ERASMUS SCHOOL OF ECONOMICS

BACHELOR THESIS ECONOMETRICS AND OPERATIONS RESEARCH

An extensive overview of machine learning methods for forecasting inflation

Author
J.I.P. (Jip) Witteveen 498788jw

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

July 4 , 2021

Abstract

This paper built further on the work of Medeiros et al. (2021) and provides an extensive overview and analysis of ML method to obtain accurate inflation forecasts. The United States inflation forecasting predictions are made for a large monthly frequently macroeconomic data set from the FRED-MD database for a out-of-sample period of more than 20 years. Several criteria are used to compare the models, including RMSE, MAE and MAD. Moreover, the forecasts are evaluated on their bias by a Mincer Zarnowitz regression. Finally, model confidence sets are applied to test if forecasts significantly differ across the models. The results suggest that, the extensive models, LLF, BART and MOTR-BART, provide similar accurate inflation forecasts as RFs based on MCSs. However, macroeconomic variables display nonlinear features and since these models can handle high-order interactions in predictors and perform well for nonlinear time series, the models have high potential in improve forecast accuracy.

Contents

1 Introduction

Inflation evaluating, forecasting and maintaining is of crucial interest because it aids policymakers in making decisions (Zahara et al. (2020)). Various parties, such as households, businesses, and policymakers, consider many contracts relating to employment, sales, tenancy, and debt that are influenced by inflation forecasting (Medeiros et al. (2021)). For central banks, one of the main tasks is maintaining price stability, that is low and stable inflation. In order for a policy to be effective, it must have correct inflation estimates. Otherwise, the program would lose its purpose.

Obtaining accurate inflation forecasts is a challenging task and therefore multiple models are introduced. Especially, there is a growing interest in using machine learning (ML) to improve the accuracy of economic and financial time series forecasts (Medeiros et al. (2021), Pavlov (2020), Rodríguez-Vargas (2020) and Masini et al. (2020)), for instance, implementing neural networks, decision trees and random forests, support-vector machines (SVM) and boosting. These models are commonly used and have potential improvements.

This paper builds on the work of Medeiros et al. (2021), who evaluated specific ML methods. The United States consumer price index (CPI) inflation forecasting results from the paper of Medeiros et al. (2021) are replicated for an out-of-sample period spanning 1990 to 2015. In this paper, a large monthly frequently macroeconomic data set is used from the FRED-MD database on the Michael McCraken's webpage. Additionally, this study includes and compares additional models, namely Bayesian additive regression trees BART and an extension of BART, BART with model trees (MOTR-BART), which is a well-performing ML model according to Prado et al. (2021). To overcome the weakness of RFs, also a local linear forest (LLF) is introduced that handles smoothness in the estimations of regression trees.

Several criteria are used to compare the models, including the root mean squared error (RMSE), mean absolute error (MAE), and median absolute deviation from the median (MAD), as suggested by Medeiros et al. (2021). The Mincer Zarnowitz (MZ) regression is used to test if the inflation forecasts are unbiased. Furthermore, the model confidence sets (MCSs) by Hansen (2005b) test if forecasts significantly differ per model.

The results in this paper find small difference in forecasts accuracy across RFs and the models. However, as nonlinearity and high-order interactions are presented in inflation forecasts and probably other macroeconomic variables, LLFs, BART and MOTR-BART have potential to improve more precise forecast values. Hereby, the paper extensive and build further on earlier research in forecasting macroeconomic variables.

The structure of this paper is as follows. In Section 2 we present a literature review, followed by a description of the used data set in Section 3. Section 4 provides an extensive description of the models, prediction calculation and predictive comparison criteria. Thereafter, in Section 5, the results are presented. Section 6 summarizes and concludes the findings. The report is further evaluated with a discussion in Section 7.

2 Literature

Recent literature examined ML methods to forecast accurate inflation values. As Pavlov (2020) emphasizes the importance of forecasting inflation, they examine several ML methods for forecasting inflation in Russia. In particular neural networks are evaluated and compared to the standard autoregressive model and the ridge regression ML method. Other work, Chakraborty and Joseph (2017), created a modeling framework to forecast UK CPI inflation on a medium-term horizon of two years with, among others, artificial neural networks, tree models, random forests and support vector machines.

Medeiros et al. (2021) compared several models to forecast US inflation over a sample period ranging from January 1960 to December 2015. Random walk (RW), autoregressive (AR), and

unobserved components stochastic volatility (UCSV) models are taken as benchmark models and are evaluated and compared to other models. Three different statistics over 15 time horizons are used to compare the models: the root mean squared error (RMSE), the mean absolute error (MAE), and the median absolute deviation from the median (MAD). Additionally, the models are tested to see if they produce different forecasts based on the model confidence sets (MCSs) (Hansen, Lunde, et al. (2011b)), superior predictive ability (SPA) tests (Hansen (2005a)), and the multi-horizon SPA test (Quaedvlieg (2017)). They show that machine learning (ML) models combined with the use of a large set of explanatory variables improve forecasting by 30% in terms of mean squared prediction error (MSPE), which is a robust and statistically significant result. As one of the additional models, Medeiros et al. (2021) examined and analyzed the random forest model (RF), an observable ML approach that outperforms all others. The RF model has the lowest RMSE, MAE, and MAD across all horizons, as well as the greatest MCS p-values, SPA test p-values, and multihorizon MCS p-values. Furthermore, at periods of greater uncertainty, as in a recession, the RF model produces substantially superior forecasts. This conclusion, according to Medeiros et al. (2021), is related to the fact that RF models allow for nonlinearities and thus outperform competitors during recessions.

In addition, when examining the entire out-of-sample window from January 1990 to December 2015, the RF model gives the least errors in the vast majority of cases, followed by the shrinkage model, the ridge regression (RR) model (Medeiros et al. (2021)). The RR, RF, and hybrid linear-Random Forest models (adaLASSO/RF and RF/OLS) are the only models that deviate from the benchmark models for all employed horizons.

According to Medeiros et al. (2021), the three ML models, RF, RR and adaLASSO, are the best performing. These models differ in various ways. The RF model is a nonlinear approach, whereas the RR model is a linear technique. Furthermore, the adaLASSO model induces real sparsity, whereas the RR and RF models simply approximate sparsity.

Other interesting models may be macroeconomic random forests, where trees are combined with linear models in terminal nodes that performs well at macroeconomic forecasting (Goulet Coulombe (2020)). Another interest method would be varying coefficient BART (VC-BART) (Deshpande et al. (2020)) for determine vary coefficients over time by treating the covariates effect in a linear model as nonparametric functions.

Chipman et al. (2010) introduced the non-parametric model with the use of Bayesian additive regression trees (BART). The BART model is a prosperous model for multivariate time series data and for dealing with considerable uncertainty (Huber and Rossini (2020)). Moreover, the BART model creates accurate forecasts compared to the machine learning literature (Behrens et al. (2018), Prüser (2019) and Huber and Rossini (2020)).

Prado et al. (2021) proposed an extension to the BART model in which a linear function is calculated at each node, instead of the traditional. Covariates, which are used as splitting variables in the regression tree, are used to build the linear predictors. Prado et al. show that, when compared to the traditional BART model, using local linear predictors captures regression trees more effectively, reduces the number of necessary trees, and hence improves efficiency and performance.

Another model which accounts for smoothness is the local linear forest (LLF) model, introduced by Friedberg et al. (2020). As RFs (Breiman (2001)) is a popular method for forecasting economic variables (Baybuza (2018), Yoon (2021) and Kumar and Thenmozhi (2006)), Breiman (2001) introduced a new method, LLF, which improves the RF using local linear predictors to catch local trends.

3 Data

Our used data sets is obtained from the FRED-MD database on the Michael McCraken's webpage.[1] McCracken and Ng (2016) describe a huge macroeconomic monthly frequency data set for the purpose of empirical research with big data. For a more detailed description of the data set, see McCracken and Ng (2016).

For the part of this paper that replicates the paper of Medeiros et al. (2021), 122 variables are recorded covering a period from January 1960 to December 2015 (672 observations). Following the methodology of Medeiros et al. (2021), four principal component factors are selected from the variables to serve as potential predictors. Four lags of all variables and four autoregressive terms are included. To attain stationary variables, both datasets are transformed (see Medeiros et al. (2019)). The inflation is computed by a transformation of the price index, P_t, where t presents period t in months: $\pi_t = log(P_t) - log(P_{t-1})$. In accordance with Medeiros et al. (2021), P_t corresponds to the CPI.

In the supplementary materials of Medeiros et al's paper, the inflation values are presented over the whole sample. A large outlier in November 2008 is caused by the financial crisis. As this outlier may affect prediction, a dummy variable is included for the period of November 2009 in the explanatory variables for all models. The out-of-sample period ranges from January 1990 to December 2015 and is split into two subsample periods. Namely, January 1990 to December 2000 (132 observations) and January 2001 to December 2015 (180 observations). A more detailed description of the sample set is presented in the work of Medeiros et al. (2019).

4 Methodology

4.1 Benchmark models

In this paper, the two benchmarks are those evaluated by Medeiros et al. (2019). The first benchmark model is the random walk (RW) model, where the inflation forecast is $\hat{\pi}_{t+h|t} = \pi_t$ and h denotes the horizon.

The second benchmark model is the autoregressive (AR) model. The parameters are estimated by OLS and the order is denotes as p which is determined by the Bayesian information criterion (BIC). The inflation forecast formula is set up specific for every horizon where the expanded inflation forecast is estimated by a summation of individual forecasts. The formula is as follows: $\hat{\pi}_{t+h|t} = \hat{\phi}_{0,h} + \hat{\phi}_{1,h}\pi_t + ... + \hat{\phi}_{p,h}\pi_{t-p+1}$.

4.2 Bagging, complete subset regressions and random forests

Bagging and complete subset regressions (CSR) are two different ensemble methods and are replicated in this paper. An ensemble method combines predictions from different machine learning algorithms and thereby produces more accurate predictions than an individual model. Bagging is an application of the bootstrap procedure and takes the most frequent class. The number of decision trees to be included, can be chosen by increasing the number of included trees until the accuracy stops giving improvements. The CSR test all possible combinations of training sets and selects the most optimal set of regressors to make predictions. The final forecast is the average of all forecasts.

Furthermore, random forests are used, which are based on the bagging algorithm. The random forest (RF) model is introduced by Breiman (2001) in order to diminish the variance of regression trees. Random forests improves bagging by reducing the correlation between subtrees. When a considerable number of trees are constructed, the most accurate class is selected. This approach is called the random forest. Appendix A presents a further description about the random forest approach.

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

4.3 Shrinkage models

In this paper, the replicated shrinkage estimators for linear models are applied in four different ways, namely, the least absolute shrinkage and selection operator (LASSO), the adaptive least absolute shrinkage and selection operator (adaLASSO), the elastic net (ElNet) and the adaptive elastic net (adaElNet). Appendix B gives a more complex overview of these models.

4.4 Factor Models

For the purpose of reducing the model dimension, factor models are included in the research. Besides the general factor model, two different extension are used, the target factors (tfactor) and the factor boosting (bfactor) method. We follow the approach of Bai and Ng (2008) and Medeiros et al. (2019). A detailed elaboration can be found in the paper of Bai and Ng (2008).

4.5 Hybrid linear-Random forest models

For predicting and evaluating forecasts of inflation, the nonlinearity and variables selection is of large importance (Medeiros et al. (2019)). Therefore, adjustments on the RF and adaLASSO model are evaluated, namely, RF/OLS and adaLASSO/RF, introduced by Medeiros et al. (2019). As main purpose of the RF/OLS model, the linearity performance and thereby, the influence of nonlinearity, is evaluated. The variables are selected through the RF model. If the results of the estimation are approximately the same as the estimation results through the RF approach, the nonlinearity does not play a major role.

The second adjusted model is adaLASSO/RF. The adaLASSO model is used for variable selection and the model is estimated through the RF approach with the selected variables. If this approach produces approximately the same results as the RF approach, variable selection is less important and nonlinearity plays a more relevant role. Both adjusted models indicate the importance of variable selection and the relevant role of nonlinearity of the RF model.

4.6 Model trees

The BART model, introduced by Chipman et al. (2010), is a machine learning method based on regression trees and a non-parametric Bayesian algorithm for classify and forecasting time series. The method is a high performing prediction method for dealing with non-linearity and high-order interactions. The BART model uses piecewise constants at node level for splitting in the regression trees. On this part, an extension is introduced by Prado et al. (2021), the Model Trees BART (MOTR-BART) model, where at node level it considers piecewise linear functions, which includes the covariates that have been used as splitting variables at the nodes. Resulting in fewer required regression trees to obtain equal performance as the general BART model, capturing local linearity and possibly improve the prediction at node level. In this Section, an explanation of the BART model and the MOTR-BART model is given.

4.7 BART model

The BART model distinguishes itself from other tree-based machine learning method as the structure of each regression tree is determined through a prior distribution. For every iterations, the BART procedure accepts or rejects constructed regression trees. As BART is based on regression trees, it includes a sum of regression trees for analysing the univariate response variable $\mathbf{y} = (y_1, ..., y_n)^T$. Every element of the response variable is estimated as follows:

$$y_i = \sum_{f=1}^{F} g(\mathbf{x}_i; T_f, M_f) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

where $g(\mathbf{x}_i; T_f, M_f)$ represents the prediction function, which gives as output a value μ_{fl} based on $\mathbf{x}_i$ and f represents tree f. The vector $\mathbf{x_i} = (x_{i1}, ..., x_{id})$ is the i-th column of the X data matrix, the set T_f contains splitting rules for constructing the f-th tree and the vector $M_f = (\mu_{f1}, ..., \mu_{fk_f})$ contains the predicted values for each node in the f-th tree, where u_{tk_f} represents the predicted value for node k_f. The splitting rules in the set can be defines as partitions $\mathcal{P}_{fl}$, where $l = 1, ..., k_f$ representing the terminal node and $g(\mathbf{x}_i; T_f, M_f) = \mu_{fl}$ holds for all observations $i \in \mathcal{P}_{fl}$.

In the BART model is each regression constructed through a backfitting algorithm further explained in Chipman et al. (1998). Each tree can grow, prune, change or swap through the algorithm. When a regression tree grows, the terminal node is selected randomly. Thereafter, the terminal node is split in two new nodes and therefore the covariate is used, which is picked uniformly. For the prune part, a node is randomly chosen and its corresponding two terminal nodes are removed. In the change step is the splitting procedure of a randomly picked pair of terminal nodes redetermined. The swap process select randomly two nodes and interchange their splitting rules. A new regression tree is created through one of these four movements and compared with the previous regression tree by the Metropolis-Hastings algorithm that samples the full conditional distribution of T_f. For the BART algorithm, the number of MCMC iterations, the trees, hyper-parameters and partial residuals have to be initialised. Within the algorithm, the new tree T_f^* is accepted or rejected as current tree with probability $\alpha(T_f, T_f^*)$. In this paper, the number of trees are set to 10 and 1000 burn-in iterations and post-burn-in iterations.

4.8 MOTR-BART model

In this paper, we follow Prado et al. (2021), which introduced an extension of the BART model, the MOTR-BART model. Following BART, the MOTR-BART model considers the response variable as the sum of regression tree as follows:

$$y = \sum_{f=1}^{F} g(\mathbf{X}; T_f, B_f) + \epsilon,$$

where $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$, n is equal to 312 the inflation forecasts and M_t is replaced by B_f, which represents the set of parameters containing all linear predictors of the tree f. The partial residuals of MOTR-BART is presented as follows:

$$r_i|\mathbf{x}_i, \beta_{fl}, \quad \sigma^2 \sim N(\mathbf{x_i}\beta_{fl}, \sigma^2),$$

where $r_i = y_i - \sum_{j \neq t}^{m} g(\mathbf{X}; T_f, B_f)$ holds and β_{fl} represents a parameter vector associated to the terminal node l of tree f. Each observations $i \in \mathcal{P}_{fl}$ contains prediction values based on β_{fl} and X_{fl}, the values of their covariates. Resulting in possible different prediction values. The distributions for β_{fl} and the overall variance, σ^2, are as follows:

$$\beta_{fl}|T_f \sim N_q(0, \sigma^2 V),$$
$$\sigma^2|T_f \sim IG(\nu/2, \nu\lambda/2),$$

where IG denotes the Inverse Gamma distribution, $V = \tau_b^{-1} \times I_q$ and $q = p_{fl} + 1$ holds. p_{fl} is the number of covariates in the linear predictor of terminal node l and regression tree f. The V matrix has one more dimension due to an extra column filled with 1's in the X_{fl}.

τ_b has as purpose construct equal importance of each tree on the final prediction by keeping the components of β_{fl} nearly to zero. Prado et al. (2021) shows benefits from penalising the intercepts and slopes differently compared to setting $\tau_b = m$. In this sense, V is specified as a $q \times q$ diagonal matrix with $V_{1,1} = \tau_{\beta_0}^{-1}$ and $V_{j+1,j+1} = \tau_{\beta}^{-1}$, where β_0 is a vector containing the intercepts from all terminal nods of all trees and β is a vector filled with the slopes from all linear

predictors of all trees. The following distributions hold: $\tau_{\beta_0}^{-1} \sim G(a_0, b_0)$ and $\tau_{\beta}^{-1} \sim G(a_1, b_1)$, where both variances are estimated through Gibbs-sampling steps. We follow Prado et al. (2021) and estimate τ_{β_0} and τ_{β} for penalising the intercept and the slopes differently. Furthermore, we assume for β_{fl} a normal distribution with mean zero. However, it is not possible to determine the distribution of each component of β_{fl} and thereafter execute the variable selection, as the regression trees may be modified through the four different possible movements. To overcome this problem, only covariates in linear predictor for splitting the regression trees are used. To illustrate this, consider x_1, x_2 and x_3 as splitting covariates. For each linear predictor i, $X_{fi}\hat{\beta_{fi}}$, the three different explanatory variables are included for the purpose of improving the prediction, as a linear or non-linear is included. For a linear relation is the dependency caught by the linear predictor. Otherwise, for a non-linear relation is the coefficient of this corresponding covariate equal to approximately zero and thus is the impact on the prediction negligible. See Chipman et al. (2010) for a further and detailed elaboration of the method. For the MOTR-BART model, we follow Prado et al. (2021) by considering 10 trees, set $\alpha = 0.95$ and set $\beta = 2$. Furthermore, we use 250 and 1000 burn-in iterations and 500 and 1000 post-burn-in iterations. Larger number of burn-in and post-burn-in iterations may results in better forecasts, but due to the time limit, these particular number of iterations are used.

4.9 Local linear forest

To extensive the paper, LLF is introduced to overcome a leading weakness of RFs, the inability of handling smoothness in estimation of regression trees. LLF do not work well in high dimensions and therefore the forests are unable to detect local linear trends, which causes wrong predictions. Breiman (2001) introduced the local linear forest, which handles local smoothness by using local linear predictors and thereby address this weakness of RFs. Weights can be produces by RFs to use these weights in local linear regression. All in all, LLF combine the adaptability of random forests and the capability of local linear regression to capture smoothness. By these regression adjustments, improvement in the accuracy of forests on smooth signals can be realized.

Consider the training data as (X_1, Y_1), ..., (X_n, Y_n) and the equation as $Y_i = \mu(X_i) + \epsilon_i$. A random forest is used to calculate the following conditional mean equation: $\mu(x_0) = \mathbf{E}[Y|X = x_0]$. In a forest within F trees, each tree T_f contains the leaf $L_f(x_0)$ and the corresponding estimated response $\hat{\mu}_f(x_0)$. $\hat{\mu}_f(x_0)$ presents the average response of all training data points for leaf $L_f(x_0)$. The average estimated response can be calculated as follows: $\hat{\mu}(x_0) = (1/F)\sum_{f=1}^{F}\hat{\mu}_f(x_0)$. LLFs construct a kernel, called the local linear regression, within weight, which are generated by the use of random forest. An alternative approach, is using random forests as weight generators:

$$\hat{\mu}(x_0) = \sum_{i=1}^{n} \alpha_i(x_0) Y_i. \tag{1}$$

The RF weights $\alpha_i(x_i)$ are as follows defined:

$$\alpha_i(x_0) = \frac{1}{F}\sum_{f=1}^{F}\frac{1\{X_i \in L_f(x_0)\}}{|L_f(x_0)|},$$

where $L_f(x_i)$ corresponds to a leaf in tree T_f, where F is the total number of trees. For each i, it holds that $0 \leq \alpha_i(x_0) \leq 1$.

LLF takes the RF weights, $\alpha_i(x_0)$ and subsequently solve the following minimization equation:

$$\begin{bmatrix}\hat{\mu}(x_0)\\ \hat{\theta}(x_0)\end{bmatrix} = \min_{\mu,\theta}\left\{\sum_{i=1}^{n}(Y_i - \mu(x_0) - (X_i - x_0)\theta(x_0))^2\alpha_i(x_0) + \lambda||\theta(x_0)||_2^2\right\}.$$

$\theta(x_0)$ corresponds to the local trend in $(X_i - x_0)$ and $\lambda||\theta(x_0)||_2^2$ presents the ridge penalty, which helps avoiding overfitting to the local trend. Friedberg et al. (2020) analyzed different tree-splitting rules to get weights $\alpha_i(x_0)$ for improve the forecast accuracy.

For the simulation of the model, we use the defaults of the model, where for instance $\alpha = 0.05$ and the number of trees in each forest is set to 10. As Friedberg et al. (2020) recommend small values of λ for penalization on regression during forest training, $\lambda = 0.1$. In this paper, the standard LLF with the use of defaults is experienced. A further and detailed explaining of the model, can be found in the paper of Friedberg et al. (2020).

4.10 Prediction calculation

In order to forecast inflation, a rolling window for a fixed length is used for all models. However, for the first subsample period (1990-2000), the length of the rolling window is equal to $R_h = 360 - h - p - 1$ and for the second subsample period (2001-2005) the length of the rolling window is equal to $R_h = 492 - h - p - 1$, where h presents the horizon and p the number of lags in the model. The advantages of working with a rolling window, is to reduce the impact of outliers on the prediction values and to prevent problems of running performance tests among models (see Medeiros et al. (2021)). All the methods are programmed and estimated in R. For the replication part, the code of Medeiros et al. (2021) is used and is available online.[2] For the extension part, BART, MOTR-BART and the local linear forest are programmed the same way as for the replication part. The package *bartMachine* is used for BART. For the MOTR-BART model[3] and the local linear forest model [4], the used packages are downloaded and available online.

4.11 Predictive comparison criteria

In this paper, we partially follow Medeiros et al. (2019). The three different criteria are used to compare the accuracy of forecasts models. Firstly, the RMSE is calculated as follows:

$$RSME_{m,h} = \sqrt{\frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \hat{e}_{t,m,h}^2}.$$

Secondly, the mean absolute error (MAE) is considered:

$$MAE_{m,h} = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} |\hat{e}_{t,m,h}|.$$

The third criteria is the median absolute deviation from the median (MAD):

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

It holds that $\hat{e}_{t,m,h} = \pi_t - \hat{\pi}_{t,m,h}$, which is the difference between the real inflation and the corresponding inflation forecast for month t, by model m and for a forecast of h months.

Testing if there exist a difference in forecasts accuracy, the model confidence sets (MCSs) by Hansen, Lunde, et al. (2011a) is applied.

For evaluating the forecasts accuracy, the unbiasedness of the forecasts is evaluated. Rodríguez-Vargas (2020) stated forecasts should have a mean of zero otherwise forecasts under- or overestimate the true inflation value. Mincer and Zarnowitz (1969) introduced a regression for testing unbiasedness:

[2] `https://github.com/gabrielrvsc/ForecastInflation`.

[3] `https://github.com/ebprado/MOTR-BART/blob/master/MOTRbart/R/MOTR_BART.R`.

[4] `https://github.com/grf-labs/grf`.

$$\pi_{t+h} = \alpha + \beta\hat{\pi}_{t+h} + \epsilon_{t+s}.$$

π_{t+s} suggest the real value and $\hat{\pi}_{t+s}$ the estimate forecast of the target value where h presents the horizon in months and t the current time in months. A t-test is used to calculate if the α and β statistically differ from relatively zero and one.

5 Results

Table 1 summarizes the models' out-of-sample performance from 1990 to 2015. Forecasts per horizon are calculated for both subsamples. Thereafter, the forecasts are merged and the errors are calculated. The resulting table resembles that of Medeiros et al. (2019), but does not include UCSV, RR and JMA models due to their long running duration. Furthermore, the ratios are not identical to those in the study, which can be explained by the absence of cumulative forecasting errors. Furthermore, the table is extended by three models, BART, MOTR-BART and the LLF. Column (1), (2) and (3) represent the average RMSE, MAE and MAD error respectively. The average error terms are computed by first estimating the inflation forecast per subsample for each horizon. Secondly, the forecasts for both subsamples are merged and the RMSE, MAE and MAD are calculated per horizon. The mean is then calculated over each horizon's error term and normalized to the RW. Column (4), (5) and (6) report the maximum RMSE, MAE and MAD value respectively, followed by column (7), (8) and (9) displaying the minimum RMSE, MAE and MAD value respectively. To make a comprehensible table, all ratios are standardized to the RW model. Column (10), (11) and (12) demonstrate how many times the model achieved the lowest RMSE, MAE and MAD. As last, column (12) and (13) demonstrate the average p-value for the MZ regression. To test if RF, BART, MOTR-BART and LLF statistically produce different inflation forecasts and outperforms the RW, Table 2 and 3 presents the average p-values obtained for square losses and absolute losses for the MCS based on t_{max} statistic as described in Hansen (2005b). For every horizon, the p-value is estimated and subsequently the average p-value is calculated. As last, to analyse the time series of the forecasts for RF, BART, MOTR-BART and LLF, inflation forecasts are plotted against the real value.

5.1 Replicated results

The following may be seen in Table 1 when looking at the replicated results. First and foremost, the RW is outperformed by all other models. In comparison to the RW, ML models and the use of a broad collection of predictors result in a significant improvement in inflation forecast accuracy.

In comparison to other models, not only the RW, but also the AR, bagging, factor, tfactor, and bfactor yield the greatest errors in term of RMSE and MAE. Furthermore, comparing the results of the RF, RF/OLS and adaLASSO/RF, we observe more similar results for the RF and the adaLASSO/RF than for the RF and the RF/OLS, indicating that nonlinearity plays a more important role than variable selection. As most observing result, the RF is the most accurate model based on the average RMSE, MAE and MAD. The ratio's for RF are the smallest for each error term. Additionally, RFs statistically outperforms the RW for both square and absolute losses based on the test statistic results of the MCS in Table 2 and 3. Moreover, the advancement of RF based on the errors terms, is a gain of approximately 30%. The well-performance of the RF is due to the variable selection procedure in the method as well as since the RF is an accurate model for dealing with nonlinearities, as the relation of inflation and the predictor contains nonlinearities.

5.2 Random forest compared to BART, BART with model trees and local linear forest

In the previous Subsection, it is concluded that the RF produces the most accurate inflation forecast based on the error terms, RMSE, MAE and MAD, compared to the replicated models. In this paper, the overview of machine learning method for forecasting inflation, is extended with BART, MOTR-BART, for 250 and 1000 burn-in iterations and 500 and 1000 post-burn-in iterations, and LLF. Summary statistics can be observed in the last rows in Table 1. For RF, BART, MOTR-BART and LLF, a MZ regression is applied to test with a t-test if the forecasts are statistically unbiased, that is testing if $\alpha = 0$ and $\beta = 1$ statistically holds. The average p-value for the intercept and the coefficient are presented in Table 1. Appendix C provides detailed results for each horizon for the test on MZ regression.

To test whether models significantly differ in forecasting inflation, the MCS by Hansen (2005b) is considered and presented in Table 2 and 3. The null hypothesis states equal forecasting performance and thus rejecting the null hypothesis means that a model is outperformed by the other model. The p-values for the MCS based on the t_{max} statistic as described in Hansen (2005b), are estimated for square losses in Table 2 and for absolute losses in Table 3.

Furthermore, the inflation forecasts are plotted against the real inflation value. In this Section, the graphs for a horizon of 1 month and of 12 months are given, evaluated and compared. In Appendix D, F and G show graphs for each other horizon for respectively the models RF, BART, MOTR-BART and LLF.

5.2.1 Evaluating Table 1

In Table 1 are relative large forecasting precision values observed for the MOTR-BART with 250 burn-in and 500 post-burn-in iteration. Even a few ratio's, for instance for the average MAE and MAD with the values of respectively 1.11 and 1.16, are larger than 1.00, indicating the RW outperforms the MOTR-BART with 250 burn-in and 500 post-burn-in iterations. Therefore, the MZ results, the MCCs results and the graphs are not produced for the MOTR-BART with 250 burn-in and 500 post-burn-in iterations.

The summary statistics in Table 1 show that the RF and LLF models' error terms differ. Namely, the RMSE of LLF is smaller than that of the RF, while the RF outperforms the LLF in terms of MAE and MAD. Also the maximum and minimum error terms differ across the models and type of error term. However, all maximum and minimum error terms for the RF are smaller than for the LLF, showing that the RF has more accurate forecasts. The findings of the MZ test are confirmed by these outcomes. The average p-values of the LLF model in the MZ regression lead to a rejection of the null hypothesis for for both the intercept and the coefficient with a significance level of 0.05 and 0.01 respectively. For the RF, the intercept is significantly not different from zero, but the coefficient differs from one, as the null hypothesis is rejected for a significance level of 0.1 and 0.05. These findings imply that the RF forecasts inflation more accurately than the LLF. However, the test results of the MCS, which can be observed in Table 2 and 3, indicate no rejecting of the null hypothesis and thus no different forecasts values across the models.

The BART model shows no improvement in forecasts accuracy compared to RF. The BART model seems to produce relative poor results compared to RF, as the errors terms are larger for RMSE, MAE and MAD. Moreover, these observations can be confirmed by the test results of MCS, which conclude that BART is outperformed by RF for absolute losses. The summary test statistics of BART and MOTR-BART are almost equal, which is also confirmed by the test results of MCSs. For both square and absolute losses is the null hypothesis of equal forecast values not rejected and can be concluded that adding piecewise linear functions in terminal nodes does not improve the forecast accuracy.

Evaluating the summary test statistics results in Table 1 for MOTR-BART with 1000 burn-

in and post-burn-in iterations, this particular model produces relative larger RMSE, MAE and MAD errors terms compared to RF and LLF. However, analyzing the minimum RMSE, MAE and MAD errors terms for the MOTR-BART, these are smaller than the errors terms of LLF and for the RFs RMSE error term. For square losses, the test statistics results of MCS show statistically no different predictions values for the MOTR-BART compared to RF and LLF. However, for absolute losses, the MCSs test results reject the null hypothesis of equal forecasts values, indicating that the RF and LLF outperforms the MOTR-BART. These results may be declared by the settings of the burn-in and post-burn-in iterations. Larger iterations may yield more accurate inflation forecasts. Overall, it can be concluded that RFs, BART, MOTR-BART and LLFs outperform the benchmark model, RW, and thereby produce more accurate inflation forecasts values.

	Forecasting precision												Mincer Zarnowitz	
Model	(1) ave. RMSE	(2) ave. MAE	(3) ave. MAD	(4) max. RMSE	(5) max. MAE	(6) max. MAD	(7) min. RMSE	(8) min MAE	(9) min. MAD	(10) #min. RMSE	(11) #min. MAE	(12) #min. MAD	(13) ave. p-value Intercept (α)	(14) ave. p-value Coefficient (β)
RW	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	0	0	-	-
AR	0.80	0.84	0.74	0.76	1.11	0.66	0.90	0.87	0.74	0	0	0	-	-
LASSO	0.77	0.77	0.72	0.73	0.74	0.67	0.84	0.83	0.78	0	0	0	-	-
adaLASSO	0.77	0.76	0.74	0.73	0.73	0.70	0.83	0.81	0.76	0	0	0	-	-
ElNet	0.76	0.77	0.72	0.73	0.75	0.67	0.83	0.82	0.77	0	0	4	-	-
adaElnet	0.77	0.76	0.73	0.73	0.75	0.74	0.84	0.82	0.75	0	0	1	-	-
Bagging	0.81	0.84	0.88	0.80	0.80	0.81	0.84	0.85	0.85	0	0	0	-	-
CSR	0.79	0.79	0.77	0.76	0.77	0.71	0.85	0.84	0.8	0	0	0	-	-
Factor	0.80	0.83	0.82	0.78	0.82	0.76	0.87	0.88	0.87	0	0	0	-	-
T.Factor	0.80	0.83	0.84	0.78	0.80	0.77	0.88	0.87	0.88	0	0	0	-	-
B.Factor	0.80	0.84	0.90	0.75	0.79	0.79	0.95	0.96	0.98	0	0	0	-	-
RF	0.74	0.73	0.70	0.69	0.67	0.63	0.85	0.82	0.70	2	9	5	0.12	0.02**
RF/OLS	0.76	0.76	0.78	0.72	0.73	0.74	0.82	0.80	0.83	1	1	0	-	-
AdaLASSO/RF	0.75	0.74	0.72	0.73	0.71	0.65	0.85	0.82	0.71	0	0	1	-	-
BART	0.78	0.81	0.82	0.75	0.76	0.74	0.90	0.90	0.88	0	0	0	0.00***	0.00***
MOTR-BART (250,500)	0.99	1.11	1.16	1.33	1.56	1.82	0.92	0.70	0.91	0	1	0	-	-
MOTR-BART (1000,1000)	0.79	0.80	0.80	0.79	0.77	0.70	0.82	0.82	0.77	0	0	0	0.00***	0.00***
LLF	0.73	0.75	0.78	0.71	0.69	0.68	0.89	0.88	0.85	9	1	1	0.03**	0.00***

Table 1: Forecasting results for the out-of-sample period from 1990 to 2015. NOTE: The table provides over 12 different time horizons summary statistics. Columns (1), (2) and (3) present respectively the average RMSE, MAE and MAD. Columns (4), (5) and (6) show the maximum error term value of respectively RMSE, MAE and MAD. Columns (7), (8) and (9) demonstrate the maximum error term value of respectively RMSE, MAE and MAD. Column (10), (11) and (12) give the number of achieved minimum errors terms for each horizon. The last columns, (13) and (14), display the t-test statistic results over the MZ regression, for testing $\alpha = 1$ and $\beta = 1$. The asterisk represents rejecting the null hypothesis for a significance level of 0.1 (*), 0.05 (**), 0.01 (***).

MCS SE					
	RW	RF	LLF	BART	MOTR-BART
RW	-	-	-	-	-
RF	0.01	-	-	-	-
LLF	0.01	0.51	-	-	-
BART	0.02	0.09	0.09	-	-
MOTR-BART	0.02	0.11	0.16	0.26	-

Table 2: The table presents for square losses the average p-values for the MCS based on the t_{max} statistic.

MCS AE					
	RW	RF	LLF	BART	MOTR-BART
RW	-	-	-	-	-
RF	0.00	-	-	-	-
LLF	0.00	0.37	-	-	-
BART	0.00	0.00	0.08	-	-
MOTR-BART	0.00	0.00	0.05	0.29	-

Table 3: The table presents for absolute losses the average p-values for the MCS based on the t_{max} statistic.

5.2.2 Evaluating inflation forecasts plotted against the real inflation values

Another interesting aspect when evaluating the inflation forecast of each model, is analyzing how the forecasts differ from the real inflation values and from forecasts of other models and how the models deal with nonlinear time series. For the RF, BART, MOTR-BART with 1000 burn-in and post-burn-in iterations and for the LLF are graphs constructed and analyzed in this section. Appendices D, E, F and G contain graphs for the time horizons of 2 till 11 months for respectively RF, BART, MOTR-BART with 1000 burn-in and post-burn-in iterations and LLF.

Figures 1 and 2 for RF, 3 and 4 for BART, 5 and 6 for MOTR-BART and 7 and 8 for LLF show substantial difference in the inflation forecasts when considering different time horizons. For each model, the plot of predicting one month ahead shows less wrong estimations than predicting 12 months ahead. For the RF, forecasting 12 months ahead results in wrong inflation predictions during period around crises in 2001 and 2008. Also the forecasting around 2010 show large wrong predictions when predicting 12 months ahead.

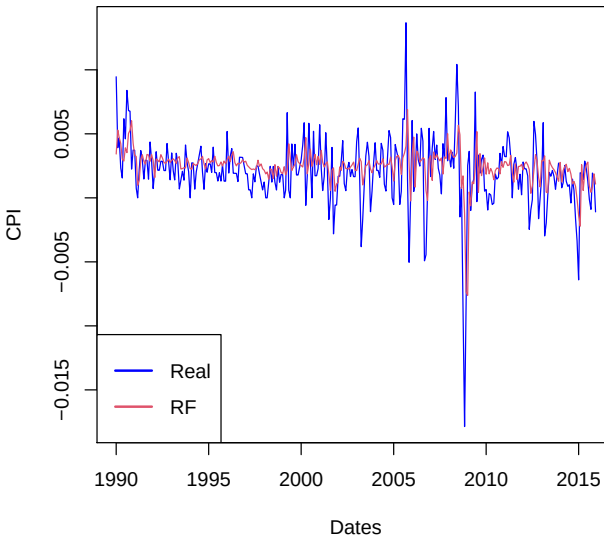

Figure 1: The real values and forecasts values for 1 horizon the RF plotted.

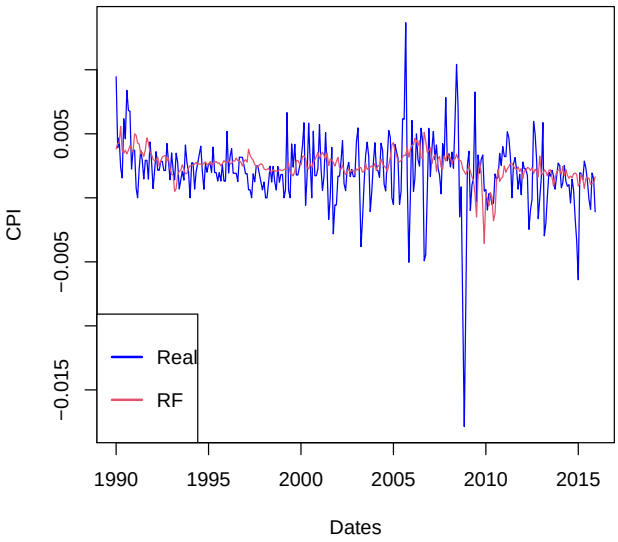

Figure 2: The real values and forecasts values for 12 horizon of the RF plotted.

Figures 3 and 4 show small difference for forecasting 1 and 12 months ahead. Substantial more large wrong forecasting values are observed when forecasting 12 months compared to forecasting 1 month ahead. Especially for the period of 2008 when prediction 12 months ahead, lower inflation forecasts are estimated too late.

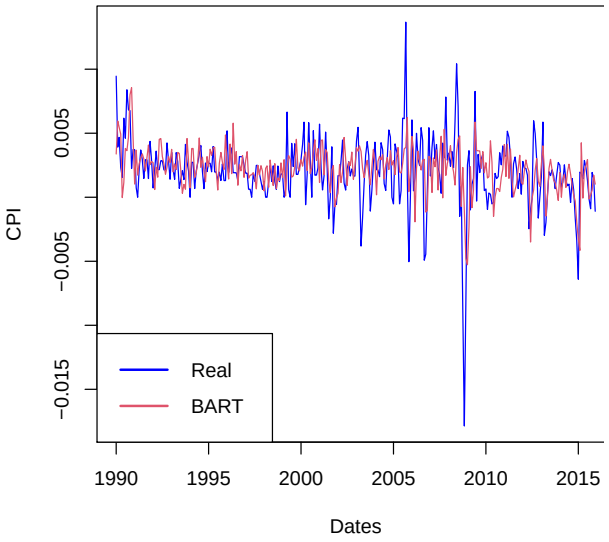

Figure 3: The real values and forecasts values for 1 horizon the BART plotted.

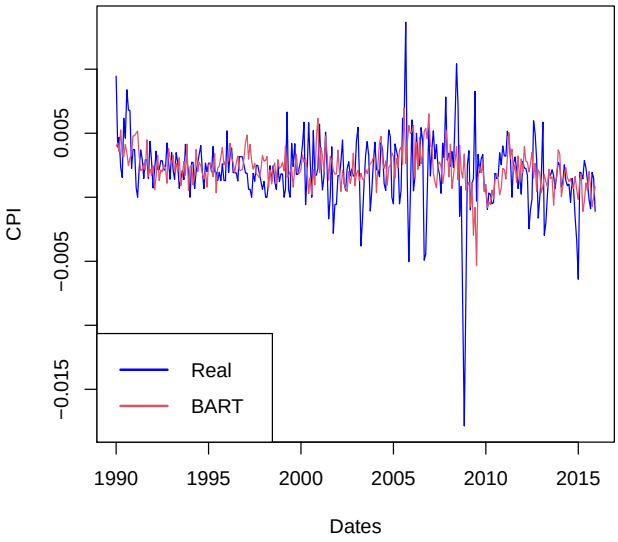

Figure 4: The real values and forecasts values for 12 horizon of the BART plotted.

Comparing Figures 5 and 6, forecasting 12 months ahead produces less accurate inflation forecasts compared to one month ahead forecasts. For instance, around 2007, too low inflation

forecasts values are predicted and around 2009 to high inflation forecasts values are predicted compared to the real inflation value. The period around 2008 shows nonlinearity caused by the financial crisis of 2007/2008.

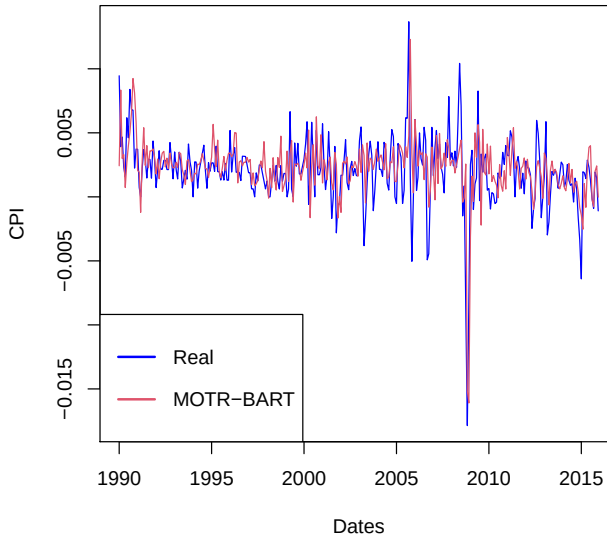

Figure 5: The real values and forecasts values for 1 horizon the MOTR-BART plotted.

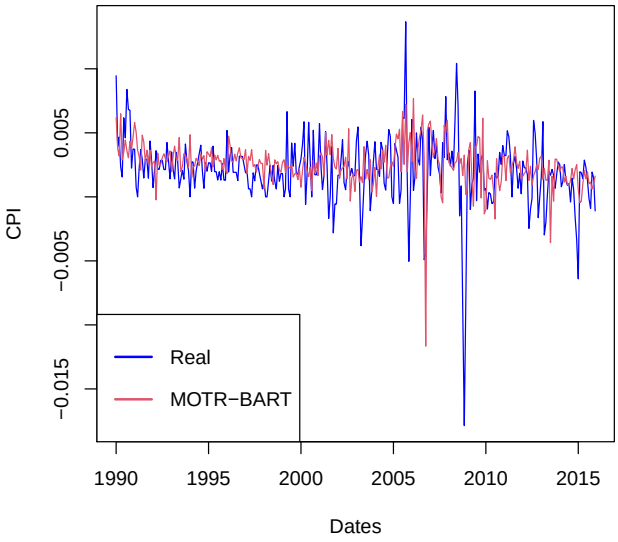

Figure 6: The real values and forecasts values for 12 horizon of the MOTR-BART plotted.

Evaluating Figures 7 and 8, we observe relative higher spikes for a horizon of one month. Inflation forecasts predicted 12 months ahead, show less non-linearity and thus more smooth prediction as small spikes are observed.

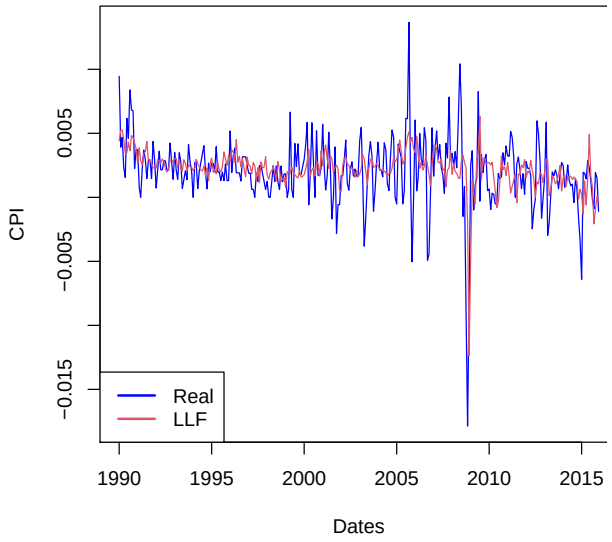

Figure 7: The real values and forecasts values for 1 horizon of the LLF plotted.

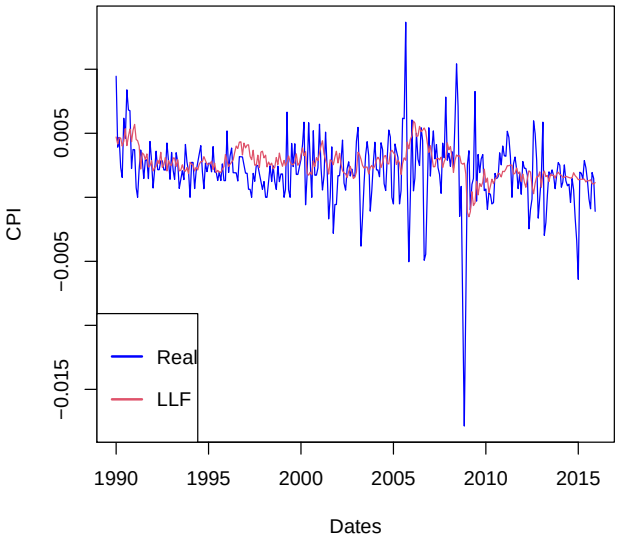

Figure 8: The real values and forecasts values for 12 horizon of the LLF plotted.

Comparing these figures across the different models, several distinctions are noted. First of all, small differences in RF and LLF are observed. Even though the time series show the same pattern of spikes, small differences are presented in the size of the spikes. These distinctions can be explained by two reasons. Firstly, in the LLF are no weights in local linear regressions included. The use of weights in LLF may results in more accurate forecasts and improve thereby RFs. The second reason can be explained by the improvement of the LLF relative to RF. Namely, LLF can handle local smoothness, which is clearly observed when comparing Figures 1 to 7 and 2 to 8. More local smoothness is observed for the LLF, as lower and wider spikes are obtained.

As BART can deal with non-linearity and high-order interactions in covariates, Figures 3 and 4 display more spikes than for the time series of RFs. Observing Figures 5 and 6, including linear predictor in the terminal nodes, instead of constants, results in a more nonlinear time series compared to the inflation forecasts time series of BART.

The forecast results of MOTR-BART in Figure 5 and 6 show relative different results compared to the RF and LLF. Larger and more spikes can be noted, which is for the reason the MOTR-BART deals with non-linearity and high-order interactions and improves thereby RFs. In these figures can be observed that there is high potential in the MOTR-BART, as it shows the same pattern as the real inflation values and as it can handle nonlinear time series.

6 Conclusion

This paper built further on the work of Medeiros et al. (2021) by evaluating and improving the accuracy of inflation forecasts. The main aim of this paper is to provide a comparative analysis of various models. New models are added in the extensive overview of Medeiros et al. (2021), namely BART and MOTR-BART, where the last model is an extension on BART by including at node level linear predictors instead of constants. Additionally, a third model is introduced to handle smoothness, which is a local linear forest. For each model, 312 inflation forecasts are predicted over 12 different monthly time horizons. As summary statistics, the RMSE, MAE and MAD are calculated, followed by test results of a MZ regression to evaluate the unbiasedness of the forecasts. For obtaining an appropriately extended comparison, the MCSs test results may conclude statistical outperformed models and graphs may display visual distinctions.

As the paper extensive earlier study, a few interesting results are observed. First of all, based on summary statistics and MCSs, LLF seems not to improve the RF when use local linear predictors in the terminal nodes. However, when observing forecasts plotted against the real inflation values, a similar time series is presented. Furthermore, adding weights to LLF may results in more accurate forecasts and thereby, LLF seems to be a potential model to produce accurate forecasts, as it may better handle smoothness.

Furthermore, the BART model and the extension of BART, called MOTR-BART, show no statistical improvement in the accuracy of inflation forecasts. However, BART and MOTR-BART show a well performance when nonlinearities are presented in inflation time series. Additionally, since the data set includes a large set of predictors and these models can handle high-order interactions, the models are interesting for further research. It is advised to further examine the use of a larger number of iterations to obtain more accurate inflation forecasts.

Overall, based on these findings it can be concluded that LLFs, BART and MOTR-BART produce statically similar accurate inflation forecast as RFs. Moreover, this paper show that these models are interesting for further research and have potential to improve the accuracy of inflation forecasts or other macroeconomic variables.

To conclude, the empirical and analytical research used in this paper, of a comparative overview, provides new and useful substantiated findings which add up to the literature on forecasting macroeconomic variables.

7 Discussion

There are many important issues for which further research is required. First of all, the results in this paper produces not exactly the same results as in the paper of Medeiros et al. (2019). This can be declared by not including RMSE, MAE and MAD errors terms of the accumulated inflation forecasts over 3, 6 and 12 months. Namely, Medeiros et al. (2019) calculated for 15 different time horizons, among others the accumulated inflation forecasts, and thereafter calculated the mean RMSE, MAE and MAD error. In this paper, only 12 different time horizons are taken into account and evaluated.

Furthermore, a few models are missing in the comparable overview. Research may further elaborate the comparison between the models.

Another important aspect is the choice of input value for the MOTR-BART model. Due to the running time limit, the value of nburn and npost are respectively set to 1000 and 1000. Larger number of iterations, may results in more accurate forecasts. Furthermore, the use of ancestors of terminal nodes in the linear predictors can be investigated in further research.

Improvements to the paper can be made in regard to the LLF model specification. Local linear forest solve the locally weighted least squares problem by the use of weights. For instance, the ridge regression can be used for determining the weights. Furthermore, the model specification can further explored by using splits based on ridge residuals as opposed to standard classification and regression trees splitting rule explained in Friedberg et al. (2020), which shows the benefit of implementing this particular splitting rule. Another aspect, is the choice of the ridge penalty. Friedberg et al. (2020) advises to use different values of λ for the forest training and local linear prediction. Additionally, small values of λ are preferable for penalize the regressions during forest training. However, in order to obtain better tuning noisy data and controlling variance, larger values of the ridge penalty are preferred.

Further research could include more models. Interesting models would be VC-BART (Deshpande et al. (2020)) which allow vary coefficients over time, soft BART for smoother functions of variables, dirichlet BART for sparse data generating and gaussian processes bayesian additive regression trees (gpBART) (Gramacy and Lee (2008)) which can include smoothness in predictions and handles well non-stationary data that may be a desirable property for inflation forecasting methods.

Further research in the field of forecasting macroeconomic variables by ML methods is expected to be promising to sharpen the obtained results.

References

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

Baybuza, I. (2018). Inflation forecasting using machine learning methods. *Russian Journal of Money and Finance*, *77*(4), 42–59.

Behrens, C., Pierdzioch, C., & Risse, M. (2018). A bayesian analysis of the efficiency of growth and inflation forecasts for germany. *Available at SSRN 3002394*.

Breiman, L. (2001). Random forests. *Mach. Learn.*, *45*(1), 5–32.

Chakraborty, C., & Joseph, A. (2017). Machine learning at central banks.

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, *93*(443), 935–948.

Chipman, H. A., George, E. I., McCulloch, R. E. et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Deshpande, S. K., Bai, R., Balocchi, C., Starling, J. E., & Weiss, J. (2020). Vcbart: Bayesian trees for varying coefficients. *arXiv preprint arXiv:2003.06416*.

Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, 1–15.

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

Gramacy, R. B., & Lee, H. K. H. (2008). Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, *103*(483), 1119–1130.

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

Hansen, P. R., Lunde, A., & Nason, J. M. (2011b). The Model Confidence Set. *Econometrica*, *79*(2), 453–497. `https://ideas.repec.org/a/ecm/emetrp/v79y2011i2p453-497.html`

Hansen, P. R. (2005a). A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics*, *23*, 365–380. `https://ideas.repec.org/a/bes/jnlbes/v23y2005p365-380.html`

Hansen, P. R. (2005b). A test for superior predictive ability. *Journal of Business & Economic Statistics*, *23*(4), 365–380.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. Springer.

Huber, F., & Rossini, L. (2020). Inference in bayesian additive vector autoregressive tree models. *arXiv preprint arXiv:2006.16333*.

Kumar, M., & Thenmozhi, M. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest. *Indian institute of capital markets 9th capital markets conference paper*.

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2020). *Machine Learning Advances for Time Series Forecasting* (Papers No. 2012.12802). arXiv.org. `https://ideas.repec.org/p/arx/papers/2012.12802.html`

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., Vasconcelos, G., Veiga, A., & Zilberman, E. (2019). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business Economic Statistics*, *39*, 1–45. `https://doi.org/10.1080/07350015.2019.1637745`

Medeiros, M. C., Vasconcelos, G. F., 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.

Mincer, J. A., & Zarnowitz, V. (1969). The evaluation of economic forecasts. *Economic forecasts and expectations: Analysis of forecasting behavior and performance* (pp. 3–46). NBER.

Pavlov, E. (2020). Forecasting inflation in russia using neural networks. *Russian Journal of Money and Finance*, *79*(1), 57–73. https://EconPapers.repec.org/RePEc:bkr:journl:v:79:y:2020:i:1:p:57-73

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

Prüser, J. (2019). Forecasting with many predictors using bayesian additive regression trees. *Journal of Forecasting*, *38*(7), 621–631.

Quaedvlieg, R. (2017). Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, *39*(1), 40–53.

Rodríguez-Vargas, A. (2020). Forecasting costa rican inflation with machine learning methods. *Latin American Journal of Central Banking*, *1*(1), 100012. https://doi.org/https://doi.org/10.1016/j.latcb.2020.100012

Yoon, J. (2021). Forecasting of real gdp growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, *57*(1), 247–265.

Zahara, S., Sugianto, & Ilmiddaviq, M. B. (2020). Consumer price index prediction using long short term memory (LSTM) based cloud computing. *Journal of Physics: Conference Series*, *1456*, 012022. https://doi.org/10.1088/1742-6596/1456/1/012022

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

Appendices

A Random Forest

Hastie et al. (2001) and Medeiros et al. (2019) use an example to explain the working of regression trees showed in Figure 9. Consider two explanatory variables X_1 and X_2 and one dependent variable Y for a given interval in a regression problem. Illustrated in the left graph, firstly, the space is split into three regions at $X_1 = s1$ and $X_1 = s3$. Secondly, the left and right space are split at $X_2 = s2$ and $X_2 = s4$ respectively. The partitioning results in five regions R_k with $k = 1, .., 5$ where in each region it is assumed that the model predicts Y with constant c_k where the constant is the estimated average of realizations of Y which applies to region R_k. As illustrated in the right figure of Figure 9, the partitioning can be reformulated into a tree where each region R_k represents a terminal node of the tree.

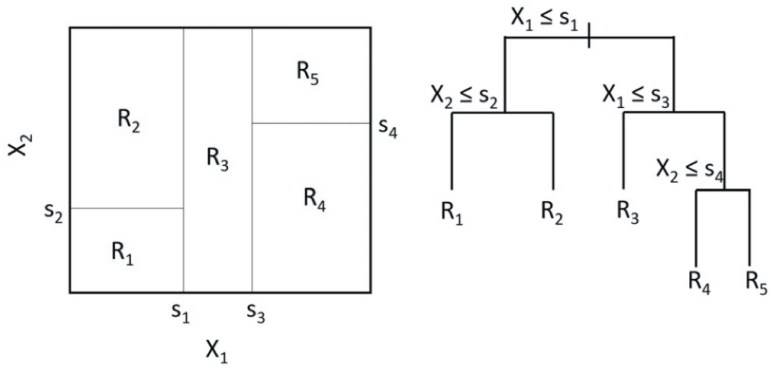

Figure 9: Example of a regression tree (Medeiros et al. (2019)).

The splits are based on the minimization of the sum of squared errors of the regression:

$$\pi_{t+h} = \sum_{k=1}^{K} c_k I_k(\mathbf{x}_t; \theta_k),$$

where the indicator function is as follows:

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

π_{t+h} presents the dependent variable, $\mathbf{x}_t$ is a set of predictors and K denotes a number of terminal nodes. θ_k is a set of parameters which specify region R_k. As described above, RF contains regression trees where each tree is specified in a bootstrap sample of the original data. We follow the paper of Medeiros et al. (2019) which use a block bootstrap for the reason that we have to deal with time series. Consider B which denotes the number of bootstrap samples. For each sample b, $b = 1, ..., B$, a randomly selected subset of the original variables is used for predicting a tree with K_b regions. K_b is defined wherefore every region has a minimum number of observation. Finally, the predicted forecast of the dependent variable $\hat{\pi}_{t+h}$ is the average of the estimated forecast for each tree and is as follows:

$$\hat{\pi}_{t+h} = \frac{1}{B} \sum_{b=1}^{B} \left[\sum_{k=1}^{K_b} \hat{c}_{k,b} I_{k,b}(\mathbf{x}_t; \theta_k) \right].$$

B Shrinkage models

The equations for shrinkage models are as follows:

$$G_h(\mathbf{x_t}) = \beta'_{\mathbf{h}}\mathbf{x_t}, \quad \hat{\beta}_{\mathbf{h}} = arg\min_{\beta_{\mathbf{h}}} \left[\sum_{t=1}^{T-h} (y_{t+h} - \beta'_{\mathbf{h}}\mathbf{x_t})^2 + \sum_{i=1}^{n} p(\beta_{h,i}; \lambda, \omega_i) \right],$$

where $p(\beta_{h,i}; \lambda, \omega_i)$ is a penalty function with λ as penalty parameter and $\omega_i > 0$ as weight.

The first approach is the least absolute shrinkage and selection operator (adaLASSO). The penalty function is formulated as:

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

The third approach replicate in this paper is the adaptive least absolute shrinkage and selection operator (adaLASSO) introduced by Zou (2006) in order to realize a consistent model selection. The penalty function is presented as follows:

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

It holds that $\omega_i = |\beta^*_{h,i}|^{-1}$ and that $\beta^*_{h,i}$ corresponds to the coefficient obtained through the first-step estimation. The adaLASSO model has in advantage the ability to work with more variables than observations and under heteroscedasticity and a non-Gaussian setting (Medeiros et al. (2019)).

The last shrinkage method is the elastic net (ElNet), which is a generalization that includes LASSO as a convex combination:

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

The last method, adaElnet, is an adaptive version of the Elnet method and follows the same approach as the adaLASSO.

C Detailed results Mincer Zarnowitz

Random forest		
Horizon (h)	Intercept (α)	Coefficient (β)
h=1	-0.388 (0.698)	-1.2267 (0.221)
h=2	2.153 (0.032)**	-3.416 (0.001)***
h=3	1.462 (0.145)	-2.552 (0.011)**
h=4	1.864 (0.063)*	-2.962 (0.003)***
h=5	1.683 (0.093)*	-2.815 (0.005)***
h=6	1.599 (0.111)	-2.677 (0.008)***
h=7	1.573 (0.117)	-2.718 (0.007)***
h=8	1.714 (0.088)*	-2.923 (0.004)***
h=9	2.118 (0.035)**	-3.391 (0.001)***
h=10	2.066 (0.040)**	-3.575 (0.000)***
h=11	2.321 (0.021)**	-3.945 (0.000)***
h=12	2.143 (0.033)**	-3.633 (0.000)***
Mean	1.692 (0.123)	-2.986 (0.0212)**

Table 4: Statistic and p-value results of the t-test for testing the intercept, $\alpha = 0$, and the coefficient, $\beta = 1$, in the MZ regression for RF.

BART model		
Horizon (h)	Intercept (α)	Coefficient (β)
h=1	3.213 (0.001)***	-5.136 (0.000)***
h=2	4.560 (0.000)***	4.560 (0.000)***
h=3	4.618 (0.000)***	-6.523 (0.000)***
h=4	4.922 (0.000)***	-6.577 (0.000)***
h=5	6.301 (0.000)***	-8.634 (0.000)***
h=6	4.795 (0.000)***	-6.889 (0.000)***
h=7	4.551 (0.000)***	-6.687 (0.000)***
h=8	4.596 (0.000)***	-6.429 (0.000)***
h=9	6.207 (0.000)***	-8.404 (0.000)***
h=10	6.186 (0.000)***	-9.062 (0.000)***
h=11	5.754 (0.000)***	-8.225 (0.000)***
h=12	5.471 (0.000)***	-7.544 (0.000)***
Mean	5.097 (0.000)***	-7.248 (0.000)***

Table 5: Statistic and p-value results of the t-test for testing the intercept, $\alpha = 0$, and the coefficient, $\beta = 1$, in the MZ regression for BART model.

Bart with model trees		
Horizon (h)	Intercept (α)	Coefficient (β)
h=1	2.940 (0.00352)***	-6.073 (0.000)***
h=2	6.155 (0.000)***	-8.859 (0.000)***
h=3	4.975 (0.000)***	-7.475 (0.000)***
h=4	6.494 (0.000)***	-10.332 (0.000)***
h=5	6.700 (0.000)***	-10.272 (0.000)***
h=6	5.639 (0.000)***	-9.146 (0.000)***
h=7	5.803 (0.000)***	-8.490 (0.000)***
h=8	5.681 (0.000)***	-8.682 (0.000)***
h=9	4.372 (0.000)***	-6.863 (0.000)***
h=10	4.625 (0.000)***	-6.909 (0.000)***
h=11	3.672 (0.000)***	-6.717 (0.000)***
h=12	3.594 (0.000)***	-6.683 (0.000)***
Mean	5.054 (0.000)***	-8.042 (0.000)***

Table 6: Statistic and p-value results of the t-test for testing the intercept, $\alpha = 0$, and the coefficient, $\beta = 1$, in the MZ regression for MOTR-BART with 1000 burn-in and post-burn-in iterations.

Local linear forest		
Horizon (h)	Intercept (α)	Coefficient (β)
h=1	2.240 (0.026)**	-3.595 (0.000)***
h=2	2.430 (0.016)**	-3.761 (0.000)***
h=3	2.567 (0.011)**	-3.911 (0.000)***
h=4	2.068 (0.040)**	-3.308 (0.001)***
h=5	2.640 (0.009)***	-3.931 (0.000)***
h=6	2.796 (0.005)***	-4.353 (0.000)***
h=7	1.850 (0.065)*	-3.377 (0.001)***
h=8	1.582 (0.115)	-3.129 (0.002)***
h=9	2.049 (0.041)**	-3.581 (0.000)***
h=10	2.322 (0.021)**	-4.067 (0.000)***
h=11	2.957 (0.003)***	-4.859 (0.000)***
h=12	3.338 (0.001)***	-5.305 (0.000)***
Mean	2.40 (0.029**)	-3.931(0.000***)

Table 7: Statistic and p-value results of the t-test for testing the intercept, $\alpha = 0$, and the coefficient, $\beta = 1$, in the MZ regression for LLF.

D Graphs for real and forecasts for every horizon for RF

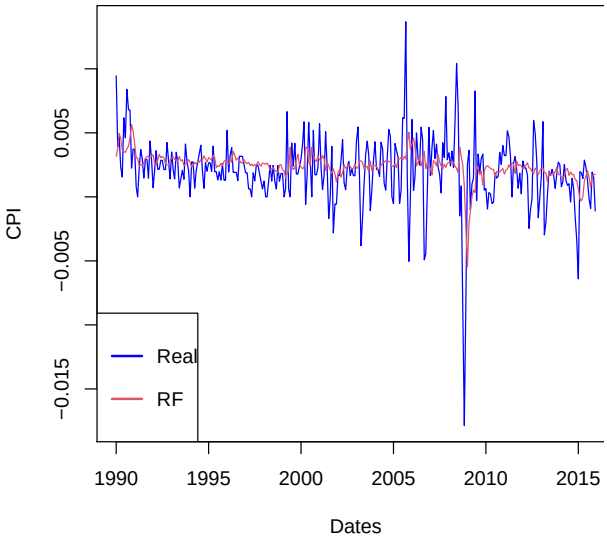

Figure 10: The real values and forecasts values for 2 horizon of the RF plotted.

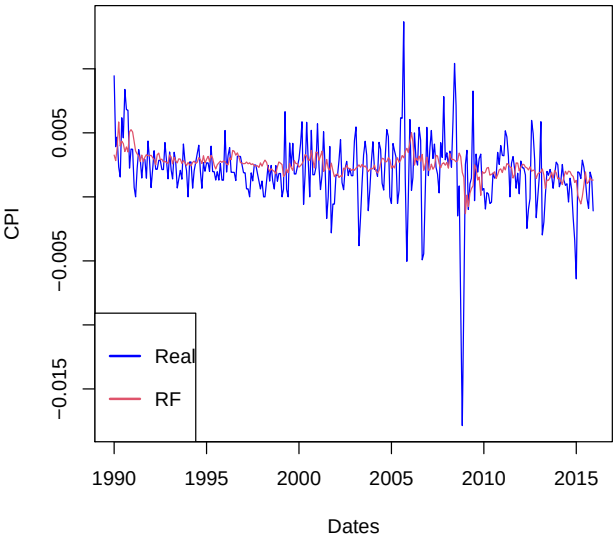

Figure 11: The real values and forecasts values for 3 horizon of the RF plotted.

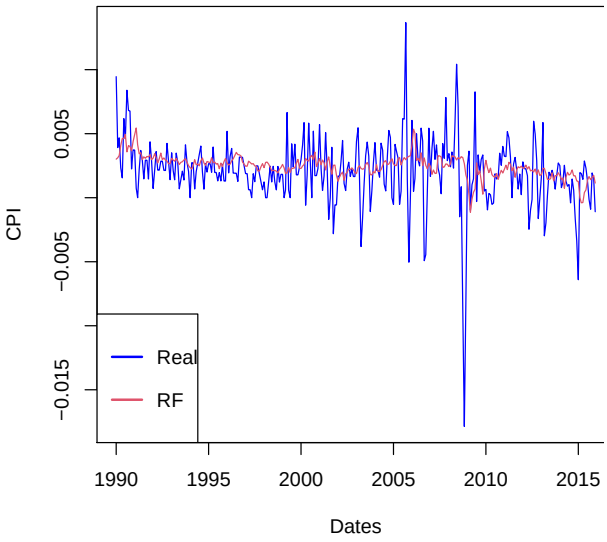

Figure 12: The real values and forecasts values for 4 horizon of the RF plotted.

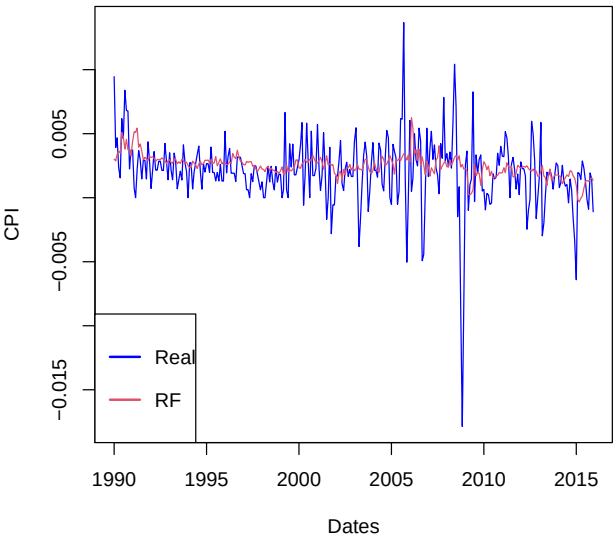

Figure 13: The real values and forecasts values for 5 horizon of the RF plotted.

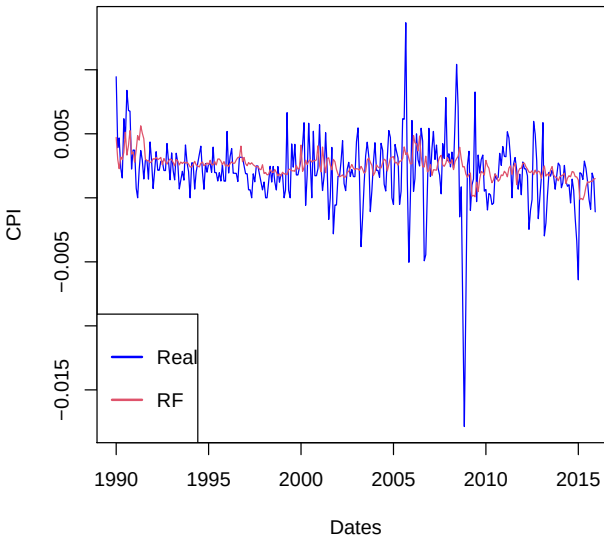

Figure 14: The real values and forecasts values for 6 horizon of the RF plotted.

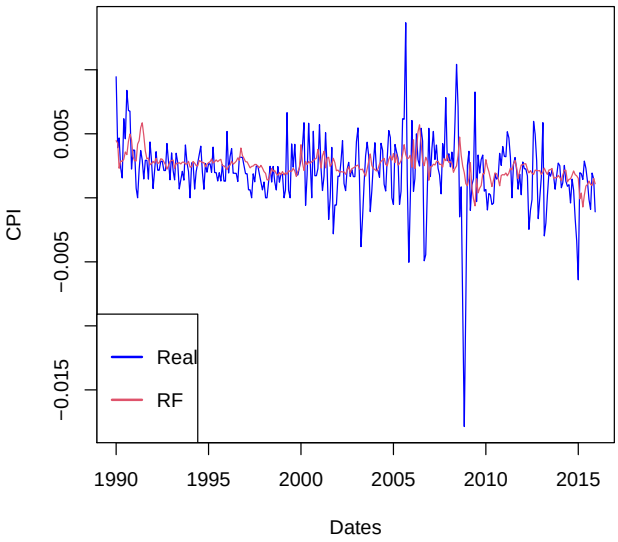

Figure 15: The real values and forecasts values for 7 horizon of the RF plotted.

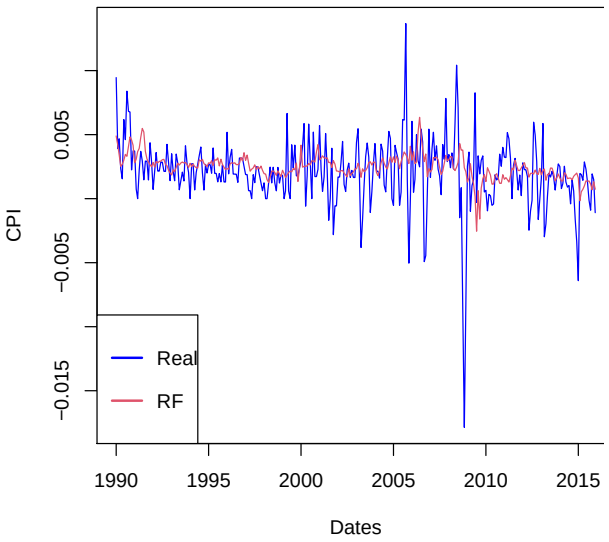

Figure 16: The real values and forecasts values for 8 horizon of the RF plotted.

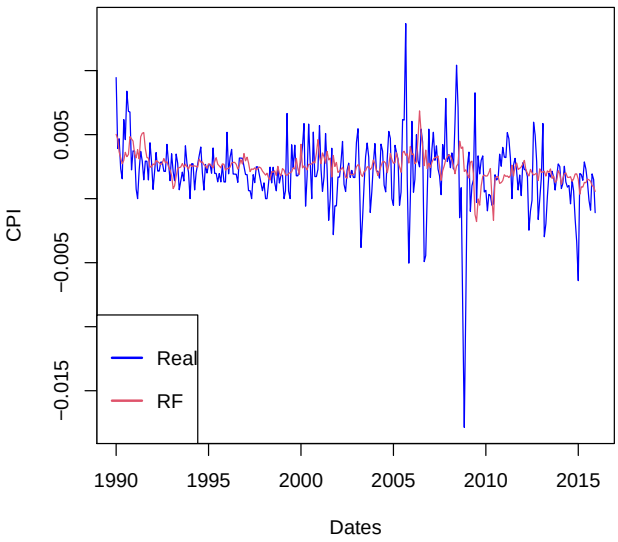

Figure 17: The real values and forecasts values for 9 horizon of the RF plotted.

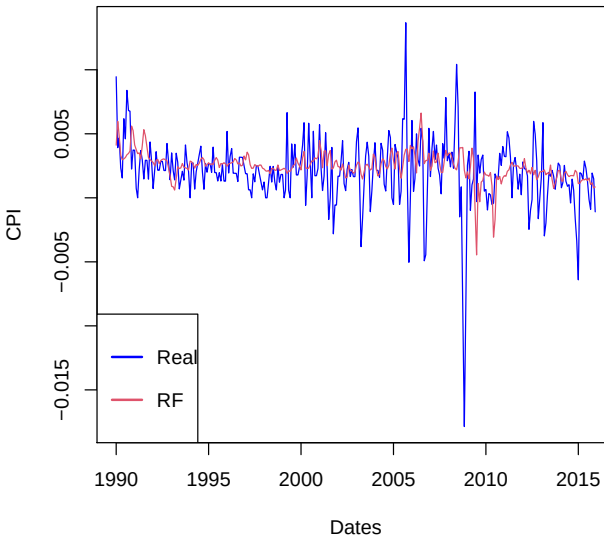

Figure 18: The real values and forecasts values for 10 horizon of the RF plotted.

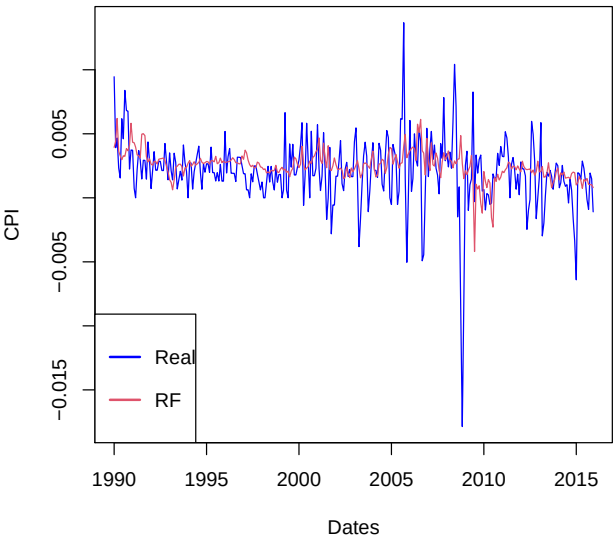

Figure 19: The real values and forecasts values for 11 horizon of the RF plotted.

E Graphs for real and forecasts for every horizon for BART

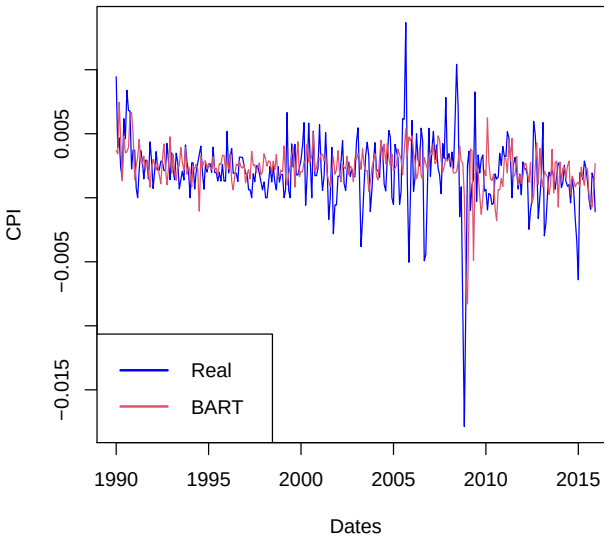

Figure 20: The real values and forecasts values for 2 horizon of the BART plotted.

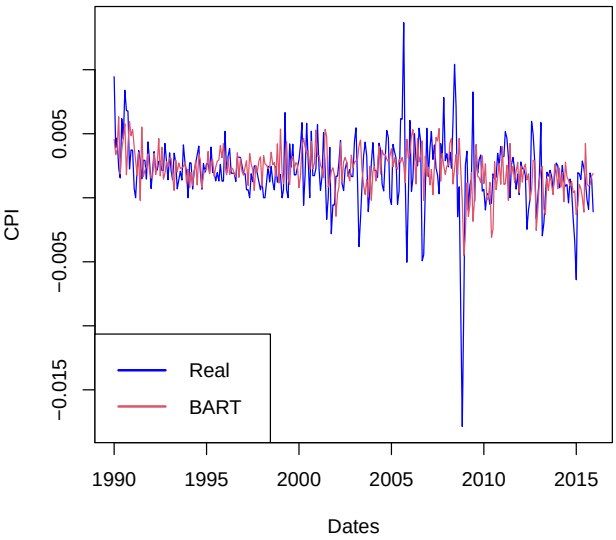

Figure 21: The real values and forecasts values for 3 horizon of the BART plotted.

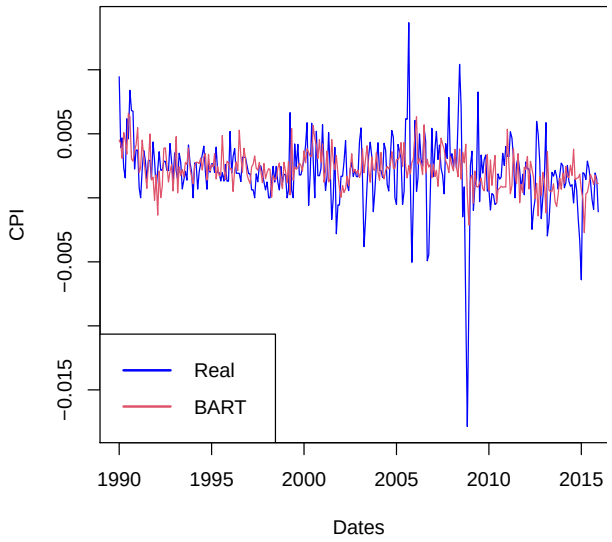

Figure 22: The real values and forecasts values for 4 horizon of the BART plotted.

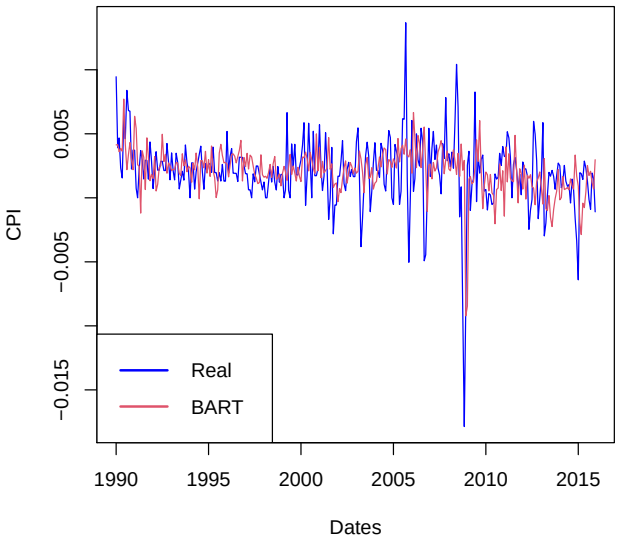

Figure 23: The real values and forecasts values for 5 horizon of the BART plotted.

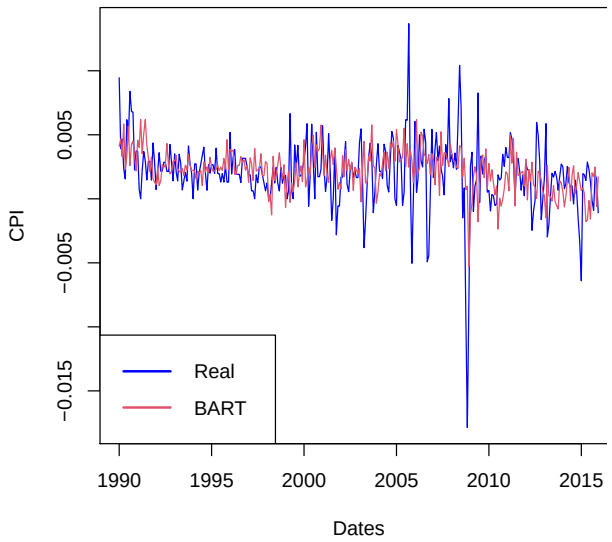

Figure 24: The real values and forecasts values for 6 horizon of the BART plotted.

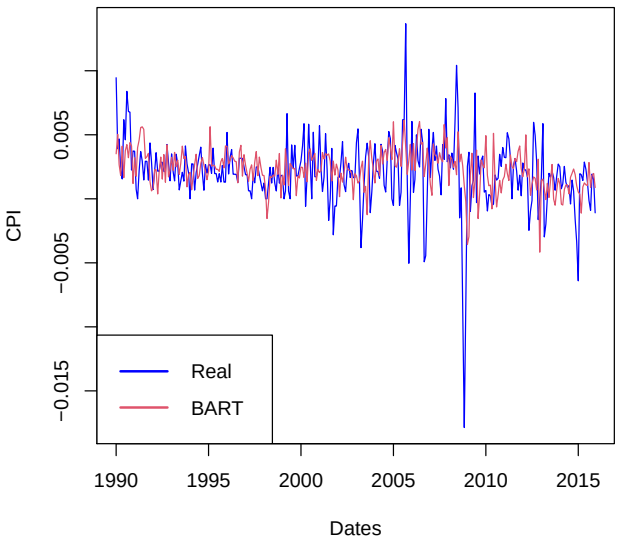

Figure 25: The real values and forecasts values for 7 horizon of the BART plotted.

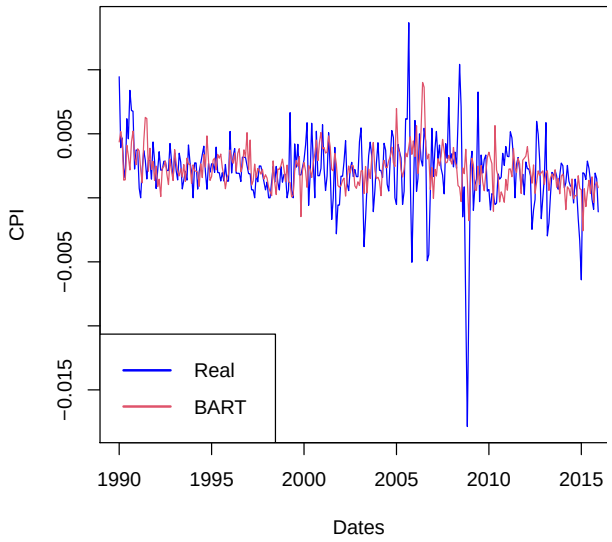

Figure 26: The real values and forecasts values for 8 horizon of the BART plotted.

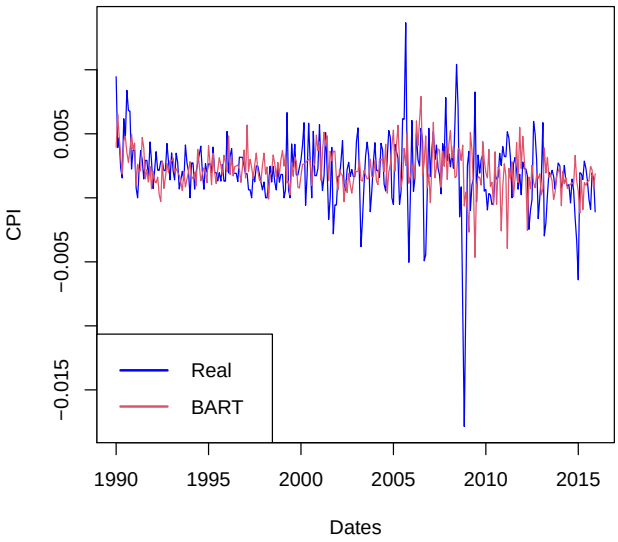

Figure 27: The real values and forecasts values for 9 horizon of the BART plotted.

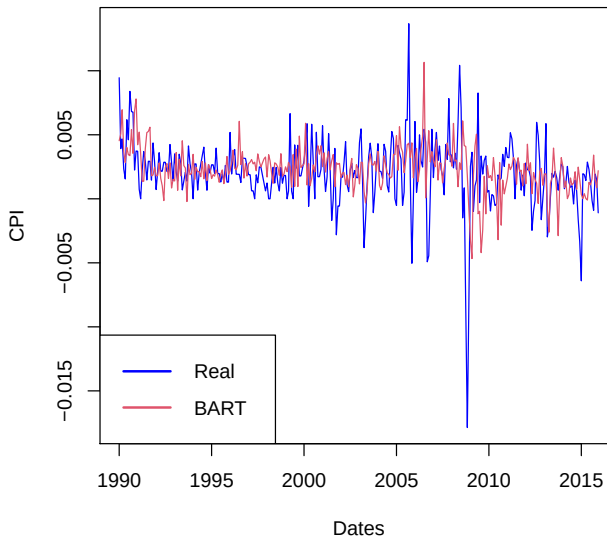

Figure 28: The real values and forecasts values for 10 horizon of the BART plotted.

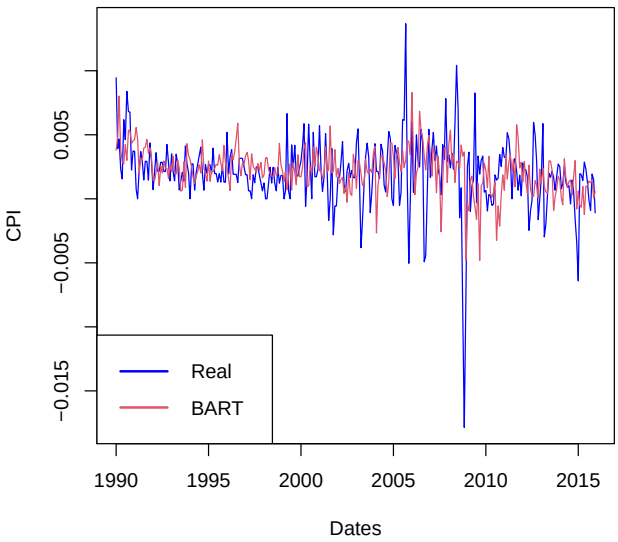

Figure 29: The real values and forecasts values for 11 horizon of the BART plotted.

F Graphs for real and forecasts for every horizon for MOTR-BART

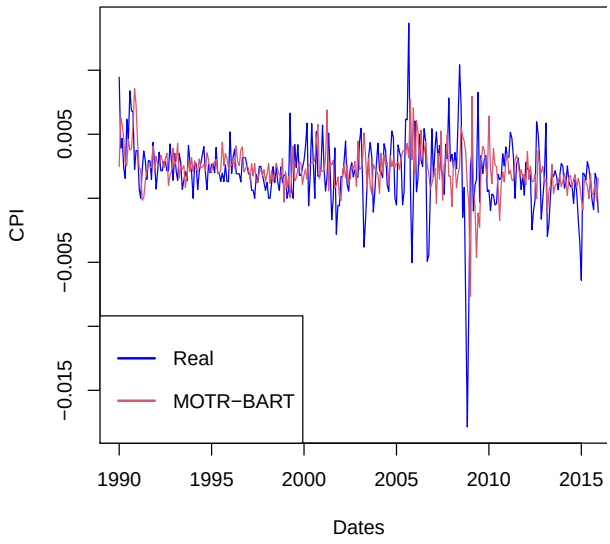

Figure 30: The real values and forecasts values for 2 horizon of the MOTR-BART plotted.

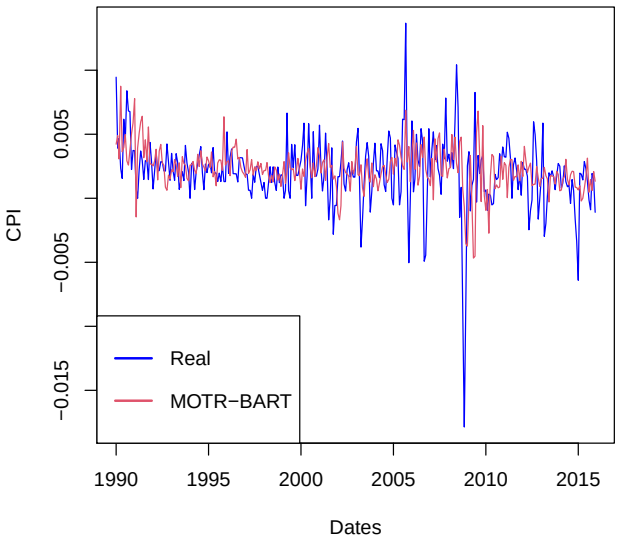

Figure 31: The real values and forecasts values for 3 horizon of the MOTR-BART plotted.

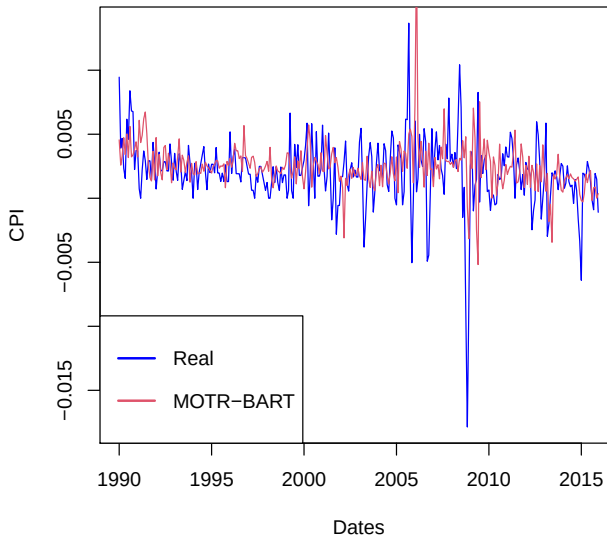

Figure 32: The real values and forecasts values for 4 horizon of the MOTR-BART plotted.

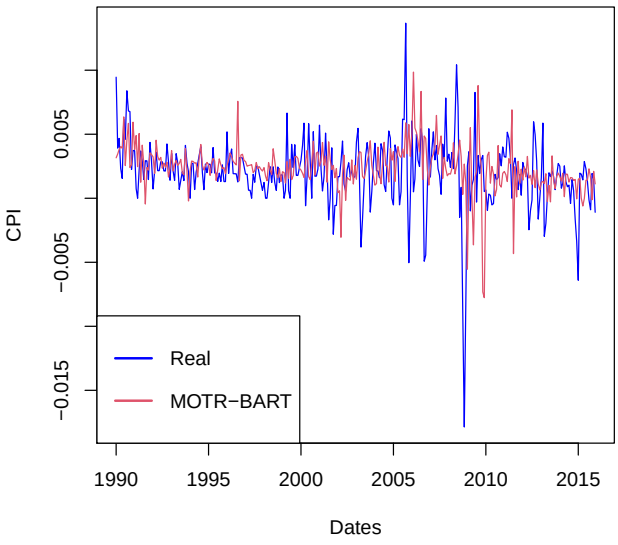

Figure 33: The real values and forecasts values for 5 horizon of the MOTR-BART plotted.

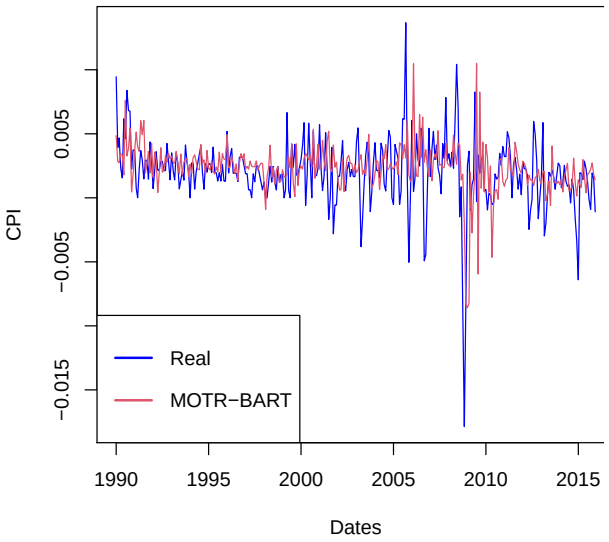

Figure 34: The real values and forecasts values for 6 horizon of the MOTR-BART plotted.

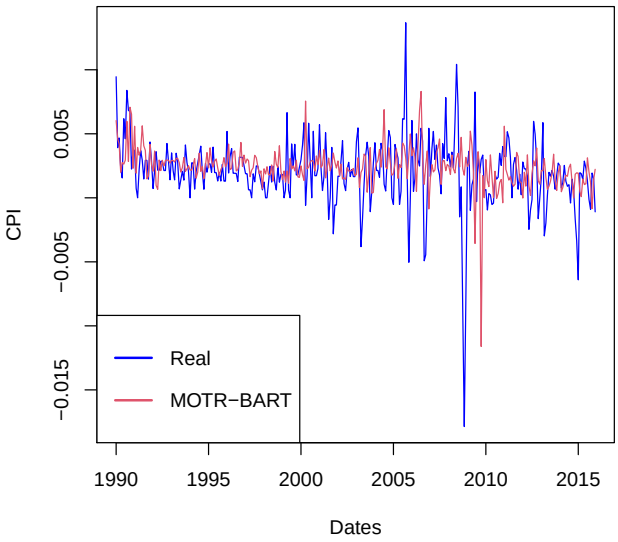

Figure 35: The real values and forecasts values for 7 horizon of the MOTR-BART plotted.

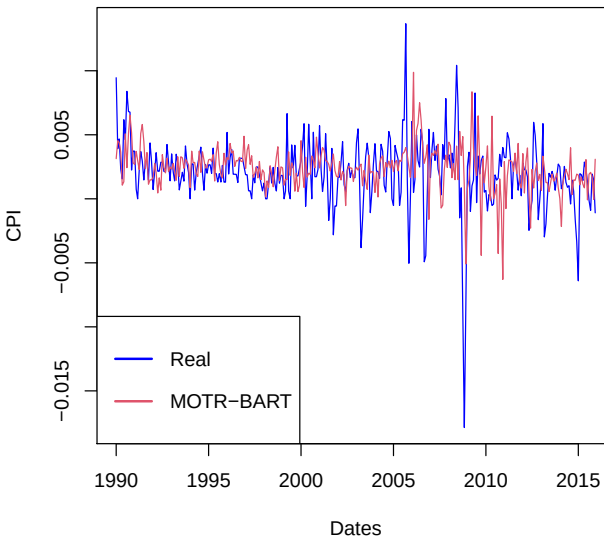

Figure 36: The real values and forecasts values for 8 horizon of the MOTR-BART plotted.

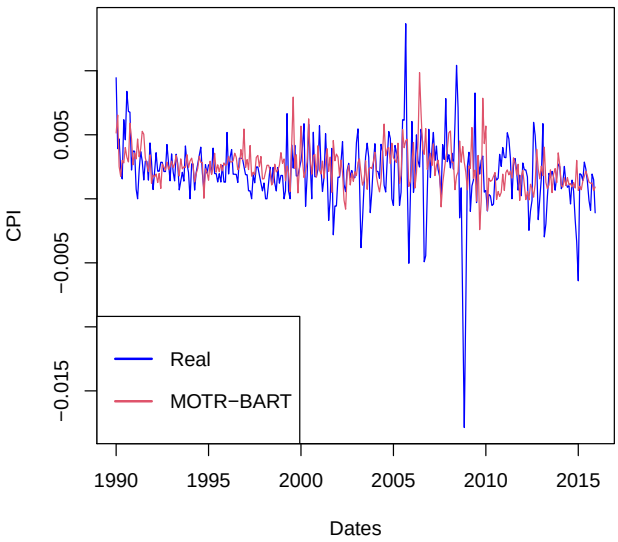

Figure 37: The real values and forecasts values for 9 horizon of the MOTR-BART plotted.

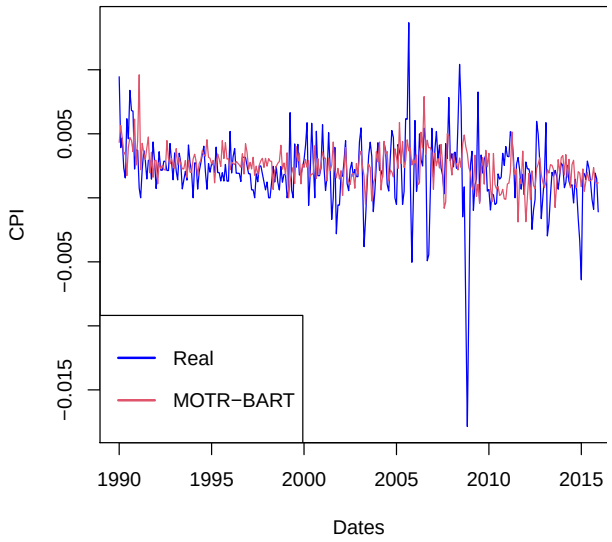

Figure 38: The real values and forecasts values for 10 horizon of the MOTR-BART plotted.

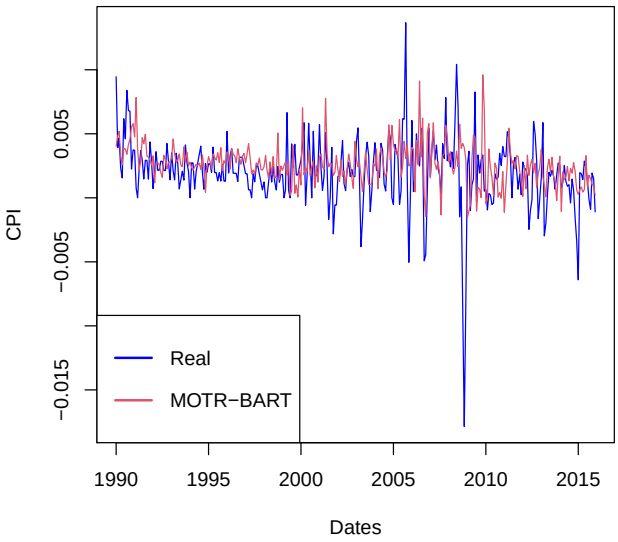

Figure 39: The real values and forecasts values for 11 horizon of the MOTR-BART plotted.

G Graphs for real and forecasts for every horizon for LLF

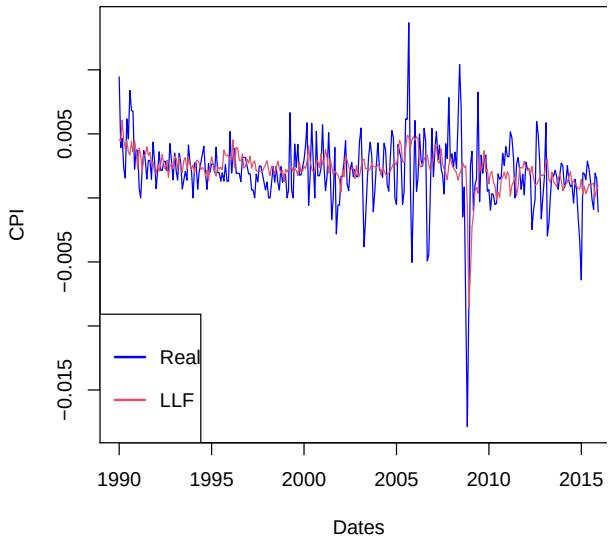

Figure 40: The real values and forecasts values for 2 horizon of the LLF plotted.

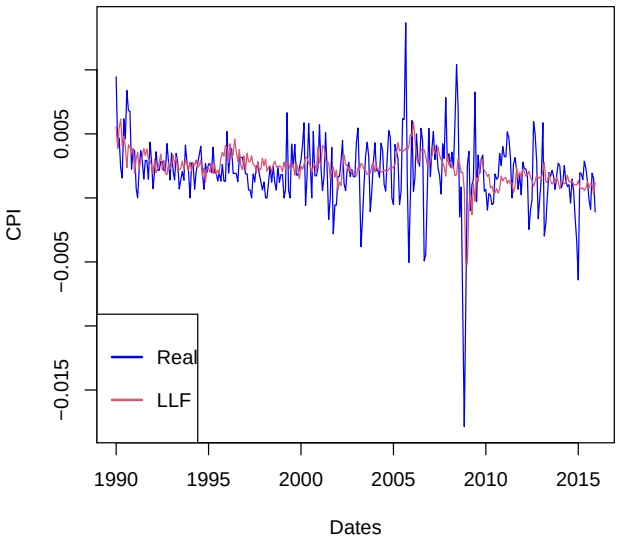

Figure 41: The real values and forecasts values for 3 horizon of the LLF plotted.

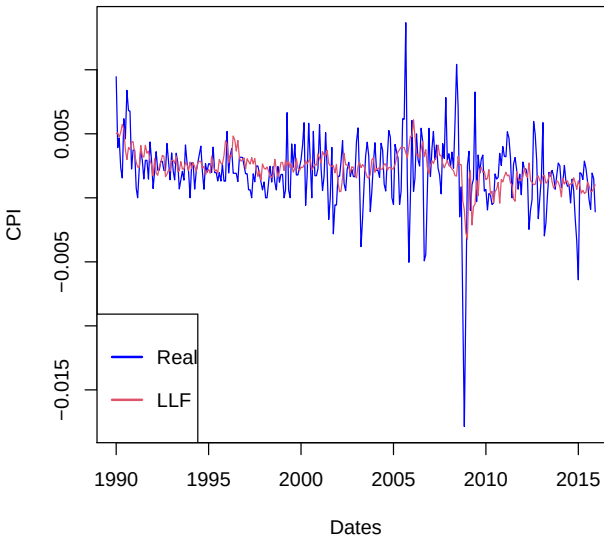

Figure 42: The real values and forecasts values for 4 horizon of the LLF plotted.

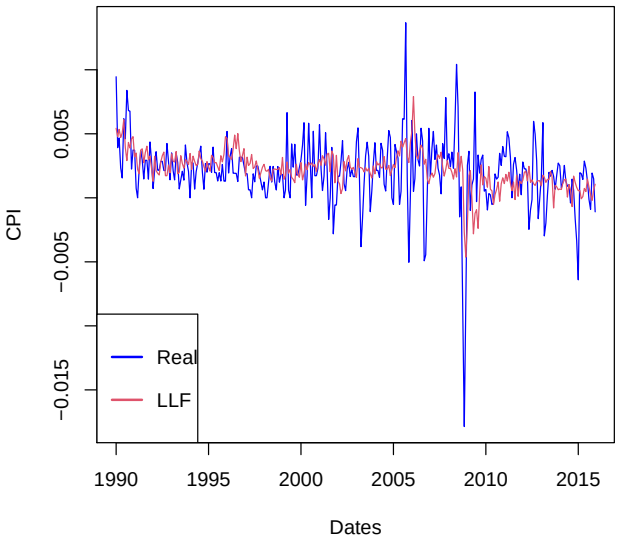

Figure 43: The real values and forecasts values for 5 horizon of the LLF plotted.

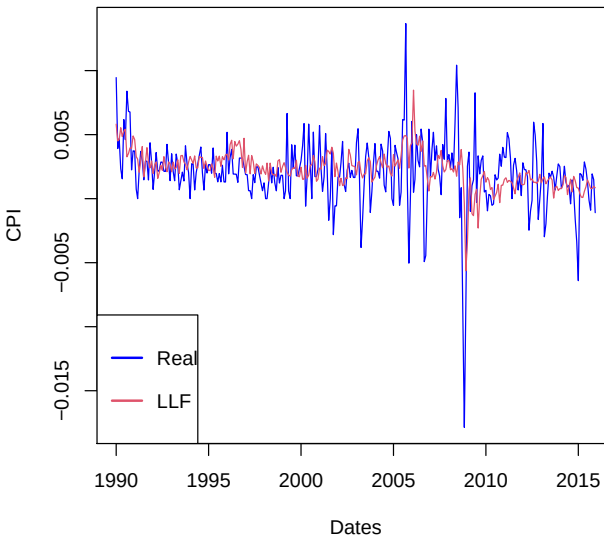

Figure 44: The real values and forecasts values for 6 horizon of the LLF plotted.

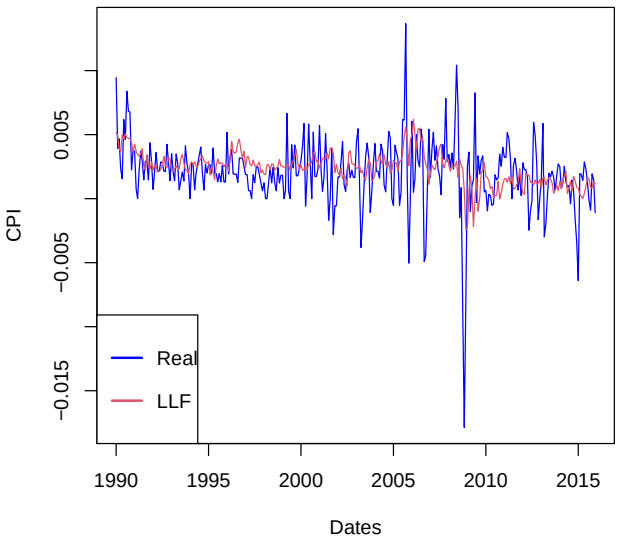

Figure 45: The real values and forecasts values for 7 horizon of the LLF plotted.

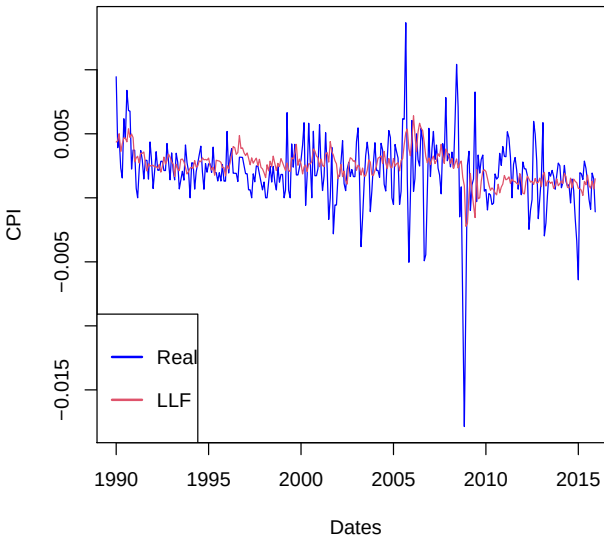

Figure 46: The real values and forecasts values for 8 horizon of the LLF plotted.

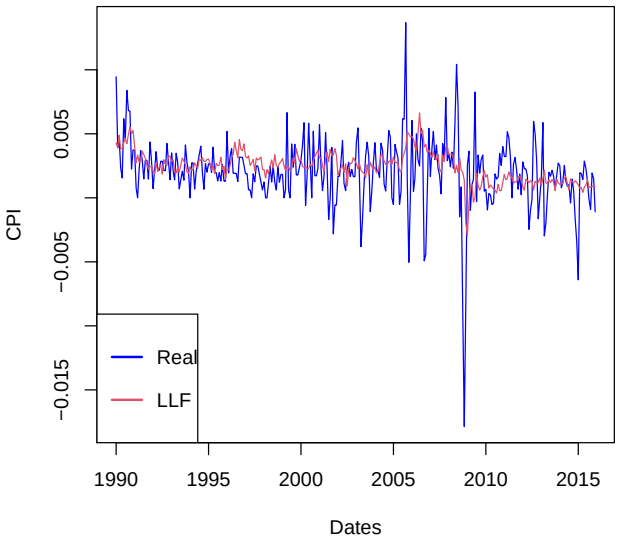

Figure 47: The real values and forecasts values for 9 horizon of the LLF plotted.

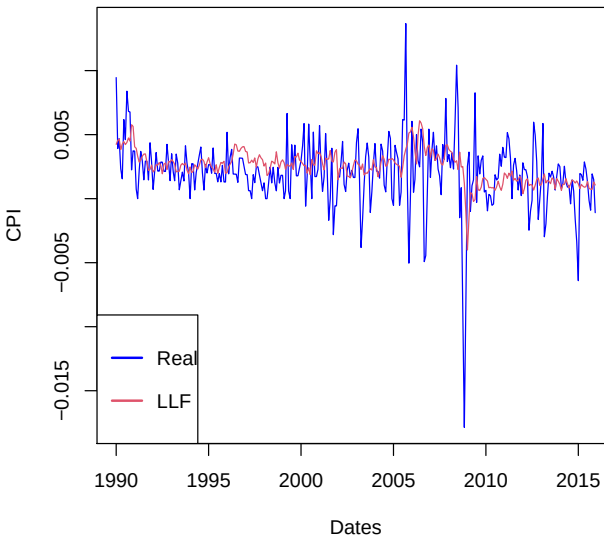

Figure 48: The real values and forecasts values for 10 horizon of the LLF plotted.

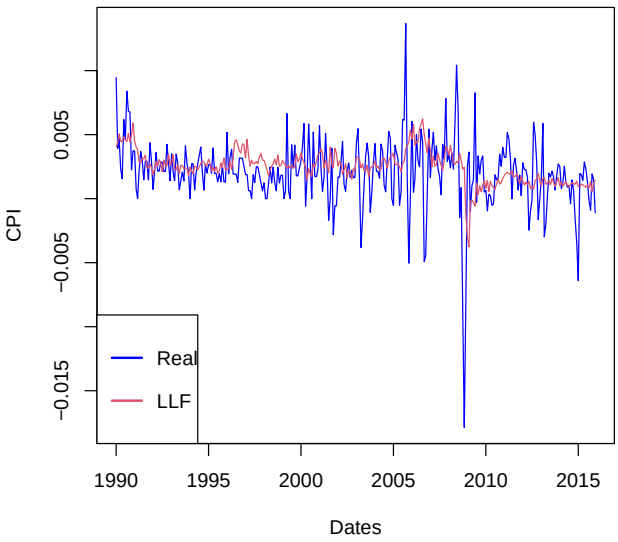

Figure 49: The real values and forecasts values for 11 horizon of the LLF plotted.