Università
Ca'Foscari
Venezia

Master's Degree Programme

in Economics and Finance
Second Cycle (D.M. 270/2004)

Final Thesis

Analysis of the volatility of high-frequency data

The Realized Volatility and the HAR model

Supervisor:

Prof. Stefano Federico Tonellato

Graduand:

Silvia Pandolfo

Matriculation Number 862830

Academic Year

2017/2018

Acknowledgement

I would like to express my sincere gratitude to my supervisor Prof. Stefano Federico Tonellato for his guidance and advice in writing this thesis.

Beside my supervisor, I would like to thank my parents Loredana and Franco, my grandmothers Gilda and Giulia, Sergio, Stella, Enrico and Erika. I am also grateful to my grandfather Lino, who would have been so proud of me.

I wish to thank all my friends who supported me all the time.

Last but not least, a special thanks to Gianluca for always being by my side.

Contents

Introduction

Volatility plays an essential role in many financial branches: it is the main interest in asset pricing and knowing its dynamics could be fundamental for hedging decisions or satisfactory risk management. Hence, in the last decades, the literature has been very active in trying to develop new efficient approaches for volatility measurement, modeling and forecasting.
In 1982, Engle introduced the Autoregressive Conditional Heteroskedasticity model (ARCH), which is still largely used in finance to capture the volatility behavior of asset returns. However, this model has been thought to be applied to low frequency data, like daily or weekly returns. Indeed, traditionally, data were low frequency because of the high costs in collecting and analyzing the transactions, but today, thanks to the technology evolution and integration of computers in financial markets, detailed information about transactions and quotes at a high frequency level are much easier to obtain. These kind of data show features that are not present in lower frequency and that standard models are not able to reproduce. Then, how can all this additional information be used to improve the volatility modeling?

A possible approach is to estimate the variance without making any parametric assumption on the dynamics of the process of the returns, but relying instead only on its moments. A recently introduced milestone in financial econometrics has been the concept of Realized Volatility. Consider the integrated volatility, which is a natural volatility measure computed as the integral of the instantaneous volatility over an interval of interest. Andersen, Bollerslev, Diebold and Labys (2001) proposed to construct the non parametric realized volatility, summing the squared intraday high-frequency returns of a stock, in order to get an estimate of the integrated volatility of its price process. Hence, under specific circumstances, this measure results to consistently estimate the price volatility over some time interval.

In 2003, Corsi introduced a new class of models : the Heterogeneous Autoregressive model of Realized Volatility (HAR-RV). His aim was to propose a conditional volatility model able to account for the typical feature of financial data and to produce good one-day-ahead forecasts of the realized volatility, relying on past intraday returns. This model has been later implemented in Andersen et al. (2007) and Corsi (2012), including in the model components related to the jump and the leverage effects in the realized volatility series.

The HAR class of models will be the focal point of my thesis, which is divided into four chapters. In Chapter 1, I will explain the main characteristics of the high-frequency data, as the non-synchronous trading, the effect of the bid-ask spread, the diurnal patterns and others. In Chapter 2, the concept of volatility is defined and its properties are listed. Moreover, the parametric and non parametric approaches in modelling volatility are compared. Regarding the first type, the Autoregressive Conditional Heteroscedasticity (ARCH) and the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) are the most used methods to model volatility in a time series. On the other hand, the non-parametric realized volatility has often lead, under specific circumstances, to better performances in avoiding the high frequency data complications, as advocated also in Andersen and Bollerslev (1998), Andersen, Bollerslev, Diebold and Labys (2001) and Meddhai (2002). In Chapter 3, the HAR model for the realized volatility is introduced, together with its improvements regarding the jump component and the leverage effect. Finally, in Chapter 4, I will analyze the high frequency price series of three stocks from the italian stock exchange: Enel, Generali and Intesa San Paolo. After computing the returns, I will calculate the realized volatility series for each stock. Then, I will apply to them the HAR models to assess their ability in capturing the features of the realized volatility. Morevorer, their forecasting performances are compared to the ones of ARMA and ARFIMA models. Data have been downloaded from Bloomberg and all the analysis have been computed in R.

Chapter 1

Characteristics of high-frequency data

In finance, by "high frequency data" we mean a collection of observations taken in an extremely fine time scale. In this chapter, I am going to explain the most important characteristics of high-frequency data that do not appear in lower frequencies, such as the nonsynchronous trading, the bid-ask spread, the diurnal pattern, the movement of trading prices and the trading intensity.

1.1 Nonsynchronous trading

One of the crucial features that mark intraday data is that the observations are collected at random time. Indeed, in a limit order market, investors set the minimum or maximum price at which they are willing to sell or buy, and the transactions occur as soon as the actual market price falls into the decided range. It results that the customers' orders take place at random times and so the stocks are traded in a non-synchronous manner, meaning that different stocks do not have the same trading frequency and a single stock's intensity could vary during the day. For example, when we consider closing prices, we incorrectly assume that the values are all equally spaced by an interval of 24 hours, while in reality the closing prices do not occur always at the same instant. This can lead to several problems regarding standard econometric models: the data are seldom identically distributed, periodic effects are difficult to detect and forecasting is not very straightforward.

In particular, for daily stock returns, the nonsynchronous trading can lead to:

- cross-correlation at lag 1 between the stock returns;
- serial correlation at lag 1 in a portfolio return;
- sometimes, negative serial correlations of the return series of a single stock.

Consider two independent stocks A and B and assume that A is traded more frequently than B. Then, if news affecting the aggregate stock market occurs a few minutes before the closing time, it is more likely that the effect appears immediately only on A, while it could be delayed to the next day on B. The lagged response of B induces lag-1 cross correlation between the stock returns, no matter if the stocks are independent. Moreover, considering a portfolio consisting of the securities A and B, a serial dependence would be exhibited. Regarding the third point, many studies have been computed. As an example, I am going to introduce a simplified version proposed by Tsay (2013) of the nonsynchronous trading model of Lo and MacKinlay (1990).

1.1.1 A model of nonsynchronous trading

Being r_t the continuously compounded return that represents the changes in a security's value at time t, we assume that r_t is a sequence of i.i.d. random variables with mean $E[r_t] = \mu$ and variance $Var(r_t) = \sigma^2$, and that r_t^0 is the observed return. As we stated earlier, there is the possibility that a security is not traded during a certain interval of time; then we call π the probability that this happens. The same can occur in the subsequent period: there is a chance with probability π that the security is not traded. It is assumed that this mechanism is independent and identically distributed, meaning that the fact that a security is traded or not in period t does not affect its chance of being traded in the following periods. The observed return at time $t+1$ can be considered as the sum of r_{t+1} and all its virtual returns for the past consecutive period for which the security has not been traded. Hence, if the security is traded at time $t+1$ and has been traded at time $t-k-1$, but it has not been traded in the period between $t-k$ and t, its observed return results to be the sum of all the virtual returns from $t-k$ to $t+1$. If no trade occurs at time t, then the observed return for period t is simply zero, since there is not information available. We can then summarize the relationship between r_t and r_t^0 as:

$$r_t^0 = \begin{cases} 0 & \text{with probability } \pi \\ r_t & \text{with probability } (1-\pi)^2 \\ r_t + r_{t-1} & \text{with probability } (1-\pi)^2\pi \\ r_t + r_{t-1} + r_{t-2} & \text{with probability } (1-\pi)^2\pi^2 \\ \vdots & \vdots \\ \sum_{i=0}^{k} r_{t-i} & \text{with probability } (1-\pi)^2\pi^k \\ \vdots & \vdots \end{cases}$$

Indeed, $r_t^0 = r_t + r_{t-1}$ if and only if trades occurred at t and $t-1$; $r_t^0 = r_t + r_{t-1} + r_{t-2}$ if trades occurred at t and $t-2$, but not at $t-1$; and so on. Obviously, the sum of the probabilities is equal to 1:

$$\pi + (1-\pi)^2[1 + \pi + \pi^2 + \ldots] = \pi + (1-\pi)^2 \frac{1}{1-\pi} = \pi + 1 - \pi = 1$$

The expectation of r_t^0 results to be:

$$\mathbb{E}(r_t^0) = \mu, \tag{1.1}$$

the variance is:

$$Var(r_t^0) = \sigma^2 + \frac{2\pi\mu^2}{1-\pi}, \tag{1.2}$$

while the autocovariance at lag 1 is equal to:

$$Cov(r_t^0, r_{t-1}^0) = -\pi\mu^2. \tag{1.3}$$

All the computations leading to 1.1, 1.2 and 1.3 are reported in **Appendix A**. It can be noticed that the nonsynchronous trading affects the variance of r_t^0, but not the mean.

Assuming $\mu \neq 0$, the lag-1 autocorrelation induced by the nonsyncronous trading is negative and equal to:

$$\rho_1(r_t^0) = \frac{-(1-\pi)\pi\mu^2}{(1-\pi)\sigma^2 + 2\pi\mu^2}.$$

Extending it generally, we obtain:

$$Cov(r_t^0, r_{t-j}^0) = -\mu^2\pi^j, \qquad j \geq 1.$$

Summing up, provided that $\mu \neq 0$, the nonsyncronous trading causes negative correlations to observed return series. The same conclusion can be obtained analysing the return series of a portfolio including N securities; see Campbell et al. (1997). A possible solutions could be the construction of a series with equally spaced observations outdistanced by intervals of length Δ. Anyway, this requires assumptions: there could be no observed transactions at time $i\Delta$, so that in its place it is necessary to put an additional value (i.e. equal to the previous one). In case of quotes data, the observations are available continuously, making it simpler to create an equally spaced time series.

1.2 The bid-ask spread

In the case of an informationally efficient market, with zero trading costs, all the relevant information are contained in the market price. The price would change if and only if the market participants receive unanticipated information. On the other hand, when the transactions are costly, there is a market maker to compensate. One of the main interests of an investor is the liquidity of a market, meaning the capability to trade quickly, anonimously and in big quantities a security, without a substantial price impact. In some stock exchanges, the trades are facilitated by market makers, who are individuals that maintain liquidity buying and selling every time somebody wants to sell or buy a security. Doing so, the market makers update in an optimal way the bid and the ask prices, so that all public information and remaining uncertainty are reflected into the prices. In return for the risk assumed trading against potentially better informed agents, market makers have the right to buy and sell an asset at two different prices, which are respectively the bid price P_b and the ask price P_a. Being $P_a > P_b$, the positive difference is called bid-ask spread, a small region of price which brackets the underlying value of the security and represents the profit of the market makers.

The bid-ask spread groups three economic sources:

- order processing costs, meaning the setup and operating costs of trading and recordkeeping;
- inventory costs, which regard the carrying of risky inventory;
- adverse-selection costs, which arise because of the possibility that the market maker could be less informed than some investors, and hence could have a loss.

In an efficient market, the bid-ask average fluctuates randomly. However, the observed market price changes are no longer identically distributed because the transactions do not occur at the average value, but either at the bid or at the ask values. One of the main impacts that the bid-ask spread has on the asset prices is called *bid-ask bounce*, which shows a negative lag-1 serial correlation in the asset return. Indeed, as random buys and sells occur, the prices move back and forth between the bid price and the ask price, creating spurious volatility and serial correlation.

1.2.1 The Roll model

In 1984, Richard Roll developed a simple model in order to infer the bid-ask spread directly from the time series of market prices.

Assume a frictionless economy; let P_t^* be the value of a security at time t and s be the bid-ask spread. If no new information occurs and the transaction at time $t-1$ was a sale, Figure 1.1 shows the possible equally likely paths of the observed price. If at $t-1$ there was a purchase instead of a selling, the graph would have been similar but asymmetrically opposite. We can define the observed market price of the security P_t as following:

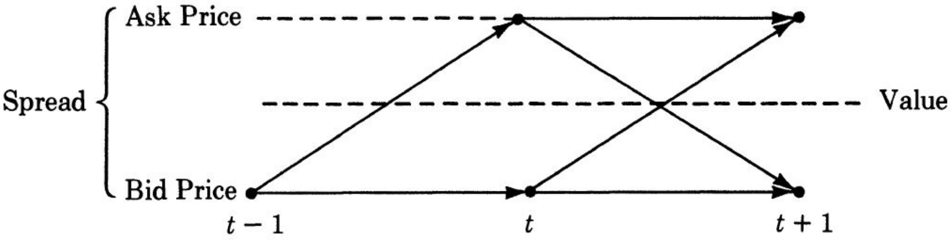


Figure 1.1: Possible paths for market price from time $t-1$ to time $t+1$.

$$P_t = P_t^* + I_t \frac{s}{2},$$

where

$$I_t \sim IID \begin{cases} +1 & \text{with probability } \frac{1}{2} \text{ (buyer initiated)} \\ -1 & \text{with probability } \frac{1}{2} \text{ (seller initiated).} \end{cases}$$

I_t is defined as an indicator function that establishes whether the random transaction is executed by a buyer or a seller, i.e. at the ask or at the bid price. Note that $E(I_t)$ has to be equal to zero, since P_t^* is the fundamental price. If there are no new information, and hence no changes in P_t^*, then $P_t^* = P_{t-1}^*$ is true at each t and the price change process is:

$$\Delta P_t = \Delta P_t^* + (I_t - I_{t-1})\frac{s}{2} = (I_t - I_{t-1})\frac{s}{2}.$$

Having assumed that I_t is IID, we have that:

$$Var(\Delta P_t) = \frac{s^2}{2}, \tag{1.4}$$

$$Cov(\Delta P_{t-1}, \Delta P_t) = -\frac{s^2}{4}, \tag{1.5}$$
$$Cov(\Delta P_{t-k}, \Delta P_t) = 0, \qquad k > 1$$

$$Corr(\Delta P_{t-1}, \Delta P_t) = -\frac{1}{2}. \tag{1.6}$$

Even though ΔP_t^* is fixed through time, we can notice that ΔP_t displays volatility and negative lag-1 serial correlation. This is due to the bid-ask bounce. Intuitively, if the fundamental value is fixed, the price can assume only two values: the bid and the ask price. If, currently, the price is either the bid or the ask, then the price change between the current value and the previous value can only be equal to zero or to the value of the spread s, and the price change between the current and the next price can only be zero or s. Both the volatility and the covariance depend on the amount of the spread, so that the bigger the spread, the bigger the variance and the smaller the autocovariance. Anyway, they change in a proportional way, so that the correlation remains equal to $-\frac{1}{2}$, for every value of s.

Now, let us relax the assumption about P_t^* fixed and consider the case in which P_t^* follows a random walk, meaning that its increments are not autocorrelated and are independent of I_t or, in other words, $\Delta P_t^* = P_t^* - P_{t-1}^* = \epsilon_t$, where $\epsilon_t \sim WN(0, \sigma^2)$. Then, 1.5 still holds, but the variance 1.4 and, therefore, the first-order serial correlation 1.6 changes into:

$$Var(\Delta P_t) = \sigma^2 + \frac{s^2}{2},$$

$$Corr(\Delta P_{t-1}, \Delta P_t) = -\frac{s^2/4}{(s^2/4) + \sigma^2} \leq 0.$$

For a positive spread, the amount of the autocorrelation is reduced, but it still remains negative.

1.3 Other empirical characteristics

Numerous studies have empirically shown that financial data, if recorded at high frequency, presents some particular features: the returns are usually leptokurtic or fat tailed, as the frequency gets higher the kurtosis tends to increase, the series shows intra-day volatility clustering, asset returns are usually serially correlated and volatility has a deterministic intra-day behaviour.

The present section discusses the major empirical properties of high-frequency data:

1. *Discreteness of prices and returns.*
   Transaction prices are quoted in a discrete way. The variance of a process over long time horizon is usually quite large if compared to the magnitude of the minimum change. However, this does not apply also for high frequency data since, for many data sets, the changes in price take only a handful of values. The minimum amount at which a price can move is called $tick$ and the price changes have to be a multiple of it. Moreover, regulators can set constraints that limit the price change. As a result, price changes are often limited to only a few possible outcomes. Therefore, discreteness can affect the computation of volatility, correlation and other measures that are small relative to the size of the tick. Furthermore, it can raise the kurtosis of the series of data. Indeed, high frequency data often show large kurtosis.

2. *Multiple transaction at the same second.*
   It can happen that more than one transaction, not necessarily with the same price, occur within a single second.

3. *The impact of macroeconomic news*
   When a major macroeconomic announcement occurs, prices modify very quickly, raising also the volatility and the volume. This causes volatility clustering, which has been empirically shown to be a typical

feature of high-frequency data.[1] In a perfect market, new information would be simultaneously acquired by every market participant and the prices would consequently change, reaching a new equilibrium value. However, in practice, not all the relevant news are available to everyone at the same moment and not all the market participants react at the information at the same speed. This results in variable time lag between a macroeconomic announcement and its reaction in the market.

4. *Correlation and persistency*

   Apart from return, high frequency trading variables are usually strongly auto-correlated. On the contrary, " intraday returns from traded assets are almost uncorrelated, with any important dependence usually restricted to a negative correlation between consecutive returns". (Taylor, 2005). Price discreteness and bid-ask bounce effect are the major causes of this dependence. Another factor could be the behaviour of some traders, who carry out many small transactions at the same time instead of a large one, aspiring to a better price overall. Indeed, this can cause a sequence of trades that shift the price in the same direction, leading to positive serial correlations at longer horizon. Being most of the high frequency characteristics persistent over time, showing dependence over a long range, there is the need of models that allow for long memory behaviour.

5. *Existence of diurnal patterns*

   Almost all high-frequency financial data exhibit intraday periodicities. Usually, for the majority of stock markets, the seasonality pattern of volatility, spreads, volume and the frequency of trades assume a U-shape. Volatility shows more activity after the opening of the market and just prior the closure, while in the lunch hour the intensity is lower. This cyclical pattern characterizes also the volume and the spreads. The time between the transactions, i.e. the durations, are usu-

[1] See, for example, Engle et al (1990) regarding foreign exchange rate returns or Hama et al (1990) for equity index returns.

> ally shorter near the open and the close time, meaning that the trades are more frequent in those time zones. The foreign exchange market, instead, works all day long, every day of the week. There is neither opening nor closing of this market, but it shows diurnal patterns too, usually following the more active periods of the day.

As an example, I considered the tick data for Starbucks during a trading day (2011/07/01). The total number of observations is 9331, spread over 6 hours and a half, since the trading hours go from 9:30 to 16:00. I obtained these data from the R package "highfrequency". They represent the changing in the value of the price every time a transaction occurs. I plotted the data in 1.2. We can notice from the figure that this difference in the price seems to be stationary around zero.

Then, I considered the number of transactions that occurred every five minutes, obtaining a time series of length 78 intervals, shown in figure 1.3. The plot shows the typical U-shape of high-frequency measures. This means that the trades are more frequent near the opening and closing times of the market, while in the lunch hours the tradings are thinner.

If we, instead focus on the duration, which is the period of time from a transaction to another one, we can notice the opposite behavior. In figure 1.4, we can see that the time interval between two trades is longer during the lunch hours than during the beginnin or ending hours of the trading day.

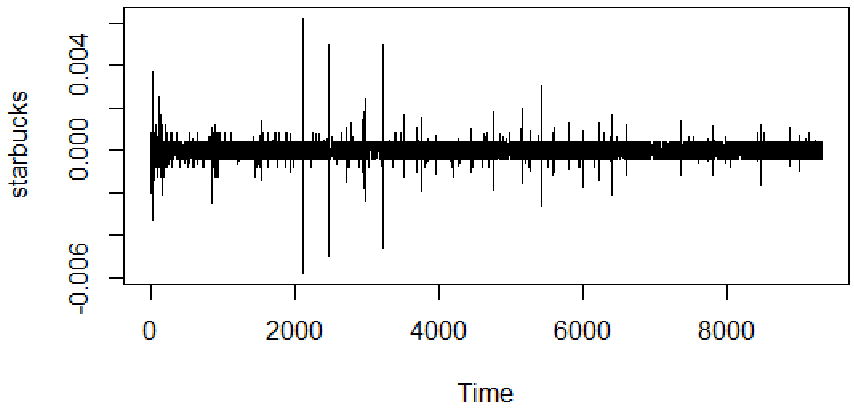


Figure 1.2: Tick data for Starbucks (2011/07/01).

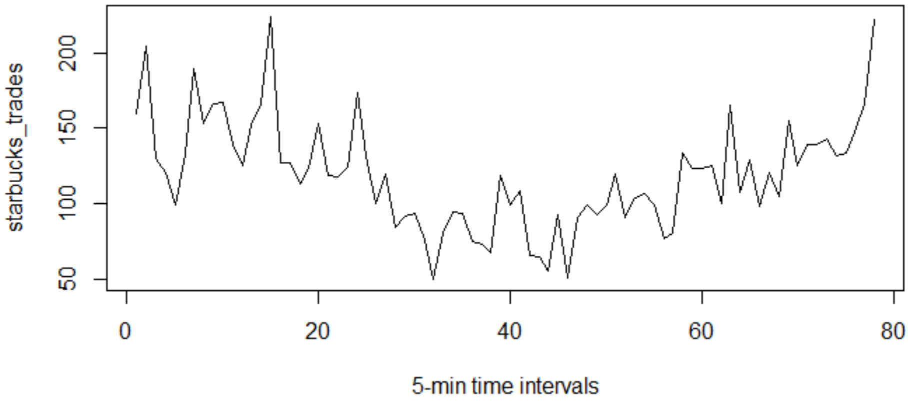


Figure 1.3: Number of transactions every 5-minute time interval for Starbucks (2011/07/01).

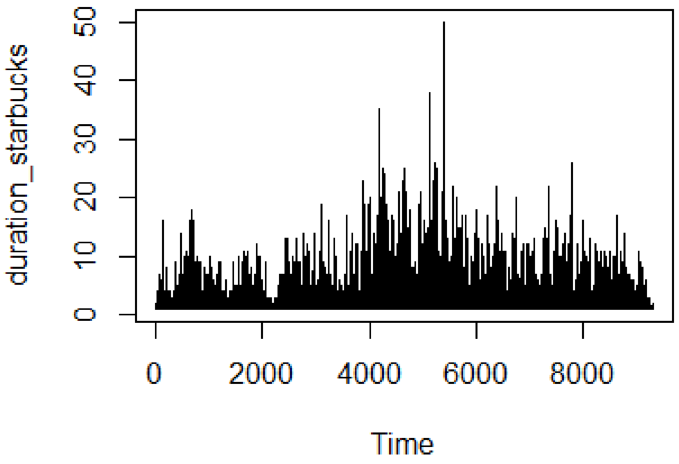


Figure 1.4: Time duration between transaction for Starbucks (2011/07/01).

Chapter 2

Volatility of high frequency data

The volatility is a statistical measure of the degree of variability of a trading price over some time periods: securities with high volatility are considered riskier. Indeed, a large variability means a large range of values that the security can potentially assume. Hence, its price can suddenly change in either directions. On the other hand, when the volatility is low, the value of the security, is more steady, without many fluctuations. In this case, the security is considered safer.

When computed using high frequency returns, the amount of the volatility changes consistently over the trading day and it is correlated with the high frequency volatility of trading volume and bid-ask spreads.

After recalling the theory about volatility, in this chapter I am going to compare the parametric and non parametric methods used in volatility modeling, in particular the ARCH and GARCH models for the first type and the Realized Volatility for the second one.

2.1 A recap of fundamental theory

The volatility measures the dispersion of the returns series of a security. It can be measured through the standard deviation or the variance between the returns.

Being X a continuous random variable, its expected value is defined as:

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx \tag{2.1}$$

with $f(x)$ the probability density function of X.
The variance of the random variable is defined as:

$$\begin{aligned}\sigma^2 &= \mathbb{V}\mathrm{ar}(X) \\ &= \mathbb{E}[(X-\mu)^2] \\ &= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx \\ &= \mathbb{E}(X^2) - \mu^2 \end{aligned} \tag{2.2}$$

Important properties of the variance of a random variable are listed below. Assume that a and b are constants and X and Y are two independent random variables. Then,

- $\mathbb{V}\mathrm{ar}(X) \geq 0$;
- $\mathbb{V}\mathrm{ar}(a + bX) = b^2\,\mathbb{V}\mathrm{ar}(X)$;
- $\mathbb{V}\mathrm{ar}(X + Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y)$.

Moreover, it can be proved that

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - [E(X)]^2. \tag{2.3}$$

The standard deviation is the positive square root of the variance and it is often denoted by σ.
The covariance is a measure of joint variability between two random variables X and Y. When its value is positive, the two random values show similar behavior, meaning that X and Y tend to move in the same direction. On the other hand, if it is negative, the two random variables tends to move in opposite directions. The covariance between X and Y is defined as:

$$Cov(X, Y) = \mathbb{E}[(X - \mu_x)(Y - \mu_y)]. \tag{2.4}$$

Considering the random variables X, Y,Z and the constant values a, b, c, and d, the main properties of the covariance are the following:

- $Cov(a + bX, c + dY) = bdCov(X, Y)$;

- $\mathbb{V}\mathrm{ar}(X, Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y) + 2Cov(X, Y)$;

- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$;

- $Cov(X, X) = \mathbb{V}\mathrm{ar}(X)$;

- $Cov(X, Y) = Cov(Y, X)$;

- $Cov(X, Y) = 0$, if X and Y are independent.

Another statistical measure of the degree of linear dependence between two variables is the correlation, which is defined as

$$\rho = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\,\mathbb{V}\mathrm{ar}(Y)}} \tag{2.5}$$

Its value can vary between -1 (perfect negative linear relationship) and 1 (perfect positive linear reationship); a value near zero indicates weak linear depencence. For the correlation coefficient it values that

$$Corr(a + bX, c + dY) = sign(bd)Corr(X, Y) \tag{2.6}$$

$$\text{where} \quad sign(bd) = \begin{cases} \;\;\, 1 \;\text{ if }\; bd > 0 \\ \;\;\, 0 \;\text{ if }\; bd = 0 \\ -\;1 \;\text{ if }\; bd < 0. \end{cases} \tag{2.7}$$

In time series analysis, data are collected sequentially over time and are modelled by a stochastic process, which is a sequence of random variables $\{Y_t : t = 0, \pm1, \pm2, \pm3, \dots\}$. For a stochastic process, the mean function μ_t, the autocovariance function $\gamma_{t,s}$ and the autocorrelation function $\rho_{t,s}$ are respectively:

- $\mu_t = \mathbb{E}(Y_t)$ for $t = 0, \pm1, \pm2, \dots$

- $\gamma_{t,s} = Cov(Y_t, Y_s)$
  $= \mathbb{E}[(Y_t - \mu_t)(Y_s - \mu_s)]$
  $= \mathbb{E}(Y_tY_s) - \mu_t\mu_s$ for $t, s = 0, \pm1, \pm2, \dots$

- $\rho_{t,s} = Corr(Y_t, Y_s) = \dfrac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}}$
  $= \dfrac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$ for $t, s = 0, \pm 1, \pm 2, \dots$

2.2 Parametric methods

Let $\{Y_t\}$ be a time series. Then, the conditional variance of Y_t given the past values (Y_{t-1},Y_{t-2},$\dots$) corresponds to the measure of the variability of the deviation of Y_t from $\mathbb{E}(Y_t|Y_{t-1}, Y_{t-2}, \dots)$, which is its conditional mean. The conditional variance is not constant, but can change with the current and past values of Y_t, being itself a random process.
In the latest years, the financial econometrics of volatility dynamics has seen an enormous growth, mainly due to the methodological advances in empirical finance. At first, developments were parametric, but recently literature has moved towards nonparametric approaches to volatility modeling. Parametric approaches rely on explicit models of the expected volatility; the main examples are the ARCH and GARCH class of models.

2.2.1 The ARCH model

The main assumption of the basic version of linear regression model is homoskedasticity, meaning that the expected value of the squared error terms is always constant. In case of heteroskedasticity, the error terms are expected to be larger in certain points than other. With a least squares regression the coefficient would still be unbiased, but the standard errors and the confidence intervals would be too restrictive. The Autoregressive Conditional Heteroscedasticity (ARCH) model, first proposed by Engle in 1982, treats heteroskedasticity as a variance to be modeled.
Consider a return series $\{r_t\}$, with $\mathbb{E}(r_t) = 0$. Let $\sigma^2_{t|t-1}$ be the conditional variance of r_t given past returns until time $t - 1$. Since r_t is observed, we can use the squared return r_t^2 as an unbiased estimator of $\sigma^2_{t|t-1}$. The ARCH model can be expressed as a regression with the conditional variance as the

dependent variable and the lagged squared returns as the regressors. As an example, the ARCH(1) model is defined as:

$$r_t = \sigma_{t|t-1}\epsilon_t \tag{2.8}$$

$$\sigma^2_{t|t-1} = \omega + \alpha r^2_{t-1} \tag{2.9}$$

where ϵ_t is a white noise with zero mean and unit variance, independent of r_{t-j}, $j = 1, 2, \ldots$, and α and ω are two unknown parameters. In order to ensure the nonnegativity of the squared returns, the conditions $0 < \alpha < \infty$ and $0 < \omega < \infty$ have to be satisfied. The weak stationarity is achieved with the condition $0 \leq \alpha < 1$ (Ling and McAleer, 2002).[1] Since $\sigma_{t|t-1}$ is known and ϵ_t has unit variance and is independent from past returns, we can show that:

$$\begin{aligned}
\mathbb{E}(r^2|r_{t-j}) &= \mathbb{E}(\sigma^2_{t|t-1}\epsilon^2_t|r_{t-j}) \\
&= \sigma^2_{t|t-1}\,\mathbb{E}(\epsilon^2_t|r_{t-j}) \\
&= \sigma^2_{t|t-1}\,\mathbb{E}(\epsilon^2_t) \\
&= \sigma^2_{t|t-1} \qquad\qquad j = 1, 2, \ldots
\end{aligned}$$

In the ARCH(1) model, the forecast of the future conditional variances only depend on the most recent squared return. Intuitively, adding more lagged squared returns in the model could increase the accuracy of the forecasting. It is then possible to parametrize the conditional volatility as an autoregressive model, in which it is defined as a linear combination of the p most recent squared returns, resulting in the ARCH(q) model:

$$\sigma^2_{t|t-1} = \omega + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + \cdots + \alpha_q r^2_{t-q} \tag{2.10}$$

2.2.2 The GARCH model

A more general class of processes was introduced by Bollerslev in 1986 as an extension to the ARCH process: the Generalized Autoregressive Con-

[1] A process can show conditional heteroscedasticity even if it is weakly stationary. Indeed, the weak stationarity imply that the mean function is constant over time and that the autocovariance function, for a fixed lag, remains the same at each point in time ($\gamma_{t,t-k} = \gamma_{0,k}$, for all time t and lag k)

ditional Heteroskedastic process (GARCH). This class of models manages to have both a much more flexible lag structure and a longer memory. It is constructed adding to the ARCH(q) model the p lagged values of the conditional variance, resulting in

$$r_t = \sigma_{t|t-1}\epsilon_t$$

$$\sigma^2_{t|t-1} = \omega + \beta_1\sigma^2_{t-1|t-2} + \cdots + \beta_p\sigma^2_{t-p|t-p-1} + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + \cdots + \alpha_q r^2_{t-q}$$

where p is the number of lags of the conditional variance and q is the ARCH order. We can notice that, if $p = 0$, the model reduces to an ARCH(q), while, if both p and q are null, r_t is equal to a white noise. The GARCH(p,q) model can be re-expressed through the backshift L notation [2] as

$$(1 - \beta_1 L - \cdots - \beta_p L^p)\sigma^2_{t|t-1} = \omega + (\alpha_1 L + \cdots + \alpha_q L^q) r^2_t \qquad (2.11)$$

Often the coefficients are restricted to positive values in order to assure a nonnegative conditional variance. Moreover, a necessary and sufficient condition for the GARCH(p,q) process to be weakly stationary is the following:

$$\sum_{i=1}^{max(p,q)} (\beta_i + \alpha_i) < 1.$$

2.3 Nonparametric methods

In contrast to the parametric methods, the nonparametric ones are data-driven measurements that manage to consistently estimate the volatility without any functional form assumption regarding the stochastic process governing the return series. The most obvious of these estimates is computed as the ex-post squared return. Anyway, even though it is unbiased, it also contains a lot of noise, especially when data are collected at a high frequency. Generally, under the stationarity and ergodicity assumptions, the nonparametric

[2] The backshift or lag operator L allows to shift time back of a lag, creating a new time series, $LY_t = Y_{t-1}$.

volatility estimates are consistent when obtained as the sample averages of squared returns, which are sampled at increasingly low frequency. As an example, I am going to discuss about the Realized Volatility.

2.3.1 The realized volatility

In order to obtain a quite accurate estimation of the volatility, we can rely on returns collected at high frequency. Anyway, the results assume a continuous sample path diffusion process, which in practice cannot be satisfied by the high-frequency return series. The nonparametric realized volatility has been empirically showed to be able to model volatility avoiding the data complications, outperforming the results of the Garch model.[3]

2.3.1.1 Price processes and financial returns

When data are collected over short intervals, the underlying price process has to be modeled in continuous time. Usually, in the asset pricing theory, the logarithmic price process is assumed to take the form of an Ito's semimartingale. Let $W(t)$, $t \geq 0$, be a brownian motion. Then, an Ito process is a stochastic process of the form:

$$X(t) = X(0) + \int_0^t \Delta(u) dW(u) + \int_0^t \Theta(u) du \tag{2.12}$$

where $X(0)$ is non random and $\Delta(u)$ and $\Theta(u)$ are adapted stochastic processes. In differential form, it is expressed as

$$dX(t) = \Delta(t)dW(t) + \Theta(t)dt \tag{2.13}$$

A process is said to be a semimartingale when it is the sum of a finite variation càdlàg drift process and an adapted càdlàg local martingale, where:

- a càdlàg is a function which is everywhere right-continuous and left bounded;

[3] As suggested by Andersen and Bollerslev (1998), Andersen, Bollerslev, Diebold and Labys (2001), Barndorff-Nielsen and Shephard (2002).

- a drift is the change of the mean value of a stochastic process;
- a martingale is a process for which the conditional expectation of its future value is equal to the present value, given the information of all past values.

Then, the evolution of logarithmic asset price $p(t)$ is assumed to be equal to

$$dp(t) = \mu(t)dt + \sigma(t)dW \tag{2.14}$$

or

$$p(t) - p(t-1) \equiv r(t) = \int_{t-1}^{t} \mu(s)ds + \int_{t-1}^{t} \sigma(s)dW(s) \qquad s \geq 0 \tag{2.15}$$

where W_t denotes a standard brownian motion, $\mu(s)$ is a finite variation càdlàg drift process and $\sigma(s)$ is the adapted càdlàg volatility process associated to $p(t)$.

Let the unit interval be one trading day, T the number of days considered, $t = \{1, 2 \ldots, T\}$, m the number of times per day the prices have been sampled, such that $m \cdot T$ is the total number of returns. Then, the asset return over $[\tau - \frac{1}{m}, \tau]$ is given by:

$$r_\tau = p(\tau) - p(\tau - \frac{1}{m}) \qquad \tau = \frac{1}{m}, \frac{2}{m}, \ldots, T.$$

Note that when m is higher than one, the returns are computed at a high-frequency, while when m is lower than one, we obtain interdaily returns.

2.3.1.2 Quadratic variation and integrated variance

Being X and Y two semimartingales, then the quadratic variation and covariation are respectively defined as:

$$[X, X](T) = \lim_{\|\Pi\| \to 0} \sum_{k=1}^{n} (X(t_k) - X(t_{k-1}))^2$$

$$[X, Y](T) = \lim_{\|\Pi\| \to 0} \sum_{k=1}^{n} [X(t_k) - X(t_{k-1})][Y(t_k) - Y(t_{k-1})],$$

where Π is the full grid containing all the observation points, $\Pi = \{t_0, t_1, \ldots, t_n\}$, and $0 = t_0 < t_1 < \cdots < t_n = T$.
It is well known that:

- for most securities, the drift is close to zero when the time interval is small;
- considering a Brownian motion $W(t)$, then, its quadratic variation is equal to $[W, W](T) = T$ for all $T \geq 0$ almost surely [4]
- considering and Ito integral $I(t)$, $I(t) = \int_0^t \Delta(u)dW(u)$, where $\Delta(t)$ is an adapted stochastic process. Then, its quadratic variation is equal to $[I, I](T) = \int_0^t \Delta^2(u)du$.

Having assumed that the logarithmic asset price follows the Ito semimartingale of equation 2.15 , we can then conclude that the corresponding quadratic variation is:

$$Qvar(t,h) \equiv [p,p](t) - [p,p](t-h) = \int_{t-h}^{t} \sigma^2(s)ds \tag{2.16}$$

where h is an arbitrarily fixed positive quantity. Since we consider variables measured on a daily interval, for notational convenience we suppress the h subscript ($h = 1$). Then,

$$Qvar(t) \equiv [p,p](t) - [p,p](t-1) = \int_{t-1}^{t} \sigma^2(s)ds$$

A natural volatility measure is the integrated volatility, which is computed as the integral of the instantaneous volatility over an interval of interest. Being the return process 2.15 continuous, we can notice that its quadratic variation 2.16 is equal to the integrated volatility:

$$IV(t) \equiv \int_{t-1}^{t} \sigma^2(s)ds = Qvar(t).$$

[4] The terminology *almost surely* indicates that there is a chance for the equality to be not true, but this chance has zero probability.

2.3.1.3 The realized volatility as an estimator of the quadratic variation

If we choose an arbitrarily high sampling frequency, we can get the *realized volatility* summing the intraday squared returns over the period we are interested. Then, the daily realized volatility is defined as:

$$RV_t = \sum_{j=0}^{m-1} r^2_{t-j\cdot\frac{1}{m}}$$

where m indicates the number of observations collected in a day. For any fixed m, the realized volatility is directly observable and, as it follows by the theory of the quadratic variation [5], as m tends to infinity, it converges uniformly in probability to the quadratic variation:

$$\lim_{m\to\infty} RV_t \to Qvar(t),$$

in the mean square sense. It results that, if we can sample frequently enough, it is possible to estimate consistently the integrated volatility through the realized volatility.
Andersen and Bollerslev (1998), Andersen, Bollerslev, Diebold and Labys (2001), Barndorff-Nielsen and Shephard (2002a,b), and Meddhai (2002), among others, suggested using the nonparametric realized volatility in order to avoid high-frequency data complications, keeping however most of the relevant information useful for measuring, modeling and forecasting volatilities. It has also been empirically shown that simple realized volatility models perform better than the GARCH models in out-of-sample forecasting.

2.3.1.4 Handling microstructure noise

According to the computations of the previous section, higher the frequency of the data and better is the estimation of the integrated volatility provided by the realized volatility. On the contrary, in the empirical finance

[5] See Andersen and Bollerslev (1998), Andersen, Bollerslev, Diebold and Labys (2001), Barndorff-Nielsen and Shephard (2002), Comte and Renault (1998)

literature it is often stated that the return series should not be sampled too often, in order to avoid that market microstructure noise causes a relevant bias. Indeed, as the frequency get higher, the returns change less, but the microstructure noise remains of the same level. It results that the observed changes in the return process are very contaminated by the noise and the realized volatility does not converge to the integrated volatility. For this reason, many authors suggest to use data that are sampled over longer time horizons, so that the impact of microstructure noise is reduced and the estimates obtained are more reasonable. Usually, one selects an arbitrary sparse sampling frequency, such as one every 5 minutes, but other methods have been proposed in the literature in order to reduce errors due to microstructure noise in calculating the realized volatility.

A first method (see Zhang et al., 2005; Bandi and Russel, 2008) consists in finding the optimal sampling interval that minimizes the mean square error when estimating the integrated volatility. It is a way to find a compromise in the trade-off between the accuracy in computing the realized volatility and the bias created by the microstructure noises. The number of sparse sampling intervals n_{sparse} should, then, be chosen such that $\partial MSE/\partial n_{sparse} \approx 0$. If n_{sparse} results to be higher than the original number of sampling intervals n, one should simply use n.

Another sampling method is a two-scale procedure proposed in Zhang et al. (2005). It is shown that it is possible to obtain a consistent and asymptotically unbiased estimator of the integrated volatility proceeding in two steps: first, we use sub-sampling and, then, we correct the bias. Sub-sampling means that, instead of using the entire sample of observations to compute the high frequency returns (for example, at a 1 second interval), we construct more than one time series of logarithmic returns at a lower frequency (5 minutes). Then, the first time series starts with the first observation, the second time series with the second observation and so on. In this way, we will get K time series of logarithmic returns with the same increment of five minutes, but different starting points, as we can see in Figure 2.1. Obviously, the number of time series K is equal to the number of observations in the first 5 minutes. We then compute for each of time series the realized volatility and take a simple average of the estimates obtained. This averaged realized volatility RV_t^{avg} is then computed as

$$RV_t^{avg} = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{n_k} r_{k,j}^2$$

where K is the number of subsamplings, n_k the number of logarithmic returns for the k-th subsampling, r_j the logarithmic return for the j-th transaction and $r_{k,j}$ the j-th logarithmic returns of the k-th subsampling. However, in Zhang et al. (2005) it is shown that RV_t^{avg} is a biased estimator, hence a correction is needed. Let n be the total number of logarithmic returns, $\bar{n}$ be the average number of n_k, such that $\bar{n} = \frac{n}{K}$. Define ϵ as the noise component, such that, summed to the efficient price, which is the price governed by the Ito process, gives out the observed price. Then, the bias of RV_t^{avg} is equal to $2n\,\mathbb{E}(\epsilon^2)$, which can be consistently approximated through the realized volatility computed using all the data available, $RV_t^{(all)} = \sum_{j=1}^{n} r_j^2$. Indeed,

$$\mathbb{E}(\epsilon^2) \sim \frac{1}{2n} RV_t^{(all)}.$$

We can finally obtain a bias-adjusted estimator, named Two Scale Realized Volatility, as

$$TSRV_t = RV_t^{avg} - \frac{\bar{n}}{n} RV_t^{(all)}.$$

The average of the realized variance is used in order to get a better estimate since more prices are used, while the quantity $\frac{\bar{n}}{n} RV_t^{(all)}$ is needed as a bias correction. This estimator reduces the impact of microstructure noise and is consistent even in presence of jumps.

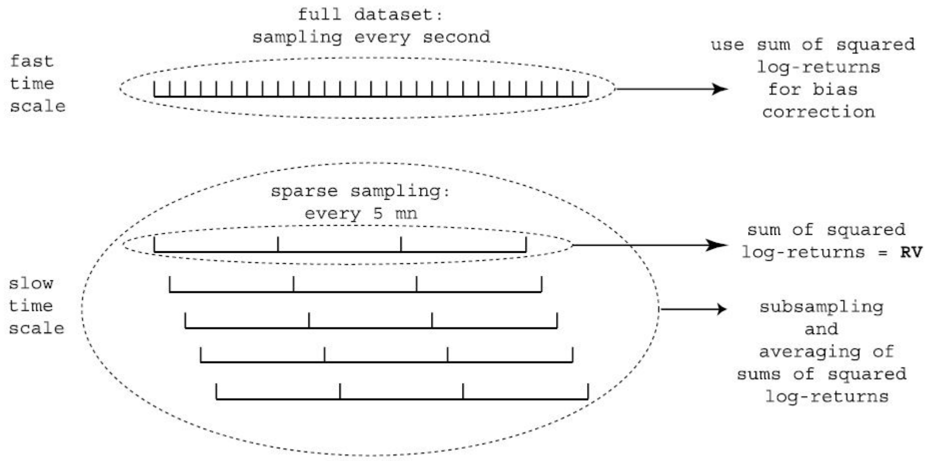


Figure 2.1: Construction of the Two Scale Realized Volatility

Chapter 3

Modelling volatility: the HAR model

The realized volatility provides a model-free measure of the quadratic variation. Since the purpose of this thesis is to use high-frequency data information to predict future volatilities, once a realized volatility time series is obtained we also need a model to be applied to it.
In this chapter I am going to discuss an interesting model that has been proposed by Fulvio Corsi: the *Heterogeneous Autoregressive model of the Realized Volatility* (HAR-RV). Its purpose is to be parsimonious and easy to estimate, but, at the same time, to reproduce the main characteristics of high-frequency data, with good out-of-sample forecasting performances.

3.1 The HAR-RV model

The *Heterogeneous Autoregressive model for Realized Volatility* has been proposed by Fulvio Corsi in 2003. The motivation for this model stands on the *Heterogeneous Market Hypothesis* presented by Müller et al. (1993). This hypothesis states that "The market is heterogeneous, with a “fractal” structure of the participants’ time horizons as it consists of short-term, medium-term and long-term components. Each such component has its own reaction time to news, related to its time horizon and characteristic dealing frequency." Indeed, the market partecipants can be of different kind, like the foreign exchange dealers or market makers, usually related to high dealing frequencies, or central banks, commercial organizations and pension funds investors, who deal with lower frequencies. Hence, Corsi arrived at the conclusion that "agents with different time horizons perceive, react to, and cause different types of volatility components. We can identify three primary volatility components: the short-term traders with daily or higher trading frequency, the medium-term investors who typically rebalance their positions weekly, and the long-term agents with a characteristic time of one or more months." Empirical experience suggests that this heterogeneous market structure creates a volatility cascade from lower to higher frequencies. This is because volatility over long time intervals has more influence on volatility over short intervals than conversely, showing an asymmetric behaviour. From an economic point of view, since long-term volatility affects the forecasting of trend and risk, its variations change the behaviour of short-term agents too, generating short-term volatility. To the contrary, long term agents do not modify their trading strategies according to short-term volatility changes. This reason has induced Corsi to propose a volatility cascade model composed by three heterogeneous volatility components.

3.1.1 The model

With this model, the realized volatility is parametrized as a linear combination of the lagged realized volatilities computed over different horizons. The author proposed a simplified model with only three intervals, one day ($1d$), one week ($1w$) and one month ($1m$), but more components could be easily added. The daily, weekly and monthly *latent partial volatility*, which is the volatility created by a particular market component, is respectively $\tilde{\sigma}_t^{(d)}$, $\tilde{\sigma}_t^{(w)}$ and $\tilde{\sigma}_t^{(m)}$.

The high-frequency return process is defined by the component with the highest frequency as

$$r_t = \sigma_t^{(d)} \epsilon_t,$$

where $\sigma_t^{(d)} = \tilde{\sigma}_t^{(d)}$ is the integrated volatility and $\epsilon_t \sim NID(0,1)$.
The unobserved partial volatility $\tilde{\sigma}_t^{(.)}$ at each time interval is modelled as a function of the previous period realized volatility computed at the same time interval and of the expectation of the next period longer-term partial volatility. Then, we can say that the partial volatilities follow an "almost AR(1)" model, meaning that, having on the left-hand side the latent volatility, on the right-hand side the corresponding realized volatility appears instead. Hence, the first term regards the AR(1) component, while the second consists in the hierarchical component. Let $RV_t^{(d)}$, $RV_t^{(w)}$ and $RV_t^{(m)}$ be respectively the daily, weekly and monthly realized volatilities. They are defined as the average of the daily quantities to allow the comparison between them. Then,

$$\begin{aligned} RV_t^{(w)} &= \frac{1}{5}(RV_t^{(d)} + RV_{t-1d}^{(d)} + \cdots + RV_{t-4d}^{(d)}) \\ RV_t^{(m)} &= \frac{1}{22}(RV_t^{(d)} + RV_{t-1d}^{(d)} + \cdots + RV_{t-21d}^{(d)}). \end{aligned}$$

The HAR-RV model is defined as

$$\begin{aligned} \sigma_{t+1m}^{(m)} &= c^{(m)} + \phi^{(m)} RV_t^{(m)} + \tilde{\omega}_{t+1m}^{(m)}, \\ \sigma_{t+1w}^{(w)} &= c^{(w)} + \phi^{(w)} RV_t^{(w)} + \gamma^{(w)}\, \mathbb{E}_t[\sigma_{t+1m}^{(m)}] + \tilde{\omega}_{t+1w}^{(w)}, \\ \sigma_{t+1d}^{(d)} &= c^{(d)} + \phi^{(d)} RV_t^{(d)} + \gamma^{(d)}\, \mathbb{E}_t[\sigma_{t+1w}^{(w)}] + \tilde{\omega}_{t+1d}^{(d)}, \end{aligned} \tag{3.1}$$

where "the volatility innovations $\tilde{\omega}_{t+1m}^{(m)}$, $\tilde{\omega}_{t+1w}^{(w)}$ and $\tilde{\omega}_{t+1d}^{(d)}$ are contemporaneously and serially independent zero-mean nuisance variates with an appropriately truncated left tail to guarantee the positivity of partial volatilities" (Corsi, 2003). The economical explanation is that each volatility component is based on the current realized volatility and on the expectation of the longer horizon volatility, which, as it has been said before, affects the future values of the shorter-term volatilities, due to the asymmetric propagation of the volatility.

Being $\sigma_t^{(d)} = \tilde{\sigma}_t^{(d)}$, by substitution the cascade model 3.1 can be written as a three-factor stochastic volatility model, with the past realized volatilities computed at different frequencies as factors:

$$\sigma_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \tilde{\omega}_{t+1d}^{(d)}. \tag{3.2}$$

with $c = c^d + c^w + c^m$ and $\tilde{\omega}_{t+1d}^{(d)}$ incorporates all the noises.
We can notice that, ex post, it is possible to define $\sigma_{t+1d}^{(d)}$ as

$$\sigma_{t+1d}^{(d)} = RV_{t+1d}^{(d)} + \omega_{t+1d}^{(d)}, \tag{3.3}$$

where $\omega_{t+1d}^{(d)}$ includes both the latent daily volatility computation and estimation errors. From equation 3.3, it is clear that the realized volatility is not considered an error-free estimator. However, in order to ensure that $\omega_{t+1d}^{(d)}$ is a zero mean nuisance the property of consistency of the realized is not sufficient. Indeed, it values that $\mathbb{E}[\omega_{t+1d}^{(d)}] = 0$ only if $\mathbb{E}[\sigma_{t+1d}^{(d)}] = \mathbb{E}[RV_{t+1d}^{(d)}]$. To make it possible the realized volatility has to be an unbiased estimator of the integrated volatility, hence it is necessary a correction of the microstructure effects when computing the realized volatilities. If we substitute 3.3 in 3.2, we obtain the time series representation

$$RV_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1d}, \tag{3.4}$$

with $\omega_{t+1d} = \tilde{\omega}_{t+1d}^{(d)} - \omega_{t+1d}^{(d)}$. Equation 3.4 can be seen as a particular autoregessive model in the realized volatility, that considers realized volatilities computed over different time intervals. To estimate the parameters, we can consider as observed the realized volatilities on the right-hand side of eq. 3.4 and apply a simple linear regression. In this way, the consistency of the resulting OLS regression estimators is ensured. Moreover, they are normally distributed.

3.2 The jump effect

In economics, the process most used to model prices is the Brownian motion, which represents a smooth and slowly mean-reverting continuous sample path process. Anyway, there is no satisfying economic theory that proves that prices must follow a process characterized by a continuous sample path. Indeed, recent parametric studies have proposed to allow for jumps when estimating stochastic volatility process or when pricing derivatives.[1] Hence, in this section, we are going to add a less persistent jump component to the Brownian semimartingale process.

3.2.1 The jump processes

Let $p(t)$ be a logarithmic asset price at time t and $\kappa(t) \equiv p(t) - p(t-)$ the size of the discrete jumps. The stochastic differential equation usually used to express the continuous-time jump diffusion process is

$$dp(t) = \mu(t)dt + \sigma(t)dW(t) + \kappa(t)dq(t), \qquad 0 \leq t \leq T,$$

where "$\mu(t)$ is a continuous and locally bounded variation process, $\sigma(t)$ is a strictly positive stochastic volatility process with a sample path that is right continuous and has well-defined left limits (allowing for occasional jumps in volatility), $W(t)$ is a standard Brownian motion, and $q(t)$ is a counting process[2] with (possibly) time-varying intensity $\lambda(t)$."(Andersen at al. (2007)). Knowing that the cumulative return process is $r(t) \equiv p(t) - p(0)$, we obtain its quadratic variation as

$$[r, r]_t = \int_0^t \sigma^2(s)ds + \sum_{0<s\leq t} \kappa^2(s),$$

We can notice that the summation of the squared jumps that occurred during the time interval $[0, t]$ equals zero in the case of absence of jumps, so that the quadratic variation becomes the integrated volatility of the continuous sample path component.

[1] See, for example, Andersen, Benzoni, and Lund (2002), Bates (2000), Chan and Maheu (2002), Chernov, Gallant, Ghysels, and Tauchen (2003), Drost, Nijman, and Werker (1998), Eraker (2004), Eraker, Johannes, and Polson (2003), Johannes (2004), Johannes, Kumar, and Polson (1999), Maheu and McCurdy (2004), Khalaf, Saphores, and Bilodeau (2003), and Pan (2002).

[2] A counting process is a stochastic process with values that are non-negative, integer and non-decreasing.

Let Δ be the length of a high frequency period, such that $\Delta = \frac{1}{m}$, where m is the number of observations in a day. Then,recall that the realized variance can be defined as

$$RV_{t+1}(\Delta) \equiv \sum_{j=1}^{1/\Delta} r^2_{t+j\cdot\Delta}, \tag{3.5}$$

where $r_t \equiv p(t) - p(t-\Delta)$ is the discretely sampled Δ-period return at time t. For ease of notation, the daily time interval is normalized to unity and $1/\Delta$ is assumed to be an integer. Then, for $\Delta \to 0$,

$$RV_{t+1}(\Delta) \to \int_t^{t+1} \sigma^2(s)ds + \sum_{t<s\leq t+1} \kappa^2(s). \tag{3.6}$$

Hence, the realized variance is a consistent estimator of the integrated variance only in the case of absence of jumps.

3.2.2 The realized bipower variation

In Andersen et al. (2007), the nonparametric realized variance approach is further advanced by using a procedure that separately considers and measures the two components of the quadratic variation process: the continuous sample path variation and the jump part. This is possible thanks to the asymptotic properties proved by Barndorff-Nielsen and Shephard (2004). This procedure involves the so-called *bipower variation*, which is a measure constructed by summing cross products of adjacent high-frequency absolute returns. The standardized realized bipower variation is defined as

$$BV_{t+1}(\Delta) \equiv \mu_1^{-2} \sum_{j=2}^{1/\Delta} |r_{t+j\cdot\Delta}||r_{t+(j-1)\cdot\Delta}|, \tag{3.7}$$

where μ_1 is the standardization factor corresponding to the mean of the absolute value of Z, which is considered to be a standard normally distributed random variable: $\mu_1 \equiv \sqrt{(2/\pi)} = \mathbb{E}(|Z|)$. As shown in Barndorff-Nielsen and Shephard (2004), it values that

$$BV_{t+1}(\Delta) \to \int_t^{t+1} \sigma^2(s)ds, \tag{3.8}$$

for $\Delta \to 0$.

Then, we can use the realized bipower variation to solve the consistency problem of the realized variation in presence of jumps. Indeed, if we combine the results of equations 3.6 and 3.8, we obtain

$$RV_{t+1}(\Delta) - BV_{t+1}(\Delta) \to \sum_{t<s\leq t+1} \kappa^2(s) \tag{3.9}$$

Hence, the difference between the realized and the bipower variations is a consistent estimator for the jump component.

We can easily note that, despite the jump component is defined as a summation of squared variables, the left-hand side of equation 3.9 could be negative. In order to ensure the non-negativity of the daily estimates, Barndorff-Nielsen and Shephard (2004) suggested a truncation as follows

$$J_{t+1}(\Delta) \equiv max[RV_{t+1}(\Delta) - BV_{t+1}(\Delta), 0]. \tag{3.10}$$

3.2.3 The HAR-RV-J model

To take into account the jump component, the HAR-RV model for one-day volatilities, described by equation 3.4, has been modified just by adding the corresponding time series. Moreover, it can be extended to longer horizons, so that the dependent variable becomes $RV_{t,t+h}$. Then, the new model resulting, the so-called HAR-RV-J, is defined as

$$RV^{(d)}_{t+h} = c + \beta^{(d)} RV^{(d)}_t + \beta^{(w)} RV^{(w)}_t + \beta^{(m)} RV^{(m)}_t + \beta^{(j)} J_t + \omega_{t+h} \tag{3.11}$$

"With observations every period and longer forecast horizons, or $h > 1$, the error term will generally be serially correlated up to (at least) order $h - 1$. This will not affect the consistency of the regression coefficient estimates, but the corresponding standard errors of the estimates obviously need to be adjusted" (Andersen et al. (2007)). To this purpose,the Barlett/Newey-West heteroskedasticity consistence covariance matrix estimator is used.

3.2.4 Measurement error correction

Thus far, jumps have been estimated as the difference between the realized variation and the bipower variation, relying on increasingly finer temporal grid ($\Delta \to 0$). Practically, a fixed sampling frequency ($\Delta > 0$) is adopted, hence the measurements are subject to errors. Part of these errors is eliminated when the nonnegativity truncation of equation 3.10 is applied. Anyway, there could be some positive nonzero jumps that are so small that should be considered as measurement errors, or part of the continuous component.

In order to achieve this distinction, we can rely on the distributional results of Barndorff-Nielsen and Shephard (2004, 2006) and of Barndorff-Nielsen, Graversen, Jacod, et al. (2006). Under regularity, frictionless market conditions, they proved that, in case of no jumps, for $\Delta \to 0$,

$$\Delta^{-1/2} \frac{RV_{t+1}(\Delta) - BV_{t+1}(\Delta)}{[(\mu_1^{-4} + 2\mu_1^{-2} - 5) \int_t^{t+1} \sigma^4(s)ds]^{1/2}} \Rightarrow N(0,1). \tag{3.12}$$

Then, a large standardized difference between the realized and the bipower volatility has to be considered as evidence of a significant jump.
We can notice that, in the denominator of the fraction in 3.12, the component $\int_t^{t+1} \sigma^4(s)ds$ is the *integrated quarticity*, which is not observable. It can be consistently estimated through the *realized tripower quarticity*, which is the normalized sum of the product of n subsequent returns, in absolute value, with $n \geq 3$, raised to the power of $\frac{4}{n}$:

$$TQ_{t+1}(\Delta) \equiv \Delta^{-1}\mu_{4/3}^{-3} \sum_{j=3}^{1/\Delta} |r_{t+j\cdot\Delta,\Delta}|^{4/3} |r_{t+(j-1)\cdot\Delta,\Delta}|^{4/3} |r_{t+(j-2)\cdot\Delta,\Delta}|^{4/3}, \tag{3.13}$$

with $\mu_{4/3} \equiv 2^{2/3} \cdot \Gamma(7/6) \cdot \Gamma(1/2)^{-1} = \mathbb{E}(|Z|^{4/3})$. It is possible to show that

$$TQ_{t+1}(\Delta) \Rightarrow \int_t^{t+1} \sigma^4(s)ds. \tag{3.14}$$

From equations 3.12, 3.13 and 3.14, we can obtain the test statistic

$$W_{t+1}(\Delta) \equiv \Delta^{-1/2} \frac{RV_{t+1}(\Delta) - BV_{t+1}(\Delta)}{[(\mu_1^{-4} + 2\mu_1^{-2} - 5) TQ_{t+1}(\Delta)]^{1/2}} \tag{3.15}$$

to compare to a normal distribution in order to detect the significant jumps. The null hypothesis is the absence of jumps in the price process, while the alternative hypothesis is the presence of jumps.
Anyway, this test statistic has a negative side: in Huang and Tauchen (2005), a Monte Carlo analysis showed that the $W_{t+1}(\Delta)$ statistic has a pitfall when applying the asymptotic approximation over an entire sample. Indeed, the microstructure noise biases the test against detecting jumps causing a tendency of over-rejecting the null hypothesis. They proposed another statistic which can be approximated very closely to a standard normal distribution, defined as

$$Z_{T+1}(\Delta) \equiv \Delta^{-1/2} \frac{[RV_{t+1}(\Delta) - BV_{t+1}(\Delta)]RV_{t+1}(\Delta)^{-1}}{[(\mu_1^{-4} + 2\mu_1^{-2} - 5)max\{1, TQ_{t+1}(\Delta)BV_{t+1}(\Delta)^{-2}\}]^{1/2}}$$

As reported in Andersen at al. (2007), "we naturally identify the significant jumps by the realizations of $Z_{t+1}(\Delta)$ in excess of some critical value, say Φ_α,

$$J_{t+1,\alpha}(\Delta) \equiv I[Z_{t+1}(\Delta) > \Phi_\alpha] \cdot [RV_{t+1}(\Delta) - BV_{t+1}(\Delta)], \tag{3.16}$$

where $I[\cdot]$ denotes the indicator function". In other words, if the test is higher than the critical value, the jump is significant and equal to the difference between the realized volatility and the bipower volatility, otherwise such difference is considered as part of the continuous component of the price process. Since the total realized volatility has to converge to the sum of the continuous and the jump components (3.6), we can obtain the continuous component as

$$C_{t+1,\alpha}(\Delta) \equiv I[Z_{t+1}(\Delta) \leq \Phi_\alpha] \cdot RV_{t+1}(\Delta) + I[Z_{t+1}(\Delta) > \Phi_\alpha] \cdot BV_{t+1}(\Delta). \tag{3.17}$$

Roughly speaking, this means that when the test $Z_{t+1}(\Delta)$ is lower than the critical value Φ_α we consider the entire amount $RV_{t+1}(\Delta)$ as the continuous component, while if the test is higher than the critical value, it is assumed that there is a significant jump, so the continuous path is composed only by the bipower variation, which can be thought as the difference between the realized volatility and the jump.

Market microstructure noise do not only affect the consistency of the realized volatility, as explained in section 2.1.4. Indeed, if $\Delta \to 0$, also the bipower variation of equation 3.7 is biased. Moreover, the microstructure effects lead to the presence of first order autocorrelation in the return series, resulting in another source of bias in 3.7. As reported in Huang and Tauchen (2005), this implies that the jump test statistic $W_{t+1}(\Delta)$ of equation 3.15 will be negatively biased in finding jumps. They proved that, for small values of Δ, the test tends to under-reject the null hypothesis and this behavior is worsened by the size of the variance of the noise term $v(t)$.
In Andersen, Bollerslev and Diebold (2007), a way to obviate to the problem of the presence of spurious autocorrelation in the equation of observed returns is illustrated. They introduced the use of a modified realized bipower variation, in which the returns are staggered, leading to the following equation:

$$BV_{1,t+1}(\Delta) \equiv \mu_1^{-2}(1-2\Delta)^{-1} \sum_{j=3}^{1/\Delta} |r_{t+j\cdot\Delta,\Delta}||r_{t+(j-2)\cdot\Delta,\Delta}|, \tag{3.18}$$

in which $(1-2\Delta)^{-1}$ is the normalization factor due to the loss of 2 observations because of the staggering. If we increase the lag length between the returns, it is possible to break higher-order autocorrelations. Analogously, the tripower quarticity used to estimate the integrated quarticity can be computed using staggered returns in the following way:

$$\mathrm{TQ}_{1,t+1}(\Delta) \equiv \Delta^{-1}\mu_{4/3}^{-3}(1-4\Delta)^{-1} \sum_{j=5}^{1/\Delta} |r_{t+j\cdot\Delta,\Delta}|^{4/3} |r_{t+(j-2)\cdot\Delta,\Delta}|^{4/3} |r_{t+(j-4)\cdot\Delta,\Delta}|^{4/3}. \tag{3.19}$$

It is shown that these staggered measures, in case of no noise term, are still consistent estimators, and so, the test statistic $Z_{1,t+1}(\Delta)$, obtained by substituting in the equation of $Z_{t+1}(\Delta)$ the corresponding staggered measures, is still asymptotically distributed as a standard Normal.

3.2.5 The HAR-RV-CJ model

It is possible to further extend the HAR-RV-J model 3.11 dividing explicitly the realized volatilities used as explanatory variables of the regression into the continuous sample path and the jump components obtained in eq. 3.17 and 3.16. We can define the multiperiod continuous path and jump components, over h days, respectively as

$$C_t^{(h)} = h^{-1}[C_{t+1} + C_{t+2} + \cdots + C_{t+h}] \tag{3.20}$$

$$J_t^{(h)} = h^{-1}[J_{t+1} + J_{t+2} + \cdots + J_{t+h}] \tag{3.21}$$

When considering the weekly and monthly components, h is respectively equal to 5 and 22. Then, the HAR-RV-CJ model is expressed as

$$\begin{aligned} RV_{t+h}^{(h)} = c + & \beta^{(d)}\hat{C}_t + \beta^{(w)}\hat{C}_t^{(5)} + \beta^{(m)}\hat{C}_t^{(22)} + \\ & + \alpha^{(d)}\hat{J}_t + \alpha^{(w)}\hat{J}_t^{(5)} + \alpha^{(m)}\hat{J}_t^{(22)} + \epsilon_t^{(h)} \end{aligned} \tag{3.22}$$

with $\{c, \beta^{(d,w,m)}, \alpha^{(d,w,m)}\}$ real numbers and an IID error term $\epsilon_t^{(h)}$.

3.3 The leverage effect

It can be empirically showed that, in equity markets, the volatility of stock returns rises more after a negative shock in the price than after an increase of the same magnitude.[3] This negative relationship between lagged negative returns and volatility is known as leverage effect.

3.3.1 The LHAR-CJ model

In Corsi and Renò (2012), a new model has been proposed, which accounts for the leverage effect in order to obtain a better volatility forecast. It is called the *Leverage Heterogeneous Auto-Regressive with Coninuous*

[3] See, for example, Christie (1982), Campbell and Hentschel (1992), Glosten et al. (1989), Bollerslev et al. (2006), Bollerslev et al. (2009), Martens et al. (2009).

volatility and Jumps model (LHAR-CJ model) and it is composed by a combination of the three characteristics: volatility, leverage and jumps. The leverage innovation of the model regards the inclusion as additional explanatory variables of lagged negative returns at three different frequencies. To do so, we define $r_t^{(h)-} = min(r_t^{(h)}, 0)$, in order to select only the negative returns. Finally, the model results to be:

$$\begin{aligned} RV_{t+h}^{(h)} = \beta_0 + & \beta^{(d)} \hat{C}_t + \beta^{(w)} \hat{C}_t^{(5)} + \beta^{(m)} \hat{C}_t^{(22)} + \\ & + \alpha^{(d)} \hat{J}_t + \alpha^{(w)} \hat{J}_t^{(5)} + \alpha^{(m)} \hat{J}_t^{(22)} \\ & + \gamma^{(d)} r_t^{-} + \gamma^{(w)} r_t^{(5)-} + \gamma^{(m)} r_t^{(22)-} + \epsilon_{t+h}^{(h)}. \end{aligned} \tag{3.23}$$

The parameters $\{c, \beta^{(d,w,m)}, \alpha^{(d,w,m)}, \gamma^{(d,w,m)}\}$ are real numbers and the error term $\epsilon_t^{(h)}$ is an IID noise. We can notice that if $\alpha^{(d,w,m)}$ and $\gamma^{(d,w,m)}$ are both equal to zero, the model reduces to the HAR of Corsi(2009). Instead, if only $\gamma^{(d,w,m)}$ is null, the model consists in the continuous and the discontinuous components, resulting in the HAR-CJ model of Andersen et al. (2007). Lastly, when $\alpha^{(d,w,m)} = 0$, $\hat{C}_t = \hat{V}_t$ and the model is called LHAR model.

Chapter 4

Empirical analysis

In this chapter, I am going to consider three stocks and compute the realized volatilities of their return series. Then, I am going to model them through the HAR models, in order to perform forecasts. Finally, the forecasted values are compared to the ones obtained through ARMA and ARFIMA models. All the analysis have been computed through R. The most used R packages have been *highfrequency* and *HARModel.*

4.1 Characteristics of the price and return series

For this analysis, I downloaded from Bloomberg the 5-minutes stock prices of Enel, Intesa San Paolo and Generali. As suggested in the literature, a time interval of 5 minutes is often chosen in order to reduce the microstructure noise. Indeed, lower frequencies risk to be too much affected by the noises, compromising all the results. Being the italian stock market open from 9:00 to 17:30, the observations considered are 103 per day. The time series go from 2017-09-18 9:00 to 2018-12-05 17:30, with a total number of observations of 32033.

In figure 4.1 the three time series of the prices are plotted. We can notice that in May 2017 all of the three time series show an important fall. The

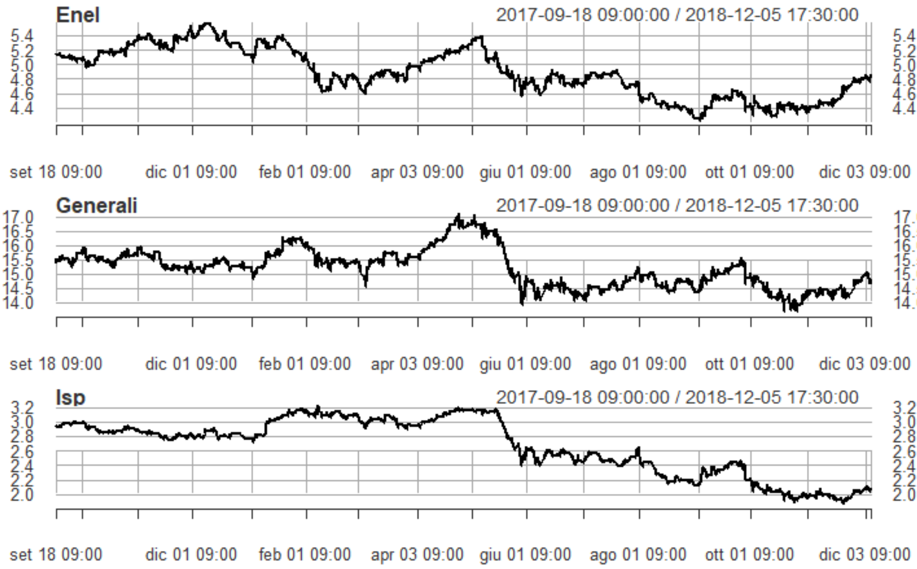


Figure 4.1: Plots of the 5-minutes prices of, respectively, Enel, Generali and Intesa San Paolo.

Prices	**Enel**	**Generali**	**ISP**
Minimum	4.220	13.650	1.871
Median	4.848	15.230	2.796
Mean	4.885	15.190	2.664
Maximum	5.585	17.120	3.227

Table 4.1: Minimum, median, mean and maximum values of the 5-minutes prices of Enel, Generali and Intesa San Paolo.

minimum and maximum values, the mean and the median of the three series are summarized in table 4.1.

The exploratory analysis continues with the computation of the logarithmic returns for each of the series. Being $p(t)$ the price at time t, the logarithmic return $r(t)$ is defined as

$$r(t) = ln\left[\frac{p(t)}{p(t-h)}\right] = ln[p(t)] - ln[p(t-h)]$$

where $h = 5$ minutes. The plots and the characteristics of the return series are shown in Figure 4.2 and Table 4.2. From the plots we can see that the returns are centered around zero, with some volatility clustering, especially

Returns	**Enel**	**Generali**	**ISP**
Minimum	-0.0323	-0.0480	-0.0635
Median	0	0	0
Mean	$-1.940 \cdot 10^{-6}$	$-1.354 \cdot 10^{-6}$	$-1.0710 \cdot 10^{-5}$
Maximum	0.0267	0.0258	0.0450
Variance	$1.911 \cdot 10^{-6}$	$1.641 \cdot 10^{-6}$	$3.207 \cdot 10^{-6}$
Std. Deviation	0.0014	0.0012	0.0018
Skewness	-1.0172	-1.3987	-0.6361
Kurtosis	51.5780	90.4545	123.2594

Table 4.2: Summary statistics of the 5-minutes returns of Enel, Generali and Intesa San Paolo.

in the second half of May and September 2017, where also the prices series saw a fall. The sample mean values are very close to zero for all the three series.

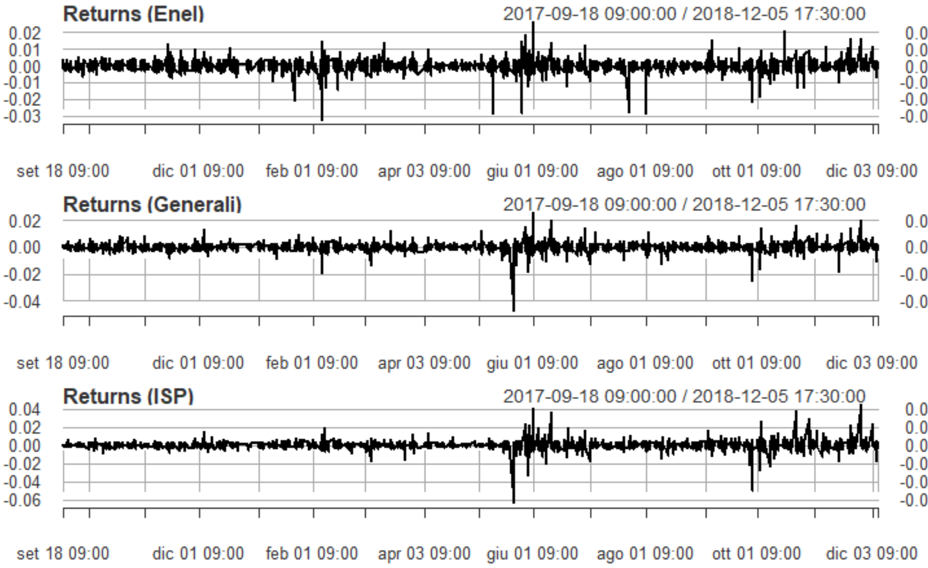


Figure 4.2: Plots of the 5-minutes returns of, respectively, Enel, Generali and Intesa San Paolo.

It is possible to deduce from the skewness and kurtosis values that the returns are not normally distributed, since they are not close respectively to zero and three, the value that a normal distribution would assume. In particular, the

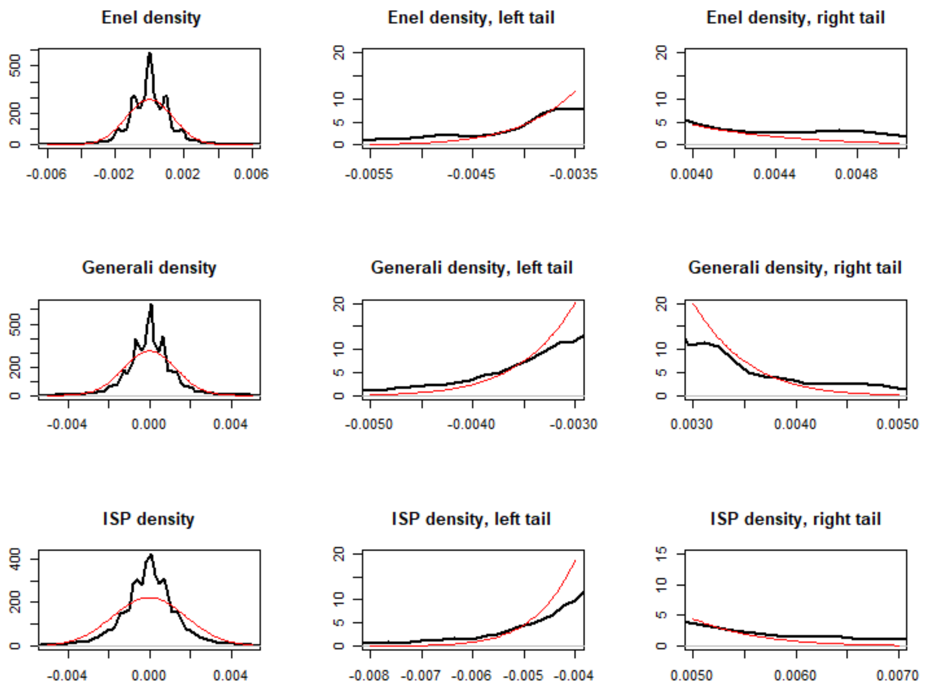


Figure 4.3: Plots of the density functions of the return series of Enel, Generali and Intesa San Paolo, compared to the normal ones, with focuses in the left and right tails.

very high values of the kurtosis for all the three series is a typical feature of the high frequency data, denoting a leptokurtic distribution. Figure 4.3 shows for each series the comparison between the density functions of the returns and the one of the corresponding normal density functions, with also a zoom on the left and right tails. As the values of the kurtosis suggested, the three distributions have fatter tails than the one of the normal distribution, as the ones of high-frequency data usually are.

To improve the exploratory analysis, we can better check for the normality assumption with a Normal probability plot, which compares the sample quantiles with the corresponding quantiles of the standard normal distribution. In Figure 4.4 the Q-Q plots of the three series are represented. The right and the left parts of the patterns show deviation from the straight lines, meaning that the two tails are heavier than the normal ones. Hence, from

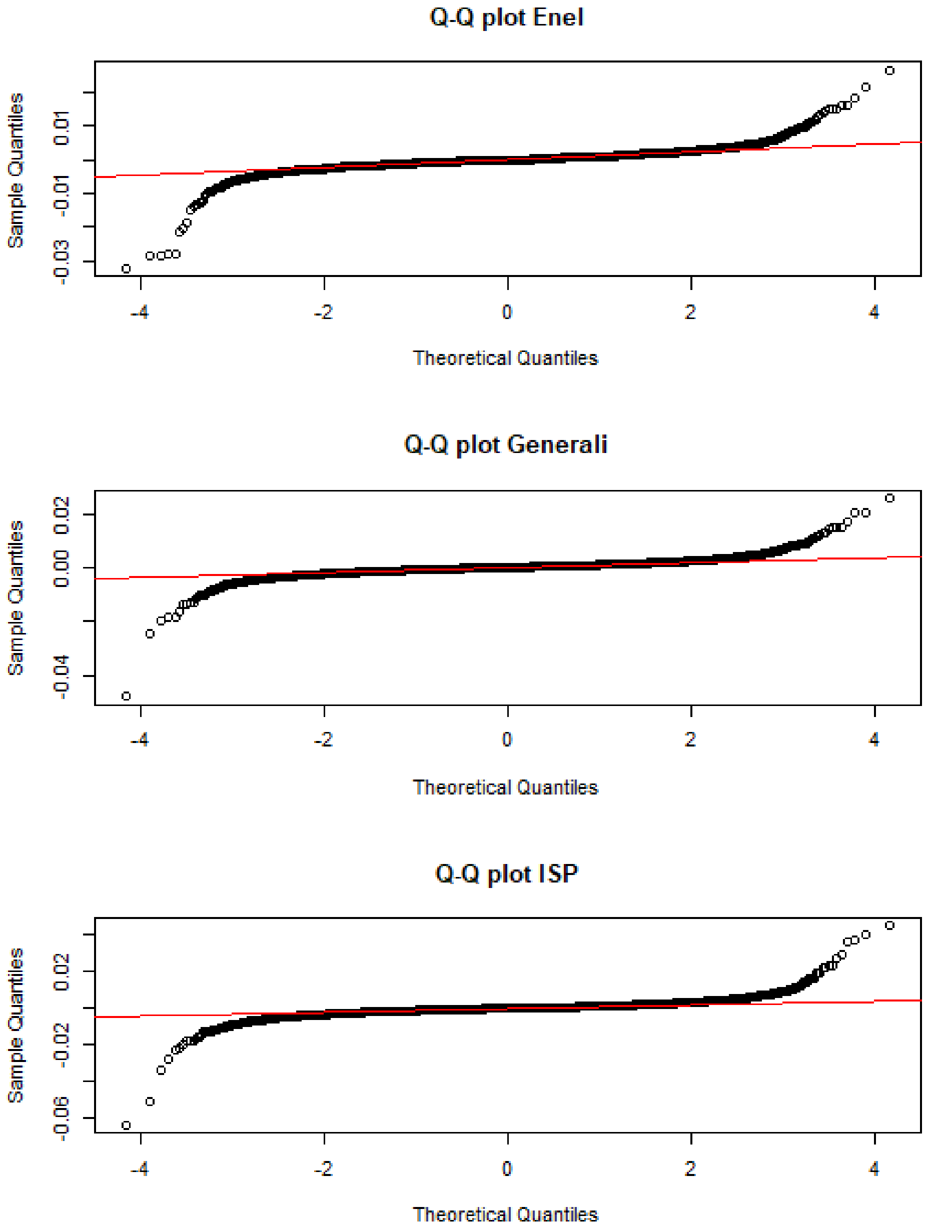


Figure 4.4: Q-Q plots of Enel, Generali and Intesa San Paolo.

these graphs, the assumption of normality does not seem to be confirmed.

As stated in Chapter 1, high frequency returns usually display a negative first-order dependence, which is mainly due to the bid-ask spread and the

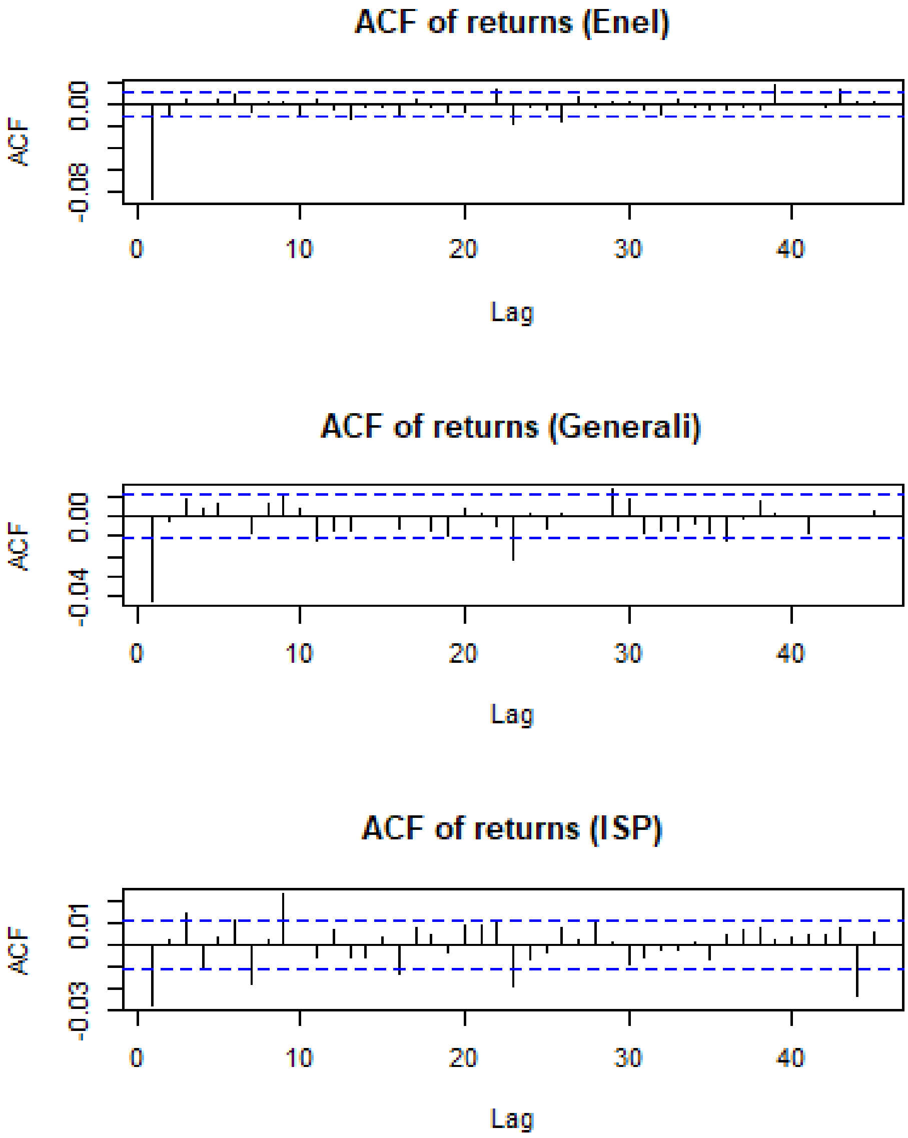


Figure 4.5: Autocorrelation functions of the return series of Enel, Generali and Intesa San Paolo.

price discreteness. Figure 4.5 confirms this feature, showing for all the three series a significant negative autocorrelation at the first lag. However, it seems to be apparently modest. On the other hand, the autocorrelation functions of the squared returns in Figure 4.6 display a slowly decaying behavior, with significant values for many lags.

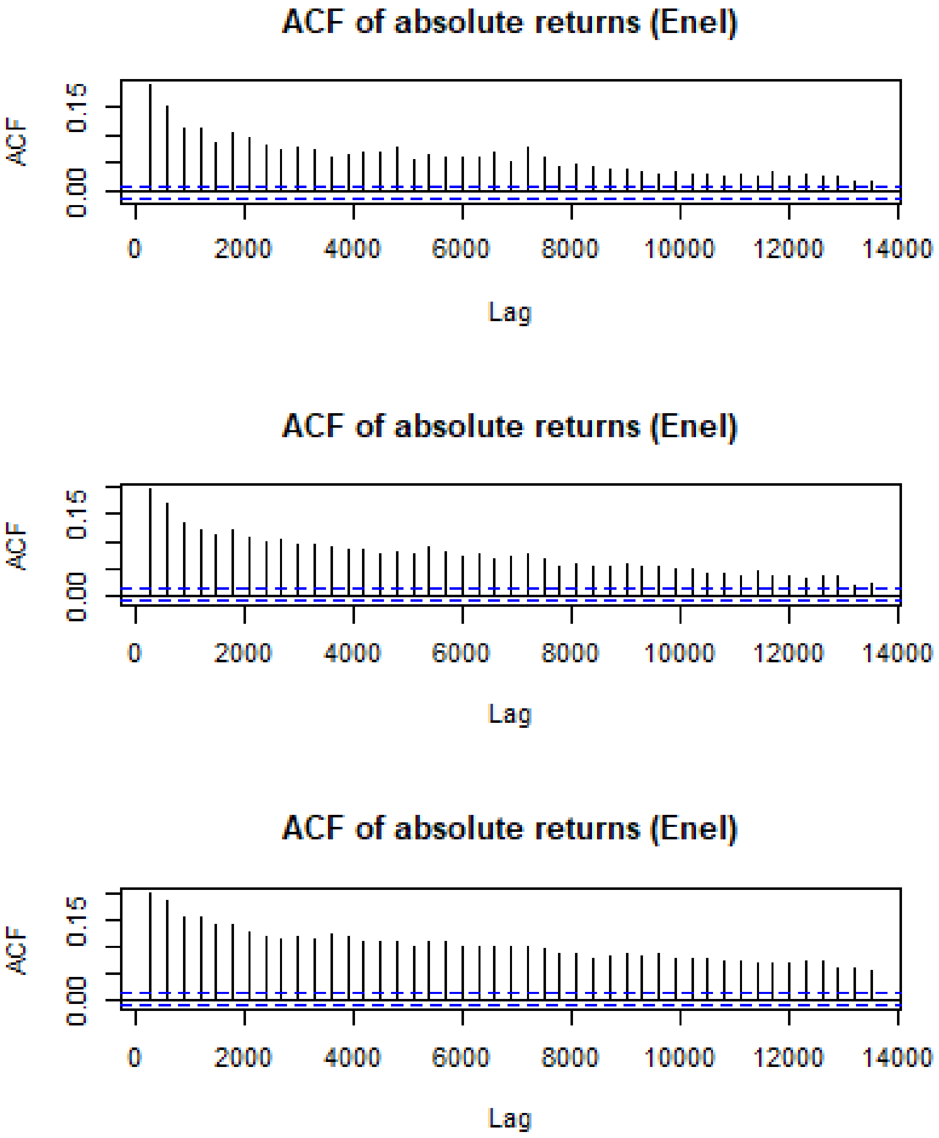


Figure 4.6: Autocorrelation functions of the absolute return series of Enel, Generali and Intesa San Paolo.

RV	**Mean**	**St. Dev.**	**Skewness**	**Kurtosis**
Enel	$1.968 \cdot 10^{-4}$	$1.885 \cdot 10^{-4}$	4.0615	21.5831
Generali	$1.69 \cdot 10^{-4}$	$1.941 \cdot 10^{-4}$	6.4331	60.9359
ISP	$3.30 \cdot 10^{-4}$	$4.605 \cdot 10^{-4}$	4.6050	27.4170

Table 4.3: Summary statistics of the 5-minutes realized volatilities of Enel, Generali and Intesa San Paolo.

4.2 Characteristics of the realized volatility series

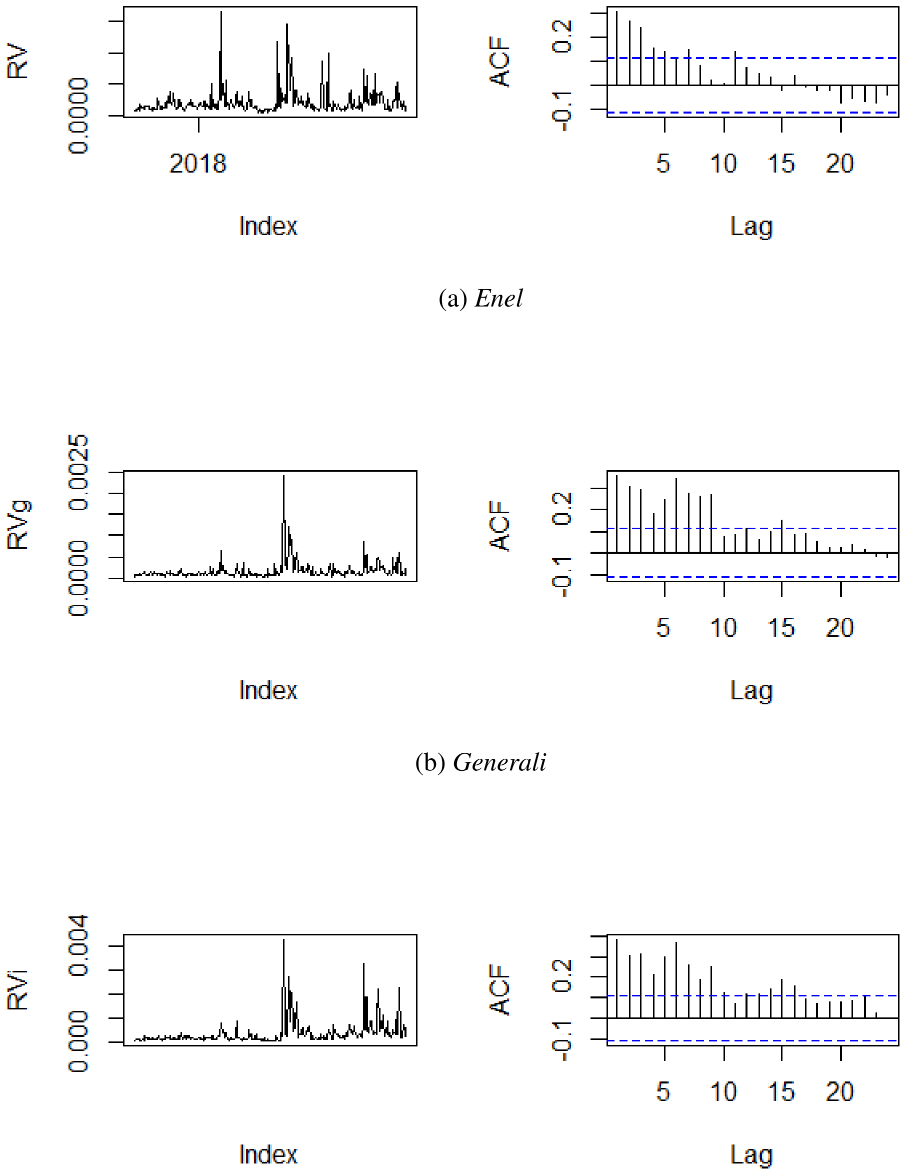
After analyzing the return series, I obtained the daily realized volatility (RV) series for each stock, through the computation explained in Chapter 2. Table 4.3 reports, for each realized volatility series, the sample mean, the standard deviation, the skewness and the kurtosis, together with the maximum and minimum values. This standard summary of statistics helps in outlining the unconditional distributions of the three realized volatility series. Overall, the most volatile is the one of Intesa San Paolo, followed by the one of Generali. Looking at the skewness, it is clear that all the series are right-skewed. Obviously, the normal distribution is a poor approximation.

In Figure 4.7 all the series are plotted, together with their autocorrelation functions.

The temporal extent of the autocorrelation is clearly exhibited by the quantity of significant values in the ACF of each series. This is supported also by the results of the Ljung-Box test, reported in table 4.4.

This test has been computed for the first twenty lags ($K = 20$). Being $\hat{\rho}_k$, $k = 1, \dots, K$, the sample autocorrelation at lag k, the hypothesis of the test are the following:

- H_0: $\rho_k = 0, k = 1, \dots, K$;
- H_1: $\exists \rho_k \neq 0, k = 1, \dots, K$.



(a) *Enel*

(b) *Generali*

(c) *Intesa San Paolo*

Figure 4.7: Plots of the realized volatility series and their corresponding autocorrelation functions of Enel, Generali and Intesa San Paolo.

RV	LB	p-value
Enel	138.79	2.76e-10
Generali	292.66	< 2.2e-16
ISP	308.3	< 2.2e-16

Table 4.4: Test statistics and p-values of the Ljung-Box test computed over 50 lags, for the realized volatility series of Enel, Generali and Intesa San Paolo.

The test statistic is computed as

$$Q = n(n+2)\Big(\frac{\hat{\rho}_1^2}{n-1} + \frac{\hat{\rho}_2^2}{n-2} + \cdots + \frac{\hat{\rho}_K^2}{n-K}\Big)$$

where n is the sample size and . Under H_0, the test statistic follows a $\chi^2_{(K)}$. The p-values are all very small and the Q-statistics are higher than the critical value, which, for a significance level of $\alpha = 0.05$, is equal to 67.50. Hence, we reject the null hypothesis and the series result to be autocorrelated.

4.3 Estimation of the HAR models

To model the realized volatility through the HAR-RV model, the daily series have been aggregated to create the weekly and monthly series, as explained in Chapter 3. In this way, the measures are comparable even if computed over different horizons. Figures 4.8, 4.9 and 4.10 show, for each stock, the scatterplots that compare the daily realized volatility to the lagged weekly and monthly ones. The majority of the points plotted lies on the lower corner on the left-hand side of the graphs, meaning that to small values of a variable correspond small values of the other variable. However, when the comparison is done against the lagged monthly realized volatility, the dispersion of the points is more evident and the link between the two variables seems weaker.

Table 4.5 includes all the parameters estimated through a standard OLS regression of the model:

$$RV_{t+1}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \epsilon_{t+1}.$$

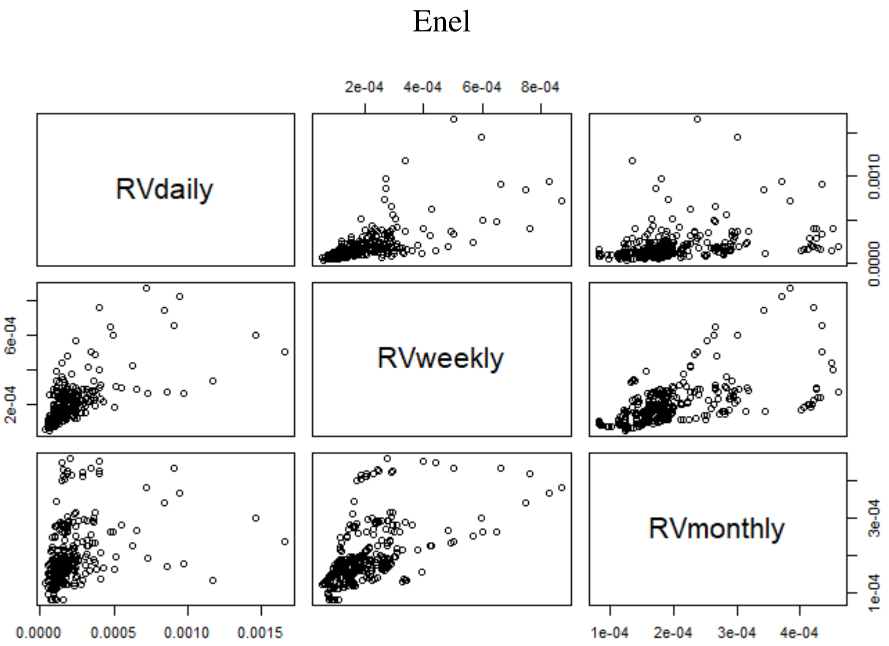


Figure 4.8: Scatterplots of the daily realized volatility and the lagged weekly and monthly realized volatility for Enel.

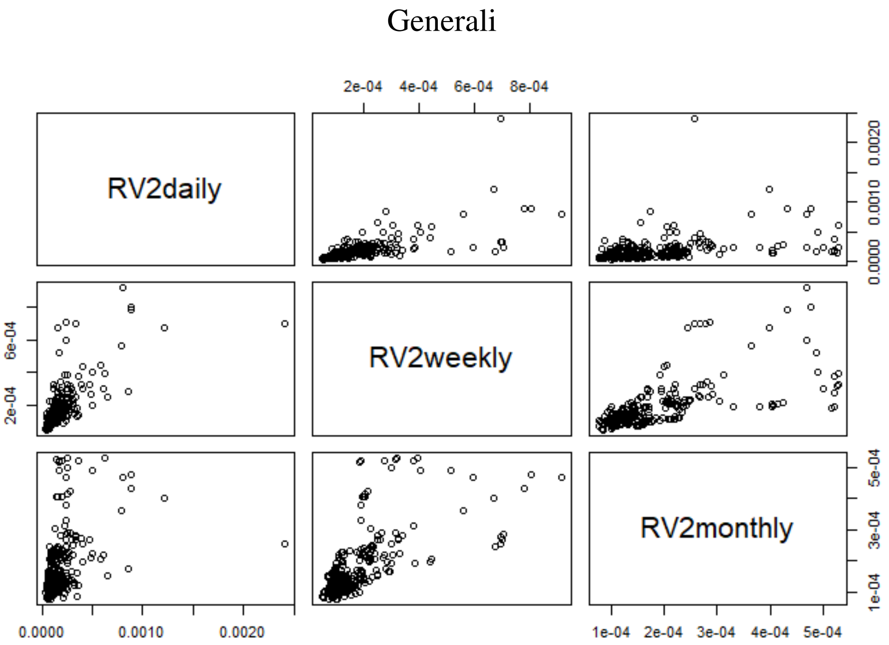


Figure 4.9: Scatterplots of the daily realized volatility and the lagged weekly and monthly realized volatility for Generali.

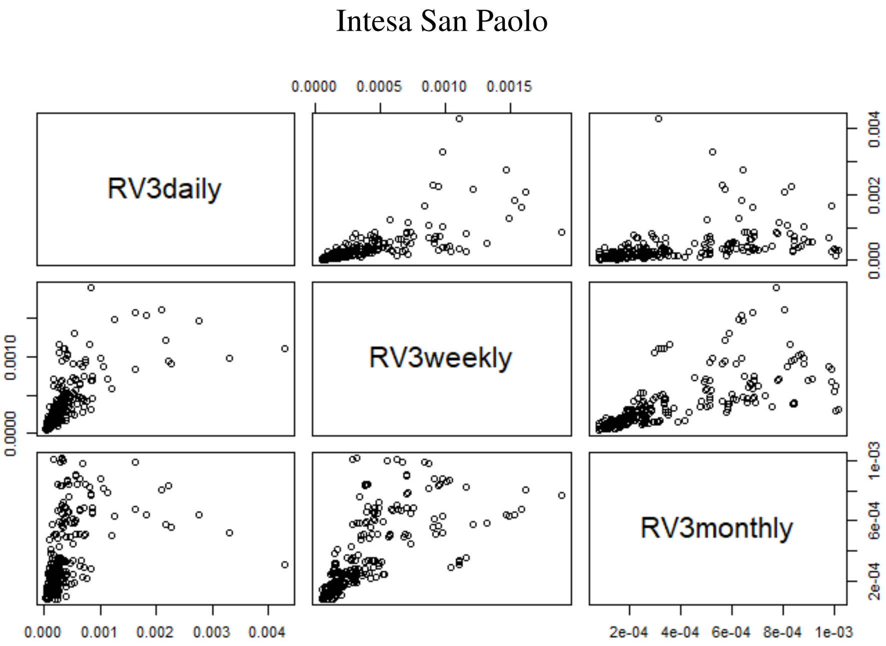


Figure 4.10: Scatterplots of the daily realized volatility and the lagged weekly and monthly realized volatility for Intesa San Paolo.

HAR-RV	Enel		Generali		ISP	
c	$1.03 \cdot 10^{-4}$	(0.0006)	$5.60 \cdot 10^{-5}$	(0.0093)	$9.66 \cdot 10^{-5}$	(0.0237)
$\beta^{(d)}$	0.1503	(0.0322)	0.1650	(0.0179)	0.1696	(0.0146)
$\beta^{(w)}$	0.3997	(0.0017)	0.3815	(0.0033)	0.3807	(0.0034)
$\beta^{(m)}$	−0.0565	(0.7299)	0.1382	(0.3397)	0.1852	(0.1992)

Table 4.5: Parameter estimation through HAR-RV model. Numbers in brackets are the corresponding p-values of the significance test. The colored cells indicate the p-values of non-significant estimates.

HAR-RV	Enel		Generali		ISP	
c	$9.57 \cdot 10^{-5}$	$(5.24 \cdot 10^{-6})$	$6.75 \cdot 10^{-5}$	$(1.62 \cdot 10^{-4})$	$1.24 \cdot 10^{-4}$	$(7.97 \cdot 10^{-4})$
$\beta^{(d)}$	0.1523	(0.0292)	0.1616	(0.0201)	0.1659	(0.0169)
$\beta^{(w)}$	0.3779	$(6.07 \cdot 10^{-4})$	0.4543	$(1.68 \cdot 10^{-5})$	0.4821	$(3.72 \cdot 10^{-6})$

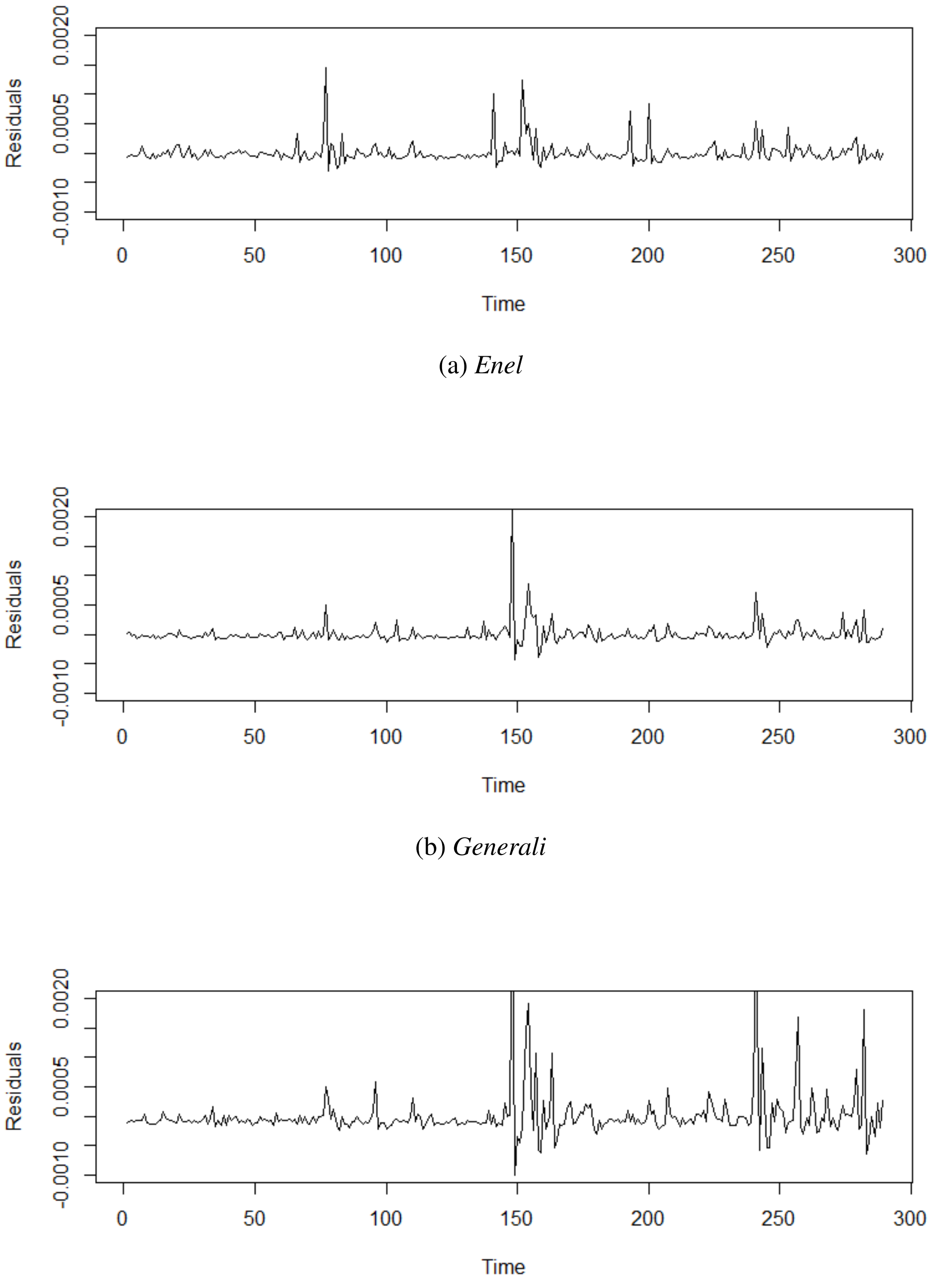
Table 4.6: Parameter estimation through HAR-RV model, without the monthly realized volatility. Numbers in brackets are the corresponding p-values of the significance test.

The numbers reported in brackets are the corresponding values of the t-statistics. We can notice that the p-values of the monthly realized volatility estimates, in the grey cells, are all higher than 0.05. Hence, those estimates are not significant and I can remove them from the regression. The results from the new model:

$$RV_{t+1}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \epsilon_{t+1}$$

are reported in Table 4.6. In this case, all the parameters are significant. Economically speaking, this means that the the daily future realized volatility does not change significantly if the previous monthly realized volatility changes, but it depends on the behavior of the previous daily and weekly ones.

If the models are correctly specified and the parameters estimates are sufficiently close to the true values, than we should obtain residuals with behaviors similar to the ones of a white noise. From the plots of the residuals in Figure 4.11, we can notice that for the Generali's and Intesa San Paolo's series, the models seems not to be able to adequately capture the increases of the volatility. On the other hand, if we look at the autocorrelation functions, in Figure 4.12 together with the p-values from the Ljung Box, we can notice that the model seems to fit quite well only to the Enel realized volatility series. Indeed, for Generali and Intesa San Paolo series, the autocorrelation functions show some significant values and the p-values are lower than the significance value of 0.05 for many lags..

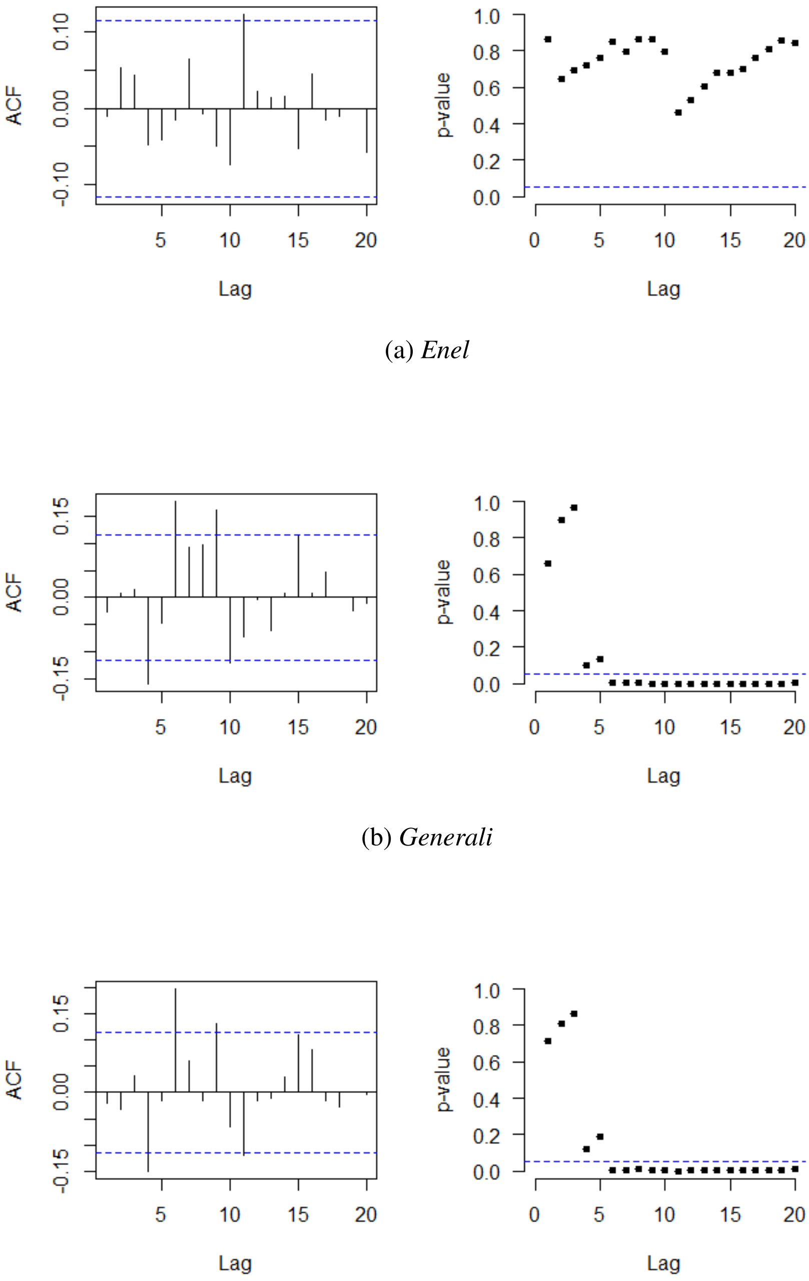


(a) *Enel*

(b) *Generali*

(c) *Intesa San Paolo*

Figure 4.11: Plots of the residuals of the HAR models applied to the realized volatility series of Enel, Generali and Intesa San Paolo.



(a) *Enel*

(b) *Generali*

(c) *Intesa San Paolo*

Figure 4.12: Plots of the autocorrelation functions and the p-values from the Ljung-Box test on the residuals of the HAR-RV models.

HAR-CJ	**Enel**		**Generali**		**ISP**	
c	$9.86 \cdot 10^{-5}$	(0.0026)	$4.30 \cdot 10^{-5}$	(0.168)	$5.35 \cdot 10^{-5}$	(0.3179)
$\beta^{(d)}$	0.3237	(0.015)	0.8269	$(9.31 \cdot 10^{-8})$	1.0500	$(1.91 \cdot 10^{-10})$
$\beta^{(w)}$	0.4878	(0.049)	0.1345	(0.558)	0.0551	(0.8322)
$\beta^{(m)}$	-0.1719	(0.623)	-0.0817	(0.815)	-0.0145	(0.9684)
$\alpha^{(d)}$	0.0164	(0.893)	-0.0582	(0.466)	-0.1638	(0.0531)
$\alpha^{(w)}$	0.0886	(0.800)	0.1835	(0.393)	0.2395	(0.2983)
$\alpha^{(m)}$	0.0058	(0.994)	0.2410	(0.706)	0.1911	(0.7273)

Table 4.7: Parameter estimation through HAR-RV-CJ model. Numbers in brackets are the corresponding p-values of the t-statitics.

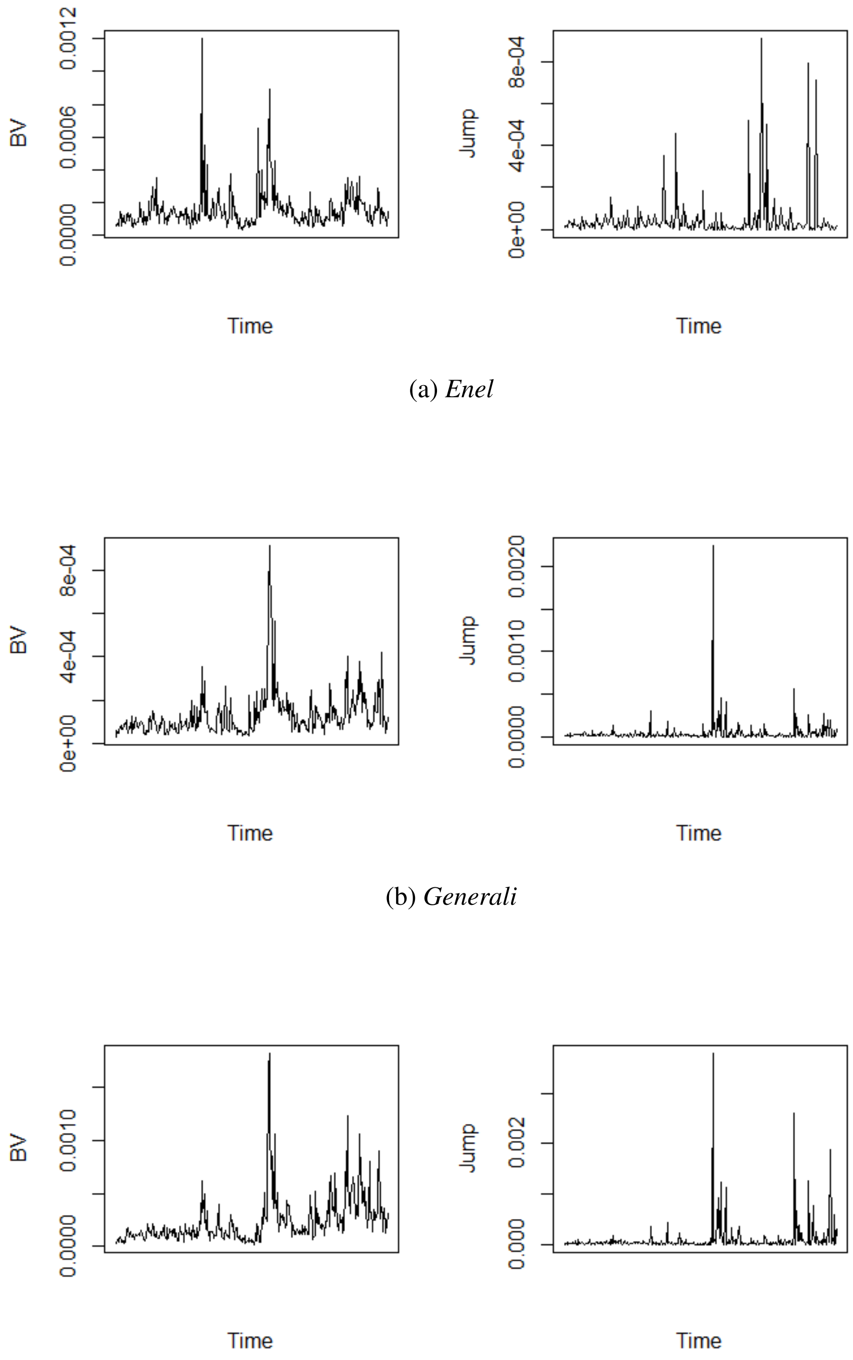
In order to improve the models, after computing the bipower realized variation, I added to the basic model also the jump component, to check if the behavior of the future realized volatility of the data could depend on the series of the jumps. As already explained, the realized volatility can be divided into a continuous component, represented by the bipower variation, and a jump component. Figure 4.13 show for each stock these two components. The jump series have been tested as in Chapter 3, in order to select only the significant values, i.e.

$$\begin{cases} J_t^{(d,w,m)} = 0 & \text{if not significant} \\ J_t^{(d,w,m)} > 0 & \text{if significant.} \end{cases}$$

In Table 4.7 we can find the parameters estimated through the HAR-RV-CJ models:

$$RV_{t+1}^{(d)} = c+\beta^{(d)}C_t^{(d)}+\beta^{(w)}C_t^{(w)}+\beta^{(m)}C_t^{(m)}+\alpha^{(d)}J_t^{(d)}+\alpha^{(w)}J_t^{(w)}+\alpha^{(m)}J_t^{(m)}+\epsilon_{t+1}.$$

It resulted that all the jump components are not significant and even removing them one-by-one from the regression, the situation did not get better. Moreover, even running an F test to check if the restrictions from an HAR-RV-CJ to an HAR-RV model were valid, I obtained results lower than the critical values, leading me to accept the null hypothesis, which states that the additional coefficients are not jointly significant. Then, I concluded that



(a) *Enel*

(b) *Generali*

(c) *Intesa San Paolo*

Figure 4.13: Bipower realized variation and jump component of the return series of Enel, Generali and Intesa San Paolo.

the future realized volatilities do not depend on the jump series.

Finally, I tried to insert into the HAR-RV regression the negative past returns computed through a period of 1, 5 and 22 days, in order to check if the leverage effect conditions the future realized volatility. However, running the model:

$$\begin{aligned}RV_{t+1}^{(d)} =& c + \beta^{(d)} RV_t + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \\ &+ \gamma^{(d)} r_t^{-} + \gamma^{(w)} r_t^{(5)-} + \gamma^{(m)} r_t^{(22)-} + \epsilon_{t+h}^{(h)}\end{aligned}$$

and reducing it removing one-by-one the regressors of which the parameters' estimates were not significant, I obtained exactly the standard HAR-RV model, without the monthly component, of Table 4.6. This means that neither the jump or the leverage components have any effect on the daily future realized volatility.

As I previously showed, it is not possible to state that residuals from the applied HAR-RV models behave very well. A possible modification to obtain a better fit could consist in transforming the dependent and independent variables of the HAR regressions through a logarithmic transformation, as suggested in Andersen et al. (2007). To do so, I recomputed all the OLS estimations, collecting the significant parameter estimates, with their corresponding p-values from the significance test, in Table 4.8. The logarithmic jump components resulted to still be not significant for all the three stocks, hence, the estimates come from the following regression model:

$$\begin{aligned}logRV_{t+1}^{(d)} =& c + \beta^{(d)} logRV_t^{(d)} + \beta^{(w)} logRV_t^{(w)} + \beta^{(m)} logRV_t^{(m)} + \\ &+ \gamma^{(d)} log(r_t^{-}) + \gamma^{(w)} log(r_t^{(5)-}) + \gamma^{(m)} log(r_t^{(22)-}) + \epsilon_{t+1}\end{aligned}$$

In Figure 4.14, the residuals of these models are plotted. It is clear the difference from Figure 4.11: in this case the residuals seems to have a much more constant variability. Moreover, looking at Figure 4.15, it is easy to notice that their autocorrelation functions show less significant values and the p-values obtained from the Ljung-Box test are almost all higher than the

Log L-HAR	**Enel**		**Generali**		**ISP**	
c	-3.557	$(7.45 \cdot 10^{-10})$	-3.291	$(9.56 \cdot 10^{-10})$	-3.873	$(3.14 \cdot 10^{-8})$
$\beta^{(d)}$	0.342	$(2.45 \cdot 10^{-6})$	0.225	(0.0031)	0.309	$(1.84 \cdot 10^{-5})$
$\beta^{(w)}$	0.254	(0.0045)	0.417	$(8.70 \cdot 10^{-7})$	0.276	(0.0030)
$\beta^{(m)}$	-	-	-	-	-	-
$\gamma^{(d)}$	-	-	-17.061	(0.0005)	-10.486	(0.0014)
$\gamma^{(w)}$	-	-	-	-	-	-
$\gamma^{(m)}$	-	-	-	-	-43.598	(0.0032)

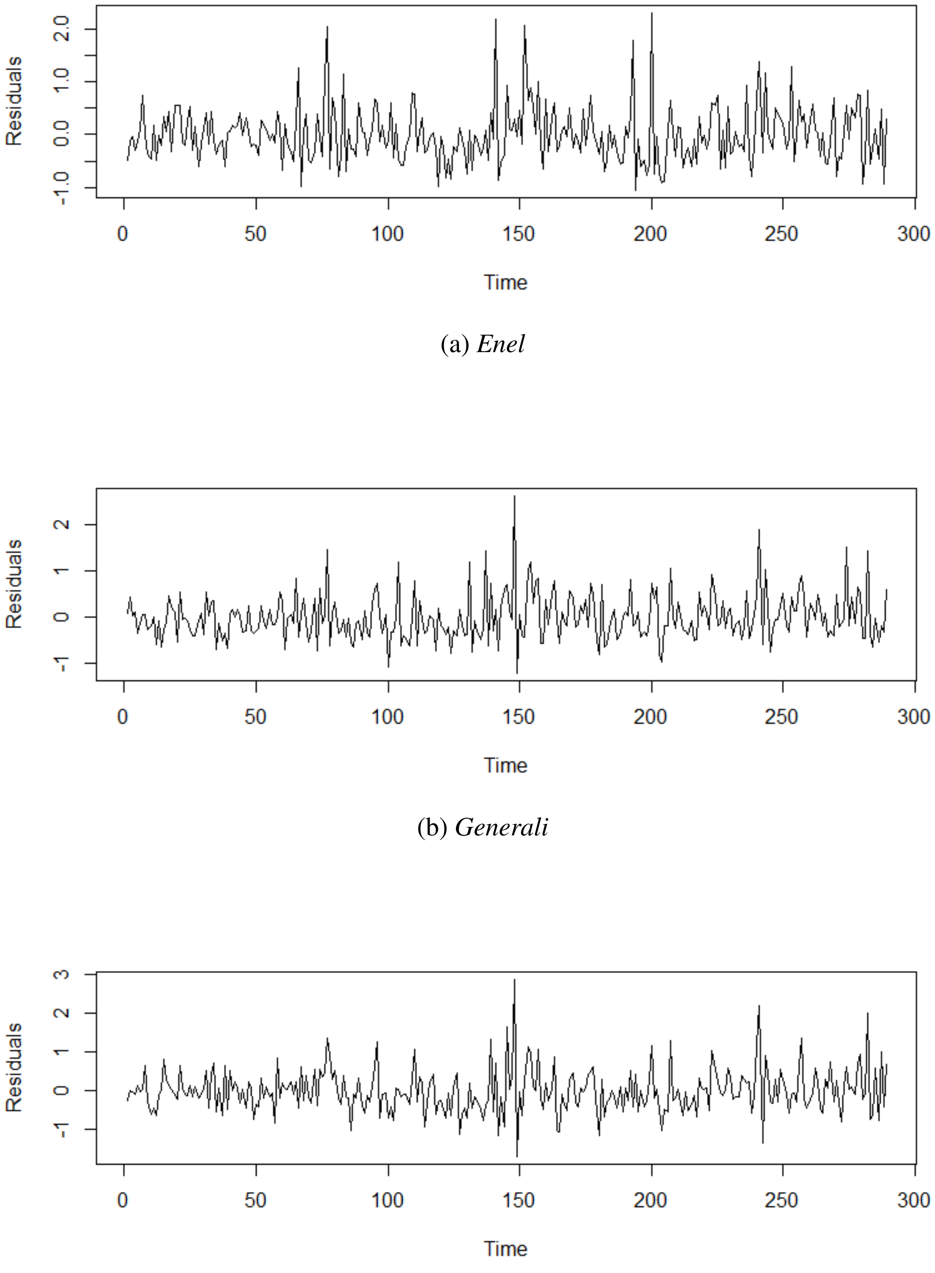
Table 4.8: Parameter estimation through LHAR-RV model with the logarithmic transformation of the variables. Numbers in brackets are the corresponding p-values from the significance test.

significance level.

A further proof that the logarithmic transformation lead to residuals that behave more similarly to a white noise is demonstrated in Figure 4.16 , which displays the density functions and the Q-Q plots, compared to the ones of the normal. We can notice that the residuals for each stock have a distribution quite similar to the one of the normal, apart from the right tails that are still a little fatter.

4.4 Forecasting performances

From the residual analysis previously computed, the HAR models applied seemed to not be enough adequate in capturing the behavior of the realized volatility, with the exception of the Enel series, and I tried to improve the results through the logarithmic transformation of the variables in the regression.
In this section, I am going to inspect the forecasting performances of the final models previously estimated. To do so, I am going to compute the one-step-ahead forecasts, both in-sample and out-of-sample, of the realized volatility for all the models and compare them with ARMA and ARFIMA models.

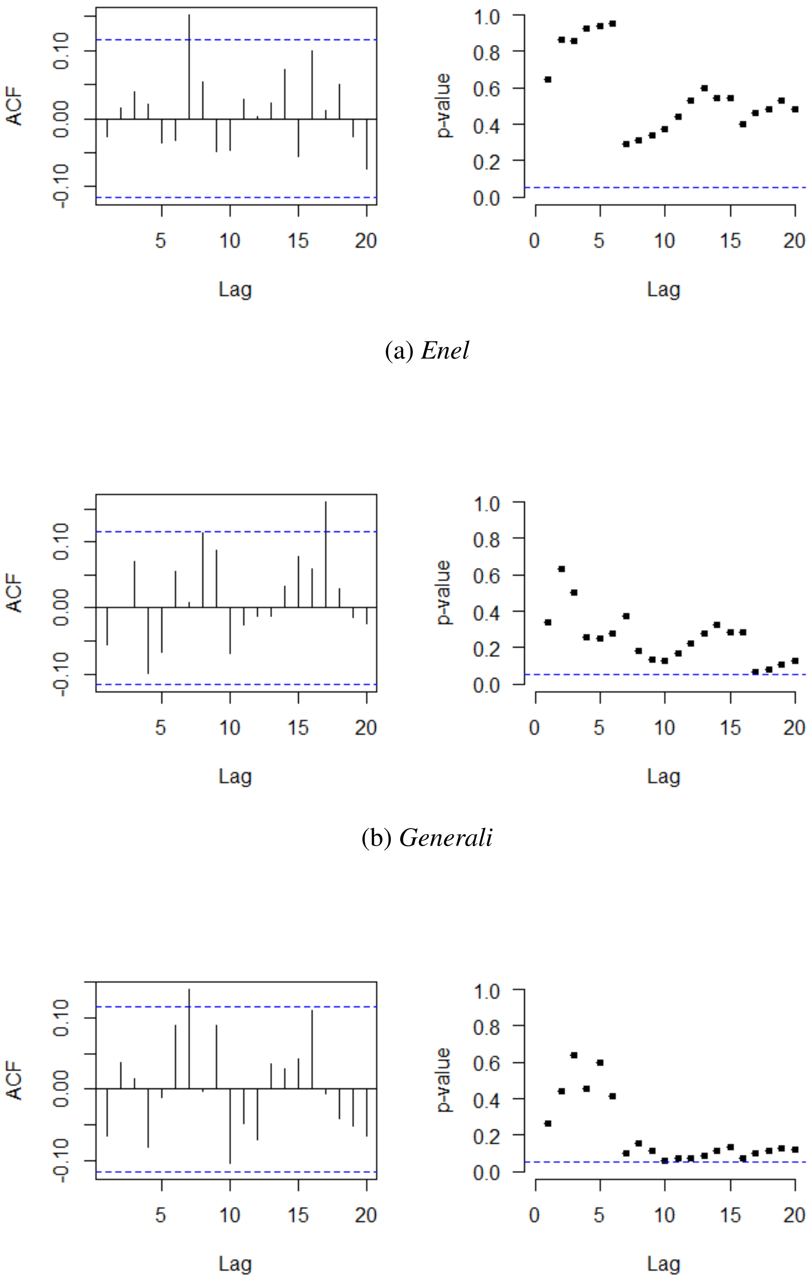


(a) *Enel*

(b) *Generali*

(c) *Intesa San Paolo*

Figure 4.14: Plots of the residuals of the HAR models applied to the logarithmic realized volatility series of Enel, Generali and Intesa San Paolo.

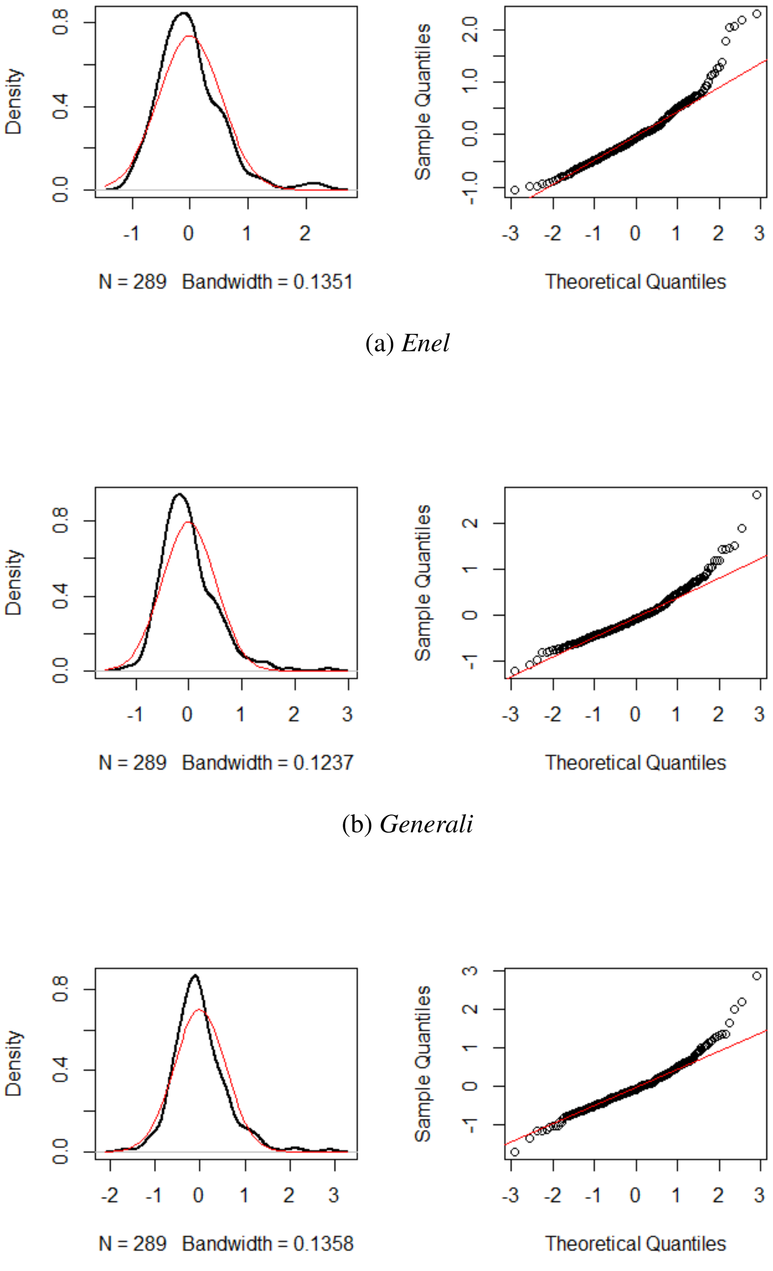


(a) *Enel*

(b) *Generali*

(c) *Intesa San Paolo*

Figure 4.15: Plots of the autocorrelation functions of the residuals from the HAR models applied to the logarithmic realized volatility series of Enel, Generali and Intesa San Paolo, together with their corresponding p-values from the Ljung-Box test.



(a) *Enel*

(b) *Generali*

(c) *Intesa San Paolo*

Figure 4.16: Density functions and Q-Q plots of the residuals from the logarithmic HAR models. The red lines are the corresponding normal ones.

4.4.1 Model fitting

The fitted values of the final models considered give out the one-step-ahead in-sample forecasts. Hence, for each type of realized volatility series of each stock, the model is estimated on the entire sample of data and, every day, the parameters' estimates obtained are used to get the realized volatility value of the day after. Figures 4.17 and 4.18 show a visual comparison between actual and in-sample forecasted values of, respectively, the HAR models and the logarithmic HAR models applied to the Enel, Generali and Intesa San Paolo stocks. The blue straight lines are the forecasted values, while the red dotted lines are the actual values. Obviously, the deviation between the blue and red lines are are the residuals inspected in the previous section.

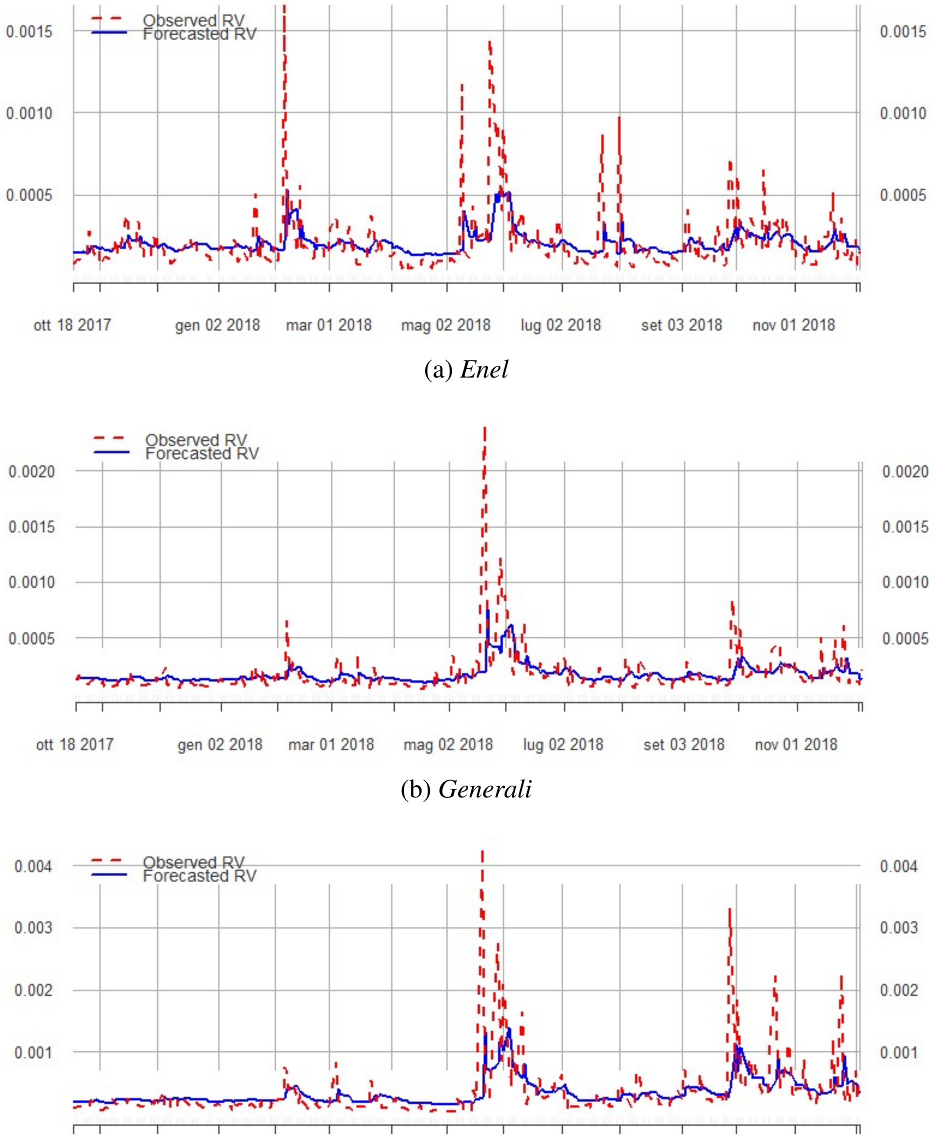
Then, I computed for each applied model, the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE), which are both measures computed on the difference between the predicted and the actual values. This means that the smaller are their amounts and the better are considered the predictions. Let Z be the forecasting period, such that $z = 1, 2, \ldots, Z$. Denote as y_z and $\hat{y}_z$, respectively, the actual and the forecasted values at time z. Then, the Root Mean Square Error is expressed as:

$$RMSD = \sqrt{\frac{\sum_{z=1}^{Z}(\hat{y}_z - y_z)^2}{Z}}$$

and the Mean Absolute Error is:

$$MAE = \frac{\sum_{z=1}^{Z}|\hat{y}_z - y_z|}{Z}.$$

In Table 4.9 are reported the results of the RMSE and MAE for all the models. We can notice that the logarithmic transformation lead to better values in terms of MAE, but not in terms of RMSE.

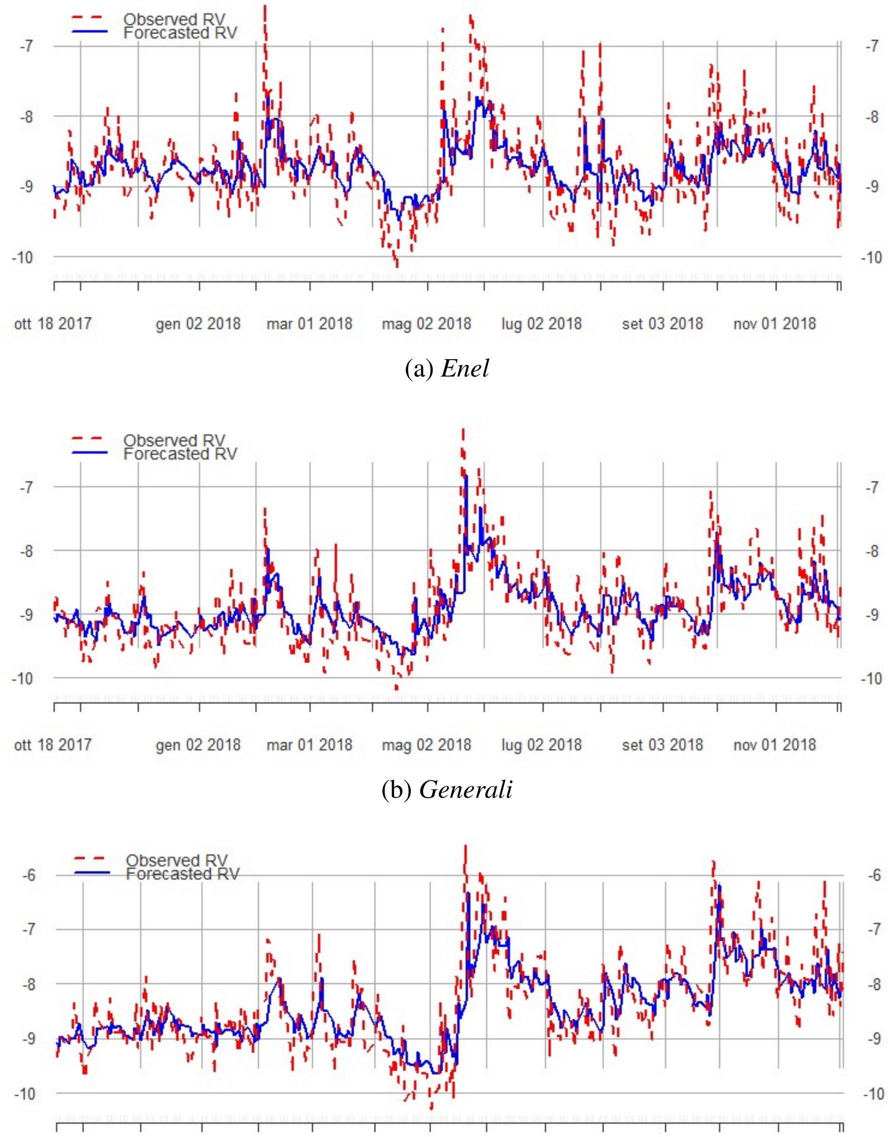


(a) *Enel*

(b) *Generali*

(c) *ISP*

Figure 4.17: Comparison between the one-day-ahead in-sample prediction and actual values (dotted line) of the HAR models applied to the Enel, Generali and Intesa San Paolo series.



(a) *Enel*

(b) *Generali*

(c) *ISP*

Figure 4.18: Comparison between the one-day-ahead in-sample prediction and actual values (dotted line) of the logarithmic HAR models applied to the Enel, Generali and Intesa San Paolo series.

Fitted Values Performances	**Levels**		**Logarithmic**	
	RMSE	**MAE**	**RMSE**	**MAE**
Enel	$1.810 \cdot 10^{-4}$	$0.991 \cdot 10^{-4}$	$1.840 \cdot 10^{-4}$	$0.894 \cdot 10^{-4}$
Generali	$1.808 \cdot 10^{-4}$	$0.865 \cdot 10^{-4}$	$1.847 \cdot 10^{-4}$	$0.752 \cdot 10^{-4}$
ISP	$4.216 \cdot 10^{-4}$	$1.974 \cdot 10^{-4}$	$4.251 \cdot 10^{-4}$	$1.668 \cdot 10^{-4}$

Table 4.9: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) of the fitted values of the models, in both levels and logarithmic form, applied to the realized volatilities of Enel, Generali and Intesa San Paolo.

Another measure representing the goodness-of-fit of an applied model is the R-squared (or Coefficient of Determination), which is computed as

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

where y_i are the observed values of the dependent variable, $\hat{y}$ is its mean and $\hat{y}_i$ are the values fitted by the model. The numerator corresponds to the Explained Sum of Squares (ESS), while the denominator is the Total Sum of Squares (TSS). It indicates the percentage of the variance in the dependent variable that the independent variables are able to explain. In Table 4.10, the results obtained for the R-squared of the HAR models are reported. We can notice that these values are all quite small, revealing a low goodness-of-fit of the models. The values slightly increase when considering the logarithmic transformation of the models.

4.4.2 Out-of-sample forecasts

The original samples of the realized volatility series go from 2017-09-18 to 2018-12-05, with 311 observations each. In order to compute the rolling out-of-sample forecasts, I considered subsamples ending in 2018-11-07, which are then composed by 291 values, excluding the last 20 days of

R^2	**Levels**	**Logarithmic**
Enel	0.1258	0.2551
Generali	0.1799	0.3930
ISP	0.2033	0.5578

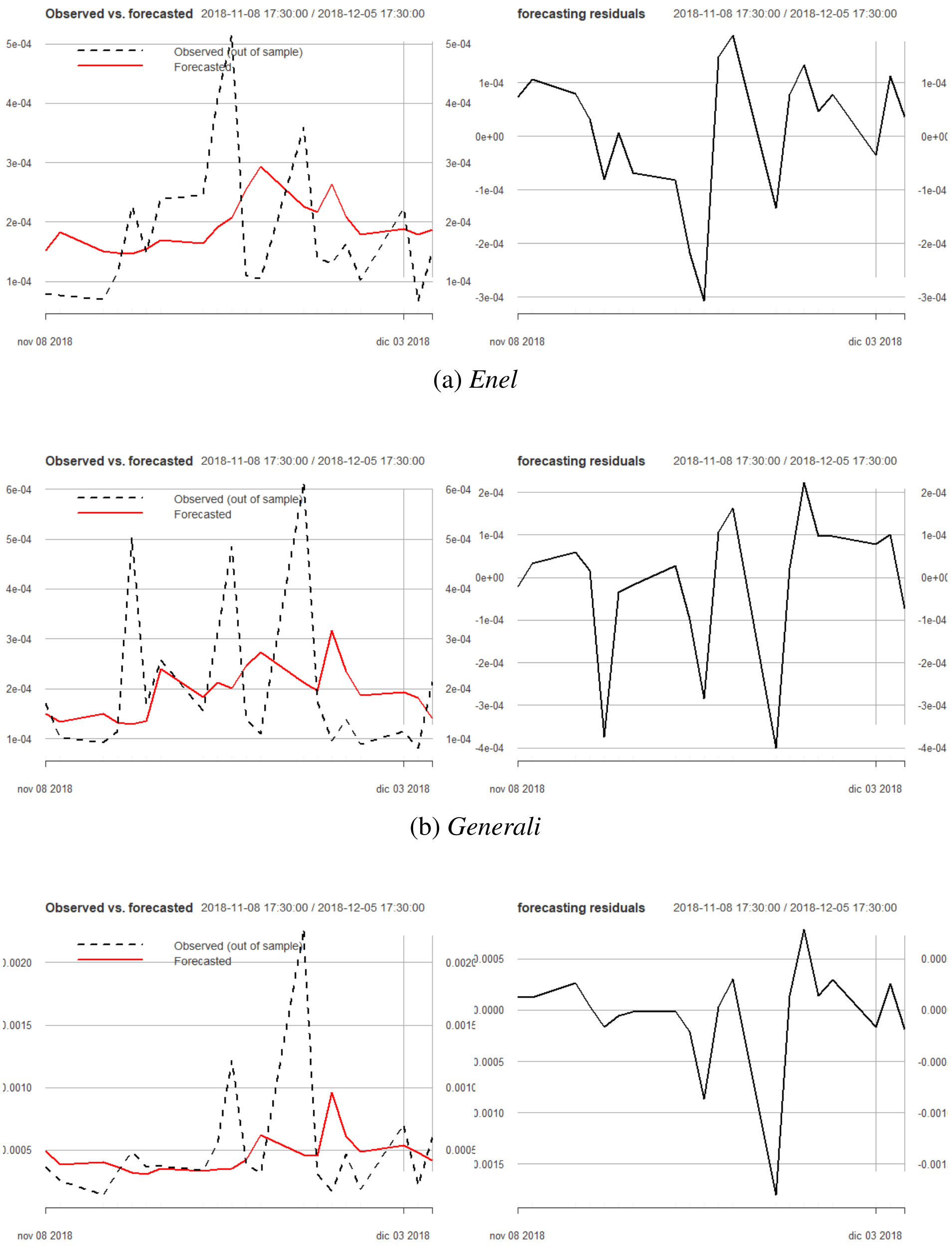
Table 4.10: Coefficients of determination for the HAR models, in both levels and logarithmic forms, applied to the Enel, Generali and Intesa San Paolo series.

Out-of-sample Performances		**Enel**	**Generali**	**ISP**
HAR (Levels)	**RMSE**	$1.236 \cdot 10^{-4}$	$1.623 \cdot 10^{-4}$	$5.071 \cdot 10^{-4}$
	MAE	$1.021 \cdot 10^{-4}$	$1.163 \cdot 10^{-4}$	$3.002 \cdot 10^{-4}$
	MAPE	0.7055	0.6390	0.7438
HAR (Logarithmic)	**RMSE**	$1.260 \cdot 10^{-4}$	$1.681 \cdot 10^{-4}$	$5.190 \cdot 10^{-4}$
	MAE	$0.963 \cdot 10^{-4}$	$1.105 \cdot 10^{-4}$	$2.902 \cdot 10^{-4}$
	MAPE	0.5814	0.5241	0.6353

Table 4.11: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) of the 20 days out-of-sample forecasts of the models, both in levels and in logarithmic form, applied to the realized volatilities of Enel, Generali and Intesa San Paolo.

observations. The models' parameters are estimated on the sample given by the first 291 values of realized volatility and used to forecast the 292nd one. Then, this resulting value is added into the sample and the regression coefficient estimates are updated on 292 values, in order to forecast the 293rd one. This rolling procedure is repeated until all the 20 missing days of observations are forecasted. Figures 4.19 and 4.20 show the comparison between actual and out-of-sample forecasted values, and the corresponding errors, of the HAR models, both in levels and logarithmic form, applied to the Enel, Generali and Intesa San Paolo realized volatility series, respectively.

Table 4.11 reports the values of the RMSE and the MAE of the out-of-

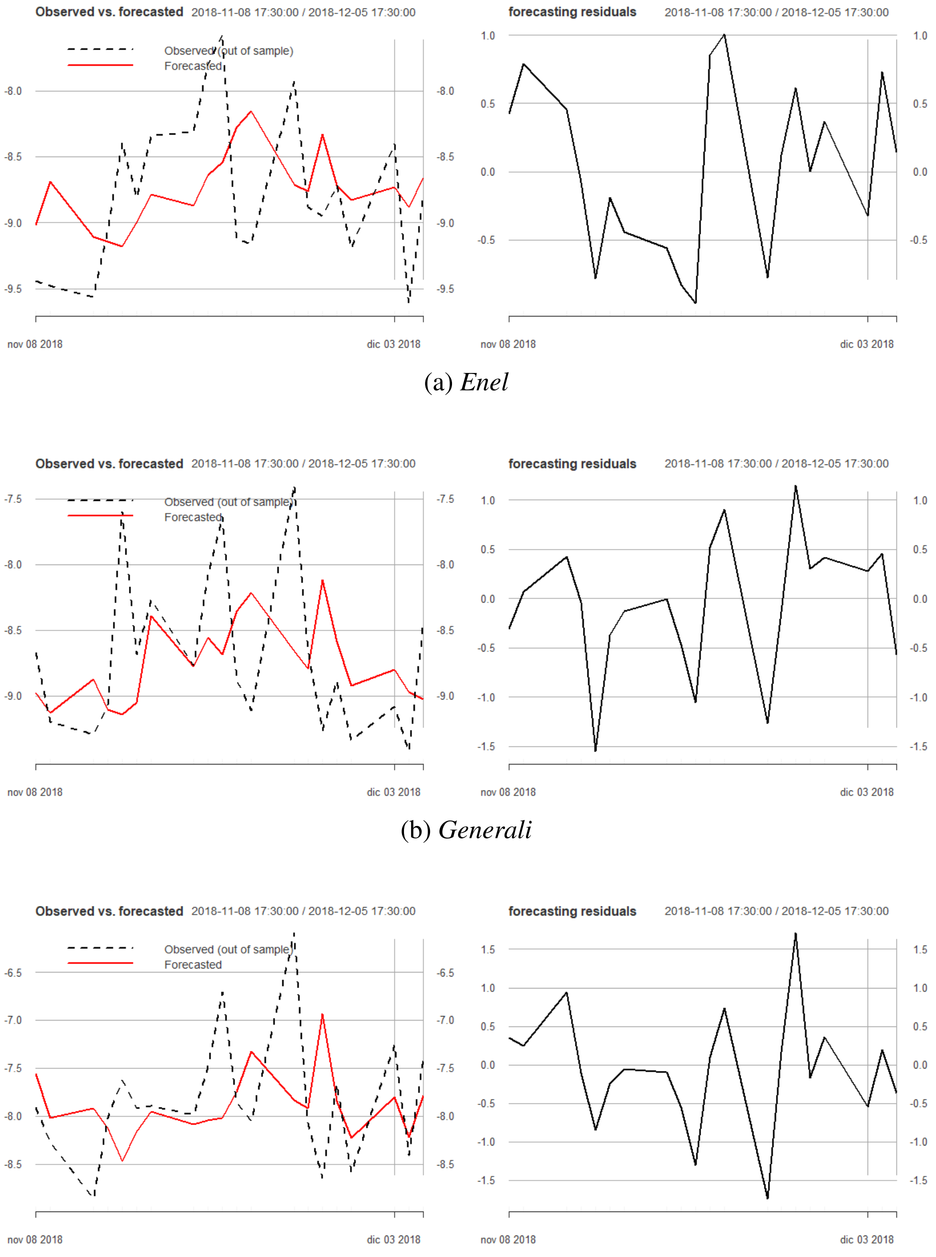


(a) *Enel*

(b) *Generali*

(c) *ISP*

Figure 4.19: Comparison between the out-of-sample predictions over 20 days and actual values (dotted line), together with the forecasting residuals, of the HAR models applied to the Enel, Generali and Intesa San Paolo series.



(a) *Enel*

(b) *Generali*

(c) *ISP*

Figure 4.20: Comparison between the out-of-sample predictions over 20 days and actual values (dotted line), together with the forecasting residuals, of the logarithmic HAR models applied to the Enel, Generali and Intesa San Paolo series.

sample forecasts, together with the Mean Absolute Percentage Error (MAPE). This measure is computed as following

$$MAPE = \frac{1}{Z}\sum_{z=1}^{Z}\left|\frac{y_z - \hat{y}_z}{y_z}\right|$$

where y_z and $\hat{y}_z$ are, respectively, the actual and the forecasted values at time z, $z = 1, \ldots, Z$. We can notice that the models applied to the Enel series lead to better results in terms of both RMSE and MAE, while, looking at the MAPE, the model with the better performances is the one applied to the Generali series. The logarithmic transformation allows improvements regarding the MAE and MAPE measures, but not the RMSE. The values taken by the MAPE are generally quite unsatisfactory.

4.5 Comparisons

To check if the HAR is a good model to forecast the realized volatility series[1] I compared its out-of-sample performances with the ones of the ARMA and ARFIMA models of realized volatility. The features of these models are better explained in **Appendix B**.

The order of the ARMA models have been chosen such to be the ones that minimize the BIC values. For all the three stocks it resulted to be ARMA(1,1). On the other hand, to find the optimal orders of the ARFIMA models I used a three-steps procedure. First of all, I estimated the fractional parameter d for each realized volatility series, using the method of Geweke and Porter-Hudak (GPH). Then, I differenced the three series with their corresponding d. Finally, I searched for the best ARIMA model to apply to the differenced series. For all the three stocks, the best order of the ARFIMA model to apply to the realized volatilities resulted to be ARFIMA(1,d,1).

After applying the ARMA(1,1) and ARFIMA(1,d,1) models to the realized

[1] Note that, from now on, I am not going to consider anymore the realized volatility in standard deviation and in logarithmic forms, since I found out that the raw realized volatility lead to models with better forecasting performances.

		Enel	Generali	ISP
HAR (Levels)	**RMSE**	$1.236 \cdot 10^{-4}$	$1.623 \cdot 10^{-4}$	$5.071 \cdot 10^{-4}$
	MAE	$1.021 \cdot 10^{-4}$	$1.163 \cdot 10^{-4}$	$3.002 \cdot 10^{-4}$
	MAPE	0.7055	0.6390	0.7438
HAR (Logarithmic)	**RMSE**	$1.260 \cdot 10^{-4}$	$1.681 \cdot 10^{-4}$	$5.190 \cdot 10^{-4}$
	MAE	$0.963 \cdot 10^{-4}$	$1.105 \cdot 10^{-4}$	$2.902 \cdot 10^{-4}$
	MAPE	0.5814	0.52405	0.6353
ARFIMA(1,d,1)	**RMSE**	$1.166 \cdot 10^{-4}$	$1.589 \cdot 10^{-4}$	$4.860 \cdot 10^{-4}$
	MAE	$0.947 \cdot 10^{-4}$	$0.993 \cdot 10^{-4}$	$0.2561 \cdot 10^{-4}$
	MAPE	0.6660	0.4269	0.4995
ARMA(1,1)	**RMSE**	$1.275 \cdot 10^{-4}$	$1.906 \cdot 10^{-4}$	$5.437 \cdot 10^{-4}$
	MAE	$0.915 \cdot 10^{-4}$	$1.415 \cdot 10^{-4}$	$3.525 \cdot 10^{-4}$
	MAPE	0.7640	0.8969	0.8272

Table 4.12: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) of the 20 days out-of-sample forecasts of the HAR (in both levels and logarithmic forms), ARFIMA and ARMA models applied to the realized volatilities of Enel, Generali and Intesa San Paolo.

volatility series of Enel, Generali and Intesa San Paolo, I performed the out-of-sample forecasts for the last 20 days of the samples, as I did with the HAR models. Then, I compared the forecasted values to the actual ones and computed the RMSEs, the MAEs and the MAPEs, which are reported in Table 4.12. Overall, the better performances are obtained by the ARFIMA(1,d,1) model.

4.6 Final remarks

From the results of this chapter, I can conclude that the HAR models do not fit very well to the data of Enel, Generali and Intesa San Paolo. Indeed, the residuals of the models applied did not seem to behave like a white noise: in their plots we could see some changing in the variability and their autocorrelations showed some significant values for the Generali and Intesa San

Paolo series. Moreover, the forecasting performances are not very satisfactory. The behavior of the residuals is improved when the models are applied through the logarithmic transformation. In this case, also the forecasting results seems to be a little better, since the model is able to capture a bit of the trend of the actual values. However, these not so accurate results could derive both from the way the HAR models are constructed, and from the computation of the realized volatility. As explained in Chapter 2, the realized volatility should ideally be calculated through the intraday returns of the efficient prices governed by an Ito process. However, this is practically not possible because the prices are contaminated by some microstructure noise. This noise component induces the realized volatility to be biased. In order to reduce this problem, it is commonly used an interval of five minutes to compute the return series, instead of higher frequency. In order to diminish the microstructure effects, I also tried to reduce the frequency of the data until 30 minutes, obtaining returns not serially correlated and unbiased realized volatility series. However, this alternative analysis lead to a general worsening of the results, both in terms of goodness-of-fit and in term of forecasting performances. Obviously, this procedure allowed me to gain in the unbiasedness of the estimator, but not without any drawback. Indeed, sampling at lower frequencies forced me to renounce to many observations and this lost of information could have lead to an inefficient estimate of volatility, compromising the analysis.

Conclusions

In the last decades, the availability of high frequency data has presented new challenges to econometricians. Indeed, they show features that classical parametric models are not able to capture.
In this context, one of the most important developments has been the proposal of a new daily volatility estimator, based on high frequency returns: the realized volatility. The HAR class of model, discussed by Corsi (2003) and Andersen et al. (2007), attempts to forecast the next day's realized volatility relying on high-frequency returns of the past. The purpose was to obtain a parsimonious model that was easy to estimate, but at the same time able to reproduce the persistence of the volatility of financial data.

Throughout this thesis, I explained the computation of the realized volatility estimator and the formulation of the HAR class of models, with its developments regarding the jump and leverage components. Then, I computed the realized volatility for the return series of Enel, Generali and Intesa San Paolo, relying on the returns over intervals of 5 minutes. When applying the HAR class of models, the results were not as expected: the models did not fit well the data and, as a consequence, the forecasting performances were not satisfying. Indeed, the root mean square error and the absolute error of an ARFIMA model's forecasts resulted to be lower than the ones obtained from the forecasts of the HAR models. The non satisfactory goodness-of-fit and forecasts could be the consequence of the difficulty in finding an equilibrium in the trade-off between the removal of the microstructure noise (reducing the frequency) and the goodness of the approximation of the quadratic variation (that requires a high frequency). However, if we do not aim to get an accurate estimator, but rather an indicator of the trend of the volatility, the forecasting results of Figure 4.20 could be helpful.

A possible approach to obtain better results could have been computing the two scale realized volatility of Zhang et al. (2005). However, this unbiased estimator of volatility need tick-by-tick data to be calculated, which are difficult to obtain, as well as very expensive.

Appendix A

A model of nonsynchronous trading: computations.

A well known property, obtained by taking the first derivative of the geometric series $g(\pi) = 1 + \pi + \pi^2 + \pi^3 + \cdots$, is that:

$$1 + 2\pi + 3\pi^2 + 4\pi^3 + \cdots = \frac{1}{(1-\pi)^2}$$

Then, the result of equation 1.1 has been achieved as following:

$$\begin{aligned}
\mathbb{E}(r_t^0) &= (1-\pi)^2\,\mathbb{E}(r_t) + (1-\pi)^2\pi\,\mathbb{E}(r_t + r_{t-1}) + \cdots \\
&= (1-\pi)^2\mu + (1-\pi)^2\pi 2\mu + (1-\pi)^2\pi^2 3\mu + \cdots \\
&= (1-\pi)^2\mu[1 + 2\pi + 3\pi^2 + 4\pi^3 + \cdots] \\
&= (1-\pi)^2\mu\frac{1}{(1-\pi)^2} \\
&= \mu.
\end{aligned}$$

Regarding the computation of the variance in equation 1.2, assuming that the returns are serially independent, we need to know the following rules:

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=0}^{k} r_{t-i}\right)^2 &= \mathbb{V}\mathrm{ar}\left(\sum_{i=0}^{k} r_{t-i}\right) + \left[\mathbb{E}\left(\sum_{i=0}^{k} r_{t-i}\right)\right]^2 \\
&= (k+1)\sigma^2 + [(k+1)\mu]^2
\end{aligned}$$

and

$$1 + 4\pi + 9\pi^2 + 16\pi^3 + \cdots = \frac{2}{(1-\pi)^3} - \frac{1}{(1-\pi)^2}.$$

The second rules is derived considering $H = 1 + 4\pi + 9\pi^2 + 16\pi^3 + \cdots$ and $G = 1 + 3\pi + 5\pi^2 + 7\pi^3 + \cdots$.
Then, $(1-\pi)H = G$ and

$$\begin{aligned}(1-\pi)G &= 1 + 2\pi + 2\pi^2 + 2\pi^3 + \cdots \\ &= 2(1 + \pi + \pi^2 + \cdots) - 1 \\ &= \frac{2}{(1-\pi)} - 1.\end{aligned}$$

Hence, we obtain the variance as

$$\begin{aligned}\mathbb{V}\mathrm{ar}(r_t^0) &= \mathbb{E}[(r_t^0)^2] - [\mathbb{E}(r_t^0)]^2 \\ &= (1-\pi)^2\,\mathbb{E}[(r_t)^2] + (1-\pi)^2\pi\,\mathbb{E}[(r_t + r_{t-1})^2] + \cdots - \mu^2 \\ &= (1-\pi)^2[(\sigma^2 + \mu^2) + \pi(2\sigma^2 + 4\mu^2) + \pi^2(3\sigma^2 + 9\mu^2) + \cdots] - \mu^2 \\ &= (1-\pi)^2\{\sigma^2[1 + 2\pi + 3\pi^2 + \cdots] + \mu^2[1 + 4\pi + 9\pi^2 + \cdots]\} - \mu^2 \\ &= \sigma^2 + \mu^2\left[\frac{2}{1-\pi} - 1\right] - \mu^2 \\ &= \sigma^2 + \frac{2\pi\mu^2}{1-\pi}.\end{aligned}$$

To compute the covariance at lag one of Equation 1.3, it is useful to define the product $r_t^0 r_{t-1}^0$ as

$$r_t^0 r_{t-1}^0 = \begin{cases} 0 & \text{with probability } 2\pi - \pi^2 \\ r_t r_{t-1} & \text{with probability } (1-\pi)^3 \\ r_t(r_{t-1} + r_{t-2}) & \text{with probability } (1-\pi)^3\pi \\ r_t(r_{t-1} + r_{t-2} + r_{t-3}) & \text{with probability } (1-\pi)^3\pi^2 \\ \vdots & \vdots \\ r_t\left(\sum_{i=0}^{k} r_{t-i}\right) & \text{with probability } (1-\pi)^3\pi^{k-1} \\ \vdots & \vdots \end{cases}$$

Moreover, for $j > 0$, we know that $\mathbb{E}(r_t r_{t-j}) = \mathbb{E}(r_t)\,\mathbb{E}(r_{t-j}) = \mu^2$. Hence, we obtain that

$$\begin{aligned}
Cov(r_t^0, r_{t-1}^0) &= \mathbb{E}(r_t^0 r_{t-1}^0) - \mathbb{E}(r_t^0)(r_{t-1}^0) \\
&= \mathbb{E}(r_t^0 r_{t-1}^0) - \mu^2 \\
&= (1-\pi)^3 \Bigg\{ \mathbb{E}(r_t r_{t-1}) + \pi\,\mathbb{E}[r_t(r_{t-1} + r_{t-2})] + \\
&\quad + \pi^2\,\mathbb{E}\left[r_t \left(\sum_{i=1}^{3} r_{t-i} \right) \right] + \cdots \Bigg\} - \mu^2 \\
&= (1-\pi)^3 \mu^2 (1 + 2\pi + 3\pi^2 + \cdots) - \mu^2 \\
&= (1-\pi)\mu^2 - \mu^2 \\
&= -\pi\mu^2.
\end{aligned}$$

Appendix B

ARMA and ARFIMA models

The Autoregressive Fractionally Integrated Moving Average, or ARFIMA, model can be used in modelling the long-run behavior of a time series, since it is able to capture it without the problems that an ARMA model would face.

Let $\{y_t\}$ be a discrete time real-valued process, L the lag operator and ϵ_t a process with zero mean and no autocorrelation.
The class of processes mostly used to model time series is the ARMA(p, q) model, defined as:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}$$

or, through the use of the lag operator L:

$$(1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p) Y_t = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_p L^p) \epsilon_t,$$

Let $\Phi(L)$ and $\Theta(L)$ be the following polynomials of respectively order p and q:

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$$
$$\Theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots - \phi_q L^q.$$

Then, we can rewrite the ARMA(p,q) model as:

$$\Phi(L) Y_t = \Theta(L) \epsilon_t \qquad \epsilon_t \sim (0, \sigma_\epsilon^2) \tag{4.1}$$

where the roots of the two polynomials have to stand outside the unit circle to induce stationarity. Anyway, the long-run behavior of time series is not well captured by this class of models for many reasons:

- the long-run behavior is captured by parameters that are near the boundary of the parameter space. This means that the sampling distributions are not well approximated by the asymptotic distributions so that inference is not reliable;
- if a model manage to capture the long-run behavior of the data, it imposes restriction on the short-run. Indeed, if an AR's parameter models the correlation at high lags, it has to model it also at lower lags.
- the parameter are estimated by MLE in a way such that may sacrifice the long-run behavior to obtain a better fit for the short-run. Indeed we can read in Sowell (1992) that "as we pointed out in Cochrane (1988) maximum likelihood (asymptotically) chooses parameter values to minimize the difference between the periodogram of the data and the spectral density of the parametric model weighted at different frequencies. (...) There is no way to direct the fit of an AR or an MA parameter to the long-run characteristics of a series, even though a researcher may be investigating long-run behavior".

When the polynomials in 4.1 have positive and real unit roots, the function Y_t should become stationary if differenced. After defining the backward difference operator ∇ as

$$\nabla Y_t = Y_t - Y_{t-1} = (1-L)Y_t, \tag{4.2}$$

it is possibile to rewrite eq. 4.1 as

$$\Phi(L)\nabla^d Y_t = \Theta(L)\epsilon_t. \tag{4.3}$$

The differenced function ∇Y_t is said to be an Autoregressive Integrated Moving Average (ARIMA) model of order $(p-1, d, q)$. The number of differentiations needed to induce stationarity to the model is specified by the parameter d:

- when $d = 0$, Y_t results to be stationary and eq. 4.3 results to be equal to eq. 4.1;
- when $d = 1$, Y_t is non stationary, but its first difference is stationary;
- when $d = 2, 3, \ldots$, Y_t has to be differenced d-times to become stationary;
- when d is a real value, the model becomes an *Autoregressive Fractionally Integrated Moving Average*, or ARFIMA(p,d,q).

Regarding the fourth case, it is possible to rewrite eq. 4.2 replacing ∇Y_t with an infinite order autoregressive process such that

$$\nabla^d Y_t = (1-L)^d Y_t = \sum_{k=0}^{\infty} (-1)^k \begin{pmatrix} d \\ k \end{pmatrix} L^k Y_t$$

Hence, the ARFIMA specification can be defined as

$$\Phi(L)(1-L)^d Y_t = \Theta(L)\epsilon_t \tag{4.4}$$

The ARFIMA model exhibits an infinite lag order dependence and could be more suitable to capture the long-run behavior of a time series. To satisfy the stationarity and invertibility conditions the value of the parameter d has to be lower in modulus than 0.5. Indeed, when $d > -0.5$ the process is invertible and has a linear representation (Wold representation); when $d < 0.5$ the process is weakly stationary. If $d \geq 0.5$, it is proved in Granger and Joyeux (1980) that the variance of the process is infinite, causing its nonstationarity. However, as explained in Sowell(1992), "long-range dependence is associated with all nonzero d>0, which allows capturing the long-run behavior without being 'close to the boundary' of the parameter space. This long-run dependence is achieved with less restrictions on the higher frequency behavior of the time series."

When $d > 0$, the process is said to be long memory because, being ρ_j the autocorrelation at lag j, the limit $lim_{k\to\infty} \sum_{j=-k,k} |\rho_j|$ is not converging to a finite number. As stated in Baillie et. al. (1996), "the ARFIMA model essentialy disentangles the short-run and the long-run dynamics, by modelling the short-run behaviour through the conventional ARMA lag polynomials, $a(L)$ and $b(L)$, while the long-run characteristic is captured by the fractional differencing parameter, d."

Bibliography

[1] Aït-Sahalia, Y. and Hansen, L. (2010). *Handbook of financial econometrics: tools and techniques. Volume 1*. North-Holland. Amsterdam.

[2] Aït-Sahalia, Y. and Jacod, J. (2014). *High-frequency financial econometrics*. Princeton University Press. Princeton, New Jersey.

[3] Aït-Sahalia, Y., Mykland, P., Zhang, L., (2005). *How often to sample a continuous-time process in the presence of market microstructure noise*. The Review of Financial Studies, 351-416.

[4] Andersen, T. G. and Bollerslev, T. (1997). *Intraday periodicity and volatility persistence in financial markets*. Journal of Empirical Finance 4, 115-158.

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

[6] Andersen, T. G., Bollerslev, T., and Diebold, F.X. (2002). *Parametric and Nonparametric volatility measurement*. in Handbook of Financial Econometrics, ed. by L. P. Hansen and Y. A-Sahalia, Amsterdam: North Holland, 2010.

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

[8] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). *The distribution of realized exchange rate volatility*. Journal of the American Statistical Association, 96, 42-55.

[9] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). *Modeling and forecasting realized volatility*. Econometrica 71, 579-625.

[10] Andersen, T. G., Bollerslev, T., and Meddhai, N. (2004). *Analytic evaluation of volatility forecasts*. International Economic Review 45, 1079-1110.

[11] Back, K. (1991). *Asset prices for general processes*. Journal of mathematical economics, 20, 317-395.

[12] Baillie, R.T.Bollerslev, T., Mikkelsen, H.O. (1996). *Fractionally integrated generalized autoregressive conditional heteroskedasticity*. Journal of Econometrics 74, pp. 3-30.

[13] Bandi, F., Russell, J. R., (2008). *Microstructure noise, realized volatility, and optimal sampling*. The Review of Economic Studies, 2, 339-369.

[14] Barndorff-Nielsen, O. E., Hansen, P., Lunde, A., Shephard, N., (2008). *Designing realized kernel to measure the ex-post variation of equity prices in the presence of noise*. Econometrica, 76, 1481-1536.

[15] Barndorff-Nielsen, O. E. and Shephard, N., (2002a). *Econometric analysis of realized volatility and its use in estimating stochastic volatility models*. Journal of the Royal Statistical Society, Series B, 64, Part 2, 253-280.

[16] Barndorff-Nielsen, O. E. and Shephard, N., (2002b). *Estimating quadratic variation using realized variance*. Journal of Applied Econometrics 17, 457-478.

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

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

[19] Bollerslev, T., Mikkelsen, H.O., (1996). *Modeling and pricing long memory in stock market volatility*. Journal of Econometrics 73, pp. 151-184.

[20] Campbell, J.Y., Lo, A.W.,and MacKinlay, A.C.(1997). *The econometrics of financial markets*. Princeton University Press. Princeton, New Jersey.

[21] Cochrane, J. H., (1988).*How big is the random walk in GNP?*. Journal of Political Economy 96, 893-920.

[22] Corsi, F. (2009). *A simple approximate long-memory model of realized volatility*. Journal of Financial Econometrics, vol.7, no. 2, pp.174-196.

[23] Corsi, F., Renò, R. (2012). *Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling*. Journal of Business & Economic Statistics, vol. 3, pp. 368-380.

[24] Cryer, J. D., and Chan K. (2008). *Time series analysis with applications in R*. Second edition. Springer. New York.

[25] Ellis, C., (1999). *Estimation of the ARFIMA(p,d,q) fractional differencing parameter (d) using the classical rescaled adjusted range technique*. International Review of Financial Analysis 8, 53-65.

[26] Emil Sjoerup (2018). HARModel: Heterogeneous Autoregressive Models. R package version 0.1. https://CRAN.R-project.org/package=HARModel

[27] Engle R. F. (1982). *Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation*. Econometrica, 50, 987-1007.

[28] Engle R. F. (2001). *GARCH 101: The use of ARCH/GARCH models in applied econometrics*. Journal of Economic Perspectives, 15, 157-168.

[29] Engle, R. F. (2002). *New frontiers for ARCH models*. Journal of Applied Econometrics 17, pp. 425-446.

[30] Granger, C. W. J., Joyeux, R. (1980). *An introduction to long-memory time series models and fractional differencing*. Journal of Time Series Analysis 1, 15-39.

[31] Hautsch, N. (2012). *Econometrics of financial high-frequency data*. Springer.

[32] Huang, X., Tauchen, G., (2005). *The relative contribution of jumps to total price variance*. Journal of Financial Econometrics 3, 456-499.

[33] Hull, J., White, A. (1987), *The pricing of options on assets with stochastic volatilities*. Journal of Finance, 42, 381-400.

[34] Jacod, J, Li, Y. Mykland, P. Podolskij, M., Vetter, M. (2009). *Microstructure noise in the continuous case: the preaveraging approach*. Stochastic Processes and their Applications, 119, 2249-2276.

[35] Kris Boudt, Jonathan Cornelissen and Scott Payseur (2018). highfrequency: Tools for Highfrequency Data Analysis. R package version 0.5.3. https://CRAN.R-project.org/package=highfrequency

[36] Lo, A, MacKinlay, A.C., (1990).*An econometric analysis of nonsynchronous trading*. Journal of Econometrics, 45, 181-212.

[37] Meddhai, N. (2002). *A theoretical comparison between integrated and realized volatility*. Journal of Applied Econometrics 17, 479-508.

[38] Müller, U., Dacarogna, M., Dav, R., Pictet, O., Olsen, R., and Ward, J. (1993). *Fractals and intrinsic time - A challenge to econometricians*. 39th International AEA Conference on Real Time Econometrics, 14–15 October 1993, Luxembourg.

[39] Pigorsch, C., Pigorsch, U., Popov, I., (2012). *Volatility estimation based on high-frequency data*.In: Duan, J. C., Härdle, W. K., Gentle, J. E., Handbook of Computational Finance, Springer-Verlag Berlin Heidelberg, 335-369.

[40] Shumway, R. H., and Stoffer, D. S. (2011). *Time series analysis and its applications*. Third edition. Springer. New York.

[41] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[42] Roll, R. (1984). *A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market*. The Journal of Finance, 39(4), 1127-1139.

[43] Sowell, F.B. (1992). *Modeling long-run behaviour with the fractional ARIMA model*. Journal of Monetary Economics 29, pp.277-302.

[44] Taylor, S. (2005). *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press. Princeton, New Jersey.

[45] Tsay, R. S. (2002). *Analysis of financial time series*. Wiley. New York.

[46] Tsay, R. S. (2013). *An introduction to analysis of financial data with R*. Wiley. Hoboken, New Jersey.

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