University of Konstanz
Department of Economics

Forecasting Aggregates with Disaggregate Variables: Does Boosting Help to Select the Most Relevant Predictors?

Jing Zeng

Working Paper Series
2014-20

Forecasting aggregates with disaggregate variables: Does boosting help to select the most relevant predictors? *

Jing Zeng†

September 23, 2014

Abstract

Including disaggregate variables or using information extracted from the disaggregate variables into a forecasting model for an economic aggregate may improve the forecasting accuracy. In this paper we suggest to use the boosting method to select the disaggregate variables which are most helpful in predicting an aggregate of interest. We conduct a simulation study to investigate the variable selection ability of this method. To assess the forecasting performance a recursive pseudo-out-of-sample forecasting experiment for six key Euro area macroeconomic variables is conducted. The results suggest that using boosting to select relevant predictors is a feasible and competitive approach in forecasting an aggregate.

Keywords: aggregation, macroeconomic forecasting, componentwise boosting, factor analysis

JEL classification: C22, C43, C52, C53, C82

---

*We thank participants in the 1st Vienna Workshop on High Dimensional Time Series in Macroeconomic and Finance 2013, the DIW Berlin Macroeconometric Workshop 2013, the Asian Meeting of the Econometric Society 2014, the Jahrestagung des Vereins für Socialpolitik 2014, the Jahrestagung der Statistischen Woche 2014 and the research seminar of the Center of Quantitative Methods and Survey Research Konstanz for helpful comments and suggestions. Financial support by the Deutsche Forschungsgemeinschaft, project number BR 2941/1-1, is gratefully acknowledged.

†University of Konstanz, Department of Economics, Box 129, 78457 Konstanz, Germany, email: jing.zeng@uni-konstanz.de

1 Introduction

The challenge of forecasting macroeconomic aggregate variables is gaining increasing importance with the creation of European Monetary Union (EMU). Many Euro area economic series are contemporaneous aggregates of disaggregate country-specific variables. In some real-life situations the forecaster encounters the problem of choosing an appropriate econometric procedure to forecast a contemporaneously aggregated variable. Traditionally, the aggregate variable can be forecasted by specifying a time series model for this aggregate variable, where the forecast is then computed by using only the historical values of this aggregate variable. In this approach, none of the disaggregate components is considered. Another approach builds a separate time series model for each disaggregate variable and forecasts the disaggregate variables first. The forecast of the aggregate variable is then computed by pooling the disaggregate forecasts together. In this approach, the aggregate variable itself is not considered. Choosing between these two strategies is not an easy task in real forecasting applications.

The theoretical literature provides inconclusive results regarding the ranking of these two approaches (Lütkepohl (2010)). In practice, in forecasting four macroeconomic Euro area aggregate variables over the period 1982-1997 Marcellino et al. (2003) find that pooling country-specific forecasts produces the most accurate forecasts for the aggregate variables. Hubrich (2005), however, shows that forecasting the Euro area year-on-year inflation using only the aggregate data results in higher forecast accuracy. In the course of further developments in this discussion, Hendry and Hubrich (2006, 2011) show analytically that taking disaggregate variables or information into account in the forecasting model should be helpful for reducing the forecast mean squared error. However, this improvement in the forecast accuracy can be effected by estimation uncertainty and model selection. Based on the analysis for aggregates with time-varying weights, Lütkepohl (2011) and Brüggemann and Lütkepohl (2013) also point out that incorporating disaggregate information contained in the stochastic weights in the forecasting model will generally lead to forecast improvements.

Including all disaggregate variables is often inappropriate since too many parameters have to be estimated, which makes the forecast inefficient. This paper focuses on how to include the relevant disaggregate variables in the forecasting model which originally uses only the lags of the aggregate variable. One popular existing approach of exploiting the high-dimensional disaggregate data is the factor method that summarizes the information contained in a large number of series in just a few unobservable common factors. If the factors can be estimated accurately and precisely, then forecasting an aggregate using disaggregate variables as an additional source of information is obtained by using the estimated factors as additional predictors. Another way of exploiting the high-dimensional disaggregate data is to select observed variables as predictors from a large feasible set which are most relevant in predicting an aggregate of interest. This paper suggests to use boosting as such a variable selection device.

Boosting stems from the machine learning and biostatistics literature for analyzing high-dimensional data and has proven to be very competitive in terms of prediction accuracy (Bühlmann and Hothorn (2007)). Until now, there are some applications in the macroeconometric forecasting literature. Bai and Ng (2009) use boosting to pick out the most relevant predictors in the factor-augmented forecasting model. The results indicate that some form of the boosting procedure outperforms the standard factor-augmented forecasts. Buchen and Wohlrabe (2011) evaluate the forecasting performance of boosting and compare it with the forecast combination schemes and dynamic factor models. Using the same data presented in Stock and Watson (2006) they find that boosting is a viable and computationally efficient approach to improve the forecasting accuracy. In the later work (Wohlrabe and Buchen (2014)) they confirm this result by examining the data for the United States, the Euro area and Germany. By investigating the forecasting performance of multivariate models, Carriero et al. (2011) also include multivariate boosting in their forecast comparison. They show that boosting performs best in forecasting CPI inflation one month ahead. Robinzonov et al. (2012) apply different boosting strategies on nonlinear autoregressive time series and find that boosting of high-order autoregressive time series is very competitive in terms of function estimation, and forecasting with boosting may successfully improve the forecasting quality of the German industrial production. In a related exercise, Taieb and Hyndman (2014) propose a new forecasting strategy for producing multi-step forecasts which uses boosting to avoid the difficult choice between the iterated and direct forecasts. They demonstrate from theoretical and empirical perspective that their method is quite attractive for the multi-step forecasting.

In terms of forecasting an aggregate variable, the boosting method has the advantage that it allows us to select, among a large number of disaggregate variables, the predictors which are most relevant in predicting the aggregate. In this paper, we investigate the properties of boosting as a selection device by conducting simulation experiments that mimic the empirical problem of selecting disaggregate components. The alternative approach, the factor method, performs a dimension reduction as well. However, it estimates only the information contained in the disaggregate variables and uses unobservable factors as predictors.

To evaluate the empirical forecasting performance of boosting, this paper compares the performance of the forecasting model which uses the boosting method to select relevant disaggregate variables with other three commonly used models in forecasting an aggregate: the forecast of an aggregate based only on the lagged values of this aggregate variable, forecasting the disaggregate variables first and then aggregating all the disaggregate forecasts and including the disaggregate information summarized in factor series in the forecasting model. The main purpose of this study is to answer two empirical questions: First, does using disaggregate information or variables as additional predictors in the forecasting model improve the forecasting performance? Second, can boosting help to select the most relevant disaggregate variables which can be incorporated in the forecasting model? In the empirical application to the Euro area macroeconomic key variables such as real GDP and consumer price index, we find that substantial improvements in terms of forecasting accuracy can be achieved when using the relevant disaggregate variables selected by the componentwise boosting as additional predictors in the forecasting model based on the historical values of the aggregate.

The structure of the remaining paper is organized as follows. Section 2 describes briefly the commonly used forecasting models for an aggregate variable. Section 3 outlines the boosting procedure, especially the componentwise boosting. The simulation study is discussed in Section 4. Section 5 describes the empirical analysis and Section 6 concludes.

2 Forecasting strategies for an aggregate variable

In this section three different forecasting models which can be used to forecast an aggregate variable are described and discussed. Consider a univariate aggregate variable of interest y_t^{AGG} which can be defined as the weighted sum of N disaggregate component variables $x_t = (x_{1,t}, \ldots, x_{N,t})'$. Thus, the contemporaneous aggregation relationship can be written as

$$y_t^{AGG} = w'x_t = w_1 x_{1,t} + \ldots + w_N x_{N,t},$$

where $w = (w_1, \ldots, w_N)'$ is the weighting vector with fixed values[1]: $w_i > 0$ and $\sum w_i = 1$ for $i = 1, \ldots, N$. The value of the aggregate variable h periods ahead y_{t+h}^{AGG} should be predicted. The optimal forecast at horizon h is the conditional mean:

$$y_{t+h|t}^{AGG} = \mathrm{E}[y_{t+h}^{AGG} | y_t^{AGG}, y_{t-1}^{AGG}, \ldots, x_{1,t}, x_{1,t-1}, \ldots, x_{N,t}, x_{N,t-1}, \ldots].$$

It is assumed that possible transformations are taken so that y_t and x_t are stationary series.

To obtain the multi-period ahead forecast $y_{t+h|t}^{AGG}$, a separate time series model for each forecasting horizon is estimated. Then the forecast is computed directly by using the estimated model.

Univariate forecast of the aggregate

To forecast the aggregate, the most commonly used technique is the univariate autoregressive model:

$$y_t^{AGG} = c_h + \sum_{j=1}^{p} \alpha_j y_{t+1-j-h}^{AGG} + \varepsilon_t.$$

[1]Brüggemann and Lütkepohl (2013) investigate forecasting aggregate variables with time-varying aggregation weights. This case is not considered in this paper.

p denotes the number of the lags of the aggregate variable y_t^{AGG}. Using the information set available until the period t, the parameters c_h and α_j ($j = 1, \ldots, p$) can be estimated by the ordinary least squares method (OLS). The estimated forecast of the aggregate variable can be computed as:

$$\hat{y}_{t+h|t}^{AGG} = \hat{c}_h + \sum_{j=1}^{p} \hat{\alpha}_j y_{t+1-j}^{AGG}.$$

This forecasting model does not take any disaggregate variables or information into account. The predictors are only the p lagged values of the aggregate. We refer to this method as AR in the following.

Aggregation of disaggregate forecasts

This approach estimates first for each disaggregate variable $x_1, \ldots, x_N$ the univariate autoregressive model:

$$x_{i,t} \quad = \quad c_{i,h} + \sum_{j=1}^{p_i} \alpha_{ij} x_{i,t+1-j-h} + \varepsilon_{i,t},$$

where $i = 1, \ldots, N$. p_i denotes the lags of the disaggregate variable $x_{i,t}$. Thus N individual forecasts for N disaggregate variables can be obtained with the estimated coefficients:

$$\hat{x}_{i,t+h|t} \quad = \quad \hat{c}_{i,h} + \sum_{j=1}^{p_i} \hat{\alpha}_{ij} x_{i,t+1-j}, \;\; i = 1, \ldots, N.$$

Then these forecasts are aggregated using the weighting vector w to construct the forecast of y_t^{AGG}:

$$\hat{y}_{t+h|t}^{AGG} = w_1 \hat{x}_{1,t+h|t} + \ldots + w_N \hat{x}_{N,t+h|t}.$$

This forecasting method is referred to as FD in the following.

Direct forecast using disaggregate information summarized in factors

A widely discussed method to extract the information contained in a large number of predictors uses a factor model which has been proposed in Stock and Watson (2002a,b). In their approach, the information contained in the N-dimensional stationary disaggregate variables $x_t = (x_{1,t}, \dots, x_{N,t})'$ can be summarized by a small number of r unobserved common factors F_t and an idiosyncratic part:

$$x_t = \Lambda F_t + e_t, \quad t = 1, \dots, T.$$

Λ is a $N \times r$ matrix of factor loadings, F_t is the $r \times 1$ vector of common factors and e_t is an $N \times 1$ vector of idiosyncratic components.

The forecast for the aggregate variable y_t^{AGG} by using factors as additional predictors can be carried out in a two-step process. In the first step, principle components of $\tilde{x}_t$ are calculated as estimated factors. $\tilde{x}_t$ is obtained by standardizing the disaggregate variables $x_t = (x_{1,t}, \dots, x_{N,t})'$ to have zero mean and unit variance (Stock and Watson (2002b)). The choice of the number of factors r may be based on the proposed information criteria of Bai and Ng (2002) or on the fraction of total variance in $\tilde{x}_t$ explained by the first r factors. Thus, r estimated factors can be obtained from this step. Stock and Watson (2002a) show that under mild conditions the principal components of $\tilde{x}_t$ are consistent estimators of the true unobservable factors. In the second step, the disaggregate information contained in the estimated factors serves as explanatory variables in the model:

$$y_t^{AGG} = c_h + \sum_{i=1}^{p} \alpha_i y_{t+1-i-h} + \sum_{i=1}^{r} \sum_{j=1}^{p_i} \beta_i \hat{F}_{i,t+1-j-h} + \varepsilon_t.$$

The h-step-forecast $\hat{y}_{t+h|t}^{AGG}$ can be then obtained with the estimated coefficients:

$$\hat{y}_{t+h|t}^{AGG} = \hat{c}_h + \sum_{i=1}^{p} \hat{\alpha}_i y_{t+1-i} + \sum_{i=1}^{r} \sum_{j=1}^{p_i} \hat{\beta}_i \hat{F}_{i,t+1-j}.$$

This forecasting model (referred to as FF1) has the advantage that it considers information from both, the aggregate variable (through the lags of the aggregate) and its disaggregate components (through the factors).

Since the target variable y_t^{AGG} is aggregated by using the disaggregate variables, it may be interesting to check whether taking only factors extracted from the disaggregate variables into account is helpful for improving the forecasting performance, thus the second model variant which has the following form can be considered (FF2):

$$y_t^{AGG} = c_h + \sum_{i=1}^{r}\sum_{j=1}^{p_j} \beta_i \hat{F}_{i,t+1-j-h} + \varepsilon_t.$$

The lagged values of the aggregate variable are not included in the forecasting model. Only the estimated factors are considered as predictors.

3 The boosting method

This paper suggests to use the boosting method to select a small subset of the most relevant predictors from a large set of candidate variables. Starting point is a forecasting model which has the following form:

$$\begin{aligned} y_t^{AGG} &= c_h + \sum_{i=1}^{p} \alpha_i y_{t+1-i-h}^{AGG} + \sum_{i=1}^{N}\sum_{j=1}^{p_i} \beta_{i,j} x_{i,t+1-j-h} + \varepsilon_t \\ &= c_h + \gamma' z_{t-h} + \varepsilon_t. \end{aligned}$$

In this model the lagged values of the aggregate, all the disaggregate variables and their lags are considered as candidate predictors. So there are $(p + p_1 + \cdots + p_N)$ candidate predictors in total, which may be summarized in z_t. By using the boosting procedure the predictors which have the largest contribution to the improvement of the model fit will be selected.

3.1 Review of the boosting procedure

The basic idea of boosting is to combine some simple function estimates such that the performance of each single function estimate can be enhanced. More precisely, it estimates an unknown function $f(x_t)$ for the target variable y_t as a sum of M estimated functions:

$$\hat{y}_t = \hat{f}^M(z_t) = f^{(0)} + \nu \sum_{m=1}^{M} \hat{g}^{(m)}.$$

$f^{(0)}$ represents the initial function value. z_t contains the $(p + p_1 + \cdots + p_N)$-dimensional predictor variables which are stationary. Originally, Bühlmann and Yu (2003) and Bühlmann and Hothorn (2007) assume independent and identically distributed predictors for the boosting algorithm. Bai and Ng (2009) extend the boosting idea to handle time-series problems. In their approach the lagged values of the predictors are allowed to be considered as candidate predictors as well.[2] The estimated functions $\hat{g}^{(m)}$ are derived by using a base learner. A base learner is a simple fitting procedure and repeatedly fits each candidate predictor or a small subset of the predictors against new refitted data. These refitted data are obtained based on the minimization of some loss functions. Thus in each iteration m, boosting adds the estimated base learner $\hat{g}^{(m)}$ providing the smallest loss. Without modifying the coefficients of those learners which are previously added to the model, only the parameters of the last base learner need to be estimated in each step. M is the final number of iterations. Friedman (2001) introduce the shrinkage parameter ν into the boosting algorithm to prevent overfitting. ν should be chosen small ($0 < \nu < 1$) so that the effect of the base learner $\hat{g}^{(m)}$ in each iteration step is weak and shrinks towards zero and an optimal solution can be achieved in small steps.

According to the specification of the target variable y_t, the base learners and the loss functions, different boosting algorithms are designed in literature. The well-known AdaBoost of Freund and Schapire (1996) can be applied for binary classification. If the function of interest $f(x_t)$ is the conditional mean $f(x_t) = E(y_t|x_t)$, the L_1-Boosting or L_2-Boosting can be used based on the form of the loss functions. Different from the L_1-loss, which is not differentiable at some points, the L_2-Boosting uses the squared errors as the loss function. It is very useful for regressions especially with many variables as candidate predictors. The loss function of the L_2-Boosting has the form[3]:

$$L(y_t, f) = \frac{1}{2}(y_t - f)^2.$$

[2] Bai and Ng (2009) propose two methods to deal with the lags. In the component-wise L_2 boost, each lag of each variable is treated as a separate candidate predictor, while the block-wise L_2 boost treats each variable and all its lags as a group.

[3] This term is scaled by the factor 1/2 so that the negative gradient vector can equal the residuals.

Componentwise L_2-Boosting

The specification of the base learner $g^{(m)}$ is of great importance. The L_2-Boosting chooses the linear least squares as the base learner for linear models. In each iteration the base learner is applied to each candidate variable in turn and only one variable will be selected in the sense of ordinary least squares fitting (hence componentwise). The componentwise boosting treats the lags of one predictor as separate predictors so that the variables and lags are selected simultaneously from a large set of candidates for forecasting (Bai and Ng (2009)). Note that for each iteration the same predictor variable or the lag can be selected. In the following we will only consider this boosting variant for our simulation and empirical study.

Componentwise L_2-Boosting algorithm

1. Start with $\hat{f}_t^{(0)} = \bar{y}$. Set $m = 0$.

2. Increase m by 1:

   - Compute the negative gradient of the loss function $-\frac{\partial L(y_t,f)}{\partial f}$ and evaluate at the estimate of the previous iteration $\hat{f}_t^{(m-1)}$:
     $u_t^{(m)} = y_t - \hat{f}_t^{(m-1)}$ for $t = 1, \ldots, T$ (current residual).
   - Regress $u_t^{(m)}$ on each predictor $z_{(k),t-h}$, $k = 1, \ldots, (p+p_1+\cdots+p_N)$ and compute $SSR_{(k)} = \sum_{t=1}^{T}(u_t - z_{(k),t-h}\hat{\theta}_{(k)})^2$.
   - Choose the predictor $z_{(k^*),t-h}$ which minimizes SSR.
   - Set $\hat{g}^{(m)} = \hat{\theta}_{(k^*)} z_{(k^*),t-h}$.
   - Update $\hat{f}^{(m)} = \hat{f}^{(m-1)} + \nu \hat{g}^{(m-1)}$.
   - Iterate step 2 until $m = m_{stop}$.

In the first step of the componentwise boosting, the mean of the target variable y_t is used as the initial value. So this starting value will be boosted with the later M iterations to improve the forecast. Next, the negative gradient of the loss function is computed and this is evaluated at the estimate of the previous iteration. As explained before the squared errors scaled by the factor 1/2 are used as the loss function, so we obtain for this step the current residual which is just the difference between the actual data y_t and the fitted value up to that iteration $\hat{f}_t^{(m-1)}$. The deviation of the fitted data in the previous step from the true data is then estimated by regressing $u_t^{(m)}$ on each of the $(p + p_1 + \cdots + p_N)$ candidate variables in turn. The variable $z_{k^*}^{(m)}$ which produces the smallest sum of squared errors is selected and can enter the next iteration with its fitted value $\hat{\theta}_{(k^*)}^{(m)} z_{(k^*),t-h}^{(m)}$ into the next step. Thus, $\hat{\theta}_{(k^*)}^{(m)} z_{(k^*),t-h}^{(m)}$ can be viewed as an approximation of the current residual u_t in the iteration step m. Suppose the variable z_{k^*} is selected in the iteration step m, it will then receive less attention in the step $m+1$ due to $u_t^{(m+1)} = y_t - \hat{f}_t^{(m)} = y_t - \hat{f}^{(m-1)} - \nu\hat{\theta}_{(k^*)}^{(m)} z_{(k^*),t-h}^{(m)}$, while the other non-selected variables have relatively larger chance to be selected in the step $m + 1$.[4] The choice of a small shrinkage parameter ν ensures that the improvement through $\hat{\theta}_{(k^*)}^{(m)} z_{(k^*),t-h}^{(m)}$ takes only limited contribution to the updates. With $\nu = 0.1$ for example, the variables which are selected 10 times (or more than 10 times) can contribute to the fit completely. Hence, the discrepancy between the true and fitted data is then continuously corrected in small steps and the most relevant predictors can be selected to improve the forecast performance.

The algorithm terminates when the final iteration step m_{stop} is reached. The estimate for the target variable y_t is:

$$\hat{y}_t = \bar{y} + \nu \sum_{m=1}^{m_{stop}} \hat{\theta}_{(k^*)}^{(m)} z_{(k^*),t-h}^{(m)}.$$

The forecast of the aggregate variable h-period ahead can be estimated as follows:

$$\hat{y}_{t+h|t}^{AGG} = \bar{y} + \nu \sum_{m=1}^{m_{stop}} \hat{\theta}_{(k^*)}^{(m)} z_{(k^*),t}^{(m)}.$$

Key parameters

There are two key parameters which should be considered carefully when applying the boosting algorithm: the shrinkage parameter ν and the final

[4]However, it is absolutely possible that the same variable will be selected in the next iteration step $m + 1$.

iteration step m_{stop}. As explained before, Friedman (2001) uses the shrinkage parameter ν ($0 < \nu < 1$) to ensure that the effect of the base learner $\hat{\theta}_{(k^*)} z_{(k^*)}$ is weak so that the deviations from the true data can be modified in small steps. Thus, if ν is chosen large, then in each step the selected variable $z_{(k^*)}$ will have larger contribution to the update with the value $\nu\hat{\theta}_{(k^*)} z_{(k^*)}$. In the following step other variables which are not selected previously receive more attention, thus they have an increasing chance to be selected. This can lead to overfitting, that is, too many predictors will be selected. On the other hand, if ν is chosen too small, the change of the current residuals in two steps $u_t^{(m+1)} - u_t^{(m)}$ is small as well so that the algorithm needs more iteration steps to identify a different predictor. Thus, a smaller value of the shrinkage parameter ν results in a larger number of optimal iteration steps m_{stop}. The larger m_{stop} becomes, more irrelevant predictors will be considered and selected. This leads again to the problem of overfitting. For the choice of the shrinkage parameter ν Friedman (2001) and Bühlmann and Hothorn (2007) suggest to use a sufficiently small value like $\nu = 0.1$ based on empirical evidence.[5]

For determining the optimal iteration number m_{stop} Bühlmann and Hothorn (2007) propose to use the corrected version of the Akaike Information Criterion (AIC)(Akaike (1973), Hurvich and Tsai (1989), Hurvich et al. (1998)) or the resampling techniques.

The corrected AIC for the iteration step m has the following form[6]:

$$AIC_c(m) = \log(\hat{\sigma}^2_{(m)}) + \frac{1 + df(m)/T}{1 - [df(m) + 2]/T}.$$

$\hat{\sigma}^2_{(m)}$ is the estimated residual variance and $df(m)$ are the degrees of freedom in the iteration step m. The iteration step number which minimizes $AIC_c(m)$ with $m = 1, \ldots, m^{max}$ will be selected as the optimal iteration number m_{stop}. The main problem of this approach is that the true degrees of freedom are

[5]For example, Buchen and Wohlrabe (2011), Wohlrabe and Buchen (2014) and Robinzonov et al. (2012) use the value of 0.1 for the shrinkage parameter, while Taieb and Hyndman (2014) set the value of ν to 0.2.

[6]The corrected AIC adds an additional nonstochastic penalty term to the standard AIC in order to have better small sample performance. More details can be found in Hurvich and Tsai (1989) and Hurvich et al. (1998).

unknown and should be estimated. Bühlmann and Hothorn (2007) suggest to estimate $df(m)$ by using the trace of the boosting hat matrix[7] in the iteration step m. However, Hastie (2007) acknowledges that this estimate generally underestimates the true degrees of freedom and thus the use of the corrected AIC always lead to select m_{stop} too large.[8]

In the literature it is also advised to use the resampling methods such as bootstrap, k-fold cross validation or subsampling to choose an appropriate number of iterations (Bühlmann and Hothorn (2007)). When using the bootstrap method to determine the m_{stop}, the bootstrapped samples are drawn with replacement from the original data. The data which are not drawn are used as the test sample. The k-fold cross validation divides the original dataset into k roughly equal subsamples and uses $k-1$ of them as the evaluation sample. The remaining one is retained for validation. In subsampling the dataset are randomly divided into two subsamples with the same size, one for evaluation and the other one for validation. Obviously, none of these resampling methods captures the dependence structure in the time series data appropriately. Thus selecting the optimal number of iterations m_{stop} based on the resampling techniques may be problematic when applying the boosting algorithm to the time series data and therefore and we do not use it here.[9]

3.2 Forecasting an aggregate variable using disaggregate variables selected by the componentwise L_2-Boosting

As in forecasting models FF1 and FF2 where factors are used as additional predictors we consider three variants here. The first variant (FB1) includes the lagged values of the aggregate variable and all the disaggregate variables

[7] Details are given in Bühlmann and Hothorn (2007).

[8] Based on this corrected version of AIC Chang et al. (2010) propose a change point detection method to avoid the problem of running all iterations under consideration, while Mayr et al. (2012) modify the approach of Chang et al. (2010) by using a sequential stopping rule for boosting.

[9] It should be noticed that Buchen and Wohlrabe (2011) and Wohlrabe and Buchen (2014) still use the resampling methods for analyzing time series data.

and their lags:

$$y_t^{AGG} = c_h + \sum_{i=1}^{p} \alpha_i y_{t+1-i-h}^{AGG} + \sum_{i=1}^{N} \sum_{j=1}^{p_i} \beta_{i,j} x_{i,t+1-j-h} + \varepsilon_t.$$

The variables or the lagged values which are most helpful to improve the model fit are selected by using the componentwise boosting algorithm discussed above. Then they are used as predictors to obtain the forecast of the aggregate variable $\hat{y}_{t+h|t}^{AGG}$.

The second variant (FB2) considered in this forecasting model focuses only on the disaggregate variables and their lags:

$$y_t^{AGG} = c_h + \sum_{i=1}^{N} \sum_{j=1}^{p_i} \beta_{i,j} x_{i,t+1-j-h} + \varepsilon_t.$$

The componentwise L_2-Boosting is used to select the most relevant predictors. The idea behind this variant is the same as in FF2: The aggregate variables are constructed by aggregating the disaggregate variables. When combining the aggregate and disaggregate information together, the information set is used twice. So it may be interesting to investigate whether only using information from the disaggregate variables provides a better forecasting performance.

The third variant (FB3) uses the partial boosting idea put forward in Tutz and Binder (2007). More precisely, we distinguish between the 'must have' predictors which must be included in the forecasting model and optional variables which can only stay in the forecasting model if they pass the boosting selection. The 'must have' predictors considered in FB3 are the lags of the aggregate variable:

$$y_t^{AGG} = c_h + \underbrace{\sum_{i=1}^{p} \alpha_i y_{t+1-i-h}^{AGG}}_{'must\ \ have'\ \ predictors} + \sum_{i=1}^{N} \sum_{i=1}^{p_i} \beta_{i,j} x_{i,t+1-j-h} + \varepsilon_t.$$

Thus with this variant we may check whether including additional selected disaggregate variables or lags to the forecasting model AR can help to improve the forecast accuracy. Taking into account the dependence structure of the time series data, this approach can be carried out in three steps. First, we estimate the following model by using the OLS method:

$$y_t^{AGG} = c_h^{AGG} + \sum_{i=1}^{p} \alpha_i y_{t+1-i-h}^{AGG} + u_t^{AGG}.$$

As a result, the estimated values of the coefficients $\hat{\alpha}_i (i = 1, \ldots, p)$, the constant $\hat{c}_h^{AGG}$ and residuals $\hat{u}_t^{AGG}$ can be obtained. Then each lag of the disaggregate variables is regressed on the same lags of the aggregate variables:

$$x_{i,t+1-j-h} = c_{ij,h} + \sum_{i=1}^{p} \alpha_i y_{t+1-i-h}^{AGG} + u_{i,t+1-j-h},$$

where $i = 1, \ldots, N$ and $j = 1, \ldots, p_i$. With the estimated residuals for each lag of the disaggregate variables we can use the componentwise L_2-Boosting for the following model:

$$\hat{u}_t^{AGG} = c + \sum_{i=1}^{N} \sum_{i=1}^{p_i} \gamma_{i,j} \hat{u}_{i,t+1-j-h} + \varepsilon_t.$$

Using the results from this boosting step the forecasting value for the residuals $\hat{u}_{t+h|t}^{AGG}$ can be computed. The estimated forecast $\hat{y}_{t+h}^{AGG}$ is obtained by:

$$\hat{y}_{t+h|t}^{AGG} = \hat{c}_h^{AGG} + \sum_{i=1}^{p} \hat{\alpha}_i y_{t+1-i}^{AGG} + \hat{u}_{t+h|t}^{AGG}.$$

4 Simulation study

In this section a simulation study is conducted which aims at 1) investigating whether the componentwise L_2-Boosting can select the disaggregate components of an aggregate variable correctly and 2) investigating whether the componentwise L_2-Boosting can select the most relevant predictors of an aggregate variable among a lot of candidate predictors.[10]

[10]For the componentwise L_2-Boosting we use the add-on package *mboost* of the statistical software *R* (version 2.15.2). Other programs for forecasting analysis are written in Gauss 12.

In this study 50 variables $x_{1,t}, \ldots, x_{50,t}$ with $t = 1, \ldots, T$ are generated by stationary $AR(1)$ processes without constant:

$$x_{i,t} = \alpha_i x_{i,t-1} + \varepsilon_{i,t} \quad i = 1, \ldots, 50,$$

where the coefficients α_i for $i = 1, \ldots, 50$ are uniformly distributed random numbers strictly between 0 and 1 and the residuals $\varepsilon_{i,t}$ are i.i.d. standard normal distributed random numbers. Hence, all the variables $x_{i,t}$ for $i = 1, \ldots, 50$ are stationary and independent.

Three variables $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ are used as relevant disaggregate components for computing the aggregate variable. The other 47 variables serve as irrelevant variables. The weighting matrix is set as $w = (0.2,\ 0.3,\ 0.5)'$. For each simulated time series the first 50 values are discarded to avoid the starting-up impact. Then the aggregate variable y_t^{AGG} is computed as[11]:

$$y_t^{AGG} = 0.2x_{1,t} + 0.3x_{2,t} + 0.5x_{3,t}.$$

At first we investigate whether using the componentwise L_2-Boosting method can indeed help to select the true disaggregate variables $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ among a lot of candidate variables, where the corrected AIC is used to find the optimal iterations number m_{stop} and the shrinkage parameter is chosen as $\nu = 0.1$[12]. m^{max} is set as 100. Different sample sizes $T = 100$, 500, 1000 and 2000 are generated. 5000 simulation runs are considered in this experiment.

To check the selection ability of the componentwise L_2-Boosting for the disaggregate components, we find that the true disaggregate components $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ are selected in all simulation runs for all considered different sample sizes (from small to large), while all other irrelevant variables cannot pass the boosting selection. This perfect variable selection ability of boosting can be explained by taking a deeper insight into the boosting algorithm and

[11]We have tried more than three relevant disaggregate variables for the aggregation and used other values for the weighting vector. The results are quite similar, so they are not explicitly reported here.

[12]Other values (e.g. 0.01, 0.05 and 0.2) for the shrinkage parameter have also been considered in our experiment. They affect mostly the optimal number of iterations m_{stop}. The main results are similar.

the aggregation structure of y_t^{AGG}. If y_t^{AGG} is aggregated by using only the variables $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$, then these three variables are correlated with y_t^{AGG}. Since all the generated series $x_{1,t}, \dots, x_{50,t}$ are independent, so the other irrelevant variables $x_{4,t}, \dots, x_{50,t}$ should not be correlated with y_t^{AGG}.

In the first iteration step of the componentwise boosting, using the mean of the aggregate variable $\bar{y}^{AGG}$ as the initial value, the current residuals $u_t^{(m=1)}$ can be calculated as:

$$\begin{aligned} u_t^{(m=1)} &= y_t^{AGG} - \bar{y}^{AGG} \\ &= 0.2x_{1,t} + 0.3x_{2,t} + 0.5x_{3,t} - \bar{y}^{AGG}. \end{aligned}$$

Since $\bar{y}^{AGG}$ is a constant value, thus the correlation coefficients between y_t^{AGG} and the variables should be equal to the correlation coefficients between $u_t^{(m=1)}$ and the variables. The results for the correlation coefficients between $u_t^{(m=1)}$ and the considered variables in Table 1 indicate that with an increasing T the correlations between y_t^{AGG} and its disaggregate components $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ converge to their real values which are significant different from zero, while the correlations between y_t^{AGG} and irrelevant variables go towards zero. With $T = 2000$, the mean of the absolute value of the correlation coefficients from 5000 simulation runs between the current residuals in the first iteration step $u_t^{(m=1)}$ and the true disaggregate variables are 0.335, 0.479 and 0.766. For three arbitrarily chosen irrelevant variables $x_{8,t}$, $x_{25,t}$ and $x_{46,t}$ these values are around 0.027, which indicates that $x_{8,t}$, $x_{25,t}$ and $x_{46,t}$ are asymptotically uncorrelated with $u_t^{(m=1)}$.

The variable which is mostly correlated with $u_t^{(m=1)}$ should provide the best least squares fitting for $u_t^{(m=1)}$. In our simulation example, the variable $x_{3,t}$ which has the largest correlation coefficient[13] is then selected in the first step and can enter with its fitted value $\hat{\theta}_3^{(m=1)} x_{3,t}$ into the second iteration. The fitted value for the aggregate in the first step is obtained as: $\hat{f}_t^{(m=1)} = \bar{y}^{AGG} + \nu \hat{\theta}_3^{(m=1)} x_{3,t}$. In the second step, the variable $x_{3,t}$ will be less weighted

[13]The variable $x_{3,t}$ has the largest weight (0.5) by aggregating, so it is mostly correlated with the aggregate variable for all considered sample sizes.

in $u_t^{(m=2)}$ because of:

$$\begin{aligned} u_t^{(m=2)} &= y_t^{AGG} - \hat{f}_t^{(m=1)} \\ &= y_t^{AGG} - \bar{y}^{AGG} - \nu\hat{\theta}_3^{(m=1)} x_{3,t} \\ &= 0.2x_{1,t} + 0.3x_{2,t} + 0.5x_{3,t} - \bar{y}^{AGG} - \nu\hat{\theta}_3^{(m=1)} x_{3,t} \\ &= 0.2x_{1,t} + 0.3x_{2,t} + [0.5 - \nu\hat{\theta}_3^{(m=1)}]x_{3,t} - \bar{y}^{AGG}. \end{aligned}$$

Hence, compared with the correlations with $u_t^{(m=1)}$, in the second iteration step $x_{3,t}$ is less correlated with $u_t^{(m=2)}$ and the other two true variables $x_{1,t}$ and $x_{2,t}$ have relatively larger correlation coefficients with $u_t^{(m=2)}$, while the irrelevant variables are still not correlated with $u_t^{(m=2)}$ and therefore cannot contribute to the fit of $u_t^{(m=2)}$. As long as $x_{3,t}$ is mostly correlated with the current residuals in the iterations, it will be selected by boosting. The selection of the same variable in different iteration steps leads to an improvement of the fit, namely, the discrepancy between the true and fitted data are corrected. Once another variable has the largest correlation coefficient with the corresponding current residuals, this variable will be selected. If m_{stop} is chosen appropriately, then the irrelevant variables which are not used for the aggregation have no chance to be selected by boosting since they are not correlated with y_t^{AGG} and therefore not correlated with the current residuals in the iterations.[14]

The strength of the correlation between the aggregate variable and its disaggregate components depends obviously on the aggregation weights. If a disaggregate variable has a very small weight in aggregating, then it will be weakly correlated with the aggregate variable at the beginning of the boosting iteration. However, this variable can still be selected by boosting. This is because the relevant variables which are already selected in the former iterations will be less weighted in the current residuals $u_t^{(m)}$ of the later iterations. Thus the disaggregate variable which is originally weakly correlated with y_t^{AGG} increases its correlations with the current residuals and can be selected, although more iteration steps are needed. The irrelevant variables cannot affect its correlations with the current residuals with an increasing

[14]However, if m_{stop} is chosen too large, then $u_t^{(m)}$ will be so small that even irrelevant variables can be selected.

number of iteration steps.[15]

Note that for this simulation experiment the corrected AIC always selects $m_{stop} = 100$ and in our simulation for applying the corrected AIC the m_{max} is set as 100. In reality, in order to identify the three disaggregate variables, less than 100 iteration steps are needed. More iterations are suggested by the corrected AIC so that the discrepancy between the true and fitted data can be more precisely corrected. To be more clearly, by examining an arbitrarily selected simulation example with $T = 100$, the boosting method requires only 17 iteration steps to identify the three true disaggregate variables $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$. With these 17 iteration steps, the estimated relationship is:

$$\hat{y}_t^{AGG} = -0.024 + 0.019x_{1t} + 0.111x_{2t} + 0.376x_{3t}.$$

If the maximal possible iteration numbers m^{max} is set as 100, then the corrected AIC suggests $m_{stop} = 100$ as well. With 100 iteration steps these true disaggregate variables are still selected, however, with another estimated relationship:

$$\hat{y}_t^{AGG} = -0.001 + 0.191x_{1t} + 0.290x_{2t} + 0.493x_{3t}.$$

If m^{max} is set as 200, then the corrected AIC selects $m_{stop} = 200$ and the boosting method provides the following estimated relationship:

$$\hat{y}_t^{AGG} = -0.00004 + 0.1997x_{1t} + 0.2998x_{2t} + 0.4998x_{3t}.$$

Thus with more iterations the weights for the disaggregate variables can be more precisely estimated.

Although the boosting method can select the true disaggregate variables of an aggregate quite correctly, a problem will occur when forecasting this aggregate, namely, only the past values can be used as predictors. Since y_t^{AGG} is aggregated by $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ which are stationary $AR(1)$ series, we can expect that $x_{1,t-1}$, $x_{2,t-1}$ and $x_{3,t-1}$ can be used as reasonable predictors for forecasting y_t^{AGG}. Furthermore, the sum of three stationary $AR(1)$

[15] In the simulation experiments, we have also considered very small values (e.g. 0.01) for the weights, the selection rate is still 1 for these disaggregate variables with small weights.

processes can be represented by an $ARMA(p,q)$ process where $p \leq 2$ and $q \leq 2$ (Granger and Morris (1976)), so using the lagged values of y_t^{AGG} as predictors makes sense as well. Now, from the group of the candidate predictors y_{t-1}^{AGG} and $x_{1,t-1}, \ldots, x_{50,t-1}$, we check wether y_{t-1}^{AGG}, $x_{1,t-1}$, $x_{2,t-1}$ and $x_{3,t-1}$ can be detected by using the componentwise L_2-Boosting.[16] Note that this simulation setup also mimics the empirical forecasting exercise in the later section.

Table 2 shows the selection rates from 5000 simulation runs for the relevant predictors y_{t-1}^{AGG}, $x_{1,t-1}$, $x_{2,t-1}$ and $x_{3,t-1}$. Three irrelevant candidate predictors $x_{8,t-1}$, $x_{25,t-1}$ and $x_{46,t-1}$ are arbitrarily chosen to see how these irrelevant predictors are selected by the boosting method. Furthermore, the average selection rates for all 47 irrelevant candidate predictors are reported in the last column of Table 2.

First of all, it can be seen that the lagged values of y_t^{AGG} are selected most often. If the sample sizes are larger than 100, in more than 90% of all 5000 simulation runs y_{t-1}^{AGG} is selected by boosting. The selection rate of y_{t-1}^{AGG} does not depend heavily on the number of iterations. Even a very small number of iterations $m_{stop} = 20$ results in a selection rate of 79.8% for y_{t-1}^{AGG} by $T = 100$, while with 80 iterations the selection rate increases slightly to 80.9%. For other sample sizes, similar results can be observed.

Compared to y_{t-1}^{AGG}, a relatively poor selection rate of the relevant predictors $x_{1,t-1}$, $x_{2,t-1}$ and $x_{3,t-1}$ can be observed in Table 2. For a small sample size $T = 100$, these relevant predictors are selected in less than 60% of all 5000 simulation runs. With an increasing sample size the selection rates for these predictors increase as well, however, they are still less than 90% even for $T = 2000$. This is due to the selection of the variable y_{t-1}^{AGG} which is aggregated by $x_{1,t-1}$, $x_{2,t-1}$ and $x_{3,t-1}$. If y_{t-1}^{AGG} is not used as a candidate predictor, then $x_{1,t-1}$, $x_{2,t-1}$ and $x_{3,t-1}$ have a much more larger chance to be selected[17].

For the irrelevant variables which should not be considered as predictors for forecasting y_t^{AGG}, relatively large selection rates can be observed in Table

[16] In the simulation study, we also included more lags in the candidate group. The results are quite robust.

[17] Detailed results are not presented here, but are available on request.

2, especially if the corrected AIC is used to determine m_{stop}. For example, considering $m_{stop}1$ which is suggested by the corrected AIC using $m^{max} = 100$, from 5000 simulation runs the variable $x_{25,t-1}$ are selected 1640 times which results in a selection rate of 32.8% for a small sample size of $T = 100$. If the sample size is increased to $T = 2000$, boosting still selects the variable $x_{25,t-1}$ in 27.3% of all 5000 simulation runs. When increasing m^{max} to 200, using $m_{stop}2$ leads to a selection rate of 40.6% of all irrelevant predictors in average. This is associated with the problem of using the corrected AIC. As explained in Section 3, the true degrees of freedom in each iteration step $df(m)$ are underestimated by using the trace of the boosting hat matrix suggested in Bühlmann and Hothorn (2007), thus the corrected AIC always overestimates the true m_{stop} which leads to overfitting. In Figure 1, the m_{stop} estimated by using the corrected AIC criterion for 200 simulations runs for different sample sizes are plotted. The left column shows the results of $m_{stop}1$ which is determined by using the corrected AIC, when m^{max} is set as 100, while the right column plots $m_{stop}2$ based on $m^{max} = 200$. It can be seen that if m^{max} is set as 100, then in many iteration runs the corrected AIC chooses the maximum number m^{max} as m_{stop}, especially when a large sample size $T = 2000$ is considered. However, using $m^{max} = 200$, the corrected AIC tends to choose the optimal iteration numbers between 100 and 200. With more iteration numbers, the irrelevant predictors have a larger chance to be selected.

Therefore we also add results for the selection rates in Table 2 if smaller numbers of iterations $m_{stop} = 20$, 40, 60 and 80 are used. As expected, a smaller m_{stop} leads to sparse models. For large sample sizes $T = 1000$ and $T = 2000$ the irrelevant variables are selected in less than 60 simulations runs with $m_{stop} = 20$. However, the selection rates for the relevant predictors $x_{1,t-1}$, $x_{2,t-1}$ and $x_{3,t-1}$ decrease as well if smaller numbers of iterations are used for boosting, but they are still much more larger than the selection rates of the irrelevant predictors.

From the simulation experiments discussed above, we find that using the boosting method can detect the disaggregate components of an aggregate variable quite exactly. With respect to the selection of the relevant predictors for this aggregate, boosting selects most likely the lagged values of the aggregate itself and the relevant disaggregate predictors have always much more larger chance to be selected than the irrelevant predictors.

5 Empirical Analysis

We now apply the the componentwise L_2-Boosting on the Euro area aggregate variables and evaluate the forecasting performance of this method.

5.1 Data

To compare the forecast accuracy of the model which uses the boosting method to select the most relevant predictors with other models described in Section 2, six macroeconomic key variables for the Euro area with quarterly frequency from 1970Q1 to 2011Q4 are considered: the real GDP (YER), the consumer price index (CPI), the GDP deflator (YED), the exchange rate against the US-Dollar (EER) and short- and long-term interest rates (STN and LTN).

The data for the aggregated macroeconomic variables are obtained from the Area Wide Model (AWM) database maintained at the Euro Area Business Cycle Network[18]. The Area Wide Model is proposed by Fagan et al. (2001) and Fagan et al. (2005) who suggest to construct the aggregates as the weighted average of log-level data for 11 European countries. The fixed weights used in aggregating the individual country series are the nominal GDP shares for 1995[19]. The three largest European countries Germany, France and Italy obtain the largest weights 0.305, 0.210 and 0.203 due to their economic leading roles in Euro area, while quite small weights are assigned to countries like Luxembourg and Finland.

[18]http://www.eabcn.org. The 12th update of the AWM database is used which extends the data up to 2011Q4. It should be noted that the AWM database is not an official ECB database.

[19]Weights used in aggregation: Belgium 0.039, Germany 0.305, Spain 0.102, France 0.210, Ireland 0.011, Italy 0.203, Luxembourg 0.002, Netherlands 0.056, Austria 0.030, Portugal 0.024, Finland 0.017. Since Greece, which joined the EMU in 2001 and therefore is not considered in Fagan et al. (2001) and Fagan et al. (2005), has the weight zero in the forecasting comparison.

The individual member countries' time series data for the same period from 1970Q1 to 2011Q4 are taken from the OECD quarterly national accounts database.[20] Twelve eurozone countries are considered. (Greece, which joined the EMU in 2001, is also included.) All the series are seasonally adjusted and transformed so that they have the same base year. Thus, for each macroeconomic aggregate variable 12 disaggregate European countries data are observed ($N = 12$).

Logarithms are taken on the real GDP, the CPI, the GDP deflator and the exchange rate EER, while the short- and long-term interest rates are left in levels. A formal unit root analysis has been conducted and all the time series for all countries can be characterized as $I(1)$ processes.[21] To ensure the stationarity the first differences of the variables enter the forecasting models. To estimate the common factors, the first differences of the series are standardized to have zero mean and unit variance. For the time series with missing observations which are listed in Table 4, the expectation-maximization (EM) algorithm proposed in Stock and Watson (2002a) is applied to construct a balanced panel[22].

5.2 Comparison methodology

A recursive pseudo-out-of-sample forecasting experiment is conducted for comparing the performance of the forecasting models discussed in Section 2 and Section 3. That is, the starting point of the estimation period is fixed at 1970Q1. In the first stage we use the data from 1970Q1 to 2002Q4. Factors estimation and variable selection via boosting are all carried out with this information set. Using the forecasting horizons $h = 1, 2$ and 4, forecasts for 2003Q1, 2003Q2 and 2003Q4 can be computed. In the next stage, the estimation period is enlarged with one observation up to 2003Q1. For this new estimation period factors are then re-estimated and variables are re-selected by boosting. The obtained estimates are used for forecasting the values at 2003Q2, 2003Q3 and 2003Q4. We continue this process until the estimation period ends in 2010Q4 and forecasts can be calculated for 2011Q1,

[20]The data are obtained via Thomson Reuters Datastream 5.0.

[21]The results for the unit root tests are available on request.

[22]A list of variables with missing observations can be found in Table 4 in Appendix.

2011Q2 and 2011Q4. Hence, the forecasting covers the period from $2003Q1$ to $2011Q4$.

To evaluate the forecasting accuracy, the mean squared forecast error (MSFE) is used as loss function, where the forecast error is defined as the difference between the estimated forecasting value and its actual value $e_{t+h} = \hat{y}^{AGG}_{t+h|t} - y^{AGG}_{t+h}$. The first forecasting model (AR) with a fixed lag length $p = 4$ is used as benchmark. To assess the forecasting accuracy, the MSFE obtained from other forecasting models (and their variants) will be expressed relative to the MSFE obtained from the first forecasting model. Thus, any value of the relative MSFE less than one indicates an improvement of the forecasting performance against the benchmark.

For the empirical results reported here, a fixed number of four lags for the autoregressive part and two lags for all disaggregate variables are used. Results for the other choices of the lag lengths are not reported, since similar results have been obtained and they are quite robust with respect to this choice.

The forecasting variants FF1 and FF2 extract factors by using only the country information on the aggregate variable to be predicted. For instance, when forecasting the Euro area real GDP, the factors are estimated from the dataset that only includes real GDP from the 12 member countries and no other variables are taken into account. The number of factors used in FF1 and FF2 has to be determined. When using the information criteria proposed by Bai and Ng (2002), we found that the maximum number of factors are suggested by all three criteria. Giving the relatively low number of disaggregate variables ($N = 12$), this is not a meaningful choice. Therefore, we decided to explore the cumulative percentage of the variance in the dataset that can be explained by the factors. For all possible estimation periods which start in 1970Q1 and end in $2002Q4, \ldots, 2010Q4$, it can be found that for each variable at least 35% of the total variance in the respective data set can be explained by two factors. Accordingly two factors with respectively two lags are used in FF1 and FF2.[23]

For the three variants FB1, FB2 and FB3 which use the componentwise

[23]Only one lag or more than two lags for the factors are also considered in our experiment. The main results do not depend on the choice of the number of factor lags.

L_2-Boosting to select the most relevant predictors, the shrinkage parameter ν is set to the commonly used value of 0.1 (Bühlmann and Hothorn (2007))[24]. Four lags for the aggregate variable and two lags for all disaggregate variables are considered[25]. Thus for the model variant FB1 there are 28 candidate predictors, while for the model variant FB2 the boosting method selects from 24 candidate variables. For the last model variant FB3, the 'must have' predictors include the four lags of the aggregate variable and relevant predictors should be selected from the 24 lags of the disaggregate components. Although we found from our simulation study that the use of the corrected AIC for determining the optimal iteration numbers m_{stop} is likely to select irrelevant predictors, however, by reducing m_{stop} the true predictors will also have less chance to be selected. Therefore, in this empirical study the crucial parameter m_{stop} is determined by the corrected version of AIC, since we do not include irrelevant variables and boosting should select the most relevant ones from all possible candidate predictors.[26] Regarding the effect of the maximum iteration number m^{max} on m_{stop}, we consider $m^{max} = 100$ and $m^{max} = 200$ for the forecasting comparison.

5.3 Empirical results

Detailed results from the forecasting comparison for all the considered macroeconomic aggregate variables are presented in Table 3. For each variable and forecasting model variant, we report the MSFE of the respective model variant relative to the MSFE of the forecasting model which uses only the lagged values of the aggregate variable as predictors (AR). Forecasting horizons $h = 1, 2$ and 4 are considered. The estimation periods start at 1970Q1 and the forecasting evaluation covers the periods from $2003Q1$ to $2011Q4$.

[24]By choosing ν less than 0.1, the corrected AIC chooses in most cases $m_{stop} = 100$ if $m^{max} = 100$.

[25]Results for the other choices of the lag length are similar.

[26]We have also tried smaller m_{stop} like 20, 40, 60 and 80 as in the simulation study. With a smaller number of iteration numbers, the relevant predictors have less chance to be selected by the boosting method, thus using smaller m_{stop} the variants FB1, FB2 and FB3 show worse forecasting results compared to the use of the corrected AIC for determining m_{stop}.

The smallest MSFE ratio for each variable and forecasting horizon is in bold. Table 3 suggests several conclusions.

First, combining the disaggregate forecasts does not necessarily provide more accurate forecasts than the forecasts based on the aggregate variable when forecasting the macroeconomic key variables for the Euro area. For example, when forecasting the growth rate of the real GDP (YER), the relative MSFEs for the forecasting model FD at the horizons $h = 1$ and $h = 4$ are larger than one. Similar results hold for some other variables especially for YED and STN where on all horizons the model FD provide worse forecasts compared to the benchmark AR. Only for the variables CPI and LTN, gains can be observed by pooling all the country-specific forecasts together.

Second, overall there is evidence that taking into account disaggregate information in form of factors or selected disaggregate variables can improve the forecasting performance. For the variable YER, forecasting models and variants based on factors or boosting procedure lead for almost all horizons to smaller MSFEs. For the other variables, the use of disaggregate information or variables may not lead to sizable gains in forecasting precision in all cases, however, for some forecasting horizons such gains are always observable.

Third, it is difficult to find an appropriate way to incorporate the disaggregate information or disaggregate variables in the forecasting model. This depends mainly on the forecasting problem and the variables which are investigated. In our empirical exercise, in some cases using factors as predictors tends to have lower MSFEs, for example for YER at $h = 1$. However, in many cases the boosting method can indeed improve the forecasting performance compared to the factor approach. Furthermore, in Table 3 the smallest MSFE ratio for each variable and forecasting horizon is in bold. It can be seen that especially for the variables CPI, EER and interest rates, using boosting to select the predictors yields for all horizons the smallest relative MSFEs.

Fourth, when comparing the model variants FF1 with FF2 and FB1 with FB2, it can be found that there is no compelling evidence that using only the disaggregate information or variables, without considering the lagged values of the aggregate variables in the forecasting model, is helpful in improving the forecasting precision, since the second variant of the forecasting model which uses only factors as predictors and the second variant of the boosting procedure perform in most cases worse than the respective first variant. For example, when forecasting CPI at all forecasting horizons, using factors extracted from all disaggregate variables as single predictors lead to larger MSFE than the variant FF1 where the lagged values of CPI are used as predictors as well. Similarly, using the model variant FB1 can improve the forecasting performance in most cases. Generally, Table 3 indicates that even if the aggregate variables are constructed by aggregating the disaggregate variables, the information hidden in the aggregate variables is important and therefore should be considered for forecasting.

Fifth, the model variant FB3 is an extension of the benchmark forecasting model AR, since in FB3 the lagged values of the aggregate variable are treated as 'must have' predictors. The boosting method selects among the disaggregate variables and adds the selected ones as additional predictors. Thus, by comparing AR and FB3 it can be seen that for YER at horizons $h = 2$ and $h = 4$, for CPI at $h = 1$ and $h = 2$, for YED at $h = 1$ and EER and interest rates at all horizons, using FB3 leads to an improvement in the forecasting accuracy. This means that adding some additional selected predictors to the model AR can be helpful. Compared to the model variants FB1 and FB2, using FB3 cannot achieve better forecasting results.

Sixth, the optimal iteration number m_{stop} determined by the corrected AIC is a key parameter which has an impressive effect on the forecasting results. If m^{max} is chosen as 200, then the boosting procedure stops with more iteration numbers. In our empirical experiment, it can be seen that the use of $m_{stop}2$ leads to an improvement in the forecasting performance in almost all cases. The reason for this is that we do not include any irrelevant variables in the forecasting models. All the lags of the aggregate and disaggregate variables are reasonable candidate predictors, thus with more iterations the boosting method selects the most relevant ones from them.

6 Conclusion

The main purpose of this paper is to check whether taking into account disaggregate information or variables is helpful for improving the forecasting accuracy of an aggregate variable. A related question is how to include the disaggregate information or the relevant disaggregate variables in the direct forecasting model appropriately. A new variable selection method, the boosting procedure is introduced and discussed.

A simulation study is performed to investigate the variable selection performance of the componentwise L_2-Boosting method. We find that the true disaggregate components of an aggregate variable can be detected by the boosting method. Moreover, when using boosting to select the most relevant predictors, care must be taken. If the lagged values of the aggregate variable are used as candidate predictors, then they will be most likely selected by boosting. Once the lagged values of the aggregate variable are already selected as predictors, the lagged values of the disaggregate variables have smaller chance to be selected. However, compared to the irrelevant predictors, all relevant predictors have an reasonable chance to be selected.

By forecasting the six macroeconomic variables for the Euro area, four forecasting models (and their variants) are compared: (1) the forecast of an aggregate which uses only the lagged values of the aggregate as predictors, (2) first forecasting the disaggregate variables respectively and then aggregating the disaggregate forecasts, (3) including disaggregate information with the help of the factor model as additional predictors in the forecasting model and (4) including disaggregate variables selected by boosting in the forecasting model. The empirical results indicate that in comparison to the forecasting model which uses only past values of the aggregate variable, using disaggregate information summarized in the factors or using disaggregate variables selected by boosting as additional predictors may provide more accurate forecasts. Moreover, it is shown that using boosting to select disaggregate variables outperforms the benchmark for all considered variables. Compared to the factor approach, boosting is a competitive approach of exploiting the high-dimensional disaggregate data.

Table 1: The mean of the absolute value of the correlation coefficients from 5000 simulation runs between the current residuals u_t of the first three iteration steps and the true disaggregate and the irrelevant variables, respectively.

	$x_{1,t}$	$x_{2,t}$	$x_{3,t}$	$x_{8,t}$	$x_{25,t}$	$x_{46,t}$	AIV
$T = 100$							
$u_t^{(m=1)}$	0.328	0.484	0.779	0.116	0.118	0.116	0.116
$u_t^{(m=2)}$	0.345	0.506	0.760	0.115	0.118	0.116	0.116
$u_t^{(m=3)}$	0.363	0.528	0.739	0.115	0.118	0.116	0.116
$T = 500$							
$u_t^{(m=1)}$	0.329	0.477	0.774	0.054	0.054	0.054	0.054
$u_t^{(m=2)}$	0.348	0.500	0.756	0.054	0.054	0.054	0.054
$u_t^{(m=3)}$	0.366	0.521	0.737	0.053	0.054	0.053	0.054
$T = 1000$							
$u_t^{(m=1)}$	0.332	0.481	0.767	0.038	0.039	0.038	0.038
$u_t^{(m=2)}$	0.349	0.504	0.749	0.038	0.039	0.038	0.038
$u_t^{(m=3)}$	0.367	0.525	0.730	0.038	0.038	0.038	0.038
$T = 2000$							
$u_t^{(m=1)}$	0.335	0.479	0.766	0.027	0.027	0.028	0.027
$u_t^{(m=2)}$	0.352	0.501	0.748	0.027	0.027	0.028	0.027
$u_t^{(m=3)}$	0.370	0.522	0.729	0.027	0.027	0.027	0.027

Note: The true disaggregate variables are $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$. $x_{8,t}$, $x_{25,t}$ and $x_{46,t}$ are arbitrarily chosen from 47 irrelevant variables. The mean of the absolute value of the correlation coefficients from 5000 simulation runs between u_t of the iteration steps $m = 1$, $m = 2$ and $m = 3$ and all irrelevant variables are reported in average in the last column. AIV refers to the average of all irrelevant variables. The sample sizes are chosen as $T = 100$, 500, 1000 and 2000.

Table 2: The selection rates for relevant and irrelevant candidate predictors.

	y_{t-1}^{AGG}	$x_{1,t-1}$	$x_{2,t-1}$	$x_{3,t-1}$	$x_{8,t-1}$	$x_{25,t-1}$	$x_{46,t-1}$	AIV
$T = 100$								
$m_{stop} = 20$	0.798	0.258	0.309	0.354	0.071	0.074	0.070	0.073
$m_{stop} = 40$	0.806	0.367	0.378	0.396	0.172	0.166	0.171	0.168
$m_{stop} = 60$	0.808	0.435	0.428	0.421	0.242	0.233	0.245	0.239
$m_{stop} = 80$	0.809	0.486	0.472	0.447	0.300	0.290	0.301	0.294
$m_{stop}1 = AIC_c$	0.810	0.511	0.499	0.462	0.334	0.328	0.335	0.330
$m_{stop}2 = AIC_c$	0.812	0.562	0.555	0.511	0.407	0.396	0.409	0.404
$T = 500$								
$m_{stop} = 20$	0.946	0.365	0.371	0.401	0.016	0.021	0.023	0.020
$m_{stop} = 40$	0.954	0.596	0.512	0.461	0.107	0.116	0.111	0.111
$m_{stop} = 60$	0.955	0.667	0.579	0.490	0.190	0.191	0.191	0.197
$m_{stop} = 80$	0.955	0.713	0.622	0.506	0.261	0.262	0.256	0.267
$m_{stop}1 = AIC_c$	0.955	0.738	0.653	0.521	0.312	0.315	0.310	0.318
$m_{stop}2 = AIC_c$	0.956	0.772	0.699	0.553	0.402	0.404	0.396	0.408
$T = 1000$								
$m_{stop} = 20$	0.960	0.384	0.390	0.391	0.009	0.012	0.010	0.010
$m_{stop} = 40$	0.970	0.662	0.572	0.465	0.078	0.078	0.075	0.081
$m_{stop} = 60$	0.971	0.734	0.661	0.499	0.164	0.158	0.166	0.170
$m_{stop} = 80$	0.971	0.770	0.707	0.517	0.242	0.236	0.244	0.246
$m_{stop}1 = AIC_c$	0.971	0.792	0.738	0.529	0.297	0.291	0.297	0.302
$m_{stop}2 = AIC_c$	0.972	0.824	0.781	0.561	0.407	0.394	0.394	0.406

(continued on the next page)

Table 2 (continued)

	y_{t-1}^{AGG}	$x_{1,t-1}$	$x_{2,t-1}$	$x_{3,t-1}$	$x_{8,t-1}$	$x_{25,t-1}$	$x_{46,t-1}$	AIV
$T = 2000$								
$m_{stop} = 20$	0.974	0.398	0.396	0.395	0.003	0.004	0.003	0.004
$m_{stop} = 40$	0.984	0.742	0.627	0.480	0.056	0.053	0.053	0.053
$m_{stop} = 60$	0.985	0.808	0.727	0.525	0.138	0.128	0.146	0.137
$m_{stop} = 80$	0.986	0.833	0.771	0.543	0.216	0.210	0.218	0.215
$m_{stop}1 = AIC_c$	0.986	0.854	0.799	0.558	0.278	0.273	0.278	0.279
$m_{stop}2 = AIC_c$	0.986	0.876	0.835	0.582	0.407	0.399	0.410	0.406

Note: y_{t-1}^{AGG}, $x_{1,t-1}$, $x_{2,t-1}$ and $x_{3,t-1}$ are relevant predictors. $x_{8,t-1}$, $x_{25,t-1}$ and $x_{46,t-1}$ are arbitrarily chosen from 47 irrelevant variables and serve as candidate predictors which are actually not relevant. The average selection rates for all irrelevant predictors are reported in the last column. AIV refers to the average of all irrelevant variables. The number of simulation runs is 5000. $m_{stop}1 = AICc$ is determined by using the corrected AIC, when m^{max} is set as 100. $m_{stop}2 = AICc$ is determined by using the corrected AIC, when m^{max} is set as 200.

Table 3: MSFEs of different forecasting models relative to the first forecasting model AR

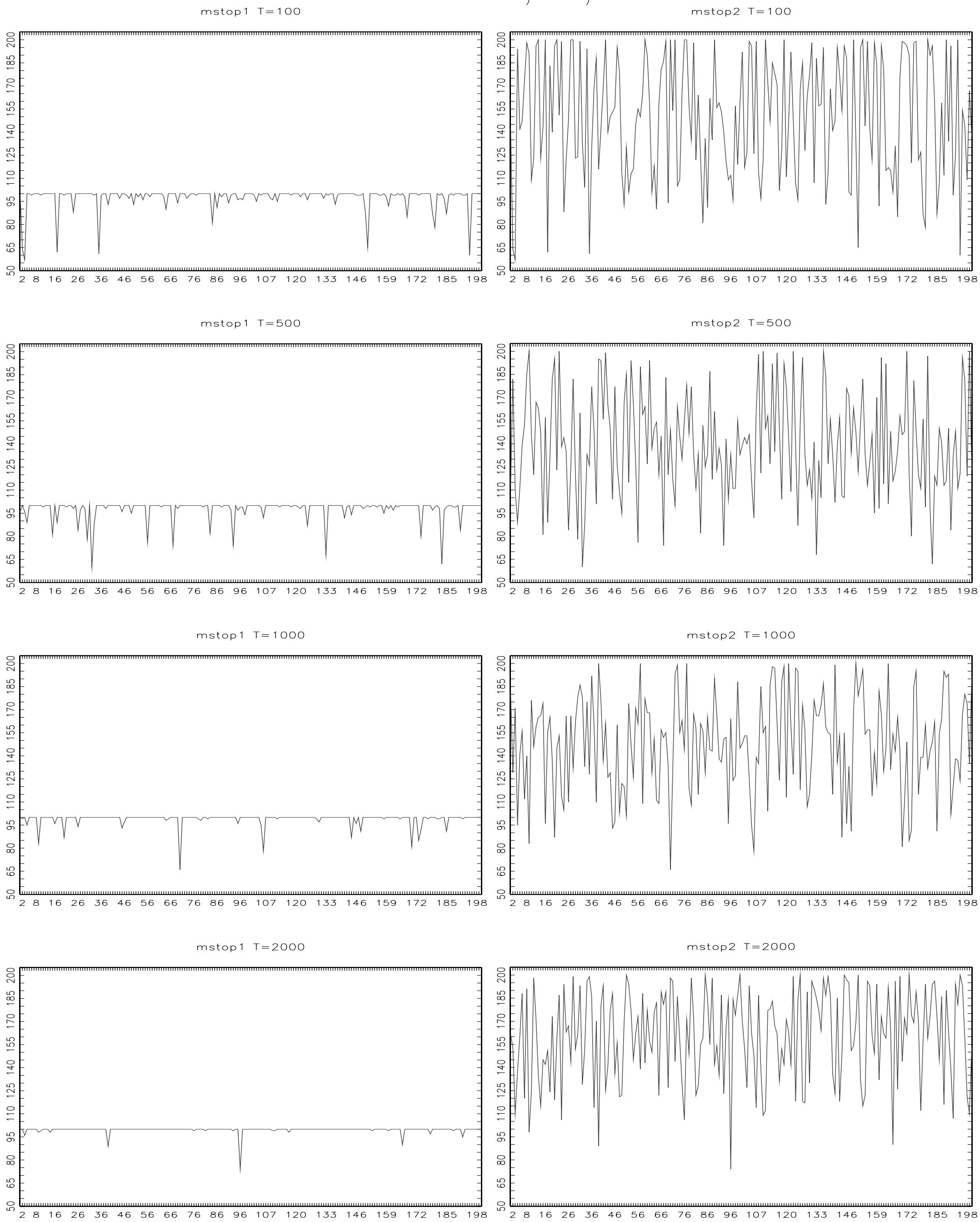
	FD	FF1	FF2	FB1	FB2 ($m_{stop}1$)	FB3	FB1	FB2 ($m_{stop}2$)	FB3
					YER				
$h=1$	1.010	0.833	**0.825**	0.907	0.907	1.000	0.877	0.858	0.993
$h=2$	0.972	0.964	0.975	0.961	0.963	0.980	0.946	**0.944**	0.971
$h=4$	1.020	0.970	0.993	0.931	0.942	0.930	0.924	0.938	**0.912**
					CPI				
$h=1$	0.965	1.002	1.008	0.974	1.010	0.916	0.959	0.998	**0.903**
$h=2$	0.943	1.008	1.019	0.931	0.937	0.972	**0.906**	0.920	0.971
$h=4$	0.970	1.015	1.085	0.967	0.993	1.020	**0.953**	0.988	1.006
					YED				
$h=1$	1.254	**0.890**	0.896	0.908	0.982	0.944	0.967	0.974	0.920
$h=2$	1.371	**0.985**	1.084	1.124	1.313	1.022	1.149	1.319	1.010
$h=4$	1.264	1.013	1.065	0.958	1.013	1.285	**0.941**	1.001	1.259
					EER				
$h=1$	0.994	1.004	1.014	0.955	0.957	0.989	**0.952**	0.954	0.986
$h=2$	0.996	1.008	0.981	**0.924**	0.925	0.980	**0.924**	**0.924**	0.979
$h=4$	1.010	0.992	0.990	0.975	0.977	0.982	**0.972**	**0.972**	0.979
					STN				
$h=1$	1.041	1.013	0.999	0.979	1.017	0.987	**0.965**	1.007	0.981
$h=2$	1.029	1.021	1.031	0.991	1.000	0.995	**0.985**	0.988	0.994
$h=4$	1.017	0.992	1.023	0.989	0.999	0.975	0.989	1.000	**0.967**
					LTN				
$h=1$	0.980	1.004	1.053	0.919	0.907	0.977	**0.868**	0.883	0.966
$h=2$	0.963	1.046	1.055	0.934	0.950	0.944	**0.921**	0.936	0.926
$h=4$	0.987	1.139	1.121	0.888	0.896	0.916	**0.881**	0.889	**0.881**

Note: $m_{stop}1$ is determined by using the corrected AIC, when m^{max} is set as 100. $m_{stop}2$ is determined by using the corrected AIC, when m^{max} is set as 200. The smallest MSFE ratio for each variable and forecasting horizon is in bold.

Table 4: List of variables with missing observations for some countries

Countries	Series	Periods with missing observations
Greece	EER	1970Q4-1994Q4
	STN	1970Q4-1994Q4
	LTN	1970Q4-1994Q4
Ireland	YER	1970Q1-1989Q4
	CPI	1970Q1-1975Q4
	YED	1970Q1-1989Q4
	EER	1970Q1-1989Q4
	STN	1970Q1-1989Q4
	LTN	1970Q1-1989Q4
Italy	STN	1970Q1-1970Q4
Spain	STN	1970Q1-1976Q4
Luxembourg	LTN	1970Q1-1993Q3

Figure 1: The optimal iteration number estimated by using the corrected AIC for 200 simulation runs for $T = 100$, 500, 1000 and 2000.



Note: $m_{stop}1$ is determined by using the corrected AIC, when m^{max} is set as 100. $m_{stop}2$ is determined by using the corrected AIC, when m^{max} is set as 200.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In 2nd International Symposium on Information Theory*, 267–281.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 71*(1), 135–172.

Bai, J. and S. Ng (2009). Boosting diffusion indices. *Journal of Applied Econometrics 24*, 607–629.

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science 22*, 477–505.

Bühlmann, P. and B. Yu (2003). Boosting with the l2 loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–339.

Brüggemann, R. and H. Lütkepohl (2013). Forecasting contemporaneous aggregates with stochastic aggregation weights. *International Journal of Forecasting 29*(1), 60–68.

Buchen, T. and K. Wohlrabe (2011). Forecasting with many predictors: Is boosting a viable alternative? *Economic Letters 113*, 16–18.

Carriero, A., G. Kapetanios, and M. Marcellino (2011). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics 26*(5), 735–761.

Chang, Y.-C. I., Y. Huang, and Y.-P. Huang (2010). Early stopping in L_2 boosting. *Computational Statistics & Data Analysis 54 (10)*, 2203–2213.

Fagan, G., J. Henry, and R. Mestre (2001). An area-wide model for the euro area. Working Papers 42, European Central Bank.

Fagan, G., J. Henry, and R. Mestre (2005). An area-wide model for the euro area. *Economic Modelling 22*, 39–59.

Freund, Y. and R. Schapire (1996). Experiments with a new boosting algorithm. *International Conference on Machine Learning*, 148–156.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 1189–1232.

Granger, C. W. J. and M. J. Morris (1976). Time series modelling and interpretation. *Journal of the Royal Statistical Society. Series A (General) 139 (2)*, 246–257.

Hastie, T. (2007). Comment: Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22 (4)*, 513–515.

Hendry, D. F. and K. Hubrich (2006). Forecasting economic aggregates by disaggregates. Working Papers 589, European Central Bank.

Hendry, D. F. and K. Hubrich (2011). Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of Business and Economic Statistics 29*(2), 216–227.

Hubrich, K. (2005). Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy? *International Journal of Forecasting 21*(1), 119–136.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike Information Criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) Vol. 60 (2)*, 271–293.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76 (2)*, 297–307.

Lütkepohl, H. (2010). Forecasting aggregated time series variables: A survey. *OECD Journal: Journal of Business Cycle Measurement and Analysis 2010/2*, 37–62.

Lütkepohl, H. (2011). Forecasting nonlinear aggregates and aggregates with time-varying weights. *Journal of Economics and Statistics 231*(1), 107–133.

Marcellino, M., J. H. Stock, and M. Watson (2003). Macroeconomic forecasting in the euro area: Country specific versus area-wide information. *European Economic Review 47*(1), 1–18.

Mayr, A., B. Hofner, and M. Schmid (2012). The importance of knowing when to stop. A sequential stopping rule for component-wise gradient boosting. *Methods of Information in Medicine 51*, 178–186.

Robinzonov, N., G. Tutz, and T. Hothorn (2012). Boosting techniques for nonlinear time series models. *Advances in Statistical Analysis 96*, 99–122.

Stock, J. H. and M. W. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Stock, J. H. and M. W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*, 147–162.

Stock, J. H. and M. W. Watson (2006). Forecasting with many predictors. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, pp. 515–554. North Holland, Amsterdam.

Taieb, S. and R. Hyndman (2014). Boosting multi-step autoregressive forecasts. *International Conference on Machine Learning (ICML) forthcoming.*

Tutz, G. and H. Binder (2007). Boosting ridge regression. *Computational Statistics & Data Analysis 51 (12)*, 6044–6059.

Wohlrabe, K. and T. Buchen (2014). Assessing the macroeconomic forecasting performance of boosting: Evidence for the united states, the euro area, and germany. *Journal of Forecasting 33 (4)*, 231–242.

University of Konstanz
Department of Economics

UNIVERSITY OF KONSTANZ
Department of Economics
Universitätsstraße 10 - 78464 Konstanz - Germany
Phone: +49 (0) 7531-88-0, Fax: +49 (0) 7531-88-3688

www.wiwi.uni-konstanz.de/econdoc/working-paper-series/