Modeling and Forecasting Realized Volatility: Evidence from Brazil*

Marcos Vinício Wink Junior**
Pedro L. Valls Pereira***

Abstract

Using intraday data for the most actively traded stocks on the São Paulo Stock Market (BOVESPA) index, this study considers two recently developed models from the literature on the estimation and prediction of realized volatility: the Heterogeneous Autoregressive Model of Realized Volatility (HAR-RV), developed by Corsi (2009), and the Mixed Data Sampling model (MIDAS-RV), developed by Ghysels et al. (2004). Using measurements to compare in-sample and out-of-sample forecasts, better results were obtained with the MIDAS-RV model for in-sample forecasts. For out-of-sample forecasts, however, there was no statistically significant difference between the models. We also found evidence that the use of realized volatility induces distributions of standardized returns that are closer to normal.

Keywords: Realized Volatility, HAR, MIDAS, High Frequency Financial Data.

JEL Codes: C22, C53, C58.

*Submitted in July 2012. Revised in January 2013. The authors would like to thank one reviewer and also the participants of the 2011 Brazilian Economics Meeting (Encontro Nacional de Economia – ANPEC). The second author would like to acknowledge the financial support from CNPq.

**Escola de Economia de São Paulo-FGV and Fundação de Economia e Estatística. Rua Itapeva 474 sala 1214, 01332-000, São Paulo, SP. Brasil. E-mail: marcos@fee.tche.br

***Escola de Economia de São Paulo – FGV and CEQEF-FGV. Rua Itapeva 474 sala 1017, 01332-000, São Paulo, SP. Brasil. E-mail: pedro.valls@fgv.br

1. Introduction

Modeling the return volatility of stocks has been a growing area of research in recent years. Models that estimate realized volatility based on intraday data have been the most commonly studied. Hypotheses have emerged for these models with respect to the behavior of agents and the attempt to verify stylized facts with respect to financial series. Returns standardized by the realized volatility of financial assets are believed to have distributions that are close to normal, which is relevant for the modeling of risk and the use of Value at Risk (V@R) models. This study compares models for the estimation of realized volatility of five stocks from the São Paulo Stock Market (BOVESPA). The models that will be considered in this study are the Heterogeneous Autoregressive Model of Realized Volatility (HAR-RV), developed by Corsi (2009), and the Mixed Data Sampling model (MIDAS-RV), developed by Ghysels et al. (2004).

The models' ability to fit and predict data is compared using in-sample and out-of-sample comparison measurements. For in-sample forecasts, we compare the in-sample adjusted R^2 and the mean squared errors (MSE). For out-of-sample comparisons, we use the out-of-sample MSE and a modified Diebold-Mariano test. This study will contribute to the literature on the prediction of realized volatility with high frequency data for emerging countries, similarly to the study by Chung et al. (2008) for the Taiwan market and to the studies by Carvalho et al. (2006) and Sá Mota and Fernandes (2004) for Brazil. In general, these studies compared predictions from realized volatility models with GARCH class models. None of these studies, however, used any form of control for the effects of the microstructure of the market. To control for this possible bias, this study uses an $MA(q)$ model to filter data, as suggested by Hansen et al. (2008). In addition, this study is also the first to apply the MIDAS methodology to Brazilian data.

As asserted by Carvalho et al. (2006), forecasting returns is not an easy task. Modeling the volatility of returns, however, is easier and therefore generates more reliable predictions. Thus, the literature on volatility modeling has expanded in recent years.

Volatility modeling began with the estimation of ARCH-GARCH models and with the use of stochastic volatility models. However, as discussed in Corsi (2009), these models suffer from a "double weakness." First, they do not represent certain specific empirical characteristics of financial data[1] and second, they tend to have nontrivial estimates, particularly when utilizing models of stochastic volatility.[2]

The next phase in the literature on volatility modeling was an attempt to

[1]Most models are able to replicate some of the stylized facts such as: heavy tails, volatility clustering, high persistence, long memory, but only some models replicate the leverage effect and arrival of news in the market.

[2]The exact estimation of the stochastic volatility model is computationally intensive; however, the estimation by quasi maximum likelihood is easy to implement using the Kalman Filter, but the estimates will be biased (see Harvey et al., 1994).

calculate volatility through the use of the squared daily return of a stock. Thus, the estimation of volatility would be given from an observed variable that was a proxy for true volatility, which is a latent variable. Andersen and Bollerslev (1998), however, observed that the estimation of volatility through daily data generates problems due to the large amount of noise associated with these series. Therefore, the authors suggested that intraday data would bring greater precision to the estimation of volatility, which has become known in the literature as the calculation of realized volatility. The methods for estimating realized volatility that will be considered in this study are the HAR-RV and the MIDAS.

The success of the HAR-RV method in estimating realized volatility is based on modeling the behaviors of long-term memory volatility simply and efficiently. This model was inspired by the Heterogeneous Market Hypothesis and also by asymmetry in the propagation of volatility. Corsi (2009) claims that there are different types of agents in heterogeneous markets, that these agents are content with different prices and that these agents may therefore decide to perform transactions at different moments. For the development of the realized volatility estimation model, Corsi (2009) considers only the different time horizons of the transactions as a source of heterogeneity. The evidence from Corsi (2009) indicates that the performance of the out-of-sample HAR-RV model is comparable to that of the ARFIMA model, which is much more complex for estimations and is superior to the autoregressive model, particularly for longer time horizons, such as two weeks.

The literature on time series normally uses models involving samples with identical frequencies. However, the MIDAS model developed by Ghysels et al. (2004) considers different levels of sample frequency, which is preferable to the normal procedure in which data are pre-filtered and potentially useful information may be discarded. As posited by Ghysels et al. (2004), the MIDAS model has the advantage of a frugal estimation with a reduced number of parameters to be estimated. As further demonstrated by Ghysels et al. (2004), the MIDAS regressors also have desirable properties such as efficiency and lack of bias when the sample tends to be more frequent.

The use of high frequency data also has a disadvantage. Andersen et al. (2007) note that obtaining the estimator of the daily realized volatility is biased as a result of the significant volume of noise associated with the market microstructure observed in intraday data. The authors posit that this noise is mainly associated with the fact that the prices observed are not continuous and are quoted on a discrete grid of values. Therefore, as we will discuss below, the observed intraday price is not a single market price at a precise instant in time, but a price with market microstructure noise instead. Andersen et al. (2007) discuss likely sources of market microstructure, positing that the most frequent sources are the differences between the prices for buyers and sellers (bid-ask spread) and different prices among different market participants because of beliefs, information, and the decision to buy or sell. To control for this possible bias, our study uses the data

filtering process suggested by Hansen et al. (2008). In addition, our study is also the first to apply the MIDAS methodology to Brazilian data.

Following this Introduction, Section 2 presents the theoretical framework on realized variance, Section 3 describes the models used, Section 4 provides the descriptive statistics of the data used, Section 5 presents the results for the estimations, and Section 6 concludes.

2. Realized Volatility

According to McAleer and Medeiros (2008) and Andersen et al. (2003), it is assumed that the logarithmic price process p_t of a particular stock in continuous time obeys the following diffusion:

$$dp_t = \sigma_t \; dW_t \quad , t = 1, 2... \tag{1}$$

where p_t is the logarithm of the price over time t, σ_t is the instantaneous volatility, strictly stationary, and W_t is a standard Brownian motion. We assume that the process described in (1) does not contain a drift component.

The sample returns with M observations per period may be calculated using:

$$r_{(M),t} = p_t - p_{t-1/M} = \int_0^{1/M} \sigma_{t-1/(M+\tau)} dW_{t-1/(M+\tau)} \; , t = \frac{1}{M}, \frac{2}{M}, ... \tag{2}$$

Defining expected returns as equal to zero for any time horizon, standardizing the time interval by M intraday observations, and also assuming that W_t and σ_t are independent and conditioning the mathematical expectation in the sample trajectory of volatility $\{\sigma_{t+\tau}\}_{\tau=0}^{h}$ the return variance for h periods may be described as follows:

$$\sigma_{t,h}^2 = \int_0^h \sigma_{t+\tau}^2 d\tau \tag{3}$$

Equation [3] describes the so-called integrated variance (IV), which is a measurement of ex-post volatility. Thus, the volatility for h periods is identical to the integral of past intraday volatilities. The IV, however, is not observed and, because it is the object of interest, it must be estimated. The intraday return in period m and on day t is defined as follows:

$$r_{t,m} = p_{t,m} - p_{t,m-1} \quad \text{for } m = 1, ..., M \quad \text{and } t = 1, ..., n \tag{4}$$

The daily realized variance (RV) may be defined as follows:

$$RV_t = \sum_{m=1}^{M} r_{t,m}^2 \tag{5}$$

Andersen et al. (2003) have demonstrated that, under certain conditions involving a lack of autocorrelation of returns, the realized variance defined in equation (5) is a consistent estimator of integrated variance, $RV_t \xrightarrow{p} IV_t$. Realized Volatility ($RV$) is the square root of the realized variance. Barndorff-Nielsen and Shephard (2002) derived the asymptotic distribution of the integrated variance estimator as follows:

$$\sqrt{n}\frac{1}{\sqrt{2IQ_t}}(RV_t - IV_t) \xrightarrow{d} N(0,1) \tag{6}$$

where IQ_t is the integrated quarticity and is defined by:

$$IQ_t = \int_0^1 \sigma^4(t+\tau)d\tau \tag{7}$$

Under the hypothesis of a lack of correlation of the intraday returns, IQ_t may be consistently estimated by the realized quarticity, defined by:

$$RQ_t = \frac{1}{3}\sum_{i=1}^{n} r_{i,t}^4 \tag{8}$$

3. HAR-RV and MIDAS

The models considered here will be the $HAR-RV$ and $MIDAS$. The $HAR-RV$ model proposed by Corsi (2009) provides a method for estimating volatility using different interval sizes. $\widetilde{\sigma}_t^{(.)}$ is defined as the partial volatility generated by certain market components. The model is proposed as an additive cascade of partial volatilities that follows an autoregressive-type process. Three different components of volatility generated by different time horizons are considered; more specifically, in this example, $\widetilde{\sigma}_t^{(d)}$, refers to one day, $\widetilde{\sigma}_t^{(w)}$, refers to one week and, $\widetilde{\sigma}_t^{(m)}$, refers to one month. Let us assume that the daily return is given by:

$$r_t = \sigma_t^{(d)}\ \varepsilon_t \tag{9}$$

where $\varepsilon_t \sim NI(0,1)$ and $\sigma_t^{(d)}$ is the daily integrated volatility that satisfies $\widetilde{\sigma}_t^{(d)} = \widetilde{\sigma}_t^{(d)}$.

The process of partial volatility $\widetilde{\sigma}_t^{(.)}$ for each time scale is a function of past realized volatility, on an identical time scale, and the expectation of partial volatility for the subsequent period with a longer time frame. Because the longest time scale is monthly, we have the following models:

$$\widetilde{\sigma}_{t+1m}^{(m)} = c^{(m)} + \phi_m RV_t^{(m)} + \widetilde{\omega}_{t+1m}^{(m)} \tag{10}$$

$$\widetilde{\sigma}_{t+1w}^{(w)} = c^{(w)} + \phi^{(w)} RV_t^{(w)} + \gamma^{(w)} E_t[\widetilde{\sigma}_{t+1m}^{(m)}] + \widetilde{\omega}_{t+1w}^{(w)} \tag{11}$$

$$\widetilde{\sigma}_{t+1d}^{(d)} = c^{(d)} + \phi^{(d)} RV_t^{(d)} + \gamma^{(d)} E_t[\widetilde{\sigma}_{t+1w}^{(w)}] + \widetilde{\omega}_{t+1d}^{(d)} \tag{12}$$

where $RV_t^{(m)}$, $RV_t^{(w)}$ and RV_t^d are the monthly, weekly, and daily ex-post realized volatilities and the volatility innovations $\widetilde{\omega}_{t+1m}^{(m)}$, $\widetilde{\omega}_{t+1w}^{(w)}$ and $\widetilde{\omega}_{t+1d}^{(d)}$ are contemporaneously and serially uncorrelated, with a mean of zero and with a truncated distribution in the lower tail that ensures that the partial volatilities are positive.

Through recursive substitutions of partial volatilities, the model may be written as follows, bearing $\widetilde{\sigma}_t^{(d)} = \widetilde{\sigma}_t^{(d)}$ in mind:

$$\sigma_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \widetilde{\omega}_{t+1d}^{(d)} \tag{13}$$

We can write the ex-post partial volatility as follows:

$$\sigma_{t+1d}^{(d)} = RV_{t+1d}^{(d)} + \omega_{t+1d}^{(d)} \tag{14}$$

where $\omega_{t+1d}^{(d)}$ represents the estimation error and the measurement of latent daily volatility.

Therefore, substituting equation (14) into (13), we find:

$$RV_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1d} \tag{15}$$

where $\omega_{t+1d} = \widetilde{\omega}_{t+1d}^{(d)} - \omega_{t+1d}^{(d)}$

Equation (15) therefore describes the realized volatility through a simple autoregressive process. As used in Corsi (2009), $RV_t^{(m)}$ corresponds to the mean monthly volatility, as is true for the week in $RV_t^{(w)}$.

The $HAR-RV$ model of one-day volatility may be easily extended, as shown by Andersen et al. (2007) and by Forsberg and Ghysels (2006) for h time horizons, indexed in t and $RV_{t,t+h}$. We define the realized volatility for multiple periods as follows:

$$RV_{t,t+h} = \frac{1}{h}(RV_{t,t+1} + RV_{t+1,t+2} + ... + RV_{t+h-1,t+h}) \tag{16}$$

where $RV_{t,t+h}$ refers to the increase in the RV from t to $t+h$ periods, with $h = 1, 5, 10, 15$ and 20, indicating, respectively, one day, one week, two weeks, three weeks, and one month. Therefore, the $HAR-RV$ model for multiple periods is given by:

$$RV_{t,t+h} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t,t+h} \tag{17}$$

Corsi (2009) further suggests that the model be estimated in its logarithmic form because of the lognormal distribution of the realized volatility. Therefore, the estimated model in this study is described by equation (18):

$$\ln RV_{t,t+h} = c + \beta^{(d)} \ln RV_t^{(d)} + \beta^{(w)} \ln RV_t^{(w)} + \beta^{(m)} \ln RV_t^{(m)} + \omega_{t,t+h} \tag{18}$$

The MIDAS models with polynomial lag structures, introduced by Ghysels et al. (2007), involve regressors with different frequencies and thus are not autoregressive. In the MIDAS model, the dependent variable Y_t has a fixed frequency (annual, quarterly, monthly, or daily) called the reference interval. Let $X_t^{(m)}$ be a sample of m intervals of time, for example, and we then have, with annual data, that $X_t^{(4)}$ corresponds to the sample with quarterly data. In this regard, a MIDAS regression is

$$Y_t = \alpha + B(L^{1/m}) X_t^{(m)} + \varepsilon_t^{(m)} \tag{19}$$

where $B(L)$ is a finite or an infinite polynomial in the lag operator, generally parameterized as a small set of hyperparameters.[3] $B(L^{1/m}) = \sum_{j=0}^{j\max} B_j L^{j/m}$ is a $j\max$ polynomial (possibly infinite) in the operator $L^{j/m}$ and $L^{j/m} X_t^{(m)} = X_{t-j/m}^{(m)}$. The operator $L^{j/m}$ therefore produces values of $X_t^{(m)}$ lag over j/m periods.

Following Forsberg and Ghysels (2006) and Chung et al. (2008), we developed a particular specification of the MIDAS-RV model based on the beta function for h time horizons.

$$RV_{t,t+h} = \alpha + \phi \sum_{k=0}^{50} b(k, \theta_1, \theta_2) RV_{t-k-1,t-k} + \varepsilon_{t,t+h} \tag{20}$$

$$b(k, \theta_1, \theta_2) = \frac{f(k/50; \theta_1, \theta_2)}{\sum_{j=0}^{50} f(j/50; \theta_1, \theta_2)} \tag{21}$$

$$f(x; \theta_1, \theta_2) = \frac{x^{\theta_1 - 1}(1-x)^{\theta_2 - 1}}{\beta(\theta_1, \theta_2)} \tag{22}$$

[3] This model is similar to the distributed lag model.

where $\beta(\theta_1, \theta_2)$ is the beta function, i.e., $\beta(\theta_1, \theta_2) = \frac{\Gamma(\theta_1)\Gamma(\theta_2)}{\Gamma(\theta_1+\theta_2)}$ and $f(x; \theta_1, \theta_2)$ is the probability density functions of the beta distribution. The beta function defines the 50 weights that determine the memory of the realized volatility process.[4] Parameters θ_1 and θ_2 determine the shape of the weight function. The higher θ_2 is the more rapidly the function declines. The parameter θ_1 defines the initial trajectory of the function; thus, if $\theta_1 > 1$, the function jumps before beginning to decline.[5] Because long-term memory processes generally depend more strongly on more recent observations, what generally occurs is $\theta_1 \simeq 1$ and $\theta_2 > 1$.

4. Database

The study uses the five most liquid stocks traded on the BOVESPA during the period, which are Bradesco (BBDC4), Petrobrás (PETR4), Vale (VALE5), Telemar (TNLP4), and Usiminas (USIM5). The sample includes intraday data for three different time periods, at intervals of 5, 15, and 30 minutes. The period analyzed is from 11/01/2007 to 04/30/2010.

As discussed in McAleer and Medeiros (2008), there is a debate in the literature regarding the selection of the frequency of the intraday data. As the frequency increases (large M), the precision also increases, as does the probability that noise will be associated with a microstructure, such as the lack of negotiation. Andersen et al. (2001) propose 5-minute intervals. Oomen (2002) argues that the optimal interval frequency is 25 minutes. Giot and Laurent (2004) observe an optimal frequency of data every 15 minutes. In the face of this debate, we chose to use data in 5-minute, 15-minute, and 30-minute intervals. Table 1 shows the number of observations for each stock in each intraday period. There are different numbers of observations because we are using the stock price negotiated immediately after the time period.[6] Thus, we are obtaining the price of the first trade after the determined time period, and more liquid stocks will therefore have more observations because they will be more likely to be traded right after the end of the negotiation period.

Figure 1 shows the graph of daily realized volatility using 5-minute time intervals for each stock. An increase can be perceived in volatility in the second half of 2008. This increase occurred because of the subprime crisis. Jumps in volatility can also be observed for all the stocks.[7]

Table 2 presents the descriptive statistics of the daily returns of the five selected stocks. As commonly found in the literature, the returns display an excess of

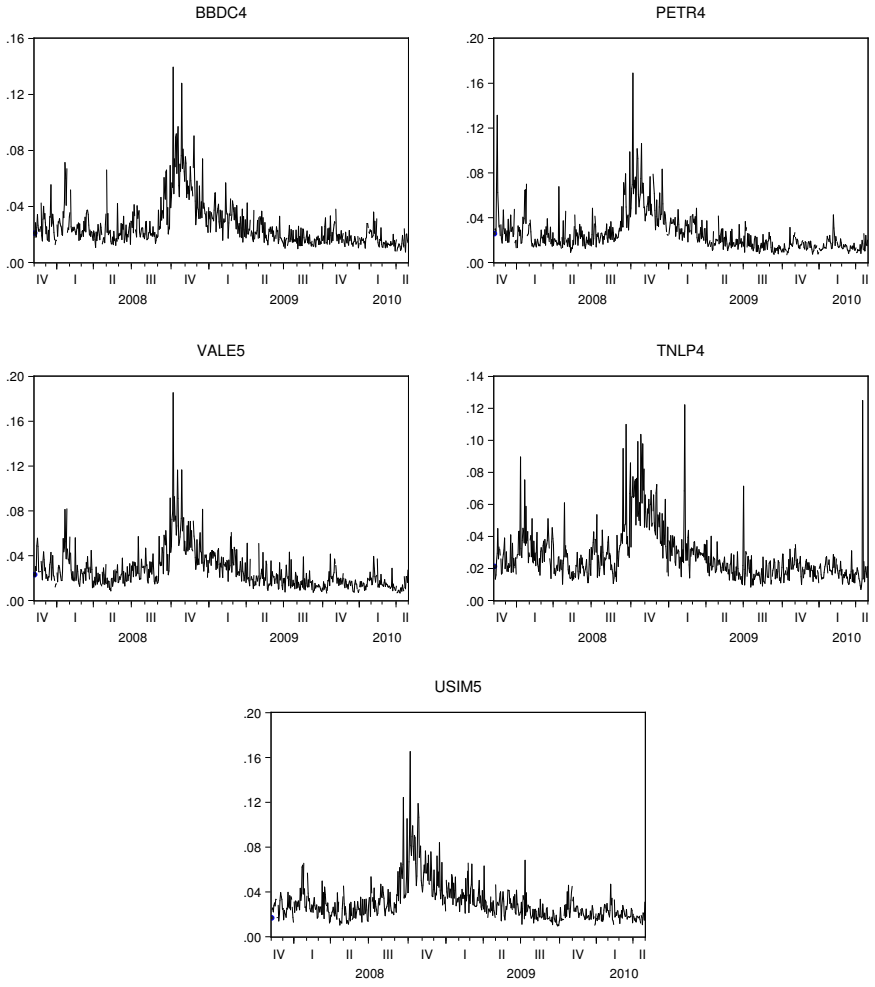
[4]The choice of $k = 50$ has been suggested by Ghysels et al. (2006). According to the authors, the results obtained using $k > 50$ are very similar.

[5]For more details, see Ghysels et al. (2007).

[6]In general, the trade occurs in the second following the end of the negotiation period, but in some cases, the trading may take up to 30 seconds.

[7]Andersen et al. (2007) state that most jumps in volatility occur due to macroeconomic announcements.

Figure 1
Daily realized volatilities



Source: BM&FBOVESPA.
Note: Realized Volatility estimated with a 5-minute period.

Table 1
Number of observations

Stocks	5-min period	15-min period	30-min period
BBCDC4	41.900	15.663	8.330
PETR4	45.858	16.142	8.456
VALE5	45.532	16.138	8.458
TNLP4	32.800	14.407	7.947
USIM5	40.413	15.475	8.280
Source: BM&FBOVESPA.			
Note: The extraction was performed according to the first trade			
after 5, 15, or 30 minutes.			

kurtosis, implying that the returns are heavy-tailed and therefore do not have normal distributions, as observed on the Jarque-Bera test. It is noteworthy that all the stocks displayed negative mean returns, primarily because the sample covers the period of the subprime crisis.

Table 2
Descriptive Statistics of Daily Returns

Stock	Mean	S.D.	Ass	Kurt	Jarque-Bera
BBDC4	−0.00015	0.029	0.678	8.484	0.000
PETR4	−0.00014	0.031	−0.034	6.322	0.000
VALE5	−0.00023	0.032	−0.153	6.284	0.000
TNLP4	−0.00061	0.031	0.470	7.541	0.000
USIM5	−0.00012	0.035	0.008	6.271	0.000
Source: Own calculations.					
Note: The Jarque-Bera value corresponds to the test's p-value.					

To select the most accurate estimator of realized volatility among the different frequencies, we determined the average size of the 95% confidence interval of the estimator, calculated from its variance described in equation (7). The results are presented in Table 3 and suggest that the smallest confidence interval is generated by the 5-minute frequency for all the stocks.[8] Therefore, from this point onward, we will use intraday data with a 5-minute frequency.

Table 4 shows the descriptive statistics of the standardized returns. Daily volatilities were estimated by the following five different methods: GARCH,

[8]If the intraday returns were independent from the time period, the lowest-frequency interval would be the smallest-sized one; however, as they depend on time period, this result may not be reliable. Therefore, the choice based on the lowest mean size seems to be the most appropriate one.

Table 3
Average size of confidence interval of daily realized volatility

Stock	5-min period	15-min period	30-min period
BBDC4	0.00016	0.00045	0.00044
PETR4	0.00016	0.0030	0.00047
VALE5	0.00019	0.00032	0.00049
TNLP4	0.00021	0.00034	0.00046
USIM5	0.00023	0.00039	0.00059
Source: Own calculations.			

EGARCH, EWMA (using the estimated decay parameter), GJR[9] and realized volatility. The order of the models and the conditional distributions of the errors were selected using the Schwarz information criterion.

The results described in Table 4 suggest that the estimation of realized volatility induces distributions in the standardized returns that are closer to a normal distribution in relation to the estimation of daily volatility by other methods. This result is useful for risk management through V@R models. Another important result presented in Table 4 is the superiority of the asymmetric models of volatility (EGARCH and GJR) compared to GARCH models in the generation of standardized returns with normal distributions,[10] with the exception of VALE5 and TNLP4, which did not show improved results.

5. Data Filtration

As discussed above, realized variance is a consistent estimator of integrated variance if the returns are not autocorrelated. Andersen et al. (2001), however, show that the market microstructure generates autocorrelation of the returns. To demonstrate this, let us assume that the observed logarithmic price of a stock $p_{t,i}$ is given by:

$$p_{t,i} = p^*_{t,i} + \varepsilon_{t,i} \tag{23}$$

where $p^*_{t,i}$ is the latent efficient price and $\varepsilon_{t,i}$ is the noise associated with the market microstructure. The return on stock $r_{t,i}$ will be given by:

$$r_{t,i} = p_{t,i} - p_{t,i-1} = r^*_{t,i} + \varepsilon_{t,i} - \varepsilon_{t,i-1} \tag{24}$$

where $r^*_{t,i} = p^*_{t,i} - p^*_{t,i-1}$ is the efficient return.

This shows that the noise of the microstructure generates autocorrelation in the returns and therefore generates a biased estimation of the integrated variance.

[9]GJR model is also known as Threshold GARCH or TGARCH.

[10]The same result has been observed by Carvalho et al. (2006) for Brazilian stock returns.

Table 4
Descriptive Statistics of Standardized Daily Returns

Stock	Mean	S.D.	Ass	Kurt	Jarque-Bera
GARCH					
BBDC4	$0,008$	1.003	0.216	3.824	0.000
PETR4	-0.007	0.998	-0.140	3.351	0.077
VALE5	0.006	1.000	-0.224	4.380	0.000
TNLP4	-0.013	1.010	-0.412	7.781	0.000
USIM5	0.010	1.002	0.078	3.408	0.088
EGARCH					
BBDC4	$-0,002$	1.000	0.096	3.352	0.130
PETR4	-0.005	1.000	-0.130	3.236	0.207
VALE5	0.060	1.000	-0.224	4.380	0.000
TNLP4	-0.020	1.019	-0.602	1.001	0.000
USIM5	0.010	1.003	0.084	4.380	0.093

Table 4(cont): Descriptive Statistics of Standardized Daily Returns

Stock	Mean	S.D.	Ass	Kurt	Jarque-Bera
EWMA					
BBDC4	$0,010$	1.038	0.205	3.792	0.000
PETR4	-0.008	1.036	-0.160	3.527	0.008
VALE5	0.007	1.047	-0.338	4.944	0.000
TNLP4	-0.016	1.069	-0.806	1.145	0.000
USIM5	0.016	1.047	0.120	3.508	0.018
GJR					
BBDC4	$0,000$	1.001	0.103	3.387	0.087
PETR4	-0.007	1.000	-0.123	3.301	0.146
VALE5	0.002	1.000	-0.211	4.067	0.000
TNLP4	-0.019	0.909	-0.407	7.478	0.000
USIM5	0.011	1.001	0.056	3.230	0.437
REALIZED VOLATILITY					
BBDC4	$0,024$	0.949	0.140	2.601	0.049
PETR4	0.050	1.047	-0.015	2.729	0.388
VALE5	0.055	1.040	0.016	2.588	0.114
TNLP4	0.021	0.964	0.189	2.742	0.070
USIM5	0.031	1.028	0.039	2.500	0.038

Source: Own calculations.

Note: The Jarque-Bera value corresponds to the test's p-value.

To correct the bias caused by the market microstructure, we use the daily realized volatility filtered by the moving average method (MA filter), which was used by Ebens (1999) and Andersen et al. (2001). These authors suggested the use of a first-order moving average process to model the intraday returns of the stocks and that the filtering would result from the use of the residuals of this estimation. Hansen et al. (2008), however, showed that the $MA(1)$ filter would be valid only under the hypothesis that the noise associated with the market microstructure is independently and identically distributed (IID); therefore, these researchers proposed a filtering method in which the hypothesis of IID noise is not necessary.

Let us assume that the intraday return of a stock follows an $MA(q)$ process given by $r_{t,i} = \varepsilon_{t,i} - \theta_1\varepsilon_{t,i-1} - ...\theta_q\varepsilon_{t,i-q}$. Hansen et al. (2008) show that, under certain conditions, for the estimation of RV based on the $MA(q)$ process, $RV_{t,MA(q)}$ is given by:

$$RV_{t,MA(q)} = (1 - \widehat{\theta}_1 - ... - \widehat{\theta}_q)^2 \sum_{i=1}^{M} \widehat{\varepsilon}_{t,i}^2 \tag{25}$$

We have the following

$$E[RV_t] = E\left[\sum_{i=1}^{M} r_{t,i}^2\right] = (1 + \theta_1^2 + ... + \theta_q^2)E\left[\sum_{i=1}^{M}\sum_{j=0}^{q} \varepsilon_{t,i-j}^2\right] \tag{26}$$

Because $(1 - \widehat{\theta}_1 - ... - \widehat{\theta}_q)^2 \sum_{i=1}^{M} \widehat{\varepsilon}_{t,i}^2 \xrightarrow{p} IV$ we then have

$$\frac{E[RV_t]}{IV} \simeq \frac{(1 + \theta_1^2 + ... + \theta_q^2)}{(1 - \theta_1 - ... - \theta_q)^2} \tag{27}$$

Equation (27) suggests a bias-corrected estimator for the realized variance, $\widetilde{RV}_{t,MA(q)}$, given by:

$$\widetilde{RV}_{t,MA(q)} = \frac{(1 + \widetilde{\theta}_1^2 + ... + \widetilde{\theta}_q^2)}{(1 - \widetilde{\theta}_1 - ... - \widetilde{\theta}_q)^2} RV_t \tag{28}$$

The estimations obtained in this study were performed from the realized volatility filtered by the MA(q) method described in equation (28), in which the order of the MA was 2, selected under the information criterion

6. Results of the Estimations

In this section, we present the results of the $HAR - RV$ and $MIDAS - RV$ models for the five stocks and for the time horizons $h = 1, 5, 10, 15$ and 20, as described in equations (18) and (20), respectively. For both models, the realized volatility that was used was filtered, as per equation (28), to correct for problems

associated with the microstructure. The estimated coefficients of the HAR-RV models are presented in Table 5, with their respective p-values.

Table 5 shows that the estimated coefficients β^d, β^w and β^m are significant, confirming the hypothesis of the high persistence of volatility. In addition, Table 5 shows that, on average, the estimations of β^d and of β^w decline along the time horizons, as opposed to β^m, which tends to be relatively more significant for long-term volatility, as previously shown by Andersen et al. (2007) and Chung et al. (2008). As in our study, Andersen et al. (2007) did not observe a common pattern for all betas in the different time horizons. It is noteworthy that only the PETR4 stock displayed a drop in the estimations of β^d and β^w and an increase in the estimations of β^m, according to the increase in the forecast horizon for all estimations.

The estimated coefficients for the MIDAS-RV models are presented in Table 6 with their respective p-values.

The results described in Table 6 demonstrate that all estimated coefficients differ significantly from zero. The main parameter of interest is ϕ because it determines whether future realized volatility is a function of past-realized volatilities. This coefficient proved to be significant for all stocks and for all forecast horizons. As discussed above, the parameters θ_1 and θ_2 determine the shape of the weight function.

As expected, θ_1 is close to 1 in all the estimations, indicating that the weight functions decline from the first lag. The parameter θ_2 indicates the velocity of the decline of the weight function. Because the estimated values for θ_2 are relatively high, the weight function of the realized volatility declines rapidly, being close to zero before lag 50.

7. In-Sample Results

Comparing the in-sample results of the $HAR-RV$ and $MIDAS-RV$ models, we follow Forsberg and Ghysels (2006) and Chung et al. (2008) and use the adjusted R^2 and the in-sample MSE. Because the adjusted R^2 is comparable only between models with the identical dependent variable, we calculate an R^2 for the $HAR-RV$ using the logarithm of the variables. The MSE was calculated through the estimation of the models with all observations, comparing the series of the estimated RV with that observed for the entire sample. As in the case of R^2, we also adjusted the MSE because the $HAR-RV$ model had been estimated with variables in their logarithmic forms. The results of the adjusted R^2 are shown in Table 7.

The results presented in Table 7 show that the $MIDAS-RV$ models for the identical stock and the identical forecast horizon display adjusted R^2 that are slightly better than those of the $HAR-RV$ model. It is noteworthy that the adjusted R^2 for $h=1$ are always lower and increase until two weeks. This occurs because the dependent variable is the mean of the realized volatility of the forecast

Table 5
Estimated Coefficients of the HAR-RV Model

Horizon	c	p-value	β^d	p-value	β^w	p-value	β^m	p-value
BBDC4								
1 day	−0.277	0.02	0.165	0.00	0.472	0.00	0.299	0.00
1 week	−0.339	0.03	0.171	0.00	0.387	0.00	0.354	0.00
2 weeks	−0.444	0.01	0.130	0.00	0.401	0.00	0.352	0.00
3 weeks	−0.529	0.00	0.129	0.00	0.379	0.00	0.353	0.00
4 weeks	−0.627	0.00	0.119	0.01	0.395	0.01	0.322	0.00
PETR4								
1 day	−0.290	0.01	0.198	0.00	0.461	0.00	0.273	0.00
1 week	−0.363	0.02	0.174	0.00	0.413	0.00	0.320	0.00
2 weeks	−0.456	0.01	0.158	0.00	0.364	0.00	0.359	0.00
3 weeks	−0.529	0.00	0.145	0.00	0.345	0.00	0.402	0.00
4 weeks	−0.599	0.00	0.130	0.00	0.307	0.00	0.407	0.00

Table 5 (cont): Estimated Coefficients of the HAR-RV Model

Horizon	c	p-value	β^d	p-value	β^w	p-value	β^m	p-value
VALE5								
1 day	−0.307	0.02	0.186	0.00	0.476	0.00	0.267	0.00
1 week	−0.376	0.02	0.151	0.00	0.436	0.00	0.315	0.00
2 weeks	−0.469	0.01	0.130	0.00	0.430	0.00	0.316	0.00
3 weeks	−0.559	0.00	0.127	0.00	0.385	0.00	0.342	0.00
4 weeks	−0.655	0.00	0.114	0.00	0.378	0.00	0.355	0.00
TNLP4								
1 day	−0.297	0.03	0.182	0.00	0.291	0.02	0.459	0.00
1 week	−0.448	0.01	0.100	0.02	0.300	0.00	0.482	0.00
2 weeks	−0.564	0.00	0.088	0.02	0.345	0.00	0.417	0.00
3 weeks	−0.625	0.00	0.097	0.00	0.287	0.02	0.450	0.00
4 weeks	−0.709	0.00	0.083	0.01	0.309	0.02	0.418	0.00
USIM5								
1 day	−0.282	0.04	0.206	0.00	0.303	0.01	0.422	0.00
1 week	−0.350	0.05	0.131	0.00	0.323	0.00	0.451	0.00
2 weeks	−0.459	0.01	0.096	0.00	0.395	0.00	0.383	0.00
3 weeks	−0.570	0.00	0.098	0.00	0.372	0.00	0.372	0.00
4 weeks	−0.691	0.00	0.093	0.01	0.375	0.00	0.341	0.00

Source: Own calculations.
Note: Coefficients and their respective p-values resulting from the estimation of the HAR-RV models are shown, as developed by Corsi (2009) and described in equation (18).
The data were filtered according to (28).
Standard errors were modified with a Newey-West correction.

Table 6
Estimated Coefficients of the MIDAS-RV Model

Horizon	c	p-value	β^d	p-value	β^w	p-value	β^m	p-value
BBDC4								
1 day	−0.277	0.02	0.165	0.00	0.472	0.00	0.299	0.00
1 week	−0.339	0.03	0.171	0.00	0.387	0.00	0.354	0.00
2 weeks	−0.444	0.01	0.130	0.00	0.401	0.00	0.352	0.00
3 weeks	−0.529	0.00	0.129	0.00	0.379	0.00	0.353	0.00
4 weeks	−0.627	0.00	0.119	0.01	0.395	0.01	0.322	0.00
PETR4								
1 day	−0.290	0.01	0.198	0.00	0.461	0.00	0.273	0.00
1 week	−0.363	0.02	0.174	0.00	0.413	0.00	0.320	0.00
2 weeks	−0.456	0.01	0.158	0.00	0.364	0.00	0.359	0.00
3 weeks	−0.529	0.00	0.145	0.00	0.345	0.00	0.402	0.00
4 weeks	−0.599	0.00	0.130	0.00	0.307	0.00	0.407	0.00

Table 6(cont): Estimated Coefficients of the MIDAS-RV Model

Horizon	c	p-value	β^d	p-value	β^w	p-value	β^m	p-value
VALE5								
1 day	−0.307	0.02	0.186	0.00	0.476	0.00	0.267	0.00
1 week	−0.376	0.02	0.151	0.00	0.436	0.00	0.315	0.00
2 weeks	−0.469	0.01	0.130	0.00	0.430	0.00	0.316	0.00
3 weeks	−0.559	0.00	0.127	0.00	0.385	0.00	0.342	0.00
4 weeks	−0.655	0.00	0.114	0.00	0.378	0.00	0.355	0.00
TNLP4								
1 day	−0.297	0.03	0.182	0.00	0.291	0.02	0.459	0.00
1 week	−0.448	0.01	0.100	0.02	0.300	0.00	0.482	0.00
2 weeks	−0.564	0.00	0.088	0.02	0.345	0.00	0.417	0.00
3 weeks	−0.625	0.00	0.097	0.00	0.287	0.02	0.450	0.00
4 weeks	−0.709	0.00	0.083	0.01	0.309	0.02	0.418	0.00
USIM5								
1 day	−0.282	0.04	0.206	0.00	0.303	0.01	0.422	0.00
1 week	−0.350	0.05	0.131	0.00	0.323	0.00	0.451	0.00
2 weeks	−0.459	0.01	0.096	0.00	0.395	0.00	0.383	0.00
3 weeks	−0.570	0.00	0.098	0.00	0.372	0.00	0.372	0.00
4 weeks	−0.691	0.00	0.093	0.01	0.375	0.00	0.341	0.00

Source: Own calculations.
Note: Coefficients and their respective p-values resulting from the estimation of the MIDAS-RV models are shown, as developed by Ghysels et al. (2004), and described in equation (20).
The data were filtered according to (28).

Table 7
Adjusted R^2

Horizon	BBDC4	PETR4	VALE5	TNLP4	USIM5
HAR-RV					
1 day	0.62	0.60	0.57	0.48	0.56
1 week	0.74	0.75	0.71	0.70	0.71
2 weeks	0.73	0.74	0.70	0.71	0.72
3 weeks	0.70	0.71	0.67	0.68	0.69
4 weeks	0.66	0.68	0.63	0.65	0.65
MIDAS-RV					
1 day	0.64	0.62	0.58	0.50	0.57
1 week	0.78	0.77	0.75	0.72	0.73
2 weeks	0.76	0.76	0.74	0.74	0.75
3 weeks	0.74	0.74	0.71	0.73	0.72
4 weeks	0.69	0.70	0.67	0.70	0.68
Source: Own calculations.					
Note: The adjusted $R^2$of the $HAR-RV$ models were constructed to be comparable to those of the $MIDAS-RV$ models.					

horizon, and therefore, as discussed by Forsberg and Ghysels (2006), it is easier to forecast longer horizons because these series are relatively smoother. In this regard, for all data in question, a better forecast horizon was obtained with $h = 10$.

Table 8 presents the in-sample MSE.

The in-sample MSE suggests that the MIDAS-RV had better in-sample performance for all estimations when identical stocks for the same time horizon are compared. Because the identical factor makes the adjusted R^2 larger in longer horizons, the MSE also displays decreasing behavior until the two-week horizon

8. Out-of-Sample Results

To compare the performance of the out-of-sample models, we follow Chung et al. (2008). We use the out-of-sample MSE and compare the out-of-sample forecast errors through the modified Diebold-Mariano test. The out-of-sample forecast errors are calculated for 60 steps ahead from the estimation of the $HAR-RV$ and $MIDAS-RV$ models without the last 60 observations (3 months). As in the case of the in-sample MSE, the out-of-sample MSE of the $HAR-RV$ model is also transformed using the logarithms of the variables for the statistical analysis. The modified Diebold-Mariano test, proposed by Harvey et al. (1997), compares the difference in the forecast errors through a specific function $g(e)$ that, in this study, as in Harvey et al. (1997), uses the quadratic function. Therefore, let us assume that we have two forecast errors, (e_1, e_2); $t = 1, ..., n$. The null hypothesis

Table 8
In-Sample Mean Squared Errors (MSE)

Horizon	BBDC4	PETR4	VALE5	TNLP4	USIM5
HAR-RV					
1 day	8.13	9.83	12.00	11.01	11.69
1 week	4.16	4.63	5.65	4.27	5.49
2 weeks	4.12	4.43	5.51	3.69	4.91
3 weeks	4.40	4.69	5.77	3.86	5.19
4 weeks	4.84	5.08	6.16	4.11	5.70
MIDAS-RV					
1 day	7.26	8.93	10.84	9.83	10.75
1 week	3.35	3.97	4.66	3.60	4.74
2 weeks	3.40	3.78	4.48	3.06	4.18
3 weeks	3.61	3.99	4.68	3.11	4.38
4 weeks	4.11	4.49	5.21	3.33	4.94
Source: Own calculations.					
Note: The MSE are multiplied by 10^5.					

of the test is the following:

$$H_0 : E[g(e_1) - g(e_2)] = 0 \tag{29}$$

The test statistic is given by:

$$S_1 = \left[\frac{n + 1 - 2\tau + n^{-1}\tau(\tau - 1)}{n}\right]^{1/2} [\widehat{V}(\overline{d})]^{1/2}\overline{d} \tag{30}$$

where n is the number of observations, τ is the number of future periods for which the predictions will be made, $d_t = g(e_1) - g(e_2)$; $t = 1, ...n$, $\overline{d} = n^{-1}\sum_{t=1}^{n} d_t$ and $\widehat{V}(\overline{d})$ is the variance of $\overline{d}$, which is given by:

$$\widehat{V}(\overline{d}) = n^{-1}\left[\widehat{\gamma}_0 + 2\sum_{k=1}^{\tau-1}\widehat{\gamma}_k\right] \tag{31}$$

where $\widehat{\gamma}_k$ is the estimation of the kth autocovariance of d_t.

Table 9 presents the out-of-sample MSE.

The MSE presented in Table 9 are similar between the models, with no apparent pattern that allows determining which model displays better out-of-sample performance. It is also noteworthy that the behavior of the models is similar in relation to the outliers, as in the case of the out-of-sample MSE of TNLP4 in the

Table 9
Out-of-Sample Mean Squared Errors (MSE)

Horizon	BBDC4	PETR4	VALE5	TNLP4	USIM5
HAR-RV					
1 day	1.96	2.45	3.22	18.60	1.52
1 week	0.79	1.19	1.28	4.39	1.07
2 weeks	1.05	1.18	1.55	1.54	0.95
3 weeks	1.38	1.40	1.80	1.00	1.17
4 weeks	1.58	1.56	1.95	0.93	1.19
MIDAS-RV					
1 day	2.10	2.42	3.23	18.97	3.23
1 week	0.81	1.17	1.35	4.55	1.16
2 weeks	1.10	1.15	1.67	3.43	1.04
3 weeks	1.47	1.38	1.94	0.93	1.30
4 weeks	1.77	1.62	2.16	0.91	1.38
Source: Own calculations.					
Note: The MSE are multiplied by 10^5.					

$h = 1$ horizon. Another factor related to the value of the out-of-sample MSE in relation to the in-sample MSE is that the out-of-sample MSE are lower, independent of the estimated model or of the forecast horizon, with the exception of TNLP4 for horizon $h = 1$. This is because the 2008 subprime crisis is represented only in the in-sample errors, while the out-of-sample errors are calculated only in the last 60 observations.

Table 10 presents the p-value of the modified Diebold-Mariano test.

Table 10
Out-of-Sample Diebold-Mariano Test (p-value)

Horizon	BBDC4	PETR4	VALE5	TNLP4	USIM5
1 day	0.990	0.997	0.999	0.970	0.994
1 week	0.998	0.998	0.994	0.987	0.993
2 weeks	0.997	0.997	0.991	0.991	0.993
3 weeks	0.993	0.998	0.989	0.994	0.989
4 weeks	0.985	0.996	0.983	0.998	0.985
Source: Own calculations.					
Note: The test was performed using (30) and a quadratic loss function.					

The p-values of the Diebold-Mariano test, shown in Table 10, confirm that the out-of-sample forecast errors between both models are not significantly different. Therefore, for out-of-sample forecasts, there is no evidence that the performance

of the $MIDAS-RV$ model for the data used is better than that of the $HAR-RV$ model, as opposed to the results of the in-sample analysis.

As the results of the out-of-sample forecasts are similar, we constructed Figure 2 from 60 out-of-sample forecasts for each stock in a time horizon. In addition, the realized volatility of the last day of the estimation is presented and kept constant for all forecasts. We observe that the $HAR-RV$ and $MIDAS-RV$ models display similar behavior throughout the forecasts.

9. Conclusions

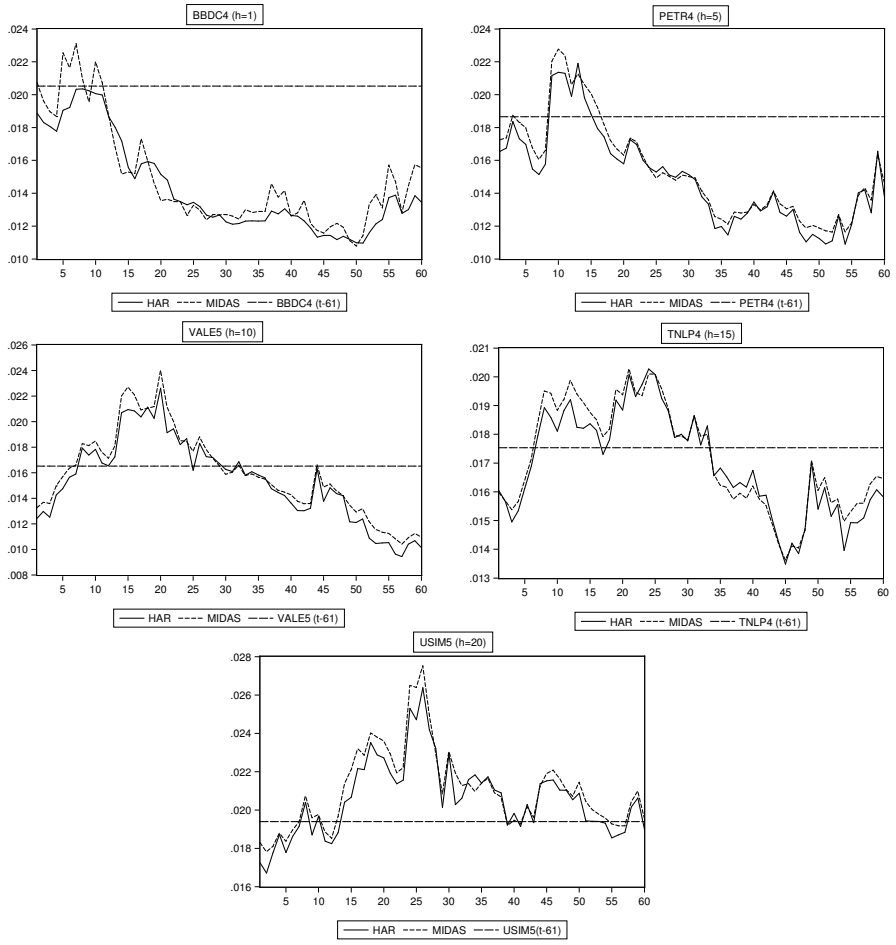
Using the theoretical framework of realized volatility proposed by Andersen et al. (2003), the goal of this study was to compare two models for volatility forecasting using the five most liquid stocks on the BOVESPA using intraday data with a 5-minute frequency during a measured period. We used the following stocks in our study: Bradesco (BBDC4), Petrobrás (PETR4), Vale (VALE5), Telemar (TNLP4), and Usiminas (USIM5). To correct for the bias of the market microstructure, we used the $MA(q)$ filter proposed by Hansen et al. (2008). The models considered in this study were the $HAR-RV$, proposed by Corsi (2009), and the $MIDAS-RV$, proposed by Ghysels et al. (2004).

The comparison between the models was performed using in-sample and out-of-sample measurements of realized volatility. Considering the MSE and the adjusted R^2 suggests that the $MIDAS-RV$ model is better than the $HAR-RV$ model only for in-sample forecasting for the stocks used in our study. For out-of-sample forecasts, according to the modified Diebold-Mariano test for the comparison of MSE, there was no significant difference between the models. Thus, our study suggests that the $HAR-RV$ model should be used for out-of-sample forecasts because of the greater ease of its estimation.

Another important result of this study is the evidence that the use of realized volatility results in distributions of standardized returns that are closer to normal distributions, as previously observed by Carvalho et al. (2006) for Brazilian data.

There are numerous possibilities for studies to complement our work here, mainly related to studies with Brazilian data. For example, we recommend that future studies consider the effects of the jumps in volatility that we observed in the data, test other types of filters for correction of the market microstructure, and apply other models of realized volatility.

Figure 2
Out-of-Sample Forecasts



References

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

Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89:701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61:43–76.

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

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

Carvalho, M. R. C., Freire, M. A. S., Medeiros, M. C., & Souza, L. R. (2006). Modeling and forecasting the volatility of Brazilian asset returns: A realized variance approach. *Revista Brasileira de Finanças*, 4:321–343.

Chung, H. M., Huang, C. S., & Tseng, T. C. (2008). Modeling and forecasting of realized volatility based on high-frequency data: Evidence from Taiwan. *International Research Journal of Finance and Economics*, 22:178–191.

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

Ebens, H. (1999). Realized stock volatility. Mimeo, Department of Economics, Johns Hopkins University.

Forsberg, L. & Ghysels, E. (2006). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics*, 5:31–67.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). The Midas touch: Mixed data sampling regression models. Minemo, CIRANO.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131:59–95.

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

Giot, P. & Laurent, S. (2004). Modelling daily Value-at-Risk using realized volatility and ARCH type models. *Journal of Empirical Finance*, 11:379–398.

Hansen, P. R., Large, J., & Lunde, A. (2008). Moving average-based estimators of integrated variance. *Econometric Reviews*, 27:79–111.

Harvey, A. C., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13:281–291.

Harvey, A. C., Ruiz, E., & Shephard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies*, 61:247–264.

McAleer, M. & Medeiros, M. C. (2008). Realized volatility: A review. *Econometric Reviews*, 27:10–45.

Oomen, R. (2002). Modelling realized variance when returns are serially correlated. Discussion paper, Warwick Business School.

Sá Mota, B. & Fernandes, M. (2004). Desempenho de estimadores de volatilidade na Bolsa de Valores de São Paulo. *Revista Brasileira de Finanças*, 58:429–448.