LUND UNIVERSITY
School of Economics and Management

Crude Volatility

Investigation of the HAR-RV model and implied volatility in Brent Crude futures

Bachelor Thesis in Financial Economics

Lund University, School of Economics & Management

Spring 2015

Author: Niklas Lindeke

Supervisor: Rikard Green

Abstract

Title: Crude Volatility – Investigation of the relationship between the HAR-RV model and Implied Volatility

Course: NEKH01, Bachelors Thesis in Financial Economics

Author: Niklas Lindeke

Contact: Lindeke.niklas@gmail.com

Supervisor: Rikard Green, PhD

Aim: Investigate the relationship and respective properties of the quantitative volatility model HAR-RV and the qualitative volatility estimation of option-implied volatility in the oil market with Brent Crude futures as proxy.

Methodology: The HAR-RV model is a three-step cascade or Realized Variance model with and intuition adjacent to a AR model. Parameters are optimized using OLS and Implied Volatility is derived through the Black Scholes Merton option pricing formula. The two volatilities are then compared using regression tests.

Theory: The HAR-RV model would outperform previously tested models such as GARCH-types and regular RV because of its long memory characteristics, and would be a more effective predictor of volatility. The relationship between IV and HAR-RV is therefore assumed to contain more information between the two.

Conclusion: The HAR-RV model yield superior results in comparison to previous research but proved (as seen in previous research) to be an inefficient and biased estimator of IV. Which is caused by the presence of a risk premium in the option pricing formula.

Table of Content

List of abbreviations

ARCH – Autoregressive Conditional Heteroscedasticity

BSM – Black, Scholes and Merton pricing model

DW – Durbin-Watson test

GARCH – Generalized Autoregressive Conditional Heteroscedasticity

HAR – Heterogeneous Autoregressive

HAC – Heteroscedasticity and Autocorrelation

HMH – Heterogeneous Market Hypothesis

ICE – InterContinental Exchange

IV – Implied Volatility

NW – Newey West estimator for autocorrelation

PCP – Put-Call-Parity

RMSE – Root Mean Squared Error

RV – Realized Variance

1. Introduction

Commodities would probably constitute the first traded good in history, and commodity derivatives can even be traced back as far as 4500 BC in ancient Sumer, where tokens for future delivery of a good was promised over a fixed price today (Banerjee (2013)). A few thousand years later the Chicago Board of Trade was founded as the world's first market for derivatives trading.

One of the most widely traded good amongst all today is oil. The oil trade is spread out on different levels of quality and geography, the three most famous oil futures being Brent, Dubai and WTI Crude – are working as the leading benchmarks for oil pricing today.

Movements in the oil price are of enormous significance for many actors on the market. Everything from freight, airline, and manufacturing companies to the LEGO Company (since oil is the main component in the production of plastic) is highly dependent on the oil price in managing their costs.

Today the oil market is in disarray because of recent events causing quite large price movements. After a few years of a stable pricing of the commodity, a group of events late 2014 caused the price to tumble.

Amongst these events is the high output of the American hydraulic fracturing shale industry, causing lower demand on foreign oil in the United States. A sustained output from the OPEC cartel and output from war torn countries such as Iraq and Libya, and also a statement of declining demand from the IEA as a reaction to a halting development of Chinese growth, all of this resulting in a major downturn of the oil price, dropping almost 60 % in just a couple of months (Craig (2015) and Raval (2014)).

The importance of this thesis lies within the fact that the world's oil dependence coincides with practically every industrial sector, and because of the fact of the very volatile price, research on how to handle volatility in the oil market is principal for a vast majority of market actors in the world.

In this thesis we aim to investigate the relationship between the Heterogeneous

Autoregressive Realized Variance (HAR-RV) model and Implied Volatility (IV) in Brent Crude futures. Recent advancements in volatility modelling have found a lot of promise of the HAR-RV model and its ability to mimic the properties of long memory and fat tails in financial data in a parsimonious manner, which is well established by the literature (See Corsi (2009)). However applications of this model in the commodity markets has been scarce, and this thesis will therefore aim to provide this perspective on the Brent Crude oil future, and compare its forecast to the IV of its options.

It is widely believed that IV holds informational advantage over any method of historical volatility estimation. This belief has been tested in previous literature by Canina and Figalewski (1993), Christensen and Prabhala (1997) and Bendi and Perron (2006). Where Canina and Figalewski (1993) and Christensen and Prabhala (1997) reached diametric conclusions regarding the informational content and efficiency, according to Christensen and Prabhala (1997), this was because of different research designs.
Bendi and Perron (2006) on the other hand, argue that the relationship is a fractional cointegration, and this is the reason to impose long memory models, where they apply narrow band spectral methods, such as FIGARCH and ARFIMA (See, Bendi and Perron (2006)).

In the recent decade, a lot of research in the field of historical volatility has been made, and a lot of work with high-frequency models has been presented. The framework suggested by Bollerslev et al (2003) on how Realized Variance significantly outperforms GARCH-type models, and the long memory extension presented by Corsi (2009) would indicate that this is a highly efficient model. But there are of course a lot of variants of this model that have been presented, such as the HAR-CV-JV model by Chan et al. 2008, which was supposed to separate jump and non-jump components (Haugom et al. (2011)). However Haugom et al. (2011) concludes in their paper that their mixed model HAR-CV-JV-EX is the optimal volatility-forecasting model for the energy market, but the differences in Root mean squared error (RMSE) was not significantly large enough for us to change the focus of this thesis.

The Heterogeneous Autoregressive Realized Variance (HAR-RV) model is still a highly efficient and parsimonious model with long memory behaviour and an intuition that is easy to evaluate. Because of this we would argue for the strength in applying the long memory HAR-RV model as proposed by Corsi (2009), and look at the relationship with the Implied Volatility as a necessary next step in the light of previous research on the subject.

Chung, Sun, Shih (2008) applied a HAR model to index options and compared it with IV, which proved significantly higher values of fit. Hence applying this model in the oil market, by testing it on Brent Crude Futures, and investigating its qualities in forecasting. By comparing it to the volatility implied by the Black, Scholes Merton (BSM) pricing model – we will not only have a unique contribution to the field of volatility theory in the oil market, but also maybe find some answers or clues on how the volatility structure in the oil market works.

Firstly this thesis will cover the theoretical and statistical framework behind the HAR-RV model, option theory and of course implied volatility. Followed by our empirical application and estimations of our quantitative model, and subsequent tests for evaluating the strength and forecasting power. Lastly we will test the models relation and theoretical implication to the more qualitative volatility model of Implied Volatility.

We found that there is an extremely high fit between the HAR-RV model and its corresponding IV, but the tests show that it suffers from what Chernov (2008) calls the "unbiasedness puzzle", which caused by the presence of volatility premia in the BSM pricing model. We found that there is an issue when applying the HAR-RV model as proposed by Corsi (2009), because it assumes some properties suggested by Müller et al. (2003) which are not on par with the properties of oil market.

2. Theoretical Framework and methodology

Following chapter will cover the methodology on which this thesis is based upon. The first part will cover some basics within the field of finance and time series analysis, where I later on will proceed with explaining the model in further detail.

2.1. Returns

When analysing financial time series the assumptions are based on statistical information apprehended via statistical methods from the price. But the price in itself has a stochastic nature and a non-constant mean. To overcome this, Ruppert (2010) suggests that one should look at the returns, which reflects the revenue-stream relative to the size of itself, and is hence a fraction of the price. Log returns (or continuously compounded returns) are approximately equal to normal price returns, but holds significant benefits in simplicity in multi-period returns (Ruppert (2010)). Log returns are defined as the following.

$$r = \log\left(\frac{P_t}{P_{t-1}}\right), \qquad t = 1 \dots T \tag{2.1}$$

Since we aim to model high frequency data, this thesis will primarily use this formulation of returns.

2.2. Time series

A time series is a sequence of values homogenously defined over time, and a stochastic process is a type of time series made up of random variables. When you look at a stochastic process almost all movements and fluctuations seem to happen at random, but it often follows the same stochastic behaviour, and by doing so they often appear to have a certain pattern in their mean, standard deviation, and correlation. This would imply that there are methods to approximate the behaviour of these processes by pinpointing the behaviour in these probability properties.

One of the most common features mentioned in time series analysis and econometrics is the phenomenon of stationarity. A process is called stationary if some aspects of its behaviour remain unchanged over time. Specifically, a process is weakly stationary if:

$$E(Y_i) = \mu \text{ , } for\ all\ i \tag{2.2}$$
$$Var(Y_i) = \sigma^2 \text{ , } for\ all\ i$$
$$Corr(Y_i, Y_j) = \rho\ (|\text{i} - j|) \text{ , } for\ i\ and\ j\ for\ some\ function\ \rho(h)$$

Another assumed property in time series is white noise, which is the simplest example of a stationary process (Ruppert (2010)). In modelling, it is used as a proxy for the residual noise and has function of being the effects that we cannot observe or predict.
White noise in itself has a few characteristics that are important to keep in mind. A sequence is considered a weak white noise process if it holds the two first characteristics of (2.2) but the correlation is equal to zero.

The most simple stationary process in time series is known as the Autoregressive model with one lag or AR (1) in short, or the more general formulation AR(P) for an undefined amount of lags, which is defined in equation (2.3).

$$X_t = w + \sum_{i=1}^{P} \varphi_i X_{t-i} + \varepsilon_t \tag{2.3}$$

An issue when utilising the autoregressive model in volatility, is that it will assume an unconditional variance, hence constant, which we know is not the case in neither nature nor financial data. To account for these effects the model called Autoregressive Conditional Heteroscedasticity (ARCH) was proposed (Ruppert (2012)).

The intuition behind ARCH is basically that for a model to be able to account for heteroscedasticity or volatility clustering, it has to have a component of a variance conditional to past information, so that it can account for that supposed volatility clustering.

Because of their theoretical appeal, volatility modelling has for a long time put a lot of focus on ARCH and GARCH-type models. But these models have of course also encountered a couple of issues. Among these is that the decay rate is too high, resulting in an exponential reversion to the unconditional variance. However, one solution for this is called fractional differentiation, where the exponent will be a non-integer value to account for the polynomial decay of the autocorrelation as found in financial time series (See, Ruppert 2012). However realized variance has not seen as much issues with this because of it being derived from quadratic variation and is therefore realized and not latent.

2.3. Option pricing

The option of options is of paramount importance for risk managers all over the world. Early on, the pricing of these had been based on a system for estimating the value of the option over a binary system of possible outcomes, discounting the effects into an end-price.

However the most commonly used model for pricing an option is the so-called Black-Scholes-Merton formula (BSM). Originally published in 1973 by Myron Scholes and Fischer Black and later extended by Robert Merton. Previous researchers had attempted this but failed in calculating the proper discount rate. Black and Scholes used a Capital Asset Pricing Model (CAPM) to determine the relation between the value of the stock and its option, which of course was not an easy task since it is dependent on both time and price (Hull (2012)).
The solution they begat was the Black-Scholes-Merton model and became the justification for option pricing to become a respected academic pursuit.

The formula is based on a differential equation with a lot of assumptions such as a log normal distribution and a risk neutral valuation (Hull (2012)). But it has been proven over the years to stand quite strong, but additions and variations have been made depending on what kind of asset the option will be derived from.

The BSM formula for estimating a standard European call option is presented below in the three equations represented in (2.4).

$$C = N(d_1)S - N(d_2)Ke^{-r(T-t)} \tag{2.4}$$

$$d_1 = \frac{1}{\sigma\sqrt{T-t}}\left[\ln\left(\frac{S}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)\right]$$

$$d_2 = d_1 - \sigma\sqrt{T-t}$$

Let C denote the price of the call option, S is the price of the underlying asset, K is the strike price for the option, T - t is the time to maturity, r stands for the risk free rate and σ^2

is the unconditional standard deviation.

A European put option is similarly valuated through the following formula:

$$P = N(-d_2)Ke^{-r(T-t)} - N(-d_1)S \qquad (2.5)$$

A parameter in the Black-Scholes-Merton formula that cannot directly be observed is a phenomenon called implied volatility; this is the volatility that the market assumes about the future affecting the underlying asset. These values cannot be estimated from any analysis of historical values, but they can be observed directly in the option price (Hull 2012). Theory also indicates that implied volatility involves to some extent an amount of premium for the risk that the investor exposes herself to.

This implies that the implied volatility holds some information about the market volatility that is accumulated from the assumptions of all of the involved market actors. This is the reason as to why it is considered to be the foremost method of measuring volatility for some specific asset (Christensen and Prabhala (1997)), and also one of the reasons for the invention of the VIX volatility index.

There is a lot of different ways to derive the implied volatility (IV), and this causes some issues, and as explained by Bendi and Perron (2001) and Christensen and Prabhala (1997), poses a central issue when researching the relationship between IV and other methods for estimating volatility.

However to illustrate, a simplified model for approximating IV of an option is given by the Bharadia, Christopher and Salkin model and is presented in equation (2.6).

$$\sigma = \sqrt{\frac{2\pi}{T}} \left[\frac{(c - \delta)}{(P - \delta)}\right] \qquad (2.6)$$

Where in (2.6), c is the price of the call, S is the price of the underlying asset and gamma is (S-X)/2 where X is the discounted strike price (Chambers et al. (2001)). But this is just one out of many examples on how to approximate Implied Volatility.

Implied volatility is different for different levels of the strike price, resulting in a convex

curve often referred to as the Volatility Smile, and because of put-call-parity, the smile will be the same for both put and call options. This relates to the options so called *moneyness*, which is the difference between the price of the underlying asset and the strike price. The relation is either called to be at-, in-, or out-of-the-money, which denotes the price relation to be either equal to, in positive relation to, or in negative relation to the option price, which differs depending on if you are holding a put or a call option.
In this thesis we will primarily study at-the-money options (ATM), where the price of the option is the same as the price of underlying asset. To only be looking at ATM options reduces the effects of the smile and will function as a more comparable metric to our futures data.

2.4. The HAR-RV model

In 1980 Merton showed the idea of an estimation of variance by summing intra-day squared returns as an efficient approximation of daily variance (Cornish (2007)). French, Schwartz and Stambaugh (1987) showed how monthly volatility could be estimated by adding the squared daily returns of the corresponding month.

Based on assumptions of continuous time arbitrage free price processes, the proposed model is developed from the theory of quadratic variation, which proposes that under appropriate circumstances, realized volatility is a highly efficient estimator of the variance in the return. Quadratic variation in itself is a rather deep and convoluted concept (Shreve (2004)), so for simplification: Define a symmetric binomial random walk M, where the outcome is either 1 or -1, with an equal probability of p and (1 – p), then consider the squared sum of all j = 1 to T.
The quadratic variation is therefore defined as in equation (2.7).

$$[M, M]_T = \sum_{j=1}^{T} \left(M_j - M_{j-1}\right)^2 \tag{2.7}$$

By computing this step-by-step, squaring and summing said steps, we can see that $\left(M_j - M_{j-1}\right)^2 = 1$ regardless of $M_j - M_{j-1}$ being 1 or -1 (because of the square), for the entire series, and therefore coming to the conclusion that the quadratic variation $[M, M]_T$ is equal to the variance $Var(M_T)$ of the same process.

The difference is that $Var(M_T)$ is computed as an overall difference between the squared expected values and the expected values squared, meaning that it will take some probability distribution into account, which if the random walk is not symmetric, affects the $Var(M_T)$ (Shreve 2008).

The idea of quadratic variation thus proposes a framework where the volatility is based on, in contrast to an average of paths of standard variance calculations, a single path, realized over the sampled intraday return. Thus treating volatility as something observed rather than latent.

The idea of realized variance is based on these assumptions and Andersen et al. (2003) provides us with an explanation in further detail on how efficient an estimator of volatility the Realized volatility is, and moreover how it outperforms traditional GARCH-type models.

Equation (2.8) presents the model for estimating a daily Realized Variance (RV), where r is high-frequency intraday log-returns as described in equation (2.1).

$$RV_t^{(d)} = \sqrt{\sum_{j=0}^{M-1} r_{t-j}^2} \tag{2.8}$$

Based on the Heterogeneous Market Hypothesis (HMH) Corsi (2009) proposes the HAR-RV as a model that will utilise three realized volatility components in an autoregressive manner, which all represent some time dependent market component for the model. The following equations 2.9 and 2.10 consider the RV over the complementing horizons. They are quite simply the average of the daily RV, so for a weekly RV we simply extend the model as following:

$$RV_t^{(w)} = \frac{1}{5}\left(RV_t^{(d)} + RV_{t-1d}^{(d)} + \cdots + RV_{t-4d}^{(d)}\right) \tag{2.9}$$

And the same definition for monthly volatility but over 22 daily periods:

$$RV_t^{(m)} = \frac{1}{22}\left(RV_t^{(d)} + RV_{t-1d}^{(d)} + \cdots + RV_{t-21d}^{(d)}\right) \tag{2.10}$$

The added sum of these three volatilities can be regarded as an additive cascade of volatilities each representing different components of market volatility. From this, it gets an almost long memory AR type of character (with lags one, five and twenty-two), but not strictly (See Corsi 2009). The model that he proposes would so forth simply just add up these three volatility components as a consequence of what they represent as economic effects in terms of the HMH.

The hypothesis was proposed by Müller et al. (1995), after they had observed different fractal properties in high-frequency intra-day data, such as a scaling law and behaviour of absolute price changes. They came to the conclusion that we can observe patterns in a market based on trading strategies with varying trading frequencies, which depends on levels and structures of risk aversion with the investors and market actors. As an explanation they propose the Heterogeneous market hypothesis (HMH), which divides the market actors in three types, the long-term, the mid-term and the short-term investor, and characterises movements in the markets based on these three types of market actors. (See Müller et al. (1995)).

The HAR-RV model tries to unify the HMH with the framework of realized variance so that it captures the efficiency and jump properties of realized variance, with the long-memory properties that will be attained when accounting for monthly and weekly volatility.

The deduction as proposed by Corsi (2009) is derived through a recursive substitution of their expected values. The return process is given by the highest frequency component in the cascade, which would be the daily high-frequency returns.

$$r_t = \sigma_t^{(d)} \varepsilon_t \tag{2.11}$$

Where epsilon is normal white noise, and through recursive substitution, the expected value of the monthly volatility is directly added in the equation for weekly volatility and the expected value of weekly volatility is directly added in the equation for daily volatility, as in the following system of equations:

$$\begin{aligned}
&\text{i.} \quad \sigma_{t+1m}^{(m)} = c^{(m)} + \varphi^{(m)} RV_t^{(m)} + \omega_{t+1m}^{(m)} && (2.12)\\
&\text{ii.} \quad \sigma_{t+1w}^{(w)} = c^{(w)} + \varphi^{(w)} RV_t^{(w)} + \gamma^{(w)} E_t\left[\sigma_{t+1m}^{(m)}\right] + \omega_{t+1w}^{(w)} &&\\
&\text{iii.} \quad \sigma_{t+1d}^{(d)} = c^{(d)} + \varphi^{(d)} RV_t^{(d)} + \gamma^{(d)} E_t\left[\sigma_{t+1w}^{(w)}\right] + \omega_{t+1d}^{(d)} &&
\end{aligned}$$

By expanding the expected values and utilising straight forward recursive substitution, the volatility model will be given by a three step cascade and has a form of something similar to three AR processes (For further detail, see Corsi (2009)).

$$\sigma_{t+1}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1d}^{(d)} \tag{2.13}$$

Now given (2.13), all variables are directly observable and available in the data set. The parameters will be able to be estimated through a simple Ordinary least squares estimation (OLS). However, because of possible serial correlation, a Newey-West (NW) covariance correction will be applied, since effects of Covariance and autocorrelation must be considered in the estimation.

3. Estimation

The estimation of the HAR-RV model is according to Corsi (2009) best done with Ordinary Least Square (OLS) estimation accompanied by a correction of a Newey-West estimator for the covariance matrix.

The OLS method is a linear regression where the goal is to minimize the sum of the squared distance from the line (Verbeek (2012)). Where a multiple linear model is traditionally formulated as the inner relationship in (3.2) and the Least Square method estimates the values of the Beta according to (3.1).

$$S(\beta) \equiv \sum_{i=1}^{N}(y_i - x_i'\beta)^2 \tag{3.1}$$

This will result in some parameters that will describe the relationship between the different variables in the models right hand side and their ability to describe the left hand side. This is one of the most famous and widely used statistical methods and it is very useful because we can gather a lot of interesting information from it and the estimation of the parameters. For example, by seeing if the parameter is either positive or negative, we gather that, given its significance, the associated variable will negatively affect our left hand side for changes in that variable. There is also an ocean of different tests one can apply to see what the relationship between the variables themselves is and so forth. But most importantly in this case, the regression gives us information on how the process works and that is useful for us to make a prediction of future events.

3.1. Linear Prediction

Forecasting in general is an important part of evaluating a model's usefulness, and development of proper forecasting techniques are hence of great value. In statistics the forecast is often denoted by its prediction and predictor, as being the means and result of the forecast, and theory behind these vary a bit between stochastic calculus, time series analysis and regression analysis. But in this thesis we are going to focus on the latter. We will also primarily focus on In-Sample forecasting, which is a simple process where the parameters for the entire period are estimated and one-day front predictions are made.

The best prediction method is always the one who minimizes the expected squared error of E (Y | X), but this relationship is seldom a linear function and is subsequently difficult to compute. The solution to this problem is called Linear Prediction, and considers only the linear functions of X as possible predictions (Ruppert (2010)).
This will basically be an expectation of the relationship in equation (3.1) and the expectation of said regression results in the following Best Linear Prediction.

$$\mathbb{E}\left\{ Y_i - \left(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} \right) \right\}^2 \tag{3.2}$$

And to be able to derive a prediction of Y from this equation, we need to rewrite it in respects to Y and consequently minimizes the argument. The partial derivatives are set to zero and will constitute the solvable system of equation, which result in the forecasting equation when solved.

However, the HAR-RV model has more than just one explanatory parameter, and the derivation of the forecasting equation is therefore slightly more comprehensive. So with the help of the methodology of derivation as proposed in Ruppert (2010), we will look for an equation that satisfies the minimization of this relationship:

$$\mathbb{E}\left\{ Y_i - \left(\beta_0 + \beta_1 RV_{d,1} + \beta_2 RV_{w,1} + \beta_3 RV_{m,p} \right) \right\}^2 \tag{3.5}$$

After taking the partial derivatives yields the following system of equation:

i. $E(Y) = \beta_0 + \beta_1 E(RV_d) + \beta_2 E(RV_w) + \beta_3 E(RV_m)$ (3.6)

ii. $E(RV_d Y) = \beta_0 E(RV_D) + \beta_1 E(RV_d^2) + \beta_2 E(RV_d RV_w) + \beta_3 E(RV_d RV_m)$

iii. $E(RV_w Y) = \beta_0 E(RV_w) + \beta_1 E(RV_d RV_w) + \beta_2 E(RV_w^2) + \beta_3 E(RV_w RV_m)$

iv. $E(RV_m Y) = \beta_0 E(RV_m) + \beta_1 E(RV_d RV_m) + \beta_2 E(RV_w RV_m) + \beta_3 E(RV_m^2)$

Resulting in this equation for the best linear prediction:

$$\hat{y} = E(Y) + \frac{\sigma_{Y,RV_d}}{\sigma_{RV_d}}(RV_d - E(RV_d)) + \frac{\sigma_{Y,RV_w}}{\sigma_{RV_w}}(RV_w - E(RV_w)) + \frac{\sigma_{Y,RV_d}}{\sigma_{RV_d}}(RV_m - E(RV_m)) \quad (3.7)$$

Equation (3.7) will therefore be the relationship that describes our method for performing our in-sample forecasting function.

4. Data and Estimation

4.1. Futures

Sample data on the futures contract written on Brent Crude were retrieved from the Bloomberg Professional Terminal. The dataset consists of raw ticker data in one-month front contracts (i.e. February data for contract expiring March) as traded on the Intercontinental Exchange (ICE), which is considered the most liquid.

Liquidity falls considerably outside of general European trading hours. So in accordance with Dacorogna (2001), an interval between 0800 and 1855 GMT were chosen, yielding 132 five-minute observations per day.

Homogenous time series was generated through a VBA routine that chose last price for each relevant minute. Contract prices are set in dollars per barrel, which corresponds to 119.24 litres and are represented below in figure 1.

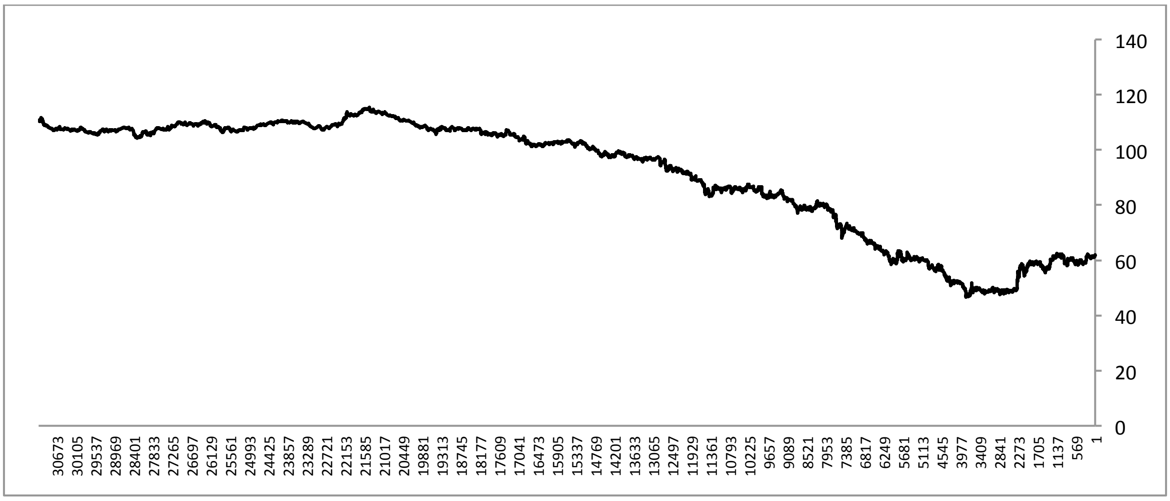


Fig. 1 - Shows the development of the price of the contracts during the chosen horizon, 1 being today, and 31206 is approximately one year before.

As we can see there is a major decline in price that started somewhere in second last half of 2014. This was a consequence of a reduced expectation of demand from Asian countries, mainly China, as well as a large rise in American inventory as a consequence of shale oil, resulting in a large reduction in international demand. This was also matched by a decision from the OPEC cartel to not reduce their output. The previous years the price had been quite stable at around 100 $/bbl., which mainly was because of the OPEC strategy of matching restraint in production every time some supply shock was introduced to the market. This is also one of the reasons for this sudden spike in crude volatility, because the market did not really expect Saudi Arabia to diverge from their previous path. But according to Saudi Arabia and the leaders of the OPEC cartel, this was a part of a strategy to gain back lost market share that was won by the rise of American shale oil and the sands in Canada (Raval (2014)).

As a consequence of this rapid and sizeable reduction in price, volatility followed. And of course, theory of variance and volatility says that variance is by its nature mean reverting, but because of the peculiar time frame the data set spans over, the said effect is not present.

4.2.Options

All data on options was retrieved from the ICE website. The implied volatility was derived from a Black and Scholes methodology with all available information present and these calculations where supplied by the ICE. As previous literature suggests only at-the-money options one month front contracts where considered in accordance to Bendi and Perron (2003). This was also the optimal choice since liquidity in said options is significantly higher than with the other relative periods, with the exception of the last few days before execution, but they trade with a generally higher volume.

4.3. Estimation

When estimating the daily, weekly and monthly volatility components a VBA-routine was used to systematically sum up the squared high-frequency five minute returns for each day, which resulted in 256 daily volatility observations. Secondly the sum of five of these daily volatility estimations divided by five was then calculated as the weekly volatility, and subsequently the monthly volatility was estimated by the average of the 22 lagging RV.

Below are the charts displaying these three volatilities estimated over the period march 2014 to march 2015 and also the chart for the price during the period for comparing the levels of volatility to the development in the price of the one-month front Brent Crude Futures.

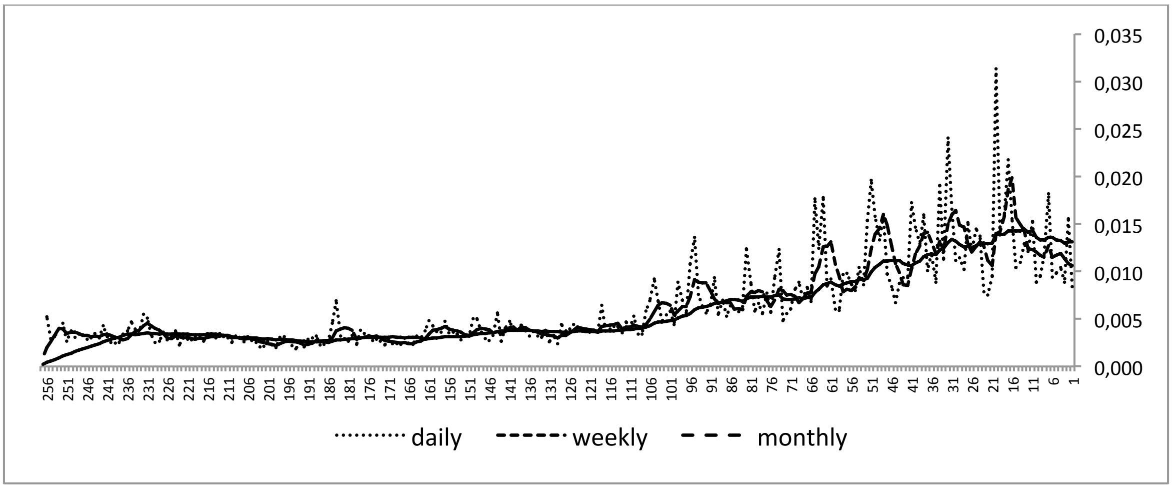


Fig. 2 – The three realized volatilities realized

In table 1 you will find the descriptive statistics of the brent futures contract, the log-returns of said contracts, the three measures of historical volatility and also Implied Volatility.

Table 1 – Descriptive Statistics

	Price	Log-return	$\sigma_{Implied}$	RV^d	RV^w	RV^m
N. Obs	31206	31206	256	256	256	256
Mean	91.2539	-0.008**	26.252	0.0063	0.0062	0.0058
Median	102,18	0	17.455	0.7719	0.2455	3.1439
Min	46.58	-0.0155	11.06	0.0018	0.00096	0.00021
Max	115.49	0.0246	71.66	0.031	0.019	0.014
Std. Dev	20.8661	0.0709*	15.0234	0.4613*	0.4070*	0.3822*
Skewness	-0.7996	0.9437	1.0673	1.8094	1.1045	0.9794
Kurtosis	2.1297	115.613	2.7908	4.0120	0.0911	-0.4125

*Multiplied with 100

**Multiplied with 1000

The estimation of the parameters where done by OLS estimation in accordance to the methodology proposed by Corsi (2009). Consequently the estimation is done with a correction for heteroscedasticity and auto-correlation. In Matlab an estimation of Newey-West covariance matrix was imposed to redeem possible auto-correlation.

Table 2 – HAR-RV (3) OLS Estimation MatLab

	β	NW	T-Test
Daily	0.2877	0.0982	2.9299
Weekly	-0.0197	0.2085	-0.0947
Monthly	0.7719	0.2455	3.1439
R = 0.677			

As shown in table 2 where t-NW shows the t-statistic for Newey-West estimates adjusted for heteroscedastic auto correlation. We can from this also see that the weekly volatility is not statistically significant. The level of fit in the model is very high.

To confirm these results a similar test was run in the Enthought Canopy environment for Python programming, and with the Statsmodels module for OLS estimation the beta parameters could be estimated. The model was also, like in Matlab, corrected for heteroscedasticity and auto-correlation using the robust covariance method. This reached the same results when done with the stats OLS module with Newey-West corrections in the Pandas toolkit.

Table 3 – HAR-RV (3) estimation of Beta parameters in Python

	β	p-value	SE
Intercept	0.0005	0.024	
Daily	0.2823	0.008	0.105
Weekly	-0.0486	0.803	0.195
Monthly	0.7450	0.002	0.234
R = 0.677	DW =	2.045	

The parameters differ slightly from the run in MatLab, which is strange but not too surprising. I will however choose to use the parameters attained from Python since we are a lot more confident in its accuracy. Only the monthly and daily parameters where significant at the 1% level, the intercept at a 5 % level, and alas the weekly not significant even at a 10 % level.

A Durbin Watson test (DW) was performed with the statistic of 2.045 as presented in table 3, rejecting the null hypothesis of there being any autocorrelation after our NW and HAC corrections. And lastly, there is a high fit with an adjusted R-value at 0,677 indicating that this model indeed is good.

The condition number test ran high over 30 indicating a high multicollinearity, which might be troublesome. Multicollinearity is the phenomenon where two or more variables in a multiple regression are highly correlated, which implies that one of these variables might not provide any useful information to the model. In the case of the OLS estimation, it is said that if there is perfect or too high multicollinearity, information about the beta is rendered useless. However our result is not too unexpected, which will be discussed further on in this thesis.

From this there are two options available for the model. Looking at chart (3), it is quite obvious that it is the weekly Realized Volatility that is causing this effect, and since it is also the only one that is not significant. A rerun of the tests will be done for a model containing only daily and monthly Realized Volatility.

So in accordance to equation 12, OLS estimation is done with the same corrections for auto-correlation in Python and Matlab.

Table 4 – HAR-RV (2) parameter estimation in Matlab

	β	NW	T-Test
Daily	0.2835	0.1018	2.7838
Monthly	0.7556	0.1154	6.5488

R = 0.677

From table 4, we can see that the daily Newey-West standard errors increases a little, but more significantly, the monthly standard errors decreases a lot and subsequently achieves a higher t-statistic, and as confirmed in the run in Python, a much higher significance.

Table 5 – HAR-RV (2) parameter estimation in Python

	β	p-value	SE
Intercept	0.0005	0.032	
Daily	0.2823	0.009	0.103
Monthly	0.7450	0.000	0.137
R = 0.678	DW =	2.038	

From table 5, in the Python OLS estimation with adjustments for heteroscedasticity and autocorrelation (HAC), we can see that the adjusted R even increases marginally when the weekly volatility variable is lost.

To lose the weekly component would alter the original idea behind the model as proposed by Corsi (2009). The components in themselves are nothing more than moving averages on weekly and monthly periods, with the objective to replicate the long memory behaviour that often is lost in standard GARCH and ARMA- type models. They are constructed as proxies for market effects based on lagged effects in the volatility structure, and specification of the lag structure in Corsi (2009) is designed for currency markets. Possibly, some other lag structure might be better suited for the oil market.

5.1. Implied volatility

As mentioned earlier, implied volatility is considered to be the most accurate estimation of the volatility of its underlying asset. To model a historical interpretation of volatility that would follow that process in an efficient manner, would be a high achievement.

After our tests we found that there is a very high level of fit (0.897) between our HAR-RV model and the IV. To illustrate why we find such a high fit between them we plotted IV against HAR-RV and the one-day ahead HAR-RV In-sample prediction.

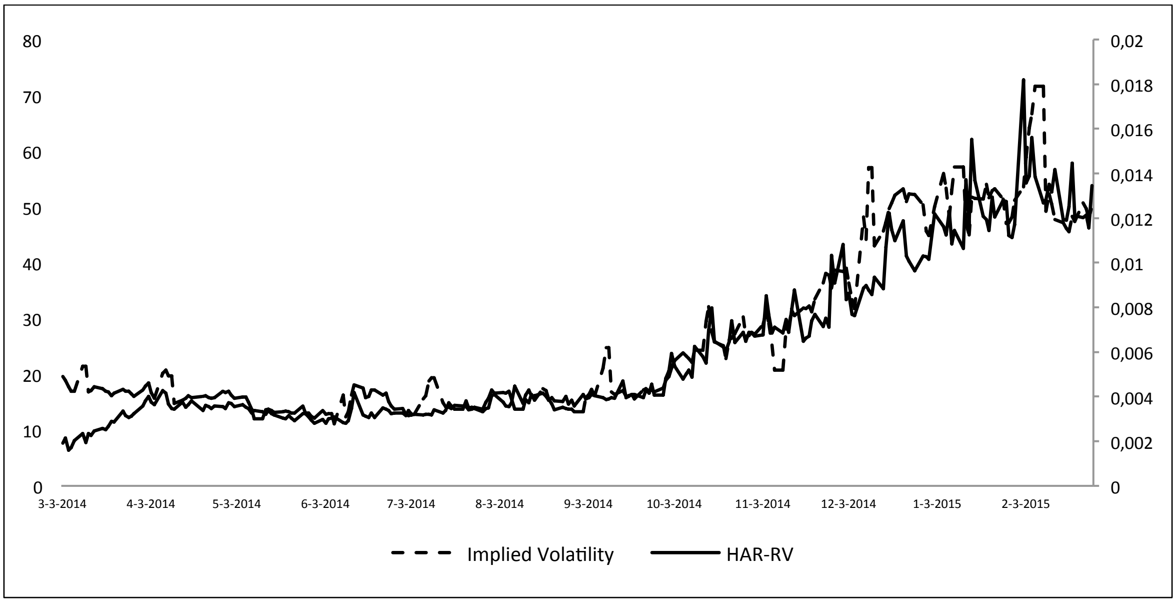


Fig. 3 – Implied Volatility plotted against HAR-RV

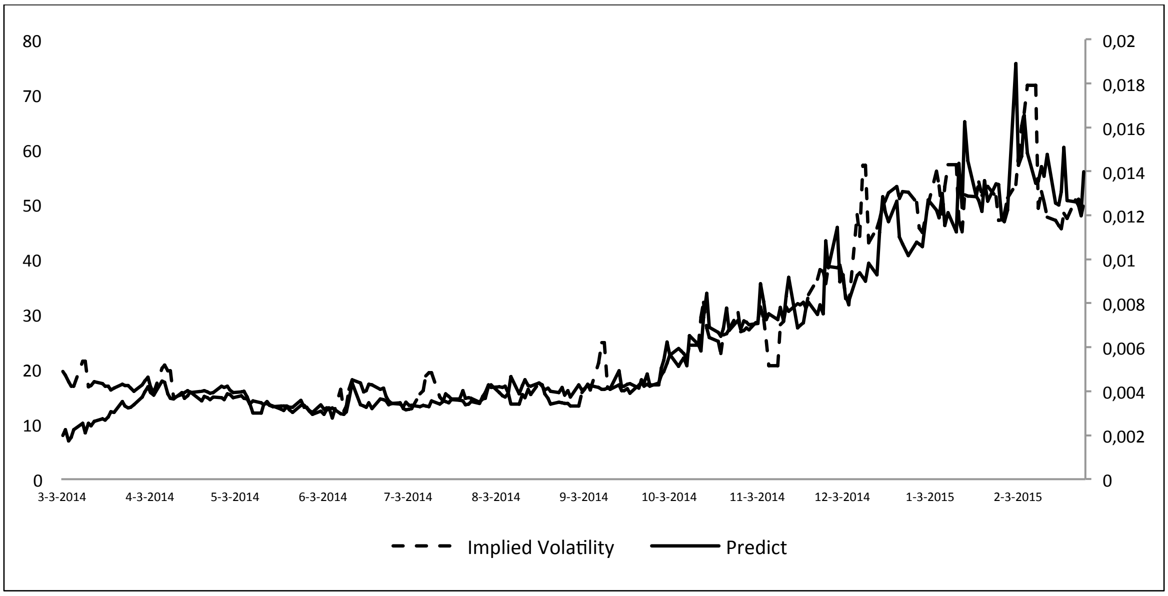


Fig. 4 – Implied volatility against in sample prediction

Seeing the two examples of variation next to each other, one could argue for deleting the first 22 days of the sample for estimation since the accumulation of the monthly RV is just an average of the last 22 observations and is therefore not at a suitable level during the first 22 observations, which might damage the results in the regression. We tried this, but there were no significant changes in the results in general after that.

5.2. Informational content

To evaluate the informational content of the relationship between IV and HAR-RV previous literature suggested a regression model to test their effects on each other, and three subsequent hypothesis tests for a more statistically rigid ground for our conclusions to stand on. The regression in itself is very simple and is formulated in equation (5.1).

$$h_t = \alpha_0 + \alpha_i i_t + e_t \tag{5.1}$$

Where h denotes the historically estimated volatility over time *t* and i denotes the implied volatility over time t. The three hypothesis tests as proposed by Christensen and Prabhala (1997) can be tested from this regression to see whether there is any informational content, bias and if the relation is efficient.

First test states that if implied volatility contains information about future volatility, the parameter for implied volatility should be nonzero. Second test states that if implied volatility is an unbiased predictor of realized volatility, the intercept and the implied volatility parameter should be zero and one respectively. And lastly, if implied volatility is efficient, residuals should be white noise (See, Christensen and Prabhala 1997).

The test of regression (5.1) was run in Python, and the resulting parameters are presented in table 6.

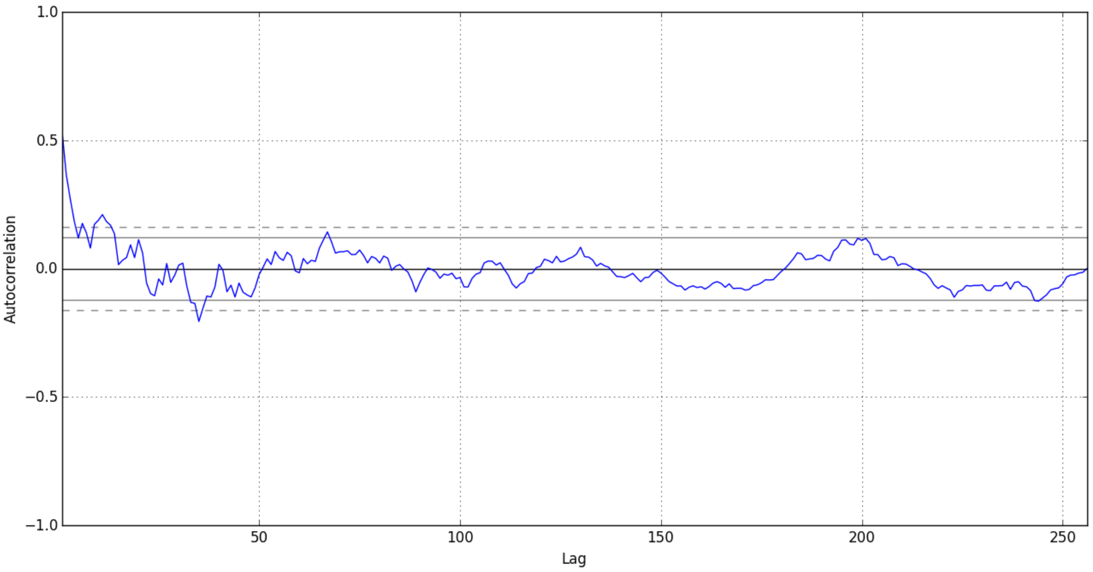
Table 6 – OLS Regression of equation (5.1)

	α	p-value	SE
Intercept	0.0000	0.9988	0.0002
Implied Vol	0.0002	0.0000	0.0000
R 0.897	RMSE	0.0012	

Our first test, which states that α_i should be non-zero, is accepted, meaning that implied volatility can say something about future volatility as formulated by the HAR-RV model. The information is however suggested to be biased since it is significantly less than unity. But on the other hand, the intercept is zero accompanied with a p-value that confirms it; hence a joint F-test is made.

Lastly, the tests on the residuals are made. If the residuals are white noise they are required to have a zero mean, a finite variance and no correlation to any of the variables or functions in the model. We should so forth be able to deduce whether or not the residuals are white noise by studying the plot of the residuals autocorrelation function.

Fig. 5 – Plot of the autocorrelation function of the residuals in regression (5.1)



However from this we realise that a visual analysis is not enough since it crosses the confidence intervals, which could also still just be random effects, but we need to be sure, so we apply a Ljung-Box test for a more formal approach.

The null-hypothesis states that the residuals are individually distributed and their correlation is zero, and a rejection of this hypothesis means that they have correlation.

We found that the residuals indeed are correlated in almost all of the lags; meaning that we can reject our third hypothesis of the IV RV relationship and finally conclude that there is information in Implied Volatility, however, this information is neither efficient nor unbiased.

The first test was accepted with a p-value indicating a very high significance. This means that implied volatility does contain information (potentially a lot) about the historically estimated volatility. The very high R^2 (0.897) confirms this by telling us that the movements of the two coincide a lot.

The second and third test where rejected meaning that implied volatility is not an unbiased or efficient forecaster of realized volatility. This has been an issue for previous literature as well; Chernov (2007) talks about this as the "unbiasedness puzzle" and explains this bias as a problem caused by the risk-premium (See, Chernov (2007)). Volatility derived from option prices contains expectations on future events in the market, and market participants' willingness and appetite to bear that consequential risk, which in it self is associated to the transaction costs of the subsequent dynamic hedging, which is reflected in a risk premium (Guo (1998)).

After testing the effects from the volatility implied in the options, we will test the effects from the volatility realized by the HAR-RV model. Equation 5.2, which is nothing more than the reverse of the previously researched relationship, should so forth test for those effects.

$$i_t = \alpha_0 + \alpha_h h_t + e_t \tag{5.2}$$

Table 6 – OLS Regression of equation (5.2)

	α	p-value	SE
Intercept	2.7030	0.0000	0.5847
HAR-RV	3730.01	0.0000	79.3108
R 0.897	RMSE	4.8312	

The first hypothesis is accepted since α_h is non-zero with a very small p-value, meaning that predicted HAR-RV volatility contains some information about implied volatility. The second test is rejected meaning that HAR-RV is a biased forecaster of implied volatility. Lastly, we again perform the Ljung-Box test, and reject the null hypothesis to conclude that HAR-RV is not an efficient estimator of Implied Volatility.

5.3. Discussion

In previous chapter all estimations where presented and a few interesting notes where taken, the two most interesting results where that the weekly parameter was not statistically significant (not even close) and seemed to cause some strong multicollinearity that we first thought could be quite troublesome.

The first case of the non-significant weekly effect can be interpreted in various ways, and the most obvious is of the absence of a weekly effect from the trading perspective in the crude oil future. If so, it renders the heterogeneous market hypothesis a lot less central to this theory, or at least, calls for a need of revision in specific markets, especially since the model proved such a high fit value even in the absence of that parameter and variable.

Regarding the multicollinearity, the issue might not be as grave as originally thought. The Realized volatility is derived from the array of price returns, and the weekly and monthly components are basically moving averages of the daily component. With that in mind, that presence of multicollinearity is not surprising at all. The issue disappeared when excluding the weekly component from the model, which is quite expected since the weekly and daily components are really similar in their correlation from just a visual point of view.

So because of this we would like to call for some fractal analysis of intraday oil futures in line with the framework proposed by Müller et al (2003) to estimate the corresponding intrinsic market components which would affect the oil market accordingly, because it is obviously not governed by the same principals as currency or equity markets, which where the test subjects for Corsi (2009) and Müller et al. (2003).

When we tested the effects between the volatility implied and the historical model we found some divergence from previous literature. Mainly the degree of fit was almost doubled from what previous literature concluded, even the most recent by Chung, Sun and Shih (2008) which also applied the HAR-RV model. One reason this for deviation might be because that previous authors had used a too wide horizon. Christensen et al. (1997) used a period of 12 years divided into two sub-periods for estimating, Chung, Sun and Shih (2008) uses a period of ten years in their estimation.

But why we believe this causes issues is because we do not believe in any market effect causing such long-term effects on the market that would justify such parameters. This conclusion, about yearly parameter estimation being optimal, also comes from the fact that the IEA, the leading authority on the subject of energy market forecasting, puts out annual global energy outlook reports once a year. These reports are one of the main factors that set the global perception of oil supply and oil demand throughout that year.

With that in mind one might argue for a model that would estimate its parameters based on other types of signals and time bound events, but with this general framework with new lags derived from a new fractal analysis, since it provides us with a very simple, intuitive and parsimonious model.

6. Conclusions

In this thesis we have investigated the relationship between the implied volatility and the historically estimated HAR-RV model on Brent Crude oil futures. Results have deviated slightly from previous research, in which we ascribe some incongruence in the suggested models component assumptions in accordance the oil market.

We used the ordinary least square estimation method with corrections for auto-correlation and heteroscedasticity with Newey-West. The usage of programming in this thesis was essential since high frequency modelling is very data heavy. The languages used were primarily Python, but some procedures and tests were also done in VBA and Matlab, and the code can be found in the appendix.

We performed regression tests to investigate the relationship between the realized and implied volatilities. Our results were mostly in line with previous literature, meaning that there is some informational content, but it is neither unbiased nor efficient. This is because of the issues regarding risk premium in the volatility implied in the options pricing model. Our volatility model however displayed a significantly higher level of fit in comparison to previous applications, which can be a consequence of the shorter horizon in the data in comparison to previous research. When estimating the model we obtained results that differed slightly from previous literature, the weekly component in the HAR-RV was not statistically significant, which we ascribe to be caused by the fact that the volatility lag structure of Brent Crude differs from the volatility lag structure in the researched assets in previous literature.

For future research, a new model for the oil market and methodology ought to be proposed, with a new lag structure derived by the fractal approach proposed by Müller et al. (2003) and then reapplied in a similar manner. But instead of estimating each parameter over several years as previous literature has done, the model should take into account the outlook reports submitted by the IEA and other agencies, and utilise some year-by-year re-estimation for the parameters, since the supply and demand structure changes in a similar manner.

7. References

Andersen, T.G. et al. (2003). *Modelling and Forecasting Realized Volatility.* Econometrica **71**(2), 579-625.

Bandi, F., and Perron B. (2006). *Long Memory and the Relation Between Implied and Realized Volatility.* Journal of Financial Econometrics **4**(4), 636-670.

Banjeree, J. (2013). *Origins of Growing Money. (Online).* Forbesindia.com/printcontent/34515.

Bollerslev, T., Gibson, M., Zhou, H. (2011). *Dynamic Estimation of Volatility Risk Premia and Inverstor Risk Aversion from Option-Implied and Realized Volatilities.* Journal of Econometrics **160**(1), 235-245.

Canina, L. Figlewski, S. (1993). *The Informational Content of Implied Volatility*. The Review of Financial Studies **6**(3), 659-681.

Chernov, M. (2007). *On the Role of Risk Premia in Volatility Forecasting*, Journal of Business & Economic Statistics, **25**(4), 411-426.

Christensen, B.J., and Prabhala, N.R. (1998). *The Relation between implied and Realized Volatility.* Journal of Financial Economics **50**(1998) 125-150.

Chung, H., Sun, E. Y., and Shih, K. C. (2008). *Do HAR and MIDAS Models Outperform Implied Volatility model? Evidence from Range-Based Realized Volatility.* Manuscript, Oversea Chinese Institute of Technology.

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

Cornish, R. (2007). *A Comparison of the Properties of Realized Variance for the FTSE 100 and FTSE 250 Equity indices.* Forecasting Volatility in the Financial Markets (A volume in Quantitative Finance). Third Edition. Elsevier Finance. 73-100.

Craig, S. (2014). *China plays big role in oil's slide.* (Online). Market Watch. Available: http://www.marketwatch.com/story/china-plays-big role-in-oilsslide-2014-11-30.

Dacorogna, M.M. et al. (2001*). An Introduction to High-Frequency Finance.* San Diego, CA: Academic Press.

Dowd, K. (2007). *Measuring Market Risk.* Second Edition. Wiley Finance

French, K. R., Schwert, G. W. and Stambaugh, R. F. (1986) *Expected Stock Returns and Volatility.* Journal of Financial Economics **19**, 3-29.

Guo, D. (1998). *The Risk Premium of Volatility implicit in Currency Options. In Computational Intelligence for Financial Engineering.* Proceedings of the IEEE/IAFE/INFORMS 1998 Conference on IEEE 1998, 224-251.

Haugom, E., Westgaard, S., Solibakke, P. B., and Lien, G. (2011). *Realized Volatility and the Influence of Measures on Predictability: Analysis of Nord Pool Forward Electricity Data.* Energy Economics **33**(6) 1206-1215

Hawley, P. (2015). *The big question is: What are the Saudis Intentions?* (Online). Financial Times. Available: http://www.ft.com/intl/cms/s/0/3fc35380-ad22-11e4 bfcf-00144feab7de.html.

Hull, J. C. (2012). *Options, Futures and Other Derivatives.* 8th edition. Pearson Global Edition.

Müller, U.A. et al. (2003). *Fractals and Intrinsic Time – A Challenge to Econometricians.* 39th International AEA Conference on Real Time Econometrics, 14- 15 October 1993, Luxembourg.

Raval, A. (2014). *Opec leader vows not to cut oil output even if price hits $20* (Online). Financial Times. Available: http://www.ft.com/intl/cms/s/0/3fc35380ad2211e4bfcf00144feab7de.html.

Raval, A. (2015). *Saudi Claims oil price strategy success,* (Online). Financial Times. Available: http://www.ft.com/intl/cms/s/2/69350a3e-f970-11e4-be7b 00144feab7de.html.

Ruppert, D. (2010). *Statistics and Data Analysis for Financial Engineering.* 1st Edition. Springer Texts in statistics

Shreve, S.E. (2004). *Stochastic Calculus for Finance II.*1st edition. Springer Finance Textbooks.

Verbeek, M. (2012). *A guide to Modern Econometrics.* 4th edition. Wiley.

8. Appendix

8.1. HAR-RV estimator and tester

```
# -*- coding: utf-8 -*-
from __future__ import print_function, division
import xlrd as xl
import xlwt as xlw
import numpy as np
import scipy.stats as ss
import scipy as sp
import pandas as pd
import statsmodels.formula.api as smf
from statsmodels.datasets.longley import load_pandas
import statsmodels.api as sm
import matplotlib.pyplot as plt
import itertools as it

file_loc = "/Users/NiklasLindeke/Python/dataset_3.xlsx"
workbook = xl.open_workbook(file_loc)
sheet = workbook.sheet_by_index(0)
tot = sheet.nrows

data = [[sheet.cell_value(r, c) for c in range(sheet.ncols)] for r in range(sheet.nrows)]

rv1 = []
rv5 = []
rv22 = []
rv1fcast = []
T = []
price = []
time = []
retnor = []

model = []

for i in range(1, tot):
        t = data[i][0]
        ret = data[i][1]
        ret5 = data[i][2]
        ret22 = data[i][3]
        ret1_1 = data[i][4]
        retn = data[i][5]
        #t = xl.xldate_as_tuple(t, 0)
        rv1.append(ret)
        rv5.append(ret5)
        rv22.append(ret22)
        rv1fcast.append(ret1_1)
        retnor.append(retn)
        T.append(t)

df = pd.DataFrame({'RVFCAST':rv1fcast, 'RV1':rv1, 'RV5':rv5, 'RV22':rv22,})
df = df[df.RVFCAST != ""]
df = df.astype(float)
Model = smf.ols(formula='RVFCAST ~ RV1 + RV5 + RV22', data = df).fit(use_correction=True)
mdl = Model.get_robustcov_results(cov_type='HAC', maxlags=1, use_correction=True)
#print(mdl.summary());
pdmdl = pd.stats.ols.OLS(y=df['RVFCAST'], x=df[['RV1', 'RV5', 'RV22']], nw_lags=1)
```

```
params = mdl.params
pred = pd.DataFrame(mdl.predict())
y = pd.DataFrame.to_csv(pred, 'prediction_insample.csv', sep=',')

#y = pd.DataFrame.to_csv(actual, '123.csv')
#fig, ax = plt.subplots(figsize=(12,8))
#fig = sm.graphics.plot_ccpr_grid(mdl, fig=fig)
#fig, ax1 = plt.subplots()
#ax1.plot(pred, 'r')
#acf = sm.tsa.stattools.acf(actual)
#ax2 = ax1.twinx()
#ax2.plot(retnor, 'b-')
#plt.plot(retnor, 'b-')
#plt.show()

#ax.plot(df.RV1, pred, 'o', label="Data")
#ax.plot(df.RV1, df.RVFCAST, 'b-', label="True")
#ax.plot(np.hstack((df.RV1, x1n)), np.hstack((pred)), 'r', label="OLS prediction")
#ax.legend(loc="best");

#norm_x = df.RV1.values
#for i, name in enumerate(df.RV1):
# if name == "const":
# continue
# norm_x[:,i] = X[name]/np.linalg.norm(X[name])
#norm_xtx = np.dot(norm_x.T,norm_x)

#fix, ax = plt.subplots(figsize=(12,14))
#fig = sm.graphics.plot_partregress("RVFCAST", "RV1", ["RV22"], data=df, ax=ax)
```

8.2. Regression tester IV-RV

```
i = genfromtxt('xxx.csv', delimiter=',')
rv = genfromtxt('yyy.csv', delimiter=',')
rw.columns = ['rw']
#rw = rw.rw.reindex(index=rw.index[::-1])
rw = rw.replace(np.nan,0, regex=True)


df = pd.DataFrame(rv, columns = ['rv'])
df2 = pd.DataFrame(i, columns = ['i'])
frame = [df, df2, rw]
df = pd.concat(frame, axis = 1)

Model = smf.ols(formula='rw ~ i', data = df).fit()
pdmdl = pd.stats.ols.OLS(y=df['rw'], x=df[['i']])

resid = Model.resid
kde = sm.nonparametric.KDEUnivariate(resid)
kkk = kde.fit()

fig, ax = plt.subplots(figsize=(8,6))
ax2 = ax.twinx()
x = np.linspace(0,256,len(rw))
ax.plot(x, i, 'r', label="Data")
ax2.plot(x, rw, 'b', label="Predicted")
ax2.plot(x, rv, 'y', label="Predicted")
legend = ax.legend(loc="best")

fig = plt.figure(figsize=(12,8))
ax = fig.add_subplot(111)
ax.hist(resid, bins=25, normed=True, color='red')
ax.plot(kde.support, kde.density, lw=2, color='grey')
```

```
acf = sm.tsa.acf(resid, 90)
test = sm.stats.diagnostic.acorr_ljungbox(resid)
print(Model.summary())
print(pdmdl)

x = resid
#rg = genfromtxt('sunspots/sp.dat')
#x = rg[:,1] # Just use number of sun spots, ignore year
h = 20 # Number of lags
lags = range(h)


h, pV, Q, cV = lbqtest(x, range(1, 10), alpha=0.1)
print 'lag p-value Q c-value rejectH0'
for i in range(len(h)):
print "%-2d %10.3f %10.3f %10.3f %s" % (i+1, pV[i], Q[i], cV[i], str(h[i]))
import pandas as pd
import numexpr
```

8.3. Optiondatacleaner

```
df = pd.DataFrame.from_csv('options2014.csv', header=0, sep=',')
df = df.query('RelativeStrike=="ATM"')
df = df.query('RelativePeriod=="M1"')
x = pd.DataFrame.to_csv(df, 'optionz14.csv', sep=',')

df2 = pd.DataFrame.from_csv('options2015.csv', header=0, sep=',')
df2 = df2.query('RelativeStrike=="ATM"')
df2 = df2.query('RelativePeriod=="M1"')
y = pd.DataFrame.to_csv(df2, 'optionz15.csv', sep=',')
```

8.4. VBA high-frequency

```
Sub dataclean()
'Niklas Lindeke, spring 2015
Dim year, month, day As Integer
Dim hour, minute, second As Integer
Dim datea, p_date As Date
Dim isExecuted As Boolean
Dim l As Long
Set shSource = ThisWorkbook.Sheets("Sheet1")
Set shDest = ThisWorkbook.Sheets("Sheet2")

l = 1
For j = 3 To 25
   For i = 4 To 65000
         'Select date
         datea = shSource.Cells(i, j)
         For k = 0 To 50

            'variable for previous date for checking highest value in minute
            p_date = shSource.Cells(i + k, j)

            'Parse year, month, day, hour, minute, second
            year = Val(Mid(datea, 1, 4))
            month = Val(Mid(datea, 6, 2))
            day = Val(Mid(datea, 9, 2))
            hour = Val(Mid(datea, 12, 2))
            minute = Val(Mid(datea, 15, 2))
            second = Val(Mid(datea, 18, 2))
            p_day = Val(Mid(p_date, 9, 2))
            p_hour = Val(Mid(p_date, 12, 2))
            p_minute = Val(Mid(p_date, 15, 2))
            p_second = Val(Mid(p_date, 18, 2))
            'The routine chooses the value associated with the minute with the highest value of the variable
second
            Do While p_hour >= 8 And p_hour <= 18
               If minute = 0 Or minute = 5 Or minute = 10 Or minute = 15 Or minute = 20 Or minute = 25 Or
minute = 30 Or minute = 35 Or minute = 40 Or minute = 45 Or minute = 50 Or minute = 55 Then
                  If minute = p_minute And day = p_day And second >= p_second Then
                     If Not isExecuted Then
                        shDest.Cells(l, 2) = shSource.Cells(i, j + 1)
                        shDest.Cells(l, 1) = shSource.Cells(i, j)
                        isExecuted = True
                        l = l + 1
                     End If
                  End If
                  If p_minute <> minute Then
                     i = k + i - 1
                     isExecuted = False
                     Exit For
                  End If
               End If
            GoTo line1
            Loop
line1:
         Next k
   Next i
   j = j + 1
Next j
End Sub
```

8.5. VBA Calculation of the RV

```
Sub harrv()
'Niklas Lindeke spring 2015
Dim day, p_day, month, year, sameday As Integer
Dim uniday, uniday_p As Date

Dim rv As Variant

Set sh1 = ThisWorkbook.Sheets("Sheet1")

m = 1
l = 1
j = 1

'the loop for going through the dates
For j = j To 320000
   rv = 0
   nobs = 0
   uniday = sh1.Cells(j, 1)
   uniday_p = sh1.Cells(l, 1)

   day = Val(Mid(uniday, 9, 2))
   p_day = Val(Mid(uniday_p, 9, 2))

   x = 0
'counting and consequently the summing of returns
   Do While day = p_day
      uniday = sh1.Cells(j, 1)
```

```
        uniday_p = sh1.Cells(l, 1)
        day = Val(Mid(uniday, 9, 2))
        p_day = Val(Mid(uniday_p, 9, 2))
        If p_day = day Then
            If rv = 0 Then Cells(m, 34) = Cells(l, 4)
            ret = sh1.Cells(l, 4)
            rv = rv + ret ^ 2
            Cells(m, 7) = rv

            Cells(m, 6) = uniday
            n = sh1.Cells(l, 5)
            nobs = nobs + n
            Cells(m, 10) = nobs

        End If
        l = l + 1
    Loop
    Cells(m, 7) = Sqr(rv)
    l = l - 1
    m = m + 1
    j = l
    If m = 258 Then GoTo line1
Next j
line1:
End Sub
```