A forest full of HAR models

- Preliminary Version -

Onno Kleen[1] and Anastasija Tetereva[1]

[1]Erasmus University Rotterdam

February 1, 2022

Abstract

We propose a heterogeneous autoregressive (HAR) model with time-varying parameters in the form of a local linear random forest. In contrast to conventional random forests that approximate the volatility nonparametrically using local averaging, the building blocks of our forest are panel HAR models. The local panel HAR models cover the established linear relationship in realized variances while the trees model nonlinearities and interaction effects. Our approach allows the model coefficients to depend on idiosyncratic stock information and overall changing market conditions. We observe superior forecasting performance of the HAR forest for daily 5- and 22-day-ahead forecasts in an empirical analysis on the realized variances of 186 S&P 500 constituents from 2000 until 2019. By evaluating several variable importance measures, we conclude that the gains in forecast performance are mainly driven by conditioning on past values of realized variances. The contributions of stock returns' higher-order moments, semivariances, and lower-frequency covariates are negligible.

Keywords: Volatility forecasting, local linear random forest, pooled estimation

JEL Classification: C32, C53, C58, G17

1 Introduction

In this paper, we apply the idea underlying random forest (Breiman, 2001) to modeling time-varying parameters in stock price volatility models. Forecasting volatility of asset returns is relevant in many areas of risk management. In portfolio management applications, investors try to accurately predict volatilities to maximize Sharpe ratios of their portfolios. Managing value-at-risk also involves volatility modeling and is important for banks to comply with the requirements of financial regulations. One of the most popular models for managing dynamics of the realized volatility is the parsimonious heterogeneous autoregressive (HAR) model by Corsi (2009). In the HAR model, the predicted realized volatility is a linear function of its own lagged value, the average realized volatility over the previous week, and the average realized volatility over the previous month. Besides its very good forecast performance, the HAR model is also economically motivated by capturing decisions of investors with different investment horizons. Since work by Corsi (2009), a lot of extensions were suggested in the literature; see, for example, Bekaert and Hoerova (2014), Patton and Sheppard (2015), and Bollerslev et al. (2016). However, as documented in Audrino et al. (2019), none of these models is able to uniformly outperform the simplest HAR specification by Corsi (2009). A partial explanation is the empirical evidence by Audrino and Knaus (2016) who apply penalized linear regression to autoregressive models with up tp 100 lags. They find that the relevance of lagged realized variances is substantially varying over time with market conditions. This is our motivation to let the HAR parameters be state-dependent.

Covering a wide range of machine learning (ML) techniques, Gu et al. (2020) document that regression trees and neural networks perform best in asset pricing. For volatility forecasting, we prefer trees over neural networks due to interpretability and their ability to capture interactions across different regimes. We employ random forests as a ML method for modeling time-varying HAR coefficients based on observed state variables. Our HAR random forest model allows for both stock-specific state variables and market-related variables like the Cboe volatility index (VIX). However, in contrast to conventional regression trees considered in Christensen et al. (2021), leaf nodes of the HAR trees constitute local linear HAR models instead of simple averages across observations. In this way, we use the HAR model's parsimonious structure as a building block for a semiparametric approach. Even though consisting of local parametric models, our HAR forest makes no assumptions about the dependence of the HAR parameters on the state variables. As a combination of ordinary least squares regression and decision trees, our method is easy to estimate. It does not cause computational issues like vanishing and exploding gradients in neural networks.

The idea of constructing local linear models by means of decision trees is not entirely new in the financial forecasting literature. For example, Audrino and Bühlmann (2001) introduce a generalized autoregressive conditional heteroskedasticity (GARCH) tree model. Their model can be seen as a GARCH regime switching model in which regimes are driven by the interest rate, the inflation rate, and the real economic activity. However, this GARCH tree model is computationally expensive which hurts its extension to flexible tree-based methods like random forests. More recent work by Goulet Coulombe (2021) examines the advantage of local linear forests in the field of macroeconomic forecasting by estimating Phillips curves with generalized time-varying parameters.

Our work contributes to the literature on machine learning and forecasting in financial econometrics in multiple ways. We are the first to apply local linear forests (Athey et al., 2019; Friedberg et al., 2020) to realized volatility forecasting. Second, we make the estimation of local linear trees feasible by replacing individual linear models in the leaf nodes of a tree by panel HAR models in the spirit of Bollerslev et al. (2018). This increases the number of observations in the leaves allowing for deeper trees which is often beneficial in random forests. Moreover, the panel structure alleviates joint signals across assets while reducing estimation error. Employing the HAR model as a panel has the additional upside that the estimation is less prone to a small number of days with exceptionally high volatility. Our HAR forest is able to capture different states, regime switches, and other possible nonlinearities in the parameters of the HAR model. Exemplary scenarios where state-dependent coefficients might make a difference are days of very high volatility. On such days we know that realized variance estimates have a high estimation error so we should reduce the weight put on these daily observations relative to calm periods. The method proposed in this paper employs economically motivated linear models per node but allows complex interactions between a large number of covariates that might drive the parameters. A major benefit of including state variables by means of decision trees is that we can employ covariates observed on different frequencies without further adjustments. This links our work to the literature on volatility forecasting using mixed-frequency data sampling (MIDAS) introduced by Ghysels et al. (2005). Even though Ghysels et al. (2019) demonstrate that MIDAS regressions sometimes perform better than the HAR model in forecasting, MIDAS estimation is computationally more expensive and might suffer from identification issues. This is why we choose a local HAR model instead of a local MIDAS model. Another upside of our model is that decision trees don't require potential covariates being measured on a scale that can be transformed to the realized variance scale. Such a transformation is straightforward for variables like semivariances or the VIX but our model also allows for the inclusion of stock market betas or size. Last, variable importance techniques allow us to identify the most important covariates driving the state-dependent coefficients in the HAR forest.

The empirical analysis of our paper addresses forecasting realized volatilities of 186 continuous S&P 500 constituents on data spanning 2000 until 2019. We compare our model with the nested panel HAR model of Bollerslev et al. (2018) and individually estimated HAR models. The pseudo out-of-sample analysis shows that the HAR forest consistently outperforms both benchmark models at mid-horizon volatility forecasts. For forecast evaluation we make use of loss functions that are robust to the measurement error present in volatility estimates; that is the squared error (SE) and the so-called QLIKE loss (Patton, 2011). In the following, all loss ratios are calculated relative to the panel HAR model which performs worse than the individual HAR models and our HAR forests on average. At the 5-day-ahead forecast horizon, we observe that our HAR forest results in a 7 percentage points (pp) lower average QLIKE loss ratio when compared to the stock-specific univariate HAR model. Superior forecasting ability of the HAR forest is even more pronounced at the 22-day-ahead horizon when the observed gain in QLIKE loss ratio is 13pp in comparison to the individual HAR models. We test whether these improvements in forecast performance are also statistically significant by employing the model confidence set (MCS) procedure (Hansen et al., 2011). Again, the HAR forest is particularly good at the 22-day-ahead horizon. It is included in the MCS for 96% of stocks for the QLIKE loss function at the monthly horizon whereas the panel HAR is only included for 15% of the stocks. The individual HAR models are included in the MCS for 48% of stocks. The results for the SE are largely in line with the results for the QLIKE loss but overall less pronounced.

The variable importance measures suggest that the volatility states in the right-hand side of the HAR equation are the most relevant ones for changes in model parameters. The most important variable of the three is the average weekly volatility. Other variables that include higher-order moments of asset returns and low-frequency variables like the beta of a stock have almost no influence. Interestingly, we achieve no gains in forecast performance by including the leverage effect. The variable importance measures of the lagged return imply no significant contribution of the lagged return to forecast performance. One conclusion that can be drawn from our results is that basically all information necessary to improve realized variance forecasts can be inferred from realized variances. This is an interesting finding in the light of works by Bollerslev et al. (2016) who add realized quarticity to the HAR model by means of an interaction effect. Their motivation is to weigh down realized variances that carry a lot of measurement error. In contrast, our HAR forest puts almost no weight on realized fourth-order moments.

The remainder of this paper is organized as follows: Section 2 discusses the concept of the realized volatility and introduces our panel HAR random forest. Empirical results are presented in Section 3. Finally, Section 4 discusses the main finding of the paper and concludes.

2 Modeling and forecasting realized variances

Let $r^2_{t-j\cdot\Delta}$ denote 5-minute log-returns of an asset where $\Delta = 1/M$ and M denotes the number of intraday observations at day t. For example, on a regular trading day at the New York Stock Exchange we have 78 5-minute returns. We aim to forecast the quadratic variation of stock returns in line with the volatility forecasting literature (Andersen and Bollerslev, 1998). As the quadratic variation is not directly observable, we employ the realized variance (RV) defined as the sum of squared 5-minute returns as a proxy for the quadratic variation,

$$RV_t = \sum_{j=0}^{M-1} r^2_{t-j\cdot\Delta}.$$

At time t, we aim to forecast the average realized variance over the upcoming h days; that is $RV_{t+1:t+h} = \frac{1}{h}\sum_{i=1}^{h} RV_{t+i}$. The 5-minute RV is a consistent estimator of the quadratic variation but not robust to microstructure noise (Andersen et al., 2011). However, it has been shown to be a fairly robust choice as a trade-off between using high-frequency data and obstructing micro-structure noise related estimation errors (Liu et al., 2015; Aït-Sahalia and Xiu, 2019).

Even though it has been shown that RV is highly persistent, the degree of persistence might be different across stocks. In Figure 1, we depict the empirical autocorrelation functions (ACF) for realized variances across 186 S&P 500 constituents.[1] The autocorrelations were estimated using the instrumental variables estimator suggested in Hansen and Lunde (2014). We employ their preferred specification, a two-stage least-squares estimator in which lagged realized variances of order 4 to 10 are used as instrumental variables (see Hansen and Lunde, 2014, p.82). The ACF of Apple Inc. and Walmart Inc. are depicted in green and blue while all other ACFs are depicted in light gray. The average ACF is depicted in black. We observe that the persistence in daily RVs varies across stocks. For example, the Apple-ACF is above the Walmart-ACF. One explanation could be that rising short-term market uncertainty or political uncertainty induces increases in the RV of both stocks. However, the RV of Walmart is going down earlier because of the crisis-resistant business model. This is one motivation for us to relax the pooled panel HAR restrictions in Bollerslev et al. (2018) in our proposed HAR forest.

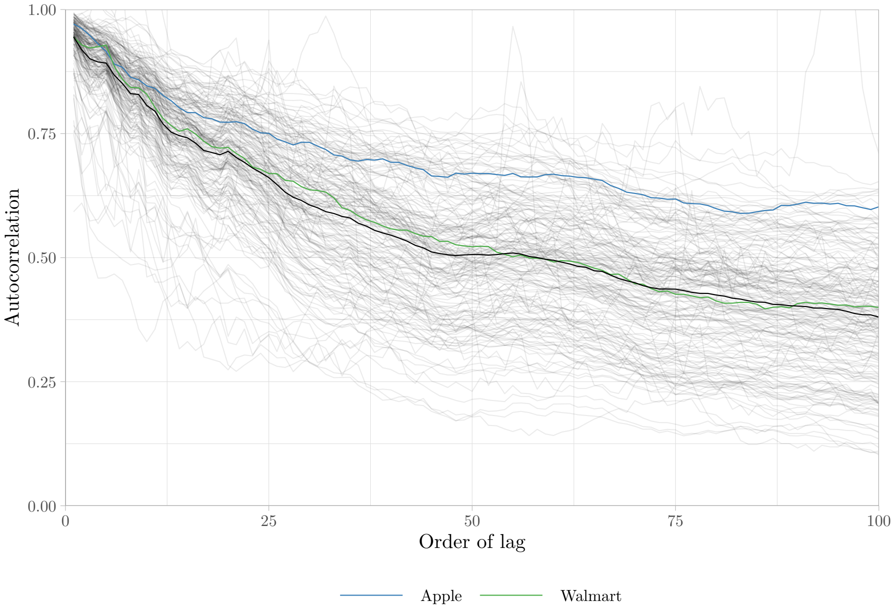
In Section 2.1, we describe the HAR model in more detail, in Section 2.2 we introduce the tree HAR model, and in Section 2.3 we present the aggregated HAR forest.

2.1 Heterogeneous autoregressive models for realized variances

Corsi (2009) introduced the HAR model, which is an additive model with heterogeneous components related to the realized volatility literature. The author motivates his model using the heterogeneous market hypothesis formulated by Müller et al. (1993). This model assumes that each market agent forms expectations for the next period volatility based on different time scales. The AR(1) part of the model characterizes the current realized volatility that is experienced at the same time scale. The hierarchical components mimic the expectations for longer horizons. In other words, the market participants can be

[1] The underlying data will be described in detail in Section 3.1.

Figure 1: Autocorrelation function across stocks.



Notes: We depict the empirical autocorrelation function (ACF) for the RV of 186 S&P 500 constituents. The ACF is estimated using the instrumental variables regression proposed by Hansen and Lunde (2014) with 4 to 10 lagged values of RV.

classified into three heterogeneous groups of agents according to their trading horizons; that is, daily, weekly, and monthly. Each group creates its own component of latent volatility. This cascade of time scales in the volatilities motivates the HAR equation:

$$RV_{t+1} = \overline{RV} + \beta_d(RV_t - \overline{RV}) + \beta_w(RV_t^w - \overline{RV}) + \beta_m(RV_t^m - \overline{RV}) + \varepsilon_{t+1}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_d, \beta_w, \beta_m)^T$ is a parameter vector, the weekly and monthly averages of volatility are defined as $RV_t^w = RV_{t+1:t+5}$ and $RV_t^m = RV_{t+1:t+22}$. $\overline{RV}$ is the average RV of the stock. Note that in preparation of our application to a panel of stocks, we define the HAR equation in terms of deviation from the long-run variance $\overline{RV}$, ε_{t+1} is an innovation term. Predictions for forecast horizons larger than one day can be obtained when substituting RV_{t+1} with $RV_{t+1:t+h}$ on the left-hand side of Eq. (1). We focus on 5- and 22-day-ahead forecasts in our paper.

A broad comparison of the vanilla HAR model to its extensions within a volatility forecasting exercise was presented in Audrino et al. (2019). Their analysis includes competitor models such as AR(p) with the order selected by AIC, the AR-lasso by Audrino and Knaus (2016), the HARQ model by Bollerslev et al. (2016), and the HAR-S-RV-J model by Chen and Ghysels (2011) and are compared by root-mean-square errors. Audrino et al. (2019) conclude that the weighting of the individual components in the HAR model depend on the market conditions. In general, the authors don't find a model that improves out-of-sample forecast performance consistently. This conclusion by Audrino et al. (2019) is supported by means of model confidence sets (Hansen et al., 2011). Therefore,the HAR model remains a valid benchmark for new models in terms of forecasting realized volatilities. This motivates our choice to introduce it to trees and forests.

2.2 Tree-structured HAR models

The question is how to best incorporate time-varying weights into the HAR model. For example, one solution to incorporate time dependence is to fit models on smaller set of data. Model-averaging can be a solution in the presence of structural breaks (Pesaran and Pick, 2011). However, it is not straightforward to capture both short-term market conditions and regime switches by only adjusting the size of the estimation window. In general, a typical estimation window for HAR models is between 4–10 years (see, for example Corsi, 2009; Corsi and Renò, 2012; Bollerslev et al., 2016; Ghysels et al., 2019). However, even short estimation windows of one year as in Kostrov and Tetereva (2019) can not react to very short-lived signals in the data like estimated measurement error from realized quarticities.

One way of addressing the state-dependence of parameters is to consider dynamic HAR models which allow the coefficients in the HAR model to depend on market conditions. However, they require parametric assumptions about the dependence of the coefficients on the market conditions. One such model is the HARQ model (Bollerslev et al., 2016) which employs the realized quarticity for quantifying the measurement error in daily RV estimates. Adding a larger number of covariates in such parametric models might be practically feasible by using regularization techniques. Nonetheless, regularized regressions still need the data to be transformed to be (approximately) linearly related to RV. These kind of transformation might not be possible if one is also interested in interaction effects among covariates driving the HAR coefficients.

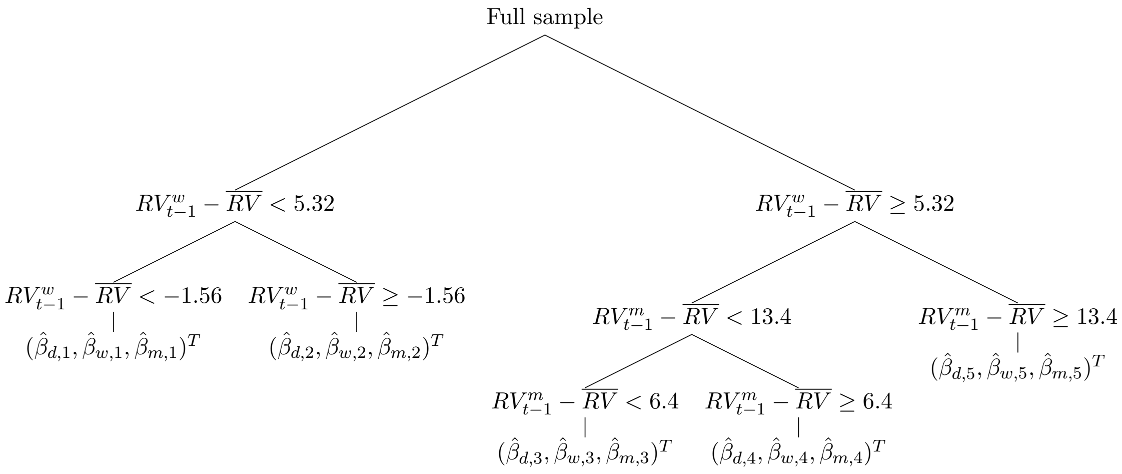
We think that machine learning approaches—in particular random forests—might be helpful to address the issue of time-varying coefficients of the HAR model. In the majority of time series applications, machine learning algorithms are used as nonparametric modeling tools. In the case of random forests, this nonparameric approximation is given by the average of many stepwise functions. In this study, we propose to shift the focus from predicting the values of the realized volatilities by some nonparametric function to modeling parameters $\boldsymbol{\beta} = (\beta_d, \beta_w, \beta_m)^T$ in Eq. (1) as a function of market conditions. On the one hand, it makes the HAR model more flexible without imposing any parametric or distributional assumptions on the parameters and makes it possible to capture regime switches and threshold effects. This is in contrast to latent regime switching models in which just a couple of regimes are feasibly estimated and the number of regimes are difficult to determine. Out forest approach makes it easy to flexibly model evolving parameters in a data-driven way and capture even interactions of different regimes with respect to different market and asset-specific conditions. On the other hand, our local HAR models for the realized volatility are economically meaningful while allowing for time-varying coefficients via trees. This approach relies on the idea of a binary local linear tree where every terminal node parametrizes a local HAR linear model. Each local model is valid for a partition cell of the splitting covariate space. The fitting of such a tree is computationally inexpensive since it relies on minimizing the sum of squares of two linear regression models. This local linear model is in the spirit of a sieve approximation with parsimonious parametric models for nonparametric function. Such an approximation builds on the simple model as a building block and relies on the belief that the classical HAR model is appropriate despite its simplicity. It allows increasing model complexity in a flexible way and improves poor performance at edges, including high values of volatility which are of particular interest in practice.

More precisely, we assume that all splitting rules in the tree $\mathcal{T}$ are based on the state of J splitting variables $\mathcal{J}$. The state vector of these J splitting variables is denoted by $Z_t \in \mathbb{R}^J$. Now, $\mathcal{T}$ assigns each possible value of Z_t one of K terminal nodes denoted by $R_1, \ldots, R_K$. The local model in terminal node R_l is given by

$$RV_{t+1:t+h} = \overline{RV} + \beta_{d,l}(RV_t - \overline{RV}) + \beta_{w,l}(RV_t^w - \overline{RV}) + \beta_{m,l}(RV_t^m - \overline{RV}) + \varepsilon_t. \tag{2}$$

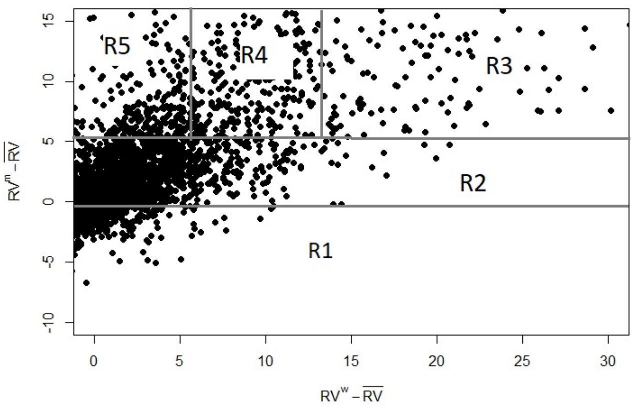
An example of a tree with local HAR models is given in Figure 2. Here, the splitting covariates vector $\mathcal{J}$ contains weekly and daily volatiliies, i.e. RV_t^w and RV_t. The final partitioning contains 5 regions or a set of 5 different linear models. Given that the splitting rules of the tree $\mathcal{T}$ are fixed and the final node assignment is determined by the state vector Z_t, the tree implies a mapping $\widetilde{\mathcal{T}}$ so that $\boldsymbol{\beta}_t = \widetilde{\mathcal{T}}(Z_t)$.

Figure 2: Example of a tree with HAR models in each leaf.


Full sample
$RV_{t-1}^w - \overline{RV} < 5.32$
$RV_{t-1}^w - \overline{RV} \geq 5.32$
$RV_{t-1}^w - \overline{RV} < -1.56$
$(\hat{\beta}_{d,1}, \hat{\beta}_{w,1}, \hat{\beta}_{m,1})^T$
$RV_{t-1}^w - \overline{RV} \geq -1.56$
$(\hat{\beta}_{d,2}, \hat{\beta}_{w,2}, \hat{\beta}_{m,2})^T$
$RV_{t-1}^m - \overline{RV} < 13.4$
$RV_{t-1}^m - \overline{RV} \geq 13.4$
$(\hat{\beta}_{d,5}, \hat{\beta}_{w,5}, \hat{\beta}_{m,5})^T$
$RV_{t-1}^m - \overline{RV} < 6.4$
$(\hat{\beta}_{d,3}, \hat{\beta}_{w,3}, \hat{\beta}_{m,3})^T$
$RV_{t-1}^m - \overline{RV} \geq 6.4$
$(\hat{\beta}_{d,4}, \hat{\beta}_{w,4}, \hat{\beta}_{m,4})^T$

Notes: At each node, we decide how to go through the tree based on the value of a splitting variable; for example $RV^w - \overline{RV}$ at the parent node. Each leaf of the tree consists of a set of coefficients that constitute the local HAR model for this leaf.

Figure 3: Example of the number of observations per leaf from Figure 2.



Notes: Visualizations of a series of splitting rules from Figure 2. Every split is aligned with one of the feature axes. Each split corresponds to drawing a line parallel to one of the axis. For a feature space of size 2, the space is divided into 5 regions, $(R_1, R_2, R_3, R_4, R_5)$, each of which is a 2-dimensional rectangular. Parameters of the panel HAR model are estimated from the data of the corresponding rectangular. For example, all observation in the region R_5 will have the HAR coefficients $(\hat{\beta}_{d,5}, \hat{\beta}_{w,5}, \hat{\beta}_{m,5})^T$. This concept generalises straightforwardly to dimensions greater than two.

It is important to mention that the set of splitting covariates can include the independent variables from the HAR model (i.e., RV, RV^w, RV^m) as in Figure 2 or variables characterising market conditions which are potentially not included in the linear model itself; for example, the Cboe VIX. Moreover, splitting covariates can also be observed on a lower frequency than daily; for example, monthly stock market betas.

The model in Eq. (2) is estimated in analogue to simple classification and regression trees. In every node, the objective is to find the splitting point that minimizes some objective function. Such a minimum is searched over all possible values in all splitting covariates. In other words, the following sum of two residual sum of squares needs to be minimized:

$$\min_{c \in \mathbb{R},\ j \in \mathcal{J}} \left(\min_{\beta} \sum_{t | Z_{j,t} < c} \widehat{\varepsilon}_t^2 + \min_{\beta} \sum_{t | Z_{j,t} \geq c} \widehat{\varepsilon}_t^2 \right), \tag{3}$$

where $\widehat{\varepsilon}_t$ are estimated residuals generated by the HAR model (Eq. (1)) and c is a potential threshold which is chosen from the values of all splitting covariates $Z_{j,t}, j \in \mathcal{J}$.

2.3 HAR forests

Although a tree of HAR models is more flexible than a single HAR model, it inherits some particular drawbacks from conventional classification and regression trees. First, one needs to prune the tree or to choose an appropriate stopping rule to avoid overfitting the data. Usually, this step requires tuning of additional hyperparameters. Second, they are typically unstable in the sense that small changes in the data can change the estimated tree a lot. In order to deal with these unsatisfactory side effects, two sources of randomness can be introduced in analogue to random forests by Breiman (2001). First, B bootstrap samples are used instead of original samples and B forecasts are generated, one for each bootstrap sample. This technique is called bagging—short for bootstrap aggregation—and was first introduced by Breiman (1996). With bagging, individual trees are allowed to become large. Bagging leads to the trees being more distinct and, as a consequence, reduces the correlation among the trees. Second, at each each split only a randomly selected subset of splitting covariates $\mathcal{J}^- \subset \mathcal{J}$ is considered in deciding which to split on. This approach is used since trees constructed based on bootstrap samples might still be highly correlated, especially because our target variable exhibits long memory. Usually, the square root of the number of variables in the splitting covariate space $\sqrt{|\mathcal{J}|}$ is chosen as the size of the randomly selected set. The random forest predictions are obtained by averaging all tree-specific predictions. When switching from individual trees to the forest, one can construct large trees and does not need to perform pruning.

It is important to mention that the introduced HAR forest approach is different from applying conventional regression trees to volatility modeling. For example, Christensen et al. (2021) apply regression trees among other machine learning algorithms to model the realized volatility as a nonparametric, piecewise constant, function of its past values. Our approach is different. Instead of predicting the realized volatility as a simple average of all observations in the leaf, we predict it by a leaf-specific HAR model. If there is a linear relationship in the data as it is in the case of RV, simple regression trees need to run very deep for a good fit whereas local linear trees can be shallow. Consequently, there is a smaller chance of overfitting for such data which otherwise harms out-of-sample performance. This is why we introduce a regime-dependent HAR model whose parameters are subject to regimes governed by changing market conditions and stock-specific characteristics.

The fact that in every leaf the linear HAR model needs to be estimated instead of computing the average leads to two technical issues that need to be addressed. First, fitting linear models in every node is computationally more costly than just computing the mean. This is of minor importance when the realized volatility of one asset needs to be predicted but might be more crucial when the realized volatility of, for example, all S&P 500 stocks are of interest to an investor. Second, a higher minimum number of observations is required in every node due to larger amount of parameters to be estimated. In contrast to taking the simple average, the simple HAR model (Eq. (1)) involves three parameters to be estimated which is still of moderate size. Adding large numbers of regressors to the HAR equation is not feasible as the number of observations in each leaf become small for trees targeting only one stock.

Potentially, one can encounter the curse of dimensionality as in Goulet Coulombe (2021) by adding a penalization term to the linear model. This, however, increases the computational time and involves the tuning of additional hyperparameters. We suggest a straightforward alternative approach and switch from the individual HAR models to a panel HAR model. It is a well-known fact that pooled estimators result in some bias due to heterogeneity but provide substantial efficiency gains from pooling. We follow Bollerslev et al. (2018) and suggest to estimate a panel of HAR models instead of individual HAR models. This is why we employ demeaned realized volatilities instead of realized volatilities in absolute value. As a consequence, we can use simple pooled estimators without explicit fixed effect estimation.

Hence, the model of each individual tree targeting the average h-step-ahead volatility suggested in this study is

$$RV_{i,t+1:t+h} - \overline{RV_i} = \big(\beta_{d,l}(RV_{i,t} - \overline{RV_i}) + \beta_{w,l}(RV^{w}_{i,t} - \overline{RV_i}) + \beta_{m,l}(RV^{m}_{i,t} - \overline{RV_i})\big)\, I_{\mathcal{T}(Z_{\cdot,i,t})\in R_l} + \varepsilon_{i,t}, \quad (4)$$

where $i = 1, \dots, N$ with N being the number of assets in the model. Entries in the splitting covariates $Z_t \in \mathbb{R}^{J\times N}$ might depend on t or/and i like $RV_{i,t}$, VIX_t, or $\overline{RV_i}$. In the case of a forest full of panel HAR models, the splitting objective also needs to be slightly adjusted; that is, the sum of losses of all (pooled) HAR models in the panel has to be considered at every split:

$$\min_{c\in\mathbb{R},\ j\in\mathcal{J}} \left(\min_{\beta} \sum_{i,t|Z_{j,i,t}<c} \widehat{\varepsilon}^{2}_{i,t} + \min_{\beta} \sum_{i,t|Z_{j,i,t}\geq c} \widehat{\varepsilon}^{2}_{i,t} \right), \quad (5)$$

where $\widehat{\varepsilon}^{2}_{i,t}$ are derived from Eq. (4). In analogue to individual HAR random forests, the final forecast is the average of B forecasts delivered by $b = 1, \dots, B$ individual panel HAR trees (Eq. (4)), where each tree is constructed on a bootstrap sample of original data with a random subset of splitting variables at each node. More details on estimating our model can be found in Section 3.2.

3 Empirical analysis

In this section, we present the empirical analysis which includes describing the S&P 500 constituents data in Section 3.1 and discussing the estimation techniques in Section 3.2. Thereafter, we evaluate the forecasts generated from the local linear HAR models relative to two benchmark models in Section 3.4 and discuss the importance of different splitting covariates in Sections 3.5.

3.1 Data

Our data set contains a range of different variables for the 186 permanent constituents of the S&P 500 index from 2000–2019 that are classified as common stocks in the Center of Research in Security Prices (CRSP) data and for which we have enough data on each day to calculate all intraday measures like realized skewness. Similar to Bollerslev et al. (2019) and Bollerslev et al. (2021), we merge daily CRSP data with New York Stock Exchange (NYSE) transactions and quotes (TAQ) intraday data. Open and close prices per day are taken from the daily CRSP data files whereas all other intraday transaction data is obtained from the NYSE TAQ. Merging the two data sets is carried out via the WRDS linking tables.[2] This intraday data is cleaned according to Barndorff-Nielsen and Shephard (2002) and we include only trades from the exchange that is referenced in the daily CRSP data. Our target variable is the daily intraday RV but we want to include the lagged return as a leverage effect in our forests which should be adjusted for dividend payments that are only available in the CRSP data.

We include a large number of potential splitting variables in our local linear forest. First, the splitting variables include the RV^d, RV^w, and RV^m that constitute the right-hand-side equation in our local HAR models. Second, we include additional daily stock-specific information derived from intraday data: Semivariances of positive and negative returns which are denoted by RV^+ and RV^- (Barndorff-Nielsen et al., 2010; Patton and Sheppard, 2015),

$$RV_{i,t}^{+} = \sum_{j=0}^{M-1} r_{i,t-j\cdot\Delta}^2 I_{\{r_{i,t-j\cdot\Delta}\geq 0\}} \text{ and } RV_{i,t}^{-} = \sum_{j=0}^{M-1} r_{i,t-j\cdot\Delta}^2 I_{\{r_{i,t-j\cdot\Delta}< 0\}},$$

and a jump-robust measure of volatility denoted by MedRV (Andersen et al., 2012),

$$MedRV_{i,t} = \frac{\pi}{6-4\sqrt{3}+\pi}\left(\frac{M-1}{M-3}\right)\sum_{j=1}^{M-2}\mathrm{Median}(|r_{i,t-(j-1)\cdot\Delta}|, |r_{i,t-j\cdot\Delta}|, |r_{i,t-(j+1)\cdot\Delta}|).$$

Higher-order moments included in this analysis are the realized skewness and realized kurtosis (Amaya et al., 2015),

$$RSkew_{i,t} = \frac{\sqrt{M-1}\sum_{j=0}^{M-1} r_{i,t-j\cdot\Delta}^3}{RV_{i,t}^{3/2}} \text{ and } RKurt_{i,t} = \frac{(M-1)\sum_{j=0}^{M-1} r_{i,t-j\cdot\Delta}^4}{RV_{i,t}^2}.$$

One splitting variable that is common to all stocks at each day t is the Cboe volatility index VIX which is also employed in HAR models since Bekaert and Hoerova (2014) who forecast aggregate stock market volatility instead of individual stocks. Monthly idiosyncratic volatility, beta, and momentum are taken from Gu et al. (2020) and merged via the CRSP permno identifier. Summary statistics of all

[2] https://wrds-www.wharton.upenn.edu/

variables can be found in Table 1.

Table 1: Average full sample summary statistics.

	Obs	Mean	Median	Std	Skew	Kurt	Min	Max
Panel A: Daily variables								
RV^d	4987	3.436	1.621	8.094	12.913	394.238	0.134	270.973
RV^w	4987	3.438	1.735	6.401	6.777	87.160	0.279	109.188
RV^m	4987	3.451	1.814	5.453	4.666	34.561	0.442	56.108
RV^+	4987	1.732	0.794	4.364	13.679	415.122	0.051	155.661
RV^-	4987	1.704	0.784	4.173	12.990	405.309	0.047	140.157
MedRV	4987	3.167	1.485	7.343	12.554	371.330	0.104	233.097
Return	4987	0.055	0.047	2.046	0.164	20.353	-20.732	21.408
RSkew	4987	0.036	0.022	1.056	0.092	5.509	-5.802	6.176
RKurt	4987	2.064	1.645	1.371	3.508	22.817	0.703	17.672
VIX	4987	19.501	17.310	8.494	2.172	10.498	9.140	80.860
Panel B: Monthly variables								
Beta	238	0.034	0.030	0.014	0.882	2.661	0.019	0.065
Idiovol	238	0.854	0.861	0.286	-0.093	2.619	0.303	1.442
Mom	238	0.114	0.101	0.292	0.577	5.714	-0.546	1.319
Panel C: Cross-sectional variable								
$\overline{\text{RV}}$	186	3.434	2.820	1.720	1.420	4.993	1.272	9.881

Notes: This table reports average summary statistics across all 186 S&P 500 constituents included in our analysis. We report the cross-sectional average of: number of observations, mean, median, standard deviation, skewness, kurtosis, minimum, and maximum across stocks. The sample period is 2000–2019. Note that the first and last month are not included in this summary statistic due to the missing observations by aggregating RV over up to 22 days.

3.2 Estimation

For estimation and subsequent evaluation, we employ a rolling window estimation scheme with yearly reestimation. That is, the first estimation sample comprises ten years of daily data from 2000–2009. Both the forests and benchmark models are estimated on this estimation sample and the model parameters are employed throughout forecasts issued in 2010. At the end of 2010, we reestimate all models on a new estimation window spanning 2001–2010. We employ the rolling window scheme for two reasons: First, it accounts for possible fundamental changes (e.g. if the average RV of a stock de-/increases over time). Second, we can employ Diebold-Mariano-type tests for forecast comparison even though the models are nested (Diebold and Mariano, 1995; Giacomini and White, 2006).

Bagging: For the HAR forests, the estimation scheme follows the general idea of random forests but we add slight adjustments to the bagging procedure. Bagging means that for each tree in the forest, we select a random sample of the data with replacement. However, given the panel structure of our data we don't sample randomly from the entire panel but we sample dates only. Our reasoning for this sampling scheme is the following: We include the VIX as a cross-sectional variable that has the the same value for all stocks on a given day. The optimal split w.r.t. the VIX is determined via a grid search over all possible VIX values in the nodes' subsample. With this in mind, our calendar-time sampling procedure that at each node we have enough observations per VIX value to fit a local linear model and, as a side effect, we preserve the full cross-sectional structure at each node. In order to save some computational time, we only consider the splitting variable's 1%- to 99%-percentile as a grid if there are more than 100 potential splitting points.

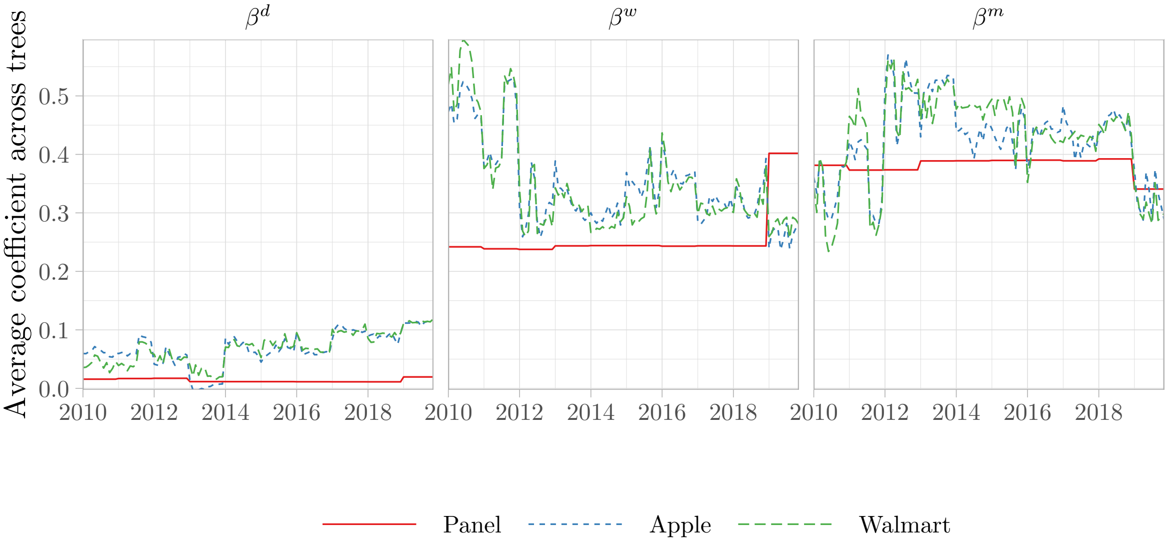
Feature subsampling: Breiman (2001) proposed to de-correlate the trees that form a forest. The reason is that we calculate our final predictions by averaging across predictions derived from the individual trees; that is, de-correlated trees help with reducing the variance of the averaging step. We follow the arguments of Goulet Coulombe (2021), and draw random subsets of features $\mathcal{J}^- \subset \mathcal{J}$ instead of $\mathcal{J}$ in Eq. (5). In contrast to Friedberg et al. (2020) and Goulet Coulombe (2021), we refrain from including a ridge regression penalty term in Eq. (5). The reason is that we have relatively large sample sizes in our final leaves with minimum sizes of either 2,000 or 500 observations which we regard to be sufficient for getting good ordinary least squares estimates for our local HAR models. This has the additional upside that we do not need to tune an additional hyperparameter. Across estimation periods, aggregating 200 trees per forest turns out to be more than sufficient for generating stable out-of-sample forecasting results.

The benchmark stock-specific HAR models and the Panel HAR models by Bollerslev et al. (2018) are estimated via ordinary least squares regression. Last, all employed HAR models do not ensure that all predicted RVs are positive. Hence, we sanitize all predictions per stock i by capping predictions at the empirical minimum and maximum value of the rolling window in-sample target RV of stock i to ensure that our results are not mainly driven by a small number of observations. The filter is invoked for only a handful of observations. Nonetheless, we need strict positive forecasts for employing the QLIKE loss in Section 3.4.

3.3 Time-varying parameters

One advantage of our local linear forest over other ML techniques like neural networks is that we can depict the state-implied time-varying HAR coefficients. At each point in time and for each stock, we can go through the trees in the forest to obtain the coefficients depending on Z_t. The average coefficients across trees for each day are presented in Figure 4 for two exemplary stocks: Apple Inc. and Walmart Inc. We choose these two stocks as both are large companies but their industries are differently affected by short-term market expectations. Figure 4 consists of three panels depicting the three HAR coefficients β^d, β^w, and β^m. The figure is based on the HAR forest with a minimum size of 2000 and for local HAR models targeting the 22-day-ahead forecast horizon. For comparison, we include the piecewise constant rolling window estimates from the plain panel HAR.

Figure 4: State-dependent time-varying HAR coefficients for two stocks along fixed panel HAR coefficents.



Notes: This figure depicts the state-implied coefficient for the Forest (2000) and the 22-day-ahead forecast horizon. Coefficients are averaged across trees and aggregated to a monthly frequency. We report these number for two exemplary stocks: Apple Inc. and Walmart Inc.

As for independently estimated HAR models, we observe that the weight on the lagged daily RV is considerably lower than the weight on the weekly and monthly aggregated RV. The panel HAR has almost constant coefficients over the entire sample. Only for the last rolling window estimation at the end of 2018 we see large changes in the weekly and monthly coefficient because the heights of the financial crisis at the end of 2008 drop out of the rolling window estimation period. The forest-implied time-varying coefficients of the two exemplary stocks clearly deviate from the panel HAR estimates. We observe a higher weight on the lagged daily observation which can be explained by the fact that both stocks are Dow Jones Industrial Average constituents and, thus, attain higher news coverage than stocks that are in the S&P 500 but not in the S&P 100. In general, we see that both the rolling window estimation and the state-dependence of the coefficients affect the coefficients over time.

3.4 Forecast evaluation

In the remainder of this analysis, we follow the arguments in Patton (2011) and use the SE and QLIKE loss as evaluation criteria. For a forecast of the average volatility over next h days $\widehat{RV}_{t+1:t+h|t}$ and its realization $RV_{t+1:t+h}$, the SE and QLIKE are defined as

$$\begin{aligned}\text{SE}\left(RV_{t+1:t+h}, \widehat{RV}_{t+1:t+h}\right) &= \left(RV_{t+1:t+h} - \widehat{RV}_{t+1:t+h}\right)^2, \\ \text{QLIKE}\left(RV_{t+1:t+h}, \widehat{RV}_{t+1:t+h}\right) &= RV_{t+1:t+h}/\widehat{RV}_{t+1:t+h} - \ln\left(RV_{t+1:t+h}/\widehat{RV}_{t+1:t+h}\right) - 1.\end{aligned}$$

As discussed in Patton (2011), the QLIKE is less sensitive with respect to extreme observations than the squared error loss but we report both for completeness. Further, it can be shown that the moment conditions required for Diebold and Mariano (1995) or Giacomini and White (2006) type tests are weaker under QLIKE than under squared error loss (see Patton, 2006).

We consider the following forecasting schemes. Based on the information available on day t, cumulative volatility forecasts are computed for horizons of 5 days and 22 days. Forecast evaluation is based on the volatility proxy $RV_{i,t+1:t+h}$ and the respective forecast of model j for stock i is denoted by $\widehat{RV^j}_{i,t+1:t+h|t}$. For each loss function L and with hindsight, we can measure the average loss of model j for stock i across time as

$$L_i^j = \frac{1}{|OOS|} \sum_{t \in OOS} L(RV_{i,t+1:t+h}, \widehat{RV^j}_{i,t+1:t+h|t}). \tag{6}$$

where $|OOS|$ denotes the number of days in the out-of-sample (OOS) period. We denote the stock specific loss of the benchmark forecast (Panel HAR) by L_i^B. As a measure for the forecast accuracy of a particular model j relative to the benchmark, we consider the following statistics: The average loss ratio relative to the benchmark model,

$$AL^j = \frac{1}{N} \sum_{i=1}^{N} \frac{L_i^j}{L_i^B} \tag{7}$$

and the loss rate which reports the share of stocks for which model j outperforms the benchmark,

$$LR^j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{L_i^j/L_i^B<1}. \tag{8}$$

Besides the two aggregated measures AL^j and LR^j, we formally test for superior predictive ability. We base our analysis on the MCS approach introduced by Hansen et al. (2011) that is obtained as follows: Denote by $\mathcal{M}$ the set of all competing models. For each loss function L and stock i we define

$$d_{i,t+1:t+h}^{j,l} = L(RV_{i,t+1:t+h}, \widehat{RV^j}_{i,t+1:t+h|t}) - L(RV_{i,t+1:t+h}, \widehat{RV^l}_{i,t+1:t+h|t})$$

as the difference in the respective loss of models j and l for the cumulative forecast horizon h. We compute the average loss difference per stock, $\bar{d}_i^{j,l}$, and calculate the test statistic

$$t_i^{j,l} = \bar{d}_i^{j,l} / \sqrt{\widehat{\mathbf{Var}}\left(d_i^{j,l}\right)} \text{ for all } j, l \in \mathcal{M}. \tag{9}$$

The MCS test statistic is then given by $T_{i,\mathcal{M}} = \max_{j,l \in \mathcal{M}} |t_i^{j,l}|$ and has the null hypothesis that all models have the same expected loss. Under the alternative, there is some model j that has an expected loss greater than the expected loss of all other models $l \in \mathcal{M} \setminus j$ for stock i. If the null hypothesis is rejected, the worst performing model is taken out of the set of models under consideration. The test is performed iteratively, until no further model can be eliminated. We denote the final set of surviving models for stock i by $\mathcal{M}_{i,MCS}$. This final set contains the best forecasting model with confidence level $1 - \nu$. We set $\nu = 0.1$. This choice is common practice in the literature. See, for example, Laurent et al. (2013) and Liu et al. (2015).

Since the asymptotic distribution of the test statistic $T_{i,\mathcal{M}}$ is nonstandard, we approximate it by block-bootstrapping as proposed by Hansen et al. (2011) with a conservative block length equal 44 which equals twice the maximum forecast horizon. In our analysis, 5,000 bootstrap replications at each stage were sufficient in order to obtain stable results.[3] As an aggregate measure, we calculate the share of stocks for which model j is in the model confidence set by

$$MCSR^j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{j \in \mathcal{M}_{i,MCS}}.$$

Fourth, we regress the h-step-ahead average log-return on the square-root of the h-step-ahead variance forecast and look at the corresponding R^2 coefficient to examine the economic importance of our new model.

We report the aggregated measures in Table 2. In Panel A we report the results for 5-day-ahead forecasts and in Panel B for 22-day-ahead forecasts. The different models listed are the full panel model (Panel), the individual HAR models (Indiv), and the local HAR forests with a minimum size of observations equal to 2000 or 500. As AL^j and LR^j are defined relative to the benchmark panel HAR model, we only list dedicated numbers for this model w.r.t. the MCSR and Ret-R^2 statistic.

Overall, the results are relatively favorable for the HAR forest when evaluated with the QLIKE loss whereas the SE results are less consistent across forecast horizons. When examining the SE loss for 5-day-ahead forecasts, we observe that the ALs for all three models are very similar but the SE-LRs are considerably lower for the forest. However, these differences in SE-LR do not seem to be statistically significant most of the time as one of the forests has an SE-MCSR that is only 5pp lower than the SE-MCSR of the individual HAR models. The QLIKE forecast evaluation at the 5-day-ahead horizon makes a strong point for employing our HAR forest models. We observe QLIKE-ALs that are 7pp lower and QLIKE-LRs that are 8pp higher than the AL and LR of the individual HAR models. What is interesting to see is that even though both forests perform relatively similar in terms of AL, it is the case that the two forest models have QLIKE-MCSRs that differ up to 44pp. This can be explained by the sequential nature of the MCS. The MCS is based on *test statistics* of pairwise comparison. This means that a model might

[3] For implementing the MCS procedure, we use the R package *rugarch* (Ghalanos, 2018) which includes the implementation used in the MFE Matlab toolbox by Kevin Sheppard. See: https://www.kevinsheppard.com/MFE_Toolbox.

Table 2: Unconditional forecast evaluation results.

	SE			QLIKE			Ret-R^2
Model	AL	LR	MCSR	AL	LR	MCSR	
Panel A: 5-day-ahead							
Panel	—	—	0.667	—	—	0.097	0.247
Indiv	0.982	**0.613**	**0.855**	0.868	0.892	0.441	0.265
Forest (2000)	0.987	0.371	0.661	0.794	0.973	0.554	0.280
Forest (500)	**0.981**	0.409	0.806	**0.776**	**0.984**	**0.995**	**0.313**
Panel B: 22-day-ahead							
Panel	—	—	0.505	—	—	0.151	0.811
Indiv	0.922	**0.753**	0.656	0.847	0.898	0.473	0.858
Forest (2000)	**0.857**	0.683	**0.941**	**0.719**	**0.962**	**0.962**	0.987
Forest (500)	0.892	0.597	0.511	0.728	**0.962**	0.887	**1.080**

Notes: We report the average loss ratio (AL) relative to the Panel HAR benchmark model, the frequency of one model performing better than the benchmark model (LR), and the MCS inclusion rate (MCSR) across stock; see Section 3.4. The last column (Ret-R^2) reports the average R^2 coefficients (in percentages) from regressing the h-step-ahead average log-return on the square-root of the h-step-ahead variance forecast. Results for individually fitted HAR models are found in the rows labelled "Indiv". Forest (2000) and Forest (500) deviate in the minimum number of observations per leaf; that is, Forest (500) is derived from an ensemble of deeper trees than Forest (2000). The number of trees per forest is 200. The forecasts are issued daily and, hence, overlapping. The out-of-sample period is 2010–2019. Averages are taken across 186 stocks. Numbers in bold indicate the lowest AL, highest LR, and highest MCSR.

be excluded from the MCS even though it does not have the lowest average loss because it is marginally but consistently outperformed by one of the better models. QLIKE-MCS inclusion rates of almost 100% for the Forest (500) make a point that the forest models perform better than both the individual HAR models and the simple panel HAR with a QLIKE-MCSR slightly below 1%. Of course, the average R^2 from regressing returns on volatilities is small in magnitude with the two R^2s based on the HAR forests being the highest. However, even small increases in the predictive R^2 can imply considerable increases in utility for a mean-variance investor (Campbell and Thompson, 2008).

When we turn to the 22-day-ahead forecast results in Table 2, we see that the QLIKE results are very much in line with QLIKE results from the 5-day-ahead horizon. Also for the SE, the results become more favorable for the forest models. The SE-ALs decrease by up to 14% for the forest models but only 6% for the individual HAR models. The SE-LR of the individual HAR models is still the highest but the SE-MCSRs are well in favor of the Forest (2000). Nonetheless, the SE results are still a mixed bag and need further investigation. Regarding QLIKE evaluation, the forests turn out to have 13pp lower QLIKE-ALs, and QLIKE-LRs that are 6pp higher than the QLIKE-LR of the individual HAR model. The QLIKE-MCSRs are in line with these observations. The individual HAR is included in the QLIKE-MCS for only 48% of stocks but the forests are included in the QLIKE-MCS for up to 96% of the stocks. Overall, it is fair to say that HAR forests are the better choice for forecasting stock-specific RVs if the relevant evaluation criterion is QLIKE and/or the forecast horizon of interest is 22 days. This translates into higher average predictive R^2s with a maximum value of 1.1% for the Forest (500). For all models, we see a threefold increase in Ret-R^2 when advancing the forecast horizon from 5 to 22 days. This resembles the widely documented empirical observation that return predictability increases in the forecast horizon. An open question remaining for further research is how to endogenously determine the optimal number of trees without an expensive grid search.

3.5 Variable importance

In order to gain insight which covariates contribute the most to our favorable forecasting results, we turn to variable importance measures which are based on reshuffling one of the splitting variables at a time. These measures are calculated for each tree and, then, aggregated across trees in the forest by averaging. There are typically two types of variable importance measures: First, each tree has its own bagged sample it was fitted on which implies that for each tree we also have an out-of-bag (OOB) sample. We shuffle one variable at a time inside this OOB sample and calculate the relative increase or decrease in SE and QLIKE. It's important to note that analogically to bagging, the daily and the monthly variables are shuffled with respect to time within the stock. Only mean realized volatility is shuffled cross-sectionally. As in our forecast evaluation, we calculate the relative change in OOB error per stock, average across stocks as we did for AL^j, and average across trees per forest in the last step. Note that our panel setting again adds an additional aggregation step by not pooling prediction errors across stocks but only relative forecast performance by averaging loss ratios. Last, we average across this forest-specific variable importance over time as we employ a rolling window estimation scheme. We denote this aggregated measure of variable importance as VI_{OOB}. However, if we are interested in the importance of a variable on actual OOS performance, VI_{OOB} might be misleading because even though each OOB sample has no overlap with the estimation sample per tree, the OOB samples overlap across trees. Hence, we can examine the same procedure as for VI_{OOB} but apply it to the OOS period per forest. We denote this measure by VI_{OOS}. For the variable importance discussion we focus on the 22-day-ahead forecast horizon.

Table 3 and Table 4 report the two variable importance measures. In Table 3, we see that for both forest specifications the highest VI_{OOB} is found for the variables that form the right-hand side in the HAR equation. In our sample, RV^w is the most important variable. The in-sample errors increase by 100–150% if one rearranges the RV^w observations. In general, VI_{OOB} is higher for the Forest (500) model than for the Forest (2000) model. This can be explained by the fact that the variance across trees is higher if one allows for deeper trees. As a consequence, deeper trees are also more sensitive to reshuffling the splitting variables because of overfitting the data. What is interesting is that other measures derived from intraday data—like realized semivariances or skewness—have VI_{OOB}- and VI_{OOS}-values that are hardly distinguashable from zero. Hence, in the context of our HAR forest we don't find evidence that there is much to derive from intraday data beyond the realized variance itself. One external variable that has a high variable importance of up to 13% is the VIX. This means that it is important to have time-varying parameters that depend on the market condition. Similarly, splitting the data cross-sectionally by $\overline{\text{RV}}$ leads to VI_{OOS}-values of up to 35%. Hence, even though the pooled estimation might be beneficial overall, it is still important to allow for some cross-sectional heterogeneity. This is something we also see in our forecasting results where the simple panel model perform worse than the individual HAR models. Incorporating the leverage effect by including the daily lagged return or a momentum covariate has almost no influence. Unfortunately, the contribution of lower-frequency variables like monthly beta estimates is also low.

Table 3: OOB variable importance VI_{OOB}.

	Forest (2000)		Forest (500)	
	SE	QLIKE	SE	QLIKE
Panel A: Daily variables				
RV^d	1.129	1.255	1.185	1.377
RV^w	**1.968**	**2.544**	**1.925**	**2.916**
RV^m	1.781	2.045	1.821	2.194
RV^+	1.003	1.005	1.006	1.016
RV^-	1.001	1.006	1.006	1.021
MedRV	1.002	1.005	1.006	1.022
Return	1.001	1.002	1.003	1.006
RSkew	1.000	1.000	1.000	1.000
RKurt	1.001	1.000	1.002	1.002
VIX	1.015	1.058	1.036	1.132
Panel B: Monthly variables				
Beta	1.001	1.006	1.007	1.018
Idiovol	1.000	1.001	1.001	1.003
Mom	1.001	1.005	1.008	1.013
Panel C: Cross-sectional variable				
$\overline{\text{RV}}$	1.001	1.013	1.013	1.031

Notes: In this table, we report the variable importance measure VI_{OOB}. Because we have a yearly rolling window estimation scheme, we report the average VI_{OOB} across ten forests. For more details on the construction of this table see Section 3.5.

4 Discussion

In this paper, we propose to combine the machine learning technique random forest with the well-established and economically motivated HAR model for forecasting realized volatilities in a large cross

Table 4: OOS variable importance VI_{OOS}.

	Forest (2000)		Forest (500)	
	SE	QLIKE	SE	QLIKE
Panel A: Daily variables				
RV^d	0.998	1.030	1.048	1.068
RV^w	**1.065**	**1.174**	1.068	**1.241**
RV^m	1.041	1.065	1.091	1.103
RV^+	1.006	1.003	1.014	1.010
RV^-	1.003	1.003	1.015	1.008
MedRV	1.004	1.003	1.018	1.011
Return	1.002	1.002	1.006	1.003
RSkew	1.002	1.001	1.003	1.000
RKurt	1.004	1.001	1.012	1.002
VIX	1.001	1.015	1.005	1.027
Panel B: Monthly variables				
Beta	1.000	1.000	1.002	1.001
Idiovol	1.000	1.000	1.001	1.001
Mom	1.002	1.004	1.007	1.006
Panel C: Cross-sectional variable				
$\overline{\text{RV}}$	1.046	1.027	**1.355**	1.051

Notes: In this table, we report the variable importance measure VI_{OOS}. Because we have a yearly rolling window estimation scheme, we report the average VI_{OOS} across one-year OOS periods. For more details on the construction of this table see Section 3.5.

section of individual stocks. Following the empirical evidence of time-varying coefficients in the HAR model, we propose to model the coefficients as nonparametric functions of state variables that can vary across stocks and/or time like realized volatility, realized skewness and kurtosis, semivariances, the VIX, stock market betas, momentum returns, or monthly idiosyncratic volatility. This is achieved by employing these characteristics as splitting variables in individual regression trees to allow for market-conditions-dependent parameters in the leaf-specific HAR model. This is an extension over previous work on machine learning in financial forecasting where the local model is only a constant equal to the simple empirical average across the terminal node's observations. In our regression setting this is equivalent to a local linear model which includes an intercept only. We suggest to extend this piecewise constant model to a local linear model building on the large literature documenting the strong predictive performance of HAR models. Because individual trees have low bias but high variance, we turn to random forest forecasts instead of individual trees in a second step. Our HAR forest makes use of bootstrap aggregation and feature subsampling.

For random forests it is often beneficial to go for deeper trees but this decreases the number of observations per terminal node in each tree. This becomes a challenge as in our time series setting with autoregressive local linear models we typically need more observations per leaf than in classical regression trees. We address this issue by estimating panel HAR models in the spirit of Bollerslev et al. (2018) in each terminal node. In addition to addressing data sparsity issue, it contributes to better forecasting properties due to joint signals in multiple realized volatility time series.

In an empirical application to S&P 500 constituents from 2000 until 2019, we show that the proposed model can outperform the benchmarks under various loss functions and across different forecast horizons. The time-varying parameters of the HAR forest are found to be distinctively different from the fixed coefficients of the panel HAR. Our HAR forest model demonstrates superior forecasting properties in the sense of achieving smaller QLIKE and SE losses. The statistical significance of our results is demonstrated by means of model confidence sets (Hansen et al., 2011). By evaluating variable importance measures, we show that the superior forecasting ability is mostly due to more efficient usage of realized volatility estimates. Specifically, average weekly volatility is found to be the most important splitting variable in the HAR forest. Moreover, our model is computationally robust and simple to implement in contrast to other machine learning methods like neural networks.

References

Aït-Sahalia, Y. and D. Xiu (2019): "A Hausman test for the presence of market microstructure noise in high frequency data," *Journal of Econometrics*, 211, 176–205.

Amaya, D., P. Christoffersen, K. Jacobs, and A. Vasquez (2015): "Does realized skewness predict the cross-section of equity returns?" *Journal of Financial Economics*, 118, 135–167.

Andersen, T. G. and T. Bollerslev (1998): "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts," *International Economic Review*, 39, 885–905.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2011): "Realized volatility forecasting and market microstructure noise," *Journal of Econometrics*, 160, 220–234.

Andersen, T. G., D. Dobrev, and E. Schaumburg (2012): "Jump-robust volatility estimation using nearest neighbor truncation," *Journal of Econometrics*, 169, 75–93.

Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized random forests," *Annals of Statistics*, 47, 1179–1203.

Audrino, F. and P. Bühlmann (2001): "Tree-structured generalized autoregressive conditional heteroscedastic models," *Journal of the Royal Statistical Society: Series B*, 63, 727–744.

Audrino, F., C. Huang, and O. Okhrin (2019): "Flexible HAR model for realized volatility," *Studies in Nonlinear Dynamics & Econometrics*, 23.

Audrino, F. and S. D. Knaus (2016): "Lassoing the HAR Model: A model selection perspective on realized volatility dynamics," *Econometric Reviews*, 35, 1485–1521.

Barndorff-Nielsen, O. E., S. Kinnebrock, and N. Shephard (2010): "Measuring downside risk—realized semivariance," *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*, 1–22.

Barndorff-Nielsen, O. E. and N. Shephard (2002): "Econometric analysis of realized volatility and its use in estimating stochastic volatility models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 253–280.

Bekaert, G. and M. Hoerova (2014): "The VIX, the variance premium and stock market volatility," *Journal of Econometrics*, 183, 181–190.

Bollerslev, T., B. Hood, J. Huss, and L. H. Pedersen (2018): "Risk everywhere: Modeling and managing volatility," *Review of Financial Studies*, 31, 2730–2773.

Bollerslev, T., S. Z. Li, and B. Zhao (2019): "Good volatility, bad volatility, and the cross section of stock returns," *Journal of Financial and Quantitative Analysis*, 55, 751–781.

Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2016): "Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting," *Journal of Econometrics*, 192, 1–18.

——— (2021): "Realized semibetas: Disentangling "good" and "bad" downside risks," *Journal of Financial Economics*, forthcoming.

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

——— (2001): "Random forests," *Machine Learning*, 45, 5–32.

Campbell, J. Y. and S. B. Thompson (2008): "Predicting excess stock returns out of sample: Can anything beat the historical average?" *Review of Financial Studies*, 21, 1509–1531.

Chen, X. and E. Ghysels (2011): "News—good or bad—and its impact on volatility predictions over multiple horizons," *Review of Financial Studies*, 24, 46–81.

Christensen, K., M. Siggaard, and B. Veliyev (2021): "A machine learning approach to volatility forecasting," *Working paper.*

Corsi, F. (2009): "A simple approximate long-memory model of realized volatility," *Journal of Financial Econometrics*, 7, 174–196.

Corsi, F. and R. Renò (2012): "Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling," *Journal of Business and Economic Statistics*, 30, 368–380.

Diebold, F. X. and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13, 253—-263.

Friedberg, R., J. Tibshirani, S. Athey, and S. Wager (2020): "Local linear forests," *Journal of Computational and Graphical Statistics*, 30, 503–517.

Ghysels, E., A. Plazzi, R. Valkanov, A. R. Serrano, and A. Dossani (2019): "Direct versus iterated multi-period volatility forecasts: Why MIDAS is king," *Annual Review of Financial Economics*, 11, 173–195.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2005): "There is a risk-return trade-off after all," *Journal of Financial Economics*, 76, 509–548.

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

Goulet Coulombe, P. (2021): "The macroeconomy as a random forest," *Available at SSRN: 3633110.*

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 33, 2223–2273.

Hansen, P. R. and A. Lunde (2014): "Estimating the persistence and the autocorrelation function of a time series that is measured with error," *Econometric Theory*, 30, 60–93.

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

Kostrov, A. and A. Tetereva (2019): "Forecasting realized correlations: a MIDAS approach," *Available at SSRN: 3346492.*

Liu, L. Y., A. J. Patton, and K. Sheppard (2015): "Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes," *Journal of Econometrics*, 187, 293–311.

Müller, U. A., M. M. Dacorogna, R. D. Davé, O. V. Pictet, R. B. Olsen, and J. R. Ward (1993): "Fractals and intrinsic time: A challenge to econometricians," *Unpublished manuscript, Olsen & Associates, Zürich*, 130.

Patton, A. J. (2011): "Volatility forecast comparison using imperfect volatility proxies," *Journal of Econometrics*, 160, 246–256.

Patton, A. J. and K. Sheppard (2015): "Good volatility, bad volatility: Signed jumps and the persistence of volatility," *Review of Economics and Statistics*, 97, 683–697.

Pesaran, M. H. and A. Pick (2011): "Forecast combination across estimation windows," *Journal of Business and Economic Statistics*, 29, 307–318.