Financial Volatility Forecasting by Nonlinear Support Vector Machine Heterogeneous Autoregressive Model: Evidence from Nikkei 225 Stock Index

Md. Ashraful Islam Khan

Institute of Economic Research, Graduate School of Economics, Hitotsubashi University

Tokyo 186-8603, Japan

&

Department of Population Science and Human Resource Development, Rajshahi University

Rajshahi-6205, Bangladesh

E-mail: khan75ru@yahoo.com

Received: January 19, 2011 Accepted: February 18, 2011 doi:10.5539/ijef.v3n4p138

Abstract

Support vector machines (SVMs) are new semi-parametric tool for regression estimation. This paper introduced a new class of hybrid models, the nonlinear support vector machines heterogeneous autoregressive (SVM-HAR) models and aimed to compare the forecasting performance with the classical heterogeneous autoregressive (HAR) models to forecast financial volatilities. It was observed through empirical experiment that the newly proposed hybrid (SVM-HAR) models produced higher predicting ability than the classical HAR model.

Keywords: Realized volatility, Bi-power variation, Jump, Tri-power variation, Heterogeneous Autoregressive model, Nonlinear Support Vector Machine, High frequency Nikkei-225 data

1. Introduction

Volatility, the standard deviation of the continuously compounded returns of a financial instrument over a specific time horizon, is both the boon and bane of all traders, you can't live with it and you can't really trade without it. Most of the financial researchers are mainly concerned with modeling and forecasting volatility in asset returns to quantify the risk of financial instruments over a particular time period so that the risk manager and practitioners can realize whether their portfolio will decline in the future and they may want to cell it before it becomes too volatile. Therefore, volatility plays the key roles in the theory and applications of asset pricing, optimal portfolio allocation, and risk management.

Researches on time varying volatility using the time series models have been active ever since Engle (1982) introduced the ARCH model. The GARCH model, generalized by Bollerslev (1986), has been extended in various directions and these extensions recognize based on the various researcher's empirical evidences that there may be important nonlinearity, asymmetry, and long memory properties in the volatility process. The popular extensions can be referred to Nelson's (1991) EGARCH model, Glosten, Jaganathan, and Runkle's (1993) GJR-GARCH which both account for the asymmetric relationship between stock returns and changes in variance (see, e.g., Black 1976, the beginning study of the asymmetric effect and Engle and Ng, 1993 for further discussion). Engle's (1990) AGARCH; Ding, Granger and Engle's (1993) APARCH; Zakoian's (1994) TGARCH; and Sentana's (1995) QGARCH models also have been developed for the flexibility of the models. The stochastic volatility (SV) modeling capitalized on and often contributed in turn to the concurrent development in the Bayesian statistical analysis using Markov chain Monte Carlo procedure (see, e.g., Shephard (2005)).

When GARCH type and SV latent volatility models are used, a well established result in the financial time series literature is that the standardized returns do not have a Gaussian distribution. The excess kurtosis factor of time series motivates the use of heavy-tailed distributions. For example, Student's t distribution has been used by Bollerslev (1987), GED by Nelson (1991), both Student's t and GED by Hsieh (1989) as alternative distributional models for innovations. The researchers have found that returns usually exhibit empirical regularities including thick tails, volatility clustering, leverage effects (see, e.g., Bollerslev et al. 1994). Andersen et al. (2000a, b, 2001, 2003) showed that the distribution of the standardized exchange rate series was almost Gaussian when the realized volatility (RV) was used. Furthermore, the logarithm of the realized volatilities was also nearly Gaussian. It was also

corroborated for stock returns in Andersen et al. (2001a). Other literatures on realized volatility can be referred along with many researchers to Aït-Sahalia and Mancini (2006), Ghysels and Sinko (2006), Corradi et al. (2006), and recently Corsi et al. (2008).

In addition, there is significant evidence of long memory in the time series, which has been conventionally modeled as an ARFIMA (*p,d,q*) process (see, e.g., Andersen et al. 2000a,b, 2001a, 2003). A large number of papers in the RV literature employ the ARFIMA model without a conditionally heteroskedastic error specification to fit daily RV series (see, e.g., Oomen 2001, Giot and Laurent 2004).

Corsi et al. (2001) and Corsi (2009) proposed the Heterogeneous Autoregressive Realized Volatility (HAR-RV) model as an alternative to the ARFIMA model and it has quickly become popular for modeling the dynamics of RV and other related volatility measures due to its ease estimation and extendability of the baseline model. The HAR-RV model employs a few predictor terms, the past daily RVs averaged over different horizons (typically a day, a week, and a month), and is capable to producing slow-decay patterns in autocorrelations exhibited by many RV series. Another recent development in the RV literature is the approach due to Barndorff-Nielsen and Shephard (2004, 2006), Andersen et al. (2003, 2007) of decomposing the RV into the contribution of continuous sample path variation and that of jumps. Extending the theory of quadratic variation of semimartingales, Barndorff-Nielsen et al. (2006) provided an asymptotic statistical foundation for this decomposition procedure under very general conditions.

However, all of the models do require specified distribution of innovations in order to estimate the model specification and to appropriately forecast future values. The semi-parametric approaches do not require any assumptions on data property (return distribution). These models have been successfully used for modeling and forecasting time series including volatility. One of such models is Support Vector Machine (SVM), introduced by Vepnik (1995), that guarantees to obtain globally optimal solution (see, e.g., Cristianini and Shawe-Taylor, 2000), which solves the problems of multiple local optima in which the neural network usually get trapped into. Pẻrez-Cruz et al. (2003) predicted GARCH (1,1) based volatility by SVM and showed that the SVM-GARCH(1,1) model yielded better predictive ability than the parametric GARCH(1,1) model. Chen et al. (2008) proposed recurrent SVM as a dynamic process to model GARCH (1,1) based Volatility and showed through simulated and real data that the model produced better performance than MLE based GARCH (1,1) model. More recently, Ou and Wang (2010) proposed GARCH-LSSVM, EGARCH-LSSVM and GJR-LSSVM hybrid models based on modification of Suykens and Vandewalle (1999) to forecast the leverage effect volatilities of ASEAN stock markets. They showed that these models provided improved performances in forecasting the leverage effect volatilities especially during the recently global financial market crashes in 2008.

This paper, closer to Andersen et al. (2003, 2007), aims to apply the SVM approach on HAR-RV models to forecast empirically the daily RV of the Nikkei 225 index. Watanabe and Yamaguchi (2007), Ishida and Watanabe (2009) among other researchers studied the RV of the Nikkei 225 index and reported empirical findings. But, to the author's knowledge, this paper is the first to apply the SVM-HAR-RV model to RV literature.

The plan for the rest of the paper is as follows. In section 2, we briefly discuss the realized volatility, realized bi-power variation, and jump component extraction. Section 3 describes the data and summary statistics. Section 4 describes the SVM volatility model. Section 5 describes different HAR-RV models. Section 6 reports the forecasting results of the RV and Section 7 concludes with suggestions for further research.

2. Realized volatility, realized bi-power variation and jump component extraction

If we consider a simple diffusion process

$$dp(s) = \mu(s)dt + \sigma(s)dW(s) \tag{1}$$

where $p(s)dt$ is the instantaneous log-price, $W(s)$ is a standard Brownian process and $\sigma(s)$ is the standard deviation of $dp(s)$, which may be time-varying but is assumed to be independent of $dW(s)$. Then the volatility for day *t* is defined as the integral of $\sigma^2(s)$ over the interval $(t, t+1)$ i.e., $\int_t^{t+1} \sigma^2(s)ds$, which is known as integrated volatility and it is unobserved. Let the discretely sampled Δ-period returns be denoted by, $r_{t,\Delta} = p(t) - p(t-\Delta)$. If the process (in our case the log of Nikkei 225 index level process) is a continuous semimartingale then under mild regularity conditions,

$$RV_t \equiv \sum_{j=1}^{1/\Delta} \left| r_{t+j\Delta,\Delta} \right|^2 \xrightarrow{p} \int_t^{t+1} \sigma^2(s)ds \quad \text{as } \Delta \downarrow 0 \tag{2}$$

RV_t is the *t*-th day realized variance since *t* has the daily unit and $\left(\frac{1}{\Delta}\right)$ is integer. We will hereafter use the terms *realized volatility* or *realized variance* interchangeably, or their common abbreviation RV.

Again, if the process is semimartingale with finite-activity jumps, i.e., only a finite number of jumps occurring in any finite time interval, such as Poisson jumps, then the realized variance converges to the quadratic variation, which can be decomposed as,

$$RV_t \xrightarrow{p} \int_t^{t+1} \sigma^2(s)ds + \sum_{t<s\le t+1} k^2(s) \text{ as } \Delta \downarrow 0 \tag{3}$$

where *k(s)* refers to the size of the jump occurring at time *s*. Barndorff-Nielsen and Shephard(2004, 2006) showed that even in the presence of jumps the *bipower variation*

$$BV_t \equiv \mu_1^{-2} \sum_{j=2}^{1/\Delta} |r_{t+j\Delta,\Delta}||r_{t+(j-1)\Delta,\Delta}| \tag{4}$$

where $\mu_1 \equiv \sqrt{2/\pi}$, holds under mild conditions and proposed to use

$$RV_t - BV_t \xrightarrow{p} \sum_{t<s\le t+1} k^2(s) \tag{5}$$

or

$$J_t \equiv max[(RV_t - BV_t), 0] \tag{6}$$

as an estimator for $\sum_{t<s\le t+1} k^2(s)$. J_t is known to take non-zero, small values very frequently due to measurement and possibly due to the presence of jumps infinite-activity types.

Andersen et al. (2007) introduced shrinkage estimator for the jump contribution based on the asymptotic distribution theory developed by Barndorff-Nielsen and Shephard (2004, 2006) and Barndorff-Nielsen et al. (2006) as

$$SJ_t \equiv I(Z_t > \Phi_\alpha).(RV_t - BV_t) \tag{7}$$

where I is an indicator function, $Z_t \equiv \Delta^{-1/2} \frac{(RV_t - BV_t)RV_t^{-1}}{\sqrt{(\mu_1^{-4} + 2\mu_1^{-2} - 5)max[1, TQ_t BV_t^{-2}]}}$ is asymptotically standard normally distributed under the null hypothesis of no jumps, $\mu_1 \equiv \sqrt{2/\pi}$, $\Phi_\alpha = \Phi(\alpha)$, the standard normal distribution function where α is usually set to the values such as .999 so that J_t can picks up only "significance jumps" and the *realized tripower variation*

$$TQ_t \equiv \Delta^{-1}\mu_{4/3}^{-3} \sum_{j=3}^{1/\Delta} |r_{t+j\Delta,\Delta}|^{4/3} |r_{t+(j-1)\Delta,\Delta}|^{4/3} |r_{t+(j-2)\Delta,\Delta}|^{4/3} \xrightarrow{p} \int_t^{t+1} \delta^4(s)ds \quad as \ \Delta \downarrow 0 \tag{8}$$

where $\mu_{4/3} \equiv 2^{2/3}\Gamma(7/6)\Gamma(1/2)^{-1}$. The convergence result holds even in the presence of jumps.

Andersen et al. (2007) introduced the shrinkage estimator for the continuous sample path variation as

$$C_t \equiv I[Z_t \le \Phi_\alpha]RV_t + I[Z_t > \Phi_\alpha]BV_t \tag{9}$$

Andersen et al. (2007) also proposed microstructure-noise-robust versions of BV_t and TQ_t as

$$BV_t \equiv \mu_1^{-2}(1 - 2\Delta)^{-1} \sum_{j=3}^{1/\Delta} |r_{t+j\Delta,\Delta}||r_{t+(j-2)\Delta,\Delta}| \tag{10}$$

$$TQ_t \equiv \Delta^{-1}\mu_{4/3}^{-3}(1 - 4\Delta)^{-1} \sum_{j=5}^{1/\Delta} |r_{t+j\Delta,\Delta}|^{4/3} |r_{t+(j-2)\Delta,\Delta}|^{4/3} |r_{t+(j-4)\Delta,\Delta}|^{4/3} \tag{11}$$

The definitions of J_t and C_t will be modified as well.

3. Data Description and summary statistics

3.1. Calculation of intraday returns and related realized volatility measures from minute-by-minute Nikkei 225 data

This paper measures the realized volatility of the Nikkei 225 index for the sample of the period 11 March 1996 to 30 September 2009. First, constract a "five- minute (percentage) returns" series by taking the five-minute log differences multiplied by hundred from the minute-by-minute data. This choice is made to mitigate the effect of microstructure related noise and increase the precision of volatility measures. (see, e.g.,Ishida and Watanabe, 2009; Watanabe and Yamaguchi, 2007).

The Tokyo Stock Exchange is open only for 9:00-11:00 (Morning Session) and 12:30-15:00 (Afternoon Session). Our database includes every minute prices of the Nikkei 225 stock index for both sessions. This paper first extracts prices for 9:01, 9:05, 9:10,.....,11:00 in the morning session and for 12:31, 12:35, 12:40,…….,15:00 in the afternoon session. Sometimes, the last transaction price for morning (and/or afternoon) session is observed slightly after 11:00 (and/or 15:00). In such cases, the last prices instead of prices at 11:00 (and/or 15:00) are used. Next using these prices the five-minute returns as mentioned in section 2 are calculated. There are 54 five-minute returns for a typical trading day in total, 24 from the morning session and 30 from the afternoon session.

Given the recent literature on the market microstructure noise effect on realized volatility estimation, the optimal choice of sampling frequency as studied by Bandi and Russell (2003, 2008) has been considered here. The sampling frequency M^{opt} (the number of observations per day) is calculated as (see also Zhang et al., 2005 and Clements et al. 2008)

$$M^{opt} = \left(\frac{\bar{\hat{Q}}_i}{\hat{\alpha}}\right)^{\frac{1}{3}} \tag{13}$$

where $\hat{\alpha} = \left(\frac{\sum_{t=1}^{T}\sum_{j=1}^{M} r_{j,t}}{TM}\right)^2$ and $\bar{\hat{Q}} = \frac{1}{T}\sum_{t=1}^{T}\hat{Q}_t$, $\hat{Q}_t = \frac{M_{15}}{3}\sum_{j=1}^{M_{15}} r_{j,t}^4$. M_{15} is the 15-minute returns and M is the highest frequency at which data are available. In our case, it is 1-minute returns. The 15-minute intraday returns are being considered to calculate realized volatility as well.

We cannot calculate the 5-minute, 15-mimute and optimally-sampled returns for the non-trading hours including lunch time and overnight period though we can calculate the lunch time and overnight returns by considering the last price of the morning session and the first prices of the afternoon session, and the last price of the afternoon session and the first price of the next morning session but following Hansen and Lunde (2005), we drop this idea and scale the realized volatility as follow,

$$RV_t \equiv C^* RV_t^{(0)}$$

where $C^* \equiv \sum_{t=1}^{T}(R_t - \bar{R})^2 \Big/ \sum_{t=1}^{T} RV_t^{(0)}$, where $\bar{R} \equiv \sum_{t=1}^{T} R_t / T$, and T is the number of complete trading days.

In my sample period, the first trading in the second session from January 1, 2006 to April 21, 2006 observed at 13:01. Therefore, I remove these trading days along with the sessions from half trading days including the first and the last trading days of each year. The remaining number of complete trading days, T is 3279. We calculate RV_t and J_t by using this 3279 days data for the four series.

3.2. Properties of the realized volatility and related measures

Summary statistics of daily returns, the daily RV, it's standard deviation form, i.e., $RV^{1/2}$, the logarithmic form i.e, *ln(RV)*, the daily jump, microstructure noise robust version of daily jump (MSNR-Jump) (where BV_t has been calculated according to Andersen et al., 2007) series and their standard deviation and logarithmic) are presented on Table 1a. The summary statistics of continuous path component, significant jump series, the microstructure noise robust version of the continuous path component (MSNR-C) and significant jump (MSNR-SJ) series due to Andersen et al. (2007), and their standard deviation and logarithmic form are presented on Table 1b. In addition to the sample skewness and kurtosis, the Jarque-Bera (JB) statistic for testing normality and the Ljung-Box statistics of order 5, 10 and 22 (corresponding to roughly one week, two weeks and a month) for testing serial correlations up to their respective order are also presented on the Tables.

From Table 1a, we observed that the unconditional distribution of the daily return series is negatively skewed but highly significantly nonnormal with high positive kurtosis. The LB statistics also indicate that the series is significantly serially correlated. We also observed that the daily RV, Jump and MNSR-Jump series are highly unconditionally nonnormally distributed with large positive values of skewness and kurtosis and highly significantly serially correlated. The average value of Jump and MSNR-Jump are 1.471 and 1.504 respectively with positive minimum values, that implies more than one jump occurred in every single days.

The square-root transformation reduces the deviation from normality but still huge. The log transformation brings down the sample skewness and kurtosis values for the RV series but still significantly nonnormal. All the transformed series remain highly significantly serially correlated.

Looking at the summary statistics from Table 1b, where $\alpha = 0.999$ in (7) and (9), we observed that the average of the significant jump and MSNR-SJ are slightly reduced but still greater than one while the minimum values reduced to zero. The Jauque-Bera statistic shows the strong evidence of highly significant nonlinearity for all series and the LB statistic shows the strong evidence of highly significant serial correlation.

Figure 1. shows the daily RV, Jump, MSNR-Jump, Significant Jump and MSNR-SJ. We visually observe few big jumps in the initial stage and the biggest jumps in the ending part of our sample period, the period of global financial market crashes.

4. The Support Vector Machines (SVMs)

The Support Vector Machines (SVMs) were introduced by Vapnik (1995) based on the statistical learning theory, which had been developed over the last three decades by Vapnik, Chervonenkis and others (see, e.g., Vapnik 1982,

1995) from a nonlinear generalization of the Generalized Portrait algorithm. SVMs were developed to solve the classification problem, but recently they have been extended to the domain of regression problems (e.g.,Vapnik et al.1997). The SVMs usually map data to a high-dimensional feature space and apply a simple linear method to the data in that high-dimensional space nonlinearly related to the input space. Moreover, even though we can think of SVMs as a linear algorithm in a high-dimensional space, in practice, it does not involve any computations in that high-dimensional space (see, e.g., Karatzoglou and Meyer 2006). The terminology for SVMs can be slightly confusing in the literature. In few literatures, SVM refers to both classification and regression with support vector methods. In this paper, the tern SVM will be used for the Nonlinear Support Vector Regression (NL-SVR). The mathematical formulation of SVM is as follows,

In the ϵ-insensitive support vector regression of Vapnik (1995), our goal is to find a function $f(x)$ that has an ϵ deviation from the actually obtained targets y_t for all training data, and at the same time, is as flat as possible. Suppose $f(x)$ takes the following form

$$f(x) = k\langle\omega, x\rangle + b \quad \text{with } \omega \in X, b \in \mathbb{R} \tag{14}$$

where X is the space of the input patterns and $k\langle .,. \rangle$ denotes the kernel function. Flatness of the above model means need to find the small ω. One way to ensure this is to minimize the Euclidean norm, i.e.,$\|\omega\|^2$ (see, e.g., Smola 1998). By applying the *soft margin* formulation of Cortes and Vapnik, (1995), and the Karush-Kuhn-Tucker (KKT) conditions (Karush, 1939, Kuhn and Tucker 1951) one can estimate the above model as

$$f(x) = \sum_{t=1}^{T}(\alpha_t - \alpha_t^*)k\langle x_t, x\rangle + \hat{b} \tag{15}$$

where b can be computed as

$\hat{b} = y_t - k\langle\omega, x_t\rangle - \epsilon \quad for \ \alpha_t \in (0, C)$

$\hat{b} = y_t - k\langle\omega, x_t\rangle + \epsilon \quad for \ \alpha_t^* \in (0, C)$

where, $C > 0$ determines the trade-off between the flatness of the $f(x)$ and the amount up to which derivations larger than ϵ are tolerated and $\alpha_t, \alpha_t^* \geq 0$. See,e.g., Smola and Schölkopf (1998) for further discussion. A several numbers (see, e.g., Kernlab in R, MATLAB, etc) of statistical software are available to handle SVM method.

According to Cortes and Vepnik(1995), any symmetric positive semi-definite function that satisfies the Mercer's conditions can be used as a kernel function in the SVMs context. The Mercer's conditions are

$\int\int K(x,y)\, g(x)g(y)dxdy > 0$ and $\int g^2(x)dx < \propto$,

where $K(x,y) \equiv \sum_{t=1}^{\propto} \alpha_t\, \psi(x)\, \psi(y),\ \alpha_t \geq 0$

This paper used the Polynomial kernel function (used for out-of-sample forecast) and Laplacian kernel function (used for in-sample forecast) for SVMs. The general form of the Polynomial kernel function is

$$K(x,y) \equiv (\text{scale}.\langle x, y\rangle + \text{offset})^{degree}$$

and the Laplacian kernel function is

$$K(x,y) \equiv exp\left(-\frac{\|x-y\|}{\sigma}\right)$$

See,e.g., Smola and Schölkopf (1998) for further discussion.

5. HAR-RV models

The HAR-RV class volatility models proposed by Corsi (2003) on the basis of a straightforward extension of the so-called Heterogeneous ARCH (HARCH) class of models analyzed by Müller et al.(1997).

To sketch the HAR-RV model, define the multi-period realized volatilities by the normalized sum of the one-period volatilities,

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

Note that, by definition of the daily volatilities, $RV_{t,t+1} \equiv RV_{t+1}$. Also, provided the expectations exist, $E\left(RV_{t,t+1}\right) \equiv E(RV_{t+1})$ for all *h*. (see, e.g., Andersen et al. 2003, 2007). Also *h=5* and *h=22* will produce the weekly and monthly volatilities, respectively. The daily HAR-RV model of Corsi (2003) may then be expressed as

$$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \in_{t+1} \tag{17}$$

where $t = 1,2, \ldots\ldots\ldots\ldots, T$.

Andersen et al.(2003, 2007) included the jump component, which has been explained in the Section 2, as an explanatory variable to the above model and introduced the new model as

$$RV_{t,t+h} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \beta_J J_t + \in_{t,t+h} \tag{18}$$

The standard deviation and logarithmic form of the above model respectively are

$$\left(RV_{t,t+h}\right)^{1/2} = \beta_0 + \beta_D (RV_t)^{1/2} + \beta_W \left(RV_{t-5,t}\right)^{1/2} + \beta_M \left(RV_{t-5,t}\right)^{1/2} + \beta_J (J_t)^{1/2} + \in_{t,t+h} \quad (19)$$

and $$log\left(RV_{t,t+h}\right) = \beta_0 + \beta_D log(RV_t) + \beta_W log\left(RV_{t-5,t}\right) + \beta_M log\left(RV_{t-22,t}\right) + \beta_J log(1 + J_t) + \in_{t,t+h} \quad (20)$$

After introducing the so-called shrinkage and microstructure-noise-robust estimator for the significance jump and continuous sample path variation, those have been discussed in the Section 2, Andersen et al. (2007) represented the HAR-RV-CJ model as

$$RV_{t,t+h} = \beta_0 + \beta_{CD} C_t + \beta_{CW} C_{t-5,t} + \beta_{CM} C_{t-22,t} + \beta_{JD} SJ_t + \beta_{JW} SJ_{t-5,t} + \beta_{JM} SJ_{t-22,t} + \epsilon_{t,t+h} \quad (21)$$

where, $C_{t,t+h} = h^{-1}(C_{t+1} + C_{t+2} + \cdots + C_{t+h})$ and $SJ_{t,t+h} = h^{-1}(SJ_{t+1} + SJ_{t+2} + \cdots + SJ_{t+h})$

The standard deviation and logarithmic version of this model respectively are

$$\left(RV_{t,t+h}\right)^{1/2} = \beta_0 + \beta_{CD} C_t^{1/2} + \beta_{CW} \left(C_{t-5,t}\right)^{1/2} + \beta_{CM} \left(C_{t-22,t}\right)^{1/2} + \beta_{JD} SJ_t^{1/2} + \beta_{JW} \left(SJ_{t-5,t}\right)^{1/2} + \beta_{JM} \left(SJ_{t-22,t}\right)^{1/2} + \epsilon_{t,t+h} \quad (22)$$

and $$log\left(RV_{t,t+h}\right) = \beta_0 + \beta_{CD} log(C_t) + \beta_{CW} log\left(C_{t-5,t}\right) + \beta_{CM} log\left(C_{t-22,t}\right) + \beta_{JD} log(1 + SJ_t) + \beta_{JW} log\left(1 + SJ_{t-5,t}\right) + \beta_{JM} log\left(1 + SJ_{t-22,t}\right) + \epsilon_{t,t+h} \quad (23)$$

See, Andersen et al. (2003, 2007) for further discussion.

6. Modeling and forecasting RV with HAR-RV and SVM-HAR-RV models

This paper compared the forecasting performance of the SVM-HAR-RV class models with the classical HAR-RV class model. For this comparison, the in-sample period considered from March 11, 1996 to December 29, 2004 and out-of-sample period from January 5, 2005 to September 30, 2009, the period including global financial market crashes. First, estimated model (17) (using RV, standard deviation and logarithm of RV series), (18), (19), (20), (21), (22) and (23) by ordinary least squares (OLS) method and next, by SVM setting the values $C = 1$ and $\epsilon = 0.1$ to these models and named SVM-HAR-RV models. The R 2.12.0-win32 and R 2.12.0-win32's Kernlab package were used for both HAR-RV and SVM-HAR-RV class models.

Both class of models for horizons *h* = 1, 5, and 22 days were estimated. To compare the forecasting performance, the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Root Mean Square Percentage Error (RMSPE) and Mean Absolute Percentage Error (MAPE) were computed, which defined as follow:

$$RMSE \equiv \sqrt{\frac{1}{N}\sum_{t=1}^{T}\left(RV_t - \hat{R}V_{t|t-1}\right)^2}, \qquad MAE \equiv \frac{1}{N}\sum_{t=1}^{T}\left|RV_t - \hat{R}V_{t|t-1}\right|,$$

$$RMSPE \equiv \sqrt{\frac{1}{N}\sum_{t=1}^{T}\left(1 - \frac{\hat{R}V_{t|t-1}}{RV_t}\right)^2}, \text{ and} \qquad MAPE \equiv \sqrt{\frac{1}{N}\sum_{t=1}^{T}\left|1 - \frac{\hat{R}V_{t|t-1}}{RV_t}\right|}.$$

where $\hat{R}V_{t|t-1}$ denotes one-day ahead realized volatility forecast. We evaluate these errors for 5 days ahead and 22-days ahead volatility forecast as well.

To save space, this paper did not include the estimation results of all models. The values of R^2 for different models are presented in Table 2a and Table 2b while the forecasting errors are presented in Table 3a and 3b.

6.1 Empirical Results

Let us first compare the R^2 results. It is observed from Table 2a, presents values of R^2 for different models, that the value of R^2 successively increases for the standard deviation of RV series than RV series and for the log RV series than standard deviation series for all the models and all different horizons but successively decreases for higher horizons in each and every series. In each series and horizon, the in-sample forecasting performance of SVM-HAR-RV models is remarkably better than HAR-RV models for each and every series and horizon. The out-of sample forecasting of SVM-HAR-RV models is also higher that the HAR-RV models for standard deviation and logarithmic series. Only the values of out-of-sample R^2 of HAR-RV model for the RV series are slightly higher than the SVM-HAR-RV. Almost similar results (differ in values) have been observed to compare the forecasting performance of the HAR-RV-J (and MSNR-J) and SVM-HAR-J (and MSNR-J) models. It is also observed for both classes of models that the model performances improved after adding the jump (and/or MSNR-J) component as explanatory variable. The MNSR-Jump remarkably improves the predictive ability for the SVM-HAR-RV-MSNR-J class models but not for the HAR-RV-J class models.

Table 2b presents values of R^2 for the HAR-RV-RV-CJ, HAR-RV-RV-MSNR-CJ, SVM-HAR-RV-CJ and SVM-HAR-RV-MSNR-CJ models. This table also produced the similar results as Table 2a. The value of R^2 successively increases for the standard deviation of RV series than RV series and for the log RV series than standard deviation series for all the models and all different horizons but successively decreases for higher horizons in each and every series. In the in-sample case, The SVM-HAR-RV-CJ (and/or SVM-HAR-RV-MSNR-CJ) class of models performed well that the HAR-RV-CJ (and/or HAR-RV-MSNR-CJ) class of models. The out-of-sample performances of SVM-HAR-RV class models are also satisfactory.

The logarithmic transformed series produced better performances compared to RV and standard deviation of RV series for both classes of models. For both class of models, the best performances observed when 5-minute intraday returns are used to estimate the realized volatility.

Next, the different errors are calculated for the logarithmic transformed series for both classes of models.

Let us now compare the results based on different above defined error squares. Table 3a represents the forecasting errors for HAR-RV, SVM-HAR-RV, HAR-RV-J, SVM-HAR-RV-J, HAR-RV-MSNR-J and SVM-HAR-RV-MSNR-J models while and 3b presents the forecasting errors for HAR-RV-CJ, SVM-HAR-RV-CJ, HAR-RV-MSNR-CJ and SVM-HAR-RV-MSNR-CJ models. It is observed that in the in-sample case, the SVM-HAR class models completely defeat the HAR-RV class models for every series, horizon and intraday returns series. For the Out-of-sample case, the performance of SVM-HAR class models is also satisfactory compared to HAR-RV class models. Figure 2 presents the out-of-sample forecasting performances of the above models when 5-minute intraday returns are used.

7. Concluding remarks

This paper combined the Support Vector Machine (SVM) regression with Heterogeneous Autoregressive (HAR) model as a hybrid model (SMV-HAR model) to improve the volatility forecasting ability. It is examined the realized volatility forecasting ability of the models for Nikkei 225 stock returns. The empirical results presented here are suggestive for several interesting extensions. First, the values $C = 1$ and $\epsilon = 0.1$ for the SVM-HAR-RV class models were set and observed better forecasting ability. The appropriate choice of the value C and ϵ could be helpful to improve the forecasting ability.

Second, the Polynomial and Laplaceian kernel were considered for the SVMs and observed better performances. The appropriate choice of other existing kernels in SVM literature or an appropriate new kernel could improve the forecasting ability.

Third, the optimally sampled sampling frequencies are considered to mitigate the market microstructure noise. This choice failed to improve the forecasting performances. It would be interesting to consider the other market microstructure noise mitigation techniques.

Those topics are left for further research.

References

Aït-Sahalia, Y., & Mancini, L. (2006). Out of Sample Forecasts of Quadratic Variation. *Journal of Econometrics*, 109, 33-65.

Andersen, T.G., Bollerslev, T., Diebold, F.X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71, 529-626.

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. (2000a). Exchange rate returns standardized by realized volatility are (nearly) Gaussian. *Multinational Finance Journal*, 4, 159-179.

Andersen, T.G., Bollerslev, T., Diebold, F.X., & Labys, P. (2000b). Great realizations. *Risk*, 13, 105-108.

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

Bandi, F.M., & Russell, J.R. (2003). Microstructure noise, realized volatility, and optimal sampling. *Working Paper*.

Bandi, F.M., & Russell, J.R., (2008). Microstructure noise, realized volatility, and optimal sampling. *Review of Economic Studies*, 75, 339-369.

Barndorff-Nielsen, O.E., Graversen, S.E., Jacod, J., Podolskij, M. & Shephard, N. (2006). A central limit theorem for realised power and bipower variation of continuous semimartingales. In Kabanov,Y., Lipseter,R., Stoyanov,J. (Eds.), *Stochastic Analysis to Mathematical Finance*, Festschrift for Albert Shiryaev, Springer Verlag: Berlin.

Barndorff-Nielsen, O.E., & Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2, 1-37.

Barndorff-Nielsen, O.E., & Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4, 217-252.

Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the 1976 Meetings of the Business and Economics Statistics Section*, *American Statistical Association*, 177-181.

Bollerslev, T. (1986). Generalized Auto Regressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31, 307-327.

Bollerslev, T. (1987). A conditional heteroskedastic time series model for speculative prices and rates of return. *Review of Economic Statistics*, 69, 542-547.

Chen, S., Jeong, K., & Hardle, W. (2008). Support Vector Regression Based GARCH Model with Application to Forecasting Volatility of Financial Returns. *SFB 649 Discussion Paper of Economic Risk, Berlin.*

Clements M.P., Galvão A.B., & Kim J.H. (2008). Quantile forecasts of daily exchange rate returns from forecasts of realized volatility. *Journal of Empirical Finance*, 15(4), 729-750.

Corradi, V., Distaso, W., & Swanson, N.R. (2006). Predictive Inference for Integrated Volatility. Working Paper, Rutgers University

Corsi, F., Kretschmer,U., Mittnik,S., & Pigorsch, C.(2005). The volatility of realized volatility. Discussion paper, no. 2005/33, Centre for Financial Studies, University of Frankfurt.

Corsi, F., Mittnik, S., Pigorsch, C., & Pigorsch, U. (2008). The volatility of realized volatility. *Econometric Reviews*, 27, 46-78.

Corsi, F., Zumbach, G., Müller, U. and Dacorogna, M., (2001). Consistent high-precision volatility from high frequency data. *Economic Notes*, 30, 183-204.

Corsi, F. (2009). A Simple Long Memory Model of Realized Volatility. *Journal of Financial Econometrics*, 7(2), 174-196.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *M. Learning*, 20, 273-297.

Cristianini, N., and Showe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* London, Cambridge University Press.

Ding, Z. C., Granger, W.J., & Engle, R.F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1, 83-106.

Donaldson, R.G., & Kamstra, M., (1997). An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Finance*, 4(1), 17-46.

Engle, R. F., & Ng, V. K. (1993). Measuring and Testing the Impact of News on Volatility. *Journal of Finance*, 48, 1749-1778.

Engle, R.F. (1982). Autoregressive conditional heteroskedasticity with estimates of variance of UK inflation. *Econometrica*, 50, 987-1008.

Engle, R. F. (1990). Discussion: Stock market volatility and the crash of 87. *Review of Financial Studies*, 3, 103-106.

Ghysels, E., & Sinko, A. (2006). Volatility Forecasting and Microstructure Noise. Working Paper, University of North Carolina

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

Glosten, L., Jagannathan,R., & Runkle, D. (1993). On the relationship between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 46, 1779-1801.

Hansen, P.R., & Lunde, A. (2005). A Forecast Comparison of Volatility Models: Does anything Beat a GARCH(1,1). *Journal of Applied Econometrics*, 20(7), 873-889.

Hsieh, D. A. (1989). The statistical properties of daily foreign exchange rates. *J. International Economics*, 24, 129-145.

Ishida, I., and Watanabe, T. (2009). Modeling and Forecasting the Volatility of the Nikkei 225 realized Volatility using the ARFIMA-GARCH model. Working paper, Institute of Economic Research, Hitotsubashi University.

Karatzolou, A., & Meyer, D. (2006). Support Vector Machines in R. *Journal of Statistical Software*, 15(9).

Karush, W. (1939). Minima of functions of several variables with inequalities as side constraints. Master's thesis, Dept. of Mathematics, University of Chicago.

Kuhn, H.W., & Tucker, A.W. (1951). Nonlinear programming. In Proceedings of. 2nd Berkeley Symposium on Mathematical Statistics and Probabilities, 481-492, Berkeley, University of California Press.

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Puctet, O. V., & von Weizsacker, J. (1997). Volatilities of different time resolutions: Analyzing the dynamics of market components. *Journal of Empirical Finance*, 4, 213-239.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, 59, 347-370.

Oomen, R. C. A. (2001). Using high frequency stock market index data to calculate, model & forecast realized return variance. *Working paper,* European University Institute.

Ou, Phichhang & Wang, Hengshan (2010). Financial Volatility Forecasting by Least Square Support Vector Machine Based on GARCH, EGARCH and GJR Models: Evidence from ASEAN Stock Markets. *International Journal of Economics and Finance (Online)*, 2 (1).

Pérez-Cruz, F., Afonso-Rodriguez, J. A., & Giner, J. (2003). Estimating GARCH models using support vector machines. *Journal of Quantitative Finance*, 3(3), 163-172.

Sentana, E. (1995). Quadratic ARCH models. *Review of Economic Studies*, 62(4), 639-661.

Shephard, N. (2005). *Stochastic volatility: Selected Readings*, Oxford University Press: Oxford.

Smola, A. J., & Schölkopf, B. (1998). A Tutorial on Support Vector Regression. *NeuroCOLT2 Tenical Report Series*, *NC2-TR-1998-030*.

Suykens J. A. K. & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Networks,* 9, 293-300

Smola, J. A. (1998). Learning with Kernels. Ph.D. thesis, Technische Universität Berlin.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer-Verlag, New York.

Vapnik, V., Golowich, S., & Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. In Mozer, M., Jordan, M., & Petsche,T., (Eds.),Editors, *Advances in Neural Information Processing Systems*, 9 (pp. 281-287). Cambridge, MA, MIT Press.

Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer Verlag, Berlin.

Watanabe, T., Yamaguchi, K. (2007). Measuring, Modeling and Forecasting Realized Volatility in the Japanese Stock Market. *Discussion Paper,* Research Unit for Statistical and Empirical Analysis in Social Science, Institute of Economic Research, Hitotsubashi University.

Zakoian, J. M. (1994). Threshold Heteroscedastic Models. *Journal of Economic Dynamics and Control*, 18, 931-955.

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

Table 1a. Summary Statistics for Nikkei 225 Daily Returns, Realized Volatility and Jumps

	Series	Mean	Std. Dev.	Skew.	Kurt.	Min.	Max.	Jarque-Bera	LB(5)	LB(10)	LB(22)
	R	-0.024	1.621	-0.163	8.351	-11.953	12.912	3898.857	13.589	24.305	41.57
5-Minute	***RV***	2.620	3.311	7.785	98.225	0.135	59.537	1263095.720	5691.204	9640.407	15588.90
	$RV^{1/2}$	1.476	0.665	2.413	15.917	0.367	7.716	25796.526	7190.833	12522.802	21726.91
	ln (RV)	0.609	0.817	**0.030**	**3.719**	-2.005	4.087	70.550	7710.639	13667.646	24950.93
	J	1.471	1.700	6.542	71.337	0.069	26.163	656785.453	5247.666	8861.280	14569.52
	$J^{1/2}$	1.111	0.486	2.082	12.789	0.263	5.115	15350.865	6597.038	11487.356	20190.51
	ln (1+J)	0.792	0.426	1.262	6.220	0.067	3.302	2270.538	6669.961	11674.345	20728.59
	MSNR-J	1.504	1.760	6.149	62.688	0.063	26.759	503859.151	5387.071	9204.820	15255.27
	$MSNR\text{-}J^{1/2}$	1.120	0.499	2.063	12.121	0.252	5.173	13595.248	6455.825	11337.073	19998.88
	ln (1+MSNR-J)	0.800	0.436	1.273	6.090	0.061	3.324	2174.681	6477.022	11425.650	20373.58
15-Minute	***RV***	2.621	4.014	7.868	96.357	0.073	69.809	1216009.157	4754.751	7958.047	12526.44
	$RV^{1/2}$	1.431	0.757	2.600	16.160	0.271	8.355	27164.726	5830.113	10110.552	17369.59
	ln (RV)	0.496	0.927	**0.117**	**3.615**	-2.614	4.246	58.760	6210.035	11092.093	20348.31
	J	1.434	2.208	6.658	70.959	0.036	36.743	650619.489	3818.669	6281.099	10107.56
	$J^{1/2}$	1.045	0.585	2.367	13.050	0.189	6.062	16741.788	4596.683	7995.894	13779.03
	ln (1+J)	0.724	0.496	1.558	6.509	0.035	3.631	2988.122	4549.871	7996.372	13955.17
	MSNR-J	1.448	2.562	8.836	122.242	0.015	52.421	1971380.206	3640.938	6077.458	9693.37
	$MSNR\text{-}J^{1/2}$	1.036	0.611	2.776	17.646	0.123	7.240	33282.970	4459.291	7767.385	13325.38
	ln (1+MSNR-J)	0.714	0.507	1.678	7.266	0.015	3.978	3997.658	4370.037	7679.203	13377.17
Optimally Sampled	***RV***	2.623	5.065	10.918	182.604	0.040	110.555	4440972.411	3180.071	5372.162	8803.31
	$RV^{1/2}$	1.389	0.834	3.108	22.342	0.200	10.515	55989.186	4604.499	8037.433	14005.35
	ln (RV)	0.385	1.027	**0.087**	**3.635**	-3.220	4.706	58.809	4620.977	8306.383	15354.54
	J	1.600	3.776	12.819	242.346	0.016	92.986	7861031.989	2291.443	3700.800	6474.52
	$J^{1/2}$	1.045	0.712	3.526	27.203	0.126	9.643	86216.756	3480.552	6049.514	10717.92
	ln (1+J)	0.716	0.557	1.839	8.158	0.016	4.543	5443.596	3469.273	6093.157	10854.49
	MSNR-J	1.508	3.977	14.713	306.319	0.009	100.773	12599157.884	2188.317	3478.987	5945.87
	$MSNR\text{-}J^{1/2}$	0.997	0.717	3.809	32.012	0.093	10.039	122060.547	3206.824	5415.597	9962.59
	ln (1+MSNR-J)	0.673	0.557	1.923	8.672	0.009	4.623	6371.333	3082.669	5473.863	9818.06

Key: The sample of the period 11 March 1996 to 30 September 2009, there are total 3281 Daily observations. The 5% critical values for Jarque-Bera (i.e.,$\chi^2(k)$) and LB (k) are 5.991 (k=2), 11.070 (5), 18.924 (10) and 33.924 (22) respectively.

Table 1b. Summary Statistics for Nikkei 225 Continuous Path Components and Significant Jumps

	Series	Mean	Std. Dev.	Skew.	Kurt.	Min.	Max.	Jarque-Bera	LB(5)	LB(10)	LB(22)
5-Minute	***C***	1.231	2.209	12.430	237.994	0.051	59.537	7575647.613	3346.519	5690.879	9141.11
	$C^{1/2}$	0.982	0.518	3.469	28.321	0.227	7.716	93510.670	5891.396	10293.427	17739.03
	ln (C)	-0.242	0.883	0.227	3.899	-2.969	4.087	137.688	6959.732	12354.487	22352.85
	MSNR-C	1.211	2.204	12.687	244.084	0.051	59.537	7972480.660	3097.589	5308.704	8504.47
	$MSNR\text{-}C^{1/2}$	0.973	0.514	3.537	29.407	0.225	7.716	101396.154	5757.799	10023.882	17252.24
	ln (MSNR-C)	-0.260	0.884	0.213	3.937	-2.979	4.087	143.745	6971.435	12274.654	22152.03
	SJ	1.389	1.559	6.169	68.703	0.000	25.769	606303.402	4387.062	6875.528	11179.50
	$SJ^{1/2}$	1.066	0.502	1.361	9.708	0.000	5.076	7109.668	3677.125	6158.408	10861.25
	ln (1+SJ)	0.759	0.432	0.990	5.408	0.000	3.287	1317.949	4262.316	7306.657	13040.37
	MSNR-SJ	1.409	1.592	5.814	60.312	0.000	24.516	463961.957	4397.535	6950.453	11354.16
	$MSNR\text{-}SJ^{1/2}$	1.072	0.510	1.361	9.308	0.000	4.951	6402.570	3585.086	6035.599	10675.92
	ln (1+MSNR-SJ)	0.764	0.439	1.001	5.320	0.000	3.239	1273.191	4122.729	7099.981	12725.01
15-Minute	***C***	1.963	3.571	10.090	147.981	0.013	69.809	2906897.518	3329.980	5689.228	8654.77
	$C^{1/2}$	1.214	0.699	3.002	21.498	0.116	8.355	51310.543	4130.561	7216.278	12357.42
	ln (C)	0.129	1.009	**0.007**	**3.595**	-4.315	4.246	48.091	4048.278	7218.492	13376.87
	MSNR-C	1.999	3.580	10.012	146.198	0.023	69.809	2836334.540	3427.439	5711.740	8618.89
	$MSNR\text{-}C^{1/2}$	1.231	0.697	3.022	21.614	0.152	8.355	51960.071	4311.802	7382.469	12551.65
	ln (MSNR-C)	0.167	0.986	**0.039**	**3.653**	-3.765	4.246	58.698	4355.868	7677.756	14154.20
	SJ	0.659	1.738	7.276	100.100	0.000	36.743	1307863.558	249.240	436.390	766.69
	$SJ^{1/2}$	0.413	0.699	1.938	7.708	0.000	6.062	5044.426	126.215	247.349	427.75
	ln (1+SJ)	0.301	0.532	1.956	6.727	0.000	3.631	3960.783	164.648	327.387	562.13
	MSNR-SJ	0.622	1.684	7.068	92.235	0.000	35.369	1107413.511	258.262	451.908	795.90
	$MSNR\text{-}SJ^{1/2}$	0.397	0.681	2.020	8.086	0.000	5.947	5722.957	134.555	258.920	448.19
	ln (1+MSNR-SJ)	0.287	0.519	2.073	7.316	0.000	3.594	4858.892	176.726	345.796	593.51
Optimally Sampled	***C***	2.310	3.900	10.411	200.702	0.005	106.605	5361480.209	2093.133	3675.979	5958.24
	$C^{1/2}$	1.319	0.755	2.592	17.459	0.072	10.325	32007.917	3665.997	6353.470	11121.59
	ln (C)	0.285	1.037	**-0.097**	**3.654**	-5.25	4.669	63.064	3639.924	6575.333	12291.61
	MSNR-C	2.324	3.890	10.465	202.603	0.009	106.605	5464595.264	2068.260	3671.110	6015.31
	$MSNR\text{-}C^{1/2}$	1.327	0.751	2.606	17.659	0.093	10.325	32835.600	3826.086	6642.548	11674.55
	ln (MSNR-C)	0.304	1.023	**-0.091**	**3.730**	-4.741	4.669	76.875	3974.559	7171.076	13384.52
	SJ	0.313	2.994	20.818	540.025	0.000	92.986	39360820.507	434.935	614.928	1348.21
	$SJ^{1/2}$	0.102	0.550	8.106	92.877	0.000	9.643	1131560.309	140.412	205.727	523.96
	ln (1+SJ)	0.074	0.376	6.184	46.879	0.000	4.543	281965.718	77.473	116.710	316.65
	MSNR-SJ	0.299	3.047	22.560	625.270	0.000	100.773	52808995.417	480.992	641.747	1364.41
	$MSNR\text{-}SJ^{1/2}$	0.098	0.538	8.576	105.621	0.000	10.039	148626.054	161.191	225.555	568.80
	ln (1+MSNR-SJ)	0.070	0.365	6.396	50.508	0.000	4.623	328395.034	88.303	128.442	337.79

Key: The sample of the period 11 March 1996 to 30 September 2009, there are total 3281 Daily observations. For continuous sample path variation and significant jump measures, we set$\alpha = 0.999$. The 5% critical values for Jarque-Bera (i.e.,$\chi^2(k)$) and LB (k) are 5.991 (k=2), 11.070 (5), 18.924 (10) and 33.924 (22) respectively

Table 2a. The R^2-Values for HAR, SVM-HAR, HAR-J, SVM-HAR-J, HAR-MSNR-J and SVM-HAR-MSNR-J models

Horizon Day(s)			1						5						22					
Model			HAR	SVM-HAR	HAR		SVM-HAR		HAR	SVM-HAR	HAR		SVM-HAR		HAR	SVM-HAR	HAR		SVM-HAR	
					Jump	MNRJ	Jump	MNRJ			Jump	MNRJ	Jump	MNRJ			Jump	MNRJ	Jump	MNRJ
In-Sample	5	RV_t	0.353	**0.572**	0.359	0.353	0.573	**0.582**	0.206	**0.431**	0.210	0.206	0.444	**0.451**	0.060	**0.301**	0.060	0.060	**0.310**	0.300
		$RV_t^{1/2}$	0.453	**0.653**	0.456	0.454	**0.663**	0.648	0.299	**0.505**	0.300	0.299	**0.520**	0.514	0.106	**0.384**	0.106	0.106	**0.382**	0.373
		$lnRV_t$	0.482	**0.644**	0.483	0.483	0.652	**0.662**	0.342	**0.547**	0.343	0.343	**0.534**	**0.534**	0.160	**0.459**	0.161	0.163	0.444	**0.458**
	15	RV_t	0.326	**0.550**	0.326	0.326	**0.563**	0.554	0.186	**0.433**	0.189	0.190	**0.445**	0.443	0.061	**0.275**	0.062	0.062	**0.230**	0.292
		$RV_t^{1/2}$	0.396	**0.610**	0.396	0.396	**0.604**	0.600	0.262	**0.469**	0.265	0.266	**0.491**	0.485	0.094	**0.354**	0.096	0.094	0.338	**0.339**
		$lnRV_t$	0.414	**0.605**	0.414	0.414	**0.600**	0.586	0.297	**0.514**	0.298	0.299	**0.507**	0.499	0.137	**0.440**	0.141	0.139	**0.410**	0.405
	Opt	RV_t	0.262	**0.489**	0.263	0.262	**0.504**	0.511	0.149	**0.400**	0.153	0.156	0.401	**0.402**	0.045	**0.251**	0.048	0.045	**0.286**	0.267
		$RV_t^{1/2}$	0.325	**0.540**	0.325	0.325	0.539	**0.550**	0.215	**0.445**	0.219	0.218	**0.454**	0.452	0.074	**0.305**	0.075	0.074	**0.304**	0.298
		$lnRV_t$	0.323	**0.498**	0.324	0.324	**0.521**	0.518	0.233	**0.449**	0.234	0.234	**0.456**	0.443	0.105	**0.336**	0.107	0.106	0.339	**0.361**
Out-of-sample	5	RV_t	**0.598**	0.597	**0.610**	0.598	0.605	0.603	**0.391**	0.388	0.379	**0.390**	0.387	0.387	**0.106**	0.093	**0.105**	**0.107**	0.091	0.090
		$RV_t^{1/2}$	0.732	**0.734**	0.735	0.735	**0.738**	0.737	0.567	**0.575**	0.565	0.567	**0.575**	**0.575**	0.250	**0.263**	0.251	**0.263**	**0.263**	0.262
		$lnRV_t$	0.743	**0.747**	0.745	0.745	0.747	**0.748**	0.604	**0.610**	0.606	0.610	**0.612**	**0.612**	**0.357**	0.352	0.363	0.362	**0.372**	0.368
	15	RV_t	**0.527**	0.507	**0.527**	0.526	0.507	0.512	**0.344**	0.333	0.338	**0.345**	0.330	0.334	**0.086**	0.064	**0.083**	0.081	0.064	0.060
		$RV_t^{1/2}$	**0.643**	**0.643**	0.645	0.645	0.645	**0.647**	0.507	**0.512**	0.505	0.505	**0.512**	0.511	0.203	**0.214**	-ve	0.201	0.209	0.213
		$lnRV_t$	0.624	**0.631**	0.626	0.625	0.630	**0.632**	0.508	**0.521**	0.511	0.511	**0.522**	0.520	0.275	**0.294**	0.287	0.283	**0.297**	0.297
	Opt	RV_t	0.393	0.374	0.390	0.393	0.373	0.373	**0.279**	0.247	0.274	**0.271**	0.245	0.245	**0.070**	0.039	**0.061**	0.067	0.038	0.038
		$RV_t^{1/2}$	**0.566**	**0.566**	**0.567**	**0.567**	**0.567**	0.565	**0.471**	0.460	0.465	**0.467**	0.459	0.460	**0.197**	0.190	0.194	**0.197**	0.190	0.188
		$lnRV_t$	0.552	**0.553**	0.554	0.554	**0.555**	0.553	0.464	**0.464**	0.465	0.465	**0.466**	0.465	**0.267**	0.252	**0.275**	0.272	0.257	0.252

Key: The sample of the period 11 March 1996 to 30 September 2009, there are total 3279 Daily observations. The Table reports the R^2-Values those have been calculated for daily (h=1), weekly (h=5) and monthly (h=22) horizons. The out-of-sample R^2 value of HAR-RV-J model of 22-day ahead horizon for standard deviation of RV series observed negative. This implies that HAR-RV-J model is not appropriate model for those data sets.

Table 2b. The R^2-Values for HAR-CJ, SVM-HAR-CJ, HAR-MNR-CJ and SVM-HAR-MNR-CJ models for different horizons

Horizon Day(s)			1				5				22			
Model			HAR		SVM-HAR		HAR		SVM-HAR		HAR		SVM-HAR	
			CJ	MNR-CJ	CJ	MNR-CJ	CJ	MNR-CJ	CJ	MNR-CJ	CJ	MNR-CJ	CJ	MNR-CJ
In-Sample	5	RV_t	0.354	0.354	**0.593**	0.584	0.212	0.213	0.446	**0.455**	0.071	0.069	0.307	**0.307**
		$RV_t^{1/2}$	0.457	0.456	**0.637**	0.630	0.305	0.304	0.511	**0.528**	0.115	0.113	0.371	**0.375**
		$lnRV_t$	0.480	0.481	0.641	**0.646**	0.341	0.341	**0.554**	0.549	0.152	0.152	0.445	**0.445**
	15	RV_t	0.329	0.328	0.540	**0.541**	0.062	0.062	0.260	**0.282**	0.062	0.062	0.260	**0.270**
		$RV_t^{1/2}$	0.399	0.399	0.544	**0.561**	0.266	0.266	**0.445**	0.441	0.097	0.096	**0.280**	0.272
		$lnRV_t$	0.413	0.412	**0.567**	0.561	0.301	0.299	**0.487**	0.480	0.140	0.139	**0.383**	0.378
	Opt	RV_t	0.269	0.269	**0.482**	0.478	0.158	0.158	0.376	**0.385**	0.042	0.048	0.285	**0.273**
		$RV_t^{1/2}$	0.269	0.270	0.473	**0.472**	0.224	0.224	0.419	**0.435**	0.077	0.077	**0.293**	0.316
		$lnRV_t$	0.324	0.325	0.520	**0.522**	0.236	0.237	0.465	**0.466**	0.106	0.107	0.379	**0.384**
Out-of-sample	5	RV_t	0.583	0.589	**0.594**	0.595	0.340	0.342	0.378	**0.379**	0.068	0.076	0.062	**0.077**
		$RV_t^{1/2}$	0.729	**0.731**	**0.731**	0.730	0.544	0.547	0.561	**0.564**	0.222	0.232	0.232	**0.244**
		$lnRV_t$	0.715	0.714	**0.723**	0.720	0.566	0.565	**0.577**	0.575	0.308	0.310	0.306	**0.312**
	15	RV_t	**0.519**	0.518	0.504	0.504	**0.079**	0.077	0.062	0.060	**0.079**	0.077	0.062	0.060
		$RV_t^{1/2}$	0.641	0.643	0.643	**0.645**	0.502	0.499	0.511	**0.519**	0.193	0.189	**0.205**	0.203
		$lnRV_t$	0.602	0.603	0.611	**0.614**	0.488	0.484	**0.501**	0.500	0.249	0.246	**0.268**	0.265
	Opt	RV_t	0.384	**0.385**	0.359	0.357	**0.205**	0.201	0.195	0.191	0.051	**0.047**	0.028	0.026
		$RV_t^{1/2}$	0.384	**0.385**	0.359	0.357	**0.423**	0.422	0.415	0.412	0.187	**0.188**	0.176	0.175
		$lnRV_t$	0.547	**0.548**	**0.548**	0.547	0.454	0.454	0.453	**0.455**	0.264	**0.266**	0.248	0.248

Key: The sample of the period 11 March 1996 to 30 September 2009, there are total 3279 Daily observations. The Table reports the R^2-Values those have been calculated for daily (h=1), weekly (h=5) and monthly (h=22) horizons.

Table 3a. Forecasting Errors of HAR, SVM-HAR, HAR-J, SVM-HAR-J, HAR-MSNR-J and SVM-HAR-MSNR-J models for the logarithmic series

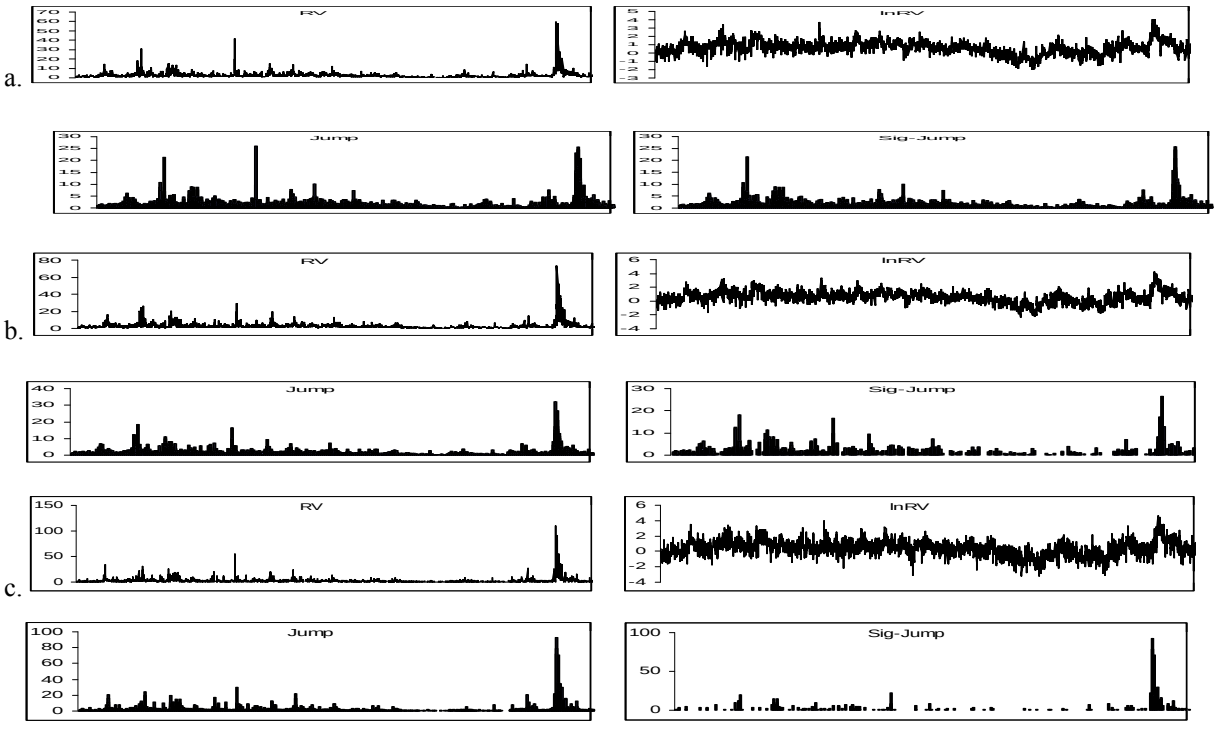
Horizon Day(s)			1						5						22					
Model			HAR	SVM-HAR	HAR		SVM-HAR		HAR	SVM-HAR	HAR		SVM-HAR		HAR	SVM-HAR	HAR		SVM-HAR	
					Jump	MNRJ	Jump	MNRJ			Jump	MNRJ	Jump	MNRJ			Jump	MNRJ	Jump	MNRJ
In-Sample	5	RMSE	1.837	**1.486**	1.828	1.837	**1.473**	1.482	0.548	**0.459**	0.547	0.547	**0.447**	0.460	0.626	**0.505**	0.626	0.625	**0.498**	0.501
		MAE	1.032	**0.609**	1.030	1.032	**0.599**	0.616	0.421	**0320**	0.421	0.421	**0.308**	0.323	0.450	**0.361**	0.489	0.489	**0.354**	0.358
		RMSPE	0.764	**0.390**	0.753	0.764	**0.379**	0.403	14.40	**9.670**	13.50	14.35	**9.131**	9.816	**16.78**	22.88	15.77	**15.46**	18.27	20.61
		MAPE	0.702	**0.495**	0.699	0.702	**0.489**	0.499	1.446	**0.566**	1.435	1.445	**0.559**	0.578	1.568	**0.923**	1.554	1.549	**0.852**	0.883
	15	RMSE	0.589	**0.489**	0.589	0.589	**0.485**	0.486	0.646	**0.536**	0.646	0.646	**0.539**	0.540	0.722	**0.597**	0.720	0.721	0.600	**0.590**
		MAE	0.461	**0.345**	0.461	0.461	**0.345**	**0.345**	0.502	**0.374**	0.502	0.502	**0.379**	0.380	0.567	**0.428**	0.566	0.567	0.431	**0.419**
		RMSPE	35.25	**13.09**	35.6	35.28	15.04	**13.23**	36.18	**14.58**	36.58	36.51	**15.31**	16.6	33.01	**17.68**	32.71	32.64	17.76	**16.55**
		MAPE	1.800	**0.587**	1.806	1.803	**0.561**	0.595	1.879	**0.716**	1.890	1.890	0.729	**0.640**	1.934	**0.576**	1.932	1.933	0.618	**0.604**
	Opt	RMSE	0.711	**0.621**	0.721	0.721	**0.606**	0.607	0.769	**0.652**	0.765	0.769	**0.648**	0.655	0.835	**0.719**	0.834	0.835	0.718	**0.706**
		MAE	0.562	**0.446**	0.562	0.562	**0.427**	0.432	0.596	**0.463**	0.596	0.596	**0.460**	0.467	0.650	**0.517**	0.649	0.650	0.517	**0.503**
		RMSPE	15.84	**12.18**	15.71	15.44	**12.54**	12.93	20.42	**19.79**	20.8	20.65	**15.71**	15.06	20.33	**23.40**	**19.59**	19.79	27.46	24.29
		MAPE	1.595	**0.669**	1.594	1.589	**0.594**	0.608	1.701	**0.418**	1.716	1.704	0.655	**0.631**	1.767	**---**	1.764	1.763	---	**----**
Out-of-sample	5	RMSE	3.001	**2.975**	**2.957**	3.000	2.975	2.983	0.620	**0.615**	0.618	0.618	0.614	**0.613**	0.787	**0.782**	0.783	0.784	**0.778**	0.780
		MAE	1.118	**1.026**	1.122	1.118	1.026	**1.024**	0.489	**0.485**	0.488	0.488	0.483	**0.483**	0.605	**0.599**	0.602	0.603	0.595	**0.597**
		RMSPE	0.979	**0.738**	1.014	0.980	0.738	**0.718**	**8.355**	8.693	8.738	**8.560**	9.013	8.937	6.333	**6.312**	6.582	6.684	6.524	**6.510**
		MAPE	0.818	**0.713**	0.833	0.819	0.713	**0.705**	**1.432**	1.437	1.451	1.442	1.451	**1.447**	1.437	**1.422**	1.450	1.457	**1.432**	1.436
	15	RMSE	0.669	**0.662**	0.667	0.667	0.663	**0.661**	0.765	**0.755**	0.763	0.763	0.754	**0.756**	0.929	**0.917**	0.922	0.924	**0.915**	0.915
		MAE	0.525	**0.518**	0.523	0.523	0.518	**0.516**	0.605	**0.595**	0.603	0.604	**0.594**	0.595	0.727	**0.711**	0.728	0.721	**0.708**	0.709
		RMSPE	**15.15**	15.29	**14.81**	15.87	15.35	15.07	8.311	**7.941**	8.463	8.386	8.020	**7.967**	11.41	**11.06**	10.3	10.83	**9.945**	10.39
		MAPE	**1.436**	1.443	**1.435**	1.436	1.443	1.445	1.361	**1.349**	1.370	1.371	1.354	**1.353**	1.455	**1.412**	1.435	1.448	**1.395**	1.403
	Opt	RMSE	**0.807**	**0.807**	0.806	0.805	**0.805**	0.807	0.884	**0.883**	0.883	0.883	**0.881**	o.883	**1.033**	1.044	**1.028**	1.030	1.040	1.041
		MAE	0.637	**0.553**	0.636	0.636	**0.635**	0.637	0.696	**0.695**	0.695	0.695	**0.694**	0.695	**0.806**	0.816	**0.800**	0.802	0.812	0.813
		RMSPE	14.46	**14.02**	14.10	14.02	**13.99**	14.55	**4.880**	4.997	5.096	4.997	4.949	**4.895**	8.074	**07.19**	8.830	8.429	7.828	**7.638**
		MAPE	1.472	**1.471**	1.472	**1.471**	1.483	1.485	**1.319**	1.325	1.331	1.325	**1.320**	**1.318**	1.389	**1.366**	1.407	1.395	1.387	**1.377**

Key: The sample of the period 11 March 1996 to 30 September 2009, there are total 3279 Daily Observations.

Table 3b. Forecasting Errors of HAR-CJ, SVM-HAR-CJ, HAR-MNR-CJ and SVM-HAR-MNR-CJ models for different horizons for the logarithmic series

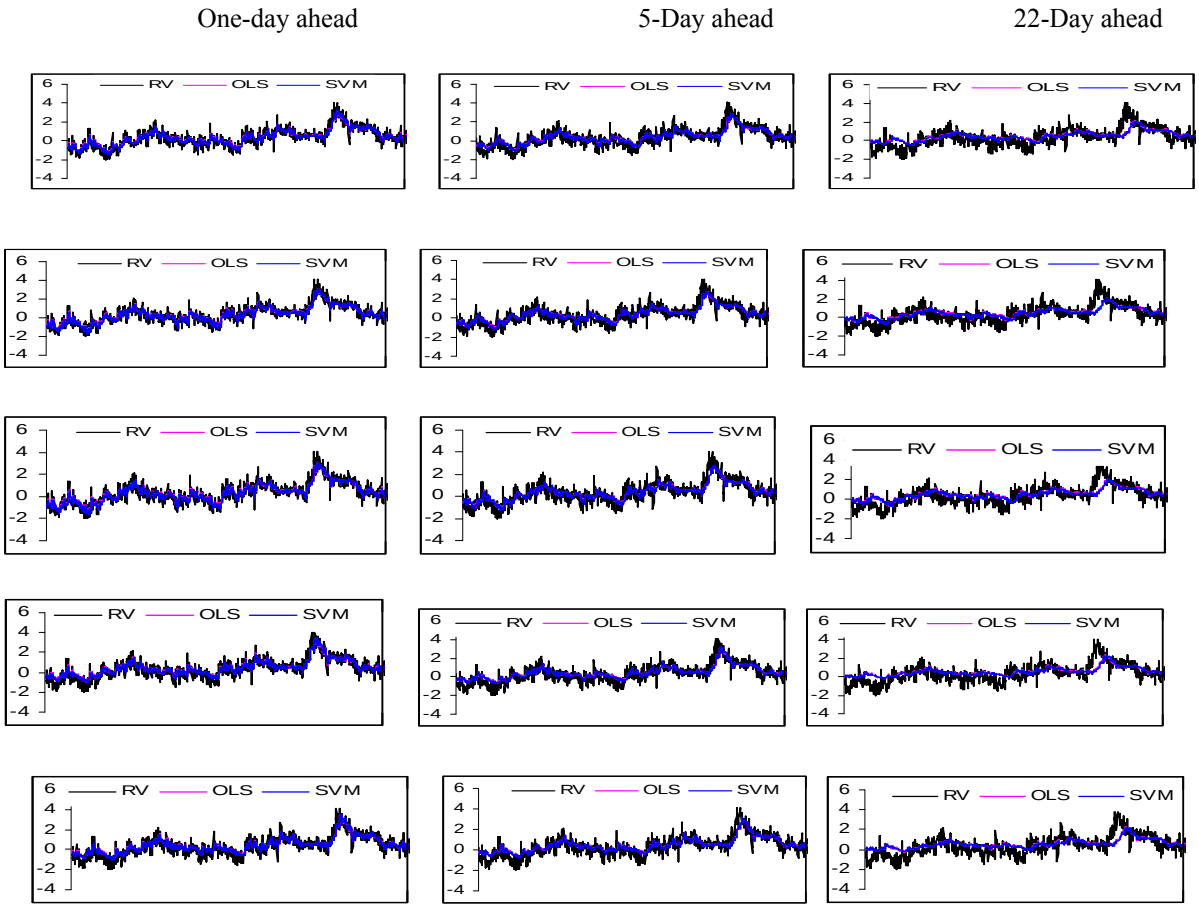
Horizon Day(s)			1				5				22			
Model			**HAR**		**SVM-HAR**		**HAR**		**SVM-HAR**		**HAR**		**SVM-HAR**	
			CJ	MNR-CJ	CJ	MNR-CJ	CJ	MNR-CJ	CJ	MNR-CJ	CJ	MNR-CJ	CJ	MNR-CJ
In-Sample	**5**	**RMSE**	0.485	0.485	0.403	**0.401**	0.548	0.548	**0.451**	0.454	0.629	0.629	**0.509**	0.509
		MAE	0.377	0.376	0.289	**0.287**	0.421	0.422	**0.317**	0.321	0.491	0.492	**0.365**	0.365
		RMSPE	**8.442**	8.974	23.583	25.214	**9.995**	10.438	11.487	11.468	**20.663**	21.570	30.927	34.853
		MAPE	1.276	1.292	**0.846**	0.874	1.356	1.377	0.646	**0.645**	1.613	1.620	**1.004**	1.048
	15	**RMSE**	0.590	0.590	**0.507**	0.511	0.645	0.645	**0.552**	0.556	0.720	0.721	**0.610**	0.612
		MAE	0.460	0.461	**0.367**	0.372	0.500	0.500	**0.398**	0.402	0.567	0.567	**0.446**	0.450
		RMSPE	36.591	36.721	35.753	**34.950**	36.448	36.366	35.447	**34.055**	35.198	34.422	23.406	**15.450**
		MAPE	1.811	1.810	----	----	1.882	1.878	---	---	1.950	1.943	0.268	**0.542**
	Opt	**RMSE**	0.721	0.720	0.607	**0.606**	0.768	0.767	0.642	**0.641**	0.834	0.834	0.696	**0.693**
		MAE	0.561	0.560	0.429	**0.428**	0.594	0.594	0.457	**0.456**	0.650	0.650	0.500	**0.496**
		RMSPE	16.518	16.459	13.032	**12.603**	19.615	**19.347**	26.736	27.105	**17.621**	17.781	29.809	29.953
		MAPE	**1.601**	**1.606**	**0.620**	0.633	1.683	**1.681**	-----	**----**	1.723	1.727	**0.023**	----
Out-of-sample	**5**	**RMSE**	0.526	0.527	0.519	**0.521**	0.649	0.650	**0.641**	0.642	0.816	0.815	0.817	**0.814**
		MAE	0.415	0.415	0.409	**0.410**	0.515	0.515	**0.507**	0.508	0.632	0.632	0.628	**0.627**
		RMSPE	10.542	10.238	9.946	**9.730**	6.603	**6.485**	6.717	6.560	6.144	6.198	**5.749**	5.814
		MAPE	1.353	1.346	1.333	**1.329**	1.385	1.383	1.374	**1.371**	1.451	1.458	**1.419**	1.424
	15	**RMSE**	0.688	0.687	0.680	**0.678**	0.781	0.784	**0.771**	0.773	0.946	0.947	**0.934**	0.936
		MAE	0.545	0.544	0.536	**0.534**	0.619	0.621	**0.609**	0.611	0.741	0.742	**0.727**	0.728
		RMSPE	13.301	12.675	13.722	**13.077**	8.048	7.970	**7.482**	7.482	10.902	10.980	**10.720**	10.732
		MAPE	1.372	1.367	1.377	**1.372**	1.353	1.355	**1.325**	1.329	1.474	1.476	**1.421**	1.424
	Opt	**RMSE**	0.812	**0.811**	**0.811**	**0.811**	0.892	0.892	0.893	**0.891**	1.035	**1.034**	1.046	1.046
		MAE	0.643	**0.642**	0.643	0.643	0.702	0.703	**0.703**	0.703	0.805	**0.804**	0.815	0.815
		RMSPE	16.454	16.749	**15.194**	16.041	4.284	4.322	**4.166**	4.256	7.741	7.710	6.695	**6.585**
		MAPE	1.490	1.493	**1.486**	1.494	1.280	1.283	1.268	**1.032**	1.342	1.340	1.322	**1.318**

Key: The sample of the period 11 March 1996 to 30 September 2009, there are total 3279 Daily observations. The Table reports the $\boldsymbol{R^2}$-Values those have been calculated for daily (h=1), weekly (h=5) and monthly (h=22) horizons.



Key: Panel-a, Panel-b, Panel-c, and Panel-d show the daily RV, log-RV, Jump and Significance Jump series of 5-min, 10-min, 15-min and optimally sampled intraday return data. The significant jumps have been calculated using a cutoff value α= 0.999.

Figure 1. Realized Volatility, Log-Realized Volatility, Jumps and Significant Jumps series.



The first panels show the HAR-RV/SVM-HAR-RV, the second is HAR-RV-J/SVM-HAR-RV-J, the third is HAR-RV-MSNR-J/SVM-HAR-RV-MSNR-J, the fourth is HAR-RV-CJ/SVM-HAR-RV-CJ and the fifth is HAR-RV-MSNR-CJ/SVM-HAR-RV-MSNR-CJ model's out-of-sample forecasts for the logarithmic transformed series.

Figure 2. Daily, Weekly and Monthly out-of-sample realized volatility forecasts from HAR-RV, SMV-HAR-RV R, HAR-RV-J and SVM-HAR-RV-J models

ISSN 1916-971X E-ISSN 1916-9728