Open access • Journal Article • DOI:10.1002/FOR.2293

Assessing the Macroeconomic Forecasting Performance of Boosting: Evidence for the United States, the Euro Area and Germany — Source link

Klaus Wohlrabe, Teresa Buchen

Institutions: Ifo Institute for Economic Research

Published on: 01 Jul 2014 - Journal of Forecasting (John Wiley & Sons, Ltd)

Topics: Boosting (machine learning)

The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

Working Papers
www.cesifo.org/wp

Assessing the Macroeconomic Forecasting Performance of Boosting

Evidence for the United States, the Euro Area, and Germany

Teresa Buchen
Klaus Wohlrabe

CESIFO WORKING PAPER NO. 4148
CATEGORY 6: FISCAL POLICY, MACROECONOMICS AND GROWTH
MARCH 2013

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4148*

Assessing the Macroeconomic Forecasting Performance of Boosting

Evidence for the United States, the Euro Area, and Germany

Abstract

The use of large datasets for macroeconomic forecasting has received a great deal of interest recently. Boosting is one possible method of using high-dimensional data for this purpose. It is a stage-wise additive modelling procedure, which, in a linear specification, becomes a variable selection device that iteratively adds the predictors with the largest contribution to the fit. Using data for the United States, the euro area and Germany, we assess the performance of boosting when forecasting a wide range of macroeconomic variables. Moreover, we analyse to what extent its forecasting accuracy depends on the method used for determining its key regularisation parameter, the number of iterations. We find that boosting mostly outperforms the autoregressive benchmark, and that K-fold cross-validation works much better as stopping criterion than the commonly used information criteria.

JEL-Code: C320, C520, C530, E370.

Keywords: macroeconomic forecasting, component-wise boosting, large datasets, variable selection, model selection criteria.

Teresa Buchen
Ifo Institute – Leibniz-Institute
for Economic Research
at the University of Munich
Poschingerstraße 5
81679 Munich
Germany
buchen@ifo.de

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

1 Introduction

There has been a recent upswing of interest using large datasets for macroeconomic forecasting. An increasing number of time series describing the state of the economy are available that could be useful for forecasting. Also, computational power to handle an immense amount of data has steadily increased over time. Thus, researchers now attempt to improve their forecasting models by exploiting a broader information base.

Conventional econometric methods are not well suited to incorporating a large number of predictors: depending on the number of time-series observations, it is either impossible or inefficient to estimate the respective forecasting model. To overcome these problems without losing relevant information, new forecasting methods were developed. Eklund and Kapetanios (2008) classify the methods for forecasting a time series into three broad, partly overlapping, categories. The first group includes methods that use the whole dataset for forecasting, such as Bayesian regression and factor methods. The second group consists of forecast combination methods that use subsets of the data to produce multiple forecasts, which are then averaged. Component-wise boosting belongs to the third category. The latter assembles variable selection methods (LASSO and least angle regression are other examples) that also use subsets of the data, but produce only one forecast based on the optimal set of variables. More specifically, component-wise boosting is a stage-wise additive modelling procedure, that sequentially adds the predictor with the largest contribution to the fit without adjusting the previously entered coefficients.

Boosting has attracted much attention in machine learning and statistics because it can handle large datasets in a computationally efficient manner and because it has proven excellent prediction performance in a wide range of applications (Bühlmann and Hothorn, 2010). However, only recently has the method found its way into the macroeconometric literature. Apart from several financial applications (Audrino and Barone-Adesi, 2005; Gavrishchaka, 2006; Audrino and Trojani, 2007; Andrada-Félix and Fernández-Rodríguez, 2008), there are only few macroeconometric studies on the forecasting per-

formance of boosting (Shafik and Tutz, 2009; Bai and Ng, 2009; Hyun Hak and Swanson, 2011; Buchen and Wohlrabe, 2011). Results with respect to the predictive accuracy of boosting are promising. However, all these studies are confined to U.S. data and use only few target variables.[1]

We add to this literature by analysing the performance of boosting when forecasting a wide range of macroeconomic variables using three datasets for the United States, the euro area, and Germany. Moreover, we investigate to what extent the forecasting performance of boosting depends on the specification of the boosting algorithm concerning the stopping criterion for the number of iterations.

Careful choice of the stopping criterion of boosting is crucial since the number of iterations M is the key parameter regularising the tradeoff between bias and variance, on which the forecasting performance hinges. Small values of M yield a parsimonious model with a potentially large bias. The larger M becomes, the more one approaches a perfect fit, increasing the variance of the forecasting model. There are several methods for estimation the optimal number of iterations. The information criteria proposed by Bühlmann (2006) are wide-spread because they are computationally attractive,[2] but they tend to lead to overfitting (Hastie, 2007). Alternatively, resampling methods, such as K-fold cross-validation, can be applied. We evaluate whether the various stopping criteria result in relevant differences in the predictive performance of boosting when forecasting macroeconomic aggregates.

The remainder of this paper is organised as follows. Section 2 explains the boosting algorithm, especially how it handles the tradeoff between bias and variance. Section 3 sums up our empirical analysis. Section 4 concludes.

[1] An exception is Carriero, Kapetanios, and Marcellino (2011) who compare different methods that can be used in a VAR framework for forecasting the whole dataset consisting of 52 macroeconomic variables, including several reduced-rank models, factor models, Bayesian VAR models, and multivariate boosting. The latter is an extension of the standard boosting method developed by Lutz and Bühlmann (2006), where the predictors are selected according to a multivariate measure of fit. The results indicate that the forecasting performance of multivariate boosting is somewhat worse than that of the standard boosting approach.

[2] They are used, for instance, by Bai and Ng (2009), Shafik and Tutz (2009), and Hyun Hak and Swanson (2011).

2 The Boosting Algorithm

Boosting was originally designed as a classification scheme (Freund and Schapire, 1995, 1996) and later extended to regression problems (Friedman, Hastie, and Tibshirani, 2000; Friedman, 2001).[3] It is based on the machine learning idea, meaning that it is a computer programme that "learns from the data" (Hastie, Tibshirani, and Friedman, 2009). Instead of estimating a "true" model, as is traditionally done in statistics and econometrics, it starts with a simple model that is iteratively improved or "boosted" based on the performance with training data. As Bühlmann and Yu (2003) put it, "for large dataset problems with high-dimensional predictors, a good model for the problem is hard to come by, but a sensible procedure is not."

2.1 Forward Stage-wise Modelling

Boosting estimates a sequence of nested models, resulting in an additive model:

$$\hat{f}_M(\boldsymbol{x}_t) = \sum_{m=1}^{M} b(\boldsymbol{x}_t; \hat{\beta}_m),$$

where $m = 1, 2, ..., M$ denote the iteration steps and $b(\boldsymbol{x}_t; \hat{\beta}_m)$ are simple functions of the input vector $\boldsymbol{x}_t$, called *learner* in boosting terminology. The fitting method used to determine $b(\boldsymbol{x}_t; \hat{\beta}_m)$ is also part of the learner.

More specifically, boosting performs forward stage-wise modelling: it starts with the intercept and in each iteration m adds to the model the learner that most improves the fit, without modifying the parameters of those previously entered. The learners are selected according to a loss function $L(y_t, \hat{f}_m(\boldsymbol{x}_t))$, given the current model $\hat{f}_{m-1}(\boldsymbol{x}_t)$. Since in each iteration, only the parameters of the last learner need to be estimated, the algorithm is computationally feasible even for high-dimensional data. Generally, a forward stage-wise modelling procedure can be summarised as follows.

1. Initialise $\hat{f}_0(\boldsymbol{x}_t) = \bar{y}$.

[3] For an overview of boosting methods, see Bühlmann and Hothorn (2007a).

2. For $m = 1$ to M:

   (a) Compute

$$\hat{\beta}_m = \underset{\hat{\beta}}{\operatorname{argmin}} \sum_{t=1}^{T} L(y_t, \hat{f}_{m-1}(\boldsymbol{x}_t) + b(\boldsymbol{x}_t; \hat{\beta})).$$

   (b) Set

$$\hat{f}_m(\boldsymbol{x}_t) = \hat{f}_{m-1}(\boldsymbol{x}_t) + b(\boldsymbol{x}_t; \hat{\beta}_m).$$

2.2 Component-wise L_2-Boosting

Generally, boosting can accommodate all sorts of nonlinearities, but for high-dimensional datasets, it is advisable to engage in variable selection so as to reduce the complexity of the learner (Bühlmann and Yu, 2003). This can be achieved by estimating a (generalised) linear model. With so-called component-wise boosting, instead of a function of predictors, one variable is chosen and fitted in each step. In regression problems with the random variable $Y \in \mathbf{R}$, squared error loss (L_2-loss) is a common choice for the loss function,[4]

$$L(y_t, \hat{f}_m(\boldsymbol{x}_t)) = 1/2(y_t - \hat{f}_m(\boldsymbol{x}_t))^2.$$

With L_2-loss, the boosting algorithm repeatedly fits the learner to the current residuals u_t:

$$\begin{aligned} L(y_t, \hat{f}_m(\boldsymbol{x}_t)) &= L(y_t, \hat{f}_{m-1}(\boldsymbol{x}_t) + b(\boldsymbol{x}_t; \hat{\beta})) \\ &= 1/2(y_t - \hat{f}_{m-1}(\boldsymbol{x}_t) - b(\boldsymbol{x}_t; \hat{\beta}))^2 \\ &= 1/2(u_t - b(\boldsymbol{x}_t; \hat{\beta}))^2. \end{aligned}$$

Note that in a time-series context the predictor vector $\boldsymbol{x}_t$ contains p lags of the target variable y_t as well as p lags of the exogenous variables $z_{j,t}$, where

[4]The loss function is scaled by the factor $1/2$ in order to ensure a convenient representation of the first derivative.

$j = 1, ..., N$:

$$\boldsymbol{x}_t = (y_{t-1}, y_{t-2}, ..., y_{t-p}, z_{1,t-1}, z_{1,t-2}, ..., z_{1,t-p}, ..., z_{N,t-1}, z_{N,t-2}, ..., z_{N,t-p}).$$

Thus, component-wise boosting simultaneously selects variables and lags. From all potential predictor variables $x_{k,t}$, where $k = 1, ..., p(1+N)$, it selects in every iteration m one variable $x_{k_m^*,t}$—but not necessarily a different one for each iteration—which yields the smallest sum of squared residuals (SSR).

The algorithm for component-wise boosting with L_2-loss can be summarised as follows.

1. Initialise $\hat{f}_0(\boldsymbol{x}_t) = \bar{y}$.
2. For $m = 1$ to M:
   (a) Compute the residual $u_t = y_t - \hat{f}_{m-1}(\boldsymbol{x}_t)$.
   (b) For $k = 1, ..., p(1 + N)$, regress the residuals u_t on $x_{k,t}$ to obtain $\hat{\beta}_k$ and compute $\text{SSR}_k = \sum_{t=1}^{T}(u_t - x_{k,t}\hat{\beta}_k)^2$.
   (c) Choose $x_{k_m^*,t}$ such that $\text{SSR}_{k_m^*} = \min \text{SSR}_k$.
   (d) Update $\hat{f}_m(\boldsymbol{x}_t) = \hat{f}_{m-1}(\boldsymbol{x}_t) + \nu b(x_{k_m^*,t}; \hat{\beta}_{k_m^*})$, where $0 < \nu < 1$.

The parameter ν was introduced by Friedman (2001) who showed that the prediction performance of boosting is improved when the learner is shrunk toward zero. The final function estimate is then the sum of the M learners multiplied by the shrinkage parameter ν:

$$\hat{f}_M(\boldsymbol{x}_t) = \bar{y} + \sum_{m=1}^{M} \nu b(x_{k_m^*,t}; \hat{\beta}_{k_m^*}).$$

2.3 Controlling the Bias-Variance Tradeoff

Both the number of iterations M and the shrinkage parameter ν regulate the tradeoff between bias and variance that arises when fitting a model and that influences its forecasting performance. Suppose the data arise from the true but unknown model $Y = f(\boldsymbol{X} + \varepsilon)$, where Y is a random target variable

and $\boldsymbol{X}$ is the vector of random predictors. Under the assumption that the error has $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$ we can derive the expected forecast error $\mathrm{Err}(\boldsymbol{x_t})$ at an arbitrary predictor vector $\boldsymbol{x}_t$ of a forecasting model $\hat{f}(\boldsymbol{x}_t)$, using squared error loss:

$$\begin{aligned}
\mathrm{Err}(\boldsymbol{x}_t) &= \mathrm{E}[(Y - \hat{f}(\boldsymbol{x}_t))^2 | \boldsymbol{X} = \boldsymbol{x}_t] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}[\hat{f}(\boldsymbol{x}_t)] - f(\boldsymbol{x}_t)]^2 + \mathrm{E}[\hat{f}(\boldsymbol{x}_t) - \mathrm{E}[\hat{f}(\boldsymbol{x}_t)]]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(\boldsymbol{x}_t)) + \mathrm{Var}(\hat{f}(\boldsymbol{x}_t)).
\end{aligned}$$

The first term of this decomposition of the expected forecast error is the noise, that is, the variance of the target series around its true mean $f(\boldsymbol{x_t}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}_t)$. It is irreducible, even if we knew the true model. The second term is the squared bias, the amount by which the average model estimate differs from the true mean. In contrast to a simple OLS regression, where you assume that the true model is known, thus $\mathrm{E}[\hat{f}(\boldsymbol{x}_t)] = f(\boldsymbol{x}_t)$, this term is not zero but depends on the model complexity. Typically, it will be larger if the model is not complex enough so that we omit important variables. The third term is the variance of the forecasting model, the expected squared deviation of $\hat{f}(\boldsymbol{x}_t)$ around its mean. This term increases with model complexity. If we fit the training data harder, the model will generalise less well to unseen data and the forecasting performance deteriorates. Thus, the model must be chosen such that bias and variance are balanced to minimise the expected forecast error (Hastie, Tibshirani, and Friedman, 2009).

One way of avoiding overfitting with boosting is to employ a *weak* learner, that is, one that involves few parameters and has low variance relative to bias (Bühlmann and Yu, 2003). This can be achieved, for instance, by shrinking the learner toward zero because doing so reduces its variance. The other way of controlling the bias-variance tradeoff is to restrict the number of boosting iterations. The shrinkage parameter ν and the number of iterations M are connected; the smaller ν, the more iterations are needed to achieve a given prediction error (Hastie, Tibshirani, and Friedman, 2009). Empirical work finds that the exact size of the shrinkage parameter is of minor importance, as

long as it is "sufficiently small", i.e., $0 < \nu \leq 0.1$ (Friedman, 2001; Bühlmann and Hothorn, 2007a; Hastie, Tibshirani, and Friedman, 2009). Thus, the optimal number of iterations M^* is the main regularisation parameter of boosting.

There are several ways of estimating M^*, the most prominent being resampling methods and information criteria. Resampling methods estimate the expected forecast error directly by running the boosting algorithm multiple times with datasets drawn randomly from the original dataset.[5] K-fold cross-validation, for instance, randomly allocates the data into K roughly equal-sized parts. For the kth part, the model is fit to the other $K-1$ parts and the forecast error with respect to the kth part is calculated. After repeating this for all K parts, the average forecast error yields the cross-validation estimate for the expected forecast error. With information criteria, on the other hand, the estimated forecast error is composed of two parts, one term capturing model fit and one term penalising model complexity, measured by the degrees of freedom.

For a linear model estimated by OLS, the degrees of freedom are simply the number of fitted parameters (Hastie, Tibshirani, and Friedman, 2009). For boosting, the degrees of freedom must be determined as a function of the number of iterations. With growing m, the complexity of the fitted procedure does not increase by constant amounts, but by exponentially decreasing amounts. This is largely due to the nature of forward stage-wise fitting; the learner that is added to the model each iteration depends on the performance of the current model (Bühlmann and Yu, 2003). In fact, there is no exact expression for the degrees of freedom of boosting (Bühlmann and Hothorn, 2007b). But Bühlmann (2006) develops an approximation for L_2-boosting, which defines the degrees of freedom of boosting in iteration m as the trace of the boosting hat matrix $\mathcal{B}_m$:

$$\mathrm{df}(m) = \mathrm{trace}(\mathcal{B}_m), \tag{1}$$

[5]To be valid for time-series data, the mboost package implemented in R uses a model-based approach, assuming i.i.d. residuals. The idea is to fit the model first, and to subsequently resample from the residuals. For details about how to construct the new samples from the residuals, see Efron and Tibshirani (1986).

where $\mathcal{B}_m$ is a projection matrix that yields the fitted function $\hat{f}_m(\boldsymbol{x}_t)$ when post-multiplied by the realisations y_t:

$$\hat{f}_m(\boldsymbol{x}_t) = \mathcal{B}_m y_t.$$

Bühlmann (2006) proposes to insert (1) into the corrected Akaike criterion:

$$\text{cAIC}(m) = \log(\hat{\sigma}^2) + \frac{1 + \text{df}(m)/T}{(1 - \text{df}(m) + 2)/T} \text{ , where}$$
$$\hat{\sigma}^2 = T^{-1} \sum_{t=1}^{T} (y_t - \hat{f}_m(\boldsymbol{x}_t))^2.$$

An alternative method is to use the gMDL criterion (Minimum Description Length criterion using a g-prior), which bridges the AIC and the BIC in a data-driven manner and adaptively selects the better among the two (Bühlmann and Hothorn, 2007a):[6]

$$g\text{MDL}(m) = \log(S) + \frac{\text{df}(m)}{T} \log(F), \text{ where}$$
$$S = \frac{T\hat{\sigma}^2}{T - \text{df}(m)}, \; F = \frac{\sum_{t=1}^{T} y_t^2 - T\hat{\sigma}^2}{\text{df}(m) S}.$$

Finally, the estimate for the optimal number of boosting iterations is given by:

$$\hat{M}^* = \underset{1 \leq m \leq M^{max}}{\text{argmin}} \; \text{IC}(m),$$

where M^{max} is a large upper bound for the candidate number of boosting iterations and IC is one of the information criteria.

These information criteria are computationally attractive, but they tend to lead to overfitting. Hastie (2007) shows that the trace of the boosting hat matrix is only a poor approximation since it treats the model at stage m as if it was computed by a predetermined sequence of linear updates.[7]

[6] For details, see Hansen and Yu (2001).

[7] In that case, Equation (1) would be an exact measure of the degrees of freedom (Hastie, Tibshirani, and Friedman, 2009).

But the sequence of updates is adaptively chosen, and the cost of searching for the variable with the best fit is ignored. Hence, the penalty term of the information criteria tends to be too small, resulting in the procedure being stopped too late. As an alternative, Hastie (2007) suggests approximating the degrees of freedom of boosting by the size of the active set, that is, the number of selected variables until iteration m, or using K-fold cross-validation to estimate the expected forecast error. In the following empirical application, we evaluate whether the various methods of determining the stopping criterion result in relevant differences in the predictive accuracy of boosting when forecasting macroeconomic aggregates.

3 Empirical Analysis

3.1 Data

For our empirical analysis, we use three large-scale datasets with monthly frequency—one each for the United States, the euro area, and Germany. All three datasets reflect various aspects of the respective economy and contain information typically taken into consideration by central banks. The variables can be grouped into the following categories: real economy (such as industrial production, orders, labour market, and housing market), money and prices (such as monetary aggregates, wages, consumer prices, producer prices, and commodity prices), financial markets (such as exchange rates, interest rates, term spreads, and stock indices), and surveys. The datasets vary in size (both with respect to T and N), but all three cover the recent economic crisis.

For the United States, we use an updated version of the dataset employed by Giannone, Reichlin, and Sala (2004) containing 168 time series from 01/1970 to 12/2010.[8] For Germany, we use the dataset by Drech-

[8]For a full list of the series, see Giannone, Reichlin, and Sala (2004). Three series were not available (series 104, 126 and 132) and two series are quarterly (series 172 and 173), so they are excluded. We used the monthly analogues of the authors' stationarity transformations, i.e., transformation 2 is the monthly difference, transformation 3 is the monthly annualised growth rate and transformation 4 is the yearly growth rate in the

sel and Scheufele (2012), which contains 217 time series from 01/1992 to 05/2011. In addition to the categories mentioned above, the German dataset also contains information on governmental indicators (such as tax revenue and customs duties) as well as a range of international indicators (such as survey indicators or share indices of export partners).[9] The smallest dataset is the one for the euro area. It contains 78 time series from 02/1994 to 10/2010, which are listed with the respective stationarity transformation in the Appendix.

3.2 Forecasting Approach

The component-wise boosting procedure applied in this study uses ordinary least squares as learner and a squared error loss function to estimate an autoregressive distributed lag (ADL) model:[10]

$$y_{t+h} = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_t + \varepsilon_t = \alpha + \sum_{i=1}^{12} \gamma_i y_{t-i} + \sum_{j=1}^{N} \sum_{i=1}^{12} \delta_{ji} z_{j,t-i} + \varepsilon_t.$$

The variables and lags not selected have a zero coefficient. To save computational time, the size of the shrinkage parameter ν is set to 0.1, the upper bound of the interval suggested by the literature (Bühlmann and Hothorn, 2007a; Hastie, Tibshirani, and Friedman, 2009). The optimal number of iterations M^* is estimated with several stopping criteria: the corrected AIC and the gMDL criterion as information criteria—both with the trace of the boosting hat matrix and the size of the active set as measures for the degrees of freedom—and 10-fold cross-validation as a resampling method. All results

respective month.

[9]For a list of the series and the stationarity transformations, see Drechsel and Scheufele (2012). To ensure that all series have the same length, we discarded the following variables. Real economic indicators: WTEXMOG, WHTCFWH, WHTCHEH, WHTCNMH, WHTSLGH, USLA01B, RVN, RETTOTG, EMPTOTO, EMPOWHH. Finance: SPR-NF2AE, SPR-NF3BE, SPR-P3BE, SPR-EUCU, VDAXNEW, VDAXIDX, MLNF2AE, MLNF3BE, MLNP3BE, MLHEUCU, TSD304B. Survey indicators: IFOMTLQ, IFOMTKQ, IFOMTAQ, IFOMCAQ, IFOMCLQ, IFOMCKQ, IFOBDOQ, IFOBDQQ, IFOBDPQ, IFOWHIQ, IFOWHAQ, IFORTIQ, IFORTHQ, CONSNT, EUSVCIQ, PMIBD, PMIBDS, PMIEUR. International indicators: POEUSESIG, CZEUSESIG, CHOL0955R.

[10]For estimation, we employed the mboost package implemented in R.

are compared for the case when the maximal number of iterations is set to $M^{max} = 50$ and 100.

We produce forecasts for the horizons $h = 1, 3, 6$, and 12 months. All forecasts are computed directly and pseudo-out-of-sample using a rolling estimation window. The forecast period starts in 01/1990 for the United States, and in 01/2000 for the euro area and Germany. Since our aim is to arrive at a broad picture of the predictive performance of boosting in a macroeconomic context, we forecast all the variables in the datasets. The specific form of the target variable depends on its stationarity transformation and can be either the (log) level in the respective month, the monthly first difference, the monthly first (second) log difference, or the yearly log difference. Due to computational considerations, all variables were centered for the respective estimation window. We assess the forecasting accuracy of boosting relative to the standard autoregressive model, where the lag length is determined by the Bayesian information criterion (BIC).

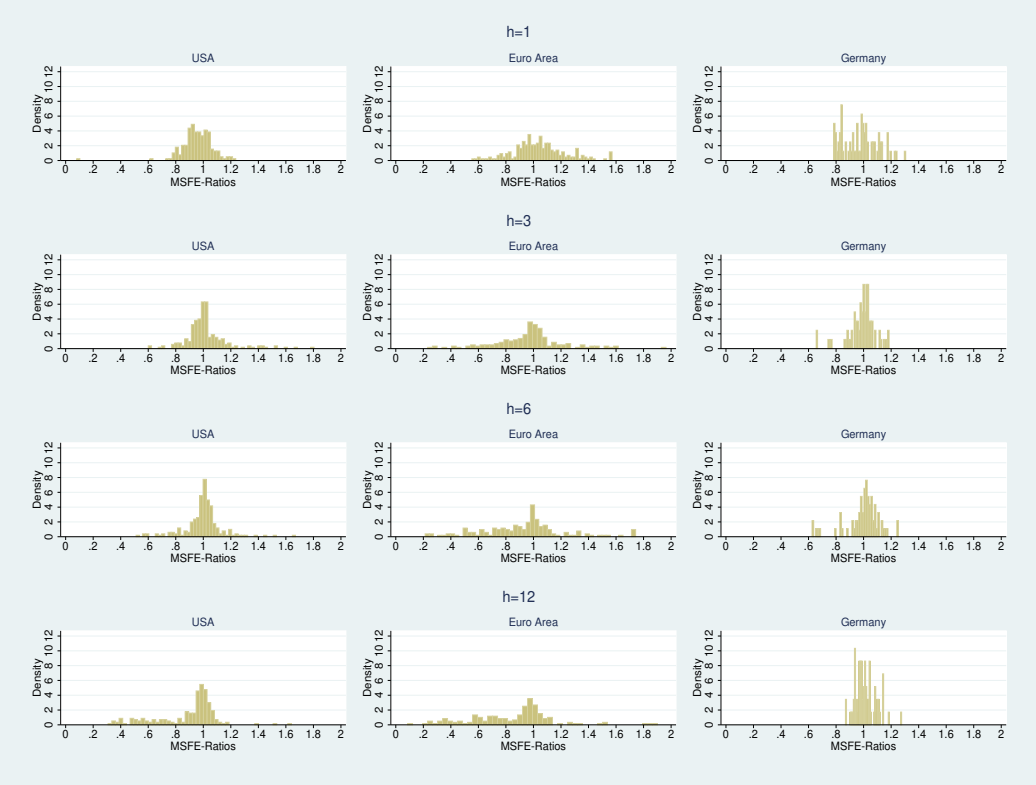
To summarise the overall forecasting accuracy, we employ a multivariate version of the mean squared forecast error (MSFE) as proposed by Christoffersen and Diebold (1998). The multivariate MSFE is given by $\text{MSFE} = \text{E}(\boldsymbol{e}'_{t+h}\boldsymbol{W}\boldsymbol{e}_{t+h})$, where $\boldsymbol{e}_{t+h}$ is the vector of the h-step-ahead forecast errors and $\boldsymbol{W}$ is an $N \times N$ weighting matrix with N being the number of target variables. In accordance with Carriero, Kapetanios, and Marcellino (2011), we choose a diagonal matrix $\boldsymbol{W}$ with the elements of the diagonal being the inverse of the variances of the target series. Consequently, a series that has large variance—and is thus less predictable—is given less weight.

3.3 Results

To give a first impression of the forecasting performance of boosting, we plot the distribution of the mean squared forecast errors across all target variables. It is shown for all three datasets and for different forecast horizons in Figure 1, where we use $M^{max} = 50$ and K-fold cross-validation as stopping criterion. The largest spikes tend to be close to one. But in most cases, more than 50% of the MSFE ratios are below one, that is, boosting performs better than the

AR benchmark. For the United States and Germany, most of the ratios are concentrated between roughly 0.8 and 1.2, while for the euro area the spread is somewhat wider.[11]

Figure 1: Histograms of MSFE-Ratios



Notes: This figure displays the frequency distribution of the MSFE ratios of boosting using $M^{max} = 50$ and K-fold cross-validation as stopping criterion, relative to the AR benchmark model for all target variables. Outliers larger than 2 are excluded.

The multivariate MSFE ratios in Table 1 summarise the forecasting accuracy of boosting across all variables in the datasets while taking into account the predictability of the target series. These aggregate results give us insights in how boosting generally performs when forecasting any kind of macroeconomic time series. First of all, it is confirmed that boosting beats the AR benchmark on average. Many of the multivariate ratios are close to one, but boosting can lead to improvements relative to the benchmark of up to 47%. Second, its relative forecast accuracy tends to improve with increasing

[11]Due to graphical reasons, outliers larger than 2 are excluded.

forecasting horizon, although for the U.S. data, boosting also performs very well at a forecast horizon of one month. Third, the number of boosting iterations leading to the smallest MSFEs seems to be quite small, at most 50. The different stopping criteria are not completely robust to the choice of the maximum number of iterations. Instead, the estimated optimal number of iterations tends to rise with larger M^{max}. However, the differences vary across criteria and are largest when the degrees of freedom to be employed in the computation of the information criteria is approximated by the trace of the boosting hat matrix (cAIC Trace or gMDL Trace). Estimating the degrees of freedom by the size of the active set (cAIC Actset or gMDL Actset) delivers better and more robust results. But finally, using 10-fold cross-validation (CV) appears to be the dominant stopping criterion. Not only does it yield the smallest multivariate MSFE ratio for a given forecast horizon (entries in bold) in most of the cases, but it is also very robust to the choice of the candidate number of iterations.

Figure 2 gives more insights into the functioning of the different stopping criteria. It compares the multivariate MSFE ratios as well as the average across variables and time of the estimated optimal number of iterations M^* and of the number of variables that enter the models. Basically, we use the same forecasting approach as described in Section 3.2, but we compare the results for a wider range of choices regarding the maximum number of iterations ($M^{max} = 10, 20, ..., 500$). Due to computational reasons, we only compute the forecasts for the last 120 months, which leads to a smaller evaluation sample especially for the U.S. dataset. Here, we display exemplarly the results for a forecast horizon of one month.

Table 1: Multivariate MSFE Ratios

		USA				Euro Area				Germany			
		h=1	h=3	h=6	h=12	h=1	h=3	h=6	h=12	h=1	h=3	h=6	h=12
Criterion	M^{max}	Ratio Relative to the AR(p) Benchmark											
cAIC Trace	50	0.842	0.995	0.951	0.899	1.051	0.997	0.836	0.559	0.949	0.939	0.902	0.835
cAIC Actset	50	0.843	0.993	0.949	0.896	**1.025**	0.946	0.783	0.529	**0.943**	0.922	0.882	0.808
gMDL Trace	50	0.842	0.995	0.951	0.899	1.051	0.997	0.836	0.559	0.949	0.939	0.902	0.835
gMDL Actset	50	0.844	0.995	0.951	0.899	1.041	0.982	0.821	0.546	0.950	0.937	0.896	0.820
K-fold CV	50	**0.779**	0.981	**0.936**	**0.884**	**1.025**	**0.943**	**0.779**	**0.528**	0.946	**0.913**	**0.875**	**0.805**
cAIC Trace	100	0.883	0.985	1.020	0.969	1.112	1.074	0.902	0.591	0.981	0.977	0.942	0.879
cAIC Actset	100	0.879	0.973	1.004	0.954	1.031	0.953	0.789	0.530	0.946	0.925	0.886	0.811
gMDL Trace	100	0.883	0.985	1.020	0.969	1.112	1.074	0.901	0.591	0.981	0.977	0.942	0.879
gMDL Actset	100	0.881	0.982	1.016	0.965	1.073	1.021	0.854	0.556	0.967	0.959	0.918	0.839
K-fold CV	100	0.784	**0.956**	0.989	0.940	1.039	0.964	0.797	0.538	0.955	0.928	.886	0.820

Notes: This table reports multivariate MSFE ratios for various boosting methods relative to the AR benchmark, where the boosting methods differ with respect to the stopping criterion: the corrected Akaike information criterion (cAIC) and the Minimum Description Length criterion using a g-prior (gMDL), both when the trace of the boosting hat matrix (Trace) and the size of the active set (Actset) is used to approximate the degrees of freedom of the respective model, and K-fold cross-validation (CV). A value smaller than one indicates that boosting delivers on average a smaller MSFE, taking into account the predictability of the series. Entries in bold indicate the best MSFE ratio for a given forecast horizon.

The figure confirms what was already indicated by Table 1; on average, the MSFE ratios are smallest when the number of iterations is highly restricted. With rising M^{max}, the forecasting performance of boosting deteriorates. But this deterioration is most pronounced for the trace criteria, and much less important for K-fold cross-validation and for the gMDL criterion when using the size of the active set to estimate model complexity. Overall, K-fold cross-validation yields the best results. While the cAIC actset stopping criterion often leads to smaller forecast errors at lower numbers of iterations allowed for, they rise strongly at larger values of M^{max} (see panels in first column).

The reason for this differing forecasting accuracy can be seen from the panels in the second column of Figure 2. They display the optimal number of iterations M^* as estimated by the various stopping criteria as a function of M^{max}. When using the trace information criteria, the chosen number of iterations tends to go to the limit allowed for.[12] Thus, these criteria indeed seem to overfit and lead to very large models with many variables (see panels in third column).[13] On the contrary, the gMDL trace criterion and K-fold cross-validation are much less sensitive with respect to the choice of M^{max}.

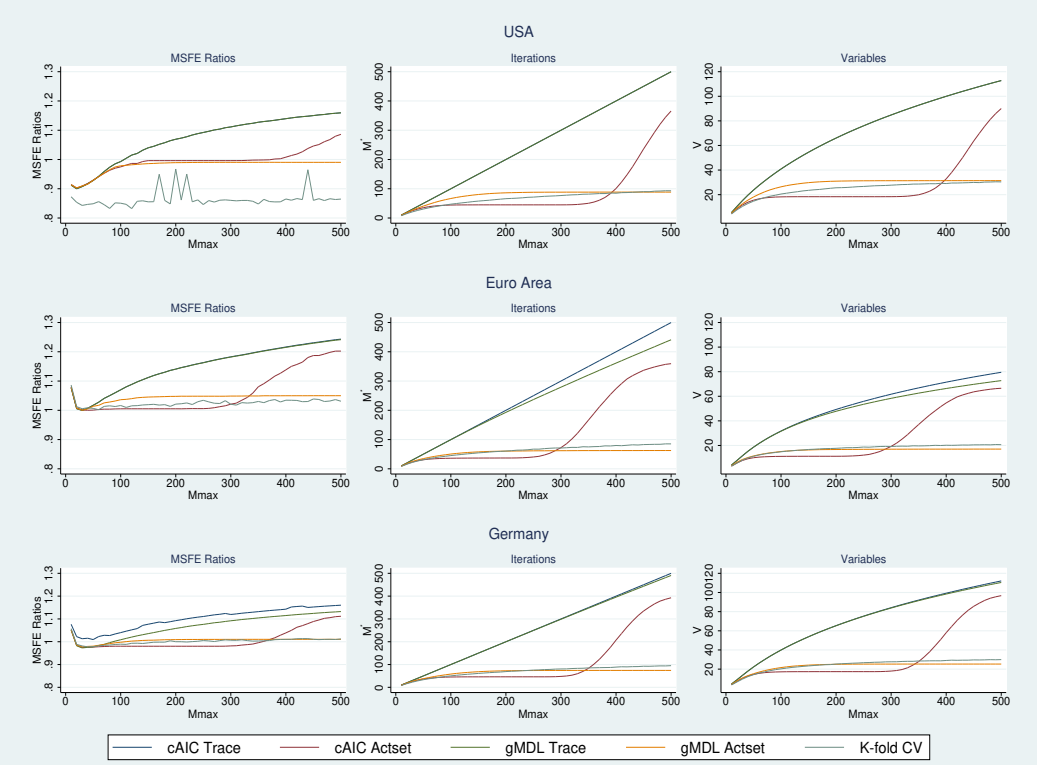
4 Conclusion

Component-wise boosting is a variable selection method that can be used in a data-rich environment. It starts with a simple model that is iteratively updated by adding the predictor with the largest contribution to the fit. We assess whether the predictive qualities of boosting that have been reported in many other areas can be confirmed when forecasting a wide range of macroeconomic variables. To that aim, we use large-scale datasets for the United States, the euro area, and Germany. Moreover, we analyse to what extent the forecasting accuracy of the boosting algorithm depends on the

[12] In that case, M^{max} should actually be set higher.

[13] Keep in mind that the contribution of each variable is shrunk and that each variable can be chosen several times. So while some of the variables are fitted completely (that is, with a shrinkage factor of $\nu = 0.1$, they are selected 10 times), some are only chosen once or twice.

Figure 2: Comparison across Stopping Criteria



Notes: This figure compares across various stopping criteria the multivariate MSFE ratios as well as the average (across variables and time) estimated optimal number of iterations M^* and the average number of variables that enter the models for different choices of the maximum number of iterations M^{max}, where $M^{max} = 10, 20, ..., 500$. The stopping criteria are the following: the corrected Akaike information criterion (cAIC) and the Minimum Description Length criterion using a g-prior (gMDL), both when the trace of the boosting hat matrix (Trace) and the size of the active set (Actset) is used to approximate the degrees of freedom of the respective model, as well as K-fold cross-validation (CV). The forecasts are computed for the last 120 months of the respective dataset and we only display the results for a forecast horizon of one month.

method chosen to determine its key regularisation parameter, the number of iterations.

Indeed, we find that boosting performs well in macroeconomic forecasting; it outperforms the benchmark in most cases. Furthermore, the choice of the stopping criterion determining the number of iterations has an important influence on the forecasting performance of boosting. We compare information criteria based on the trace of the boosting hat matrix as a measure of model complexity, which were proposed by Bühlmann (2006) and are widely used, with information criteria based on the size of the active set as well as with K-fold cross-validation as an example of a resampling method. Our results confirm the critique by Hastie (2007) and suggest that the trace criteria indeed underestimate model complexity. Thus, the boosting procedure is stopped too late and overfits, which is reflected in larger forecasting errors. Using the number of selected variables as a measure of model complexity, as proposed by Hastie (2007) abates the problem. But overall, K-fold cross-validation is the dominant stopping criterion.

To conclude, component-wise boosting is a powerful method that can be used to forecast a wide range of macroeconomic variables. However, to achieve the best possible results, it is important to choose the model selection criterion carefully and not to adopt it by mere convention.

References

Andrada-Félix, J., and F. Fernández-Rodríguez (2008): "Improving moving average trading rules with boosting and statistical learning methods," *Journal of Forecasting*, 27, 433–449.

Audrino, F., and G. Barone-Adesi (2005): "Functional gradient descent for financial time series with an application to the measurement of market risk," *Journal of Banking & Finance*, 29, 959–977.

Audrino, F., and F. Trojani (2007): "Accurate short-term yield curve forecasting using functional gradient descent," *Journal of Financial Econometrics*, 5, 591–623.

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

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

Bühlmann, P. (2006): "Boosting for high-dimensional linear models," *Annals of Statistics*, 34, 559–583.

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

——— (2007b): "Rejoinder: Boosting algorithms: Regularization, prediction, and model fitting," *Statistical Science*, 22, 516–522.

——— (2010): "Twin Boosting: Improved feature selection and prediction," *Statistics and Computing*, 20, 119–138.

Bühlmann, P., and B. Yu (2003): "Boosting with the L2 loss: Regression and classification," *Journal of the American Statistical Association*, 98, 324–339.

Carriero, A., G. Kapetanios, and M. Marcellino (2011): "Forecasting large datasets with Bayesian reduced rank multivariate models," *Journal of Applied Econometrics*, 26, 715–734.

Christoffersen, P., and F. Diebold (1998): "Cointegration and long-run forecasting," *Journal of Business and Economic Statistics*, 16, 450–458.

Drechsel, K., and R. Scheufele (2012): "Bottom-up or direct? Forecasting German GDP in a data-rich environment," Swiss National Bank Working Papers 2012-16.

Efron, B., and R. Tibshirani (1986): "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical Science*, 1, 54–75.

Eklund, J., and G. Kapetanios (2008): "A review of forecasting techniques for large data sets," *National Institute Economic Review*, 203, 109–115.

Freund, Y., and R. Schapire (1995): "A decision-theoretic generalization of on-line learning and an application to boosting," in *Computational Learning Theory*, pp. 23–37. Berlin/New York: Springer.

Freund, Y., and R. Schapire (1996): "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148–156. Citeseer.

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

Friedman, J., T. Hastie, and R. Tibshirani (2000): "Additive logistic regression: A statistical view of boosting," *Annals of Statistics*, 28, 337–407.

Gavrishchaka, V. (2006): "Boosting-based frameworks in financial modeling: Application to symbolic volatility forecasting," *Advances in Econometrics*, 20, 122–151.

Giannone, D., L. Reichlin, and L. Sala (2004): "Monetary policy in real time," *NBER Macroeconomics Annual*, 19, 161–200.

Hansen, M., and B. Yu (2001): "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, 96, 746–774.

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

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The elements of statistical learning: Data mining, inference, and prediction*. Berlin: Springer.

Hyun Hak, K., and N. Swanson (2011): "Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence," Working Paper 2011, 19, Department of Economics, Rutgers University.

Lutz, R., and P. Bühlmann (2006): "Boosting for high-multivariate responses in high-dimensional linear regression," *Statistica Sinica*, 16, 471–494.

Shafik, N., and G. Tutz (2009): "Boosting nonlinear additive autoregressive time series," *Computational Statistics and Data Analysis*, 53, 2453–2464.

Appendix

Table 2: List of Variables, Euro Area

Series	Transformation
REAL ECONOMY	
Eurostat, manufacturing, production, total	3
Eurostat. manuf., prod., textiles	5
Eurostat, manuf., prod., food, beverages and tobacco products	3
Eurostat, manuf., prod., motor vehicles, trailers, semi-trailers and other transport	3
Eurostat, manuf., prod., machinery and equipment	3
Eurostat, manuf., prod., basic pharmaceutical products and pharmaceutical preparations	3
Eurostat, manuf., prod., coke and refined petroleum products	3
Eurostat, manuf., prod., chemicals and chemical products	3
Eurostat, manuf., prod., basic metals	3
Eurostat, manuf., prod., rubber and plastic products	3
Eurostat, manuf., prod., intermediate goods	3
Eurostat, manuf., prod., consumer goods	3
Eurostat, unemployment	3
Eurostat, unemployment rate	2
Eurostat, manufacturing, order books	2
OECD, manufacturing, export order books or demand	2
OECD, retail trade volume	3
MONEY AND PRICES	
ECB, M1	4
ECB, M2	4
ECB, M3	4
HWWI, total index, average	4
HWWI, agricultural raw materials index, average	4
HWWI, crude oil index, average	4
HWWI, industrial raw materials index, average	4
HWWI, energy raw materials index, average	4
ECB, consumer prices, index	4
ECB, consumer prices excluding energy and unprocessed food	6
Eurostat, domestic producer prices, manufacturing	4
Eurostat, dom. prod. prices, energy	4
Eurostat, dom. prod. prices, food products and beverages	4

Table 2: List of Variables, Euro Area

Series	Transformation
Eurostat, dom. prod. prices, tobacco products	4
Eurostat, dom. prod. prices, chemicals and chemical products	4
Eurostat, dom. prod. prices, motor vehicles, trailers and semi-trailers	4
Eurostat, dom. prod. prices, intermediate goods	4
Eurostat, dom. prod. prices, capital goods	4
Eurostat, dom. prod. prices, durable consumer goods	4
Eurostat, dom. prod. prices, non-durable consumer goods	4
FINANCIAL MARKETS	
OECD, real effective exchange rate, EUR	3
OECD, EUR/US$ exchange rate, monthly average	3
Eurostat, interbank rates, 3 month, yield, average	2
ECB, government benchmarks, bid, 2 year, yield, average	2
ECB, government benchmarks, bid, 3 year, yield, average	2
ECB, government benchmarks, bid, 5 year, yield, average	2
ECB, government benchmarks, bid, 7 year, yield, average	2
ECB, government benchmarks, bid, 10 year, yield, average	2
ECB, term spread, government benchmarks, 5-3 years	1
ECB, term spread, government benchmarks, 7-3 years	1
ECB, term spread, government benchmarks, 10-3 years	1
STOXX Limited, STOXX, broad index, end of month	3
STOXX Limited, STOXX 50, end of month	3
SURVEYS	
CEPR, EuroCOIN, industry sector	1
DG ECFIN, economic sentiment indicator	1
DG ECFIN, manufacturing, industrial confidence indicator	1
DG ECFIN, construction confidence indicator	1
DG ECFIN, retail trade confidence indicator	1
DG ECFIN, manufacturing, export order books	1
DG ECFIN, manufacturing, order books	1
DG ECFIN, construction, order books	1
DG ECFIN, retail trade, employment expectations	1
DG ECFIN, construction, employment expectations	1
DG ECFIN, manufacturing, employment expectations	1
DG ECFIN, services, expectation of demand over next 3 months	1
DG ECFIN, manufacturing, production expectations	1
DG ECFIN, manufacturing, selling-price expectations	1

Table 2: List of Variables, Euro Area

Series	**Transformation**
DG ECFIN, consumer surveys, consumer confidence indicator	1
DG ECFIN, cons. surv., general economic situation over next 12 months	1
DG ECFIN, cons. surv., unemployment expectations over next 12 months	1
DG ECFIN, cons. surv., price trends over next 12 months	1
DG ECFIN, cons. surv., financial situation of households over next 12 months	1
DG ECFIN, cons. surv., major purchases at present	1
DG ECFIN, cons. surv., major purchases over next 12 months	1
DG ECFIN, cons. surv., savings at present	2
DG ECFIN, cons. surv., savings over next 12 months	1
OECD, total leading indicator, quantum, normalised	1
OECD, total leading indicator, trend restored	2
OECD, total leading indicator, amplitude adjusted	1

Transformation - 1: x_t, 2: $x_t - x_{t-1}$, 3: $\ln(x_t/x_{t-1})$, 4: $\ln(x_t/x_{t-1}) - \ln(x_{t-1}/x_{t-2})$.