# The Macroeconomy as a Random Forest

Philippe Goulet Coulombe[*]
University of Pennsylvania

First Draft: November 15, 2019
This Draft: June 24, 2020
Latest Draft Here

## Abstract

Over the last decades, an impressive amount of non-linearities have been proposed to reconcile reduced-form macroeconomic models with the data. Many of them boil down to have linear regression coefficients evolving through time: threshold/switching/smooth-transition regression; structural breaks and random walk time-varying parameters. While all of these schemes are reasonably plausible in isolation, I argue that those are much more in agreement with the data if they are combined. To this end, I propose Macroeconomic Random Forests, which adapts the canonical Machine Learning (ML) algorithm to the problem of flexibly modeling evolving parameters in a linear macro equation. The approach exhibits clear forecasting gains over a wide range of alternatives and successfully predicts the drastic 2008 rise in unemployment. The obtained generalized time-varying parameters (GTVPs) are shown to behave differently compared to random walk coefficients by adapting nicely to the problem at hand, whether it is regime-switching behavior or long-run structural change. By dividing the typical ML interpretation burden into looking at each TVP separately, I find that the resulting forecasts are, in fact, quite interpretable. An application to the US Phillips curve reveals it is probably not flattening the way you think.

[*] Department of Economics, gouletc@sas.upenn.edu. For many helpful discussions, I would like to thank (but not implicate) Karun Adusumilli, Edvard Bakhitov, Francesco Corsello, Frank Diebold, Maximilian GÃűbel, Frank Schorfheide, Dalibor Stevanovic and Boyuan Zhang.

# 1 Introduction

In recent years, the rise of Machine Learning (ML) lead to great excitement in the econometrics community. In empirical macroeconomics, a first wave of papers took ML algorithms of the shelf and applied them to classical problems in the literature – mostly forecasting. Such enterprises were faced with varying levels of success. This was not met with enormous surprise given that macroeconomic data (and even time-series data in general) is known to have very different properties than what most ML algorithms are designed and optimized for. An increasingly popular and arguably more fruitful research agenda is to understand what these algorithms really do, pin down the specific economic problems that they can be helpful with, and finally, adapt current macroeconomic modeling approaches accordingly. This paper lies in this line of work.

**NON-LINEAR TIME SERIES & EMPIRICAL MACRO.** Before attempting to explain facts with a structural model, one must know the facts. As it turns out, not so many statistical consensus emerge from the use of time series econometrics on macroeconomic data. Even before addressing the delicate question of identification, one must wonder "are all the relevant variables included in the model" and at a higher level of sophistication, "is linearity a valid approximation of reality?". The first question led to the development of factor models and large Bayesian Vector Autoregressions over the last 20+ years. To address the second question, applied macroeconomic researchers have proposed many non-linear modeling schemes based on reasonable reduced-form economic intuition. A great amount of them amounts to have regression coefficients $\beta_t$ in

$$y_t = X_t\beta_t + \varepsilon_t$$

evolving through time in some way or another. To name a few, we have threshold/switching regressions (Hansen (2011)), smooth transition (Teräsvirta (1994)), structural breaks (Perron et al. (2006), Stock (1994)) and random walk time-varying parameters (Sims (1993), Cogley and Sargent (2001), Primiceri (2005)).[1] While one can reasonably state that factor models and large VARs have gone a long way in achieving what they were designed for, less victorious statements are available for the various time-variation proposals. Why?

**OBSERVABLE TIME-VARIATION VIA INTERACTION TERMS.** A natural approach to create time-variation in a linear equation is to use of interaction terms. Those have many refinements, but the point made here generalizes easily. A way to obtain switching regimes based on an observed regressor is a threshold model where the threshold variable $q_t$ is somehow cleverly chosen by the researcher. However, using readily available FRED US macro data reveals that

[1] Naturally, we now also have at our disposal approaches that combine the two, like large time-varying VAR (Giraitis et al. (2018), Koop and Korobilis (2013)).

there is a great number of candidates for $q_t$. Choosing the right one in a data-driven way requires some serious grid search work. Additionally, the model may have multiple regimes that interact in ways a researcher cannot easily design based on economic a prioris. Another possibility is that the right $q_t$ (or one of them) is a complicated and unknown function of available predictors. Structural breaks or slow exogenous variation could also get in the way. The list goes on. This renders a credible exploration of the threshold structures space impossible. As a result, the enterprise of manually specifying the model is very much compromised.

This discussion is not merely of theoretical consideration, its implications in empirical work being clearly visible. For instance, Auerbach and Gorodnichenko (2012b) and Ramey and Zubairy (2018) pick some transformation of GDP/unemployment as $q_t$ in a model where the effect of fiscal policy smoothly transits between that of high and low states. Batini et al. (2012) expand the former to generate fiscal policy effects that depend on whether it happened during expansion vs recession and whether it was implemented via revenue or spending.[2] It rapidly occurs to applied econometricians that the possibilities are endless.[3] Given currently available time-series models, this is more of a curse than a blessing.

**Latent Time-Variation.** A different view on all this can be found in methods that fall under the umbrella of "latent change". In this broad line of work, the evolving $\beta_t$ either follows a law of motion (random walk, Markov process) or is subject to discrete breaks. Structural breaks and smoothly varying parameters consider time-variation as exogenous. In Markov-switching models, it is rather an endogenous product of the whole model. However, all share the common feature that time-variation is a latent state. At first glance, this appears to solve many of the problems detailed for models using refinement of interaction terms. By treating $\beta_t$ as a state to be filtered/estimated within the model, the complexity of characterizing its path correctly out of abundant data seems to vanish. Alas, estimating the path $\beta_t$ implies a great number of parameters which inevitably comes with the use of strong regularization. That regularization is the law of motion itself, a choice which is far from innocuous. Whether it is latent regime-switching, exogenous breaks or slow change, none of them can easily accommodate for the additional presence of the other.[4] Yet, these models are routinely fitted *separately* on the *same data*. Consequently, one observes that regime-switching models often suggest evidence of regime-switching behavior in the data, whereas smooth time variation models will find smooth

[2] Bognanni (2013) studies the same issue with a Bayesian Markov-Switching VAR and find that fiscal multipliers are likely *smaller* in recessions, contradicting previous evidence.

[3] Ramey (2016) surveys a finite number of them for monetary policy shocks.

[4] For instance, if switching behavior is present in the data, neglecting it will seriously compromise the successful detection of structural change. Effectively, regime-switching behavior could be modeled (very inefficiently) as a multitude of breaks but this would fail to leverage the recurrent nature of the process and provide the false impression of an ever-changing economic structure. Random-walk time-varying parameters only provide a smoother version of the same roadblock.

time variation (or nothing).

Additionally, while some of these approaches rationalize the data quite well in-sample, many of them will struggle to outperform the simple autoregression *out-of-sample*. Often, the very nature of the assumed law of motion guiding $\beta_t$ makes forecasting difficult: successfully detecting a structural break, for instance, is much harder without great amount data on both sides of it. By construction, exogenous structural change creates forecasting headaches and should be, in some sense, a time-variation of last resort. For instance, if the slope of the Phillips curve has changed because an economy became more much open than it used to be, including an interaction term with some measure of national trade is wildly more efficient than modeling the whole $\beta_t$ path non-parametrically. As we have seen in the discussion of interaction term methods, this option also has its own catalog of pitfalls. To get out of this deadlock, we can use a little help from Machine Learning.

**Macroeconomic Random Forest**. I introduce Macroeconomic Random Forests (MRF) which adapt the canonical Machine Learning algorithm to the reality of macro data. I extend Friedberg et al. (2018) Local Linear Forests to macroeconomic data and create specific data transformations that boost RF performance. The approach's main output are Generalized Time-Varying Parameters (GTVPs) which address many of the issues discussed above. Additionally, it procures an explanation for forecast originating from a black box model via its representation as a linear macro model with time-varying coefficients. To boost performance and interpretability in a time series context, I propose the use of Moving Average Factors (MAFs) as a simple way to compress ex-ante the information contained in the lags of a given predictor. I argue that this transformation is helpful to avoid running out of splits (when growing trees) and is motivated by the rather large literature on constraining/regularizing lag polynomials. Additionally, I provide a regularization scheme better suited for time series that procures desirable smoothness (with respect to time) of the estimates. A variant of the Bayesian Bootstrap provides credible regions which are instrumental for the interpretation of GTVPs.

**Preview of Results**. The tool does well on simulated data and confirms intuition about when we expect the approach to outperform standard models. In a forecasting application, the gains from using specific MRFs will be present for almost all variables and horizons under study, which is rarely the case for non-linear forecasting approaches. For instance, a simple Autoregressive Random Forest will very rarely be worse than its OLS counterpart. A combination of the Autoregressive Diffusion Indexes (henceforth ARDI, basically a factor model with factors extracted by principal component analysis) with RF generate very accurate forecasts of the 2008 downturn for both GDP and unemployment rate (UR). Inspection of resulting GTVPs reveals a noticeably different behavior from random walk TVPs. For instance, parts of the UR equation clearly alternate between two states. In contrast, inflation is subject to various sorts of time-variation. The long-run mean and the persistence evolved slowly and in an exogenous fashion. The contribution from real activity depends positively on the strength of well-known leading indicators, like those that pertain to the state of the housing market. Following this lead, I complete the analysis by looking at a traditional Phillips' curve formulation. I report that the inflation/unemployment trade-off coefficient decreased significantly since the 1980's, and also varies strongly along the business cycle. Indeed, the trade-off itself could very well depend on other activity indicators (like Capacity Utilization). Overall, this suggest inflation can rise from a positive unemployment gap, but much more timidly does it go down from economic slack. These findings are made possible by combining different tools within the new framework, such as credible intervals for the GTVPs, new variable importance measures specifically designed for MRF and small surrogate trees as an interpretative devices for GTVPs.

**Roadmap.** In section 2, I first introduce MRF and motivate its use with respect to available alternatives. Secondly, I present simulations results and forecasting results in sections and 3 and 4 respectively. Section 5 looks at GTVPs in various ways and interpret some key forecasts. Section 6 concludes.

## 2 Macroeconomic Random Forests

This section introduces Macroeconomic Random Forest (MRF). I first motivate the use of trees as basis functions by casting relatively standard switching structures for autoregressions as special cases. Second, I detail the MRF mechanics and how it in fact estimates generalized time-varying parameters (GTVPs). Third, I discuss how the approach relates to both standard RF and traditional random walk TVPs. Fourth, I discuss potential for interpretation and a way to assess parameter's uncertainty.

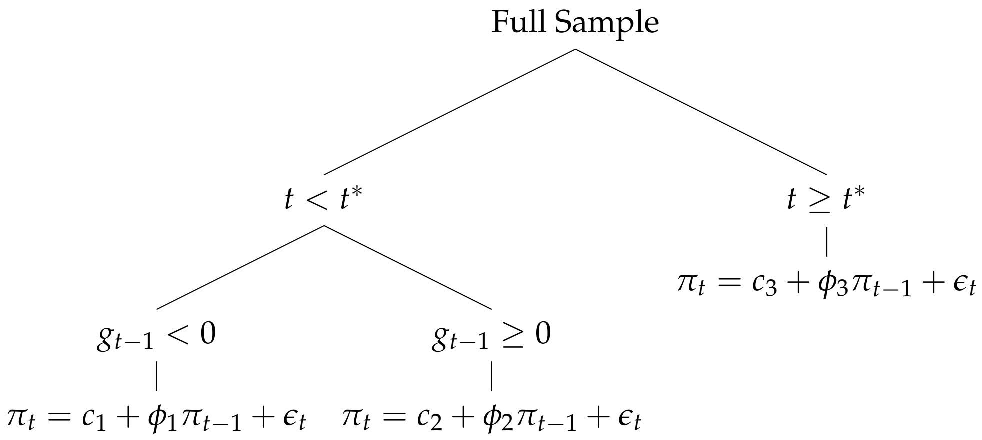
### 2.1 Traditional Macro Non-Linearities as Trees

Whether it is empirical macro or forecasting, the use of *state-dependent* models is usually accompanied by some economic narrative. The first line of literature will proceed to test if treatment effects/IRFs are statistically different from each other. The second will also consider testing (Hansen (2011)) but will also typically run a forecasting experiment and hope for a rejection of the null of a Diebold and Mariano (2002) test. When dealing with *structural change*, in typical empricial macro practice, one would drop early observations or consider subsamples (Champagne and Sekkel (2018) is a recent example). The forecasting literature will consider estimating models using a rolling-window. This way of addressing structural change can be understood as TVPs implemented via a very specific kernel.

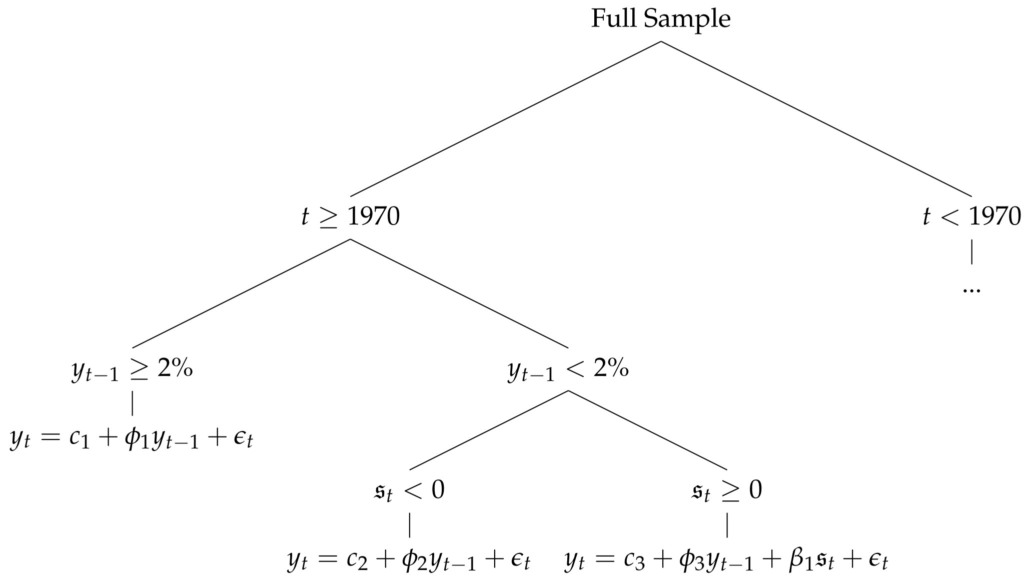
The goal of this section is to show that regression trees can encompass and automatize all the above – and more. The method at the heart of this paper aims to eliminate the arbitrary search for an econometric specification. It aims at creating a unified view so that the myriad of time-variations suggested separately can now be tackled jointly. It has the potential to uncover time-varying patterns that were unsuspected based on rough economic intuition. It can limit substantially look-ahead bias that inevitably creeps in when one does manual specification search.

To provide a simple motivation for the Macroeconomic RF approach, I now run through two simple examples that display how common non-linearities have a tree structure for an AR process, without loss of generality. This chain of deduction will eventually lead to the proposition of the simple Autoregressive Random Forest (ARRF).

Let us first consider the behavior of inflation in a country where inflation targeting (IT) was implemented at a publicly known date – like in Canada. Let $\pi_t$ be inflation at time $t$ and $t^*$ is the inflation targeting implementation date. Additionally, $g_t$ is some measure of output gap. A plausible model of inflation can then have the form of the following tree graph. The story is straightforward. Inflation behaved differently before vs after IT. After IT, it is a simple AR process. Before IT, it was a switching AR process that depended on the sign of the output gap. Both the mean and the persistence depend on $g_t$.


Full Sample
$t < t^*$
$t \geq t^*$
$\pi_t = c_3 + \phi_3 \pi_{t-1} + \epsilon_t$
$g_{t-1} < 0$
$g_{t-1} \geq 0$
$\pi_t = c_1 + \phi_1 \pi_{t-1} + \epsilon_t$
$\pi_t = c_2 + \phi_2 \pi_{t-1} + \epsilon_t$

Here is a second (more intricate) example. Loosely inspired by Auerbach and Gorodnichenko (2012a), Ramey and Zubairy (2018) and others, let $y_t$ be GDP growth at time $t$. Let $\mathfrak{s}_t$ be some measure of government spending shock.


Full Sample
$t \geq 1970$
$t < 1970$
...
$y_{t-1} \geq 2\%$
$y_t = c_1 + \phi_1 y_{t-1} + \epsilon_t$
$y_{t-1} < 2\%$
$\mathfrak{s}_t < 0$
$y_t = c_2 + \phi_2 y_{t-1} + \epsilon_t$
$\mathfrak{s}_t \geq 0$
$y_t = c_3 + \phi_3 y_{t-1} + \beta_1 \mathfrak{s}_t + \epsilon_t$

The above tree tells us that only data post 1970 is of "current" interest. The high-growth environment of pre-1970 being characterized by a different process. The effect of spending $\mathfrak{s}_t$ on growth $y_t$ depends on two variables: previous growth $y_{t-1}$ (the state of the economy) and whether government spending $\mathfrak{s}_t$ is expanding or contracting. Hence, this tree allows for different mean/dynamics of growth and different treatment effects of spending conditional on three variables: $t$, $y_{t-1}$ and $\mathfrak{s}_t$.

These are two stories out of many that trees can characterize. This is both good and bad news. It highlights the flexibility of the tree structure. It also suggests that designing the "true one" purely from economic deduction is a daunting task – the space of economic stories being arguably infinite-dimensional. The problem calls for an alternative approach.

**Autoregressive Random Forest.** A general way of writing the problem of unknown time-varying parameter structure for the autoregressive example is

$$\begin{aligned} y_t &= \mu_t + \phi_t y_{t-1} + \epsilon_t \\ \mu_t &= \mathcal{F}_\mu(S_t; \Psi) \\ \phi_t &= \mathcal{F}_\phi(S_t; \Psi) \end{aligned} \tag{1}$$

where $S_t$ are the state variables governing time-variation. If we knew the threshold variables and values (as parametrized by the generic $\Psi$), obtaining $\phi_t$ and $\mu_t$ amounts to run separate regressions on different subsamples.[5] Another possibility is to write it as a linear regression with interaction terms. If the structure of the operator $\mathcal{F}$ is unknown, then one could resort to a global grid search. However, that is infeasible if either $S_t$ is large or if we want to consider

[5] As this will be discussed later, constraining $\phi_t = 0$ leads to fit a standard RF.

more than a few splits. These are unfortunately the conditions facing the (purely agnostic) macroeconomic modeler.

A natural way forward is to proceed with recursive partitioning of the data set with trees.[6] With the latter known for its high variance (Friedman et al. (2001)), one must do Bootstrap Aggregation – *Bagging* – of many de-correlated trees in order to keep the generalization error in check. This is the famous Random Forests proposition of Breiman (2001). From this deduction chain, we arrive at the natural proposition of fitting an Autoregressive Random Forest (ARRF), a special case of MRFs. That is, the operator $\mathcal{F}$ above is approximated by a forest – an ensemble of trees.[7] This autoregressive example can be generalized for any macro model we want to be time-varying.

## 2.2 Generalized Time-Varying Parameters

The general model is

$$y_t = X_t\beta_t + \epsilon_t$$

$$\beta_t = \mathcal{F}_\beta(S_t; \Psi)$$

where $S_t$ are the state variables that determine time-variation. The exact composition of $S_t$ optimized for macro forecasting is motivated in section 2.6 and laid out explicitly in section 4.1.4. $\Psi$ parametrizes the structure of the many tree functionals. $X$ determines the *linear* model that we want to be time-varying. As in Friedberg et al. (2018), the tree fitting procedure is modified to

$$\min_{j\in\mathcal{J}^-,\ c\in\mathbb{R}} \left[ \min_{\beta_1} \sum_{\{t\in l|S_{j,t}\leq c\}} (y_t - X_t\beta_1)^2 + \lambda\|\beta_1\|_2 \right. \\ \left. + \min_{\beta_2} \sum_{\{t\in l|S_{j,t}>c\}} (y_t - X_t\beta_2)^2 + \lambda\|\beta_2\|_2 \right]. \tag{2}$$

The purpose of this problem is to find the optimal variable $S_j$ (so, finding the best $j$ out of the sub-sample of predictors indexes $\mathcal{J}^-$) to split the sample with and at which value $c$ of that variable should we split. All these values, along with the recursive structure are encoded in $\psi$. We start with the leaf $l$ being the full sample. Then, we perform a split according to the minimization problem, which procures us with 2 sub-samples. Within each of these two newly created sub-samples, we run equation (2) again and obtain a new set of $l$'s. Doing so recursively until a stopping criteria is met generates a tree.

Recursively splitting $\beta_0$ into $\beta_1$ and $\beta_2$ eventually leads to $\beta_t$. However, $\beta_t$, by construc-

[6] Autoregressive Trees (ART) were proposed in Meek et al. (2002).

[7] As we will see in section 2.7, this will correspond to the posterior on a tree functional $\mathcal{T}$.

tion, has very little company within its terminal node/leaf. As result, a single tree has low bias but also very high variance for both $\beta_t$ ad the prediction and it characterize. A Random Forest is a clever re-sampling scheme that creates many trees and then averages them. First, the many trees are "grown" on bootstrapped samples, which is commonly referred to as Bagging (for Bootstrap Aggregation, Breiman (1996)). For tree $b$ out of a total of $B$ trees, we draw observations with replacement from the true sample. Since recursive partitioning is non-linear in observations' weights, gains from averaging can be large as pointed out in Grandvalet (2004).[8] Second, to maximize the benefits of Bagging, Breiman (2001) suggest averaging trees that are de-correlated. Averaging reduces variance at a much faster rate if its components are un-correlated. This is obtained by growing trees stochastically. In equation (2), this is made operational by having $\mathcal{J}^- \subset \mathcal{J}$ rather than $\mathcal{J}$ itself. That is, at each step of the recursion, $\mathcal{J}^-$ is a different random sub-sample of predictors that are considered for the split. Finally, the fraction of randomly selected predictors is a tuning parameter typically referred to as `mtry` in the literature (and all software) with a default value of $\frac{1}{3}$ for regression settings. Stacking all specific trees' parametrization $\psi$ in one vector leads to $\Psi$ of the general model above. Unlike Friedberg et al. (2018) where $S_t = X_t$, the information sets can differ, which more natural when the approach is motivated from a TVP perspective. Indeed, forcing their equivalence is not feasible nor desirable in a macro forecasting environment.

The standard Random Forest has many practical qualities that are transferable to MRFs. It is easy to implement and to tune. That is, it has few tuning parameters that are usually of little importance to the overall performance – robustness. It is relatively immune to the adverse effects of including many irrelevant features (Friedman et al. (2001)). Given the standard ratio of regressors to observations in macro data, this is a non-negligible advantage. Furthermore, with a sufficiently high `mtry`, it can adapt nicely to sparsity and discard useless predictors (Olson and Wyner (2018)). Finally, its vanilla version already shows good forecasting performance for US inflation (Medeiros et al. (2019)) and macro data in general (Chen et al. (2019), Goulet Coulombe et al. (2019)).

## 2.3 Adapting Regularization for Time Series

Equation (4) uses Ridge shrinkage which implies that each time-varying coefficient is implicitly shrunk to 0 at every point in time. $\lambda$ and the prior it entails can have a significant influence. For instance, if a process is highly persistent (AR coefficient lower than 1 but nevertheless quite high) as it is the case for SPREAD, shrinking heavily the first lag to 0 could incur serious bias.[9]

[8] For reasons detailed in section 2.7, a more sophisticated bootstrapping procure is preferable when it comes to time series data.

[9] An intuitive solution to this specific problem could be to change prior mean in the spirit of a Minnesota prior for Bayesian VARs. However, the fact that $S_t$ can itself include members of $X_t$ makes a coherent specification of

That is, like any Ridge regression, the specification of previous sections implies that if $\lambda$ grows large, $\forall t \;\; \beta_t = 0$. The need for a large $\lambda$ depends on the size of the resulting leaves which themselves depend on the original time series' length. In a macro setup, the need for regularization is present even if we do not attempt any sample-splitting. Thus, in MRFs, the importance of the prior we postulate for $\beta_t$ does not vanish at all, it is magnified. $\beta_i = 0$ is the natural choice of stochastic constraint in Friedberg et al. (2018)'s cross-sectional setting. However, its time series translation $\beta_t = 0$ can easily be sub-optimal. The traditional regularization employed for stochastic coefficients in macro is rather the random walk law of motion

$$\beta_t = \beta_{t-1} + u_t.$$

Thus, it is desirable to transform (4) so that it implements the prior that coefficients evolve smoothly, which is just shrinking $\beta_t$ to be in the neighborhood of $\beta_{t-1}$ rather than 0. The random walk regularization ensure that the parameter's path will be smooth, to some extent. This is more in line with the view that economic states (as expressed by $\beta_t$ here) last for at least a certain number of consecutive periods. Such shrinkage will greatly facilitate interpretation of resulting GTVPs.

The question is how to implement such shrinkage. A look at (4) reveals that $\beta_t = \beta_{t-1} + u_t$ makes the $T$ problems linked together, which was not the case before. To remedy that, I will rather implement the desired regularization by taking the rolling-window view of time-varying parameters. That is, the tree, instead of solving a plethora of small ridge problems, will rather solve many weighted least squares problems with additional close-by observations. The latter will be selected by being in the neighborhood (in time) of those in the current leaf. For simplicity's sake and to keep computational demand low, the kernel implicitly used by WLS will be rather rudimentary: it will look like a symmetric 5-step Olympic podium. Informally, the kernel puts a weight of 1 on observation $t$, a weight of $\zeta < 1$ for observations $t-1$ and $t+1$ and a weight of $\zeta^2$ for observations $t-2$ and $t+2$. Since some specific $t$'s will come up many times (for instance, if both observations $t$ and $t+1$ are within the same leaf, kernels overlap), I take the maximal weight allocated to $t$ as the final weight $w(t;\zeta)$. Formally, define a subset of observations as considered by the tree algorithm as §. At each splitting step, for the evaluated

such a prior difficult. Furthermore, the proposition of this section is more generally applicable to any model that does or does not have a VAR structure.

subsample §, I use

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

for $\zeta < 1$, a tuning parameter guiding the level of time-smoothing. Let

$$§_1(j,c) \equiv \{t \in l | S_{j,t} \leq c\} \quad \text{and} \quad §_2(j,c) \equiv \{t \in l | S_{j,t} > c\}.$$

Finally, the expanded sets can be defined as

$$\text{for } i = 1,2: \quad §_i^{RW}(j,c) \equiv §_i(j,c) \cup [§_i(j,c) + 1 \cup §_i(j,c) - 1] \cup [§_i(j,c) + 2 \cup §_i(j,c) - 2].$$

The splitting rule becomes

$$\min_{j \in \mathcal{J}^-, \ c \in \mathbb{R}} \left[ \min_{\beta_1} \sum_{t \in §_1^{RW}(j,c)} w(t;\zeta) \left(y_t - X_t\beta_1\right)^2 + \lambda \|\beta_1\|_2 \right. \\ \left. + \min_{\beta_2} \sum_{t \in §_2^{RW}(j,c)} w(t;\zeta) \left(y_t - X_t\beta_2\right)^2 + \lambda \|\beta_2\|_2 \right]. \tag{3}$$

Note that the Ridge penalty is kept in anyway, so the final model has in fact two sources of regularization. The traditional shrinkage of $\beta_t$ to 0 with an increasing $\lambda$ and smoothness of the $\beta_t$ path with the additional neighboring observations weighted by $w(t;\zeta)$. With $\zeta \rightarrow 0$, we are heading back to pure Ridge.

Although not considered in the main applications of this paper, it is possible – without additional computational burden – to consider models with a larger linear part. For instance, one could estimate equation by equation in a high-dimensional VAR which in practice would just require higher values of $\lambda$, $\zeta$ and a larger minimum size of the terminal leaves to behave well. Nonetheless, the empirical benefits from this strategy could prove to be limited. Put shortly, the time-varying constant in MRF can be seen as a complex misspecification function (in the deep learning jargon, it is effectively called the bias) that adaptively controls for omitted variables in a way that is both non-linear and strongly regularized via randomization (again, see discussion in Friedman et al. (2001)). Of course, this treats the extra regressors as exogenous which could be at odds with some researchers' will to investigate a large web of impulse response functions. Anyhow, both approaches are possible within MRF and it is up to the researcher to decide which one is more appropriate for a specific situation. As it turns out, large VAR specifications also deliver good results within MRF (see section 4.2.1). For instance, the high-dimensional VAR MRF (HD-VARRF) provides the best tracking of one-quarter ahead unemployment, as well as the best 1-year ahead forecasts for both unemployment and GDP – signaling a (albeit small) recession up to a year ahead.

## 2.4 Relationship to Random Walk Time-Varying Parameters

The specific use of Random Forest to obtain time-varying parameters comes with many advantages over standard alternatives. First, the algorithm selects by itself the relevant variables for time-variation, which is, as mentioned before, not the case for TAR, STAR and similar models. Second, unlike the reputedly flexible random walk parameters (Granger (2008)), it provides time-variation in the most efficient way. If the true DGP is some switching mechanism determined by an observable, random walk parameters will provide a very inefficient estimate of it (Aruoba et al. (2017)).[10]. Some dimensionality reduction techniques – for instance reduced-rank restrictions (Stevanovic (2016), Chan et al. (2018) and Goulet Coulombe (2019)) – can help, but nothing in that paradigm can come close to the parsimony obtained by simply interacting relevant variables. Third, even though the methodology is remarkably flexible in the model building step, it has low variance. Finally, it creates a data-driven hybrid sort of time-variation that pools *both* latent *and* observable time-variation via the composition of $S_t$. The fact that $t$ can be included in $S_t$ allows for what economists usually think of as exogenously evolving parameters and their usual characterization as latent states in a state space model.

Random Forest usually averages about 500 trees which makes inspection of individual trees impossible. A natural way to start is to look at TVPs themselves and their credible regions. On that regard, the new methodology is superior to RF and not better nor worse than standard TVPs. With MRF being sort of a hybrid between latent time variation and interactions terms approaches, there is more to do than simply starring at evolving coefficients. There exist many tools to open the black box and those will be used to deliver some insights in section 5.3. However, unlike the latter, we have much more at our disposal than a time series of $\phi_t$ to attempt interpretation. In fact, we have both the time series of $\phi_t$ and the complete structure that generated it.

Econometrically, The easiest way to connect this paradigm to recent work on time-varying parameters is to adopt the view that Random Forest are adaptive kernel estimators as in Meinshausen (2006), Athey et al. (2019) and Friedberg et al. (2018). The RF is seen as a machine that generates kernel weights. Once the weights $\alpha_t$ are obtained, estimation amounts to weighted

[10] In fact, consistency results for kernel versions Ãă la Giraitis et al. (2014) rely on smoothness assumptions that exclude breaks.

least squares (WLS) problem with a Ridge penalty. That is, by running (2) recursively, one obtains terminal nodes/leaves $L_b()$ to construct kernel weights

$$\alpha_t\left(\mathrm{x}_0\right)=\frac{1}{B}\sum_{b=1}^{B}\frac{1\left\{\mathrm{X_t}\in L_b\left(\mathrm{x}_0\right)\right\}}{\left|L_b\left(\mathrm{x}_0\right)\right|}$$

to use in

$$\forall t \quad: \operatorname{argmin}_{\beta_t}\left\{\sum_{\tau=1}^{T}\alpha_t\left(\mathbf{s}_\tau\right)\left(Y_\tau-X_\tau\beta_\tau\right)\right)^2+\lambda\|\beta_t\|_2\right\}. \tag{4}$$

As shown in Goulet Coulombe (2019), standard random walk TVPs are in fact a smoothing splines problem for which a reproducing kernel exists (Dagum and Bianconcini (2009)). Giraitis et al. (2014) drop the random walk altogether and proposed to use kernels directly. However, in both cases, the only variable entering the kernel is $t$. In other words, only information about proximity in time is considered for the clustering of observations. As mentioned earlier, this makes the seemingly flexible estimator in fact quite restrictive and dependent on the smoothness prior. Standard kernel methods are known to break down well before we include 10 variables in them (Friedberg et al. (2018)), hence augmenting $t$ with additional regressors is not an option in the current paradigm. However, no such constraint binds on the RF approach.

## 2.5 Relationship to Standard Random Forest

The standard RF is a restricted version of MRF where $X_t = \iota$, $\lambda = 0$ and $\zeta = 0$. In words, the only regressor is a constant and there is no within-leaf shrinkage. The previous sections motivated ARRF and the more general MRF starting from non-linear macro and time-series models. At this point, a reasonable question to ask is why we need MRFs rather than go with full blown RF and estimate everything non-parametrically. One reason is statistical efficiency. The other is potential for interpretation. Thus, the current section goes in the opposite direction and motivates the use of MRFs over that of standard RFs.

### 2.5.1 Smooth Relationships are Hard Relationships (to estimate)

In finite samples, RF can have a hard time learning smooth relationships – like a AR(1) process. This is bad news for time series applications. For prediction purposes, estimating by OLS

$$y_t=\phi y_{t-1}+\varepsilon$$

implies a single parameter. However, approximating the same relationship with a tree (or an ensemble of them) is far more consuming in terms of degrees of freedom. In fact, to get close to the straight line once parametrized parsimoniously by $\phi$, we now need a succession of many step functions.[11] This analogy has one takeaway: for small samples available in macroeconomics, modeling any smooth/linear relationships with step functions is a luxury one cannot afford. More generally, this is perhaps one of the main reasons why most successful non-linear time series models all include a linear part. Of course, a standard RF with a great number of observations and suitable tuning parameters can estimate anything non-parametrically. However, in a small sample setting, the complexity constraint is binding, and RF will waste a lot of splits on something that could have been achieved by a handful of coefficients in a linear model. It will run out of them before it gets to focus on true non-linearities – that is, more subtle phenomena that really cannot be captured by linear regression. In a language more familiar to economists, this simply means running out of degrees of freedom. If the true DGP has a linear and a non-linear part, then modeling the first part in the most concise way leaves more degrees of freedom left for the latter part. Since the complexity of the fitted function depends on the number of splits, which is itself limited by the sample size, the resulting (estimated) partially linear model could be, in fact, more non-linear than the fully non-parametric one.

This paper is not the first to recognize the potential need for a linear part in tree-based models. For instance, both Alexander and Grimshaw (1996) and Wang and Witten (1996) proposed linear regressions within a leaf of a tree, respectively denominated "Treed Regression" and "Model Trees". More focused on real activity forecasting, Woloszko (2020) and Wochner (2020) blend insights from macroeconomics to build better-performing tree-based models.[12] On a different end of the econometrics spectrum, Friedberg et al. (2018) proposed to improve the non-parametric estimation of treatment effect heterogeneity by combining those ideas developed for trees into a forest.[13] To my knowledge, this paper is the first to exploit explicitly the link between this strand of work and the sempiternal search for the "true" state-dependence in empirical macroeconomic models. The point is to leverage these new technologies to shed some more consensual light on evolving macroeconomic relationships.

[11] Indeed, in standard regression setup, nobody would model a continuous variable as a an ordinal one unless some wild non-linearities are suspected.

[12] Specifically, Wochner (2020) also note that using trees in conjunction with factor models can improved on GDP forecasting. An analogous finding will be reported in section 4.

[13] More broadly, this is extending to trees and ensemble of trees the "classical" non-parametrics literature's knowledge that local linear regression usually has much better properties (especially at the sample boundaries) than the Naradaya-Watson estimator.

### 2.5.2 A Note on Interpretability

The interpretation of Machine Learning outputs is now a field of its own (Molnar (2019)). Random Forests (unlike their tree components) are often considered as a prime example of a black box model which needs to be interpreted using an external device.

MRFs partially circumvent that problem by providing time series $\beta_t$ that can themselves be examined and have a clear meaning as time-varying parameters for the macro model. Hence, all the quantities reported in Cogley and Sargent (2001) and Primiceri (2005) can be produced with a VAR version of MRF. For instance, the evolving long-run growth of GDP and predictability of inflation can be computed exactly in the same fashion as in the original papers using as primary input $\beta_t^{VARRF}$. Time-varying IRFs are also possible. Section 5.1 makes a first step in that direction by looking more in depth at GTVPs of the forecasting experiment's most successful models.

A rather popular approach to interpret a standard RF by using surrogate tree models (Molnar (2019)) to try to replicate in part the black-box model's fit (that is, the whole conditional mean as one block) with interpretable models. This idea can be transferred to MRFs. In fact, partial linearity has the potential to facilitate such an exercise. The linear part in MRF splits the non-parametric atom into different pieces ($X_{t,k}\beta_{t,k}$) which can be analyzed separately. If $X_{t,k}$ are chosen such as to have a specific meaning in the final model, then each time series $\beta_{t,k}$ can be dissected with its own surrogate model. Meaningful combination/transformations of coefficients can also be considered.

**On Overfitting.** A relatively under-appreciated result is Breiman (2001)'s proof that the generalization error of RF converges to the true error as the number of tree grows. In short, with a proper amount of randomization in the fitting process, the algorithm does not overfit.[14] This is remarkable property. It is even more so when it comes to times series since dependence and structural change pose challenges to hyperparameter tuning. In the case of plain RF, given a large enough $B$, a reasonable `mtry` and standard sub-sampling rate, we can be confident that the out-of-bag prediction exclude fitted noise. Hence, it is re-assuring that the out-of-bag $\beta_t$ will mechanically inherit the great RF properties. In this context, it means the sample will not be over-split and we are not going to see time-variation when it is not there. Naturally, the credible regions proposed in section 2.7 will also help in that regard. This enviable property will be illustrated in section 3.2.4.

[14] Another equally remarkable result is that of Scornet et al. (2015) that shows RF is consistent, without having the leaf size grow large – as the latter are usually minuscule in practice.

## 2.6 Engineering $S_t$: Sparse, Dense and Other Unnecessary Dilemmas

$S_t$ is extremely wide (call it $K$ by 1 with a large $K$) but we do not have many observations: the curse of dimensionality. For most non-linear methods, this incurs both computational and statistical difficulties. The former is avoided in the case of RF since it does not rely on inverting a matrix. However, the statistical curse of dimensionality, a feature of the $K$ to $T$ ratio, remains a difficulty to overcome.

There are two extreme ways of reducing dimensionality: sparse or dense. The former selects a small number of features out of the large pool in a supervised way (e.g. LASSO), the latter compresses the data in a set of latent factors that span (hopefully) most of the $S_t$ space. This is often seen as a necessity to choose *one* of them.[15] However, in a regularized model, both can be included, and we can let the algorithm select an optimal combination of original features and factors. To appreciate this point, let us look at a linear model. Suppose we have $S_t = [X_t \;\; F_t]$ and by construction the factors are some linear combination of original features ($F_t = X_t R$). We can estimate

$$y_{t+1} = X_t\beta + X_t R\gamma + u_t$$

using LASSO. Of course, this would not run with OLS because of perfect collinearity, which is the standard motivation for not mixing dense and sparse approaches. By Frisch-Waugh-Lowell (FWL) theorem and using the factor model

$$X_t = \Lambda F_t + e_t$$

the above is equivalent to

$$y_{t+1} = e_t\beta + F_t\gamma + u_t.$$

At first sight, this has more parameters than either the dense or sparse approach. However, with some proper penalization of $\beta$ and $\gamma$, the model can balance a proper mix of dense and sparse. For instance, activating some $\beta$'s "corrects" the overall prediction when the factor model representation is too restrictive for the effect of a specific regressor $X_k$ on $y_{t+1}$.[16] This representation has been studied in Hahn et al. (2013) and Hansen and Liao (2019) to enhance hard-thresholding methods' performance (like LASSO) in the presence of highly correlated regressors. When it comes to RF, this suggest both the original data and its rotation can be included in $F_t$. This also suggest it is relatively costless to explore alternative rotation possibilities.

In time series, the number of potential predictors is always large given that not only $X_t$ may

[15] In recent macro forecasting work using RF, Goulet Coulombe et al. (2019) follow a dense approach by only including factors in the regression while Borup et al. (2020) put their money on sparsity by proposing a Lasso pre-processing step using the raw data.

[16] That problem has been documented in Bai and Ng (2008) and others.

be of interest, but also its (potentially numerous) lagged values. The dense approach amounts to extract factors out of the cross section of available data, generate lags of them and use the resulting matrix as the input in RF. Since the number of factors is usually small, including a certain number of lags of each of them is not problematic. Another approach is to use the raw data and lags of it, then let RF select by itself the relevant features. While RF will not overfit given bagging and suitable tuning parameters, it can underfit if (i) the data set is too compressed (factors) or if (ii) it is not compressed enough. As a result, a purely dense or sparse approach (as those above) could easily be suboptimal to a more nuanced solution.

From a purely predictive standpoint, autocorrelation of residuals means leaving forecasting power on the table. To get rid of it in a VAR, one may need to include many lags. Given the symmetric nature of the VAR, that can quickly lead to overfitting. A standard solution is to resort to Bayesian estimation and use priors in the line of Doan et al. (1984), which is specially designed for the VAR structure – blocks of lags in particular. Outside of the VAR paradigm, there is a whole strand of older literature that seeks to estimate restricted lag polynomials in Autoregressive Distributed Lags (ARDL) models (Almon (1965), Shiller (1973)). More recently, these methods have found new applications in mixed-frequency models (Ghysels et al. (2007)) where the very design of the model leads to an explosion of parameters and a need for regularization. RF experiences an analogous situation. A tree may waste many splits trying to efficiently extract information out of a lag polynomial: for instance, splitting on the first lag, then the 7th one, then the 3rd one. In linear parametric models, the above methods can extract the relevant information out of a lag polynomial without sacrificing too many degrees of freedom. A significant roadblock to this enterprise in the RF paradigm is that there are no lag polynomials of coefficients to penalize.

**Moving Average Factors**. To extract the essential information out of the lag polynomial of a specific variable, a linear transformation of regressors can do the job. Consider forming a panel of $P$ lags of variable $j$:

$$X_{t,j}^{1:P} \equiv [X_{t-1,j} \; ... \; X_{t-P,j}] \; .$$

We want to form weighted averages of the $P$ lags so that it summarizes most efficiently the temporal information of the feature indexed by $j$.[17] The weighted averages with that property will be the first factors extracted by PCA on $X_{t,j}^{1:P}$. This can be seen as the time-dimension analog to the traditional cross-sectional factors. The latter are defined such as to maximize their capacity to replicate the cross-sectional distribution of $X_{t,j}$ fixing $t$ while the Moving Average Factors (MAFs) proposed here seek to represent the temporal distribution of $X_{t,j}$ for a fixed $j$ in a lower dimensional space.[18] By doing so, our goal to summarize the information of $X_{t,j}^{1:P}$ without

[17] $P$ is a tuning parameter the same way the set of included variables in a standard factor model is one.

[18] In the spirit of the Minnesota prior, one can assign decaying (in $p$) weights to each lag before running PCA.

modifying the RF algorithm is achieved: rather than using the numerous lags as regressors, we can use the MAFs which compress information ex-ante. As it is the case for standard factors, MAF are designed to maximize the explained variance in $X_{t,j}^{1:P}$, not the fit of the final target. It is the job of the RF part to select the relevant linear combinations in the resulting $S_t$ to maximize the fit.

The take-away from this subsection can be summarized in three key points. First, there is no need to choose ex-ante between sparse and dense when the model performs selection/regularization. We can let the algorithm find the optimal balance. Second, to make the inclusion of many lags useful, we need to regularize the lag polynomial. Third, such compression can be achieved without regularizing the polynomial's parameter directly. The proposed MAFs can be easily created ex-ante and fed in MRF (or any model). This is especially convenient for RF (and any non-parametric method) because there is no explicit lag polynomial of parameters to penalize. Finally, MAFs also further facilitate interpretation. As these are moderately sophisticated averages of a single time-series, they can be viewed as a smooth index for a specific (but tangible) economic indicator. This is arguably much easier to interpret than a plethora of lags coefficients.

## 2.7 Quantifying Uncertainty of $\beta_t$'s Estimates

Taddy et al. (2015) and Taddy et al. (2016) interpret the forest as the posterior distribution on the tree functional $\mathcal{T}$ which is obtained by a Bayesian bootstrap.[19] Their view of $\mathcal{T}$ as a Bayesian non-parametric statistic (independently of the DGP) is of even greater interest in the case of MRF.[20] It could provide inference for meaningful time-varying parameters $\beta_t$ rather than an opaque conditional mean function. Such techniques, that originate from Ferguson (1973), have seldomly found applications in econometrics, such as Chamberlain and Imbens (2003) for instrumental variable and quantile regressions. However, while the Bayesian Bootstrap desirably does not assume many things about the underlying data, it yet makes the assumption that $Z_t = [y_t \ \ X_t \ \ S_t]$ is an *iid* random variable. Thus, it cannot be used directly as a proper theoretical motivation for using the bag of trees directly to conduct inference. Additionally, neglecting this dependence could weaken the claim that MRF will not overfit. Fortunately, the natural extension that is a Block Bayesian Bootstrap (BBB) makes Taddy et al. (2015)'s convenient ap-

This has the analogous effect of shrinking more heavily the distant lags and less so the recent ones.

[19] The connection between Breiman (1996)'s bagging and Rubin (1981)'s Bayesian Bootstrap was acknowledged earlier in Clyde and Lee (2001).

[20] An alternative (frequentist) inferential approach is that of Friedberg et al. (2018). However, their asymptotic argument requires estimating the linear coefficients and the kernel weights on two different sub-samples. This is hard to reconciliate with our goal of modeling time-variation and different regimes throughout the entire sample. Furthermore, when the sample size is small, splitting the sample in such a way carries binding limitations on the complexity of the estimated function.

proach amenable to this paper's setup.

## 2.8 Block Bayesian Bootstrap

BBB consists of a conceptual workaround to reconcile time series data with multinomial sampling. I first briefly review the standard Bayesian Bootstrap. Let all the available data be cast in the matrix $Z_t = [y_t \;\; X_t \;\; S_t]$. $Z$ is considered as a discrete *iid* random variable with $T$ support points. Define $N_t = \sum_{\tau=1}^{T} I\left(Z_\tau = z_t\right)$, which is the number of occurrences of $z_t$ in the sample. To make inference on $\beta_t = \mathcal{T}(\boldsymbol{\theta})$ by considering the latter as a posterior functional, we need to characterize the posterior distribution of $\theta$

$$\pi(\theta|\mathbf{z}) = \frac{f(\mathbf{z}|\theta)\pi(\theta)}{\int f(\mathbf{z}|\theta)\pi(\theta)d\theta}.$$

Conditional on $\theta$, the likelihood of the data is multinomial. The prior is Dirichlet. Since Dirichlet is the conjugate prior of the multinomial distribution, the posterior is also Dirichlet. That is, it can be shown that combining the likelihood

$$f(\mathbf{z}|\theta) = \frac{N!}{N_1! \cdots N_T!} \prod_{t=1}^{T} \theta_t^{N_t}$$

with prior distribution

$$\pi(\theta) = \frac{1}{B(\alpha_{1:T})} \prod_{t=1}^{T} \theta_t^{N_t+\alpha_t-1}$$

gives rise to the posterior distribution

$$\pi(\theta|\mathbf{z}) = \frac{1}{B(\bar{\alpha}_{1:T})} \prod_{t=1}^{T} \theta_t^{N_t+\alpha_t-1} \quad .$$

where $\bar{\alpha}_t = \alpha_t + N_t$ and $B(\bar{\alpha}_{1:T}) = \frac{\prod_{t=1}^{T} \Gamma(\bar{\alpha}_t)}{\Gamma\left(\sum_{t=1}^{T} \bar{\alpha}_t\right)}$. Using the uninformative (and improper) prior $\alpha_t = 0 \;\forall t$, we can simulate draws from the (proper) posterior using $\theta_t \sim \text{Exp}(1)$. The object of scientific interest is typically not $\theta$ *per se* but rather a functional of it In Taddy et al. (2015), the functional of interest is a tree and inference is obtained by computing $\mathcal{T}(\theta_{1:T})$ for each $\theta_{1:T}$ draw.

BBB is a simple re-definition of $Z$ so that it is plausibly *iid*. Hence, in the spirit of traditional frequentist block bootstrap (MacKinnon (2006)), blocks of a well-chosen size will be exchangeable. Thus, a new variable can be defined $Z_{\mathfrak{b}} \equiv [y_{\underline{\mathfrak{b}}:\bar{\mathfrak{b}}} \;\; X_{\underline{\mathfrak{b}}:\bar{\mathfrak{b}}} \;\; S_{\underline{\mathfrak{b}}:\bar{\mathfrak{b}}}]$. There will be a total of $\mathfrak{B} = {}^{T}/_{\text{block size}}$ fixed and non-overlapping blocks. Under covariance stationarity, $\tilde{Z}_{\mathfrak{b}} = vec(Z_{\mathfrak{b}})$

are *iid*, for a properly chosen block length.[21] The derivations above can be carried by replacing $t$ by $\mathfrak{b}$ and $T$ by $\mathfrak{B}$. Practically, this implies drawing $\theta_{\mathfrak{b}} \sim \text{Exp}(1)$ which means observations within the same block ($\underline{\mathfrak{b}} : \bar{\mathfrak{b}}$) share the same weight. As an alternative to this BBB that would also be valid under dependent data, Cirillo and Muliere (2013) provide a more sophisticated urn-based approach with theoretical guarantees. It turns out their approach contains the well-known non-overlapping block bootstrap as a special case, which the above is only its Bayesian rendition.

### 2.8.1 About Heteroscedasticity and Serial Correlation

It is reasonable to wonder whether plain sub-sampling or Dirichlet weighting procedures become ill-suited in the presence of heteroscedasticity. Fortunately, the non-parametric bootstrap/sub-sampling that RF uses is in fact the "pairs" bootstrap of Freedman et al. (1981) which is valid under general forms of heteroscedasticity (MacKinnon (2006)).[22] From a Bayesian point of view, Lancaster (2003) show that the obtained variance for OLS from using such a bootstrap is asymptotically equivalent to that of White's heteroscedasticity-robust sandwich formula. Poirier (2011) propose better priors and Karabatsos (2016) incorporate such ideas into a generalized ridge regression. Hence, in the spirit of heteroscedasticity-robust estimation, no attempt will be made at directly modeling stochastic volatility (which is a GLS approach) but it will rather be reflected in larger bands for periods of high volatility.

BBB, by construction, is already the "hammer" solution to serial dependence in residuals. However, it is worth noting that for most applications, that problem is already taken care of by the very specification of ARRF and MRFs in general: *it can include lags of the dependent variable*. Augmenting a specification with more lags is the traditional fix for autocorrelation in time series. The equivalence has been known at least since Cochrane and Orcutt (1949) and it is now the standard view that autocorrelation being wiped out by including lags of $y_t$ is a *condicio sine qua non* a well-specified model. Hence, ARRF and more generally MRFs, by putting the accent on modeling explicitly autoregressive behavior, are immune to the usual concern about the use of plain bootstrap aggregation on time series data. This is much less likely to be true for plain RF. For the *direct* forecasting of variables at horizon higher than $h = 1$, the presence of autocorrelation cannot be ruled out. In those cases – which include time-varying Jordà (2005) local projections, the block approach is a necessity.

[21] In practice, I will use block of two years for both quarterly or monthly data.

[22] For a purely prediction point of view, Grandvalet (2004) also stresses the point that bagging provides important improvements when there is "badness" in the data, that is, the presence of uninformative leverage points. Those improvements are shown to be especially meaningful for unstable algorithms such as regression trees.

# 3 Simulations

Simulations are divided in two parts. The first (and main) part aims at showing that Autoregressive Random Forest (ARRF) will show forecasting improvements over standard non-linear time series model when the true DGP mixed both endogenous and exogenous time-variation. This will help at rationalizing the real forecasting results that will show ARRF supplanting the $\sim$TAR family of models for the great majority of targets. The second part looks at slightly simpler DGPs and focus on $\beta_t$ itself and its credible bands. The main point is to visually show that (i) GTVPs adapts nicely to a wide range of DGPs and (ii) are not prone to crying wolf on time-variation.

## 3.1 Setup

In short, I consider 6 DGPs: Autoregression (AR), two Self-Exciting Threshold ARs (SETAR), SETAR with a structural break, AR with a structural break and finally a SETAR model that switches once (a break) to be AR. The point of choosing these DGPs is to span the space of time-variations I wish to investigate: endogenous, exogenous and both together. I compare the RMSPE of these models out-of-sample with respect to an oracle AR(2) – a model that knows the true law of motion $\beta_t$. The simulated series sample size is either $T = 150$ or $T = 300$. The last 40 observations of each sample consist the hold-out sample for evaluation. I forecast 4 different horizons: $h = 1, 2, 3, 4$. Models are estimated once at the last available data points.

### 3.1.1 Models

First, these simulations will depict how ARRF does better very quickly when the DGP slightly departs from what $\sim$TAR assumes. That is, when the switching/threshold variable is not $y_{t-1}$. Of course, $\sim$TAR model's in general can handle any threshold variable one may wish to use. However, as discussed in section 2.1, choosing the right threshold variable out of many is computationally prohibitive. Furthermore, the architecture of the model may be much more complicated that of a single threshold. In the latter case, $\sim$TAR are simply misspecified. Second, it will also document the failings of RF when the AR part is pervasive, which was the main concern expressed in section 2.5.1. That is, I want to show the empirical superiority of ARRF over RF when there is truly an AR part.

The DGPs will include two types of switching variable: $y_{t-1}$ and $t$. Since a structural break is just a threshold effect with respect to variable $t$, I can conclude without loss of generality that similar results would be obtained having any other (additional) switching variable. In all simulations, $S_t$ includes 8 lags of $y_t$ and a time trend, which match what will be referred to in the empirical section as "Tiny ARRF" (see section 4.1.5).

Models include SETAR, Rolling-Window (RW) AR, Random Forest (RF) and Autoregressive Random Forest (ARRF). Iterated SETAR forecasts are obtained via the standard bootstrap method (Clements and Smith (1997)) and all the others are generated via direct forecasting. That is, in the latter case, I fit the model directly on $y_{t+h}$ rather than iterating forward the one-step ahead forecast. To display that the observed differences between SETAR and other models is not merely due to the choice of iterated vs direct forecasts (which is a non-trivial choice in many environments (Chevillon (2007))), I additionally include SETAR-d where "d" means those forecasts have rather been obtained by direct forecasting.

### 3.1.2 Performance Metric

The statistic I report is the root Mean Square Prediction Error (MSPE). In simulation $s$, for the forecasted value at time $t$ made $h$ ahead, I compute

$$RMSPE_{s,h,m} = \sqrt{\frac{1}{40 \times S} \sum_{s=1}^{S} \sum_{t \in OOS} (y_t^s - \hat{y}_{t-h}^{s,h,m})^2}.$$

I consider $S = 100$ which means the total number of squared errors being averaged for a given horizon and model is 100*40=4000. The bar plots report a $RMSPE_{s,h,m}$ that is relative to the oracle. Specifically, the reported measure is

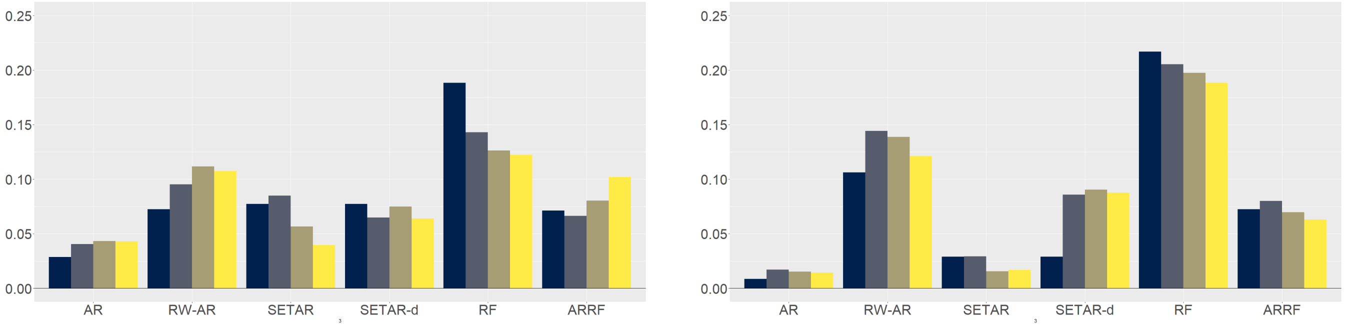
$$\Delta_o \overline{RMSPE}_{1:S,h,m} = \frac{\overline{RMSPE}_{1:S,h,m}}{\overline{RMSPE}_{1:S,h,o}} - 1.$$

It is worth specifying what do I mean by "oracle". The so-called oracle knows perfectly the law of motion of time-varying parameters $\beta_t$. If the model has a break and a switching variable, it knows exactly the break points, thresholds and AR parameter values in each regime. However, it does not know the future shocks ($\epsilon_{t+h}$) and neither the future evolution of parameters ($\beta_{t+h}$) unless the latter is purely deterministic.

## 3.2 DGPs in detail and Results

The ∼TAR packages in `R` provide functions to simulate from these models. I used them and combine them in multiple ways to obtain a reasonable range of DGPs. For all DGPs, $X_t = [1 \;\; y_{t-1} \;\; y_{t-2}]$.

$$\begin{aligned} y_t &= X_t \beta + \epsilon_t, \quad \epsilon_t \sim N(0, 0.25^2) \\ \beta &= [0.7 \quad -0.2] \end{aligned}$$



(a) 150 obs

(b) 300 obs

Figure 1: DGP 1: Plain AR(2). This figure displays increases in relative RMSPE with respect to the oracle. Four horizons are reported, from 1 (dark blue) to 4 (yellow).

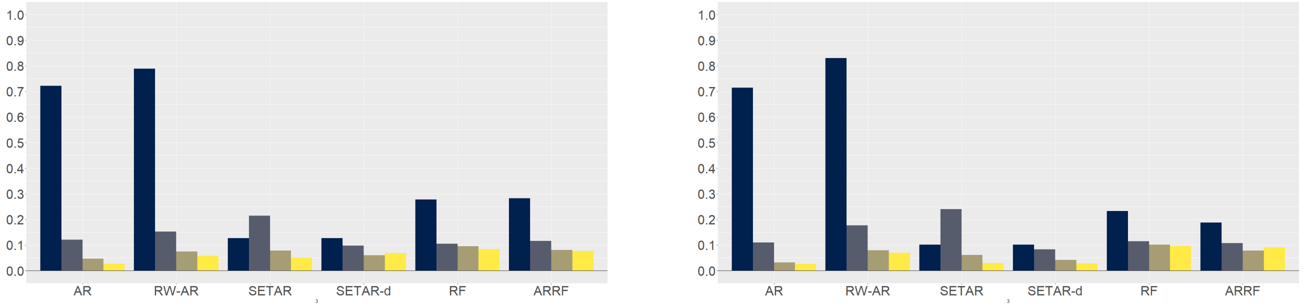
**DGP 1: Plain AR(2).** Given the incredible resilience of AR models in any macroeconomic forecasting exercise, the first DGP being considered is an autoregressive process of order 2. As it should be, the best model is the AR for all horizons and all sample sizes. The RW-AR suffers from high variance and it is assumed that tuning the window length in a data-driven way would help, but that is not the point here. Plain RF struggles, irrespective of the sample size.[23] For the smaller sample, ARRF performs as well as the tightly parametrized SETARs. Their marginal increase in RMSPE with respect to the oracle are typically less than 10%, which is small in contrast to simulations yet to come. More observations generally helps AR, the iterated SETAR and ARRF especially at longer horizons.

### 3.2.1 Endogenous TV

I now turn to consider example where parameters vary endogenously according to previous values of $y$ itself.

$$y_t = X_t\beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, 0.5^2)$$
$$\beta_t = \begin{cases} [2 \;\; 0.8 \;\; -0.2], & \text{if } y_{t-1} \geq 1 \\ [0 \;\; 0.4 \;\; -0.2], & \text{otherwise} \end{cases}$$

[23] This will be a recurring theme. If the DGP is linear, RF never ever perform well. This strength of this finding is only magnified when the true $X_t$ dimension grows, which goes in line with the discussion in section 2.5.1.



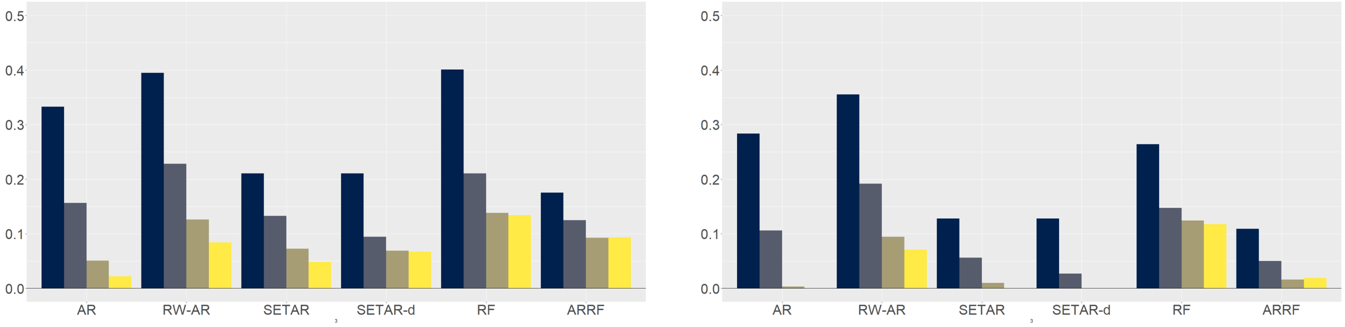
(a) 150 obs

(b) 300 obs

Figure 2: DGP 2: SETAR. This figure displays increases in relative RMSPE with respect to the oracle. Four horizons are reported, from 1 (dark blue) to 4 (yellow).

**DGP 2: SETAR.** This DGP represent what one could think of as an endogenous switching process for real activity variables: it includes high and low regimes and mildly different dynamics in each of them. In this first SETAR example, AR models are doing badly by not capturing the change in mean and dynamics. In this DGP, predictive power quickly vanishes after $h = 1$ which is why we observe little performance heterogeneity at longer horizons and results close to that of the oracle. Specifically tailored for this class of DGPs, the two SETARs are offering the best performance. A less trivial observation is that both ARRF and RF, while much more general, performs only marginally worse than SETARs. The tie between ARRF and RF is attributable the importance of the switching constant in the current DGP, which both models allow for.

$$
\begin{aligned}
y_t &= X_t \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, 0.5^2) \\
\beta_t &= \begin{cases} [2 \;\; 0.8 \;\; -0.2], & \text{if } y_{t-1} \geq 0 \\ [0.25 \;\; 1.1 \;\; -0.4], & \text{otherwise} \end{cases}
\end{aligned}
$$



(a) 150 obs

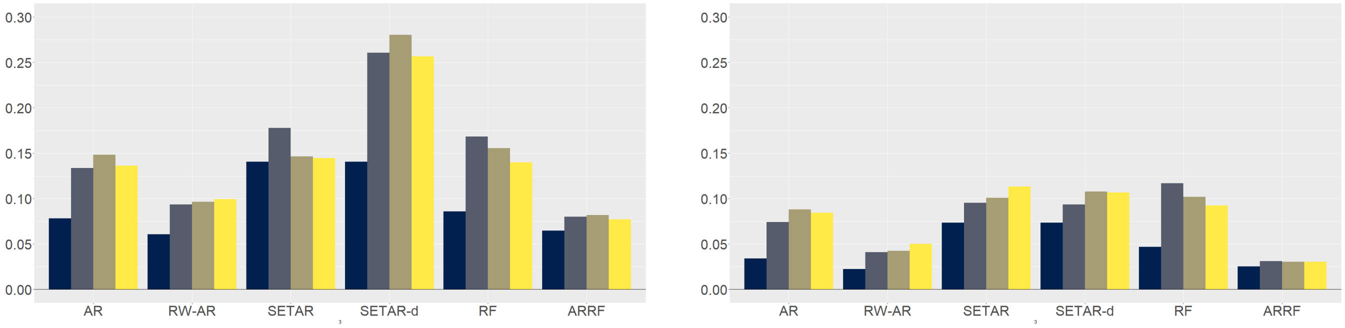
(b) 300 obs

Figure 3: DGP 3: More Persistent SETAR. This figure displays increases in relative RMSPE with respect to the oracle. Four horizons are reported, from 1 (dark blue) to 4 (yellow).

**DGP 3: More Persistent SETAR.** The increased persistence makes results at higher horizons of greater interest. In the previous SETAR, the forecasting ability of the oracle was practically null beyond $h = 2$. For all horizons and sample sizes considered, ARRF is practically as good as SETAR, the optimal model in this context. With the increased importance of changing dynamics relative to that of a changing mean, RF is now trailing behind with RW-AR. The former nevertheless improves substantially at shorter horizons when the sample size increase. AR is resilient at longer horizons but is much worse than ARRF and SETAR at shorter ones.

### 3.2.2 Exogenous TV

$$y_t = X_t\beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, 0.3^2)$$
$$\beta_t = \begin{cases} [0 \ \ 0.7 \ \ -0.35], & \text{if } t < T/2 \\ [0.15 \ \ 0.6 \ \ 0], & \text{otherwise} \end{cases}$$



(a) 150 obs

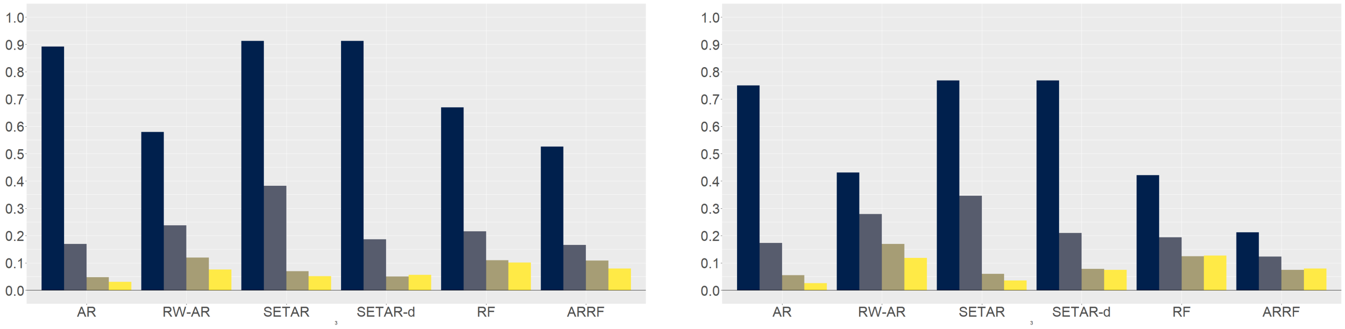
(b) 300 obs

Figure 4: DGP 4: AR(2) with break. This figure displays increases in relative RMSPE with respect to the oracle. Four horizons are reported, from 1 (dark blue) to 4 (yellow).

**DGP 4: AR(2) WITH A BREAK IN DYNAMICS AND MEAN.** In this setup, RW-AR is expected to have an edge, with the estimation window excluding pre-break data. At horizon 1, both RW-AR and ARRF are the best model, beating the robust AR by a thin margin. For $h > 1$, ARRF emerges as the best model at both 150 and 300 sample sizes. RW-AR is naturally always close behind.[24] As expected, the two models are better than the remaining alternatives by allowing for exogenous structural change (which SETARs and AR do not) and explicitly modeling the autoregressive part (which RF does not).

### 3.2.3 Exogenous & Endogenous TV

$$\text{DGP 4} = \begin{cases} \text{DGP 2}, & \text{if } t < T/2 \\ \text{DGP 3}, & \text{otherwise} \end{cases}$$



(a) 150 obs

(b) 300 obs

Figure 5: DGP 5: SETAR with a structural break. This figure displays increases in relative RMSPE with respect to the oracle. Four horizons are reported, from 1 (dark blue) to 4 (yellow).

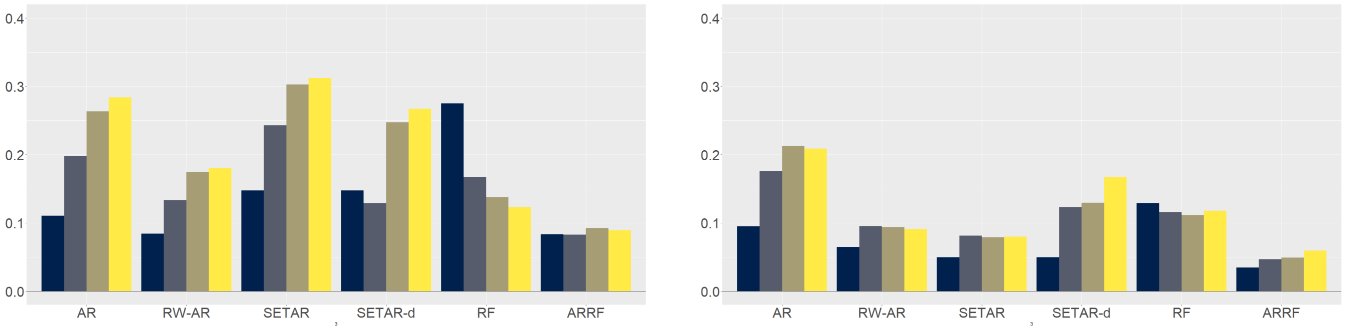
**DGP 5: SETAR WITH A STRUCTURAL BREAK.** Until now, I have focused on dynamics that can be captured successfully by currently available time series models. The point of this paper is that most of these models may suffer from serious misspecification issues when estimated on real data. Hence, I introduce here what is possibly the simplest example where structural breaks and switching interacts.[25] SETARs are expected to fail because they are not designed to catch breaks. RW-AR is also expected to fail because it does not model switching. RF can work but is anticipated to be inefficient and thus unreliable in samples of small and medium size. All these heuristics arguments are verified in Figure 5: ARRF is the better model followed closely

[24] Although not reported here, I considered a simple linear model where I search for a single break (in time) and use the data after the break for forecasting. This option does as well as ARRF for this particular DGP.

[25] Without loss of generality, the threshold according to $t$ could be replaced by a threshold according to any other variable.

by RW-AR and RF for short horizons. With 300 observations, the lead of ARRF and the second position of RF are strengthened. At longer horizons, all models perform poorly (including the oracle) due to the fundamental unpredictability of the law of motion for $\beta_t$. For these horizons, misspecification only plays a minor role in total forecast error variance, explaining the small and homogeneous decrease in performance with respect to the oracle at longer horizons.

$$\text{DGP 6} = \begin{cases} \text{DGP 2}, & \text{if } t < T/2 \\ \text{DGP 1}, & \text{otherwise} \end{cases}$$



(a) 150 obs

(b) 300 obs

Figure 6: DGP 6: SETAR that morphs in AR(2). This figure displays increases in relative RMSPE with respect to the oracle. Four horizons are reported, from 1 (dark blue) to 4 (yellow).

**DGP 6: SETAR THAT MORPHS INSTANTLY IN AR(2).** The goal of this last DGP is to gain consider a mixture of endogenous and exogenous time-variation, but with more interesting results at longer horizons. Further, this last DGP can rightfully be hypothesized for some economic time series: complex dynamics up until the mid-1980's followed by a very simple autoregressive structure during the Great Moderation. ARRF comes out as the best model for all horizons in the smaller sample. For horizon 1, RW-AR does equally well, which is expected in this DGP. With respect to the plain exogenous time-variation scenario in Figure 4, both SETAR and RF's performance has further deteriorated in the smaller sample size.

**ABOUT MISSPECIFICATION IN ARRF.** Most of the reported gains from using ARRF come from its capacity to avoid being misspecified when a more complex DGP arise. What happens if the arbitrary linear part in ARRF, $X_t$ is itself misspecified? Figure 24 in the appendix report corresponding results. For all DGPs under consideration in this section, an ARRF where $X_t$ has been replaced with two white noise series (instead of the first two lags) and it performs similarly well (or bad) as RF.[26]

**Summary.** In short, this set of simulations display what one would be expecting to find. First, when the true DGP is that of the tightly parametrized classical non-linear time series model, the latter usually perform better than a more flexible approach. Second, when it is not the case, the more flexible ARRF does better. Third, when there are pervasive linear autoregressive relationships, plain RF struggles. Fourth, ARRF and RF relative performance both increase with the number of observations but ARRF's one increases faster if the linear part is well-chosen. The balance of all these forces determines which model will be successful and which one will not. The nature of that balance is an empirical question.

### 3.2.4 A Look at GTVPs when $S_t$ is Large

A notable difference between the simulations presented up to now and the applied work being carried in later sections is the size of $S_t$. In many macro applications, there is no shortage of variables to include in the tree part of MRF. For instance, the FRED-QD data base contains more than 200 potential predictors, which is much more than just lags of $y$ and a time trend. RF will benefit from more data through increased randomization in the creation of trees, the latter preventing overfitting.

The additional simulations go as follow. First, I simplify the analysis by looking at a static model with mutually orthogonal but autocorrelated regressors $X_1$ and $X_2$, both driving $y_t$ according to some process. I simulate each of them for 1000 periods and estimate the models over the first quarter of it. The remaining 750 observations are used to evaluate the out-of-sample performance. The signal to noise ratio is calibrated to 2/3 which is about what is found (out-of-sample) for most models in the empirical section.

The only remaining questions are that of the constitution of $S_t$ and the generation of $\beta_t$'s. I create two autocorrelated (but not cross-correlated) factors. Out of each of them, I create 50 series with a varying amount of additional white noise.[27] Adding an exogenous time trend and the lags of $y$, the size of $S_t$ is slightly above 100. Finally, $\beta_t$'s are functions of the underlying *first* factor which (like the second) is not directly included in the data set. In certain DGPs, some $\beta_t$'s will also be a pure function of $t$ (like random walks, structural breaks).[28] Table 1 summarizes the six DGPs in words. More illustratively, Figure 25 plots one example of each DGPs as well as

[26] This result may not hold, however, when the law of motion for $\beta_t$ is highly complex and requires a great number of split (unlike what is considered here). The reason for this is that the size of linear part restricts the potential depth of the tree, especially if the number observations is small. In practice, it is a safer bet to use a small linear part if uncertainty around its composition is high. More on this and the effect of hyperparameters can found in section A.3.

[27] To be precise, their standard deviation is $U[0.5, 3]$ % that of the original factor standard deviation.

[28] To clarify, the second factor and underlying series are completely useless to the true DGP – arguably mimicking the inevitable when using a data base of the size of FRED-QD.

Table 1: Summary of Data-Rich Simulations DGPs

| DGP # | Constant | $\beta_t^{X_1}$ | $\beta_t^{X_2}$ | Residuals Variance |
|---|---|---|---|---|
| 1 | Switching | Switching | Switching | Flat |
| 2 | Flat | Switching | Slow Change (function of $t$) | Flat |
| 3 | Flat | Switching | Structural Break | Flat |
| 4 | Flat | Latent factor directly | Slow Change (function of $t$) | Flat |
| 5 | Flat | Random Walk | Random Walk | Flat |
| 6 | Flat | Flat | Flat | Stochastic Volatility |

the estimated GTVPs and their credible region (as discussed in section 2.7). It is visually shown that GTVPs are adaptive in the sense that it can discover which kind of time-variation is present in the data while estimating it.

I consider the out-of-sample performance of MRF in this context comparing it to OLS, Rolling-Window OLS (RW-OLS) and plain RF. Figure 26 report the corresponding forecasting performances of those models in terms of RMPSE differentials with respect the oracle (the forecast that knows the $\beta_t$'s law of motion). As expected, MRF outperforms all alternatives by a wide margins for most DGPs. By construction, for DGP 5 (random walks) and DGP 6 (constant parameters), RW-OLS and OLS also perform well. Nevertheless, it is reassuring to see that MRF either performs much better than OLS or worse by a thin margin (in cases with no time-variation).

# 4 Forecasting

## 4.1 Setup

This subsection presents the data and the design of the pseudo-out-of-sample forecasting experiment.

### 4.1.1 Data, Transformations and Forecasting Targets

In this section, I present results for quarterly frequency using the dataset FRED-QD, publicly available at the Federal Reserve of St-Louis's web site. It contains 248 US macroeconomic and financial aggregates observed from 1960Q1. The series transformations to induce stationarity are indicated in McCracken and Ng (2020). The variables of interest are: real GDP, Unemployment rate (UR), Consumer Price Index (CPI, which will be transformed and called INF), difference between 10-year Treasury Constant Maturity rate and Federal funds rate (SPREAD), housing starts (HOUST) and 1-Year Treasury Constant Maturity Rate (IR). Forecasting horizons are 1, 2, 4, 6 and 8 quarters. These are representative macroeconomic indicators of the US economy which is based on Goulet Coulombe et al. (2019) exercise for many ML models, itself based on **??** and a whole literature of extensive horse races in the spirit of Stock and Watson (1998b). The unemployment rate is considered $I(1)$ and I target the first difference without logs. The spread is $I(0)$. CPI, IR and HOUST are considered $I(1)$.

### 4.1.2 Pseudo-Out-of-Sample Experiment Design

The POOS period starts in 2003Q1 and ends 2014Q4. I use expanding window estimation from 1961Q3. I use direct forecasts except for the ∼TAR models where the iterated forecasts are computed.[29] ∼TAR forecasts are calculated using the block-bootstrap method which is standard in the literature. hyperparameters are optimized and the model re-estimated every 8 quarters.

### 4.1.3 Forecast Evaluation Metrics

Following a standard practice in the forecasting literature, I evaluate the quality of our point forecasts using the root Mean Square Prediction Error (MSPE). For the forecasted value at time $t$ of variable $v$ made $h$ steps before, I compute

$$RMSPE_{v,h,m} = \sqrt{\frac{1}{\#OOS} \sum_{t \in OOS} (y_t^v - \hat{y}_{t-h}^{v,h,m})^2}.$$

The standard Diebold-Mariano (DM) test procedure is used to compare the predictive accuracy of each model against the reference (AR(4)) model. Alternative loss functions could be considered $RMSPE$ is the most natural and only consistent one given that all models are trained to minimize the squared loss in-sample.

### 4.1.4 Exact Composition of $S_t$

It has been argued in section 2.6 that feature engineering matters crucially when the number of raw features clearly exceeds the sample size. The composition of $S_t$, the set of variables from which the RF can select, reflects that. It considers both cross-sectional and time factors that both at compressing information along their respective dimensions. The exact composition of $S_t$ is

1. 8 lags of $y_t$;
2. $t$ for structural breaks/exogenous time-variation;
3. 2 lags of all variables in FRED, call them $Z$;

[29] Calculating univariate iterated would imply fixing the parameters at the last value (since we are not forecasting all the elements of $S_t$) The direct forecast does it implicitly.

4. $F$ to summarize the cross-section variation: 8 lags of 5 factors extracted from $Z$ by PCA;
5. for each variable $Z_j$, I generate two $M_{t,j}$ moving-average factors that summarize the information contained in its distributed lags. Done by PCA on 8 lags.

I also report data-poor versions that I call "tiny" RF and ARRF. These have $S_t^-$ which is composed of 1 and 2 only. This could be seen as a slightly more general SETAR model that (i) consider more than one lag of $y_t$ as a potential switching variable and (ii) also accommodate for the possibility of structural breaks.

### 4.1.5 Models Considered

I consider many models for the main quarterly results. To better understand where the gains from MRF are coming from, I include models that use different subsets of ideas developed in earlier sections. I include both data-rich models (as they tend to be more competitive) and classical non-linear time series models since they share an obvious familiarity with ARRF. First, here is the batch of relatively standard models.

- **AR(4)**: the benchmark
- **ARDI**: 2 lags of 3 factors extracted by PCA, 4 lags of $y$, shown to in a vast array of papers to be hard to beat.
- **Ridge**: regressors are $S_t$, to distinguish gains obtained specifically from using $S_t$ (that includes MAFs) in more standard (linear) regressions that can handle high-dimensionality.
- **LASSO**: regressors are $S_t$, idem.
- **RW-AR**: AR(4) estimated on a Rolling-window of 10 years
- **TV-AR**: AR(4) with random walk TVPs
- $\sim$**TAR**: All of them with threshold variable is $y_{t-1}$, as benchmark classical non-linear time series models. 4 lags in each regimes.
- **RF**: using 8 lags of the full raw data set.
- **Tiny RF**: using $S_t^-$. The point of including this model is to gauge the pertinence of the data-rich environment when it comes to plain RF – to see how it fares with an information set comparable to that of $\sim$TARs.

Finally, the following is the list of MRFs considered in the exercise.

- **ARRF**: AR(2) MRF using $S_t$, already thoroughly discussed.
- **Tiny ARRF**: same as above but using $S_t^-$. This is again to gauge the pertinence of more data.
- **ARDIRF**: small ARDI MRF: two factors, 1 lag each, + 2 lags of $y$. As discussed in McCracken and Ng (2020), the first factor mostly loads on real activity variables while the second is a composite of forward-looking indicators like term spreads, permits and inventories.[30]

[30] The choice of including only two factors is in the interest of parsimony, especially that we are fitting one model in every single leaf.McCracken and Ng (2020) note that the first two factors account for 30% of the variation in the data while adding two more only bumps it up to 41%, making the last two presumably more disposable in this specific context.

- **VARRF**: small VAR: 1 lag of 3 variables (GDP, inflation, interest rate) + 2 lags of $y$.
- **RF-MAF**: plain RF using $S_t$, to quantify the effect of mixing MAFs with traditional RF.
- **AR+RF**: A restricted version of ARRF obtained by first running AR(4) and then fitting plain RF on residuals (this constraints the linear dynamics to be time-invariant).

## 4.2 Main Results

The violin plots aim at summarizing the information contained in large appendix tables. These contain the RMSPE for each horizon and variables with respect to a benchmark model. Hence, I report the distribution of $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$ for horizons $h = 1, 2, 3$.[31]

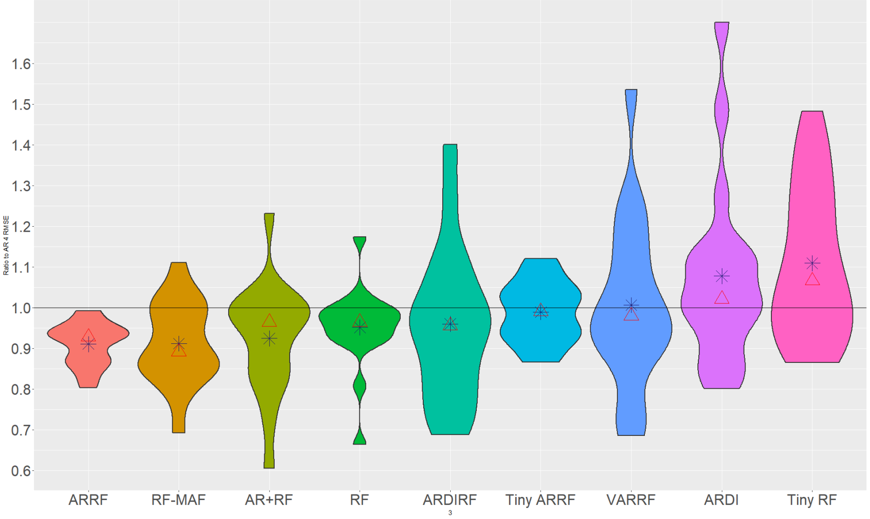


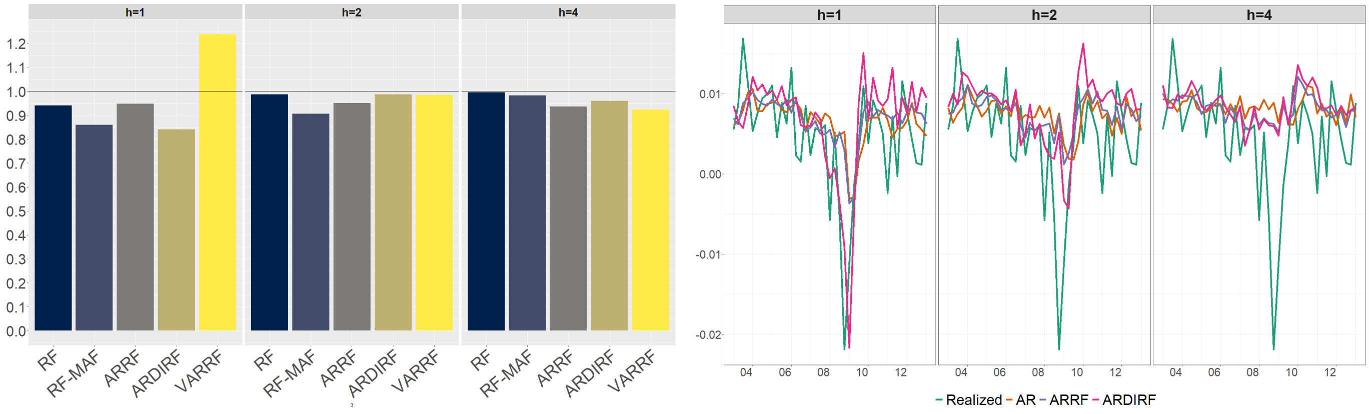
Figure 7: The distribution of $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$. The star is the mean and the triangle is the median.

This informs us about which model is the best overall. Of course, it does not mean it has to be the best model for every $h$ and $v$, but it rather means that on average, for any target considered, it performs better. First, we see that almost any form of RF is better than ARDI (linear factor model with PCA). The best RF is ARRF which has noticeably a very small mass above 1, suggesting it almost never does worse than AR(4), which is a somewhat uncommon property for a non-linear time series model. It is followed closely by RF-MAF, a simplification of ARRF that keeps $S_t$ but rather runs a plain RF with it. Another simplification, AR+RF, ranks third with a density that suggests its performance is much more spread out. Fourth, we have the plain RF that does not use $S_t$, with much more conservative gains over the benchmark. The importance of $S_t$ is further demonstrated in appendix A.2 by comparing workhorse models that handle high-dimensional data nicely (plain versions of RF, LASSO, Ridge) with different

[31] The remaining longer horizons results are reported in appendix tables. However, for the graphical analysis, I focus on horizons up to a year since the short run is the usual playground for statistical models.

information sets. It is shown to be particularly useful for tree-based models, as conjectured in section 2.6. In positions 5 and 6, we have ARDIRF and VARRF. Both can provide substantial improvements at times, but also can fail badly. This is just a reflection of empirical properties of the underlying model: ARDI will mostly work well for real activity variables while AR work reasonably well for everything. Thus, it is not surprising to also see ARDIRF inherit these uneven properties.

As discussed in earlier sections, ARRF connects to the wider family of non-linear autoregressive models. In Figure 29, I carry the same violin plot exercise where I rather compare ARRF to other non-linear time series models. It clearly does better on average than the considered SETAR and Smooth-Transition TAR. This advantage is in great part attributable to the use of much more data to characterize the AR coefficients' law of motion: ARRF does much better than Tiny ARRF. Nevertheless, the latter is still better than the ∼TAR group. Linking this result to the conclusion of simulations, this means that none of ∼TAR's model is likely the true model. Indeed, it has been shown that ARRF is very unlikely to beat SETAR if the true DGP is SETAR.



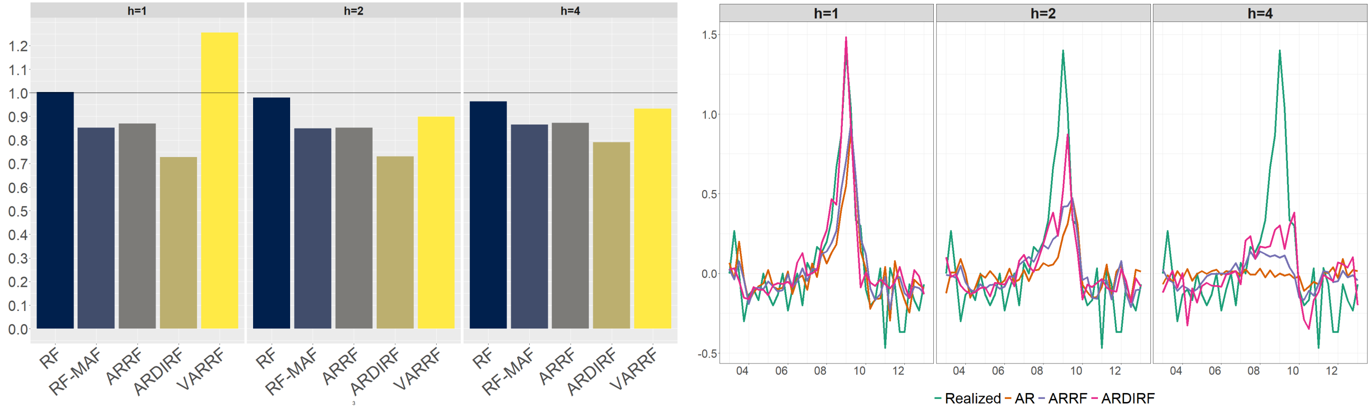
(a) $RMSPE_{GDP,h,m}/RMSPE_{GDP,h,AR(4)}$

(b) A look at forecasts

Figure 8: GDP results in detail

The first thing that catches the eye for GDP is how well does ARDIRF grasp the 2008 drop one quarter ahead. ARDIRF exhibits a bit less than a 20% drop in RMSPE over the quite robust and competitive AR(4).[32] Results in section 4.2.1 will reinforce the view that a data-rich model within MRF are successful at catching the recession as it unfolds and raising some serious flags up to year ahead.

[32] Diebold and Rudebusch (1994) proposed a regime-switching factor which was rather empirically successful. Given that line of work and more recent results in Wochner (2020), the ARDIRF's success as reported here is not an anomaly.



(a) $RMSPE_{UR,h,m}/RMSPE_{UR,h,AR(4)}$

(b) A look at forecasts

Figure 9: UR results in detail

ARDIRF dominates even more strongly for UR. Table 4 reveals it is the best model for all horizons but the last one (8 quarters ahead, where the encompassed RF-MAF is the best). Clearly, at an horizon of one quarter, the preferred model successfully predicts the unprecedented rise in unemployment that occurred during the Great Recession. Rather than responding with delay to negative shocks as they come (which is what we observe from AR and ARRF), the model visibly predicts them. As a result, improvements in RMSPE are between 25% and 30% over AR(4) for all horizons. Specifically, predicting UR with ARDIRF at $h = 1$ yields an unusually high out-of-sample $R^2$ of about 80%. The nearly perfect overlap of the green and magenta lines in Figure 9 reveals the noticeable absence of a one-step ahead shock around 2008 – conditional on using the proper model. For $h = 2$, the quantitative rise is nowhere near the realized one, but it nevertheless reveals 6 months ahead the arrival of a significant economic downturn. For $h = 4$, ARRF and ARDIRF reveal one year ahead the arrival of a rise in unemployment, which is a quality shared by very few models. The barplot in Figure 9 provides a natural decomposition of ARDIRF's gains (ARDIRF $\succ$ RF $\sim$ AR). Adding the MAFs to an otherwise plain RF procures an improvement of roughly 15% across all horizons (RF-MAF $\succ$ RF). The linear ARDI part and the rest of improvements discussed in section 2 provide an additional reduction of 10% to 15% depending on the forecasted horizon (ARDIRF $\succ$ RF-MAF).[33]

VARRF only shines for SPREAD (Figure 30) and captures very well its movements, even up to a year ahead. For this very persistent variable, the improvements of using RF over AR(4) are reasonable but can be magnified by opting for VARRF. Gains over the benchmark are up to 40 % and are up to 20% with respect to RF-MAF.[34] The simpler AR+RF also does remarkably well.

[33] The good results for $h = 1$ are mechanically close to impossible with a plain RF since it cannot extrapolate – predict values of $y_t$ that did not occur in-sample. In contrast, this is absolutely feasible within MRF via the use of a linear part.

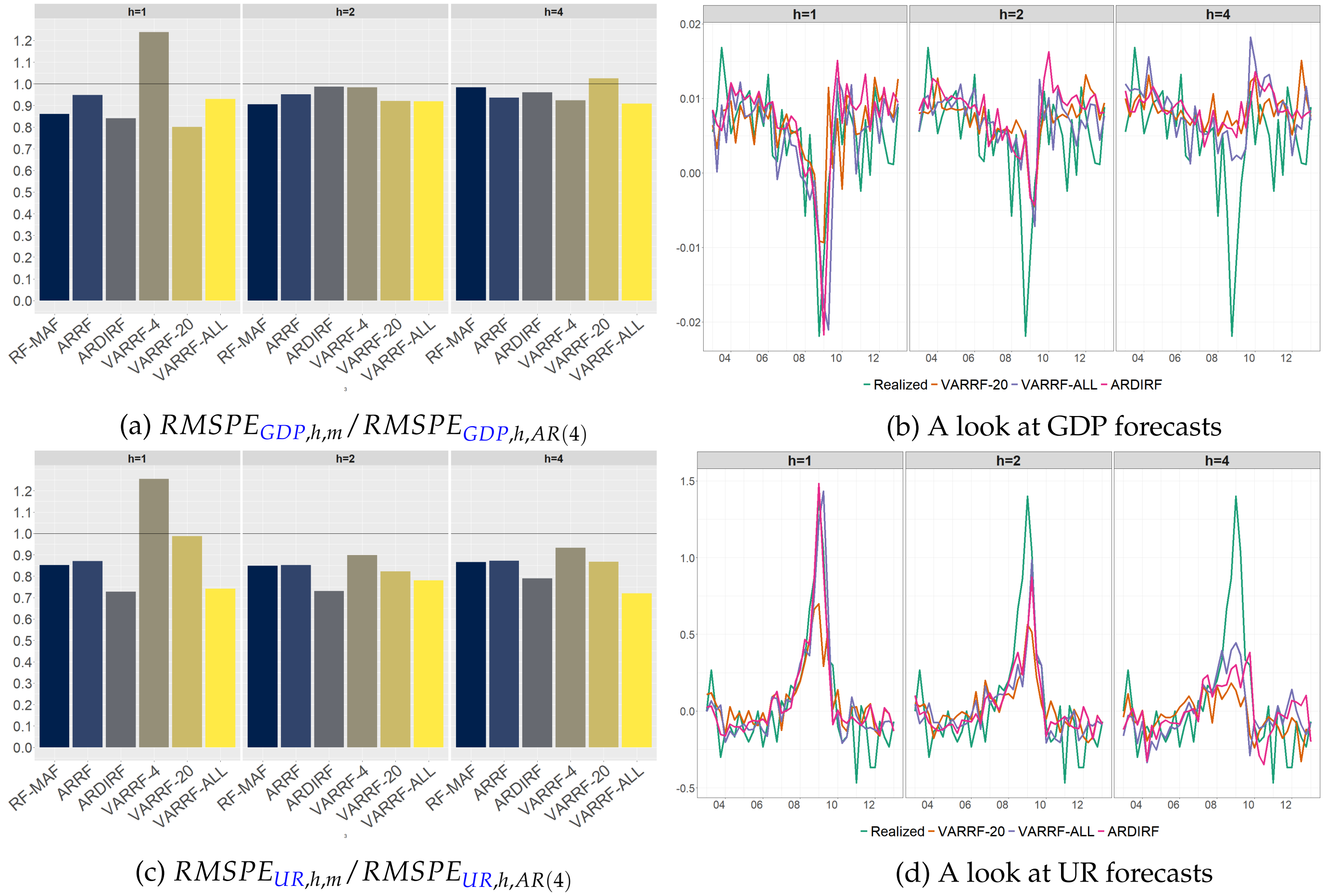
[34] Goulet Coulombe (2019) also found that a small TVP-VAR (rather than AR or ARDI) provide consistently better results for the spread at any horizon.

Overall, these results highlight the importance of the autoregressive part, which is no surprise given SPREAD's persistence. For INF, Table 4 displays that RF-MAF (the simpler MRF that only employs MAFs in a plain RF) is the leading model with an average reduction in RMSPE of 15% with respect to AR(4). Figure 31 shows how Tiny ARRF successfully recuperate the later part of the 2008 inflation drop. This unsurprisingly points out that including lags of the dependent variable and a flexible exogenous trend goes a long way when modeling inflation. Nonetheless, the marginal success of RF-MAF is visually easy to explain: it is a flexible time-varying constant that captures well the long-run trajectory of inflation in the forecasting window.

### 4.2.1 Can Larger VARs Help?

As argued earlier, an advantage of MRF over plain RF is that by taking the TVP view of non-linearities, we are in a much better position to attempt an interpretation of the successful model. One could rightfully retort that while ARDIRF performs nicely, its potential for interpretation is spoiled using factors rather than raw data. While this critique is partially addressable by putting names on factors such as 'real activity' and 'forward looking' factors, it is worthwhile to consider alternative dimensionality reductions schemes that keep the data in the original space. Since the 4 variables VARRF results are not necessary sterling, the expectations for VARRF of bigger size are rather low. Regardless, Figure 10 show promising results for both a VARRF with 20 variables (in the spirit of Bańbura et al. (2010)'s medium VAR) and a VARRF that includes *all* FRED-QD 200+ variables. In both cases, especially the later, regularization must be stringent to keep estimation variance low. Indeed, in VARRF-ALL, the linear part has more regressors than observations even before any split is attempted. This implies a much higher value of $\lambda$, $\zeta = 1$ and the use of a single lag in the linear part. Quite strikingly, Figure 10d report that VARRF-ALL tracks UR at $h = 1$ remarkably well, in addition to providing forecasts that clearly hint at a recession for both UR and GDP up to a year ahead. In that latter regard, VARRF-ALL is the best model of the whole lot. Hence, larger models, while not the center of interest for this paper, can be handled by MRF and provide excellent forecasts given proper regularization.

This subsection's results point out that the tool can be easily (and desirably) extended to estimate large GTVP-VARs. The dynamic coefficients can be estimated by either fitting MRF equation by equation. Another possibility (left for future research) is to simply modify the splitting rule in (3) to be multivariate so that each tree is fitted jointly for all equation – pooling time-variation across equations. Finally, elements of the covariance matrix of residuals can be fitted separately with a plain RF, which is very fast.



(a) $RMSPE_{GDP,h,m}/RMSPE_{GDP,h,AR(4)}$

(b) A look at GDP forecasts

(c) $RMSPE_{UR,h,m}/RMSPE_{UR,h,AR(4)}$

(d) A look at UR forecasts

Figure 10: Larger VAR Results

## 4.3 Monthly Data Results

Will the general message of the preceding sections hold when changing (part of) the data and the sampling frequency? I run a similar exercise as in Goulet Coulombe et al. (2019) which is very close to what has been done for quarterly data. I now use FRED-MD which has less series but more observations. It contains 134 monthly US macroeconomic and financial indicators observed from 1960M01 to 2017M12. The details on the dataset and the series transformation are all in McCracken and Ng (2016). To match exactly the experimental design of Goulet Coulombe et al. (2019), GDP is replaced by Industrial Production (IP) and GS1 is dropped. The horizons of interest are $h = 1, 3, 9, 12, 24$ months. The forecast target is the average growth rate $\sum_{h'}^{h} y^{v}_{t+h'}/h$ which is much less noisy than the monthly growth rate. For instance, in the case of inflation 24 months ahead, this means we are targeting the average inflation rate over the next two years rather than the quarterly inflation rate in 2 years, as was done in the quarterly exercise. The OOS period is the same.

The ARDIRF is still very competitive. Like in the quarterly exercise, ARRF provides great in-

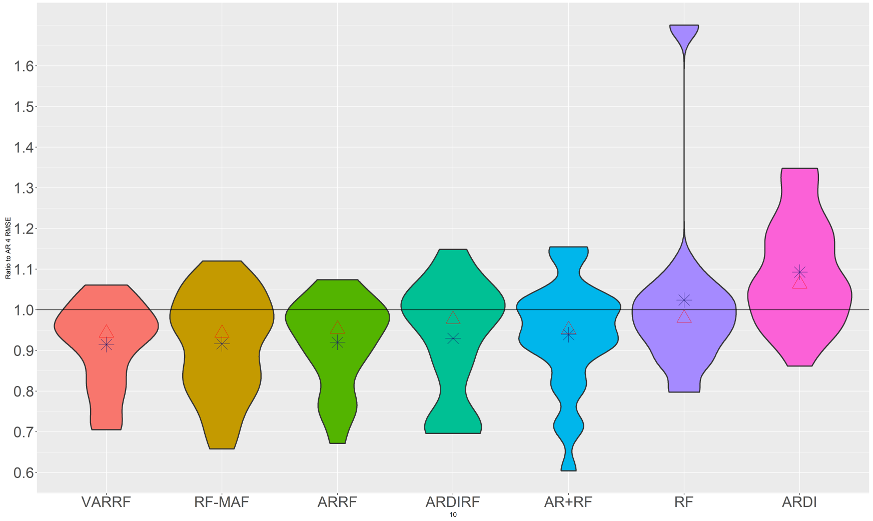


Figure 11: The distribution of $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$ for monthly data and horizons 1,3,9,12 and 24 months. The star is the mean and the triangle is the median.

surance against doing worse than a plain AR counterpart (here AR(12)).[35] The last two models do not have the MAFs and are clearly outperformed by the rest that do, unsurprisingly suggesting that lag polynomial compression can be of even greater use at the monthly frequency.

Classic tables of specific $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$'s with DM tests showing results for all monthly models simultaneously (see Table 5) are in the appendix. Broadly, they show that (i) MAFs are without any doubt the major improvement for the first three variables (IP, UR, SPREAD), (ii) ARDIRF can improve on RF-MAF, but not by a whole lot and (iii) ARDIRF is often close to or the best model for INF at all horizons (except for $h = 9$ where ARRF wins by a thin margin). Naturally, point (iii) is the most striking as ARDIRF can be thought of as a Phillips' curve forecast, which recurrent failures are well documented (Atkeson et al. (2001), Stock and Watson (2007)).[36] For that particular reason, Table 5 includes forecasts inspired by the contribution of Atkeson et al. (2001): 1, $h$ and 12 months moving averages are considered (where $h$ is the targeted horizon). As in the original paper, the AO-12 forecast proves remarkably resilient, but is bested with a sizable margin at each horizon by both ARRF and ARDIRF. This finding is further explored in section 5.3.2. Finally, Tiny ARRF is also shown to perform well at forecasting inflation at longer horizons, like 1 and 2 years. This is sensible given that by restricting $S_t$, Tiny ARRF puts the emphasis on modeling long-run exogenous change, the usual suspect for inflation.

[35] This is also true for the more parsimonious AR(4), see Table 5.

[36] Additionally, it is reported that the plain ARDI, in contrast, does really bad for inflation.

### 4.4 External Validity

Much attention has been paid to the prediction of US economic aggregates. An even greater challenge is that of forecasting the future state of a small open economy. Such an application is beyond the scope of this paper but is considered in Goulet Coulombe et al. (2020). The study considers the prediction of more than a dozen key economic variables for Canada and QuÃľbec using the large Canadian data base of Fortin-Gagnon et al. (2018). The forecasts from about 50 models and different averages of them are compared, with ARRF and ARDIRF among them. The MRFs provide substantial improvements especially at the one-quarter horizon for numerous real activity variables (Canadian GDP, QuÃľbec GDP, industrial production, real investment). In such cases, ARRF or ARDIRF provide reductions (with respect to autoregressive benchmark) that are sizable and statistically significant, going up to 32% in RMSPE. That performance is sometimes miles ahead from the next best option among Complete Subset Regression, Factor models, Neural Networks, Ridge, Lasso, plain RF, different models averages. Goulet Coulombe et al. (2020)'s results suggest that MRFs forecasting abilities generalize beyond the traditional exercise of predicting US macro aggregates.

## 5 Analysis

Based on forecasting results from the previous section, I focus on the small factor model MRF (ARDIRF). I first dissect the ARDIRF forecast for UR around 2008 and see how it differs from its OLS counterpart. Second, I compare GTVPs to random walk time-varying parameters which reveals that the two can differ substantially. Finally, I use the surrogate model approach to attempt an explanation of the parameters' law of motion in terms of observed variables.

### 5.1 Forecast Anatomy

Any non-linear time series model can be represented as a time-varying parameter model (Granger (2008)). Consequently, a natural way to start the investigation on the successful R forecasts reported in section (4.2) is to look at the time-varying parameters themselves. In the case of the ARDIRF, the forecasting equation is

$$y_{t+h} = \mu_t + \phi_{1,t} y_t + \phi_{2,t} y_{t-1} + \gamma_{1,t} F_{1,t} + \gamma_{2,t} F_{2,t} + u_{t+h}.$$

Let $\beta_t$ be the $(5 \times 1)$ vector that stacks all coefficients and $X_t$ the $(5 \times 1)$ vector of regressors. To avoid overfitting, the reported $\hat{\beta}_t$'s are (as in simulations) the posterior mean over draws that did not include observations $t-4$ to $t+4$ (a two years block) in the tree-fitting process. Intuitively, this mimics in-sample the real out-of-sample experiment that starts here in 2007Q2.[37]

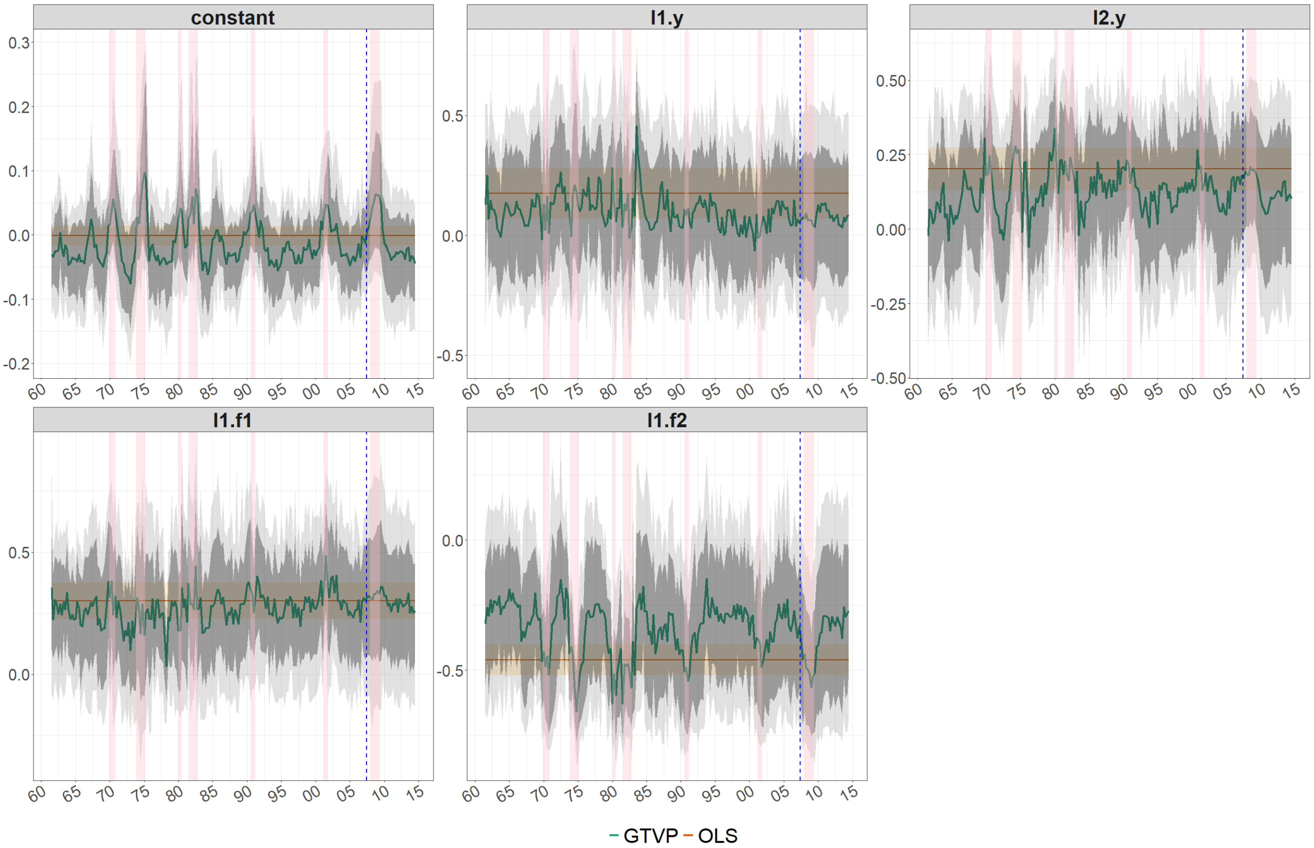


Figure 12: GTVPs of the one-quarter ahead UR forecast. The grey bands are the 68% and 90% credible region. The pale orange region is the OLS coefficient ± one standard error. The vertical dotted blue line is the end of the training sample. Pink shading corresponds to NBER recessions.

Figure 12 display the coefficients underlying the successful one-step ahead UR forecast reported earlier. The constant clearly alternates between at least two regimes and the "increasing UR" one is in effect circa 2008.[38] Also, the effect of the first lag is smaller compared to OLS while that of the forward-looking factor $F_2$ is clearly magnified during recessionary episodes. The conjunction of the movements of these two parameters explains the faster response of ARDIRF during the onset of Great Recession. $\gamma_{2,t}$ smooth-switching behavior can be best interpreted by remembering that $F_2$ is highly correlated with capacity utilization, manufacturing sector indicators, building permits and spreads (McCracken and Ng (2020)). Many of those variables are considered "leading" indicators and have often been found to increase forecasting performance, mostly before and during recession periods (Stock and Watson (1989), Estrella and Mishkin (1998)). In essence, MRF learns that the relationship between this pack of economic indicators

[37] Note that this is partially different from what gave the results reported in section 4.2, where the model was re-estimated every 2 years. Here, estimation is done once in 2007Q2.

[38] In levels, this means UR alternates between a positive and negative trend.

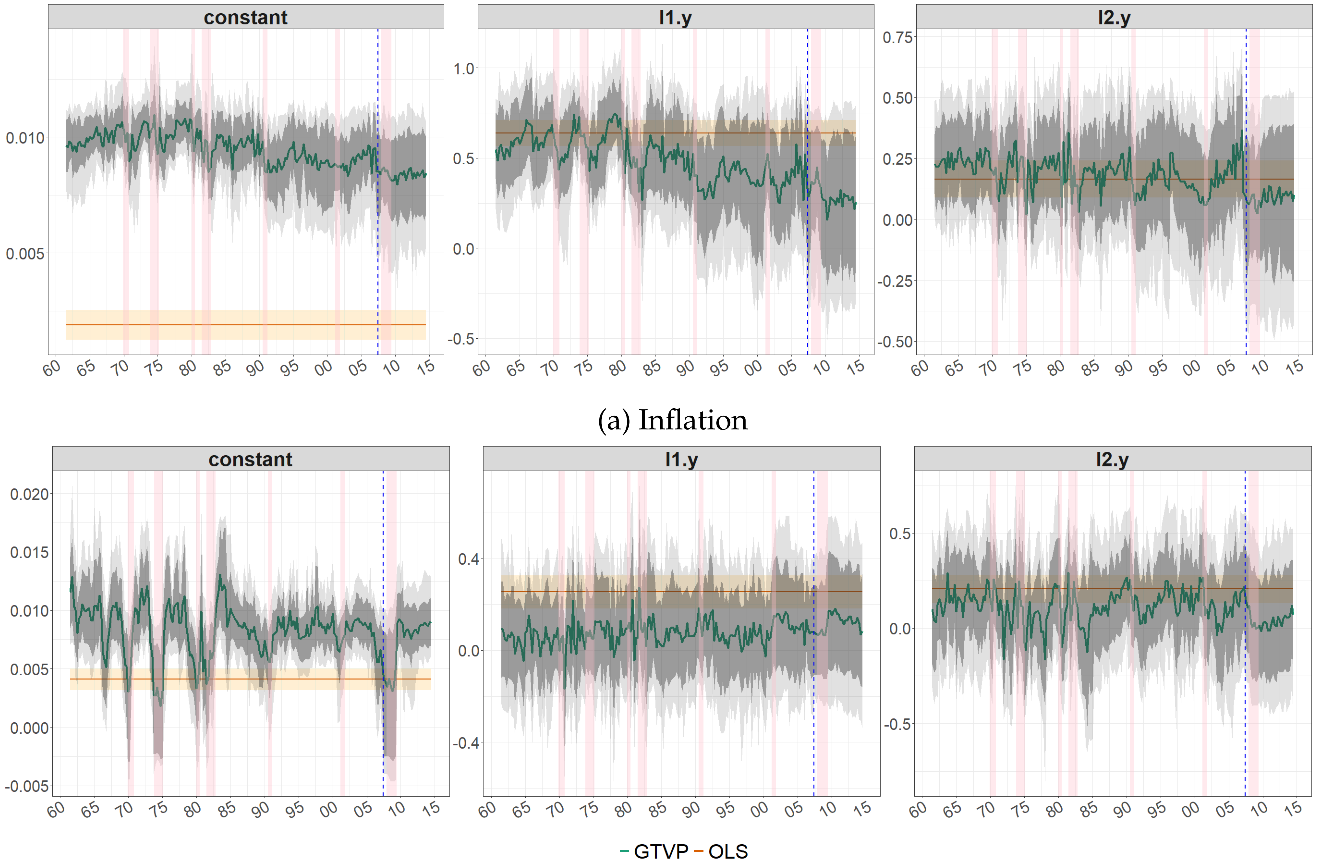
and unemployment is stronger before and during recessions than it is during expansions. OLS can only provide a clumsy average of these very different regimes-dependent $\gamma_{2,t}$'s. Section 5.3 will investigate formally the underlying variables driving this time-variation. Finally, Figure 32 displays $\beta_t$ for the prediction of GDP one quarter ahead where slow and relatively mild long-run change is observed.

### 5.1.1 Why and When MRF Can Fail to Deliver Better Forecasts

MRF can sometimes be outperformed by available alternatives, like standard RF that incorporate MAFs. When that occurs, it is usually due to inadequacy of the assumed linear model rather than GTVPs themselves. Unlike traditional TVPs, GTVPs very rarely provide a model worse than that of the plain linear part itself. Additionally, in contrast to pure black-box models, MRF provides the option of visualizing the evolving parameters, which can help greatly in understanding forecasting performance results. For instance, in the case of forecasting inflation with the quarterly data set, ARRF does not supplant RF-MAF. The critical difference between ARRF (reported in Figure 13a) and its restricted RF-MAF analog is that the two autoregressive coefficients of the former are shut to 0.[39] In Figure 13a, the estimates of ARRF broadly agree with the view that inflation persistence has substantially decreased during and following Volker disinflation (see Cogley and Sargent (2001) and Cogley et al. (2010) for instance).

In terms of anticipated forecasting performance, such decline in persistence suggests that ARRF may not fare so well with respect to a constrained version. The OOS evaluation period corresponds to the section of Figure 13a where the credible region of both autoregressive coefficients includes 0. Given that observation, the superiority of a model that does not focus on autoregressive behavior (RF-MAF) is much less surprising. An analogous finding emerges for GDP at many horizons. ARRF does not outperform RF-MAF like ARDIRF and larger VARs versions of MRF do. GTVPs showcased in Figure 13b provide a simple explanation for the phenomenon. There is a very limited role for autoregressive coefficients: they are under the OLS values most of the time and the credible 68% credible region frequently includes 0. Essentially, the forecast is a time-varying constant, which is what plain RF does. The limited gains from using ARRF on GDP can thus be explained by the fact that autoregressive behavior is of lesser importance once we allow for the constant to be time-varying in a flexible way. In sum, unlike many members of its model family, MRF provide a way to explain its successes and failings via a time-varying parameter interpretation. The helpfulness of this attribute cannot be overstated when it comes to interpreting the forecasts and thinking about further model improvements.

[39] Of course, lags of INF can still enter the forest part for $\mu_t$, so RF-MAF does not suppress entirely the link between current and recent inflation.



(a) Inflation

(b) GDP

Figure 13: GTVPs of the one-quarter ahead forecasts using ARRF. The grey bands are the 68% and 90% credible region. The pale orange region is the OLS coefficient ± one standard error. The vertical dotted blue line is the end of the training sample. Pink shading corresponds to NBER recessions.

## 5.2 Comparing Generalized TVPs with Random Walk TVPs

The relationship between random walk TVPs and GTVPs was evoked earlier. I compare them for the small ARDI model that is reported to be successful in forecasting experiments, especially for UR. Thus, I estimate standard TVPs using the ridge regression technique developed in an earlier paper.[40] The level of smoothness in that approach is also guided by a ridge $\lambda$ tuned by k-fold cross-validation (which is valid under no serial correlation (Bergmeir et al. (2018)).

As Figure 12 suggested for $\mu_t$ and the coefficient on $F_2$, parameters can be subject to recurrent, rapid and statistically meaningful shifts. Such behavior creates difficulties for random-walk TVPs that rather excel at smooth and persistent structural change by construction. Figure 33 confirms this conjecture. Standard TVPs look for long-run change when regime-switching

[40] In that work, I show with simulations that this much easier approach and a traditional Bayesian TVP-VAR perform similarly well for models that the latter is able to estimate (computational constraints limiting its use for bigger models).

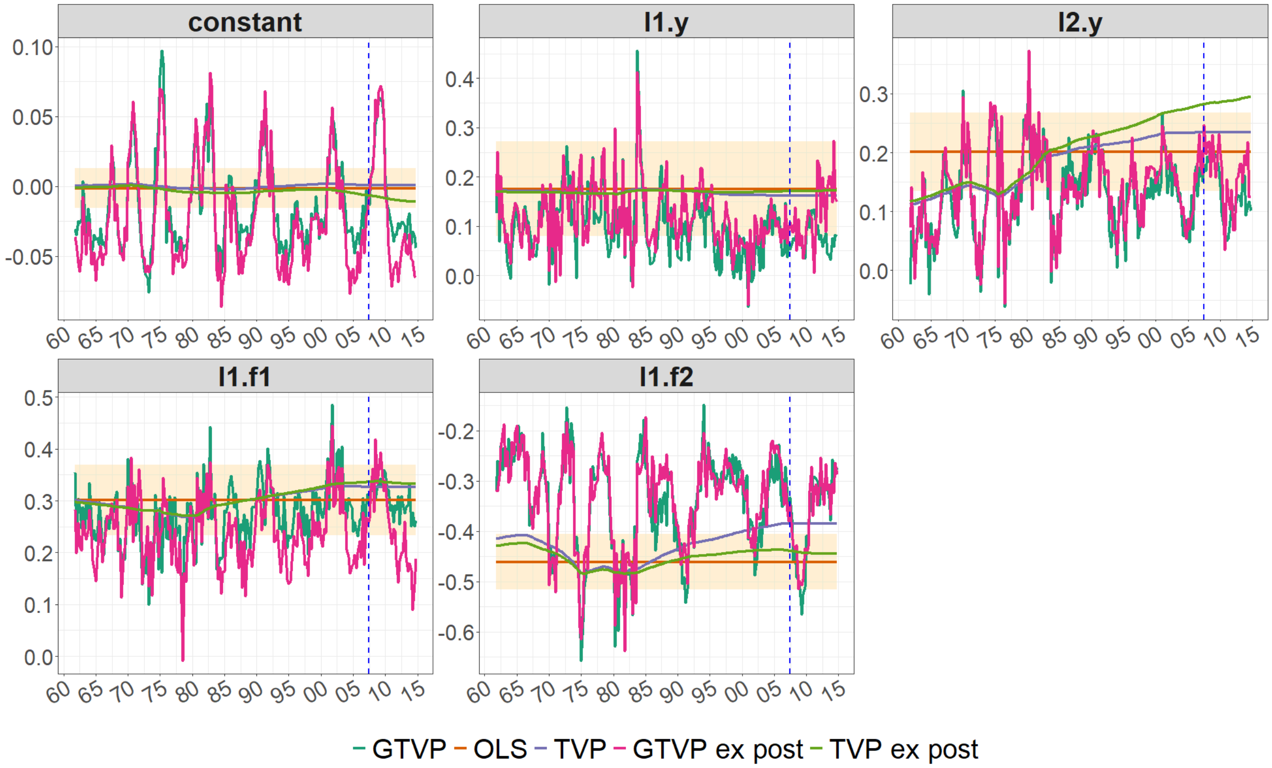


Figure 14: GDP equation $\beta_t$'s obtained with different techniques. TVPs estimated with a ridge regression as in Goulet Coulombe (2019) and the parameter volatility is tuned with k-fold cross-validation. Ex Post TVP means using the full sample for estimation and tuning as opposed to only using pre-2002 data as for GTVPs. The pale orange region is the OLS coefficient ± one standard error. Pink shading corresponds to NBER recessions.

behavior is the main driving force. As a result, they are flat and within OLS confidence bands, as often reported in the literature (D'Agostino et al. (2013))). Of course, more action will mechanically be obtained for TVPs when considering a smaller amount of smoothness than what cross-validation proposed. In appendix A.8, I report the same figures, but using the optimal smoothing parameters (as picked by CV) divided by 1000. This provides much more volatile random walk TVPs that are inclined, at certain specific moments, to follow the GTVPs. However, it is clear in Figure 33 that the end-of-sample/revision problem is worsen by the artificially soft smoothing imposed. It is known in the traditional TVP literature that there is a balance between flexible (but often erratic) $\beta_t$ paths and very smooth ones where time-variation may simply vanish. In yet another incarnation of the bias and variance trade-off, the second option is usually preferred.[41] Since random-walk TVPs are unfit for many forms of the time-variation present in macroeconomic data, high bias estimates are usually reported as only them can keep

[41] In the case of ridge regression-based TVPs, cross-validation is just a data-driven way of backing this necessary empirical choice.

variance below a manageable level. This can have serious implications. Relying too much on time-smoothing can create a mirage of long-run change and/or dissimulate parameters that mostly (but not solely) vary according to expansions/recessions.

As we have seen, one concern is that random-walks TVPs are not flexible enough. Another, particular to the act of forecasting with such models, is the boundary problem. As mentioned before, random-walk TVP models forecasts can suffer greatly from it because by construction, the forecasts are always made at the boundary of the variable on which the kernel is based – time $t$. Of course, one can deploy a 1-sided kernel, but this only alleviate a few pressing symptoms, without attempting any cure for the true underlying problem. In sharp contrast, GTVPs use a large information set $S_t$ to create the kernel, which implies that the likelihood of making a forecast at the boundary of the kernel is rather low, unless the RF part constantly selects $t$ as splitting variable. Figures 33 and 14 show, for both random walk and generalized TVPs, their full-sample versions (up to the end of 2014, "ex post") and their version with a training sample ending in 2007Q2 (the dashed blue line). There are two main observations. First, GTVPs are much less prompt to "rewrite history" than random-walk TVPs. Indeed, the green line and the magenta one closely follows each other all the way up to end of the training sample. That is less true of random walk TVPs as there are clear examples where the two version differ for a long period of time (for instance, the constant and the coefficient on $F_2$ in the GDP equation). Second, while GTVPs can change long after the 2007Q2 boundary (like the GDP constant), they are generally very close at the boundary especially when the time-variation is statistically meaningful (like the coefficient on $F_{2,t}$), which is what matters for forecasting. This is much less true of random walk TVPs. [42]

## 5.3 Cutting Down the Forest, One Tree at a Time

An inevitable irritant of GTVPs (that also pertains to traditional TVPs) is that parameters change – and sometimes quite fast. This can limit macroeconomists in their ability to use the model for counterfactuals. Policy makers will rightfully complain about the limited use for a model in which tomorrow's parameters are unknown. Fortunately, GTVPs may be the result of an opaque ensemble of trees, but they are made out of observables rather than a multiplicity of latent states. That is, they change, but according to a *fixed* structure.[43] Hence, the reduced-form coefficients could change for reasons evoked in Lucas (1976) and yet remain completely predetermined as long as the tree structure itself is stable. Additionally, in the RF paradigm, there

[42] In (real) practice, all models would be re-estimated each quarter. However, it is worth pointing out that re-estimating every period is much more important for random-walk TVP than it is for GTVPs. For such reasons, the TV-AR in section 4 was the sole model estimated every period rather than every two years.

[43] This also means that $\beta_{t+1}$ is forecasted by a RF, which lends itself to much higher hopes than a random walk forecast.

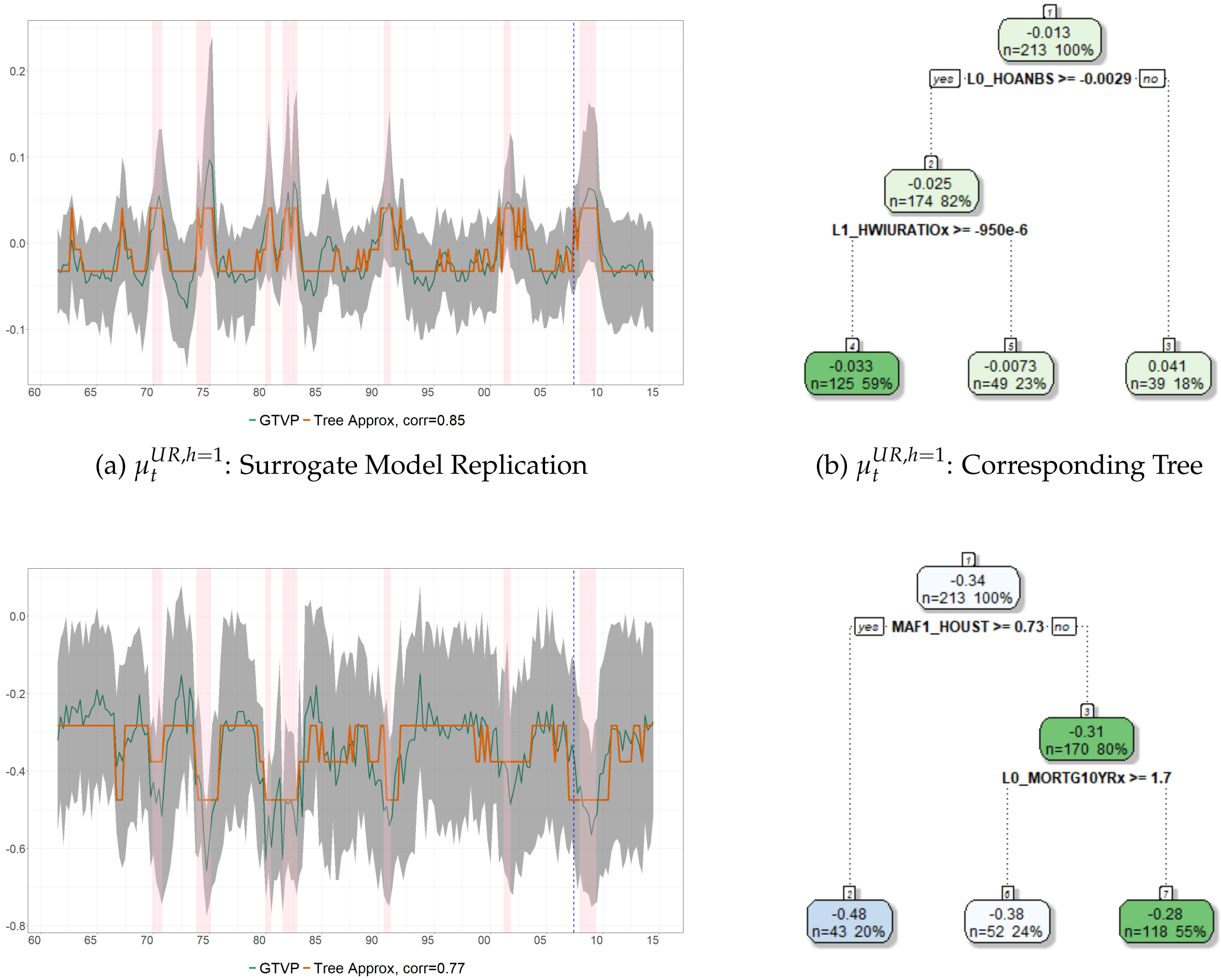
exist well-established built-in measures of Variable Importance (VI) as originally proposed in Breiman (2001). Those usually focus on extracting features driving the overall prediction. Fortunately, they can be adapted for the inquiry of linear parameters themselves. Then, one can capitalize on VI's insights to build interpretable small trees in order to parsimoniously approximate each parameter $\beta_{t,k}$'s path in terms of observed variables.

The construction of upcoming graphs consists of two steps. First, I compute 3 different VI measures for each variables. They are used to first uncover a set of reasonably potent predictors for subsequent steps of this descriptive analysis. As a potential data set for the construction of the tree, I consider the union of the sets of the 20 most important variables as highlighted by either $VI_{OOB}$ (out-of-bag predictive performance), $VI_{OOS}$ (out-of-sample predictive performance) or $VI_{\beta}$ (for a specific coefficient rather than the whole prediction). The tree is pruned with a cost-complexity factor (usually referred to as `cp`) of 0.075. The latter tuning parameter is chosen such as to balance the capacity of the tree to mimic the original parameter and its potential for interpretation. Appendix A.1 contains a detailed explanation of VI measures as well as a discussion of how the current approach relates to recent work in the ML interpretability literature.

### 5.3.1 Unemployment

I first focus on the ARDIRF and its results for UR at the quarterly frequency. I fit a standard regression tree to the switching constant and the coefficient driving the forward-looking factor's impact on UR. I center the analysis around those parameters for two reasons: the significance of their time-variation and the potential for the path to be reasonably well characterized by a single tree. Figure 12 suggest a clear switch of values for both $\mu_t$ and $\gamma_{t,F_2}$ around all recessions. Figures 15b and 15d point out that an important part of each parameter path can be explained by a handful of predictors. $\mu_t$ essentially alternates between two states which are determined by a cut-off on lagged non-farm business sector hours (HOABNS): 0.041 (increasing unemployment) and -0.025 (decreasing). This first layer basically classifies recession vs expansions in a very parsimonious way, which is inevitably crude and imperfect as revealed by the second layer of the tree. The additional split on HWIRATIOx (help-wanted index ratio to unemployed, a measure of labor market tightness) provides a more refined classification: there are in fact more of less three states. The time series plot shows the alternation between two symmetrically opposed states of 0.041 and -0.033 (respectively entering and exiting a recession) and a transitory (and seldomly visited) middle ground around 0. This finding is in accord with the view that a great part of UR in *levels* can be modeled by the alternation of two trends.

The impact of $F_2$ on UR switches in a significant way most of the action can be captured by the first MAF of Housing Starts. The leading indicator's movement downwards – which



(a) $\mu_t^{UR,h=1}$: Surrogate Model Replication

(b) $\mu_t^{UR,h=1}$: Corresponding Tree

(c) $\gamma_{t,F_2}^{UR,h=1}$: Surrogate Model Replication

(d) $\gamma_{t,F_2}^{UR,h=1}$: Corresponding Tree

Figure 15: Surrogate $\beta_{t,k}$ Trees. Shade is 68% credible region. `cp`=0.075. Trees drawn with Rattle. Pink shading corresponds to NBER recessions.

usually commence from the very onset of a recession and sometimes even before – can double the effect of $F_2$ on UR in absolute terms. Since $F_2$ is already composed mostly of forward-looking variables (which include housing sector indicators), this hints at an overall non-linear effect of forward-looking variables, at least for the three pre-1985 and 2008 recession episodes. The other also see a more modest increase in $\gamma_{t,F_2}$ which is driven by the ratio of the 30 years Mortgage Rate to the 10-year treasury rate. Mortgage rates usually follow the 10-year treasury rate quite tightly, but the ratio can surge preceding downturns like it did in the mid 70's and the early 2000's.

Given the points raised earlier in 2.1, it is more appropriate to see these surrogate trees as suggestive of one potential explanation. It is an open secret that their exact structure is sensible to small changes in the estimated path. For instance, little variation in $\beta_t$ is needed to observe a change in the exact choice of variables itself. As a result, some of them may rightfully seem exotic when singled out in such a simple tree. GTVPs, as the product of a forest, will more often than not rely on a multitude of indicators from a specific group (which we observe in 34a) rather than a single indicator. Nevertheless, combined with a harsh screening of variables backed by different measures of VI used in the much more robust MRF, we can be confident that these trees provide very useful indicative information about what MRF is actually doing. Finally, it is obvious that the very use of factors in the linear part is an obstacle to deeper economic explanations: MRF's interpretability is bounded above by that of the linear model it encompasses. Hence, more insights can be gained if, for instance, the coefficient on $F_1$ has a deeper economic meaning, like it would in a Phillips' curve.

### 5.3.2 A Look at Inflation

As discussed briefly earlier, the ARDIRF turns out to be a very competitive model for *monthly* inflation forecasting at almost all horizons. Being composed of two lags of inflation and lags of $F_1$ and $F_2$, the resulting model has the familiar flavor of a Philipps' curve.[44] This is an interesting finding given that Phillips' curves have at best a very uneven forecasting performance record (Stock and Watson (2008)). Recently, Kotchoni et al. (2019) conclude that an ARMA(1,1) is the best modeling option for inflation *growth* except in recessionary periods where a data-rich environment can be helpful.[45] Faust and Wright (2013) report that Phillips curves, TVP-VARs and factor augmented regression all provide results that are usually inferior to that of the benchmark for CPI (which is the series used here). Medeiros et al. (2019) obtain significant improvements combining plain Random Forest with inflation subcomponents. From these studies, a pattern emerges: future inflation is best predicted by past inflation and not much else.

An improved understanding of all these findings can be achieved by standing on the shoulders of Atkeson et al. (2001) and Stock and Watson (2008). Simple autoregressive/random walk/historical mean benchmarks are hard to beat. The relationship between real activity and inflation is either time-varying in a way that annihilates its forecasting potential (Stock and Watson (2008)) or may have flatten to the point of predictive desuetude (Blanchard et al. (2015), Blanchard (2016)). An adjacent literature (for instance, Dolado et al. (2005)) rather stipulates that the curve of high interest is in fact non-linear. Hence, the success of the ARDIRF is presumably attributable to incorporating different elements of the aforementioned approaches. Among other things, it includes the slowly moving mean that Faust and Wright (2013) rightfully claim

[44] As noted in Stock and Watson (2008), the plethora of output gap indicators used in literature makes the use of a common statistical factor a credible alternative.

[45] Obviously, leveraging the power of big data in this scenario requires a recession/expansion forecast, which makes the finding of limited practical value. Through GTVPs on $F_1$ and $F_2$, the ARDIRF can exploit the insight of Kotchoni et al. (2019) and activate the data-rich part when it is most beneficial to do so. Figure 16 suggests this is exactly what is happening – especially with the use of leading indicators to activate the coefficient of $F_1$.

to be necessary. Further, it includes a real activity factor rather than a single indicator.[46] I now turn to investigate those claims more systematically.

The variable importance measures reported in Figure 36 paint a somewhat different picture than that of real activity variables considered earlier. First, there is a clear "consensus" subset of variables that seem to matter for inflation at all horizons. Three recurrent variables are an exogenous trend, MAF of building permits / housing starts and MAF of business inventories. The leading role for the trend suggests that exogenous time-variation as implemented by random walk TVP is important to explain inflation. That is indeed no surprise given that modeling an exogenously slowly time-varying inflation process was one of the main motivation for the development of drifting parameters models (Cogley and Sargent (2001)). However, it is not the full story. VI measures also display another very interesting fact. For both the 1 month and 12 months ahead forecasts, we see that exogenous time-variation is important only for the constant and the two autoregressive lags, which is, again, in line with the view that inflation's mean and persistence have been evolving in an "exogenous" fashion – structural change. The long-run change in autoregressive lags is clearly visible in Figure 38 and especially Figure 39.[47]

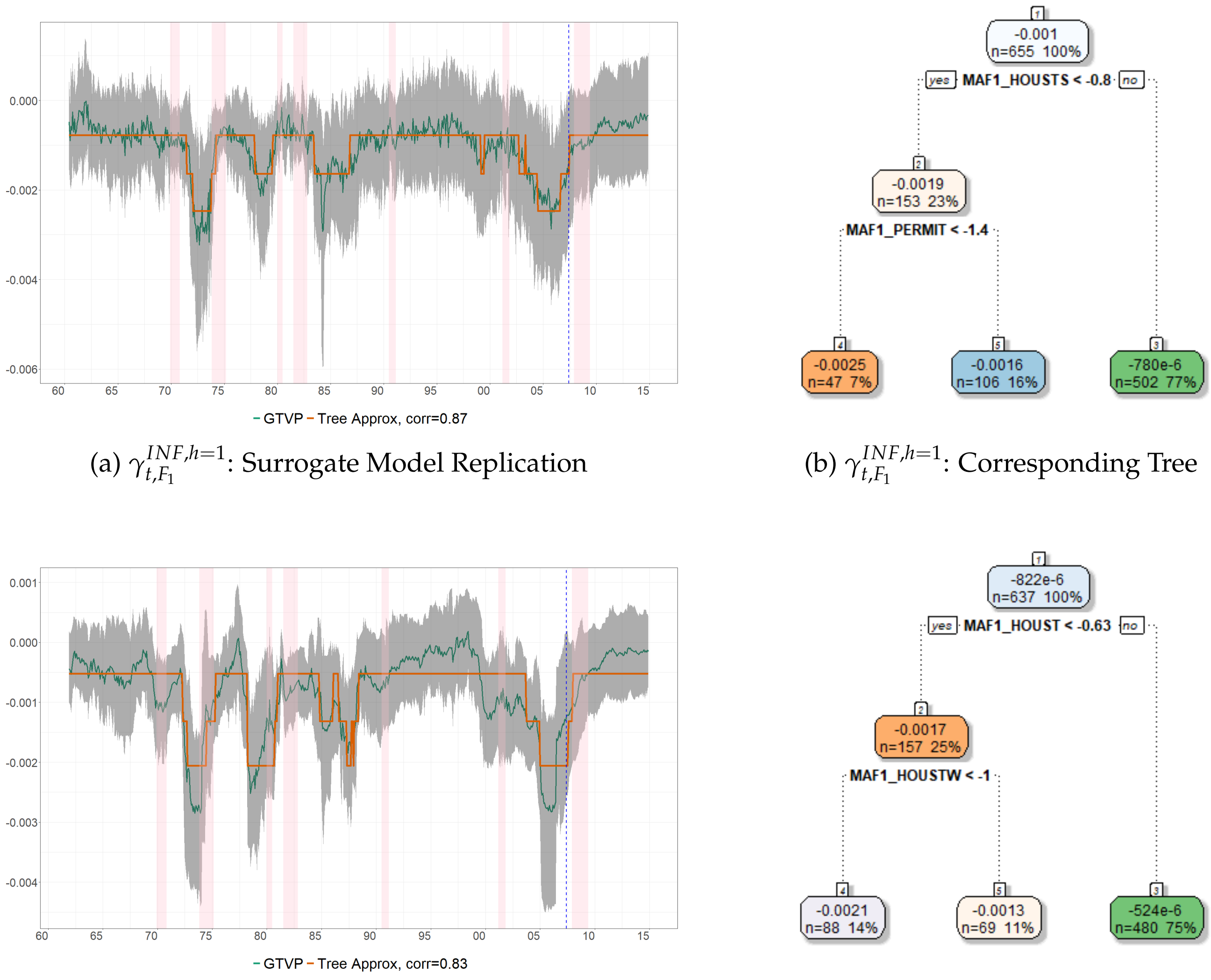
The coefficients on $F_1$ (real activity factor) and $F_2$ (forward-looking factor, can be understood as another predictor proxying for expectations along with lags) are also influenced by the trend, but in a much milder way. Overall, Figure 36 highlights that the most important variables are housing sector-related (permits, housing starts) which suggest that the coefficients may rather have cyclical behavior. Figures 38 and 39 confirm that. Both coefficients on the activity indicator ($\gamma_{t,F_1}^{INF,h=1}$ and $\gamma_{t,F_1}^{INF,h=12}$) follow a distinct but common pattern. It points that the positive relationship between inflation and economic activity is episodic and usually prevails before recessions – but not all. Figure 16 proposes a clear-cut answer: inflation responds to real activity when there is overheating on the housing market – as characterized by the MAFs of housing starts and permits (see Figure 37, the rotation is such that "overheating" correspond to when MAF is low).[48] It is worth reminding the reader that while ARDIRF superiority over the simpler ARRF (or even RF-MAF) occurs for some horizons in 5, it is sometimes by a rather small margin. Figure 38 rationalize that: the credible regions of $\gamma_{t,F_1}^{INF,h=1}$ and $\gamma_{t,F_1}^{INF,h=12}$ much more frequently 0 than the autoregressive lags and the constant. Hence, the focus on the real activity indicator is rather for its meaning than for an outstanding predictive power contribution.

An interesting finding is the importance of the housing sector in driving the strength of the

[46] Results reported in Table 5 also point in the direction that a plain factor-augmented regression (termed ARDI) does not deliver the goods, supporting the view that non-linearities matter.

[47] Moreover, period of higher volatility like the late 70's are clearly reflected in the bands.

[48] Some of that effect could be offset by the constant moving in the opposite direction and being driven by related variables. However, we see from VI measures that it is driven by alternative sources (especially an exogenous trend). Additionally, in the $h = 12$ case, the constant is much smoother (no switching) and yet, we still observe the same phenomena for $\gamma t, F_1$.



(a) $\gamma_{t,F_1}^{INF,h=1}$: Surrogate Model Replication

(b) $\gamma_{t,F_1}^{INF,h=1}$: Corresponding Tree

(c) $\gamma_{t,F_1}^{INF,h=12}$: Surrogate Model Replication

(d) $\gamma_{t,F_1}^{INF,h=12}$: Corresponding Tree

Figure 16: Surrogate $\beta_{t,k}$ Trees for Inflation. Shade is 68% credible region. cp=0.075. Trees drawn with Rattle. Pink shading corresponds to NBER recessions.

output/inflation trade-off. Clearly, this is not the first time that the role of building permits and housing starts is brought up in the forecasting literature. Stock and Watson (1998a) identifies permits as leading indicators with considerable predictive power for output. Stock and Watson (1999) note that replacing unemployment with housing starts can provide better inflation forecasts for the 1979-1997 period.[49] Leamer (2007) makes a vibrant case for the major role of the housing sector in predicting (and causing) economic downturns. However, when it comes to forecasting inflation, the available evidence points out that including leading indicators (like

[49] They also point out that using an index of 61 real activity variables (the rightful ancestor of $F_1$) rather than unemployment also yields improvements.

permits) does not remedy Phillips' curve forecasts failures.[50] The ARDIRF differs from the above by that the role of permits as a leading indicator is not as a replacement and/or additional output gap proxy. Rather, its role is to increase the curvature of the famous curve when the housing market is booming. The importance of housing starts and permits can be understood as proxying for future economic activity, suggesting a Philipps curve that is non-linear in real activity. By looking at predictive performance results *ex-post*, Stock and Watson (2008) report that Phillips' curve forecasts usually outperform univariate benchmarks around turning points but suffer a reversal of fortune when the output/unemployment gap is close to 0. They note that the finding is nonoperational because of the two-sided nature of gap measure being used. Allegedly, MRF recognize this potential and relies on leading indicators to activate the Phillips' curve when the time is right.

Of course, the predictive Philipps' curve under study here differ in many aspects to those studied, for instance, in Blanchard et al. (2015). Most importantly, $F_1$ summarize indicators that are (for most of them) in first differences. A typical output/unemployment gap measure will be much more persistent. Economically, this means the gap can remain negative for many years following a downturn. In contrast, $F_1$, which is strongly correlated with the first difference of UR, will go back up as soon as UR stops growing. Additionally, unlike a traditional measure of gap, $F_1$ is highly skewed rather symmetric around 0. To validate current insights and obtain new ones, I complete this section by looking at a prototypical Phillips' Curve.

## 5.4 A Look at a More Traditional Phillips' Curve

Much attention has been given lately to the hypothesized flattening of the Phillips' curve (Blanchard et al. (2015), Galí and Gambetti (2019), Del Negro et al. (2020)). In this recent body of work, a strong point is made that the reduced-form Phillips' curve coefficient (or any of its multiple incarnations) has substantially declined over the last few decades. The focus on slow structural change is motivated ex-ante by some implicit economic assumptions and is operationalized by the modelling strategy – either random walk TVPs or sample splitting at a specific date. I contribute to the literature by fitting a MRF which linear part corresponds to a New Keynesian Phillips' curve. As should be clear by now, the benefit of doing so is that MRF could uncover more subtle time-variation structures that would go undetected by plain sample splitting or equivalent approaches. The additional heterogeneity may also help at pinning down potential causes of the apparent decline. The linear equation is inspired by what Blanchard

[50] Stock and Watson (2007) report the Atkeson et al. (2001) forecast (a one-year moving average) to be hard to beat for the period of 1984 to 2004. Among others, a Phillips' curve with permits cannot beat that benchmark.

et al. (2015) (henceforth BCS) considers:

$$\pi_t = \theta_t \hat{\pi}_t^{LR} + (1-\theta_t)\hat{\pi}_t^{SR} + \phi_t u_t^{GAP} + \psi_t \pi_t^{IMP} + \epsilon_t, \tag{5}$$

where $\pi_t$ stands for CPI inflation, $\hat{\pi}_t^{LR}$ and $\hat{\pi}_t^{SR}$ respectively for long-run and short-run inflation expectations. $u_t^{GAP}$ represents the (negative) unemployment gap and $\pi_t^{IMP}$ is import prices inflation. I translate this to the MRF framework by making $\mu_t = \theta_t \hat{\pi}_t^{LR}$ the time-varying constant, letting $\beta_{t,1} = 1 - \theta_t$ and by obtaining $u_t^{GAP}$ by means of Hodrick-Prescott filtering with $\lambda_{HP} = 10^5$.[51] As in BCS, $\hat{\pi}_t^{SR}$ is the average inflation over the last four quarters. Hence, the estimated equation

$$\pi_t = \mu_t + \beta_{1,t}\hat{\pi}_t^{SR} + \beta_{2,t} u_t^{GAP} + \beta_{3,t}\pi_t^{IMP} + \varepsilon_t$$

does not impose the constraint implied by $\theta_t$ in equation (5). However, estimation results will desirably have $\beta_{1,t} \in [0,1]$ at almost any point in time. $S_t$ is the same as that considered in the forecasting section. The data set runs up to 2019Q4.

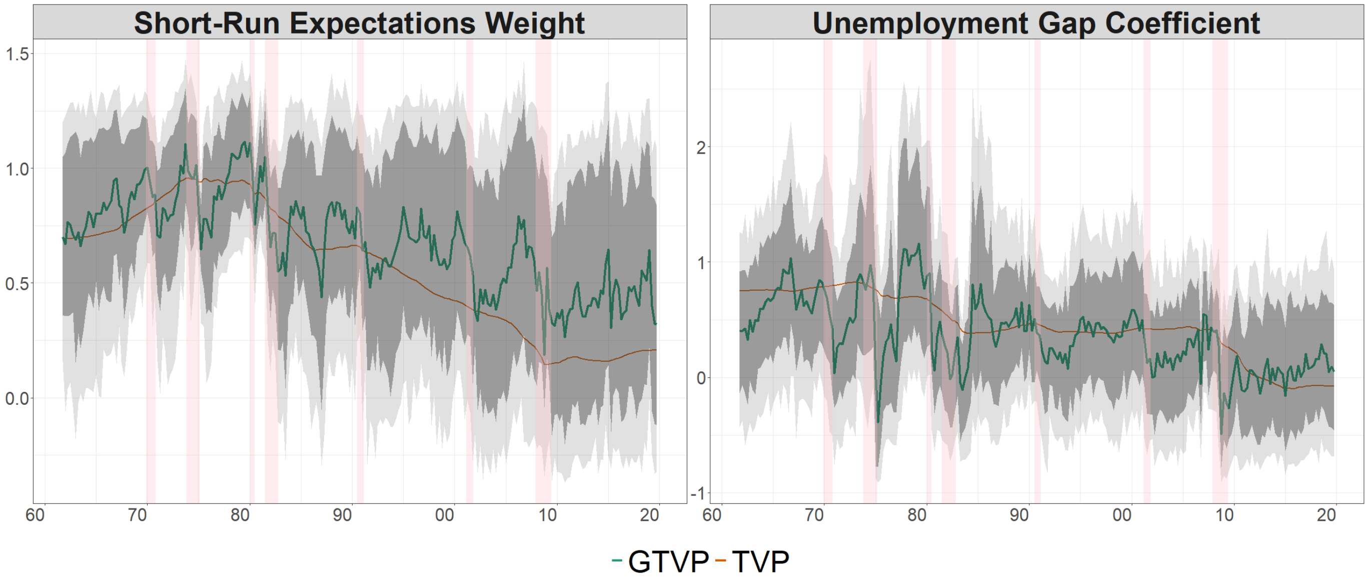


Figure 17: The grey bands are the 68% and 90% credible region. Pink shading corresponds to NBER recessions.

Figure 17 reports the GTVPs of interest: the weight on short-run expectations and the output gap coefficient. Additionally, it contains traditional TVP estimates as mean of comparison. Those convey the usual wisdom: inflation expectations slowly start to be more anchored from the mid 1980's. Around the same time, the unemployment/inflation trade-off begins its slow

[51] Specifically, both this gap and that of BCS get out of negative territory around 2014 (see Figure 9 in BCS). After that, the HP-based gap is mildly positive until the end of the sample.

collapse. The updated data shows that the TVP-based Phillips' curve has further flatten to plain 0 in the last decade.[52]

For $\beta_{1,t}$, the weight on short-run expectations, both methods agree that it has been decreasing steadily after the 1983 recession. GTVPs highlight an additional and interesting pattern for the importance of $\hat{\pi}_t^{SR}$: it tends to increase during economic expansions, collapse during recessions then start increasing again until the next downturn. Note that the phenomenon is also observed in Figure 13b for the simpler ARRF on quarterly inflation. The decrease in the coefficient (usually of about 0.25) is observed for *every* recessions and usually last for some additional quarters after the end of it. The linear rise in the coefficient occurs for all expansions except those preceding the early 90's and 2000's recessions, where the pattern is punctuated with additional peaks and troughs. The increased importance of short-run expectations with the age of the expansion is also observed for the last expansionary periods (including the one that ended in 2019Q4). Hence, the phenomenon is not merely a matter of the 70's and 80's recessions being preceded by a sharp acceleration of inflation. From a statistical point a view, the sharp decline in $\beta_{1,t}$ following every recessions suggest that in the aftermath of an important downward shock, the long-run inflation expectation is a more reliable predictor as it is only affected in a minimal way by recent events. As the expansion slowly progress (and recessionary data points get out of the short-run average), $\hat{\pi}_t^{SR}$ becomes a more up-to-date and reliable barometer of future inflation conditions. Figure 43c presents a small tree attempt at replicating downward path of $\beta_{1,t}$. Given the smoothness of this GTVP, the tree crudely approximate structural change with one break and the upward movements before some recessions with a MAF of retail sales.

When it comes to the low-frequency movements the unemployment gap coefficient, both methods agree about a significant decline starting in the 80's. However, GTVPs uncover some significant additional heterogeneity in the parameter of interest. **First** and most strikingly, $\beta_{2,t}$ gets very close to 0 following every recessions. This suggest a non-linear Philipps' curve where inflation respond strongly to a positive $u_t^{GAP}$ but not so much to a negative one. **Second**, the historically high $\beta_{2,t}$ of the 70's and early 80's are attributed to a series of peaks (before the first three recessions of the sample) rather than a sustained high coefficient. As obvious from Figure 17, a traditional TVP taking an average over time of such peak and troughs can hide that the high output-inflation trade-off that used to prevail in the 70's and 80's is attributable to a series of specific inflationary spirals. Such pre-recession accelerations still occur in the latter part of the sample but in a much milder way. **Third**, Figure 43d tries to rationalize the phenomenon with a simple surrogate tree. Again, we see that unlike some results presented in section 5.3, these trees

[52] This difference can be explained by the additional 6 years of data and the fact that traditional TVPs tend to rewrite (mostly recent) history as they are re-estimated on additional data. Indeed, stopping the estimation in 2014 (as in BCS) gets the TVP version of output gap coefficient to be 0.2 rather than in the vicinity of 0.

struggle to visually match the smooth GTVP path as nicely as one could hope for, suggesting the forest is performing some desirable smoothing. Nevertheless, the peaks in $\beta_{2,t}$ are clearly pinned down by two highly correlated indicators: capacity utilization of all industries (TCU) and manufacturing sector (CUMFNS). In fact, plain correlation between $\beta_{2,t}$ and TCU is 0.8 and the correspondence between the two variables is striking in Figure 18. This suggests a clear predicament, a positive $u_t^{GAP}$ will lead to inflationary pressures when it is accompanied by rising capacity utilization. Hence, the conjunction of rising demand for both capital and labor inputs push the inflation level upward.[53] Many notable increases in $\beta_{2,t}$ are nicely matched by TCU (between the two 70's recessions and before 2008). However, as Figure 18 shows, that characterization is imperfect since TCU by itself cannot explain some important spikes in the GTVPs (end of 70's, mid 80's) and predicts a higher $\beta_{2,t}$ in the years following the 2008-2009 recession. In fact, the 70's and 80's peaks can be successfully captured by a MAF of employment growth which highlights the exceptionally rapid rise of employment exiting the 1974 and 1981 recessions. This, again, points in the direction of a non-linear Phillips' curve.

This above pattern remains when adding controls in the linear part for supply shocks and monetary policy shocks. Those are the usual confounding factors suspected of blurring the relationship of interest by introducing a positive correlation between unemployment and inflation.[54] The economic suspicion particular to this application is that omitting or down-weigthing them could generate a downward bias in $\beta_{2,t}$ that only occurs locally, partially generating the observed pattern. As it turns out, the controls help making the case for the above explanations visually even cleaner in Figure 44. The strong correlation with TCU remains but is visibly imperfect in the second half of the sample where $\beta_{2,t}$ is visibly more cyclical, with clear accelerations that peak before the 2001 and 2008 recessions. Of course, with the additional regressions targeted at purging out policy shocks and supply shocks, this last iteration is much closer to estimating the effect of an index of demand shocks on inflation. The clear cyclical pattern being present for both the plain and augmented Phillips curve is in line with Galí and Gambetti (2019) that show very similar historical paths for the wage Phillips' curve and their proposed semi-structural counterpart.

GTVPs in Figure 17 suggest there is not one but *two* phenomena responsible for the exceptionally mild response of inflation during the Great Recession and the subsequent recovery. The collapse of $\beta_{2,t}$ when negative shocks hit the economy is not unique to 2008: it happened following *every* recession since 1960. Overall, this correlation between the state of the economy and the steepness of the Phillips' curve suggest an important empirical regularity: inflation

[53] Given the co-cyclical relationship between TCU and $u_t^{GAP}$, this is also evidence in favor of a non-linear (albeit slowly decreasing) Phillips' Curve.

[54] While the time-varying constant can go a long way at controlling for such factors – being a RF in itself, including them in the linear part makes them "stand out" as everything going through the constant is inevitably heavily regularized.

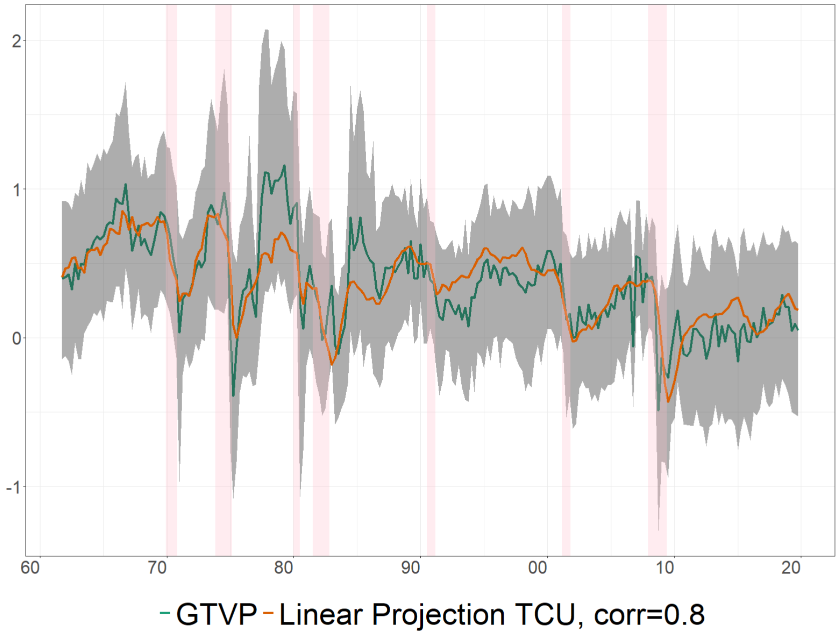


Figure 18: "What Goes Around Comes Around": Capacity Utilization is substantially correlated with the inflation-unemployment trade-off. The grey bands are the 68% credible region. Pink shading corresponds to NBER recessions.

will rise when the economy is running above its potential, but much more timidly will it go down from economic slack. Recently, Lindé and Trabandt (2019) have shown that this exact phenomenon can be rationalized by a New Keynesian DSGE model. Indeed, by allowing for additional strategic complementarity in firms price- and wage-setting behavior and solving the non-linear model (rather than considering the linear approximation around the steady state), the authors remarkably obtain a state-dependent Phillips' curve that becomes very flat during large downturns – as reported here. This can explain both the small coefficient during recessions and its subsequent timid increase.

Nonetheless, with $\beta_{2,t}$ clearly trending down, this is unlikely to be the full story. In line with previous work championing the slow death hypothesis, it is visually clear in Figure 17 that $\beta_{2,t}$ remaining in the vicinity of 0 for much of the 2010's is partly due to how low it was before the 2008 drop. Hence, the exceptionally weak response of inflation to $u_t^{GAP}$ in the years following 2008 is most likely due to both state-dependent demand and secular decrease. Thus, an interesting question is to wonder how this realization helps at sorting out potential causes for the large decline of $\beta_{2,t}$ in traditional TVP models. Much of the commentary regarding the slow weakening of the inflation-output relationship has to do with globalization (Del Negro et al. (2020)). MRF points at a complementary interpretation. The high coefficient usually reported until the mid-1980's is clearly sustained by the abundance of inflation spirals and exceptionally fast recoveries. This is clearly not the case for the last 3 recessions in the sample. By construction, the Great Moderation itself made the upward non-linearities of the plain Phillips' curve less solicited in the last 3 decades. Thus, on average, the coefficient must be lower. This certainly does not invalidate the view that the coefficient has been going down in structural/long-term fashion – we observe it as well with GTVPs. What it suggests, however, is that much of the quantitative decline usually reported with traditional TVPs is likely due to the absence of inflation/growth spirals starting from the Great Moderation era.

# 6 Conclusion

I proposed a new time series model that **(i)** expands multiple non-linear time series models, **(ii)** adapts Random Forest for Macro forecasting and **(iii)** can be interpreted as Generalized Time-Varying Parameters. On the empirical front, the methodology provides substantial empirical gains over RF and competing non-linear TS models. The resulting Generalized TVPs have a very distinct behavior vis-Ãă-vis standard random walk parameters. For instance, they adapt nicely to regime-switching behavior that seems pervasive for unemployment – while not neglecting potential long-run change. This finding is facilitated by the fact that GTVPs lend themselves much more easily to interpretation than either standard RF or random-walk TVPs. Indeed, rather than trying to open the back-box of an opaque conditional mean function (like one would with plain RF), MRFs can be compartmentalized in different components of the small macro model. Furthermore, the TVPs characterizing completely the forecast of a MRF can be visualized with standard time series plots and credible intervals are provided by a variant of the Bayesian Bootstrap. Unlike standard TVPs, these paths can also be constructed from observed quantities rather than latent states only. By looking at different variable importance measures and fitting simple trees to each TVP in isolation, I provide a way to explain the output of a seemingly opaque prediction machine as a sum of simple (albeit reduced-form) economic relationships.

When looking at Phillips' curves in general, MRF finds both structural change in the persistence and regime-dependent behavior in the economic activity/inflation trade-off. In particular, a recurrent theme across all specifications is that the slowly decaying curve is also much steeper when the economy is overheating – in line with the convexity/non-linearity hypothesis. Hence, MRF can be of great help sorting out what is plausible and what is not when it comes to macroeconomic equations with a history of controversy. Since there is no shortage of those, MRF holds many possibilities for future research.

# References

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

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, pages 178–196.

Aruoba, S. B., Bocola, L., and Schorfheide, F. (2017). Assessing dsge model nonlinearities. *Journal of Economic Dynamics and Control*, 83:34–54.

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

Atkeson, A., Ohanian, L. E., et al. (2001). Are phillips curves useful for forecasting inflation? *Federal Reserve bank of Minneapolis quarterly review*, 25(1):2–11.

Auerbach, A. J. and Gorodnichenko, Y. (2012a). Fiscal multipliers in recession and expansion. In *Fiscal policy after the financial crisis*, pages 63–98. University of Chicago Press.

Auerbach, A. J. and Gorodnichenko, Y. (2012b). Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy*, 4(2):1–27.

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

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Batini, N., Callegari, G., and Melina, G. (2012). Successful austerity in the united states, europe and japan.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Blanchard, O. (2016). The phillips curve: Back to the'60s? *American Economic Review*, 106(5):31–34.

Blanchard, O., Cerutti, E., and Summers, L. (2015). Inflation and activity–two explorations and their monetary policy implications. Technical report, National Bureau of Economic Research.

Bognanni, M. (2013). An empirical analysis of time-varying fiscal multipliers.

Borup, D., Christensen, B. J., Mühlbach, N. N., Nielsen, M. S., et al. (2020). Targeting predictors in random forest regression. Technical report, Department of Economics and Business Economics, Aarhus University.

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

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chamberlain, G. and Imbens, G. W. (2003). Nonparametric applications of bayesian inference. *Journal of Business & Economic Statistics*, 21(1):12–18.

Champagne, J. and Sekkel, R. (2018). Changes in monetary regimes and the identification of monetary policy shocks: Narrative evidence from canada. *Journal of Monetary Economics*, 99:72–87.

Chan, J. C., Eisenstat, E., and Strachan, R. W. (2018). Reducing dimensions in a large TVP-VAR. CAMA Working Papers 2018-49, Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University.

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

Chevillon, G. (2007). Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, 21(4):746–785.

Cirillo, P. and Muliere, P. (2013). An urn-based bayesian block bootstrap. *Metrika*, 76(1):93–106.

Clements, M. P. and Smith, J. (1997). The performance of alternative forecasting methods for setar models. *International Journal of Forecasting*, 13(4):463–475.

Clyde, M. and Lee, H. (2001). Bagging and the bayesian bootstrap. In *AISTATS*.

Cochrane, D. and Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American statistical association*, 44(245):32–61.

Cogley, T., Primiceri, G. E., and Sargent, T. J. (2010). Inflation-gap persistence in the us. *American Economic Journal: Macroeconomics*, 2(1):43–69.

Cogley, T. and Sargent, T. J. (2001). Evolving post-world war ii us inflation dynamics. *NBER macroeconomics annual*, 16:331–373.

D'Agostino, A., Gambetti, L., and Giannone, D. (2013). Macroeconomic forecasting and structural change. *Journal of Applied Econometrics*, 28(1):82–101.

Dagum, E. B. and Bianconcini, S. (2009). Equivalent reproducing kernels for smoothing spline predictors. In *Proceedings of the American Statistical Association, Business and Economic Statistics Section*.

Del Negro, M., Lenza, M., Primiceri, G. E., and Tambalotti, A. (2020). WhatâĂŹs up with the phillips curve? Technical report, National Bureau of Economic Research.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Diebold, F. X. and Rudebusch, G. D. (1994). Measuring business cycles: A modern perspective. Technical report, National Bureau of Economic Research.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100.

Dolado, J. J., Marıa-Dolores, R., and Naveira, M. (2005). Are monetary-policy reaction functions asymmetric?: The role of nonlinearity in the phillips curve. *European Economic Review*, 49(2):485–503.

Estrella, A. and Mishkin, F. S. (1998). Predicting us recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, 80(1):45–61.

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In *Handbook of economic forecasting*, volume 2, pages 2–56. Elsevier.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Fortin-Gagnon, O., Leroux, M., Stevanovic, D., and Surprenant, S. (2018). A large canadian database for macroeconomic analysis. Technical report, Department of Economics, UQAM.

Freedman, D. A. et al. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2018). Local linear forests. *arXiv preprint arXiv:1807.11408*.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Galí, J. and Gambetti, L. (2019). Has the us wage phillips curve flattened? a semi-structural exploration. Technical report, National Bureau of Economic Research.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). Midas regressions: Further results and new directions. *Econometric Reviews*, 26(1):53–90.

Giraitis, L., Kapetanios, G., and Yates, T. (2014). Inference on stochastic time-varying coefficient models. *Journal of Econometrics*, 179(1):46–65.

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

Goulet Coulombe, P. (2019). Time-varying parameters: A machine learning approach.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., Surprenant, S., et al. (2019). How is machine learning useful for macroeconomic forecasting? Technical report, CIRANO.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., Surprenant, S., et al. (2020). Prévision de lâĂŹactivitĂl' Ăl'conomique au québec et au canada Ãă lâĂŹaide des méthodes âĂIJmachine learningâĂİ. Technical report, CIRANO.

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

Granger, C. W. (2008). Non-linear models: Where do we go next-time varying parameter models? *Studies in Nonlinear Dynamics & Econometrics*, 12(3).

Hahn, P. R., Carvalho, C. M., and Mukherjee, S. (2013). Partial factor modeling: predictor-dependent shrinkage for linear regression. *Journal of the American Statistical Association*, 108(503):999–1008.

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

Hansen, C. and Liao, Y. (2019). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory*, 35(3):465–509.

Ishwaran, H. and Malley, J. D. (2014). Synthetic learning machines. *BioData mining*, 7(1):28.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American economic review*, 95(1):161–182.

Karabatsos, G. (2016). A dirichlet process functional approach to heteroscedastic-consistent covariance estimation. *International Journal of Approximate Reasoning*, 78:210–222.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter vars. *Journal of Econometrics*, 177(2):185–198.

Kotchoni, R., Leroux, M., and Stevanovic, D. (2019). Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics*, 34(7):1050–1072.

Lancaster, T. (2003). A note on bootstraps and robustness. *Available at SSRN 896764*.

Leamer, E. E. (2007). Housing is the business cycle. Technical report, National Bureau of Economic Research.

Lindé, J. and Trabandt, M. (2019). Resolving the missing deflation puzzle.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46.

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

McCracken, M. and Ng, S. (2020). Fred-qd: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research.

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

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., and Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, (just-accepted):1–45.

Meek, C., Chickering, D. M., and Heckerman, D. (2002). Autoregressive tree models for time-series analysis. In *Proceedings of the 2002 SIAM International Conference on Data Mining*, pages 229–244. SIAM.

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

Molnar, C. (2019). *Interpretable machine learning*. Lulu.com.

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

Perron, P. et al. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics*, 1(2):278–352.

Poirier, D. J. (2011). Bayesian interpretations of heteroskedastic consistent covariance estimators using the informed bayesian bootstrap. *Econometric Reviews*, 30(4):457–468.

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

Ramey, V. A. (2016). Macroeconomic shocks and their propagation. In *Handbook of macroeconomics*, volume 2, pages 71–162. Elsevier.

Ramey, V. A. and Zubairy, S. (2018). Government spending multipliers in good times and in bad: evidence from us historical data. *Journal of Political Economy*, 126(2):850–901.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rubin, D. B. (1981). The bayesian bootstrap. *The annals of statistics*, pages 130–134.

Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Shiller, R. J. (1973). A distributed lag estimator derived from smoothness priors. *Econometrica (pre-1986)*, 41(4):775.

Sims, C. A. (1993). A nine-variable probabilistic macroeconomic forecasting model. In *Business cycles, indicators and forecasting*, pages 179–212. University of Chicago press.

Stevanovic, D. (2016). Common time variation of parameters in reduced-form macroeconomic models. *Studies in Nonlinear Dynamics & Econometrics*, 20(2):159–183.

Stock, J. H. (1994). Unit roots, structural breaks and trends. *Handbook of econometrics*, 4:2739–2841.

Stock, J. H. and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, 4:351–394.

Stock, J. H. and Watson, M. W. (1998a). Business cycle fluctuations in us macroeconomic time series. Technical report, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (1998b). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. Technical report, National Bureau of Economic Research.

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

Stock, J. H. and Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39:3–33.

Stock, J. H. and Watson, M. W. (2008). Phillips curve inflation forecasts. Technical report, National Bureau of Economic Research.

Taddy, M., Chen, C.-S., Yu, J., and Wyle, M. (2015). Bayesian and empirical bayesian forests. *arXiv preprint arXiv:1502.02312*.

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672.

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

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

Wei, P., Lu, Z., and Song, J. (2015). Variable importance analysis: a comprehensive review. *Reliability Engineering & System Safety*, 142:399–432.

Wochner, D. (2020). Dynamic factor trees and forests–a theory-led machine learning framework for non-linear and state-dependent short-term us gdp growth predictions.

Woloszko, N. (2020). Adaptive trees: a new approach to economic forecasting.

# A Appendix

## A.1 More on Surrogate $\beta_t$ Trees

### A.1.1 Link to Literature

The approach described in section 5.3 belongs to a family of methods usually referred to as "surrogate models" (Molnar (2019)). Attempting to fit the whole conditional mean obtained from a black-box algorithm using a more transparent model is a global surrogate. An obvious critique of this approach is that if the complicated model justifies its cost in interpretability with its predicting gains, it is hard to believe a simple model can reliably recreate its predictions. Conversely, if the surrogate model is quite successful, this casts some doubts about the relevance of the black box itself. In this line of work, a more promising avenue is a local surrogates model as proposed in Ribeiro et al. (2016), which fits interpretable models *locally*. By following Granger (2008)'s insights, we already have this: by looking at the $\beta_t$ paths directly, we effectively have a local model – in time. The purpose of surrogate models is to learn about the model, not the data. The former is much easier in MRF than in standard RF since the vector $\beta_t$ fully characterizes the prediction at a particular point in time.[55] Moreover, the coefficients are attained to predictors that can have themselves a specific economic meaning. Considering this and the earlier discussion of section 2.1, it is natural in a macro time series context to fit surrogate models to time-varying parameters themselves – a blatant divide-and-conquer strategy.

### A.1.2 About $VI_{OOB}$, $VI_{OOS}$ and $VI_\beta$

I now explain the motivation and mechanics behind the different VI measurements. The first measure, $VI_{OOB}$, is the standard out-of-bag (hence OOB) VI permutation measure widely used in RF applications (Wei et al. (2015)). It consists of randomly permuting one feature $S_j$ and comparing predictive accuracy to the full model on observations that were not used to fit the tree.[56] This pseudo evaluation set is convenient because it is a direct byproduct of the construction of the forest. Under a well-specified model that includes enough lags of $y_t$, autocorrelation of residuals will not be an issue. This condition is likely to be met here since the analysis focuses on results for $h = 1$. [57] $VI_{OOS}$ considers a different testing set more natural for time series data:

[55] More generally, any partially linear model in the spirit of MRF has a potential for local surrogate analysis along the linear regression space rather than the observations line.

[56] This is thought as the equivalent for a black-box model to setting a specific coefficient to 0 in a linear regression and then comparing fits. However, VI as implemented here (and in most applications) does not re-estimate the model after dropping $S_j$. This differs from a t-test since it is well known that the latter is equivalent to comparing two $R^2$'s – the original one and that of a re-estimated model, under the constraint.

[57] Notwithstanding, at longer horizons, $VI_{OOB}$ could paint a distorted picture in the presence of autocorrelation – the same way K-fold cross validation can be inconsistent for time series data (Bergmeir et al. (2018)). This worry

the real OOS, which in this section spans from 2007q2 to the end of 2014. By construction, this measure focuses on finding variables which contribution paid off during a specific forecasting experiment, rather than throughout the whole sample. This is not bad *per se* but is a different concept that can be of independent interest. Finally, both $VI_{OOB}$ and $VI_{OOS}$ focus on overall fit. $VI_{\beta}$ implements the same idea as $VI_{OOB}$ but is calculated using a different loss function. That is, $VI_{\beta_{k,j}}$ reports a measure of how much the path of $\beta_k$ is altered (out-of-bag) when variable $S_j$ is randomly permuted in the forest part. Finally, I use the various VI measurements as devices to narrow down the set of predictors for the construction of intuitive trees.

I restrict the number of considered variables (for the next step) to be 20 for each VI criteria. When VI suggest that a parsimonious set of variables matter, it is very rarely more than 3 or 4 variables. Thus, restricting it to 20 is a constraint that only binds if all variables contribute, but marginally, in the spirit of a Ridge regression (Friedman et al. (2001)). When it comes to that, the cut-off is simply the natural reflection of a trade-off between interpretability and fit.

## A.2 Further Investigation of the Importance of $S_t$

Do MAFs matter? For 3 standard ML models (standard RF, LASSO, Ridge) that can handle high-dimensional data sets, I investigate the usefulness of the MAFs advocated in section 2.6. The codes to describe the different information sets are

- **"CSF"**: Only standard cross-sectional factors (5 factors, 8 lags of them)
- **"MAF"**: $S_t$
- **"ALL"**: $S_t$ + all the raw data (8 lags)
- **"X"**: 8 lags of the raw data

Figure 19 summarizes results over 18 targets (6 variables and the first 3 horizons). The first striking fact is that the four best models are RF, followed by the LASSO block, Ridge and ARDI. This suggests, with an unprecedented level of surprise, that models matter. For RFs, the best model is the one using $S_t$ followed closely by the one that also adds the raw data to it. However, if we drop the MAFs, we incur a significant loss and obtain the worst of RFs, (so-called RF-X). The RF with cross-sectional factors only performs quite well in an unequal fashion.

can be alleviated by using a block approach like in section 2.8.

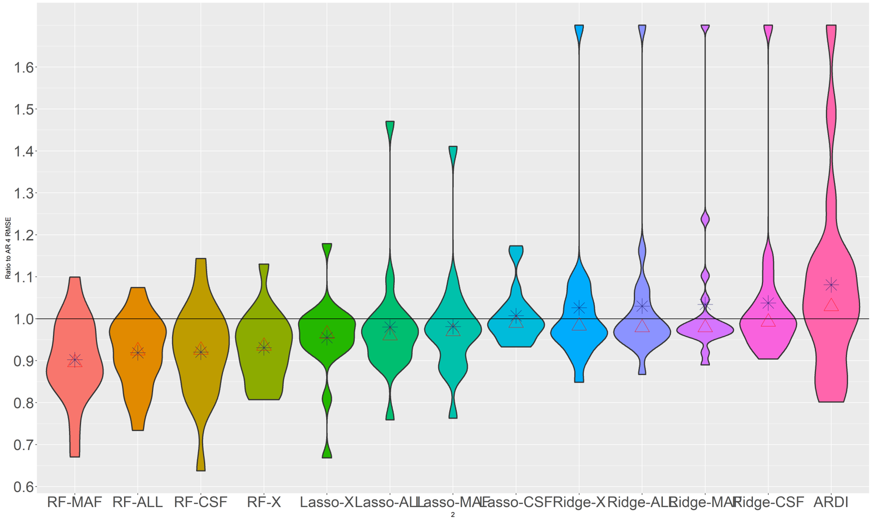


Figure 19: The usefulness of $S_t$.

The best LASSO models must include the raw data. Models with either standard factors or MAFs only do not perform as well. This is not true for Ridge where the best model is the one that uses $S_t$. It is however important to note at this point of the ranking that these models are already lagging RFs in a significant way.

## A.3 On the Relative Irrelevance of Tuning Parameters

One reason among others for RF popularity is that great performance is achieved without much or any tuning. This is not true of Boosting and even less so of Neural Networks. Nevertheless, as discussed in Ishwaran and Malley (2014), the minimum node size of trees can matter (especially in data sets with many observations – not our case here). If the number of relevant feature is very small with respect to the total number of them, `mtry` the fraction of randomly selected predictors could matter. For the application of standard RF to Quarterly data, it turns out they do not.[58]

The relevant question is whether the tuning parameters specifically introduced by MRF can have a significant impact on the observed performance. Since some of them were not discussed explicitly in the main text, I quickly review them here.

- `MLF`: stands for Minimum Leaf Fraction. It is the HP in MRF that has a role equivalent to that of minimum node size in standard RF. The so-called "fraction" is the ratio of parameters in the linear part to that of observations in any node (which includes most importantly the terminal ones). To clarify things, I proceed with an example. Set `MLF` $= 2$, the

[58] Results are available upon request.

linear part has 3 parameters, and we are trying to split 15 observations apart. This setting implies that any split that result in having less than 6 observations in the children note will not be considered. This specific setting ensures that the ratio of parameters to observations never exceeds 1/2 in any node. This ensure stability, especially if the following 3 tuning parameters are set to 0. However, when RWR and RL are active, it is possible to consider MLF $= 1$ or even lower, like for the large VARs specifications of section **??**. The extra regularization allows in the latter case to have base regressions that have parameters/observations ration exceeding 1 (high-dimensional setting). This is very desirable in a quarterly macro setting because setting MLF $> 2$ or higher seriously restrict the depth of the trees being grown.

- RWR: stands for Random Walk Regularization strength as discussed in 2.3. It is the $\zeta$ in equation (3).
- RL: stands for Ridge Lambda ($\lambda$) in equation (2).
- HRW: stands for Hierarchical Regularization weight. This HP works as a middle ground between a small and a large MLF and relax the need for a RL or RWR that would be too strong. The intuition is rather straightforward. Let us say that terminal nodes $B$ and $C$ both have as a parent node $A$. Then, $\beta_A$ is necessarily some convex combination of $\beta_B$ and $\beta_C$. It is expected to be somewhat close to both $\beta_B$ and $\beta_C$ and have a smaller variance by construction. Thus, rather than using $\beta_C$ directly, there may be stability gains from using $\beta_{\tilde{C}} = \texttt{HRW}\beta_A + (1 - \texttt{HRW})\beta_C$.

The tuning parameters evaluation exercise is a simplified and limited (in scope) re-edition of section (4.2). The estimation of each model is done for the last time at 2007Q2 and their performance is evaluated starting from that date until 2014 for a total of 30 quarters. I limit my attention to the 4 key variables (GDP, UR, SPREAD, INF) and 3 horizons (1,2 and 4 quarters) that I evaluated in detail earlier. For each of these, a grand total of 81 models will be evaluated. The first (and the benchmark) is the AR(4). The 80 models a different combinations of HPs. That is, each $m$ is a combination of the following list of possibilities.[59]

- MRF $\in$ {ARRF, VARRF1, VARRF2, ARDIRF2, ARDIRF1}
- MLF $\in \{1, 2\}$
- RWR $(\zeta) \in \{0, 0.7\}$
- RL $(\lambda) \in \{2, 1\}$

[59] While those values may all seem rather loose at first look, it is worth reminding the reader that the final output is an average of many such trees, which is itself a strong source of regularization. This is in the same spirit as Breiman (2001) encouraging the use of fully-grown trees in standard RF.

- `HRW` $\in \{0.01, 0.2\}$

The singular organization of Figure 20 serves the purposes of making one specific point: what matters is the choice of the linear part and not much else. The within color block heterogeneity is nowhere as large as that of between. For instance, the reported gains of the baseline ARDIRF (violet) for UR (all $h$'s) and GDP ($h = 1$) in section 4.2 can be obtained by most if not all HP combinations. That is also true of VARRF for the SPREAD. The parsimony of lags (1 rather than 2) is helpful for both VARRF and ARDIRF. Whenever there is significant gain, it comes from the smaller model.

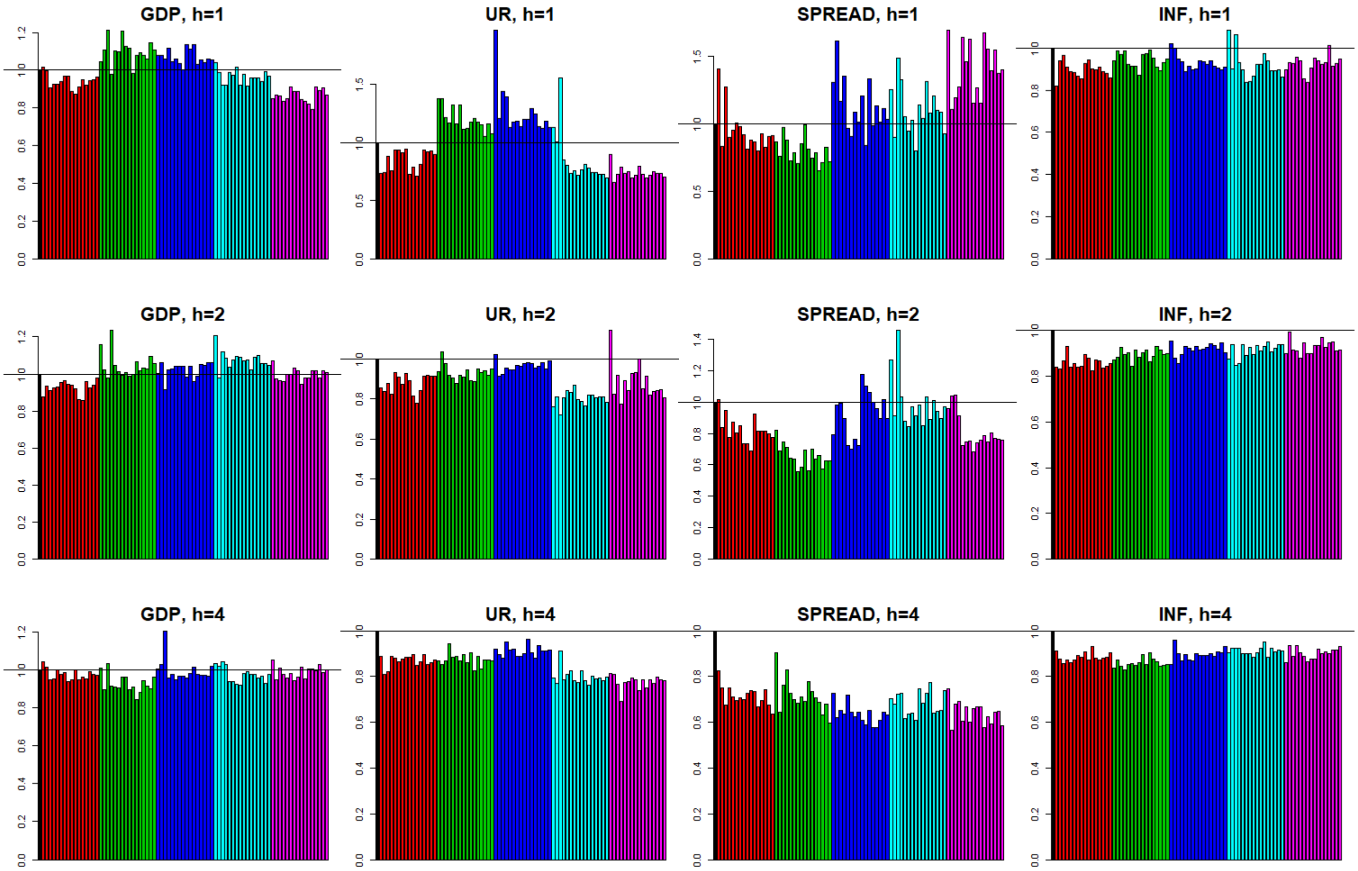


Figure 20: All $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$ for the Quarterly data test. The different color blocks are different MRF models. The red block is ARRF, the green block is VARRF with 1 lag, the blue block is VARRF with 2 lags, the turquoise block is ARDIRF with 2 lags and lastly the violet block is ARDIRF with 1 lag. The different bars within each block represent a specific combination of $\{$`MLF`, `RWR`, `RL`, `HRW`$\}$. The first black line is the AR(4) and is normalized to 1.

To make a final point and digest all the available information in 20, I run a RF on the results themselves (rMPSE) and use as explanatory variables the features $\{h, v,$ MRF, `MLF`, `RWR`, `RL`, `HRW`$\}$. These are just a way to decompose the $m$ in $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$. While such an analysis of variance could be done with a regression, the numerous possible interactions of all features makes a complete analysis a daunting task. Thus, I let RF do that. An analogous exercise was done in Goulet Coulombe et al. (2019) to identify Machine Learning features that truly matters in a sea of RMSPE's. The VI calculation technique is the standard random permutation approach.

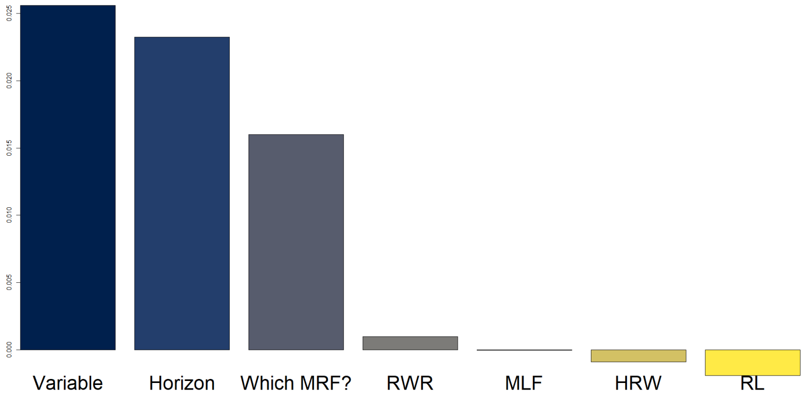


Figure 21: Variable Importance of MRFs features to explain $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$'s for the Quarterly data test.

Unsurprisingly, $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$ are strongly explained by $h$ and $v$. In third place but still quite important, is the choice of MRF. Far behind, we get the 4 tuning parameters which have a very (relative) negligible influence. Negative values are possible in VI calculations the same way an out-of-sample $R^2$ can be negative.

Nevertheless, some marginal gains can be obtained from choosing the right combination of tuning parameters. Table 2 informs us about the best (out of 80) combination of MRF and tuning parameters. Except for the very persistent SPREAD, `RWR` is universally preferred to be loose. The best MRF match that what is reported in the main tex. This is line with the previous discussion that highlights the negligible influence of tuning parameters compared to that of models.

As indicative evidence, I report here results for 4 models that span the space of possibilities quite well. The statistic is the mean difference with respect to the standard RMSE reduction for a given model/variable/horizon. The so-called "clever" specification is using a mix of different regularizations while allowing for node that runs regressions with as many parameters as observations. It is the best of the four specifications, on average. Note that units are small given that different MRF give about 20% gains on the AR(4).

The relative irrelevance result of HPs has several advantages. First, this reduces dramatically computing demand. Second, tuning properly a ML model on time series data is notoriously difficult (discussed in Goulet Coulombe et al. (2019) and many others) and requires a great amount of care in the implementation. Thus, avoiding the systemic need for HP optimization in MRF is highly desirable.

Table 2: Best Model and Tuning Parameters Combination by Targets

| | MRF | MLF | RWR | RL | HRW |
|---|---|---|---|---|---|
| **GDP** | | | | | |
| h=1 | ARDIRF1 | Tight | Loose | Tight | Tight |
| h=2 | ARRF | Tight | Loose | Tight | Tight |
| h=4 | VARRF1 | Tight | Loose | Tight | Loose |
| **UR** | | | | | |
| h=1 | ARDIRF1 | Loose | Loose | Loose | Tight |
| h=2 | ARDIRF2 | Loose | Loose | Tight | Loose |
| h=4 | ARDIRF1 | Loose | Loose | Tight | Tight |
| **SPREAD** | | | | | |
| h=1 | VARRF1 | Tight | Tight | Loose | Loose |
| h=2 | VARRF1 | Loose | Tight | Tight | Loose |
| h=4 | ARDIRF1 | Loose | Loose | Loose | Tight |
| **INF** | | | | | |
| h=1 | ARRF | Loose | Loose | Loose | Loose |
| h=2 | ARRF | Tight | Loose | Tight | Loose |
| h=4 | VARRF1 | Loose | Loose | Tight | Tight |

Table 3: Average Performance of Key HP combinations

| **Name** | **Values** | **Mean rMSPE Change** |
|---|---|---|
| Tight | $\{\texttt{MLF} = 2, \texttt{RWR} = 0.75, \texttt{RL} = 0.01, \texttt{HRW} = 0.2\}$ | -0.0166 |
| Quasi-OLS | $\{\texttt{MLF} = 2, \texttt{RWR} = 0, \texttt{RL} = 0.001, \texttt{HRW} = 0.01\}$ | 0.0012 |
| Clever | $\{\texttt{MLF} = 1, \texttt{RWR} = 0.75, \texttt{RL} = 0.01, \texttt{HRW} = 0.2\}$ | -0.0313 |
| All Hell Breaks Loose | $\{\texttt{MLF} = 1, \texttt{RWR} = 0, \texttt{RL} = 0.001, \texttt{HRW} = 0.01\}$ | 0.0742 |

## A.4 Extension: Eye of the Beholder MRF

Fixing the $X_t$ part of a MRF based on some small macro model (like the VAR or the ARDI) has interpretative virtues and delivers satisfactory forecasting results. However, in many instances, it may be desirable that the crucial choice of the linear part in MRF be selected automatically.

Following the intuition of section 2.5.1 that motivated the very existence of the linear part, we know that if a conditional mean has a strong linear component, the corresponding vanilla RF fit will spend a great number of splits on the members of $X_t$. For instance, fitting a RF to a highly persistent target leads to Variable Importance (VI) suggesting the high importance of AR terms. This pattern is especially clear when looking at variable importance measurements for inflation and the very persistent spread. It is also quite clear from VI calculation on simulated AR model data.

Thus, I implement the *Eye of the Beholder* MRF by first running a standard RF (as a screening step) on the data at hand and then construct the MRF based on VI recommendations. I implement the following simple rule. First, $X_t$ consist of the $\tilde{K}$ most important variables according to VI. The number of selected variables is determined by looking VI contribution differentials. The latter are ordered (like the eigenvalues would be for selecting the number of factors) and I pick the first $\tilde{K}$ before that largest drop. $\tilde{K}$ is forced to be less or equal to 10. Finally, I randomly select one third of the $\tilde{K}$ regressors to constitute the linear part in each tree, which adds another layer of randomization in MRF. The second stage $S_t$ will consist of variables in the first stage $S_t$ that have a strictly positive VI measure. This step does not impact the fit by a lot, but rather helps in decreasing computational time. Also, to further decrease computing time, all the models in this section have $\zeta = 0$ and use sub-sampling rather than BBB. Since no attempt will be made (and none should) at interpreting the coefficients of EOTB's, these features are reasonably disposable. I compare the Eye of the Beholder (EOTB) MRF to other Random Forests reported in 4.2. Furthermore, I include EOTB-ARRF which constructs a similar forest as EOTB-MRF but imposes that half of them are autoregressive ones (ARRF). This can be understood as having a prior on the VI procedure that favor AR terms, the same way, for instance, AR terms are less strongly penalized in a BVAR. In terms of overall performance, EOTB-MRF outdoes both ARDIRF and VARRF but is yet behind the more conservative ARRF, RF and AR+RF. EOTB-ARRF that blends a data-driven selection of $X_t$ and a prior favorable to AR terms is the best model overall. Specifically, the latter inherits most of the gains from EOTB-MRF without some of its important failures as shown by the tail of EOTB-MRF between 1.1 and 1.4.

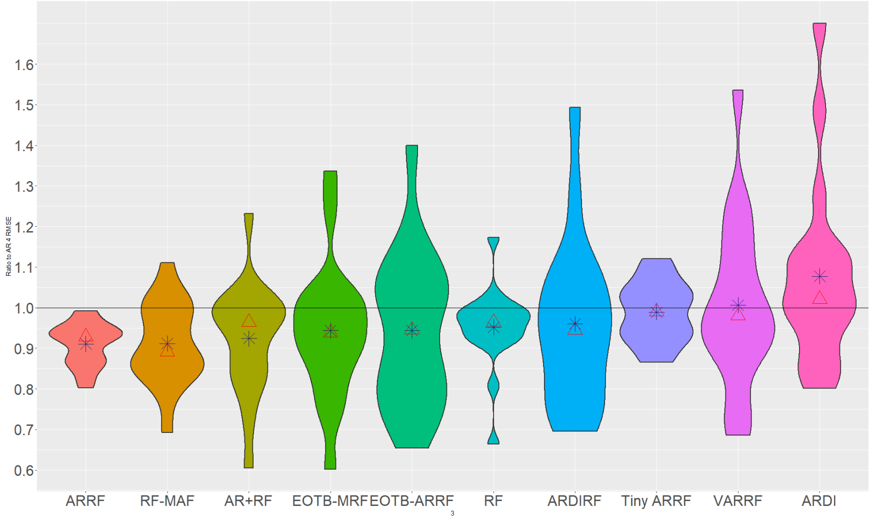
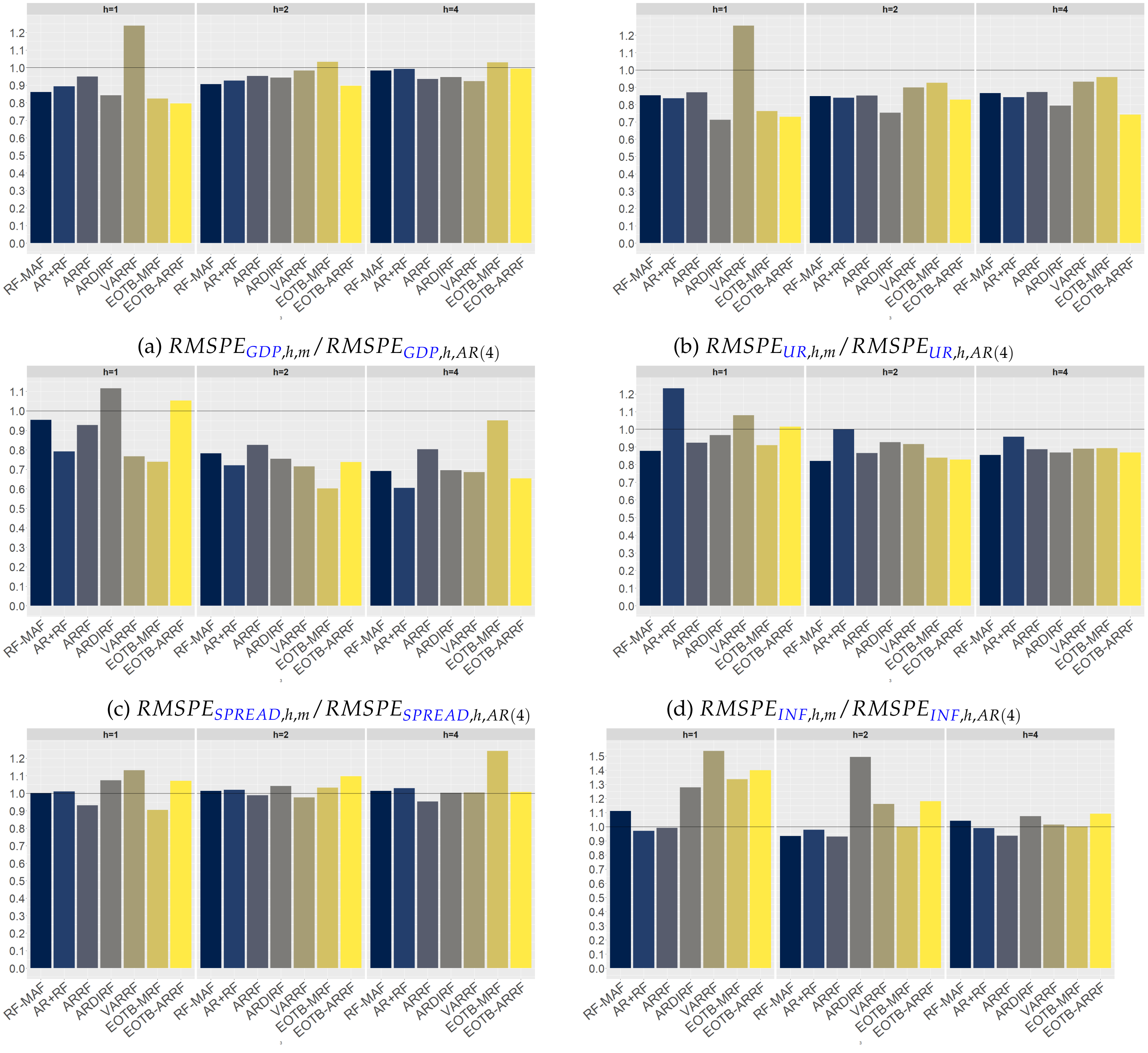


Figure 22: The distribution of $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$ for monthly data and horizons 1, 2 and 4 quarters. The star is the mean and the triangle is the median.

The bar plots display specific results. The key takeaway is that the EOTB procedure coupled with ARRF procures results that are usually as good as the best model of the lot. This is not always the case, however, as longer horizons UR results show. Nevertheless, EOTB is a great addition to this paper's toolbox since it is not hard to imagine applications where an obvious choice of a linear part may not be available. Accordingly, a variant of EOTB-MRF is deployed in Goulet Coulombe et al. (2020) and is shown to provide the best Canadian unemployment one-quarter-ahead forecast out of 50 models with a reduction of 26% in RMSPE over the benchmark. I believe a finer tuning and formalization of the procedure could easily yield even better results, but that is beyond the scope of this paper.



(a) $RMSPE_{GDP,h,m}/RMSPE_{GDP,h,AR(4)}$

(b) $RMSPE_{UR,h,m}/RMSPE_{UR,h,AR(4)}$

(c) $RMSPE_{SPREAD,h,m}/RMSPE_{SPREAD,h,AR(4)}$

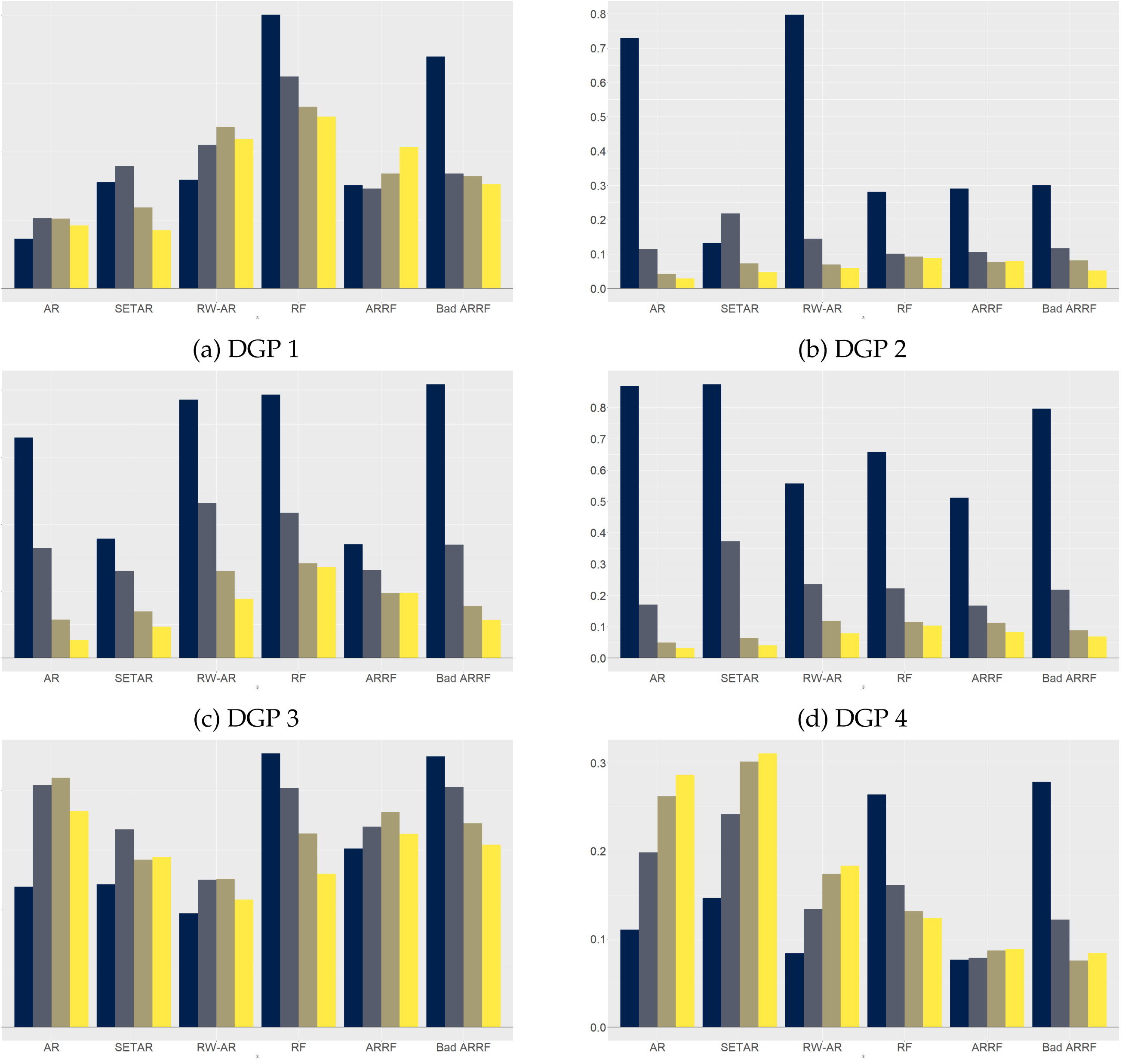
(d) $RMSPE_{INF,h,m}/RMSPE_{INF,h,AR(4)}$

(e) $RMSPE_{HOUST,h,m}/RMSPE_{HOUST,h,AR(4)}$

(f) $RMSPE_{IR,h,m}/RMSPE_{IR,h,AR(4)}$

Figure 23: All detailed results for EOTB-MRFs.

## A.5 Additional Simulations Graph(s)



(a) DGP 1

(b) DGP 2

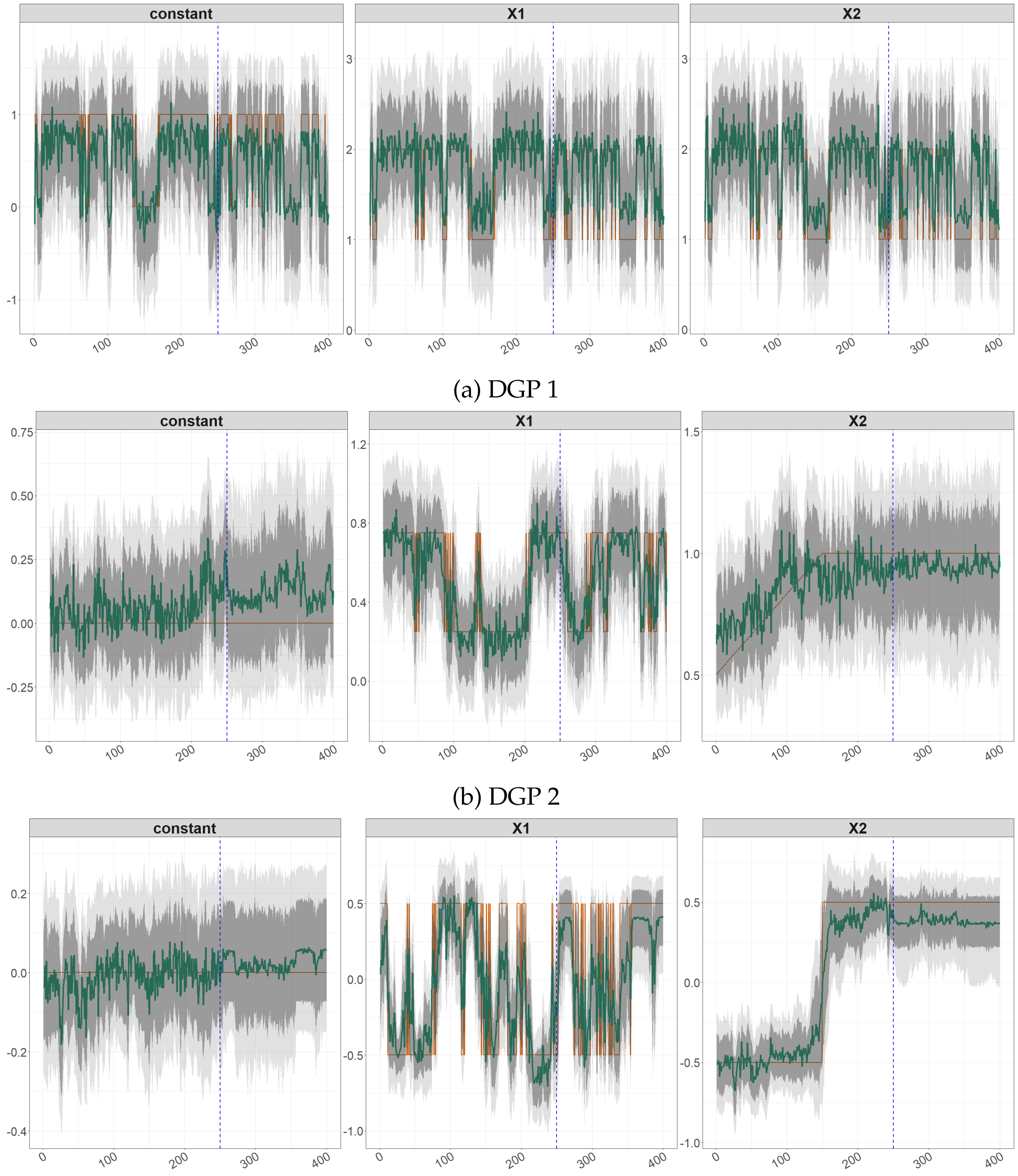
(c) DGP 3

(d) DGP 4

(e) DGP 5

(f) DGP 6

Figure 24: Investigation of the consequences of $X_t$'s misspecification, as exemplified by ’Bad ARRF’. Instead of the first two lags of $y_t$, $X_t$ is replaced by randomly generated *iid* (normal) variables. Total number of simulations is 50, and the total number of squared errors is thus 2000. Four horizons are reported, from one to 4 four.

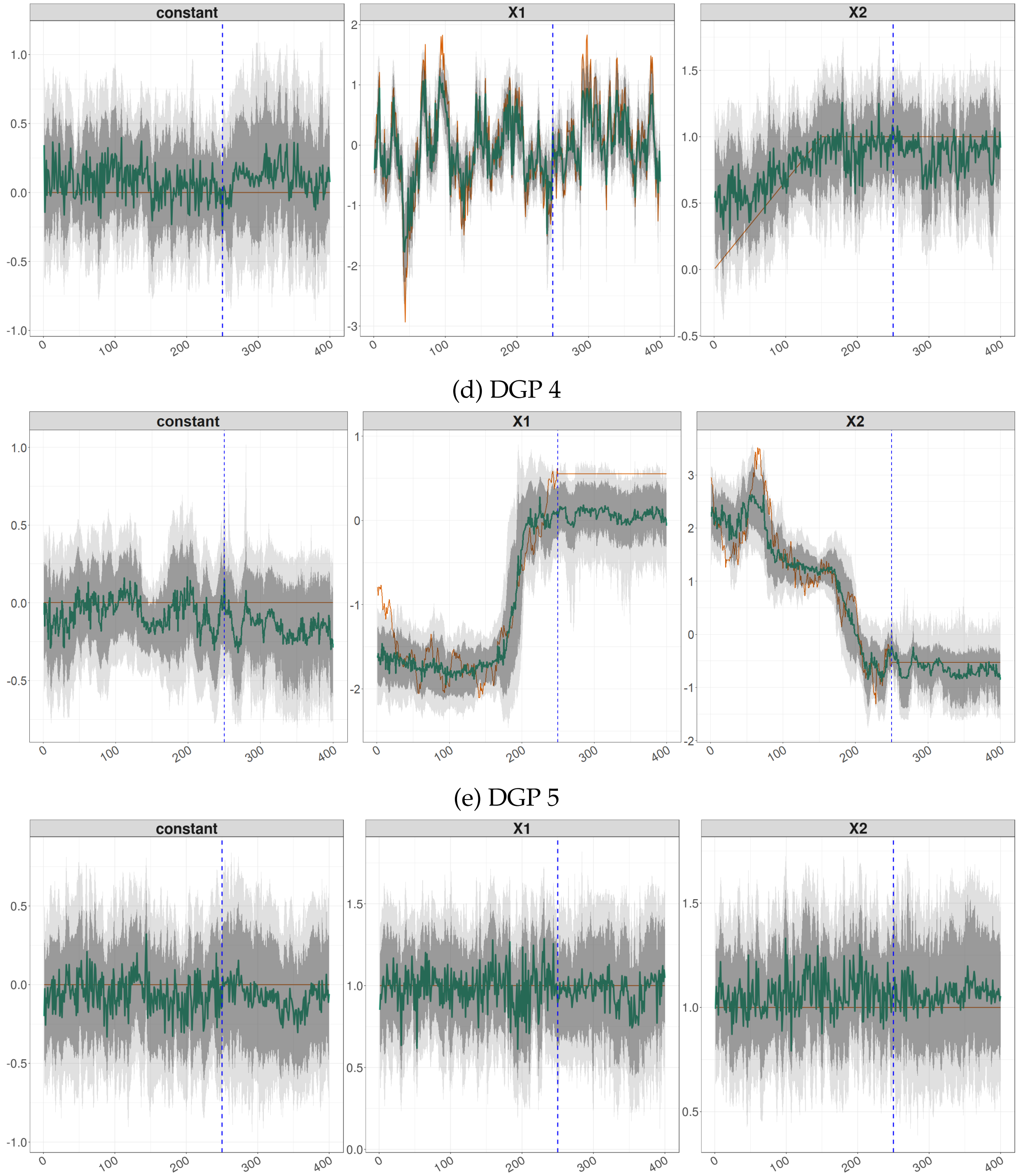


(a) DGP 1

(b) DGP 2

(c) DGP 3

Figure 25: The grey bands are the 68% and 90% credible region. After the blue line is the hold-out sample. Green line is the posterior mean and orange is the truth. The plots include only the first 400 observations for visual convenience.



(d) DGP 4

(e) DGP 5

(f) DGP 6

Figure 25: (Continued) The grey bands are the 68% and 90% credible region. After the blue line is the hold-out sample. Green line is the posterior mean and orange is the truth. The plots include only the first 400 observations for visual convenience.

## A.6 Monthly Results Details

## A.7 Tables of $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$ and DM tests

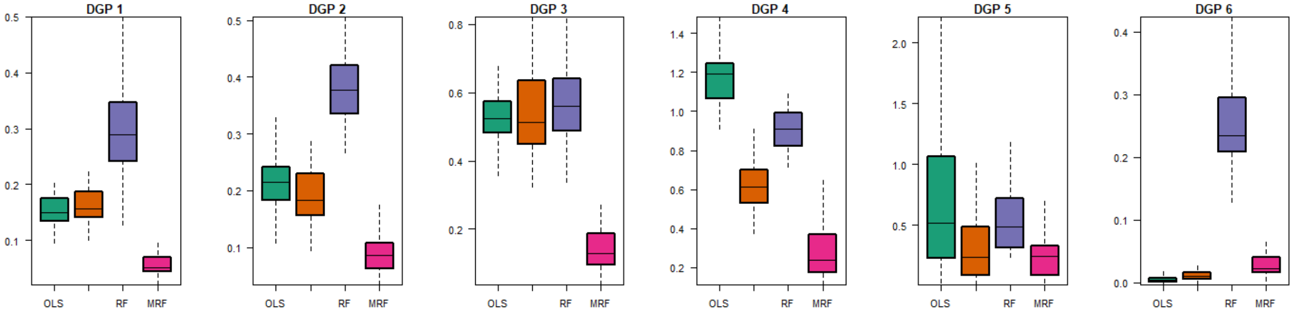


Figure 26: The distribution of RMSE dis-improvements with respect to the oracle's forecast for 4 models: OLS, Rolling-Window OLS, plain RF, MRF. 50 simulations of 750 OOS forecasts each.

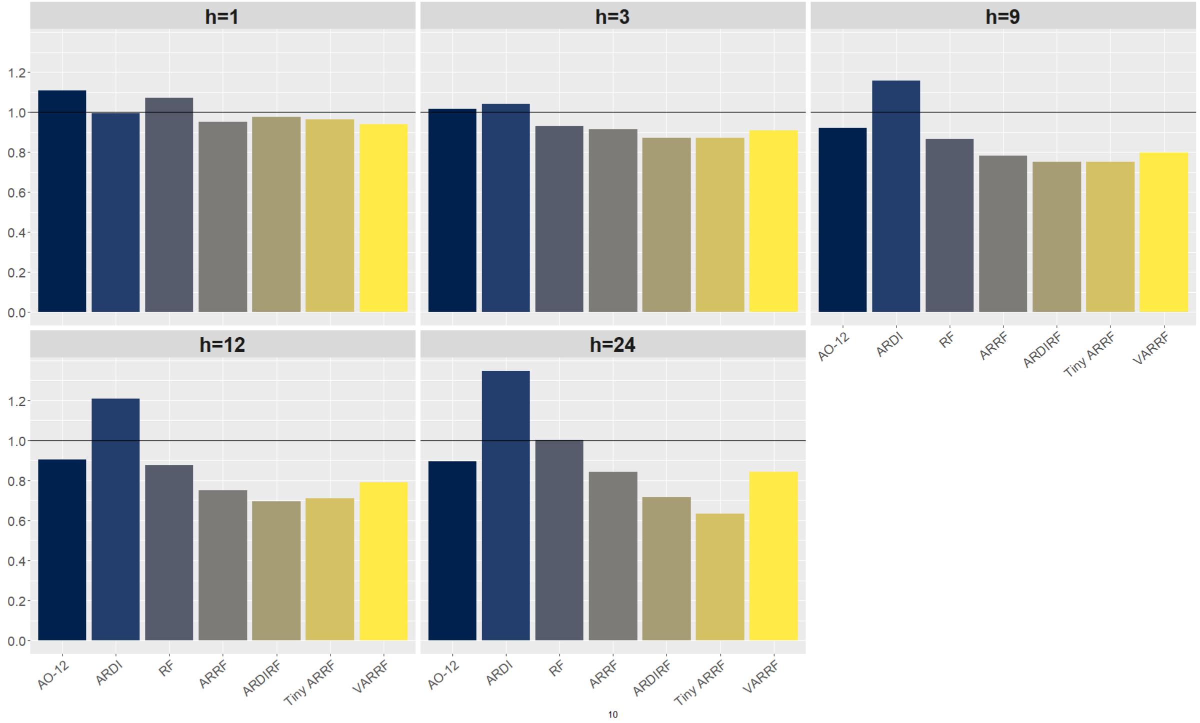


Figure 27: $RMSPE_{INF,h,m}/RMSPE_{INF,h,AR(12)}$ for monthly data

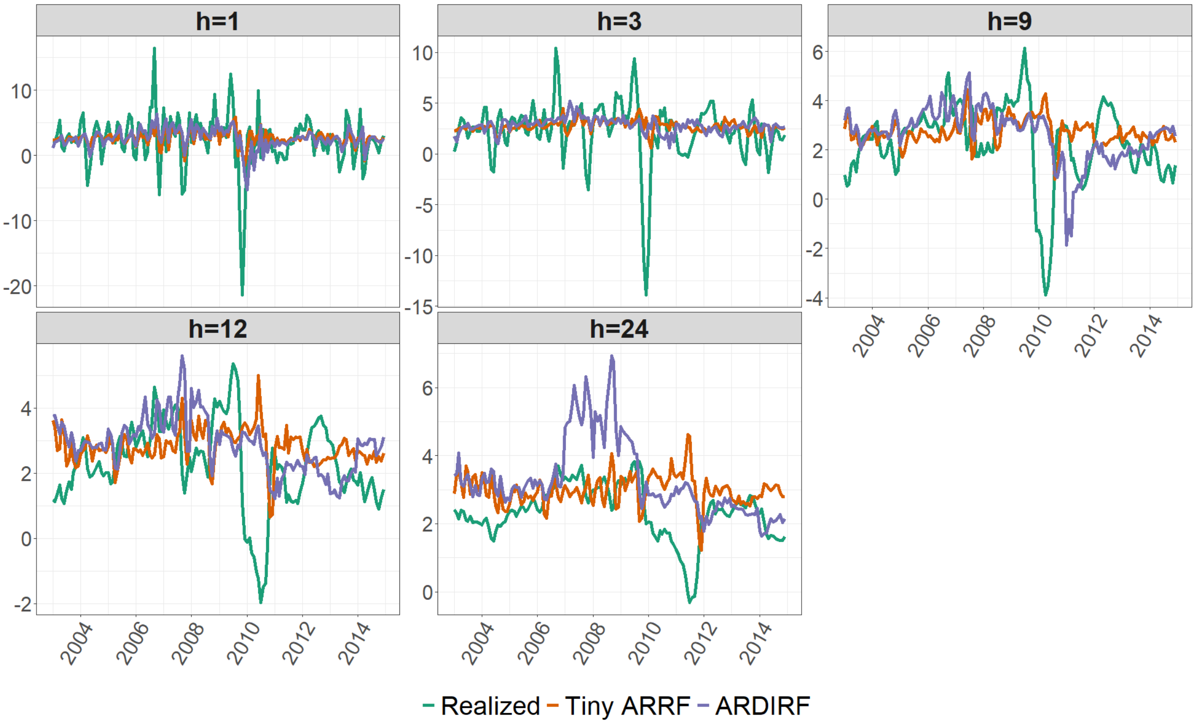


Figure 28: Selected forecasts of average inflation for monthly data

Table 4: Main Quarterly Results

| | ARDI | LASSO-MAF | Ridge-MAF | RF | RF-MAF | AR+RF | Tiny RF | ARDIRF | ARRF | Tiny ARRF | VARRF | SETAR | STAR | TV-AR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **GDP** | | | | | | | | | | | | | | |
| h=1 | 1.0158 | 0.9592 | 0.8899** | 0.9406 | 0.8603 | 0.8932 | 1.0296 | **0.8405** | 0.9487 | 1.043 | 1.2387 | 1.0104 | 1.0282 | 0.9948 |
| h=2 | 0.9641 | 0.9776 | 0.9751 | 0.9874 | **0.9055** | 0.9259 | 1.0069 | 0.9864 | 0.9516** | 1.0263 | 0.9839 | 0.9745 | 0.9789 | 1.0335 |
| h=4 | 1.0267 | 0.9803 | 0.9871*** | 0.9961 | 0.9831 | 0.9933 | 1.0273 | 0.9597 | 0.9358 | 0.9785 | **0.9235** | 0.971*** | 0.9595*** | 0.9605 |
| h=6 | 1.3583 | 0.9785 | 0.9772 | 0.9818 | 0.9952 | 1.0026 | 1.0753 | 1.0019 | 0.9838 | 0.9782 | 1.0202 | 0.9811 | **0.9548** | 0.983 |
| h=8 | 1.372 | 1.0049 | 0.9907 | 0.9899 | 0.9889 | **0.9611** | 1.1516 | 1.0508 | 0.9842 | 1.0057 | 1.0003 | 0.9997 | 0.9714 | 0.9982 |
| **UR** | | | | | | | | | | | | | | |
| h=1 | 0.827 | 0.9865 | 0.9869 | 1.0032 | 0.8527* | 0.8359 | 1.2379** | **0.7277** | 0.8701** | 1.0025 | 1.2554 | 1.1813 | 1.0996 | 1.0041 |
| h=2 | 0.8008 | 0.9758 | 0.9228* | 0.9804 | 0.8492 | 0.8392 | 1.1452* | **0.7299** | 0.852 | 0.9552 | 0.8986 | 1.0276 | 0.9741 | 0.9923 |
| h=4 | 0.8831 | 0.9565*** | 0.9432** | 0.9631* | 0.8654* | 0.8428* | 1.3664 | **0.7904** | 0.8729* | 0.9239 | 0.9326 | 1.0187 | 1.0085 | 1.3389 |
| h=6 | 1.176* | 0.983 | 0.9824 | 1.0062 | 0.9386 | 0.9014 | 1.5955* | **0.894** | 0.9254 | 0.969 | 0.9754 | 1.0706 | 1.0424 | 1.1426 |
| h=8 | 1.2461 | 0.9789 | 1.0123 | 1.0143 | **0.9491** | 0.9535 | 1.5716 | 1.008 | 0.9842 | 0.9778 | 1.0256 | 1.093 | 1.0591 | 1.1143** |
| **SPREAD** | | | | | | | | | | | | | | |
| h=1 | 1.2751 | 2.158*** | 0.932 | 0.9096 | 0.9539 | 0.7915** | 0.9608 | 1.125 | 0.9267* | 1.057 | **0.7672**** | 1.5079*** | 1.5297*** | 0.9792 |
| h=2 | 1.1347 | 1.1986 | 0.7689 | **0.6636**** | 0.7824 | 0.7211*** | 0.9294 | 0.8214 | 0.8249** | 1.1062 | 0.7158*** | 1.1893 | 1.2039 | 1.0352 |
| h=4 | 0.8633 | 0.9458 | 1.0078 | 0.808 | 0.6919** | **0.6054**** | 1.4821* | 0.6879** | 0.8027* | 1.0715 | 0.6855** | 1.0411 | 1.0617 | 1.3048 |
| h=6 | 1.505 | 0.7976* | 1.1315 | 0.9814 | 0.8032 | 0.8049 | 1.4343 | **0.6903***** | 0.815 | 1.05 | 0.7411** | 1.034 | 1.0565 | 1.1871 |
| h=8 | 1.2845 | **0.7633**** | 0.9583 | 0.924 | 0.834 | 0.8884 | 1.3614 | 0.8251 | 0.7703* | 0.9861 | 0.8157 | 1.1142 | 1.1364 | 0.9865 |
| **INF** | | | | | | | | | | | | | | |
| h=1 | 1.0057 | 0.9304 | 0.9523 | 0.9758 | 0.8768 | 1.2315 | 0.8976 | 1.0579 | 0.9236 | **0.8656*** | 1.0791 | 1.05 | 1.0021 | 0.9314 |
| h=2 | 1.0065 | 0.9642 | 0.9238 | 0.9244 | **0.8202** | 1.0003 | 0.8813 | 0.8928 | 0.8643* | 0.8724 | 0.9159 | 0.8601* | 0.8552 | 0.8916 |
| h=4 | 1.0775 | 0.9238 | 0.873 | 0.9352 | **0.8546**** | 0.9571 | 0.865 | 0.9289 | 0.8865* | 0.9466* | 0.8902* | 0.9031* | 0.8701* | 0.9051 |
| h=6 | 1.3193 | 0.9634 | 0.8971 | 1.005 | 0.8804 | 0.9972 | **0.8619** | 0.9067 | 0.8873 | 0.9235** | 0.8664 | 0.943 | 0.8892 | 0.9777 |
| h=8 | 1.2103 | 0.978 | 1.2677 | 1.4389 | **0.8773*** | 0.9412 | 0.8845 | 0.9454 | 0.9424 | 0.9409 | 0.9339 | 0.9644 | 0.9224 | 0.9821 |
| **HOUST** | | | | | | | | | | | | | | |
| h=1 | 1.1278 | 1.0395 | 0.9394* | **0.9162*** | 1.0014 | 1.0104 | 1.2407*** | 1.1106* | 0.931** | 0.9505 | 1.1318 | 1.0073 | 0.9866 | 1.0014 |
| h=2 | 1.1297 | 0.989 | 0.9444** | 0.9503* | 1.0141 | 1.0202 | 1.103* | 1.0057 | 0.9886 | 1.0207 | 0.9763 | **0.9379** | 0.9746 | 1.0071 |
| h=4 | 1.1089 | 0.9755** | 0.9703* | 0.967 | 1.0133 | 1.0296 | 1.1199 | 0.9908 | 0.9526*** | 1.0238 | 1.0046 | **0.9511** | 0.9634 | 1.0781 |
| h=6 | 1.4045 | 0.9603 | 0.9591 | 0.957 | 0.9601*** | 1.0117 | 1.1629 | 1.0138 | 0.9896 | 0.9975 | 1.0062 | **0.9476** | 0.9565 | 0.9912 |
| h=8 | 1.0394 | 0.9507 | 0.9506 | 0.9501 | 0.9882 | 1.0244 | 1.4355 | 1.0054 | 0.9999 | 1.0069 | 0.9795 | **0.9472** | 0.9476 | 1.0288 |
| **IR** | | | | | | | | | | | | | | |
| h=1 | 1.8457 | 1.0161 | 1.5506 | 1.173 | 1.1104 | 0.9716 | 0.9856 | 1.2673 | 0.9923 | **0.9198** | 1.5357 | 1.3899 | 1.2026 | 0.9745 |
| h=2 | 1.4852 | 0.9591 | 1.0061 | 1.0042 | 0.9339 | 0.9784 | 1.2924*** | 1.4005 | 0.9298 | **0.9151** | 1.1619 | 1.1521 | 1.1137 | 1.0408 |
| h=4 | 0.9619 | 1.0027 | 1.0318 | 1.0324 | 1.0412 | 0.9903 | 1.3921* | 0.9515 | **0.9361** | 1.1202 | 1.0159 | 1.0796 | 1.066 | 1.0938 |
| h=6 | 1.8672 | 0.9539 | 0.9945 | 0.9983 | **0.9266** | 0.9297 | 1.2289* | 1.0686 | 0.9503 | 1.0699 | 1.2807 | 1.1862 | 1.143 | 1.0609** |
| h=8 | 1.5808 | 0.9846 | 1.0192 | 1.0253 | **0.9592** | 0.9605 | 1.2009 | 1.0691 | 1.0064 | 1.0978 | 1.023 | 1.2496** | 1.2049** | 1.0566 |

The numbers represent the root MSPE of the model with respect to the root MSPE of an AR(4). A number in bold means that the specific model is the best model of the row (including constant parameters). DM test is for each model against the AR(4). ′*′, ′**′ and ′***′ means p-values of below 10%, 5% and 1%.

Table 5: Main Monthly Results

| | RW | AO-12 | AO-h | AR4 | ARDI | Ridge-MAF | LASSO-MAF | RF | RF-CSF | RF-MAF | AR+RF | ARRF | ARDIRF | Tiny ARRF | VARRF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **IP** | | | | | | | | | | | | | | | |
| h=1 | 1.3379* | 1.1078* | 1.1356 | 1.0026 | 0.9631 | 0.999 | 0.981 | 1.0266 | 0.9921 | **0.9356*** | 0.9716 | 0.9748 | 0.9446 | 1.0026 | 0.9428 |
| h=3 | 1.109** | 1.174* | 1.0234 | 1.0157 | 0.989 | 0.9993 | 0.9816 | 1.1187 | 1.0547 | 0.9763 | **0.9634** | 1.0033 | 1.0414 | 1.0025 | 1.0608 |
| h=9 | 0.9864 | 1.0372 | 1.0257 | 1.006 | 1.0643 | **0.9761** | 1.0109 | 1.0242 | 1.0777 | 1.055 | 1.0176 | 1.0701 | 1.1485 | 0.9772 | 1.0078 |
| h=12 | 0.961 | 0.9961 | 0.9961 | 1.0055 | 1.054 | 0.9592 | 1.2361 | 0.9894 | 1.0258 | 0.9736 | **0.9141** | 0.9766 | 1.0715 | 0.9943 | 0.9709 |
| h=24 | **0.8297** | 0.8424 | 0.843 | 0.9959 | 1.1693* | 0.9801 | 1.1925** | 0.9237 | 0.9963 | 0.8613 | 0.855 | 0.9199 | 0.9797 | 1.0123 | 0.9583 |
| **UR** | | | | | | | | | | | | | | | |
| h=1 | 1.2912*** | 1.0299 | 1.0923 | 1.007 | 0.9457 | 0.9255*** | 0.9039*** | 0.9698 | 0.9283** | **0.8738***** | 0.9491 | 0.91*** | 0.9248 | 0.9801 | 0.9423** |
| h=3 | 1.1001** | 1.102 | 1.0459 | 0.9997 | 0.8615 | 0.92* | 0.8735* | 1.0527 | 1.0073 | **0.806***** | 0.9223 | 0.8752** | 0.8523* | 0.9985 | 0.8766*** |
| h=9 | 1.08 | 1.114 | 1.1029 | 0.9906 | 0.9195 | 0.9768 | 1.0095 | 1.0197 | 1.0611 | 0.9621 | **0.9105** | 0.9814 | 0.9732 | 1.0603* | 0.9815 |
| h=12 | 1.0479 | 1.0746 | 1.0746 | 0.9934 | 0.9612 | 0.9884 | 1.0573 | 0.9719 | 1.0164 | 0.9586 | **0.9086** | 0.985 | 1.0265 | 1.1089* | 0.9589 |
| h=24 | 1.0094 | 1.0243 | 1.0256 | 1.0157 | 1.0595 | 0.9803 | 1.1236 | 0.9074 | 0.867* | 0.8378 | **0.8108** | 0.8608 | 0.9929 | 1.1332** | 0.8946 |
| **SPREAD** | | | | | | | | | | | | | | | |
| h=1 | 0.9992 | 2.8757*** | 1.2271*** | 0.991 | 1.2149** | 2.5884*** | 1.1492** | 3.5224*** | 1.2137*** | 1.0743 | **0.9057***** | 1.0738 | 1.0133 | 1.0902** | 0.9651 |
| h=3 | 0.9321 | 1.6797*** | 1.0683 | 1.0147 | 1.2467 | 1.3704*** | 0.9253 | 1.6939*** | 0.982 | 0.8222** | **0.8081***** | 1.0301 | 0.8696** | 1.028 | 0.8178*** |
| h=9 | 0.973 | 1.3586 | 1.2722 | 1.0145 | 1.0634 | 0.9318 | 1.095 | 0.9429 | 0.8567 | 0.7328** | 0.7174** | 0.7772** | **0.7067***** | 1.0345 | 0.7326*** |
| h=12 | 1.0208 | 1.275 | 1.275 | 1.0225 | 1.0463 | 0.8471** | 1.0166 | 0.8023*** | 0.7391** | 0.6576*** | **0.6037***** | 0.6714*** | 0.7132*** | 1.014 | 0.7049*** |
| h=24 | 1.254 | 1.3355 | 1.3412 | 1.0297 | 0.9603 | 0.8545* | 0.9613 | 0.7975** | 0.7762* | 0.6994** | 0.7108* | 0.73** | **0.6957**** | 0.9183 | 0.7107*** |
| **INF** | | | | | | | | | | | | | | | |
| h=1 | 1.0918 | 1.1068* | 1.1844* | 1.0159 | 0.9946 | 1.0175 | 1.0001 | 1.0724 | 1.0272 | 1.0645* | 1.0128 | 0.9525 | 0.976 | 0.9631 | **0.9389*** |
| h=3 | 1.372** | 1.0159 | 1.2409* | 1.0408 | 1.0396 | 0.895 | 0.945 | 0.9288 | 0.9008 | 0.8803 | 1.0502 | 0.9137 | **0.871** | 0.8728* | 0.9092 |
| h=9 | 1.7653*** | 0.9212 | 1.0138 | 1.074 | 1.1561 | 0.9017 | 1.1787 | 0.8638 | 0.8042 | 0.7783 | 1.1506 | 0.7815 | 0.7523 | **0.7522*** | 0.7983 |
| h=12 | 1.8604** | 0.9055 | 0.9055 | 1.0915 | 1.2097 | 0.9016 | 1.1008 | 0.8778 | 0.8062 | 0.7944 | 1.1544 | 0.7515 | **0.696** | 0.7103* | 0.7903 |
| h=24 | 2.1229** | 0.8962 | 0.8583 | 1.044 | 1.3474 | 1.0853 | 1.5838 | 1.0015 | 0.8228 | 1.1197 | 1.1197 | 0.8436 | 0.7153 | **0.6339**** | 0.8429 |
| **HOUST** | | | | | | | | | | | | | | | |
| h=1 | 1.7327*** | 1.1017** | 1.3475*** | **0.9958** | 1.0673 | 1.0641* | 1.031 | 1.0772** | 1.0236 | 1.0171 | 1.0034 | 0.998 | 1.0122 | 1.0064 | 1.0157 |
| h=3 | 2.5999*** | 1.0553 | 1.3361*** | **0.9624**** | 1.1467 | 1.0701 | 1.1157 | 1.0318 | 1.011 | 1.0672 | 1.0337 | 1.0091 | 1.0329 | 0.9834 | 1.0115 |
| h=9 | 3.4514*** | 1.0516 | 1.1165 | **0.9766*** | 1.3475 | 1.1483 | 1.2893 | 0.979 | 1.0133 | 1.0152 | 1.0141 | 1.0383 | 1.1284 | 1.0161 | 1.0316 |
| h=12 | 3.6597*** | 1.047 | 1.047 | 0.9787 | 1.3213 | 1.1293 | 1.5051 | **0.9506** | 0.9866 | 0.9978 | 1.0066 | 0.9584 | 1.0756 | 1.0491 | 1.0529 |
| h=24 | 3.9969*** | 1.0949 | 1.0713 | 0.9549 | 1.1652 | 0.9849 | 1.0422 | 0.8705 | **0.8626** | 0.9422 | 0.9513 | 0.9165 | 1.0283 | 1.1042 | 0.9435 |

The numbers represent the root MSPE of the model with respect to the root MSPE of an AR(4). A number in bold means that the specific model is the best model of the row (including constant parameters). DM test is for each model against the AR(4). '*', '**' and '***' means p-values of below 10%, 5% and 1%.

## A.8 Additional Empirical Graphs

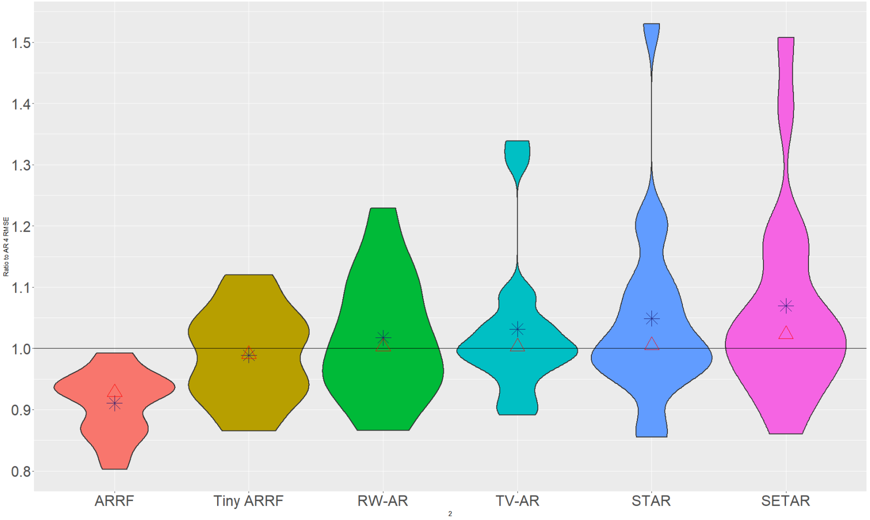
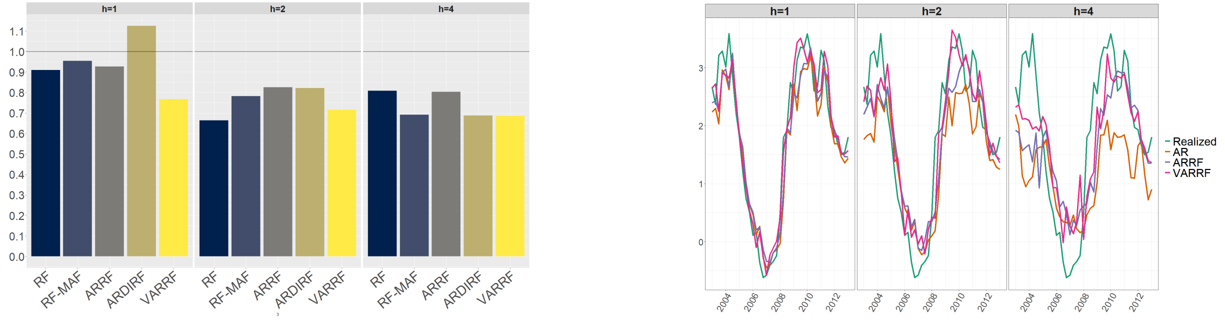


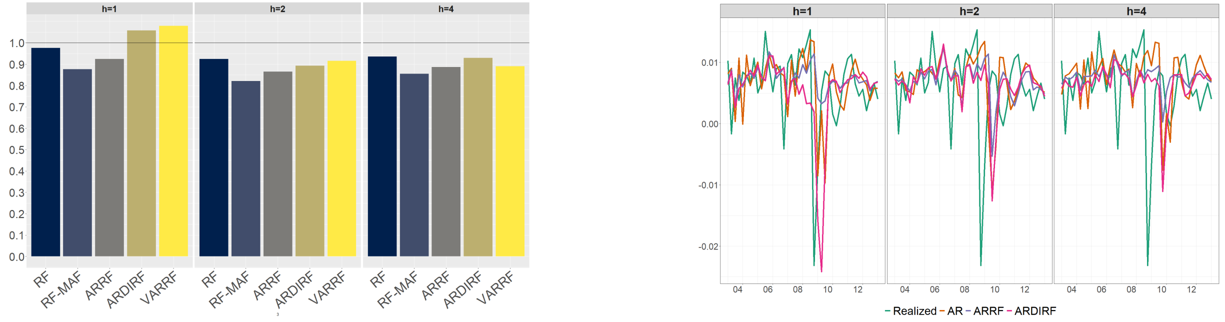
Figure 29: The distribution of $RMSPE_{v,h,m}/RMSPE_{v,h,AR(4)}$. The star is the mean and the triangle is the median.



(a) $RMSPE_{SPREAD,h,m}/RMSPE_{SPREAD,h,AR(4)}$

(b) A look at forecasts

Figure 30: SPREAD results in detail



(a) $RMSPE_{INF,h,m}/RMSPE_{INF,h,AR(4)}$

(b) A look at forecasts

Figure 31: INF results in detail

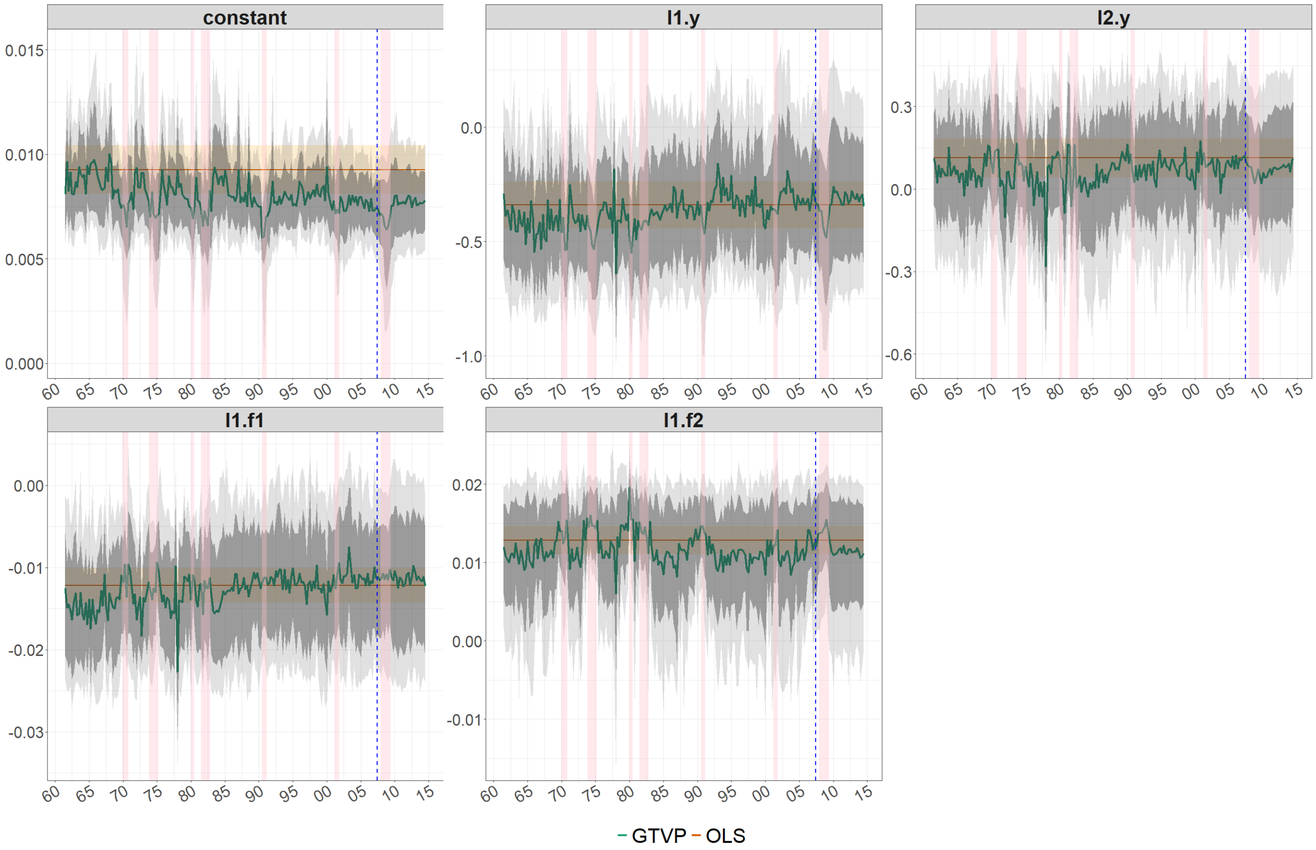


Figure 32: GTVPs of the one-quarter ahead GDP forecast. The grey bands are the 68% and 90% credible region. The pale orange region is the OLS coefficient ± one standard error. The vertical dotted blue line is the end of the training sample. Pink shading corresponds to NBER recessions.

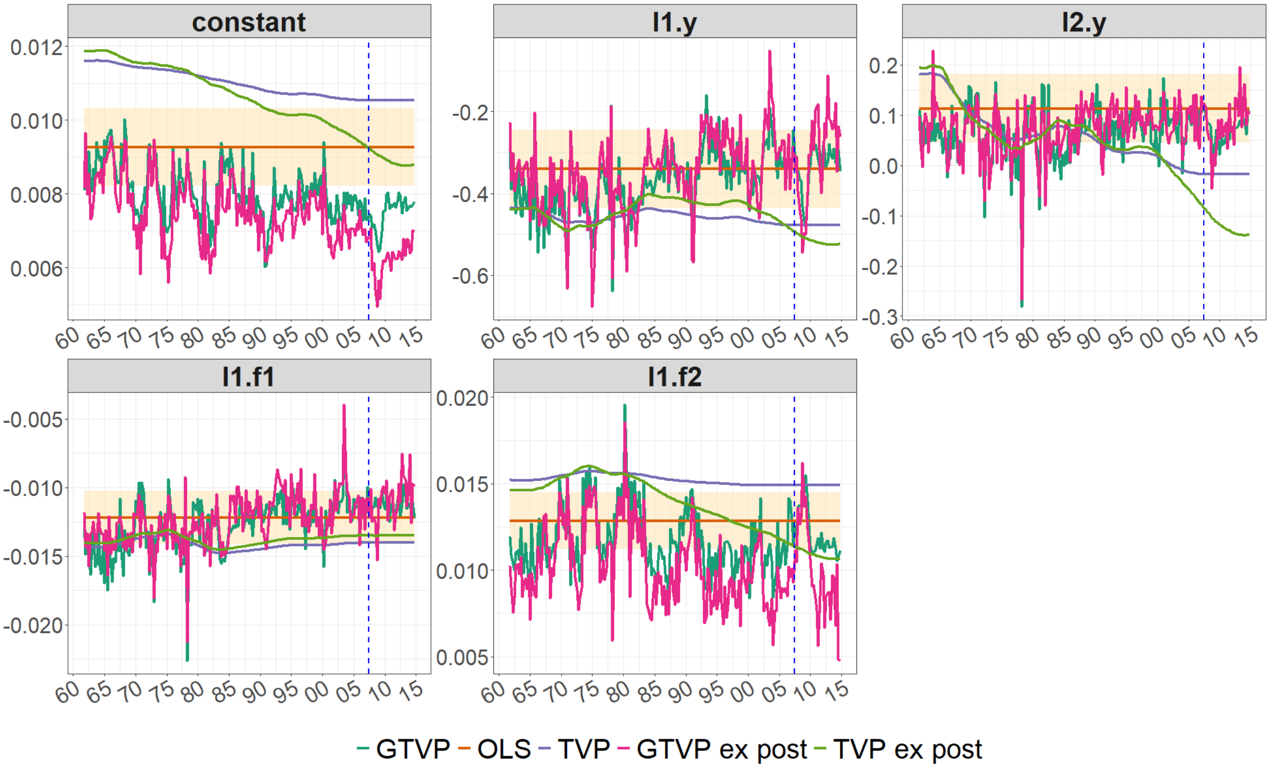
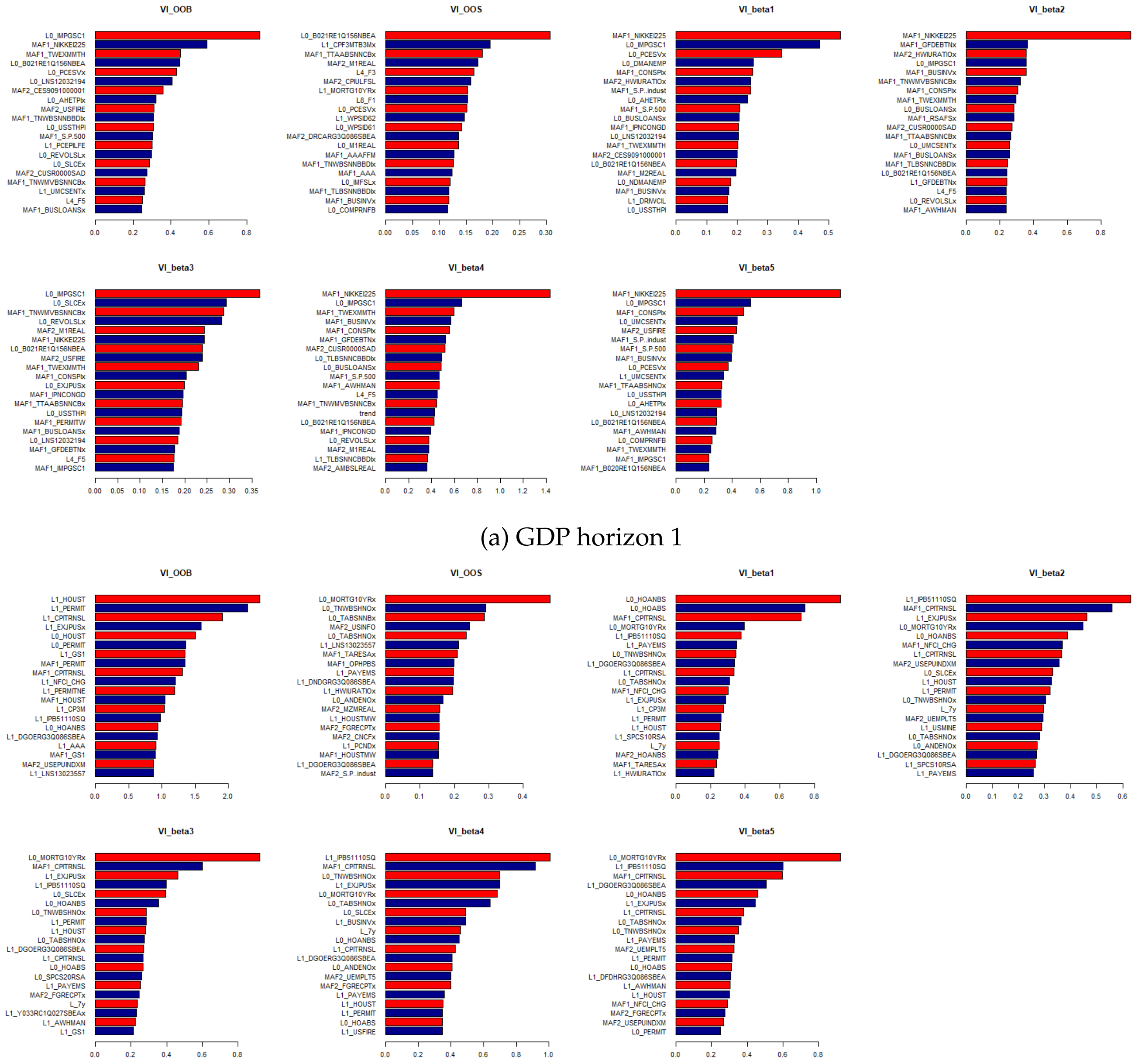


Figure 33: UR equation $\beta_t$'s obtained with different techniques. TVPs estimated with a ridge regression as in Goulet Coulombe (2019) and the parameter volatility is tuned with k-fold cross-validation. Ex Post TVP means using the full sample for estimation and tuning as opposed to only using pre-2002 data as for GTVPs. The pale orange region is the OLS coefficient ± one standard error. Pink shading corresponds to NBER recessions.



(a) GDP horizon 1

(b) UR horizon 1

Figure 34: 20 most important series according to the various variable importance (VI) criteria. Units are relative RMSPE gains (in percentage) from including the specific predictor in the forest part. $VI_{OOB}$ means VI for the out-of-bag criterion. $VI_{OOS}$ is using the hold-out sample. $VI_{\beta}$ is an out-of-bag measure of how much $\beta_{t,k}$ varies by withdrawing a certain predictor.

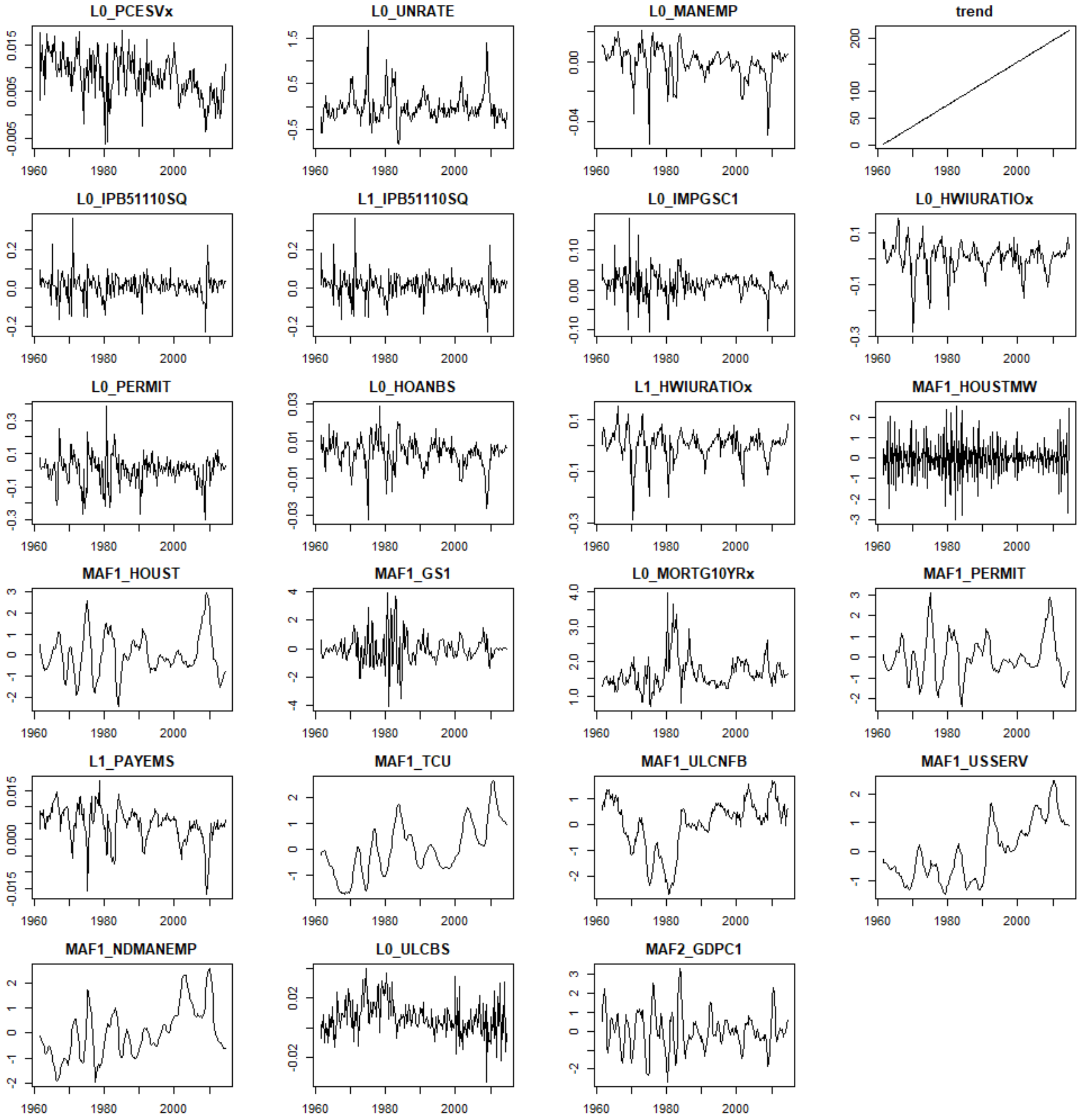
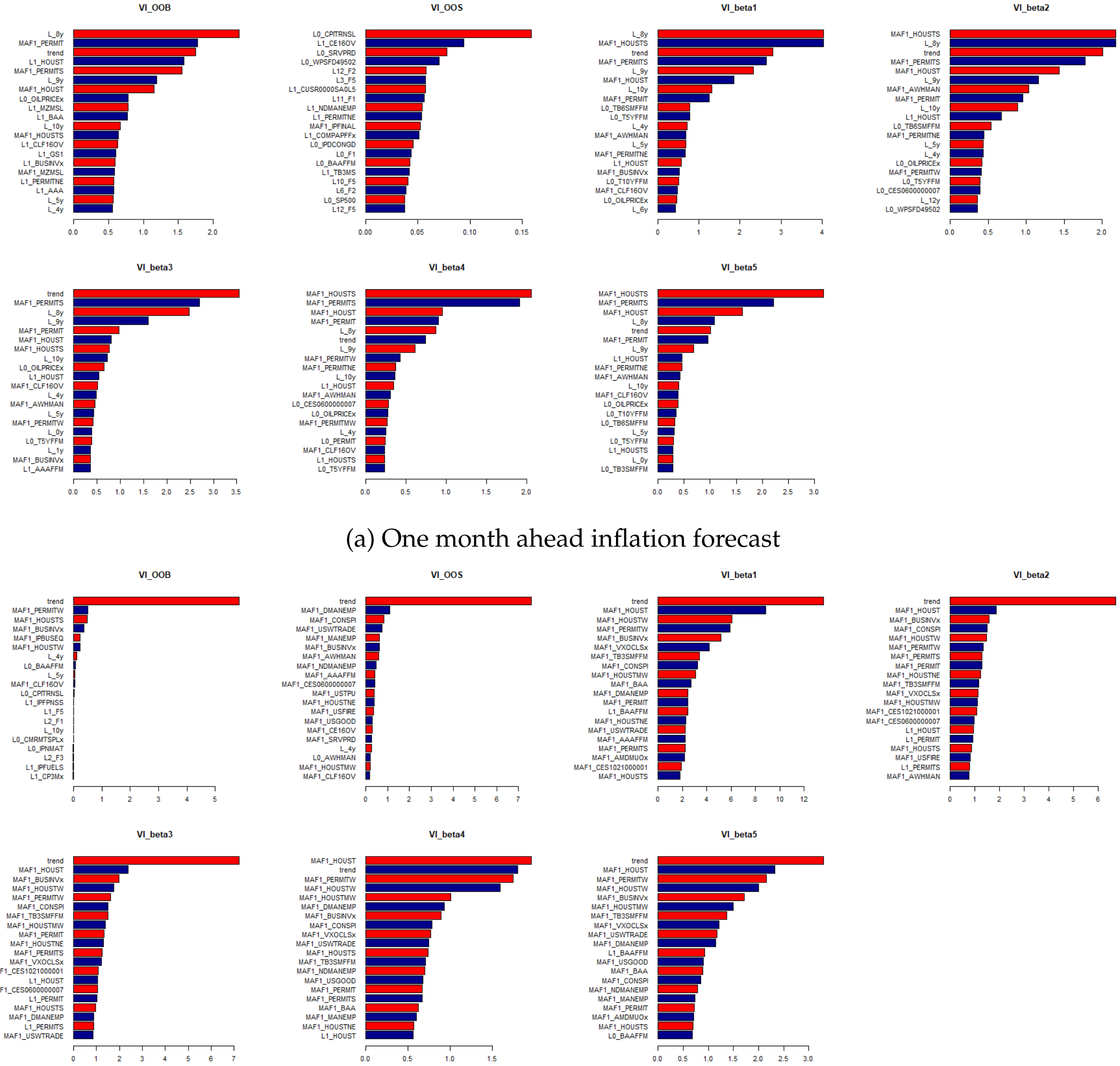


Figure 35: Series Underlying the Trees in Figure 15 and those of other (not plotted) GTVPs



(a) One month ahead inflation forecast

(b) Average inflation over the next 12 months

Figure 36: 20 most important series according to the various variable importance (VI) criteria. Units are relative RMSPE gains (in percentage) from including the specific predictor in the forest part. $VI_{OOB}$ means VI for the out-of-bag criterion. $VI_{OOS}$ is using the hold-out sample. $VI_{\beta}$ is an out-of-bag measure of how much $\beta_{t,k}$ varies by withdrawing a certain predictor.

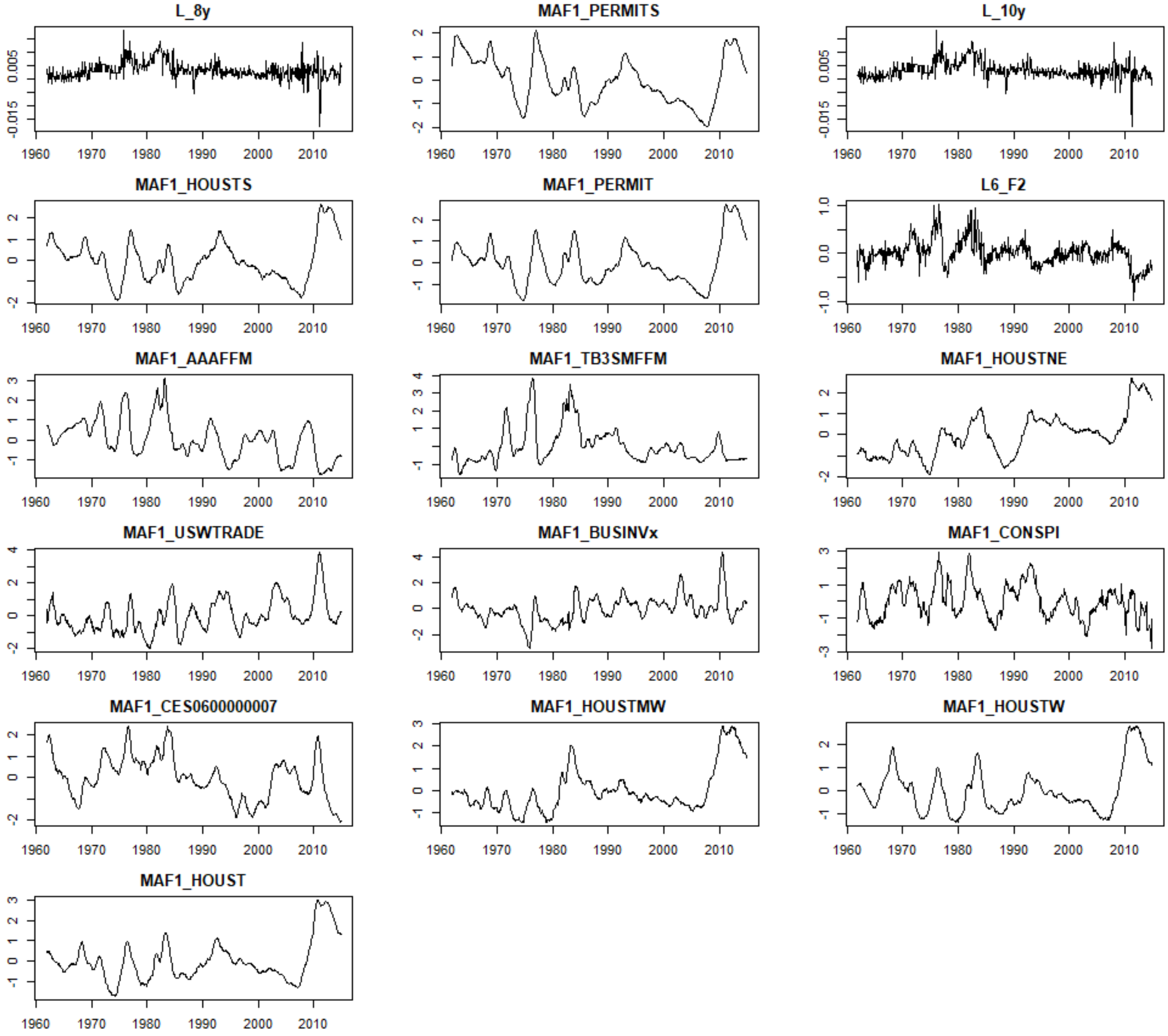


Figure 37: Series Underlying the Trees in Figure 16 and those of other (not plotted) GTVPs

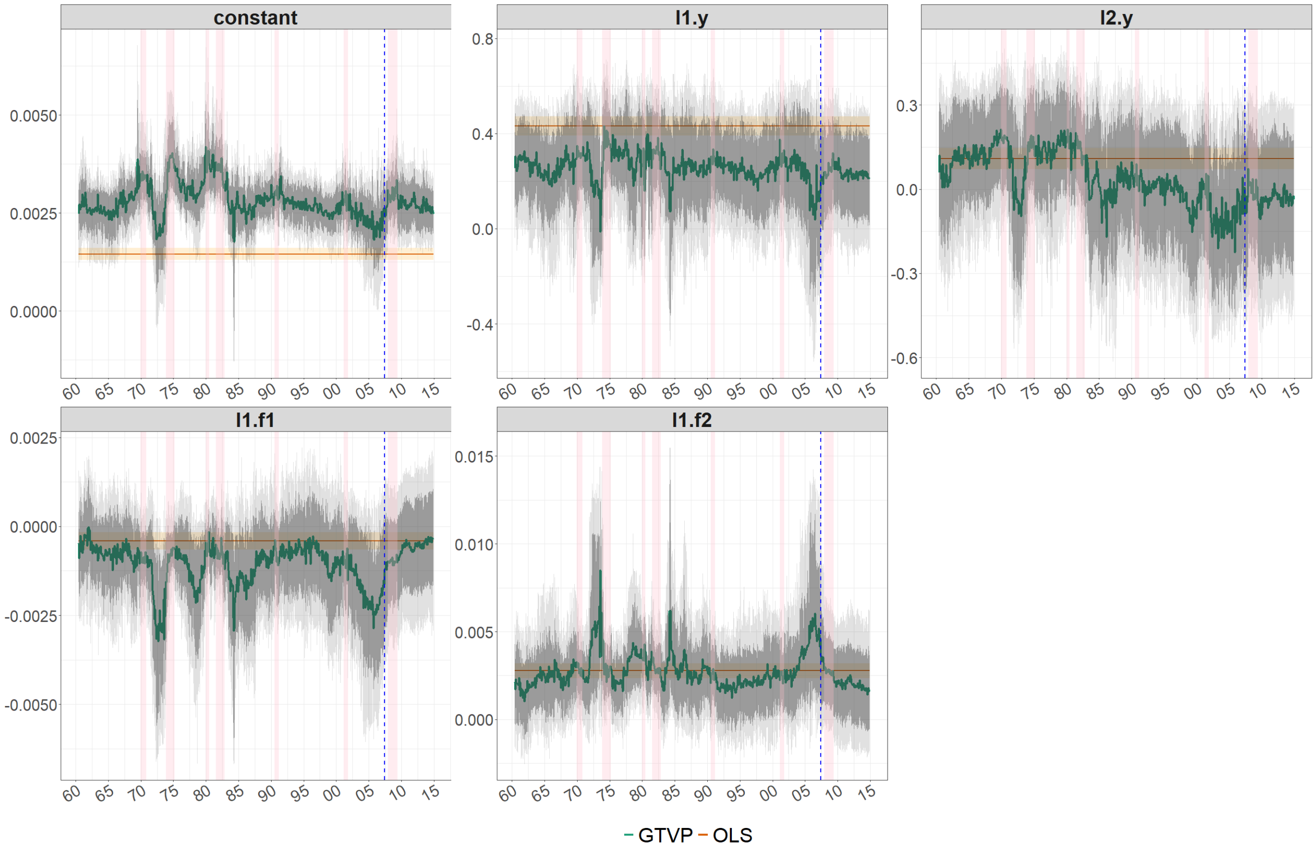


Figure 38: GTVPs of the one-month ahead INF forecast. The grey band is the 68% credible region. The pale orange region is the OLS coefficient ± one standard error. The vertical dotted blue line is the end of the training sample. Pink shading corresponds to NBER recessions.

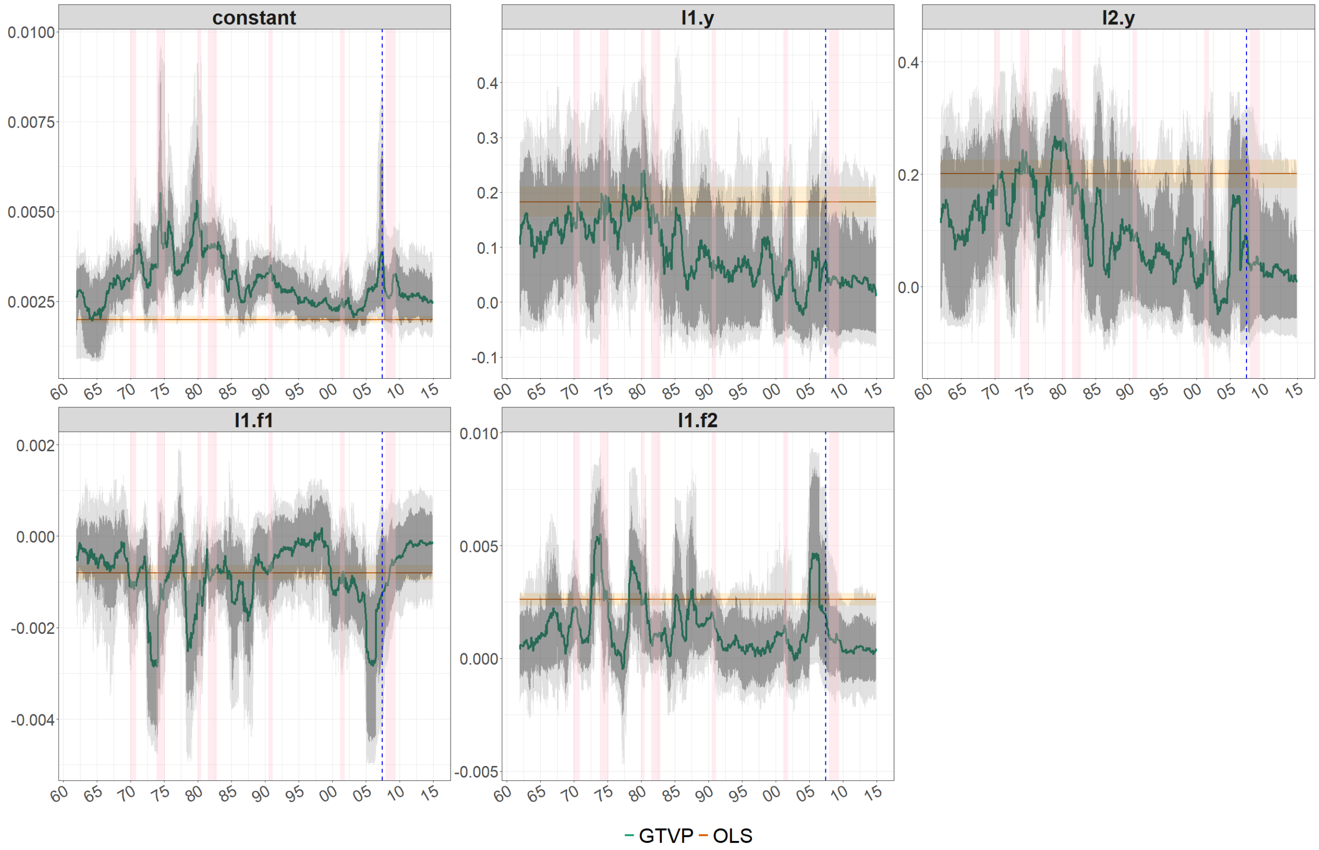
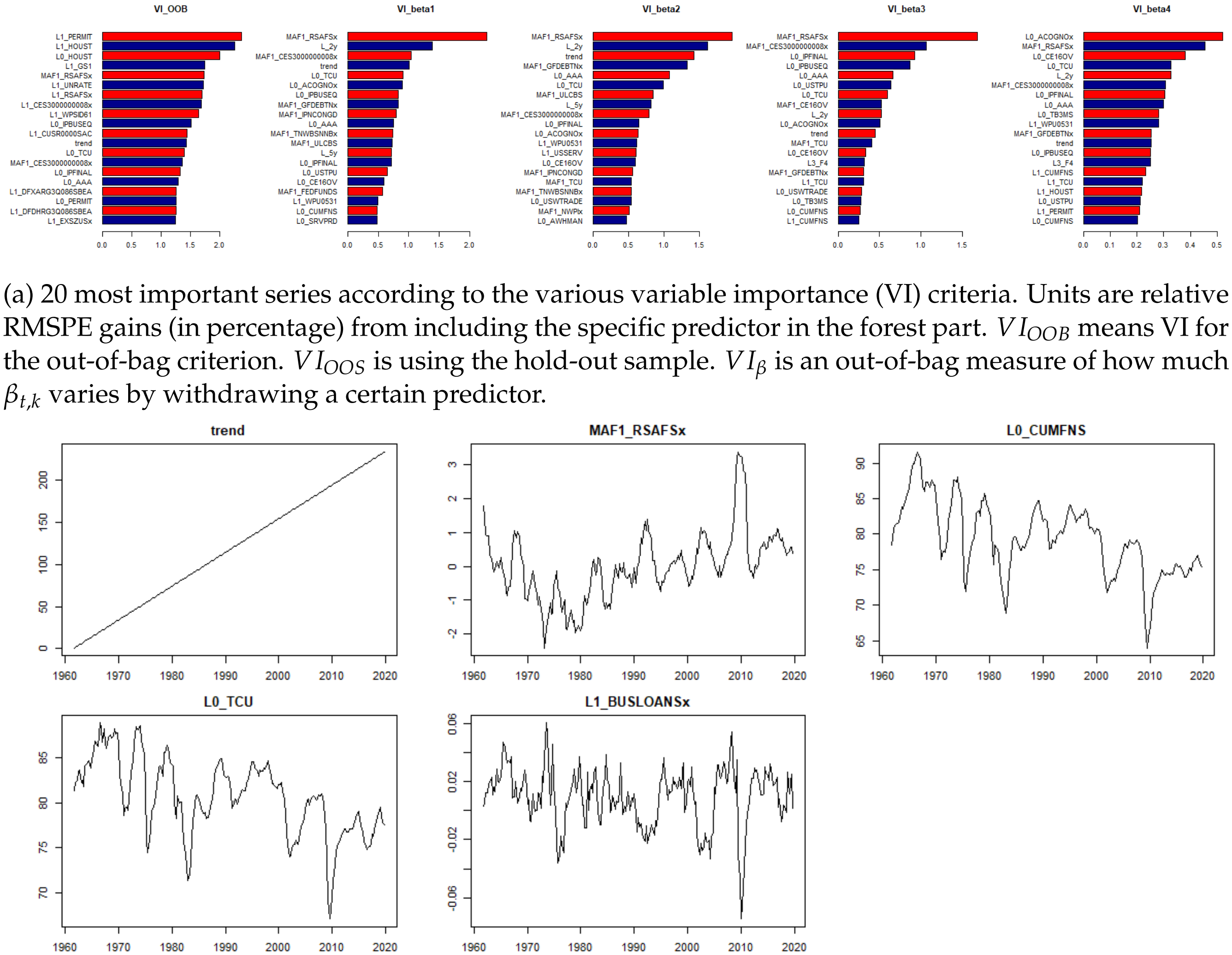


Figure 39: GTVPs of the 12-months ahead average INF forecast. The grey band is the 68% credible region. The pale orange region is the OLS coefficient ± one standard error. The vertical dotted blue line is the end of the training sample. Pink shading corresponds to NBER recessions.



(a) 20 most important series according to the various variable importance (VI) criteria. Units are relative RMSPE gains (in percentage) from including the specific predictor in the forest part. $VI_{OOB}$ means VI for the out-of-bag criterion. $VI_{OOS}$ is using the hold-out sample. $VI_{\beta}$ is an out-of-bag measure of how much $\beta_{t,k}$ varies by withdrawing a certain predictor.

(b) Series Underlying the surrogate trees

Figure 40: Complementary graphs for BCS GTVP analysis

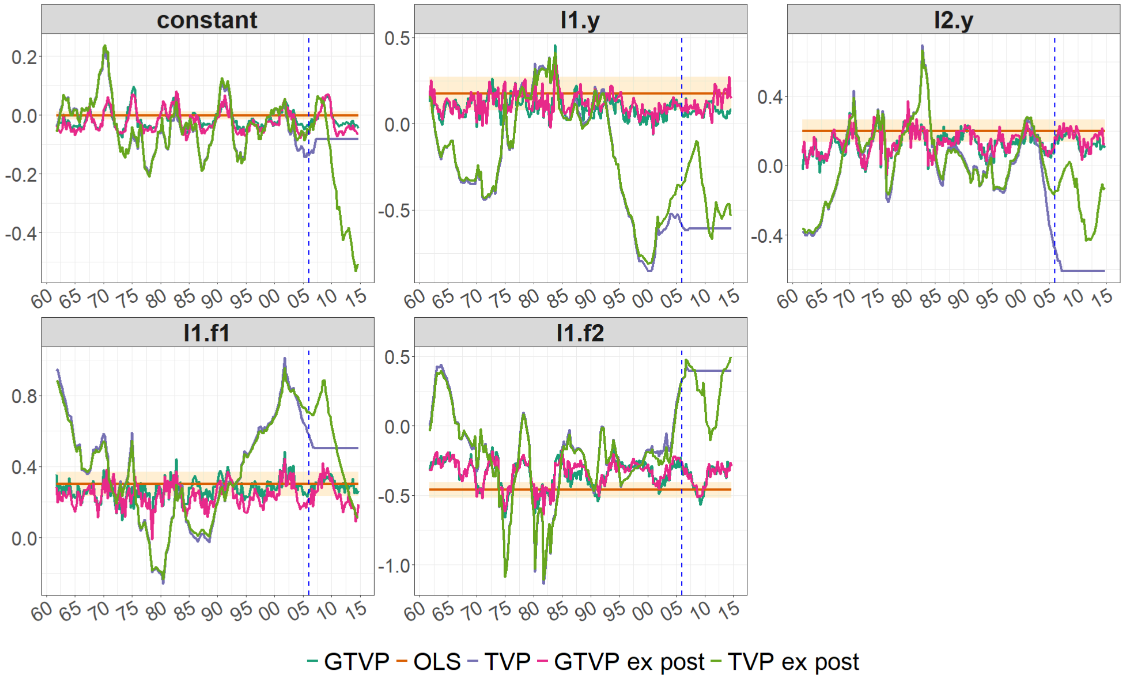


Figure 41: UR equation $\beta_t$'s obtained with different techniques. TVPs estimated with a ridge regression as in Goulet Coulombe (2019) and the parameter volatility $\lambda$ is tuned with k-fold cross-validation, **then divided by 1000**. Ex Post TVP means using the full sample for estimation and tuning as opposed to only using pre-2002 data as for GTVPs. The pale orange region is the OLS coefficient ± one standard error.

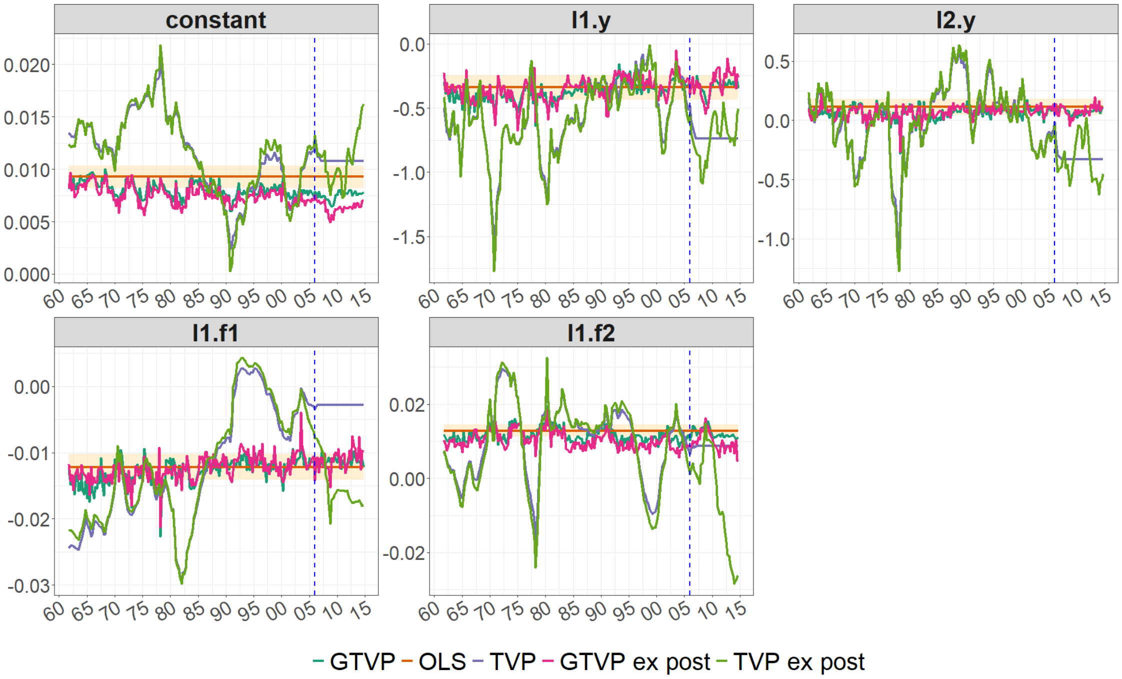
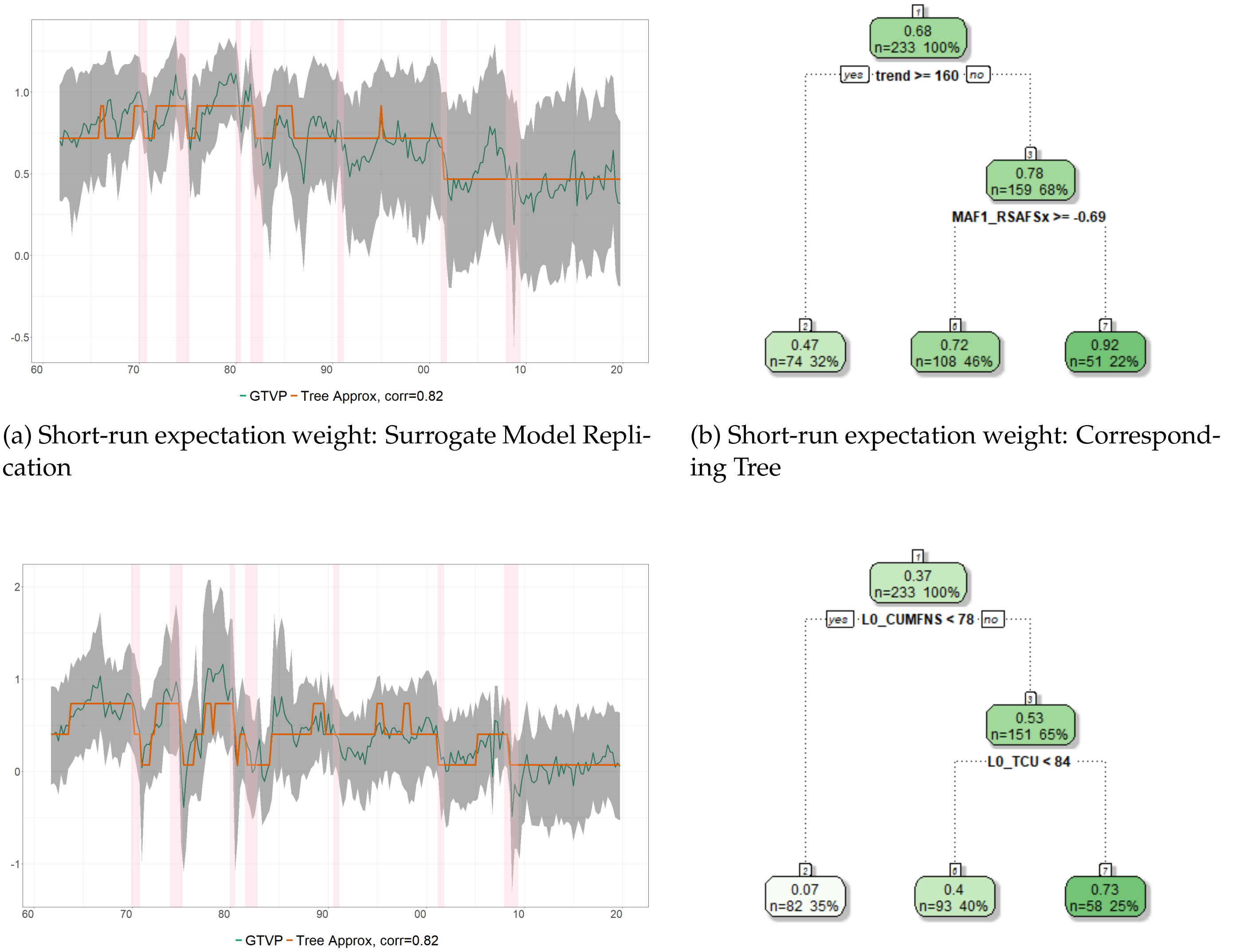


Figure 42: GDP equation $\beta_t$'s obtained with different techniques. TVPs estimated with a ridge regression as in Goulet Coulombe (2019) and the parameter volatility $\lambda$ is tuned with k-fold cross-validation, **then divided by 1000**. Ex Post TVP means using the full sample for estimation and tuning as opposed to only using pre-2002 data as for GTVPs. The pale orange region is the OLS coefficient ± one standard error.



(a) Short-run expectation weight: Surrogate Model Replication

(b) Short-run expectation weight: Corresponding Tree

(c) Unemployment Gap: Surrogate Model Replication

(d) Unemployment Gap: Corresponding Tree

Figure 43: Surrogate Trees for BCS Philipps' curve. Shade is 68% credible region. `cp`=0.075. Trees drawn with Rattle. Pink shading corresponds to NBER recessions.

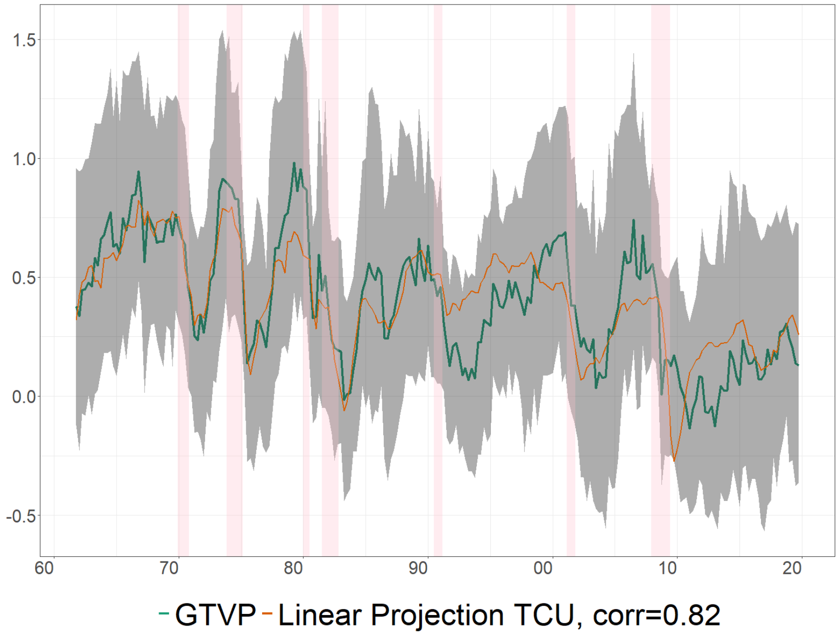


Figure 44: The grey bands are the 68% credible region. Pink shading corresponds to NBER recessions.