1 - \delta_{it}) \cdot \Delta U_{it}^{clas})], \qquad (2)$$

where $\gamma_{it} = 0$ if the last two months' predictions of agency i have been wrong and $\gamma_{it} = 1$ otherwise. In case the last two prediction errors coincide in sign (=bias type), δ_{it} is set to be 1 and predictions are adjusted (ϕ_{it}^{adj}) for the current forecast error. If the signs of the last two prediction errors differ (=variance type), δ_{it} equals zero and the respective assessments are replaced by the current classified unemployment change (ΔU^{clas}).

Figure 5 shows the resulting seasonally adjusted indicator using the reference group as weights (solid line). While its development still resembles that of the conventional index (dashed line), some differences are visible at first sight. The novel indicator clearly shifted to the left, especially during and after the Great

[7] However, adjusted prediction is restricted not to exceed 2 in absolute value.

Figure 5: Unemployment leading indicator, bias-adjusted



Recession where herding might have led respondents into the wrong direction. We argue that the new leading indicator described in equation (4) is flexible as it allows considering a time-varying number of (un-)reliable respondents and adjusting for collective over- or underestimation. The new index also seems to be more volatile, probably reflecting the nature of unemployment changes rather than the level. In order to get a first impression of the success of our sorting-out procedure, we take 3-month differences of seasonally adjusted aggregate unemployment and regress them on the either lagged conventional or lagged novel unemployment leading indicator (see the random walk version of equation (3) below). Applying the sorting-out procedure, the resulting mean squared prediction error (MSPE) is reduced from $3.57 \cdot 10^9$ to $1.69 \cdot 10^9$, i.e. by more than 50 percent. Tables 2 to 5 and 6 to 9 show that this is not a singular case. The novel unemployment leading indicator typically produces lower MSPEs than a conventionally aggregated one. The general question whether the new leading indicator helps to (significantly) improve forecasting unemployment and, if so, at which horizons, is treated by means of forecast evaluation in the next chapter.

4 Forecast evaluation

4.1 Comparison to autoregressive benchmarks

This subsection compares the forecasting performance of purely autoregressive models to that of AR models enhanced by the new unemployment leading indicator. As a consequence, the parsimonious benchmark model is nested in the larger model, which is of crucial importance in tests of equal predictive accuracy. Clark and West (2007) argue that the mean squared prediction error (MSPE) of the larger model is upward-biased due to additional noise stemming from the need to estimate a parameter which - under the null hypothesis of equal predictive performance - (1) is zero in population[8] and which (2) is correctly set to zero in the parsimonious model. In a sense, the smaller benchmark model is more efficient and hence benefits from not carrying the burden of estimating the parameter of a redundant variable to zero. Consequently, usual tests in the style of Diebold and Mariano (1995) are undersized and have poor power in a nested model environment. Therefore, we implement the nested-model test described in Clark and West (2007), applying a one-sided test for equal predictive accuracy with the alternative hypothesis being worse forecast performance of the nesting model. Since multiperiod-ahead forecast errors are usually autocorrelated, we use the heteroskedasticity and autocorrelation robust covariance estimator proposed in Newey and West (1987) in case of multiple-step forecasts. Inference based on asymptotic critical values - as proposed in McCracken (2004) or Clark and McCracken (2001) - might not be appropriate in case of small sample sizes. Therefore, the fixed regressor bootstrap method proposed in Clark and McCracken (2012a,b) is implemented. We allow for horizon-specific and model type-specific sets of critical values and find them in most cases to be larger than their asymptotic counterparts. While this makes rejection of the null more difficult, we argue that bootstrapping considerably strengthens the validity of our test results.

For computing multi-step forecasts we use direct, lead time-dependent forecasts. At least in theory, direct forecasts are more immune to model misspecification than iterated forecasts since they use the chosen model only once. On the other hand, parameter estimates are more efficient in the iterated approach because it usually allows eliminating residual autocorrelation. As a consequence, it is an empirical question which approach should be used. Literature on this topic is ambiguous, ranging from emphasizing the advantages of direct forecasts (e.g. Klein (1968)) over mixed results (e.g Kang (2003)) to the finding of an empiric study on 170 U.S. macroeconomic variables that iterated forecasts typically outperform direct forecasts (Marcellino et al. (2006)). In applying direct forecasts we avoid forecasting the indicator variable itself and modeling feedback effects to our target variable. Furthermore, the asymptotic theory of the specific nested model test we use in our application requires the forecasts to be linear functions of parameters which applies in direct forecasts but not in iter-

[8] For a discussion of the difference between a null hypothesis of equal accuracy in the *population* vs. *finite sample*, see e.g. Clark and McCracken (2009, 2012a).

ated approaches. The general lead time-dependent estimation specification for regressing aggregate unemployment[9] U on a constant, two autoregressive lags and one lag of the novel unemployment leading indicator introduced in equation (4) follows

$$U_{t+h} = \alpha_0 + \alpha_1 \cdot U_t + \alpha_2 \cdot U_{t-1} + \beta \cdot I_t^{sel,j} + \epsilon_{t+h}, \tag{3}$$

with h denoting the forecast horizon and ϵ the error term. We do not index the coefficients by h for simplicity.

The following paragraph discusses the choice of the underlying parsimonious benchmark model. One could think of models relying solely on the own past such as AR(p)-models or random walk (RW). In their GDP growth application, Clark and West (2007) use an AR(1) with constant as benchmark model, Clark and McCracken (2009) use models with just a constant in order to predict stock returns. Sometimes AR models of higher order, determined by in-sample information criteria such as AIC or SC, are used. We argue that relative performance of a model including a leading variable considerably depends on the choice of the parsimonious benchmark model. The additional variable in question might perfectly complement an AR(1) or RW specification but simply be in the way when using AR models of higher order as benchmark. Instead of relying on a single benchmark model, we present a choice of models that seem plausible in the light of the time series properties of the underlying unemployment variable. The Bayesian Information Criterion (BIC) with monthly unemployment data from 1998 to 2007 (and hence excluding data from our estimation and evaluation periods) gives evidence for using AR models with order not higher than 2.[10] Due to the high persistence we also check the respective unit-root equivalents, i.e. models in first differences.

Table 1 shows all six benchmark models we initially rely on, together with the respective restrictions on equation (3):

Table 1: Set of benchmark models

benchmark model	restrictions on eq (3)	forecasting equation
AR(h)	$\alpha_2 = 0$	$U_{t+h} = \alpha_0 + \alpha_1 \cdot U_t + \epsilon_{t+h}$
AR(h+1)		$U_{t+h} = \alpha_0 + \alpha_1 \cdot U_t + \alpha_2 \cdot U_{t-1} + \epsilon_{t+h}$
RW	$\alpha_0 = \alpha_2 = 0$, $\alpha_1 = 1$	$U_{t+h} - U_t = \epsilon_{t+h}$
RW with drift	$\alpha_1 = 1$, $\alpha_2 = 0$	$U_{t+h} - U_t = \alpha_0 + \epsilon_{t+h}$
dAR(h)	$\alpha_0 = 0$, $\alpha_1 = 1 - \alpha_2$	$U_{t+h} - U_t = -\alpha_2 \cdot d(U_t) + \epsilon_{t+h}$
dAR(h) with drift	$\alpha_1 = 1 - \alpha_2$	$U_{t+h} - U_t = \alpha_0 - \alpha_2 \cdot d(U_t) + \epsilon_{t+h}$

All respective benchmark models do not include the unemployment leading indicator ($\beta = 0$). Since we use the direct approach, the model type changes

[9] Throughout the evaluation process we target forecasts of aggregate unemployment figures (and hence not forecasts of the unemployment rate).

[10] This result remains valid in case data after 2007 are included.

with forecast horizon. For instance, the first model becomes an $AR(1)$ for 1-step-ahead forecasts and an AR(2) without the first lag for 2-step-ahead forecasts. As any direct h-step-ahead forecasting equation implies a MA(h-1) error structure, we also considered the respective ARMA models. However, out-of-sample performance of these models turned out to be worse than that of their AR counterparts such that we do not report ARMA results.

In this paper we focus on linear single-equation models, taking the leading indicators under consideration as exogenous.[11] The limited time range for which the underlying survey data are available seems to be in the way of more sophisticated non-linear models that treat periods of booms and recessions in different ways.[12] The limited time range makes it also questionable to follow rolling window approaches resting upon changing but equally long estimation periods. Instead, we follow the recursive approach in order to fully exploit all available information for estimation purposes.

Since the survey is explicitly designed for a short forecast horizon, we concentrate on 1-, 2- and 3-step-ahead forecasts. In addition, we take 6-step-ahead forecasts in order to get evidence for higher forecast horizons. We divide the sample into an estimation period which is consistently updated, and an evaluation period. Hansen and Timmermann (2012) and Clark and McCracken (2012a) find that the optimal sample split results in an evaluation period being relatively large compared to the initial estimation period. We chose the split parameter Π to be approximately 2, signifying an evaluation period twice as large as the initial estimation period. For instance, the initial estimation period for 1-step-ahead forecasts based upon the AR(1) model ranges from December 2008 to February 2010 using data from November 2008 as initial observations. Our evaluation period ranges from March 2010 to July 2012 in case of 1-step-ahead forecasts and from August 2010 to December 2012 in case of 6-step-ahead forecasts. As a consequence, the evaluation period consists of 29 forecasts for all 4 forecast horizons. The estimation period is regularly updated by adding the month that has become recently available (recursive scheme). Hence, the last estimation period ends in June 2012. Each time the forecasts are calculated, the respective forecasting model is re-estimated first. Since we use lead time-dependent forecasts, the necessary number of initial observations differs not only across model types but also across forecast horizons.

Tables 2 to 5 show the test results for the 1-, 2-, 3- and 6-step-ahead forecasts, respectively. We abstain from reporting test results of the first two model types (AR(h) and AR(h+1)) because for all investigated forecast horizons, and both for the benchmark and the indicator models, the respective MSPEs turned out to be substantially higher than in the unit root cases. This speaks in favor of modeling unemployment in differences so that we concentrate on the remaining four model types (RW without/with drift, dAR(h) without/with drift). The

[11] Alternatively, univariate benchmark models could also be confronted with bivariate VAR or dVAR models (e.g. Clements and Hendry (1996), Christoffersen and Diebold (1998)).

[12] For logit/probit models, Markov-switching models or smooth-transition models, see e.g Hamilton and Perez-Quiros (1996), Granger et al. (1993). For an application of non-linear methods to forecast the U.S. unemployment rate, see e.g. Golan and Perloff (2004).

Table 2: Evaluation of monthly 1-step-ahead unemployment forecasts

		with I^{all}		with I^{adj}	
model type	$MSPE_1$	$MSPE_2$	$CW_{2,1}$	$MSPE_3$	$CW_{3,1}$
RW	0.63	0.50	*2.90***	0.29	*4.74****
RW with drift	0.70	0.50	*3.71****	0.30	*3.74****
dAR(h)	0.31	0.31	*-0.27*	0.27	*2.62****
dAR(h) with drift	0.32	0.36	*0.03*	0.28	*2.45****

Notes: $MSPE_1$ is the out-of-sample MSPE of the parsimonious benchmark model. $MSPE_2$ is the out-of-sample MSPE of the alternative larger model including the lagged conventional indicator I^{all}. $CW_{2,1}$ is the Clark/West test statistic comparing $MSPE_2$ to $MSPE_1$. $MSPE_3$ is the out-of-sample MSPE of the alternative larger model including the adjusted unemployment leading indicator I^{adj}. $CW_{3,1}$ is the Clark/West test statistic comparing $MSPE_3$ to $MSPE_1$. The MSPEs are to be multiplied by 10^9. *, **, *** denote significance at the 10, 5, 1 percent level, respectively. Critical values are calculated following the fixed regressor bootstrap procedure proposed in Clark and McCracken (2012b) using 99,999 replications. The heteroskedasticity and autocorrelation robust covariance estimator proposed in Newey and West (1987) was used in case of multiple-step forecasts.

Table 3: Evaluation of monthly 2-step-ahead unemployment forecasts

benchmark model	$MSPE_1$	$MSPE_2$	adj. term	$\Delta MSPEadj$ *(test statistic)*
RW	2.18	0.92	1.05	2.31 *(2.99)****
RW with drift	2.54	1.01	1.60	3.13 *(2.16)***
dAR(h)	1.15	0.96	0.10	0.29 *(3.40)****
dAR(h) with drift	1.21	1.04	0.13	0.29 *(2.51)****

Notes: see table 2.

Table 4: Evaluation of monthly 3-step-ahead unemployment forecasts

benchmark model	$MSPE_1$	$MSPE_2$	adj. term	$\Delta MSPEadj$ *(test statistic)*
RW	4.24	1.69	2.05	4.59 *(3.05)****
RW with drift	4.95	1.82	3.65	6.78 *(2.15)***
dAR(h)	2.49	1.92	0.24	0.81 *(3.53)****
dAR(h) with drift	2.49	2.04	0.36	0.82 *(2.64)****

Notes: see table 2.

Table 5: Evaluation of monthly 6-step-ahead unemployment forecasts

benchmark model	$MSPE_1$	$MSPE_2$	adj. term	$\Delta MSPEadj$ *(test statistic)*
RW	12.01	4.88	3.29	10.43 *(3.42)****
RW with drift	13.56	5.61	10.83	18.78 *(2.32)****
dAR(h)	7.19	6.03	0.35	1.51 *(2.27)****
dAR(h) with drift	6.91	6.81	0.90	1.00 *(1.46)**

Notes: see table 2.

first column displays the type of the benchmark model. The second column shows the corresponding MSPEs, whereas the third column displays MSPEs of the alternative larger model including the new unemployment leading indicator. One can see that both for the benchmark ($\beta = 0$) and indicator models, restricting α_0 to zero leads to lower MSPEs compared to the respective models with drift. In all cases, the MSPE of the benchmark model exceeds that of the larger model. The novel unemployment indicator seems to go particularly well with the plain RW model type, especially in case of 2-, 3- and 6-months-ahead forecasts. Adjusted for the upward bias (fourth column), all resulting test statistics are significantly positive at least at the 5 percent level, the only exception being the $dAR(h)$ model with drift for 6-step-ahead forecasts with 10 percent. Hence, the test results show that the null hypothesis of equal predictive accuracy can be rejected and that models enhanced by the new unemployment leading indicator outperform their benchmark counterparts.

Table 6: Conventional unemployment indicator: 1-step-ahead forecasts

benchmark model	$MSPE_1$	$MSPE_2$	adj. term	$\Delta MSPEadj$ *(test statistic)*
RW	0.63	0.50	0.06	0.20 *(2.90)***
RW with drift	0.70	0.50	0.96	1.16 *(3.71)***
dAR(h)	0.31	0.31	0.00	-0.00 *(-0.27)*
dAR(h) with drift	0.32	0.36	0.06	0.03 *(0.72)*

Notes: see table 2.

Table 7: Conventional unemployment indicator: 2-step-ahead forecasts

benchmark model	$MSPE_1$	$MSPE_2$	adj. term	$\Delta MSPEadj$ *(test statistic)*
RW	2.18	1.78	0.17	0.57 *(1.75)***
RW with drift	2.54	2.36	3.58	3.75 *(2.28)***
dAR(h)	1.15	1.20	0.00	-0.04 *(-1.48)*
dAR(h) with drift	1.21	1.91	0.25	-0.46 *(-1.29)*

Notes: see table 2.

Table 8: Conventional unemployment indicator: 3-step-ahead forecasts

benchmark model	$MSPE_1$	$MSPE_2$	adj. term	$\Delta MSPEadj$ *(test statistic)*
RW	4.24	3.57	0.23	0.90 *(1.66)**
RW with drift	4.95	6.06	7.03	5.92 *(1.99)**
dAR(h)	2.49	2.65	0.02	-0.15 *(-1.06)*
dAR(h) with drift	2.49	5.00	0.91	-1.60 *(-1.17)*

Notes: see table 2.

Table 9: Conventional unemployment indicator: 6-step-ahead forecasts

benchmark model	$MSPE_1$	$MSPE_2$	adj. term	$\Delta MSPEadj$ *(test statistic)*
RW	12.01	12.42	0.09	-0.32 *(-0.57)*
RW with drift	13.56	71.40	42.86	-14.97 *(-0.60)*
dAR(h)	7.19	9.09	0.49	-1.41 *(-0.93)*
dAR(h) with drift	6.91	49.00	29.62	-12.47 *(-1.12)*

Notes: see table 2.

The fact that results are considerably stable across forecast horizons and model types is a consequence of the sorting-out procedure described above. Tables 6 to 9 display test results using the conventional unemployment indicator constructed according to equation (1). In all cases forecasts with the conventional indicator produce higher MSPEs compared to our favored unemployment leading indicator. Especially in cases of forecasts based on the two dAR(h) models, the null of equal predictive power cannot be rejected.

Furthermore, the conventional leading indicator does not significantly outperform any benchmark model in case of a forecast horizon of 6 months. Even for 3-step-ahead forecasts, test statistics are only weakly significant at most. Hence, we conclude that distinguishing between *temporarily reliable* and *temporarily unreliable* agencies and adjusting for a negative or positive bias in response behaviour is an appropriate method not only to improve short-term forecasts of unemployment, but also to enable more accurate forecasts at higher forecast horizons and hence look further into the future.

4.2 Comparison to other leading indicators

After investigating predictive accuracy compared to purely autoregressive benchmarks, the following paragraphs focus on alternative leading indicators. To our knowledge there is no leading indicator on the market that explicitly aims at predicting German unemployment development. However, there are some economic variables and survey data that can be expected to have direct and indirect links to our target variable.

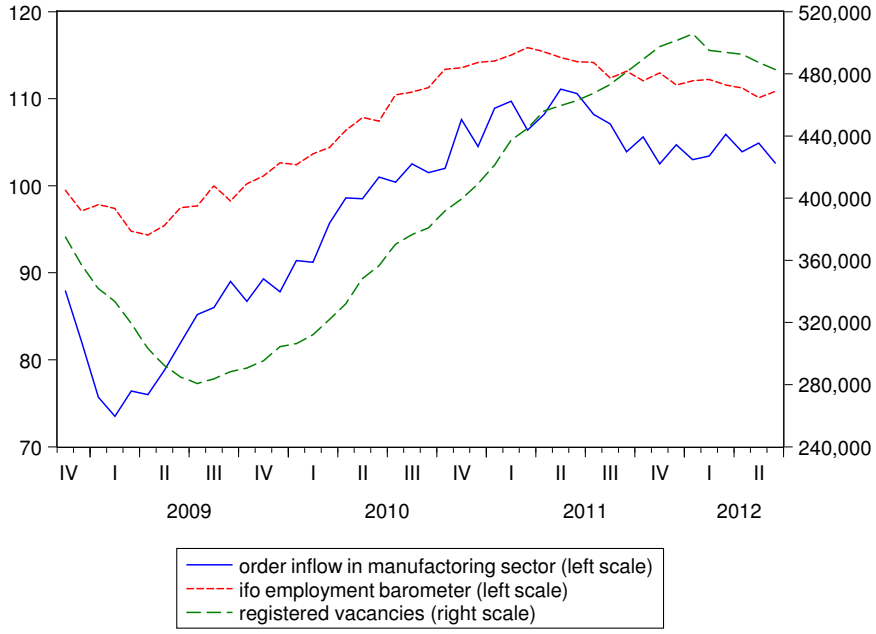
Order inflow

We consider a business cycle indicator like order inflow (OI) in the manufacturing sector as potential candidate for leading unemployment.[13] The index of incoming orders is constructed on a monthly basis using data from the statistical offices of the German Länder. It comprises the monthly value exclusive of VAT of all accepted orders of manufacturing companies with more than 50 employees, indexed to a base year. The solid line in figure 6 shows the seasonally adjusted order inflow from November 2008 to June 2012 as published by the Federal Statistical Office. The latest index value is made available only around the beginning of the next month but one, together with revisions for previous months.[14] Consequently, $I_t^{sel,j}$ in (3) is replaced by OI_{t-1}. Hence, the index of incoming orders enters the estimation equation with a delay of one month compared to the respective lag order of the AR-term. Since it behaves countercyclically with respect to unemployment, one would expect β to be negative.

[13] We also considered industrial production as a natural indicator for labor demand. However, the industrial production index typically produced higher MSPEs than new orders. The paper focuses on the stronger competitor only.

[14] These revisions could advantage the index of order inflow compared to the novel unemployment indicator. For out-of-sample tests taking into account the real time nature in case of data revisions, see Clark and McCracken (2007).

Figure 6: Order inflow, ifo employment barometer and registered vacancies



ifo employment barometer

Since employment and unemployment are highly (and negatively) correlated, another promising approach would be to use employment leading indicators to forecast unemployment. The ifo employment barometer is a survey-based indicator for predicting employment development. It uses a question that captures hiring and firing plans of the responding companies within a three-month horizon. Contrary to the order inflow, the ifo employment barometer is published without delay so that equation (3) is employable analogously. The way the question is asked is similar to the survey design of the FEA described in section 2. However, the answer options in the style of a Likert scale comprise three instead of five categories. The ifo employment barometer is depicted as dotted line in figure 6. There is a strong negative correlation with our favored unemployment leading indicator constructed according to equation (4). The resulting cross-correlogram shows a maximum negative correlation of 0.74 where the new unemployment indicator leads the ifo employment barometer by four months.

Registered vacancies

Following classic matching theory (see e.g. Mortensen and Pissarides (1994), Petrongolo and Pissarides (2001), Shimer (2007) and Yashiv (2007)) taking registered vacancies as unemployment leading indicator is another natural choice. The crucial question is whether there is just a contemporary comovement between unemployment and (inverse) vacancies or whether vacancies have predictive power. The dashed line in figure 6 shows registered vacancies as published by the FEA. The variable comprises all job offers that employers report to the respective local agencies and that are approved for placement. Consequently, the chosen variable does not cover the whole job market. Day of the count is at the middle of the month but figures are published at the end of the month, together with the publication of unemployment figures. The correlation between registered vacancies and seasonally adjusted unemployment (see figure 3) is highly negative.

Comparing different predictors using various underlying autoregressive specifications leads to a high number of competing models. Therefore, we follow an approach recently established by Hansen et al. (2011) to compare predictive accuracy of a large number of models: the model confidence set (MCS). MCSs are comparable to confidence intervals for estimation parameters and comprise the best forecasting model with a chosen confidence level. The strategy is to sort out models with poor out-of-sample performance and hence reduce the large number of models to a smaller set. We investigate whether models including the new unemployment leading indicator survive the selection process and succeed to stay put in the MCS, and if so, for which forecast horizons. An MCS is generated through an iterated two-step procedure. The first step, i.e. the equivalence test, is applied to all (remaining) models. The null hypothesis states that they perform equally well. In case of rejection, the second step is employed to drop an inferior model from the set. The two steps are repeated until the equivalence test cannot be rejected any more. The remaining choice of models is the MCS. Note that not all models surviving the elimination procedure necessarily

have lower sample MSPEs than those excluded from the MCS. As in any other significance test, it is possible not to reject the null (and hence to stay in the MCS) due to high variance.

For all competing indicators, AR(h) and AR(h+1) models turned out to perform worse than their unit root counterparts which is why we use the same four model types as introduced in section 4.1. Furthermore, we follow the h-step-ahead forecast procedure as described for the unemployment leading indicator in equation (3) analogously for the ifo employment barometer and registered vacancies. In case the order inflow is used, we adjust equation 3 for the delayed availability of the indicator as described above. [15]

Table 10: Model confidence set for monthly 1-step-ahead unemployment forecasts

model type	leading indicator	MSPE ($*10^9$)	MCS p-value
dAR(h)	unemployment leading indicator	0.27	1.0000
dAR(h)with drift	unemployment leading indicator	0.28	0.4833
RW	unemployment leading indicator	0.29	0.4034
RW with drift	unemployment leading indicator	0.30	0.4034

Notes: Results were calculated with the OX MulCom package version 2.00, significance level: $\alpha = 0.1$, number of models: $l = 16$, sample size: $n = 29$, loss function: MSPE, test statistic: MaxT, bootstrap parameters: $B = 10000$ (resamples), $d = 2$ (block length). For robustness checks, we also used block lengths of $d = 1$, $d = 3$ and $d = 5$. Results do not substantially change, though.

Tables 10 to 13 display all models surviving the elimination procedure in case of 1-, 2-, 3- and 6-step ahead forecasts at a significance level of 0.1, together with the respective MSPEs and MCS p-values. The latter are connected to the null hypothesis of the equivalence test stating that all remaining models perform equally well. Hence, the model with the lowest p-value is the first not being eliminated from the MCS at a significance level of ten percent. Although the models are ranked according to their MCS p-values, a MCS is silent about which model is the best - instead it comprises the best model with a 90 percent confidence probability. In case of a forecast horizon of one month, the selection procedure is able to exclude 12 out of 16 models from the MCS. All four models including the new unemployment leading indicator stay put in the MCS. Two additional leading indicators enter the MCS for the 2- and 3-step-ahead forecasts: the ifo employment barometer and order inflow. In case of a forecast horizon of six months, the MCS comprises all four investigated leading indicators, where the MSPEs of the novel unemployment indicator are still clearly lowest.

One should take into consideration that none of the competing indicators

[15] Hansen et al. (2011) use rolling window schemes instead of recursive estimation approaches. However, they point out that recursive approaches lead to MCS results that are very similar to those generated by rolling window approaches.

Table 11: Model confidence set for monthly 2-step-ahead unemployment forecasts

model type	leading indicator	MSPE ($*10^9$)	MCS p-value
RW	unemployment leading indicator	0.92	1.0000
dAR(h)	unemployment leading indicator	0.96	0.7190
RW with drift	unemployment leading indicator	1.01	0.5429
dAR(h) with drift	unemployment leading indicator	1.04	0.5429
dAR(h)	ifo employment barometer	1.22	0.4060
dAR(h)	order inflow	1.22	0.1021

Notes: Results were calculated with the OX MulCom package version 2.00, significance level: $\alpha = 0.1$, number of models: $l = 16$, sample size: $n = 29$, loss function: MSPE, test statistic: MaxT, bootstrap parameters: $B = 10000$ (resamples), $d = 2$ (block length). For robustness checks, we also used block lengths of $d = 1$, $d = 3$ and $d = 5$. In case of $d = 1$, the dAR(h) model including order inflow drops out of the MCS. In case of $d = 5$, the RW model with order inflow and the dAR(h) model with vacancies stay put in the MCS, too.

Table 12: Model confidence set for monthly 3-step-ahead unemployment forecasts

model type	leading indicator	MSPE ($*10^9$)	MCS p-value
RW	unemployment leading indicator	1.69	1.0000
RW with drift	unemployment leading indicator	1.82	0.6485
dAR(h)	unemployment leading indicator	1.92	0.2982
dAR(h) with drift	unemployment leading indicator	2.04	0.2982
dAR(h)	ifo employment barometer	2.52	0.1319
dAR(h)	order inflow	2.54	0.1023

Notes: Results were calculated with the OX MulCom package version 2.00, significance level: $\alpha = 0.1$, number of models: $l = 16$, sample size: $n = 29$, loss function: MSPE, test statistic: MaxT, bootstrap parameters: $B = 10000$ (resamples), $d = 2$ (block length). For robustness checks, we also used block lengths of $d = 1$, $d = 3$ and $d = 5$. In case of $d = 1$, the dAR(h) model with vacancies stays put in the MCS, too. In case of $d = 3$ and $d = 5$, the dAR(h) model including order inflow drops out of the MCS.

Table 13: Model confidence set for monthly 6-step-ahead unemployment forecasts

model type	leading indicator	MSPE ($*10^9$)	MCS p-value
RW	unemployment leading indicator	4.88	1.0000
RW with drift	unemployment leading indicator	5.61	0.6599
dAR(h)	unemployment leading indicator	6.03	0.4179
dAR(h) with drift	unemployment leading indicator	6.81	0.4179
dAR(h)	ifo employment barometer	7.20	0.4057
dAR(h)	order inflow	7.56	0.3732
dAR(h) with drift	order inflow	11.95	0.1561
dAR(h)	vacancies	9.02	0.1509

Notes: Results were calculated with the OX MulCom package version 2.00, significance level: $\alpha = 0.1$, number of models: $l = 16$, sample size: $n = 29$, loss function: MSPE, test statistic: MaxT, bootstrap parameters: $B = 10000$ (resamples), $d = 2$ (block length). For robustness checks, we also used block lengths of $d = 1$, $d = 3$ and $d = 5$. In case of $d = 1$, the dAR(h) model with vacancies drops out of the MCS. In case of $d = 3$, the RW model with order inflow stays put in the MCS, too. In case of $d = 5$, all RW models without drift survive the selection procedure, too.

aims at *directly* signaling unemployment changes. However, they probably still perform well in forecasting other target variables. For instance, Abberger (2007) concludes that the ifo employment barometer is a valid leading indicator of actual employment changes. Furthermore, we note that many other business tendency surveys collect all data necessary to implement the reliability checks and bias adjustments presented in this paper. Especially survey-based indicators with short forecast horizons could be further improved by exploiting information about serial correlation and bias in response behavior.

5 Robustness checks

Subsection 4.1 discussed the advantages of an indicator that accounts for the agencies' reliability and distinguishes between the variance and bias type over a simple weighted average of all responses. In a sense, the latter is a limiting case of the more sophisticated version of the novel unemployment leading indicator. One can see that equation (2) collapses to equation (1) if all agencies are treated equally and thus considered reliable ($\gamma_{it} = 1 \; \forall \; i, t$). The first and last four rows of table 14 summarize the test results of tables 2 to 5 and 6 to 9. They show that using I^{adj} leads to a substantial reduction in MSPE compared to the conventional unemployment indicator I^{all}. This section focuses on two additional options while constructing the leading indicator that are worth investigating.

No bias-adjustment

This alternative construction method is another limiting case of equation (2).

Table 14: Out-of-sample performance of alternative unemployment leading indicators

indicator	model type	forecast horizon			
		1 month	2 months	3 months	6 months
I^{adj}	RW	0.29	0.92	1.69	4.88
	RW with drift	0.30	1.01	1.82	5.61
	dAR(h)	0.27	0.96	1.92	6.03
	dAR(h) with drift	0.28	1.04	2.04	6.81
$I^{\delta=0}$	RW	0.34	1.17	2.28	6.51
	RW with drift	0.34	1.25	2.52	8.46
	dAR(h)	0.30	1.13	2.34	7.35
	dAR(h) with drift	0.30	1.17	2.45	7.98
$I^{\gamma 8cat}$	RW	0.43	1.53	2.99	9.99
	RW with drift	0.49	2.07	4.65	23.03
	dAR(h)	0.31	1.19	2.60	8.98
	dAR(h) with drift	0.34	1.47	3.32	17.32
I^{all}	RW	0.50	1.78	3.57	12.42
	RW with drift	0.50	2.36	6.06	71.40
	dAR(h)	0.31	1.20	2.65	9.09
	dAR(h) with drift	0.36	1.91	5.00	49.00

Notes: The table displays MSPEs which are to be multiplied by 10^9.

It sets $\delta_{it} = 0 \; \forall \; i, t$ and thus investigates the effects of not adjusting predictions of biased agencies but instead treating bias and variance types alike. Hence, equation (2) collapses to:

$$I_t^{\delta=0,j} = \sum_{i=1}^{176} \omega_{it}^j \cdot [\phi_{it} \cdot \gamma_{it} + \Delta U_{it}^{clas} \cdot (1 - \gamma_{it})], \tag{4}$$

The second section of table 14 shows out-of-sample performance of $I^{\delta=0}$ (equation (4)). Forecast accuracy worsens for all horizons and models. This shows the importance of taking into account the sign of forecast errors as in our favored indicator.

Non-dichotomous reliability parameter

The second alternative allows the reliability parameter γ_{it} to be non-dichotomous and hence to take on values between 0 and 1 depending on the size of recent prediction errors of agency i. Since the last two forecasting errors are considered (each of which ranges between 0 and 4 in absolute value), the accumulated forecasting error can take on values from 0 to 8. Accounting for the size of past prediction errors thus allows the reliability parameter to be non-dichotomous, e.g.:

$$\gamma_{it}^{8cat} \in (0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1) \quad (5)$$

This way, an agency would only be considered completely unreliable ($\gamma_{it}^{8cat} = 0$) if both of its past two prediction errors equaled the highest possible integer (=4) in absolute value. Applying equation (5) to equation (4) gives an alternative unemployment leading indicator ($I^{\gamma 8cat}$) the out-of-sample performance of which is summarized in the third section of table 14. A comparison of $I^{\gamma 8cat}$ to $I^{\delta=0}$ shows that non-dichotomous reliability weights do not improve forecasting power of the unemployment leading indicator. This speaks in favor of entirely excluding assessments of temporarily unreliable agencies until they prove to be reliable again.

6 Conclusion

This paper aimed at closing a gap and constructing the first leading indicator that *directly* signals changes in Germany's aggregate unemployment figures. For this purpose, we use a new survey conducted by the FEA among the CEOs of local agencies. A comparison of reported expectations and actual unemployment changes at the regional level reveals serial correlation in response behavior and time-varying reliability in the agencies' reports. We find that an aggregate unemployment indicator that adjusts for over- or underestimation of temporarily unreliable agencies in order to effectively exploit serial correlation in the response behavior has the potential to outperform a simple weighted average of all agencies.

Results from forecast comparison tests for nested models confirm our expectations. In most of the cases, forecasts relying on a simple weighted average of all survey responses produce higher sample MSPEs compared to our favored unemployment leading indicator that distinguishes between temporarily reliable and non-reliable regional employment agencies. Out-of-sample tests including this new unemployment leading indicator show that the null hypothesis of equal predictive accuracy can be rejected in general and that models enhanced by the new unemployment leading indicator typically outperform their benchmark counterparts.

Comparisons of forecasting performance of the new unemployment indicator to other established leading indicators such as the ifo employment barometer, order inflow and registered vacancies are made with the help of model confidence sets. Our results show that models including the new indicator survive the selection procedure at a significance level of 10 percent. Moreover, they tend to be rather dominant in MCSs for all four investigated forecast horizons.

The new FEA survey is a unique data set covering labor market expectations with significant potential for further investigations. Our findings show that for some survey-based indicators, it could be worth investigating the response behavior in detail. In case the results show serial correlation in response behavior, and in case the survey design allows assessing the correctness of the respondents' predictions after a reasonably short time span, our methods can be

a useful guide for constructing a more efficient leading indicator. We argue that our construction method is flexible enough to adapt to any other environment: It allows considering a time-varying number of (un-) reliable respondents, and it allows the respondents to learn from their mistakes (i.e., to re-enter the sample once they stop their mistakes). We find that an effective sorting-out procedure that captures serial correlation and systematic over- or underestimation in response behavior can improve forecast accuracy and allow looking further into the future.

Prospective research could benefit from an increasing number of observations, allowing for a detailed analysis of recessions and expansions, e.g. with the help of nonlinear models. Since survey data are available in a balanced panel format, it would also be interesting to learn more about spatial dependencies of unemployment expectations. Furthermore, additional questions in the survey which are not focus of this paper could be investigated with respect to their leading indicator properties once they meet a critical number of observations.

References

Abberger, K. (2007). Qualitative business surveys and the assessment of employment - A case study for Germany. *International Journal of Forecasting 23*(2), 249–258.

Askitas, N. and K. F. Zimmermann (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly 55*(2), 107–120.

Christoffersen, P. F. and F. X. Diebold (1998). Cointegration and Long-Horizon Forecasting. *Journal of Business &Economic Statistics 16*(4/1998), 450–458.

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

Clark, T. E. and M. W. McCracken (2007). Tests of Equal Predictive Ability with Real-Time Data. Research Working Papers 06, The Federal Reserve Bank of Kansas City, Economic Research Department.

Clark, T. E. and M. W. McCracken (2009). Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy. Working Paper Series 2009-050A, Federal Reserve Bank of St. Louis, Research Division.

Clark, T. E. and M. W. McCracken (2012a). Advances in forecast evaluation. Working Paper 1120, Federal Reserve Bank of Cleveland.

Clark, T. E. and M. W. McCracken (2012b). Reality Checks and Comparisons of Nested Predictive Models. *Journal of Business and Economic Statistics 30*, 53–66.

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

Clements, M. P. and D. F. Hendry (1996). Multi-Step Estimation for Forecasting. *Oxford Bulletin of Economics and Statistics 58*(4/1996), 657–684.

Diebold, F. X. and R. S. Mariano (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics 13*(3/1995), 253–263.

Golan, A. and J. M. Perloff (2004). Superior Forecasts of the U.S. Unemployment Rate Using a Nonparametric Method. *Review of Economics & Statistics 86*(1), 433–438.

Granger, C. W., T. Teräsvirta, and H. M. Anderson (1993). Modelling Nonlinearity over the Business Cycle. In *Business Cycles, Indicators, and Forecasting*, pp. 311–326. Chicago/London: James H. Stock and Mark W. Watson.

Hamilton, J. D. and G. Perez-Quiros (1996). What do the Leading Indicators Lead? *Journal of Business 69*(1/1996), 27–49.

Hansen, P. R., A. Lunde, and J. M. Nason (2011, March). The Model Confidence Set. *Econometrica 79*(2), 453–497.

Hansen, P. R. and A. Timmermann (2012). Choice of Sample Split in Out-of-Sample Forecast Evaluation. Technical report, Stanford University.

Kang, I.-B. (2003). Multi-period forecasting using different models for different horizons: an application to U.S. economic time series data. *International Journal of Forecasting 19*, 387–400.

Klein, L. R. (1968). An Essay on the Theory of Economics Prediction. Technical report, Sanomaprint, Helsinki, Finland (Yrjö Jansen lectures).

Marcellino, M., J. H. Stock, and M. W. Watson (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics 135*, 499–526.

McCracken, M. W. (2004). Asymptotics for Out of Sample Tests of Granger Causality. Technical report, University of Missouri.

Möller, J. (2010). Germany's job miracle in the world recession. Shock-absorbing institutions in the manufacturing sector. In *The Economy, Crises, and the Labor Market. Can Institutions Serve as a Protective Shield for Employment? (Applied Economics Quarterly Supplement)*, Volume 56, pp. 9–28. Klaus F. Zimmermann and Christian Wey (Publisher).

Mortensen, D. T. and C. A. Pissarides (1994). Job Creation and Job Destruction in the Theory of Unemployment. *Review of Economic Studies 61*, 397–415.

Newey, W. K. and K. D. West (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica 55*, 703–708.

Petrongolo, B. and C. A. Pissarides (2001). Looking into the Black Box: A Survey of the Matching Function. *Journal of Economic Literature XXXIX*.

Schanne, N. (2012). The formation of experts' expectations on labour markets. Do they run with the pack? IAB Discussion Paper 25, Institute for Employment Research.

Schanne, N., R. Wapler, and A. Weyh (2010, October). Regional unemployment forecasts with spatial interdependencies. *International Journal of Forecasting 26*(4), 908–926.

Shimer, R. (2007). Mismatch. *American Economic Review 97*(4).

Yashiv, E. (2007). Labor search and matching in macroeconomics. *European Economic Review 51*.