ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus School of Economics

BACHELOR THESIS IN QUANTITATIVE FINANCE

Analyzing the Performance of Tree-Based Machine Learning Methods in Forecasting Inflation

Author
Julia van Lent
Student ID number
507202

Supervisor
Dr. E.P. (Eoghan) O'Neill
Second Assessor
Dr. A.A. (Andrea) Naghi

August 7, 2021

Abstract

Inflation is one of the most important indicators of economic activity. An adequate forecasting model, however, still remains to be found. This paper extends the work of Medeiros, Vasconcelos, Veiga, et al. (2021) and studies the performance of tree-based ML methods in forecasting US inflation using high-dimensional data. The data set used consists of a large number of macroeconomic variables obtained from the FRED-MD database. We compare and contrast the performance of the standard RF to honest RFs, local linear forests (LLFs) and macroeconomic random forests (MRFs). Furthermore, we investigate whether our results hold for two smaller out-of-sample periods. The comparison is based on the RMSE, MAE and MAD in combination with Diebold-Mariano tests and Model Confidence Sets. Our results suggest that the forecasts obtained by the LLF are similarly accurate to the RF forecasts. The MRF, on the contrary, provides less accurate results. These results also hold for the two alternative samples investigated. The honest RFs only provide forecasts that are similarly accurate to the RF for one of the subsamples. Even though the tree-based methods considered in this paper do not outperform the standard RF, their ability to capture complex interactions of high-dimensional data sets makes them a promising approach to further improve prediction accuracy for forecasting inflation.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics. or Erasmus University Rotterdam.

Contents

1 Introduction

Finding an adequate model for forecasting inflation is a recurrent problem in macroeconomics. Inflation forecasting is especially key for policy makers and economic agents as inflation is an important measure for economic performance. For central banks, for example, being able to predict future inflation is crucial for adjusting their monetary policy in order to control inflation. This high demand for accurate inflation forecasts, has led many economists to search for an accurate model. Stock and Watson (1999), for example, have investigated the predictive performance of the Philips curve, Aron and Muellbauer (2013) have evaluated various vector autoregressive (VAR) models and Banerjee et al. (2014) have used factor-augmented error correction models to forecast inflation. However, due to the complex dynamics of inflation, a universally reliable model still remains to be found.

As Medeiros, Vasconcelos, Veiga, et al. (2021) have shown that the random forest (RF) model outperforms many other machine learning (ML) models in forecasting inflation, we investigate whether an adjustment to this model leads to improvements in forecasting accuracy. More specifically, we focus on tree-based methods which were not used in Medeiros, Vasconcelos, Veiga, et al. (2021), but have recently received increasing attention in the field of economics. We compare and contrast the predictive performance of three different extensions of the RF model, namely honest RFs, local linear forests (LLFs), as proposed by Friedberg et al. (2020), and macroeconomic random forests (MRFs), as proposed by Goulet Coulombe (2020).

As stated by Araujo and Gaglianone (2020), the literature on the use of ML methods for macroeconomic forecasting is relatively new and rather limited. While the MRF has already been considered in the context of forecasting inflation, the LLF has not yet been used in this context to our knowledge. Therefore, this paper adds to existing literature by broadening the range of ML methods considered for forecasting inflation. We also extend the existing research concerning the MRF for forecasting inflation by using a substantially larger data set.

The main research question this paper considers is whether the LLFs and the MRFs can outperform the standard RF approach in forecasting inflation. Furthermore, we investigate whether the obtained results hold for different sample periods. We use a high-dimensional data set that consists of 122 monthly macroeconomic variables, originating from the FRED-MD database presented by McCracken and Ng (2016). In order to compare the different models, we compute the root mean squared error (RMSE), the mean absolute error (MAE) and the mean absolute deviation (MAD) for various forecasting horizons. Furthermore, we perform Diebold-Mariano (DM) tests and construct the Model Confidence Sets (MCSs) as described by Hansen et al. (2011).

Our results suggest that the LLF performs similarly to the standard RF model, but is slightly less accurate. The forecasts obtained from MRFs are not as accurate as the RF forecasts. However, as a result of their ability to forecast using large data sets and both the LLF and the MRF appear to be a promising approach for forecasting inflation using a large number of macroeconomic covariates. These results also hold for the two subsamples we have investigated.

The outline of the paper is as follows. In Section 2 we present a literature review, followed by a description of our data set in Section 3. Section 4 provides an extensive description of the forecasting models and the comparison methods. In Section 5, we present our results and we terminate our paper with a conclusion and discussion of our results in Section 6.

2 Literature

As recently highlighted by Goulet Coulombe et al. (2020), ML methods are very useful for macroeconomic forecasting due to their ability to capture important nonlinearities. While Ang et al. (2005) state that the four main methods to forecast inflation are time-series models, structural models, asset price models and methods using survey-based measures, some promising recent advances have been made in the area of forecasting inflation using ML methods. Ülke et al. (2018), for example, show that ML methods beat various time series models in prediction accuracy when forecasting the core personal consumption expenditure (PCE) inflation. In their analysis, they use neural networks (NN), artificial neural networks (ANN) and support vector machines (SVM). Furthermore, Medeiros, Vasconcelos, and Freitas (2016) and Garcia et al. (2017) both show that the least absolute shrinkage and selection operator (LASSO) model performs well for forecasting Brazilian inflation. Cheng et al. (2021) uses ML in order to aggregate individual forecasts for improving US inflation predictions.

A key advantage of ML methods is their ability to capture nonlinearities in large data sets. Many methods have been proposed to capture nonlinearities in data, such as threshold autoregressive models (Hansen, 2011), smooth transition models (Teräsvirta, 1994) and random-walk time-varying parameters (Primiceri, 2005). However, most of the proposed methods only perform well for little data. As stated by Fortin-Gagnon et al. (2018), the use of regression trees is one of the most promising approaches to introduce nonlinearity in the predictive equation in a data-rich macroeconomic context.

One of the most popular ML methods for regression and classification problems, is the RF approach as proposed by Breiman (2001). This tree-based method randomizes both the training set and the splitting directions. RFs have been investigated extensively over the last decade, mainly due to their ability to perform well for high dimensional data and their good predictive performance. Additionally, RFs are convenient to construct as they only depend on few tuning parameters which can be configured by sensible heuristics (Scornet, 2017 and Liaw and Wiener, 2001). Olson and Wyner (2018) have shown that with a sufficient number of randomly sampled parameters available for splitting at each node, the RF model is able to adapt well to sparsity and can successfully eliminate useless predictors. This is in alignment with Friedman et al. (2001), who state that RFs are relatively immune to the effects of including a large number of irrelevant features. Moreover, Chen et al. (2019) and Goulet Coulombe et al. (2020) have shown that the model performs well for macro data in general. However, most of the the theoretical papers require assumptions that restrict the number of covariates relative to the number of observations. The paper by Medeiros, Vasconcelos, Veiga, et al. (2021), who have proven RFs to be very successful in forecasting US inflation, does include a large number of covariates.

One of the key weaknesses of RFs, however, is their inability to capture smoothness. Also Zhang, Nettleton, et al. (2019) note several drawbacks to the use of RFs. They state that as RFs are a fully nonparametric predictive algorithm, they lack in incorporating known relations between the response and the predictor variables. Another shortcoming of the RF model is their inability to extrapolate. RFs may be unable to accurately predict if predictions are required at values that fall outside the range of the training data, as the RF predictions are an average of the previously observed values.

Since the introduction of the RF model, there have been many methodological and theoretical advances in this approach, such as quantile regression forests (Meinshausen, 2006), multivariate random forests (Segal and Xiao, 2011), extremely randomised regression forests (Geurts et al., 2006) and regression-enhanced s (Zhang, Nettleton, et al., 2019).

Another extension of the RF is the LLF model, which elaborates on the view of RFs as an adaptive kernel method, as originally proposed by Breiman (2000). This link between RFs and kernel methods has been studied by many others. Hothorn et al. (2004), for example, have proposed using weights from survival trees resulting in interesting simulation results, and Athey et al. (2019) have suggested to use generalised RFs in order to solve heterogeneous estimating equations. Scornet (2016) provides empirical evidence that estimates obtained through kernels based on RFs show improvements over RF estimates. LLFs expand this literature by taking the kernel to estimate a linear model.

While local linear regressions can capture smoothness in low-dimensional models, they have shown to lose their functionality for high-dimensional data. RFs, on the contrary, perform particularly well in high-dimensional environments. Banfield et al. (2007) have shown that RFs provide accurate predictions for high dimensional data. Friedberg et al. (2020) argue that the combination of local linear regressions and RFs results in a method in which both the adaptivity of RFs and the smoothness capturing ability of linear regressions is combined. They find that local linear forests improves substantially over RFs in both predictive performance and accuracy.

Friedberg et al. (2020) state that honesty may increase predictive performance when working with large sample sizes and weak signals due to its ability to stabilise forests. This is further substantiated in Appendix B of Wager and Athey (2018), which states that the traditional tree construction approach may over-represent outliers in the corners of the predictor space when separating outliers from the data, resulting in a bias especially for large data sets. Honest prediction avoids this bias by splitting the predictor space into two samples. As macroeconomic data sets are often available in large sample sizes and are likely to contain weak signals due to their high dimensionality, honesty could be very well suited for this type of data. However, in practice, honesty can also worsen predictions. Denil et al. (2014) have shown this for a number of data sets from the UCI repository.

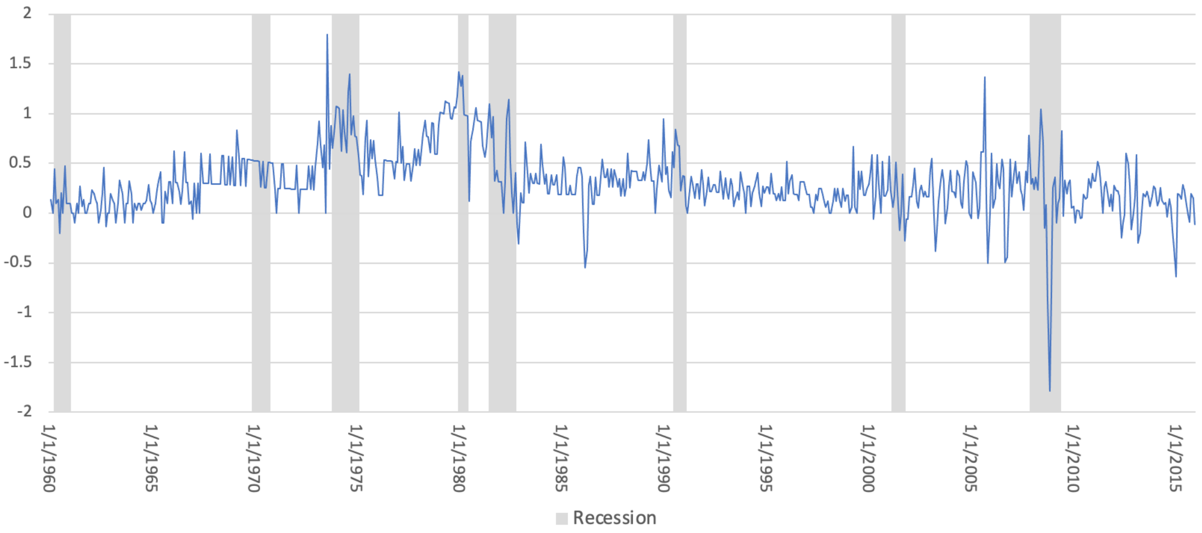
Another ML model that uses trees with a linear part is the MRF, as proposed by Goulet Coulombe (2020). This extension to the traditional RF that does not only provide gains in forecasting accuracy, but also has the benefit of delivering an output that can be easily interpreted. Goulet Coulombe has shown this method to outperform many alternative models and to perform particularly well for inflation. The key difference between the MRF and the RF is that the plain RF only contains an intercept in each leaf, while the MRF also includes a linear part. Other economic researchers have also considered incorporating linear parts in trees before, such as Alexander and Grimshaw (1996) and Wang and Witten (1996). However, these papers only use the linear part as a way to improve the efficiency of nonparametric estimation, whereas the linear part of the MRF shows to be much more meaningful when viewing the regression coefficients as a combination of nonlinear time series models.

3 Data

Similarly to the research of Medeiros, Vasconcelos, Veiga, et al. (2021), we use the monthly variables from the FRED-MD database as presented by McCracken and Ng (2016). This data set is updated in real-time by means of the FRED database and can also be obtained from McCracken's website.[1]

From this data set, only the variables that have no missing observations have been used and all variables have been treated to obtain stationarity using the transformations as described in Medeiros, Vasconcelos, Veiga, et al. (2021). The used data set ranges from January 1960 to December 2015, of which only the variables without missing observations have been taken into consideration (122 variables). Moreover, this data set is extended by including the corresponding four principal component factors, four lags of each variable and four autoregressive terms. Therefore, the resulting data set includes 508 variables with 671 observations. A more detailed explanation of the used variables and their corresponding transformations can be found in Tables S.1-S.8 of the Supplementary Material of Medeiros, Vasconcelos, Veiga, et al. (2021). The out-of-sample period extends from January 1990 to December 2015. The inflation in month t, denoted by π_t, is computed as $\pi_t = \log(P_t) - \log(P_{t-1})$, where P_t denotes the CPI price index in period t. The course of the CPI monthly inflation for the sample of January 1960 to December 2015 is given in Figure 1. Noteworthy are the high peak in August 1973, reaching a value of 1.79%, which DeLong (1996) states to be a result of the memory of the Great Depression. This memory caused the left and centre to be strongly averse to risking unemployment and eliminated any mandate the Federal Reserve had for controlling inflation that formed a risk to the unemployment rate. Without this interference of the Federal Reserve, the inflation reached very high levels during the 1970s. After this period, however, the Federal Reserve regained its mandate to control inflation by risking unemployment. Figure 1 shows that the inflation slowly declines again, with the exception of the peak during the recession of 1982. The most prominent outlier we observe, is the is the sharp decline during the financial crisis of 2008.

Figure 1. Monthly CPI inflation in percentages from January 1960 to 2015 December.



[1] https://research.stlouisfed.org/econ/mccracken/fred-databases/

Similar to the methods described by McCracken and Ng (2016), we treat the outlier of November 2008 by including a dummy variable for November 2008 in all models estimated after that date. From the graph in Figure 1 it is evident that the inflation volatility during the years 1990 to 2000 is much lower than during the years 2001 to 2015. This is supported by the standard deviation, which is equal to 0.17% during the first period and 0.32% during the second period. With this dummy, our data contains a mixture of continuous and categorical variables, for which tree-based methods perform well. Following McCracken and Ng (2016), the 122 variables used can be divided into eight groups, each containing variables that are closely related: (1) output and income; (2) labor market; (3) housing; (4) consumption, orders and inventories; (5) money and credit; (6) bond and exchange rates; (7) prices; and (8) stock market. As many of the variables are closely related, local regression adjustments might be useful as stated by Friedberg et al. (2020).

4 Methodology

Following Medeiros, Vasconcelos, Veiga, et al. (2021), we consider the following model for the relation between inflation and other macroeconomic variables:

$$\pi_{t+h} = G_h(\boldsymbol{x}_t) + u_t, \text{ for } h = 1, ..., H, \ t = 1, ..., T. \tag{1}$$

Here, π_{t+h} denotes the inflation in month $t+h$, $\boldsymbol{x}_t = (x_{1t}, ..., x_{nt})'$ an $n \times 1$ vector of independent variables and $G_h(\cdot)$ denotes the function describing the relation between the predictors and the future inflation. Furthermore, u_{t+h} denotes an error term with a mean of zero. For each forecast horizon h, the mapping $G_h(\cdot)$ differs. Moreover, the function $G_h(\boldsymbol{x}_t)$ can be either a single model or an ensemble of different models. The forecasts are computed using a rolling-window of fixed length. The inflation forecasts are specified by

$$\hat{\pi}_{t+h} = \hat{G}_{h,t-R_h+1:t}(\boldsymbol{x}_t), \tag{2}$$

where $\hat{G}_{h,t-R_h+1:t}$ denotes the target function that is estimated based on observations from time $t - R_h + 1$ to time t. The variable R_h denotes the size of the window, which depends on the forecasting horizon and the number of lagged variables that are included in the model.

For all methods, the sample is split into two subsets. The first subset, which doesn't include a dummy variable for the observation of November 2008, ranges from January 1960 till December 2000. From this subset, predictions are made for the sample 1990-2001, using a rolling-window of size $R_h = 360 - h - p - 1$, where p denotes the number of lags in the model. The second subset ranges from January 1960 till December 2015 and predicts inflation for the sample 2001-2015, using a rolling-window of size $R_h = 492 - h - p - 1$.

As a part of this paper, a replication is done of the results of Medeiros, Vasconcelos, Veiga, et al. (2021). Of the three benchmark models that they use, we investigate the random walk (RW) and the autoregressive (AR) model. Furthermore, the shrinkage methods investigated in this paper are the ridge regression (RR), the least absolute shrinkage and selection operator (LASSO), the adaptive LASSO (adaLASSO), the elastic net (ElNet) and the adaptive ElNet (adaElNet).

The factor models considered, in addition to the traditional factor model, are models that use target factors (T. Factor) and boosted factors (B. Factor). Moreover, we evaluate the averaging ensemble methods bagging and complete subset regressions (CSR). Lastly, we investigate the performance of the RF, the honest RF, three types of the LLF, and the MRF. For each model, we consider twelve forecasting horizons, taking $h = 1, ..., 12$, and compute the 3-, 6- and 12-month ahead accumulated forecasts. In order to compute the accumulated inflation forecast over the following h-months, the individual forecasts are aggregated similarly for all models except the random walk. All models and methods have been implemented using R.

4.1 Benchmark models

The two benchmark models we investigate are the RW and the AR model. The random walk forecasts are simply computed as $\hat{\pi}_{t+h|t} = \pi_t$, where h denotes the forecasting horizon. The AR model takes the inflation forecasts to depend on a linear combination of the previous inflation values and a stochastic term and is often used as a benchmark model when forecasting inflation. (Stock and Watson, 2008, Hanif and Malik, 2015 and Zhang, Chan, et al., 2020). The order p of the AR model is determined by evaluating the Bayesian Information Criterion, after which the model is estimated by ordinary least squares (OLS). A more detailed description of the benchmark models can be found in Appendix B.2.

4.2 Shrinkage models

Following the approach of Medeiros, Vasconcelos, Veiga, et al. (2021), five different shrinkage models have been considered. The first is the ridge regression (RR), as proposed by Hoerl and Kennard (1970). The second and third models are the least absolute shrinkage and selection operator (LASSO), which was introduced by Tibshirani (1996), and the adaptive LASSO (adaLASSO), which was proposed by Zou (2006). The adaLASSO model was introduced with the aim of achieving consistency in model selection, wherefore it includes a weighting parameter originating from a first-step estimation. The fourth shrinkage model is the ElNet approach, which was introduced by Zou and Hastie (2005) and is a combination of the LASSO and the RR. Lastly, we consider the adaElNet method. A more detailed description of the benchmark models can be found in Appendix B.3.

4.3 Factor models

In addition to the traditional factor model approach, as thoroughly explained in Bai (2003), two adapted factor models have been considered in this paper. The first approach was introduced by Bai and Ng (2008) and uses a subset of the predictors used in the traditional factor model approach, such that only the variables with high prediction power are used for forecasting. An extensive description of the procedure can be found in Medeiros and Vasconcelos (2016), who have shown that the use of target factors slightly increases the forecasting performance of CPI inflation relative to the standard factor model. The second adaptation of the factor model, aims at finding an optimal selection of factors for the predictive regression. We follow the procedure proposed by Bai and Ng (2008), which uses a boosting algorithm to select both the factors and the number of lags and in the model. A detailed description of the factor models can be found in Appendix B.4.

4.4 Ensemble methods

The ensemble methods that we used are bagging and CSR. The bagging approach was introduced by Breiman (1996) and combines predictions from models in different bootstrap samples. The CSR approach was proposed by Elliott et al. (2013) and Elliott et al. (2015) with the aim of finding a more efficient way to select an optimal subset of the regressors. A more detailed description of these methods can be found in Appendix B.5.

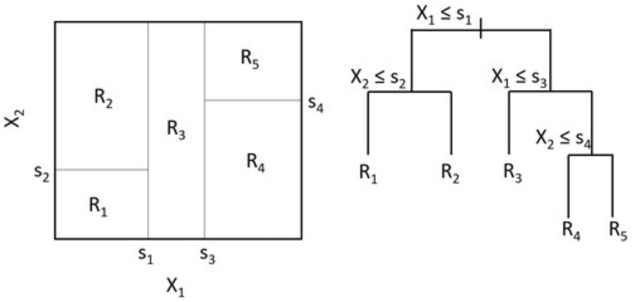
4.5 Random forests

The RF model has first been proposed by Breiman (2001) as an extension on his prior work on bagging (Breiman, 1996). For this tree-based method, all trees depend on a collection of random variables. That is, the forest is grown by combining many randomly constructed regression trees that form an ensemble, a process called bootstrap aggregating (bagging). This process reduces variance and also helps to avoid overfitting. The regression trees used are nonparametric models constructed by a binary recursive partitioning of the predictor space that approximate an unknown nonlinear function.

In order to understand how the regression tree works, we refer to the example of Hastie et al. (2001) as reproduced and adjusted by Medeiros, Vasconcelos, Veiga, et al. (2021), which is shown in Figure 2. Here, a regression problem is displayed, where X_1 and X_2 are the predictor variables and the dependent variable is given by Y. In this example, the predictor space is split into two regression regions at $X_1 = s_1$, followed by a split at $X_2 = s_2$ in the region to the left of $X_1 = s_1$ and a split at $X_1 = s_3$ in the region to the right of $X_1 = s_1$. Lastly, a split at $X_2 = s_4$ divides the regression region to the right of $X_1 = s_3$ into two regions, resulting in a partitioning of five regions in total, denoted by R_k, for $k = 1, ..., 5$. The aim of the RF is to estimate a prediction function $f(x) = E[Y|X = x_t]$. In each of the k regions, the dependent variable Y is assumed to be predicted by the sample average of its realisations that are contained in region R_k, denoted by c_k.

This recursive binary partition can be represented as a tree, as shown in Figure 2. Here, the root node is made up of the entire predictor space and the leaves form the final partition of the predictor space.

Figure 2. Visualisation of a regression tree. Reproduction by Medeiros, Vasconcelos, Veiga, et al. (2021) of part of Figure 9.2 in Hastie et al. (2001).



The splits are determined by considering a random subset of m predictors and then selecting the split that minimizes the sum of squared residuals of the following regression model:

$$\pi_{t+h} = \sum_{k=1}^{K} c_k I_k(\boldsymbol{x}_t; \boldsymbol{\theta}_k), \tag{3}$$

where π_{t+h} denotes the dependent variable, $\boldsymbol{x}_t$ a set of predictor variables and K the number of terminal nodes. The indicator function $I_k(\boldsymbol{x}_t; \boldsymbol{\theta}_k)$ is defined as

$$I_k(\boldsymbol{x}_t; \boldsymbol{\theta}_k) = \begin{cases} 1 & \text{if } \boldsymbol{x}_t \in R_k(\boldsymbol{\theta}_k), \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\boldsymbol{\theta}_k$ denotes the set of parameters included in the k-th regression. The construction process of RFs contains two elements of randomness. As discussed in this section, the first element of randomness manifests itself in the splitting of the nodes, as a random subset of m predictor variables is considered, which is seperately drawn for each split. Secondly, the RF is constructed as a collection of regression trees that are all grown from an independent bootstrap sample of the original data set. We denote the number of bootstrap samples by B. Then for a random subset of the original covariates, a tree with K_b regions is estimated for each sample b, with $b = 1, ..., B$. Applying the forecasts of each tree to the original data and taking the average, gives the final forecast of the RF, which is specified as

$$\hat{\pi}_{t+h} = \frac{1}{B} \sum_{b=1}^{B} \left[\sum_{k=1}^{K_b} \hat{c}_k I_k(\boldsymbol{x}_t; \boldsymbol{\theta}_{k,b}) \right]. \tag{4}$$

For the RF, there are three tuning parameters that can be adjusted to improve its performance. These are the the number of randomly sampled predictor variables available for splitting at each node, denoted by m, the number of trees grown in the forest, B, and the tree size.

4.6 Local Linear Forests

The local linear forests (LLF) as proposed by Friedberg et al. (2020) are an extension on the RFs that adapts to smoothness, where the RFs are considered as an adaptive weight generator. This method, deviating from the traditional approach viewing RFs as an ensemble method, complements the preceding literature by Athey et al. (2019), Hothorn et al. (2004) and Meinshausen (2006). Suppose we have training data $(X_1, Y_1), ..., (X_n, Y_n)$, with X_i the set of predictor variables and Y_i the inflation realisations, and we want to estimate the prediction function $\pi_{t+h}(x_t) = E[Y|X = x_t]$. For each tree $b = 1, ..., B$ in a forest, we define L_b to be the set of training data that falls in the same leaf as x_t. Whereas we had for the RF that the average prediction was equal to $\hat{\pi}_{t+h} = \frac{1}{B}\sum_{b=1}^{B} \hat{\pi}_{b,t+h}$, this alternative approach uses RF as an adaptive weight generator. Instead of taking the average of predictions made by individual trees, we specify the RF predictions as

$$\hat{\pi}_{t+h} = \sum_{i=1}^{n} \alpha_i(x_t) Y_i, \tag{5}$$

in which the forest weight represents the frequency with which the i-th training data point falls into the same leaf as x_t. This weight $a_i(x_t)$ is equal to

$$\alpha_i(x_t) = \frac{1}{B} \sum_{b=1}^{B} \frac{I_b(X_i)}{|L_b(x_t)|}, \quad \text{where } \; I_b(X_i) = \begin{cases} 1 & \text{if } X_i \in L_b(x_t), \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Here, n denotes the number of observations in the training data and $I_b(X_i)$ is an indicator function. From this formulation follows that $0 \leq \alpha_i(x_t) \leq 1$. These weights sum to 1 when given that there exists a nonempty cell containing x and are all equal to zero otherwise. A derivation of Equation 5 can be found in Appendix B.6.1.

In order to obtain local linear forests, the weights $a_i(x_t)$ are used in a local linear regression model, which was first introduced by Stone (1977). In this paper, we specify the local linear regression model as follows:

$$\begin{pmatrix} \hat{\pi}_{t+h}(x_t) \\ \hat{\theta}(x_t) \end{pmatrix} = \text{argmin}_{\pi,\theta} \left\{ \sum_{i=1}^{n} \alpha_i(x_t)(Y_i - \pi_{t+h}(x_t) - (X_i - x_t)\theta(x_t))^2 + \lambda ||\theta(x_t)||_2^2 \right\}. \tag{7}$$

Here, the parameter $\hat{\pi}_{t+h}(x_t)$ estimates the prediction function $\pi_{t+h}(x_t)$ and the parameter $\hat{\theta}(x_t)$ estimates the correction $\theta(x_t)$ for the local trend in $X_i - x_t$, for $i = 1, ..., n$. The last term in this regression model, $\lambda||\theta(x_t)||_2^2$, represents the ridge penalty, where $|| \cdot ||_2^2$ denotes the squared Euclidean norm. This ridge penalty shrinks the coefficients, reducing overfitting to the local trend. If this parameter is parameter is not set to a constant, but tuned automatically, its value gives an indication whether the smoothness assumption is satisfied for the used data. If λ is very small, the local regression step is necessary. If it is very large, the squared term in Equation 7 is zeroed out and a standard RF is returned.

Solving Equation (7) provides a closed-form solution, which can be written as

$$\begin{pmatrix} \hat{\pi}_{t+h}(x_t) \\ \hat{\theta}(x_t) \end{pmatrix} = (\Delta^T A \Delta + \lambda J)^{-1} \Delta^T A Y \tag{8}$$

in matrix form. Here, A is a diagonal matrix with the values of $a_i(x_t)$ on its diagonal and J denotes a $(d+1) \times (d+1)$ diagonal matrix, for which holds that $J_{1,1} = 0$ and $J_{i+1,i+1} = 1$. Δ denotes a regression matrix that is centered and includes an intercept, with $\Delta_{i,1} = 1$ and $\Delta_{i,j+1} = x_{i,j} - x_{t,j}$.

In this paper, we investigate three types of LLFs. The first type (LLF1) uses the same Classification and Regression Trees (CART) splitting procedure as the traditional RFs, which was proposed by Breiman et al. (1984). To improve the performance of this first local linear forest model, the procedure of obtaining the weights $\alpha_i(x_0)$ is adjusted. This tree-splitting procedure is modified in such a way that it accounts for the use of the estimate of $\pi_{t+h}(x_t)$ from the local linear regression. Whereas for the first type of LLF we use the CART procedure, in which the splits are chosen to minimize the sum of squared errors of the regression in Equation (3), we use a different procedure for the other two LLFs. In the root node P, the inflation π_{t+h} is estimated using a ridge regression, as introduced by Hoerl and Kennard (1970), and can be specified as

$$\hat{\pi}_{t+h} = \hat{\alpha}_P + x_i^T \hat{\beta}_P, \tag{9}$$

where $\hat{\alpha}_P$ denotes the intercept in the parent node and $\hat{\beta}_P$ is computed as $\hat{\beta}_P = (x_P^T x_P + \lambda J)^{-1} x_P^T Y_P$. The residuals $Y_i - \hat{\pi}_{t+h}$ are taken to find the split and construct two child nodes, using a CART splitting procedure. Afterwards, the resulting child nodes are taken as parent nodes and for each of these, the ridge regression in Equation 9 and the subsequent CART procedure are repeated. This procedure regresses out any signal that hypothetically would be covered by the ridge regression performed in the prediction step.

The difference between the second and third LLFs manifests itself in the ridge penalty λ in the prediction step in Equation 7. Whereas the ridge penalty for prediction is set to $\lambda = 0.1$ for the second LLF (LLF2), it is left unrestricted and tuned automatically for the third LLF (LLF3). For the first LLF, with weights $\alpha_i(x_0)$ determined by a CART splitting procedure, the ridge penalty λ is also left unrestricted.

4.7 Honest Random Forests

In addition to building the RFs using the traditional binary recursive way of sub-sample splitting as described in Section 4.5, we also grow trees using an alternative procedure described in Procedure 1 of Wager and Athey (2018) called honesty. In contrast to the CART splits, the trees obtained by this double-sample algorithm use a different set of data points from the training data for constructing the splits and predicting the response Y in each leaf. For each tree, the training sample is divided into two sub-samples; $\mathcal{I}_b$ and $\mathcal{J}_b$. Subsequently, the splits are chosen based on any data in $\mathcal{J}_b$, but only the covariates X in $\mathcal{I}_b$. For the first step in this splitting procedure, the tree T_b is chosen based on data from the $\mathcal{J}_b$ sample, writing the boolean indicating whether the points x_t and x' fall into the same leaf of T_b as $x_t \leftrightarrow_b x'$. Secondly, we define the set of training data that falls in the same leaf as x_t to be equal to $L_b(x_t) = \{i \in \mathcal{I}_b : x_t \leftrightarrow_b X_i\}$ and determine the forest weights with Equation 6. Finally, the inflation predictions are made using only the sample $\mathcal{I}_b$. While the LLFs can also be constructed without honestly, we follow the approach of Friedberg et al. (2020) and grow all LLFs investigated in this paper using honesty.

4.8 Macroeconomic random forests

4.8.1 General Specification

The MRF has been proposed by Goulet Coulombe (2020) as a new model with time-varying parameters (TVP) to forecast macroeconomic variables. The general model is specified as

$$Y_t = X_t \theta_t + \varepsilon_t, \tag{10}$$

$$\beta_t = \mathcal{F}(\mathcal{S}_t). \tag{11}$$

Here, Y_t is a dependent variable and θ a time-varying coefficient. Whereas we set $X_t = 1$ for the standard RF, we now specify X_t to determine the time-varying linear model. X_t is taken such that $X_t \subset \mathcal{S}_t$ and is substantially smaller than $\mathcal{S}_t$. Furthermore, $\mathcal{S}_t$ denotes a set of observed state variables and $\mathcal{F}$ denotes a random forest. If we let $\mathcal{J}^-$ be a subset of prediction variables, and l the parent node containing the full sample, the splits in the trees can be determined by solving the following problem:

$$\begin{aligned} \min_{j \in \mathcal{J}^-,\, c \in \mathbb{R}} \Bigg\{ & \min_{\theta_1} \sum_{\{t \in l | S_{j,t} \leq c\}} (Y_t - X_t\theta_1)^2 + \lambda||\theta_1||_2^2 \\ & + \min_{\theta_2} \sum_{\{t \in l | S_{j,t} > c\}} (Y_t - X_t\theta_2)^2 + \lambda||\theta_2||_2^2 \Bigg\}. \end{aligned} \tag{12}$$

The solution to this problem provides an optimal j^* and c^* that determine the optimal predictor variable $\mathcal{S}_j$ for splitting and the corresponding split value c of this variable. Afterwards, the parent node l is split into two child nodes containing a subsample of the predictor space, denoted by l_1 and l_2. For each child node, the next split is determined using Equation 12. By recursively repeating this process, a tree is grown. As proposed by Breiman (2001), we consider "decorrelated" trees, meaning that at each step in the recursion, a different set of regressors is taken as potential candidates for the split. This prevents the algorithm from repeatedly choosing the same optimization route, which further diversifies the trees and reduces computation times.

This recursive splitting of θ_0 into θ_1 and θ_2 eventually results in θ_t. However, due to the fact that θ_t has very few neighbours in its terminal leaf, it has a very low bias, but a very high variance in a single tree. Therefore, we consider a sufficiently diversified ensemble of trees using the bagging method of Breiman (1996) in order to avoid overfitting. This approach, in which the trees are all grown from an independent bootstrap sample of the original data, can result in large improvements, as has been shown by Breiman (1996) and Grandvalet (2004). Similarly to the RF approach, the simple average of the single trees can be taken to construct the MRFs. The bagging method used for the construction of the MRFs is the nonparametric bootstrap as proposed by MacKinnon (2006). However, to correct for the dependence inherent to time series data, the bootstrapping procedure is slightly adjusted. As first proposed by Taddy, Chen, et al. (2015) and Taddy, Gardner, et al. (2015), we view a RF as a sample from a posterior over trees obtained by a Bayesian bootstrap. From this perspective, the RF predictions are the approximate posterior mean of a tree functional $\mathcal{T}$, where $\mathcal{T}$ is viewed as a Bayesian nonparametric statistic. However, since this approach requires that the assumption that $Z_t = [y_t \; X_t \; \mathcal{S}_t]$ is an independent and identically distributed random variable is satisfied, we follow Goulet Coulombe (2020) by using a block extension in order to be able to use this approach. This block extension is called the Block Bayesian Bootstrap (BBB) and redefines Z such that it is plausibly independent and identically distributed. The new variable Z is specified as $Z_{\mathfrak{b}} = [y_{\underline{\mathfrak{b}}:\bar{\mathfrak{b}}} \; X_{\underline{\mathfrak{b}}:\bar{\mathfrak{b}}} \; \mathcal{S}_{\underline{\mathfrak{b}}:\bar{\mathfrak{b}}}]$, wherefore it holds that the $\mathfrak{B}$ blocks are fixed and non-overlapping and $\mathfrak{B} = \frac{T}{\text{block size}}$. A more detailed description of the BBB procedure can be found in Appendix A.2 of Goulet Coulombe (2020).

4.8.2 Model adjustments

The Ridge shrinkage used in Equation 12 causes each time-varying coefficient to be shrunk to zero at every time t. However, shrinking θ_t to zero can provoke a substantial bias when, for example, a process is highly persistent. Therefore, we transform Equation 12 such that the coefficients change smoothly over time, by shrinking θ_t to be in the neighbourhood of θ_{t-1} and θ_{t+1} instead of zero. As noted by Goulet Coulombe (2020), this is in alignment with the view that macroeconomic states, denoted by θ_t, stay approximately constant for at least a few consecutive periods. This regularisation transformation is done by, instead of solving many small ridge problems in each tree, solving many weighted least squares (WLS) problems that include close-by observations. Close-by observations are in the neighbourhood, with regard to time, of observations within the current leaf. These observations are included in the estimation of the WLS problem, but are assigned a smaller weight. This approach has also been recently investigated by Giraitis et al. (2018) and Petrova (2019) to estimate large VAR models with TVPs. The details on this regularisation transformation procedure can be found in Appendix B.6.2.

4.8.3 Choice of X_t and $\mathcal{S}_t$

In order to construct $\mathcal{S}_t$, we partly follow the approach of Goulet Coulombe (2020). In addition to all 508 variables used by Medeiros, Vasconcelos, Veiga, et al. (2021), which includes principal component factors, lags and autoregressive terms of a set of variables, we also add a time trend t in order to construct $\mathcal{S}_t$. As setting $X = 1, Y_{t-1}, Y_{t-2}$ has shown to produce accurate results in Goulet Coulombe (2020), we also choose this as our X_t.

4.9 Hybrid linear-random forests

The hybrid linear-random forest models considered are the RF/OLS model and the adaLASSO/RF model. Following Medeiros, Vasconcelos, Veiga, et al. (2021), we use these combined approaches to be able to make a distinction between the effects of variable selection and nonlinearity in forecasting US inflation. The RF/OLS approach is used to test whether nonlinearity plays a role in the predictive performance of the RF model. In the first step of this method, the forecasts $\hat{\pi}^b_{t+h}$ are determined for each bootstrap sample $b = 1, ..., B$. This is done by first growing a tree with k nodes for each sample b, while saving the $N \leq k$ split variables, and then performing OLS on the selected variables. Following Medeiros, Vasconcelos, Veiga, et al. (2021), we use $k = 20$ nodes. In the second step of the RF/OLS method, the final forecasts are determined with $\hat{\pi}_{t+h} = \frac{1}{B}\sum_{b=1}^{B} \hat{\pi}^b_{t+h}$.

If the results of the RF/OLS model are very similar to the RF model, nonlinearity shows to be irrelevant, the performance of the RF model can be explained by variable selection and nonlinearity shows to be irrelevant. However, if the RF/OLS performs better than other linear models such as bagging, but worse than the nonlinear RF approach, both nonlinearity and variable selection are important justifications for the performance of the RF model. For the adaLASSO/RF model, the adaLASSO is used to select the variables and then the forecasts are estimated by running a RF model on these selected variables. If the results of this hybrid approach are very close to those of the RF, this indicates that the variable selection in RF is not very important, whereas nonlinearity plays an important role.[2]

4.10 Comparison Criteria and Tests

In order to compare the performance of all models, we used three different comparison criteria: the root mean squared error (RMSE), the mean absolute error (MAE) and the median absolute deviation from the median (MAD). These RMSE and MAE are specified as

$$RMSE_{m,h} = \sqrt{\frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \hat{e}^2_{t,m,h}} \quad \text{and} \quad MAE_{m,h} = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} |\hat{e}_{t,m,h}|.$$

As the effect of each error on the RMSE is proportional to the size of the squared error, it gives errors with large absolute values more weight than errors with small absolute values. By definition, the MAE is smaller than the RMSE. Therefore, evaluating not only the RMSE, but also the values of the MAE and MAD gives a more complete representation of the forecasting accuracy and prevents the results to depend solely on a few large forecasting errors. The MAD, which is robust to outliers and asymmetric data properties, is specified as

$$MAD_{m,h} = \text{median}[|\hat{e}_{t,m,h} - \text{median}(\hat{e}_{t,m,h})|].$$

Following the approach of Medeiros, Vasconcelos, Veiga, et al. (2021), we use the model confidence sets (MCS) proposed by Hansen et al. (2011) to test whether the forecasts of the models differ. For each forecasting horizon, we compute the MCSs based on both the $t_{\max}$ and the $t_{\max}$ statistic as described by Hansen et al. (2011). The MCS procedure orders all models on forecasting accuracy and, where the model with the highest p-value performs best. Similar to the application of our paper, Hansen et al. (2011) also apply the MCSs to inflation forecasts. Furthermore, we evaluate the relative forecasting accuracy between different models through the Diebold-Mariano (DM) test, introduced by Diebold and Mariano (1995) and a correction on this test for small samples proposed by Harvey et al. (1997). The DM test has also been used by Goulet Coulombe (2020), who uses ML methods and Castañeda-Fuentes et al. (2018) and Gneiting and Thorarinsdottir (2010), who both use this test for evaluating the predictive performance of models in forecasting inflation.

4.11 Alternative forecasting samples

To evaluate if our results are stable over time and for different sample sizes, we follow Medeiros, Vasconcelos, Veiga, et al. (2021) by comparing the predictive performance of the above models for two different forecasting samples. In existing literature, this difference in performance has often been observed. Stock and Watson (2008), for example, show that Philips curve predictions show to be more accurate than the UCSV during periods of high volatility, while it is outperformed in periods of low volatility. d'Agostino et al. (2008) have shown a similar change in results over

[2] As the RF and the LLF are similar in approach, we did not repeat this experiment for the LLF case, as the results are likely to be the same.

different time periods. Similarly to the methods of Medeiros, Vasconcelos, Veiga, et al. (2021), for which they included the results in Tables S.10 and S.11 of the supplementary material, we split the full out-of-sample period into two. We compare a period of low inflation volatility, namely January 1990 to December 2000, to a period with high inflation volatility, namely from January 2001 to December 2015, and evaluate whether the relative performance of each of the models stays the same and our results are robust to periods of high volatility.

5 Results

Table 1 shows the forecasting results of the RW model, the AR model, two RFs constructed using different packages[3], a RF constructed using honest estimation, the hybrid linear RF models, the three types of LLFs and the MRFs constructed with 1 and 8 trees respectively. All values are obtained by first forecasting inflation over the two out-of-sample periods 1990-2000 (132 predictions) and 2001-2015 (180 predictions)[4] and then computing the three evaluation criteria using all 312 predictions. The first three columns show the average RMSE, MAE and the MAD of the inflation forecasts over forecasting horizons $h = 1, ..., 12$ and of the 3-, 6- and 12-month ahead accumulated forecasts. Columns (4), (5) and (6) show the maximum RMSE, MAE and MAD over the 15 forecasting horizons and columns (7), (8) and (9) report the minimum RMSE, MAE and MAD. All values displayed in columns (1)-(9) have been normalised such that the values of the RW are equal to one. The lowest values for the criteria are shown in bold. Columns (10), (11) and (12) display the frequency with which each model achieved the lowest value for the RMSE, the MAE and the MAD respectively. The last two columns report the average p-values for the MCSs of Hansen et al. (2011) based on the $t_{\max}$ statistic for absolute and square losses respectively. The results of the MCSs based on the t_R statistic can be found in Table 7 in Appendix A.1, which provide similar findings as the MCSs based on the $t_{\max}$ statistic.

Table 1

Forecasting results for the out-of-sample period from 1990-2015.

	Average			Maximum			Minimum			# Minimum			MCS	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Model	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	abs	sq
RW	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	0	0	0.02	0.03
AR	0.84	0.86	0.78	1.22	1.22	1.16	0.75	0.76	0.60	0	0	0	0.14	0.12
RF	0.74	0.74	**0.70**	**0.85**	**0.81**	0.84	0.69	**0.67**	0.58	0	4	4	1.00	0.68
RF grf	**0.72**	**0.73**	**0.70**	**0.85**	**0.81**	0.86	0.67	**0.67**	**0.53**	7	7	4	1.00	0.89
Honest RF	0.79	0.83	**0.70**	1.03	1.17	**0.81**	0.72	0.72	0.58	0	0	6	0.77	0.25
RF/OLS	0.76	0.78	0.81	0.94	0.97	1.08	0.71	0.71	0.66	1	1	0	0.89	0.54
adaLASSO/RF	0.75	0.75	0.73	**0.85**	0.82	0.87	0.70	0.68	0.58	0	1	0	0.90	0.57
LLF1	**0.72**	0.75	0.76	0.88	0.89	0.88	**0.66**	**0.67**	0.63	7	2	0	0.95	0.82
LLF2	0.74	0.78	0.82	0.90	0.90	1.06	0.70	0.72	0.67	0	0	0	0.58	0.41
LLF3	0.73	0.77	0.81	0.91	0.92	1.05	0.67	0.70	0.65	2	0	0	0.88	0.55
MRF (B=1)	1.03	1.09	1.07	1.94	1.81	1.79	0.83	0.88	0.79	0	0	0	0.31	0.19
MRF (B=8)	0.77	0.79	0.75	0.94	1.00	1.02	0.72	0.72	0.58	0	0	1	0.70	0.38

When looking at columns (1), (2) and (3), we find that both standard RFs and LLF1 perform

[3] We use both packages in order to be able to make an adequate comparison of the standard RF and the honest RF, which is also implemented with the **grf** package. The RF constructed with the **RandomForest** package is denoted by "RF" in Table 1 and the RF constructed using the **grf** package is denoted by "RF grf".

[4] For the second sample (2001-2015), a dummy is added for the observation in November 2008.

best in terms of the average of the three evaluation criteria. The RF grf model performs best for all three criteria, closely followed by the RF and the LLF1 model. The minimum of the criteria over all horizons, which is reported in columns (7), (8) and (9), shows a similar, but slightly different pattern. Again we find that the RF and the RF grf perform very similar and that LLF1 performs best in terms of RMSE. When we look at the maximum of the comparison criteria over all 15 horizons, we find that both standard RFs and the adaLASSO/RF perform best. Again, the RFs are approximately equally accurate for all three criteria.

Furthermore, it is notable that the honest RF achieves the lowest value of the MAD when looking at the average and the maximum MAD over all models and all 15 horizons. Moreover, it has a relatively low value for the minimum MAD and achieves the lowest minimum MAD for 6 of the 15 horizons. This might be caused by the fact that the MAD is more robust to outliers than the RMSE and the MAE, indicating that the honest RF forecasts contain a substantial amount of outliers and are probably less biased.

Whereas the RF and the RF grf perform very similar over all criteria, we find from columns (11)-(14) that the RF grf gives slightly more accurate results, closely followed by the RF and the LLF1 model. The RF grf achieves the minimum value for each of the three evaluation criteria over all 15 horizons with the highest frequency and has the highest p-values for the MCSs for both absolute and square losses. The RF achieves the minimum MAE and MAD for 4 out of 15 horizons, and belongs to the three best models indicated by the MCS tests, with a p-value of 1.00 for the MCSs for absolute losses and a p-value of 0.68 for square losses. While LLF1 performs very well in terms of the average and minimum RMSE, MAE and MAD, it has relatively high maximum values. With a p-value of 0.95 for the MCSs for absolute losses and a p-value of 0.82 for square losses, the LLF1 model also belongs to the three best models according to the MCSs.

The MCSs also indicate that both LLF2 and LLF3 perform worse than LLF1. The LLF3 model, in which the ridge penalty for prediction is left unspecified and tuned by the forest, performs slightly better than the LLF2 model with a ridge penalty λ set to 0.1. That the models with local linear splits based on ridge residuals performs worse than the LLF with CART splits suggests that the local linear regression step is not necessary for obtaining accurate results.

Looking at the performance of the two MRFs, we find that the forecasting accuracy improves substantially as the number of trees increases. Both models perform worse than most of the other models investigated, which could have been expected since the LLFs and RFs are built using 2000 and 500 trees respectively, while the MRFs are built using a maximum of 8 trees.

Lastly, we find that, based on the average p-values for the MCSs in columns (13) and (14), all models perform better than the two benchmark models as they all have higher p-values.

Tables 2 and 3 show the average p-values of the standard Diebold Mariano tests for all 15 horizons. If the null hypothesis of equal forecasting accuracy between two models is rejected at a 95% confidence interval, this is indicated by one asterisk. If the null hypothesis is rejected at a 99% confidence interval, this is indicated by two asterisks. In Tables 24 and 25 in Appendix A.3, the values of the Diebold Mariano test with a Harvey, Leybourne and Newbold (HLN) correction can be found. As these results do not differ substantially from the results in Tables 2 and 3, the same conclusions can be drawn as from the standard Diebold Mariano tests.

Table 2

The p-values of the Diebold Mariano tests for absolute losses.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.02*										
RF	0.00**	0.01**									
RF grf	0.00**	0.01**	0.41								
honest RF	0.04*	0.14	0.06	0.10							
RF/OLS	0.03*	0.06	0.20	0.21	0.64						
adaLASSO/RF	0.00**	0.02*	0.43	0.36	0.25	0.39					
LLF1	0.00**	0.10	0.65	0.48	0.33	0.33	0.47				
LLF2	0.00**	0.31	0.11	0.09	0.51	0.47	0.23	0.10			
LLF3	0.01**	0.18	0.27	0.16	0.56	0.52	0.30	0.28	0.38		
MRF (B=1)	0.06	0.52	0.02	0.02	0.33	0.27	0.06	0.15	0.30	0.45	
MRF (B=8)	0.07	0.12	0.07	0.21	0.59	0.40	0.23	0.24	0.49	0.57	0.20

Table 3

The p-values of the Diebold Mariano tests for square losses.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.05*										
RF	0.01*	0.05*									
RF grf	0.01*	0.07	0.39								
honest RF	0.03*	0.14	0.07	0.18							
RF/OLS	0.03*	0.10	0.24	0.21	0.69						
adaLASSO/RF	0.01*	0.09	0.40	0.42	0.32	0.34					
LLF1	0.02*	0.13	0.49	0.44	0.38	0.33	0.39				
LLF2	0.02*	0.26	0.63	0.49	0.63	0.49	0.59	0.36			
LLF3	0.02*	0.23	0.68	0.61	0.57	0.49	0.57	0.49	0.53		
MRF (B=1)	0.04*	0.38	0.04	0.04	0.31	0.58	0.17	0.13	0.24	0.31	
MRF (B=8)	0.02*	0.14	0.16	0.21	0.52	0.64	0.25	0.26	0.40	0.49	0.34

The results in Table 3 suggest that the forecasts of all models are significantly more accurate than the forecasts of the RW with a 95% confidence level. The p-values based on absolute errors provide slightly different results. With p-values of 0.06 and 0.07 respectively, the null hypothesis of equal accuracy is not rejected for the MRF built with 1 tree and the MRF built with 8 trees. Furthermore, we find that the only model that performs significantly better than the AR model for both the absolute and square losses, is the RF. Between all other models than the two benchmarks and the RF, the null hypothesis of equal forecasting accuracy can not be rejected for the test for square losses, indicating that the forecasts of these models are equally accurate. Furthermore, Table 2 suggests that the LLF1 forecasts are very similar to the RF and RF grf forecasts, as the DM test provides relatively high p-values when comparing these models.

5.1 Alternative forecasting samples

Tables 4 and 5 show the forecasting results over a period with low inflation volatility (1990-2000) and high inflation volatility (2001-2015) respectively. From columns (10), (11) and (12) we find that the RF grf achieves the minimum value of the RMSE, MAE and MAE 7, 11 and 6 times respectively for the first sample, while it only achieves the minimum 5, 8 and 1 times for the second sample. This decrease in the number of times the model reaches the mimimum RMSE, MAE and MAD also holds for the RF. This suggests that the improved performance of the the RFs over the other models is more evident for periods of low inflation volatility compared to the other models. However, we do still find for both samples that the RF grf and the RF provide the best results. Furthermore, we find that the honest RF and the LLF3 model perform better relative to the other models during periods of high inflation volatility than during the first sample period with low inflation volatility as the number of times these models achieve the minimum RMSE, MAE and MAD increases. This finding is supported by the p-values for the MCSs, which increase from 0.15 and 0.16 to 0.91 and 1.00 for the absolute and square losses respectively for the honest RF and from 0.81 and 0.89 to 0.93 and 0.96 for the LLF3 model. This might suggest that the RF and the LLF3 perform better for periods of high volatility than for periods of low volatility. However, further research is needed to validate this hypothesis as there are many other possible explanations for the observed difference in performance between the two samples. It might also be related to the small size of the sample, resulting in too many parameters to be estimated relative to the number of number of observation.

We do find that the LLF2 and LLF3 perform better for the two subsamples than for the full sample in Table 1. The relatively weak performance of the LLF2 and LLF3 model for the full sample compared the the two subsamples might be caused by the fact that this LLF extrapolates in cases where it shouldn't. This extrapolation could be less of an issue for smaller samples, causing the forecasting accuracy to be higher.

Table 4

Forecasting results for the out-of-sample period from 1990-2000.

	Average			Maximum			Minimum			# Minimum			MCS	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Model	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	abs	sq
RW	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1	0	0.40	0.32
AR	0.90	0.96	0.76	1.24	1.36	1.26	0.78	0.80	0.59	0	0	1	0.60	0.70
RF	0.81	0.85	0.67	0.98	1.15	0.88	**0.71**	0.75	0.56	2	1	4	1.00	1.00
RF grf	**0.79**	**0.83**	0.69	0.93	**1.09**	0.89	0.72	**0.72**	0.55	7	11	6	1.00	1.00
honest RF	0.96	1.10	**0.64**	1.41	1.82	**0.80**	0.80	0.91	**0.53**	0	0	2	0.15	0.61
RF/OLS	0.84	0.90	0.87	1.06	1.23	1.15	0.74	0.76	0.70	0	0	0	0.84	0.86
adaLASSO/RF	0.83	0.87	0.72	1.00	1.16	0.88	0.72	0.74	0.54	0	0	0	0.94	1.00
LLF1	0.80	0.86	0.77	0.94	1.13	1.09	0.72	0.75	0.63	5	2	0	1.00	1.00
LLF2	0.86	0.92	0.83	1.04	1.26	1.15	0.74	0.80	0.66	0	0	0	0.89	0.96
LLF3	0.84	0.90	0.88	**0.91**	**1.09**	1.35	0.74	0.79	0.75	1	0	0	0.81	0.89
MRF (B=1)	0.93	0.99	0.80	1.21	1.47	1.18	0.80	0.81	0.65	0	0	1	0.29	0.34
MRF (B=8)	0.89	0.95	0.80	1.17	1.40	1.21	0.77	0.82	0.64	0	0	1	0.50	0.55

Table 5

Forecasting results for the out-of-sample period from 2001-2015.

	Average			Maximum			Minimum			# Minimum			MCS	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Model	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	abs	sq
RW	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	0	0	0.63	0.89
AR	0.84	0.82	0.80	1.21	1.17	0.94	0.73	0.70	0.62	0	0	0	0.31	0.43
RF	0.72	0.70	0.70	**0.87**	**0.81**	0.94	0.68	**0.63**	0.50	0	1	0	1.00	1.00
RF grf	**0.71**	**0.69**	0.68	**0.87**	**0.81**	0.84	0.69	**0.63**	0.53	5	8	1	1.00	1.00
honest RF	0.77	0.74	**0.66**	0.94	0.95	0.81	0.70	0.65	0.52	0	3	2	0.91	1.00
RF/OLS	0.75	0.74	0.69	0.92	0.89	0.83	0.70	0.67	0.54	1	1	3	0.97	0.91
adaLASSO/RF	0.74	0.71	0.69	0.86	**0.81**	0.89	0.69	0.65	0.54	1	1	2	1.00	0.98
LLF1	**0.71**	0.71	0.69	0.91	0.90	0.99	0.78	0.74	0.84	6	0	2	0.93	0.99
LLF2	0.72	0.74	0.71	0.92	0.92	0.97	0.66	**0.63**	**0.49**	1	0	2	0.74	0.95
LLF3	0.72	0.72	0.70	0.92	0.91	0.93	**0.65**	0.66	0.58	1	1	3	0.93	0.96
MRF (B=1)	0.78	0.76	0.70	0.91	0.89	**0.80**	0.72	0.69	0.52	0	0	0	0.62	0.63
MRF (B=8)	0.76	0.74	0.70	0.88	0.87	0.83	0.71	0.68	0.53	0	0	0	0.86	0.91

The Diebold Mariano test results for the two different samples can be found in Tables 20-23 in Appendix A.2.

We also included the Diebold Mariano with HLN correction for the two different subsamples in Tables 26-29 Appendix A.3. Again, we find that the results do not differ substantially from the results of the standard Diebold Mariano test.

5.2 MCS per horizon

Tables 8-19 in Appendix A.1 show the results of the MCSs per horizon based on both the $t_{\max}$ and t_R statistics for absolute and square losses for the out-of-sample period 1990-2015, the sample period 1990-2000 and the sample period 2001-2015. These results provide insights in the difference in performance of the models over different horizons. From Tables 8-11 we find that the adaLASSO/RF performs best for the 8-month ahead forecasts for the full out-of-sample period 1990-2015. Furthermore, it is notable that the hybrid linear random forests RF/OLS and adaLASSO/RF provide the most accurate 1-month ahead forecasts for this sample. Overall, we find that the RF grf and the RF provide the most accurate forecasts most often, closely followed by the LLF1 model.

The results for the out-of-sample period 1990-2000, shown in Tables 12-15, indicate that the LLF1 model performs best for the 3-month ahead accumulated forecasting horizon and the RW model performs best for the 6-month accumulated forecasting horizon. Again, we find that the RF grf and the RF provide the most accurate results for most horizons, followed by the LLF1 model.

The out-of-sample period of 2001-2015 provides different results, which are shown in Tables 16-19 of the Appendix. Whereas the honest RF did not provide the most accurate forecasts for any of the horizons for the other two out-of-sample periods, it does provide the most accurate results for several horizons for the out-of-sample period 2001-2015, which is in line with the improved performance of the honest RF found in Table 5. The LLF2 model performs best for the 12-month accumulated forecasts during the out-of-sample period of 2001-2015. Furthermore, similar to the findings for the forecasting period 1990-2015, we find that the two hybrid linear random forests provide the best forecasts for the 1-month ahead horizon.

None of the horizon-specific results found for the full out-of-sample period 1990-2015 is also true for both subsamples, except for the results that the RF grf and the RF provide the most accurate forecasts most often, closely followed by the LLF1 model.

5.3 Performance analysis

As also found by Medeiros, Vasconcelos, Veiga, et al. (2021), the forecasting results of the two hybrid linear random forests in Table 6 of Appendix A show that the RF/OLS forecasts are more accurate than bagging and less accurate than the RF. Therefore, both variable selection and nonlinearity play important roles in the performance of the RF.

The performance of the LLF models is dependent on whether the assumption that the data contains smoothness is satisfied. If this assumption is satisfied, the local regression step could potentially improve forecasting accuracy. The ridge penalty λ of Equation 7 gives an indication for the amount of smoothness in our data. Whereas we set λ to 0.1 for LLF2, we left it unspecified and automatically tuned for LLF1 and LLF3. For LLF1, the average ridge penalty λ of Equation 7 over all horizons is equal to 0.14. For LLF3, λ has a mean value of 0.18. This value is not large enough to cancel out the squared term in Equation 7. This is an indication that the local regression step could potentially be useful and that the data contains a substantial amount of smoothness.

6 Conclusion

In this paper, we extend the work of Medeiros, Vasconcelos, Veiga, et al. (2021) by analysing the relative ability of various ML methods to forecast US inflation in a data-rich environment. The main aim was to investigate whether adaptations to the RF model could improve forecasting accuracy relative to the standard RFs, providing an extensive overview and comparison of these methods. Our results suggest that the LLF1 model gives similar, though slightly less accurate forecasts than the standard RF model, based on on a comparison of the RMSE, MAE and MAD, the DM test and the MCSs. As the LLF2 and LLF3 model perform worse, our results imply that the ridge regression for estimation is not necessary in order to obtain accurate results. The good performance of the LLF1 model could be explained by the fact that the data contains a substantial amount of smoothness.

From our results we find that the forecasts of the MRF are less accurate than the standard RF. This can be explained, however, by the fact that we used a very small number of bootstrap samples B to build the MRFs. Whereas we use $B = 8$ trees for the MRF, we use $B = 500$ for the standard RF, possibly explaining the difference in forecasting accuracy between the two models. We expect that increasing the number of trees used to build the forests will improve forecasting accuracy. From this, in combination with the fact that both models can capture complex interactions of large data sets, we conclude that the LLF and the MRF are both promising approaches that have high potential to further improve forecasting accuracy in a data-rich environment.

This same conclusion that the LLF performs similar to the RF and the MRF produces relatively weak forecasts holds for the two alternative subsamples investigated.

We find that the forecasting performance of the honest RF is relatively weak compared to the performance of the standard RF. This weak performance is especially evident during periods of low inflation volatility. For the period 2001-2015, with higher inflation volatility, we show the honest RF to perform better, though still slightly worse than the RF.

Improvements to our paper could be made in regard to the tuning parameters of the LLF models. If the tuning parameters model are set perfectly, the efficiency as well as the forecasting accuracy of the used models could be improved. For the LLF, alterations to the split cutoff parameter could potentially increase performance. Although decreasing this parameter results in a loss of efficiency, it causes the tree to run a regression in smaller leaves instead of using the regression coefficients of the full data set, which could possibly provide more accurate predictions. Moreover, the LLFs could have been tuned in such a way that the ridge penalty is standardized by covariance instead of penalizing all covariates equally.

The MRFs could also be improved by setting the tuning parameters differently. For the MRF, a larger number of bootstrap samples will very likely result in more accurate predictions. Furthermore, Goulet Coulombe (2020) highlights that the gains from using MRF are a result of a correct specification of X_t for a complex DGP. Therefore, the relatively weak performance of the MRF could also be caused by a misspecification in X_t. Investigating MRFs with other specifications of X_t and a larger number of bootstrap samples, could potentially provide better forecasting results.

That the honest RFs did not perform as well as the standard RFs may be a result of the tuning parameters used. Especially increasing the number of parameters used for determining the splits could potentially increase the predictive ability of this method. The fact that local linear forests do perform well under honest prediction can be explained by the ability of its local linear corrections to reduce the loss of expressive power caused by honesty.

What should be noted is that our results particularly hold for this data set and sample. Forecasting inflation using different covariates could lead to very different results. The use of different intervals, data sets, inflation measures or comparison criteria could also result in a different outcome. Even though Medeiros, Vasconcelos, Veiga, et al. (2021) have also shown the results to be similar for the inflation CPI, PCE and CPI core inflation, other research, such as Fisher et al. (2002), suggests the contrary. Furthermore, other data frequencies could be investigated. In this paper, we used monthly data, but other results could be obtained from high frequency data or quarterly data. For making a more general comparison between the performance of the different forecasting models in forecasting inflation, a multitude of data sets, data samples and data frequencies should be analysed.

We expect further research in the field of RFs as adaptive kernel methods or other variations on the RF model to provide promising results that could potentially improve the predictions of the RF model. Moreover, other methods that can capture smoothness and linear functions could be investigated. Soft Bayesian Additive Regression Trees (soft BART), for example, which have been introduced by Linero and Yang (2017), can capture smoothness and sparsity. Other variations on BART, such as the Model Trees BART, proposed by Prado et al. (2021), or the BART with Targeted Smoothing of Starling et al. (2020) can also capture local linearities and could be interesting to further investigate in the context of inflation forecasting with high-dimensional data.

Further research that could also be insightful is to consider and compare variable importance for each of the models in order to form a better understanding of the performance results.

The difference in results in this paper and the paper of Medeiros, Vasconcelos, Veiga, et al. (2021) can be explained by the fact that the codes provided by Medeiros, Vasconcelos, Veiga, et al. (2021) were not entirely complete. The default settings of the **randomForest** package have been used to estimate the RFs. As these default parameter settings might change over time, this could explain the difference between our results and those of Medeiros, Vasconcelos, Veiga, et al. (2021). Furthermore, the random seed has not been specified, allowing for deviations in results due to randomness. Medeiros, Vasconcelos, Veiga, et al. (2021) have used a dummy variable for the observation of November 2008. However, this could give an inaccurate representation of the forecasting ability of the models as forecasters would not have known before the event that it was important. This might make the trained forest less able to predict events. To make sure that this is not the case, this research could be repeated without the inclusion of the dummy variable.

References

Alexander, W. P., & Grimshaw, S. D. (1996). Treed regression. *Journal of Computational and Graphical Statistics*, *5*(2), 156–175.

Ang, A., Bekaert, G., & Wei, M. (2005). *Do macro variables, asset markets or surveys forecast inflation better?* (Working Paper No. 11538). National Bureau of Economic Research.

Araujo, G. S., & Gaglianone, W. P. (2020). *Machine learning methods for inflation forecasting in brazil: New contenders versus classical models* (tech. rep.). Mimeo.

Aron, J., & Muellbauer, J. (2013). New methods for forecasting inflation, applied to the us. *Oxford Bulletin of Economics and Statistics*, *75*(5), 637–661.

Athey, S., Tibshirani, J., Wager, S. et al. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171.

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*(2), 304–317.

Banerjee, A., Marcellino, M., & Masten, I. (2014). Forecasting with factor-augmented error correction models. *International Journal of Forecasting*, *30*(3), 589–612.

Banfield, R., Hall, L., Bowyer, K., & Kegelmeyer, W. (2007). A comparison of decision tree ensemble creation techniques. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *29*, 173–180.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2000). Some infinite theory for predictor ensembles.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Castañeda-Fuentes, J. C., Valle-Samayoa, H. A., Catalán-Herrera, J. C., Arriaza-Herrera, J. C., Gutiérrez-Morales, M. J., Castillo-Maldonado, C. E., Galindo-Gonzáles, D. N., Hurtarte-Aguilar, G., & Ortiz-Cardona, E. R. (2018). *Evaluation of inflation forecasting models in guatemala* (tech. rep.). IDB Working Paper Series.

Chen, J. C., Dunn, A., Hood, K., Driessen, A., & Batch, A. (2019). Off to the races: A comparison of machine learning and alternative data for predicting economic indicators. *Big data for 21st century economic statistics*. University of Chicago Press.

Cheng, K., Huang, N., & Shi, Z. (2021). Survey-based forecasting: To average or not to average. *Behavioral predictive modeling in economics* (pp. 87–104). Springer.

d'Agostino, A., Giannone, D., & Surico, P. (2008). (un) predictability and macroeconomic stability.

DeLong, J. B. (1996). America's only peacetime inflation: The 1970s. *NBER Working Paper*, (h0084).

Denil, M., Matheson, D., & De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. *International conference on machine learning*, 665–673.

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

Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, *177*(2), 357–373.

Elliott, G., Gargano, A., & Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, *54*, 86–110.

Fisher, J. D., Liu, C. T., & Zhou, R. (2002). When can we forecast inflation? *Economic Perspectives-Federal Reserve Bank of Chicago*, *26*(1), 32–44.

Fortin-Gagnon, O., Leroux, M., Stevanovic, D., & Surprenant, S. (2018). *A Large Canadian Database for Macroeconomic Analysis* (CIRANO Working Papers 2018s-25). CIRANO.

Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2020). Local linear forests.

Friedman, J., Hastie, T., Tibshirani, R. et al. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Garcia, M. G., Medeiros, M. C., & Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, *33*(3), 679–693.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*(1), 3–42.

Giraitis, L., Kapetanios, G., & Yates, T. (2018). Inference on multivariate heteroscedastic time varying random coefficient models. *Journal of Time Series Analysis*, *39*(2), 129–149.

Gneiting, T., & Thorarinsdottir, T. L. (2010). Predicting inflation: Professional experts versus no-change forecasts. *arXiv preprint arXiv:1010.2318*.

Goulet Coulombe, P. (2020). The macroeconomy as a random forest. *Available at SSRN 3633110*.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting?

Grandvalet, Y. (2004). Bagging equalizes influence. *Machine Learning*, *55*(3), 251–270.

Hanif, M. N., & Malik, M. J. (2015). Evaluating performance of inflation forecasting models of pakistan.

Hansen, B. E. (2011). Threshold autoregression in economics. *Statistics and its Interface*, *4*(2), 123–127.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, *13*(2), 281–291.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. *Aug, Springer*, *1*.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hothorn, T., Lausen, B., Benner, A., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in Medicine*, *23*(1), 77–91.

Liaw, A., & Wiener, M. (2001). Classification and regression by randomforest. *Forest*, *23*.

Linero, A. R., & Yang, Y. (2017). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *arXiv preprint arXiv:1707.09461*.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics. *Economic Record*, *82*, S2–S18.

McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

Medeiros, M. C., & Mendes, E. F. (2016). 1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, *191*(1), 255–271.

Medeiros, M. C., Vasconcelos, G., & Freitas, E. (2016). Forecasting brazilian inflation with high-dimensional models. *Brazilian Review of Econometrics*, *36*(2), 223–254.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119.

Medeiros, M. C., & Vasconcelos, G. F. (2016). Forecasting macroeconomic variables in data-rich environments. *Economics Letters*, *138*, 50–52.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(35), 983–999.

Olson, M. A., & Wyner, A. J. (2018). Making sense of random forest probabilities: A kernel perspective. *arXiv preprint arXiv:1812.05792*.

Petrova, K. (2019). A quasi-bayesian local likelihood approach to time varying parameter var models. *Journal of Econometrics*, *212*(1), 286–306.

Prado, E. B., Moral, R. A., & Parnell, A. C. (2021). Bayesian additive regression trees with model trees. *Statistics and Computing*, *31*(3), 1–13.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, *72*(3), 821–852.

Scornet, E. (2016). Random forests and kernel methods. *IEEE Transactions on Information Theory*, *62*(3), 1485–1500.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, *60*, 144–162.

Segal, M., & Xiao, Y. (2011). Multivariate random forests. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 80–87.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., & Scott, J. G. (2020). Bart with targeted smoothing: An analysis of patient-specific stillbirth risk. *The Annals of Applied Statistics*, *14*(1), 28–50.

Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, *44*(2), 293–335.

Stock, J. H., & Watson, M. W. (2008). Phillips curve inflation forecasts.

Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, 595–620.

Taddy, M., Chen, C.-S., Yu, J., & Wyle, M. (2015). Bayesian and empirical bayesian forests.

Taddy, M., Gardner, M., Chen, L., & Draper, D. (2015). A nonparametric bayesian analysis of heterogeneous treatment effects in digital experimentation.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the american Statistical association*, *89*(425), 208–218.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Ülke, V., Sahin, A., & Subasi, A. (2018). A comparison of time series and machine learning models for inflation forecasting: Empirical evidence from the usa. *Neural Computing and Applications*, *30*(5), 1519–1527.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.

Zhang, B., Chan, J. C., & Cross, J. L. (2020). Stochastic volatility models with arma innovations: An application to g7 inflation forecasts. *International Journal of Forecasting*, *36*(4), 1318–1328.

Zhang, H., Nettleton, D., & Zhu, Z. (2019). Regression-enhanced random forests.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

Appendices

A Additional results

Table 6

Forecasting results for the out-of-sample period from 1990-2015.

	Average			Maximum			Minimum			# Minimum		
Model	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD
RW	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	0	0
AR	0.84	0.86	0.78	1.22	1.22	1.16	0.75	0.76	0.60	0	0	0
LASSO	0.78	0.80	0.74	0.98	1.05	0.90	0.72	0.73	0.61	0	0	0
adaLASSO	0.78	0.79	0.76	0.95	0.96	0.98	0.72	0.71	0.63	0	0	0
ElNet	0.78	0.8	0.73	0.98	1.06	0.89	0.73	0.72	0.60	0	0	2
adaElNet	0.78	0.78	0.76	0.96	0.98	0.97	0.73	0.71	0.61	0	0	2
RR	0.76	0.77	0.79	0.89	0.93	1.03	0.70	0.71	0.67	0	0	0
BVAR	0.80	0.82	0.80	1.07	1.09	1.14	0.74	0.73	0.64	0	0	0
Bagging	0.79	0.84	0.89	0.83	0.90	1.16	0.74	0.78	0.74	0	0	0
CSR	0.82	0.82	0.80	1.13	1.11	1.09	0.76	0.74	0.67	0	0	0
Factor	0.84	0.87	0.87	1.17	1.21	1.25	0.78	0.78	0.71	0	0	0
T. Factor	0.83	0.88	0.89	1.17	1.23	1.26	0.77	0.80	0.70	0	0	0
B. Factor	0.83	0.90	0.99	1.17	1.32	1.60	0.74	0.75	0.73	0	0	0
RF	0.74	0.74	0.70	0.85	0.81	0.84	0.69	0.67	0.57	11	13	8
RF/OLS	0.76	0.78	0.81	0.94	0.97	1.08	0.71	0.71	0.66	1	1	0
adaLASSO/RF	0.75	0.75	0.73	0.85	0.82	0.87	0.70	0.68	0.58	3	1	3

A.1 Results Model Confidence Sets

Table 7

The average p-values for the MCSs based on the t_R statistic over all horizons.

	Out-of-sample period					
	1990-2015		1990-2000		2001-2015	
Model	abs	sq	abs	sq	abs	sq
RW	0.02	0.11	0.28	0.28	0.34	0.84
AR	0.10	0.34	0.23	0.29	0.17	0.39
RF	0.68	0.90	0.56	0.70	0.75	0.98
RF grf	0.87	0.99	0.99	0.97	0.91	0.97
honest RF	0.25	0.40	0.00	0.04	0.80	0.83
RF/OLS	0.57	0.79	0.47	0.57	0.78	0.75
adaLASSO/RF	0.59	0.73	0.67	0.66	0.72	0.84
LLF1	0.79	0.90	0.86	0.92	0.77	0.92
LLF2	0.33	0.70	0.42	0.38	0.43	0.88
LLF3	0.52	0.83	0.58	0.65	0.71	0.97
MRF (B=1)	0.15	0.23	0.15	0.16	0.35	0.39
MRF (B=8)	0.33	0.39	0.20	0.28	0.56	0.55

Table 8

The p-values for the MCSs based on the $t_{\max}$ statistic for absolute losses obtained from forecasts over the out-of-sample period from 1990-2015.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00
AR	0.22	0.00	0.19	0.23	0.00	0.18	0.33	0.59	0.19	0.21	0.00	0.00	0.00	0.00	0.00
RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
RF grf	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
honest RF	0.17	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00	0.67	0.99	0.50	0.27	0.00
RF/OLS	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	0.80	0.68	1.00	0.97	0.00
adaLASSO/RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.43
LLF1	0.26	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF2	0.00	0.27	0.33	0.43	0.68	0.80	0.43	0.49	1.00	1.00	0.45	0.47	0.33	1.00	1.00
LLF3	0.00	1.00	1.00	0.80	0.66	0.75	0.97	0.98	1.00	1.00	1.00	1.00	0.99	1.00	1.00
MRF (B=1)	0.46	0.00	0.26	0.99	0.00	0.89	0.27	0.97	0.15	0.59	0.00	0.00	0.00	0.00	0.00
MRF (B=8)	1.00	0.58	0.95	1.00	0.62	1.00	1.00	0.80	1.00	1.00	1.00	0.15	0.30	0.11	0.00
Best model	RF	RF	RF	RF grf	RF grf	RF	RF grf	RF	RF	RF grf	RF grf	RF	RF	RF grf	RF grf

Table 9

The p-values for the MCSs based on the t_R statistic for absolute losses obtained from forecasts over the out-of-sample period from 1990-2015.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00
AR	0.00	0.00	0.06	0.05	0.00	1.00	0.07	0.26	0.03	0.06	0.00	0.00	0.00	0.00	0.00
RF	0.80	0.45	1.00	0.52	0.63	1.00	0.23	0.45	1.00	1.00	1.00	1.00	0.14	0.90	0.03
RF grf	1.00	0.45	0.96	1.00	1.00	1.00	1.00	1.00	0.86	1.00	1.00	0.61	0.14	1.00	1.00
honest RF	0.00	0.45	0.84	0.10	0.14	1.00	0.07	0.45	0.17	0.24	0.08	0.08	0.12	0.00	0.00
RF/OLS	1.00	0.45	0.96	1.00	0.74	1.00	0.23	0.45	0.80	0.93	0.37	0.38	0.14	0.06	0.00
adaLASSO/RF	1.00	0.45	0.96	1.00	0.74	1.00	0.23	1.00	0.86	0.05	0.00	0.61	0.14	0.82	0.00
LLF1	0.02	1.00	0.96	1.00	0.74	1.00	0.23	0.45	0.86	1.00	1.00	0.61	1.00	0.90	1.00
LLF2	0.00	0.00	0.03	0.02	0.33	1.00	0.07	0.16	0.73	0.93	0.37	0.20	0.01	0.06	1.00
LLF3	0.00	0.45	0.96	0.10	0.14	0.07	0.07	0.45	0.86	1.00	1.00	0.61	0.14	0.90	1.00
MRF (B=1)	0.09	0.00	0.03	0.52	0.00	1.00	0.06	0.45	0.02	0.08	0.00	0.00	0.00	0.00	0.00
MRF (B=8)	0.02	0.18	0.54	1.00	0.06	1.00	0.23	0.35	0.54	0.24	0.81	0.00	0.04	0.00	0.00
Best model	RF/OLS	LLF1	RF	RF grf	RF grf	RF	RF grf	adalasso/RF	RF	RF grf	RF grf	RF	LLF1	RF grf	RF grf

Table 10

The p-values for the MCSs based on the $t_{\max}$ statistic for square losses obtained from forecasts over the out-of-sample period from 1990-2015.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.20	0.00	0.00	0.37	0.26
AR	0.13	0.00	0.39	0.41	0.34	0.30	0.39	0.67	0.65	0.84	0.65	0.00	0.11	0.12	0.00
RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RF grf	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
honest RF	0.92	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.86	0.82	0.00
RF/OLS	1.00	1.00	1.00	0.82	0.79	0.83	0.78	0.96	0.94	1.00	1.00	0.93	1.00	1.00	0.12
adaLASSO/RF	1.00	0.90	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF1	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF2	0.25	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.59	1.00	1.00	1.00
LLF3	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00
MRF (B=1)	0.48	0.35	0.21	0.77	0.16	0.18	0.00	0.74	0.00	1.00	0.64	0.00	0.29	0.61	0.00
MRF (B=8)	1.00	0.74	0.93	0.84	0.51	0.96	0.99	0.35	0.27	1.00	1.00	0.00	0.58	0.83	0.00
Best model	adalasso/RF	RF	RF	RF grf	RF grf	RF	LLF1	adalasso/RF	LLF1	RF	RF	RF	RF	RF	RF grf

Table 11

The p-values for the MCSs based on the t_R statistic for square losses obtained from forecasts over the out-of-sample period from 1990-2015.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.36	0.00	0.00	0.67	0.32
AR	0.04	0.00	0.53	1.00	0.11	0.37	0.51	0.56	0.71	0.74	0.09	0.00	0.13	0.28	0.00
RF	0.90	0.94	1.00	1.00	0.98	1.00	0.78	1.00	1.00	1.00	0.99	0.58	1.00	1.00	0.35
RF grf	1.00	1.00	1.00	1.00	1.00	0.99	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
honest RF	0.27	0.69	0.79	1.00	0.48	0.42	0.31	0.58	0.40	0.65	0.10	0.13	0.17	0.04	0.00
RF/OLS	1.00	0.88	0.91	1.00	0.84	0.75	0.94	0.90	0.90	1.00	0.96	0.58	0.69	0.28	0.17
adaLASSO/RF	0.87	0.72	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.32	0.14	0.58	0.69	0.53	0.09
LLF1	0.04	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.58	1.00	1.00	1.00
LLF2	0.03	0.04	0.24	1.00	0.48	0.99	0.94	1.00	1.00	0.99	0.96	0.48	0.69	0.67	1.00
LLF3	0.00	0.94	1.00	0.11	1.00	0.99	1.00	1.00	0.90	1.00	1.00	0.58	1.00	1.00	1.00
MRF (B=1)	0.04	0.54	0.15	1.00	0.01	0.09	0.00	0.25	0.00	0.84	0.43	0.00	0.16	0.00	0.00
MRF (B=8)	0.18	0.83	0.76	1.00	0.12	0.46	0.37	0.09	0.11	0.99	1.00	0.00	0.01	0.00	0.00
Best model	RF/OLS	RF grf	RF grf	LLF1	LLF1	LLF1	LLF1	adalasso/RF	LLF1	RF grf	RF grf	RF grf	RF grf	LLF1	RF grf

Table 12

The p-values for the MCSs based on the $t_{\max}$ statistic for absolute losses obtained from forecasts over the out-of-sample period from 1990-2000.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.16	0.00	0.52	0.00	0.00	0.17	0.00	0.00	0.57	0.83	0.93	0.00	0.89	1.00	1.00
AR	1.00	0.96	1.00	0.71	0.67	1.00	1.00	1.00	0.85	0.62	0.00	0.00	0.19	0.00	0.00
RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RF grf	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
honest RF	0.88	0.42	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.43	0.00	0.00	0.00	0.00	0.00
RF/OLS	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.21	1.00	0.49	0.10
adaLASSO/RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.15	1.00	1.00	0.99	1.00
LLF1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF2	1.00	1.00	0.98	0.62	0.85	0.99	0.83	0.90	0.96	1.00	1.00	0.99	1.00	0.46	0.76
LLF3	0.30	0.99	1.00	0.60	0.48	1.00	0.72	0.13	0.87	1.00	1.00	1.00	1.00	1.00	1.00
MRF (B=1)	0.21	0.00	0.41	0.00	0.54	0.44	0.32	0.56	0.61	0.89	0.34	0.00	0.00	0.00	0.00
MRF (B=8)	1.00	0.36	1.00	0.17	0.44	1.00	0.36	0.00	1.00	0.40	1.00	0.25	0.45	0.00	0.00
Best model	RF	RF	RF grf	RF grf	RF grf	RF	RF grf	RF grf	RF grf	RF grf	RF	RF	LLF1	RW	RF grf

Table 13

The p-values for the MCSs based on the t_R statistic for absolute losses obtained from forecasts over the out-of-sample period from 1990-2000.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.02	0.00	0.49	0.00	0.00	0.10	0.00	0.00	0.10	0.30	0.35	0.00	0.92	1.00	1.00
AR	0.02	0.57	0.76	0.16	0.03	0.55	0.18	1.00	0.02	0.00	0.00	0.00	0.21	0.00	0.00
RF	0.04	0.83	0.76	0.16	0.25	0.55	0.18	1.00	0.10	0.98	0.99	1.00	0.99	0.60	0.01
RF grf	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00
honest RF	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RF/OLS	0.04	0.42	1.00	0.16	1.00	0.55	0.18	1.00	0.10	0.85	0.56	0.12	0.51	0.56	0.00
adaLASSO/RF	0.04	0.83	0.76	1.00	1.00	0.55	0.18	1.00	1.00	0.00	0.01	0.98	0.96	0.99	0.69
LLF1	0.04	1.00	0.76	1.00	1.00	1.00	0.18	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00
LLF2	0.04	0.14	0.33	0.05	0.18	0.00	0.18	1.00	0.10	1.00	0.96	0.56	0.92	0.11	0.69
LLF3	0.04	0.83	0.76	0.16	0.25	0.55	0.18	0.00	0.10	1.00	0.99	0.98	0.92	0.99	1.00
MRF (B=1)	0.02	0.00	0.28	0.00	0.18	0.02	0.13	1.00	0.01	0.30	0.24	0.00	0.00	0.00	0.00
MRF (B=8)	0.04	0.06	0.76	0.06	0.19	0.55	0.15	0.00	0.10	0.00	0.91	0.04	0.09	0.00	0.00
Best model	RF grf	LLF1	RF grf	RF grf	RF grf	RF grf	RF grf	RF grf	RF grf	RF grf	RF grf	RF	LLF1	RW	RF grf

Table 14

The p-values for the MCSs based on the $t_{\max}$ statistic for square losses obtained from forecasts over the out-of-sample period from 1990-2000.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.00	0.36	0.00	0.00	0.13	0.00	0.00	0.47	0.25	0.60	0.00	0.95	1.00	0.97
AR	0.78	0.96	1.00	0.95	0.98	1.00	1.00	1.00	1.00	0.99	0.33	0.00	0.48	0.00	0.00
RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RF grf	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
honest RF	0.89	1.00	0.92	0.61	0.57	1.00	0.96	0.29	0.87	0.94	0.49	0.00	0.60	0.00	0.00
RF/OLS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.40	1.00	0.29	0.28
adaLASSO/RF	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.98	1.00
LLF1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF2	1.00	1.00	0.90	0.80	0.91	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.86	0.95
LLF3	0.20	1.00	1.00	0.99	1.00	1.00	0.88	0.42	0.91	1.00	1.00	1.00	1.00	1.00	1.00
MRF (B=1)	0.10	0.19	0.95	0.72	0.50	0.23	0.22	0.27	0.38	1.00	0.39	0.00	0.11	0.00	0.00
MRF (B=8)	1.00	0.49	1.00	0.89	0.23	1.00	0.27	0.00	1.00	0.41	1.00	0.18	0.72	0.00	0.00
Best model	RF	RF	RF	RF grf	RF grf	RF	RF grf	RF	RF grf	RF grf	RF	RF	RF	RF grf	RF grf

Table 15

The p-values for the MCSs based on the t_R statistic for square losses obtained from forecasts over the out-of-sample period from 1990-2000.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.00	0.42	0.00	0.00	0.14	0.00	0.00	0.37	0.22	0.42	0.00	0.91	0.87	0.86
AR	0.03	0.10	0.94	0.45	0.05	0.95	0.64	0.35	0.37	0.22	0.01	0.00	0.18	0.00	0.00
RF	0.09	0.73	0.94	0.91	0.75	1.00	0.66	0.21	0.56	1.00	0.99	1.00	1.00	0.53	0.09
RF grf	1.00	0.73	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.87	1.00
honest RF	0.00	0.10	0.03	0.07	0.03	0.19	0.13	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
RF/OLS	0.09	0.53	0.94	0.45	1.00	0.95	0.66	0.35	0.56	0.98	0.88	0.35	0.70	0.08	0.07
adaLASSO/RF	0.09	0.73	0.45	1.00	1.00	0.95	0.58	1.00	1.00	0.11	0.59	0.97	0.90	0.48	0.07
LLF1	0.09	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.87	0.86
LLF2	0.09	0.53	0.11	0.01	0.02	0.00	0.05	0.21	0.56	1.00	0.88	0.46	0.90	0.14	0.68
LLF3	0.09	0.10	0.94	0.45	0.75	0.95	0.21	0.11	0.37	1.00	0.99	0.97	1.00	1.00	0.86
MRF (B=1)	0.06	0.10	0.50	0.06	0.01	0.01	0.21	0.11	0.21	0.98	0.10	0.00	0.08	0.00	0.00
MRF (B=8)	0.09	0.03	1.00	0.45	0.03	0.95	0.10	0.00	0.56	0.04	0.88	0.01	0.09	0.00	0.00
Best model	RF grf	LLF1	RF grf	LLF1	RF grf	RF	LLF1	RF grf	RF grf	RF grf	LLF1	RF	LLF1	LLF3	RF grf

Table 16

The p-values for the MCSs based on the $t_{\max}$ statistic for absolute losses obtained from forecasts over the out-of-sample period from 2001-2015.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.43	0.61	0.46	0.71	0.79	0.81	0.71	0.67	0.62	0.71	0.76	0.64	0.90	0.69
AR	0.00	0.00	0.11	0.51	0.25	0.18	0.29	0.58	0.46	0.55	0.31	0.44	1.00	0.00	0.00
RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RF grf	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
honest RF	0.19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.44
RF/OLS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.56
adaLASSO/RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF1	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF2	0.00	0.44	0.48	0.93	1.00	1.00	0.64	0.40	1.00	1.00	0.69	0.84	0.73	1.00	1.00
LLF3	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MRF (B=1)	0.50	0.66	0.93	1.00	0.13	1.00	0.78	1.00	0.51	1.00	0.48	0.14	0.77	0.40	0.00
MRF (B=8)	0.48	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.43	0.10
Best model	RF/OLS	RF	honest RF	RF grf	RF grf	adaLASSO/RF	RF grf	RF	RF	RF grf	RF	RF	RF	RF	RF grf

Table 17

The p-values for the MCSs based on the t_R statistic for absolute losses obtained from forecasts over the out-of-sample period from 2001-2015.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.36	0.39	0.34	0.35	0.55	0.51	0.59	0.45	0.14	0.12	0.25	0.69	0.31	0.11
AR	0.00	0.00	0.01	0.36	0.12	0.07	0.13	0.23	0.27	0.53	0.02	0.03	0.77	0.00	0.00
RF	0.43	1.00	0.98	0.87	0.89	0.59	0.70	0.99	1.00	0.97	0.82	0.87	0.77	0.31	0.11
RF grf	1.00	1.00	0.98	1.00	1.00	0.59	1.00	0.94	0.97	1.00	1.00	1.00	1.00	1.00	0.11
honest RF	0.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.77	0.31	0.01
RF/OLS	1.00	1.00	0.98	0.87	0.78	0.59	0.70	0.87	0.97	1.00	1.00	0.87	0.77	0.22	0.11
adaLASSO/RF	1.00	1.00	1.00	1.00	0.64	0.59	0.70	1.00	1.00	0.78	0.12	0.87	0.77	0.31	0.01
LLF1	0.00	1.00	0.98	1.00	1.00	0.59	1.00	0.98	1.00	1.00	1.00	0.87	0.77	0.31	0.11
LLF2	0.00	0.04	0.08	0.75	0.86	0.59	0.36	0.23	0.98	0.55	0.26	0.38	0.10	0.31	1.00
LLF3	0.00	1.00	0.98	0.87	1.00	0.59	0.70	0.99	1.00	1.00	1.00	0.38	0.77	0.31	0.11
MRF (B=1)	0.21	0.58	0.13	1.00	0.01	0.59	0.47	0.92	0.25	0.77	0.06	0.00	0.21	0.00	0.00
MRF (B=8)	0.02	0.89	0.62	1.00	0.65	0.59	0.70	0.98	0.75	1.00	0.81	0.08	0.24	0.00	0.00
Best model	RF/OLS	honest RF	honest RF	RF grf	RF grf	honest RF	RF grf	adaLASSO/RF	RF	RF grf	RF grf	RF grf	RF grf	RF grf	LLF2

Table 18

The p-values for the MCSs based on the $t_{\max}$ statistic for square losses obtained from forecasts over the out-of-sample period from 2001-2015.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.44	1.00	0.97
AR	0.27	0.17	0.44	0.50	0.44	0.31	0.51	0.68	0.59	0.68	0.63	0.60	0.17	0.25	0.16
RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RF grf	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
honest RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.98
RF/OLS	1.00	1.00	1.00	0.81	0.75	0.70	0.87	0.92	0.94	1.00	1.00	1.00	1.00	0.91	0.79
adaLASSO/RF	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	0.99	0.81	1.00	1.00	1.00	1.00
LLF1	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF2	0.30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLF3	0.34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MRF (B=1)	0.91	0.71	0.54	0.94	0.51	0.62	0.32	0.91	0.00	0.94	0.66	0.13	0.83	0.63	0.73
MRF (B=8)	1.00	0.98	0.87	0.97	0.96	0.95	1.00	0.87	0.45	1.00	1.00	0.99	1.00	0.70	0.85
Best model	adalasso/RF	RF	RF	RF grf	RF grf	RF	LLF1	LLF1	LLF1	RF grf	RF	RF	RF	RF grf	RF

Table 19

The p-values for the MCSs based on the t_R statistic for square losses obtained from forecasts over the out-of-sample period from 2001-2015.

	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.00	0.95	1.00	0.98	0.93	0.97	0.97	0.91	0.94	0.96	1.00	0.82	0.42	0.95	0.77
AR	1.00	0.21	0.32	0.61	0.16	0.30	0.57	0.57	0.68	0.67	0.12	0.11	0.13	0.29	0.16
RF	1.00	1.00	1.00	1.00	1.00	1.00	0.84	1.00	1.00	1.00	1.00	0.82	1.00	1.00	0.98
RF grf	1.00	1.00	1.00	1.00	1.00	0.68	0.87	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.98
honest RF	1.00	0.99	1.00	0.99	0.98	0.92	0.87	0.98	1.00	0.96	0.66	0.82	0.78	0.33	0.19
RF/OLS	1.00	0.98	0.73	0.58	0.68	0.58	0.87	0.89	0.94	0.96	0.99	0.82	0.46	0.39	0.43
adaLASSO/RF	1.00	0.92	1.00	1.00	0.96	0.99	0.97	1.00	1.00	0.90	0.36	0.82	0.97	0.53	0.16
LLF1	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.99	0.82	1.00	1.00	0.98
LLF2	1.00	0.15	0.61	0.99	1.00	1.00	0.87	0.98	1.00	1.00	0.99	0.68	1.00	0.95	1.00
LLF3	1.00	1.00	1.00	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	1.00	0.95	0.98
MRF (B=1)	1.00	0.79	0.05	0.87	0.18	0.38	0.08	0.81	0.00	0.58	0.57	0.05	0.52	0.01	0.02
MRF (B=8)	1.00	0.97	0.37	0.84	0.64	0.51	0.97	0.55	0.25	1.00	1.00	0.08	0.08	0.00	0.02
Best model	RF/OLS	RF grf	adalasso/RF	LLF1	LLF1	LLF1	LLF1	LLF3	LLF1	RF grf	RF grf	RF grf	RF grf	LLF1	LLF2

A.2 Results Diebold Mariano tests

Table 20

The p-values of the Diebold Mariano tests for absolute losses for the out-of-sample period from 1990-2000.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.24										
RF	0.07	0.04*									
RF grf	0.04*	0.05	0.32								
honest RF	0.31	0.21	0.00**	0.01**							
RF/OLS	0.16	0.31	0.32	0.22	0.10						
adaLASSO/RF	0.18	0.27	0.47	0.30	0.08	0.47					
LLF1	0.20	0.33	0.65	0.52	0.05	0.51	0.62				
LLF2	0.25	0.53	0.26	0.20	0.34	0.47	0.38	0.15			
LLF3	0.26	0.46	0.21	0.18	0.33	0.27	0.48	0.31	0.69		
MRF (B=1)	0.28	0.41	0.02	0.01	0.50	0.14	0.13	0.10	0.40	0.42	
MRF (B=8)	0.20	0.48	0.10	0.22	0.25	0.36	0.24	0.30	0.47	0.51	0.27

Table 21

The p-values of the Diebold Mariano tests for square losses for the out-of-sample period from 1990-2000.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.14										
RF	0.08	0.13									
RF grf	0.04*	0.14	0.38								
honest RF	0.15	0.53	0.00**	0.02*							
RF/OLS	0.15	0.34	0.37	0.29	0.13						
adaLASSO/RF	0.14	0.38	0.46	0.35	0.24	0.52					
LLF1	0.13	0.30	0.75	0.76	0.09	0.45	0.57				
LLF2	0.23	0.47	0.36	0.27	0.53	0.57	0.44	0.08			
LLF3	0.12	0.47	0.25	0.24	0.52	0.39	0.35	0.20	0.49		
MRF (B=1)	0.16	0.36	0.05*	0.03*	0.51	0.10	0.15	0.93	0.37	0.36	
MRF (B=8)	0.12	0.50	0.13	0.29	0.44	0.34	0.29	0.22	0.46	0.45	0.20

Table 22

The p-values of the Diebold Mariano tests for absolute losses for the out-of-sample period from 2001-2015.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.58										
RF	0.04*	0.02*									
RF grf	0.05*	0.03*	0.42								
honest RF	0.07	0.06	0.51	0.44							
RF/OLS	0.34	0.12	0.33	0.35	0.46						
adaLASSO/RF	0.06	0.04*	0.41	0.32	0.48	0.40					
LLF1	0.06	0.10	0.65	0.63	0.59	0.44	0.45				
LLF2	0.38	0.29	0.16	0.19	0.25	0.50	0.22	0.17			
LLF3	0.12	0.16	0.45	0.43	0.52	0.61	0.45	0.48	0.32		
MRF (B=1)	0.52	0.34	0.14	0.16	0.18	0.50	0.17	0.27	0.35	0.56	
MRF (B=8)	0.71	0.09	0.31	0.35	0.35	0.63	0.42	0.44	0.56	0.44	0.34

Table 23

The p-values of the Diebold Mariano tests for square losses for the out-of-sample period from 2001-2015.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.41										
RF	0.55	0.06									
RF grf	0.41	0.09	0.39								
honest RF	0.75	0.13	0.48	0.42							
RF/OLS	0.74	0.12	0.32	0.25	0.58						
adaLASSO/RF	0.61	0.10	0.44	0.43	0.48	0.35					
LLF1	0.40	0.18	0.49	0.44	0.53	0.39	0.42				
LLF2	0.39	0.25	0.59	0.54	0.61	0.44	0.60	0.54			
LLF3	0.31	0.24	0.64	0.59	0.60	0.43	0.54	0.56	0.55		
MRF (B=1)	0.56	0.36	0.09	0.11	0.29	0.52	0.22	0.22	0.27	0.35	
MRF (B=8)	0.73	0.13	0.23	0.25	0.48	0.59	0.31	0.30	0.43	0.39	0.39

A.3 Results Diebold Mariano tests with HLN correction

Table 24

The p-values of the Diebold Mariano tests with HLN correction for absolute losses for the out-of-sample period from 1990-2015.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.02*										
RF	0.00**	0.00**									
RF grf	0.00**	0.01**	0.40								
honest RF	0.04*	0.13	0.05	0.09							
RF/OLS	0.03*	0.05*	0.19	0.19	0.63						
adaLASSO/RF	0.00**	0.02*	0.42	0.35	0.23	0.38					
LLF1	0.00**	0.09	0.64	0.47	0.31	0.32	0.46				
LLF2	0.00**	0.30	0.10	0.08	0.50	0.46	0.22	0.09			
LLF3	0.01**	0.17	0.26	0.14	0.54	0.50	0.29	0.26	0.36		
MRF (B=1)	0.06	0.50	0.02*	0.02*	0.32	0.25	0.06	0.14	0.29	0.43	
MRF (B=8)	0.07	0.11	0.07	0.19	0.58	0.58	0.22	0.23	0.47	0.56	0.19

Table 25

The p-values of the Diebold Mariano tests with HLN correction for square losses for the out-of-sample period from 1990-2015.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.04*										
RF	0.01**	0.04*									
RF grf	0.01**	0.06	0.37								
honest RF	0.02*	0.13	0.06	0.17							
RF/OLS	0.02*	0.09	0.22	0.19	0.68						
adaLASSO/RF	0.01*	0.08	0.39	0.41	0.31	0.33					
LLF1	0.01*	0.12	0.47	0.43	0.36	0.31	0.37				
LLF2	0.02*	0.25	0.62	0.48	0.62	0.48	0.58	0.35			
LLF3	0.02*	0.21	0.67	0.59	0.56	0.48	0.56	0.47	0.52		
MRF (B=1)	0.04*	0.36	0.03	0.04	0.30	0.39	0.16	0.12	0.23	0.29	
MRF (B=8)	0.02*	0.13	0.15	0.19	0.51	0.63	0.24	0.25	0.39	0.38	0.33

Table 26

The p-values of the Diebold Mariano tests with HLN correction for absolute losses for the out-of-sample period from 1990-2000.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.22										
RF	0.06	0.03*									
RF grf	0.03*	0.05*	0.30								
honest RF	0.29	0.19	0.00**	0.00**							
RF/OLS	0.14	0.29	0.29	0.19	0.07						
adaLASSO/RF	0.16	0.25	0.45	0.29	0.06	0.45					
LLF1	0.17	0.31	0.63	0.50	0.04*	0.49	0.61				
LLF2	0.22	0.52	0.23	0.17	0.32	0.44	0.34	0.13			
LLF3	0.23	0.43	0.20	0.16	0.30	0.24	0.29	0.29	0.68		
MRF (B=1)	0.26	0.38	0.01*	0.01**	0.48	0.10	0.11	0.08	0.38	0.40	
MRF (B=8)	0.17	0.47	0.09	0.19	0.23	0.34	0.26	0.27	0.44	0.48	0.27

Table 27

The p-values of the Diebold Mariano tests with HLN correction for square losses for the out-of-sample period from 1990-2000.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.12										
RF	0.07	0.12									
RF grf	0.03*	0.12	0.35								
honest RF	0.13	0.51	0.00**	0.01*							
RF/OLS	0.14	0.32	0.34	0.26	0.10						
adaLASSO/RF	0.12	0.36	0.44	0.33	0.23	0.49					
LLF1	0.10	0.27	0.74	0.75	0.07	0.42	0.56				
LLF2	0.20	0.44	0.32	0.24	0.50	0.55	0.40	0.07			
LLF3	0.10	0.44	0.22	0.21	0.49	0.36	0.32	0.17	0.47		
MRF (B=1)	0.14	0.34	0.04*	0.02*	0.64	0.08	0.13	0.09	0.35	0.43	
MRF (B=8)	0.10	0.48	0.11	0.26	0.42	0.32	0.22	0.20	0.43	0.43	0.18

Table 28

The p-values of the Diebold Mariano tests with HLN correction for absolute losses for the out-of-sample period from 2001-2015.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.57										
RF	0.03*	0.01*									
RF grf	0.04*	0.02*	0.39								
honest RF	0.53	0.05	0.50	0.41							
RF/OLS	0.31	0.10	0.30	0.32	0.43						
adaLASSO/RF	0.04*	0.03*	0.33	0.30	0.32	0.38					
LLF1	0.05	0.08	0.63	0.62	0.58	0.42	0.43				
LLF2	0.36	0.27	0.14	0.17	0.23	0.49	0.20	0.15			
LLF3	0.11	0.14	0.43	0.41	0.50	0.59	0.42	0.47	0.30		
MRF (B=1)	0.50	0.31	0.13	0.15	0.17	0.48	0.15	0.25	0.33	0.53	
MRF (B=8)	0.62	0.07	0.30	0.32	0.33	0.61	0.39	0.42	0.54	0.41	0.33

Table 29

The p-values of the Diebold Mariano tests with HLN correction for square losses for the out-of-sample period from 2001-2015.

Model	RW	AR	RF	RF grf	honest RF	RF/OLS	adaLASSO/RF	LLF1	LLF2	LLF3	MRF(B=1)
AR	0.38										
RF	0.53	0.05									
RF grf	0.39	0.07	0.36								
honest RF	0.06	0.11	0.45	0.40							
RF/OLS	0.73	0.10	0.29	0.23	0.56						
adaLASSO/RF	0.60	0.09	0.42	0.42	0.48	0.33					
LLF1	0.39	0.16	0.46	0.41	0.52	0.37	0.40				
LLF2	0.36	0.22	0.57	0.52	0.59	0.42	0.57	0.53			
LLF3	0.29	0.21	0.63	0.57	0.58	0.41	0.52	0.53	0.54		
MRF (B=1)	0.52	0.33	0.08	0.09	0.27	0.51	0.21	0.20	0.25	0.32	
MRF (B=8)	0.71	0.11	0.22	0.23	0.46	0.57	0.29	0.28	0.41	0.24	0.38

B Model specification

B.1 Tuning parameters and packages

For all replication models shown in Table 6, we used the tuning parameters as described in Medeiros, Vasconcelos, Veiga, et al. (2021). In order to obtain similar results for the RF obtained from the **grf** package as for the RF obtained from the **RandomForest** package, we set the number of trees used for each forest to 500, the number of trees grown on each subsample equal to 1 and the number of variables tried for each split equal to the number of covariates divided by 3. These same settings were used for the honest RF. For the LLFs, we used the default parameters of the **grf** package for the LLF1 model. For the LLF2, we set the tuning parameters such that the model remained efficient for the large data set we used. Therefore, we set the ridge penalty for prediction to $\lambda = 0.1$ and the split cutoff parameter to 50. For LLF3, we also set the split cutoff parameter to 50, but we set the. Setting the split cutoff parameter to 50 means that when leaves reach a size of 50, regression coefficients of the full data set are used for splitting. For the other tuning parameters, the default settings were used. The **MRF** package by Goulet Coulombe (2020) was used for the MRFs. We set the trend.push parameter to 4, as Goulet Coulombe (2020) states that this is sensible for macro data. Furhtermore, we set the fraction of all variables in $\mathcal{S}_t$ considered for each split to 0.15, as this is useful for efficient estimation when $\mathcal{S}_t$ contains many correlated variables. We also increase efficiency by setting the quantile.rate variable to 0.3, such that one out of every three splitting points is considered for splitting instead of all splitting points.

B.2 Benchmark models

B.2.1 Random walk model (RW)

The first benchmark is given by the random walk (RW) model, which is defined as $\hat{\pi}_t = \pi_{t-h} + u_t$. Here the error terms u_t are an independent and identically distributed white noise series with a mean of zero. The forecasts of the RW model are specified as $\hat{\pi}_{t+h|t} = \pi_t$, for $h = 1, ..., 12$. Here, $\hat{\pi}_{t+h}$ is the h-month ahead inflation forecast and π_t represents the inflation at time t. The cumulative h-month ahead forecast is given by $\hat{\pi}_{t+1:t+h|t} = \pi_{t-(h-1):t}$, with $\pi_{t-(h-1):t}$ the accumulated inflation over the h months preceding month t.

B.2.2 Autoregressive model (AR)

As a second benchmark model, an autoregressive (AR) model is used with lag order p. This model is specified as follows:

$$\pi_t = c + \phi_{1,h}\pi_{t-1} + ... + \phi_{p,h}\pi_{t-p} + \varepsilon_t, \text{ for } h = 1, ..., 12. \tag{13}$$

The residuals ε_t are assumed to follow a white noise process. In order to determine the number of lags p, ordinary least squares (OLS) is performed and the values of the Bayesian Information Criterion (BIC) are compared for different lags. The inflation forecasts of the AR model for each horizon are specified as $\hat{\pi}_{t+h|t} = c + \phi_{1,h}\pi_t + ... + \phi_{p,h}\pi_{t-p+1}$. Aggregating these individual forecasts gives the accumulated forecasts.

B.3 Shrinkage models

For the shrinkage models, we formulate $G_h(\boldsymbol{x}_t)$ from Equation 1 as a linear combination of covariates, thus $G_h(\mathbf{x}_t) = \boldsymbol{\beta}_h' \boldsymbol{x}_t$. We specify the estimate of the parameter $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}_h} \left[\sum_{t=1}^{T-h} (y_{t+h} - \boldsymbol{\beta}_h' \mathbf{x}_t)^2 + \sum_{i=1}^{n} p(\beta_{h,i}; \lambda, \omega_i) \right], \tag{14}$$

where horizon h takes on values $h = 1, ..., 12$, T denotes the total number of observations and n the number of covariates. Moreover, $p(\beta_{h,i}; \lambda, \omega_i)$ denotes a penalty function depends on both a penalty parameter λ and a weight $\omega_i > 0$.

B.3.1 Ridge regression (RR)

For the RR, proposed by Hoerl and Kennard (1970), the penalty is specified as

$$p(\beta_{h,i}; \lambda, \omega_i) = \lambda \sum_{i=1}^{n} \beta_{h,i}^2. \tag{15}$$

The advantages of the RR are that it has an exact solution that can be easily computed and that has to ability to shrink the coefficients of less-relevant variables to almost zero.

B.3.2 Least absolute shrinkage and selection operator (LASSO)

The LASSO model proposed by Tibshirani (1996) is given by

$$p(\beta_{h,i}; \lambda, \omega_i) = \lambda \sum_{i=1}^{n} |\beta_{h,i}|. \tag{16}$$

LASSO has the advantage of shrinking irrelevant variables to zero. A disadvantage, however, is that the consistency of the model selection only applies under very strict conditions.

B.3.3 Adaptive least absolute shrinkage and selection operator (adaLASSO)

The adaLASSO was proposed by Zou (2006) as a way to achieve consistency for model selection. To obtain this consistency, a weight parameter originating from a first-step estimation is added to the LASSO penalty. This results in the following penalty:

$$p(\beta_{h,i}; \lambda, \omega_i) = \lambda \sum_{i=1}^{n} \omega_i |\beta_{h,i}|, \tag{17}$$

where $\omega_i = |\beta^*_{h,i}|$, with $\beta^*_{h,i}$ the the first-step estimation coefficient. According to Medeiros and Mendes (2016), advantages of the adaLASSO over the LASSO method are that it can handle a larger amount of variables than observations and that it performs well in non-Gaussian environments and under heteroskedasticity.

B.3.4 Elastic net (ElNet)

ElNet is a generalization that combines both the RR and the LASSO, forming a convex combination of the ℓ_1 and ℓ_2 norms as described by Zou and Hastie (2005). The penalty is given by

$$p(\beta_{h,i}; \lambda, \omega_i) = \alpha\lambda \sum_{i=1}^{n} \beta^2_{h,i} + (1-\alpha)\lambda \sum_{i=1}^{n} |\beta_{h,i}|, \tag{18}$$

where $\alpha \in [0, 1]$. The penalty of the adaptive ElNet approach works similar to the adaLASSO.

B.4 Factor models

Bai (2003) provide an extensive theory for factor models. Factor models reduce the dimension of the model by combining the common terms from all covariates into common factors. These factors are constructed by computing the principal components of the set of variables $_t$, such that $\boldsymbol{P}_t = \boldsymbol{A}\boldsymbol{z}_t$. Here, $\boldsymbol{P}_t$ denotes a vector of principal components and $\boldsymbol{A}$ denotes a rotation matrix. If we use this for the model in Equation 1, we get that $\boldsymbol{x}_t$ is equal to π_{t-j}, for $j = 0, 1, 2, 3$, plus $\boldsymbol{f}_{t-j}$, for j=0,1,2,3. Here, $\mathbf{f}_t$ is defined as a vector containing the first four principal components of $\boldsymbol{z}_t$.

B.4.1 Target factors

Target factors were proposed by Bai and Ng (2008) as a method to improve the prediction ability of factor models. The key concept of this model is that it does not include all variables in $\boldsymbol{z}_t$ in order to construct the factors, but selects only the relevant variables with a substantial forecasting power. In Section 4.3.1 of Medeiros, Vasconcelos, Veiga, et al. (2021), the forecasting algorithm of this model is described in detail.

B.4.2 Factor boosting

Bai and Ng (2008) also describe a factor model called factor boosting. This model selects the relevant factors instead of the relevant variables in $\boldsymbol{z}_t$. Following the approach in Section 4.3.2

of Medeiros, Vasconcelos, Veiga, et al. (2021), we take the boosting procedure of Bai and Ng (2008) for selecting both the factors and the number of lags in the model.

B.5 Ensemble methods

The term ensemble methods includes all models that compute forecasts as a (weighted) average of the forecasts of several methods.

B.5.1 Bagging

Bootstrap aggregating, or bagging, has been introduced by Breiman (1996) as a way to combine the forecasts of an ensemble of methods. In the algorithm, three steps are carried out for B bootstrap samples. As a first step, OLS is performed with all candidate variables and only the variables with an absolute t-statistic above threshold c are selected. Secondly, the selected variables are used to perform an OLS regression. Thirdly, the OLS coefficients from the second step are used to compute forecasts on the original data set.

As the number of variables is larger than the number of observations, performing OLS in the first step is unfeasible. To account for this, we follow Medeiros, Vasconcelos, Veiga, et al. (2021) by, for each sample B, dividing the variables into randomly created groups and carrying out the first step for each of the groups.

B.5.2 CSR

In order to select an optimal subset of predictors in a computationally efficient way that did not require testing all possible combinations of regressors, Elliott et al. (2013) and Elliott et al. (2015) proposed the CSR method. The key concept is to select a number qn and run regressions with all possible combinations with size q of the n variables. The final prediction is constructed as an average of the forecasts of these regressions with q covariates. Since this algorithm loses efficiency for a large number of covariates, an adjustment is necessary. We follow Medeiros, Vasconcelos, Veiga, et al. (2021) and add two steps prior to performing the CSR algorithm. First, the t-statistic for the regression of each of the candidate covariates is determined. Second, the covariates are ranked based on the absolute t-statistics and the $\tilde{n}$ most relevant are selected.

B.6 Derivations

B.6.1 LLF weight equation

The predictions of the RF are computed by firstly determining the sum of all values of Y_i that correspond to a training data point X_i that falls into the same leaf as x and dividing this by the total number of training data points falling in the same leaf as x. Secondly, we take the mean of all forests $b = 1, ..., B$, which results in the following final formulation for the RF predictions:

$$\hat{\pi}_{t+h} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} Y_i \frac{I_b(X_i)}{|L_b(x)|} = \sum_{i=1}^{n} Y_i \frac{1}{B}\sum_{b=1}^{B}\frac{I_b(X_i)}{|L_b(x)|} = \sum_{i=1}^{n}\alpha_i(x)Y_i, \tag{19}$$

where the indicator function $I_b(X_i)$ is specified as

$$I_b(X_i) = \begin{cases} 1 & \text{if } X_i \in L_b(x), \\ 0 & \text{otherwise.} \end{cases}$$

B.6.2 Regularisation Transformation

The kernel used by WLS assigns a weight of 1 to observation t, a weight of $\zeta < 1$ to observations $t-1$ and $t+1$ and a weight of ζ^2 to observations $t-2$ and $t+2$. All other observations are assigned a weight of zero. As some t's might qualify to be assigned multiple weights, we take the maximal allocated weight as the final weight. For a given subsample in leaf l, final weights $w(t;\zeta)$ can thus be specified as

$$w(t;\zeta) = \begin{cases} 1, & \text{if } t \in l, \\ \zeta, & \text{if } t \in (l_{+1} \cup l_{-1})/l, \\ \zeta^2, & \text{if } t \in (l_{+2} \cup l_{-2})/(l \cup (l_{+1} \cup l_{-1})), \\ 0, & \text{otherwise.} \end{cases}$$

Here, l_{-1} denotes the set containing each observation from l, but lagged one time period. l_{+1} denotes the set containing the one step forwarded observations from l. l_{-2} and l_{+2} denote the two-steps lagged and forwarded observations respectively. Furthermore, the parameter ζ indicates the degree of time-smoothness, for which it holds that $\zeta < 1$. Whereas we had the two splitting sets $l_1(j,c) \equiv \{t \in l | S_{j,t} \leq c\}$ and $l_2(j,c) \equiv \{t \in l | S_{j,t} > c\}$ in Equation 12, we expand these for regularisation by introducing the following two splitting sets:

$$l_i^R(j,c) \ \equiv \ l_i(j,c) \ \cup \ l_i(j,c)_{-1} \ \cup \ l_i(j,c)_{+1} \ \cup \ l_i(j,c)_{-2} \ \cup \ l_i(j,c)_{+2},$$

with $l_1^R(j,c)$ and $l_2^R(j,c)$ the splitting sets after regularisation. The splits can be determined by solving the following problem:

$$\min_{j \in \mathcal{J}^-,\, c \in \mathbb{R}} \left\{ \min_{\theta_1} \sum_{t \in l_1^R(j,c)} w(t;\zeta)(Y_t - X_t\theta_1)^2 + \lambda ||\theta_1||_2^2 \right.$$
$$\left. + \min_{\theta_2} \sum_{t \in l_2^R(j,c)} w(t;\zeta)(Y_t - X_t\theta_2)^2 + \lambda ||\theta_2||_2^2 \right\}.$$

Note that this problem is equal to the problem in Equation 12 when $\zeta \to 0$ and equal to the standard RF when $X_t = \iota$ and $\lambda = \zeta = 0$. That is, the standard RF does not have within-leaf shrinkage and only regresses on a constant.