LUND UNIVERSITY

"Symmetry or Asymmetry?"

A model comparison between different ARCH-class volatility models using Bitcoin returns

Author
Hannes Wiklund

Supervisor
Simon Reese

Department of Economics
Master Thesis - NEKN01
Submission Date: 25-05-2022

Table of Contents

1. Introduction

Forecasting is a valuable tool for economic actors and is one of the primary goals of time series analysis. Being able to predict future outcomes is especially important in the financial sector, as Poon & Granger (2003) notes. For example, when pricing options, it is essential to know the volatility of the underlying asset throughout its maturity period. Moreover, Poon & Granger also point out that volatility forecasting is required by financial regulations. As with the introduction of the Basel accords, banks need to evaluate the volatility of their assets to assess their value-at-risk exposure, as this metric determines the amount of reserve capital these organizations need to hold to guard themselves against price downturns. However, when an analyst takes on the task to forecast volatility, a central question will arise in the main phases of the examination: *which model should I choose?*

A salient feature of financial time series is that they exhibit asymmetry in their volatility. Asymmetric volatility entails that either bad or good news, or, in other words, negative or positive shocks respectively, have a larger effect on asset volatility. Black (1976) argued that stocks are affected by a leverage effect. Meaning that when a stock experiences negative returns, their financial leverage is increased, which inturn increases their volatility. This insight led Nelsson (1991) to create the EGARCH model. The EGARCH model is an ARCH-class volatility model that incorporates this asymmetric volatility effect into its specification. Starting with the EGARCH model of Nelson (1991), a whole host of other asymmetric models have been added to the tool box of volatility analysts. Empirical findings have also shown that these asymmetric models perform better than symmetrical ones. Hansen & Lunde (2005) and Awartani & Corradi (2005) both were able to show that asymmetric models gave the best volatility forecast when modeled on stock returns. But the same could not be said for exchange rate data. Hence, asymmetric models are not a catch-all and there needs to exist evidence to motivate their usage compared to more parimsonlsy models.

Lately, the above question regarding what kind of volatility model to choose, has been discussed in regards to cryptocurrencies. The focus on cryptocurrencies, in some aspects, comes at a surprise as it has been disregarded by investors and economists alike. However, on the other hand, the usage of cryptocurrencies have increased in recent years. A survey from HSB (2020) showed that 36 percent of small and midsize businesses in the U.S. accepted

cryptocurrencies as a form of payment. Furthermore, Bitcoin can now be traded either using options or futures. Which have led some to suggest that these tradable goods have “matured” (Baur & Dimpfl, 2018:150) and have become more accepted by the overall market. Yet, as the survey from HSB (2020) clearly states, it is unclear if holders of these kinds of assets understand the financial risk surrounding them. Cryptocurrencies have exhibited high amounts of volatility in their returns, compared to other forms of currencies (Peng et al. 2017). The price of Bitcoin swung between being valued below 1000 dollars in February of 2017, to 20,000 dollars in December of the same year (Ftiti et al. 2021). Hence, there exists a need to evaluate which kind of models fit these kinds of assets the best.

This thesis will in turn evaluate the forecast performance of different ARCH-type models' forecast ability using Bitcoin returns from 01-04-2015 to 01-04-2022. More specifically, it is of interest to see if a simple GARCH(1,1) model can outperform more sophisticated models that incorporate the asymmetry in volatility. Besides the GARCH(1,1) model, other models used in this thesis were the IGARCH, EGARCH, GJR-GARCH, TGARCH and APARCH models, where the EGARCH, GJR-GARCH, TGARCH and APARCH are asymmetric volatility models. To compare the different models' forecast abilities, the Hansen et al. (2011) *Model Confidence Set* test was utilized on the two loss functions: *Mean Square Error* and *Mean Absolute Error*. Further, there is evidence that the Bitcoin returns exhibit structural breaks, i.e. changes in the dynamic. To deal with the problem of possible structural breaks in the data generating process, and to check if the results hold over time, the period studied has been divided into 2 sub periods of equal length where the models are evaluated. As a further robustness check, the evaluation period was also doubled from 150 observations to 300. This will have the effect of increasing the power of the Model confidence set test, as well as reducing the number of observations for the estimation window, minimizing the risk of structural breaks affecting the conclusion even more. Although research has already compared volatility models using Bitcoin returns, this thesis will expand the literature in two ways. First of all, this thesis will use new data not considered by past research. All of the papers highlighted in this thesis, even the most recently published, have only analyzed pre 2020 data. As the volatility process of Bitcoin returns is susceptible to change, it is necessary to see if the results hold over time and, hence, it is imperative to analyze more recent data. Secondly, although all of the past papers have utilized asymmetric models in their comparison, none of them have specifically wanted to see if asymmetric models outperforms the symmetric GARCH(1,1).

The results showed that the case for using asymmetric models to explain Bitcoin returns are limited, as the GARCH(1,1) were often selected by the Model confidence set. Although the overall performances of asymmetric models were poor at all forecast horizons, the GJR-GARCH model performed quite admirably, as it was selected more times by the model confidence set test than the GARCH(1,1). On the other hand, even though the GJR-GARCH performed better than the GARCH(1,1), this is not sufficient evidence to warrant the use of asymmetric models, as the other asymmetric models were, more often than not, selected by the model confidence set procedure. Despite the steps taken to minimize the risk of structural breaks, the GARCH(1,1) exhibited strong IGARCH effects, which can be the result of structural breaks during the estimation window. Further, this also had the effect that the GARCH(1,1) and IGARCH model often produced similar forecasts.

The structure of this thesis is as follows: in section two, findings of past research regarding volatility forecasting are highlighted, both regarding cryptocurrency and other assets, to acclimate the reader to the findings in this field. Moving forwards to section three, the GARCH models utilized, the tests used and forecasting strategy applied in this thesis are introduced and explained to the reader. Section four gives an overview of the selected data and the results from the model confidence set test are reported. Lastly, in section five, the main results from the paper are highlighted and compared to past findings.

2. Literature Overview

Researchers have compared in the past the performance of different volatility models. Hansen & Lunde (2005) and Awartani & Corradi (2005) both wanted to see if a more simple GARCH(1,1) could give better forecast predictions than more complicated volatility models. Hansen & Lunde applied the analyses on IBM stock returns and the Deutsche Mark-Dollar exchange rate data, whereas Awartani & Corradi only looked at the stock return of the S&P-500 stock index. Although the two papers investigated different time periods, used different forecasting evaluation tests and different forecast horizons, their results pointed to the same conclusion. When it came to stock returns, volatility models that incorporated an asymmetric parameter outperformed the simple GARCH(1,1) model. Furthermore, Awartani & Corradi also showed that this result held for forecast horizons between one day and 30 days. On the other hand, for the exchange rate, Hansen & Lunde concluded that no model, neither any of the asymmetric models tested, was able to outperform the GARCH(1,1). As Awartani & Corradi explains, these results are due to the fact that stock returns are affected by leverage effect, whereas exchange rate returns do not.

When it comes to cryptocurrencies, the asymmetric property is harder to find in their returns. Baur & Dimpfl (2018) looked at the asymmetric properties of 20 different cryptocurrencies returns between 2013 to 2018 (the starting date changed depending on which date the respective currency was introduced). Baur & Dimpfl reported that nearly all of the assets had a negative asymmetric parameter using the TGARCH. Meaning, in contrast to stocks return, positive shocks give rise to higher volatility. Not discussed by Baur & Dimpfl, but shown in their reported results, almost all of the parameter estimations for the different cryptocurrencies were insignificant at the 5% level. Only three of the currencies had a significant asymmetric parameter at 5% level, none of which was a major traded cryptocurrency such as Bitcoin. These findings are supported by Bouri et al. (2017). Bouri et al. found that in between August 2011 and April 2016, Bitcoin only showed asymmetry in its volatility between August 2011 to November 2013, i.e. before the Bitcoin price crash of December 2013. Studying the returns of the S&P 500, it displayed asymmetric volatility throughout the whole period, indicating that results for Bitcoins volatility were not due to a market phenomenon.

In spite of the evidence for an asymmetric parameter in cryptocurrencies is limited, research has shown that these asymmetric models often achieved better performance than symmetric volatility models. For example, even though Bouri et al. (2017) found limited evidence for asymmetric properties in the volatility process for Bitcoin, the result from an in-sample comparison, using the Schwarz information criterion to assess performance, showed that the asymmetric model fitted the data better compared to the GARCH. Nevertheless, it has also been shown that the out-of-sample performance of asymmetric models beats the conventional GARCH. Looking at Bergsil et al. (2022), the results demonstrated that the APARCH and EGARCH models performed better than symmetric models, when using Bitcoin returns between 2014 and 2018. Peng et al. (2017) also found similar results for Bitcoin, looking at a one year period between 2016 and 2017. On the other hand, the outcome was that the GJR-GARCH performed a bit better than a GARCH model. It should be noted that Peng et al. did not include the APARCH model into their analysis.

As brought up in the introduction, there are signs that cryptocurrency experience structural changes in their volatility. For example, Bouri et al. (2017) results indicated that there was a significant change in the asymmetric parameter after the Bitcoin Crash of 2013. This finding is further supported by Bouri et al. (2018), which tried to identify structural breaks in Bitcoins volatility between 2011 to 2016. Bouri et al. also found evidence of a structural break during November of 2013. Furthermore, Bouri et al. also spotted a change in the volatility dynamics of Bitcoin during January of 2015. As discussed by Lamoureux & Lastrapes (1990), changes in the volatility process can lead to an upward bias on the parameter estimations of the GARCH model. Lamoureux & Lastrapes warned that when a GARCH model is estimated over a long time period, thus increasing the risk of estimating the model during the presence of a shift in the conditional variance, the parameters summation will go towards unity, amplifying the persistence of shocks. Persistence, in other words, denotes that past volatility shocks have a longer lasting effect on current volatility. Hence, structural changes can lead to model misspecification and poor estimates, which can have an adverse effect on forecasting ability.

The choice of the estimation window is further complicated by the fact that there exist no determined optimal window length to gain good forecasting results. For example, Poon & Grangers (2003) showed that the estimation window can vary from a 12 year to a one year window in their overview of the volatility forecasting literature. When Akgiray (1986)

compared ARCH-class volatility models using stock data, the sample was divided into subperiods, each consisting of about 1500 observations, or 6 years worth of daily data. Akgiray noted that this amount of observation was adequately large to gain precise parameter estimates for volatility forecasting. Bronwlees et al. (2011), on the other hand, wanted to specifically see how the length of the estimation window affected models forecast performance. Comparing window sizes of 4, 8 and 11 years worth of daily data Brownlees et al. found that the 11 year window yielded the best forecast when using a rolling window method. Still, the four year window generated only slightly less accurate estimates compared to the longest window size. Brownlees et al. argued that this was due to parameter instability in the estimated process. Even though the shortest window got less precise estimates due to the limited amount of data, it had the ability to faster capture changes in the parameters and, therefore, was able to compete with the models using the 11 year data window. Although the authors noticed parameter drift in the process, Brownlees et al. did not find any suggestion of any breaks during the studied time interval.

There are other methodology choices that, on the surface, can seem arbitrary, but can have an effect on the conclusion of the analysis. First of all, as discussed by Poon & Granger (2003), the true test of a model's ability to explain its data is its out-of-sample forecasting ability, i.e. can the model predict observations not included outside of the estimation window. Poon & Granger (2003) explains that out-of-sample mirrors reality more closely, and that in-sample evaluation implicitly assumes that volatility remains stable over time, which the above discussion has shown, might not be the case. Secondly, in Bergsli et al. (2022), different loss functions lead to different results when comparing models. Loss functions quantify the size of the forecasting error from a model prediction compared to the actual value. Bergsli et al. (2022) applied both loss functions used in this thesis, Mean Squared Error (MSE) and Mean Absolute Error (MAE), among others. In the case of Bergsli et al. (2022), which also used the Model Confidence Set test, MAE led to a smaller set of good performing models, compared to the MSE.

3. Methodology

In *Methodology*, the models and evaluation procedure applied in the thesis will be made clear. To start off, the volatility models used in this thesis will be presented in section 3.1. In section 3.2, the applied loss functions, that quantify each model's forecasting accuracy, will be discussed. Furthermore, in 3.2 the volatility proxy, *realized variance,* will be presented. In 3.3, the *Model Confidence Set* test, devised by Hansens et al. (2011), will be introduced, as it will be used to determine if a model's forecasting ability is statistically significant from others. Lastly, in section 3.4, the size of estimation, evaluation windows and subsamples will be specified and discussed.

3.1 GARCH Models & the GARCH(1,1)

ARCH-class models were first introduced by Engle (1982) when he specified the ARCH (AutoRegressive Conditional Heteroscedasticity) model. The ARCH model was proposed to capture volatility clustering and the excess kurtosis found in financial time series, by using p past realization of the variance to forecast future realizations of the conditional volatility. The ARCH model was expanded upon by Bollerslev (1986) when he put forward the GARCH (Generalized ARCH) model. The GARCH extends the ARCH model by introducing q lags of the conditional variance itself into the ARCH model, resulting in a model with a more parsimony lag structure.

The GARCH(1,1) model is defined in equation (3):

$$r_t = \mu_t + \epsilon_t \tag{1}$$

$$\epsilon_t = \upsilon_t \sigma_t \tag{2}$$

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{3}$$

Equation (1) defines the mean equation, where r_t denotes the return at time t and μ is the conditional mean specification of r_t. The residual from the mean equation, ϵ_t, is defined in equation (2) using υ_t, an IID term with 0 in mean and 1 in variance, and the conditional

standard deviation, σ_t. For GARCH(1,1) equation, the conditions, $\omega > 0$, $\alpha \geq 0$ and $\beta \geq 0$ are enforced to ensure that the conditional variance produced by the model is non-negative.

The GARCH(1,1) is widely used as it is an uncomplicated model that can capture the main characteristics of financial time series (Francq & Zakoian, 2019:19). As the *Literature Overview* showed, it has its limitations in capturing the dynamics of some financial time series. One of these features, not yet discussed, is that financial time series often exhibit a high level of persistence. Therefore, another symmetric model, the IGARCH, will also be used in this thesis as a competing model, since this model is specifically modeled to capture high persistence in the volatility process.

3.1.1 IGARCH

The IGARCH(p,q) (Integrated GARCH) replaces equation (3) in the GARCH(1,1) in the following way:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{4}$$

The IGARCH has the same structure as the GARCH(p,q) model, but it is constrained so that the coefficient sum of the p-lags of alphas and q-lags of betas add up to one. This restriction on the parameters has the effect that past shocks have a permanent effect on future volatility forecasts. The motivation for this, as Engle & Bollerslev (1986) explained when they presented the IGARCH, was due to the fact that financial time series often show high levels of persistence in their volatility, and Bitcoin is no exception. Bouri et al. (2018, 2017) noted that Bitcoin returns showed high levels of persistence. As discussed in section 2, high levels of persistence in time series can be the result of structural breaks. But, Lamoureux & Lastrapes (1990) makes it clear that persistence is still a feature of financial time series, their findings suggest that the level of persistence in financial time series might be overstated. Therefore, the IGARCH model is also included in this thesis. Yet, neither the GARCH(1,1) or the IGARCH contain an asymmetric parameter. As shown in section 2, asymmetric models

have shown promise in explaining the volatility process of some financial time series. Hence, these kinds of models will be presented next.

3.1.2 EGARCH

The EGARCH(p,q) (Exponential GARCH), proposed by Nelson (1991), defines (3) as:

$$\log \sigma_t^2 = \omega + \sum_{i=1}^{p} [\gamma_i \frac{\epsilon_{t-i}}{\sigma_{t-i}} + \alpha_i (|\frac{\epsilon_{t-i}}{\sigma_{t-i}}| - E[|\frac{\epsilon_{t-i}}{\sigma_{t-i}}|]) + \sum_{j=1}^{q} \beta_j \log \sigma_{j-i}^2 \tag{5}$$

The log-linear form of the EGARCH allows the coefficient parameters to be negative, thus the constraints applied to the parameters of the GARCH model are not needed for the EGARCH, making the model more flexible. The γ coefficient captures the asymmetric effects in the volatility. If $\gamma < 0$, then a negative shock results in a larger effect on the volatility than positive shocks. From the Literature Overview, this model has proven to be able to explain Bitcoin volatility, and hence has been included in this thesis.

3.1.3 GJR-GARCH

The GJR-GARCH(p,q) (Glosten, Jagannathan and Runkle GARCH) model give (3) the form of:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} (\alpha_i \epsilon_{t-i}^2 + \gamma_i D_{i,t-1} \epsilon_{t-i}^2) + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{6}$$

$$D_{t-1} = \begin{cases} 1, & \text{if } \epsilon_{t-1} < 0 \\ 0, & \text{if } \epsilon_{t-1} \geq 0 \end{cases}$$

The constraints, $\omega > 0$, $\alpha_i \geq 0$, $\alpha_i + \gamma_i \geq 0$ and $\beta_j \geq 0$ are applied to ensure positive conditional volatility.

The GJR GARCH model, suggested by Glosten et al. (1993), includes an indicator function, D_{t-1}. D_{t-1} takes value 1 if the residual is negative and zero if it is positive. Consequently,

ϵ_{t-1} acts as a threshold. Any shocks bigger than ϵ_{t-1} will have different effects on the conditional variance. If $\gamma > 0$ then negative shocks will have a larger effect on volatility, and vice versa. Another fact about the GJR-GARCH model is that it nest another model applied in this thesis, the GARCH model. Nesting implies that the GJR-GARCH model can be reduced to a GARCH in its specifications. This happens when $\gamma = 0$. As with the EGARCH model, this model has shown promise in being able to model cryptocurrency volatility. Therefore, the GJR-GARCH was also included in the set of models utilized in the thesis.

3.1.4 TGARCH

The TGARCH(p,q) (Threshold GARCH) is a variation of the GJR-GARCH. It differs from the GJR-GARCH by modeling the conditional standard deviation, instead of the conditional variance. Furthermore, it replaces that squared residuals by their absolute value in the model specification. Thus, (3) is defined for the TGARCH as:

$$\sigma_t = \omega + \sum_{i=1}^{p}(\alpha_i|\epsilon_{t-i}| + \gamma_i D_{i,t-1}|\epsilon_{t-i}|) + \sum_{j=1}^{q}\beta_j\sigma_{t-j}^2 \tag{7}$$

$$D_{t-1} = \begin{cases} 1, & \text{if } \epsilon_{t-1} < 0 \\ 0, & \text{if } \epsilon_{t-1} \geq 0 \end{cases}$$

As in the GJR-GARCH case, the constraints, $\omega > 0$, $\alpha_i \geq 0$, $\alpha_i + \gamma_i \geq 0$ and $\beta_j \geq 0$ are applied to provide positive conditional volatility.

Devised by Zakoian (1994), the result of using absolute value of the residuals is that they act as a better estimate of the variance, compared to when they are squared. Hence, this specification should improve on the former GJR-GARCH model. The TGARCH incorporates the asymmetric parameter the same way the GJR-GARCH model does. Brownlees et al. (2011) found this model to perform the best, even when the forecasted period spanned the initial unstable phase of the Great Recession. Hence, this model has been proven to perform well under turbulent periods, and, hence, has been included in this thesis as a possible candidate.

3.1.5 APARCH

The APARCH(p,q) (Asymmetric Power ARCH), created by Ding et al. (1993), has the preceding structure for (3):

$$\sigma_t^{\delta/2} = \omega + \sum_{i=1}^{p} \alpha_i (|\epsilon_{t-i}| - \gamma_i \epsilon_{t-i})^{\delta} + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^{\delta} \tag{8}$$

The parameters , $\omega > 0$, $\delta > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ and $-1 \leq \gamma_i \leq 1$ are utilized to provide positive conditional variance.

The δ parameter is what primarily defines the APARCH model, as it frees up the model, making it more adaptable to the data. As autocorrelation for the absolute residuals can be larger than for the squared residuals, the δ parameter can account for these two different situations by changing its value (Francq & Zakoan, 2019:97). Asymmetry, as in the other models, are captured by the γ parameter. If $\gamma > 0$ then negative shocks lead to a larger effect on the volatility. The APARCH model nest several other models applied in this thesis. If $\delta = 1$, then it is reduced to an TGARCH model. On the other hand, if $\delta = 2$ and $\gamma_i = 0$ then it turns into a GARCH. Lastly, if $\delta = 2$ then the APARCH is specified as the GJR-GARCH. This model as well has demonstrated good forecasting ability. In Hansen & Lunde (2005) extensive analysis, it was shown that the APARCH model to be the best performing for stock return volatility.

3.2 Loss Functions & Volatility Proxy

To measure the accuracy of the different models' forecast abilities, two error statistics will be used: (i) *Mean Square Error* (MSE) (ii) *Mean Absolute Error* (MAE). These two error statistics are presented below:

$$MSE = \frac{1}{n} \sum_{t=1}^{n} (\sigma_t^2 - \hat{\sigma}_t^2)^2 \tag{9}$$

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |\sigma_t^2 - \hat{\sigma}_t^2| \tag{10}$$

σ_t^2 denotes the forecasted conditional variance and $\hat{\sigma}_t^2$ stands for the actual variance for time t. There are several reasons for using several measurement errors when evaluating forecast performance. As we saw in the *Literature Overview,* different measurement errors can lead to different rankings of the forecasts, a result also highlighted by Diebold & Lopez (1996). This is due to the error statistics putting different weights on different kinds of forecast errors. For example, the MSE loss function puts more weight into unexplained outliers, compared to the MAE. Despite the fact that different measurement errors can lead to different conclusions, Hansen & Lunde (2005) notes that there is no consensus regarding which measurement error is preferred over the other. Therefore, as Hansen & Lunde (2005) also suggest, it is better to use several methods to quantify the errors generated by different models forecasts.

When using measurement errors such as MSE and MAE, a choice needs to be made regarding how to measure the volatility of the process, i.e. $\hat{\sigma}_t^2$. Different volatility proxies have been used in the litterature. For example, Awartani & Corradi (2005) used the daily squared returns as their volatility proxy. Yet, the usage of squared returns as a volatility proxy have been criticized in the litterature (Hansen & Lunde, 2005: Poon & Granger, 2003: Poon, 2005). Poon (2005:21) stated unambiguously that "[t]he practice of using daily squared returns to proxy daily conditional variance has been shown time and again to produce wrong signals in model selection". For example, Hansen & Lunde (2006) showed that inferior models can be chosen during out-of-sample evaluation, when using squared returns as the volatility proxy. Hence, Hansen & Lunde (2006) suggest using *Realized Variance* (RV) as a proxy instead, since it gives a more precise measure of the volatility.

RV uses intraday returns to calculate the conditional variance, i.e. returns that occur during trading. The intraday return are calculated as follows:

$$r_{t,i,m} = \log(p_{t-(i-1)/m}) - \log(p_{t-i/m}), \text{for } i =, ..., m \tag{11}$$

$r_{t,i,m}$ is then the logged return for the time interval with a length of $1/m$ on the given day t. RV, with the frequency of m, is then defined as the summation of the squared intraday returns:

$$RV_t^m = \sum_{i=1}^{m} r_{t,i,m}^2 \tag{12}$$

There is no agreed time interval for choosing m. Poon & Granger (2003) explains that defining m in between a 5- to 15-minute interval is commonly used in the litterature. Hansen & Lunde (2005), for example, used a 5-minute interval for calculating the realized variance in their study. But, both Poon & Granger (2003) and Hansen & Lunde (2005) warn against choosing an interval lower than 5-minute as this can make the returns to become correlated, since they will be affected by market microstructure effects. With these facts in mind, the RV proxy will be defined by using 15-min intraday logarithmic returns.

3.3 Evaluating Forecast Performance, the Model Confidence Set Test

When evaluating the performance of forecast predictions, a natural inclination would be to simply judge them based on their loss function value. And one could be convinced that this method would be adequate, as this approach has been used time and again in the litterature (Diebold, 2015). However, as Diebold (2015) states, you cannot merely announce a winner based on the fact that it achieved a lower loss function value than the others, you also need to show if the difference was statistically significant. As the comparison is made on a sample realization and may not be indicative for the population as a whole. Hence, Diebold and Mariano (1995) proposed their Diebold-Mariano (DM) test to compare forecasts between two selected models.

Although the DM test is extensively used in the litterature, the DM test suffers some shortfalls. Diebold (2015) notes that the DM test was never meant to answer the question "which model is the best", instead its purpose is to see if the two diffrent *forecasts* differ significantly. Further, the DM test can perform poorly when comparing a model that is nested in the other, as noted by Clark and McCracken (2001). As discussed above, the APARCH nest the GARCH model. If these two models were to be compared to each other, and the analyzed volatility process was created by a GARCH model, then the loss function of the fitted GARCH model would be smaller than the APARCH, even though they should perform

equally. As the APARCH models would induce noise into its prediction with its included redundant parameters (Clark & West, 2007). Lastly, White (2000) pointed out when executing pair-wise comparison of several models forecasting abilities for the same dataset, you can run the risk of multiple testing problems, or what White refers to as "data snooping" (White, 2000:1098), i.e. that you run the risk of selecting a model that, on the surface, seems to explain the data, but is in fact "useless" (White, 2000:1097).

As a result, this thesis will use the *Model Confidence Set* (MCS) created by Hansen et al. (2011). The MCS wants to find the set of the best performing models, M^*, from the set of used models, M. Hansen et al. (2011) explains that MCS-procedure works through an algorithm that applies an equivalence test and an elimination rule. The MCS-procedure first carries out the equivalence test. This test checks if the forecast errors from the models differ significantly from each other. Hence, the equivalence tests checks if the models offer equal predictive ability (EPA), or $H_{0,M} : E[d_{ij,t}] = 0$ for a given significance level α, where $d_{ij,t}$ is the difference in loss between model i and j at time t. To test the null of EPA, the following t-statistic in equation (13) is used:

$$t_i = \frac{\bar{d}_i}{\sqrt{\hat{var}(\bar{d}_i)}} \tag{13}$$

$\bar{d}_i = \frac{1}{m} \sum_{j \in M} \bar{d}_{ij}$, where m is the amount of models in M and $\bar{d}_{ij}$ is the average of $d_{ij,t}$ during the evaluation period. Hence, $\bar{d}_i$ is the sample loss of model i compared to the mean losses for the models contained in M (Elliot & Timmermann, 2016:416). $\hat{var}(\bar{d}_i)$ is the sample estimated variance. As many t-test are generated with equation (13), the following test statistics in equation (14) is used to evaluate the EPA null (Elliot & Timmermann, 2016:416):

$$T_{Max,M} = \max_{i \in M} t_i \tag{14}$$

As the test statistic in (14) follows a non-standard distribution, the critical values can be found using a bootstrap procedure, similar to the bootstrap method outlined by White (2000). In this thesis, the bootstrap-procedure was done with 10,000 resample using a a block length equal to the amount of significant parameters for an autoregressive model applied on all the $d_{i,j}$ terms. If the null of EPA is rejected, then at least one of the model's forecast errors performs significantly worse than another. When this occurs, then the elimination rule is applied:

$$e_{max,M} = \arg\max_{i \in M} t_i \tag{15}$$

In combination with equation (14), the elimination rule removes the model with the largest t-statistic relative to the mean loss of the set of models considered (Elliot & Timmermann, 2016:416). The above steps are sequentially performed until the EPA null is not rejected. The resulting set of models is the $\hat{M}^*_{1-\alpha}$, which contains the best performing models with a confidence interval of $(1-\alpha)$. A $MCS_{p-value}$ is also given for each included model in the $\hat{M}^*_{1-\alpha}$. These $MCS_{p-value}$ values are the threshold that determines if a model is included into the $\hat{M}^*_{1-\alpha}$ set, as a model is only included if and only if $MCS_{p-value} > \alpha$. Hence, as Hansen et al. (2011) explains "an object with a small MCS p-value makes it unlikely that it is one of the best alternatives" (Hansen et al., 2011:455). Albeit, the $MCS_{p-value}$ does not indicate the probability that the included model is the best out of the selected.

There are several aspects of the MCS that need to be highlighted. In contrast to other tests that compare several models and adjust for possible data snooping (such as the *Superior Predictive Ability* and *Reality Check* tests) it does not make use of a specified benchmark model to which the other models are compared to. Furthermore, the MCS is more flexible since it gives a set of models with favorable performance. This feature of the MCS procedure can result in a situation where many or all models are selected as the best performing ones. On the face of it, it might seem that this aspect of the test is a deficiency. But Hansen et al (2011) suggest that this eventuality shows that the test recognizes the quality of the data. If many or all models are selected, then this shows that the data is of lesser quality or less

“informative” (Hansen et al., 2011:454). Further, such results can also suggest that none of the models applied to the data can capture the dynamics of the process. In other words, the models utilized are equally bad in predicting the return volatility, and hence are all included into the best performing model set. Hence, the MCS test is flexible in its applicability and its results offer more information about the models tested and data set used.

3.4 Evaluation Strategy

To evaluate which model fits the Bitcoin returns the best, the daily trading and intraday prices of Bitcoin, denoted in USD, between 01-04-2015 to 01-04-2022 have been retrieved from Bitstamp, an online exchange for cryptocurrencies. Bitstamp is not the largest Bitcoin exchange house in terms of volume traded, but it was the first Bitcoin exchange to become a fully regulated payment institution in the EU (Shin, 2016). In comparison to stock exchanges, Bitcoins exchanges are open every day for every hour of the day. Hence, trading occurs for 365 days (366 days during leap years), in contrast to 252 days for which stock exchanges are open on average per year. Although Bitcoin has been traded since its inception in the late 2011s, the availability of regular intraday data only exists since 2014 via Bitstamp (Bergsli et al., 2022), hence the need to limit the time period analyzed to acquire percies volatility estimations using the RV procedure. However, to exclude the structural break identified by Bouri et al. (2018), observations before 01-04-2015 have been excluded in this thesis. Despite regular intraday data being available during the studied time period, 112 intraday prices are missing in the dataset. The longest continuous period of missing observations is a four and half hour period during 25-04-2020. These missing values have been replaced with the latest known value. These altercation of the data will not bias the RV metric, as the log return of these observations will be zero.

The period analyzed has been divided into two sub periods of equal length to minimize the risk of structural breaks confounding the result of the analysis, but also giving enough observations in each subsample to make precise estimates. Hence, subsample I will contain observations from 01-04-2015 to 30-09-2018, whereas sub sample II continues from 01-10-2018 until 01-04-2022. Each subperiod includes 1279 observations, where the last 150 observations, or around 12% of the observations in the subsample, are used for forecasting evaluation. Although the structural breaks identified by Bouri et al. (2018) have been

excluded in this thesis, there still exists the possibility of structural breaks in data considered by the thesis. It should be noted that the division was made arbitrarily, akin to Akgiray (1989). Akgiray also divided his sample randomly, but likewise as in this study, did it to increase the likelihood of having homogeneous sub periods. Furthermore, the choice of limiting the amount of sub periods to two was to ensure that the length of the estimation period was sufficient to get relatively good performing models. Note that these estimation periods are slightly shorter than the shortest period considered by Brownlees et al. (2011) in terms of years. The estimation window contains 1129 observations, or about 3 years worth of data. But, as Bitcoins are traded everyday, this estimation window specification equates to nearly four and half years of stock data. Hence, the estimation window specified should be adequate. To act as a robustness check, the models will also be evaluated when the evaluation period has been doubled. This will have several effects. Firstly, as a consequence, the estimation window will be reduced to 979 observations, or 2,5 years worth of data. Secondly, the expanded evaluation window will raise the power of the MCS-procedure, as the number of observations used for the procedure is increased. Lastly, it will also include a part of a turbulent period during the first subsample, which will become clearer in section 4.2. Making it possible to see which model performs the best under more uncertain periods.

The procedure of which the model specification was determined and how the forecast was created is as follows. For each subperiod, the lag length of the models, except for the GARCH(1,1) model, was determined using the Akaike and Baynesian information criterion (AIC and BIC respectively) during the first phase of the estimation period. However, the result of this step led to all of the models getting their p- and q -lags equal to 1. Furthermore, the conditional mean, that is μ in equation (1), has been specified as a constant for all sub periods and models. The forecasts are created using a rolling window method. Meaning that $[1, n]$ of observations are used to create the first t-step ahead forecast. For the next t-step ahead forecast, the parameters for the models are estimated using $[2, n + 1]$, i.e. we drop the last observation in the first estimation window, but add the actual return at time t into the window. The reason for choosing a rolling window was to further account for possible changes in the time series process. Brownlees et al. (2011) found that this method, in comparison to a recursive window where new observations are added but old ones are not dropped, was better suited when the estimated process exhibits instability. Six forecast horizons were considered, when t is equal to: 1, 5, 10 and 15.

4. Data Description & Results

In section 4, the overall structure of the Bitcoin data and the result from analysis will be presented. Firstly, in part 4.1, the price data and log-returns for Bitcoin from 01-04-2015 to 01-04-2022 will be analyzed to make the reader informed about the structure of the data. In the next part, 4.2, the results from the sub period analysis will be shown, as well as the result from the robustness check, when the evaluation period is increased from 150 observations to 300 observations.

4.1 Data Description

Via an ocular inspection of figure 1, the daily closing price of Bitcoin exhibits signs of non-stationarity throughout the sample. The result from a Dickey-Fuller test, performed on the closing price process from the starting date 01-04-2015 to the last date 01-04-2022, also supports this assertion. As the null is not rejected, we cannot exclude the possibility that the process does follow a random walk. Hence, the data was transformed by the following equation:

$$R_t = \log(P_t) - \log(P_{t-1}) \tag{16}$$

Where P_t is the closing price at period t and where R_t denotes the logarithmic returns. The process for the logarithmic returns is also reported in figure 1. The resulting process clearly displays stationarity as it means reverting around the value zero, an observation which the Dickey-Fuller test confirms. Furthermore, the process also shows signs of volatility clustering, as one can see periods of both high and low volatility indicating volatility persistence, making it suitable for volatility modeling. Engles famous ARCH LM test, which checks for ARCH-effects in the process, is also significant at the reported 10 lags as shown in table 1. Figure 2 instead shows the process of the realized variance throughout the sample. Notable in the two figures is the upward spike in volatility and downward spike in returns around the beginning of 2020. This negative return, of around 49%, at 12-03-2020 is more than twice as big in absolute value than any other returns during the whole sample. As such values can have an adverse effect on parameter optimization when applying the models to the

data. Therefore, this observation value has been replaced by the mean return for the whole sample. Data trimming of outliers is not an uncommon occurrence in volatility forecasting, as it has been suggested that such extreme observations follow their own distribution, and are not indicative of the process as a whole (Poon, 2005:7-8). Note that this modification of the data will not have an effect on the main result of this paper, as this observation is outside of all evaluation periods. The summary statistics for the logged returns for the entire sample is reported in table 1, where the negative spike at has been excluded as these outliers can seriously affect values reported for kurtosis and skewness.

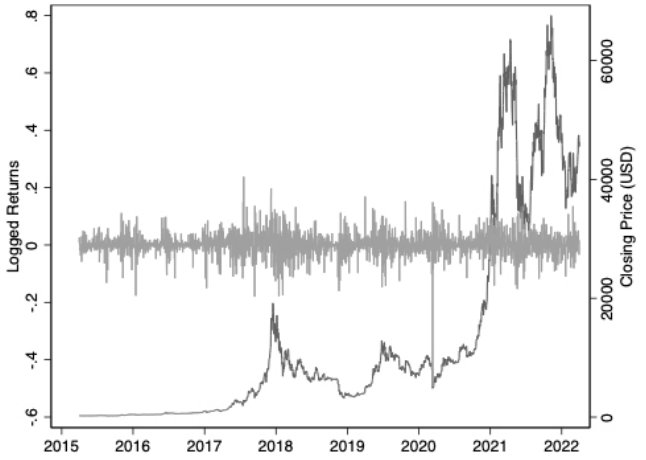

Figure 1 - Closing Price and Logarithmic Returns for the whole sample. The logarithmic returns are shown in gray and the daily closing price of Bitcoin in USD are shown in black. Period runs from 01-04-2015 to 01-04-2022

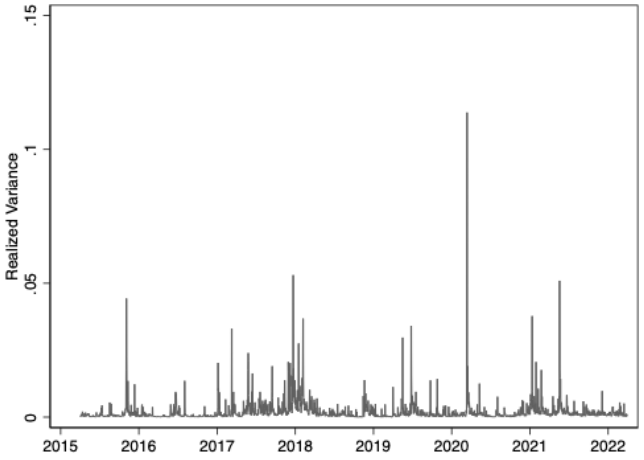

Figure 2 - Realized variance for the whole sample. The realized variance of Bitcoin has been calculated using the summation of the 15-min intraday logarithmic returns. Period runs from 01-04-2015 to 01-04-2022

Period	Mean	Std. Dev.	Min	Max	Skew.	Kurt.	ARCH-LM
Whole Sample	.002	.038	-.18	.238	-.05	6.707	139.76***

Table 1 - Descriptive Statistics for logarithmic returns for the whole sample - Shows the mean, standard deviation (Std. Dev.), minimum value, maximum value, skewness (Skew.), kurtosis (Kurt.), as well as the test statistics from the ARCH LM test using 10 lags, for the logarithmic returns during the period 01-04-2015 to 01-04-2022. The large negative return at 12-03-2020 has been replaced by the mean. *** denotes significance at the 1% level.

Figures 3 to 4 show the realized variance for subperiod I and II, respectively. From the figures we can see that the first subsample is more irregular when it comes to its volatility. In this period, Bitcoin saw its first price growth during the middle of 2017 to the beginning of 2018 where the price dramatically dropped, leading to a period of higher volatility in regards to the rest of the subsample. On the other hand, as seen in figure 4, subperiod II is more homogeneous in its volatility. There are also periods of increasing volatility, especially

around the end of 2020, as Bitcoin prices saw once again a sharp increase. A common occurrence in both samples is that there are large spikes in volatility compared to each respective sub sample. Still, the first evaluation period of 150 observations for both subsamples (this start has been marked with a red line) do not exhibit any remarkable jumps in volatility. These two periods are relatively calm, compared to each subperiod's respective estimation windows. Although, when it is increased to 300 observations, we see that a part of the turbulent period during 2018 is included into the evaluation period. Table 2 shows the summary statistics for the logged returns during the two sub periods. As seen in the table, period I has more excess kurtosis than period II, further indicating that period I was more turbulent than period II. Albeit, the logarithmic returns are fairly symmetrical in both periods. As for the whole sample, both periods exhibit ARCH-effects. Hence, modeling of Bitcoin returns volatility using ARCH-class models is applicable in both superiods.

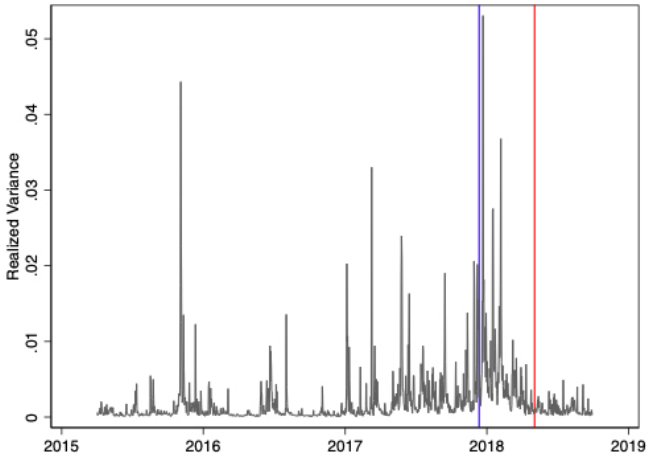

Figure 3 - Realized variance for subperiod I. Note: the start of the evaluation period of 150 observations has been marked with a red line. While the start of the prolonged evaluation period of 300 observations has been marked with a blue line.

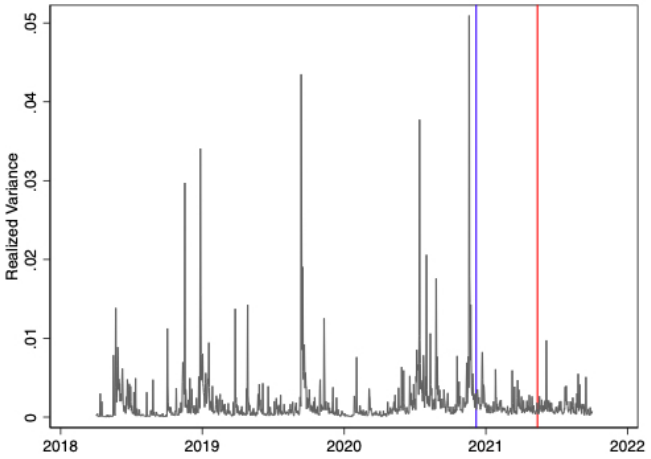

Figure 4 - Realized Variance for subperiod II. The evaluation period of 150 observations has been marked with a red line. The start of the prolonged evaluation period using 300 observations has been marked with a blue line. The volatility spike at 12-03-2020 has been excluded from the figure

Period	Mean	Std. Dev.	Min	Max	Skew.	Kurt.	ARCH-LM
Subperiod I	.003	.038	-.18	.238	-.152	7.585	110.04***
Subperiod II	.002	.038	-.16	.178	.056	5.788	36.99***

Table 2 - Descriptive statistics for subperiod I & II. Shows the mean, standard deviation (Std. Dev.), minimum value, maximum value, skewness (Skew.), kurtosis (Kurt.), as well as the test statistics from the ARCH LM test using 10 lags, for the logarithmic returns during the two specified subperiods. The large negative return at 12-03-2020 has been replaced by the mean return. *** denotes significance at the 1% level.

4.2 Results

Table 3 reports the MSE and MAE errors for the models at the four different forecast horizons during the first subsample. As seen, during most periods the GJR-GARCH model achieved the lowest forecast errors, followed by the simple GARCH(1,1) model. The forecast errors for GARCH(1,1) and GJR-GARCH models do not deviate in any substantial way. However, the remaining asymmetric models performed the worst for the whole selection of models during all sub samples. The EGARCH model was, by a large margin, the worst performing model. Giving MSE errors that are at least twice as big as the best performing model for all forecast horizons. Though, the relative difference gets smaller when MAE loss values are contrasted. The values for the GARCH(1,1) and IGARCH models in table 3 indicates that volatility persistence is high during this time period. The GARCH and IGARCH models forecast errors are about the same during all forecast horizons. Suggesting that the α_i and β_j coefficients of the GARCH(1,1) model are very close to unity, leading the GARCH(1,1) model to near the IGARCH model in its specification and, hence, resulting in similar volatility predictions. Furthermore, the TGARCH and APARCH models performed similarly. Not reported in the thesis, the δ parameter in the APARCH model is stable around 1, reducing the APARCH to the TGARCH model, explaining the similar volatility predictions.

	h=1		h=5		h=10		h=15	
Model	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GARCH	**0.6838**	0.5631	0.8457	0.6666	1.0052	0.7565	**0.8449**	0.6913
IGARCH	0.6859	0.5655	0.8514	0.6705	1.0175	0.7630	0.8538	0.6982
EGARCH	1.5266	1.0129	1.8404	1.1249	1.9623	1.1909	2.3266	1.2661
GJR-GARCH	0.7045	**0.5551**	**0.8028**	**0.6228**	**0.7966**	**0.6469**	0.8852	**0.6869**
TGARCH	1.3498	0.9575	1.7411	1.0846	1.9922	1.1923	2.4707	1.2986
APARCH	1.3319	0.9473	1.7072	1.0711	1.9319	1.1694	2.4265	1.2866

Table 3 - MSE & MAE errors for subperiod I. Reports the MSE & MAE values for each respective model during the evaluation period, 04-05-2018 to 30-06-2018, consisting of 150 observations. The errors reported for MSE has been scaled up with 10^6, while the errors from the MAE have been scaled with 10^3. Values marked in bold indicate the lowest forecast loss for each respective forecast horizon and loss function.

Studying table 4, the MCS result for subperiod I, the overall picture seen in table 3 is confirmed. The GARCH(1,1), IGARCH and GJR-GARCH models were included in nearly all of the different forecast horizons and specified loss functions for the $\hat{M}^*_{95}$ set. Only at the 10-step ahead forecast horizon were the GJR-GARCH model the only model selected into the $\hat{M}^*_{95}$ for both loss functions. The reason for the two loss functions generating the same result

using the MCS could be the lack of outliers in the evaluation period, leading to similar results for both MSE and MAE. Also reported in table 4, the set remains the same when the significance level is increased to 10%. Studying the parameter estimates for the GARCH(1,1) model through the sample for the one-step ahead forecast in figure 5, one can note that the parameter estimates nearly sums up unity, verifying the above explanation for the nearly identical GARCH(1,1) and IGARCH results. Furthermore, the parameter estimates are relatively stable over the evaluation period. The beta parameter grows slightly, increasing the effect of past volatility on future volatility predictions, increasing the persistence of volatility. In figure 6 the coefficients of the GJR-GARCH model are reported, as well as the 95% confidence interval for the asymmetry parameter, gamma. As seen in the figure, the gamma parameter is negative over the entire period, resulting in that past positive shocks leads to a higher volatility. Furthermore, the parameter is significant at 5% throughout the entire period. These findings rule out the possibility that the GJR-GARCH converged to a GARCH(1,1) model in its specifications.

	h=1		h=5		h=10		h=15	
Model	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GARCH	1.0000*	0.7775*	0.4358*	0	0	0	1.0000*	1.0000*
IGARCH	1.0000*	0.5152*	0.2761*	0	0	0	1.0000*	0.1516*
EGARCH	0	0	0	0	0	0	0	0
GJR-GARCH	0.4911*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	0.1176*	1.0000*
TGARCH	0	0	0	0	0	0	0	0
APARCH	0	0	0	0	0	0	0	0

Table 4 - Model Confidence Set results for subperiod I. Values reported in the table are the $MCS_{p-value}$ for the $\hat{M}^*_{95}$ set. A value of 0 indicate that the respective model was not included into $\hat{M}^*_{95}$. While * denotes if the model was also included into the $\hat{M}^*_{90}$ set. Period runs from 04-05-2018 to 30-06-2018, consisting of 150 observations

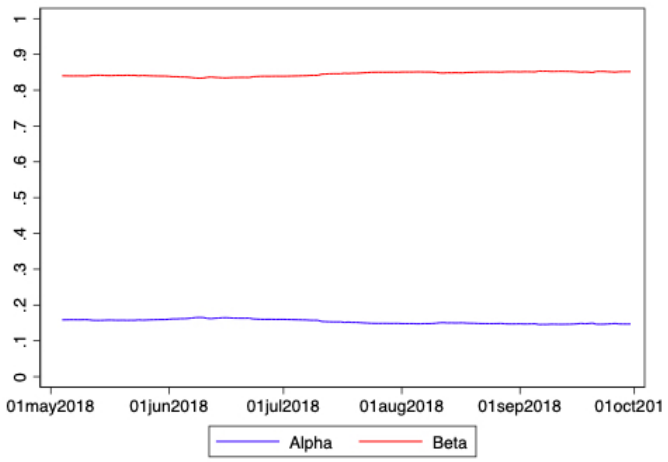

Figure 5 - Parameter values for the GARCH(1,1) during subperiod I. The parameter values are from the one step ahead rolling procedure.

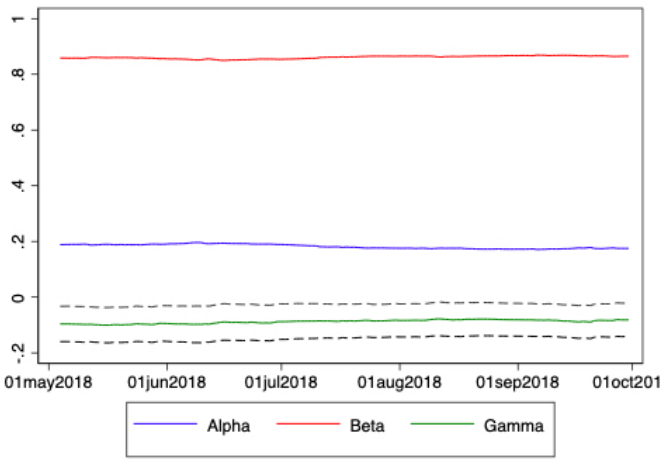

Figure 6 - Parameter values for the GJR-GARCH model during subperiod I. Estimates are from the one-step ahead rolling procedure. The 95% confidence interval for gamma is shown as dashed lines.

Table 5 reports the MSE and MAE values for the subsequent subsample. As seen in the table, GARCH(1,1) gave the lowest forecast losses for all forecasthorizions and loss functions, except for MAE at 15-steps-ahead forecast, where the GJR-GARCH performed the best. Further, for all subsequences horizons and loss functions, the EGARCH performed the worst. But, the relative difference between the top model and the EGARCH has enlarged in this subsample, compared to the previous subperiod. In some cases, the MSE errors generated by the EGARCH forecast were more than double that of the GARCH model for several forecast horizions. Also, in contrast to the first subperiod, the MSE errors are larger for all the models, implying that there are larger outliers in this period than the past evaluation window. At the same time, the MAE errors remain largely the same in value in both periods. As in subsample I, the errors for the GARCH and IGARCH models are nearly identical for all periods, indicating once again IGARCH effects in the volatility. A main difference between the two evaluation periods is the relative performance of the TGARCH and the APARCH compared to the best performing models. Their relative performance improved substantially, at some points nearly even matching the best performing one. As in the result for subsample I, these two models performed similarly.

	h=1		h=5		h=10		h=15	
Model	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GARCH	**1.3534**	**0.7887**	**1.3765**	**0.7887**	**1.4619**	**0.8348**	**1.6470**	0.9076
IGARCH	1.3602	0.7940	1.3852	0.7940	1.4750	0.8420	1.6646	0.9159
EGARCH	4.6221	1.2830	3.6237	1.2829	2.6182	1.1882	8.3444	1.9234
GJR-GARCH	1.4151	0.7928	1.4336	0.7927	1.5102	0.8258	1.6628	**0.8893**
TGARCH	1.6782	0.9653	1.7413	0.9653	1.8234	0.9996	2.1885	1.1416
APARCH	1.6484	0.9466	1.7011	0.9466	1.7853	0.9845	2.0524	1.0881

Table 5 - MSE & MAE errors for subperiod I. Reports the MSE and MAE values for each respective model during the evaluation period, 03-11-2021 to 01-04-2022, consisting of 150 observations. The errors reported for MSE has been scaled up with 10^6, while the errors from the MAE have been scaled with 10^3. Values marked in bold indicate the lowest forecast loss for each respective forecast horizon and loss function.

The $MCS_{p-value}$ for the $\hat{M}^*_{95}$ for the second evaluation window are shown in table 6. Not surprisingly, the result from the MCS-procedure is not as clear cut as in the first subsample. All models are included into the set when considering the MSE loss function. Even the EGARCH model that produced notably higher forecast error is included in nearly all of $\hat{M}^*_{95}$ sets for the MSE loss function. A reason for this result could be that the 150 observations used to create the set of superior models is too low, leading to false negatives in the equivalence test. The MAE leads to a more parsimonious selection of models. The difference between the result using the MSE and MAE could be caused by all the models having a hard time forecasting large increases in volatility. Interestingly, we see that the MAE leads mostly to the

same selection of models as we saw in the first subperiod, i.e. the GARCH(1,1), IGARCH and GJR-GARCH. These models are also included into all $\hat{M}^*_{95}$ sets for the MSE. Furthermore, the $MCS_{p-values}$ for the remaining models included into the set are relatively low. Hence, when the significance level is increased to 10%, these models are excluded from the set. Increasing the significance level, α, leads to a more narrow selection of models, as the EPA null hypothesis in the equivalence test is more easily rejected. But, ofcourse, increases the chance of rejecting a true null hypothesis.

	h=1		h=5		h=10		h=15	
Model	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GARCH	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	0.0939
IGARCH	1.0000*	0.9340*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	0
EGARCH	0.1215*	0	0.2232*	0.0930	0	0	0.2208*	0
GJR-GARCH	1.0000*	0.7049*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*
TGARCH	1.0000*	0	1.0000*	0.1706	0	0	1.0000*	0
APARCH	1.0000*	0	1.0000*	0.6653	0.0559	0.0664	1.0000*	0

Table 6 - Model Confidence set results for subperiod II. Values reported in the table are the $MCS_{p-value}$ for the $\hat{M}^*_{95}$ set. A value of 0 indicate that the respective model was not included into $\hat{M}^*_{95}$. While * denotes if the model was also included into the $\hat{M}^*_{90}$ set. Period runs from 03-11-2021 to 01-04-2022, consisting of 150 observations.

Lastly, the robustness checks for each respective sub period, where the evaluation windows are doubled, are reported in table 7 and 8, for subperiod I and II respectively. As the evaluation period for subsample I is extended to included more turbulent periods, we see that the $\hat{M}^*_{95}$ set for the MSE includes all models for all forecast horizons. Once again demonstrating that when the errors for unexplained outliers are given more weight, the models have the same prediction ability. Demonstrating that all of the models have a hard time explaining the large outliers included into the extended evaluation period. As seen in the table, the EGARCH, TGARCH and APARCH achieve lower $MCS_{p-values}$ then the GARCH(1,1), IGARCH and GJR-GARHC. On the other hand, the MAE leads to the same model selection when the number of observations was 150, i.e. the GARCH(1,1), IGARCH and the GJR-GARCH. This selection is also stable when significance level is increased to 10%. Examining table 8, we see that all of the models have been selected into the $\hat{M}^*_{95}$ for both loss functions. Looking at the $MCS_{p-values}$ for the models we can note that it has dropped for the EGARCH, as well as the TGARCH and APARCH. Whereas the $MCS_{p-values}$ for the GARCH(1,1), IGARCH and GJR-GARCH remains stable at 1.0000 throughout the table. As seen in figure 4, in section 4.1, the realized variance during this extended evaluation window remained relatively low. This fact supports the suggestion that

the result in table 6 was primarily a cause of low power of using 150 observations. Hence, when studying the results for $\hat{M}^*_{90}$ we see that the GARCH(1,1), IGARCH and GJR-GARCH are selected for almost all forecast horizons and loss function.

	h=1		h=5		h=10		h=15	
Model	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GARCH	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*
IGARCH	1.0000*	0.2617*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*
EGARCH	0.5073*	0	0.4974*	0	0.4342*	0	1.0000*	0
GJR-GARCH	1.0000*	0	1.0000*	0.1489*	1.0000*	0.8723*	1.0000*	0.6616*
TGARCH	0.3410*	0	0.3986*	0	0.4781*	0	0.6505*	0
APARCH	0.2686*	0	0.3172*	0	0.5874*	0	0.6976*	0

Table 7 - Model Confidence Set result for subperiod I using 300 observations. Values reported in the table are the $MCS_{p-value}$ for the $\hat{M}^*_{95}$ set. A value of 0 indicate that the respective model was not included into $\hat{M}^*_{95}$. While * denotes if the model was also included into the $\hat{M}^*_{90}$ set. The prolonged period runs from 05-12-2017 to 30-06-2018, consisting of 300 observations

	h=1		h=5		h=10		h=15	
Model	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GARCH	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000	1.0000*	1.0000
IGARCH	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000	1.0000*	1.0000
EGARCH	0.0822	0.0506	0.0963	0.0657	0.1174*	0.0870	0.2674*	0.0716
GJR-GARCH	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*
TGARCH	0.4632	0.0544	0.6255	0.0938	0.5562*	0.3097	0.5771*	0.2238
APARCH	0.4422	0.0553	0.6379	0.0874	0.5618*	0.3022	0.5797*	0.2160

Table 8 - Model Confidence Set result for subperiod II using 300 observations..Values reported in the table are the $MCS_{p-value}$ for the $\hat{M}^*_{95}$ set. A value of 0 indicate that the respective model was not included into $\hat{M}^*_{95}$. While * denotes if the model was also included into the $\hat{M}^*_{90}$ set. The prolonged period runs from 06-06-2021 to 01-04-2022, consisting of 300 observations

5. Discussion and Conclusions

The results gathered from the MCS procedure have shown that the result gathered from the initial subsample analysis, shows that the best performing models seem to be the GARCH(1,1), IGARCH and GJR-GARCH models. The most decisive results for this conclusion can be found in subperiod I, where these models were the only ones selected. However, the data was not as informative in subperiod II. The results from the MCS-procedure, using the MSE loss function, created a $\hat{M}^*_{95}$ set that included nearly all of the utilized models for each forecast horizon. Besides from the EGARCH model, all the models generated a $MCS_{p-value}$ of 1, indicating equal performance. On the other hand, when the MAE errors were used for the MCS, the generated results were nearly identical to the first evaluation period. The GARCH(1,1), IGARCH and GJR-GARCH was included in all of the $\hat{M}^*_{95}$ sets, except for the IGARCH in the 15-day ahead forecast horizon. Although other models were also included in some of the forecast horizons considered, they generated low $MCS_{p-value}$, often just above the threshold of being included into the set of best performing models.

When the evaluation was doubled, the result changed in some aspects. But the overall insights gained from the first two evaluation periods were strengthened. The MSE loss function for the first period led to all the models being selected into all of the $\hat{M}^*_{95}$ set. Yet, the $MCS_{p-value}$ for the EGARCH, TGARCH and APARCH models were considerably lower in comparison to the best performing models selected in the initial evaluation window. Interestingly, the MAE led to a similar pick during the evaluation window of 150 observations. The difference between these two loss functions could be due to the longer evaluation period including a more turbulent period with bigger volatility spikes, compared to the initial window, that are unexplained by the models. As the MSE weighs these forecast errors more heavily, the difference between the models might have been minimized, leading to a generous $\hat{M}^*_{95}$ set of models. Hence, there are some signs that the difference in forecast performance between the models during turbulent time. Yet, for the second subsample, the main results became even more shrouded. Here, for both loss functions, all models were included into $\hat{M}^*_{95}$ set. However, as before, the EGARCH, TGARCH and APARCH models $MCS_{p-value}$ was considerably lower, compared to the initial evaluation window, and was therefore

subsequently excluded from nearly all of the $\hat{M}_{90}^{*}$ sets. This shows that reported equal performance during the initial evaluation period, regarding the MSE errors, was possibly a consequence of low power in the test. However, the reason for the more liberal selection, with respect to the MAE loss function, is less obvious. Although the robustness checks for the two periods did not fully support the initial results, it still pointed in the same general direction. That the GARCH(1,1), IGARCH and GJR-GARCH are the most perferred models in terms of explaining Bitcoins volatility.

Hence, the evidence that asymmetric models better explain Bitcoins volatility than a simple symmetric GARCH(1,1) model is poor. The GARCH(1,1) model was almost always included into the best performing model set. The only other model that was included more into the superior model set was the GJR-GARCH model. A proposed reason for this could have been the GJR-GARCH model being reduced to a GARCH(1,1), as the former nest the latter. However, this was not the case as the asymmetric parameters for the GJR-GARCH were significant and negative during the evaluation period, ruling out this possible explanation for the comparable performances. Although the GJR-GARCH seems to pick up asymmetric volatility in bitcoin returns and, in one aspect, outperforms the GARCH(1,1), as it was included more times into the best performing set over both periods, these facts does give enough evidence that asymmetric models fit Bitcoin returns volatility the best. It is clear from the result that, overall, asymmetric models lead to worse volatility forecasting than the symmetric ones. In comparison to research done on stock return volatility, Awartani & Corradi (2005) found that the GARCH(1,1) was beaten by all of the asymmetric models considered at the 5% significance level. This result held when the forecast horizon was increased, but the difference diminished.

The result gained from this thesis seem to contradict the result gained from past research analyzing the Bitcoin returns volatility. For example, Bergsli et al. (2022) concluded that the best performing ARCH-class models were the EGARCH and APARCH. The conflicting result between Bergsli et al. (2022) and this thesis is striking, as these two models performed less than admirable in this thesis. The reason for the divergence in the result is hard to see. Bergsli et al. uses the same evaluation test as this thesis, the same amount of observations in the estimation window and the same length of forecast horizons. The only considerable differences are the time periods analyzed and the length of the evaluation period. Bergsli et al.

data started earlier, 01-01-2014, compared to this thesis. As noted in section 2, this earlier period had signs of structural breaks in the Bitcoin returns and therefore could have confounded the result. Bersli et al. used a longer evaluation window, consisting of 500 observations, it is hard to see how this choice could lead to such a difference in the results. As the result from the robustness checks in this thesis showed that the evidence for the EGARCH and APARCH models being included into the best performing set got lower when the number of observations in the evaluation window was increased. But some of the results in this thesis are also supported by past findings. Peng et al. (2017) also found that there was evidence for the GJR-GARCH outperforming the GARCH model.

There are limitations and issues regarding the results found in this thesis. Although steps were taken in this thesis to minimize the effects of structural breaks on the results, due to the GARCH(1,1) exhibiting IGARCH effects there is a possibility of structural breaks in the return process. However, it has been noted by Bouri et al. (2018, 2017), which studied Bitcoins volatility at different time periods, that the volatility process of Bitcoin exhibited high levels of persistence. Hence, there is a need to further evaluate to see if the persistence found in Bitcoin returns volatility is an effect of structural breaks, or is a part of the dynamics of the return process. This can be done by studying if models that try to account for structural breaks would have generated better forecasting results. This suggestion further highlights another limitation of this thesis, that it is silent about the vast majority of volatility models. But the primary interest of this thesis was to examine the effectiveness of asymmetric models compared to the GARCH(1,1), there are many asymmetric models left out from the analysis of this thesis.

In conclusion, the result of this thesis gives no support for using more sophisticated asymmetric models compared to a GARCH(1,1) model. The asymmetric GJR-GARCH model gave good forecasting results, and has so before in other research papers, the overall performance of asymmetric models suggest that incorporating asymmetric effects into the volatility model gives no advantage compared to a GARCH(1,1). Although issues surrounding the results of this thesis have been brought up, and there are many more volatility models left to be analyzed, the point view of this thesis that when it comes to volatility process of Bitcoin returns “nothing beats the GARCH(1,1)” (Hansen & Lunde 2005:888).

References

Akgiray, V. (1989). Conditional Heteroscedasticity in Time Series of Stock Returns: Evidence and Forecasts, *The Journal of Business,* Vol. 62, No. 1, pp.55-80.

Awartani, B., & Corradi, V. (2005). Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries, *International Journal of Forecasting*, Vol. 21, No. 1, pp. 167-183.

Baur, D. G., & Dimpfl, T. (2018). Asymmetric volatility in cryptocurrencies, *Economics Letters*, Vol. 173, pp. 148-151

Bergsli, L. Ø., Lind, A. F., Molnár, P., & Polasik, M. (2022). Forecasting volatility of Bitcoin, *Research in International Business and Finance,* Vol. 59, pp.1-30.

Black, F. (1976). Studies of stock prices volatility changes, *Proceeding of the 976 Meeting of the American Statistical Association, Business and Economic Statistics Section,* pp. 177-181.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics*, Vol. 31, pp. 307-327.

Bouri, E., Gil-Alana, L. A., Gupta, R., & Roubaud, D. (2018). Modelling long memory volatility in the Bitcoin market: Evidence of persistence and structural breaks, *International Journal of Finance & Economics,* Vol. 24, No. 1, pp. 412-426.

Bouri, E., Azzi, G., & Dyhrberg A. H. (2017) On the return-volatility relationship in the Bitcoin Market Around the Price Crash of 2013, *Economics: The Open-Access, Open-Assessment E-Journal,* Available online: http://www.economics-ejournal.org/economics/journalarticles/2017-2/ [Accessed 10 April 2022]

Brownlees, C., Engle, R. F., & Kelly, B. (2011). A practical guide to volatility forecasting through calm and storm, *The Journal of Risk,* Vol. 14, No. 2, pp. 3-22.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models, *Journal of Econometrics*, Vol. 105, No. 1, pp. 85-110.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics*, Vol. 138, No. 1, pp. 291-311.

Diebold, F. X., & Lopez, J. A. (1996). Forecast evaluation and combination, in S. Maddala & C. R. Rao (eds.), *Handbook of Statistics*, Vol. 14, Amsterdam: North-Holland, pp. 241-268.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy, *Journal of Business & Economic Statistics,* Vol. 13, No. 3, pp. 134-144.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the use and Abuse of Diebold-MAriano Tests, *Journal of Business & Economic Statistics*, Vol. 33, No. 1, pp. 1-9.

Ding, Z., Granger, C., & Engle, R. F. (1993). A long memory property of stock market returns and a new model, *Journal of Empirical Finance*, Vol. 1, No. 1, pp. 83-106.

Elliot, G., & Timmermann, A. (2016). *Economic Forecasting*, New Jersey: Princeton University Press.

Engle, R. F. (1982). Autoregressive conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica*, Vol. 50, No. 4, pp. 987-1007.

Engle, R. F., & Bollerslev, T. (1986). Modelling the Persistence of Conditioned Variance, *Econometric Reviews,* Vol. 5, pp

Francq, C., & Zakoian, J. (2019). *GARCH Models: Structure, Statistical Inference and Financial Applications*, 2nd ed., : John Wiley & Sons.

Ftiti, Z., Louhichi, W., & Ben Ameur, H. (2021). Cryptocurrency volatility forecasting: What can we learn from the first wave of the COVID-19 outbreak?, *Annals of Operations Research.*

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks, *The Journal of Finance,* Vol. 48, No. 5, pp. 1779-1801.

Hansen, P. R., & Lunde, A. (2005). A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?, *Journal of Applied Econometrics*, Vol. 20, No. 7, pp. 873-889

Hansen, P. R., & Lunde, A. (2006). Consistent ranking of volatility models, *Journal of Econometrics*, Vo. 131, No.1.2, pp.97-121.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The Model Confidence Set, *Econometrica*, Vol. 79, No. 2, pp. 453-497.

HSB. (2020). One-Third of Small Businesses Accept Cryptocurrency: Do They Understand the Cyber and Financial Risks? Available online: https://www.munichre.com/hsb/en/press-and-publications/press-releases/2020/2020-01-15-one-third-of-small-businesses-accept-cryptocurrency.html [Accessed 7 April 2022]

Lamoureux, C. G., & Lastrapes, W. D. (1990). Persistence in Variance, Structural Change, and the GARCH model, *Journal of Business & Economic Statistics,* Vol. 8, No. 2, pp.225-234.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach, *Econometrica,* Vol. 50, No. 2, pp. 347-370.

Peng, Y., Albuquerque, P.H.M., Sá, J.M.C.D, Padula, A.J.A., & Montenegro, M.R. (2017). The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with Support Vector Regression, *Expert Systems With Applications,* Vol. 97, No. 1, pp. 177-192

Poon, S., & Granger, C. W.J. (2003). Forecasting Volatility in Financial Markets: A Review, *Journal of Economic Literature,* Vol. 41, No. 2, pp. 478-539.

Poon, S. (2005). *A Practical Guide to Forecasting Financial Market Volatility,* New York: John Wiley & Sons.

Shin, L. (2016). Bitstamp Becomes First nationally Licensed Bitcoin Exchange; License Applies in 28 EU Countries, Forbes, 25 April, Avaiable Online: https://www.forbes.com/sites/laurashin/2016/04/25/7886/?sh=5ba0a49e6056 [Accessed on 26 April 2022]

Zakoian, J. (1994). Threshold heteroskedastic models, *Journal of Economic Dynamics and Control*, vol. 18, no. 5, pp. 931-955