Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis BSc Econometrics and Operations Research

Enhancing the Forest: A Study on Variable Selection Methods for the Targeted Random Forest

David Chen (541762)

Abstract

Random Forests are widely applied in many fields of research for their high performance in high dimensional settings and their ability to detect non-linearities and interactions. In this paper, different variable selection techniques are considered to target a specific number of variables, prior to performing Random Forest. We find the largest gains from targeting for random forests for one-year ahead forecasts, of which the adaptive LASSO method generally returns the best results. The obtained forecasts are evaluated by means of mean squared error and various tests of equal predictive ability, such as the Diebold-Mariano test, the fluctuation test and model confidence sets. Finally, we find drastic improvements in forecasting performance from targeting during highly volatile periods, such as the Covid-19 financial crisis.

Supervisor:	EP O'Neill
Second assessor:	A Pick
Date:	27th June 2023

1 Introduction

Advancements in machine learning have been at large in recent years, with application in many fields, such as medicine, economics and finance. Random forests (RF) by Breiman (2001) have been one of the models at the forefront of this, showing to be one of the most successful general purpose regression models Biau and Scornet (2016). Borup et al. (2023) attempt to improve on the RF model by first performing variable selection to reduce dimensionality, and then applying RF to the selected variables. They provide a theoretical perspective on the advantages of targeting variables for RF and find a tree strength-correlation trade-off. Even though stronger trees are always grown with the inclusion of targeting, the tress become more highly correlated as well.

Their methodology is applied to a number of datasets to predict equity premium of the S&P500 stock index and macroeconomic variables, such as industrial production growth (IP), employment growth (EMP) and consumer price index acceleration (CPI). For variable selection, they use LASSO to subset the variables used for random forest, otherwise referred to as targeted RF (TRF). This is done for different degrees of targeting with h = 1, 3 and 12 months ahead forecasts and considered over an expanding window with forecasts for 1985 to 2021, as well as additional forecasts for subsamples corresponding to NBER recessions and expansions. Borup et al. (2023) find that, in settings with high signal-to-noise ratio (SNR), targeting improves the RF forecast more than in settings with low SNR. Note that LASSO performs well in high SNR settings (Fan and Li, 2001), meaning that their positive results could be partly attributed to their choice of variable selection method.

Borup et al. (2023) unfortunately do not go into further exploration of other potential variable selection methods. That is why in this research, the focus lies on investigating different variable selection methods to determine its effects on TRF and after that, to test for which method leads to the best predictions. In this way, the paper extends on the original paper of Borup et al. (2023) by contributing to the research of optimal variable selection method. The main research question is formulated as: *'What is the best targeting method for variable selection that improves the predictability of industrial production growth using Targeted Random Forests?'*. Furthermore, answers are sought after as to what the gains are of different variable selection methods on the forecasting performance of TRF and whether there are viable alternatives to the TRF, such as the targeted gradient boosting machine (TGB) and reverse TRF, where RF is deployed for targeting and the usual targeting method is used for forecasting. As well as, its performance in highly volatile environments, such as the Covid-19 financial crisis.

This paper mainly contributes to existing literature in machine learning and macroeconomic forecasting. Additionally, this research makes contributions to existing literature in regression models and forecasting with its high focus on variable selection. Lastly, the methods used in this research are applicable to other cases of high dimensional data, such as volatility forecasting (Luong and Dokuchaev, 2018) and Bitcoin price direction forecasting (Basher and Sadorsky, 2022), where random forest has been shown to excel.

Prior research on targeting variables for RF include the work of Medeiros et al. (2021) who explore the use of adaptive LASSO to select variables for their RF, as well as variable selection with RF for OLS estimation. They conclude that the good performance of RF is driven by variable selection, as well as non-linearity. Borup and Schütte (2022) use Elastic Net to target predictors for employment growth predictions using RF. Next, Charles and Darné (2022) also perform variable selection prior to RF to create sparse RF models. They find significant forecasting improvements over the full RF model, when Elastic Net and Distance Correlation Sure Independence Screening are used for variable selection. On the other hand, Mentch and Zhou (2022) state that one should be wary of variable targeting prior to using random forest, as they find that adding irrelevant variables in the form of randomly generated white noise variables vastly improves the performance of random forest, suggesting that irrelevant variables should not be removed by means of targeting.

2 Methodology

This section introduces the methodology applied to obtain the results of this research. First, the general procedure of the targeted random forest is presented, along with the forecasting framework, general notation and variants of the TRF. Then, the methodology used to replicate the results in Borup et al. (2023) is specified. Finally, the methods used to extend on the findings of Borup et al. (2023) are described in detail.

2.1 Targeted Random Forest

The targeted random forest proposed in Borup et al. (2023) is essentially a 2-step Random Forest regression, where a set of variables is selected using a variable selection method in the first targeting step, according to some importance measure. Borup et al. (2023) use the Least Absolute Shrinkage and Selection Operator (LASSO) as in Tibshirani (1996), however any method that can form a ranking of variable importance can be used, such as adaptive LASSO, random forest and Ordinary Least Squares. Afterwards, RF regression is done using the smaller subset of variables. In general, the TRF can be denoted in the form V_s/M, where V_s denotes the variable selection method V, targeting s amount of variables, and M denotes the model used for forecasting using the targeted set of variables. For example, using LASSO at the first step, followed by RF would be denoted as LASSO/RF. Forecasts are made using the TRF models over an h-step ahead forecasting horizon. In the case of this research, horizons of size h = 1, 3 and 12 months ahead are considered, along with an expanding window framework for the replication as in Borup et al. (2023) and a rolling window for the extension results[1]. The initial window size is set to 180 observations, which equates to the first 15 years of the sample. Lastly, variants of the TRF are also considered as an extension on the paper of Borup et al. (2023). One will be referred to as reverse TRF, which is of the form RF_s/M, where RF is used for the initial targeting step and the models that would typically be used for variable selection are used for forecasting instead. The other will be referred to as Targeted Gradient Boosting Machine (TGB), which follows the same procedure as a TRF, but using a gradient boosting machine (GBM) algorithm (Friedman, 2001) instead for the forecasting step.

[1] A rolling window is required for the implementation of the additional forecast evaluation tests, model confidence sets and fluctuation test, further discussed in Sections 2.4.6 and 2.4.7

2.2 Notation

First, denote the following:

- $T = T_{IS} + T_{OS}$, the time horizon T, where T_{IS} denotes the intital in-sample time period used for model estimation and T_{OS} denotes the-initial out-of-sample time period used for forecasting.
- y_t: Value of the observed dependant variable at time t = 1,..., T.
- $x_{k,t}$: Denotes the variable $x_{k,t}$ for k = 1,...,K at time t = 1,..., T. Also referred to as predictor, regressor and feature.
- β_k: Denotes the effect size of the predictor x_k on the dependent variable y.
- s: Denotes the amount of variables targeted in the targeting step to obtain the subset of variables used for TRF. It follows that $s \leq k$.

2.3 Methodology - Replication

In this subsection, the TRF procedure as in Borup et al. (2023) is described, who forecast S&P500 equity premium, industrial production growth, employment growth, and CPI acceleration over 1, 3, and 12 month ahead horizons using LASSO/RF.

2.3.1 Least Absolute Shrinkage and Selection Operator

Consider the LASSO minimization by Tibshirani (1996) as in Equation 1.

$$\hat{\beta}_{LASSO} = \min_{\tilde{\beta}} \left(\sum_{t=1}^{T} (y_t - \sum_{j=1}^{K} x_{j,t}\beta_j)^2 + \lambda \sum_{j=1}^{K} |\beta_j| \right) \tag{1}$$

Here, λ is the tuning, or hyper-, parameter that controls for the strength of the L^1 penalty: the sum of absolute coefficients β_j. LASSO performs variable selection by setting coefficients to 0, leading to sparse models, wherein the degree of sparsity is determined by the value of λ. Borup et al. (2023) target a subset of variables containing $s = 5, 10, 20, 30, 50$ variables.

This is replicated in this paper by constructing the regularization path (Friedman et al., 2010) for 100 values of λ, which correspond to a certain number of non-zero variables. The λ parameter is tuned using cross-validation. As no specific rules were set by Borup et al. (2023) in cases where more or no values of λ correspond to a specific number of variables s, the following is assumed:

1. When multiple λ correspond to a specific target s, the λ with the smallest mean square error (MSE) is selected.
2. When no λ corresponds to a specific target s, the λ corresponding to the smallest integer larger than s is chosen. For example, when s = 5 is not in the computed regularization path, but s = 6 is, then the λ for s = 6 will be chosen.

Afterwards, the dependant variable y_t is regressed using LASSO for the λ corresponding to specified degree of targeting. The selected non-zero variables are then passed on to the random forest regression step.

2.3.2 Random Forest

The RF machine learning algorithm by Breiman (2001) approximates non-linear functions through utilization of decision trees. There are many types of decision trees, such as classification trees and regression trees, which are used for binary and continuous variables respectively. As continuous variables are used, the decision tree of interest is the regression tree. The tree grows by optimally splitting on each node in a tree, creating new nodes from the parent node. This optimal splitting point consists of an optimal variable and observation and is chosen by means of maximizing the impurity decrease as in Breiman (2017).

The RF model as noted in Coulombe et al. (2021) is given in Equation 2.

$$\hat{y}_{t+h} = \sum_{m=1}^{M} c_m I_{R_m \in Z_t} \tag{2}$$

with R_m for $m = 1,...,$ M being a partition of the feature space Z_t and M being the number of terminal nodes of the regression tree. A subset of variables is randomly sampled for potential splitting at each node. Then, by means of bootstrap aggregating (bagging), the RF algorithm grows many regression trees on subsamples of the data. After which, a forecast is made as the average of all forecasts in the bootstrap. The RF regression is considered for the various targeted subsets of variables, which obtains the TRF model, as well as the full set of variables, which will be referred to as the full RF.

2.3.3 Forecasting and Evaluation

As mentioned before, Borup et al. (2023) adopt an expanding window framework for a dataset ranging from 1970 January to 2018 December. They set their initial estimation window between 1970 January and 1984 December. Meaning that the first point forecast is made for the h-th month of 1985, for horizons h = 1, 3 and 12 months ahead and the last point forecast is made for December 2018. The TRF is re-estimated at every expansion of the window, that is after each point forecast.

Predictions are then evaluated based on the widely used MSE, denoted as:

$$MSE = \frac{1}{T_{OS}} \sum_{t=T_{OS}}^{T} e_{t+h}^2$$

with e_{t+h}^2 being the squared forecast error at time t+h, summed over the out-of-sample period, which runs from $t = T_{OS}$ until T

Next MSE ratio's are computed as:

$$Ratio = \frac{MSE_1}{MSE_2}$$

with MSE_1 the MSE of the first forecasting model, $LASSO_s/RF$ in the case Borup et al. (2023) with s number of variables and MSE_2 the MSE of the second forecasting model, which is the full RF in Borup et al. (2023). A ratio smaller than 1 indicates that model 1 has a lower MSE and thus a better performance.

To test whether the TRF is significantly better, an one-sided Diebold-Mariano (DM) test is used, as in Diebold and Mariano (2002). Denote the loss difference of squared forecast errors between two forecasts from models i and j at time t as:

$$d_{ij,t} = e^2_{t+1|t,i} - e^2_{t+1|t,j} \tag{3}$$

The DM test statistic is then constructed using HAC standard errors and a Bartlett kernel as done in Borup et al. (2023). The null is that of equal performance against the alternative that the TRF performs better than the full RF:

$$H_{0,DM} : E(d_{ij,t}) = 0$$
$$H_{1,DM} : E(d_{ij,t}) < 0$$
$$DM = \frac{\bar{d}_{ij}}{\sqrt{V(\hat{d}_{ij})}}$$

Here $\bar{d}_{ij} = T_{OS}^{-1} \sum_{t=T_{IS}}^{T_{OS}} d_{ij,t}$ is the sample mean loss difference and $V(\hat{d}_{ij})$ is the variance of the loss differences.

2.4 Methodology - Extension

This subsection presents the methodology applicable to the extension of Borup et al. (2023). First, a number of penalty-based and tree-based selection methods are considered and discussed in more detail below. Then, the forecasting procedure is described along with additional tests for forecasting performance.

2.4.1 Adaptive LASSO

One problem of LASSO is that it can introduce bias in the estimated coefficients, which is especially apparent in smaller sample sizes. Multicollinearity is another problem, where in cases of high correlation between certain variables, LASSO will arbitrarily select one of those variables, while setting the others to zero (Altelbany, 2021). Lastly, in LASSO, coefficients are all equally penalized in the L^1 penalty given in Equation 1.

To solve this, Zou (2006) proposes the adaptive LASSO (adaLASSO) minimization problem denoted in Equation 4.

$$\hat{\beta}_{adaLASSO} = \min_{\tilde{\beta}} \left(\sum_{t=1}^{T} (y_t - \sum_{j=1}^{K} x_{jt}\beta_j)^2 + \lambda \sum_{j=1}^{K} w_j |\beta_j| \right) \tag{4}$$

for $w_j \geq 0$ and $\lambda > 0$. Here w_j is the j-th weight given to the j-th coefficient β_j, which controls the degree of penalization for that coefficient.

Zou (2006) finds adaLASSO performing better than the preceding LASSO, while also adhering to the so-called oracle property, meaning that it is consistent in both parameter estimation and variable selection. adaLASSO additionally tends to produce sparser models than LASSO, showing that it is more effective in selecting the relevant variables. Furthermore, adaptive LASSO is more robust to multicollinearity in the data, as it penalizes weak regressors more strongly than strong regressors.

Initial weights for adaLASSO are computed by optimizing Ridge regression (McDonald, 2009) coefficients according to Bayesian Information Criterion (BIC). The Ridge minimization problem is given as:

$$\min_{\tilde{\beta}} \left(\sum_{t=1}^{T} (y_t - \sum_{j=1}^{K} x_{jt}\beta_j)^2 + \lambda \sum_{j=1}^{K} \beta_j^2 \right)$$

Much like LASSO, λ controls the strength of the penalty, which is the sum of squared coefficients β_j in this case.

For variable selection in the targeting step, regularization paths are constructed for 100 values of λ, where the λ parameter is tuned by means of cross-validation. The same assumptions as in Section 2.3.1 are followed for the selection of λ in case more or no values correspond to the variable target amount s.

Finally, adaLASSO is also used as a forecasting model in the reverse TRF setting: $RF_s/adaLASSO$. Again, initial weights are first determined by means of Ridge regression for the RF targeted subset of variables, optimzied for BIC. Afterwards, adaLASSO coefficients are estimated with the λ parameter tuned such that it minimizes MSE. The model is then used for forecasting.

2.4.2 Smoothly Clipped Absolute Deviation

Next, SCAD is considered as a penalty-based variable selection method with a concave penalty, opposed to the convex penalties in LASSO and adaLASSO. The SCAD minimization problem and corresponding penalty as proposed in Fan and Li (2001) are given in Equation 5 and Equation 6, respectively.

$$\hat{\beta}_{SCAD} = \min_{\tilde{\beta}} \left(\sum_{t=1}^{T} (y_t - \sum_{j=1}^{K} x_{jt}\beta_j)^2 + \lambda \sum_{j=1}^{K} p_\lambda(\beta_j) \right) \tag{5}$$

for $a > 2$, $\beta > 0$ and with the continuous differentiable penalty $p'_\lambda(\beta)$ function given as:

$$p'_\lambda(\beta) = \lambda \left(I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right) \tag{6}$$

This corresponds to a quadratic spline function with knots at λ and $a\lambda$ (Fan and Li, 2001). The penalty $p_\lambda(\beta)$ may then be represented according to the following piecewise function:

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda \\ \frac{2a\lambda|\beta| - \beta^2 - \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta| \leq a\lambda \\ \frac{\lambda^2(a+1)}{2} & \text{otherwise.} \end{cases} \tag{7}$$

An advantage of SCAD is that it, like adaLASSO, adheres to the oracle property, meaning that it is consistent in both parameter estimation and variable selection, being especially apparent in large sample data. Fan and Li (2001) also state in their research that SCAD does not create excessive bias in its estimated coefficients, unlike LASSO. This is due to the piecewise form of the penalty function, resulting in smoothing of coefficient penalization.

In terms of targeting step, SCAD follows the same procedure as described in Sections 2.3.1 and 2.4.1 by constructing regularization paths. The same assumptions are also applicable here. Lastly, SCAD is also one of the models applied in the reverse TRF setting, denoted as: $RF_s/SCAD$. Much like, adaLASSO, SCAD coefficients for the RF targeted subset of variables are estimated for the dependant variable with λ tuned to minimize MSE, after which forecasts are made.

2.4.3 Random Forest Variable Selection

Aside from being the main regression method in the second step of TRF, Random forest is used as one of the tree-based variable selection methods in the targeting step. This is done by first obtaining the RF regressions using the full set of variables, upon which the variables are ranked by means of the percentage increase in MSE of variable k (Archer and Kimes, 2008). This measure is calculated by first calculating the MSE of the regression RF, then the observations of a given variable x_k are shuffled, which is also known as permuting. After which the MSE of the tree with the shuffled values is calculated. The percentage increase in MSE of variable k can then be denoted as:

$$\%MSEinc_k = 100\% \cdot \frac{MSE_{1,k} - MSE_0}{MSE_0}$$

Here, $MSE_{1,k}$ represents the MSE after permuting variable k and MSE_0 represents the non-shuffled random forest MSE. The percentage increase in MSE is computed for all x_k, k = 1,..., K variables with high values indicating a higher importance.

2.4.4 Gradient Boosting Machine

Apart from RF variable selection, Gradient Boosting Machine (GBM) is also considered as one of the tree-based variable selection methods. First proposed in Friedman (2001), gradient boosting is a method to approximate non-linear functions using decision trees, similar to RF. The difference is that in RF, trees are built independently and the results are combined for the eventual forecast, while GBM builds the tree additively. Which is to say that the next tree is built upon the previous tree.

The GBM model by Friedman (2001) is given as follows in Equation 8.

$$y_{t-h}^{(n+1)} = y_{t-h}^{(n)} + \rho_{n+1} f(Z_t, c_{n+1}) \tag{8}$$

with $(c_{n+1}, \rho_{n+1}) = \arg\min_{\rho,c} \sum_{t=1}^{T} (e_{t+h}^{(n)} - \rho_{n+1} f(Z_t, c_{n+1}))^2$, and $c_{n+1} = (c_{n+1,m})_{m=1}^{M}$ consisting of the parameters used in the regression trees. The total number of trees that are summed together is indexed by n and m indexes the successive increments, based on the preceding sequence of steps.

GBM is applied both as variable selection method in the targeting step, as well as forecasting model in the forecasting step. For targeting, the set-up is similar to that of RF, discussed in Section 2.4.3. First, a GBM is estimated over the full set of features, after which the features are ranked based on the gain as in Chen and Guestrin (2016). This measure represents the fractional contribution of a feature x_k to the model, according to the total gains of the variable's split. The gain of each feature is computed and ranked, with high gain indicating higher importance. The first $s \leq K$ variables are then subsetted into the targeted set of variables for forecasting.

GBM is used as forecasting model in the targeted gradient boosting machine setting, which can be denoted as: V_s/GB. In which it uses the targeted subset of variables of size s to compute point forecasts. The set-up of TGB is identical to that of TRF, except that GBM is the forecasting model of interest.

2.4.5 Forecasting and model evaluation

Forecasts using the methodology applied as an extension on Borup et al. (2023) is set in a rolling window framework with a window size of 15 years (180 observations), instead of an expanding window framework. On one hand, older observations are disregarded as these observations likely are less or no longer important for a forecast years later. On the other hand, there is a potential worsening of the of the RF model forecasts as less data is considered per forecast. Decision trees are not able to extrapolate as explained in Zhang et al. (2019). The loss of observations could be problematic for the forecast in periods where a forecast is made for an observation of which its actual value is not within the highest and lowest observed value bounds for that window. However, as stationary variables are considered, this is not deemed a large issue for our setting. Lastly, the original forecast horizons of h = 1, 3 and 12 months ahead are retained for the extension and all models are re-estimated at the next time interval, after a forecast has been made.

In terms of performance measures and evaluation, MSE ratio's and Diebold-Mariano tests as introduced in Section 2.3.3 are also applicable when comparing two settings, such as TGB with differing degrees of targeting vs a full GBM or a reverse TRF vs a standard TRF. In cases like the latter example, where a targeted model is compared against another targeted model, a two-sided Diebold-Mariano test is considered, rather than an one-sided test. This is achieved by re-formulating the alternative hypothesis as $H_{1,DM} : E(d_{ij,t}) \neq 0$. This methodology of reversing the order of variable selection and forecasting method was previously applied in Medeiros et al. (2021). In their research, they compare the performance of RF/OLS and adaLASSO/RF with a full RF. They note that the results can be interpreted as a test for importance of non-linearity in the RF/OLS case and importance of variable selection in the adaLASSO/RF case.

2.4.6 Model Confidence Sets

Next, Model Confidence Sets (MCS) by Hansen et al. (2011) is used in order to find the best predictive models. MCS attempts to find the superior set of models (SSM) $MCS^*_{1-\alpha}$ from an initial set of models MCS based on a given level of confidence $1 - \alpha$ and an user-specified loss function, such as the squared forecast error. The optimal set $MCS^*_{1-\alpha}$ is then found by sequentially testing over a bootstrap sample B. The null hypothesis is that of equal forecasting performance, against the alternative of unequal performance:

$$H_{0,MCS} : E(d_{ij,t}) = 0 \text{ for all i,j} \in S$$
$$H_{1,MCS} : E(d_{ij,t}) \neq 0 \text{ for all i,j} \in S$$

with $d_{ij,t}$ the loss difference as defined in Equation 3. The corresponding test statistic $T_{R,MCS}$ is then defined as:

$$T_{R,MCS} = \max_{i,j \in MCS} |t_{ij}| \text{ with} \tag{9}$$

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{V(\hat{d}_{ij})}} \tag{10}$$

which shows that the test statistic to test $H_{0,MCS}$ is given by the absolute maximum of all Diebold-Mariano test statistics considered for every model $i, j \in MCS$.

MCS then removes the models that performs the worst relatively and performs the test again, until all models in the set have a p-value larger than α, after which the SSM is obtained. Larger levels of α result in a smaller subset of models. Additionally, the models within the SSM are also ranked based on their relative performance.

2.4.7 Fluctuation Test

Finally, to test whether there were periods where the models performed statistically the same and periods where one model significantly outperformed the other, the fluctuation test by Giacomini and Rossi (2010) is used to test the null hypothesis of equal performance at each time $t \in T_{OS}$. This test is designed for unstable environments, for which our dataset is extended to include the recent Covid-19 financial crisis.

The null hypothesis and its test statistic are presented below in Equation 11:

$$H_0 : E(d_{ij,t}) = 0 \text{ for each } t \in T_{OS} \tag{11}$$

$$F_{t,R}^{OS} = \hat{\sigma}^{-1} R^{-1/2} \sum_{s=t}^{t+R} d_{ij,s} \tag{12}$$

where R is the size of the rolling window, σ^{-1} is a HAC-estimator of variance of the loss difference $d_{ij,t}$ and the test statistic $F_{t,R}^{OS}$ represents the relative performance of the two models at time t. If at any point, this statistic crosses a critical value K_F, there is a significant indication that one of the models outperformed the other at that given moment.

Finally, Giacomini and Rossi (2010) show that the critical values K_F depend on the size of the rolling window, relative to the sample period, that is $\mu = \frac{R}{T}$. The authors find that the critical value K_F decreases as μ, and thus the rolling window size, becomes larger.

3 Data

Two datasets are considered to obtain the results. Both datasets follow the setting as in Borup et al. (2023), which uses variables from the FRED-MD database by McCracken and Ng (2016). This dataset consists of 127 monthly series of macroeconomic indicators of the US. Two sample periods are considered, the first is between 1970-2018[2] as in Borup et al. (2023). The second sample period extends on the original dataset by including the Covid-19 period, adding data from 2019, 2020 and 2021[3].

3.1 Industrial Production Growth

The variable of interest is industrial production, corresponding to panel B of Borup et al. (2023). The series are treated as an $I(1)$ variable, meaning that there is an unit root in the level data. For this reason, the forecasting objective in the form of industrial production growth is considered, accumulated over horizon h:

$$y_{t+h} = \log i_{t+h} - \log i_t \tag{13}$$

with i_t being the industrial production at time t.

To verify that the series are now non-stationary, an augmented Dickey-Fuller (ADF) is done for y_{t+h} based on BIC with the null hypothesis being that the series contain an unit root. The test returns p-values of $p = 0.0000...$ for all horizons h = 1, 3 and 12, indicating that, indeed, the cumulative industrial production growth compunded over horizon h is non-stationary.

3.2 Data Processing

As the raw data consist of mostly non-stationary series, the data are transformed according to the recommended transformations from McCracken and Ng (2016). Next, the data are further processed by removing all variables that have missing observations from consideration. This results in a dataset of 125 variables for the 1970-2018 sample and a set of 102 variables for the 1970-2021 sample. In Borup et al. (2023) they state that their dataset consists of exactly 100 predictors, however it is not disclosed which vintage of the FRED-MD dataset is used for their research.

4 Results

Results of the various settings of the targeted forecasting models are reported below. Section 4.2 presents the results obtained from TRF, replicating the set-up as in Borup et al. (2023). The following sections give the results in the rolling window framework for TRF, TGB, reverse TRF and TRF with extended dataset, respectively. Finally, a robustness check is performed to analyse the importance of the window size in the rolling window framework. Note that a bootstrap sample of $B = 10000$ is used for MCS.

[2] Vintage of October 2020 is taken from https://research.stlouisfed.org/econ/mccracken/fred-databases/
[3] Vintage of October 2022 is taken from https://research.stlouisfed.org/econ/mccracken/fred-databases/

4.1 Replication Targeted Random Forest

Table 1 gives the MSE ratios for $LASSO_s/RF$ in the expanding window framework of Borup et al. (2023), with s = 5, 30, 50 and fully grown, deep trees, found in Table 5 of their Appendix. The replication results deviate somewhat as random forest is inherently random. Other than the inherent randomness in random forest, another reason the results deviate is because different data sets were used. As mentioned before, it is not stated what vintage of the dataset is used in the original paper. This could be an issue as the FRED-MD is backwards updated as data becomes corrected, meaning different values are likely in place for some variables. Secondly, the dataset in the replication consists of 125 variables, which is what is obtained after following the procedure described in , while the original authors obtained a set of 100 variables instead. The consequence of this is that LASSO will certainly have differences in hyper-parameters and selected variables compared to Borup et al. (2023), which then leads to different random forests and forecasts. The largest deviation is seen in the three- and twelve-step ahead TRF with 5 variables selected, where the replicated full RF performs much better than the TRF and the ratio's deviate largely from the original ratios. Other than that, similar results are found overall, with the three-step ahead TRF targeting s = 30 and 50 variables being the closest to the original. Other than the twelve-step ahead TRF with s = 5, the same directions in terms of improvement are found in both replication and original.

Panel: Industrial Production Growth						
	Replication			Original		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	1.145	1.351	1.179	1.122	1.192	**0.953**
30	1.063	**0.959**	**0.950**	1.050	**0.968**	**0.919**
50	1.031	**0.983**	**0.928***	1.025	**0.948**	**0.895****

Table 1: This Table reports MSE ratios between each TRF and full RF, for h = 1, 3 and 12 step forecast horizons. The sample runs from 1970 January to 2018 December. Forecasts are made under an expanding window framework. Bold numbers indicate improvement from targeting and *, **, *** indicate statistical significance at levels of 10%, 5% and 1%, respectively, based on an one-sided Diebold Mariano test.

4.2 Targeted Random Forest

Table 2 reports significantly improved forecasting performance of TRF for forecast horizons of h = 12 and s = 5, 30 and 50 variables selected for LASSO, adaLASSO and SCAD variable selection methods. This indicates that targeting helps significantly at the 5% to 10% level for longer horizons where forecasting is more inaccurate. In general, larger improvements are found when targeting a larger number of variables. Next, barely any improvements are found when targeting using RF and GBM. When there are improvements, they are not found to be significant. Thus, it is likely that the tree-based targeting methods are not selecting optimal variables for forecasting. Finally, regular RF with a full set of variables mostly outperforms TRF for h = 1 month ahead forecast horizon, across the board, showing that the full RF performs very well in forecasts with less uncertainty.

Table 3 reports the set of superior models obtained from MCS. MCS are conducted for all

Panel: Industrial Production Growth									
	LASSO/RF			adaLASSO/RF			SCAD/RF		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	1.042	**0.974**	**0.934***	1.129	1.053	**0.671***	1.042	**0.991**	**0.935**
30	1.038	**0.944**	**0.846****	1.0365	**0.864**	**0.732****	1.032	1.008	**0.777****
50	1.000	**0.962**	**0.848****	**0.988**	**0.905***	**0.733****	1.027	**0.977**	**0.780****
	RF/RF			GBM/RF					
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12			
5	1.151	1.189	**0.952**	1.130	1.078	**0.967**			
30	1.050	1.015	1.024	1.078	**0.972**	1.005			
50	1.037	1.004	1.013	1.030	**0.982**	**0.997**			

Table 2: This Table reports the MSE ratio between each TRF and full RF, for h = 1, 3 and 12 step ahead forecast horizons. The sample runs from 1970 January to 2018 December. Bold numbers indicate improvements from targeting and *, **, *** indicate statistical significance at confidence levels of 10%, 5% and 1%, respectively, based on an one-sided Diebold-Mariano test.

	h = 1		
s	$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
5	{Full, LAS, SCAD	{Full, LAS, SCAD}	{Full}
30	{Full, SCAD, ada}	{Full}	{Full}
50	{ada, LAS, Full, SCAD}	{ada, Full, LAS}	{ada}
	h = 3		
s	$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
5	{LAS, Full, ada}	{LAS, Full, ada}	{LASSO}
30	{ada, LAS, GBM, RF, Full}	{ada}	{ada}
50	{ada, LAS, GBM, Full, RF}	{ada}	{ada}
	h = 12		
s	$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
5	{ada, LAS, RF, Full, GB, SCAD}	{ada}	{ada}
30	{ada}	{ada}	{ada}
50	{LAS, SCAD, ada, GB, RF, Full}	{ada, SCAD}	{ada}

Table 3: MCS for all versions of TRF and full RF at confidence levels of $1-\alpha=0.7$, 0.4 and 0.1. The Table results give the SSM at each forecast horizon h, each selected variable target s and each level of $1-\alpha$. The models in the SSM are ranked from best (left) to worst (right).

combinations of the three horizons, three selection targets and three confidence levels of $1-\alpha=0.7$, 0.4 and 0.1, meaning that in total 27 MCS are performed, consisting of 5 TRFs and the full RF each. As $1-\alpha$ decreases, the cardinality of the SSM decreases until, in this case, there is only the best surviving model in the SSM. The results suggest that the full RF model likely performs the best for short horizons, while *adaLASSO/RF* performs better in the longer forecasting horizons. As $1-\alpha$ is decreased, one may expect that the worst models in the SSM are removed in order, however this is not always the case, for example in the MCS with h = 12 and s = 50. As $1-\alpha$ decreases from 0.7 to 0.4, LASSO/RF is removed from the SSM, while being the best performing model for $1-\alpha=0.7$. While the reason to this is not entirely clear, one explanation could be that LASSO/RF was performing very well relative to GB/LASSO, RF/LASSO and full RF, but as those models are removed, LASSO/RF loses its relative strength and is removed.

Next, all TRFs are combined into one large MCS of 16 models for each forecasting horizon,

h = 1		
$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
$\{LAS_{50}$, ada_{50}, $Full$, $SCAD_{50}$, GB_{50}, $SCAD_{30}$, ada_{30}, LAS_{5}, $SCAD_{5}$, RF_{50}, LAS_{30}, $RF_{30}\}$	$\{ada_{50}$, $Full$, $LAS_{50}\}$	$\{ada_{50}\}$
h = 3		
$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
$\{ada_{50}$, ada_{30}, LAS_{30}, LAS_{50}, GB_{30}, LAS_{5}, $SCAD_{50}$, GB_{50}, $SCAD_{5}$, $Full$, RF_{50}, RF_{30}, $ada_{5}\}$	$\{ada_{50}$, ada_{30}, LAS_{30}, LAS_{50}, LAS_{5}, GB_{30}, GB_{50}, $Full$, RF_{50}, $RF_{30}\}$	$\{ada_{30}\}$
h = 12		
$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
$\{LAS_{50}$, ada_{30}, LAS_{30}, $SCAD_{30}$, $SCAD_{50}$, ada_{50}, ada_{5}, RF_{5}, GB_{50}, GB_{30}, $Full$, RF_{50}, RF_{30}, LAS_{5}, $GB_{5}\}$	$\{ada_{5}\}$	$\{ada_{5}\}$

Table 4: Model Confidence set for all versions of TRF and full RF at confidence levels of $\alpha =$ 0.3, 0.6 and 0.9. The Table results give the SSM at each forecast horizon h, and each level of α. The models in SSM are ranked from best (left) to worst (right).

still for confidence levels of $1-\alpha = 0.7$, 0.4 and 0.1. AdaLASSO/RF is found to be the only remaining model in the SSM at higher confidence levels. Notice that the optimal degree of targeting is inversely proportional to the forecasting horizon. This means that for h = 1, the high degree of targeting s = 50 is preferred, while for h = 12, a low amount of targeting is preferred. This could be due to the fact that it is harder to assess which variables are important for longer horizons, which can then lead the trees in a full RF model to split sub-optimally. Furthermore, cases are found where all the forecasting models are removed from the SSM except for one, as $1-\alpha$ decreases from 0.7 to 0.4, such as in the MCS for h = 12. This indicates that the surviving model performs much better than its competing models and that the competing models survived due to the high confidence level. The opposite is true when little models are removed as $1-\alpha$ decreases, such is the case for the h = 3 MCS. Several models are found to perform equally at mid-level values of $1-\alpha$.

4.3 Targeted Gradient Boosting Machine

TGB of the form V_s/GBM is considered as an alternative to TRF. Table 5 reports the MSE ratios and DM test results when comparing TGB to full GB. In terms of improvement, mixed results are found with targeting showing forecast improvements in roughly half the cases. Much like TRF, TGB mostly find improvements in the year-ahead forecast horizon, with adaLASSO looking like the best targeting method. Furthermore, TGB seems to improve more with tree-based targeting compared to TRF, which showed small or no improvements with RF and GBM targeting. However TGB also shows smaller improvements with penalty-based targeting methods, which showed high forecasting performance gains for TRF. This indicates that the performance of GBM does not depend as much on having an optimally targeted set of variables for forecasting, making GBM more consistent with a higher floor but also a lower ceiling in terms of performance.

Panel: Industrial Production Growth									
	LASSO/GBM			adaLASSO/GBM			SCAD/GBM		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	**0.912**	1.024	1.019	1.071	1.265	**0.743**	**0.923**	1.029	1.025
30	**0.965**	1.024	**0.921***	1.004	**0.989**	**0.812****	**0.962**	1.082	**0.891***
50	1.046	**0.992**	**0.970**	**0.999**	**0.994**	**0.943**	**0.988**	**0.949**	**0.927***
	RF/GBM			GBM/GBM					
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12			
5	**0.945**	1.168	**0.914**	1.036	1.179	**0.950**			
30	1.065	1.052	1.011	1.072	**0.949**	**0.946**			
50	**0.832***	**0.931**	1.026	1.040	1.011	**0.949**			

Table 5: This Table reports the MSE ratio between each TGB and full GB, for h = 1, 3 and 12 step ahead forecast horizons. The sample runs from 1970 January to 2018 December. Bold numbers indicate improvements from targeting and *, **, *** indicate statistical significance at confidence levels of 10%, 5% and 1%, respectively, based on an one-sided Diebold-Mariano test.

$1-\alpha = 0.7$	$1-\alpha = 0.4$	$1-\alpha = 0.1$
h = 1		
$\{RF_{50}, LAS_5, SCAD_5, LAS_{30}, RF_5, SCAD_{30}, SCAD_{50}, Full, ada_{50}, ada_{30}, GB_{50}, GB_{30}, LAS_{50}, RF_{30}, GB_5\}$	$\{RF_{50}, LAS_5, SCAD_5, SCAD_{50}, Full, ada_{50}, ada_{30}, GB_{50}, GB_{30}, RF_{30}, LAS_{50}\}$	$\{RF_{50}\}$
h = 3		
$\{GB_{30}, RF_{50}, SCAD_{50}, LAS_{50}, ada_{50}, Full, ada_{30}, GB_{50}, LAS_{30}, LAS_5, SCAD_5, RF_{30}, SCAD_{30}, ada_5\}$	$\{RF_{50}, GB_{30}, SCAD_{50}, LAS_{50}, ada_{30}, ada_{50}, Full, GB_{50}, LAS_{30}, LAS_5, SCAD_{30}, SCAD_{30}\}$	$\{RF_{50}, GB_{30}, SCAD_{50}, ada_{30}\}$
h = 12		
$\{ada_{30}, SCAD_{30}, ada_5, LAS_{30}, RF_5, SCAD_{50}, GB_{30}, GB_5, ada_{50}, GB_{50}, Full, RF_{30}, LAS_{50}, RF_{50}\}$	$\{ada_5\}$	$\{ada_5\}$

Table 6: Model Confidence set for all versions of TGB and full GBM at confidence levels of $1-\alpha$ = 0.7, 0.4 and 0.1. The Table results give the SSM at each forecast horizon h, and each level of $1-\alpha$. The models in SSM are ranked from best (left) to worst (right).

Then, another MCS procedure is performed to find the best targeting method for TGB with the results presented in Table 6. The MCS are again sub-divided in each forecast horizon, along with confidence levels of $1-\alpha = 0.7$, 0.4 and 0.1. Each MCS consist of a set of 16 models - 15 TGBs and the full GBM. MCS are also performed for the models per target s, which includes 6 models per MCS akin to the setup for Table 3. The results for this are given in

Table 12 in Appendix A. For now, our focus is directed to Table 6, which shows RF_{50}/GB and $adaLASSO_5/GB$ being the best performing models for horizons h = 1 and h = 12, respectively. For h = 3, multiple models are found in the SSM at the lowest confidence level, indicating equal performance between those 4 models. This is further supported by Table 5, where it is evident that the gains of targeting are relatively small and always insignificant. The TGB that do improve from targeting, obtain similar MSE ratios, indicating equal performance between those models.

Panel: Industrial Production Growth									
	LASSO/M			adaLASSO/M			SCAD/M		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	1.244***	1.121**	1.241**	1.349***	1.280***	1.260***	1.259***	1.107*	1.247**
30	1.321***	1.157**	1.238*	1.377***	1.220*	1.261*	1.325***	1.144**	1.304*
50	1.487***	1.098	1.301**	1.436***	1.170**	1.387*	1.367***	1.035	1.352*
	RF/M			GBM/M					
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12			
5	1.168***	1.050	1.093*	1.096	1.009	1.039			
30	1.441***	1.105	1.123	1.006	**0.902***	**0.935**			
50	1.141*	**0.988**	1.152**	1.250*	1.087	**0.925****			

Table 7: This Table reports MSE ratios between each TGB and TRF, for h = 1, 3 and 12 step ahead forecast horizons. The sample runs from 1970 January to 2018 December. Ratios larger than 1 indicate TRF to outperform, while ratios below 1 indicate TGB to outperform and *, **, *** indicate statistical significance at confidence levels of 10%, 5% and 1%, respectively, based on a two-sided DM test. The labels represent the used targeting method for RF and GBM.

Next, comparisons between TGB and TRF are made by taking MSE ratios as: $\frac{MSE_{TGB}}{MSE_{TRF}}$, which are shown in Table 7. It follows that TRF significantly outperforms TGB in terms of forecasting performance, apart from a few instances in the tree-based selection methods, which were found to already be the weakest targeters of TRF. This result is as expected as TGB showed to have lower improvements from targeting compared to TRF.

To further investigate whether TGB is viable for forecasting, an MCS is performed for every forecast horizon including all TGB, TRF, and full GBM and RF models, meaning 32 models are included for each MCS procedure. Table 8 presents the MCS results. Immediately telling is the weak performance of TGB compared to TRF, seeing that all TGB models have been removed from the SSM at the medium confidence level of $1 - \alpha = 0.4$ and those that were not removed at confidence level of $1 - \alpha = 0.7$ are generally the weakest models in the SSM. The results reinforce that TGB is not viable for forecasting, compared to TRF

Finally, notice that the SSM obtained under confidence levels $1 - \alpha = 0.4$ and 0.1 are equivalent to those obtained in Table 4, meaning that the same conclusion can be reached with regards to the targeting degree of adaLASSO and forecast horizons, where longer horizons prefer a small degree of targeting and short horizons prefer a large degree of targeting.

h = 1				
$1-\alpha=0.7$		$1-\alpha=0.4$		$1-\alpha=0.1$
$\{Full_{RF}$, GB_{50}/RF, RF_{50}/RF, LAS_{30}/RF, LAS_{50}/RF, $SCAD_{50}/RF$, RF_{30}/RF, ada_{50}/RF, ada_{30}/RF, $SCAD_{30}/RF$, LAS_5/RF, $SCAD_5/RF$, RF_5/RF, $Full_{GB}$, GB_{50}/GB, $GB_{30}/GB\}$		$\{ada_{50}/RF$, $Full_{RF}$ $LAS_{50}/RF\}$		$\{ada_{50}/RF\}$
h = 3				
$1-\alpha=0.7$		$1-\alpha=0.4$		$1-\alpha=0.1$
$\{ada_{50}/RF$, LAS_{30}/RF, ada_{30}/RF, LAS_{50}/RF, GB_{30}/RF, $SCAD_{50}/RF$, LAS_5/RF, GB_{50}/RF, RF_{50}/GB, $SCAD_5/GB$, $Full_{RF}$, $SCAD_{50}/GB$ RF_{50}/RF, RF_{30}/RF, GB_{30}/GB, $ada_5/RF\}$		$\{ada_{50}/RF$, ada_{30}/RF, LAS_{30}/RF, LAS_{50}/RF, LAS_5/RF, GB_{30}/RF GB_{50}/RF, $Full_{RF}\}$		$\{ada_{30}/RF\}$
h = 12				
$1-\alpha=0.7$		$1-\alpha=0.4$		$1-\alpha=0.1$
$\{LAS_{50}/RF$, ada_{30}/RF, LAS_{30}/RF, $SCAD_{30}/RF$, $SCAD_{50}/RF$, ada_{50}/RF, LAS_5/RF, ada_5/RF, ada_{30}/GB, RF_5/RF, GB_5/RF, GB_{50}/RF, $Full_{RF}$, GB_{30}/RF, RF_5/GB, $SCAD_{30}/GB$, RF_{50}/RF, $SCAD_{50}/GB$, RF_{30}/RF, LAS_{30}/GB, ada_{50}/GB, $Full_{GB}$, RF_{30}/GB, GB_{30}/GB, GB_{50}/GB, LAS_{50}/GB, GB_5/GB, $LAS_5/GB\}$		$\{ada_5/RF\}$		$\{ada_5/RF\}$

Table 8: Model Confidence set for all versions of TRF, full RF, TGB and full GBM at confidence levels of $1-\alpha=0.7$, 0.4 and 0.1. The Table results give the SSM at each forecast horizon h, and each level of $1-\alpha$. The models in SSM are ranked from best (left) to worst (right).

4.4 Reverse Targeted Random Forest

Reverse TRFs are considered to gauge the usefulness of variable selection and TRF forecasting by comparing reverse TRFs with their full forecasting model, for example, comparing $RF_s/adaLASSO$ to a full adaLASSO forecasting model, and by comparing reverse TRFs with the standard TRFs.

Table 9 presents the MSE ratios with corresponding DM test significance of the reverse TRF compared to their full models for adaLASSO, SCAD and GBM. Some small improvements in forecasting performance are found, but only $RF_5/adaLASSO$ finds significant improvement at the 10% level for one month ahead forecasts. Aside from RF_s/GBM with h = 1 being largely

Panel: Industrial Production Growth									
	RF/adaLASSO			RF/SCAD			RF/GBM		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	**0.905***	1.009	1.039	1.043	1.059	1.001	1.148	**0.913**	**0.833**
30	**0.944**	**0.941**	**0.949**	1.035	**0.945**	**0.983**	1.199	**0.924**	**0.836**
50	**0.964**	1.050	**0.893**	1.085	1.018	**0.995**	1.219	**0.928**	**0.836**

Table 9: This Table reports the MSE ratio between each reversed TRF, that is using RF as variable selection step, for h = 1, 3 and 12 step ahead forecast horizons and its forecasting model using full variables. The sample runs from 1970 January to 2018 December. Bold numbers indicate improvements from targeting and *, **, *** indicate statistical significance at confidence levels of 10%, 5% and 1%, respectively, based on an one-sided Diebold-Mariano test.

outperformed by full GBM, most results suggest practically equal forecasting performance for ratios close to 1, such as for $RF_5/adaLASSO$ with h = 3 and $RF_5/SCAD$ with h = 12. This seems to suggest that potentially, RF is not able to target optimal variable subsets for forecasting or removes important variables, which would explain why RF, and by extension GBM, seem to underperform when applied as targeting method.

Panel: Industrial Production Growth									
	RF/adaLASSO			RF/SCAD			RF/GBM		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	**0.954**	**0.955**	1.446	1.067	1.126	1.117	1.443***	**0.903**	**0.980**
30	1.084**	1.085	1.209**	1.070	**0.989**	1.320	1.579***	1.013	**0.946**
50	1.160**	1.157	1.077	1.126*	1.098	1.330	1.682***	1.007	**0.954**

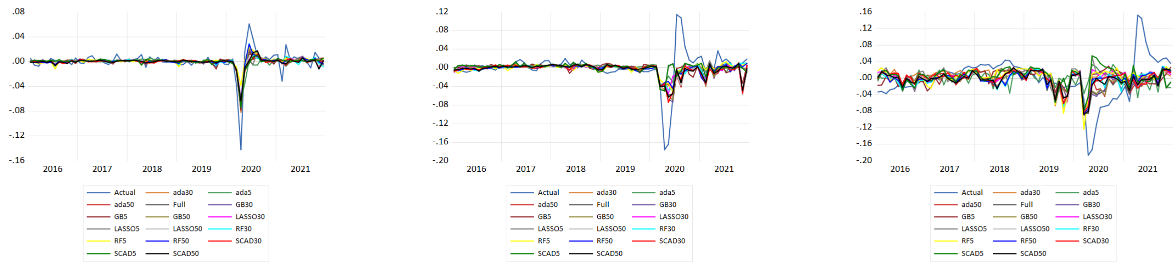
Table 10: This Table reports the MSE ratio between each reversed TRF, that is using RF as variable selection step, for h = 1, 3 and 12 step ahead forecast horizons and the original TRF used in Table 1. The sample runs from 1970 January to 2018 December. Ratios larger than 1 indicate original TRF to outperform, while ratios below 1 indicate reverse TRF to outperform and *, **, *** indicate statistical significance at confidence levels of 10%, 5% and 1%, respectively, based on a two-sided Diebold-Mariano test.

Next, results for the comparison between reverse TRF and standard TRF are given in Table 10 showing that the standard RF performs better in most cases with some results significant to even the 1% level, based on a two-sided DM test. This means that one can conclude that the standard TRF is preferable over its alternatives in TGB and reverse TRF. What causes the weak performance of reverse TRF is not entirely clear. One reason could be as stated before that RF is not doing a good job at selecting variable subsets for forecasting with previous results also being somewhat evident of this finding. Another could be that RF is simply a stronger forecasting model in this case, due to RF being able to detect non-linearities and interactions amongst features.

4.5 Covid-19 Recession

Finally, 2019 to 2021 are added to the dataset in order to analyse the effects of targeting in highly volatile environments, such as the Covid-19 crisis of 2020-2021. Table 13 in Appendix A shows MSE ratios of the standard TRF set-up with an expanding window over the sample 1970-2021. Generally speaking, similar results are found compared to Table 2 except for the RF_s/RF model, which finds improvements across all targets and horizons when the Covid-19 recession is included. As we wish to analyse the forecasting performance during the recession, forecast graphs for the three horizons are plotted in Figure 1 over the period January 2016 to December 2021. Notice the enormous peaks around 2020 February and May of the actual industrial production growth (blue line), which have been the hardest points to forecast, leading to large forecast errors. Furthermore, much more volatile movement can be seen in general during the 2020-2021 period, compared to the previous years.

Table 11 presents MSE ratios of TRF vs full RF from 2020-2021. DM tests are not performed to assess significance in case TRF outperforms due to there only being 24 observations, causing the DM test to lose power (Diebold and Mariano, 2002). Overall, the gains in performance of



(a) One-month ahead (b) Three-month ahead (c) Twelve-month ahead

Figure 1: Forecast graphs including all point forecasts for every TRF with targeting degrees s = 5, 30 and 50, for each forecast horizon h = 1, 3 and 12 months ahead. The sample runs from 1970 to 2021 and the first forecast is made for 1985.

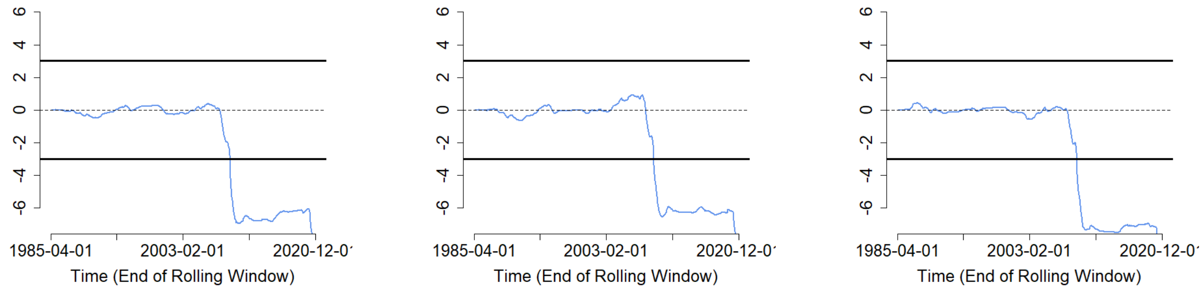
Panel: Industrial Production Growth									
	LASSO/RF			adaLASSO/RF			SCAD/RF		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	1.056	**0.789**	**0.543**	1.146	**0.849**	**0.377**	1.088	**0.799**	**0.471**
30	**0.961**	**0.915**	**0.614**	1.124	**0.943**	**0.620**	1.018	1.039	**0.754**
50	**0.953**	1.034	**0.767**	1.017	**0.933**	**0.829**	1.020	1.015	**0.828**
	RF/RF			GBM/RF					
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12			
5	**0.872**	**0.878**	**0.642**	**0.990**	1.008	**0.688**			
30	**0.933**	**0.925**	**0.788**	1.006	1.044	**0.927**			
50	**0.981**	**0.891**	**0.753**	**0.973**	1.015	**0.920**			

Table 11: This Table reports the MSE ratio between each TRF and full RF, for h = 1, 3 and 12 step ahead forecast horizons. The sample runs from 2020 January to 2021 December.

TRF do not change much for one- and three-step ahead forecasts, when isolating the sample to the Covid-19 recession This indicates that there are no additional gains from targeting in unstable environments and that the full RF is able to perform relatively well in these horizons. However, for the one-year ahead horizons, drastic improvements are found across all TRFs, in particular when targeting 5 variables. This indicates that the full RF is not able to discern the change in environment and is assigning importance to the wrong variables or ones that are no longer important during highly volatile periods. Which then leads to RF having difficulties with forecasting, causing large forecast errors. On the other hand, targeting seems to help significantly in volatile periods as they still appear to be useful in filtering out redundant predictors. This may also be the reason why a TRF with targeting degree of 5 sees such drastic improvements over a full RF, as almost all predictors are filtered out by then, leading to better predictions.

Looking deeper into it, consider the fluctuation test by Giacomini and Rossi (2010) plotted in Figure 2 for $adaLASSO_s/RF$ at the one-year ahead horizon. Additionally, fluctuation tests for all TRFs with different degrees of targeting and horizons are plotted in Appendix B. The horizontal black lines give the critical value bounds corresponding to $\mu = \frac{T_{OS}}{T}$ while the black line plots the test statistic over time as given in Equation 11. Finally, a negative test statistic indicates TRF performing better than full RF with significant outperformance at the 10% level at time t when the test statistic is outside the critical value bounds. From Figure 2, it seems that

$adaLASSO_s/RF$ and the full RF had equal performance until around 2008, which coincides with the great financial crisis, where the TRF significantly outperformed the full RF for the remainder of the sample. Another downward dive is seen around 2020, coinciding with the Covid-19 recession. The results point evidence towards targeting being especially helpful during volatile times. This pattern is generally followed in the fluctuation tests for the other TRFs at h = 12, where a downward spike is first seen during the 2008 recession followed by another downward spike during Covid-19.



(a) adaLASSO/RF, s = 5 (b) adaLASSO/RF, s = 30 (c) adaLASSO/RF, s = 50

Figure 2: Fluctuation test plots - adaLASSO/RF vs full RF at h = 12

For the one- and three-month forecasting horizons, similar movements can be observed where the performance appears mostly equal until a down or upward spike happens around the Covid-19 recession, depending on whether TRF or full RF outperformed the other, respectively. Unlike the one-year ahead fluctuation tests however, significant results are rarely found, indicating that the TRFs did not significantly outperform RF at any point in time and further supporting the notion that targeting is especially useful in longer horizon settings. Finally, changes in performance around the great recession of 2008 are not as evident for the one- and three-month ahead horizons, possibly suggesting again that the full RF's performance remains relatively equal to that of TRF during volatile periods.

4.6 Robustness Check

Lastly, to analyse the effect of the choice of initial window size, the standard TRF setting as in Section 4.2 is performed again, but with a window size of 360 observations (30 years). The MSE ratios are presented in Table 14 in Appendix A. In general, improvements from TRF appear more frequently, but with lower gains when increasing the window size. This could be the result of both targeting and the full RF forecasts improving as more data is processed in both steps. For the first, this leads to more optimally selected subsets of variables, which then leads to gains from targeting. For the second, this leads to lower forecast errors and improved MSE in both TRF and RF. Depending on the marginal gains in MSE for TRF and full RF when expanding the window, this can lead to both an increase in targeting improvement frequency, as well as a decrease in gain from targeting, which is the case here.

5 Conclusion

In this paper, we applied the methodology of Borup et al. (2023) and attempt to forecast industrial production growth over a forecasting horizon of one-, three- and twelve-months ahead. Datasets were gathered from the FRED-MD database (McCracken and Ng, 2016), which includes 127 monthly variables of macroeconomic relevance from 1970-2018. Contributions to the original findings of Borup et al. (2023) are made by extending and analyzing different possible variable selection methods, such as adaLASSO, SCAD, RF and GBM next to LASSO which was used in the original paper. Additionally, MCS are considered to find the best performing targeted models among a set of competing models. Our results suggest that targeted random forests generally are the best performing targeted forecasting model, compared to targeted gradient boosting and reverse TRF, making them non-viable alternatives to TRF. Furthermore, it seems that TRFs find larger gains from targeting by penalty-based methods, such as LASSO, adaLASSO and SCAD, suggesting that the tree-based methods in RF and GBM are not able to select the correct variables for forecasting. Finally, targeting sees the largest gains for the one-year ahead forecasting horizon.

The research question is originally formulated as: *'What is the best targeting method for variable selection that improves the predictability of industrial production growth using Targeted Random Forests?'* adaLASSO is found to be the best method for the initial targeting step for targeted random forests. Additionally, it is evident from the MCS results that the smallest amount of 5 selected variables is preferred for the one-year ahead horizon, while a target of 50 variables is preferred for one-month ahead horizons and a target of 30 variables is preferred for the middle horizon of 3 months. Finally, our methodology was applied to an extended dataset from 1970-2021, which includes the Covid-19 recession. Large forecast errors were found during the 2020-2021 period. Upon closer inspection, it is found that there are barely additional gains in performance of TRF over the full RF for one- and three-month ahead horizons. However, drastic improvements from targeting are found for the one-year ahead horizon, particularly when targeting 5 variables, indicating that RF possibly has difficulties forecasting in volatile periods, where relevant variables and interactions have likely changed.

For further research, one could analyse the frequency at which variables are selected and how this changes over time to gain an insight on the relevance of certain variables over time or how interactions may have developed. Alternatively, Shapley values like in Buckmann et al. (2022) could be considered to gain an insight on the variable importance of forecasts and to explain the predictions made by random forest. As well as to interpret non-linearities of economic relevance found by the model, something that is often disregarded due to the black box nature of machine learning models. One could also consider comparing TRF performances for different countries, such as the UK, Canada or Japan. Another possible extension to consider is that of extending the dataset with non-linearities, such as squared and absolute terms of observed values, and lags. The caveat is that this can increase the number of variables by many magnitudes, resulting in a more time-consuming TRF procedure. To that end, one could consider sure independance screening by Fan and Lv (2008), which specializes in variable selection in ultra-high dimensional settings.

References

Altelbany, Shady (2021). 'Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: a simulation study'. In: *Journal of Applied Economics and Business Studies* 5.1, pp. 131–142.

Archer, Kellie J and Ryan V Kimes (2008). 'Empirical characterization of random forest variable importance measures'. In: *Computational statistics & data analysis* 52.4, pp. 2249–2260.

Basher, Syed Abul and Perry Sadorsky (2022). 'Forecasting Bitcoin price direction with random forests: How important are interest rates, inflation, and market volatility?' In: *Machine Learning with Applications* 9, p. 100355.

Biau, Gérard and Erwan Scornet (2016). 'A random forest guided tour'. In: *Test* 25, pp. 197–227.

Borup, Daniel and Erik Christian Montes Schütte (2022). 'In search of a job: Forecasting employment growth using Google Trends'. In: *Journal of Business & Economic Statistics* 40.1, pp. 186–200.

Borup, Daniel et al. (2023). 'Targeting predictors in random forest regression'. In: *International Journal of Forecasting* 39.2, pp. 841–868.

Breiman, Leo (2017). *Classification and regression trees.* Routledge.

— (2001). 'Random forests'. In: *Machine learning* 45, pp. 5–32.

Buckmann, Marcus, Andreas Joseph and Helena Robertson (2022). *An interpretable machine learning workflow with an application to economic forecasting.* Tech. rep. Bank of England.

Charles, Amélie and Olivier Darné (2022). 'Forecasting Macroeconomic Time Series Using Sparse Random Forest Models'. In: *Available at SSRN 4111995.*

Chen, Tianqi and Carlos Guestrin (2016). 'XGBoost: A Scalable Tree Boosting System'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '16. San Francisco, California, USA: ACM, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

Coulombe, Philippe Goulet, Massimiliano Marcellino and Dalibor Stevanović (2021). 'Can machine learning catch the Covid-19 recession?' In: *National Institute Economic Review* 256, pp. 71–109.

Diebold, Francis and Robert S Mariano (2002). 'Comparing predictive accuracy'. In: *Journal of Business & economic statistics* 20.1, pp. 134–144.

Fan, Jianqing and Runze Li (2001). 'Variable selection via nonconcave penalized likelihood and its oracle properties'. In: *Journal of the American statistical Association* 96.456, pp. 1348–1360.

Fan, Jianqing and Jinchi Lv (2008). 'Sure independence screening for ultrahigh dimensional feature space'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5, pp. 849–911.

Friedman, Jerome (2001). 'Greedy function approximation: a gradient boosting machine'. In: *Annals of statistics*, pp. 1189–1232.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani (2010). 'Regularization paths for generalized linear models via coordinate descent'. In: *Journal of statistical software* 33.1, p. 1.

Giacomini, Raffaella and Barbara Rossi (2010). ‘Forecast comparisons in unstable environments’. In: *Journal of Applied Econometrics* 25.4, pp. 595–620.

Hansen, Peter, Asger Lunde and James M Nason (2011). ‘The model confidence set’. In: *Econometrica* 79.2, pp. 453–497.

Luong, Chuong and Nikolai Dokuchaev (2018). ‘Forecasting of realised volatility with the random forests algorithm’. In: *Journal of Risk and Financial Management* 11.4, p. 61.

McCracken, Michael and Serena Ng (2016). ‘FRED-MD: A monthly database for macroeconomic research’. In: *Journal of Business & Economic Statistics* 34.4, pp. 574–589.

McDonald, Gary (2009). ‘Ridge regression’. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.1, pp. 93–100.

Medeiros, Marcelo C et al. (2021). ‘Forecasting inflation in a data-rich environment: the benefits of machine learning methods’. In: *Journal of Business & Economic Statistics* 39.1, pp. 98–119.

Mentch, Lucas and Siyu Zhou (2022). ‘Getting better from worse: Augmented bagging and a cautionary tale of variable importance’. In: *Journal of Machine Learning Research* 23.224, pp. 1–32.

Tibshirani, Robert (1996). ‘Regression shrinkage and selection via the lasso’. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Zhang, Haozhe, Dan Nettleton and Zhengyuan Zhu (2019). ‘Regression-enhanced random forests’. In: *arXiv preprint arXiv:1904.10416*.

Zou, Hui (2006). ‘The adaptive lasso and its oracle properties’. In: *Journal of the American statistical association* 101.476, pp. 1418–1429.

A Appendix - Additional Tables

s	h = 1		
s	$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
5	{LAS, SCAD, RF, Full, GB}	{LAS, SCAD, RF, Full}	{LAS}
30	{LAS, SCAD, Full, ada, GB, RF}	{LAS, SCAD, Full, ada}	{LAS, SCAD, Full, ada}
50	{RF, SCAD, Full, ada, GB, LAS}	{RF}	{RF}
	h = 3		
s	$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
5	{Full, LAS, SCAD}	{Full, LAS, SCAD}	{Full}
30	{GB, ada, Full, LAS}	{GB, ada, Full, LAS}	{GB}
50	{RF, SCAD, LAS, ada, Full, GB}	{RF, SCAD, LAS, ada, Full, GB}	{RF}
	h = 12		
s	$1-\alpha=0.7$	$1-\alpha=0.4$	$1-\alpha=0.1$
5	{ada, RF, GB, Full, LAS, SCAD}	{ada, RF, GB, Full, LAS}	{ada}
30	{ada}	{ada}	{ada}
50	{ada, SCAD, GB, LAS, Full, RF}	{ada, SCAD, GB, LAS, Full}	{SCAD}

Table 12: Model Confidence set for all versions of TGB and full GB at confidence levels of $1-\alpha$ = 0.7, 0.4 and 0.1. The Table results give the SSM at each forecast horizon h, each selected variable target amount s and each level of $1-\alpha$. The models in SSM are ranked from best to worst, from left to right.(LASSO has been shortened to LAS and adaLASSO to ada)

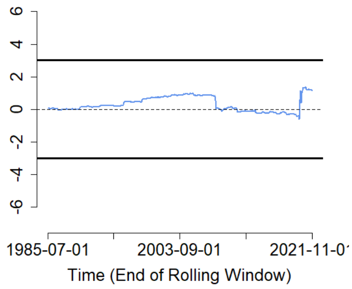
Panel: Industrial Production Growth									
	LASSO/RF			adaLASSO/RF			SCAD/RF		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	1.052	**0.848***	**0.890****	1.130	**0.923**	**0.680****	1.070	**0.848***	**0.878****
30	**0.971**	**0.942***	**0.865****	1.095	**0.928***	**0.813****	1.022	1.030	**0.841****
50	**0.970**	1.015	**0.884****	1.018	**0.937***	**0.841****	1.026	**0.997**	**0.866****
	RF/RF			GBM/RF					
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12			
5	**0.934**	**0.988**	**0.934**	1.035	1.040	**0.931**			
30	**0.957***	**0.962**	**0.994**	1.012	1.026	1.004			
50	**0.992**	**0.932**	**0.991**	**0.984**	1.008	0.994			

Table 13: This Table reports the MSE ratio between each TRF and full RF, for h = 1, 3 and 12 step ahead forecast horizons. The sample runs from 1970 January to 2021 December. Bold numbers indicate improvements from targeting and *, **, *** indicate statistical significance at confidence levels of 10%, 5% and 1%, respectively, based on an one-sided Diebold-Mariano test.

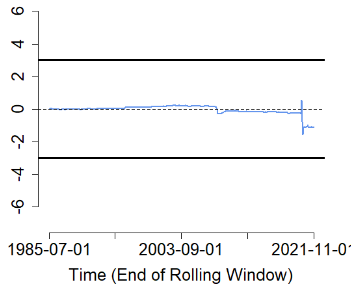
Panel: Industrial Production Growth									
	LASSO/RF			adaLASSO/RF			SCAD/RF		
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
5	1.017	**0.857**	**0.930**	**0.916**	**0.841**	**0.910**	**0.964**	**0.865**	**0.905**
30	**0.960**	**0.911***	**0.990**	1.090	**0.929***	**0.874**	**0.998**	**0.884****	**0.898****
50	1.040	**0.965***	**0.970**	1.034	**0.922***	**0.898***	1.050	**0.925***	**0.951***
	RF/RF			GBM/RF					
s	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12			
5	**0.854**	**0.897**	**0.897**	**0.940**	**0.954**	1.099			
30	**0.981**	**0.909**	1.056	**0.987**	**0.994**	1.091			
50	**0.959**	**0.872**	**0.989**	**0.974**	**0.985**	1.052			

Table 14: This Table reports the MSE ratio between each TRF and full RF, for h = 1, 3 and 12 step ahead forecast horizons. The sample runs from 1970 January to 2018 December **over a rolling window with initial size of 30 years**. Bold numbers indicate improvements from targeting and *, **, *** indicate statistical significance at confidence levels of 10%, 5% and 1%, respectively, based on an one-sided Diebold-Mariano test.

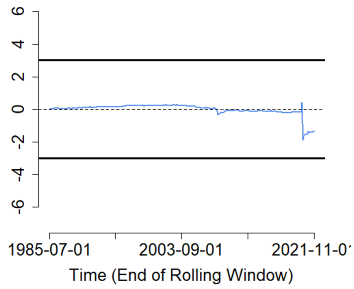
B Appendix - Fluctuation tests



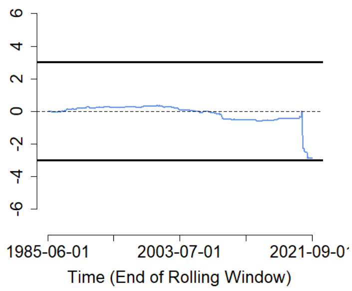
(a) LASSO/RF, s = 5, h = 1



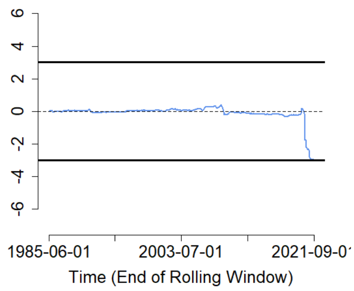
(b) LASSO/RF, s = 30, h = 1



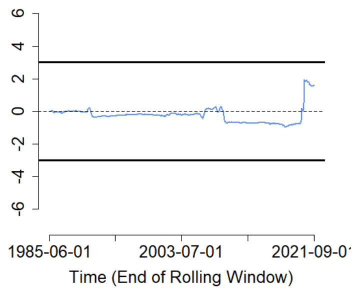
(c) LASSO/RF, s = 50, h = 1



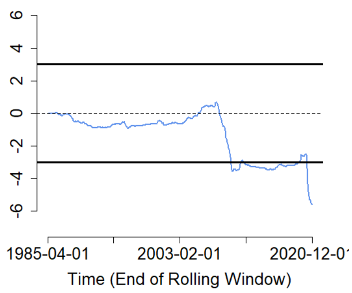
(d) LASSO/RF, s = 5, h = 3



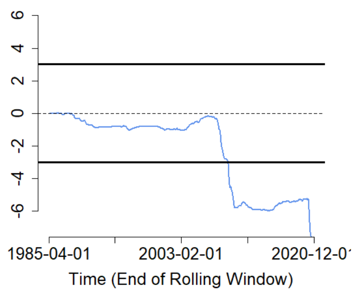
(e) LASSO/RF, s = 30, h = 3



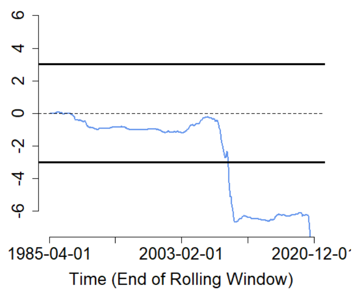
(f) LASSO/RF, s = 50, h = 3



(g) LASSO/RF, s = 5, h = 12

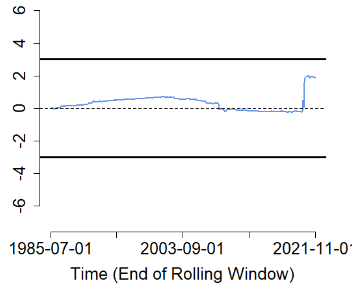


(h) LASSO/RF, s = 30, h = 12

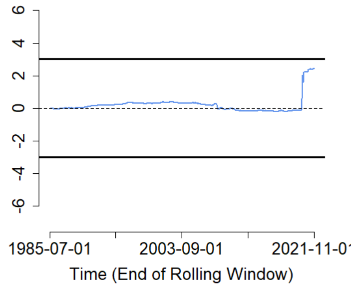


(i) LASSO/RF, s = 50, h = 12

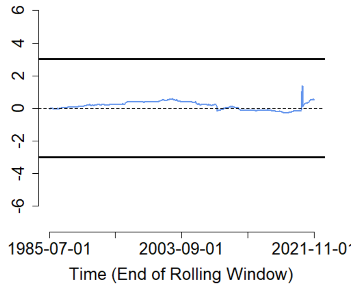
Figure 3: Fluctuation test plots - LASSO/RF vs full RF



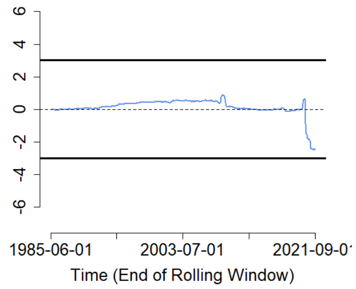
(a) adaLASSO/RF, s = 5, h = 1



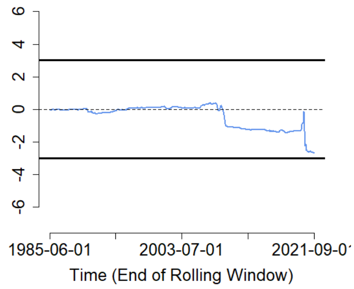
(b) adaLASSO/RF, s = 30, h = 1



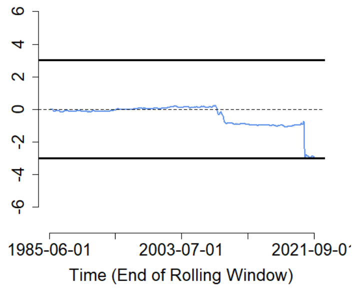
(c) adaLASSO/RF, s = 50, h = 1



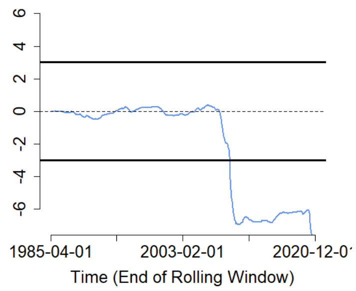
(d) adaLASSO/RF, s = 5, h = 3



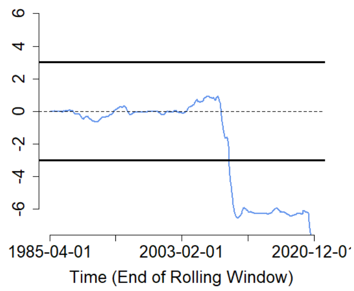
(e) adaLASSO/RF, s = 30, h = 3



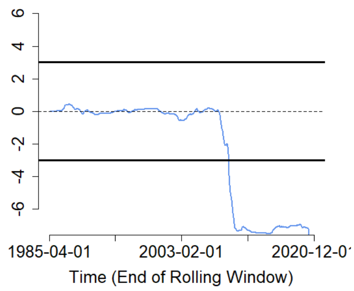
(f) adaLASSO/RF, s = 50, h = 3



(g) adaLASSO/RF, s = 5, h = 12

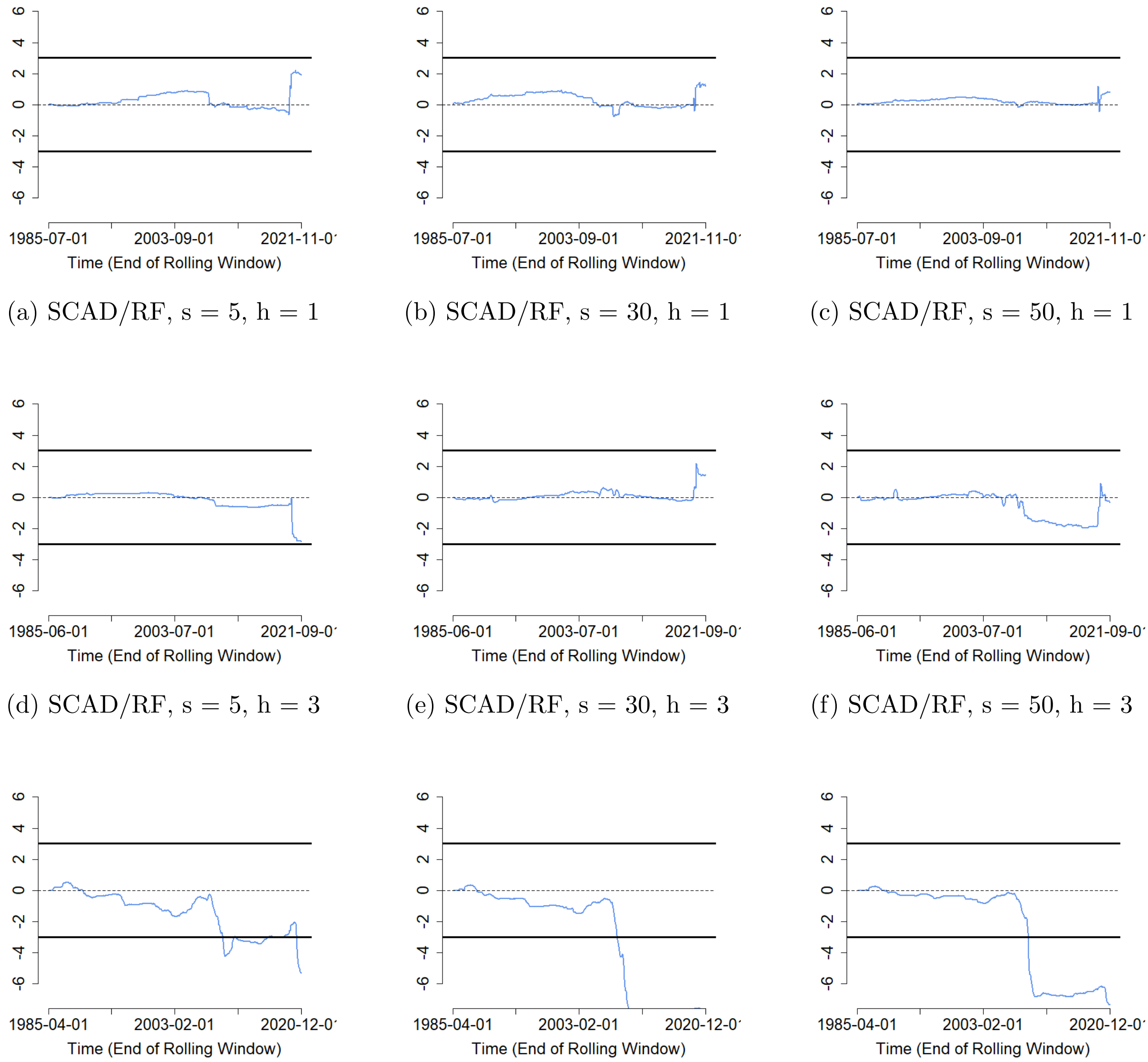


(h) adaLASSO/RF, s = 30, h = 12



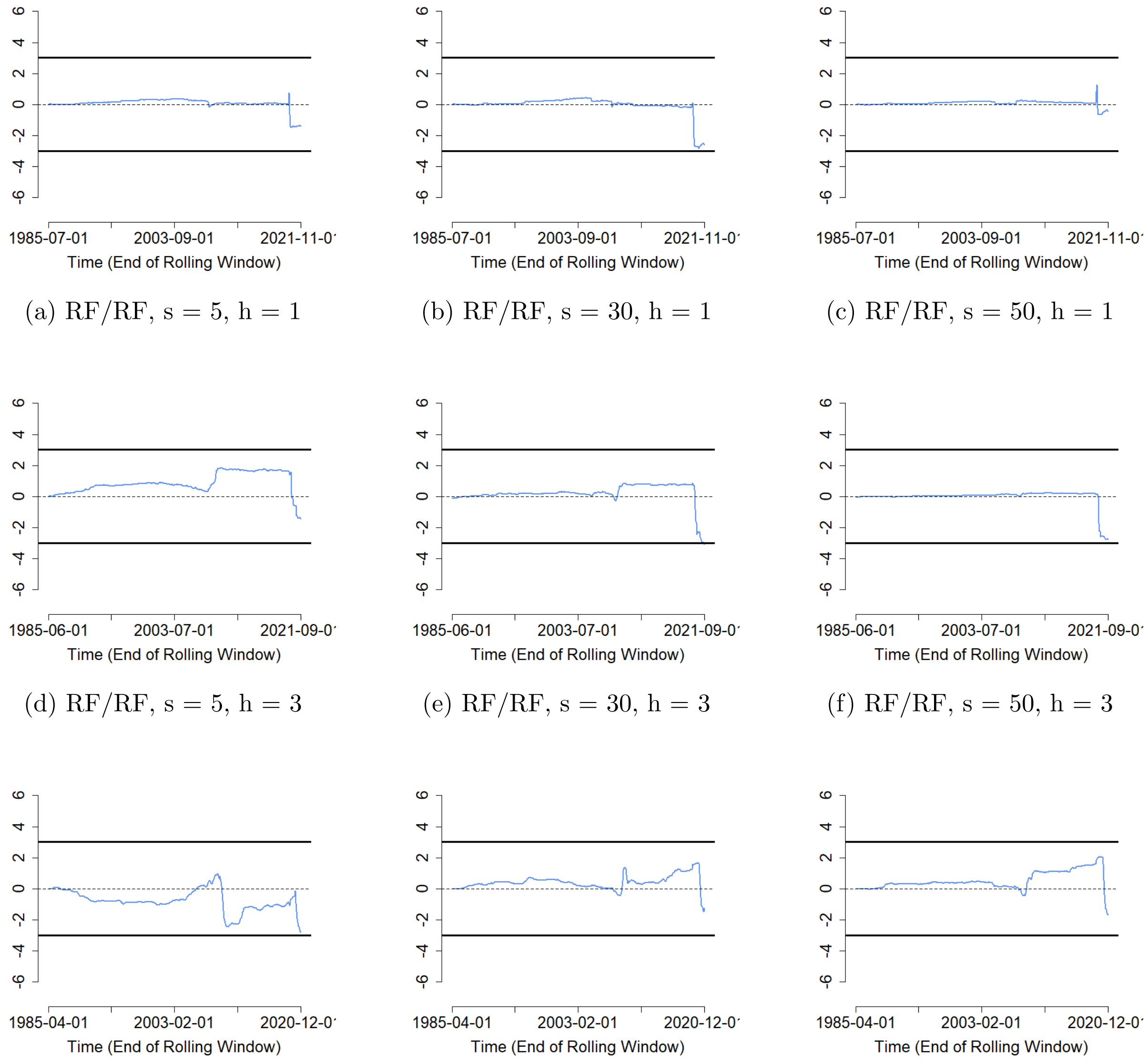
(i) adaLASSO/RF, s = 50, h = 12

Figure 4: Fluctuation test plots - adaLASSO/RF vs full RF



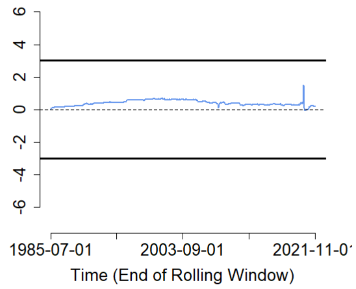
(a) SCAD/RF, s = 5, h = 1
(b) SCAD/RF, s = 30, h = 1
(c) SCAD/RF, s = 50, h = 1
(d) SCAD/RF, s = 5, h = 3
(e) SCAD/RF, s = 30, h = 3
(f) SCAD/RF, s = 50, h = 3
(g) SCAD/RF, s = 5, h = 12
(h) SCAD/RF, s = 30, h = 12
(i) SCAD/RF, s = 50, h = 12

Figure 5: Fluctuation test plots - SCAD/RF vs full RF

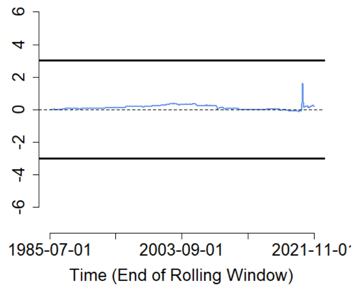


(a) RF/RF, s = 5, h = 1
(b) RF/RF, s = 30, h = 1
(c) RF/RF, s = 50, h = 1
(d) RF/RF, s = 5, h = 3
(e) RF/RF, s = 30, h = 3
(f) RF/RF, s = 50, h = 3
(g) RF/RF, s = 5, h = 12
(h) RF/RF, s = 30, h = 12
(i) RF/RF, s = 50, h = 12

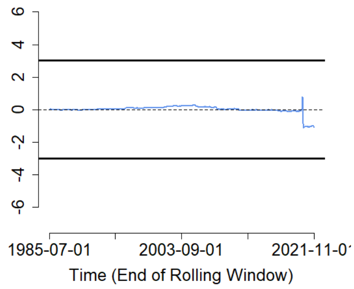
Figure 6: Fluctuation test plots - RF/RF vs full RF



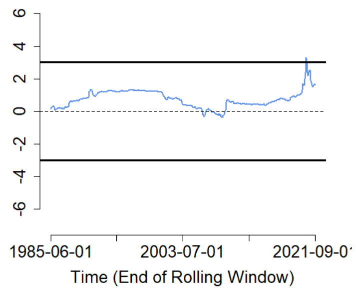
(a) GB/RF, s = 5, h = 1



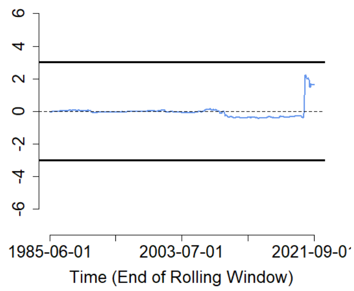
(b) GB/RF, s = 30, h = 1



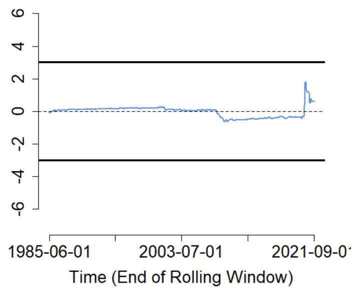
(c) GB/RF, s = 50, h = 1



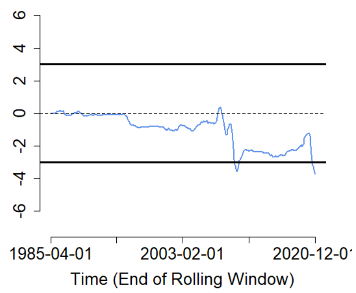
(d) GB/RF, s = 5, h = 3



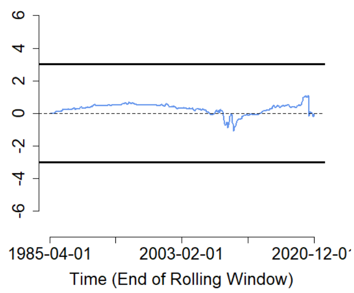
(e) GB/RF, s = 30, h = 3



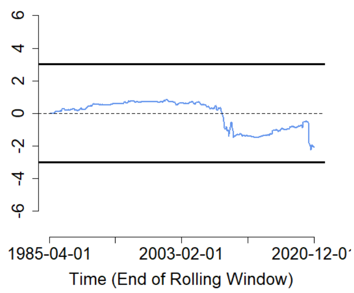
(f) GB/RF, s = 50, h = 3



(g) GB/RF, s = 5, h = 12



(h) GB/RF, s = 30, h = 12



(i) GB/RF, s = 50, h = 12

Figure 7: Fluctuation test plots - GB/RF vs full RF