Improving realized volatility forecasts using news flow

Marcelo Fernandes
São Paulo School of Economics

Murilo Pereira
Quantique

December 2022

Abstract

Economic news may contain valuable information to predict future movements in financial market prices. In this work, we explore the relative importance of news flow to forecast realized volatility. We build text-based indicators using major newspapers in Brazil. Then, we incorporate these indicators into volatility models, controlling for key empirical features, such as asymmetries and discontinuities. Our main results show that the inclusion of news-based variables significantly improve forecasting accuracy. The gains are concentrated in the most liquid stocks and in forecasting horizons above one day.

1 Introduction

Volatility forecasting is paramount for investment decisions, risk management and portfolio allocation. A key feature of volatility is that it varies according to the information arrival in the market. Several studies show that the flow of information, measured by macroeconomic news or some instrument for firm-specific news, is associated with changes in financial market prices (e.g. Andersen et al. (2007) and Barndorff-Nielsen and Shephard (2006)). Other key empirical features such as heavy tails, asymmetric behavior and long memory are crucial for the better understanding of volatility dynamics as well as other higher-order moments.

This work investigate the relative importance of accounting for news flow to forecast realized volatility, controlling for asymmetries and discontinuities in financial asset prices. To deal with the high persistence in volatility, we adopt a Heterogeneous Autoregressive (HAR) specification, as in Corsi (2009). To handle asymmetric effects in price movements, we employ Corsi and Renò (2012) approach. To cope with discontinuities, we compute preaveraged jump-robust estimators based on the ideas of Podolskij and Vetter (2009a). Then, we follow Andersen et al. (2007) and Corsi and Renò (2012) by including jump components in the HAR-type specification. The measures of information flow consider news articles from major newspapers in Brazil. We reconstruct Baker et al. (2016) Economic Policy Uncertainty (EPU) index for Brazil using a broader selection of newspapers and build news-based indicators for the arrival volume of firm-specific news. Finally, to deal with model dimensionality issues, we employ a penalized regression method.

There is extensive research on volatility estimators based on parametric methods, such as GARCH and stochastic volatility models. However, important features like quick responses to short-term shocks, heavy tails and leverage effects, are not captured by these models, as discussed by Corsi (2009). The access to high-frequency data allows researchers to estimate volatility using intraday returns. Andersen et al. (2003)) show that simple models based on realized volatility provide more accurate forecasts than standard parametric models. The HAR-type specification is able to capture the long memory of volatility and the empirical evidence show that it produces good out-of-sample forecasts. In this work, the realized volatility estimators rely on 1-minute intraday returns for the five most liquid stocks and the main Exchange Traded Fund (ETF) traded on the Brazilian stock exchange. Andersen and Bollerslev (1998) argue that the high frequency data increase the accuracy of volatility estimators. However, as the sampling frequency increases, the measurement error due to microstructure noise induces bias. To deal with market microstructure noise, we follow the preaveraging approach, introduced by Jacod et al. (2009).

The empirical evidence shows that many price process can be partitioned into a continuous component and a jump component. Jumps have important implications for derivatives pricing and parameter estimation in some volatility models (see Johannes (2004) and Andersen et al. (2002)). Corsi and Renò (2012) analyse the relationship between jumps and leverage effects, they argue that leverage effects are induced by jumps. In addition, the dynamic of the jump component may vary according to specific market features. Recent studies show that jumps in emerging markets are more severe and present higher intensity than in developed markets (e.g. Chan et al. (2014)). Hence, to evaluate the potential additional information in news-based indicators relative to standard numerical predictors, it's relevant to control for discontinuities, particularly in emerging markets.

Our results show that the inclusion of news-based indicators provide substantial gains relative to the standard HAR model in terms of out-of-sample forecasting accuracy. The improvements in forecasting performance are concentrated in the most liquid stocks and in forecasting horizons of five, ten and twenty-two days ahead. For one day ahead forecasts, the data is less informative and volatility persistence stands out. In this case, the Model Confidence Set (MCS) indicates that several specifications perform equally well, there is large number of models in the confidence set. In addition, our results for the variable selection method point that both firm-specific news and the new version of the EPU index are relevant predictors. We also find evidence that accounting for differential responses to negative returns and signed jumps matter.

We next briefly discuss some related studies. The HAR model has an important role for modeling and forecasting realized volatility, it has been widely applied in the literature. Corsi and Renò (2012) expand the standard model to include jumps and leverage effects, while

Bollerslev et al. (2016) consider dynamic coefficients and microstructure noise. Fernandes et al. (2014) rely on HAR-type models to analyse statistical properties of the VIX index. For Brazil, Wink Junior and Pereira (2011) compare the standard HAR and the MIDAS model using high-frequency data. Recently, a growing number of studies have analyzed the relationship between text data and asset prices. Antweiler and Frank (2004) is one of the first to study the relationship between sentiment analysis and stock prices, the authors apply naive bayes and support vector machines to predict market volatility and returns. Bybee et al. (2020) consider topic models, based on business news, to measure the state of the economy. They use news-attention indicators as inputs in time-series models and show that news data have additional information relative to numerical predictors. Rahimikia and Poon (2021) build a large database from news articles and limit order book data to forecast volatility from NASDAQ stocks. Ke et al. (2021) present a supervised learning approach that extract information from news articles and generate signals to predict asset returns. Several studies explore the relation between central bank communication and asset prices. Hansen and McMahon (2016) analyse the effects of central bank communication on market and real economic variables. Ehrmann and Talmi (2020) find that large changes in central banks statements generate higher market volatility, while similar statements lead to less volatility. Gentzkow et al. (2019) discuss the main features of text data, present an overview of the main statistical methods and show some applications in economics.

This work is divided in the following way. The next section presents the realized variance estimators. Section 3 introduces and describes the data. Section 4 presents the forecasting models and Section 5 discuss the results. Finally, Section 6 presents the conclusion.

2 Realized Variance Measures

The main objective of this section is to present the theoretical framework for the realized variance measures. We start with the realized variance estimator. Then, we present the preaveraged estimator, which is robust to microstructure noise. For both measures, we also present the computation of higher order moments, since it's useful in our forecasting applications. Finally, we decompose the preaveraged estimator in two parts, a continuous component and a discontinuous component.

2.1 Realized Variance

Assume that the log price process, $\tilde{p}_t$, follows a univariate continuous time diffusion

$$d\tilde{p}_t = \mu_t dt + \sigma_t dW_t \tag{2.1}$$

where μ_t is the mean process with finite variation, σ_t is the instantaneous volatility process and W_t is a standard Brownian motion. The realizations of $\tilde{p}_t$ consider intraday data for one day. The latent variable of interest is the integrated variance (IV_t), which is a measure of ex post volatility. The IV_t is defined as

$$\mathrm{IV}_t = \int_{t-1}^{t} \sigma_s^2 ds. \tag{2.2}$$

As shown in Merton (1980) and Andersen et al. (2003), it's possible to estimate the latent volatility over a given period using the sum of n intraday squared returns. Given specific assumptions, for prices sampled at sufficiently small intervals, it's possible to construct an estimator arbitrarily close to the integrated variance. The realized variance estimator converges in probability to the integrated variance when n tends to infinity. The estimator is defined such that

$$\mathrm{RV}_t = \sum_{i=1}^{n} |r_{t+i/n}|^2 \tag{2.3}$$

where $r_{t+i/n} = \tilde{p}_{t+i/n} - \tilde{p}_{t+(i-1)/n}$ and for n sufficiently large we have

$$\mathrm{RV}_t \xrightarrow{p} \mathrm{IV}_t$$
$$\text{as } n \to \infty.$$

Barndorff-Nielsen and Shephard (2002) derived the asymptotic distribution for this estimator, which is defined as

$$\frac{\sqrt{n}}{\sqrt{2\mathrm{IQ}_t}} (\mathrm{RV}_t - \mathrm{IV}_t) \xrightarrow{d} \mathcal{N}(0,1) \tag{2.4}$$

where $\mathrm{IQ}_t = \int_{t-1}^{t} \sigma_s^4 ds$ is the integrated quarticity. The authors show that a consistent estimator for the integrated quarticity is the realized quarticity (RQ_t),

$$\mathrm{RQ}_t = \frac{n}{3} \sum_{i=1}^{n} |r_{t+i/n}|^4. \tag{2.5}$$

.

2.2 Microstructure Noise

In the presence of market microstructure, the true price $\tilde{p}_t$ is contaminated by noise. Market microstructure noise results from market frictions, such bid-ask spread, asynchronous trading and discrete sampling. Assuming an additive noise, we can only observe p_t such that

$$p_t = \tilde{p}_t + u_t, \quad t \geq 0. \tag{2.6}$$

We observe n equally-spaced time points of interval $1/n$, indexed by $i = 1, ..., n$. Calculating returns we have that

$$r_{t+i/n} = \tilde{r}_{t+i/n} + u_{t+i/n} - u_{t+(i-1)/n} = \tilde{r}_{t+i/n} + v_{t+i/n} \tag{2.7}$$

where $\tilde{r}_{t+i/n} = \tilde{p}_{t+i/n} - \tilde{p}_{t+(i-1)/n}$. Assuming that u_t is a white noise, the microstructure noise induces a MA(1) structure and the realized volatility is a biased estimator for the integrated variance. Substituting the observed price in equation (2.3) we have

$$\mathrm{RV}_t = \sum_{i=1}^{n} \left(\tilde{r}_{t+i/n}\right)^2 + 2\sum_{i=1}^{n} \tilde{r}_{t+i/n} v_{t+i/n} + \sum_{i=1}^{n} v_{t+i/n}^2,$$

assuming that the noise is centered, i.i.d and independent of the true price process, then

$$\mathbb{E}\left(\mathrm{RV}_t \mid \tilde{r}\right) = \mathrm{RV}_t + 2n\mathbb{E}\left(u_{t+i/n}^2\right). \tag{2.8}$$

2.3 Preaveraged Estimator

To deal with microstructure noise, sparse sampling has been proposed (see Bandi and Russell (2008)). However, this method discard a large amount of information, since it only decreases the sampling frequency. Hence, several alternative approaches have been developed. Zhang et al. (2005) consider the subsampling method and Barndorff-Nielsen et al. (2008) propose the realized kernel estimators. Jacod et al. (2009) propose the preaveraging approach and Hautsch and Podolskij (2013) analyse empirical features and extend the theory related to the preaveraged estimator.

In this work, we rely on the preaveraging approach. The estimator is based on local moving averages and is denoted as preaveraged estimator. The objective of these moving averages is to reduce the influence of the microstructure noise. The main idea is to average an integer k of observed intraday returns such that the variance is reduced by a factor of $1/k$. According to Jacod et al. (2009), the realized variance estimator based on these preaveraged returns will be close to the latent process of interest.

As in Hautsch and Podolskij (2013), we will assume that the noise, u_t, conditional on the efficient price $\tilde{p} = \left(\tilde{p}_s\right)_{s\geq 0}$ is such that

$$\mathbb{E}\left(u_t \mid \tilde{p}\right) = 0 \quad \text{and} \quad \mathbb{E}\left(u_t u_s \mid \tilde{p}\right) = 0 \text{ for } t \neq s \tag{2.9}$$

and the conditional variance of the noise is

$$\omega_t^2 = \mathbb{E}\left(u_t^2 \mid \tilde{p}\right). \tag{2.10}$$

The model assumptions regarding the noise allows time-varying variance and dependence between $\tilde{p}_t$ and u_t. Given a sequence of integers k_n satisfying

$$k_n n^{-1/2} = \theta + o\left(n^{-1/4}\right) \tag{2.11}$$

where $\theta > 0$, the estimator is given by

$$V(p,2)_t = \sum_{i=0}^{n-k} \left| \sum_{j=1}^{k} g\left(\frac{j}{k}\right) \left(p_{t+(i+j)/n} - p_{t+(i+j-1)/n}\right) \right|^2, \tag{2.12}$$

for a nonzero real-valued function $g : [0,1] \to \mathbb{R}$. The function g must be continuous, piecewise continuously differentiable such that g' is piecewise Lipschitz, $g(0) = g(1) = 0$ and $\int_0^1 g^2(s)\mathrm{d}s < \infty$. We consider the following real numbers associate to g to construct a consistent estimator for the integrated variance:

$$\begin{aligned} \psi_1 &= \int_0^1 \left(g'(s)\right)^2 ds, \quad \psi_2 = \int_0^1 (g(s))^2 ds \\ \phi_1(s) &= \int_s^1 g'(u)g'(u-s)du, \quad \phi_2(s) = \int_s^1 g(u)g(u-s)du \\ & s \in [0,1], \\ \Phi_{ij} &= \int_0^1 \phi_i(s)\phi_j(s)ds, \quad i,j = 1,2. \end{aligned}$$

an example of g is $g(u) = \min\{u, 1-u\}$, in this case the constants are

$$\psi_1 = 1, \quad \psi_2 = \frac{1}{12}, \quad \Phi_{11} = \frac{1}{6}, \quad \Phi_{12} = \frac{1}{96} \quad \Phi_{22} = \frac{151}{80,640}.$$

Assuming $\mathbb{E}\left(|u_t|^2 \mid \tilde{p}\right)$ is locally bounded, Jacod et al. (2009) show that a consistent estimator for IV_t is

$$\overline{\mathrm{RV}}_t := \frac{1}{\theta\psi_2 n^{1/2}} V(p,2)_t - \frac{\psi_1}{2\theta^2\psi_2 n}\mathrm{RV}_t \xrightarrow{p} \mathrm{IV}_t = \int_0^t \sigma_s^2 ds. \tag{2.13}$$

In addition, assuming $\mathbb{E}\left(|u_t|^8 \mid \tilde{p}\right)$ is locally bounded, Jacod et al. (2009) show that

$$n^{1/4}\left(\overline{\mathrm{RV}}_t - \mathrm{IV}_t\right) \xrightarrow{\mathrm{st}} \mathrm{MN}\left(0, \Gamma_t\right) \tag{2.14}$$

where the convergence is stable in distribution and MN is a mixed normal distribution. The authors also propose a feasible estimator for Γ_t and for the integrated quarticity, $\int_0^t \sigma_s^4 ds$. The preaveraged realized quarticity estimator is defined as

$$\begin{aligned} \overline{\mathrm{RQ}}_t =& \frac{1}{3\theta^2\psi_2^2} \sum_{i=0}^{n-k} \left|\overline{r}_{t+i/n}\right|^4 + \frac{-2\psi_1}{2n\theta^4\psi_2^2} \sum_{i=0}^{n-2k} \left|\overline{r}_{t+i/n}\right|^2 \sum_{j=i+k+1}^{i+2k} \left|r_{t+j/n}\right|^2 + \\ &+ \frac{1}{4n}\left(\frac{\psi_1^2}{\theta^4\psi_2^2}\right) \sum_{i=1}^{n-2} \left|r_{t+i/n}\right|^2 \left|r_{t+(i+2)/n}\right|^2 \xrightarrow{p} \mathrm{IQ}_t \end{aligned} \tag{2.15}$$

where $\overline{r}_{t+i/n} = \sum_{j=1}^{k_n} g\left(\frac{j}{k_n}\right)\left(p_{t+(i+j)/n} - p_{t+(i+j-1)/n}\right)$.

2.4 Jump-robust Preaveraged Estimator

In the presence of discontinuities, the preaveraged estimator defined above does not converge to the integrated variance but to the integrated variance plus a jump component. Following Jacod et al. (2010),

$$\overline{\mathrm{RV}}_t \xrightarrow{p} \int_0^t \sigma_s^2 ds + \sum_{s \leq t} |\Delta J_s|^2 \tag{2.16}$$

where $\Delta J_s = J_s - J_{s-}$.

In a setting free of microstructure noise, a typical procedure to specify the diffusive and the jump component is to calculate the bipower variation, as shown by Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen et al. (2006). In order to obtain a jump-robust estimator in the presence of contaminated prices, Podolskij and Vetter (2009b) combined the concepts of bipower variation and preaveraging to define the following estimator

$$V\left(p,1,1\right)_t \quad = \sum_{i=0}^{n-2k+1} \left|\overline{r}_{t+i/n}\right| \left|\overline{r}_{t+(i+k)/n}\right| \tag{2.17}$$

where $\overline{r}_{t+i/n} = \sum_{j=1}^{k_n} g\left(\frac{j}{k_n}\right)\left(p_{t+(i+j)/n} - p_{t+(i+j-1)/n}\right)$.

Podolskij and Vetter (2009b) show that the asymptotic behavior of $V\left(p,1,1\right)_t$ is

$$(1/n)^{1-\frac{p^+}{4}} V\left(p,1,1\right)_t \xrightarrow{p} m_1^2 \times \int_0^t \left(\theta\psi_2\sigma_s^2 + \frac{1}{\theta}\psi_1\alpha_s^2\right)^{\frac{p^+}{2}} ds \tag{2.18}$$

where $m_1 = \mathbb{E}\left[|N(0,1)|^p\right]$ and $p^+ = 2$. Given $V\left(p,1,1\right)_t$, then we can estimate the continuous component, denoted by $\overline{\mathrm{BP}}$, and the jump component, denoted by $\overline{\mathrm{JC}}$. The estimators are

$$\overline{\mathrm{BP}}_t := \frac{1}{\theta m_1^2 \psi_2 n^{1/2}} V(p,1,1)_t^n - \frac{\psi_1}{2\theta^2 \psi_2 n}\mathrm{RV}_t \xrightarrow{p} \mathrm{IV}_t, \tag{2.19}$$

$$\overline{\mathrm{JC}}_t := \frac{1}{\theta \psi_2 n^{-1/2}}\left(V(p,2)_t^n - m_1^{-2} V(p,1,1)_t^n\right) \xrightarrow{p} \sum_{s \leq t} |\Delta J_s|^2 . \tag{2.20}$$

3 Data and Variables Construction

In this section, we present the high-frequency dataset and the news dataset. Then, we explain the construction of the news-based indicators and present a descriptive analysis.

3.1 High-Frequency Data

Our work consider the Exchange Traded Fund Ibovespa (Bova) and the five most liquid stocks traded on B3, the Brazilian stock exchange. The period of analysis is from January 03, 2016 to November 12, 2019. Two companies are from the real sector, Petrobras (Petr) and Vale do Rio Doce (Vale). The other three are from the banking sector, Banco do Brasil (Bbas), Itaú Unibanco (Itub) and Bradesco (Bbdc).

The data is from *Market Data* and we have a total of 938 trading days during the sample period. The regular trading day starts at 10:00am and ends at 5:00pm. Our empirical analysis is based on high-frequency data and the sampling frequency is 1-minute. The sampling frequency scheme results in a total of 420 tick-price observations in a regular trading day[1].

As discussed by Bandi and Russell (2008), there is a debate in the literature about the optimal sampling frequency when dealing with intraday tick-price observations. Sampling with a higher frequency increases the probability of market microstructure noise. Usual sampling intervals in empirical works are 1-minute, 5-minutes and 15-minutes. For comparative reasons we keep in this range, and since we are adjusting for microstructure noise, we compute our estimators based on a sampling frequency of 1-minute.[2] Since we are working with very liquid assets, there is no problems associated to the lack of transactions. Table 1 describes the financial assets with sector and ticker information. It also presents the daily average traded volume (R$ millions).

Company/Asset	Ticker	Sector	Avg. volume
Ibovespa ETF	BOVA11	-	263.7
Petrobras	PETR4	Basic Materials	1,029.7
Vale	VALE3	Basic Materials	610.1
Banco do Brasil	BBAS3	Financial	563.1
Itaú Unibanco	ITUB4	Financial	430.9
Bradesco	BBDC4	Financial	365.0

Table 1: Financial asset information, provided by B3. Avg. volume is the daily average transaction volume, measured in millions (R$).

Table 2 shows descriptive statistics for the preaveraged realized volatility. In the sample period, the companies Petr, Vale and Bbas have the highest volatility. The returns have high right-skew distributions and high kurtosis for all assets, typical features of financial data, as in Andersen et al. (2003) and Corsi (2009). The stocks that present greater skewness and kurtosis are Petr and Itub.

Figure 1 presents the continuous component and the jump component for Bova and the five stocks. The first two plots consider Bova and stocks from the real sector, while the

[1] We do not consider the pre-market and the after-market, since this would result in tick-price observations with a frequency lower than what we are considering.

[2] Our results are robust to a sampling frequency of 5-minutes.

third and fourth plot present our estimates for stocks from the banking sector. The figure shows two periods of high volatility and jump intensity. First, the turmoil associated with the impeachment of the president Roussef in mid-2016 was responsible for a stress on Brazilian assets. The second period of turbulence is in mid-2018, the crisis related to the lorry drivers' strike. There are also large jumps and a burst in volatility in May 2017, the same period of the tainted-meat scandal in Brazil. However, the persistence effect of this shock is very low, the assets volatility quickly return to previous levels.

	Bova	Petr	Vale	Bbas	Itub	Bbdc
Observations	938	938	938	938	938	938
Mean	0.10	0.44	0.44	0.37	0.22	0.28
Std	0.09	1.27	0.58	0.68	0.32	0.38
Median	0.08	0.24	0.26	0.25	0.17	0.22
25%-quantile	0.05	0.13	0.15	0.16	0.12	0.14
75%-quantile	0.12	0.43	0.49	0.39	0.26	0.33
Skew	5.57	21.5	5.15	14.6	19.4	17.27
Kurtosis	59.0	563.9	43.8	303.5	491.0	408.36

Table 2: Descriptive statistics for the preaveraged realized volatility. All values, except for the number of observations, skew and kurtosis, are multiplied by 10^3.

3.1.1 Jump Dynamics

Table 3 shows descriptive statistics for the jump component. The asset with the highest average jump component is Petr, followed by Bbas. The asset with the lowest average jump component is Bova. For the six financial assets, there are many days with jumps equal to zero or very close to zero and a few days with large jumps, which explain the large standard deviations and the high mean compared to quartiles. The mean jump component relative to the total quadratic variation is between 10% and 14%, close to the results of Hautsch and Podolskij (2013). The row Mean JC/RV shows the results for each asset.

Table 3 also shows that there is evidence of asymmetries in the time-series of the jump component. The last two rows in table 3 show the average jump component on days with negative returns (JC^-) and the average jump component on days with positive returns (JC^+). There is a large difference between JC^+ and JC^- for Petr, Bbas and Bbdc. On days with negative returns, the average jump component is higher than on days with positive returns. In addition, as illustrated in figure 1, jump intensity varies over time. The periods of higher intensity seems to occur at the beginning of the sample, and around July 2018.

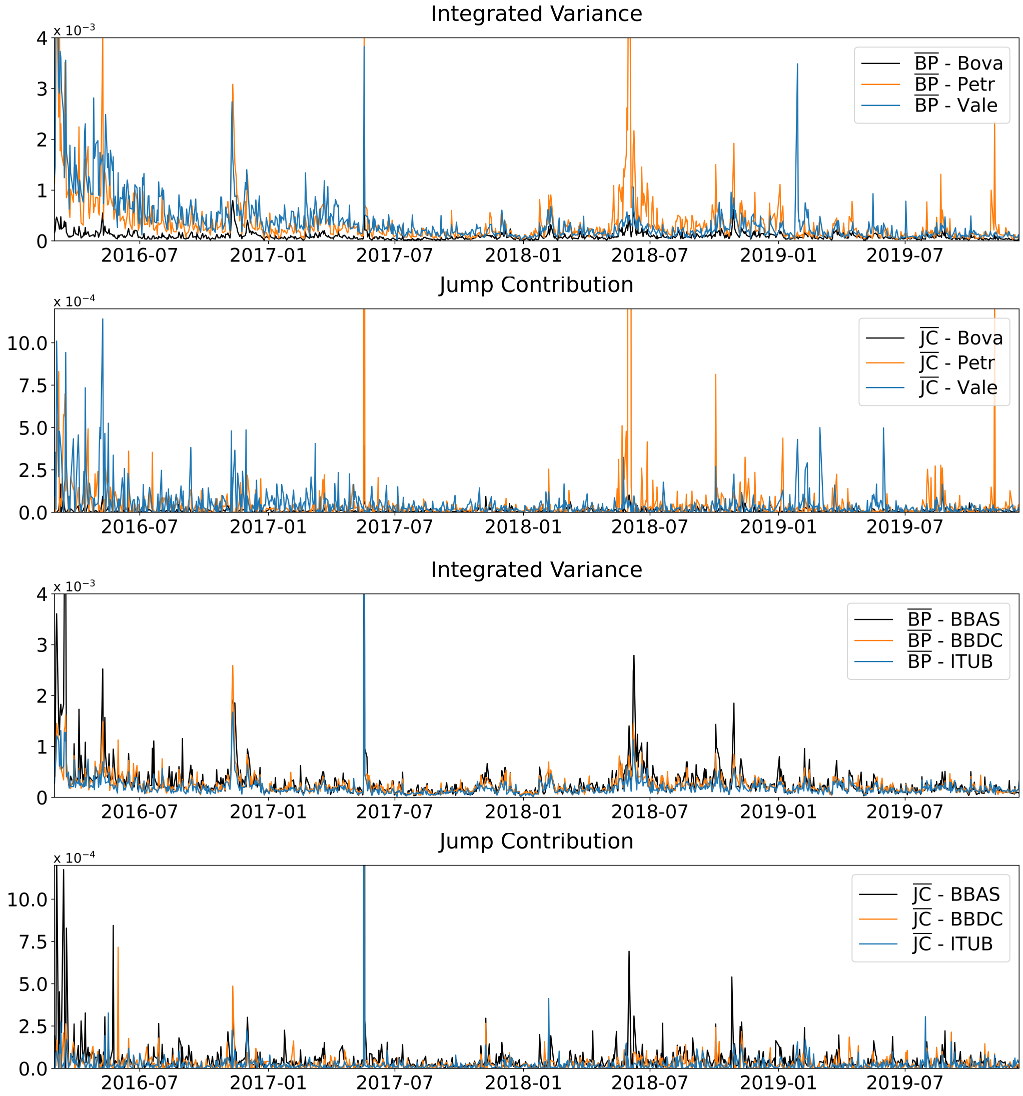

Figure 1: Integrated Variance and Jump Contribution for Bova and stocks from the real sector on the first and second plot. Integrated Variance and Jump Contribution for stocks from the banking sector on the third and fourth plot.

3.1.2 Cross-correlations

We next show evidence of leverage effects and a short-lived correlation between jumps and realized volatility. Figure 2 presents the lagged cross-correlation function between the preaveraged realized variance ($\overline{\text{RV}}_t$) and four variables, the continuous component ($\overline{\text{BP}}_t$), the jump component ($\overline{\text{JC}}_t$), positive and negative returns. The lags are in the explanatory variables. These lagged cross-correlation plots motivates the forecasting models that we present in Section 5.

	Bova	Petr	Vale	Bbas	Itub	Bbdc
Observations	938	938	938	938	938	938
Mean	0.10	0.76	0.49	0.58	0.25	0.33
Std	0.19	8.85	0.97	2.81	0.90	1.51
Median%	0.06	0.18	0.21	0.29	0.13	0.16
25%-quartile	0.00	0.03	0.05	0.11	0.02	0.01
75%-quartile	0.13	0.44	0.51	0.57	0.29	0.36
Mean JC/RV	0.10	0.10	0.11	0.14	0.10	0.10
Mean JC^{+}	0.09	0.44	0.47	0.49	0.25	0.29
Mean JC^{-}	0.12	1.12	0.52	0.69	0.26	0.37

Table 3: Descriptive statistics for the jump component. All values except for the number of observations and Mean JC/RV are multiplied by 10^4.

A well-know feature of volatility is the slow decaying autocorrelation function. As expected, the volatility persistence is mainly associated with the continuous part, figure 2 illustrate this relationship for the six financial assets. Lagged values of the jump component and the contemporaneous jump component are correlated with the realized variance estimator. However, the plots in figure 2 show that this effect is short-lived for all assets, except for Vale. For this asset, even considering five lags of the jump component, the correlation with volatility still remains close to 0.4.

The figure also shows a similar short-lived correlation for negative returns and a small correlation between volatility and positive returns. These results are different from Audrino and Hu (2016), who find evidence of slow decaying cross-correlation between volatility and negative returns.

3.2 News Data

The unstructured dataset includes news articles collected from three major newspapers in Brazil. They are Valor Econômico (Valor), Folha de São Paulo (Folha) and Estadão. We use the same news dataset as in Martins and Medeiros (2021). The news articles are from the sections related to Politics, Economics, Markets and International Affairs. The sample period is the same of the high-frequency data, from January 03, 2016 to November 12, 2019. Table 4 reports the annual number of news articles for each newspaper and their daily averages. The total number of news aggregated over newspapers is very stable over the years, with an overall aggregated daily average of 286 news articles.

Next, we present the news-based indicators. First, we show how we compute the EPU index and present the time-series for Brazil. Then, we explain and present the indicators for firm-specific news.

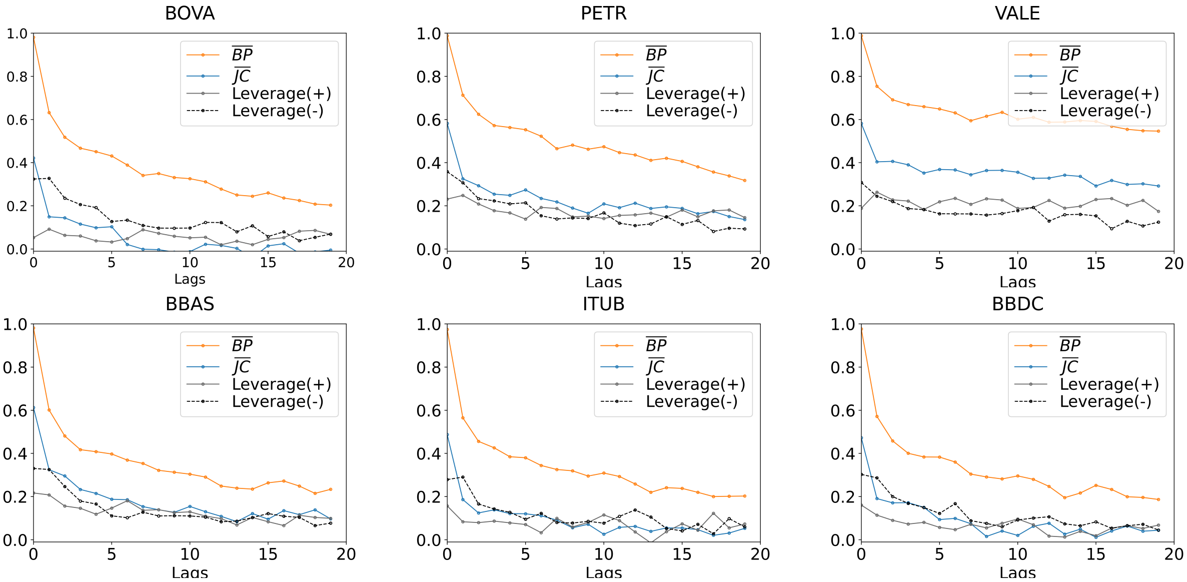

Figure 2: Lagged cross-correlation function between the preaveraged realized variance ($\overline{\text{RV}}_t$) and the continuous component ($\overline{\text{BP}}_t$), the jump component ($\overline{\text{JC}}_t$), positive returns and negative returns. We consider lagged values only for predictor variables.

Newspaper	2016	2017	2018	2019	Total	Daily average
Valor	40,031	46,195	47,001	43,899	177,126	128
Folha	18,526	18,479	25,179	15,786	77,970	57
Estadão	32,399	41,125	35,316	31,203	140,043	101
Total	90,956	105,799	107,496	90,888	395,139	286

Table 4: Annual number of news articles for each newspaper and the daily average

3.2.1 Economic Policy Uncertainty

The EPU index, developed by Baker et al. (2016), measures economic policy uncertainty based on newspaper coverage frequency. The authors show that the index spikes in relevant events, such as presidential elections, debt disputes and wars. In addition, they argue that policy uncertainty has large effects on stock price volatility, investment rates, and employment growth.

Baker et al. (2016) produce daily EPU index for the U.S using a large unstructured dataset. For Brazil, the index is computed based on monthly frequency, considering news articles only from Folha. In this work, we reconstruct Baker et al. (2016) EPU index for Brazil using a broader selection of newspapers. In order to add this index in volatility models, we compute the indicator in a daily frequency. As described earlier, we consider news articles from Folha, Estadão and Valor.

We follow the methodology in Baker et al. (2016) to compute the EPU index for

Brazil. In each day, we count the number of articles containing the terms "incerto" or "incerteza", "econômico" or "economia", and one or more of the following policy-relevant terms: "regulação", "déficit", "orçamento", "imposto", "banco central", "alvorada", "planalto", "congresso", "senado", "câmara dos deputados", "legislação", "lei", "tarifa". The first four terms are equivalent to "uncertain", "uncertainty", "economic" and "economy". While the policy-relevant terms can be translated to "regulation", "deficit", "budget", "tax", "central bank", "alvorada", "planalto", "congress", "senate", "chamber of deputies", "legislation", "law" , "tariff". Since the volume of news articles varies according to the newspaper and time period, we scale the raw EPU counts by the number of all articles in the same newspaper and day. Then, we standardize the daily newspaper-level series to unit standard deviation in our sample period and take the average across newspapers. Finally, we rescale the resulting series to a mean of 100 in our sample period. Figure 3 shows the EPU index for Brazil, we highlight some important events during the sample period.

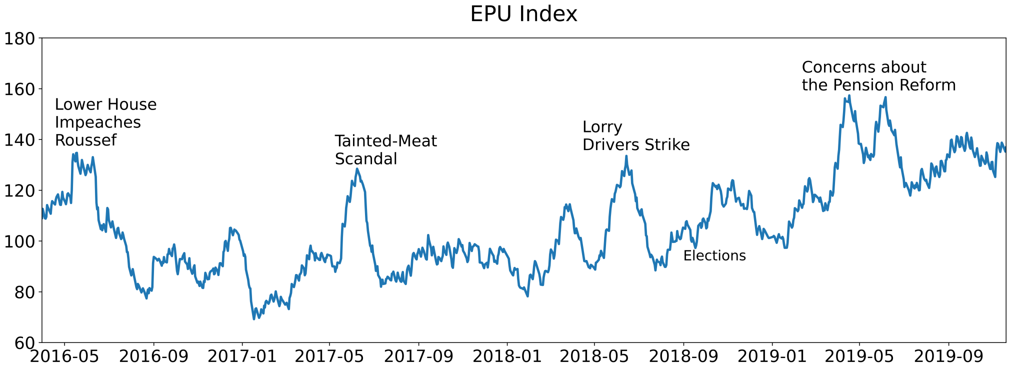

Figure 3: EPU index based on moving average of 22 days.

3.2.2 Firm-specific News

To measure the rate of information arrival that is associated to each company, we consider the volume of firm-specific news. Similar to the methodology followed in Baker et al. (2016), our approach is based on a counting procedure. We select and count news that mention specific keywords in a period of time. In our setting, this firm-specific news indicator is denoted by N_t.

Our database is not tagged with information such as the specific subject of the news or the company being mentioned in each article. To ensure that we are accounting for finance related news, we only select news articles that contain, at least once, the words: "lucro", "prejuízo", "receita" and "despesa". These words are equivalent to words "profit", "loss", "income" and "expense". In order to attribute the news to a particular firm, the name of the firm or the trading ticker must be mentioned, at least once, in the article.

	Petr	Vale	Bbas	Itub	Bbdc
Overnight	1,478	601	876	944	911
Trading hours	1,159	756	587	700	699
Total	2,637	1,357	1,463	1,644	1,610

Table 5: Number of firm-specific news during overnight period and trading hours.

In addition, to better capture the relationship between news and market movements, we split the variable N_t according to news that occur in trading hours and overnight. Table 5 shows descriptive statistics for firm-specific news. The company with the largest number of news in the sample period is Petr, while the other companies have a relatively similar number of news, with mean value of 1,518 articles. Vale is the firm with the lowest value during the sample period, totaling 1,357 articles.

4 Forecasting Models

In this section we introduce the forecasting models. Our work is based on the Heterogeneous Autoregressive (HAR) specification, first proposed by Corsi (2009), and it's extensions.

4.1 HAR Models

The HAR model is inspired on the Heterogeneous Market Hypothesis, the main idea is that agents with heterogeneous time horizons generate different types of volatility components.

Corsi (2009) considers three components, with time horizons of one day, one week and one month. The author proposes an economic interpretation based on market agents. The first component is associated with short-term traders with daily trading frequency. The second, represent medium-term investors who rebalance their positions weekly. Finally, the last component is associated with long-term market agents, such as institutional investors, with time horizon of one month or more.

The HAR model is an additive cascade with asymmetric propagation of volatility between long and short time horizons. Following Corsi (2009), and using the preaveraged estimator from Section 3, we consider the model with three components,

$$\ln \overline{\mathrm{RV}}_{t+h,h} = c + \beta_1 \ln \overline{\mathrm{RV}}_{t,1} + \beta_5 \ln \overline{\mathrm{RV}}_{t,5} + \beta_{22} \ln \overline{\mathrm{RV}}_{t,22} + \varepsilon_{t+h}. \tag{4.1}$$

where the aggregations are normalized sums of one-period realized volatilities, $\overline{\mathrm{RV}}_{t,5}$ and $\overline{\mathrm{RV}}_{t,22}$ are

$$\overline{\mathrm{RV}}_{t,5} = \frac{1}{5}\sum_{h=0}^{4} \overline{\mathrm{RV}}_{t-4,1} \quad \text{and} \quad \overline{\mathrm{RV}}_{t,22} = \frac{1}{22}\sum_{h=0}^{21} \overline{\mathrm{RV}}_{t-21,1}. \tag{4.2}$$

For forecasting horizons longer than one day, the model forecast the average of daily realized volatility for the next h days. The aggregation period is indicated in the lower subscript and for each forecasting horizon, h, we estimate a different model.

Several extension for the HAR model have been proposed in literature. Bollerslev et al. (2016) explore the asymptotic distribution theory for high-frequency realized volatility and introduce the HARQ model using the concept of realized quarticity. The authors consider a dynamic coefficient on the first component and allows it to vary over time according to the estimation error. The HARQ is defined as

$$\ln \overline{\mathrm{RV}}_{t+h,h} = c + (\beta_1 + \beta_{1Q}\overline{\mathrm{RQ}}_t^{1/2}) \ln \overline{\mathrm{RV}}_{t,1} + \beta_5 \ln \overline{\mathrm{RV}}_{t,5} + \beta_{22} \ln \overline{\mathrm{RV}}_{t,22} + \varepsilon_{t+h} \tag{4.3}$$

where $\overline{\mathrm{RQ}}$ is demeaned and therefore β_1 is the average autoregressive coefficient, directly compared to the one from the HAR model. In the class of HAR models, the degree of attenuation bias caused by the microstructure noise depends on the magnitude of the error. The persistence of the volatility process will be lower if the variance of the error is large. The main idea of the HARQ specification is to decrease the impact of the first component when the estimation error is large and increase when the estimation error is small.

In an empirical analysis, Bollerslev et al. (2016) show that the HARQ offer significant improvements to forecast realized volatility comparing with the standard HAR model. The authors also argue that even though alternative estimators, such as the preaveraged and kernel estimators, are more efficient than the standard realized volatility, the HARQ model still provides large forecasting gains relative to the HAR model. In this work, we compute the HARQ model for comparative reasons.

4.1.1 Jumps, Leverage Effects and News

Corsi and Renò (2012) include jumps and leverage effects in the HAR model. Leverage effects can be understood as the correlation between lagged negative returns and volatility. The empirical evidence shows that volatility tends to increase more after a negative shock in asset prices than after a positive shock of the same magnitude (e.g. Bollerslev et al. (2006)). Corsi and Renò (2012) include lagged negative returns at different frequencies as predictors and denote the model as LHAR-CJ. In our setting, it is defined as

$$\begin{aligned} \ln \overline{\mathrm{RV}}_{t+h,h} = {} & c + \beta_1 \ln \overline{\mathrm{BP}}_{t,1} + \beta_5 \ln \overline{\mathrm{BP}}_{t,5} + \beta_{22} \ln \overline{\mathrm{BP}}_{t,22} + \\ & \alpha_1 \ln(1 + \overline{\mathrm{JC}}_{t,1}) + \alpha_5 \ln(1 + \overline{\mathrm{JC}}_{t,5}) + \alpha_{22} \ln(1 + \overline{\mathrm{JC}}_{t,22}) + \\ & \gamma_1 r^-_{t,1} + \gamma_5 r^-_{t,5} + \gamma_{22} r^-_{t,22} + \varepsilon_{t+h} \end{aligned} \tag{4.4}$$

where $r^-_{t,h} = \min(0, r_{t,h})$. In addition, we consider two more specifications based on equation 4.4. First, the nested model HAR-CJ, where $\gamma_1 = \gamma_5 = \gamma_{22} = 0$. This model imposes a restriction on leverage effects and decompose the realized volatility in a continuous component and in a jump component. The second specification consider signed jumps, we include this model based on the evidence of asymmetries in the jump component, as shown in Section

3. The model is denoted LHAR-CSJ and we split the jumps according to the signal of daily returns. In equation 4.4, the time-series for $\overline{\mathrm{JC}}_t$ is substituted by $\overline{\mathrm{JC}}_t^+$ and $\overline{\mathrm{JC}}_t^-$, where

$$\begin{aligned} \overline{\mathrm{JC}}_t^+ &= \overline{\mathrm{JC}}_t \text{ if } r_t > 0 \text{ and } \overline{\mathrm{JC}}_t^+ = 0 \text{ otherwise} \\ \overline{\mathrm{JC}}_t^- &= \overline{\mathrm{JC}}_t \text{ if } r_t < 0 \text{ and } \overline{\mathrm{JC}}_t^- = 0 \text{ otherwise.} \end{aligned} \tag{4.5}$$

Finally, in order to investigate the forecasting ability of the news dataset, we define an augmented version of the HAR model including as additional regressors jumps, leverage effects and news-based indicators. We denote this large model as HARX.

Following the same specification as before, based on heterogeneous markets, we specify daily averages over three different horizons for all predictors. The HARX model is defined as

$$\begin{aligned} \ln \overline{\mathrm{RV}}_{t+h,h} =& c + \beta_1 \ln \overline{\mathrm{BP}}_{t,1} + \beta_5 \ln \overline{\mathrm{BP}}_{t,5} + \beta_{22} \ln \overline{\mathrm{BP}}_{t,22} + \\ & \psi_1 \mathrm{X}_{t,1} + \psi_5 \mathrm{X}_{t,5} + \psi_{22} \mathrm{X}_{t,22} + \varepsilon_{t+h} \end{aligned} \tag{4.6}$$

where X_t are the additional regressors. As before, the variables are aggregated over the most recent five trading days and twenty-two trading days for the weekly and monthly component, respectively.

Including additional predictors, we may be specifying an over-parameterized model, leading to poor out-of-sample forecasting performance. However, as we will see in the next section, this is not the case. Even so, we estimate and present the predictive results for the HARX model using the penalized regression method Adaptive Lasso.

4.2 Adaptive Lasso

Zou (2006) propose the Adaptive Lasso (AdaLasso) regression and shows that the estimator enjoys oracle properties, according to the definition proposed by Fan and Li (2001). The procedure improves the Lasso regression, from Tibshirani (1996), and is based on two stages. Zou (2006) consider adaptive weights to penalize different coefficients in the l_1-penalty.

The Adaptive Lasso estimates are defined by

$$\widehat{\boldsymbol{\beta}}^{\text{AdaLasso}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} w_j \left|\beta_j\right| \tag{4.7}$$

where $w_j = \frac{1}{|\tilde{\beta}_j|^\gamma}$, for $j = 1, \cdots, p$.

The weights are data-dependent and $\tilde{\beta}_j$ is calculated in the first-step estimation. The entire solution path for the adaptive lasso estimates can be computed using LARS algorithm from Efron et al. (2004).

As suggest by Zou (2006), we compute the $\tilde{\beta}_j$ from the best ridge regression fit, since we expect collinearity between our predictors. The hyper-parameters λ and γ are fine-tuned using K-fold cross validation, considering the time-series structure of the data.

5 Results

This section presents and discuss the main findings. We first focus on the out-of-sample forecasting results and then explore the outputs of the penalized regression.

5.1 Forecasting Results

To evaluate the out-of-sample forecasting performance we consider the Mean Absolute Forecasting Error (MAFE) and the Mean Absolute Percentage Error (MAPE), defined as

$$\text{MAFE} = \frac{1}{n_f} \sum_{t=n_e+1}^{n_e+n_f} \left| y_{t+h} - \hat{y}_{t+h|t} \right| \tag{5.1}$$

$$\text{MAPE} = \frac{1}{n_f} \sum_{t=n_e+1}^{n_e+n_f} \left| \frac{y_{t+h} - \hat{y}_{t+h|t}}{y_{t+h}} \right| \tag{5.2}$$

where n_e is the number of training observations and n_f is the number of out-of-sample observations. In our setting, n_e and n_f are equal to 496 and 442, respectively. The out-of-sample period is from 01 March, 2018 to 11 December, 2019 and the forecasting horizons are one, five, ten and twenty-two days ahead. We use expanding windows and the predictive models are re-estimated each day.

We choose the Model Confidence Set (MCS) approach, developed by Hansen et al. (2011), to statistically select the best models. The MCS methodology allows to compare a set of models. Given a loss function, the procedure rank the models and indicate whether one or a group of them perform significantly better. The interpretation is that the models in the confidence set have the same predictive ability and yield the best forecasts for a given level of confidence. Our results are compute based on a block bootstrap procedure, the number of bootstrapped samples is equal to 1000 and the confidence level of the test is 0.10.

Before we explore the results broken down by forecasting horizon, we present the aggregated results over multiple horizons. Table 6 shows the aggregated out-of-sample MAFE with all values relative to the standard HAR. The aggregation includes forecasts for one, five, ten and twenty-two days ahead. We highlight the best models for each financial asset. The results show that the inclusion of additional predictors brings substantial gains comparing to the standard HAR model. The model HARX presents the highest forecasting accuracy for the most liquid stocks, considering the penalized version for Itub. The inclusion of news-based indicators add extra information, which contributes to increase forecasting accuracy, even controlling for jumps and asymmetric effects. Our results are similar to those of other studies, such as Rahimikia and Poon (2021) and Bybee et al. (2020). It provides evidence on the importance of using alternative sources of data to forecast volatility, and more generally, in economic applications. As we will later explore, the improvements in forecasting accuracy varies according to the financial asset and the forecasting horizon. Table 6 also show that for Bova and Bbdc, the model LHAR-CJ has the lowest aggregated out-of-sample MAFE. Since we do not consider company-specific news for Bova, the potential prediction gains of models HARX and HARX (AdaLasso) are smaller. The table indicates that the inclusion of the EPU index in the LHAR-CJ model does not reduce MAFE for Bova. Regarding Bbdc, even considering firm-specific news, the LHAR-CJ model outperform competitors.

Table 6 points that the Random Walk (RW) has the worst performance for all assets. As expected, the simple RW approach fail to address important volatility features, generating poor out-of-sample forecasting results. Surprisingly, the performance of the HARQ model is very close to the standard HAR. As we will show later, this is also true when we split the results for multiple forecasting horizons. There are no significant forecasting gains adding the realized quarticity and allowing dynamic coefficients in the first component of the HAR model. Our results are different from Bollerslev et al. (2016), the authors report large improvements in forecasting accuracy using the HARQ model. One possible reason for this difference could be due to the efficiency of the preaveraged estimator to recover the integrated variance. In this case, there would be a lower possibility to explore the gains in predictive accuracy resulting from the estimation error, as discussed in Bollerslev et al. (2016).

Table 7 shows similar results using the evaluation criterion MAPE. The models with the best forecasting accuracy are the same as in table 6. The main difference is that when we use MAPE, the forecasting errors relative to the standard HAR decreases or remains very close to the ones using MAFE. Therefore, the highlighted models perform well not only when the prediction errors are larger, in absolute terms, but also when the errors are close to zero.

	Bova	Petr	Vale	Bbas	Itub	Bbdc
RW	1.050	1.028	1.217	1.082	1.047	1.060
HARQ	0.997	0.994	0.998	0.999	0.997	0.998
HAR-CJ	0.974	0.955	0.997	0.997	1.037	0.987
LHAR-CJ	0.956	0.957	0.989	0.960	1.034	0.974
LHAR-CSJ	0.958	0.954	0.992	0.956	1.036	0.981
HARX	0.963	0.916	0.956	0.936	1.017	0.994
HARX (AdaLasso)	0.966	0.944	0.969	0.970	0.983	0.979

Table 6: Out-of-sample MAFE aggregated over multiple horizons. All values are relative to the standard HAR model. Highlighted cells show the results for models with the smallest forecasting errors.

We now investigate the results for multiple forecasting horizons. Table 8 presents MAFE results in out-of-sample data for forecasts of one, five, ten, and twenty-two days ahead. We highlight cells that present the MAFE for models in the MCS with confidence level of 0.10. The performance using the evaluation criterion MAPE give close results and we show it on

	Bova	Petr	Vale	Bbas	Itub	Bbdc
RW	1.062	1.100	1.297	1.125	1.057	1.075
HARQ	0.978	0.985	0.992	0.995	0.997	0.998
HAR-CJ	0.939	0.946	1.000	0.996	1.056	0.989
LHAR-CJ	0.913	0.950	0.994	0.952	1.051	0.976
LHAR-CSJ	0.914	0.952	0.992	0.949	1.053	0.982
HARX	0.921	0.915	0.896	0.924	1.051	1.032
HARX (AdaLasso)	0.936	0.948	0.962	0.966	0.996	0.989

Table 7: Out-of-sample MAPE aggregated over multiple horizons. All values are relative to the standard HAR model. Highlighted cells show the results for models with the smallest forecasting errors.

the appendix, table 12.

Table 8 shows that the number of models in the MCS varies according to the forecasting horizon and the financial asset. For the forecasting horizon of one day, the set of best models is larger than for longer horizons. In this case, the data is less informative and volatility persistence dominate. Hence, the confidence sets include several models. There is no evidence of one single best model, the exception is Petr, where the HARX is statistically superior over competitors. The only specification that is not selected for all assets is the Randon Walk.

For horizons longer than one day, table 8 shows a different picture. The news-augmented models, HARX and HARX (AdaLasso), outperform competitors considering the four most liquid stocks, except for Itub in twenty-two days ahead. The results show that only a single model is in the MCS for forecasting horizons of five, ten and twenty-two days considering the financial assets Petr, Vale, Bbas and Itub. Comparing with the standard HAR, the table shows substantial forecasting gains for some assets. There is a reduction in MAFE of around 11% for Petr over five day horizon and 12% for Vale over twenty-two day horizon. Comparing with the model LHAR-CJ, which takes into account asymmetries and leverage effects, the forecasting gains are smaller but still significant.

The model LHAR-CJ is in the MCS for all forecasting horizons for Bova, the same happens to Bbdc, except for forecasts of twenty-two days ahead. Our results does not show significant prediction gains of including news-based indicators for these two financial assets. In addition, the performance of both assets is poorer for the forecasting horizon of twenty-two days ahead relative to other horizons.

We next analyse the cumulative absolute errors to further explore the relative importance of news flow to forecast volatility. The main objective is to analyse the behavior of the errors in each time period. Figure 4 presents the cumulative absolute error difference between models LHAR-CJ and HARX for the six financial assets. The only difference between both

	Bova	Petr	Vale	Bbas	Itub	Bbdc
h=1						
RW	1.040	1.047	1.156	1.055	1.025	1.041
HARQ	0.988	0.975	0.992	0.994	0.995	0.991
HAR-CJ	0.983	0.947	0.995	0.996	0.992	0.980
LHAR-CJ	0.960	0.933	1.005	0.949	0.977	0.978
LHAR-CSJ	0.973	0.920	1.013	0.944	0.978	0.989
HARX	0.972	0.913	1.024	0.944	0.965	0.988
HARX (AdaLasso)	0.966	0.946	1.002	0.951	0.967	0.977
h=5						
RW	1.071	1.029	1.239	1.114	1.072	1.081
HARQ	0.994	0.997	1.003	1.000	1.002	0.998
HAR-CJ	0.955	0.947	0.997	0.992	1.017	0.976
LHAR-CJ	0.928	0.925	0.995	0.940	1.010	0.945
LHAR-CSJ	0.927	0.924	0.997	0.936	1.009	0.950
HARX	0.942	0.890	0.972	0.927	0.987	0.958
HARX (AdaLasso)	0.947	0.943	0.986	0.965	0.958	0.961
h=10						
RW	1.062	1.014	1.286	1.088	1.070	1.067
HARQ	1.004	1.005	0.998	1.003	0.993	1.002
HAR-CJ	0.957	0.945	0.998	0.991	1.025	0.982
LHAR-CJ	0.949	0.957	0.990	0.961	1.030	0.957
LHAR-CSJ	0.945	0.957	0.988	0.956	1.034	0.960
HARX	0.954	0.906	0.937	0.913	1.017	0.977
HARX (AdaLasso)	0.974	0.941	0.959	0.969	0.971	0.983
h=22						
RW	1.028	1.018	1.202	1.079	1.034	1.059
HARQ	1.005	1.001	0.999	0.999	1.001	1.003
HAR-CJ	1.003	0.981	0.996	1.012	1.158	1.016
LHAR-CJ	0.989	1.020	0.964	0.999	1.179	1.018
LHAR-CSJ	0.985	1.023	0.964	0.995	1.183	1.029
HARX	0.983	0.957	0.877	0.956	1.156	1.066
HARX (AdaLasso)	0.978	0.945	0.922	1.005	1.061	0.999

Table 8: Out-of-sample MAFE for multiple forecasting horizons. All values are relative to the standard HAR model. Highlighted cells represent the MAFE for models included in the MCS with confidence level of 0.10.

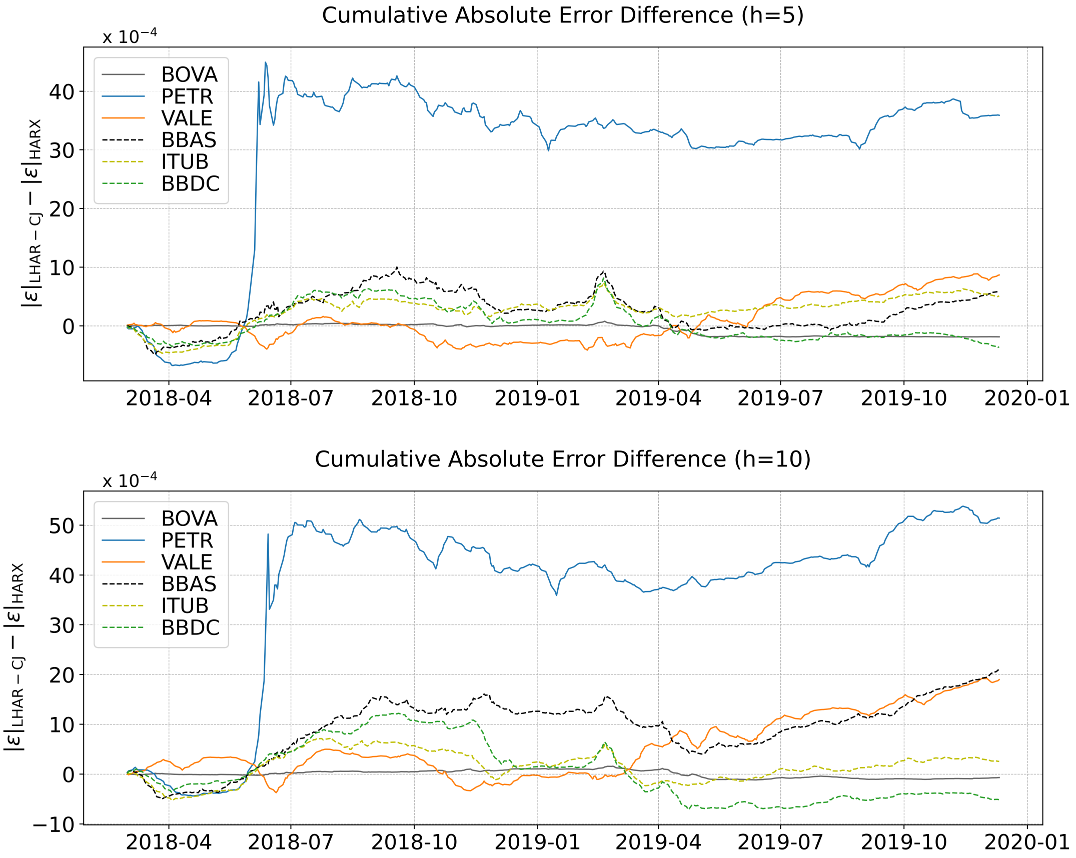

Figure 4: Cumulative absolute error difference between the LHAR-CJ and the HARX model. The upper plot consider forecasting horizon of five days ahead, while the lower plot consider ten days ahead.

models is the inclusion of news-based indicators. When we consider the entire sample period, the figure shows that the HARX outperform the LHAR-CJ for Petr, Vale, Bbas and Itub. Interestingly, figure 4 shows a joint behavior of the cumulative absolute errors difference for all assets, in particular for the ones from the banking sector and Vale. After and during the lorry drivers' strike (May, 2018), the gap between models LHAR-CJ and HARX increases. The figure exhibit a bounce in this difference for Petr, the most liquid stock in this work and the one with the largest number of news. The model HARX shows superior performance for Bbdc from April, 2018 to October, 2018. After that, with the exception to some specific periods, the LHAR-CJ model outperforms. The behavior of the errors difference for Bova is stable over the entire sample period and the inclusion of news flow does not result in an increase in predictive accuracy, complementing the results of the tables 6, 7 and 8.

We report the results considering different settings in the appendix. Table 15 presents the output for out-of-sample MAFE for training sets with rolling-windows of two years. The most striking difference with respect to expanding windows is the underperformance of news-augmented models for Vale. Table 16 shows out-of-sample results for multiple forecasting horizons using Mean Squared Forecasting Error (MSFE).

5.2 Variable Selection

In this subsection, we further investigate the results of the Adaptive Lasso regression. The main goal is to shed light on the relevant predictors of the HARX model. Tables 9 and 10 present the proportion of time that each predictor's coefficient assumes a nonzero value out of the total out-of-sample period. Table 9 shows the results for the forecasting horizon of one day ahead, while table 10 considers five days ahead. The results for longer horizons are on the appendix.

We do not report the values for the continuous components since the coefficients are always nonzero, for all forecasting horizons and financial assets. This result confirms the well-known importance of incorporating persistence effects to model and forecast volatility. Both tables show that jump components are relevant in some periods of time. Although the inclusion of signed jumps does not significantly increase forecasting accuracy, as discussed in Section 5.1, the variable selection method suggests that jumps signed with positive returns and jumps signed with negative returns have different relative importance. Tables 9 and 10 show that the coefficients for the predictor variable $\overline{\mathrm{JC}}^{+}_{t,1}$ are rarely nonzero, while the coefficients for $\overline{\mathrm{JC}}^{-}_{t,1}$ have greater relevance regardless of the forecasting horizon and financial asset. The only exception is Petr for the forecasting horizon of five days ahead. Similar to Sheppard and Patton (2011), our results suggest that signed jumps play an important role in empirical applications. In addition, the results for the penalized regression indicate a strong relationship between lagged negative returns and volatility. Tables 9 and 10 show that the fraction of time in which coefficients associated to the predictor $r^{-}_{t,1}$ are nonzero is equal to one for all financial assets, while for $r^{-}_{t,5}$ and $r^{-}_{t,22}$ the results are mixed.

The proportion of time that the model selects the EPU index varies substantially depending on the forecast horizon, the number of lags and the financial asset. Considering the forecasting horizon of one day ahead, table 9 suggests that the short-term component of the EPU index is more relevant than the other components. In this case, the coefficients for $\mathrm{EPU}_{t,1}$ assume nonzero values most of the time, while the coefficients of $\mathrm{EPU}_{t,5}$ take values equal to zero for all financial assets. This result is different when we consider the forecasting horizon of five days ahead, as we can observe in table 10.

The results of the variable selection method also show that the three components based on firm-specific news can have predictive power. Tables 9 and 10 indicate that when the forecasting horizon is equal to one day ahead, the proportions related to variable $\mathrm{N}_{t,1}$ assume value equal to one for Vale, while the ones related to $\mathrm{N}_{t,22}$ take values equal to one for Petr, Bbas and Itub. Interestingly, for Petr and Itub, the model selects the three components most of the time.

	Bova	Petr	Vale	Bbas	Itub	Bbdc
$\overline{\text{JC}}^{+}_{t,1}$	0.00	0.00	0.23	0.00	0.00	0.00
$\overline{\text{JC}}^{-}_{t,1}$	0.48	0.49	1.00	0.00	1.00	1.00
$\overline{\text{JC}}_{t,5}$	0.00	0.94	0.43	0.79	0.00	0.08
$\overline{\text{JC}}_{t,22}$	0.26	0.95	0.00	0.00	0.00	0.28
$r^{-}_{t,1}$	1.00	1.00	1.00	1.00	1.00	1.00
$r^{-}_{t,5}$	0.35	0.86	0.25	1.00	0.00	0.00
$r^{-}_{t,22}$	0.33	0.00	0.00	0.81	0.86	0.86
$\text{EPU}_{t,1}$	1.00	0.51	0.51	0.67	0.92	0.85
$\text{EPU}_{t,5}$	0.00	0.00	0.00	0.00	0.00	0.00
$\text{EPU}_{t,22}$	0.00	0.09	0.49	0.22	0.00	0.56
$\text{N}_{t,1}$	-	0.20	1.00	0.00	0.28	0.79
$\text{N}_{t,5}$	-	0.97	0.39	0.00	1.00	1.00
$\text{N}_{t,22}$	-	1.00	0.54	1.00	1.00	0.56

Table 9: Proportion that each predictor is selected out of the total out-of-sample observations in the Adaptive Lasso regression. The forecasting horizon is one day ahead.

	Bova	Petr	Vale	Bbas	Itub	Bbdc
$\overline{\text{JC}}^{+}_{t,1}$	0.00	0.61	0.45	0.05	0.00	0.00
$\overline{\text{JC}}^{-}_{t,1}$	1.00	0.59	1.00	0.82	1.00	1.00
$\overline{\text{JC}}_{t,5}$	0.00	0.13	0.00	0.48	0.00	0.00
$\overline{\text{JC}}_{t,22}$	0.00	0.00	0.00	0.00	0.00	0.00
$r^{-}_{t,1}$	1.00	1.00	1.00	1.00	1.00	1.00
$r^{-}_{t,5}$	0.00	0.86	0.00	0.27	0.00	0.00
$r^{-}_{t,22}$	0.19	0.23	0.64	0.54	0.83	0.84
$\text{EPU}_{t,1}$	0.00	0.00	0.00	0.00	0.00	0.00
$\text{EPU}_{t,5}$	0.00	0.36	0.00	0.00	0.38	0.19
$\text{EPU}_{t,22}$	0.27	0.30	0.81	0.50	0.48	0.67
$\text{N}_{t,1}$	-	1.00	0.00	0.00	0.84	0.86
$\text{N}_{t,5}$	-	0.98	0.54	0.09	0.86	0.59
$\text{N}_{t,22}$	-	1.00	0.55	1.00	1.00	1.00

Table 10: Proportion that each predictor is selected out of the total out-of-sample observations in the Adaptive Lasso regression. The forecasting horizon is five days ahead.

6 Conclusion

This dissertation rely on the theoretical framework for high frequency data and realized volatility to show that news flow can be used to improve volatility forecasts. We build indicators from major newspapers in Brazil to evaluate news-augmented models in out-of-sample data. The indicators capture broadly aspects of economic policy uncertainty and firm-specific news. Using extensions of the HAR-type model from Corsi (2009), we compare several specifications, considering key empirical features, to evaluate the forecasting gains of using an alternative source of data. Finally, we apply a penalized regression method to perform variable selection and analyse the predictor's relative importance.

The results of this work provide new evidence on the use of news flow to forecast realized volatility for Brazilian financial assets and indicate that it's possible to extract useful information from newspapers. We find that even controlling for discontinuities and asymmetric effects, the inclusion of news-based indicators bring significant forecasting gains, especially for more liquid stocks and longer forecasting horizons. An in-depth analysis, through the results of the variable selection method, highlights the relevance of our version of the EPU index and firm-specific news indicators as predictors variables. In addition, we also confirm some empirical results for the Brazilian stock market, such as the negative correlation between returns and volatility and the importance of signed jumps.

There are several possible extensions for this work. An interesting starting point would be to build alternative news variables based on statistical learning and machine learning methods, as described in Gentzkow et al. (2019). Some of them include topic models (Latent Dirichlet Allocation), Support Vector Machines and Neural Networks. These methods may also contribute to improve the methodology to select news. Alternative approaches may also consider multivariate setting (Vector HAR) and nonlinear specifications.

References

T. Andersen and T. Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905, 1998.

T. Andersen, T. Bollerslev, F. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003.

T. G. Andersen, L. Benzoni, J. Lund, D. Bates, M. Brenner, S. Das, B. Eraker, R. Gallant, and R. Green. An empirical investigation of continuous-time equity return models. *Journal of Finance*, pages 1239–1284, 2002.

T. G. Andersen, T. Bollerslev, and F. X. Diebold. Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. *The Review of Economics and Statistics*, 89(4):701–720, 2007.

W. Antweiler and M. Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.

F. Audrino and Y. Hu. Volatility forecasting: Downside risk, jumps and leverage effect. *Econometrics*, 4(1), 2016. ISSN 2225-1146.

S. R. Baker, N. Bloom, and S. J. Davis. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

F. M. Bandi and J. R. Russell. Microstructure noise, realized variance, and optimal sampling. *The Review of Economic Studies*, 75(2):339–369, 2008.

O. Barndorff-Nielsen and N. Shephard. Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4(1):1–30, 2006.

O. E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(2):253–280, 2002. ISSN 13697412, 14679868.

O. E. Barndorff-Nielsen and N. Shephard. Power and Bipower Variation with Stochastic Volatility and Jumps. *Journal of Financial Econometrics*, 2(1):1–37, 2004.

O. E. Barndorff-Nielsen, N. Shephard, and M. Winkel. Limit theorems for multipower variation in the presence of jumps. *Stochastic Processes and their Applications*, 116(5): 796–806, 2006. ISSN 0304-4149.

O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard. Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise. *Econometrica*, 76(6):1481–1536, 2008.

T. Bollerslev, J. Litvinova, and G. Tauchen. Leverage and Volatility Feedback Effects in High-Frequency Data. *Journal of Financial Econometrics*, 4(3):353–384, 2006.

T. Bollerslev, A. Patton, and R. Quaedvlieg. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1):1–18, 2016.

L. Bybee, B. T. Kelly, A. Manela, and D. Xiu. The structure of economic news. Working Paper 26648, National Bureau of Economic Research, 2020.

K. F. Chan, J. G. Powell, and S. Treepongkaruna. Currency jumps and crises: Do developed and emerging market currencies jump together? *Pacific-Basin Finance Journal*, 30(C): 132–157, 2014.

F. Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

F. Corsi and R. Renò. Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business Economic Statistics*, 30(3):368–380, 2012. ISSN 07350015.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

M. Ehrmann and J. Talmi. Starting from a blank page? semantic similarity in central bank communication and market volatility. *Journal of Monetary Economics*, 111:48–62, 2020. ISSN 0304-3932.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

M. Fernandes, M. C. Medeiros, and M. Scharth. Modeling and predicting the cboe market volatility index. *Journal of Banking Finance*, 40:1–10, 2014. ISSN 0378-4266.

M. Gentzkow, B. Kelly, and M. Taddy. Text as data. *Journal of Economic Literature*, 57(3): 535–74, 2019.

P. R. Hansen, A. Lunde, and J. M. Nason. The model confidence set. *Econometrica*, 79(2): 453–497, 2011. ISSN 00129682, 14680262.

S. Hansen and M. McMahon. Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99(S1):S114–S133, 2016.

N. Hautsch and M. Podolskij. Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: Theory, implementation, and empirical evidence. *Journal of Business & Economic Statistics*, 31(2):165–183, 2013.

J. Jacod, Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter. Microstructure noise in the continuous case: the pre-averaging approach. 2009.

J. Jacod, M. Podolskij, and M. Vetter. Limit theorems for moving averages of discretized processes plus noise. *The Annals of Statistics*, 38(3):1478 – 1545, 2010.

M. Johannes. The statistical and economic role of jumps in continuous-time interest rate models. *The Journal of Finance*, 59(1):227–260, 2004.

Z. T. Ke, B. T. Kelly, and D. Xiu. Predicting Returns With Text Data. Working papers, 2021.

Martins and Medeiros. Nowcasting with unstructured data. 2021.

R. C. Merton. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361, 1980. ISSN 0304-405X.

M. Podolskij and M. Vetter. Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli*, 15(3), 2009a. ISSN 1350-7265. doi: 10.3150/08-bej167.

M. Podolskij and M. Vetter. Bipower-type estimation in a noisy diffusion setting. *Stochastic Processes and their Applications*, 119(9):2803–2831, 2009b. ISSN 0304-4149.

E. Rahimikia and S.-H. Poon. Big Data Approach to Realised Volatility Forecasting Using HAR Model Augmented With Limit Order Book and News. Working papers, 2021.

K. Sheppard and A. Patton. Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97, 2011. doi: 10.1162/REST_a_00503.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996. ISSN 00359246.

M. V. Wink Junior and P. L. V. Pereira. Modeling and Forecasting of Realized Volatility: Evidence from Brazil. *Brazilian Review of Econometrics*, 31(2), 2011.

L. Zhang, P. A. Mykland, and Y. Ait-Sahalia. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100:1394–1411, 2005.

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

Appendix A - In-sample Results

	Bova	Petr	Vale	Bbas	Itub	Bbdc
RW	1.057	1.118	1.276	1.048	1.065	1.065
HARQ	0.989	0.999	0.993	0.999	0.982	0.997
HAR-CJ	0.966	0.961	0.996	0.953	0.907	0.950
LHAR-CJ	0.957	0.941	0.985	0.939	0.895	0.938
LHAR-CJ$^+$	0.952	0.932	0.982	0.933	0.895	0.935
HARX	0.951	0.923	0.935	0.933	0.890	0.915
HARX (AdaLasso)	0.961	0.938	0.936	0.952	0.900	0.918

Table 11: In-sample Mean Absolute Errors (MAE) aggregated over multiple horizons. All values are relative to the standard HAR model. We highlight models with lowest MAE.

Appendix B - Out-of-sample MAPE Results

	Bova	Petr	Vale	Bbas	Itub	Bbdc
			h=1			
RW	1.064	1.092	1.236	1.096	1.035	1.056
HARQ	0.964	0.965	0.985	0.983	0.989	0.991
HAR-CJ	0.953	0.915	0.992	1.000	1.017	0.984
LHAR-CJ	0.921	0.900	0.994	0.942	1.003	0.984
LHAR-CSJ	0.933	0.900	0.989	0.937	1.005	0.990
HARX	0.935	0.910	0.948	0.939	1.000	1.029
HARX (AdaLasso)	0.948	0.952	1.014	0.954	0.980	0.984
			h=5			
RW	1.080	1.118	1.318	1.163	1.081	1.095
HARQ	0.976	0.988	0.997	0.999	1.005	1.000
HAR-CJ	0.916	0.944	1.002	0.994	1.040	0.979
LHAR-CJ	0.881	0.936	1.001	0.944	1.028	0.954
LHAR-CSJ	0.880	0.940	1.001	0.940	1.027	0.959
HARX	0.904	0.922	0.924	0.944	1.039	1.026
HARX (AdaLasso)	0.928	0.957	0.986	0.981	0.986	0.986
			h=10			
RW	1.060	1.104	1.369	1.135	1.078	1.081
HARQ	0.990	0.997	0.994	1.003	0.993	1.002
HAR-CJ	0.922	0.957	1.007	0.986	1.042	0.986
LHAR-CJ	0.908	0.964	1.004	0.950	1.040	0.958
LHAR-CSJ	0.903	0.966	1.000	0.948	1.044	0.960
HARX	0.914	0.909	0.882	0.902	1.058	1.018
HARX (AdaLasso)	0.923	0.951	0.948	0.973	0.989	0.984
			h=22			
RW	1.040	1.086	1.281	1.125	1.045	1.076
HARQ	0.990	0.994	0.994	1.001	1.001	1.004
HAR-CJ	0.962	0.975	0.999	1.001	1.175	1.013
LHAR-CJ	0.943	1.011	0.978	0.982	1.193	1.010
LHAR-CSJ	0.938	1.014	0.978	0.980	1.198	1.020
HARX	0.926	0.920	0.812	0.902	1.167	1.063
HARX (AdaLasso)	0.939	0.932	0.886	0.961	1.050	1.008

Table 12: Out-of-sample MAPE for multiple forecasting horizons. All values are relative to the standard HAR. Highlighted cells represent the MAPE for models in the MCS.

Appendix C - Variable Selection

	Bova	Petr	Vale	Bbas	Itub	Bbdc
$\overline{\text{JC}}^{+}_{t,1}$	0.85	0.59	0.97	0.38	0.00	0.34
$\overline{\text{JC}}^{-}_{t,1}$	0.89	0.61	1.00	0.98	0.94	1.00
$\overline{\text{JC}}_{t,5}$	0.00	0.00	0.00	0.14	0.00	0.00
$\overline{\text{JC}}_{t,22}$	0.00	0.00	0.00	0.11	0.00	0.00
$r^{-}_{t,1}$	0.89	1.00	0.00	0.98	0.94	1.00
$r^{-}_{t,5}$	0.00	0.85	0.00	0.01	0.00	0.00
$r^{-}_{t,22}$	0.12	0.77	0.17	0.78	0.27	0.83
$\text{EPU}_{t,1}$	0.00	0.00	0.00	0.00	0.00	0.00
$\text{EPU}_{t,5}$	0.02	0.55	0.23	0.54	0.00	0.74
$\text{EPU}_{t,22}$	0.27	0.27	1.00	0.51	0.46	0.66
$\text{N}_{t,1}$	-	0.86	0.03	0.00	0.94	0.95
$\text{N}_{t,5}$	-	1.00	0.64	0.26	0.88	0.99
$\text{N}_{t,22}$	-	1.00	0.38	0.98	0.95	1.00

Table 13: Proportion each predictor is selected out of the total out-of-sample observations in the Adaptive Lasso regression. Forecasting horizon equal to ten days ahead.

	Bova	Petr	Vale	Bbas	Itub	Bbdc
$\overline{\text{JC}}^{+}_{t,1}$	0.36	0.51	1.00	0.00	0.00	0.38
$\overline{\text{JC}}^{-}_{t,1}$	0.80	0.73	1.00	0.76	0.80	0.65
$\overline{\text{JC}}_{t,5}$	0.00	0.03	0.00	0.00	0.00	0.19
$\overline{\text{JC}}_{t,22}$	0.00	0.02	0.00	0.00	0.00	0.00
$r^{-}_{t,1}$	0.80	0.01	0.00	0.76	0.65	0.01
$r^{-}_{t,5}$	0.00	0.02	0.00	0.00	0.00	0.00
$r^{-}_{t,22}$	0.04	1.00	0.67	0.71	0.18	0.17
$\text{EPU}_{t,1}$	0.00	0.10	0.00	0.00	0.00	0.00
$\text{EPU}_{t,5}$	0.26	0.35	0.20	0.49	0.43	0.39
$\text{EPU}_{t,22}$	0.10	0.47	1.00	0.60	0.32	0.58
$\text{N}_{t,1}$	-	0.92	0.05	0.00	0.03	0.07
$\text{N}_{t,5}$	-	1.00	0.55	0.76	0.80	0.65
$\text{N}_{t,22}$	-	1.00	0.16	0.24	0.60	0.00

Table 14: Proportion each predictor is selected out of the total out-of-sample observations in the Adaptive Lasso regression. Forecasting horizon equal to twenty-two days ahead.

Appendix D - Out-of-sample Results MAFE (Rolling-Window)

	Bova	Petr	Vale	Bbas	Itub	Bbdc
			h=1			
RW	1.041	1.037	1.059	1.054	1.025	1.042
HARQ	0.991	0.969	0.993	0.996	0.995	0.998
HAR-CJ	0.988	0.954	1.008	0.997	0.983	0.982
LHAR-CJ	0.967	0.941	1.079	0.945	0.979	0.988
LHAR-CSJ	0.981	0.933	1.091	0.940	0.986	0.995
HARX	0.977	0.937	1.151	0.951	0.974	0.994
HARX (AdaLasso)	0.961	0.941	1.078	0.957	0.959	0.983
			h=5			
RW	1.071	1.033	1.075	1.105	1.053	1.071
HARQ	1.002	0.992	1.001	1.002	1.001	1.005
HAR-CJ	0.974	0.973	1.038	0.987	0.977	0.962
LHAR-CJ	0.939	0.924	1.052	0.936	0.967	0.923
LHAR-CSJ	0.943	0.927	1.052	0.929	0.968	0.928
HARX	0.953	0.902	1.066	0.954	0.962	0.947
HARX (AdaLasso)	0.936	0.921	1.050	0.978	0.943	0.944
			h=10			
RW	1.074	1.021	1.092	1.087	1.054	1.058
HARQ	1.013	1.004	1.001	1.009	1.003	1.016
HAR-CJ	0.971	0.980	1.053	0.987	0.987	0.962
LHAR-CJ	0.971	0.974	1.060	0.962	0.990	0.920
LHAR-CSJ	0.972	0.976	1.060	0.950	0.996	0.924
HARX	0.982	0.929	1.094	0.966	0.991	0.946
HARX (AdaLasso)	0.935	0.945	1.041	0.964	0.954	0.952
			h=22			
RW	1.041	1.034	1.042	1.082	1.031	1.044
HARQ	1.018	0.997	0.999	1.007	1.012	1.034
HAR-CJ	1.007	1.021	1.036	0.985	1.056	0.959
LHAR-CJ	1.019	1.056	1.041	0.980	1.087	0.968
LHAR-CSJ	1.016	1.060	1.043	0.975	1.091	0.975
HARX	0.960	1.026	1.090	0.974	1.110	1.002
HARX (AdaLasso)	0.957	0.989	1.018	0.968	1.025	0.963

Table 15: Out-of-sample MAFE for training sets considering rolling windows of two years. Highlighted cells represent the MAFE for models included in the MCS.

Appendix E - Out-of-sample Results MSFE

	Bova	Petr	Vale	Bbas	Itub	Bbdc
h=1						
RW	1.036	1.009	1.054	1.024	1.021	1.027
HARQ	0.990	0.989	0.996	0.995	1.002	0.995
HAR-CJ	0.986	0.984	0.998	1.003	0.987	0.978
LHAR-CJ	0.965	0.982	1.005	0.964	0.964	0.954
LHAR-CSJ	0.976	0.981	1.019	0.954	0.965	0.961
HARX	0.976	0.972	1.051	0.952	0.951	0.950
HARX (AdaLasso)	0.974	0.984	0.998	0.982	0.967	0.967
h=5						
RW	1.082	1.006	1.214	1.065	1.058	1.067
HARQ	1.013	1.000	1.000	1.001	1.001	0.999
HAR-CJ	0.972	0.981	0.998	0.997	0.985	0.942
LHAR-CJ	0.964	0.926	1.000	0.971	0.969	0.913
LHAR-CSJ	0.964	0.924	1.004	0.955	0.969	0.912
HARX	0.964	0.836	1.006	0.954	0.937	0.901
HARX (AdaLasso)	0.973	0.910	0.993	0.986	0.941	0.934
h=10						
RW	1.066	0.997	1.289	1.052	1.056	1.058
HARQ	1.006	1.013	0.999	1.000	1.006	1.001
HAR-CJ	0.954	0.964	1.000	0.996	1.002	0.927
LHAR-CJ	0.945	0.954	0.991	0.955	1.000	0.905
LHAR-CSJ	0.942	0.956	0.989	0.936	1.002	0.907
HARX	0.946	0.878	0.985	0.924	0.957	0.898
HARX (AdaLasso)	0.957	0.918	0.978	0.977	0.933	0.924
h=22						
RW	1.024	0.994	1.297	1.034	1.024	1.032
HARQ	1.006	1.005	0.997	1.000	1.010	1.004
HAR-CJ	0.999	0.987	1.001	1.007	1.118	0.967
LHAR-CJ	0.999	0.999	0.965	0.982	1.134	0.991
LHAR-CSJ	0.997	1.002	0.965	0.975	1.135	0.999
HARX	0.995	0.927	0.926	0.979	1.109	1.019
HARX (AdaLasso)	0.992	0.926	0.936	0.985	0.987	0.964

Table 16: Out-of-sample MSFE. All values are relative to the standard HAR model. Highlighted cells represent the MAFE for models included in the MCS.