Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods

Marcelo C. Medeiros, Gabriel F. R. Vasconcelos, Álvaro Veiga & Eduardo Zilberman

To cite this article: Marcelo C. Medeiros, Gabriel F. R. Vasconcelos, Álvaro Veiga & Eduardo Zilberman (2019): Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2019.1637745

To link to this article: https://doi.org/10.1080/07350015.2019.1637745

Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods

Marcelo C. Medeiros

Department of Economics

Pontifical Catholic University of Rio de Janeiro

Rua Marquês de São Vicente 225, Gávea

Rio de Janeiro, 22451-900, BRAZIL

E-mail: mcm@econ.puc-rio.br

Gabriel F. R. Vasconcelos

Department of Electrical Engineering

Pontifical Catholic University of Rio de Janeiro

Rua Marquês de São Vicente 225, Gávea

Rio de Janeiro, 22451-900, BRAZIL

E-mail: gabrielrvsc@yahoo.com.br

Álvaro Veiga

Department of Electrical Engineering

Pontifical Catholic University of Rio de Janeiro

Rua Marquês de São Vicente 225, Gávea

Rio de Janeiro, 22451-900, BRAZIL

E-mail: alvf@ele.puc-rio.br

Eduardo Zilberman

Research Department

Central Bank of Chile

Agustinas 1180

Santiago, 867, CHILE

E-mail: ezilberman@bcentral.cl

JEL: C22, E31.

Keywords: Big data, inflation forecasting, LASSO, random forests, machine learning.

Acknowledgements: We are very grateful to the coeditor, Christian B. Hansen; the associate editor; and two anonymous referees for very helpful comments. We thank Federico Bandi, Anders B. Kock and Michael Wolf for comments during the workshop "Trends in Econometrics: Machine Learning, Big Data, and Financial Econometrics," Rio de Janeiro, October 2017. The participants of the "Third International Workshop in Financial Econometrics," Arraial d' Ajuda, October 2017, the workshop "Big data, Machine Learning and the Macroeconomy," Oslo, October 2017, and the National Symposium on Probability and Statistics (SINAPE), São Pedro, September 2018, are also gratefully acknowledged. We also thank Alberto Cavallo for an insightful discussion during the workshop “Measuring and Analyzing the Economy using Big Data” at the Central Bank of Brazil, November 2017. We also acknowledge the suggestions received during seminars at the Università Ca’ Foscari, Venice; at the School of Applied Mathematics at FGV/RJ; and at the Department of Economics at Duke University. Special thanks also to Tim Bollerslev, Jia Li, Andrew Patton, Rogier Quaedvlieg, Peter R. Hansen, Lucciano Villacorta, Mark Watson and Brian Weller for inspiring discussions. Finally, we are grateful to Michael McCracken for his help with the real-time dataset. The views expressed herein do not necessarily reflect the position of the Central Bank of Chile.

Abstract

Inflation forecasting is an important but difficult task. Here, we explore advances in machine learning (ML) methods and the availability of new datasets to forecast US inflation. Despite the skepticism in the previous literature, we show that ML models with a large number of covariates are systematically more accurate than the benchmarks. The ML method that deserves more attention is the random forest model, which dominates all other models. Its good performance is due not only to its specific method of variable selection but also the potential nonlinearities between past key macroeconomic variables and inflation.

1 Introduction

It is difficult to overemphasize the importance of forecasting inflation in rational economic decision-making. Many contracts concerning employment, sales, tenancy, and debt are set in nominal terms. Therefore, inflation forecasting is of great value to households, businesses and policymakers. In addition, central banks rely on inflation forecasts not only to inform monetary policy but also to anchor inflation expectations and thus enhance policy efficacy. Indeed, as part of an effort to improve economic decision-making, many central banks release inflation forecasts on a regular basis.

Despite the benefits of forecasting inflation accurately, improving simple models has proved challenging. As Stock and Watson (2010) emphasize, “it is exceedingly difficult to improve systematically upon simple univariate forecasting models, such as the Atkeson and Ohanian (2001) random walk model [...] or the time-varying unobserved components model in Stock and Watson (2007).” This conclusion is supported by a large literature (Faust and Wright 2013), but this literature has largely ignored the recent machine learning (ML) and “big data” boom in economics.[1] With a few exceptions, previous works either considered a restrictive set of variables or were based on a small set of factors computed from a larger number of predictors known as “diffusion indexes” (Stock and Watson 2002). In addition, most of these works focused on a time period when inflation was very persistent, which favors models that treat inflation as nonstationary.

“Big data” and ML methods are not passing fads, and investigating whether the combination of the two is able to provide more accurate forecasts is of paramount importance. Gu et al. (2018), for example, show that ML methods coupled with hundreds of predictors improve substantially out-of-sample stock return predictions. In a similar spirit, despite the previous skepticism, we argue that these methods lead to more accurate inflation forecasts. We find that the gains of using ML methods can be as large as 30% in terms of mean squared errors.[2] Moreover, this new set of models can help uncover the main predictors for future inflation, possibly shedding light on the drivers of price dynamics.

These findings are practically important in light of the large forecast errors that many central banks, international institutions and other forecasters have made recently. A striking example is the European Central Bank (ECB), whose projections have been systematically and substantially above realized inflation recently. Effective monetary policy depends on accurate inflation forecasts; otherwise, the policy stance will be tighter or looser than necessary. In addition, systematic forecasting errors may undermine central banks’ credibility and their ability to anchor inflation expectations. Furthermore, Svensson and Woodford (2004) argue that optimal monetary policy can be implemented through an inflation forecast-targeting regime in which central banks target inflation forecasts over a given horizon. Taken together, these arguments suggest potentially large welfare costs associated with failures to forecast inflation. Not surprisingly, such recent failures have fostered a debate on inflation forecasting practices.[3] This paper contributes to this debate by providing a comprehensive guide and assessment of ML methods for one important case study, US inflation.

1.1 Main Takeaways

We contribute to the literature in a number of ways. First, contrary to the previous evidence in Stock and Watson (1999, 2007), Atkeson and Ohanian (2001), and many others, our results show that it is possible to consistently beat univariate benchmarks for inflation forecasting, namely, random walk (RW), autoregressive (AR) and unobserved components stochastic volatility (UCSV) models. We consider several ML models in a data-rich environment from FRED-MD, a monthly database compiled by McCracken and Ng (2016), to forecast US consumer price index (CPI) inflation during more than twenty years of out-of-sample observations. We show that the gains can be as large as 30% in terms of mean squared errors. Our results are valid for different subsamples. Forecasting inflation is important to inform rational economic decision-making. However, as these decisions are made in real time, we check the robustness of our results by considering a real-time experiment from 2001 to 2015. The superiority of ML methods persists even in real time.

Second, we highlight the main set of variables responsible for these forecast improvements. Our results indicate that this set of variables is not sparse, which corroborates the findings of Giannone et al. (2018). Indeed, we find that ML models that do not impose sparsity are the best-performing ones. By contrast, the high level of aggregation of factor models, which have been among the most popular models for macroeconomic forecasting, is not adequate. Furthermore, either replacing standard principal component factors with target factors, as advocated by Bai and Ng (2008), or using boosting to select factors as in Bai and Ng (2009) improves the results only marginally.

Finally, we pay special attention to a particular ML model, the random forest (RF) of Breiman (2001), which systematically outperforms the benchmarks and several additional ML and factor methods: the least absolute shrinkage and selection operator (LASSO) family, which includes LASSO, adaptive LASSO (adaLASSO), elastic net (ElNet) and the adaptive elastic net (adaElNet); ridge regression (RR); Bayesian vector autoregressions (BVAR); and principal component factors, target factors (T. Factor) and linear ensemble methods such as bagging, boosted factors (B. Factor), complete subset regressions (CSR) and jackknife model averaging (JMA). RF models are highly nonlinear nonparametric models that have a tradition in statistics but have only recently attracted attention in economics. This late success is partly due to the new theoretical results developed by Scornet et al. (2015) and Wagner and Athey (2018). As robustness checks, we also compare the RF models with four nonlinear alternatives: deep neural networks (Deep NN), boosted trees (BTrees), and a polynomial model estimated either by LASSO or adaLASSO. We also compare our results to a linear model estimated with nonconcave penalties (SCAD) as in Fan and Li (2001).

There are several reasons why our results differ from those of Stock and Watson (1999, 2007), and Atkeson and Ohanian (2001). The combination of ML methods with a large dataset not considered by these authors is an obvious one. The choice of a different sample can also explain the differences. The above papers work in an environment where inflation is clearly integrated, whereas inflation is stationary in the period considered in this paper. This fact alone makes both the RW and UCSV models less attractive, as clearly they are not suitable for stationary data. Nevertheless, if the gains of the ML methods are due solely to the fact that inflation is much less persistent, we would observe competitive performance of the AR or factor models, which is not the case. Although the performance of the RW and UCSV models improves when accumulated inflation is considered, the ML methods still achieve superior results. By construction, accumulated inflation is much more persistent compared to the monthly figures. Given all of these reasons, ML methods coupled with large datasets merit serious consideration for forecasting inflation.

To open the black box of ML methods, we compare the variables selected by the adaLASSO, RR, and RF models. Following McCracken and Ng (2016), we classify the variables into eight groups: (i) output and income; (ii) labor market; (iii) housing; (iv) consumption, orders and inventories; (v) money and credit; (vi) interest and exchange rates; (vii) prices; and (viii) stock market. In addition, we consider AR terms and the factors computed from the full set of predictors. The most important variables for RR and RF models are stable across horizons but are quite different between the two specifications. For RR, AR terms, prices and employment are the most important predictors, resembling a sort of backward-looking Phillips curve, whereas RF models give more importance to disaggregated prices, interest-exchange rates, employment and housing. The RF model resembles a nonlinear Phillips curve in which the backward-looking terms are disaggregated. The adaLASSO selection is quite different across forecasting horizons and is, by construction and in opposition to RF and RR models, sparse. Only AR terms retain their relative importance independent of the horizon, and prices gradually lose their relevance until up to six months ahead but partially recover for longer horizons. Output-income variables are more important for medium-term forecasts. Finally, none of the three classes of models selects either factors or stocks, not even RR or RF, which produces a nonsparse solution. This result may indicate that the high level of cross-section aggregation of the factors is one possible cause for its poor performance.

To disentangle the effects of variable selection from nonlinearity, we consider two alternative models. The first uses the variables selected by RF and estimates a linear specification by OLS. The second method estimates RF with only the regressors selected by adaLASSO. Both models outperform RF only for one-month-ahead forecasting. For longer horizons, the RF model is still the winner, which provides evidence that both nonlinearity and variable selection play a key role in the superiority of the RF model.

There are many sources of nonlinearities that could justify the superiority of the RF model. For instance, the relationship between inflation and employment is nonlinear to the extent that it depends on the degree of slackness in the economy. Another source of nonlinearity is economic uncertainty, as this uncertainty increases the option value of economic decision delays if they have an irreversible component (Bloom 2009). For example, if it is expensive to dismiss workers, hiring should be nonlinear on uncertainty. In addition, this real option argument also makes households and businesses less sensitive to changes in economic conditions when uncertainty is high. Hence, the responses of employment and inflation to interest rate decisions are arguably nonlinear on uncertainty. The presence of a zero lower bound on nominal interest rates and the implications of this bound for unconventional monetary policy is another source of nonlinearity among inflation, employment and interest rate variables (Eggertsson and Woodford 2003). Finally, to the extent that houses serve as collateral for loans, it interacts with monetary policy (Iacoviello 2005) and financial intermediation (Mian and Sufi 2009). As in the Great Recession, a housing bubble can form, resulting in a deep credit crash (Shiller 2014). Needless to say, these interactions are highly nonlinear and arguably have nonlinear effects on inflation, employment and interest rates. In line with these arguments, we show that the gains of the RF model are larger during recessions and periods of high uncertainty, especially during and after the Great Recession.

1.2 A Brief Comparison to the Literature

The literature on inflation forecasting is vast, and there is substantial evidence that models based on the Phillips curve do not provide good forecasts. Although Stock and Watson (1999) showed that many production-related variables are potential predictors of US inflation, Atkeson and Ohanian (2001) showed that in many cases, the Phillips curve fails to beat even simple naive models. These results inspired researchers to seek different models and variables to improve inflation forecasts. Among the variables used are financial variables (Forni et al. 2003), commodity prices (Chen et al. 2014) and expectation variables (Groen et al. 2013). However, there is no systematic evidence that these models outperform the benchmarks.

Recently, due to advancements in computational power, theoretical developments in ML, and the availability of large datasets, researchers have started to consider the usage of high-dimensional models in addition to the well-established (dynamic) factor models. However, most of these studies have either focused only on a very small subset of ML models or presented a restrictive analysis. For example, Inoue and Kilian (2008) considered bagging, factor models and other linear shrinkage estimators to forecast US inflation with a small set of real economic activity indicators. Their study is more limited than ours both in terms of the pool of models and richness of the set of predictors. Nevertheless, the authors are among the few voices suggesting that ML techniques can deliver nontrivial gains over univariate benchmarks. Medeiros and Mendes (2016) provided evidence that LASSO-based models outperform both factor and AR benchmarks in forecasting US CPI. However, their analysis was restricted to a single ML method for only one-month-ahead forecasting.

The literature has mainly explored linear ML models. One explanation for this limitation is that several of the papers in the early days considered only univariate nonlinear models that were, in most cases, outperformed by simple benchmarks; see Teräsvirta et al. (2005). An exception is Nakamura (2005), who showed that neural networks outperform univariate autoregressive models for short horizons.

Recently, Garcia et al. (2017) applied a large number of ML methods, including RFs, to real-time inflation forecasting in Brazil. The results indicated a superiority of the CSR method of Elliott et al. (2013). However, an important question is whether this is a particular result for Brazil or if similar findings can be replicated for the US economy. The first difference between the results presented here and those in Garcia et al. (2017) is that the RF model robustly outperforms its competitors and CSR does not perform particularly well. With respect to the set of models considered, on the one hand, in this paper we employ more ML models than Garcia et al. (2017), but on the other hand, we do not have a real-time daily database of inflation expectations as in the case of Brazil. We also provide a much richer discussion of variable selection and the nature of the best-performing models.

Finally, it is important to contextualize our work in light of the criticisms of Makridakis et al. (2018) with respect to the ability of ML methods to produce reliable forecasts. The methods considered here are much different than those in the study of Makridakis et al. (2018). While we consider modern ML tools such as RF, shrinkage, and bagging, the authors focus on simple regression trees, shallow neural networks and support vector machines. Furthermore, the models considered in Makridakis et al. (2018) are univariate in the sense that no other variables apart from lags are used as predictors.

1.3 Organization of the Paper

Section 2 gives an overview of the data. Section 3 describes the forecasting methodology. Section 4 describes the models used in the paper. Section 5 provides the main results. Section 6 concludes. The online Supplementary Material provides additional results. Tables and figures labeled with an “S” refer to this supplement.

2 Data

Our data consist of variables from the FRED-MD database, which is a large monthly macroeconomic dataset designed for empirical analysis in data-rich environments. The dataset is updated in real time through the FRED database and is available from Michael McCraken’s webpage.[4] For further details, we refer to McCracken and Ng (2016).

We use the vintage as of January 2016. Our sample extends from January 1960 to December 2015 (672 observations), and only variables with all observations in the sample period are used (122 variables). In addition, we include as potential predictors the four principal component factors computed from this set of variables. We consider four lags of all variables, as well as four autoregressive terms. Hence, the analysis contemplates 508 potential predictors. The out-of-sample window is from January 1990 to December 2015. All variables are transformed to achieve stationarity as described in the Supplementary Material. π_t is the inflation in month t computed as $\pi_t = \log(\mathrm{P}_t) - \log(\mathrm{P}_{t-1})$, and P_t is a given price index in period t. The baseline price index is the CPI, but in the Supplementary Material we report results for the PCE and the core CPI inflation. Figure S.1 in this supplement displays the evolution of inflation measures during the full sample period. The non-core inflation measures have a large outlier in November 2008 associated with a large decline in oil prices. This outlier can certainly affect the estimation of the models considered in this paper. In order to attenuate its effects, we include a dummy variable for November 2009 in all models estimated after that date. We do not include any look-ahead bias, as the dummy is added only after the outlier has been observed by the econometrician.

We compare performance across models in the out-of-sample window and in two different subsample periods, namely, January 1990 to December 2000 (132 observations) and January 2001 to December 2015 (180 observations). The first sample corresponds to a period of low inflation volatility ($\sigma = 0.17\%$), while in the second sample, inflation is more volatile ($\sigma = 0.32\%$). However, on average, inflation is higher during 1990–2000 than 2001–2015 and much more persistent as well. Relative to the 1990-2000 period, inflation was more volatile near the recession in the early 1990s. S.9 in the Supplementary Material provides descriptive statistics and gives an overview of the economic scenario in each subsample.

As a robustness check, we also report the results of a real-time experiment using our second subsample, from 2001 to 2015. We choose this particular period because the real-time vintages are easily available from the FRED-MD database. For the real-time experiment, the number of potential regressors varies according to the vintage.

3 Methodology

Consider the following model:

$$\pi_{t+h} = G_h(\boldsymbol{x}_t) + u_{t+h}, \quad h = 1, \ldots, H, \quad t = 1, \ldots, T, \qquad (1)$$

where π_{t+h} is the inflation in month $t+h$; $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{nt})'$ is a n-vector of covariates possibly containing lags of π_t and/or common factors as well as a large set of potential predictors; $G_h(\cdot)$ is the mapping between covariates and future inflation; and u_t is a zero-mean random error. The target function $G_h(\boldsymbol{x}_t)$ can be a single model or an ensemble of different specifications. There is a different mapping for each forecasting horizon.

The direct forecasting equation is given by

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

where $\hat{G}_{h,t-R_h+1:t}$ is the estimated target function based on data from time $t - R_h + 1$ up to t and R_h is the window size, which varies according to the forecasting horizon and the number of lagged variables in the model. We consider direct forecasts as we do not make any attempt to predict the covariates. The only exception is the case of the BVAR model, where joint forecasts for all predictors are computed in a straightforward manner following the procedure described in Bańbura et al. (2010).

The forecasts are based on a rolling-window framework of fixed length. However, the actual in-sample number of observations depends on the forecasting horizon. For example, for the 1990–2000 period, the number of observations is $R_h = 360 - h - p - 1$, where p is the number of lags in the model. For 2001–2015, $R_h = 492 - h - p - 1$. We choose to work in a rolling-window scheme for two reasons: to attenuate the effects of potential structural breaks and outliers and to avoid problems of running superior predictive performance tests among nested models; see the discussion in Giacomini and White (2006). For instance, with a rolling-window scheme, the unconditional Giacomini-White (GW) test is equivalent to the traditional Diebold-Mariano (DM) test.

In addition to three benchmark specifications (RW, AR and UCSV models), we consider factor-augmented AR models, sparsity-inducing shrinkage estimators (LASSO, adaLASSO, ElNet and adaElNet), other shrinkage methods that do not induce sparsity (RR and BVAR with Minnesota priors), averaging (ensemble) methods (bagging, CSR and JMA) and RF. Bagging and CSR can be viewed as nonsparsity-inducing shrinkage estimators. With respect to the factor-augmented AR models, we consider in addition to the standard factors computed with principal component analysis a set of target factors as in Bai and Ng (2008) and boosted factors as in Bai and Ng (2009). We also include in the comparison three different model combination schemes, namely, the simple average, the trimmed average and the median of the forecasts. For both the shrinkage and factor-based methods, the set of predictors are standardized before estimation.

We find that RF, a highly nonlinear method, robustly outperforms the other methods. To disentangle the role of variable selection from nonlinearity, we also consider a linear model where the regressors are selected by RFs (RF/ordinary least squares, OLS) and an RF model with regressors preselected by adaLASSO (adaLASSO/RF).

Forecasts for the accumulated inflation over the following three, six and twelve months are computed, with the exception of the RW and UCSV models, by aggregating the individual forecasts for each horizon. In the case of the RW and UCSV models, a different specification is used to construct the forecasts (see below).

4 Models

4.1 Benchmark Models

The first benchmark is the RW model, where for $h = 1, \ldots, 12$, the forecasts are as $\hat{\pi}_{t+h|t} = \pi_t$. For the accumulated *h*-month forecast, we set $\hat{\pi}_{t+1:t+h|t} = \pi_{t-(h-1):t}$, where $\pi_{t-(h-1):t}$ is the accumulated inflation over the previous *h* months.

The second benchmark is the autoregressive (AR) model of order *p*, where *p* is determined by the Bayesian information criterion (BIC) and the parameters are estimated by OLS. The forecast equation is $\hat{\pi}_{t+h|t} = \hat{\phi}_{0,h} + \hat{\phi}_{1,h}\pi_t + \ldots + \hat{\phi}_{p,h}\pi_{t-p+1}$. There is a different model for each horizon. The accumulated forecasts are computed by aggregating the individual forecasts.

Finally, the third benchmark is the UCSV model, which is described as follows:

$$\pi_t = \tau_t + e^{h_t/2}\varepsilon_t, \quad \tau_t = \tau_{t-1} + u_t, \quad h_t = h_{t-1} + v_t,$$

where $\{\varepsilon_t\}$ is a sequence of independent and normally distributed random variables with zero mean and unit variance and u_t and v_t are also normal with zero mean and variance given by inverse-gamma priors. $\tau_1 \sim \mathrm{N}(0, V_\tau)$, and $h_1 \sim \mathrm{N}(0, V_h)$, where $V_\tau = V_h = 0.12$. The model is estimated by Markov Chain Monte Carlo (MCMC) methods. The *h*-steps-ahead forecast is computed as $\hat{\pi}_{t+h} = \hat{\tau}_{t|t}$. For accumulated forecasts, the UCSV is estimated with the accumulated *h*-month inflation as the dependent variable.

4.2 Shrinkage

In this paper we estimate several shrinkage estimators for linear models where $G_h(\boldsymbol{x}_t) = \boldsymbol{\beta}_h{}'\boldsymbol{x}_t$ and

$$\boldsymbol{\beta}_h = \arg\min_{\boldsymbol{\beta}_h}\left[\sum_{t=1}^{T-h}\left(y_{t+h} - \boldsymbol{\beta}_h{}'\boldsymbol{x}_t\right)^2 + \sum_{i=1}^{n} p(\beta_{h,i}; \lambda, \omega_i)\right]. \quad (3)$$

$p(\beta_{h,i}; \lambda, \omega_i)$ is a penalty function that depends on the penalty parameter *λ* and on a weight $\omega_i > 0$. We consider different choices for the penalty functions.

4.2.1 Ridge Regression (RR)

RR was proposed by Hoerl and Kennard (1970). The penalty is given by

$$\sum_{i=1}^{n} p(\beta_{h,i};\lambda,\omega_i) := \lambda \sum_{i=1}^{n} \beta_{h,i}^2. \quad (4)$$

RR has the advantage of having an analytic solution that is easy to compute, and it also shrinks the coefficients associated with less-relevant variables to nearly zero. However, the coefficients rarely reach exactly zero for any size of λ.

4.2.2 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO was proposed by Tibshirani (1996), where the penalty is

$$\sum_{i=1}^{n} p(\beta_{h,i};\lambda,\omega_i) := \lambda \sum_{i=1}^{n} |\beta_{h,i}|. \quad (5)$$

LASSO shrinks the irrelevant variables to zero. However, model selection consistency is achieved only under very stringent conditions.

4.2.3 Adaptive Least Absolute Shrinkage and Selection Operator (adaLASSO)

adaLASSO was proposed by Zou (2006) to achieve model selection consistency. adaLASSO uses the same penalty as LASSO with the inclusion of a weighting parameter that comes from a first-step estimation. The penalty is given by

$$\sum_{i=1}^{n} p(\beta_{h,i};\lambda,\omega_i) := \lambda \sum_{i=1}^{n} \omega_i \, |\beta_{h,i}|, \qquad (6)$$

where $\omega_i = |\beta_{h,i}^{*}|^{-1}$ and $\beta_{h,i}^{*}$ is the coefficient from the first-step estimation. adaLASSO can deal with many more variables than observation and works well in non-Gaussian environments and under heteroskedasticity (Medeiros and Mendes 2016).

4.2.4 Elastic Net (ElNet)

ElNet is a generalization that includes LASSO and RR as special cases. It is a convex combination of the ℓ_1 and ℓ_2 norms (Zou and Hastie 2005). The ElNet penalty is defined as

$$\sum_{i=1}^{n} p(\beta_{h,i};\lambda,\omega_i) := \alpha\lambda\sum_{i=1}^{n}\beta_{h,i}^{2} + (1-\alpha)\lambda\sum_{i=1}^{n}|\beta_{h,i}|; \qquad (7)$$

where $\alpha \in [0,1]$. We also consider an adaptive version of ElNet (adaElNet), which works in the same way as adaLASSO.

4.3 Factor Models

The idea of the factor model is to extract common components from all predictors, thus reducing the model dimension. Factors are computed as principal components of a large set of variables $\boldsymbol{z}_t$ such that $\boldsymbol{F}_t = \boldsymbol{A}\boldsymbol{z}_t$, where $\boldsymbol{A}$ is a rotation matrix and $\boldsymbol{F}_t$ is the vector of the principal components. Consider equation (1). In this case, $\boldsymbol{x}_t$ is given by π_{t-j}, $j = 0,1,2,3$ plus $\boldsymbol{f}_{t-j}$, $j = 0,1,2,3$, where $\boldsymbol{f}_t$ is the vector with the first four principal components of $\boldsymbol{z}_t$. The theory behind factor models can be found in Bai and Ng (2003).

4.3.1 Target Factors

To improve the forecasting performance of factor models, Bai and Ng (2008) proposed targeting the predictors. The idea is that if many variables in $\boldsymbol{z}_t$ are irrelevant predictors of π_{t+h}, factor analysis using all variables may result in noisy factors with poor forecasting ability. The idea is to compute the principal components only of the variables with high prediction power for future inflation. Let $z_{i,t}$, $i = 1,\dots,q$ be the candidate variables and $\boldsymbol{w}_t$ a set of controls. We follow Bai and Ng (2008) and use lagged values of π_t as controls. The procedure is described as follows.

(1) For $i = 1,\dots,q$, regress π_{t+h} on $\boldsymbol{w}_t$ and $z_{i,t}$ and compute the *t*-statistics for the coefficient corresponding to $z_{i,t}$.

(2) Choose a significance level α and select all variables that are significant using the computed *t*-statistics.

(3) Let $\boldsymbol{z}_t(\alpha)$ be the selected variables from steps 1–2. Estimate the factors $\boldsymbol{F}_t$ from $\boldsymbol{z}_t(\alpha)$ by the principal components.

(4) Regress π_{t+h} on $\boldsymbol{w}_t$ and $\boldsymbol{f}_{t-j},\ j=0,1,2,3$, where $\boldsymbol{f}_t \subset \boldsymbol{F}_t$. The number of factors in $\boldsymbol{f}_t$ is selected using the BIC.

The same procedure was used by Medeiros and Vasconcelos (2016) to forecast US CPI inflation. The authors showed that the target factors slightly reduce the forecasting errors.

4.3.2 Factor Boosting

The optimal selection of factors for predictive regressions is an open problem in the literature. Even if the factor structure is clear in the data, it is not obvious that only the most relevant factors should be included in the predictive regression. We adopt the boosting algorithm as in Bai and Ng (2008) to select the factors and the number of lags in the model. Define $\boldsymbol{z}_t \in \mathbb{R}^q$, the set of all *n* factors computed from the original *n* variables plus four lags of each factor. Therefore, $q=5n$.

The algorithm is defined as follows:

(1) Let $\Phi_{t,0}=\bar{\pi}$ for each *t*, where $\bar{\pi}=\frac{1}{t}\sum_{i=1}^{t}\pi_i$.

(2) For $m=1,\ldots,M$: (a) Compute $\hat{u}_t=\pi_t-\Phi_{t-h,m-1}$. (b) For each candidate variable $i=1,\ldots,q$, regress the current residual on $z_{i,t}$ to obtain $\hat{b}_i$, and compute $\hat{e}_{t,i}=\hat{u}_t-z_{i,t}\hat{b}_i$. Calculate $SSR_i=\hat{\boldsymbol{e}}_i{}'\hat{\boldsymbol{e}}_i$. (c) Select i_m^* as the index of the variable that delivers the smallest *SSR*, and define $\hat{\phi}_{t,m}=z_{i_m^*,t}\hat{b}_{i_m^*}$. (d) Update $\hat{\Phi}_{t,m}=\hat{\Phi}_{t,m-1}+v\phi_{t,m}$, where *v* is the step length. We set *v* = 0.2.

(3) Stop the algorithm after the *M*th iteration or when the BIC starts to increase.

4.4 Ensemble Methods

Ensemble forecasts are constructed from a (weighted) average of the predictions of an ensemble of methods.

4.4.1 Bagging

The term “bagging” comes from bootstrap aggregation, which was proposed by Breiman (1996). The idea is to combine forecasts from several unstable models estimated for different bootstrap subsamples. Normally, there is much more to gain from combinations of models if they are very different. The bagging steps are as follows:

(1) For each bootstrap sample, run an OLS regression with all candidate variables and select those with an absolute t-statistic above a certain threshold *c*.

(2) Estimate a new regression only with the variables selected in the previous step.

(3) The coefficients from the second regression are finally used to compute the forecasts **on the actual sample**.

(4) Repeat the first three steps for *B* bootstrap samples and compute the final forecast as the average of the *B* forecasts.

Note that in our case, the number of observations may be smaller than the number of variables, which makes the regression in the first step unfeasible. We solve this issue by, for each bootstrap subsample, randomly dividing all variables in groups and running the pretesting step for each one of the groups.

4.4.2 Complete Subset Regressions (CSR)

CSR was developed by Elliott et al. (2013, 2015). The motivation was that selecting the optimal subset of $\mathbf{x}_t$ to predict π_{t+h} by testing all possible combinations of regressors is computationally very demanding and, in most cases, unfeasible. Suppose that we have *n* candidate variables. The idea is to select a number $q \leq n$ variables and run regressions using all possible combinations *q* of *n* variables. The final forecast is the average over all forecasts.

CSR handles a small number of variables. For large sets, the number of regressions to be estimated increases rapidly. For example, with *n* = 25 and *q* = 4, there are 12, 650 regressions. Therefore, we adopt a pretesting procedure similar to that used with the target factors. We start fitting a regression of π_{t+h} on each of the candidate variables (including lags) and save the *t*-statistics of each variable.[5] The *t*-statistics are ranked by absolute value, and we select the $\tilde{n}$ variables that are more relevant in the ranking. The CSR forecast is calculated on these variables.

4.4.3 Jackknife Model Averaging (JMA)

JMA is a method to combine forecasts from different models. Instead of using simple average, JMA uses leave-one-out cross-validation to estimate optimal weights; see Hansen and Racine (2012) and Zhang et al. (2013).

Suppose we have *M* candidate models that we want to average from and write the forecast of each model as $\hat{\pi}_{t+h}^{(m)}, m = 1, \ldots, M$. Set the final forecast as

$$\hat{\pi}_{t+h} = \sum_{m=1}^{M} \omega_m \hat{\pi}_{t+h}^{(m)},$$

where $0 \le \omega_m \le 1$ for all $m \in \{1, \ldots, M\}$ and $\sum_{m=1}^{M} \omega_m = 1$.

The JMA procedure is as follows:

(1) For each observation of $(\boldsymbol{x}_t, \pi_{t+h})$: (a) Estimate all candidate models leaving the selected observation out of the estimation. Since we are in a time series framework with lags in the model, we also remove four observations before and four observations after $(\boldsymbol{x}_t, \pi_{t+h})$. (b) Compute the forecasts from each model for the observations that were removed in the previous step.

(2) Choose the weights that minimize the cross-validation errors subject to the constraints previously described.

Each candidate model in the JMA has four lags of the inflation and four lags of one candidate variable.

4.5 Random Forests

The RF model was proposed by Breiman (2001) to reduce the variance of regression trees and is based on bootstrap aggregation (bagging) of randomly constructed regression trees. In turn, a regression tree is a nonparametric model that approximates an unknown nonlinear function with local predictions using recursive partitioning of the space of the covariates (Breiman 1996).

To understand how a regression tree works, an example from Hastie et al. (2001) is useful. Consider a regression problem in which X_1 and X_2 are explanatory variables, each taking values in some given interval, and Y is the dependent variable. We first split the space into two regions at $X_1 = s_1$, and then the region to the left (right) of $X_1 = s_1$ is split at $X_2 = s_2$ ($X_1 = s_3$). Finally, the region to the right of $X_1 = s_3$ is split at $X_2 = s_4$. As illustrated in the left plot of Figure 1, the final result is a partitioning into five regions: R_k, $k = 1, \ldots, 5$. In each region R_k, we assume that the model predicts Y with a constant c_k, which is estimated as the sample average of realizations of Y that “fall” within region R_k. A key advantage of this recursive binary partition is that it can be represented as a single tree, as illustrated in the right plot of Figure 1. Each region corresponds to a terminal node of the tree. Given a dependent variable π_{t+h}, a set of predictors $\boldsymbol{x}_t$ and a number of terminal nodes K, the splits are determined in order to minimize the sum of squared errors of the following regression model:

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

where $\mathrm{I}_k(\boldsymbol{x}_t; \boldsymbol{\theta}_k)$ is an indicator function such that

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

$\boldsymbol{\theta}_k$ is the set of parameters that define the *k*-th region. $\mathrm{I}_k(\boldsymbol{x}_t;\boldsymbol{\theta}_k)$ is in fact a product of indicator functions, each of which defines one of the splits that yields the *k*-th region.

RF is a collection of regression trees, each specified in a bootstrap sample of the original data. Since we are dealing with time series, we use a block bootstrap. Suppose there are *B* bootstrap samples. For each sample *b*, $b=1,\ldots,B$, a tree with K_b regions is estimated for a randomly selected subset of the original regressors. K_b is determined in order to leave a minimum number of observations in each region. The final forecast is the average of the forecasts of each tree applied to the original data:

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

4.6 Hybrid Linear-Random Forest Models

RF/OLS and adaLASSO/RF are adaptations specifically designed to disentangle the relative importance of variable selection and nonlinearity to forecast US inflation. RF/OLS is estimated using the following steps:

(1) For each bootstrap sample $b=1,\ldots,B$: (a) Grow a single tree with *k* nodes (we used *k* = 20), and save the $N \leq k$ split variables. (b) Run an OLS on the selected splitting variables. (c) Compute the forecast $\hat{\pi}_{t+h}^{b}$.

(2) The final forecast will be $\hat{\pi}_{t+h} = B^{-1}\sum_{b=1}^{B}\hat{\pi}_{t+h}^{b}$.

The main objective of RF/OLS is to check the performance of a linear model using variables selected by the RF model. If the results are very close to those of the RF model, we understand that nonlinearity is not an issue, and the RF model is superior solely because of variable selection. However, if we see some improvement compared to other linear models, especially bagging,[6] but RF/OLS is still less accurate than RF, we have evidence that both nonlinearity and variable selection play important roles.

The second adapted model is adaLASSO/RF, where we use adaLASSO for variable selection and then estimate a fully grown RF with the variables selected by adaLASSO. If adaLASSO/RF performs similarly to RF, we understand that the variable selection in RF is less relevant and nonlinearity is more important.

adaLASSO/RF and RF/OLS together create an “if and only if” situation where we test the importance of variable selection and nonlinearity from both sides. The results indicate that nonlinearity and variable selection are both important to explain the performance of RF.

4.7 Computer Codes and Tuning Parameters

All ML methods are estimated in `R` using, in most cases, standard and well-established packages. For the linear ML methods, the following packages are used: `HDEconometrics` and `glmnet`. The RF models are estimated using the `randomForest` package. The deep networks and boosted trees used for robustness checks are estimated, respectively, with the `h2o` and `xgboost` packages.[7] The `R` codes are available online at `https://github.com/gabrielrvsc/ForecastInflation`.[8]

For all methods within the LASSO family, the penalty parameter λ is chosen by the BIC as advocated by Kock and Callot (2015) and Medeiros and Mendes (2016). The α parameter for the ElNet penalty in (7) is set to 0.5. We also tried selecting it by using the BIC, but the results are quite similar, and the computational burden is much higher. The weights of the adaptive versions of LASSO and ElNet are chosen as $\omega_i = \dfrac{1}{\left|\tilde{\beta}_i\right| + \frac{1}{\sqrt{T}}}$, where $\tilde{\beta}_i$ is the estimate from the non-adaptive version of the method. The additional term $\frac{1}{\sqrt{T}}$ gives variables excluded by the initial estimator a second chance for inclusion in the model.

The numbers of factors and lags in the factor models are set to four. Data-driven methods for selecting the number of factors and the lags usually yield smaller quantities, delivering in most cases worse forecasts. For this reason, we decided to fix the number of factors and lags at four. For the target factors, we adopt the 5% significance level ($\alpha = 0.05$). For the factor-boosting algorithm, we make the following choices. The maximum number of iterations is set to $M = 10 \times \text{number of variables} \approx 5,000$. However, the boosting algorithm is stopped whenever the BIC of the model starts to increase. The *shrinkage parameter* is set to $\nu = 0.2$ as advocated by Bai and Ng (2009).

The number of bootstrap replications for the bagging method is *B* = 100. We experimented with other values, and the results are very stable. The pre-testing procedure is conducted at the 5% level as in Inoue and Kilian (2008). Given the large number of variables, the pre-testing was performed in two steps. First, for each bootstrap sample, we divide the variables into 10 groups and perform pre-testing on each group. Then, selected variables from each group are used in the next pre-testing, which selects the final variables.

For the CSR, we set $\tilde{n} = 20$ and *q* = 4. These choices are made to avoid a huge computational burden. We tried varying both $\tilde{n}$ and *q*, and the results do not differ much. As in the bagging algorithm and the target factors, the initial pre-testing is carried out at the 5% level.

Each individual tree in the RF model is grown until there are only five observations in each leaf. The proportion of variables randomly selected in each split is 1/3. This is the default setting in the `randomForest` package in `R`. The number of bootstrap samples, *B*, is set to 500. We also experimented with other combinations of parameters, and the final results are reasonably stable.

5 Results

The models are compared according to three different statistics, namely, the root mean squared error (RMSE), the mean absolute error (MAE) and the median absolute deviation from the median (MAD), which are defined as follows:

$$\text{RMSE}_{m,h} = \sqrt{\frac{1}{T-T_0+1}\sum_{t=T_0}^{T}\hat{e}_{t,m,h}^2}, \quad \text{MAE}_{m,h} = \frac{1}{T-T_0+1}\sum_{t=T_0}^{T}\left|\hat{e}_{t,m,h}\right|, \quad \text{and}$$

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

where $\hat{e}_{t,m,h} = \pi_t - \hat{\pi}_{t,m,h}$ and $\hat{\pi}_{t,m,h}$ is the inflation forecast for month *t* made by model *m* with information up to *t* – *h*. The first two measures above are the usual ones in the forecasting literature. MAD, which is less commonly used, is robust to outliers and asymmetries. Reporting both MAE and MAD in addition to RMSE is important for confirming that the results are not due to a few large forecasting errors.

To test whether the forecasts from distinct models are different, we consider a number of tests: the model confidence sets (MCSs) of Hansen et al. (2011), the superior predictive ability (SPA) tests of Hansen (2005), and the multi-horizon SPA test of Quaedvlieg (2017).

5.1 Overview

Table 1 reports a number of summary statistics across all forecasting horizons. The results concern the sample from 1990 to 2015. Additional results for the different subsamples can be found in Tables S.10 and S.11 in the Supplementary Material. Columns (1), (2) and (3) report the average RMSE, the average MAE and the average MAD. Columns (4), (5), and (6) report, respectively, the maximum RMSE, MAE and MAD over the forecasting horizons. Columns (7), (8), and (9) report, respectively, the minimum RMSE, MAE and MAD over the 15 different horizons considered. We normalize these statistics for the benchmark RW model to one. Columns (10), (11) and (12) report the number of times (across horizons) each model achieved the lowest RMSE, MAE, and MAD, respectively. Columns (13) and (14) show the average *p*-values of the SPA test proposed by Hansen (2005). The SPA test of Hansen (2005) compares a collection of models against a benchmark where the null hypothesis is that no other model in the pool of alternatives has superior predictive ability. In the present context, for each forecasting horizon, we run Hansen's SPA test by setting each one of the models as the benchmark. A rejection of the null indicates that the reference model is outperformed by one or more competitors. Columns (15) and (16) present for square and absolute losses, respectively, the average *p*-values for the MCSs based on the t_{max} statistic as described in Hansen et al. (2011). An MCS is a set of models that is built such that it will contain the best model with a given level of confidence. An MCS is analogous to a confidence interval for a parameter. The MCS procedure also yields *p*-values for each of the models considered. The best model has the highest *p*-value. We construct MCSs for each forecasting horizon. In order to build a more general MCS that contemplates all horizons, column (17) displays the *p*-value of the multi-horizon MCS proposed by Quaedvlieg (2017). The test is based on the squared errors only.

The following facts emerge from the tables: (1) ML models and the use of a large set of predictors systematically improve the quality of inflation forecasts over traditional benchmarks. This is a robust and statistically significant result. (2) The RF model outperforms all the other alternatives in terms of point statistics. The superiority of RF is due both to the variable selection mechanism induced by the method as well as the presence of nonlinearities in the relation between inflation and its predictors. RF has the lowest RMSEs, MAEs, and MADs across the horizons and the highest MCS *p*-values. The RF model also has the highest *p*-values in the SPA test and the multi-horizon MCS. The improvements over the RW in terms of RMSE, MAE and MAD are almost 30% and are more pronounced during the second subsample, when inflation volatility is much higher. (3) Shrinkage methods also produce more precise forecasts than the benchmarks. Sparsity-inducing methods are slightly worse than nonsparsity-inducing shrinkage methods. Overall, the forecasting performance among shrinkage methods is very similar, and ranking them is difficult. (4) Factor models are strongly outperformed by other methods. The adoption of boosting and target factors improves the quality of the forecasts only marginally. The poor performance of factor models is more pronounced during the first subsample (low-volatility period). (5) CSR and JMA do not perform well either and are comparable to the factor models. (6) Forecast combination schemes (simple average, trimmed average and median) do not bring any significant improvements in any of the performance criteria considered. (7) Among the benchmark models, both AR and UCSV outperform the RW alternative. Furthermore, the UCSV model is slightly superior to the AR specification.

5.2 Random Forests versus Benchmarks

Tables 2–4 show the results of the comparison between RF and the benchmarks. Table 2 presents the RMSE, MAE and MAD ratios of the AR, UCSV and RF models with respect to the RW alternative for all forecasting horizons as well as for the accumulated forecasts over three, six and twelve months. The models with the smallest ratios are highlighted in bold. It is clear that the RF model has the smallest ratios for all horizons.

To check whether this is a robust finding across the out-of-sample period, we compute rolling RMSEs, MAEs, and MADs over windows of 48 observations. Table 3 shows the frequency with which each model achieved the lowest RMSEs, MAEs and MADs as well as the frequency with which each model was the worst-performing alternative among the four competitors. The RF model is the winning specification and is superior to the competitors for the majority of periods, including the Great Recession. By contrast, the RW model delivers the worst forecasts most of the time. Figures S.4–S.6 in the Supplementary Material show the rolling RMSEs, MAEs, and MADs over the out-of-sample period. The performance of the RW deteriorates as the forecasting horizon increases. However, the accomplishments of the RF seem rather robust.

Table 4 reports the *p*-values of the unconditional Giacomini and White (2000) test for superior predictive ability for squared and absolute errors and the multi-horizon superior predictive ability test of Quaedvlieg (2017). The latter test compares all horizons jointly. Rejections of the null mean that the forecasts are significantly different. It is clear that the RF has forecasts that are significantly different from and superior to the three benchmarks.

5.3 The Full Picture

In this section, we compare all models. Table 5 presents the results for the full out-of-sample period, whereas Tables S.12 and S.13 present the results for the 1990–2000 and 2001–2015 periods, respectively. The tables report the RMSEs and, in parentheses, the MAEs for all models relative to the RW specification. The error measures were calculated from 132 (180) rolling windows covering the 1990–2000 (2001–2015) period. Values in bold denote the most accurate model in each horizon. Cells in gray (blue) show the models included in the 50% MCS using the squared error (absolute error) as the loss function. The MCSs were constructed based on the maximum t-statistic. The last column in the table reports the number of forecast horizons in which the model was included in the MCS for the square (absolute) loss. The last two rows in the table report the number of models included in the MCS for the square and absolute losses.

We start by analyzing the full out-of-sample period. Apart from a few short horizons, where either RF/OLS or adaLASSO/RF is the winning model, RF delivers the smallest ratios in most of the cases. RF is followed closely by shrinkage models, where RR seems be superior to the other alternatives. RR, RF and the hybrid linear-RF models are the only ones included in the MCS for all horizons. Neither RF nor RR impose sparsity, which may corroborate the conclusions of Giannone et al. (2018), who provide evidence against sparsity in several applications. Factor models have very poor results and are almost never included in the MCS. When factors are combined with boosting, there is a small gain, but the results are still greatly inferior to those from the RF and shrinkage models. This is particularly curious as there is a correspondence between the factor models and RR: RR predictions are weighted combinations of all principal component factors of the set of predictors. Several reasons might explain the difference. The first potential explanation is a lack of a clear factor structure in the regressors. This is not the case as shown in Figure S.2 in the Supplementary Material, where we display the eigenvalues of the correlation matrix of regressors over the forecasting period. As shown in the figure, there is a small number of dominating factors. Second, there might be factors that explain only a small portion of the total variance of the regressors but have high predictive power on inflation. Again, we do not think this is the case, as target factors as well as boosting are specifically designed to enhance the quality of the predictions but do not bring any visible improvement in this case. Furthermore, we allow the ML methods to select factors as well, and as shown below, they are never selected. Lastly, we believe the most probable explanation is that although sparsity can be questioned, factor models are an overly aggregated representation of the potential predictors. The results of JMA are not encouraging either. All of the competing models outperform RW for almost all horizons. Finally, the forecast combination does not provide any significant gain due to the empirical fact that most of the forecasts are positively correlated; see Figure S.3.

To check whether this is a robust finding across the out-of-sample period, we compute rolling RMSEs, MAEs, and MADs over windows of 48 observations as shown in Figures S.7–S.18 in the Supplementary Material. The results corroborate the superiority of the RF model, particularly for long as well as aggregated horizons.

Focusing now on the two subsamples, the following conclusions stand out from the tables in the Supplementary Material. The superiority of RF is more pronounced during the 2001–2015 period, when inflation is much more volatile. During this period, RF achieves the smallest RMSE and MAE ratios for almost all horizons. From 1990-2000, the linear shrinkage methods slightly outperform RF for short horizons. However, RF dominates for long horizons and for the twelve-month forecasts. Among the shrinkage models and during the first period, there is no clear evidence of a single winner. Depending on the horizon, different models perform the best. Another salient fact is that there are fewer models included in the MSC during the first subsample.

Finally, we test whether the superiority of the RF model depends on the state of the economy. We consider two cases, namely, recessions versus expansions according to the NBER classification and high versus low macroeconomic uncertainty.[9] We consider the following regressions:

$$\hat{e}^2_{t+h,\mathrm{RF}} - \hat{e}^2_{t+h,\text{other model}} = \alpha_0 \mathrm{I}_{t+h} + \alpha_1 (1 - \mathrm{I}_{t+h}) + \text{error}, \qquad (8)$$

where $\hat{e}^2_{t+h,\mathrm{RF}}$ is the squared forecasting error of the RF for horizon *h*, $\hat{e}^2_{t+h,\text{other model}}$ is the squared forecasting error of the competing model, and I_t is an indicator function that equals one for periods of recession (or high macroeconomic uncertainty). The results are presented in Tables S.14 and S.15 in the Supplementary Material for periods of expansion versus recession and high versus low macroeconomic uncertainty, respectively. The tables report the estimated values of α_0 and α_1 in equation (8) as the respective standard errors. For conciseness, we display only the results for the most relevant models.

Inspecting the tables, it is clear that the majority of the statistics are negative, which indicates that the RF model is superior to its competitors. Of the 72 entries in each table, the values of the statistics are positive only in four and seven cases in Tables S.14 and S.15, respectively. However, the differences are not statistically significant during recessions. This result is not surprising, as only 34 of the 312 out-of-sample observations are labeled as recessions. Nevertheless, the magnitudes of the differences are much higher during recessions. Turning attention to periods of low and high macroeconomic uncertainty, it is evident that the RF model is statistically superior to the benchmarks and the differences are higher in periods of greater uncertainty. The gains from using RF are particularly large during and after the Great Recession (see Figures S.4–S.18). As argued above, both the degrees of slackness and uncertainty might be sources of nonlinearities in the economy. The fact that the RF model outperforms competitors in these states of the economy suggests that allowing for nonlinearities is key to improving macroeconomic forecasts.

5.4 Real-Time Experiment

To test the robustness of our results, we also conduct an experiment with the real-time database provided by FRED-MD. Due to data limitations, we compute the real-time forecasts only for the second subsample considered here. We compare the following models against the three benchmarks: RR, adaLASSO, RF, RF/OLS and adaLASSO/RF. Table 6 reports, for each forecasting horizon, the RMSE, MAE and MAD ratios with respect to the RW model. The following conclusions emerge from the table. The ML methods clearly outperform the three benchmarks. Furthermore, for the three accumulated horizons considered, RF is the best model. For the monthly forecasts, both RF and adaLASSO/RF achieve the best performance for most of the horizons. For one-month-ahead, RR seems to be the best model. The results are robust to the choice between the RMSE and MAE criteria. For MAD, RR is also competitive. These results corroborate our conclusions that ML methods should be considered seriously to forecast inflation.

5.5 Opening the Black Box: Variable Selection

We compare the predictors selected by some of the ML methods: adaLASSO, RR and RF. We select these three models for two reasons. First, they are generally the three best-performing models. Second, they have quite different characteristics. While adaLASSO is a true sparsity-inducing method, RR and RF models are only approximately sparse. In addition, RR is a linear model, whereas RF is a highly nonlinear specification.

This analysis is straightforward for sparsity-inducing shrinkage methods such as adaLASSO, as the coefficients of potentially irrelevant variables are automatically set to zero.[10] For the other ML methods, the analysis is more complex. To ensure that the results among models are comparable, we adopt the following strategy. For RR and adaLASSO, the relative importance measure is computed as the average coefficient size (multiplied by the respective standard deviations of the regressors). To measure the importance of each variable for the RF models, we use out-of-bag (OOB) samples.[11] When the *b*-th tree is grown, the OOB samples are passed down the tree, and the prediction accuracy is recorded. Then, the values of the *j*-th variable are randomly permuted in the OOB sample, and the accuracy is again computed. The decrease in accuracy due to the permutation is averaged over all trees and is the measure of the importance of the *j*-th variable in RF.

Due to space constraints, we cannot show the relative importance for each variable, each lag, each horizon or each estimation window. Therefore, as described in the Supplementary Material and following McCracken and Ng (2016), we categorize the variables, including lags, into the following eight groups: (i) output and income; (ii) labor market; (iii) housing; (iv) consumption, orders and inventories; (v) money and credit; (vi) interest and exchange rates; (vii) prices; and (viii) stock market. We also consider two additional groups, namely, the principal component factors computed from the full set of potential predictors and autoregressive terms. Furthermore, the results are averaged across all estimation windows.

Figure 2 shows the importance of each variable group for the adaLASSO, RR and RF methods for each horizon. The values in the plots are re-scaled to sum to one.

The set of the most relevant variables for the RR and RF models is quite stable across forecasting horizons but is remarkably different between them. While for RR, AR terms, prices and employment are the most important predictors, RF models give more importance to prices, interest-exchange rates, employment and housing. For adaLASSO, selection is quite different across forecasting horizons, and only AR terms retain their relative importance independent of the horizon. Other prices gradually lose their relevance until up to six months ahead and partially recover relevance when longer horizons are considered. Output-income variables are more important for medium-term forecasts. Finally, none of the three classes of models selects either factors or stocks. This result may indicate that the high level of cross-section aggregation of the factors is responsible for the poor performance of factor models.

To compare the degree of sparsity between adaLASSO and RF, we report word clouds of the selected variables in the Supplementary Material. A word cloud is an image composed of the names of variables selected by a specific model across the estimation windows; in the word cloud, the size of each word indicates its frequency or importance. The names displayed in the clouds are as defined in the third columns of Tables S.1–S.8. It is evident that the RF models are much less sparse than adaLASSO.

RR, adaLASSO and RF select different variables, which suggests non-trivial interactions among variable selection, sparsity and nonlinearity. If the econometrician is only interested in forecasting, variable selection is less of a concern. One should select (a combination of) the best-performing models. If she/he is also interested in the precise mechanisms underlying price dynamics, careful identification schemes should be considered, which is beyond the scope of this paper. Nonetheless, compared to adaLASSO and RR, the best-performing method, RF, uses disaggregated inflation as a substitute for lags of CPI inflation, thus resembling an unusual Phillips curve with heterogeneous backward-looking terms. This finding may shed light on price dynamics and can be useful for future research aimed at uncovering such mechanisms.

5.6 Opening the Black Box: Nonlinearities

The role of nonlinearity in the relative performance of the RF model is highlighted in the plots of recursive RMSEs, MAEs, and MADs in Figures S.7–S.18. First, RF is compared with the adaLASSO/RF and RF/OLS models in Figures S.16–S.18. For $h = 1$, the performances of the three models are almost identical, indicating that there are no benefits in introducing nonlinearity for very short-term forecasts. On the other hand, the superiority of the RF model becomes more evident for longer horizons and for the accumulated inflation over six and twelve months. Furthermore, the results do not seem to be due to a few outliers, as both the rolling MAEs and MADs, which are less sensitive to extreme observations, confirm the overperformance of the RF model. These findings are also corroborated when RF is compared with the other linear models; see Figures S.7–S.15.

Although is clear that nonlinearity is present in the dynamics of inflation, it is very difficult to completely uncover the nonlinear mechanism driving the forecasts, as in our experiment we estimated 12 different RF models (one for each horizon) for each rolling window. As we have 312 windows, there are total of 3,744 different models to analyze.

5.7 Robustness: Alternative Nonlinear Models and Other Penalties

As a robustness check, we compute forecasts from other nonlinear alternatives. The first is a boosted regression tree (BTree) as in Friedman (2001). The second one is based on the results of Gu et al. (2018) and is a Deep Neural Network (DeepNN) with three hidden layers and 32, 16, and 8 rectified linear units (ReLU) in each layer, respectively. The DeepNN model is estimated by the stochastic gradient descent algorithm. We also consider two second-order polynomial models with interaction terms estimated either by LASSO or adaLASSO. Finally, in addition to these nonlinear alternatives, we estimate a linear model with a SCAD penalty.

Table 7 reports, for each forecasting horizon, the RMSE, the MAE and the MAD ratios with respect to the RW model for the full out-of-sample period (1990–2015). In terms of RMSE, RF is the best model in 11 of 15 cases. Among these, in five cases there is a tie with the polynomial model estimated with adaLASSO. In general, the DeepNN model has the worst performance. SCAD is also outperformed by the RF model. The results for MAE are quite similar. On the other hand, there are some differences when MAD is considered. Although the RF model is still the best alternative, the polynomial model estimated with LASSO overperforms the one estimated with adaLASSO.

These results corroborates the superiority of RF. Furthermore, the fact that the polynomial models with interactive terms are competitive sheds some light on the nature of the nonlinearity captured by the tree-based models. We should bear in mind that regression trees are models that explore interactions among covariates.

6 Conclusions

We show that with the recent advances in ML methods and the availability of new and rich datasets, it is possible to improve inflation forecasts. Models such as LASSO, RF and others are able to produce more accurate forecasts than the standard benchmarks. These results highlight the benefits of ML methods and big data for macroeconomic forecasting. Although our paper focuses on inflation forecasting in the US, one can easily apply ML methods to forecast other macroeconomic series in a variety of countries.

The RF model deserves special attention, as it robustly delivers the smallest errors. Its good performance is due to both potential nonlinearities and its variable selection mechanism.

The selection of variables for RF models is stable across horizons. These variables are mostly selected from the following groups: prices, exchange and interest rates, and the housing and labor markets. Although it is difficult to disentangle the precise sources of the nonlinearities that the RF model uncovers, the variable selection may shed light on them. In fact, there are many theoretical arguments justifying nonlinear relationships among inflation, interest rate, labor market outcomes and housing. For example, the relationship between inflation and employment depends on the degree of slackness in the economy. Uncertainty might also induce nonlinearities. Finally, part of the out-of-sample window encompasses quarters when the zero lower bound on nominal interest rates is binding, which is another source of nonlinearity. This out-of-sample window also encompasses a period in which a housing bubble led to a credit crunch, events with highly nonlinear consequences.

RF is also the winning method in the periods of expansion and recession as well as in the periods of low uncertainty and high uncertainty. The gains from using RF are larger in periods of recession and high uncertainty. RF also outperforms other methods during and after the Great Recession, when uncertainty skyrocketed and when the zero lower bound was binding. Taken together, these results suggest that the relationships among key macroeconomic variables can be highly nonlinear. If this is the case, the linear methods applied in the profession not only to forecast variables but also to achieve other objectives such as approximate DSGE models might lead to inaccurate results, especially during bad states of the economy.

References

Atkeson, A. and Ohanian, L. (2001), ‘Are phillips curves useful for forecasting inflation?’, *Federal Reserve bank of Minneapolis Quarterly Review* **25**, 2–11.

Bai, J. and Ng, S. (2003), ‘Inferential theory for factor models of large dimensions’, *Econometrica* **71**, 135–171.

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

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

Bańbura, M., Giannone, D. and Reichlin, L. (2010), ‘Large Bayesian vector autoregressions’, *Journal of Applied Econometrics* **25**, 71–92.

Bloom, N. (2009), ‘The impact of uncertainty shocks’, *Econometrica* **77**(3), 623–685.

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

Breiman, L. (2001), ‘Random forests’, *Machine Learning* **45**, 5–32.

Chakraborty, C. and Joseph, A. (2017), Machine learning at central banks, Working Paper 674, Bank of England Staff Working Paper.

Chen, Y.-C., Turnovsky, S. and Zivot, E. (2014), ‘Forecasting inflation using commodity price aggregates’, *Journal of Econometrics* **183**, 117–134.

Dellas, H., Gibson, H., Hall, S. and Tavlas, G. (2018), ‘The macroeconomic and fiscal implications of inflation forecast errors’, *Journal of Economic Dynamics and Control* **93**, 203 – 217.

Eggertsson, G. B. and Woodford, M. (2003), ‘Zero bound on interest rates and optimal monetary policy’, *Brookings Papers on Economic Activity* (1), 139–233.

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

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

Fan, J. and Li, R. (2001), ‘Variable selection via nonconcave penalized likelihood and its oracle properties’, *Journal of the american Statistical Association* **96**, 1348–1360.

Faust, J. and Wright, J. (2013), Forecasting inflation, *in* G. Elliott and A. Timmermann, eds, ‘Handbook of Economic Forecasting’, Vol. 2A, Elsevier.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2003), ‘Do financial variables help forecasting inflation and real activity in the euro area?’, *Journal of Monetary Economics* **50**, 1243–1255.

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

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

Giacomini, R. and White, H. (2006), ‘Tests of conditional predictive ability’, *Econometrica* **74**, 1545–1578.

Giannone, D., Lenza, M. and Primiceri, G. (2018), Economic predictions with big data: The illusion of sparsity, Working paper, Northwestern University.

Groen, J., Paap, R. and Ravazzolo, F. (2013), ‘Real-time inflation forecasting in a changing world’, *Journal of Business and Economic Statistics* **31**, 29–44.

Gu, S., Kelly, B. and Xiu, D. (2018), Empirical asset pricing with machine learning, Working paper, University of Chicago.

Hall, A. S. (2018), ‘Machine learning approaches to macroeconomic forecasting’, *The Federal Reserve Bank of Kansas City Economic Review*.

Hansen, B. E. and Racine, J. S. (2012), ‘Jackknife model averaging’, *Journal of Econometrics* **167**(1), 38–46.

Hansen, P. (2005), ‘A test for superior predictive ability’, *Journal of Business and Economic Statistics* **23**, 365–380.

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

Hastie, T., Tibshirami, R. and Friedman, J. (2001), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer.

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

Iacoviello, M. (2005), ‘House prices, borrowing constraints, and monetary policy in the business cycle’, *American Economic Review* **95**(3), 739–764.

Inoue, A. and Kilian, L. (2008), ‘How useful is bagging in forecasting economic time series? a case study of U.S. CPI inflation’, *Journal of the American Statistical Association* **103**, 511–522.

Jurado, K., Ludvigson, S. and Ng, S. (2015), ‘Measuring uncertainty’, *American Economic Review* **105**, 1177–1215.

Kock, A. and Callot, L. (2015), ‘Oracle inequalities for high dimensional vector autoregressions’, *Journal of Econometrics* **186**, 325–344.

Lucas, R. (1987), *Models of business cycles*, Blackwell, Oxford.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2018), ‘Statistical and machine learning forecasting methods: Concerns and ways forward’, *Plos One* **13**, e0194889.

McCracken, M. and Ng, S. (2016), ‘FRED-MD: A monthly database for macroeconomic research’, *Journal of Business and Economic Statistics* **34**, 574–589.

Medeiros, M. and Mendes, E. (2016), ‘ℓ_1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors’, *Journal of Econometrics* **191**, 255–271.

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

Mian, A. and Sufi, A. (2009), ‘The consequences of mortgage credit expansion: Evidence from the u.s. mortgage default crisis’, *Quarterly Journal of Economics* **124**(4), 1449–1496.

Mullainathan, S. and Spiess, J. (2017), ‘Machine learning: An applied econometric approach’, *Journal of Economic Perspectives* **31**, 87–106.

Nakamura, E. (2005), ‘Inflation forecasting using a neural network’, *Economics Letters* **86**, 373–378.

Quaedvlieg, R. (2017), Multi-horizon forecast comparison, Working paper, Erasmus School of Economics.

Scornet, E., Biau, G. and Vert, J.-P. (2015), ‘Consistency of random forests’, *Annals of Statistics* **43**, 1716–1741.

Shiller, R. J. (2014), ‘Speculative asset prices’, *American Economic Review* **104**(6), 1486–1517.

Stock, J. H. and Watson, M. W. (2010), Modeling inflation after the crisis, Technical report, National Bureau of Economic Research.

Stock, J. and Watson, M. (1999), 'Forecasting inflation', *Journal of Monetary Economics* **44**, 293–335.

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

Stock, J. and Watson, M. (2007), 'Why has US inflation become harder to forecast?', *Journal of Money, Credit and Banking* **39**, 3–33.

Svensson, L. E. O. and Woodford, M. (2004), Implementing optimal policy through inflation-forecast targeting, *in* B. S. Bernanke and M. Woodford, eds, ' The Inflation-Targeting Debate', University of Chicago Press, pp. 19–92.

Teräsvirta, T., van Dijk, D. and Medeiros, M. (2005), 'Linear models, smooth transition autoregressions and neural networks for forecasting macroeconomic time series: A reexamination (with discussion)', *International Journal of Forecasting* **21**, 755–774.

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

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

Zhang, X., Wan, A. T. and Zou, G. (2013), 'Model averaging by jackknife criterion in models with dependent data', *Journal of Econometrics* **174**(2), 82–94.

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

Zou, H. and 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.

Notes

[1] See Mullainathan and Spiess (2017) for discussions of ML methods and big data in economics. In this paper, an ML model is any statistical model that is able to either handle a large set of covariates and/or describe nonlinear mappings nonparametrically. Some of these methods even predate “ machines”.

[2] To gauge the relevance of such gains, consider the rough welfare calculations in Dellas et al. (2018). The authors use a textbook macro model with nominal rigidities, in which inflation forecasting errors lead to inefficient output gap volatility. If relative risk aversion is equal to two, agents are willing to forgo 0.16% (0.34%) of their steady state consumption to avoid a forecasting deterioration of 20% (50%) in terms of mean squared errors. These figures are relatively large. As a ground for comparison, Lucas (1987), who originally proposed this method to measure the welfare cost of business cycles, found it to be 0.10%.

[3] The roles of ML methods and big data were discussed at the Norges Bank’s workshop on “big data, machine learning and the macroeconomy” in 2017, as well as at the ECB workshop on “economic forecasting with large datasets” in 2018. Banca D’Italia and Deutsche Bundesbank promoted similar workshops in 2018 and 2019, respectively. Finally, staff of central banks have just started to produce working papers with applications that combine big data, ML and forecasting, e.g. Chakraborty and Joseph (2017) from Bank of England and Hall (2018) from Kansas City Fed.

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

[5] We do not use a fixed set of controls, $\boldsymbol{w}_t$, in the pretesting procedure like we did for the target factors.

[6] Bagging and RF are bootstrap-based models; the former is linear, and the latter is nonlinear.

[7] `HDEconometrics` is available at `https://github.com/gabrielrvsc/HDeconometrics/`. The remaining packages are available at `https://cran.r-project.org/web/packages/`.

[8] See also `http://www.econ.puc-rio.br/mcm/codes/`

[9] Periods of high (low) macroeconomic uncertainty are those where uncertainty is higher (lower) than the historical average. Since the results barely change if we consider either financial or real, rather than macroeconomic, uncertainty, we do not report them for brevity. They are available upon request. These measures of macroeconomic, financial and real uncertainty are computed as in Jurado et al. (2015) and are the conditional volatility of the unforecastable part of macroeconomic, financial and firm-level variables, respectively. They are available at Sydney C. Ludvigson's webpage (https://www.sydneyludvigson.com/).

[10] Medeiros and Mendes (2016) showed, for example, that under sparsity conditions, the adaLASSO model selection is consistent for high-dimensional time series models in very general settings, i.e., the method correctly selects the "true" set of regressors.

[11] For a given data point $(y_t, \boldsymbol{x}'_t)$, the OOB sample is the collection of all bootstrap samples that do not include $(y_t, \boldsymbol{x}'_t)$.

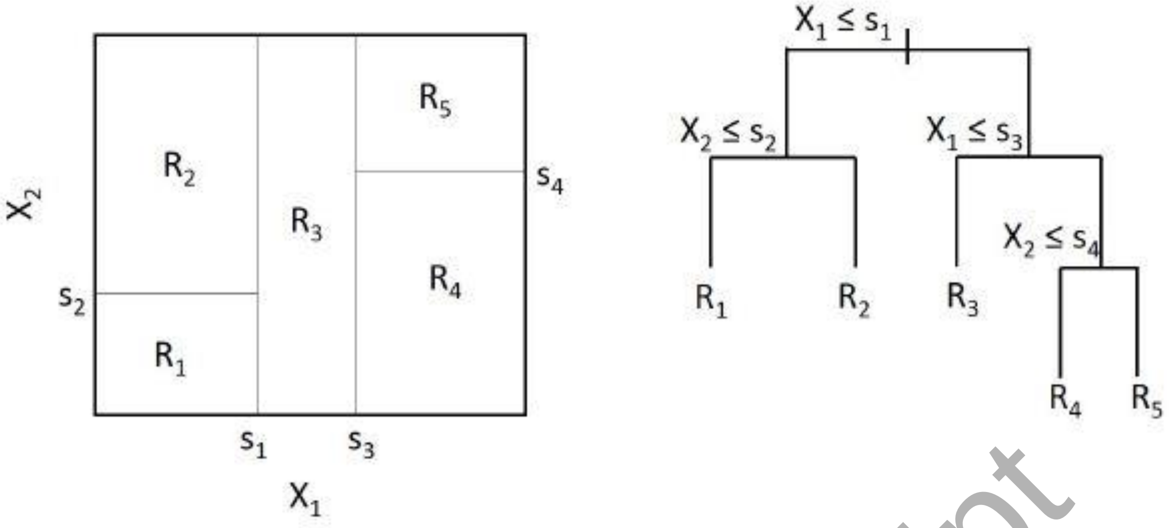

Fig. 1 Example of a regression tree. Reproduction of part of Figure 9.2 in Hastie et al. (2001).

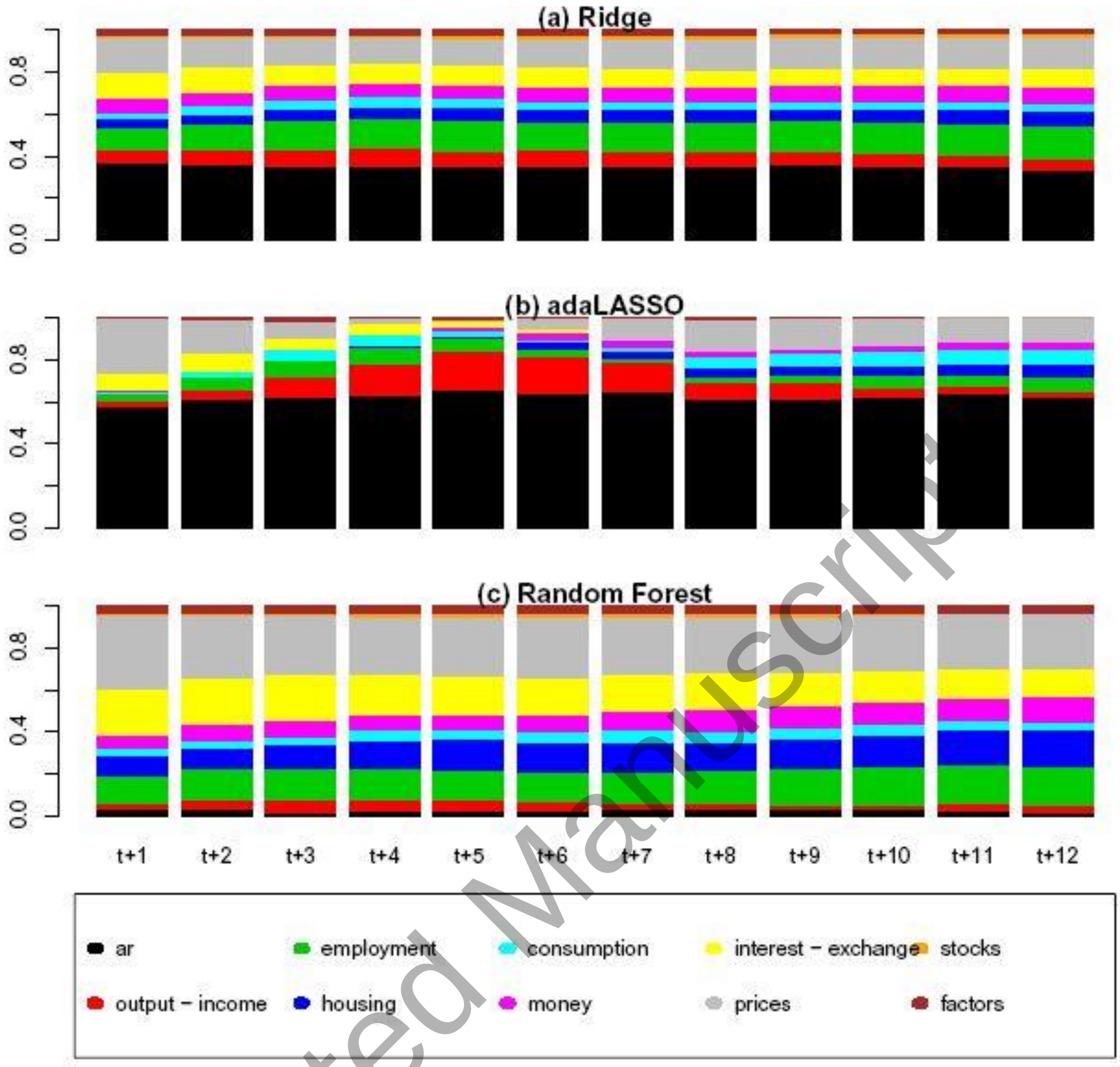

Fig. 2 Variable importance

The picture shows the importance of each variable group for the adaLASSO, RR and RF methods for all the twelve forecasting horizons. For all different methods, the values in the plots are re-scaled to sum one. For RR and adaLASSO, the relative importance measure is computed as the average coefficient size (multiplied by the respective standard deviations of the regressors). To measure the importance of each variable for the RF models, we use out-of-bag (OOB) samples. When the bth tree is grown, the OOB samples are passed down the tree and the prediction accuracy is recorded.

Then, the values of the jth variable are randomly permuted in the OOB sample, and the accuracy is again computed. The decrease in accuracy due to the permutation is averaged over all trees and is the measure of the importance of the jth variable in the RF.

Table 1 Forecasting Results: Summary statistics for the out-of-sample period from 1990–2015

The table reports for each model a number of different summary statistics across all the forecasting horizons, including the accumulated three-, six-, and twelve-month horizons. Columns (1), (2) and (3) report the average root mean square error (RMSE), the average mean absolute error (MAE) and the average median absolute deviation (MAD). Columns (4), (5), and (6) report, respectively, the maximum RMSE, MAE and MAD over the forecasting horizons. Columns (7), (8), and (9) report, respectively, the minimum RMSE, MAE and MAD over the 15 different horizons considered. Columns (10), (11) and (12) report the number of times (across horizons) each model achieved the lowest RMSE, MAE, and MAD, respectively. Columns (13) and (14) show the average *p* values of the superior predictive ability (SPA) test proposed by Hansen (2005). Columns (15) and (16) present for square and absolute losses, the average *p*-values for the Model Confidence Sets (MCS) based on the t_{max} statistic as described in Hansen et al. (2011). Column (17) displays the *p*-value of the multi-horizon MCS proposed by Quaedvlieg (2017). The test is based on the squared errors only.

	Forecasting Precision												Sup. Pred. Ability		Model Confidence Set		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
	ave.	ave.	ave.	max.	max.	max.	min.	min.	min.	# min.	# min.	# min.	ave. p.v.	ave. p.v.	ave. p.v.	ave. p.v.	p.v.

	Forecasting Precision												Sup. Pred. Ability		Model Confidence Set		
Model	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	RMSE	MAE	MAD	sq	abs	Tmax sq	Tmax abs	m.h. 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	0.00	0.13	0.07	0.00
AR	0.84	0.86	0.78	1.22	1.22	1.16	0.75	0.76	0.60	0	0	0	0.05	0.01	0.48	0.21	0.00
UCSV	0.82	0.82	0.85	0.95	0.91	1.04	0.77	0.78	0.76	0	0	0	0.07	0.04	0.52	0.37	0.00
LASSO	0.78	0.79	0.74	0.98	1.04	0.91	0.73	0.71	0.61	0	0	0	0.14	0.17	0.72	0.56	0.73
adaLASSO	0.78	0.78	0.76	0.96	0.96	0.99	0.72	0.71	0.63	0	0	0	0.15	0.31	0.68	0.66	0.18
ElNet	0.78	0.80	0.73	0.98	1.05	0.89	0.73	0.71	0.61	0	0	2	0.15	0.15	0.72	0.55	0.89
adaElnet	0.78	0.78	0.76	0.96	0.97	0.96	0.73	0.71	0.61	0	0	2	0.18	0.33	0.70	0.67	0.05
RR	0.76	0.77	0.79	0.89	0.93	1.03	0.70	0.71	0.67	0	0	0	0.43	0.46	0.77	0.69	0.40
BVAR	0.80	0.82	0.80	1.07	1.09	1.14	0.74	0.73	0.64	0	0	0	0.12	0.13	0.66	0.51	0.04
Bagging	0.79	0.83	0.90	0.83	0.89	1.29	0.74	0.78	0.76	0	0	0	0.20	0.06	0.63	0.39	0.04
CSR	0.82	0.82	0.80	1.13	1.11	1.09	0.76	0.74	0.67	0	0	0	0.13	0.06	0.58	0.37	0.00
JMA	0.86	0.91	1.00	0.99	0.99	1.39	0.76	0.83	0.78	0	0	0	0.06	0.01	0.29	0.08	0.00

	Forecasting Precision												Sup. Pred. Ability		Model Confidence Set		
Factor	0.84	0.87	0.87	1.17	1.21	1.25	0.78	0.78	0.71	0	0	0	0.04	0.01	0.34	0.13	0.00
T. Factor	0.83	0.88	0.89	1.17	1.23	1.26	0.77	0.80	0.70	0	0	0	0.01	0.00	0.30	0.09	0.00
B. Factor	0.83	0.90	0.99	1.17	1.32	1.60	0.74	0.75	0.73	0	0	0	0.02	0.00	0.58	0.25	0.00
RF	0.73	0.73	0.70	0.84	0.81	0.84	0.68	0.67	0.58	11	13	8	0.94	0.95	0.95	0.97	1
Mean	0.77	0.77	0.77	0.95	0.97	1.01	0.71	0.70	0.66	0	0	0	0.37	0.40	0.75	0.64	0.89
T.Mean	0.77	0.77	0.75	0.95	0.96	0.95	0.71	0.70	0.62	0	0	0	0.35	0.48	0.74	0.71	0.83
Median	0.77	0.77	0.75	0.94	0.97	0.99	0.71	0.70	0.63	0	0	0	0.32	0.44	0.73	0.70	0.89
RF/OLS	0.77	0.78	0.81	0.94	0.97	1.05	0.71	0.72	0.63	1	1	0	0.46	0.47	0.76	0.70	0.96
adaLASSO/RF	0.75	0.75	0.73	0.85	0.82	0.87	0.70	0.68	0.58	3	1	3	0.53	0.59	0.82	0.79	0.89

Table 2 Forecasting Results: RMSE, MAE and MAD Ratios (1990–2015)

The table reports, for each forecasting horizon, the root mean squared error (RMSE), mean absolute error (MAE) and median absolute deviation from the median (MAD) ratios with respect to the random walk model for the full out-of-sample period (1990–2015). The last three columns represent, respectively, the ratios for the accumulated three, six, and twelve-month forecasts. The statistics for the best-performing model are highlighted in bold.

	Panel (a): RMSE Ratio														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
AR	0.90	0.81	0.79	0.81	0.79	0.79	0.78	0.76	0.78	0.82	0.84	0.75	0.86	0.97	1.22
UCSV	0.95	0.82	0.80	0.81	0.78	0.78	0.78	0.78	0.77	0.80	0.83	0.78	0.87	0.86	0.91
RF	**0.84**	**0.73**	**0.71**	**0.74**	**0.71**	**0.72**	**0.72**	**0.71**	**0.72**	**0.76**	**0.77**	**0.68**	**0.71**	**0.71**	**0.77**
	Panel (b): MAE Ratio														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
AR	0.87	0.79	0.78	0.81	0.80	0.81	0.78	0.76	0.81	0.85	0.86	0.76	0.89	1.04	1.22
UCSV	0.91	0.82	0.79	0.80	0.80	0.79	0.80	0.79	0.78	0.80	0.85	0.78	0.86	0.90	0.89
RF	**0.81**	**0.72**	**0.71**	**0.75**	**0.73**	**0.73**	**0.70**	**0.68**	**0.72**	**0.75**	**0.77**	**0.67**	**0.74**	**0.77**	**0.77**

	Panel (a): RMSE Ratio														
	Panel (c): MAD Ratio														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
AR	0.74	0.70	0.82	0.82	0.83	0.75	0.66	0.68	0.77	0.70	0.77	0.60	0.87	1.16	0.89
UCSV	0.88	0.77	0.83	0.91	0.88	0.79	0.76	0.83	0.86	0.83	0.88	0.78	0.87	1.04	0.88
RF	**0.70**	**0.63**	**0.77**	**0.84**	**0.75**	**0.73**	**0.65**	**0.64**	**0.73**	**0.69**	**0.71**	**0.58**	**0.71**	**0.80**	**0.59**

Table 3 Forecasting Results: Ranking of Models (1990–2015)

The table reports the frequency with which each model achieved the best (worst) performance statistics over a rolling window period of four years (48 observations). The last three columns represent, respectively, the ratios for the accumulated three, six, and twelve-month forecasts. The statistics for the model with the highest figures are highlighted in bold.

	Panel (a): Lowest Rolling RMSE														
	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	0.04	0.00	0.00	0.06	0.08
AR	0.08	0.05	0.00	0.16	0.01	0.01	0.10	0.18	0.12	0.13	0.11	0.00	0.00	0.00	0.00
UCSV	0.02	0.05	0.21	0.10	0.18	0.11	0.00	0.03	0.19	0.11	0.09	0.00	0.21	0.31	0.24
RF	**0.89**	**0.90**	**0.79**	**0.74**	**0.81**	**0.88**	**0.90**	**0.79**	**0.69**	**0.76**	**0.75**	**1.00**	**0.78**	**0.63**	**0.68**
	Panel (b): Lowest Rolling MAE														
	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	0.09	0.00	0.03	0.11	0.04
AR	0.17	0.03	0.00	0.05	0.00	0.02	0.10	0.14	0.13	0.06	0.05	0.02	0.00	0.00	0.00
UCSV	0.15	0.18	0.26	0.23	0.15	0.15	0.00	0.07	0.24	0.23	0.09	0.02	0.21	0.24	0.20

	Panel (a): Lowest Rolling RMSE														
RF	**0.68**	**0.79**	**0.74**	**0.72**	**0.85**	**0.82**	**0.90**	**0.79**	**0.63**	**0.71**	**0.76**	**0.95**	**0.76**	**0.65**	**0.76**
	Panel (c): Lowest Rolling MAD														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	0.14	0.00	0.03	0.00	0.09	0.05	0.00	0.00	0.03	0.05	0.03	0.05	0.02	0.15	0.02
AR	0.23	0.16	0.23	0.23	0.12	0.23	0.26	0.32	0.11	0.15	0.42	0.23	0.04	0.03	0.04
UCSV	0.04	0.19	0.27	0.33	0.09	0.12	0.03	0.01	0.22	0.09	0.03	0.02	0.10	0.05	0.05
RF	**0.59**	**0.65**	**0.46**	**0.44**	**0.69**	**0.60**	**0.70**	**0.67**	**0.64**	**0.71**	**0.52**	**0.70**	**0.84**	**0.77**	**0.89**
	Panel (d): Highest Rolling RMSE														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	**0.82**	**0.98**	**0.94**	**0.99**	**0.98**	**1.00**	**1.00**	**1.00**	**0.86**	**0.80**	**0.71**	**0.85**	**0.61**	0.26	0.00
AR	0.00	0.00	0.06	0.01	0.02	0.00	0.00	0.00	0.14	0.19	0.29	0.15	0.38	**0.74**	**0.97**
UCSV	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.03
RF	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00

	Panel (a): Lowest Rolling RMSE														
	Panel (e): Highest Rolling MAE														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	**0.90**	**0.94**	**0.86**	**1.00**	**0.99**	**0.97**	**1.00**	**0.94**	**0.82**	**0.73**	**0.69**	**0.77**	**0.63**	0.26	0.03
AR	0.08	0.03	0.13	0.00	0.01	0.03	0.00	0.06	0.18	0.27	0.28	0.23	0.36	**0.74**	**0.86**
UCSV	0.02	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	0.11
RF	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
	Panel (f): Highest Rolling MAD														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RW	**0.68**	**0.94**	**0.85**	**0.80**	**0.80**	**0.66**	**0.81**	**0.94**	**0.77**	**0.80**	**0.79**	**0.92**	**0.68**	0.30	0.31
AR	0.05	0.03	0.12	0.10	0.10	0.07	0.04	0.02	0.10	0.03	0.02	0.01	0.18	**0.57**	**0.51**
UCSV	0.22	0.03	0.03	0.10	0.04	0.26	0.14	0.04	0.11	0.14	0.19	0.07	0.12	0.13	0.17
RF	0.05	0.00	0.00	0.00	0.06	0.02	0.01	0.00	0.02	0.02	0.00	0.00	0.02	0.01	0.00

Table 4 Forecasting Results: Superior Predictive Ability Test (1990–2015)

The table reports the *p*-values of the unconditional Giacomini-White test for superior predictive ability between the random forest models and each of the benchmark models for each forecasting horizon as well as for the three accumulated horizons. The test is based on the full out-of-sample period. Panel (a) presents the results for squared errors, while panel (b) shows the results for absolute errors. The GW statistics are computed with Heteroskedastic-Autocorrelation (HAC) robust variances with the quadratic spectral kernel and bandwidth selected by Andrew’s automatic method. The table also reports the *p*-values of the uniform and average multi-horizon superior predictive ability test proposed by Quaedvlieg (2017).

	Panel (a): Squared Errors															
	Giacomini-White test – Forecasting Horizon														Quaedvlieg test	
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m	Unif.	Avg.
RW	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.03	0.02	0.00	0.00	0.05	0.05	0.00	0.00
AR	0.00	0.01	0.02	0.04	0.02	0.02	0.06	0.08	0.05	0.06	0.01	0.00	0.01	0.02	0.02	0.00	0.00
UCSV	0.00	0.00	0.01	0.06	0.06	0.02	0.00	0.00	0.00	0.04	0.00	0.00	0.04	0.06	0.07	0.00	0.00
	Panel (b): Absolute Errors																

	Panel (a): Squared Errors																
	Giacomini-White test – Forecasting Horizon															Quaedvlieg test	
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m	Unif.	Avg.
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	0.05	0.05	0.00	0.00
AR	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.01	0.02	0.02	0.00	0.00
UCSV	0.00	0.00	0.01	0.03	0.00	0.01	0.00	0.00	0.01	0.05	0.01	0.00	0.04	0.06	0.07	0.00	0.00

Table 5 Forecasting Errors for the CPI from 1990 to 2015

The table shows the root mean squared error (RMSE) and, between parenthesis, the mean absolute errors (MAE) for all models relative to the random walk (RW). The error measures were calculated from 132 rolling windows covering the 1990-2000 period and 180 rolling windows covering the 2001-2015 period. Values in bold show the most accurate model in each horizon. Cells in gray (blue) show the models included in the 50% model confidence set (MCS) using the squared error (absolute error) as loss function. The MCSs were constructed based on the maximum *t*-statistic. The last column in the table reports in how many horizons the row model was included in the MCS for square (absolute) loss. The last two rows in the table report how many models were included in the MCS for square and absolute losses.

Consumer Price Index 1990–2015																
Forecasting Horizon																
RMSE/(MAE)	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m	RMSE count (MAE count)
RW	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	1
	(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)	(0)
AR	0.90	0.81	0.79	0.81	0.79	0.79	0.78	0.76	0.78	0.82	0.84	0.75	0.86	0.97	1.22	8
	(0.87)	(0.79)	(0.78)	(0.81)	(0.80)	(0.81)	(0.78)	(0.76)	(0.81)	(0.85)	(0.86)	(0.76)	(0.89)	(1.04)	(1.22)	(0)
UCSV	0.95	0.82	0.80	0.81	0.78	0.78	0.78	0.78	0.77	0.80	0.83	0.78	0.87	0.86	0.91	9

Consumer Price Index 1990–2015																
	(0.91)	(0.82)	(0.79)	(0.80)	(0.80)	(0.79)	(0.80)	(0.79)	(0.78)	(0.80)	(0.85)	(0.78)	(0.86)	(0.90)	(0.89)	(4)
LASSO	0.83	0.75	0.73	0.76	0.74	0.75	0.75	0.73	0.75	0.80	0.82	0.73	0.75	0.79	0.98	13
	(0.82)	(0.74)	(0.73)	(0.78)	(0.77)	(0.75)	(0.74)	(0.71)	(0.76)	(0.81)	(0.84)	(0.74)	(0.79)	(0.89)	(1.04)	(11)
adaLASSO	0.84	0.76	0.74	0.77	0.75	0.75	0.76	0.75	0.76	0.80	0.83	0.72	0.76	0.80	0.96	13
	(0.81)	(0.75)	(0.72)	(0.77)	(0.75)	(0.74)	(0.73)	(0.71)	(0.75)	(0.79)	(0.84)	(0.73)	(0.79)	(0.86)	(0.96)	(13)
ElNet	0.83	0.75	0.73	0.76	0.75	0.74	0.75	0.74	0.76	0.81	0.82	0.73	0.75	0.79	0.98	13
	(0.82)	(0.74)	(0.73)	(0.78)	(0.78)	(0.76)	(0.75)	(0.71)	(0.77)	(0.81)	(0.85)	(0.75)	(0.78)	(0.89)	(1.05)	(11)
adaElnet	0.84	0.75	0.73	0.77	0.75	0.75	0.75	0.74	0.76	0.80	0.81	0.73	0.76	0.81	0.96	13
	(0.82)	(0.74)	(0.72)	(0.76)	(0.75)	(0.74)	(0.73)	(0.71)	(0.75)	(0.79)	(0.83)	(0.75)	(0.79)	(0.86)	(0.97)	(13)
RR	0.85	0.73	0.72	0.75	0.74	0.75	0.75	0.73	0.74	0.77	0.78	0.70	0.73	0.77	0.89	14
	(0.83)	(0.72)	(0.72)	(0.77)	(0.76)	(0.76)	(0.73)	(0.71)	(0.74)	(0.77)	(0.79)	(0.71)	(0.77)	(0.86)	(0.93)	(14)
BVAR	0.86	0.76	0.75	0.77	0.74	0.76	0.77	0.76	0.77	0.82	0.83	0.74	0.79	0.85	1.07	12
	(0.87)	(0.73)	(0.75)	(0.79)	(0.78)	(0.78)	(0.76)	(0.76)	(0.81)	(0.83)	(0.85)	(0.76)	(0.82)	(0.93)	(1.09)	(9)

Consumer Price Index 1990–2015															
Bagging	0.83	0.76	0.76	0.80	0.78	0.79	0.83	0.81	0.78	0.82	0.83	0.74	0.74	0.76	0.82	12
	(0.84)	(0.78)	(0.79)	(0.87)	(0.86)	(0.85)	(0.83)	(0.80)	(0.80)	(0.84)	(0.86)	(0.78)	(0.79)	(0.89)	(0.88)	(7)
CSR	0.85	0.77	0.76	0.79	0.77	0.79	0.79	0.77	0.79	0.83	0.84	0.76	0.79	0.87	1.13	10
	(0.84)	(0.76)	(0.75)	(0.79)	(0.79)	(0.79)	(0.76)	(0.74)	(0.79)	(0.83)	(0.84)	(0.77)	(0.82)	(0.94)	(1.11)	(4)
JMA	0.99	0.82	0.84	0.85	0.84	0.81	0.91	0.86	0.84	0.95	0.92	0.80	0.76	0.80	0.88	4
	(0.99)	(0.85)	(0.89)	(0.94)	(0.96)	(0.90)	(0.91)	(0.87)	(0.93)	(0.96)	(0.96)	(0.83)	(0.86)	(0.96)	(0.91)	(0)
Factor	0.87	0.78	0.78	0.79	0.78	0.78	0.80	0.81	0.82	0.84	0.84	0.78	0.82	0.90	1.17	4
	(0.88)	(0.80)	(0.80)	(0.82)	(0.82)	(0.80)	(0.78)	(0.80)	(0.87)	(0.87)	(0.87)	(0.82)	(0.89)	(1.02)	(1.21)	(1)
T. Factor	0.88	0.79	0.78	0.80	0.77	0.79	0.79	0.80	0.80	0.82	0.83	0.78	0.82	0.91	1.17	3
	(0.87)	(0.82)	(0.81)	(0.84)	(0.83)	(0.84)	(0.80)	(0.80)	(0.84)	(0.87)	(0.86)	(0.80)	(0.90)	(1.06)	(1.23)	(0)
B. Factor	0.95	0.77	0.76	0.78	0.77	0.79	0.79	0.78	0.79	0.83	0.84	0.74	0.82	0.91	1.17	10
	(0.96)	(0.80)	(0.81)	(0.85)	(0.84)	(0.86)	(0.84)	(0.82)	(0.85)	(0.86)	(0.86)	(0.75)	(0.92)	(1.13)	(1.32)	(1)
RF	0.84	**0.73**	**0.71**	0.74	**0.71**	**0.72**	**0.72**	0.71	0.72	**0.76**	**0.77**	**0.68**	**0.71**	**0.71**	**0.77**	15
	(0.81)	**(0.72)**	**(0.71)**	(0.75)	**(0.73)**	**(0.73)**	**(0.70)**	**(0.68)**	**(0.72)**	**(0.75)**	**(0.77)**	**(0.67)**	**(0.74)**	**(0.77)**	**(0.77)**	(15)

Consumer Price Index 1990–2015

Mean	0.83	0.75	0.73	0.76	0.74	0.74	0.75	0.74	0.75	0.77	0.78	0.71	0.75	0.80	0.95	14
	(0.81)	(0.74)	(0.73)	(0.76)	(0.76)	(0.75)	(0.73)	(0.71)	(0.75)	(0.76)	(0.78)	(0.70)	(0.78)	(0.87)	(0.97)	(14)
T.Mean	0.84	0.74	0.73	0.75	0.74	0.74	0.75	0.73	0.74	0.78	0.79	0.71	0.75	0.79	0.95	14
	(0.81)	(0.74)	(0.72)	(0.76)	(0.75)	(0.74)	(0.72)	(0.70)	(0.74)	(0.77)	(0.79)	(0.70)	(0.78)	(0.86)	(0.96)	(14)
Median	0.84	0.75	0.72	0.76	0.74	0.74	0.75	0.73	0.74	0.78	0.79	0.71	0.75	0.79	0.94	14
	(0.81)	(0.74)	(0.72)	(0.76)	(0.76)	(0.74)	(0.73)	(0.70)	(0.74)	(0.77)	(0.79)	(0.71)	(0.78)	(0.86)	(0.97)	(14)
RF/OLS	**0.81**	0.73	0.72	0.75	0.74	0.75	0.75	0.74	0.74	0.78	0.79	0.71	0.73	0.78	0.94	14
	(0.79)	(0.73)	(0.72)	(0.76)	(0.76)	(0.76)	(0.73)	(0.72)	(0.75)	(0.78)	(0.81)	(0.72)	(0.78)	(0.86)	(0.97)	(14)
adaLASSO/RF	0.85	0.76	0.72	**0.73**	0.73	0.72	0.72	**0.71**	**0.72**	0.79	0.82	0.70	0.73	0.73	0.80	15
	(0.82)	(0.73)	(0.72)	**(0.74)**	(0.74)	(0.73)	(0.71)	(0.68)	(0.72)	(0.79)	(0.82)	(0.68)	(0.76)	(0.80)	(0.82)	(15)

Consumer Price Index 1990–2015																
RMSE count	14	15	13	17	19	19	17	15	17	18	19	7	13	17	5	
MAE count	(12)	(13)	(12)	(11)	(10)	(12)	(14)	(12)	(10)	(15)	(16)	(7)	(13)	(13)	(4)	

Table 6 Real-Time Forecasting Results: RMSE, MAE and MAD Ratios.

The table reports, for each forecasting horizon, the root mean squared error (RMSE), mean absolute error (MAE) and median absolute deviation from the median (MAD) ratios with respect to the random walk model for the period 2001–2015. The last three columns represent, respectively, the ratios for the accumulated three, six, and twelve-month forecasts. The statistics for the best-performing model are highlighted in bold. The forecasts are computed in real-time.

	Panel (a): RMSE Ratio														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
AR	0.91	0.81	0.77	0.78	0.79	0.80	0.80	0.77	0.79	0.82	0.83	0.77	0.84	0.95	1.21
UCSV	0.97	0.82	0.79	0.80	0.78	0.79	0.81	0.80	0.79	0.81	0.81	0.77	0.86	0.88	0.87
RR	**0.84**	0.74	0.72	0.75	0.76	0.78	0.79	0.76	0.77	0.79	0.80	0.74	0.72	0.80	0.96
adaLASSO	0.93	0.92	1.08	1.01	1.56	0.96	0.96	0.92	0.88	0.93	0.94	0.87	0.86	1.03	1.11
RF	0.88	0.74	**0.70**	0.72	**0.72**	0.75	0.76	0.74	**0.74**	0.79	0.80	**0.72**	**0.69**	**0.64**	**0.65**
RF/OLS	**0.84**	0.75	0.72	0.75	0.76	0.78	0.78	0.76	0.76	0.78	0.79	0.73	0.73	0.79	0.95
adaLASSO/RF	0.85	**0.73**	**0.70**	**0.71**	0.74	**0.73**	**0.75**	**0.72**	**0.74**	**0.77**	**0.78**	**0.72**	0.70	0.72	0.74
	Panel (b): MAE Ratio														
	Forecasting Horizon														

	Panel (a): RMSE Ratio														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
AR	0.87	0.80	0.73	0.75	0.80	0.81	0.79	0.75	0.77	0.80	0.84	0.78	0.83	0.98	1.19
UCSV	0.91	0.82	0.77	0.77	0.79	0.80	0.81	0.80	0.79	0.80	0.83	0.78	0.85	0.91	0.86
RR	**0.81**	0.72	0.68	0.72	0.77	0.80	0.77	0.73	0.76	0.76	0.82	0.74	0.72	0.81	0.94
adaLASSO	0.88	0.91	0.89	0.98	1.14	1.00	1.00	0.94	0.88	0.92	1.01	0.92	0.87	1.11	1.09
RF	0.83	0.72	0.66	0.72	0.74	0.76	**0.75**	0.72	**0.72**	**0.75**	0.81	**0.73**	**0.68**	**0.65**	**0.63**
RF/OLS	0.82	0.73	0.69	0.74	0.77	0.80	0.79	0.74	0.75	0.76	**0.80**	0.74	0.74	0.80	0.90
adaLASSO/ RF	**0.81**	**0.71**	**0.65**	**0.68**	**0.73**	**0.74**	**0.75**	**0.70**	**0.72**	**0.75**	**0.80**	0.74	0.69	0.71	0.71
	Panel (c): MAD Ratio														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
AR	0.80	0.78	0.66	0.80	0.76	0.76	0.73	0.74	0.81	0.77	0.83	0.73	0.83	1.13	0.95
UCSV	0.86	0.82	0.82	0.81	0.80	0.76	0.90	0.86	0.84	0.75	0.92	0.70	0.93	1.03	0.88
RR	0.77	0.70	**0.60**	0.76	0.74	0.75	**0.75**	0.72	0.77	**0.69**	**0.75**	**0.67**	0.72	0.90	0.75
adaLASSO	0.75	0.94	0.86	1.01	1.07	1.04	1.09	1.05	1.05	0.90	1.08	0.94	0.85	1.22	0.97

	Panel (a): RMSE Ratio														
RF	0.74	0.72	0.62	0.73	0.75	0.77	**0.75**	0.73	0.76	0.65	0.81	0.68	**0.72**	**0.73**	**0.53**
RF/OLS	0.77	0.71	0.66	0.76	0.77	0.78	0.81	0.77	0.83	0.72	0.85	0.69	0.76	0.84	0.77
adaLASSO/RF	**0.70**	**0.68**	**0.60**	**0.66**	**0.74**	**0.73**	0.78	**0.64**	**0.71**	0.76	0.81	**0.67**	0.69	0.80	0.56

Table 7 Forecasting Results (Alternative Models).

The table reports, for each forecasting horizon, the root mean squared error (RMSE), mean absolute error (MAE) and median absolute deviation from the median (MAD) ratios with respect to the random walk model for the full out-of-sample period (1990–2015). The last three columns represent, respectively, the ratios for the accumulated three, six, and twelve-month forecasts. The statistics for the best-performing model are highlighted in bold.

	Panel (a): RMSE Ratio														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RF	**0.84**	**0.73**	**0.71**	0.74	**0.71**	0.72	**0.72**	**0.71**	**0.72**	**0.76**	**0.77**	**0.68**	**0.71**	0.71	0.77
SCAD	0.85	0.76	0.77	0.79	0.77	0.77	0.81	0.77	0.78	0.83	0.84	0.73	0.75	0.79	0.96
BTrees	0.88	0.78	0.74	0.76	0.73	0.74	0.74	0.73	0.73	0.77	0.79	0.71	0.74	**0.70**	**0.71**
Deep NN	1.00	0.84	0.78	0.80	0.78	0.85	0.79	0.77	0.78	0.84	0.81	0.75	0.81	0.84	0.90
LASSO	0.89	0.76	0.74	0.76	0.74	0.72	0.74	0.75	0.75	0.79	0.83	0.74	0.79	0.83	1.04
adaLASSO	0.91	**0.73**	**0.71**	**0.73**	**0.71**	**0.70**	**0.72**	0.73	0.74	**0.76**	0.80	0.69	0.74	0.73	0.80
	Panel (b): MAE Ratio														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RF	**0.8**	0.7	**0.7**	**0.7**	**0.7**	0.7	0.7	**0.6**	**0.7**	**0.7**	**0.7**	**0.6**	**0.7**	0.7	0.7

	Panel (a): RMSE Ratio														
	1	2	**1**	**5**	**3**	3	0	**8**	**2**	**5**	**7**	**7**	**4**	7	7
SCAD	0.84	0.77	0.78	0.81	0.81	0.78	0.80	0.74	0.79	0.84	0.87	0.76	0.80	0.87	0.99
BTrees	0.84	0.75	0.74	0.78	0.76	0.77	0.74	0.70	0.73	0.75	0.78	0.70	0.77	**0.75**	**0.71**
Deep NN	0.97	0.84	0.83	0.85	0.84	0.87	0.79	0.79	0.83	0.85	0.84	0.79	0.83	0.93	0.91
LASSO	0.94	0.78	0.75	0.79	0.79	0.74	0.74	0.76	0.80	0.82	0.89	0.78	0.88	1.01	1.21
adaLASSO	0.86	**0.71**	**0.71**	**0.75**	**0.73**	**0.71**	**0.69**	0.71	0.75	0.76	0.82	0.69	0.75	0.80	0.85
	Panel (c): MAD Ratio														
	Forecasting Horizon														
Model	1	2	3	4	5	6	7	8	9	10	11	12	3m	6m	12m
RF	**0.70**	**0.63**	0.77	0.84	0.75	0.73	**0.65**	0.64	**0.73**	0.69	**0.71**	**0.58**	**0.71**	**0.80**	**0.59**
SCAD	0.80	0.75	0.79	0.83	0.89	0.77	0.71	0.69	0.80	0.78	0.84	0.70	0.82	1.02	0.81
BTrees	0.73	0.70	0.79	0.86	0.80	0.78	0.69	0.71	0.76	**0.68**	0.77	0.65	0.79	0.88	0.61
Deep NN	0.85	0.73	0.90	0.97	0.94	0.83	0.78	0.86	0.95	0.83	0.87	0.83	0.82	1.15	0.87
LASSO	0.83	0.68	**0.70**	**0.80**	**0.73**	**0.68**	0.66	0.64	0.76	0.73	0.73	0.60	0.77	0.88	0.65
adaLASSO	0.81	0.65	0.78	0.81	0.81	0.76	0.63	**0.63**	0.81	0.78	0.82	0.66	0.75	0.90	0.76

	Panel (a): RMSE Ratio														