Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

RIGOROSIS DIPLOMA THESIS

Forecasting realized volatility: Do jumps in prices matter?

Author: **Mgr. Štefan Lipták**

Supervisor: **PhDr. Jozef Baruník, Ph.D.**

Academic Year: **2013/2014**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, February 10, 2014

Signature

Acknowledgments

I am very grateful to my supervisor, Jozef Baruník, for his great patience and valuable suggestions. His guidance was more than necessary for the completion of this thesis. He also provided me with the data, which I am thankful for.

I would like to thank Diana for her help, support and endless love as she is always there for me.

Abstract

This thesis uses Heterogeneous Autoregressive models of Realized Volatility on five-minute data of three of the most liquid financial assets – S&P 500 Futures index, Euro FX and Light Crude NYMEX. The main contribution lies in the length of the datasets which span the time period of 25 years (13 years in case of Euro FX). Our aim is to show that decomposing realized variance into continuous and jump components improves the predicatability of RV also on extremely long high frequency datasets. The main goal is to investigate the dynamics of the HAR model parameters in time. Also, we examine whether volatilities of various assets behave differently.

Results reveal that decomposing RV into its components indeed improves the modeling and forecasting of volatility on all datasets. However, we found that forecasts are best when based on short, 1-2 years, pre-forecast periods due to high dynamics of HAR model's parameters in time. This dynamics is revealed also in a year-by-year estimation on all datasets. Consequently, we consider HAR models to be inappropriate for modeling RV on such long datasets as they are not able to capture the dynamics of RV. This was indicated on all three datasets, thus, we conclude that volatility behaves similarly for different types of assets with similar liquidity.

JEL Classification	C22, C50, C58, G17
Keywords	quadratic variation, realized volatility, realized variance, high frequency data, heterogeneous autoregressive model

Author's e-mail	`stefan.liptak86@gmail.com`
Supervisor's e-mail	`barunik@utia.cas.cz`

Abstrakt

Táto práca aplikuje heterogénny autoregresný model realizovanej volatility na päť-minútové dáta troch spomedzi najlikvidnejších finančných aktív – S&P 500 Futures index, Euro FX a ropa. Hlavný prínos tejto práce spočíva v analyzovaní mimoriadneho množstva dát, keďže pochádzajú z neobyčajne dlhého obdobia až 25 rokov, v prípade Euro FX je to 13 rokov. Jedným z cieľov je ukázať, že rozklad realizovanej variancie na spojitú a skokovú časť má pozitívny vplyv na jej predpovedateľnosť aj na vysokofrekvenčných dátach pokrývajúcich veľmi dlhé obdobia. Hlavným cieľom práce je skúmať dynamiku parametrov HAR modelu v čase, a taktiež povahu volatility u rôznych druhov finančných aktív.

Výsledky analýzy na dátach všetkých troch aktív potvrdzujú, že rozklad realizovanej variancie prispieva k vylepšeniu odhadov. Ukázalo sa však, že predpovedacia schopnosť modelu je najlepšia v prípade, že parametre boli odhadnuté na krátkych obdobiach (1-2 roky), čo je spôsobené pravdepodobne vysokou dynamikou parametrov v čase. Táto nestabilita parametrov bola odhalená aj s pomocou odhadov za jednotlivé roky, a to u všetkých súborov. Z toho vyplýva zaujímavé zistenie, a to že HAR model nie je vhodný na predpovedanie realizovanej volatility na dlhých dátach, keďže nie je schopný zachytiť dynamiku parametrov modelu. Celkovo boli výsledky pre všetky aktíva do značnej miery podobné, z čoho usudzujeme, že volatilita rôznych typov aktív nie je príliš špecifická.

Klasifikace JEL	C22, C50, C58, G17
Klíčová slova	kvadratická variácia, realizovaná volatilita, realizovaná variancia, vysokofrekvenčné dáta, heterogénny autoregresný model

E-mail autora	`stefan.liptak86@gmail.com`
E-mail vedoucího práce	`barunik@utia.cas.cz`

Contents

List of Tables

List of Figures

Acronyms

ACH	Autoregressive Conditional Hazard (model)
AR	Autoregressive (model or process)
BV	Bipower Variation
CL	Light Crude NYMEX
CME	Chicago Mercantile Exchange
EC	Euro FX
FX	Foreign Exchange
GARCH	Generalized Autoregressive Conditional Heteroskedasticity (model)
HAR	Heterogeneous Autoregressive (model)
HAR-RV	Heterogeneous Autoregressive model of Realized Variance
MSE	Mean Square Error
NYMEX	New York Mercantile Exchange
OLS	Ordinary Least Squares
QV	Quadratic Variation
RMSE	Root Mean Square Error
RV	Realized Variance
RVO	Realized Volatility
SP	Standard & Poor 500 Futures index
TQ	Tripower Quarticity
TSRV	Two-Scale Realized Variance (estimator)

Master Thesis Proposal

Author	Mgr. Štefan Lipták
Supervisor	PhDr. Jozef Baruník, Ph.D.
Proposed topic	Forecasting realized volatility: Do jumps in prices matter?

Topic characteristics Volatility in financial markets is essential for asset pricing, asset allocation and hedging or risk management. Most of the models are based on assumptions like volatility or prices following a continuous path. However, macroeconomic news, firm-specific information or other economic news can cause dramatic changes in prices over a very short time period, which is in contrast with the assumption of continuous sample price paths. Recent studies show that discontinuous price jumps are indeed important and have a significant impact on volatility and thus also on asset pricing etc. In this thesis we are going to examine the role of price jumps and their effect on volatility using high-frequency data.

Hypotheses

1. There are significant jumps in the price paths in currency markets.
2. It is possible to predict the price jumps.
3. Predicting jumps improves the predictability of volatility and prices.

Methodology The most important part of my work will be the detection of jumps in prices. As we know, significant jumps occur between the opening price of one day and closing price of the previous day. However, we will focus only on detection of intra-day jumps. To do so we will use common methods which are based on realized variation measures.

We use realized variance (sum of squared returns) instead of the unobservable quadratic variation, which consists of a term representing the continuous

price path and a term representing the within-day jumps. Using high-frequency data ensures that realized variance converges in probability to quadratic variation (Andersen et al. (2003)).

Realized bi-power variation depends on the sum of products of absolute values of consequent intra-day returns and it can be shown that this variation converges in probability to the continuous price path component of the quadratic variation (Barndorff-Nielsen (2004)).

Given the mentioned properties, it is possible to estimate the price jumps as the difference between the realized variance and the bi-power variation. In the following parts of the work, standard econometric methods will be used to perform forecasting exercise in order to test our hypotheses.

Outline

1. Introduction
2. Theory of realized variation measures
3. Estimation of jumps - methodology
4. Methodology of forecasting
5. Description of the data
6. Decomposition of prices
7. Conclusion and discussion of the results

Core bibliography

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

2. Andersen, T. G., T. Bollerslev, F. X. Diebold & H. Ebens (2001):"The distribution of realized stock return volatility." *Journal of Financial Economics* **61(1)**: pp. 43–76.

3. Andersen, T. G., T. Bollerslev, F. X. Diebold & P. Labys (2003):"Modeling and Forecasting Realized Volatility." *Econometrica* **71(2)**: pp. 579–625.

4. Andersen, T. G., T. Bollerslev & X. Huang (2010):"A reduced form framework for modeling volatility of speculative prices based on realized variation measures." *Journal of Econometrics* **160(1)**: pp. 176–189.

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

6. Bollerslev, T., T. H. Law & G. Tauchen (2008):"Risk, jumps, and diversification." *Journal of Econometrics* **144(1)**: pp. 234–256.

7. Christensen, B. J. & M. Ø. Nielsen (2005): "The Implied-Realized Volatility Relation with Jumps in Underlying Asset Prices." *Working Papers 1186*, Queen's University, Department of Economics

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

9. Fleming, J. & B. S. Paye (2010): "High-frequency returns, jumps and the mixture of normals hypothesis." *Journal of Econometrics* **160(1)**: pp. 119–128.

Author Supervisor

Chapter 1

Introduction

Volatility can be understood as a measure of riskiness of financial instruments over a given time period and is, therefore, essential for asset pricing, hedging or risk management. Great desire of traders for knowledge of future volatility has made it one of the central concerns of financial econometrics. Only a few years ago, GARCH or stochastic volatility models were used to model volatility on daily (and even coarser) frequency data. Now, however, the existence and availability of high frequency data has made it possible to observe the (until then) unobservable part of stochastic volatility models. A new non-parametric realized measure of volatility has occured, called the *Realized Volatility*, which is based on summing the squared intraday high frequency returns.

The concept of realized volatility (and realized variance) was first introduced by Andersen *et al.* (2001). Other works concerning the theoretical properties of realized volatility include Andersen *et al.* (2003) and Barndorff-Nielsen & Shephard (2004). A Heterogeneous Autoregressive model of Realized Volatility (HAR-RV) was proposed by Corsi (2004) based on the new volatility measure and Heterogeneous Market Hypothesis of Müller *et al.* (1997). This model uses three volatility components, each stemming from one of the three main types of market agents, ensuring the ability to capture the persistence of volatility.

Nevertheless, Corsi (2004) considered the price process of an asset to be continuous, while empirical findings pointed to the existence of rather incontinuous price processes – processes containing a jump component and a continuous part. Therefore, Barndorff-Nielsen & Shephard (2004) and Barndorff-Nielsen & Shephard (2006) introduce the concept of bipower variation which plays a key role in separating the jump component from the continuous part of a process. A numerous list of works on the importance of jumps includes Christensen &

Nielsen (2005), Andersen *et al.* (2007) and Bollerslev *et al.* (2008). We will further review relevant literature in the following chapter as the theory of realized variance will be introduced.

The empirical part of the thesis partly follows the work Andersen *et al.* (2010) where components of realized volatility are modeled separately. However, we are going to use a simpler tool for the modeling of RV as our contribution lies in something else. First of all, we have datasets almost twice as long as were used in the mentioned paper as we want to find out whether decomposing RV improves its modeling and forecasting also on such long datasets, i.e. we compare HAR-RV and HAR-RV-CJ models to see if jumps really do matter. Our second goal is to investigate if the HAR model is appropriate for RV modeling. To do so, we first carry out a year-by-year estimation of the parameters and then perform a one-year out-of-sample forecast on various lengths of pre-forecast periods. Moreover, three different types of highly liquid assets (a stock market index, a currency exchange rate and a commodity) are used to see if the effects are different among assets, which is our third objective.

Our results indicate that jumps in prices do matter as HAR-RV-CJ models provide a better fit than HAR-RV models. However, comparison of the forecasting performances offers no clear recommendation, which, we believe, has to do with the second objective of the thesis. Year-by-year estimations reveal significant dynamics of the parameters in time. This finding is supported by results of the predictions based on different pre-forecast periods, as the best one-year out-of-sample forecasts were obtained from parameters estimated on short pre-forecast periods. These results suggest that HAR model is not appropriate for realized variance modeling as the model is not stationary and performing OLS estimation automatically assumes parameters stable in time. All three datasets gave very similar results, therefore, there is no reason to think that volatilities of different assets behave differently and should be modeled each by a specific model.

The rest of this work is organized as follows. Chapter 2 provides an overview of the theory behind realized variance and variation measures, coupled with subsistent literature reviews. In Chapter 3 we introduce the bipower variation and jump detection test statistic, followed by a summary of the HAR class of models used in the empirical part. In the end, we present three methods for evaluation of forecasts. Datasets and the process of preparing the data for estimations are described in Chapter 4. Results of our estimations are reported in Chapter 5. Chapter 6 concludes.

Chapter 2

Theory of realized variation measures

This chapter is dedicated to the theoretical background of volatility modeling based on realized variation measures. The development of this theory was based on Back (1991). Main works that further contributed notably include Andersen *et al.* (2003) and Barndorff-Nielsen & Shephard (2004). As there is still no cohesive theory on this concept, we rely mostly on three works throughout this chapter – Andersen *et al.* (2003), Barndorff-Nielsen & Shephard (2004) and Andersen *et al.* (2010).

First, we build up the settings of the framework in which we look at the price (and return) process as a special semimartingale and we use this property to show that the return process can be decomposed into a predictable and unpredictable part. Then, the quadratic variation is introduced with its main properties followed by the definition of the realized variance (the estimator of the quadratic variation). Finally, we discuss the problem of microstructure noise and mention some of its possible solutions.

2.1 Price processes under conditions of continuous time and no arbitrage

Let us begin with the asset returns, which we assume to consist of two parts. The first one is the predictable component, which compensates the investor for the risk of holding the security. The second part is the unobservable shock, which we are not able to predict based on the available information. We also assume the absence of arbitrage opportunities which is a quite important as-

sumption as it has significant impact on modeling and measuring of variation in continuous time. We now continue with the definition of the setting.

Consider a univariate logarithmic price process p_t ($p_t = \ln P_t$, where P_t denotes the price process of an asset) of an asset, defined on a complete probability space $(\Omega, \mathcal{F}, P)$ and evolving in continuous time over the time interval $[0, T]$ (T is a positive finite integer). Further, we consider the natural information filtration, an increasing family of σ–fields $(\mathcal{F}_t)_{t\in[0,T]} \subseteq \mathcal{F}$, which satisfies the usual conditions of P-completeness and right continuity. We finally assume that the information set $\mathcal{F}_t$ contains information about all the asset prices and relevant state variables that occured from time 0 until time t.

Now let us define the continuous return of an asset as proposed in Andersen *et al.* (2003).

Definition 2.1. Let $[t-h, t]$ be a time interval, where $0 \leq h \leq t \leq T$. Then the continuously compounded asset return over $[t-h, t]$ is the difference between the logarithmic price at time t and the logarithmic price at time $t-h$.

$$r_{t,h} = p_t - p_{t-h}$$

We here establish that from now on, if not stated otherwise, by $[t-h, t]$ we will denote a time interval, where $0 \leq h \leq t \leq T$. Directly from the previous definition we have the special case of the continuously compounded return, the cumulative return from time $t_0 = 0$ up to time t, $r_t = (r_t)_{t\in[0,T]}$, which is $r_t \equiv r_{t,t} = p_t - p_0$. Furthermore, we can obtain a very simple, but important relation between the period-by-period and the cumulative returns: $r_{t,h} = r_t - r_{t-h}$, $0 \leq h \leq t \leq T$.

It is also very convenient to assume that the price process remains almost surely (henceforth *a.s.*) strictly positive and finite so that p_t and r_t are well defined over $[0, T]$ (*a.s.*). Also, there are only countable number of jumps (jump points) in the return process r_t over the time interval $[0, T]$, and both the price process and return process are squared integrable. Let us define $r_{t-} \equiv \lim_{\tau\to t-} r_\tau$ and $r_{t+} \equiv \lim_{\tau\to t+} r_\tau$. Then we are able to determine the right-continuous, left-limit (càdlàg[1]) version of the process, for which $r_t = r_{t+}$ (a.s.), and the left-continuous, right-limit (càglàd[2]) version, for which $r_t = r_{t-}$ (a.s.), $\forall t \in [0, T]$. Without loss of generality, we will work with the càdlàg version of the return process in the following text.

[1]From French "'continu à droite, limite à gauche"'.
[2]From French "'continu à gauche, limite à droite"'.

Given the previous, we impose the jumps in the return process as: $\Delta r_t = r_t - r_{t-}$, $0 \leq t \leq T$. Naturally, we have $\Delta r_t = 0$ at continuity points and also $P[\Delta r_t \neq 0] = 0$ for any arbitrary chosen $t \in [0, T]$. However, the previous assumption only implies that there is countable number of jumps in the price process but it says nothing about how often they occur. Moreover, we need the assumption that the jump process is not explosive. We will call such process a *regular* process with a finite number of jumps.

Having completed the basic introduction to price processes, we now need to impose the final standard assumptions to complete the definition of a continuous-time no-arbitrage price process. As showed by Back (1991), assuming that a return process is arbitrage-free and has a finite mean, the price process must belong to the class od special semi-martingales. These processes permit a unique canonical decomposition, as stated by a fundamental result of stochastic integration theory (e.g. Protter (1992)). Let us recall that a martingale is a process for which at each time of the realized sequence, the expected value of the next realization does not depend on the previous realizations, but is equal to the present observed value. A semi-martingale is defined as a process that can be decomposed as the sum of a local martingale and a càdlàg adapted finite-variation process. The following proposition from Andersen *et al.* (2003) characterizes the logarithmic asset price process.

Proposition 2.1. *Any arbitrage-free regular logarithmic price process may be uniquely decomposed as the sum of a finite variation and predictable mean component μ_t and a local martingale M_t:*

$$r_t \equiv p_t - p_0 = \mu_t + M_t = \mu_t + M_t^C + \Delta M_t \,, \tag{2.1}$$

where the local martingale component M_t consists of a continuous sample path, infinite variation local martingale component M_t^C and a compensated jump martingale ΔM_t, and the predictable mean process μ_t can be further decomposed into a continuous process (μ_t^C) and jump process ($\Delta\mu_t$). From the definition, we must have $\mu_0 \equiv M_0 \equiv M_0^C \equiv \Delta M_0 \equiv 0$. In addition, there is a jump risk associated with the predictable jump process $\Delta\mu_t$, meaning that if $\Delta\mu_t \neq 0$, then

$$P\left[sgn\left(\Delta\mu_t\right) = -sgn\left(\Delta\mu_t + \Delta M_t\right)\right] > 0 \,, \tag{2.2}$$

where $sgn(x) \equiv 1$ for $x \geq 0$ and $sgn(x) \equiv -1$ for $x < 0$.

Equation 2.2 means that if a predictable jump in price $\Delta\mu_t$ occurs (i.e. we know the time when the jump will occur and the size of the jump), there would be an arbitrage opportunity, if there was no jump in the martingale component ΔM_t at the same time. In addition, we need the martingale jump ΔM_t to be at least as high as $\Delta\mu_t$, but in the opposite direction, with strictly positive probability, in order to overturn the possible gain from the predictable jump.

The previous implies several characteristics of the return process r_t. It can be decomposed into a predictable and integrable mean (expected return) component and a local martingale innovation. It is clear from (2.1) that the return process r_t has the same main properties as the price process p_t. Although the finite variation and predictable mean component μ_t is predictable, it may be stochastic and display jumps, but the continuous component μ_t^C must follow a smooth path. Moreover, if a jump occurs in the predictable mean component, there must be a simultaneous jump present in the compensated jump martingale, ΔM_t. Thus, two kinds of jumps can occur in the return process – the predictable ones in the case of $\Delta\mu_t \neq 0$, and the unpredictable ones in the case of $\Delta\mu_t = 0$ but $\Delta M_t \neq 0$. In practice, the former type will occur if anticipated information become available in the market (such as macroeconomic news or company reports), the latter type may be caused by unexpected (macroeconomic or firm-specific) information that hit the market from time to time. We emphasize that in case of any uncertainty about the exact time when the jump will occur, the jump should not be considered to be predictable, and it should be, therefore, removed from the predictable mean process. A continuous sample path of μ_t (although it may be stochastic) would be a consequence of complete absence of anticipated jumps in the process.

2.1.1 Quadratic variation

We now focus on the behavior of the martingale component from the decomposition (2.1). Unfortunately, we are not able to observe the local martingale process M_t, since we would need continuous data in order to do so. Therefore, we use discrete variation measures that represent the variation process over a discrete time interval. Thus, the continuous decomposition in (2.1) takes the following discrete time form:

$$r_{t,h} = \mu_{t,h} + M_{t,h} = \mu_{t,h} + M_{t,h}^C + \Delta M_{t,h}\,, \tag{2.3}$$

where $\mu_{t,h} = \mu_t - \mu_{t-h}$, $M^C_{t,h} = M^C_t - M^C_{t-h}$ and $\Delta M_{t,h} = \Delta M_t - \Delta M_{t-h}$.

Definition 2.2. Let r_t be a semi-martingale process. Then we can define the unique *quadratic variation process*, $[r,r]_t$, $t \in [0,T]$, associated with r_t in the following manner:

$$[r,r]_t = r_t^2 - 2\int_0^t r_{s-}\, ds\,, \tag{2.4}$$

If the finite variation process μ_t from (2.1) is continuous, then directly from (2.4) we have that the quadratic variation of μ_t is zero. This implies that the predictable component has no impact on the quadratic variation of the return $(r_t)_{t\in[0,T]}$. Therefore, we are able to define the quadratic return variation (based on Andersen *et al.* (2003) and Barndorff-Nielsen & Shephard (2002b)) as follows.

Definition 2.3. Let r_t be a semi-martingale process. Then the quadratic variation $QV_{t,h}$ of the return process $(r_t)_{t\in[0,T]}$ over $[t-h,t]$ is

$$QV_{t,h} = [r,r]_t - [r,r]_{t-h} = [M^C,M^C]_t - [M^C,M^C]_{t-h} + \sum_{t-h<s\leq t} \Delta M_s^2 \tag{2.5}$$

$$QV_{t,h} = [r,r]_t - [r,r]_{t-h} = [M^C,M^C]_t - [M^C,M^C]_{t-h} + \sum_{t-h<s\leq t} \Delta r_s^2 \tag{2.6}$$

The realized variation of the return process is measured by the quadratic variation process which we are able to approximate using the *realized variance* (that will be defined later).

Most of the continuous-time models which try to model asset returns can be cast within the very general setting of (2.3). A framework to the study of the model-implied return variation (and also the square root of the variation called volatility), which constitutes one of the main interests of econometricians, is provided by quadratic variation. The integral representations for continuous sample path semi-martingales corresponding to (2.3) are rather abstract. However, it is frequently assumed in the theoretical asset and derivatives pricing literature that the continuous-time models have continuous sample paths and the corresponding diffusion processes are given in the form of stochastic differential equations. The previous assumption can be made using the following result, the martingale representation theorem, without loss of generality (Protter (1992)).

Proposition 2.2. *For any univariate, square-integrable, continuous sample path,*

logarithmic price process $(p_t)_{t\in[0,T]}$ *which is not locally riskless (this condition is not restrictive), we have over* $[t-h,t]$

$$r_{t,h} = \mu_{t,h} + M_{t,h} = \int_{t-h}^{t} \mu_s \, ds + \int_{t-h}^{t} \sigma_s \, dW_s \,, \tag{2.7}$$

where μ_s *is an integrable, predictable and finite-variation stochastic process,* σ_s *is a strictly positive càdlàg stochastic process satisfying*

$$P\left[\int_{t-h}^{t} \sigma_s^2 \, ds < \infty\right] = 1,$$

and W_t *is a standard Brownian motion.*

Let us give some examples of the discussed setting. We begin with a special case – the Black & Scholes (1973) option pricing model. In this example, the mean process is constant ($\mu_t = \mu$), the martingale jump component is absent ($\Delta M_t = 0$), and the continuous martingale component M_C is a standard Brownian motion. Thus we have

$$dp_t = \mu dt + \sigma dW_t \,. \tag{2.8}$$

Thus, the quadratic variation over $[t-h,t]$ takes a very simple form

$$QV_{t,h} = \int_{t-h}^{t} \sigma^2 \, ds = \sigma^2 h \,, \tag{2.9}$$

which implies that the return variation over a time interval of length h does not change in time.

We continue with the Merton (1976)'s jump diffusion model, which is as follows

$$dp_t = \left(\mu - \lambda\bar{\xi}\right) dt + \sigma dW_t + \xi_t dq_t \,, \tag{2.10}$$

where q_t denotes a Poisson process, which is uncorrelated with process W_t, and is governed by λ–constant jump intensity. ξ_t is responsible for the magnitude of the jumps, and is normally distributed with parameters $\left(\bar{\xi}, \sigma_\xi^2\right)$. This process has the following quadratic variation over $[t-h,t]$

$$QV_{t,h} = \int_{t-h}^{t} \sigma^2 \, ds + \sum_{t-h<s\leq t} J_s^2 = \sigma^2 h + \sum_{t-h<s\leq t} J_s^2 \,, \tag{2.11}$$

where $J_t = \xi_t dq_t \neq 0$ only in the presence of a jump in the process. The

quadratic variation in this case is again constant but it differs from the variation of Black-Scholes, represented by (2.8), in the jump variation $\sum_{t-h<s\leq t} J_s^2$.

The last model that we present herein is a jump-diffusion model which defines a very general class of stochastic volatility models. This class of models is used in the present thesis, as our work is based on the assumption that the price process contains jumps. The model has the following form

$$dp_t = \mu_t dt + \sigma_t dW_t + \xi_t dq_t \,, \tag{2.12}$$

where q_t denotes a Poisson process with the same attributes as in (2.10). We may characterize (2.12) as a Brownian semi-martingale process with finite jump activity. Moreover, it is a special case of (2.1). The quadratic variation of this process over $[t-h, t]$ is as follows

$$QV_{t,h} = \int_{t-h}^{t} \sigma_s^2 \, ds + \sum_{t-h<s\leq t} J_s^2 \,. \tag{2.13}$$

This quadratic variation consists of two components. We will call the first component the Integrated Variance $IV_{t,h}$

$$IV_{t,h} = \int_{t-h}^{t} \sigma_s^2 \, ds \,. \tag{2.14}$$

The second component $\sum_{t-h<s\leq t} J_s^2$ is then called the Jump Variation. The next part is dedicated to the definition of the already mentioned realized variance and its basic properties.

2.2 Measurement of the realized variance using high-frequency data

The availability of high frequency data has enabled a quite simple way of measuring the quadratic variation – the realized variance. However, the idea of using only return realizations for the measurement of return variation comes from not so recent times. Monthly realized variance estimates were, for example, computed from daily returns (by French *et al.* (1987)), which might have been considered as high frequency data then. We now continue with the definition of the realized variance (i.e. the estimator of quadratic variation).

Definition 2.4. Let r_t be a logarithmic return process. The realized variance $\widehat{RV}_{t,h}$ over $[t-h,t]$ is then defined as

$$\widehat{RV}_{t,h} = \sum_{i=1}^{n} r^2_{t-h+\left(\frac{i}{n}\right)h}, \tag{2.15}$$

where n denotes the number of observations from the time interval $[t-h,t]$.

The realized variance is in fact just the second sample moment of the return process over a fixed interval of length h, scaled by the number of observations n in order to provide a variance measure calibrated to the measurement interval of length h. The convergence of the realized variance measure $\widehat{RV}_{t,h}$, described by (2.15), to the return quadratic variation $QV_{t,h}$ described by (2.5) is ensured by the semi-martingale theory. Details and more theoretical properties of this important result can be found in Andersen & Bollerslev (1998), Andersen *et al.* (2001; 2003) and Barndorff-Nielsen & Shephard (2001; 2002a;b). We now state these two important results – that the realized variance is unbiased and consistent estimator of variance of the return process – mathematically.

Proposition 2.3. *Let $(r_t)_{t\in[0,T]}$ be a square-integrable return process and $\mu_t \equiv 0$. Then we have*

$$E\left[RV_{t,h}|\mathcal{F}_t\right] = E\left[M^2_{t,h}|\mathcal{F}_t\right] = E\left[\widehat{RV}_{t,h}|\mathcal{F}_t\right] \tag{2.16}$$

for all $n \geq 1$ and $h > 0$. The term $RV_{t,h}$ denotes the ex ante variance of the return process.

Proposition 2.4. *The realized variance converges uniformly in probability to the variance of the return process $(r_t)_{t\in[0,T]}$,*

$$\text{plim}_{n\to\infty}\, \widehat{RV}_{t,h} = RV_{t,h}\,, \;\; 0 \leq h \leq t \leq T\,, \tag{2.17}$$

i.e. $\widehat{RV}_{t,h}$ provides a consistent nonparametric measure of the variance.

In fact, (2.16) means that the *ex-post* realized variance $\left(\widehat{RV}_{t,h}\right)$ is an unbiased estimator of the *ex-ante* expected variance $RV_{t,h}$, while (2.17) tells us that if we let the sampling frequency go to infinity, the realized variance will be a consistent estimator of the variance over any time interval $[t-h,t]$, $h > 0$.

To avoid any confusion that may arise from the use of variance $RV_{t,h}$, quadratic variation $QV_{t,h}$ and realized variance $\widehat{RV}_{t,h}$, let us clarify the re-

lations between these concepts. $QV_{t,h}$ provides a measure of $RV_{t,h}$, but since $QV_{t,h}$ is obviously unobservable, the actual measurement of $RV_{t,h}$ is carried out by the estimator of $QV_{t,h}$ – the realized variance $\widehat{RV}_{t,h}$.

Not only does the realized variance measure the variability of the return process, but it also helps with the specification of its distribution, which is needed for the empirical modeling and forecasting of the process. However, the additional assumptions from this part are not necessary for the volatility modeling.

Proposition 2.5. *The following holds for any square-integrable arbitrage-free price process that has a sample path satisfying* (2.7) *with innovation process* W_t *independent of the conditional mean process* μ_t *and volatility process* σ_t *over* $[t-h, t]$

$$r_{t,h}|\sigma\left\{\mu_{t,h}, RV_{t,h}\right\} \sim N\left(\mu_{t,h}, RV_{t,h}\right), \tag{2.18}$$

where $\sigma\left\{\mu_{t,h}, RV_{t,h}\right\}$ *denotes the* σ*-field generated by* $(\mu_{t,h}, RV_{t,h})$.

Expression (2.18) means that the distribution of the returns over the time interval $[t-h, t]$, conditional on the mean return and variance, will be Gaussian. We bring to the reader's attention that the characterization of the return distribution in (2.18) is conditional on the *ex post* realizations of μ_t and $RV_{t,h}$, which are typically unobservable. This could imply that (2.18) is practically useless. Nevertheless, we are able to approximate the realized quadratic variation, thus also the conditional variance, from the observed high-frequency returns (as shown in e.g. Andersen *et al.* (2003)). Moreover, we can ignore the conditional mean variation for daily or weekly data, since it is negligible relative to the volatility of returns. Applying this on equation (2.18) we have that the distribution of returns (daily or weekly) is determined by a normal mixture, which is governed by the realized quadratic variation of the returns (daily or weekly).

Strictly speaking, we can only apply the normal mixture distribution if the price process is continuous and both the mean process and volatility process are independent of the innovation process. This independence implies that the returns are conditionally symmetrically distributed, which raises two major concerns. Firstly, there is evidence (e.g. in Andersen *et al.* (2002), Bates (2000), Pan (2002) and Eraker *et al.* (2003)) suggesting that discrete jumps might be present in asset prices, which would make the sample paths discontinuous. On the other hand, the same studies also suggest that jumps occur rarely and they

have difficulties coming to a consensus regarding the distribution of the size of jumps. Secondly, some asset classes might have correlated their return and volatility innovations, which may be the reason of existing evidence of leverage effects. However, it is likely that these contemporaneous correlation effects are quantitatively of little importance at the daily or weekly horizon.

Naturally, we would expect that consistency of $\widehat{RV}_{t,h}$ and normal distribution of the returns from (2.5) imply that we are able to measure the variance of the return quite simply. However, there are two issues complicating the practical use of the very convenient convergence results. We need continuous sample path of the returns for the convergence of the RV estimate to RV, since the realized variance is a consistent estimator with increasing sampling frequency, $n \to \infty$. Nevertheless, we are only able to observe discrete prices in practice, which means that discretization error is inevitable. On the other hand, return observations in practice are contaminated with market microstructure effects such as price discreteness, bid-ask spread and bid-ask bounce. This implies that we should not employ sampling of returns with very high frequency, no matter how much data we have at hand, if we want to avoid large bias from the market microstructure (this will be discussed more further in the text). There have been extensive studies that were trying to find the optimal noise-to-signal ratio. As a result optimal sampling schemes were constructed, which range from 5 to 30 minutes. More information and literature on this matter can be found for example in Zhang *et al.* (2005), Hansen & Lunde (2006), Bandi & Russell (2006a;b), Andersen & Benzoni (2007), McAleer & Medeiros (2008) and Barndorff-Nielsen *et al.* (2008).

The previously mentioned recommendation on data sampling brings another problem – discarding of a very large amount of information. For example, if we had data recorded every second, but in order to avoid microstructure noise we would use 5-minute sampling frequency. This would lead to using only one record of the data from every 300 available data points. In case of even more liquid stocks with higher sampling frequency, we would throw away even larger amounts of available data. However, it is quite hard to believe that getting rid of such amounts of data is the solution to the problem of microstructure noise. Therefore, we also mention other possible solutions, proposed by Zhang *et al.* (2005) and by Barndorff-Nielsen *et al.* (2008), which will be briefly described in the next section.

For the sake of completeness, we herein impose the definition of *realized volatility* which is just the square root of the realized variance.

Definition 2.5. The realized volatility $\widehat{RVO}_{t,h}$ over $[t-h,t]$ is defined in the following manner:

$$\widehat{RVO}_{t,h} = \sqrt{\widehat{RV}_{t,h}}, \tag{2.19}$$

where $\widehat{RV}_{t,h}$ is the realized variance defined in (2.15).

The two concepts are sometimes exchanged by mistake in the literature, therefore, in order to avoid any confusion we herein establish, that by *realized variance* we mean the equation (2.15), while when referring to the *realized volatility* we will have in mind the square root of realized variance (2.19).

2.3 The effects of microstructure noise

This section is dedicated to solutions to the problem with microstrucure noise other than throwing away large amounts of data. However, we only describe these methods briefly, as they exceed the scope of this work. Let us begin with the one proposed by Zhang *et al.* (2005).

If we consider an observed logarithmic price process, we can say that the data consist of the so-called *true* log-price process, but it also contains noise. Thus when calculating the realized variance of the observed logarithmic price process, we obtain a result contaminated with this noise, i.e. the estimated variance will be biased. Moreover, this bias grows as we increase the number of observations (in order to use all the available data and obtain a consistent estimate as $n \to \infty$).

Zhang *et al.* (2005) propose using the *Two-Scale Realized Variance* estimator (TSRV henceforth), which can be computed in the following way. Let us return to the example from the previous section, i.e. we have data set consisting of prices recorded once every second. First, we need to create equally sized subsamples, where the first subsample would start at the first observation and continue with observations taken for example every five minutes (other frequencies are, of course, also possible), the second subsample would start at the second observation and again continue with observations taken every five minutes, etc. This way, we would obtain 300 equally sized subsamples[3] (the first would consist of observations {1, 301, 601,...}, the second of observations {2, 302, 602,...}, etc.). The next step would be calculating the realized variance

[3]The number of subsamples depends directly on our choice of the sampling frequency. In the example, we chose a five-minute frequnecy, thus we have 300 subsamples, but using ten-minute frequnecy would result in 600 subsamples, etc.

from each subsample, thus, we would have 300 estimates of the realized variance. Finally, we simply calculate the average of the realized variances, which would be the so-called average estimator.

The average estimator is better than the simple estimate using all the observations, but it is still biased. Zhang *et al.* (2005) solve this by estimating the bias caused by noise. They propose, that a certain fraction of the realized variance, calculated from the original whole dataset, is a consistent estimate of the bias from noise. Therefore, TSRV can be estimated as the difference between the average estimator and certain fraction of the realized variance of the original dataset. Also, it is the bias-adjusted estimator of the *true* logarithmic price process. Further generalizations and details, which we are not going to talk about, can be found for example in Zhang (2006).

Barndorff-Nielsen *et al.* (2008) propose another solution to the problem with microstructure noise – an estimator called the realized kernel estimator. This estimator consists of the sum of two main parts. The first one is simply the estimate of the realized variance defined by (2.15). The second part depends on the realized autocovariance of the intraday return process and on the kernel function, which is of our choice. Nevertheless, we shall not go into further details in this work.

The previously mentioned estimators would enable us to estimate the realized variation consistently even from noisy data. However, we are not going to employ these estimators in the empirical part of this work. In the next chapter we proceed to the final steps – identification of jumps in the price process, estimation of realized variation measures and modeling of realized variance.

Chapter 3

Estimation of jumps and methodology of forecasting

One of the key interests of this work is detection and separation of the jumps in the price, and hence, the return process. We already showed that the quadratic variation can be decomposed into two components:

$$QV_{t,h} = \underbrace{\int_{t-h}^{t} \sigma_s^2 \, ds}_{IV_{t,h}} + \underbrace{\sum_{t-h \leq s \leq t} J_s^2}_{\text{Jump Variation}} , \tag{3.1}$$

where the integrated variation is latent and we need to use some approximation. An elegant solution of jump-detection in high-frequency data is proposed by Barndorff-Nielsen & Shephard (2004; 2006). The idea is quite simple – using two estimators of the quadratic variation. One estimator contains both the jump variation and the integrated variation, the other one only contains the integrated variation component. We are then able to obtain the jumps as the difference between the two estimators. One of the estimators has already been defined (the realized variance), the other, containing only the IV component, will be introduced in the following section.

3.1 Bipower variation and the jump detection test statistic

We begin with the formal definition of the bipower variation (as originally proposed by Barndorff-Nielsen & Shephard (2004)), which is the realized measure

of $QV_{t,h}$ containing only the integrated variation.

Definition 3.1. The bipower variation over $[t-h,t]$, for $0 \leq h \leq t \leq T$ is

$$\widehat{BV}^{1}_{t,h} = \mu_1^{-2} \sum_{i=2}^{n} \left| r_{t-h+\left(\frac{i-1}{n}\right)h} \right| \cdot \left| r_{t-h+\left(\frac{i}{n}\right)h} \right| , \tag{3.2}$$

where $\mu_a = E\left[|Z|^a\right] = 2^{r/2} \frac{\Gamma\left(\frac{1}{2}(r+1)\right)}{\Gamma\left(\frac{1}{2}\right)}$, for $Z \sim N(0,1)$, $a \geq 0$ and $\Gamma(\cdot)$ denoting the Gamma function (in this case $\mu_1 = \sqrt{2/\pi}$).

Barndorff-Nielsen & Shephard (2004) also show that $\widehat{BV}^{1}_{t,h} \to \int_{t-h}^{t} \sigma_s^2 \, ds$, which is a crucial result for us. However, in order to render the estimator robust to certain types of microstructure noise, we use the Andersen *et al.* (2010) adjustment of the original estimator in this work.

Definition 3.2. The adjusted version of the bipower variation over $[t-h,t]$, for $0 \leq h \leq t \leq T$ is

$$\widehat{BV}_{t,h} = \mu_1^{-2} \frac{n}{n-2} \sum_{i=3}^{n} \left| r_{t-h+\left(\frac{i-2}{n}\right)h} \right| \cdot \left| r_{t-h+\left(\frac{i}{n}\right)h} \right| , \tag{3.3}$$

where, as in the previous definition, $\mu_1 = \sqrt{2/\pi}$.

Naturally, $\widehat{BV}_{t,h}$ also converges to the IV component of QV. Thus, we have $\widehat{BV}_{t,h}$ – a consistent estimator of the integrated variation, and $\widehat{RV}_{t,h}$ – a consistent estimator of the sum of integrated variation and jump variation (i.e. the quadratic variation). Finally, we are able to estimate the consistent estimator of the jump variation as the difference between the realized variance and the realized bipower variation, since the following is a consequence of the convergence results of BV and RV

$$\text{plim}_{n\to\infty} \left(\widehat{RV}_{t,h} - \widehat{BV}_{t,h} \right) = \sum_{l=1}^{N_t} J^2_{t,h,l} \, , \tag{3.4}$$

where N_t denotes the number of non-zero jumps over $[t-h,t]$. However, in practice this procedure would give a non-zero jump for every day, while we only expect jumps to occur rather rarely. Therefore, we need to distinguish between significant jumps and those, that are of no importance. There are various modifications of the test statistic used for this purpose (see e.g. Barndorff-Nielsen & Shephard (2004), Christensen & Nielsen (2005) or Bollerslev *et al.*

(2008)). We use the jump detection test statistic introduced by Andersen *et al.* (2010), based on which we will be able to separate the significant jumps from the rest.

Definition 3.3. We define the jump detection test statistic $Z_{t,h}$ by

$$Z_{t,h} = \frac{\frac{\widehat{RV}_{t,h} - \widehat{BV}_{t,h}}{\widehat{RV}_{t,h}}}{\sqrt{\left(\left(\frac{\pi}{2}\right)^2 + \pi - 5\right) \frac{1}{n} \max\left(1, \frac{\widehat{TQ}_{t,h}}{\left(\widehat{BV}_{t,h}\right)^2}\right)}} \tag{3.5}$$

$Z_{t,h}$ is standard normally distributed under the null hypothesis of no within-day jumps. The realized tripower quarticity $\widehat{TQ}_{t,h}$ in (3.5) is defined by

$$\widehat{TQ}_{t,h} = n\mu_{4/3}^{-3} \left(\frac{n}{n-4}\right) \sum_{i=5}^{n} \left|r_{t-h+\left(\frac{i-4}{n}\right)h}\right|^{4/3} \cdot \left|r_{t-h+\left(\frac{i-2}{n}\right)h}\right|^{4/3} \cdot \left|r_{t-h+\left(\frac{i}{n}\right)h}\right|^{4/3}$$

, where $\mu_{4/3} = 2^{2/3}\frac{\Gamma(7/6)}{\Gamma(1/2)}$ and $\Gamma(\cdot)$ denotes the Gamma function.

Finally, we are able to identify the realized measure of the jump contribution, as well as the realized measure of the integrated variation contribution to the quadratic variation of the log-price process. We now continue with the definitions of both realized measures as introduced in Andersen *et al.* (2010).

Definition 3.4. The realized measure of the jump contribution $J_{t,h}$ to the quadratic variation of the logarithmic price process is defined by

$$J_{t,h} = I\left(Z_{t,h} > \Phi_\alpha\right) \cdot \left(\widehat{RV}_{t,h} - \widehat{BV}_{t,h}\right), \tag{3.6}$$

where $I(\cdot)$ is the indicator function and Φ_α represents the α-quantile of the standard normal distribution function.

Definition 3.5. The realized measure of the integrated variance $C_{t,h}$ is defined by

$$C_{t,h} = I\left(Z_{t,h} \leq \Phi_\alpha\right) \cdot \widehat{RV}_{t,h} + I\left(Z_{t,h} > \Phi_\alpha\right) \cdot \widehat{BV}_{t,h}, \tag{3.7}$$

where $I(\cdot)$ and Φ_α denote the same as in the previous definition.

It is worth noting that these realized measures are defined in a manner which ensures that they add up to the realized variance $\widehat{RV}_{t,h}$. Another important detail lies in the actual construction of the components of the quadratic variation which depend on our choice of the α-quantile of the normal distribu-

tion. The results presented in this work were obtained using a 99% quantile, since choosing lower quantiles results in only slightly different estimates. For simplicity reasons, we will use the following notation in the remainder of this work: RV_t will stand for realized variance, RVO_t for realized volatility and C_t and J_t will denote the continuous part and the jump process of the quadratic variation (realized variance), respectively. Now, let us take a look at the models that will be used for modeling and forecasting of realized variance.

3.2 HAR models

The primary reason for studying volatility and related concepts is the desire to predict its evolution. As the main aim of this work is no different, we need a tool for the forecasting. Therefore, in this section, we are going to describe the development of a realized volatility model first proposed by Corsi (2004), called the Heterogeneous Autoregressive model. The motivation for the development of such model is the need of a simple additive model that is able to catch the volatility of financial data, as opposed to complicated and hard-to-estimate multiplicative models with no clear economic interpretation.

The main idea behind the HAR model is based on the Heterogeneous Market Hypothesis of Müller *et al.* (1997), which states that there is distinct heterogeneity in the behavior of traders. This hypothesis offers a possible explanation of the observed positive correlation between volatility and the number of traders in a market. It says that the more traders participate in a homogeneous market, the quicker the price converges to its real market value which would result in negatily correlated volatility and number of traders. Heterogeneous agents, on the other hand, have different preferences and make different decisions in different situations, thus they create volatility in the market.

Corsi (2004) distinguishes three main types of agents from the time horizon point of view (i.e. based on the frequency of activity). The first type are agents with high intraday frequency of trading – dealers, market makers and speculators. The second type consists of agents who make decisions on a weekly basis, such as portfolio managers. Central banks, funds or commercial organizations with trading frequency on a monthly basis (or higher) constitute the third type of agents. Each of these types of agents is responsible for a different kind of volatility in the market. We call these kinds of volatilities partial and they have a structure similar to AR(1) processes (as the next observation depends on the present one). Moreover, these partial volatilities influence each other in such

way, that each partial volatility depends also on the expected partial volatility of the next longer horizon volatility, but not vice versa (i.e. the weekly partial volatility is affected by the monthly partial volatility, but the monthly partial volatility is not affected by the weekly partial volatility, etc.). These interesting findings are captured by the HAR-RV model which we are going to describe.

3.2.1 The HAR-RV model

The simple HAR-RV model by Corsi (2004) assumes that the price process contains no jumps, i.e. it considers the following stochastic volatility process, already mentioned above

$$dp_t = \mu_t dt + \sigma_t dW_t \,, \tag{3.8}$$

where p_t is a logarithmic price process, μ_t is a continuous finite variation process and σ_t is a stochastic process independent of the standard Brownian motion W_t.

If we denote the daily, weekly and monthly partial volatilities by $\tilde{\sigma}_t^{(d)}$, $\tilde{\sigma}_t^{(w)}$ and $\tilde{\sigma}_t^{(m)}$, we can express the above mentioned relations as

$$\tilde{\sigma}_{t+1m}^{(m)} = \tilde{\alpha}^{(m)} + \tilde{\beta}^{(m)} RVO_t^{(m)} + \tilde{\epsilon}_{t+1m}^{(m)} \,, \tag{3.9}$$

$$\tilde{\sigma}_{t+1w}^{(w)} = \tilde{\alpha}^{(w)} + \tilde{\beta}^{(w)} RVO_t^{(w)} + \tilde{\gamma}^{(w)} E_t \left[\tilde{\sigma}_{t+1m}^{(m)}\right] + \tilde{\epsilon}_{t+1w}^{(w)} \,, \tag{3.10}$$

$$\tilde{\sigma}_{t+1d}^{(d)} = \tilde{\alpha}^{(d)} + \tilde{\beta}^{(d)} RVO_t^{(d)} + \tilde{\gamma}^{(d)} E_t \left[\tilde{\sigma}_{t+1w}^{(w)}\right] + \tilde{\epsilon}_{t+1d}^{(d)} \,, \tag{3.11}$$

where $(1d)$, $(1w)$ and $(1m)$ denote the time horizons of one day, one week and one month, and $RVO_t^{(m)}$, $RVO_t^{(w)}$ and $RVO_t^{(d)}$ represent the observed monthly, weekly and daily realized volatility, respectively. The innovation terms $\tilde{\epsilon}_{t+1m}^{(m)}$, $\tilde{\epsilon}_{t+1w}^{(w)}$ and $\tilde{\epsilon}_{t+1d}^{(d)}$ are independent and serially uncorrelated with zero mean and truncated left tail in order to ensure only positive values of partial volatilities. Simple substitutions of (3.9) into (3.10) and immediately (3.10) into (3.11), together with the assumption that the daily integrated volatility $\sigma_t^{(d)}$ is determined by the highest frequency partial volatility (i.e. $\sigma_t^{(d)} = \tilde{\sigma}_t^{(d)}$), yield the following model

$$\sigma_{t+1d}^{(d)} = \alpha + \beta^{(d)} RVO_t^{(d)} + \beta^{(w)} RVO_t^{(w)} + \beta^{(m)} RVO_t^{(m)} + \tilde{\epsilon}_{t+1d}^{(d)} \tag{3.12}$$

Now we have three observable variables on the right hand side of (3.12) but

we have a latent variable on the left hand side. Nevertheles, knowing that the realized volatility is an estimator of the latent volatility, we have

$$\sigma_{t+1d}^{(d)} = RVO_{t+1d}^{(d)} + \epsilon_{t+1d}^{(d)} \tag{3.13}$$

Here the problem of microstructure noise becomes apparent as we need the realized volatility to be a consistent and unbiased estimator in order to assume zero mean of the error term. However, by substituting (3.13) into (3.12) we finally obtain the desired model with no latent variables

$$RVO_{t+1d}^{(d)} = \alpha + \beta^{(d)} RVO_t^{(d)} + \beta^{(w)} RVO_t^{(w)} + \beta^{(m)} RVO_t^{(m)} + \epsilon_{t+1d} \tag{3.14}$$

We need to clarify that the weekly and monthly realized volatilities are respectively computed as the average of the last week (5 days) daily volatilities and as the average of the last month (22 days) daily volatilities. That is

$$RVO_t^{(w)} = \frac{1}{5}\left(RVO_t^{(d)} + RVO_{t-1}^{(d)} + \ldots + RVO_{t-4}^{(d)}\right),$$

$$RVO_t^{(m)} = \frac{1}{22}\left(RVO_t^{(d)} + RVO_{t-1}^{(d)} + \ldots + RVO_{t-21}^{(d)}\right)$$

In the following text, we will introduce different HAR models, therefore we need to specify the notation of the parameters according to the variables – the subscript of beta parameters will indicate which variable they refer to (the usefulness of this notation will be apparent in Chapter 5). For simplicity reasons, the notation of the variables has to be also slightly changed – as we always consider the time scale to be in days, we will no longer specify this in the subscript of the variables. Thus the previously introduced HAR-RV model assumes the form

$$\begin{aligned} RVO_{t+1}^{(d)} = & \\ & \alpha + \beta_{RVO}^{(d)} RVO_t^{(d)} + \beta_{RVO}^{(w)} RVO_t^{(w)} + \beta_{RVO}^{(m)} RVO_t^{(m)} + \epsilon_{t+1} \end{aligned} \tag{3.15}$$

Similarly, we can write the HAR-RV model using realized variance instead of realized volatility

$$\begin{aligned} RV_{t+1}^{(d)} = & \\ & \alpha + \beta_{RV}^{(d)} RV_t^{(d)} + \beta_{RV}^{(w)} RV_t^{(w)} + \beta_{RV}^{(m)} RV_t^{(m)} + \epsilon_{t+1}\,, \end{aligned} \tag{3.16}$$

where the weekly ($RV_t^{(w)}$)and monthly ($RV_t^{(m)}$) realized variances are defined as

$$RV_t^{(w)} = \frac{1}{5}\left(RV_t^{(d)} + RV_{t-1}^{(d)} + \ldots + RV_{t-4}^{(d)}\right) ,$$

$$RV_t^{(m)} = \frac{1}{22}\left(RV_t^{(d)} + RV_{t-1}^{(d)} + \ldots + RV_{t-21}^{(d)}\right)$$

For the sake of completeness, we need to mention yet another modification of the HAR-RV model – the logarithmic specification, which is constructed using logarithms of the variables from the previous model

$$\ln\left(RV_{t+1}^{(d)}\right) = \alpha + \beta_{RV}^{(d)} \ln\left(RV_t^{(d)}\right) + \beta_{RV}^{(w)} \ln\left(RV_t^{(w)}\right) + \beta_{RV}^{(m)} \ln\left(RV_t^{(m)}\right) + \epsilon_{t+1} , \tag{3.17}$$

The logarithmic transform of the model has one very important property. It obviously ensures that the dependent variable assumes only positive values, which is very convenient as the realized variance is a positive random variable. However, the logarithmic model lacks economic interpretability compared to the non-logarithmic specification, which will be more distinct when we employ further models. Consequently, we use both specifications in the empirical part of this work.

Residuals of the HAR-RV models were (e.g. by Corsi (2004)) first assumed to be normal, independent, identically distributed variables. Nevertheless, various empirical studies show that this assumption is violated, and the residuals are heteroskedastic. For example Andersen *et al.* (2010) propose to augment the model by a GARCH error structure for the so-called volatility-of-volatility, with errors of the GARCH model assumed to be conditionally t-distributed, allowing for fat-tails. We present the general form of the GARCH(p,q) error structure.

$$\epsilon_{t+1} = u_{t+1}\sigma_{t+1} , \; u_t \text{ iid } N(0,1) \tag{3.18}$$

$$\sigma_{t+1}^2 = \omega + \sum_{i=1}^{p} \alpha_i u_{t+1-i}^2 + \sum_{j=1}^{q} \gamma_j \sigma_{t+1-j}^2 \tag{3.19}$$

However, heteroskedasticity is not a serious issue as our main goal is to compare the performance of different models (under heteroskedasticity the parameters are still unbiased, only the standard errors are incorrect and the prediction inefficient). Therefore, we are not going to work with GARCH augmented models. We now continue with the development of the HAR models proposed by Andersen *et al.* (2007).

3.2.2 HAR-RV-J models

The previously introduced models assume that the price process is continuous. However, as already mentioned above, actual price processes consist of a continuous and discontinuous part and thus the volatility of the price process also consists of volatility from the continuous part and volatility from the jump part. Thus, the rest of the models presented in this chapter will assume a price process containig jumps, i.e. (3.8) transforms to

$$dp_t = \mu_t dt + \sigma_t dW_t + \xi_t dq_t \,, \tag{3.20}$$

where $\xi_t dq_t$ is the term responsible for jumps. We also showed that, using the realized variation measures, we are able to separate the jump component and the continuous part of the realized variance – the estimator of volatility. Knowing this, we simply add the jump component as an extra explanatory variable to the HAR-RV model and we have the following HAR-RV-J model, proposed by Andersen *et al.* (2007)

$$\begin{aligned} RV_{t+1}^{(d)} &= \alpha \\ &+\beta_{RV}^{(d)} RV_t^{(d)} + \beta_{RV}^{(w)} RV_t^{(w)} + \beta_{RV}^{(m)} RV_t^{(m)} + \beta_J^{(d)} J_t^{(d)} + \epsilon_{t+1} \end{aligned} \tag{3.21}$$

Similarly, we can add the corresponding form of jump variable to the modified versions of the HAR-RV model. A smaller complication occurs in the logarithmic version of the model as there are days with no jumps in the price process, i.e. the jump component equals zero. A very simple solution to this problem results in the following form of the logarithmic HAR-RV-J model

$$\begin{aligned} \ln\left(RV_{t+1}^{(d)}\right) &= \alpha + \beta_J^{(d)} \ln\left(1 + J_t^{(d)}\right) \\ &+\beta_{RV}^{(d)} \ln\left(RV_t^{(d)}\right) + \beta_{RV}^{(w)} \ln\left(RV_t^{(w)}\right) + \beta_{RV}^{(m)} \ln\left(RV_t^{(m)}\right) + \epsilon_{t+1} \end{aligned} \tag{3.22}$$

However, the above mentioned problem with interpretation of the parameters becomes more apparent. Adding 1 to the jump component artificially forces the whole term to be always non-negative, while we still allow the other terms of the model to become both positive and negative. Nevertheless, we continue with the ultimate class of HAR models used in this thesis.

3.2.3 HAR-RV-CJ models

Andersen *et al.* (2007) take the extension of the model one step further. They propose a HAR-RV-CJ model which is based on the explicit decomposition of the realized variance into the continuous part and jump component. Put another way, the explanatory variables (daily, weekly and monthly volatilities) of the model are substituted by daily, weekly and monthly continuous and jump components. The weekly and monthly components are computed analogically to the weekly and monthly realized variance.

$$J_t^{(w)} = \frac{1}{5}\left(J_t^{(d)} + J_{t-1}^{(d)} + \ldots + J_{t-4}^{(d)}\right), \;\; C_t^{(w)} = \frac{1}{5}\left(C_t^{(d)} + C_{t-1}^{(d)} + \ldots + C_{t-4}^{(d)}\right),$$

$$J_t^{(m)} = \frac{1}{22}\left(J_t^{(d)} + J_{t-1}^{(d)} + \ldots + J_{t-21}^{(d)}\right), \;\; C_t^{(m)} = \frac{1}{22}\left(C_t^{(d)} + C_{t-1}^{(d)} + \ldots + C_{t-21}^{(d)}\right)$$

The HAR-RV-CJ model then assumes the form

$$\begin{aligned} RV_{t+1}^{(d)} = \alpha + \beta_C^{(d)} C_t^{(d)} + \beta_C^{(w)} C_t^{(w)} + \beta_C^{(m)} C_t^{(m)} \\ + \beta_J^{(d)} J_t^{(d)} + \beta_J^{(w)} J_t^{(w)} + \beta_J^{(m)} J_t^{(m)} + \epsilon_{t+1} \end{aligned} \tag{3.23}$$

This model allows us to observe the contribution of each component of the partial realized volatilities (daily, weekly and monthly) to the daily realized volatility. Naturally, we can employ the modified versions of the model – the realized volatility HAR-RV-CJ model,

$$\begin{aligned} RVO_{t+1}^{(d)} = \alpha + \beta_C^{(d)} \sqrt{C_t^{(d)}} + \beta_C^{(w)} \sqrt{C_t^{(w)}} + \beta_C^{(m)} \sqrt{C_t^{(m)}} \\ + \beta_J^{(d)} \sqrt{J_t^{(d)}} + \beta_J^{(w)} \sqrt{J_t^{(w)}} + \beta_J^{(m)} \sqrt{J_t^{(m)}} + \epsilon_{t+1}\,, \end{aligned} \tag{3.24}$$

and the logarithmic realized variance HAR-RV-CJ model,

$$\begin{aligned} \ln\left(RV_{t+1}^{(d)}\right) = \alpha + \beta_C^{(d)} \ln\left(C_t^{(d)}\right) + \beta_C^{(w)} \ln\left(C_t^{(w)}\right) + \beta_C^{(m)} \ln\left(C_t^{(m)}\right) \\ + \beta_J^{(d)} \ln\left(1 + J_t^{(d)}\right) + \beta_J^{(w)} \ln\left(1 + J_t^{(w)}\right) + \beta_J^{(m)} \ln\left(1 + J_t^{(m)}\right) + \epsilon_{t+1} \end{aligned} \tag{3.25}$$

Even though the HAR-RV-CJ model offers a significant improvement in the forecasting of volatility, there is still a possibility of achieving better forecasts, for example, by estimating each component of the realized volatility separately. One method of such estimation is proposed by Andersen *et al.* (2010). The continuous part is estimated by a quite simple logarithmic HAR-C model which

takes the form

$$\ln\left(C_{t+1}^{(d)}\right) = \alpha + \beta_C^{(d)} \ln\left(C_t^{(d)}\right) + \beta_C^{(w)} \ln\left(C_t^{(w)}\right) + \beta_C^{(m)} \ln\left(C_t^{(m)}\right) \qquad (3.26)$$
$$+\beta_J^{(d)} \ln\left(1 + J_t^{(d)}\right) + \beta_J^{(w)} \ln\left(1 + J_t^{(w)}\right) + \beta_J^{(m)} \ln\left(1 + J_t^{(m)}\right) + \epsilon_{t+1}$$

Modeling of the jump component, on the other hand, is carried out in a more complicated manner. The occurence of the jumps is modeled by the Autoregressive Conditional Hazard (ACH) model, while the size of the jump is modeled by the HAR-J model. Nevertheless, we are not going to talk about these models as we are not using them in this thesis. Instead, we introduce the basic tools for forecast evaluation which will be used in Chapter 5.

3.3 Evaluation of forecasts

As forecasting of economic indicatiors (such as inflation, GDP, or volatility, etc.) is the ultimate purpose of econometrics, there are many ways how to measure the accuracy of predictions. However, we only use three methods for the evaluation of the performed out-of-sample forecasts – Mincer-Zarnowitz regressions, Mean Square Error and Theil's U. Let us begin with the first one. Mincer & Zarnowitz (1969) proposed an interesting approach to assessing the performance of estimators. The basic idea is regressing the observed values of a variable on its fitted values obtained from the model (and a constatnt). If the model is accurate, the predicted values should be unbiased, i.e. we would obtain an insignificant intercept equal to zero, a significant slope coefficient equal to one and R^2 of the regression would be high. We, for example, are going to compare the performance of HAR-RV and HAR-RV-CJ models using the following Mincer-Zarnowitz style regression

$$RV_{t+1}^{(d)} = \alpha + \beta \widehat{RV}_{t+1}^{(d)} + \epsilon_{t+1}\,, \qquad (3.27)$$

where $RV_t^{(d)}$ stands for the observed daily realized variance at time t and $\widehat{RV}_t^{(d)}$ denotes its estimated value (prediction) from time $t-1$ using parameters obtained from the given model. $\alpha = 0$, $\beta = 1$ and R^2 close to 1 would mean that the model gives very accurate predictions. The regression for the logarithmic specification of the model is constructed analogically.

There are also two statistics that we use to measure the forecasting ability of the models – the Mean Square Error (MSE) and Theil's U proposed by Theil

(1966). The MSE is a common measure defined as

$$MSE = T^{-1} \sum_{t=1}^{T} \epsilon_t^2 \,, \tag{3.28}$$

where ϵ_t is the error at time t and T is the number of observations (and associated forecasts). Theil's U is defined as the squared root of the following

$$U^2 = \sum_{t=1}^{T-1} \left(\frac{f_{t+1} - y_{t+1}}{y_t} \right)^2 \cdot \left[\sum_{t=1}^{T-1} \left(\frac{y_{t+1} - y_t}{y_t} \right)^2 \right]^{-1} , \tag{3.29}$$

where y_t is the value of the observation at time t, f_t is the value of the forecast at time t and T is the number of observations (and associated forecasts). The interpretation of Theil's U is quite simple – it is a ratio of the proposed model's Root Mean Square Error (RMSE) to the RMSE of the so-called naive model ($f_{t+1} = y_t$ for all t). Values of Theil's U lower than 1 mean that the proposed model is better than the naive, while values greater than 1 indicate otherwise.

The Mincer-Zarnowitz regressions will be used only in Section 5.1 for the comparison of HAR-RV and HAR-RV-CJ models. MSE and Theil's U, on the other hand, are going to help us assess the optimal length of the pre-forecast period for the HAR-RV-CJ model forecasting in Section 5.3. Let us continue with the empirical part of this work starting with description of our datasets.

Chapter 4

Description of the data

Here we turn our attention to the description of data as we find it necessary for the reader to get a better picture of the datasets in order to follow and understand the empirical part of this thesis. As already mentioned, high frequency data are used in this work, which means we need to be very careful. We have at hand three different types of financial datasets – S&P 500 Futures index, Euro FX and Light Crude NYMEX, all of them obtained from Tick Data. What makes this work special is the large time span of available observations, which is quite unusual. The observations for S&P 500 Futures index and Light Crude NYMEX start on January 2, 1987, in case of Euro FX it is January 4, 1999, while all datasets end on December 30, 2011. Let us begin with some general description of high frequency data and the process of 'cleaning' a raw dataset until we obtain data that can be used for estimations.

4.1 High frequency data

Econometrics of high frequency data is still a rather fresh area of financial econometrics, which emerged in the last decade. Naturally, technological progress plays a key role in the availability and use of such data. The need to use high frequency data comes from the importance of the decisions made by traders who need as much information as possible. High frequency data are used mainly for intraday trading, in the determination of optimal orders or for estimations of high frequency volatility, as we do.

This work utilizes five-minute data, but different kinds of analyses require different recording-frequencies, optimal for their research (e.g. 1 minute, 10 minutes, 30 minutes, etc.). The datasets with these specific time-frequencies

must be somehow 'created' from the original raw datasets, which is not a trivial task. There are several problems (which arise directly from the nature of tick data) one needs to deal with when working with raw high frequency data and trying to filter data with the desired time-frequency. We only briefly go through some of these difficulties.

One of the problems is the length of the dataset – the number of ticks differs dramatically for different securities. Some other issues include bid-ask bounce, the possibility of errors in the recorded data (such as missing decimal parts, decimal errors, etc.), seasonal patterns in the intraday tick frequencies and perhaps the most crucial fact is that there is no exact definition of 'bad' and 'good' data points. When filtering the original dataset, we need to take into consideration that the filter has to be able to adapt to different tick frequencies of different datasets and also to different time-frequencies of the desired 'cleaned' datasets. As we can see, a rather complex filter (or even a set of filters) is necessary in order to obtain datasets cleaned from 'bad' observations, but at the same time datasets that still carry all the information relevant to different traders.

However, we have already at hand datasets that were carefully filtered, therefore we are not going to discuss the properties of a proper filter or the process of data 'cleaning', since it is not the main purpose of this work. A reader interested in this topic can find further information in the white paper *High Frequency Data Filtering*[1] by Thomas Neal Falkenberry. We now describe each dataset and the transformation of the data in a more detailed manner.

4.2 S&P 500 Futures index

The Standard & Poor 500 Futures index (henceforth SP) is a widely known free-float capitalization-weighted index, which consists of stock prices of top 500 American companies from the leading industries of the US economy. The index as we know it has been first published in 1957 and is considered by many to be the best representative of the US economy and market.

Our original dataset contains five-minute observations of SP starting on January 2, 1987 and ending on December 30, 2011 with a two-week gap between October 23, 1987 and November 8, 1987, which means we have 495 436 data points from 6 305 active trading days. The observations were recorded

[1]This white paper can be found at www.tickdata.com

throughout the time interval between 14:35 and 21:10 during the period from January 2, 1987 until October 31, 2002. Since November 1, 2002 the observations span the interval from 14:35 to 21:00. This would mean there were 80 and 78 price observations from each day, respectively, resulting in 79 and 77 intraday returns, respectively.

However, our dataset does not contain all intraday observations. Most of the times, no more than three observations from one day are missing, but on some occasions, there are less than 50 observations per day available. This could lead to inconsistency later in the estimations, arising from the definition of RV and other variation measures we make use of. The lack of observations could lead to overvalued or undervalued variation measures. For example, if there were observations missing between two distant data points (i.e. the time interval between two observations was longer than 5 minutes) and the price would move constantly up (or down) between these two observations, we would obtain an overvalued RV. On the other hand, if the price followed a chain of upward and downward movements so that the price would be the same in these two points (i.e. the return would be 0 or close to 0), the RV would be undervalued. We also need as many observations as possible for the RV to maintain its properties. Therefore, we drop the days with number of observations lower than or equal to 70, which account for just above 1 % of the total number of days. This leaves us with 6 239 days for further analysis.

We continue with the construction of the realized variation measures and components of RV. Firstly, the five-minute log-returns, which will be denoted as ${r_{t,h}}^2$, were calculated as $r_{t,h} = \ln(P_t/P_{t-h})$, where P_t denotes the intraday observation at time t and h denotes the five-minute interval between the two successive observations from one day. In accord with the definitions in the previous chapters, we were able to use the log-returns for the construction of the realized variation measures – realized variance, bipower variation and tripower quarticity. In addition, we also constructed standardized versions of the realized variance and bipower variation in the following manner

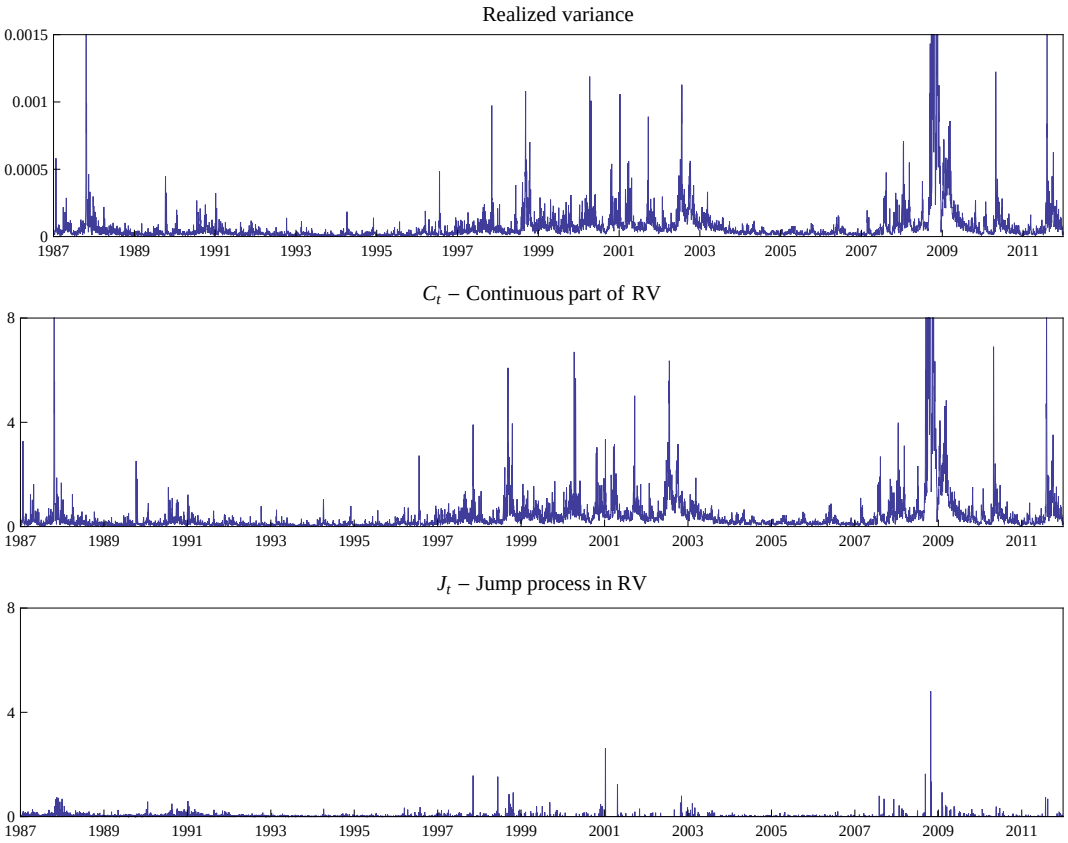
$$RV_t^{standardized} = \frac{RV_t}{\hat{\sigma}_{RV}}, \tag{4.1}$$

$$BV_t^{standardized} = \frac{BV_t}{\hat{\sigma}_{BV}}, \tag{4.2}$$

[2]We will denote five-minute intraday log-returns as $r_{t,h}$ throughout the whole chapter.

where $\hat{\sigma}_{RV}$ and $\hat{\sigma}_{BV}$ represent the standard deviations of realized variance and bipower variation, respectively.

Figure 4.1: Realized variance and variation components for SP



Source: Author's computations.

The next step in the process was the evaluation of the Z_t test statistic (which was defined in Equation 3.5). It is important to note, that we used the non-standardized realized variation measures for the test statistic. The variation components J_t and C_t (defined in Equation 3.6 and Equation 3.7, respectively), on the contrary, were constructed using the standardized measures. The standardization, inspired by Andersen *et al.* (2010), was done mainly in order to rescale the variables that appear in the logarithmic specification of the HAR-RV(-CJ) models.

To ensure a better notion of the data (together with the realized variation measures and the variation components) for the reader, we present a summary of descriptive statistics of SP intraday log-returns $r_{t,h}$, daily realized variance RV_t and its components C_t and J_t, in Table 4.1. Moreover, Figure 4.1 depicts plots of RV_t, C_t and J_t, which indicate dynamic dependencies in the series with

Table 4.1: Descriptive statistics of SP

	$r_{t,h}$	RV_t	C_t	J_t
Mean	9.67648×10^{-7}	0.00008	0.41764	0.02248
Std. dev.	0.00100	0.00018	0.98570	0.10053
Skewness	0.21898	12.4330	12.6785	24.0038
Kurtosis	28.5118	260.350	272.033	943.711
Min	-0.02389	2.94892×10^{-6}	0.01227	0
Max	0.03667	0.00578	32.5945	4.81333
Obs.	485912	6239	6239	6239

Source: Author's computations.

realized variance and the continuous part seeming much more predictable than the jump process. Among the three datasets, the price process of SP seems to be the least 'jumpy' as it has the lowest mean of the jump process, and lowest relative amount of days with jumps – 1 585 days representing 25.4 % of the total number of days.

4.3 Euro FX

The FOREX-Euro generally contains a set of reference rates of the major currencies against the Euro, which are determined on the foreign exchange market. Our dataset, the Euro FX (henceforth EC), follows the development of EUR/USD exchange rate, which is the most liquid one with the highest trading activity.

EC consists of five-minute records of the EUR/USD reference rate from the period between January 4, 1999 and December 30, 2011, i.e. we have 681 351 observations from 3 737 active trading days. This makes it the dataset with the shortest period covered, even though it represents the whole existence of the exchange rate. There are two main periods in this dataset – the first before July 1, 2003, and the second after this day. The reason is that before July 2003, only the day-session pit trading activity is recorded (i.e. there are no data from markets with electronic-only trading), while since the first trading day of July, all sessions are included (i.e. also data from Globex[3]). The observations in the

[3]Globex (or CME Globex) is an electronic trading platform initially developed for CME in 1992. The system then spread over markets becoming the first globally used electronic trading system. It runs continuously which makes trading possible in fact at all times and between traders worldwide.

first period come from the time interval starting at 13:25 and ending at 20:00, meaning we have a maximum of 80 observations per day (and a maximum of 79 intraday returns). Since the second period contains observations from all sessions, there is only one hour gap with no data each day. Therefore, we can have a maximum of 276 data points (and a maximum of 275 intraday returns) from each day.

Nevertheless, similarly to the previous dataset, some observations were missing again. Therefore, we had to find a threshold (which obviously had to be different for the two main periods) for the number of observations per day in order to filter the relevant days. In addition, it was necessary to divide each main period into two smaller periods, since the number of daily observations differed substantially even within the main periods.[4] The first main part was divided into one shorter interval (from January 4, 1999 to July 16, 1999) and a longer interval (from July 17, 1999 to June 30, 2003) with values of thresholds set to 60 and 70, respectively. The second main part was divided in a similar manner – into a short interval (from July 1, 2003 to October 31, 2003) and a longer interval (from November 1, 2003 to December 30, 2011) with values of thresholds set to 220 and 250, respectively. For the sake of clarity, we emphasize that days, containing no more observations than the value of the subsistent threshold, were dropped from the dataset and were not used for further analysis. As a result, only 3 083 days, representing 82.5 % of the original dataset, were used for the construction of the realized variation measures and components of RV, which followed.

Table 4.2: Descriptive statistics of EC

	$r_{t,h}$	RV_t	C_t	J_t
Mean	1.06666×10^{-6}	3.90820×10^{-5}	0.96989	0.04948
Std. dev.	0.00044	3.83392×10^{-5}	0.96729	0.16127
Skewness	0.09063	4.33271	4.53986	11.8609
Kurtosis	18.5440	33.7786	37.2539	253.014
Min	-0.01379	4.36115×10^{-6}	0.10648	0
Max	0.01160	0.00056	14.6749	4.59939
Obs.	629142	3083	3083	3083

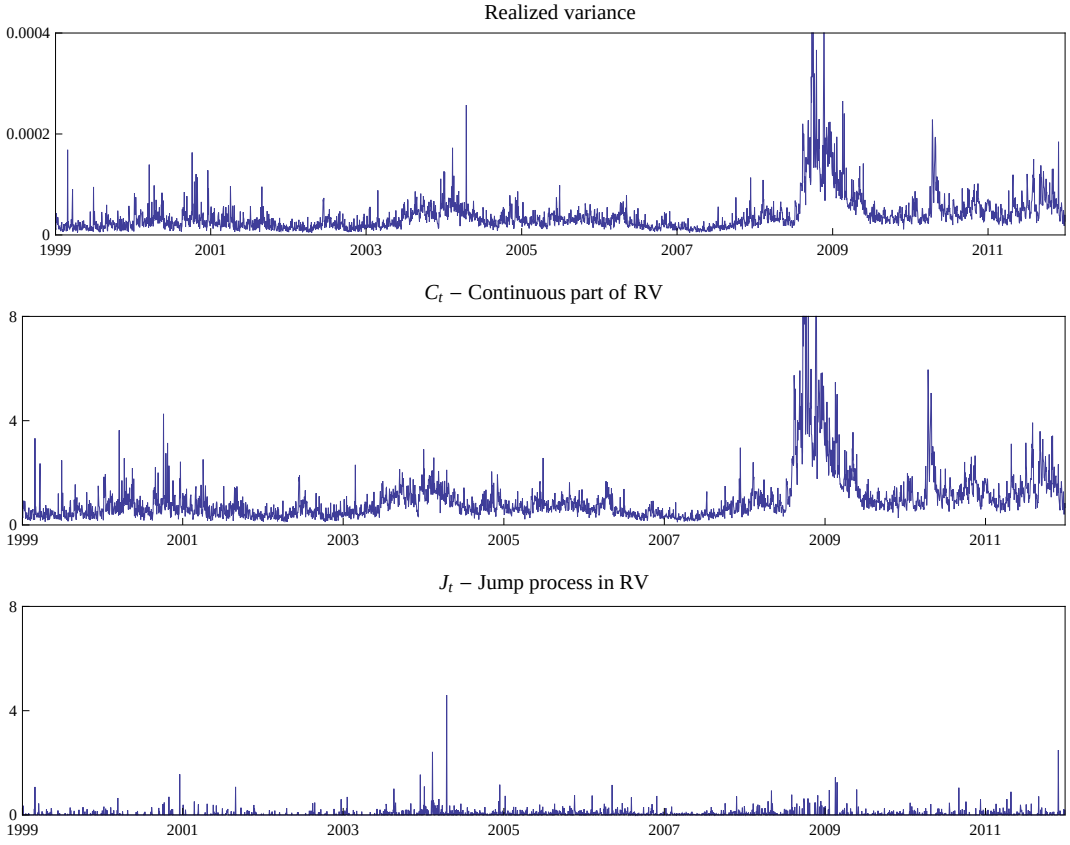
Source: Author's computations.

[4]We did not actually create subsamples from the original dataset. We only created certain intervals, where different values were applied as thresholds for the minimum number of daily observations.

Using the intraday log-returns of EC, we constructed the realized variation measures (realized variance, bipower variation and tripower quarticity) and components of RV (J_t and C_t) in the same way as we did in the previous section with SP. Again, we constructed both the standardized and non-standardized versions of RV an BV. The non-standardized measures were used for evaluation of the Z_t test statistic, while the standardized measures were used for the computation of RV components.

We also present a summary of descriptive statistics of EC intraday log-returns r_t, realized variance RV_t and its components, C_t and J_t, in Table 4.2, as well as plots of RV_t, C_t and J_t in Figure 4.2. It is clear from the plots that there are significant dynamic dependencies in the series with RV and the continuous part seeming much more predictable than the jump process. Contrary to the previous dataset, the price process of EC shows the highest relative amount of 'jumpy' days with jumps occuring on 837 days, which is 27.15 % of all relevant days. Also the mean of the jump process is almost 0.05 – about twice as high as in the other datasets.

Figure 4.2: Realized variance and variation components for EC



Source: Author's computations.

4.4 Light Crude NYMEX

Our last dataset, Light Crude NYMEX (henceforth CL), consists of five-minute prices of light crude oil futures contracts. The prices are quoted in American dollars and cents per barrel with 1 cent being the smallest possible price change.

CL covers the period starting on January 2, 1987, and ending on December 30, 2011, resulting in 807 553 observations from 6 654 active trading days. This is the largest dataset and it also required the largest modifications. We divided[5] the dataset into two main parts by the same event as the previous one – the transition from recording the day-session only to recording all sessions. Therefore, we have one period from the beginning until June 30, 2003, and another period from July 1, 2003, until the end. The observations in the first period were recorded during the time interval from 14:50 to 20:25. Therefore, we can have a maximum of 68 observations, although on the vast majority of days there were 65 data points. On the other hand, the second period contains all sessions with only a 45 minute gap each day, meaning we can have a maximum of 279 observations from each day.

As in the previous cases, there were certain observations missing and we had to determine the threshold for the minimum number of data points per day throughout the different periods, again. Moreover, the events from September 11, 2001 had a huge impact also on the recording of CL prices in the first main period of the dataset. For one week there were no observations at all and during the next three weeks the number of daily observations increased gradually from 28 to 47. In order to maintain the properties of RV, we had to throw away these three weeks. In addition, the time interval during which the data were recorded changed to 14:05 - 18:30, resulting in a maximum of only 54 observations per day. The thresholds were set to 60 in the period before September 11, 2001, and 50 during the period from October 8, 2001 to June 30, 2003 (i.e. days with number of observations equal or less than the value of the threshold were dropped). Another event, which occured on January 9, 1991, had a great impact on the evolution of the price of CL. On that day, the US and Iraq failed to reach an agreement, based on which Iraq would have withdrawn their troops from Kuwait, which resulted in escalated tension and much higher probability of war. The prices of CL reacted with extremely high volatility,

[5]The division is in the same sense as with EC – creating intervals where different values were applied as thresholds for the minimum number of daily observations.

therefore, we excluded this day from further analysis as it would cause great bias in the results.

The second main period also had to be divided into two parts. The reason is that in the first 4 years after the change in recording of the prices the number of maximum daily observations was increasing gradually from below 200 in 2003 to more than 250 in 2007. This substantial difference in the number of daily observations also required different thresholds. Therefore, we divided the second part of the dataset in the following manner: the first period covers days from July 1, 2003 to December 31, 2007 and the second period spans days from January 1, 2008 to December 30, 2011. The thresholds were set to 150 in the first period and to 250 in the second one (i.e. we threw away days with no more than 150 observations from the first period, and no more than 250 observations from the second one). Moreover, there was one extreme jump in the price on January 14, 2009, when the price changed from 37.38 to 44.55, creating an extremely high intraday return. This return had to be removed because of the bias it would have caused later in the results.

As a result of these adjustments, we used 5 947 days (almost 89.4 % of the original dataset) in the process of constructing the realized variation measures and components of RV. Using the intraday log-returns of CL, we constructed the realized variation measures (realized variance, bipower variation and tripower quarticity) and components of RV (J_t and C_t) in the same way as we did previously with SP and EC. Again, we constructed both the standardized and non-standardized versions of RV an BV. The non-standardized measures were used for the evaluation of the Z_t test statistic, while the standardized measures were used for the components of RV.

Table 4.3: Descriptive statistics of CL

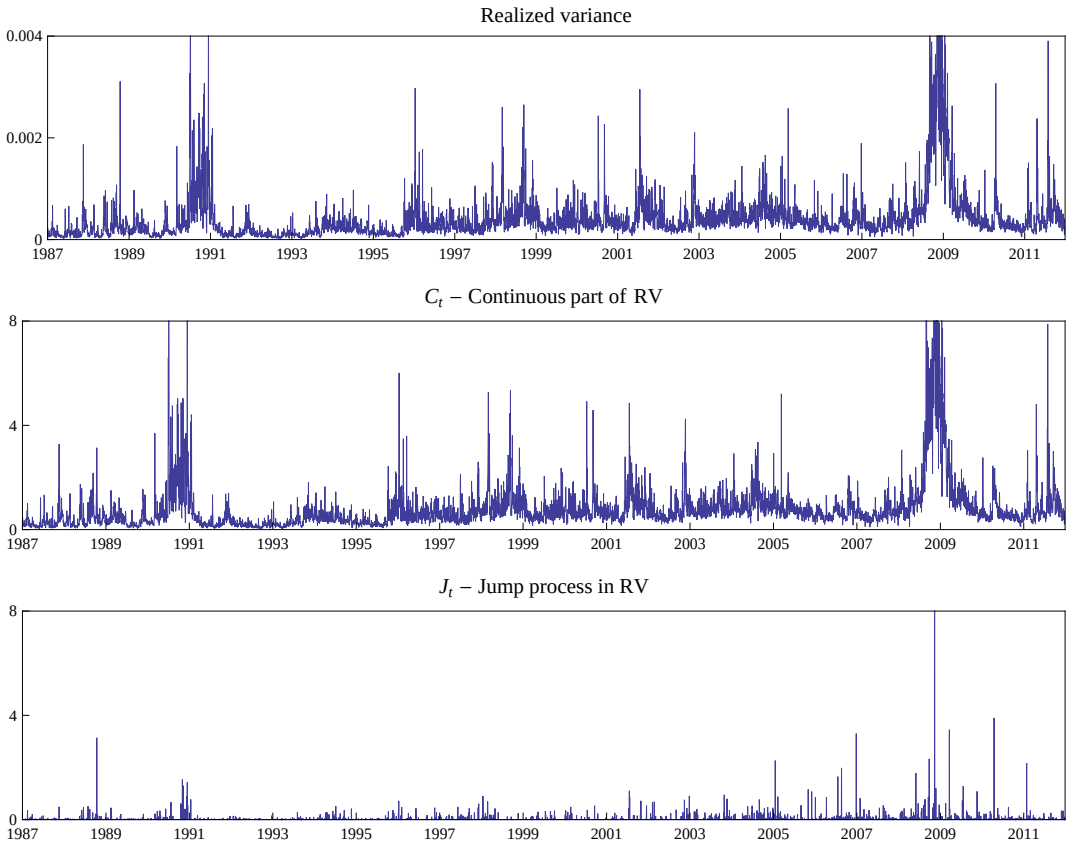
	$r_{t,h}$	RV_t	C_t	J_t
Mean	2.67806×10^{-6}	0.00042	0.78607	0.02849
Std. dev.	0.00179	0.00084	0.94125	0.19875
Skewness	2.18798	30.3912	4.87533	30.2183
Kurtosis	233.333	1314.76	36.1369	1392.05
Min	-0.04460	1.77825×10^{-5}	0.03591	0
Max	0.17548	0.03995	12.2623	10.5387
Obs.	767763	5948	5947	5947

Source: Author's computations.

In addition, a summary of descriptive statistics of CL intraday log-returns

$r_{t,h}$, realized variance RV_t and its components, C_t and J_t, can be found in Table 4.3, while plots of RV_t, C_t and J_t are shown in Figure 4.3. Similarly to the previous datasets, the plots show distinct dynamic dependencies in the series with the jump process appearing much more random than RV and the continuous part. CL has roughly the same mean of the jump process and relative number of 'jumpy' days (1 530 representing 25.73 % of the total number of days) as SP. However, a significant difference from the previous datasets is in the size of the jumps. There are several very high jumps with the maximum being more than twice as high as in case of SP or EC. This is caused probably by the importance of this commodity and the fact that there are many conflicts between countries with access to the biggest sources of oil.

Figure 4.3: Realized variance and variation components for CL



Source: Author's computations.

Having completed the first part of the empirical analysis we can move to the actual results.

Chapter 5

Discussion of the results

Finally, we arrive at the most important part of the work where our findings are reported and discussed. The chapter consists of three sections, each presenting the outcome of the same econometric method performed on our three datasets. For transparency reasons, a subsection is devoted to every dataset (SP, EC and CL) throughout the chapter.

In Section 5.1, we compare the HAR-RV and HAR-RV-CJ models (including their forecasting abilities) to show that decomposing the realized variance improves the performance of the HAR-RV model. However, we suspect that the parameters of the model change considerably over time. This suspicion is confirmed by the results of the so-called year-by-year estimations on all datasets, which are summarized in Section 5.2. The dynamics in the parameters should significantly affect the forecasting ability of the model. Therefore, one year out-of-sample forecasts were performed on each dataset using various lengths of the pre-forecast period. The results in Section 5.3 indicate that optimal length of the pre-forecast period is approximately one year, which is in contrast with the notion that the forecasting performance of a model improves with more data available for the parameters' estimation. For these reasons we do not consider the HAR model to be the most appropppriate for modeling realized variance. But let us first begin with the comparison of the models.

5.1 Comparison of HAR-RV and HAR-RV-CJ models on whole datasets

As already mentioned, this section confronts the performance of the HAR-RV and HAR-RV-CJ models. Simple OLS estimations were carried out on all

datasets coupled with out-of-sample forecasts. We forecasted the last year for each dataset and evaluated the predictions by Mincer-Zarnowitz regressions. The OLS estimates, as well as the summary statistics of every model and Mincer-Zarnowitz regression, are presented in tables. Each of them contains the estimated parameters (with standard errors in parentheses and p-values in square brackets) of both the logarithmic and non-logarithmic versions of the HAR-RV and HAR-RV-CJ models. In addition, R-squared, log-likelihood and the number of observations can be found at the bottom of each table.

5.1.1 SP

Table 5.1 contains results of the estimations performed on the first dataset - SP. Looking at the parameters of each model separately, we can say that weekly realized variance (in HAR-RV model) and weekly continuous variation (in HAR-RV-CJ model) have the biggest impact on daily realized variance. Parameters of the HAR-RV model are very similar to the continuous varia-

Table 5.1: Estimated parameters for SP

Estimated parameters of daily (d), weekly (w) and monthly (m) components in HAR-RV and HAR-RV-CJ models, reported with standard errors in parentheses and p-values in square brackets.

	$\ln(RV)$		RV	
	HAR-RV	HAR-RV-CJ	HAR-RV	HAR-RV-CJ
α	-0.150 (0.013) [0.000]	-0.148 (0.015) [0.000]	0.039 (0.010) [0.000]	0.048 (0.011) [0.000]
$\beta_{RV}^{(d)}$	0.324 (0.015) [0.000]	-	0.324 (0.015) [0.000]	-
$\beta_{RV}^{(w)}$	0.372 (0.024) [0.000]	-	0.384 (0.024) [0.000]	-
$\beta_{RV}^{(m)}$	0.256 (0.019) [0.000]	-	0.204 (0.022) [0.000]	-
$\beta_{C}^{(d)}$	-	0.315 (0.014) [0.000]	-	0.322 (0.015) [0.000]
$\beta_{C}^{(w)}$	-	0.344 (0.022) [0.000]	-	0.396 (0.024) [0.000]
$\beta_{C}^{(m)}$	-	0.238 (0.019) [0.000]	-	0.215 (0.022) [0.000]
$\beta_{J}^{(d)}$	-	-0.290 (0.130) [0.026]	-	-0.263 (0.095) [0.006]
$\beta_{J}^{(w)}$	-	0.088 (0.235) [0.707]	-	-1.381 (0.225) [0.000]
$\beta_{J}^{(m)}$	-	0.936 (0.334) [0.005]	-	1.740 (0.367) [0.000]
R^2	0.720989	0.722538	0.558552	0.570958
Log-l.	-5158.592	-5141.292	-6285.594	-6196.987
Obs.	6217	6217	6217	6217

Source: Author's computations.

tion parameters of the HAR-RV-CJ model, which is in accord with the fact

that continuous variation accounts for the significant part of RV. Overall, the HAR-RV-CJ model fits the data better than the HAR-RV model, although the improvement is only minor.

Results of the Mincer-Zarnowitz regressions in Table 5.2 insinuate that the logarithmic specification predicts realized variance almost equally for both models. While HAR-RV-CJ has values of parameters closer to the desired ones, R^2 is insignificantly higher for HAR-RV model. However, HAR-RV model seems to give more accurate forecasts than HAR-RV-CJ model in the non-logarithmic specification, as indicated by values of both parameters and R^2. Generally, the log-models perform better than the non-log-models (difference in R^2 is around 0.2).

Table 5.2: Mincer-Zarnowitz regressions for SP

Estimated parameters, evaluating one-year out-of-sample forecasts of HAR-RV and HAR-RV-CJ models, reported with standard errors in parentheses and p-values in square brackets.

	$\ln(RV)$		RV	
	HAR-RV	HAR-RV-CJ	HAR-RV	HAR-RV-CJ
α	-0.007 (0.060) [0.914]	-0.003 (0.061) [0.962]	0.002 (0.054) [0.977]	0.020 (0.055) [0.716]
β	0.966 (0.043) [0.000]	1.009 (0.045) [0.000]	1.022 (0.067) [0.000]	0.966 (0.067) [0.000]
R^2	0.676353	0.674216	0.489038	0.459438
Log-l.	-205.9252	-206.7411	-230.9762	-237.9592
Obs.	248	248	248	248

Source: Author's computations.

5.1.2 EC

Parameters estimated on CE are reported in Table 5.3. Unlike in the previous case, the monthly realized variance in HAR-RV model (and monthly continuous variation in HAR-RV-CJ model) has the greatest impact on RV for the logarithmic specification, while daily RV (and daily continuous variation) affects the realized variance most significantly for the non-log-model. Again, parameters of the continuous variation are very similar to the HAR-RV model parameters and, in comparison, the HAR-RV-CJ model provides a slightly better fit.

Mincer-Zarnowitz regressions support the previous assertion, as parameters and statistics in Table 5.4 show that HAR-RV-CJ models give better forecasts than HAR-RV models for both specifications. However, the explanatory ability

is quite low for all models, since R^2 of the logarithmic HAR-RV-CJ model slightly above 0.43 is the maximum value.

Table 5.3: Estimated parameters for EC

Estimated parameters of daily (d), weekly (w) and monthly (m) components in HAR-RV and HAR-RV-CJ models, reported with standard errors in parentheses and p-values in square brackets.

	$\ln(RV)$		RV	
	HAR-RV	HAR-RV-CJ	HAR-RV	HAR-RV-CJ
α	-0.060 (0.008) [0.000]	-0.018 (0.015) [0.220]	0.049 (0.016) [0.002]	0.080 (0.018) [0.000]
$\beta_{RV}^{(d)}$	0.210 (0.022) [0.000]	-	0.398 (0.021) [0.000]	-
$\beta_{RV}^{(w)}$	0.338 (0.038) [0.000]	-	0.283 (0.034) [0.000]	-
$\beta_{RV}^{(m)}$	0.420 (0.034) [0.000]	-	0.271 (0.030) [0.000]	-
$\beta_{C}^{(d)}$	-	0.244 (0.023) [0.000]	-	0.479 (0.023) [0.000]
$\beta_{C}^{(w)}$	-	0.296 (0.039) [0.000]	-	0.216 (0.036) [0.000]
$\beta_{C}^{(m)}$	-	0.408 (0.034) [0.000]	-	0.275 (0.030) [0.000]
$\beta_{J}^{(d)}$	-	-0.128 (0.080) [0.109]	-	-0.160 (0.069) [0.020]
$\beta_{J}^{(w)}$	-	0.355 (0.157) [0.024]	-	0.427 (0.173) [0.013]
$\beta_{J}^{(m)}$	-	-0.033 (0.264) [0.901]	-	-0.301 (0.311) [0.333]
R^2	0.661503	0.665371	0.696633	0.705581
Log-l.	-1634.136	-1616.550	-2524.652	-2478.828
Obs.	3061	3061	3061	3061

Source: Author's computations.

Table 5.4: Mincer-Zarnowitz regressions for EC

Estimated parameters, evaluating one-year out-of-sample forecasts of HAR-RV and HAR-RV-CJ models, reported with standard errors in parentheses and p-values in square brackets.

	$\ln(RV)$		RV	
	HAR-RV	HAR-RV-CJ	HAR-RV	HAR-RV-CJ
α	0.021 (0.029) [0.465]	0.019 (0.028) [0.510]	0.117 (0.114) [0.305]	0.096 (0.110) [0.384]
β	0.996 (0.076) [0.000]	0.988 (0.074) [0.000]	0.924 (0.076) [0.000]	0.941 (0.073) [0.000]
R^2	0.422249	0.433505	0.386983	0.413372
Log-l.	-80.68715	-191.3273	-6285.594	-186.1352
Obs.	236	236	236	236

Source: Author's computations.

5.1.3 CL

We present the estimates obtained from the last dataset in Table 5.5. The logarithmic model suggests that daily realized variance (and daily continuous variation) influences the realized variance most noticeably. However, monthly realized variance (and monthly continuous variation) contributes mostly to the realized variance in the non-log-model. As in the previous cases, the HAR-RV-CJ model explains the data only a little better than the HAR-RV model with parameters of the HAR-RV model being close to the parameters of the continuous variation in HAR-RV-CJ model.

Table 5.5: Estimated parameters for CL

Estimated parameters of daily (d), weekly (w) and monthly (m) components in HAR-RV and HAR-RV-CJ models, reported with standard errors in parentheses and p-values in square brackets.

	$\ln(RV)$		RV	
	HAR-RV	HAR-RV-CJ	HAR-RV	HAR-RV-CJ
α	-0.078 (0.007) [0.000]	-0.047 (0.010) [0.000]	0.039 (0.011) [0.000]	0.035 (0.011) [0.001]
$\beta_{RV}^{(d)}$	0.352 (0.015) [0.000]	-	0.280 (0.015) [0.000]	-
$\beta_{RV}^{(w)}$	0.328 (0.024) [0.000]	-	0.294 (0.027) [0.000]	-
$\beta_{RV}^{(m)}$	0.274 (0.020) [0.000]	-	0.379 (0.024) [0.000]	-
$\beta_{C}^{(d)}$	-	0.359 (0.016) [0.000]	-	0.305 (0.017) [0.000]
$\beta_{C}^{(w)}$	-	0.331 (0.025) [0.000]	-	0.348 (0.030) [0.000]
$\beta_{C}^{(m)}$	-	0.264 (0.021) [0.000]	-	0.355 (0.027) [0.000]
$\beta_{J}^{(d)}$	-	0.061 (0.071) [0.386]	-	0.134 (0.044) [0.002]
$\beta_{J}^{(w)}$	-	0.006 (0.116) [0.960]	-	-0.145 (0.107) [0.174]
$\beta_{J}^{(m)}$	-	0.010 (0.179) [0.957]	-	-0.423 (0.214) [0.048]
R^2	0.707941	0.710269	0.647845	0.653648
Log-l.	-3681.773	-3658.064	-5321.490	-5272.266
Obs.	5925	5925	5925	5925

Source: Author's computations.

Table 5.6 offers interesting results of the Mincer-Zarnowitz regressions. While parameters for both specifications indicate that HAR-RV model is the better one, R^2 is higher for the HAR-RV-CJ models. The difference in parameters is most evident in the non-log-models, where parameters α differ by 0.08 and β by 0.09, which is quite a significant contrast. Nevertheless, all models explain the data rather poorly as the highest R^2 is 0.45 in the case of logarith-

mic HAR-RV-CJ model and the non-log-models have values of R^2 under 0.3.

Table 5.6: Mincer-Zarnowitz regressions for CL

Estimated parameters, evaluating one-year out-of-sample forecasts of HAR-RV and HAR-RV-CJ models, reported with standard errors in parentheses and p-values in square brackets.

	$\ln(RV)$		RV	
	HAR-RV	HAR-RV-CJ	HAR-RV	HAR-RV-CJ
α	-0.007 (0.034) [0.843]	-0.010 (0.034) [0.759]	0.041 (0.106) [0.697]	0.119 (0.099) [0.228]
β	0.933 (0.069) [0.000]	0.920 (0.066) [0.000]	0.966 (0.100) [0.000]	0.874 (0.090) [0.000]
R^2	0.440312	0.449497	0.282379	0.284742
Log-l.	-149.9883	-148.0275	-247.4560	-247.0651
Obs.	237	237	237	237

Source: Author's computations.

Despite the ambiguity in the forecasting performances, we can conclude that the HAR-RV-CJ model outperforms the HAR-RV model, even though the difference is quite small. One of the reasons may be that we in fact only decompose the realized variance into two parts, but the model itself remains the same. A much better fit could be achieved, for example, by modeling each component separately, as already mentioned above. Another cause of the limited (if any at all) improvement might be the enormous amounts of data. Nontrivial dynamics of the parameters over time could make it almost impossible to enhance the explanatory ability of the model on such long datasets. We believe that the inconsistency of the forecasting results on our datasets is also a consequence of this dynamics in parameters (both the continuous part and jump component of RV). However, the small differences between the forecasting performances of HAR-RV model and HAR-RV-CJ model are caused probably by the dynamics of jump parameters. We think so, because the continuous variation is the prevailing component of RV, i.e. their dynamics should be very similar (positively correlated). The jump component, on the other hand, should change more individually and independently from RV parameters, creating the differences in the forecasting performances. That is, if the true jump parameters during the forecasted period are significantly different from the estimated ones, the performance of HAR-RV-CJ model should be worse than that of HAR-RV

model, and vice versa. Therefore, we believe the dynamics of the parameters is important and we shall investigate it further.

Thus, we divide each dataset into equally sized subsamples representing approximately one year and estimate the regression on them (a so-called year-by-year estimation) to find out whether parameters are stable in time or not. Generally, by performing OLS estimation (of any model), we assume that parameters of the model are constant over time. Therefore, if the parameters in fact change in time, the assumption is violated, i.e. we should not be using OLS. The following section provides a detailed discussion of our findings.

5.2 Comparison of year-by-year estimates and estimates from whole datasets

In this part, we restrict ourselves to the non-logarithmic HAR-RV-CJ model, as our main interest is to demonstrate the dynamics of the model estimates and we obtained results qualitatively similar for both investigated models. Furthermore, we find the intuition behind this specification clearer and the parameters more interpretable in comparison to the logarithmic model, therefore it is the model of our choice.

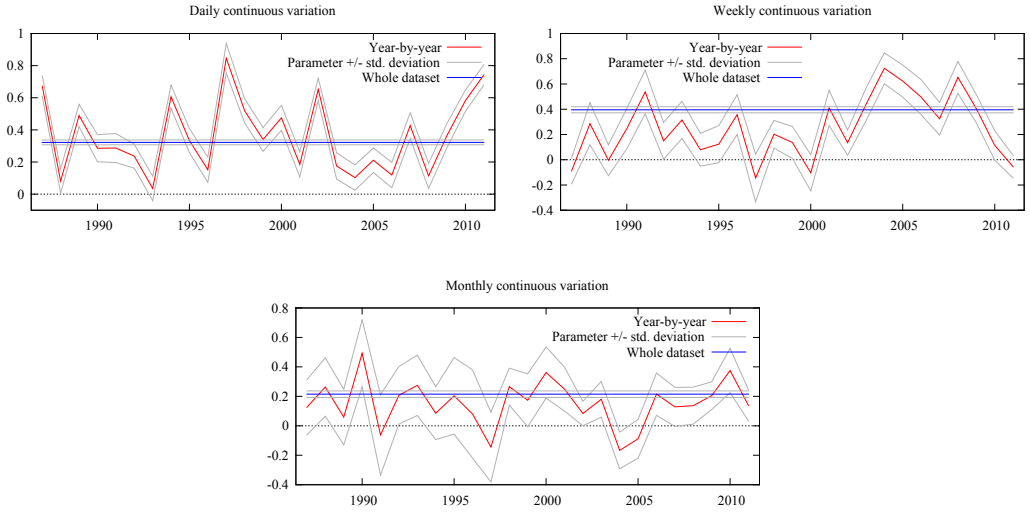
After the division of the datasets into equally sized subsets, we obtained 25 subsamples of SP containing 248 to 249 observations, 13 subsamples of EC with 235 to 236 observations and 25 subsamples of CL consisting of 237 observations. In the remainder of this section only figures showing the evolution of continuous variation parameters will be depicted for two main reasons. Firstly, these parameters change more significantly than parameters of the jump components, and secondly, the continuous part is the prevailing component of RV. Each of the figures shows one parameter (daily, weekly or monthly) estimated on the whole dataset (blue line) with a band representing one standard deviation of the parameter in both directions, and the estimates from every single year (red line) also with a band for one standard deviation. Tables with more detailed results of the estimations can be found in Appendix A, as well as figures and tables of the jump components' dynamics.

5.2.1 SP

We can deduce from Figure 5.1 that parameters of the model for SP change significantly over time, especially the daily and weekly ones (from one year to the next, they often change even by multiples of their standard errors). Moreover, a sort of pattern can be recognized in the dynamics of the daily

Figure 5.1: Year-by-year continuous parameters for SP

Dynamics of the daily, weekly and monthly parameters from year-by-year estimations compared with estimates from the whole dataset.



Source: Author's computations.

and weekly parameters – they change in the opposite direction (value of the correlation coefficient is -0.78). This tells us that usually the realized variance is driven mostly by either only the daily or only the weekly variation, i.e. either agents with high trading frequency are responsible for the volatility or agents with medium trading frequency.

R^2 of the regressions (from Table A.2) indicates that on some years the model fits the data quite well (R^2 is almost 0.75), while on some years the fit is very poor (with R^2 only slightly above 0.1). We consider this to be another proof that the HAR model is unable to capture the dynamics of realized variance sufficiently.

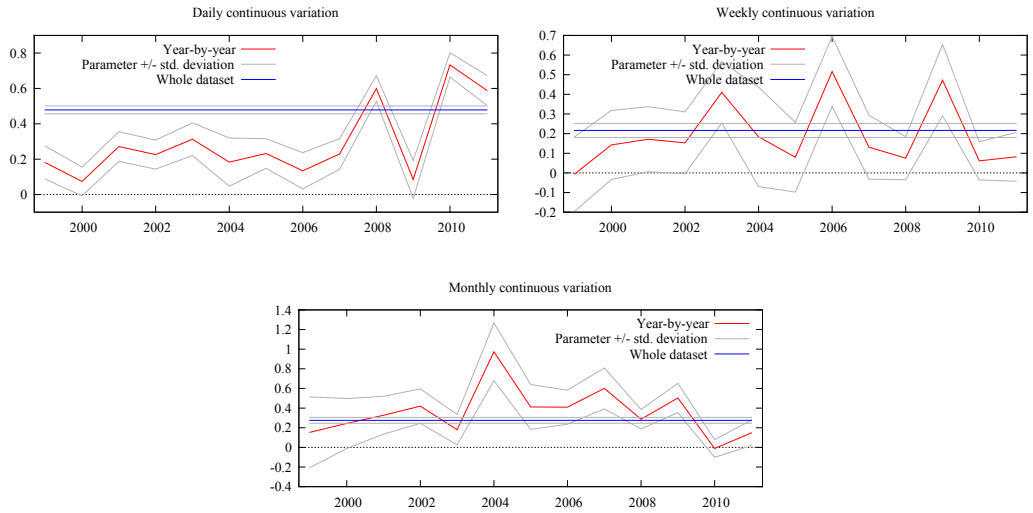
5.2.2 EC

Looking at Figure 5.2 we must admit that the dynamics in the parameters for EC is not so distinct as for SP, but this can be caused by the length of

the dataset (only 13 years). However, the weekly parameter exhibits certain variability and most importantly the daily parameter is very unstable in the last five years. The range of R^2 values, which can be found in Table A.1, is very similar to the previous one, from 0.05 to almost 0.74, supporting our view that HAR model cannot give a true picture of RV dynamics.

Figure 5.2: Year-by-year continuous parameters for EC

Dynamics of the daily, weekly and monthly parameters from year-by-year estimations compared to estimates from the whole dataset.



Source: Author's computations.

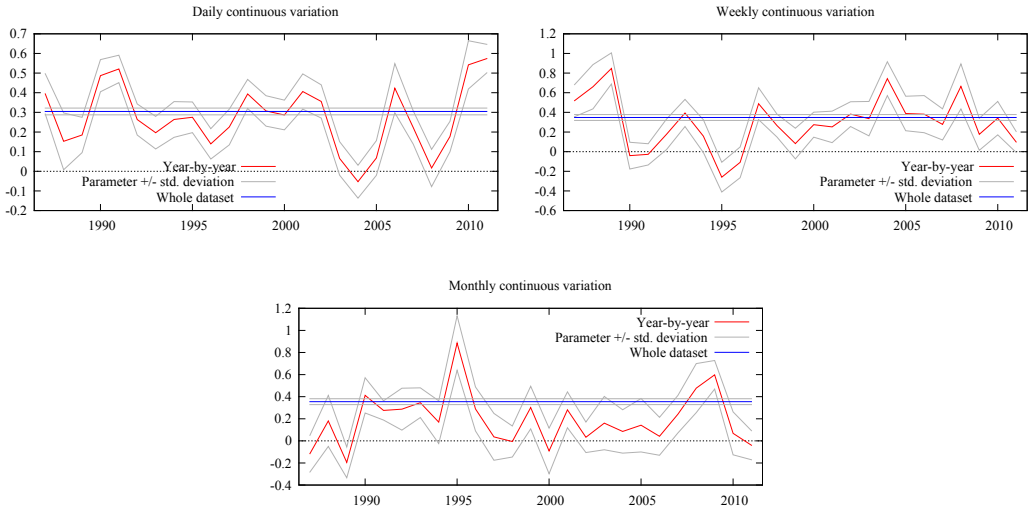
5.2.3 CL

Figure 5.3 reveals significant dynamics of the parameters also for CL. We can suspect a negative correlation between the daily and weekly coefficients and also between the weekly and monthly coefficients, but these relations are certainly not as strong (correlation coefficients around -0.5) as between the daily and weekly parameters for SP. R^2 from Table A.3 are consistent with those of the previous datasets, the model fit varies from just above 0.05 to more than 0.72.

From the previous results, we can conclude that parameters of the HAR-RV(-CJ) models are very dynamic, resulting in limited ability to improve the performance of the model by adding more observations to the dataset. Thus, using OLS for modeling RV is not an optimal choice of estimator. Furthermore, this dynamics should significantly affect the model's key purpose – forecasting of RV. Therefore, another evidence of the highly dynamic parameters is pre-

Figure 5.3: Year-by-year continuous parameters for CL

Dynamics of the daily, weekly and monthly parameters from year-by-year estimations compared with estimates from the whole dataset.



Source: Author's computations.

sented in the next part, where we investigate the forecasting performance of the model estimated on datasets of various lengths.

5.3 Forecasting

This section demonstrates the impact of the parameters' dynamics on the HAR-RV-CJ model's forecasting ability (both the logarithmic and non-logarithmic specifications). We concentrate on this because forecasting through the use of OLS models makes sense only when the parameters of the model are stable in time. That is, if the estimated parameters assumed various values from some interval (e.g. approximately from -0.2 to 1 as our parameters of continuous RV component) over time, we would obtain some kind of 'average' estimate of these parameters when estimating over the whole dataset. Consequently, the forecast would be also only an 'average' incapable of taking the dynamics and true values of parameters into account. Of course, estimating the parameters on longer datasets could make our 'average' parameters closer to the true 'average' parameters, i.e. the forecast more accurate on average, but this forecast would not be the best possible.

Thus, out-of-sample forecasts were performed on each dataset using various lengths of the pre-forecast period. All datasets were divided into two parts

– one part representing approximately the last year of the dataset and the other part consisting of the rest of the data, the so-called pre-forecast part. Then we carried out the forecasting of the last year realized variance, so that the parameters of the model were estimated on different segments of the pre-forecast period. The SP and CL parameters were estimated gradually on the last year, last 2, 3, 5, 10, 15 and 20 years and also on the whole period. In case of EC it was 1, 2, 3, 5 and 10 years and the whole period, as we have only 13 years of data for EC in total. In each subsection we present figures and tables with statistics illustrating the performance of the forecasts.

In the text, we only present two plots for each dataset to highlight the difference between the best and the worst forecast. We always choose the specification (logarithmic or non-logarithmic) of the model so that the contrast is most distinctive. However, complete plots of both specifications can be found in Appendix A for all datasets. Let us turn our attention to the actual results.

5.3.1 SP

Both MSE and Theil's U in Table 5.7 reveal that the best forecast is achieved with the model estimated on the last two years before the forecasted period,

Table 5.7: Forecast evaluation for SP

Forecast evaluation statistics comparing the accuracy of one-year out-of-sample forecasts depending on the length of pre-forecast period.

	$\ln(RV)$		RV	
	Mean Squared Error	Theil's U	Mean Squared Error	Theil's U
1	0.29612	1.5636	0.33915	0.98272
2	0.29534	1.0785	0.33004	0.85156
3	0.30020	1.1256	0.47372	1.17220
5	0.29992	1.0956	0.46874	1.07380
10	0.30487	1.1470	0.45510	1.01740
15	0.30442	1.1157	0.43142	0.98925
20	0.30834	1.1130	0.43113	0.97585
All	0.31039	1.1079	0.39940	0.93696

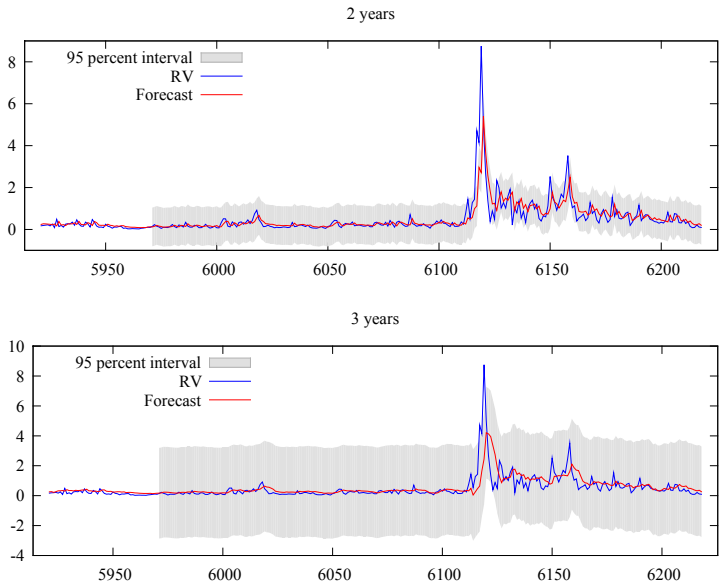
Source: Author's computations.

while the worst forecast comes from the model based on the last three years. However, all predictions of the logarithmic models are worse than the naive model as Theil's U is always above 1. Nevertheless, the performance of the

forecast improves slowly with extending the pre-forecast period, but it never outperforms the period of last 2 years before the forecast. Very similar results were obtained also for the logarithmic specification of the models. Figure 5.4 provides an illustration of the difference between the best and the worst model forecasts.

Figure 5.4: Comparison of best and worst forecast for SP

One-year out-of-sample forecasts based on 2-year pre-forecast period and on 3-year pre-forecast period.



Source: Author's computations.

5.3.2 EC

Table 5.8 suggests that the non-logarithmic models give very similar predictions, but the prediction based on the last year is slightly better than the rest. On the other hand, Theil's U for the logarithmic specification clearly shows that the forecast based on the last year is the best, while the forecast based on the whole period is the worst (even worse than the naive model). Moreover, the forecast gets worse gradually with longer pre-forecast periods, which is in contrast with the theory that the estimated parameters are more precise with more available data (i.e. also the forecast should improve) as they should asymptotically converge to the true parameters. For a better notion of the difference, the best and the worst forecasts of the logarithmic specification are depicted in Figure 5.5.

Table 5.8: Forecast evaluation for EC

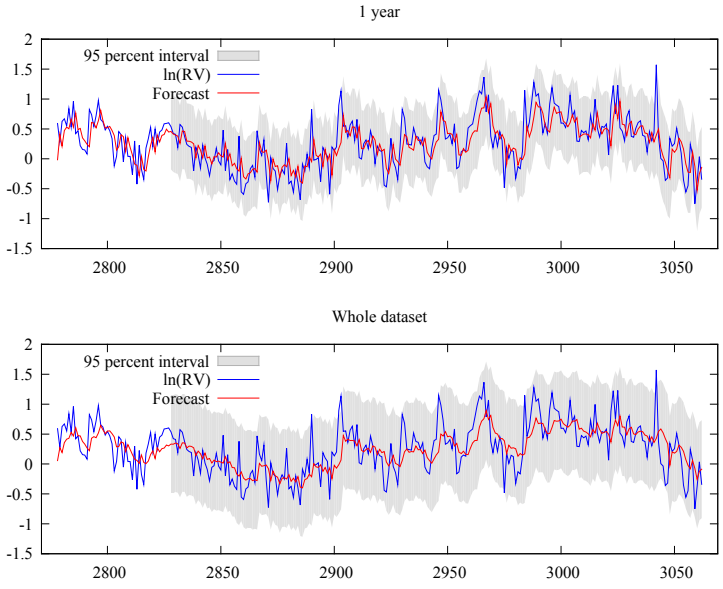
Forecast evaluation statistics comparing the accuracy of one-year out-of-sample forecasts depending on the length of pre-forecast period.

	$\ln(RV)$		RV	
	Mean Squared Error	Theil's U	Mean Squared Error	Theil's U
1	0.10107	0.54907	0.28365	0.90371
2	0.10314	0.69903	0.28612	0.90972
3	0.10411	0.81976	0.28788	0.92885
5	0.10696	0.85348	0.28431	0.92081
10	0.11211	1.00370	0.28353	0.91606
All	0.11408	1.03460	0.28525	0.92045

Source: Author's computations.

Figure 5.5: Comparison of best and worst forecast for EC

One-year out-of-sample forecasts based on 1-year pre-forecast period and on pre-forecast period containing the whole dataset.



Source: Author's computations.

5.3.3 CL

MSE and Theil's U from Table 5.9 indicate that the model based on the last year is the best for forecasting of CL realized variance, while the worst model is based on the last 3 years. Similarly to SP, the forecasting ability of the model slightly improves as we prolong the dataset, but it never reaches the performance of the prediction from the most recent data before the forecast. Interestingly, all logarithmic models give worse forecasts than the naive model as Theil's U yields values greater than 1 every time. We again plot the best and the worst forecasts of the non-logarithmic specification in Figure 5.6.

Table 5.9: Forecast evaluation for CL

Forecast evaluation statistics comparing the accuracy of one-year out-of-sample forecasts depending on the length of pre-forecast period.

	$\ln(RV)$		RV	
	Mean Squared Error	Theil's U	Mean Squared Error	Theil's U
1	0.19092	1.4181	0.43615	0.94002
2	0.20369	1.3787	0.47681	1.00200
3	0.21436	1.0533	0.54723	1.03210
5	0.21151	1.0347	0.53931	1.02720
10	0.21695	1.0197	0.52282	1.01580
15	0.21078	1.0938	0.49992	1.00290
20	0.21095	1.1444	0.49570	1.00060
All	0.20556	1.1973	0.47492	0.99353

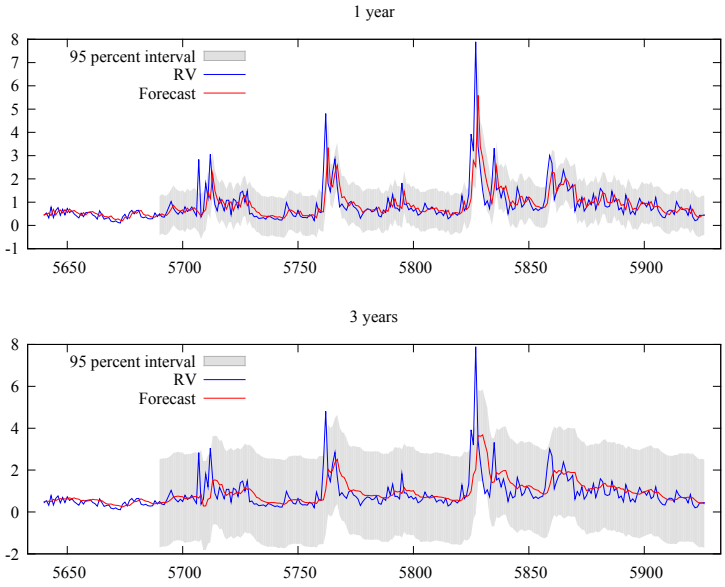
Source: Author's computations.

Results presented in this section confirm that the dynamics of the parameters is indeed significant as we saw that estimates obtained from short periods gave the best forecasts for all of our datasets. This instability of parameters suggests that the model is not stationary, meaning that HAR models used in this work are not suitable for modeling realized volatility.

To sum up, we showed that there are significant jumps in the price processes (on all three datasets) and it is reasonable to model them separately from the continuous part of the realized variance as the HAR-RV-CJ model achieved a better fit than the HAR-RV model. However, HAR model turned out to be incapable of modeling the dynamics of the model's parameters, which was confirmed by the year-by-year estimations on every dataset. Results obtained in the last section only support the previous findings as forecasts based on

Figure 5.6: Comparison of best and worst forecast for CL

One-year out-of-sample forecasts based on 1-year pre-forecast period and on 3-year pre-forecast period.



Source: Author's computations.

estimations from short periods turned out to be more accurate than those based on estimations from longer periods. Thus, we should model realized volatility with tools perhaps more sophisticated than HAR model.

Chapter 6

Conclusion

This work concentrates on the topic of modeling realized volatility, particularly on the role of jumps. Put another way, we investigate if it is of any use to account for the macroeconomic or firm-specific news when modeling volatility. Our main contribution stems from the fact that the analysis is performed on very long high frequency datasets – S&P 500 Futures index and Light Crude NYMEX spanning the time period from January 2, 1987, to December 30, 2011, and Euro FX with data starting on January 4, 1999, ending also on December 30, 2011.

In the beginning of the thesis we introduce the theory behind realized variation measures, we define realized variance, quadratic variation and the setting where we consider the price process to consist of a continuous part and a jump component. Then we continue with the bipower variation which enables us to decompose RV, followed by a jump detection test statistic which identifies the significant jumps in the return process. In the theoretical part we also mention the problem of microstructure noise and its possible solutions. After the HAR models are introduced we move to the empirical part of this work.

Before presenting and discussing the results, we carefully describe the datasets and their modifications as this knowledge is needed for a better notion of the whole work. The empirical results give answers to our main concerns. First, we want to know whether decomposing RV into its two parts significantly improves the predictions of volatility also on datasets as long as ours. Further, we investigate the dynamics of the parameters in time. The reason is that using OLS models implies that we assume parameters stable in time. However, if the parameters were significantly dynamic, it would mean that the assumption is violated, hence making the OLS not valid for modeling of RV. Finally, we

investigate if volatilities of different types of assets behave specifically or rather similarly.

The results themselves indicate that decomposing RV into its components improves the fit of the HAR model on every dataset. However, one-year out-of-sample forecasts provide us with certain ambiguity as forecasts of EC are more accurate when we decompose RV, while forecasts of SP favor the simple HAR-RV model. Moreover, forecasts of CL bring very little light into the problem as the logarithmic specification gives better forecasts for HAR-RV-CJ model, while between non-log-models HAR-RV seems to perform slightly better. Nevertheless, this unclarity might already be caused by the dynamics of the parameters which is examined in the next steps. Year-by-year estimations reveal that parameters of the HAR-RV-CJ model are highly dynamic in time, particularly parameters of the continuous components which account for the prevailing part of RV. The reason for the parameters' dynamics might be the different behavior of the main types of agents (who are responsible for volatility in financial markets) in times of stability and in times of turmoils such as the recent crisis. In addition, analysis of the impact of the length of pre-forecast period on the forecast accuracy shows that short periods (1-2 years) before the forecasted ones are the best for estimating the parameters. The previous results clearly point to the fact that parameters of our HAR models are not stable in time, therefore, we should use models that are able to capture the dynamics of RV better than HAR models. All datasets suggested roughly the same conclusions, indicating that volatility behaves similarly for various types of assets.

In conclusion, we showed that, consistently with previous studies, it is reasonable to decompose RV and model the components of volatility separately. On the other hand, we found that simple HAR-RV models used in this thesis are not approppriate for the modeling of RV as they are not able to capture its dynamics.

Bibliography

Andersen, T. & L. Benzoni (2007): "Realized Volatility. In T. Andersen, R. Davis, J. Kreiss, and T. Mikosch (Eds.)," *Volatility. In J. Birge and V. Linetsky (Eds.), Handbook of Financial Time Series.* Springer Verlag.

Andersen, T., L. Benzoni, J. Lund (2002): "An empirical investigation of continuous-time equity return models." *Journal of Finance* **(57)**: pp. 1239–1284.

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

Andersen, T., T. Bollerslev, F. Diebold, & P. Labys (2001): "The distribution of realized exchange rate volatility." *Journal of the American Statistical Association* **(96)**: pp. 42–55.

Andersen, T., T. Bollerslev, F. Diebold, & P. Labys (2003): "Modeling and Forecasting Realized Volatility." *Econometrica* **71(2)**: pp. 579–625.

Andersen, T., T. Bollerslev & F. Diebold (2007): "Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility." *NBER Working Paper Series 11775*, National Bureau of Economic Research, Inc.

Andersen, T., T. Bollerslev & X. Huang (2010): "A reduced form framework for modeling volatility of speculative prices based on realized variation measures." *Journal of Econometrics* **160(1)**: pp. 176–189.

Back, K. (1991): "Asset prices for general processes." *Journal of Mathematical Economics* **(20)**: pp. 317–395.

Bandi, F. & J. Russell (2006a): "Separating microstructure noise from volatility." *Journal of Financial Economics.* **(79)**: pp. 655–692.

Bandi, F. & J. Russell (2006b): "Volatility. In J. Birge and V. Linetsky (Eds.)", *Handbook of Financial Engineering.* Elsevier.

Barndorff-Nielsen, O., P. Hansen, A. Lunde & N. Shephard (2008): "Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise." *Econometrica* **76(6)**: pp. 1481–1536.

Barndorff-Nielsen, O. & N. Shephard (2001): "Non-gaussian Orstein-Uhlenbeck-based models and some of their uses in financial economics." *Journal of the Royal Statistical Society, Series B* **(63)**: pp. 167–241.

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

Barndorff-Nielsen, O. & N. Shephard (2002b): "Estimating quadratic variation using realized variance." *Journal of Applied Econometrics* **(17)**: pp. 457–477.

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

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

Bates, D. (2000): "Post-'87 Crash Fears in S&P 500 Futures Options." *Journal of Econometrics* **(94)**: pp. 181–238.

Black, F. & M. Scholes (1973): "The pricing of options and corporate liabilities." *Journal of Political Economy* **(81)**: pp. 637–654.

Bollerslev, T., T. H. Law & G. Tauchen (2008): "Risk, jumps, and diversification." *Journal of Econometrics* **144(1)**: pp. 234–256.

Christensen, B. J. & M. Ø. Nielsen (2005): "The Implied-Realized Volatility Relation with Jumps in Underlying Asset Prices." *Working Papers 1186*, Queen's University, Department of Economics

Corsi, F. (2004): "A Simple Long Memory Model of Realized Volatility." *Working Paper*, University of Lugano.

Eraker, B., M. Johannes & N. Polson (2003): "The Impact of Jumps in Volatility." *Journal of Finance* **(58)**: pp. 1269–1300.

Fleming, J. & B. S. Paye (2010): "High-frequency returns, jumps and the mixture of normals hypothesis." *Journal of Econometrics* **160(1)**: pp. 119–128.

French, K., G. Schwert & R. Stambaugh (1987): "Expected stock returns and volatility." *Journal of Financial Economics* **(19)**: pp. 3–29.

Hansen, P. & A. Lunde (2006): "Realized variance and market microstructure noise." *Journal of Business and Economic Statistics* **24(2)**: pp. 127–161.

McAleer, M. & M. Medeiros (2008): "Realized Volatility: A Review." *Econometric Reviews* **27(1-3)**: pp. 10–45.

Merton, R. (1976): "Option pricing when underlying stock returns are discontinuous." *Journal of Financial Economics* **(3)**: pp. 125–144.

Mincer, J. & V. Zarnowitz (1969): "The evaluation of economic forecasts." New York: National Bureau of Economic Research.

Müller, U., M. Dacorogna, R. Dav, R. Olsen, O. Picket & J. von-Weizsacker (1997): "Volatilities of different time resolutions - analysing the dynamics of market components." *Journal of Empirical Finance* **(4)**: pp. 213–239.

Pan, J. (2002): "The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time Series Study." *Journal of Financial Economics* **(63)**: pp. 3–50.

Protter, P. (1992): *Stochastic integration and differential equations: A new approcah.* New York: Springer-Verlag.

Theil, H. (1966): *Applied Economic Forecasting* Amsterdam: North-Holland.

Zhang, L. (2006): "Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach." *Bernoulli* **(12)**: pp. 1019–1043.

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

Appendix A

Results of estimations and forecasting

Table A.1: Year-by-year estimates of continuous parameters for EC

Estimated parameters of daily (d), weekly (w) and monthly (m) continuous components from the non-logarithmic HAR-RV-CJ model are reported with standard errors in parentheses and p-values in square brackets. Periods 1 to 13 stand for years 1999 to 2011, period 1-13 denotes parameters estimated on the whole dataset.

	Parameters			Summary		
Period	$\beta_C^{(d)}$	$\beta_C^{(w)}$	$\beta_C^{(m)}$	R^2	Log-l.	Obs.
1	0.180 (0.094) [0.055]	-0.006 (0.190) [0.976]	0.153 (0.359) [0.670]	0.050	-131.3	236
2	0.074 (0.081) [0.365]	0.142 (0.176) [0.419]	0.244 (0.254) [0.339]	0.055	-193.7	236
3	0.271 (0.084) [0.001]	0.171 (0.166) [0.302]	0.327 (0.193) [0.090]	0.188	-69.8	236
4	0.226 (0.082) [0.006]	0.153 (0.158) [0.334]	0.420 (0.176) [0.018]	0.226	-13.3	236
5	0.313 (0.093) [0.001]	0.410 (0.156) [0.009]	0.180 (0.154) [0.244]	0.437	-109.6	236
6	0.183 (0.136) [0.180]	0.184 (0.253) [0.469]	0.973 (0.294) [0.001]	0.322	-183.2	236
7	0.232 (0.083) [0.006]	0.080 (0.178) [0.654]	0.412 (0.228) [0.073]	0.137	-33.3	235
8	0.134 (0.103) [0.192]	0.516 (0.178) [0.004]	0.409 (0.173) [0.019]	0.401	-7.2	235
9	0.229 (0.086) [0.009]	0.131 (0.163) [0.423]	0.601 (0.208) [0.004]	0.359	-17.8	235
10	0.600 (0.073) [0.000]	0.075 (0.110) [0.494]	0.287 (0.099) [0.004]	0.738	-372.1	235
11	0.085 (0.107) [0.426]	0.472 (0.182) [0.010]	0.503 (0.149) [0.001]	0.610	-204.1	235
12	0.733 (0.069) [0.000]	0.061 (0.097) [0.529]	-0.012 (0.091) [0.898]	0.596	-176.6	235
13	0.589 (0.085) [0.000]	0.082 (0.124) [0.510]	0.148 (0.129) [0.252]	0.433	-181.6	235
1-13	0.479 (0.023) [0.000]	0.216 (0.036) [0.000]	0.275 (0.030) [0.000]	0.706	-2478.8	3061

Source: Author's computations.

Table A.2: Year-by-year estimates of continuous parameters for SP

Estimated parameters of daily (d), weekly (w) and monthly (m) continuous components from the non-logarithmic HAR-RV-CJ model are reported with standard errors in parentheses and p-values in square brackets. Periods 1 to 25 stand for years 1987 to 2011, period 1-25 denotes parameters estimated on the whole dataset.

	Parameters			Summary		
Period	$\beta_C^{(d)}$	$\beta_C^{(w)}$	$\beta_C^{(m)}$	R^2	Log-l.	Obs.
1	0.672 (0.063) [0.000]	-0.089 (0.102) [0.384]	0.126 (0.188) [0.504]	0.424	-358.8	249
2	0.082 (0.076) [0.280]	0.285 (0.167) [0.089]	0.263 (0.198) [0.184]	0.125	159.6	249
3	0.489 (0.069) [0.000]	-0.004 (0.122) [0.974]	0.059 (0.189) [0.754]	0.259	22.8	249
4	0.286 (0.084) [0.001]	0.246 (0.160) [0.126]	0.492 (0.227) [0.031]	0.232	22.6	249
5	0.287 (0.090) [0.002]	0.537 (0.173) [0.002]	-0.062 (0.273) [0.821]	0.243	193.7	249
6	0.237 (0.075) [0.002]	0.150 (0.146) [0.305]	0.207 (0.196) [0.291]	0.118	296.6	249
7	0.036 (0.076) [0.635]	0.315 (0.147) [0.033]	0.275 (0.205) [0.180]	0.124	335.0	249
8	0.606 (0.072) [0.000]	0.079 (0.130) [0.543]	0.086 (0.179) [0.633]	0.369	236.3	249
9	0.333 (0.074) [0.000]	0.123 (0.147) [0.402]	0.204 (0.260) [0.434]	0.170	312.0	249
10	0.152 (0.078) [0.053]	0.357 (0.157) [0.024]	0.081 (0.299) [0.786]	0.132	33.1	249
11	0.846 (0.092) [0.000]	-0.145 (0.188) [0.443]	-0.144 (0.237) [0.543]	0.342	-96.6	249
12	0.519 (0.075) [0.000]	0.202 (0.110) [0.067]	0.266 (0.126) [0.035]	0.557	-180.2	249
13	0.340 (0.074) [0.000]	0.137 (0.126) [0.278]	0.173 (0.179) [0.334]	0.209	-15.0	249
14	0.475 (0.078) [0.000]	-0.103 (0.143) [0.473]	0.362 (0.174) [0.038]	0.227	-261.2	249
15	0.186 (0.078) [0.018]	0.410 (0.139) [0.004]	0.249 (0.149) [0.096]	0.319	-155.9	249
16	0.652 (0.070) [0.000]	0.135 (0.100) [0.177]	0.083 (0.083) [0.317]	0.644	-194.2	249
17	0.175 (0.082) [0.033]	0.437 (0.129) [0.001]	0.180 (0.121) [0.139]	0.587	92.28	249
18	0.103 (0.079) [0.191]	0.724 (0.123) [0.000]	-0.167 (0.125) [0.181]	0.334	245.7	248
19	0.211 (0.076) [0.006]	0.623 (0.127) [0.000]	-0.088 (0.132) [0.503]	0.342	292.6	248
20	0.119 (0.079) [0.133]	0.496 (0.138) [0.000]	0.215 (0.144) [0.136]	0.333	222.9	248
21	0.427 (0.080) [0.000]	0.324 (0.129) [0.013]	0.129 (0.133) [0.333]	0.456	-53.9	248
22	0.114 (0.078) [0.146]	0.652 (0.126) [0.000]	0.137 (0.126) [0.277]	0.522	-584.3	248
23	0.365 (0.075) [0.000]	0.405 (0.112) [0.000]	0.204 (0.094) [0.031]	0.749	-147.9	248
24	0.582 (0.066) [0.000]	0.113 (0.117) [0.332]	0.375 (0.151) [0.013]	0.393	-168.2	248
25	0.742 (0.063) [0.000]	-0.056 (0.090) [0.536]	0.138 (0.105) [0.191]	0.601	-200.4	248
1-25	0.322 (0.015) [0.000]	0.396 (0.024) [0.000]	0.215 (0.022) [0.000]	0.571	-6197.0	6217

Source: Author's computations.

Table A.3: Year-by-year estimates of continuous parameters for CL

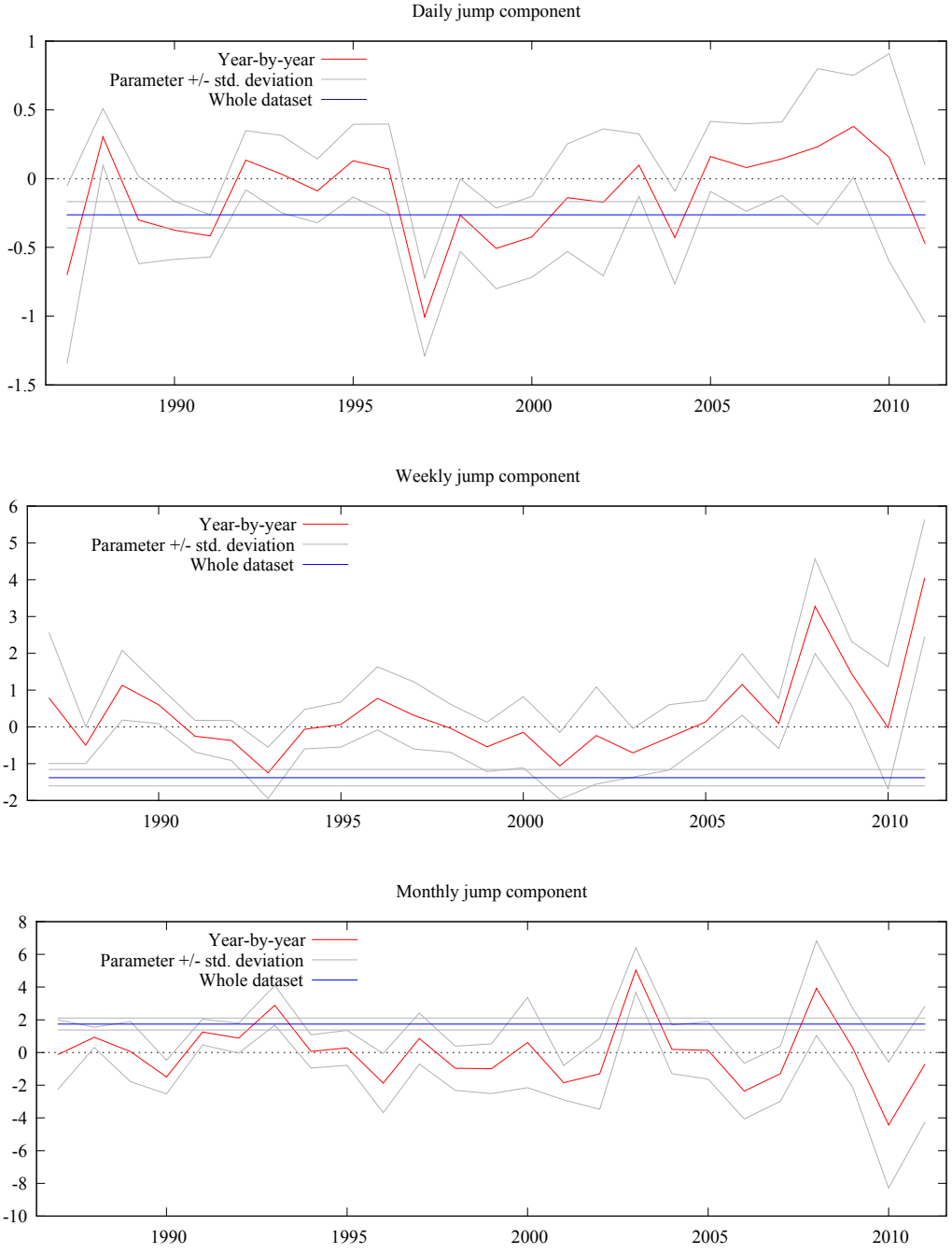
Estimated parameters of daily (d), weekly (w) and monthly (m) continuous components from the non-logarithmic HAR-RV-CJ model are reported with standard errors in parentheses and p-values in square brackets. Periods 1 to 25 stand for years 1987 to 2011, period 1-25 denotes parameters estimated on the whole dataset.

	Parameters			Summary		
Period	$\beta_C^{(d)}$	$\beta_C^{(w)}$	$\beta_C^{(m)}$	R^2	Log-l.	Obs.
1	0.394 (0.102) [0.000]	0.520 (0.163) [0.002]	-0.116 (0.166) [0.486]	0.334	-38.0	237
2	0.152 (0.145) [0.293]	0.661 (0.225) [0.004]	0.181 (0.231) [0.435]	0.199	-157.6	237
3	0.185 (0.090) [0.041]	0.847 (0.159) [0.000]	-0.195 (0.140) [0.164]	0.452	35.9	237
4	0.487 (0.082) [0.000]	-0.041 (0.136) [0.765]	0.412 (0.159) [0.010]	0.349	-389.2	237
5	0.521 (0.070) [0.000]	-0.028 (0.108) [0.799]	0.276 (0.087) [0.002]	0.490	65.7	237
6	0.264 (0.079) [0.001]	0.177 (0.155) [0.256]	0.287 (0.189) [0.132]	0.173	179.6	237
7	0.197 (0.083) [0.019]	0.394 (0.138) [0.005]	0.345 (0.134) [0.011]	0.484	38.0	237
8	0.264 (0.091) [0.004]	0.158 (0.167) [0.345]	0.169 (0.194) [0.382]	0.132	22.2	237
9	0.275 (0.078) [0.001]	-0.260 (0.152) [0.089]	0.885 (0.247) [0.000]	0.494	-123.6	237
10	0.140 (0.077) [0.072]	-0.109 (0.157) [0.488]	0.289 (0.199) [0.148]	0.054	-122.7	237
11	0.225 (0.091) [0.014]	0.488 (0.162) [0.003]	0.036 (0.212) [0.865]	0.213	-108.0	237
12	0.394 (0.074) [0.000]	0.266 (0.117) [0.024]	-0.006 (0.140) [0.966]	0.303	-227.9	237
13	0.307 (0.078) [0.000]	0.082 (0.156) [0.598]	0.301 (0.193) [0.120]	0.170	-67.0	237
14	0.287 (0.076) [0.000]	0.275 (0.127) [0.032]	-0.091 (0.206) [0.659]	0.170	-161.4	237
15	0.406 (0.090) [0.000]	0.252 (0.160) [0.117]	0.281 (0.162) [0.085]	0.411	-181.0	237
16	0.356 (0.085) [0.000]	0.382 (0.126) [0.003]	0.032 (0.138) [0.815]	0.429	-128.7	237
17	0.065 (0.085) [0.445]	0.336 (0.175) [0.057]	0.161 (0.241) [0.507]	0.079	-66.3	237
18	-0.053 (0.084) [0.528]	0.743 (0.172) [0.000]	0.085 (0.196) [0.664]	0.199	-151.6	237
19	0.068 (0.089) [0.445]	0.389 (0.176) [0.028]	0.142 (0.242) [0.559]	0.095	-141.7	237
20	0.423 (0.125) [0.001]	0.382 (0.189) [0.045]	0.041 (0.172) [0.810]	0.271	-96.8	237
21	0.223 (0.085) [0.009]	0.278 (0.159) [0.081]	0.241 (0.170) [0.158]	0.230	-26.6	237
22	0.017 (0.095) [0.860]	0.664 (0.229) [0.004]	0.477 (0.222) [0.032]	0.694	-443.9	237
23	0.177 (0.077) [0.023]	0.177 (0.163) [0.704]	0.598 (0.129) [0.000]	0.722	-263.2	237
24	0.542 (0.123) [0.000]	0.341 (0.170) [0.046]	0.068 (0.194) [0.727]	0.392	-131.2	237
25	0.574 (0.073) [0.000]	0.099 (0.107) [0.356]	-0.040 (0.131) [0.763]	0.384	-230.1	237
1-25	0.305 (0.017) [0.000]	0.348 (0.030) [0.000]	0.355 (0.027) [0.000]	0.654	-5272.3	5925

Source: Author's computations.

Figure A.1: Year-by-year jump parameters for SP

Dynamics of the daily, weekly and monthly parameters from year-by-year estimations compared to estimates from the whole dataset.



Source: Author's computations.

Table A.4: Year-by-year estimates of jump parameters for SP

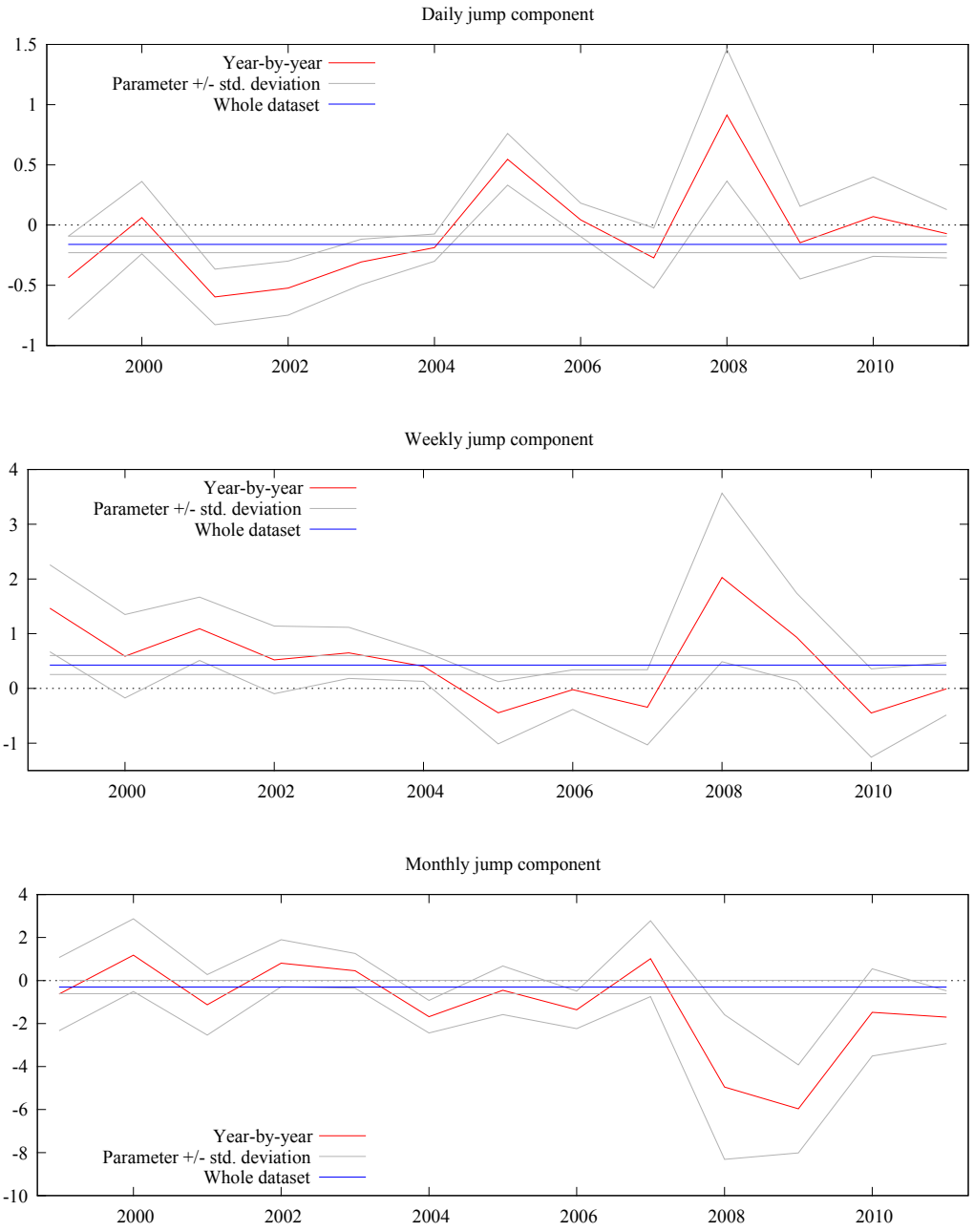
Estimated parameters of daily (d), weekly (w) and monthly (m) jump components from the non-logarithmic HAR-RV-CJ model are reported with standard errors in parentheses and p-values in square brackets. Periods 1 to 25 stand for years 1987 to 2011, period 1-25 denotes parameters estimated on the whole dataset.

	Parameters			Summary		
Period	$\beta_J^{(d)}$	$\beta_J^{(w)}$	$\beta_J^{(m)}$	R^2	Log-l.	Obs.
1	-0.697 (0.646) [0.281]	0.779 (1.774) [0.661]	-0.120 (2.115) [0.955]	0.424	-358.8	249
2	0.304 (0.206) [0.142]	-0.495 (0.501) [0.324]	0.928 (0.614) [0.132]	0.125	159.6	249
3	-0.301 (0.317) [0.343]	1.132 (0.946) [0.233]	0.060 (1.839) [0.974]	0.259	22.8	249
4	-0.375 (0.211) [0.077]	0.602 (0.513) [0.242]	-1.497 (1.029) [0.147]	0.232	22.6	249
5	-0.416 (0.154) [0.007]	-0.255 (0.432) [0.556]	1.252 (0.788) [0.113]	0.243	193.7	249
6	0.134 (0.215) [0.535]	-0.370 (0.543) [0.496]	0.878 (0.915) [0.338]	0.118	296.6	249
7	0.033 (0.281) [0.907]	-1.249 (0.700) [0.076]	2.887 (1.215) [0.018]	0.124	335.0	249
8	-0.089 (0.231) [0.700]	-0.062 (0.536) [0.908]	0.065 (1.018) [0.949]	0.369	236.3	249
9	0.131 (0.264) [0.622]	0.064 (0.612) [0.917]	0.281 (1.066) [0.792]	0.170	312.0	249
10	0.069 (0.328) [0.832]	0.775 (0.858) [0.367]	-1.871 (1.811) [0.302]	0.132	33.1	249
11	-1.006 (0.284) [0.001]	0.312 (0.913) [0.732]	0.852 (1.555) [0.584]	0.342	-96.6	249
12	-0.266 (0.264) [0.315]	-0.036 (0.653) [0.956]	-0.970 (1.352) [0.474]	0.557	-180.2	249
13	-0.507 (0.293) [0.085]	-0.542 (0.670) [0.419]	-0.989 (1.521) [0.516]	0.209	-15.0	249
14	-0.424 (0.293) [0.150]	-0.144 (0.967) [0.882]	0.608 (2.762) [0.826]	0.227	-261.	249
15	-0.138 (0.391) [0.725]	-1.060 (0.906) [0.243]	-1.845 (1.055) [0.082]	0.319	-155.9	249
16	-0.172 (0.535) [0.748]	-0.233 (1.319) [0.860]	-1.300 (2.167) [0.549]	0.644	-194.2	249
17	0.099 (0.227) [0.664]	-0.704 (0.663) [0.289]	5.049 (1.362) [0.000]	0.587	92.28	249
18	-0.428 (0.338) [0.207]	-0.285 (0.888) [0.749]	0.193 (1.492) [0.897]	0.334	245.7	248
19	0.161 (0.255) [0.527]	0.134 (0.582) [0.818]	0.133 (1.759) [0.940]	0.342	292.6	248
20	0.081 (0.318) [0.800]	1.153 (0.840) [0.171]	-2.360 (1.704) [0.167]	0.333	222.9	248
21	0.145 (0.267) [0.589]	0.095 (0.680) [0.889]	-1.297 (1.686) [0.442]	0.456	-53.9	248
22	0.233 (0.566) [0.682]	3.277 (1.285) [0.011]	3.930 (2.891) [0.175]	0.522	-584.3	248
23	0.380 (0.370) [0.307]	1.448 (0.872) [0.098]	0.302 (2.411) [0.901]	0.749	-147.9	248
24	0.154 (0.755) [0.838]	-0.024 (1.664) [0.989]	-4.434 (3.856) [0.251]	0.393	-168.2	248
25	-0.471 (0.574) [0.413]	4.033 (1.591) [0.012]	-0.737 (3.547) [0.836]	0.601	-200.4	248
1-25	-0.263 (0.095) [0.006]	-1.381 (0.225) [0.000]	1.740 (0.367) [0.000]	0.571	-6197.0	6217

Source: Author's computations.

Figure A.2: Year-by-year jump parameters for EC

Dynamics of the daily, weekly and monthly parameters from year-by-year estimations compared to estimates from the whole dataset.



Source: Author's computations.

Table A.5: Year-by-year estimates of jump parameters for EC

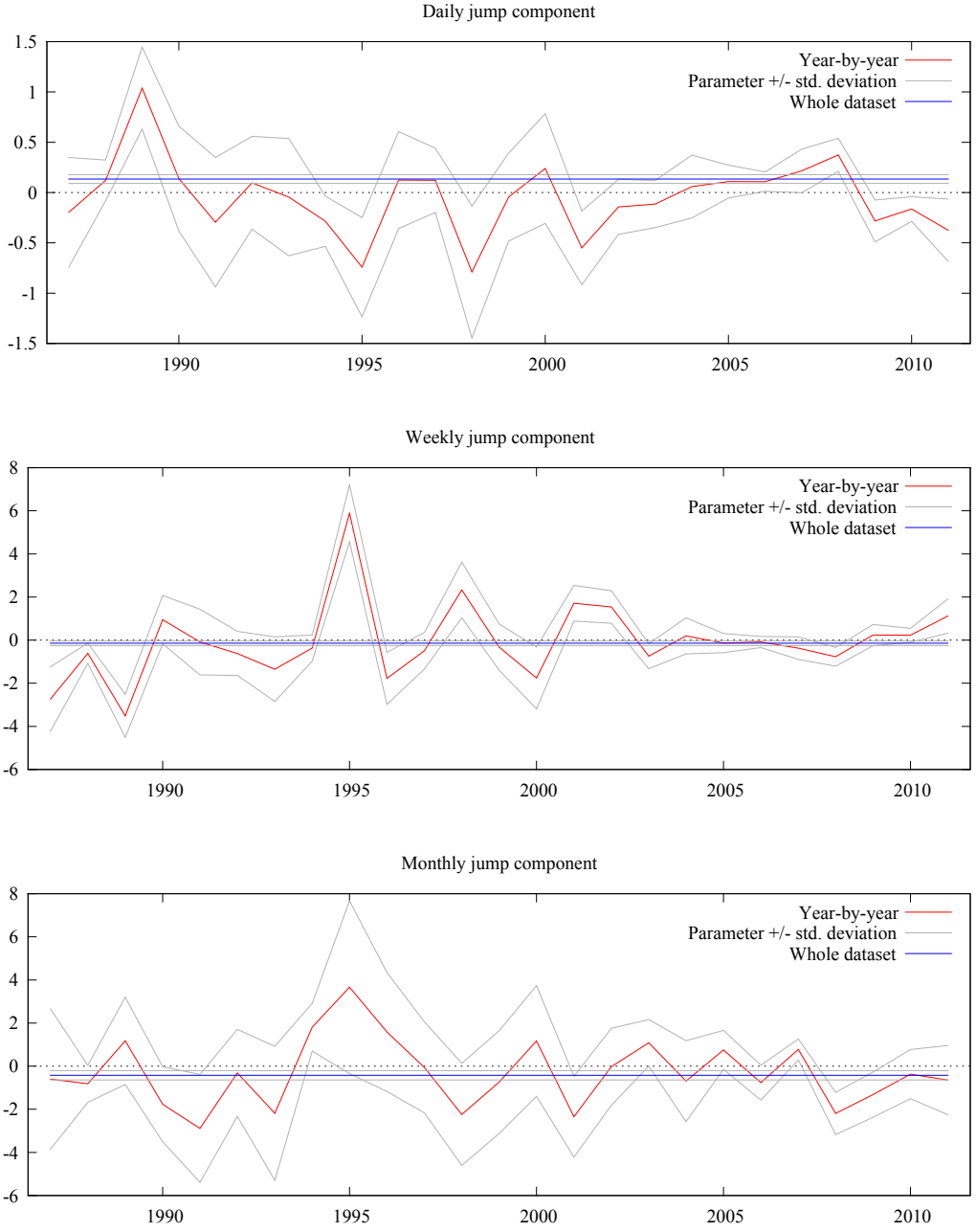
Estimated parameters of daily (d), weekly (w) and monthly (m) jump components from the non-logarithmic HAR-RV-CJ model are reported with standard errors in parentheses and p-values in square brackets. Periods 1 to 13 stand for years 1999 to 2011, period 1-13 denotes parameters estimated on the whole dataset.

	Parameters			Summary		
Period	$\beta_J^{(d)}$	$\beta_J^{(w)}$	$\beta_J^{(m)}$	R^2	Log-l.	Obs.
1	-0.435 (0.345) [0.208]	1.459 (0.793) [0.067]	-0.617 (1.701) [0.717]	0.050	-131.3	236
2	0.062 (0.300) [0.836]	0.587 (0.762) [0.442]	1.182 (1.691) [0.485]	0.055	-193.7	236
3	-0.597 (0.231) [0.011]	1.090 (0.579) [0.061]	-1.128 (1.408) [0.424]	0.188	-69.8	236
4	-0.524 (0.224) [0.021]	0.521 (0.618) [0.400]	0.804 (1.093) [0.463]	0.226	-13.3	236
5	-0.306 (0.189) [0.107]	0.650 (0.468) [0.166]	0.452 (0.807) [0.576]	0.437	-109.6	236
6	-0.187 (0.113) [0.100]	0.405 (0.277) [0.145]	-1.681 (0.754) [0.027]	0.322	-183.2	236
7	0.547 (0.215) [0.012]	-0.445 (0.568) [0.435]	-0.451 (1.129) [0.690]	0.137	-33.3	235
8	0.044 (0.139) [0.751]	-0.023 (0.362) [0.948]	-1.364 (0.872) [0.119]	0.401	-7.2	235
9	-0.273 (0.249) [0.275]	-0.344 (0.687) [0.617]	1.013 (1.763) [0.566]	0.359	-17.8	235
10	0.913 (0.548) [0.097]	2.027 (1.539) [0.189]	-4.951 (3.363) [0.142]	0.738	-372.1	235
11	-0.146 (0.301) [0.628]	0.934 (0.805) [0.247]	-5.967 (2.046) [0.004]	0.610	-204.1	235
12	0.070 (0.329) [0.832]	-0.449 (0.806) [0.579]	-1.481 (2.028) [0.466]	0.596	-176.6	235
13	-0.071 (0.201) [0.724]	-0.009 (0.480) [0.985]	-1.703 (1.231) [0.168]	0.433	-181.6	235
1-13	-0.160 (0.069) [0.020]	0.427 (0.173) [0.013]	-0.301 (0.311) [0.333]	0.706	-2478.8	3061

Source: Author's computations.

Figure A.3: Year-by-year jump parameters for CL

Dynamics of the daily, weekly and monthly parameters from year-by-year estimations compared to estimates from the whole dataset.



Source: Author's computations.

Table A.6: Year-by-year estimates of jump parameters for CL

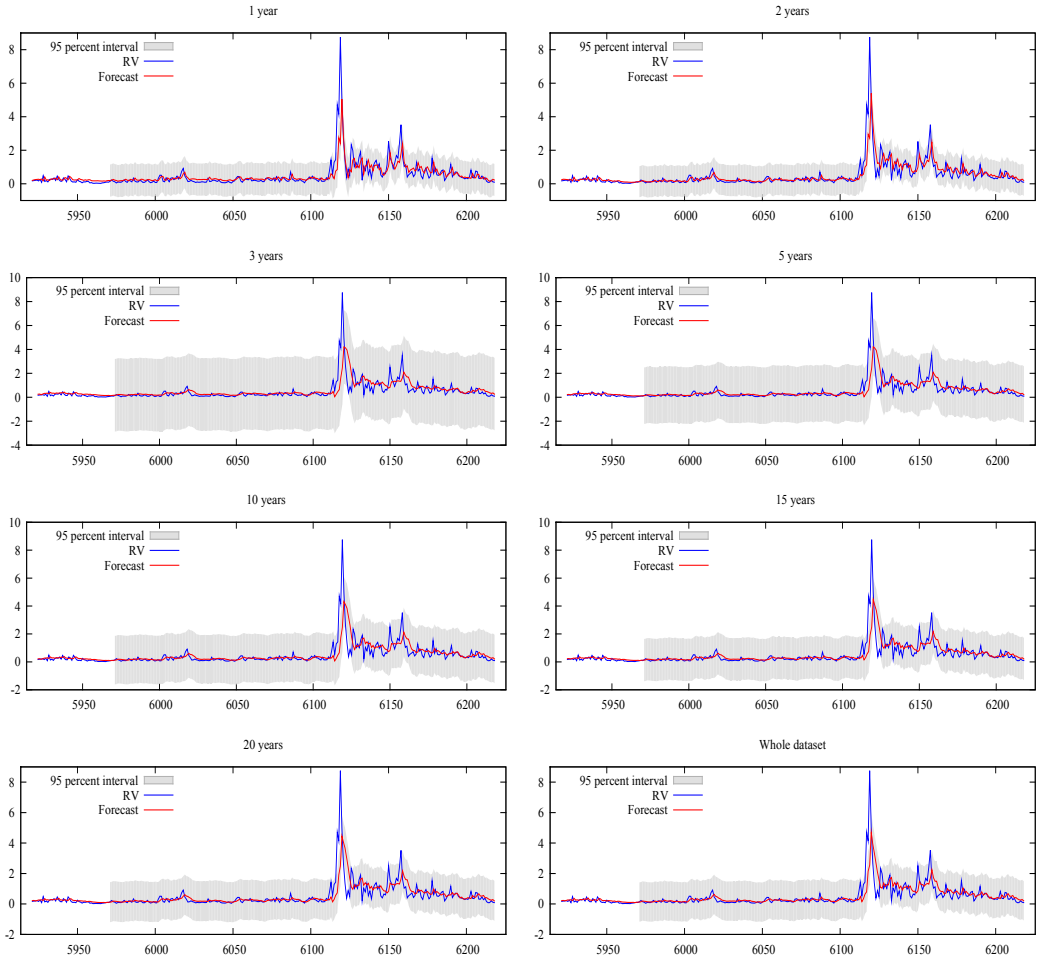
Estimated parameters of daily (d), weekly (w) and monthly (m) jump components from the non-logarithmic HAR-RV-CJ model are reported with standard errors in parentheses and p-values in square brackets. Periods 1 to 25 stand for years 1987 to 2011, period 1-25 denotes parameters estimated on the whole dataset.

	Parameters			Summary		
Period	$\beta_J^{(d)}$	$\beta_J^{(w)}$	$\beta_J^{(m)}$	R^2	Log-l.	Obs.
1	-0.196 (0.544) [0.719]	-2.736 (1.497) [0.069]	-0.597 (3.256) [0.855]	0.334	-38	237
2	0.120 (0.202) [0.553]	-0.612 (0.466) [0.191]	-0.817 (0.858) [0.342]	0.199	-157.6	237
3	1.039 (0.408) [0.012]	-3.512 (0.999) [0.001]	1.176 (2.023) [0.562]	0.452	35.9	237
4	0.140 (0.521) [0.789]	0.947 (1.123) [0.400]	-1.765 (1.739) [0.311]	0.349	-389.2	237
5	-0.295 (0.643) [0.648]	-0.087 (1.527) [0.955]	-2.886 (2.503) [0.25]	0.49	65.7	237
6	0.097 (0.461) [0.834]	-0.617 (1.023) [0.547]	-0.314 (2.019) [0.877]	0.173	179.6	237
7	-0.045 (0.583) [0.938]	-1.348 (1.500) [0.370]	-2.185 (3.109) [0.483]	0.484	38	237
8	-0.285 (0.250) [0.256]	-0.367 (0.604) [0.545]	1.810 (1.111) [0.105]	0.132	22.2	237
9	-0.742 (0.494) [0.134]	5.886 (1.322) [0.000]	3.664 (4.006) [0.361]	0.494	-123.6	237
10	0.124 (0.482) [0.797]	-1.777 (1.210) [0.143]	1.590 (2.763) [0.566]	0.054	-122.7	237
11	0.122 (0.319) [0.703]	-0.494 (0.845) [0.559]	-0.055 (2.110) [0.979]	0.213	-108	237
12	-0.790 (0.656) [0.230]	2.331 (1.289) [0.072]	-2.236 (2.365) [0.346]	0.303	-227.9	237
13	-0.045 (0.437) [0.918]	-0.324 (1.07) [0.763]	-0.737 (2.385) [0.758]	0.170	-67	237
14	0.239 (0.545) [0.662]	-1.752 (1.431) [0.222]	1.168 (2.566) [0.650]	0.170	-161.4	237
15	-0.551 (0.366) [0.134]	1.710 (0.831) [0.041]	-2.343 (1.874) [0.213]	0.411	-181	237
16	-0.142 (0.274) [0.604]	1.540 (0.756) [0.043]	-0.028 (1.788) [0.988]	0.429	-128.7	237
17	-0.115 (0.233) [0.623]	-0.743 (0.579) [0.200]	1.082 (1.077) [0.316]	0.079	-66.3	237
18	0.059 (0.313) [0.850]	0.198 (0.837) [0.813]	-0.694 (1.874) [0.711]	0.199	-151.6	237
19	0.109 (0.164) [0.506]	-0.136 (0.441) [0.758]	0.755 (0.902) [0.404]	0.095	-141.7	237
20	0.108 (0.096) [0.261]	-0.087 (0.258) [0.736]	-0.763 (0.81) [0.347]	0.271	-96.8	237
21	0.215 (0.216) [0.320]	-0.372 (0.520) [0.475]	0.777 (0.486) [0.111]	0.230	-26.6	237
22	0.374 (0.164) [0.024]	-0.773 (0.435) [0.077]	-2.191 (0.981) [0.026]	0.694	-443.9	237
23	-0.283 (0.209) [0.178]	0.234 (0.491) [0.635]	-1.321 (1.051) [0.210]	0.722	-263.2	237
24	-0.164 (0.121) [0.179]	0.228 (0.322) [0.480]	-0.369 (1.138) [0.746]	0.392	-131.2	237
25	-0.375 (0.311) [0.229]	1.116 (0.797) [0.163]	-0.646 (1.605) [0.688]	0.384	-230.1	237
1-25	0.134 (0.044) [0.002]	-0.145 (0.107) [0.174]	-0.423 (0.214) [0.048]	0.654	-5272.3	5925

Source: Author's computations.

Figure A.4: Non-logarithmic HAR-RV-CJ model forecasts for SP

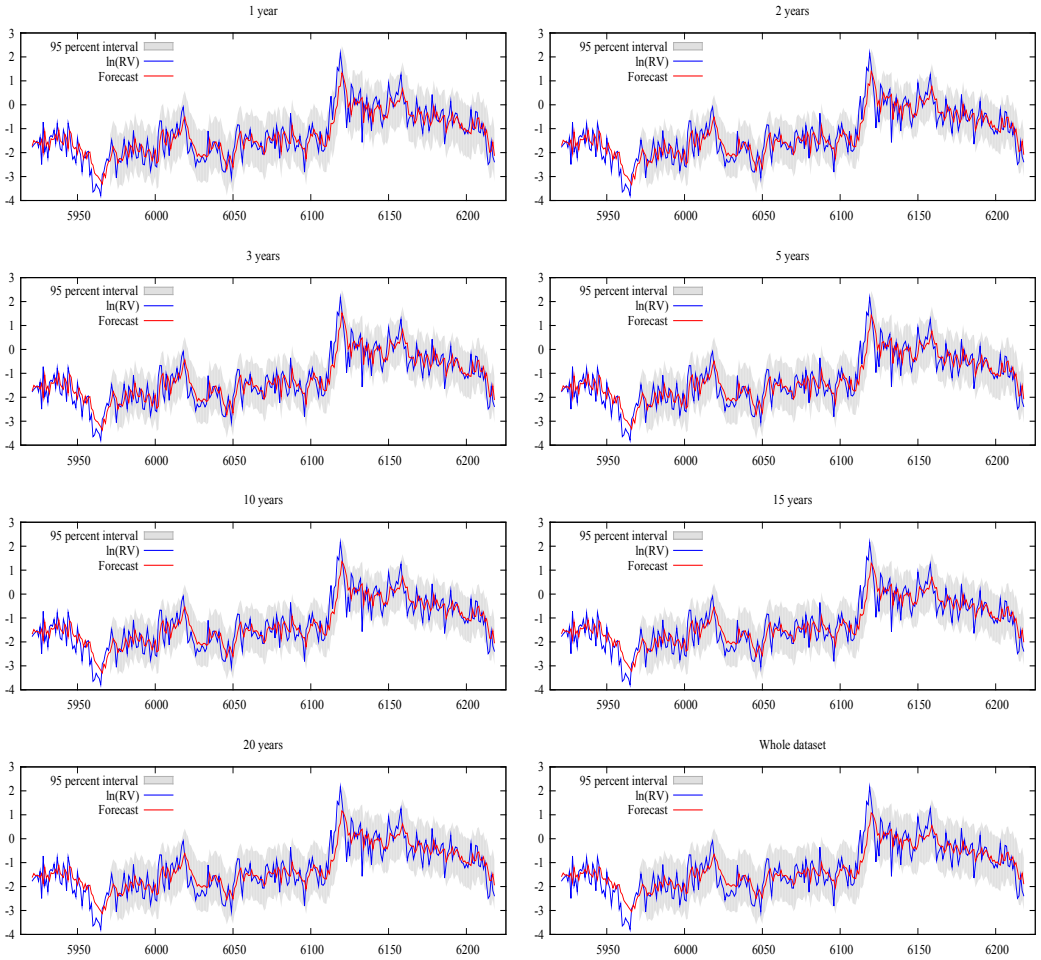
One-year out-of-sample forecasts of realized volatility based on pre-forecast periods of various lengths. The forecasted period is always the same, length of the pre-forecast period is indicated above each plot and varies from 1 year to 24 years (i.e. all available data).



Source: Author's computations.

Figure A.5: Logarithmic HAR-RV-CJ model forecasts for SP

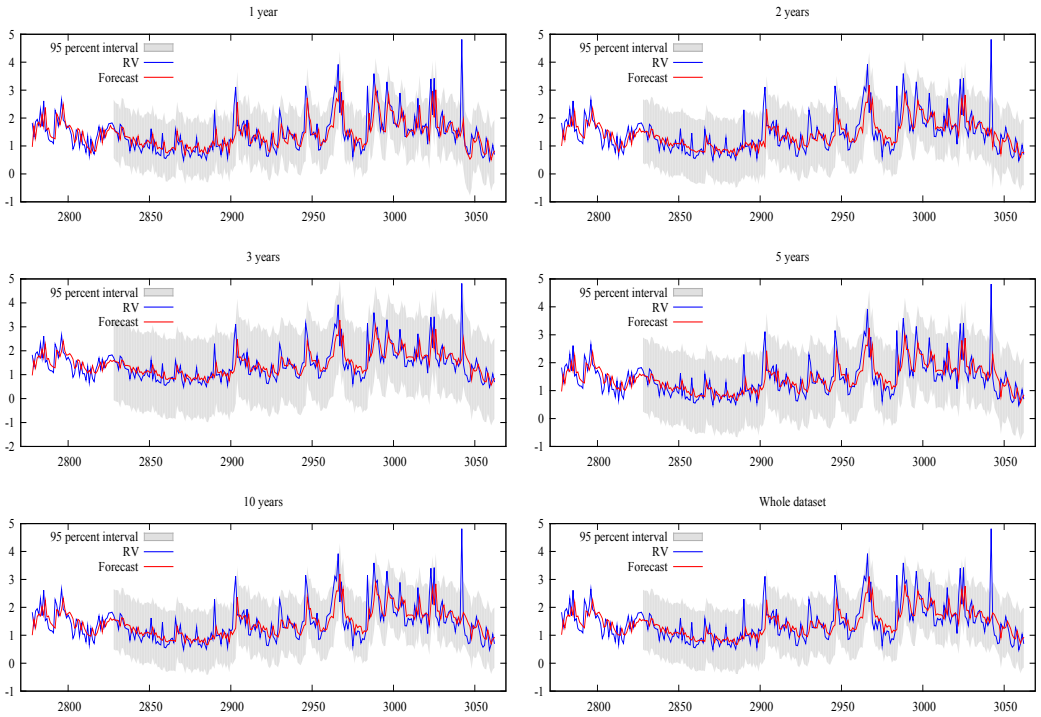
One-year out-of-sample forecasts of realized volatility based on pre-forecast periods of various lengths. The forecasted period is always the same, length of the pre-forecast period is indicated above each plot and varies from 1 year to 24 years (i.e. all available data).



Source: Author's computations.

Figure A.6: Non-logarithmic HAR-RV-CJ model forecasts for EC

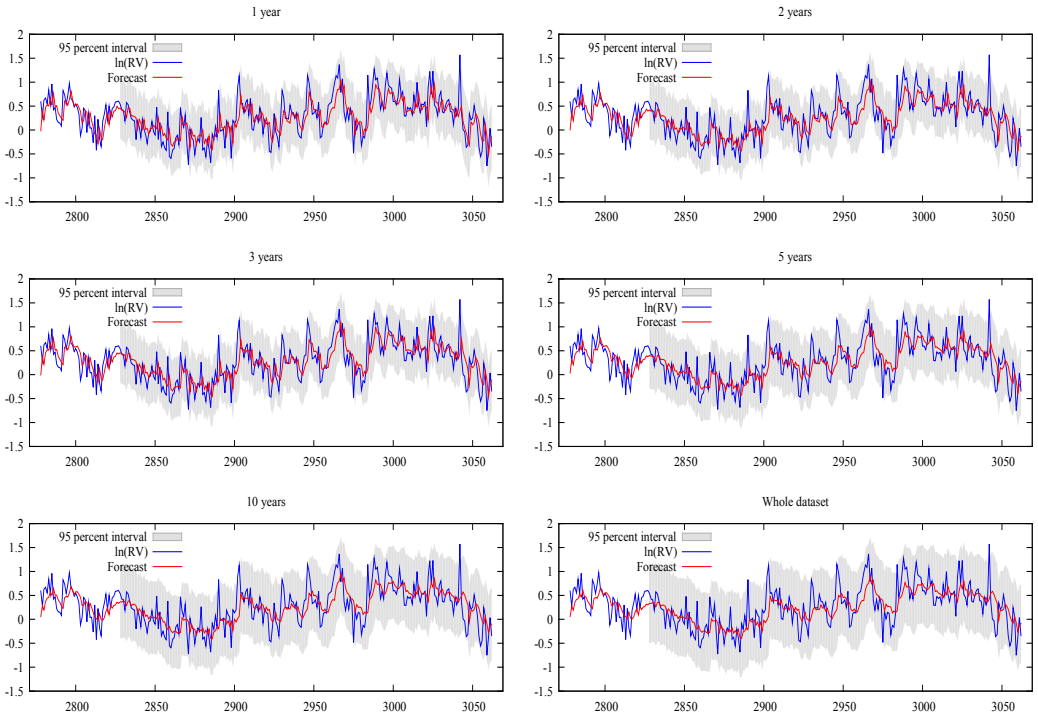
One-year out-of-sample forecasts of realized volatility based on pre-forecast periods of various lengths. The forecasted period is always the same, length of the pre-forecast period is indicated above each plot and varies from 1 year to 12 years (i.e. all available data).



Source: Author's computations.

Figure A.7: Logarithmic HAR-RV-CJ model forecasts for EC

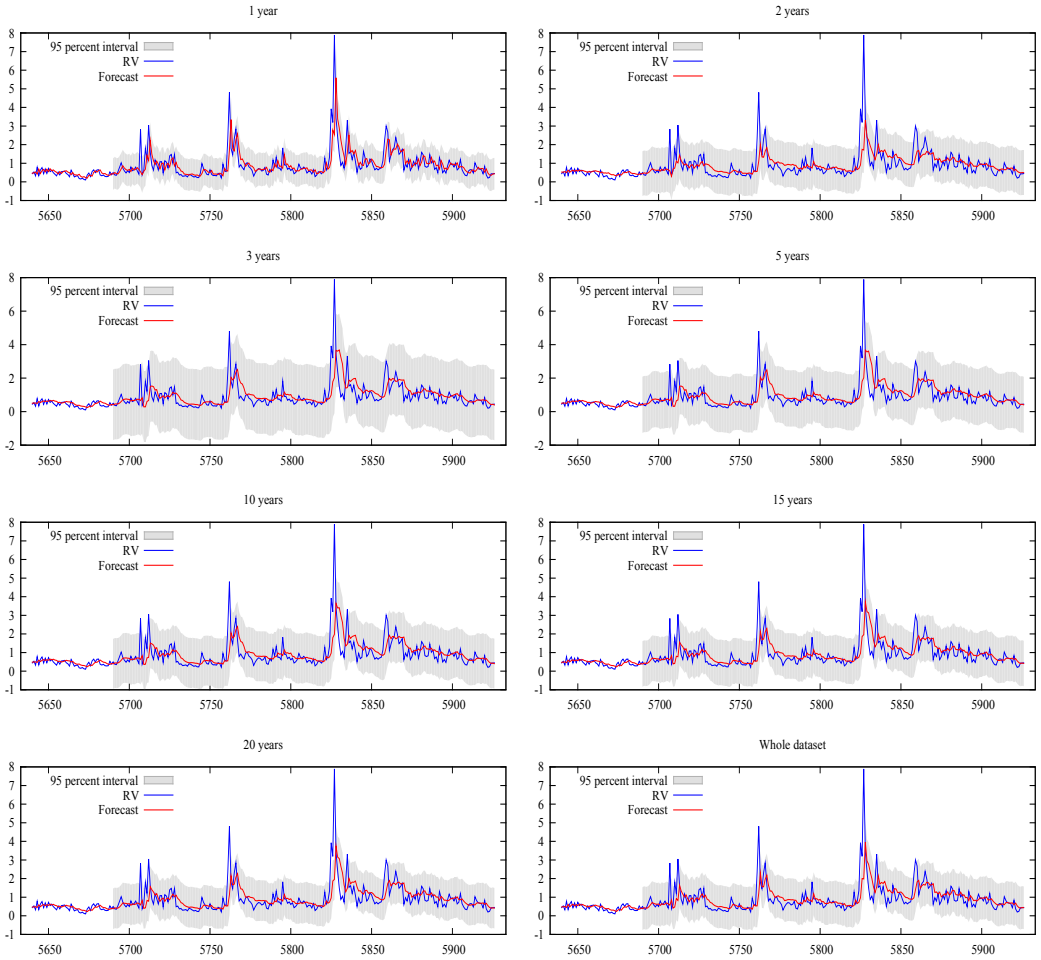
One-year out-of-sample forecasts of realized volatility based on pre-forecast periods of various lengths. The forecasted period is always the same, length of the pre-forecast period is indicated above each plot and varies from 1 year to 12 years (i.e. all available data).



Source: Author's computations.

Figure A.8: Non-logarithmic HAR-RV-CJ model forecasts for CL

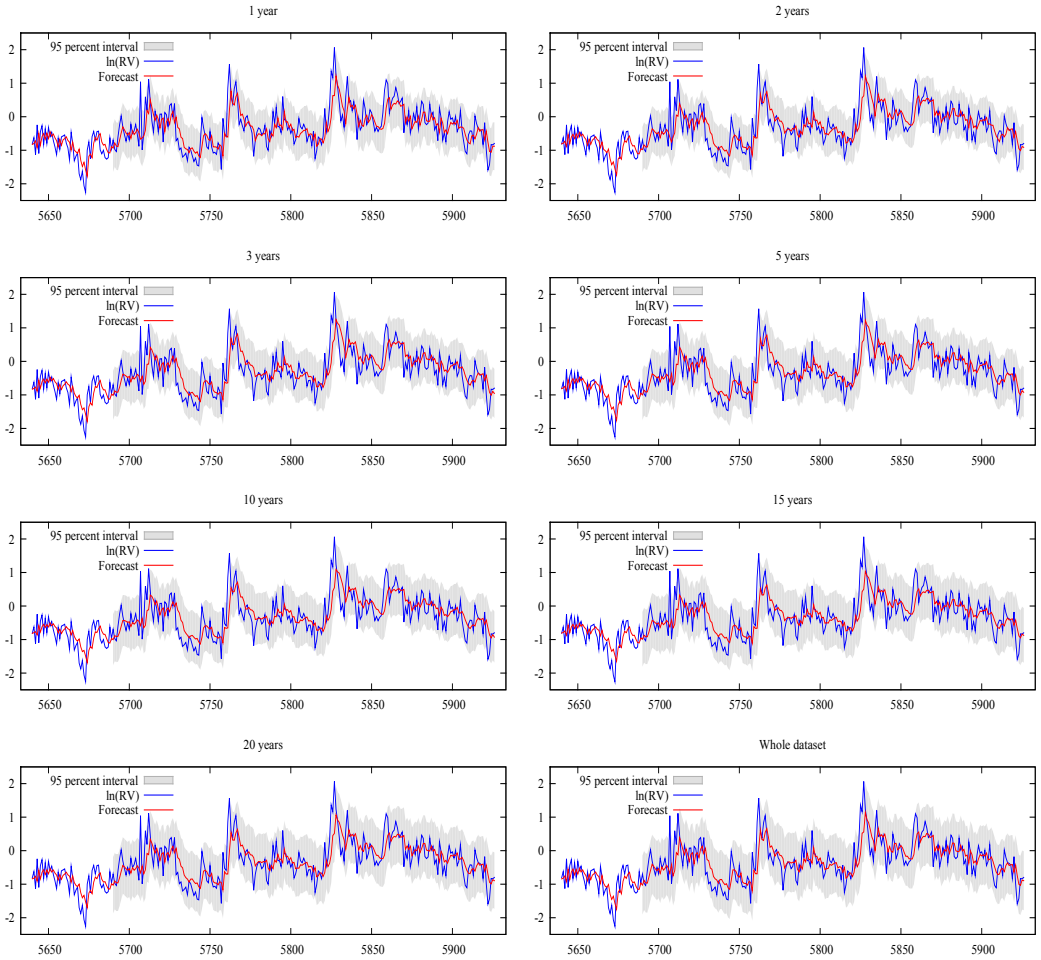
One-year out-of-sample forecasts of realized volatility based on pre-forecast periods of various lengths. The forecasted period is always the same, length of the pre-forecast period is indicated above each plot and varies from 1 year to 24 years (i.e. all available data).



Source: Author's computations.

Figure A.9: Logarithmic HAR-RV-CJ model forecasts for CL

One-year out-of-sample forecasts of realized volatility based on pre-forecast periods of various lengths. The forecasted period is always the same, length of the pre-forecast period is indicated above each plot and varies from 1 year to 24 years (i.e. all available data).



Source: Author's computations.