The Academy of Economic Studies
The Faculty of Finance, Insurance, Banking and Stock Exchange
Doctoral School of Finance and Banking

High Frequency Data in Modeling and Forecasting Volatility

Dissertation paper

Student: Varvaroi Maria

Coordinator Professor PhD Moisă Altăr

7/10/2009

Table of Contents

Abstract

Until recently, the most popular ways to model and to forecast the volatility was through the use of ARCH-type models that treat volatility as a latent variable. The present paper follows the new wave emerged in the volatility-related research which highlights the great potential embodied in the use of high frequency data. This new approach proposes the construction of ex post volatility measures that allow us to treat volatility as an *observable* variable. The idea behind this new concept is that volatility can be approximated arbitrarily well by summing intradaily returns sampled at an ever higher frequency. The Realized Volatility (this name was first time used in (Andersen T.G.and T. Bollerslev , 1998) allows us to model and to forecast the volatility directly thus bringing significant increase both in fitting and in forecasting performance.

An important issue related to high-frequency data is the contamination with microstructure noise. This paper adopts two approaches to tackle this problem: first, the sampling frequency is chosen such that the effect of contamination is benign, second-the use of corrected measures, the so-called Realized Kernels, proposed by (Barndorff-Nielsen,O.E.,P. R. Hansen, A. Lunde, and N. Shephard, 2006) with three different kernel weight functions: Bartlett, Parzen and Tukey-Hanning kernels. Thus a total of four volatility proxies are used in estimation and forecasting.

The models considered are the HAR-RV model proposed by (Corsi, 2003) and different modifications of the basic setup, built by considering different explanatory variables, yielding a total of 5 models. The models and the proxies are compared, using the Euro/USD currency pair. The conclusion is that the corrected measures bring little or no improvements as concerned the fitting performance. This is due to the fact that at the considered sampling frequency the contamination with microstructure noise is almost negligible. However the forecasts are more accurate for the corrected measures, showing that as little as it is, the influence of microstructure noise still has to be taken account for. The conclusion is that the improved forecasting performance provides a sufficient incentive to use the corrected measures, even with some loss of simplicity.

Introduction

Given the rapid growth in financial markets and the continual development of new and more complex financial instruments, there is an ever-growing need for theoretical and empirical knowledge of the volatility in financial time series. Measurement and forecasting volatility of asset prices in general and of exchange rates in particular has been a focus for researchers. But despite its importance, volatility is still an ambiguous term for which there is no unique, universally accepted, precise definition. Most frequently the volatility is defined as an indicator of the size of price movements. An interesting approach is found in (Dacorgogna,M.M., R. Gencay,U.A.Muller,R.B. Olsen,and O.V. Pictet, 2001) where the volatility is described as being "the visible "footprint" of less observable variables such as market presence and market volume".

Once one recognizes the importance of the volatility, the main issue arises: how to model and, most of all, to predict the volatility. Given its latent character, a common approach to deal with volatility is to conduct inference through strong parametric assumptions, considering, e.g. a GARCH-type model or a stochastic volatility model. The drawback of these approaches is the poor out-of-sample forecasting performance, despite the good fitting performance. An alternative approach is to rely on historical volatility measure that utilizes a backward looking rolling window of sample return volatility. The drawback of this specific alternative is its lack of dynamic updates and the equal weights that it assigns to all observations in the sample. And finally, the last procedure is to employ the existing derivatives and applying an options pricing model, to extract, using numerical procedures, the implied volatility, which is considered an unbiased forecast of the volatility. This approach has, also, some pitfalls, since it is model dependent.

The availability of high-frequency data gives us a new tool to study the volatility. The idea of using intra daily data was first introduced by Merton (1983), who noted that the volatility of a Brownian motion can be approximated to an arbitrary precision using the sum of intraday squared returns. The vast literature of the last two decades has documented important improvements in modeling and forecasting volatility via use of novel volatility proxies constructed from high-frequency data. The advantage of converting the volatility from a latent variable to an observable one lies in the fact that it allows to directly fit econometric models rather than using the more complicated GARCH-type models, required when volatility is latent.

(Andersen T.G.and T. Bollerslev , 1998) proposes the Realized Volatility as an ex post volatility measure, and it can be regarded as a seminal paper on using high-frequency data in volatility forecasting, showing the big improvement of the forecast performance using novel proxies compared to a daily GARCH model.

One of the issues related to measures of volatility using high frequency data is the microstructure noise that biases these measures. This noise is due to imperfections of the trading process e.g. bid-ask bounce, discreteness of price changes etc. There are procedures to undertake in order to mitigate the effect of the microstructure noise. One is sampling at an optimal frequency chosen through the analysis of the volatility signature plot, see (Andersen, T.G.,T. Bollerslev, F. X. Diebold and P. Labys, 1999) or other, more sophisticated methods, see (Zhang, L.,P.A.Mykland and Y. Ait-Sahalia, 2005) and (Bandi, F.M. and J. R. Russel, 2003) or staggered sampling (skip one sampling). Another is construction of new volatility proxies as proposed by (Zhang, L.,P.A.Mykland and Y. Ait-Sahalia, 2005), (Barndorff-Nielsen,O.E.,P. R. Hansen, A. Lunde, and N. Shephard, 2006) and (Hansen P.R. and A. Lunde, 2006).

The purpose of this paper is to evaluate the out-of-sample forecasting performance of various models for realized volatility of the Euro/USD exchange rate, all related to the HAR-RV model proposed by (Corsi, 2003). We use four volatility measures: a naïve estimator, the Realized Volatility, and three measures corrected for microstructure noise: realized kernels using Bartlett kernel, Parzen kernel and Tukey-Hanning kernel. The choice of this particular exchange rate is determined by two reasons: the high liquidity and the small number of studies based on this pair of currencies. The considered models are: First, the basic HAR-RV-RV model, proposed by (Corsi, 2003), then the realized variance is decomposed into its continuous sample path and jump components and each of these terms are taken as explanatory variable in the RV regression, yielding the HAR-RV-C and HAR-RV-CJ models. Finally, following (Forsberg,L.and E.Ghysels, 2006), a new explanatory variable is added based on its appealing properties: the realized power variation – RPV (p). Two cases are considered: p=1.3 and p=1.5, rendering other two models: HAR-RV-RPV (p). The models are compared both for their in-sample fitting performance and their out-of-sample forecasting accuracy.

The paper is organized as follows: section 1 provides a literature review; section 2 covers the theoretical framework followed by section 3 that introduces the models that are estimated and gives the tools to assess the forecasting performance; section 4 describes the data, the methodology employed to compute the measures used throughout the paper and subsection 4.3 provides description of the corrected measures. Section 5 discusses the properties and the stylized facts of intradaily returns, daily returns, realized volatility and jumps; the results of the estimations are presented in section 6, and section 7 concludes.

Literature Review

Early reference to the use of high-frequency data are found in Merton, 1983, who noted that the integrated volatility of a Brownian motion can be approximated arbitrarily well using the sum of intradaily squared returns sampled at an ever increasing frequency. However, given the lack of data, the research focused on using daily squared returns as measures of volatility. GARCH-type models were the most popular ones and were used to model and forecast volatility considering it a latent variable, but these models performed poorly out-of-sample. With the availability of high-frequency data, new volatility proxies can be computed, and, as shown in (Andersen T.G.and T. Bollerslev , 1998), these proxies improve significantly the out-of-sample forecasting performance of GARCH-type models and, in addition, turn the volatility into an observable variable, thus making possible to model it directly. Since then, a vast literature, focusing on the use of high-frequency data to obtain good volatility forecasts, emerged. An important contribution to the development and understanding of this new field has been brought by (Dacorgogna,M.M., R. Gencay,U.A.Muller,R.B. Olsen,and O.V. Pictet, 2001), who discusses ways of filtering the data, describes its properties, proposes operators to deal with unequal spaced data (the Convolution Operators) and more. Very influential papers are (Barndorff O. E. and N. Shephard, 2002) who define, besides the realized volatility, other measures that circumvent the data complications, while retaining most of the relevant information in the intraday data for measuring, modeling and forecasting volatility and (Barndorff-Nielsen,O.E. and N. Shephard, 2004) who provides theoretical framework to separate the continuous sample path component of the return process from the jump component.

An important issue related to high-frequency data is the noise that affects the true price. Studies showed that ignoring this noise can compromise the work and render sub-optimal results. (Andersen, T.G.,T. Bollerslev, F. X. Diebold and P. Labys, 1999) uses the volatility signature plot to determine the highest frequency at which the noise is negligible. They found that sampling at 5 minutes is sufficiently safe for highly liquid assets. Other, more formal ways to determine the optimal sampling frequencies were proposed by (Bandi, F.M. and J. R. Russel, 2003) and (Zhang L.,P. Mykland and Y. Ait-Sahalia, 2005). But sampling at lower frequencies results in discarding large amount of data. Thus, new ways to deal with the noise have been proposed. One of the first noise corrected measures was proposed by (Zhou, 1996), who incorporates the first order autocovariance of returns obtaining an unbiased estimator of variance with i.i.d. noise. (Zhang, L.,P.A.Mykland and Y. Ait-Sahalia, 2005) suggest using the Two Scales Realized Volatility (TSRV) that combines two RV measures, one computed at the highest and on at a lower sampling frequency. The realized kernel estimates proposed by (Barndorff-Nielsen,O.E.,P. R. Hansen, A. Lunde, and N. Shephard, 2006) provide an equally efficient estimate of the volatility as the TSRV depending on the choice of the kernel weighting function.

An important improvement brought by the realized variance, as shown in (Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001), is that, when realized volatility is used, the distribution of standardized daily return series is almost Gaussian. Moreover, the log-realized volatility is almost Gaussian too. Thus one can use traditional, well documented models with normality assumptions to make inference and to forecast the volatility. An important property of the realized volatility is its high persistency. Thus, even if it is a stationary time series, there is significant evidence of long-memory, which has to be modeled using appropriate specifications, e.g. ARFIMA (p, d, q), d $\in$(0, 0.5), FIGARCH. Recently, (Corsi, 2003) proposed the Heterogeneous Autoregressive model for Realized Volatility (HAR-RV) which is able to capture the long-memory property being, at the same time, easy to estimate and to interpret. Another important alternative to the fractionally integrated models is Mixed Data Sampling (MIDAS) approach proposed and described by E. Ghysels, P. Santa-Clara and R. Valkanov.

Theoretical framework

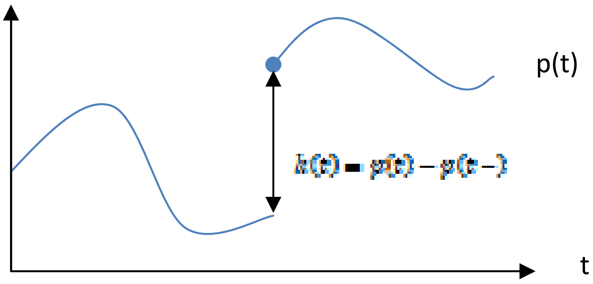
The logarithmic price $p(t) = \log P(t)$ is assumed to evolve in continuous time as a jump diffusion process of the form:

$$dp(t) = \mu(t)dt + \sigma(t)dW(t) + k(t)dq(t) \quad 0 \leq t \leq T \quad (1)$$

where $\mu(t)$ is the instantaneous drift function, $\sigma(t)$ is the instantaneous diffusion function, $W(t)$ is a standard Brownian motion[1], $q(t)$ is a Poisson jump process that describes the arrival of random events and $k(t)$ refers to the size of the jump:

$$k(t) = p(t) - p(t-)$$

where p (t-) denotes the left limit of p(t), that is $\lim_{s \downarrow t} p(s)$.



The quadratic variation of a continuous process is a measure of its volatility. Then, for the return process $r(t) = p(t) - p(0) = dp(t)$ the quadratic variation is:

$$[r,r]_t = \int_0^t \sigma^2(s)ds + \sum_{s>0}^{t} k^2(s) \qquad (2)$$

[1] A random variable W(t) is called a Brownian motion if it satisfies the following properties:

a) W(0)=0
b) W(t) is a continuous function of t
c) W has independent , normally distributed increments (i.e. if $0=t_0<t_1<\ldots t_n$ and $Y_i=(W(t_i)-W(t_{i-1}))$, i=1...n, then i)Y_i are independent
ii)$E(Y_i)=0$, $\forall i$
iii)$var(Y_i)= t_i-t_{i-1}$ (Shreve, 1997)

If the return process does not present any jumps, then the quadratic variation equals the integrated volatility of the continuous sample path component ($IV = \int_0^t \sigma^2(s)ds$).

Definition:

Realized variance - the sum of intradaily squared returns sampled at 1/Δ frequency.

$$RV_{t+1}(\Delta) - \sum_{j=1}^{\frac{1}{\Delta}} r^2_{t+j/\Delta,\Delta} \qquad (3)$$

$$r_{t,\Delta} = p_t - p_{t-\Delta}$$

According to (Andersen, T.G., T. Bollerslev, F. X. Diebold and P. Labys, 2001) “the realized variation converges uniformly in probability to the increment of the quadratic variation process as the sampling frequency increases”[2].

$$RV_{t+1}(\Delta)\lim_{\Delta\to 0}\left(\int_t^{t+1} \sigma^2(s)ds + \sum_{s>t}^{t+1} k^2(s)\right) \qquad (4)$$

In the absence of jumps, the realized volatility is a consistent estimator of the integrated volatility. But in the presence of jumps, a new robust measure is needed. (Barndorff-Nielsen,O.E. and N. Shephard, 2004) proposes the Bi-Power Variation (BV) as a robust measure in the presence of infrequent jumps:

$$BV_{t+1}(\Delta) = \mu_1^{-2} \sum_{i=2}^{1/\Delta} \left|r_{t+i\Delta,\Delta}\right| \left|r_{t+(i-1)\Delta,\Delta}\right| \qquad (5)$$

It is possible to show that:

$$BV_{t+1}(\Delta)\lim_{\Delta\to 0}\left(\int_t^{t+1} \sigma^2(s)ds\right) \qquad (6)$$

Hence, combining the results in (4) and (6) the jump component of the return process can be consistently estimated by:

[2] (Andersen, T.G., T. Bollerslev, F. X. Diebold and P. Labys, 2001)

$$RV_{t+1}(\Delta) - BV_{t+1}(\Delta)\lim_{\Delta\to 0}\sum_{s>t}^{t+1} k^2(s) \qquad (7)$$

The jump component is truncated to zero as follows:

$$J_{t+1}(\Delta) = max[RV_{t+1}(\Delta) - BV_{t+1}(\Delta), 0] \qquad (8)$$

The continuous sample path component is defined as:

$$C_{t+1}(\Delta) = RV_{t+1}(\Delta) - J_{t+1}(\Delta) \qquad (9)$$

Modeling and Forecasting the Realized Volatility

Modeling Realized Volatility

The first model to consider is the basic HAR-RV model (Heterogeneous Autoregressive model for Realized Volatility) proposed by (Corsi, 2003) and further developed and augmented in (Andersen,T.G.,T. Bollerslev and F. X. Diebold, 2007). In order to define the model, it is necessary to calculate the following measurements:

$$RV_{t,t+h} = \frac{1}{h}[RV_{t+1} + RV_{t+2} + \cdots + RV_{t+h}], h = 1, 2, \ldots \qquad (10)$$

$RV_{t,t+h}$ denotes the standardized multi-period realized variance. For h=5, we get the weekly realized volatility, and for h=22 – the monthly measure.

Given these measures, the basic setup of the HAR-RV-RV model is:

$$RV_{t,t+h} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \varepsilon_{t,t+h} \qquad (11)$$

The model can be shown to capture the long memory property of the realized variance being parsimonious and easy to estimate and providing an alternative to more difficult long memory models such as ARFIMA or FIGARCH, see (Corsi, 2003). The underlying assumptions of the model are the Heterogeneous Market Hypothesis[3] and the asymmetric

[3] (Muller, U. A., M. M. Dacorogna, R. D. Dave, O. V. Pictet, R. B. Olsen and J . R. Ward, 14-15 Oct,1993)

propagation of volatility between the short and long time horizons[4]. Thus, the market participants are separated in three categories: short term traders (e.g. market makers, intra-day traders); medium term and long term traders (e.g. portfolio managers, central banks, pension funds). Each category reacts differently to new information available in the market. Long term traders react to changes in volatility at a weekly or monthly time scale and do not respond to short term changes, while short term traders react quickly both to long term and to short term changes. Therefore, in the financial market one can observe an "informational cascade"[5] from long time horizons to short term traders. According to this theory the coefficients of weekly and monthly realized volatility should be larger than the coefficient of daily RV. Also, if the aggregated realized volatilities over different horizons are indeed good proxies for multiperiod variance, then the coefficients could also be interpreted as market component weights[6] .

The HAR-RV-RV is also estimated in the realized volatility and logarithmic realized volatility forms:

$$\sqrt{RV_{t,t+h}} = \beta_0 + \beta_D\sqrt{RV_t} + \beta_W\sqrt{RV_{t-5,t}} + \beta_M\sqrt{RV_{t-22,t}} + \varepsilon_{t,t+h} \quad (12)$$

$$\log\sqrt{RV_{t,t+h}} = \beta_0 + \beta_D \log\sqrt{RV_t} + \beta_W \log\sqrt{RV_{t-5,t}} + \beta_M \log\sqrt{RV_{t-22,t}} + \varepsilon_{t,t+h} \quad (13)$$

By decomposing the realized variance into its continuous sample path component and jump component one can account for the influence of the continuous sample path component by estimating, following (Forsberg,L.and E.Ghysels, 2006), the HAR-RV-C model, in realized variance, realized volatility and logarithmic realized volatility forms:

$$RV_{t,t+h} = \beta_0 + \beta_D C_t + \beta_W C_{t-5,t} + \beta_M C_{t-22,t} + \varepsilon_{t,t+h} \quad (14)$$

$$\sqrt{RV_{t,t+h}} = \beta_0 + \beta_D\sqrt{C_t} + \beta_W\sqrt{C_{t-5,t}} + \beta_M\sqrt{C_{t-22,t}} + \varepsilon_{t,t+h} \quad (15)$$

[4] (Zumbach G. and P.Lynch, 2001)

[5] idem

[6] (Corsi, 2003)

$$\log\sqrt{RV_{t,t+h}} = \beta_0 + \beta_D \log\sqrt{C_t} + \beta_W \log\sqrt{C_{t-5,t}} + \beta_M \log\sqrt{C_{t-22,t}} + \varepsilon_{t,t+h} \quad (16)$$

The normalized multiperiod sample path component is determined in a similar manner as the multiperiod RV:

$$C_{t,t+h} = \frac{1}{h}[C_{t+1} + C_{t+2} + \cdots + C_{t+h}], h = 1, 2, \ldots \quad (17)$$

Next, the basic HAR-RV-C model is augmented, by adding the jump component. The jump component is estimated using Z-statistic, relying on α=0,999. Introduction of HAR-RV-CJ model needs the definition of the normalized multi-period jump:

$$J_{t,t+h} = \frac{1}{h}[J_{t+1} + J_{t+2} + \cdots + J_{t+h}], h = 1, 2, \ldots \quad (18)$$

The HAR-RV-CJ model may be expressed as:

$$RV_{t,t+h} = \beta_0 + \beta_{CD} C_t + \beta_{CW} C_{t-5,t} + \beta_{CM} C_{t-22,t}$$

$$+\beta_{JD} J_t + \beta_{JW} J_{t-5,t} + \beta_{JM} J_{t-22,t} + \varepsilon_{t,t+h} \quad (19)$$

The realized volatility form:

$$\sqrt{RV_{t,t+h}} = \beta_0 + \beta_{CD}\sqrt{C_t} + \beta_{CW}\sqrt{C_{t-5,t}} + \beta_{CM}\sqrt{C_{t-22,t}}$$

$$+\beta_{JD}\sqrt{J_t} + \beta_{JW}\sqrt{J_{t-5,t}} + \beta_{JM}\sqrt{J_{t-22,t}} + \varepsilon_{t,t+h} \quad (20)$$

The logarithmic realized volatility form:

$$\log\sqrt{RV_{t,t+h}} = \beta_0 + \beta_{CD} log\sqrt{C_t} + \beta_{CW} log\sqrt{C_{t-5,t}} + \beta_{CM} log\sqrt{C_{t-22,t}}$$

$$+\beta_{JD}\log(\sqrt{J_t} + 1) + \beta_{JW}\log(\sqrt{J_{t-5,t}} + 1) + \beta_{JM}\log(\sqrt{J_{t-22,t}} + 1) + \varepsilon_{t,t+h} \quad (21)$$

According to the asymmetric propagation of volatility, for the HAR-RV-C and for HAR-RV-CJ models the coefficient of the weekly and monthly component should be larger than the coefficient of the daily component.

Following (Forsberg,L.and E.Ghysels, 2006) and (Liu,C. and J. M. Maheu, 2008) a new explanatory variable, the realized power variation, is added. It is defined as:

$$RPV_{t+1}(p) = \mu_p^{-1} \Delta^{1-\frac{p}{2}} \sum_{j=1}^{\frac{1}{\Delta}} \left| r_{t,j} \right|^p \quad p \in (0,2) \quad (22)$$

For p=2 the RPV (2) equals RV. The reasons of including this measure in the RV regression are:

- The absolute returns show a higher degree of persistence than the squared returns
- The measures based on absolute returns are less severe downward biased that those based on squared returns
- The absolute returns are immune (asymptotically) to the presence of jumps[7]

The estimated models are:

$$RV_{t,t+h} = \beta_0 + \beta_D RPV_t(p) + \beta_W RPV_{t-5,t}(p) + \beta_M RPV_{t-22,t}(p) + \varepsilon_{t,t+h}, \quad (23)$$

$$p = 1.3, 1.5$$

As pointed in (Forsberg,L.and E.Ghysels, 2006) the realized volatility is comparable to the RPV, thus one should not take the square root of RPV when estimating the realized volatility or the logarithmic realized volatility form. Thus, the non linear models are:

$$\sqrt{RV_{t,t+h}} = \beta_0 + \beta_D RPV_t(p) + \beta_W RPV_{t-5,t}(p) + \beta_M RPV_{t-22,t}(p) + \varepsilon_{t,t+h}, \quad (24)$$

$$\log\sqrt{RV_{t,t+h}} = \beta_0 + \beta_D \log RPV_t(p) + \beta_W \log RPV_{t-5,t}(p) + \beta_M \log RPV_{t-22,t}(p) + \varepsilon_{t,t+h}, \quad (25)$$

Considering all the terms entering the equations above are observable, the models are estimated using standard OLS method. The first estimation yields heteroskedastic errors with significant autocorrelation up to 20th lag, thus the Newey-West HAC estimator of the variance-covariance matrix is used with a number of lag equal to 20.

[7] (Forsberg,L.and E.Ghysels, 2006)

The out-of-sample forecast is performed as follows: the models are daily reestimated on a moving window of 1320 days yielding 520 daily forecasts.

Evaluating Alternative Volatility Forecasts

The procedures used to assess the performance of the forecasts are:

- A Mincer-Zarnowitz[8] type regression of the form:

$$v_{t+1} = \alpha_0 + \alpha_1 v^{*}_{t+1_{Model\ i}}, i = 1{:}15 \quad (26)$$

where the actual realized volatility is denoted by v_{t+1}, and the volatility forecasted by a certain model is denoted by $v^{*}_{t+1_{Model\ i}}$.

For each MZ type regression we test the joint hypothesis (using Wald coefficient test)

$H0{:}\ \alpha_0 = 0\ and\ \alpha_1 = 1$

- The loss function:

$$MSE = n^{-1}\sum_{t=1}^{n}\left(RV_t - \widetilde{RV}_t\right)^2 \quad (27)$$ - mean squared error

Where RV_t denotes the true value of the realized variance and $\widetilde{RV}_t$- the predicted value of the dependent variable. The choice of the MSE as the loss function used to assess the forecasting performance is largely explained in (Forsberg,L.and E.Ghysels, 2006). They also point to the fact that MSE obtained from different models where transformation upon the variables were made (such as square root or logarithm) are not comparable. Thus, for the model where the dependent variable is the realized volatility, the MSE is computed as follows:

$$MSE = n^{-1}\sum_{t=1}^{n}\left(RV_t - \sqrt{\widetilde{RV}_t}^{\,2}\right)^2 \quad (28)$$

[8] Mincer,J. and V. Zarnowitz;"The Evaluation of Economic Forecasts", NBER paper, 1969

In the case where the logarithmic realized volatility is the dependent variable, the MSE takes the form:

$$\mathrm{MSE} = \mathrm{n}^{-1} \sum_{i=1}^{n} \left(\mathrm{RV}_t - \exp\left(\log \sqrt{\widehat{\mathrm{RV}}_t} \right)^2 \right)^2 \quad (29)$$

These measures are comparable and can be used to judge the performance of different forecasts. However there should be a note of precaution since transforming a logarithmic forecast to level by applying exponential function yields at least suboptimal prediction[9].

Data and Methodology

Data

The empirical investigation is carried out on Eur/USD exchange rate at 5-minute frequency[10]. The choice of these currencies was determined by the liquidity and the depth of the market. Although attempts were made to work with Eur/Ron or USD/Ron, the conclusion was that these data sets were not suited for this kind of study, because of the presence of large gaps in the series that could not be filled by means of interpolation without loss of original properties. Also, an important issue to address when working with high frequency data is the noise contaminating the volatility measure and that is of much greater magnitude in an illiquid market. The explanation could be found in the fact that in a highly liquid market, the bid-ask spread is one hundredth of a percent, while in an illiquid market it can reach 2-3 percent. Due to the big spread, the bid-ask bounce effect creates the false impression of big market moves, when, in fact, the market didn't change. This spurious price change affects the realized volatility, thus making it a biased estimator for the true return variance. This is one of the motives of choosing to work with highly liquid Eur/USD exchange rate, which accounted for about 27% in 2007[11].

[9] (Granger, C. and P. Newbold, 1976)
[10] The data was provided by Olsen and Associates
[11] (Bank for International Settlements, 2007)

The data span from November 1, 2001 – April 30, 2009, that is a total of 600 000 raw observations or 2347 days. The data is filtered, so as to account only for significant days considering the number of transactions. That means that there are at least 22 days a month with at least 60 observations per day. We followed (Zhou, 1996) when defining the start and the end of a day, thus we choose 24 hours since 0:00 Greenwich Mean Time (GMT) as a day. The argument provided in the cited paper above is the fact that "0:00 GMT is 9:00 am in Tokyo time and 24:00 GMT is 7:00 pm New York time and this 24 hour period covers most activities of the world market". Although the exchange market is open 24 hours a day, 7 days a week, only small trading volumes can be observed during weekends and holidays. Following (Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001) we remove weekend returns from the sample, i.e. from Friday 21:00 GMT to Sunday 21:00 GMT, and certain inactive trading days associated with holidays: Christmas (December 24,25,26), New Year (December 31 and January 1-2), Good Friday, Easter Monday, Memorial Day, July Fourth, Labor Day and Thanksgiving and the following day. This leaves us 1,862 daily observations in total, of which 1,342 (from November 1, 2001 through April 14, 2008) form the estimation period, and the remaining 520 observations (from March 27, 2008 – April 30, 2009) are used for the out-of-sample forecast evaluation.

Methodology

The actual construction of the returns includes following steps: first, we extract the data for a single day from the file; second: we obtain the 5 – minute returns as the first difference of the logarithmic price: $r_{t,\Delta} = p(t) - p(t - \Delta)$

These returns are labeled, following Andersen, Bollerslev, Diebold and Labys, 2001, $r_{t+\Delta,\Delta}, t = \Delta, 2\Delta, 3\Delta \ldots$ 2 347, where Δ=1/288 (there are a total of 288 intra-daily returns each 24 hours for the sampling frequency = 5 minutes). These returns are saved for further analysis as we proceed to the next step, which is construction of the realized volatilities.

As pointed in the theoretical background, meaningful ex – post volatility proxies may be constructed by cumulating squared intraday returns at an appropriate frequency (i.e. 5 minutes). In particular, based on the series of intraday returns constructed above, the realized volatility for 1 day is defined by:

$$RV_{t+1}(\Delta) = \sum_{i=1}^{1/\Delta} (r_{t+i\Delta,\Delta})^2 \qquad (30)$$

The RVs on Fridays were summed up with the RVs on Sundays, thus maintaining the definition of a day.

Also, it is needed to calculate the standardized bipower variation. As shown in (Barndorff-Nielsen,O.E., Shephard,N., 2004), the difference between the realized variance defined in (1) and the standardized bipower variation consistently estimates the jump component of the return process.

$$BV_{t+1}(\Delta) = \mu_1^{-2} \sum_{i=2}^{1/\Delta} |r_{t+i\Delta,\Delta}| \, |r_{t+(i-1)\Delta,\Delta}| \qquad (31)$$

where $\mu_1 = E(|Z|) = \sqrt{2/\pi}$.

The jump component is calculated as the difference of these two measures defined above, and, since this difference can also be negative, it is necessary to truncate the actual empirical measurement at zero (Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001)):

$$J_{t+1}(\Delta) = \max\left[RV_{t+1}(\Delta) - BV_{t+1}(\Delta), 0\right] \qquad (32)$$

The continuous sample path component :

$$C_{t+1}(\Delta) = RV_{t+1}(\Delta) \quad J_{t+1}(\Delta) \qquad (33)$$

The jumps series computed using equation 3 yields a total of 1696 jumps and contains a large number of close to zero jumps, that can be atributed rather to the discreteness of the return process than to jumps. Therefore one needs a way of discerning the significant jump from the rest of the series. Thus, following (Andersen,T., Bollerslev,T.,Diebold,F.X., 2005), we employ a shrinckage estimator:

$$Z_{t+1}(\Delta) = \Delta^{-\frac{1}{2}} \frac{[RV_{t+1}(\Delta) - BV_{t+1}(\Delta)]RV_{t+1}(\Delta)^{-1}}{\left[(\mu_1^{-4} + 2\mu_1^{-2} - 5)\max\{1, TQ_{t+1}(\Delta)BV_{t+1}(\Delta)^{-2}\}\right]^{\frac{1}{2}}} \qquad (34)$$

This test statistic, under the assumption of no – jumps process, follows a standard normal distribution. The $TQ_{t+1}(\Delta)$ measure is the standardized tri – power quarticity[12]:

$$TQ_{t+1}(\Delta) = \Delta^{-1}\mu_{4/3}^{-3}\sum_{j=3}^{\frac{1}{\Delta}}\left|r_{t+j\Delta,\Delta}\right|^{\frac{4}{3}}\left|r_{t+(j-1)\Delta,\Delta}\right|^{\frac{4}{3}}\left|r_{t+(j-2)\Delta,\Delta}\right|^{\frac{4}{3}} \quad (35)$$

$$\mu_{4/3} = E\left(|Z|^{4/3}\right)$$

The idea is to compare the Z_{t+1} statistic with a threshold value, say Φ_α, determined for a significance level α. Using the shrinkage estimator defined above, the jump component at the confidence level α, $J_{t+1,\alpha}(\Delta)$ takes the form:

$$J_{t+1,\alpha} = I[Z_{t+1}(\Delta) > \Phi_\alpha] * [RV_{t+1}(\Delta) - BV_{t+1}(\Delta)] \quad (36)$$

where I[·] denotes the indicator function. The significance level considered in this paper is α = 0.999.

Adjusting for Market Microstructure Noise

It is generally accepted that, due to microstructure effect that induces autocorrelation in the return process, the simple realized volatility measure defined in 1 is a biased and inconsistent estimator of the true volatility, with a bias that grows linearly with the number of sampled observations. The bias may be removed by adding the first order autocovariance of returns as in (Zhou, 1996), but this estimator is inconsistent. A consistent estimator is proposed by (Zhang L.,P. Mykland and Y. Ait-Sahalia, 2005) who define the Two Scales Realized Volatility by combining two RV, one computed at the highest frequency available and one computed at a lower frequency. This approach however requires a much denser data base that that we were working with. But other equally consistent estimate that can be computed using the data available could be found in the literature. In that sense, following (Barndorff-Nielsen,O.E.,P. R. Hansen, A. Lunde, and N. Shephard, 2006), we use realized kernel estimates, that use kernel weight functions

[12] The tri power quarticity is a robust estimator of the integrated quarticity $\int_t^{t+1}\sigma^2(s)\,ds$

to give weights and add autocovariance terms to the naïve RV defined in 1. The general formulation of the realized kernel is:

$$RV^{q}_{t+1}(\Delta) = \sum_{i=1}^{1/\Delta}(r_{t+i\Delta,\Delta})^2 + 2\sum_{w=1}^{q} k(\frac{w}{q+1}) \sum_{i=1}^{(1/\Delta)-w} r_{\Delta i} r_{\Delta i+w} \quad (37)$$

where q is the bandwidth of k(x) and k(x) is a convenient kernel weight function, that satisfies k(0)=1 and k(1)=0. In addition, kernels that satisfy the smoothness conditions, k’(0)=k’(1)=0 are guaranteed to produce non-negative estimators In this paper we consider the following weight functions:

- Bartlett kernel: $k(x) = 1 - x$
- Parzen kernel: $k(x) = \left\{\begin{matrix} 1-6x+6x^3 & 0 \leq x \leq \frac{1}{2} \\ 2(1-x)^3 & \frac{1}{2} \leq x \leq 1 \\ 0 & x > 1 \end{matrix}\right\}$
- Tukey-Hanning$_2$ kernel: $k(x) = \sin^2(\frac{\pi}{2}(1-x)^2)$

The last two kernel weight functions satisfy the smoothness condition too.

The choice of optimal q (the bandwidth) implies the best tradeoff between the bias and the variance, but it requires the availability of the tick-by-tick data. Thus we rely on previous works of (J. Gatheral and C. A. Oomen, 2009) and (Chaboud,A., B. Chiqoine, E. Hjalmarsson and M. Loretan, March 2008). These papers study the performance of the realized kernels computed using bandwidth ranging from 1 to 30 or higher. The conclusion is that the use of large q allows sampling at very high frequencies (15 to 20 seconds), but at lower frequencies only a minimum bandwidth would improve the measures. Thus q in our case is set to 1. In addition higher value for q induce a greater variance, but do not reduce significantly the noise, therefore the optimal q can be set to one without any loss of optimality.

The bipower – variation and the tri-power quarticity are also biased by the microstructure noise, specifically the bid – ask bounce effect, which determine a first – order autocorrelation. That could be the reason of finding too few jumps. In order to mitigate this effect a simple approach has been adopted: the bi – power variation and the tri – power quarticity measures are calculated based on staggered returns (skip – one returns).

$$BV_{t+1}(\Delta) = \mu_1^{-2}(1-2\Delta)^{-1}\sum_{i=3}^{1/\Delta}\left|r_{t+i\Delta,\Delta}\right|\left|r_{t+(i-2)\Delta,\Delta}\right| \quad (38)$$

$$TQ_{t+1}(\Delta) = \Delta^{-1}\mu_{4/3}^{-3}(1-4\Delta)^{-1}\sum_{j=5}^{\frac{1}{\Delta}}\left|r_{t+j\Delta,\Delta}\right|^{\frac{4}{3}}\left|r_{t+(j-2)\Delta,\Delta}\right|^{\frac{4}{3}}\left|r_{t+(j-4)\Delta,\Delta}\right|^{\frac{4}{3}} \quad (39)$$

Using these new measures we calculate four new Z_{t+1} statistics: one using the naïve estimate of the realized volatility as in eq.30 and the staggered BV and TQ and the following three using the realized kernels and the staggered BV and TQ. Based on this new statistic, we construct four new $J_{t+1,\alpha}(\Delta)$ series.

Properties and stylized facts

1. Intra-daily returns

Figure 1 graphs the 5-minutes returns of the Eur/USD exchange rate and table 1 contains the first 4 moment of its empirical distribution and the p-value of the Ljung-Box statistics computed for the 10th lag. The data are consistent with those reported in the documented literature, that is the series is slightly asymmetric, with Skewness = 0.137184 and mean around zero. The kurtosis is 26.94424 - far greater than the kurtosis of a normally distributed random variable, which is 3. The leptokurtosis is an indicator of fat tails, meaning that an important probabilistic mass is found in the extremes of the distribution, making the 10sigma – events much more probable relative to a normal distribution. The histogram of 5 minutes returns depicted in figure 2 confirms the properties described above. To emphasize the large kurtosis of the series under discussion, the probability distribution function of a normally distributed variable is drawn too. Also, the low p-value of the LB statistics rejects the null of independence of the series. The acf plot shows a significant negative autocorrelation, which will be discussed below among the stylized facts.

The analysis of the data beyond the basic statistics confirms the properties discussed in (Dacorgogna,M.M., R. Gencay,U.A.Muller,R.B. Olsen,and O.V. Pictet, 2001), that is:

- ✓ The higher the sampling frequency, the higher the kurtosis

In order to confirm this stylized fact, we constructed two new price series: one, sampled at 15 minutes and the second sampled at 30 minutes[13]. The kurtosis determined for each series are presented in table 2. As predicted, the 5-minute series present the largest kurtosis and the 30-minute series – the smallest.

✓ The intradaily return series presents first-order negative autocorrelation.

Figure 3 graphs the autocorrelation function. The negative correlation can be observed up to lag 4, but the largest is at the first lag, hence the term (first order autocorrelation). After the fourth lag, the autocorrelations mainly lie within the 95% confidence interval of an independent and identically distributed Gaussian variable. This phenomenon is explained by the tendency of price to reverse, or bid-ask bounce in transaction price changes. When transactions randomly occur at the bid or ask quotes, the price changes in transaction prices can exhibit reversal which produces a negative first-order autocorrelation. This negative autocorrelation appears even if the "true" process for price changes lacks autocorrelation.

✓ The intradaily returns present a U-shape seasonality for every 24 hours

Figure 4 displays the autocorrelation function for the absolute intradaily return series, and we included as many lags as there are 5-minute returns in 4 days. The U-shape pattern can be observed at every 288 lags, confirming the presence of the seasonality with a period of one day. The highest autocorrelation are at the beginning and at the end of the day. The explanation, according to (Dacorogna, M.M., Gencay, R., Muller, U.A., Olsen, R.B., Pictet, O.V., 2001), lies in the overlapping of trading hours of the three major foreign exchange markets: the American, the European and the Asian markets.

2. Daily returns

The daily returns are determined as the logarithmic difference of the close – to – open prices. Table 3 shows the statistics for the daily return, and, as regards unconditional distribution, the series has zero mean and is slightly asymmetric to the right. The kurtosis is greater than 3 indicating fat tails. The Ljung – Box statistics shows no serial correlation in returns, but significant correlation in squared returns, pointing to volatility clustering. Figure 5 shows the daily return series, as well as the squared daily return series.

[13] For the 15-minute series we kept the prices at 0,15,30,45 and discarded the rest; for the 30-minute series only prices at 0 and 30 were kept.

As highlighted in (Andersen,T.G.,Bollerslev,T.,Diebold,F.X.,Labys,P., 2003) the standardized returns are well approximated by the standard normal distribution. The daily returns are standardized using all four volatility proxies, yielding four standardized series. Figure 6 plots the histogram, the empirical density function of the standardized returns and the normal density function for all four series. Table 4 contains the first four moments of the unconditional distribution and the results of the Jarque - Bera test and Kolmogorov Smirnov test statistics. Although the null cannot be accepted, still the series are fairly close to the Gaussian distribution.

3. Realized Volatilities

Table 5 summarizes the parameters of the unconditional distribution of the realized volatility. Again, as in the case of intradaily returns, the results are in line with the findings in the documented literature. The kurtosis and skewness far exceed those of a normally distributed variable, suggesting a severely right - skewed leptokurtotic distribution. Following (Andersen,T.G.,Bollerslev,T.,Diebold,F.X.,Labys,P., 2003) we also studied the distributional properties of the logarithmic return $lRV = {}^{1}/_{2}\, ln\ (RV)$ and found that the distribution of this series is indeed well approximated by the normal distribution. Table 6 presents the summary statistics. The kurtosis is in fact smaller than that of a Gaussian distribution. Figure 7 displays the histogram along with the empirical density function and the normal density and it can be seen that the empirical density conveys to normality. However the results reported in table 6 for the Jarque – Bera and Kolmogorov-Smirnov tests strongly reject the null at any level of confidence. The departure of the empirical distribution from the normal one can be explained by the unusually high level of volatility recorded since 2008 which increased the kurtosis of the series. Indeed, if the normality tests are to be applied on the series discarding the last 500 observation, the p-values increase substantially, as shown in the last two columns of table 6.

Another important feature of the realized volatility is its long memory. Figure 8 plots the autocorrelation function up to lag 100, that is 100 days, and it is evident that the autocorrelation presents a hyperbolic decay which characterizes a long memory process.

The inclusion of the Realized Power Variation as an additional explanatory variable is justified by its significant autocorrelation. It is evident that the RPV(1.3) and the RPV(1.5) display larger sample autocorrelation compared to the RV, as shown in figure 9 which plots the sample autocorrelation function for $RPV_{t+1}(1.3), RPV_{t+1}(1.5)$ and RV_{t+1} up to lag 100 and it confirms the findings described above.

4. Jumps

Figure 10 displays the histogram of Z_{t+1} for each of the four proxies and Φ_α corresponding to α= 0.9, 0.99, 0. 999 respectively. The fact that the histogram is so severely right skewed suggests that there are indeed significant jumps. Figure 11 shows exactly how the shrinkage estimator works, as it displays the initial jump series, the Z-statistics and the final jump series for the naïve RV (for the realized kernels the graphs are similar).

Figure 12 displays the 4 jump series estimated based on the shrinkage estimator Z. The estimation yielded a total of 340 jumps for the naïve estimate of the realized volatility, 378 for the realized Bartlett kernel, 348 for the realized Parzen kernel and 347 for the realized Tukey-Hanning kernel. The basic visual inspection shows that years 2004 and 2008 display jump clustering. An explanation could be found in BIS reports[14] for these years. The 2004 report points to a sharp increase of the trading volume in the exchange rate market, particularly the carry trades, and, in addition, early 2004 recorded the lowest level of the US dollar vis-à-vis the Euro. In fact, 2004 ended a downward trend of the USD that started in 2002 culminating on February 17, when Euro reached its highest level. These factors are known to have a direct impact on volatility, in the sense that a large trading volume often implies an increased volatility, as well as negative returns.

On the other hand, the record high volatility level of 2008 continued the trend started in July 2007, according to the same reference. The Euro/US dollar pair faced a heightened volatility that reached the level of exchange market volatility of September 2001. This pick-up of volatility was accompanied, as in the previous case discussed above, by higher turnover in the foreign exchange spot market, as reported by EBS[15]. The main cause of such increase in volatility, as stressed by the BIS report, was a massive dislocation in major financial markets, as the result of sharp decrease of the attractiveness for the carry trade. Thus, exchange rates involved in these leveraged trades experienced a sharp increase in volatility and a reversal of the previous trend. Also, with the unwinding of these strategies (i.e. carry trades), the focus

[14] BIS 74th and 78th annual reports
http://www.bis.org/publ/arpdf/ar2008e.htm

[15] EBS-Electronic Broking Services, which accounts for about 60% of the spot interbank market

shifted from the interest rate differentials to growth differentials and current account balances as leading indicators. The soaring current account deficit of US and the negative prospects regarding the economy brought a rapid depreciation of the US dollar.

Figure 13 plots the intradaily returns for some specific jump days. For all six days, the Z statistics is well above 10, indicating a highly significant jump. The first largest jump occurred on May 7, 2004 and the timing corresponds to the release of five important economic indicators such as hourly earnings, nonfarm payroll, unemployment rate, wholesale inventories and consumer credit. All indicators, except for the wholesale inventories and the unemployment rate, have a positive relationship with the dollar exchange rate: that is an increase in the hourly earnings or consumer credit gives a positive signal to the market, in the sense that it shows that consumers can afford large expenses, which can fuel economic growth. The wholesale inventories have a different significance: a high inventory suggests that the economy is slowing down, giving a negative sign to the market as a whole, and to the US dollar in particular. The unemployment rate shows the number of unemployed workers divided by the total civilian labor force, and a high figure (or an increase from the last period) indicates a lack of expansion within the economy and have a negative impact on the currency. All five indicators have had a positive impact leading to an appreciation of the US currency. Table 7 displays the figures related to this day and for the rest 5 days.

For January 9 and March 5, 2004 the same indicator have had a different dynamics, thus an opposite impact: the American currency fell in respect to Euro.

January 12, 2005 displays a significant jump, but as a result of two totally different macroeconomic indicators: the trade balance and the treasury budget. Both indicators have had a negative impact on the dollar: the trade balance decreased by almost 60$ bn while the budget deficit soared by 3, 4$ bn. These movements translated into a bearish market, meaning a decrease of the value of the American currency.

The next day displays the reaction of the exchange rate to announcements related to inflation (CPI[16], core CPI[17]) and more interesting, to the data provided by the surveys of worldwide homebuilders, that is housing starts and building permits. The first two indicators account for a majority of overall inflation, which is important because it may lead to a raise in

[16] Consumer Price Index

[17] the Core CPI excludes more volatile items like food and energy

the interest rates, which is the price of money. But the actual figures are below the forecasted ones, having a negative impact on the currency. As for the last two indicators, they are used as business cycle indicators, and, in the context of the crisis, the focus on these figures is explained by the fact that investors want to see signs of recovery. But the negative dynamic cannot be interpreted as a sign of upward sloping economic trend, thus the devaluation of the American dollar.

The trading session of March 18, 2009 was influenced by the current account deficit, which shrank by less than estimated and also by the FED[18] decision to buy 1,2$ trillion worth of government bonds and mortgage related securities. These announcements influenced in a negative way the exchange rate, as the current account deficit decreased by less than expected and the new buying is perceived as printing more money. The result was a devaluation of the American dollar, as shown in figure 13.

The association of highly significant jump with identifiable macroeconomic news is in direct line with the evidence in (Andersen, T.G., Bollerslev,T., Diebold, F.X. , 2007) that document the link between significant price moves and macroeconomic announcements.

Table 8 displays basic summary statistics for the jump series J, the $\sqrt{J}$ series and for the $\log(\sqrt{J+1})$ series. The last column reports the Ljung – Box statistics for up to fifth lag serial correlation and it confirms the lack of autocorrelation in the jumps series, suggesting that jump are very difficult to predict.

Empirical evidence

The models above were estimated using standard OLS, but the residuals presented heteroskedasticity, as shown by the Breusch-Pagan test, and significant autocorrelation up to lag 20. Thus, to obtain consistent estimators, it was necessary to use a heteroskedasticity and autocorrelation consistent (HAC) estimator of the variance-covariance matrix as the one proposed by Newey and West.

[18] Federal Reserve System

Table 9 reports the results of the estimation of HAR-RV-RV models. The realized volatility form performs better in terms of R^2, but the log realized volatility form is superior in terms of MSE. The estimates of the coefficients are significant thus confirming the persistence of the volatility. In addition, the coefficient of the weekly component is the largest one, followed closely by the monthly component, thus confirming the assumption of asymmetric volatility propagation.

Table 10 reports the estimates for the HAR-RV-C models, which, judging by the R^2 performs slightly better than the HAR-RV-RV model, suggesting that there may be some gain in performance by modeling separately the continuous sample path and the jump component. Again the weekly component seems to prevail.

The augmented HAR-RV-CJ model have higher R^2 and lower MSE as compared to the HAR-RV-C model, thus providing evidence that including the jump component may improve fitting performance, as pointed in (Andersen,T.G.,T. Bollerslev and F. X. Diebold, 2007). The coefficients of daily and weekly jump component are significant at 1% level at least, but the monthly coefficient is not significant at any level. The insignificance of the jump component points to the fact that the evolution of realized variance is almost entirely due to the sample path component.

The HAR-RV-RPV model does not outperform the HAR-RV-CJ model, judging both by R^2 and by MSE. But there are evidence that a higher persistence improves fitting performance, since the RPV (1.3) gives better results than the RPV (1.5), thus confirming the results of (Forsberg,L.and E.Ghysels, 2006).

In evaluating the performance of various estimation models the most important issue is its out-of-sample forecasting performance. For initial illustration, figure 14a through 14d plots the forecasts obtained from HAR-RV-C model against the actual realized volatility for all four proxies. The model is picked randomly in order to save space, since the plots are similar for all models. It is evident that the forecasted series closely follows the actual one. More formal ways to assess forecasting performance used in the paper are presented in tables 14-16. Table 14 shows the adjusted R^2 of the Mincer-Zarnowitz regression and the conclusion is that the Bartlett realized kernel performs the worst, the other three measures presenting fairly similar results. Moving to table 15 which presents the result of the Wald test that tests the joint hypothesis $\alpha1=0$ and $\alpha2=1$ one can conclude that neither of the forecasts is optimal. Still the

HAR-RV-CJ model in the realized volatility and log realized volatility forms brings the most satisfactory results, since we cannot reject the null at any level of significance. As concern the MSE of the forecasts presented in table 16 it is clear that the realized kernels bring a significant improvement compared to the naïve realized variance, which shows a MSE sometimes 5 times larger than that of the other three proxies. This result highlights the importance of accounting for the microstructure noise even at such low sampling frequency, considered, otherwise, free of contamination.

Concluding remarks

The main objective of this study is to forecast volatility of Euro/USD exchange rate by using high frequency data. The choice of this currency pair was determined by its high liquidity and depth, and by the fact that, despite the great importance of the Euro in international financial market, most of the studies are focused on US dollar.

The paper considers 5 models, all related to the basic HAR-RV (Heterogeneous AutoRegressive model for Realized Volatility) model proposed by (Corsi, 2003). To explore new possibilities related to high frequency data four different volatility proxies were adopted based on quotes sampled at 5 minute interval. The first one is a naïve estimator, which does not take account of the microstructure noise that contaminates the intradaily returns given the fact that at such low sampling frequency the contamination is benign. The other three are three different realized kernels, namely Bartlett, Parzen and Tukey-Hanning realized kernels, proposed by (Barndorff-Nielsen,O.E.,P. R. Hansen, A. Lunde, and N. Shephard, 2006) as robust estimators of volatility in presence of microstructure noise. The proxies are put to test resulting 15 model estimations for each of them, given the fact that the models were estimated in the realized variance, realized volatility and log realized volatility forms.

The results are mostly in line with previous findings regarding modeling and forecasting volatility obtained using high frequency data. It turns out that the realized variance is most difficult to model, yielding the worst results considering both fitting performance and forecast accuracy. The realized volatility and log realized volatility forms bring rather good results, with an adjusted R^2 well above 0.60 for all considered models.

As concern specific model performance, the HAR-RV-CJ model surpasses all of the other models selected proving that separating the continuous sample path component from the jump component could bring improvement in modeling and forecasting volatility. The addition of new explanatory variable such as Realized Power Variation of order p=1.3 and 1.5, given the persistence properties gives results fairly close to those obtained from modeling only the realized volatility.

Turning to various volatility estimates an interesting conclusion arises. Judging by the fitting performance, it seems that the naïve estimator performs at least as good as the realized kernel estimators, supporting the idea that contamination with microstructure noise at such low frequencies as 5 minutes is benign. But the forecasting accuracy draws a different picture, especially looking at the MSE indicator. Thus, the forecasts based on naïve estimator display a MSE as much as 5 times higher than the MSE of the realized kernel based forecasts. This highlights the importance of correcting for the microstructure noise even at low level sampling frequencies.

Hopefully, this study would bring some contribution to the literature devoted to Euro and inspire some further research, since there are many issues that are to be explored. First, the realized volatility series may display some structural changes suggesting that a model that implies two or more regimes may produce some satisfactorily results. Second, as suggested by (Andersen,T.G.,T. Bollerslev and F. X. Diebold, 2007), improvements may be achieved in forecasting accuracy by modeling separately the continuous sample path component and the jump component. An example in this direction could be found for DEM/USD in (Lanne, 2006) and it would be natural to apply the same approach to the Eur/USD pair. Third, very interesting results could be found if working with tick by tick data, which would make possible to compute other volatility estimates that exploit the whole data base. In addition, a recent study: (Chaboud,A., B. Chiqoine, E. Hjalmarsson and M. Loretan, March 2008) shows that the FX market allows sampling once every 15 to 20 seconds. Thus, the results obtained in this study could be redone using returns sampled at frequencies less than 1 minute. And last, new models developed recently and designed for realized measures, should be tested. In that sense the following paper is to be considered: (Brownlees, C. T. and G. M. Gallo, 2008).

Bibliography

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

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2001). The Distribution of Realized Exchange Rate Volatility. *Journal of the American Statistical Association,96* , 42-55.

Andersen, T.G.,T. Bollerslev, F. X. Diebold and P. Labys. (1999, october). (Understanding,Optimizing,Using and Forecasting) Realized Volatility and Correlation. *Working paper.*

Andersen,T., Bollerslev,T.,Diebold,F.X. (2005). Roughing it Up: Including Jumps Component in the Measurement, Modeling and Forecasting of return Volatility. *NBER working paper 11775* .

Andersen,T.G.,Bollerslev,T.,Diebold,F.X.,Labys,P. (2003). Modeling and Forecasting Realized Volatility. *Econometrica, 71* , 529-626.

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

Bandi, F.M. and J. R. Russel. (2003). Microstructure Noise, Realized Volatility and Optimal Sampling. *Manuscript* , University of Chicago.

Barndorff O. E. and N. Shephard. (2002). Estimating Quadratic Variation using Realized Variance. *Journal of Applied Econometrics, 17* , 457-478.

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

Barndorff-Nielsen,O.E.,P. R. Hansen, A. Lunde, and N. Shephard. (2006). Designing Realized Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise. *Manuscript, Oxford* .

Brownlees, C. T. and G. M. Gallo. (2008). Comparison of Volatility Measures: a Risk Management Perspective.

Chaboud,A., B. Chiqoine, E. Hjalmarsson and M. Loretan. (March 2008). Frequency of Observation and the Estimation of Integrated Volatility in Deep and Liquid Financial Markets. *BIS Working paper no 249* .

Corsi, F. (2003). A Simple Long Memory Model of Realized Volatility". *Manuscript* .

Dacorogna,M.M., R. Gencay,U.A.Muller,R.B. Olsen,and O.V. Pictet. (2001). *An Introduction to High-Frequency Finance.* Academic Press.

Forsberg,L.and E.Ghysels. (2006). Why Do Absolute Returns predict Volatility So Well.

Hansen P.R. and A. Lunde. (2006). Realized Variance and Market Microstructure Noise. *Journal of Business and Economic Statistics 24(2)* , 127-161.

Huang,X. and G. Tauchen. (2005). The Relative Contribution of Jumps to the Total Price Variance. *Journal of Financial Econometrics 3* , 456-499.

J. Gatheral and C. A. Oomen. (2009). Zero-Intelligence Realized Variance Estimation.

Lanne, M. (2006). Forecasting Realized Volatility by Decomposition. *European University Institute, working paper ECO2006/20* .

Liu, C. and J. M. Maheu. (2008). Forecasting Realized Volatility: A Bayesian Model Averaging. *Working Paper 313* .

Zhang L.,P. Mykland and Y. Ait-Sahalia. (2005). A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data. *Journal of American Statistical Association* , 1394-1411.

Zhou, B. (1996). High Frequency Data and Volatility in Foreign-Exchange Rates . *Journal of Business and Economic Statistics, 14* , 45-52.

Intradaily returns				
Mean	Std. dev	Skewness	Kurtosis	LB_{10} p-value
0.000001	0.000379	0.137184	26.944244	< 2.2e-16

Table 1

Frequency	5 minutes	15 minutes	30 minutes
Kurtosis	26.94424	22.08325	17.78766

Table 2

Daily return series					
Mean	Std. dev	Skewness	Kurtosis	LB_{10} p-value	LB^{2}_{10} p-value
0.000220	0.006467	0.103582	4.635859	0.2413	< 2.2e-16

The last two columns show the p-value of the Ljung-Box statistics for the daily return series, respectively for the squared daily return series.

Table 3

RV type used to standardize the returns	Mean	Std. dev	Skewness	Kurtosis	JB p-value	KS p-value
Naïve	0.008292	0.986941	0.008412	2.531097	0.0002157	0.4554
Bartlett	0.008601	0.986302	0.007408	2.468673	1.939e-05	0.3068
Parzen	0.008406	0.985680	0.007414	2.495548	5.684e-05	0.3247
TH	0.008349	0.985976	0.007692	2.509149	9.582e-05	0.3762

Table 4

RV type	Mean	Std. dev	Skewness	Kurtosis
Naïve	0.000041	0.000043	4.418956	31.914416
Bartlett	0.000041	0.000044	4.549322	31.423712
Parzen	0.000041	0.000043	4.475789	30.052765
TH	0.000041	0.000043	4.449837	29.545604

Table 5

$\log\sqrt{RV}$	Mean	Std. dev	Skewness	Kurtosis	JB	KS	JB’	KS’
Naïve	-5.189066	0.346557	0.590484	3.968051	0	0	0.06375	0.5611
Bartlett	-5.198005	0.353831	0.559127	3.913384	0	0	0.1306	0.8639
Parzen	-5.193016	0.349481	0.577726	3.948891	0	0	0.0936	0.7252
TH	-5.191258	0.348102	0.583685	3.958832	0	0	0.08009	0.5423

Table 6

Date	Announcement	Actual	Forecasted
January 9, 2004	Hourly Earning	0.2%	0.3%
	Nonfarm Payrolls	+124K	+155K
	Unemployment Rate	5.7%	5.9%
March 5, 2004	Hourly Earnings	0.2%	0.2%
	Nonfarm Payrolls	21K	120K
	Unemployment Rate	5.6%	5.6%
	Consumer Credit	$14.3 B	$ 7.5 B
May 7, 2004	Hourly Earnings	0.3%	0.1%
	Nonfarm Payrolls	288K	170K
	Unemployment Rate	5.6%	5.7%
	Wholesale Inventories	0.6%	0.4%
	Consumer Credit	$ 5.7B	$ 6.0 B
January 12, 2005	Trade Balance	-$60.3 B	-$52.3B
	Treasury Budget	-$3.4 B	-$ 0.0 B
December 16, 2008	Core CPI	0.0%	0.0%
	CPI	-1.7%	-1.5%
	Building Permits	616K	700K
	Housing Starts	625K	725K
March 18, 2009	Core CPI	0.2%	0.0%
	CPI	0.4%	0.2%
	Current Account	-$132.8B	-$137,1B

Table 7

Jump	Mean	Std.dev	Skewness	Kurtosis	LB 5th lag p-value
Naïve	0.000002	0.000007	12.897440	275.576480	0.1176
Bartlett	0.000002	0.000007	7.508456	84.774190	0.002701
Parzen	0.000002	0.000007	9.279680	131.049161	0.02416
TH	0.000002	0.000007	10.662665	175.539196	0.04712

Table 8

The models are denoted by two numbers: the first stands for the type of estimated model and the second refers to the proxy used.

Type of models:

1. Realized Variance form
2. Realized volatility form
3. Log Realized Volatility

Volatility proxy:

1. Naive RV
2. Bartlett realized kernel
3. Parzen Realized kernel
4. Tukey-Hanning realized kernel

The standard deviations (in paranthesis) are computed using the Newey – West HAC estimator with 20 lags.

Significance level:

„***” -0%,” **”-0.1%,” *”-1%,”·” -5%, „ ” -10%

Red-negative value, Black-pozitive value

			HAR-RV-RV			
	“β0”	“βD”	“βW”	“βM”	Adj R2	MSE
1.1	0.0183030	0.374400***	0.297333.	0.286631**	0.7082542	0.3044323
	(0.011979)	(0.102969)	(0.155471)	(0.091812)		
1.2	0.0193960	0.337379***	0.315517.	0.302501***	0.6836378	0.3010511
	(0.012096)	(0.098071)	(0.161592)	(0.091125)		
1.3	0.0187040	0.358978***	0.306095.	0.292157**	0.6990210	0.3028522
	(0.011959)	(0.100420)	(0.159490)	(0.091615)		
1.4	0.0185010	0.366253***	0.302276.	0.289259**	0.7036337	0.3035359
	(0.011947)	(0.101425)	(0.158033)	(0.091707)		
2.1	0.0217520	0.247831***	0.395355***	0.311428***	0.7370266	0.2791520
	(0.013251)	(0.062307)	(0.082444)	(0.061849)		
2.2	0.023347 .	0.218780***	0.411972***	0.319527***	0.7182862	0.2740827
	(0.013446)	(0.060455)	(0.086444)	(0.060658)		
2.3	0.022383.	0.234750***	0.405983***	0.312194***	0.7303769	0.2770110
	(0.013270)	(0.061787)	(0.085119)	(0.061101)		
2.4	0.0185010	0.366253***	0.302276.	0.289259**	0.7338050	0.2779938
	(0.011947)	(0.101425)	(0.158033)	(0.091707)		
3.1	0.041912***	0.158082***	0.439067***	0.367475***	0.7095344	0.2523004
	(0.011303)	(0.034517)	(0.050749)	(0.047430)		
3.2	0.047209***	0.126060***	0.466352***	0.368991***	0.6908543	0.2455768
	(0.011827)	(0.034355)	(0.053903)	(0.047776)		
3.3	0.043948***	0.142344***	0.456148***	0.364862***	0.7030317	0.2495828
	(0.011481)	(0.034554)	(0.052131)	(0.047406)		
3.4	0.042960***	0.149000***	0.449850***	0.365113***	0.7064159	0.2508637
	(0.011386)	(0.034569)	(0.051487)	(0.047361)		

Table 9

	HAR-RV-C					
	“β0”	“βD”	“βW”	“βM”	Adj R2	MSE
1.1	0.035426**	0.450572***	0.2275140	0.283267**	0.7165928	0.3060031
	(0.011369)	(0.093594)	(0.154235)	(0.095593)		
1.2	0.034535**	0.422058***	0.2371380	0.320273**	0.6879971	0.3018959
	(0.011512)	(0.098807)	(0.187587)	(0.106077)		
1.3	0.035389**	0.446056***	0.2222300	0.298473**	0.7067372	0.3043211
	(0.011092)	(0.091282)	(0.159330)	(0.096397)		
1.4	0.035964**	0.447423***	0.2262810	0.288348**	0.7106752	0.3048693
	(0.011446)	(0.091916)	(0.159509)	(0.098778)		
2.1	0.038919**	0.311881***	0.340240***	0.300246***	0.7430561	0.2834385
	(0.012689)	(0.060841)	(0.084628)	(0.064319)		
2.2	0.037540**	0.286546***	0.364102***	0.309442***	0.7231296	0.2781244
	(0.012615)	(0.065353)	(0.105879)	(0.068432)		
2.3	0.038422**	0.304371***	0.351471***	0.299113***	0.7365573	0.2815555
	(0.012336)	(0.062672)	(0.093589)	(0.065094)		
2.4	0.039196**	0.306492***	0.348898***	0.297164***	0.7399490	0.2823886
	(0.012675)	(0.061315)	(0.090809)	(0.065680)		
3.1	0.021208.	0.192883***	0.399314***	0.357941***	0.7138841	0.2594587
	(0.011288)	(0.035204)	(0.049826)	(0.047471)		
3.2	0.0163130	0.148369***	0.458578***	0.346665***	0.6960513	0.2543382
	(0.011671)	(0.037920)	(0.058882)	(0.050444)		
3.3	0.019407.	0.169719***	0.436918***	0.345355***	0.7085950	0.2580003
	(0.011213)	(0.037244)	(0.054725)	(0.048290)		
3.4	0.020798.	0.178137***	0.421818***	0.350560***	0.7119516	0.2587369
	(0.011224)	(0.036513)	(0.052395)	(0.047648)		

Table 10

				HAR-RV-CJ				
	“β0”	“βCD”	“βCW”	“βCM”	“βJD”	“βJW”	“βJM”	Adj R2	MSE
1.1	0.0328016*** (0.0092034)	0.4466412*** (0.0900008)	0.22961490 (0.1492245)	0.2853838** (0.0941760)	0.2730442** (0.0838989)	0.7946695* (0.3388726)	0.39066760 (0.3290424)	0.7187927	0.3065040
1.2	0.034891** (0.011133)	0.411738*** (0.087839)	0.2182550 (0.185127)	0.349915** (0.113515)	0.1056240 (0.142443)	0.908881. (0.488247)	0.8210580 (0.559055)	0.6910015	0.3025760
1.3	0.0338552*** (0.0099609)	0.4406273*** (0.0838216)	0.21931920 (0.1555935)	0.3061295** (0.1016716)	0.2494592*** (0.0565773)	0.8576442* (0.3658093)	0.52156370 (0.3602103)	0.7093431	0.3049075
1.4	0.0336910*** (0.0095673)	0.4413515*** (0.0846305)	0.22248920 (0.1561395)	0.2984952** (0.1028544)	0.2244304*** (0.0507171)	0.8601428* (0.3956776)	0.52513820 (0.3760195)	0.7130583	0.3054093
2.1	0.031657** (0.011313)	0.300965*** (0.058840)	0.354015*** (0.081102)	0.299882*** (0.062306)	0.064513** (0.023462)	0.142387** (0.047596)	0.0391150 (0.062186)	0.7450116	0.2831009
2.2	0.036863** (0.012422)	0.276056*** (0.060534)	0.357296*** (0.105075)	0.331853*** (0.072710)	0.0139660 (0.023911)	0.140398** (0.047158)	0.122038. (0.074037)	0.7251635	0.2790272
2.3	0.036398** (0.011606)	0.297710*** (0.060520)	0.353661*** (0.092820)	0.306109*** (0.067897)	0.038366. (0.020238)	0.123012** (0.044074)	0.0767860 (0.060422)	0.7377847	0.2818962
2.4	0.032963** (0.011684)	0.299727*** (0.059603)	0.352836*** (0.090588)	0.302254*** (0.065898)	0.041199. (0.021817)	0.118511** (0.045882)	0.0385720 (0.061752)	0.7410954	0.2821236
3.1	0.025255. (0.015350)	0.183644*** (0.034944)	0.412153*** (0.049597)	0.356268*** (0.046424)	0.094330* (0.040987)	0.220666** (0.068343)	0.0907840 (0.090384)	0.7154586	0.2582947
3.2	0.00725450 (0.0209832)	0.1423434*** (0.0368037)	0.4530443*** (0.0584459)	0.3643880*** (0.0519864)	0.01182480 (0.0355018)	0.1985661** (0.0634270)	0.1919051. (0.1085555)	0.6974769	0.2561558
3.3	0.0148740 (0.016808)	0.165206*** (0.036596)	0.437995*** (0.054499)	0.352259*** (0.048499)	0.0521290 (0.037235)	0.187264** (0.065276)	0.145258. (0.087902)	0.7095485	0.2590879
3.4	0.0216040 (0.015405)	0.172704*** (0.036161)	0.426818*** (0.052555)	0.352875*** (0.047292)	0.067680. (0.040320)	0.190354** (0.068036)	0.1065580 (0.087732)	0.7129216	0.2585035

Table 11

HAR-RV-RPV 1.3						
	“β0”	“βD”	“βW”	“βM”	Adj R2	MSE
1.1	0.2519596***	0.0240830***	0.0225336**	0.0100999*	0.6980546	0.3025111
	(0.0282561)	(0.0067165)	(0.0081115)	(0.0050457)		
1.2	0.2551428***	0.0219784**	0.0240033*	0.0107542*	0.6747313	0.2993250
	(0.0278935)	(0.0072349)	(0.0099376)	(0.0052982)		
1.3	0.2535512***	0.0230307***	0.0232685**	0.0104271*	0.6888056	0.3009077
	(0.0279978)	(0.0069025)	(0.0089825)	(0.0051367)		
1.4	0.2528919***	0.0234666***	0.0229641**	0.0102916*	0.6932479	0.3015693
	(0.0280861)	(0.0068069)	(0.0086099)	(0.0050901)		
2.1	0.2078389***	0.0107012***	0.0127523***	0.0095939***	0.7390120	0.2957358
	(0.0110163)	(0.0020419)	(0.0033107)	(0.0026216)		
2.2	0.2039657***	0.0099400***	0.0132163***	0.0099005***	0.7187328	0.2901962
	(0.0108316)	(0.0023148)	(0.0037273)	(0.0027870)		
2.3	0.2060007***	0.0103200***	0.0129908***	0.0097460***	0.7311268	0.2932011
	(0.0108902)	(0.0021614)	(0.0034981)	(0.0026923)		
2.4	0.2067854***	0.0104776***	0.0128936***	0.0096828***	0.7349488	0.2943073
	(0.0109344)	(0.0021075)	(0.0034150)	(0.0026600)		
3.1	2.269210***	0.115943***	0.385081***	0.213149***	0.7176303	0.2597955
	(0.025181)	(0.031253)	(0.044386)	(0.035525)		
3.2	2.289895***	0.101612**	0.395058***	0.222360***	0.6964922	0.2539491
	(0.027350)	(0.033465)	(0.050319)	(0.040935)		
3.3	2.278770***	0.108617***	0.390046***	0.217828***	0.7095292	0.2572893
	(0.026053)	(0.032250)	(0.047151)	(0.037992)		
3.4	2.274626***	0.111607***	0.387991***	0.215906***	0.7134840	0.2584265
	(0.025639)	(0.031810)	(0.045956)	(0.036911)		

Table 12

HAR-RV-RPV 1.5						
	“β0”	“βD”	“βW”	“βM”	Adj R2	MSE
1.1	0.148391***	0.059188***	0.043249*	0.025541*	0.7088575	0.3045460
	(0.019487)	(0.015004)	(0.018726)	(0.010768)		
1.2	0.151535***	0.054493**	0.046503*	0.027025*	0.6846241	0.3012422
	(0.019072)	(0.016744)	(0.023056)	(0.011393)		
1.3	0.149963***	0.056840***	0.044876*	0.026283*	0.6991839	0.3028833
	(0.019199)	(0.015715)	(0.020799)	(0.010999)		
1.4	0.149312***	0.057813***	0.044202*	0.025975*	0.7038103	0.3035693
	(0.019299)	(0.015379)	(0.019915)	(0.010883)		
2.1	0.2708239***	0.0259551***	0.0241301**	0.0238844***	0.7342095	0.3022057
	(0.0113140)	(0.0043864)	(0.0082415)	(0.0066202)		
2.2	0.2669749***	0.0242391***	0.0251836**	0.0245663***	0.7137531	0.2963808
	(0.0111726)	(0.0051337)	(0.0092382)	(0.0069888)		
2.3	0.2690066***	0.0250979***	0.0246699**	0.0242220***	0.7262206	0.2995318
	(0.0112183)	(0.0047209)	(0.0086946)	(0.0067809)		
2.4	0.2697848***	0.0254530***	0.0244495**	0.0240814***	0.7300820	0.3006965
	(0.0112519)	(0.0045721)	(0.0084951)	(0.0067085)		
3.1	1.424941***	0.113777***	0.318939***	0.193467***	0.7182422	0.2590906
	(0.012513)	(0.026782)	(0.037599)	(0.030312)		
3.2	1.440041***	0.100253***	0.328676***	0.201574***	0.6970294	0.2532627
	(0.013684)	(0.028607)	(0.042923)	(0.035041)		
3.3	1.431841***	0.106906***	0.323755***	0.197584***	0.7101032	0.2565922
	(0.012969)	(0.027601)	(0.040085)	(0.032466)		
3.4	1.428830***	0.109720***	0.321754***	0.195893***	0.7140735	0.2577258
	(0.012748)	(0.027239)	(0.039012)	(0.031522)		

Table 13

Mincer-Zarnowitz regression
Adjusted R^2

		Naive	Bartlett	Parzen	Tukey-Hanning
HAR-RV-RV	1	0.8950810	0.8815645	0.8907257	0.8931173
	2	0.9049928	0.8957341	0.9016784	0.9033715
	3	0.9041155	0.8964124	0.9013027	0.9027194
HAR-RV-C	1	0.9201028	0.8990014	0.9170683	0.9193151
	2	0.9173330	0.9038770	0.9140440	0.9160607
	3	0.9079164	0.8955576	0.9038662	0.9064052
HAR-RV-CJ	1	0.8981015	0.8793819	0.8967763	0.8988256
	2	0.9009779	0.8902342	0.8991075	0.9003612
	3	0.8926952	0.8839583	0.8905152	0.8912801
HAR-RV-RPC 1.3	1	0.8521560	0.8251208	0.8408019	0.8460536
	2	0.9113617	0.8965140	0.9054941	0.9083048
	3	0.9006989	0.8899413	0.8966788	0.8986691
HAR-RV-RPV 1.5	1	0.8789885	0.8514711	0.8675038	0.8728372
	2	0.9166845	0.9016403	0.9107523	0.9135979
	3	0.9055235	0.8943567	0.9013256	0.9033959

Table 14

Mincer-Zarnowitz regression
Ho: α1=0 and α1=1 (p-value)

		Naive	Bartlett	Parzen	Tukey-Hanning
HAR-RV-RV	1	0.32	0.22	0.27	0.29
	2	0.29	0.28	0.29	0.29
	3	0.016	0.019	0.019	0.018
HAR-RV-C	1	0.14	0.054	0.059	0.082
	2	0.33	0.2	0.19	0.23
	3	0.19	0.25	0.23	0.20
HAR-RV-CJ	1	0.3	0.13	0.13	0.16
	2	0.76	0.55	0.5	0.54
	3	0.77	0.55	0.38	0.37
HAR-RV-RPC 1.3	1	0.066	0.045	0.055	0.059
	2	0.15	0.14	0.14	0.14
	3	0.12	0.13	0.12	0.11
HAR-RV-RPV 1.5	1	0.063	0.042	0.051	0.056
	2	0.11	0.11	0.1	0.11
	3	0.067	0.092	0.076	0.071

Table 15

Forecast: MSE					
		Naive	Bartlett	Parzen	Tukey-Hanning
HAR-RV-RV	1	0.05579621	0.01776198	0.01793531	0.01800635
	2	0.07843692	0.01591763	0.01535754	0.01531190
	3	0.10611847	0.02824288	0.02684165	0.02661027
HAR-RV-C	1	0.04218951	0.02430843	0.02713929	0.02689419
	2	0.06507980	0.01581078	0.01608302	0.01637243
	3	0.09317418	0.01870461	0.01817107	0.01915310
HAR-RV-CJ	1	0.05307414	0.03367978	0.03543336	0.03489641
	2	0.07592465	0.02554392	0.02503671	0.02507692
	3	0.10239262	0.02753245	0.02808739	0.02964508
HAR-RV-RPC 1.3	1	0.08334444	0.02634451	0.02591997	0.02590775
	2	0.06833373	0.02454084	0.02397395	0.02389288
	3	0.10139842	0.02029693	0.01909034	0.01880543
HAR-RV-RPV 1.5	1	0.06786980	0.02201944	0.02188815	0.02200263
	2	0.06423852	0.03654996	0.03654827	0.03669380
	3	0.09806634	0.02044325	0.01936109	0.01913394

Table 16

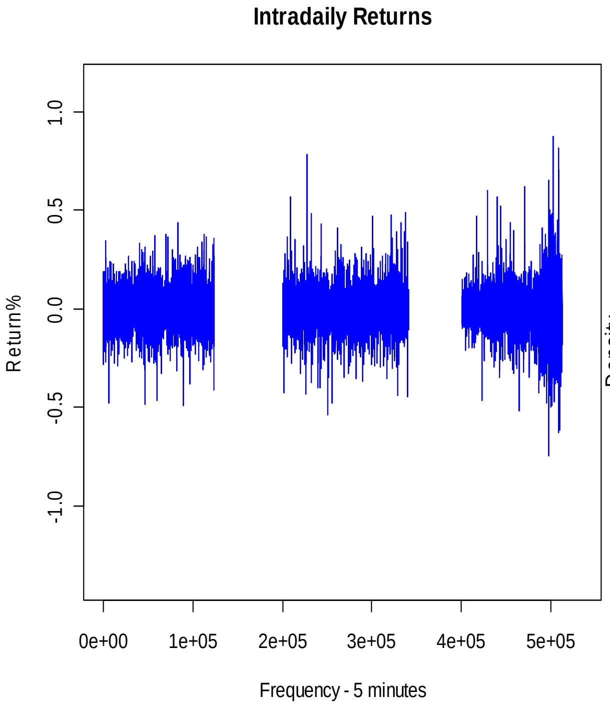


Figure 1

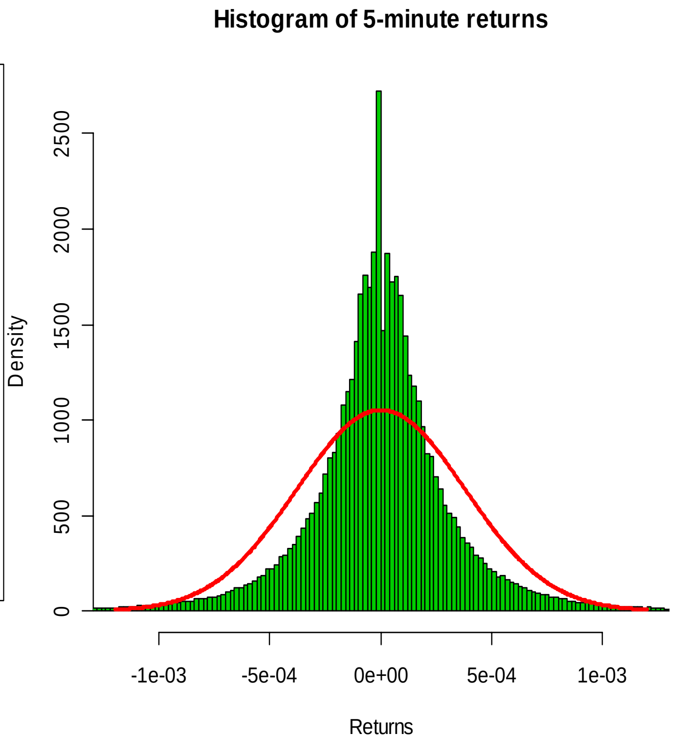


Figure 2

The red lines depicts the distribution of a normal variable

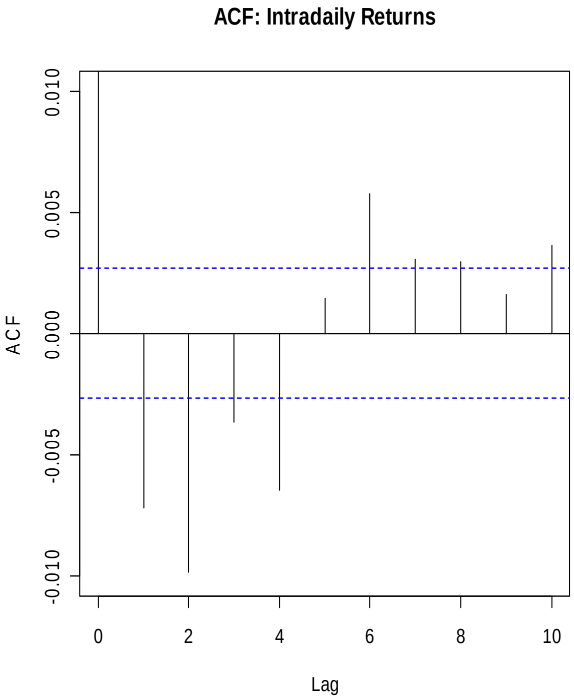


Figure 3

A significant negative correlation can be observed up to 4th lag

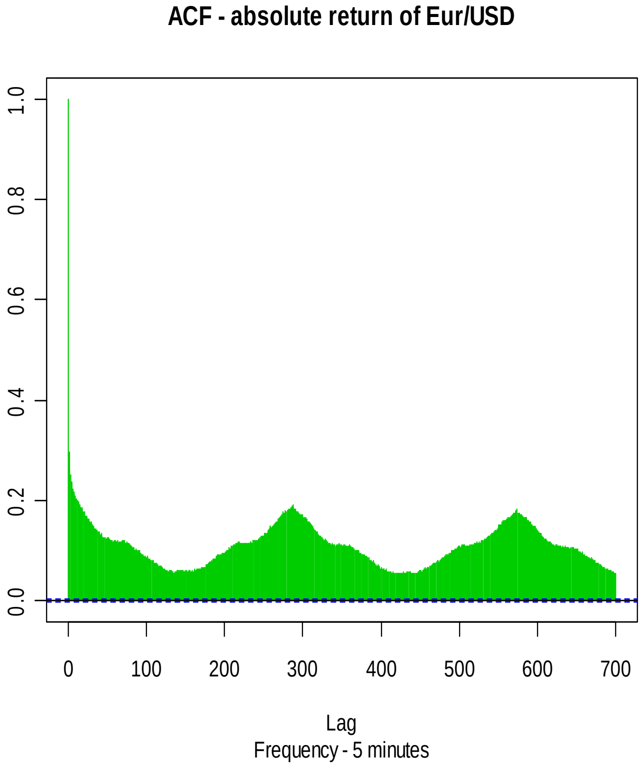


Figure 4

The autocorrelation function computed for the absolute intradaily returns.

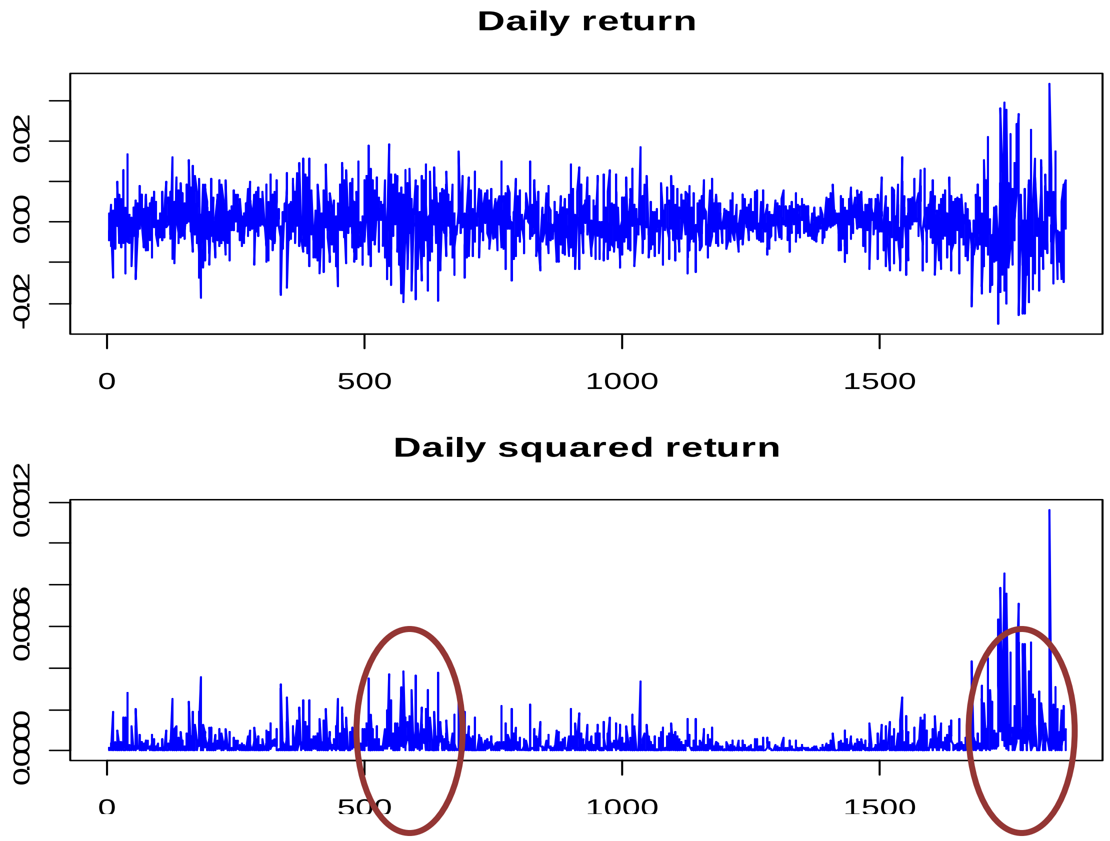


Figure 5

The red circles show the volatility clustering

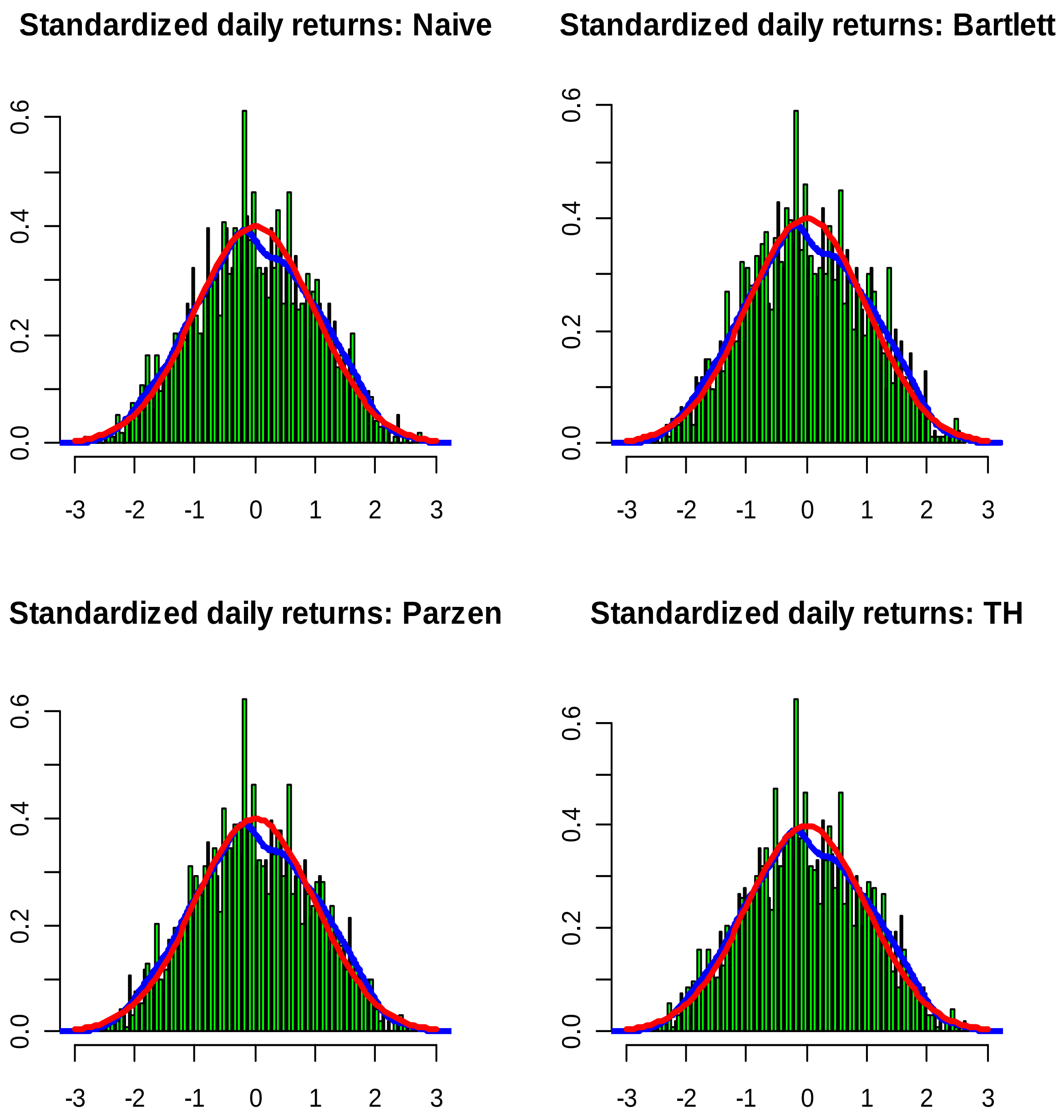

Standardized daily returns: Naive
Standardized daily returns: Bartlett
Standardized daily returns: Parzen
Standardized daily returns: TH
0.0
0.2
0.4
0.6
-3
-2
-1
0
1
2
3

Figure 6

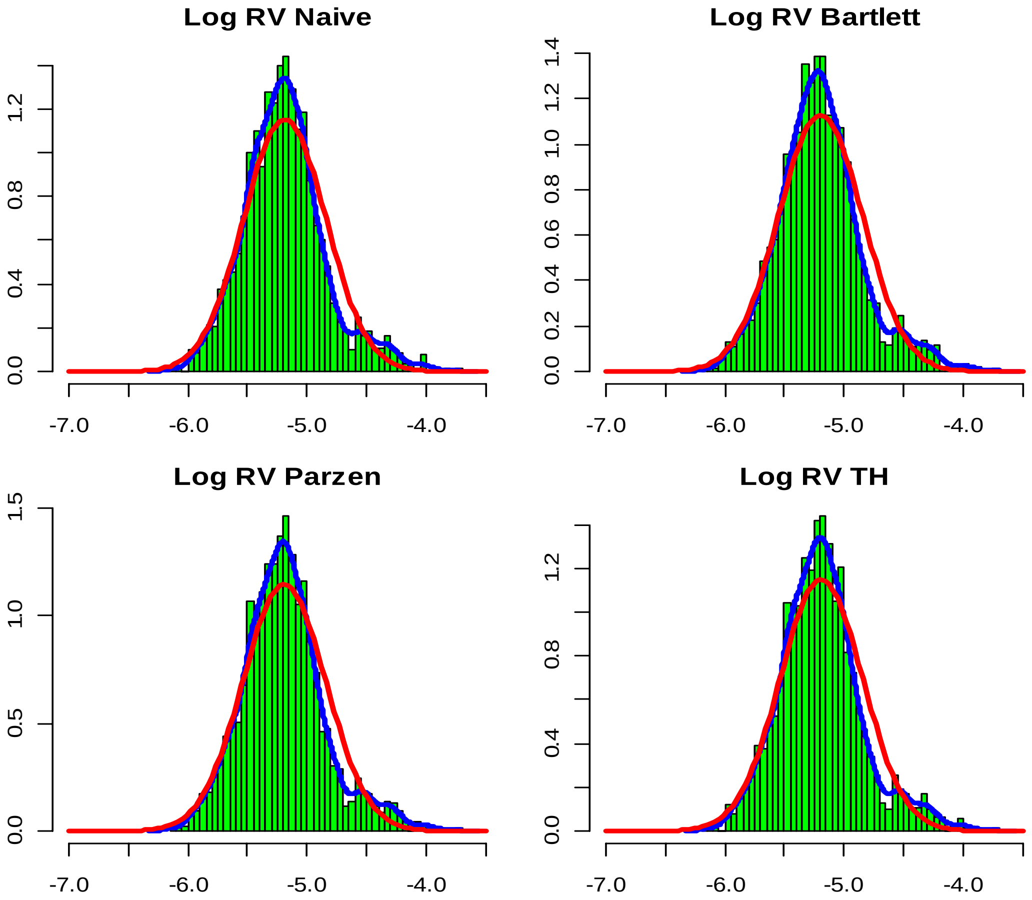


Figure 7

The blue line represents the empirical probability function of the series, and the red line – the pdf of a normally distributed variable

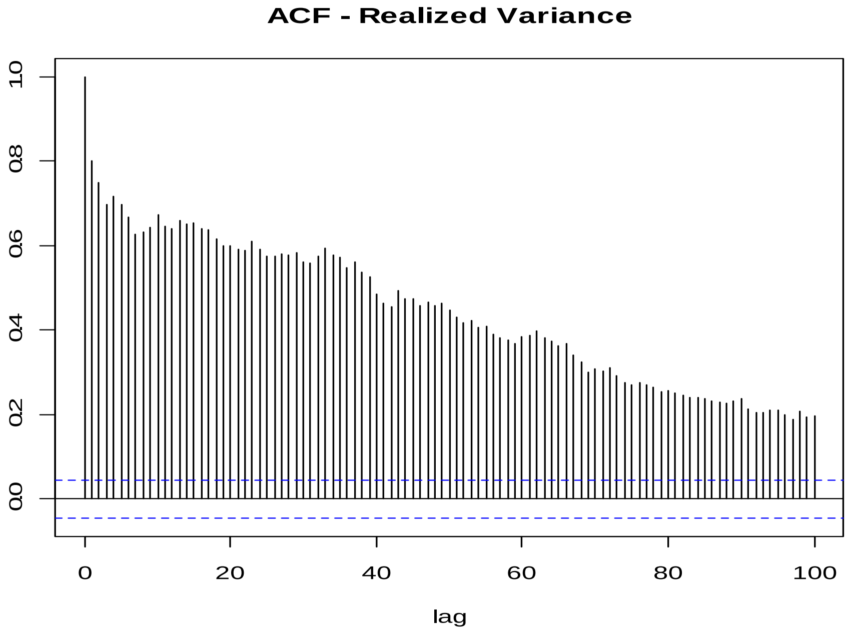


Figure 8

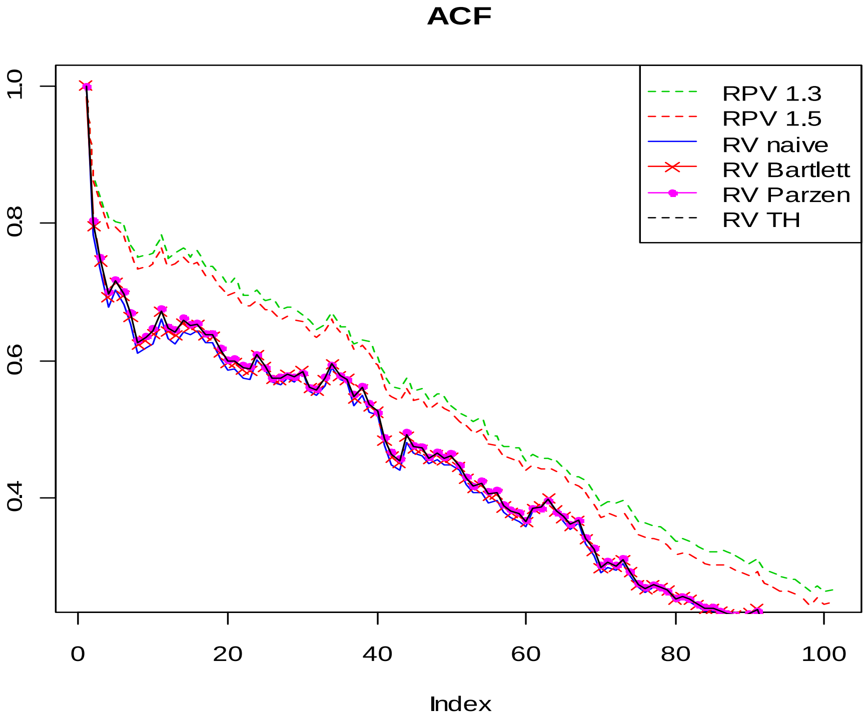


Figure 9

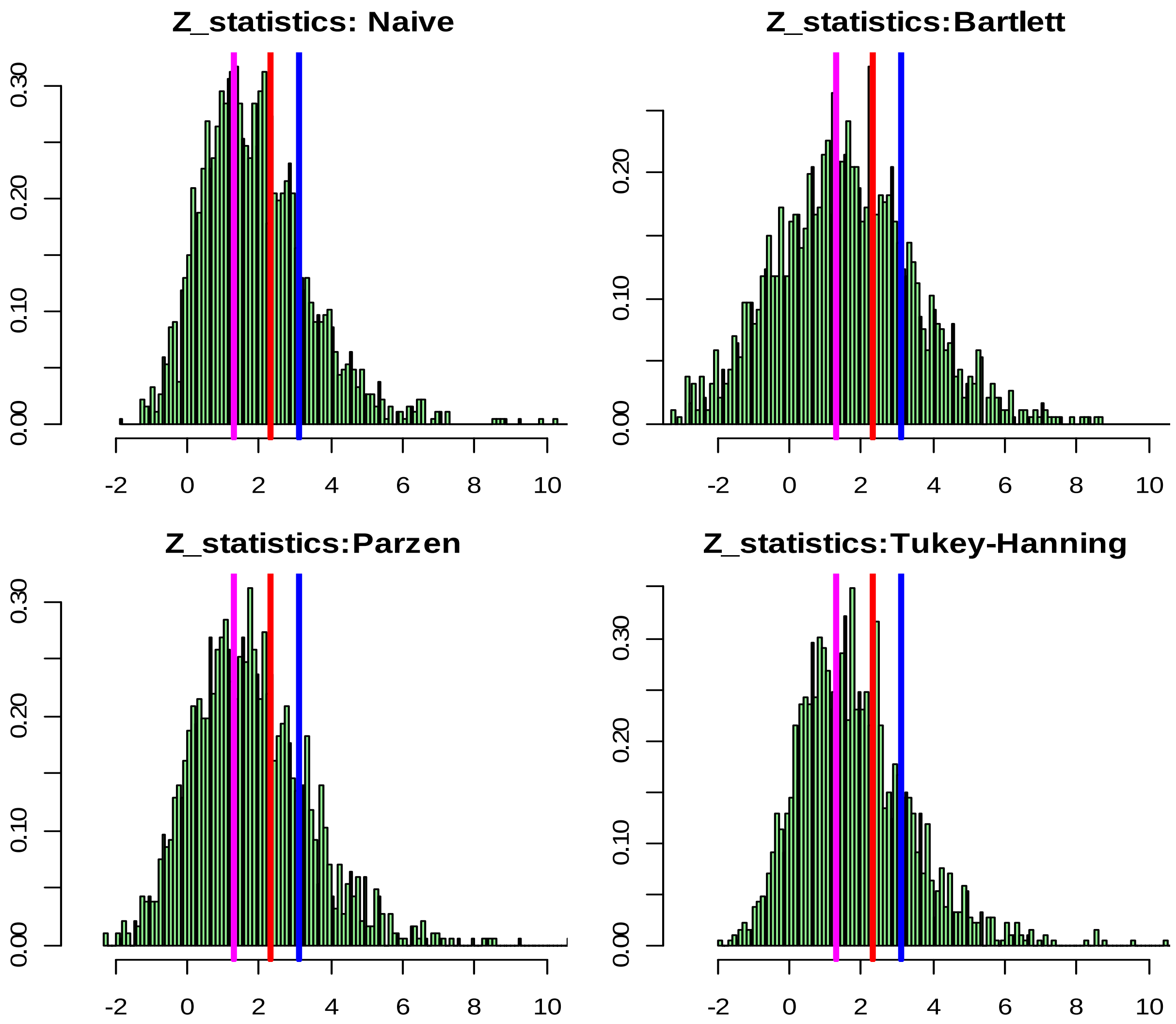

Z_statistics: Naive
0.00
0.10
0.20
0.30
-2
0
2
4
6
8
10
Z_statistics:Bartlett
0.00
0.10
0.20
-2
0
2
4
6
8
10
Z_statistics:Parzen
0.00
0.10
0.20
0.30
-2
0
2
4
6
8
10
Z_statistics:Tukey-Hanning
0.00
0.10
0.20
0.30
-2
0
2
4
6
8
10

Figure 10

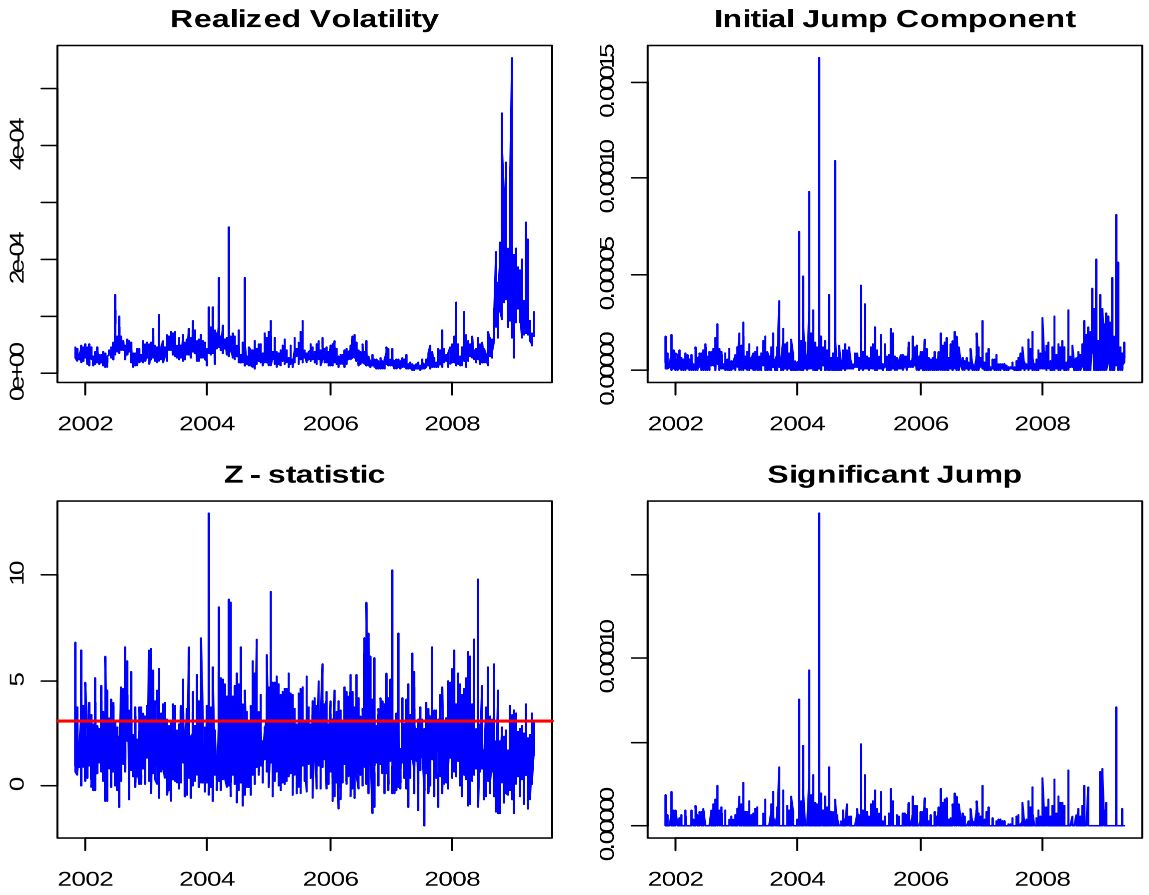


Figure 11

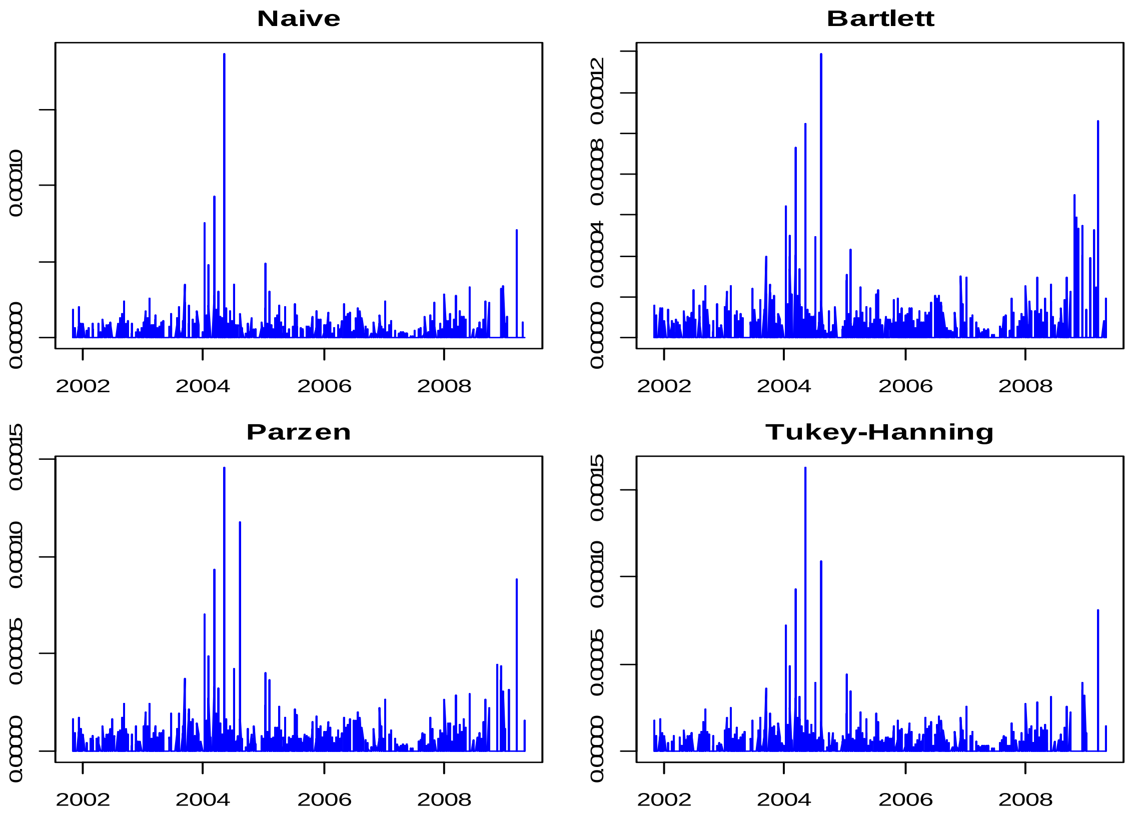


Figure 12

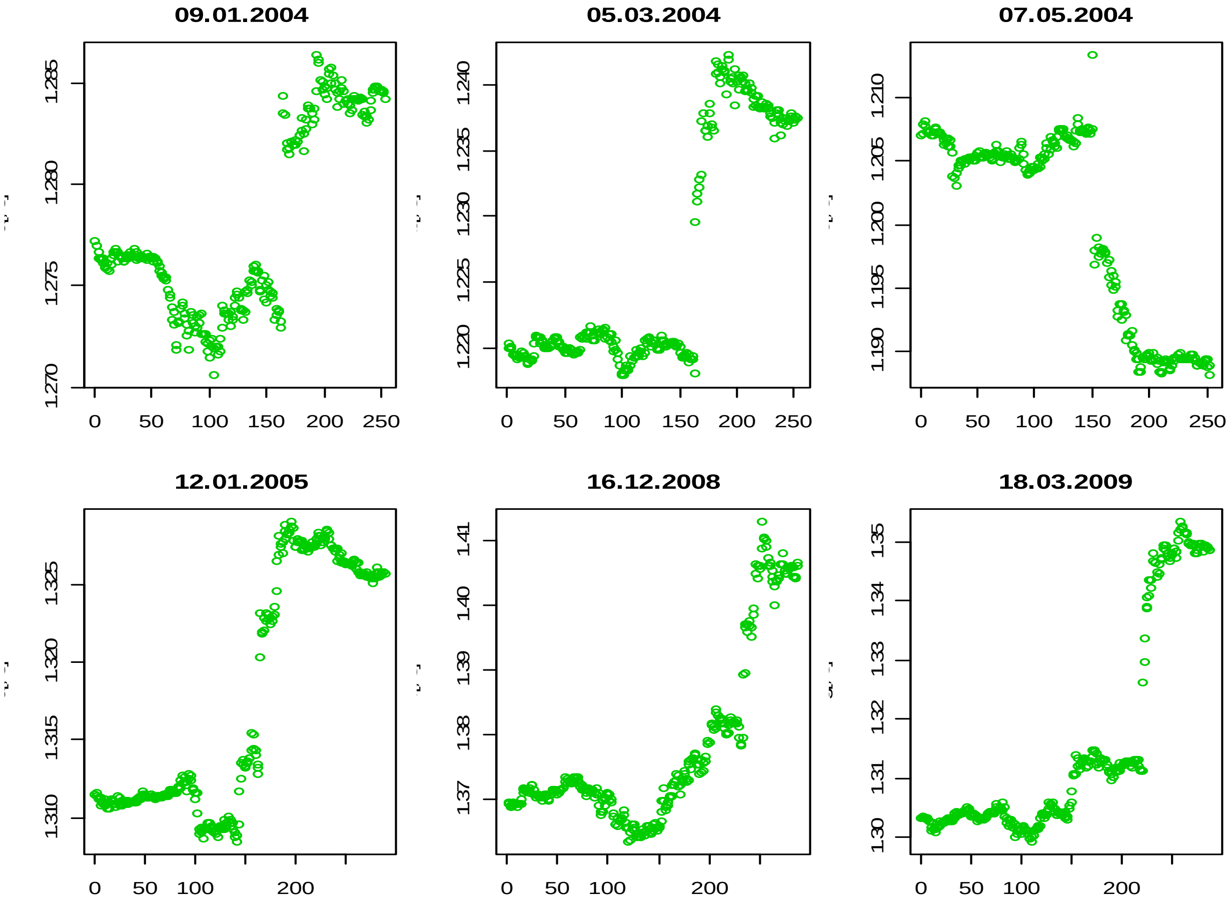

09.01.2004
1.270
1.275
1.280
1.285
0
50
100
150
200
250
05.03.2004
1.220
1.225
1.230
1.235
1.240
0
50
100
150
200
250
07.05.2004
1.190
1.195
1.200
1.205
1.210
0
50
100
150
200
250
12.01.2005
1.310
1.315
1.320
1.325
0
50
100
200
16.12.2008
1.37
1.38
1.39
1.40
1.41
0
50
100
200
18.03.2009
1.30
1.31
1.32
1.33
1.34
1.35
0
50
100
200

Figure 13

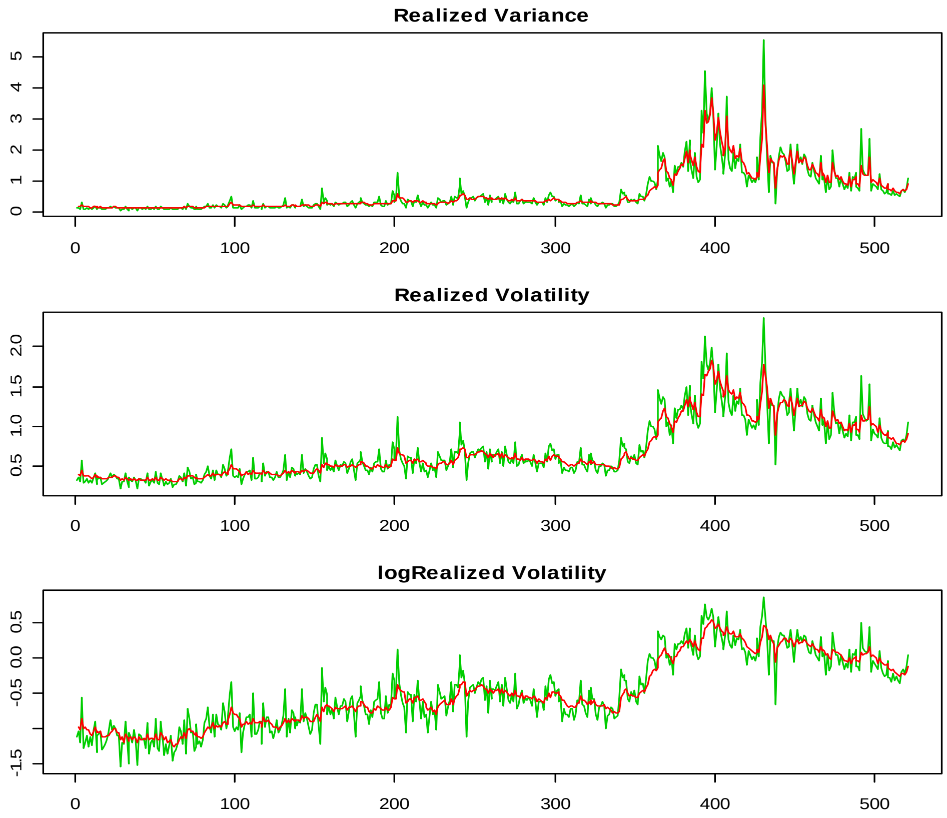


Figure 14a

HAR-RV-C forecasts for the Naïve RV; Red-forecast; Green-actual

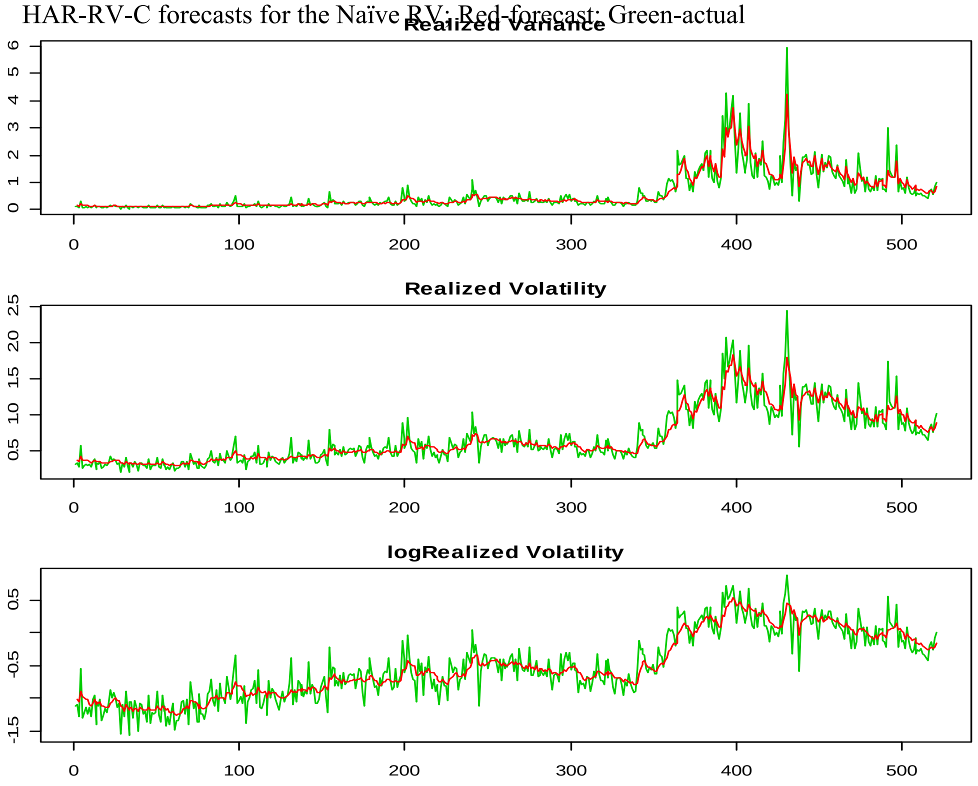


Figure 14b

HAR-RV-C forecasts for the Bartlett realized kernel; Red-forecast; Green-actual

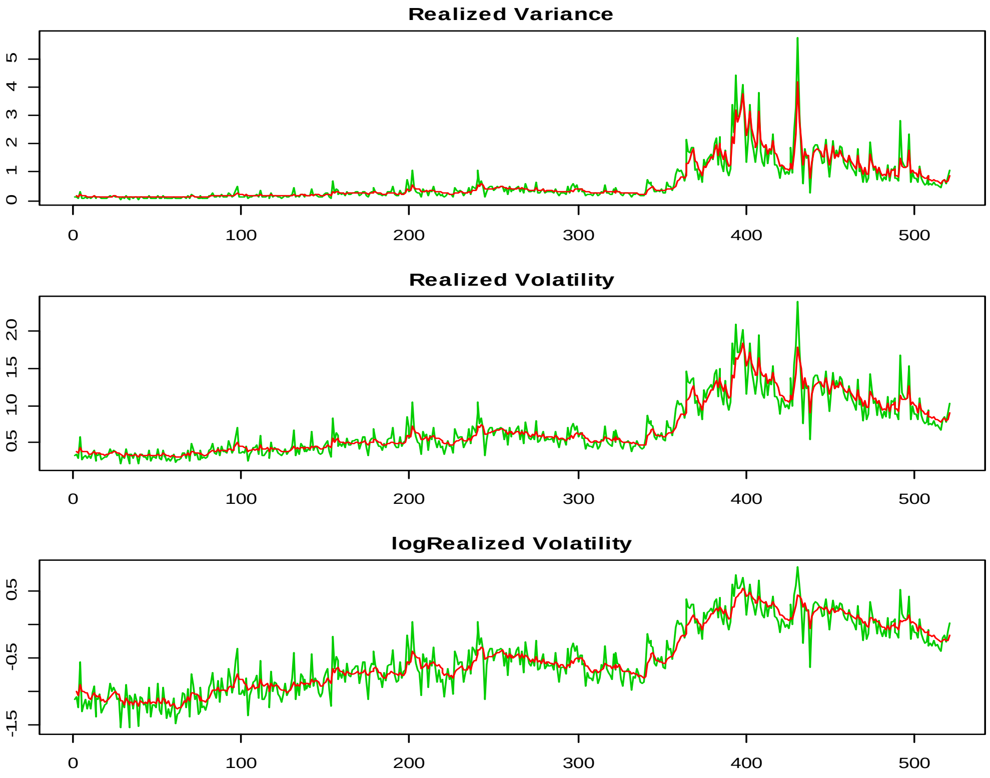


Figure 14c

HAR-RV-C forecasts for the Parzen realized kernel; Red-forecast; Green-actual

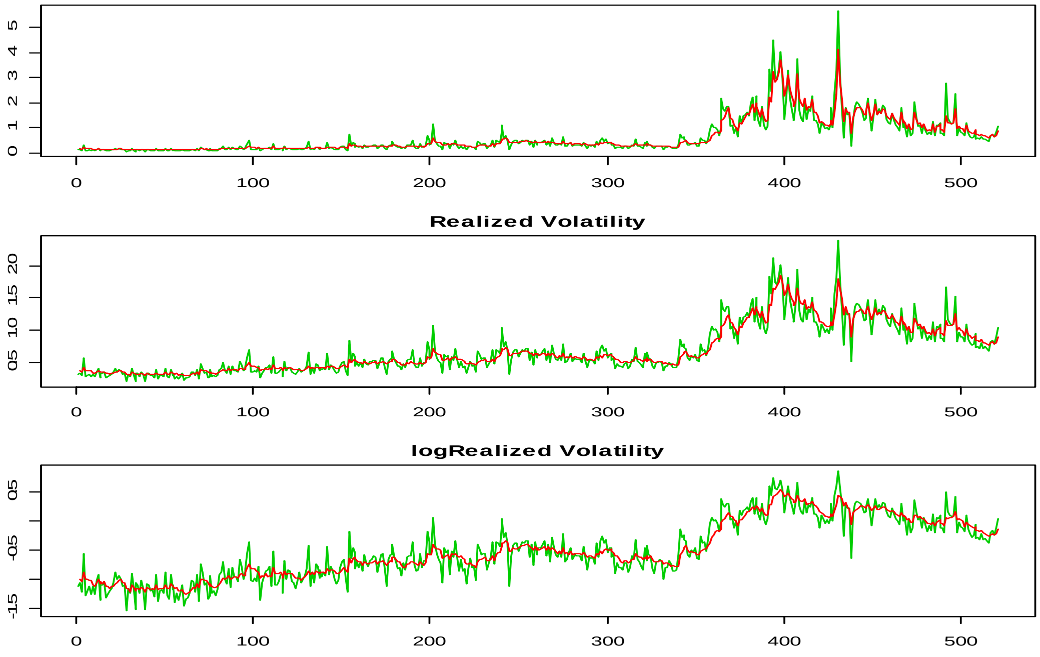


Figure 14d

HAR-RV-C forecasts for the Tukey-Hanning realized kernel; Red-forecast; Green-actual