Volatility of Cryptocurrencies: Can we model the volatility of Bitcoin with GARCH models?

Giovanni La Cagnina
École Polytechnique
Fédérale de Lausanne
Lausanne, Switzerland
Email: giovanni.lacagnina@epfl.ch

Joshua Voelkel
École Polytechnique
Fédérale de Lausanne
Lausanne, Switzerland
Email: joshua.voelkel@epfl.ch

I. Introduction

Cryptocurrencies have gained increasing attention in the past years. Bitcoin can be seen as its most prominent and successful advocate, with accounting for more than 45% of the total market cap of all cryptocurrencies in May 2022 CoinMarketCap [2022b]. But recent moves in Bitcoin prices show that there are periods of high volatility as, currently, Bitcoin is primarily used as an asset rather than a currency Glaser et al. [2014], Baek and Elbeck [2015], Bouri et al. [2017]. This also leads to the fact that the Bitcoin market is highly speculative compared to other currencies Cheah and Fry [2015] and studying its volatility is crucial. Abdalla [2012] concludes in their research that the exchange rates volatility can be adequately modelled by the class of GARCH models. This leads us to the question whether this is also true for cryptocurrencies, in particular Bitcoin. The aim of this paper, is to exactly address this quesetion. That is, we study whether it is possible to model the volatility of Bitcoin using different GARCH models. Our approach is to find the best performing models with respect to in-sample performance and out-of-sample performance (forecasting).

The paper is organised as follows. Section II briefly introduces what research has already been done on this topic. In section III, we analyse the statistical properties of the Bitcoin time series and its volatility time series. We introduce the different models we use throughout the paper in IV. In section V, we find the best performing models w.r.t. in-sample data and in section VI the best performing model w.r.t. unseen, out-of-sample data. After the out-of-sample analysis we select the best model and we test them on a test set in section VII. In section VIII, we summary our results and we give some intuitions regarding what we have observed.

II. Literature Review

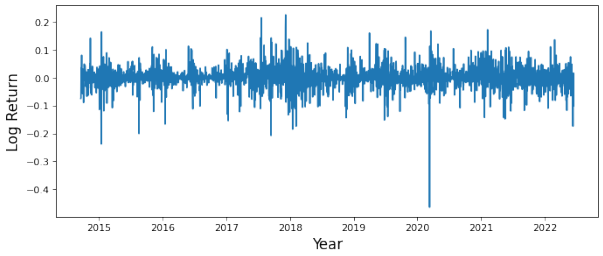
In this section, we describe the research that has already been done using different GARCH models to model the volatility of Bitcoin. While the literature around volatility modelling and forecasting is vast for most asset classes, there has been significantly less research done in the field of volatility modelling and forecasting for cryptocurrencies with a focus on Bitcoin. This could be due to the fact that it is still not clarified whether to classify Bitcoin as a commodity or currency. Furthermore, there are still many unsettled questions about the nature and utility of Bitcoin in the financial world Naimy and Hayek [2018].Baek and Elbeck [2015] analysed the volatility of Bitcoin with a focus on the question whether cryptocurrencies can be seen as an investment or just a speculative vehicle. They conclude that Bitcoin is extremely volatile and speculative and that its volatility is internally driven by demanders and sellers and not by external economic factors. Focusing on GARCH models to explain the volatility of Bitcoin, there have been multiple studies using different GARCH models trying to model the in-sample volatility of Bitcoin. Glaser et al. [2014], Gronwald [2014] used linear GARCH models, Bouri et al. [2017], Bouoiyour and Selmi [2015], Bouoiyour et al. [2016] used Threshold GARCH (T-GARCH) and Bouoiyour and Selmi [2015], Bouoiyour et al. [2016], Dyhrberg [2016] tried Exponential GARCH (EGARCH). However, all of these studies have only used one single model and did not do a comparison of different models. Katsiampa [2017] tried different GARCH models and came up with a AR-CGARCH model as the best performing model in terms of goodness-of-fit. Regarding out-of-sample performance, there has been significantly less research done. Naimy and Hayek [2018] tries simple GARCH(1,1) and EGARCH(1,1) models and compares the out-of-sample performance. Aras [2021] stacks multiple simple GARCH models to improve the predictive power, while Kim et al. [2021] compares simple GARCH models to different stochastic volatility models. However, there is no research done in comparing different GARCH models for forecasting. Hence, the aim of this study is to choose the best performing models both in-sample and out-of-sample while choosing from a large pool of different models, which we will describe in detail in section IV.

III. Data Analysis

In this section, we conduct a statistical analysis of the Bitcoin data at hand to see if there are quantitative reasons to model the

volatility using GARCH models. sIn the first step, we downloaded the daily Bitcoin price in USD (ticker: "BTC-USD") from yahoo finance, which takes the data from CoinMarketCap CoinMarketCap [2022a]. The time series starts 18-09-2014 and goes until 16-05-2022. This is the maximum range, we had access to. Using this price data, we calculated the daily log returns. These are visualized in figure 1.

Fig. 1. Daily percentage log returns of BTC-USD



Taking a look at the plot, we can clearly observe volatility clustering, that is, return volatility is not constant over time (heteroskedastic) and large (small) changes tend to be followed by large (small) changes Gourier [2022]. Furthermore, we plotted the autocorrelation function (ACF) and the partial autocorrelation function (PACF) for different lags both for the simple log returns and for the squared log returns. As we could not observe any significances for the simple log returns, we could observe that for the squared log returns, the PACF and ACF are significantly different from zero for the first, fourth and seventh lag. The respective plots can be seen in figure 2 and 3.

Fig. 2. ACF Plot of the squared daily log returns of BTC-USD

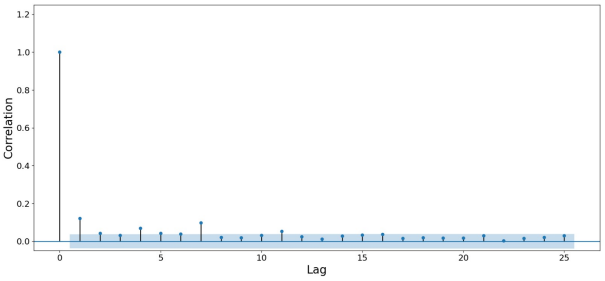
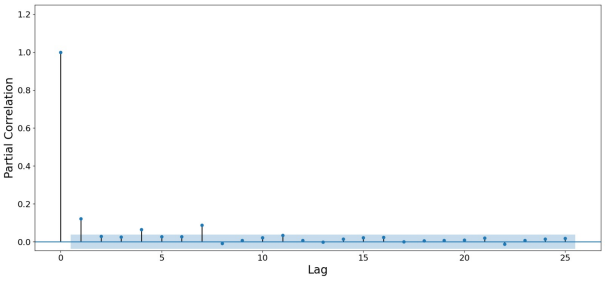


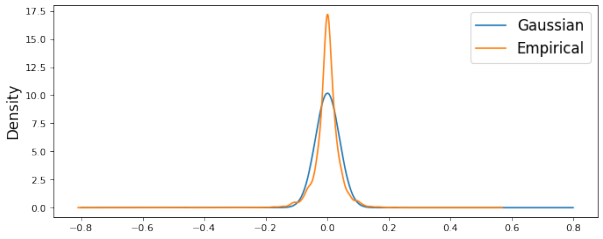
Fig. 3. PACF Plot of the squared daily log returns of BTC-USD



To not only visually but also quantitatively test for serial autocorrelation of the error terms, we applied the Ljung-Box test to the squared log returns. The test-statistics for various lags are summarized in table I and we reject the null-hypothesis that the squared log returns are independently distributed for all lags. These results are good indicators that GARCH models could potentially work well to describe the volatility of Bitcoin as GARCH models are specifically designed to model the volatility of data where the error variance follows an ARMA process. We discuss the model selection in more detail in section IV.

In the next step, we examined the distribution of log returns and plotted the empirical distribution together with a normal distribution having the same mean and variance as the empirical distribution. The plot is displayed in figure 4.

Fig. 4. Empirical distribution of log returns vs. normal distribution with same mean and variance



We can observe that the empirical distribution has heavier tails and is slightly negatively skewed. In fact, the distribution of log returns have a slight negative skewness of -0.77 and an excess kurtosis of 10.98, which verifies our visual observations. Intuitively, this means we observe more extreme returns compared to a normal distribution while the left tail is longer than the right tail indicated by the negative skewness. This can be interpreted such that the majority of log returns are slightly positive but there are a few strongly negative returns.

IV. Model Selection

This section aims to introduce the different GARCH models which we will use throughout the paper.

A. GARCH

The GARCH(p, q) process, Generalized Autoregressive Conditional Heteroskedasticity, Bollerslev [1986], is given by;

$$r_t = \mu_t + \epsilon_t, \qquad \epsilon_t = u_t \sigma_t, \qquad u_t \sim f(x),$$

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

where we indicate with $f(x)$ the distribution of the shock considered in the model. When the model was proposed for the first time the author used a Standard Normal distribution, but it is possible to use other shock distributions, as shown in III, for allow for fatter tails or additional skewness, as Standard Student-t distribution and Standard Skewed Student-t distribution. In the case where $q = 0$ the model simplifies to an ARCH(p) model. The main difference between ARCH

and GARCH models is the fact that the latter allows lagged conditional variance to influence the conditional variance σ_t^2. This model can be seen as an ARMA model for squared returns. An important aspect of this model that need to be underlined is that we have to impose positivity constraints on the parameters of the process in order to guarantee the positivity of the conditional variance estimated, that is

$$\omega > 0, \quad \alpha_i \geq 0, \quad \beta_j \geq 0$$
$$i = 1, \ldots, p, \quad j = 1, \ldots, q.$$

The GARCH model is the simplest model considered in this analysis and we can consider it as the benchmark model.

B. GJR-GARCH

The GJR-GARCH(p, o, q) model, Glosten et al. [1993], is given by:

$$r_t = \mu_t + \epsilon_t, \quad \epsilon_t = u_t \sigma_t, \quad u_t \sim f(x)$$
$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{h=1}^{o} \gamma_h \epsilon_{t-h}^2 1_{\{\epsilon_t - j < 0\}} + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2.$$

Glosten et al. [1993] modifies the GARCH model in order to take into account the leverage effect. The leverage effect describes the negative relationship between return and volatility, Dahlvid and Granberg [2017]. Therefore this process specification aims to model asymmetric volatility clustering, allowing for a different volatility response to past negative and positive returns. In this analysis we consider only GJR-GARCH models with $o = 1$, it would be also possible to consider larger values in order to take into account more than one lagged past squared return for the asymmetric response. As the GARCH models we need to impose non-negativity constraints on the coefficients to ensure the positivity of the conditional variance, that is

$$\omega > 0, \quad \alpha_i \geq 0, \quad \beta_j \geq 0,$$
$$i = 1, \ldots, p, \quad j = 1, \ldots, q.$$

Note that in this case we do not impose any constraints on the coefficients that model the asymmetric response of the variance.

C. EGARCH

The $EGARCH(p.j, q)$, Exponential Generalized Autoregressive Conditional Heteroskedasticity, Nelson [1991] is given by:

$$r_t = \mu_t + \epsilon_t, \quad \epsilon_t = u_t \sigma_t, \quad u_t \sim f(x),$$
$$\ln \sigma^2 = \omega + \sum_{i=1}^{p} \alpha \left(|u_{t-p}| - \sqrt{\frac{2}{\pi}} \right) + \sum_{r=1}^{j} \gamma_{t-r} u_{t-r} + \sum_{h=1}^{q} \beta_h \ln \sigma_{t-h}^2.$$

Nelson [1991] decides to use $j = p$. As the GJR-GARCH model, the EGARCH is an asymmetric GARCH type, in fact we can see that parameters γ_{t-j}, $j = 1, \ldots, J$ describe asymmetric response to different shocks, indeed if these coefficients are negative we could conclude that the negative shocks contribute more volatility than positive shocks, Chen et al. [2019]. One relevant fact about this class of models is that we do not have any positivity constraints to ensure positivity of conditional the variance. This can be seen as a positive feature for the reason that non-negativity constraints create difficulties in estimating GARCH models, Nelson [1991]. Despite the fact that the paper that introduced this model set $j = p$, in this analysis, after trying to fit different models we decided to use all $j = 1$. We made this choice for two reasons: by introducing larger orders of asymmetry, the corresponding parameters were not statistically different from zero, and furthermore, the performance, both in sample and out of sample confirms that EGARCH models with j = 1 give better results.

D. APARCH

The APARCH process, Autoregressive Conditional Heteroskedasticity, Ding et al. [1993] is given by:

$$\sigma_t^{\delta} = \omega + \sum_{i=1}^{p} \alpha_i (|\epsilon_{t-i}| - \gamma_i 1_{\{o<i\}} \epsilon_{t-i})^{\delta} + \sum_{j=1}^{q} \beta_{t-j} \sigma_{t-k}^{\delta}.$$

Ding et al. [1993] find that $|\epsilon_t|^{\delta}$ often displays strong and persistent autocorrelation for various values of δ, and that returns tends to have a long memory property. This model can be seen as a GARCH model where a Box-Cox power transformation is applied to the conditional volatility and to the asymmetric power residuals. The particularity of this model comes from the fact that the power used to model the conditional volatility of returns is estimated from data. As shown in the above written models also for APARCH model it is required to impose non-negativity constraints.

V. In Sample Analysis

In this section, we are examining whether it is possible to find a GARCH model that describes the past volatility of Bitcoin adequately well using the 4 different model classes introduced in section III.

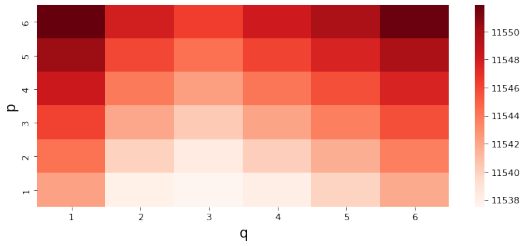
In the beginning, we divided our data set into a training set reaching from 18-09-2014 to 31-12-2020, validation set reaching from 01-01-2021 to 31-12-2021 and a test set starting from 01-01-2022 to 15-05-2022. The idea behind this was that we initially anticipated that the best performing in-sample models will also perform the best out-of-sample, which we elaborate in section VI.

As we have seen in section III, the distribution of residuals is slightly negatively skewed and has a strong positive excess kurtosis. Therefore, we wanted to test the performance of the different GARCH models using three different distributions. The standard normal distribution, the standardized Student's T distribution which has heavier tails (i.e. positive excess kurtosis) and the skewed Student's T distribution which has an additional skewness parameter which is also fitted on the data.

In the first step of determining the best performing models, we created two heat maps for each combination of model

class and distribution. The heat maps displayed the AIC and BIC values for different p and q ranging from 1 to 6. Here, p and q represent the amount of lagged shocks and conditional variance terms respectively. We chose AIC and BIC as performance measures of the model for the in-sample analysis due to the fact that these two information criteria aim to avoid overfitting of the model based on a penalty term regarding the number of parameters of the model. For our analysis avoiding overfitting is a desirable outcome since our initial idea was to use the best performing in-sample also to forecast volatility out-of-sample. If we had used only the log-likelihood value, we would have ended with a model having large number of parameters. Furthermore, it is also the most common performance measure for in-sample analysis of time series Zajic [2019]. To get an idea of how these heat maps look like, figure 5 shows an example of an AIC heat map for a GJR-GARCH(p, q) model with skewed Student's T distribution.

Fig. 5. Example of an AIC heat map for a GJR-GARCH(p, q) model with skewed Student's T distribution



In the next step, we chose for each model-distribution combination the best performing models each w.r.t. AIC and BIC. This gave us a pool of 23 potential best models.
To further evaluate the in-sample performance of these models, we computed the error between the fitted data (using the models) and the realized daily volatility. We decided to use both the mean absolute error (MAE) and the mean squared error (MSE), which are defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|,$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

where N denotes the sample size and y and $\hat{y}$ stand for the realized and fitted daily volatility respectively. Since the daily volatility is not a simply obserservable metric, we needed to come up with a proxy for it. In the literature Barndorff-Nielsen and Shephard [2005], Ahoniemi and Lanne [2013], Degiannakis and Livada [2013], a common proxy for daily realized volatility is the squared daily log return. That is

$$y_i = r_i^2,$$

where r_i denotes the log return for day i. This proxy is clearly a very rough measure and we will therefore refer to it as the *naive* proxy. Since we are having access to intra-day data, we came up with two, more precise proxies. That is, we computed the realized volatility using both minute and hourly data. Quantitatively, the realized variance is estimated using the following:

$$y_i = \sum_{j=1}^{n}(r_{ij} - \overline{r}_i)^2,$$

where n denotes the amount of minutes/hours per day, r_{ij} denotes the log return per minute/hours of at day i and $\overline{r}_i$ is the empirical mean of the log returns for day i. To visualize, how these proxies differ, we plotted the different proxies for the full data set in figure 6 and the signature plot in figure 7.

Fig. 6. Different realized volatility proxies

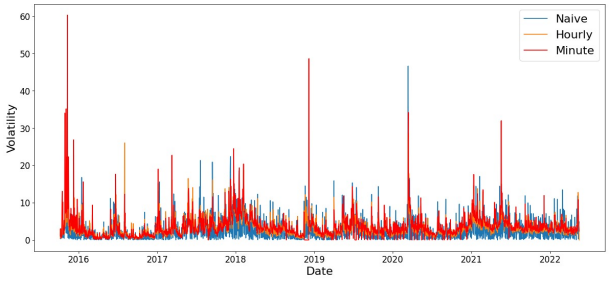
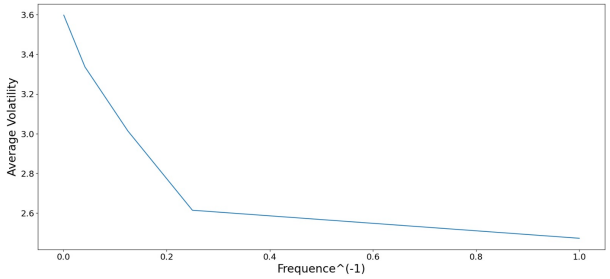


Fig. 7. Signature plot of different realized volatility proxies

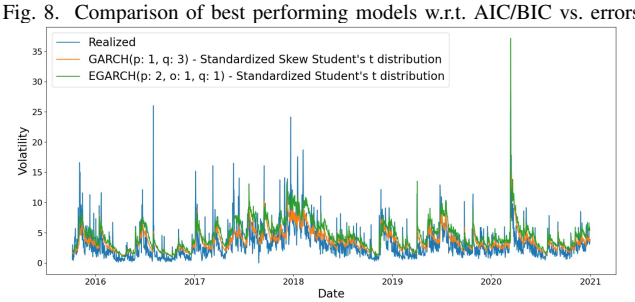


We can see that the different proxies lead to different results for the realized volatility. In fact, the proxy obtained using minutes data yields to a higher average volatility than the other two proxies, while the hour-proxy results in a higher average volatility compared to the naive proxy. This observation is also called microstructure noise and it shows that the realized volatility proxies are not very stable. Therefore it makes sense to report the losses with respect to each of these realized volatility series. Another observation we noted is that the minute data had two extreme outliers which made no sense economically. We, therefore, replaced them by the realized volatility from the previous day. On top of that, we could also observe some measurement errors in the minute data (between the years 2019 and 2021). That is, we observe that the reported minute prices were fixed, for some specific dates, for a part of the day (ranging from half day to full day). This leads to a realized volatility estimation around zero. Since this observation does not make sense economically, we decided to introduce a lowerbound of 0.5 and to replace each affected data point that lies below this threshold with the previous realized

volatility observation. We graphically illustrate this in figure 12.The intraday data is taken from Gemini Exchange Gemini. Using these three different proxies, we calculated the respective errors for the pool of models we obtained previously. The best performing model for each of the errors are summarized in table II. As described before, we manipulated the minute data fairly strong, which makes the proxies and errors a lot more unreliable. As we obtained very similar results between the minute proxy and hourly proxy, we decided to only report the hourly results and to not use the minute data as a proxy in further testing. We can observe that both GARCH$(1,3)$ and GJR-GARCH$(1,3)$ perform the best w.r.t. the different errors with standardized Student's T distribution. The two best performing models w.r.t. AIC and BIC are EGARCH$(2,1)$ with standarized Student's T distribution and the EGARCH$(1,1)$ also with standardized Student's T distribution. However, it must be kept in mind that we already chose the best performing models w.r.t. AIC and BIC to create this pool of models, so the differences in AIC and BIC are not too large for all models. Analyzing the best models, we can see that the standardized Student's T distribution seems to be the best choice to describe the insample data, as 5 out of the 6 best models use this distribution. The slight skewness, we observed in section III seems to be not significant. Additionally, we observed a negative statistically significant asymmetry parameter for the GJR-GARCH$(1,3)$ model. This result was surprising to us as it can be interpreted that that positive shocks in the returns imply a higher next period conditional variance than negative shocks of the same magnitude. This could possibly explained by the behaviour of cryptocurrency traders who get excited when the returns are positive and tend to trade it more, while they "sit out" negative returns of the same magnitude. Another interesting result is that there is no APARCH model under the best performing models. One of the main reasons for this was that they partially had convergence difficulties for some of the APARCH models, given our limited computational power, which lead to undesirable results, give also the fact that is model require to estimate the parameter for the power of the volatility numerically. Coming back to the best performing models, to further understand what it means to perform well w.r.t. AIC/BIC in comparison to the errors, we plotted in figure 8 the realized volatility using the hourly proxy and the fitted data of the EGARCH$(2,1)$ with standardized Student's T distribution (best performing w.r.t. AIC and BIC) and the GARCH$(1,3)$ with standardized Student's T distribution (best perfoming w.r.t. most of the errors).

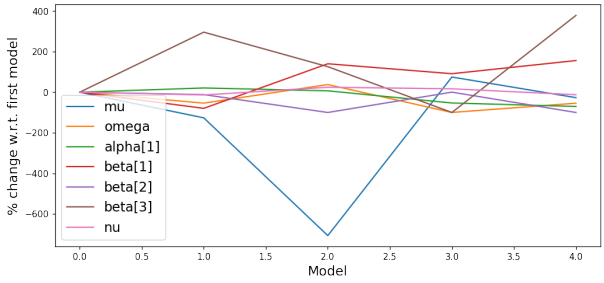
We can clearly see that the EGARCH model manages to capture the spikes adequately well, especially compared to the GARCH model. This can be explained by the nature of AIC/BIC, which aims at finding the model which fits the observed data the best using maximum likelihood. In contrast, the different error estimates compute the absolute/squared difference betweeen the realized and fitted data. This leads to the behaviour of the GARCH model, which manages to capture the general trends fairly well but is not so good at capturing the spikes.

Fig. 8. Comparison of best performing models w.r.t. AIC/BIC vs. errors



In the previous steps, we fitted the models using the whole training data. This raised the question, whether the parameters vary strongly when we using less data, i.e. subsets of the training data. Figure 9 shows an example of how the parameters change w.r.t. the initial model parameters using the GARCH$(1,3)$ model with standardized Student's T distribution. Here, we divided the training set into equally sized, non-overlapping bins and fitted the model for each of these training sets.

Fig. 9. Example of GARCH(1,3) (standardized Student's T distribution) parameter variation using bins of size 400



We can indeed see that the parameters vary strongly depending on the underlying training data. To see if this leads to different best performing models w.r.t. the errors, we split the train data in equally sized, non-overlapping bins of 400 and 600 days and computed the different errors by summing the individual errors of each of the bins. The best performing models w.r.t. the different errors are summarized in table III and IV and we can see that most of the best performing models mostly coincide with the ones using the whole training set. That is, the models that describe the in sample data the best w.r.t. the different errors are GARCH$(1,3)$ with standardized Student's T distribution and EGARCH$(5,1,1)$ with standard normal distribution. We therefore conclude that this potential problem of parameter instability turned out not to be a significant problem.

VI. Out of Sample Data Analysis

In this section we present the results and the approach used in the out-of-sample performance analysis of the GARCH models considered in the previous sections, in forecasting the Bitcoin volatility. Firstly we can explain what is the purpose of the following analysis: the out-of-sample performance analysis

aims to evaluate the performance of the GARCH models in forecasting the future volatility of BTC using past return values, indeed at each day t we aim to predict the value of the daily volatility of returns at time $t+h$, where h is called *horizon* and it represents how many day in advance we want to predict the volatility. The forecasted volatility σ_{t+h} will be a function of h, and $\mathcal{F}_t$, where the latter represent the set of past values of returns. The performance of the models in forecasting the future values of volatility is evaluated on the validation set, that includes BTC returns from 01-01-2021 to 31-12-2021.

We have forecasted the series of conditional volatility exclusively using the rolling window methods. The rolling windows method consist in the following algorithm:

1) For each date t we fit a model for the variance on w past returns, $r_t, r_{t-1}, \ldots, r_{t-w+1}$ and we obtain the estimation of parameters of the models.
2) Using the fitted model we forecast the value of σ^2_{t+h}. When $h > 1$ it is possible to use three different methods to forecast σ^2_{t+h}: Analytical, Bootstrap, and Simulation ARCH Library [2019].
   - Analytical: this methods is used in all models when $h = 1$ and it is used in longer horizons in models that evolve in terms of the squares of the models residuals. It exploit the relationship $E_t(\epsilon_{t+1}) = \sigma^2_{t+1}$. The variance forecasts are constructed as follow, where we show an example of a GARCH(1,1) model:
   $$\begin{cases} \sigma^2_{t+1} = \omega + \alpha\epsilon^2_t + \beta\sigma^2_t & h = 1, \\ \sigma^2_{t+h} = \omega + (\alpha+\beta)E_t(\sigma^2_{t+h-1}) & h > 1, \end{cases}$$
   - Simulation forecasts are obtained by simulating draws of the standardize residuals e_{t+h}. The simulated values are used to generate a pre-specified number of paths of the variance which are then averaged to produce the forecasts. Simulating k paths we obtain:
   $$\begin{cases} \sigma^2_{t+h,b} = \omega + \alpha\epsilon^2_{t+h-1,b} + \beta\sigma^2_{t+h-1,b} \\ \epsilon_{t+h-1,b} = e_{t+h-1,b}\sqrt{\sigma^2_{t+h,b}} \\ E_t(\epsilon_{t+h}) = \sigma^2_{t+h} = \frac{1}{k}\sum_{j=1}^{k}\sigma^2_{t+h,j} \end{cases}$$
   - Bootstrap forecasts are obtained in a very similar way to the Simulation forecasts, the only difference is that the standardize residuals are generated by the model using the observed data and the estimated parameters:
   $$\hat{e_t} = \frac{r_t - \hat{\mu}}{\hat{\sigma_t}}$$
   The simulated shocks are drawn with replacement from $\hat{e_1}, \ldots, \hat{e_t}$, and therefore we use only the data available at time t to simulate the variance paths.
3) After having estimated the value of the variance for time $t+h$, we can estimate σ_{t+h+1} going back to point 1) and considering $t+1$.

Note that for computational reasons in the cases where we use horizon larger than one, $h > 1$, our approach was to fit a model every h days, indeed we fit a model using as last observation r_t then we forecast $\sigma_{t+1}, \ldots, \sigma_{t+h}$. In this way we can reduce the number of times a model is fitted and therefore reducing the computational time, given the fact that the estimation of the parameters of these models are obtained numerically, which represents the most demanding part of this method.

The rolling windows method require the choice of the hyperparameter w, called *window size*, and it represents the number of observation on which a model, at each date, is estimated. The performance of each model is evaluated using the same loss functions used in the in-sample analysis, MAE and MSE and using the same two realized volatility series obtained using hourly and daily returns.

The first step undertaken in the out-of-sample performance was to estimate the performance in forecasting future values of the volatility of the best in-sample performance models obtained in the previous section V. We can see the results obtained in table V. We observe that: the best model with respect to the Naive MSE and to the 1h MAE is the GJR-GARCH$(4,1,2)$ with Standard Normal distribution, the best model with respect to the Naive MAE is the GJR-GARCH$(1,1,1)$ and finally the best model with respect to the 1h MSE is the GARCH$(1,1)$ with the Standardized Student's distribution. Given these results and the result obtained previously, reported in table II, the model that have performed the best in-sample are not the same models that obtained the best scores with respect to the losses in the out of sample analysis. In the light of this result, a change of approach to finding the model that performs best for prediction purposes is justified. The following approach is based on testing all possible parsimonious models on the validation set, considering the three possible probability distributions shown above (Normal distribution, Student-t distribution and Skewed Student-t distribution). For parsimonious models we mean that we have considered all possible GARCH, GJR-GARCH, EGARCH models with $p = 1, \ldots, 5$, $q = 1, \ldots, 5$ and, as written in IV, we set the asymmetric order for EGARCH and GJR-GARCH models equal to 1. Therefore in our approach we tested the performance of all the possible combinations of p, q and distributions for all the three models. Although this method is computationally intense permits us to determine and individuate the best model, and also to evaluate the differences in the best model when we increase or decrease the window size or the horizon of the forecasts. We have obtained the estimated volatility series using rolling windows method using different windows sizes and several horizons, indeed we have tested $w = 600, 800, 1000$ and horizon $h = 1, 2$. Considering $h = 1$, as above written, the forecasted volatility values are obtained using the Analytic method. We can see results obtained with this horizon, using different rolling windows size w in the tables VIII, VII VI (where we have ordered the models in the tables in order to have in the first row the models that performs the best with respect to the loss measure of the first column, Naive MSE, in the second row the model

that performs the best with respect to the second loss measure, Naive MAE, etc.) Given these results we can observe that:

- The best performance with respect to each loss measured is obtained in all the cases using $w = 1000$, so we can conclude that the performance of GARCH models in forecasting tends to increase with the rolling windows size.
- In the majority of the cases different loss measures based on different realized volatility estimations methods choose different models. We consider this result relevant given the fact that the absence in consistency in these scores, the choice of one benchmark loss measure and of the realized volatility method, will affect the final model choice.
- It possible to observe that MSE loss based on Naive realized volatility chooses always, with $w = 600, 800, 1000$ an EGARCH model.
- Considering $w = 1000$, table VI, we observe that the Mean Absolute Error loss tends to select as best model a GJR-GARCH model, both using Naive realized volatility and realized volatility based on intra-day hourly returns. It is has been shown that MAE loss measure is a better loss function to evaluate the average performance of a model rather than the MSE loss function, but it poorly penalizes large errors, Willmott and Matsuura [2005].
- Another relevant aspect is that our procedure chooses always model with Standard Normal shock distribution. We can consider this observation as relevant given the fact that it is in contrast with the in-sample analysis done in V, where the best models considered a Standardized Student-T distribution to model the density of shocks.
- We observe a tendency to choose distributions different from the Standard Normal when we reduce the rolling window size w.

Now we can analyse the results obtained using $h = 2$ reported in tables IX, X, XI, XII, XIII, XIV. We can observe the followings:

- Considering both forecasting methods, Simulations and Bootstrap, we see the same pattern as observed with $h = 1$: the performance tends to increase with the rolling window size, indeed the best performance have been obtained using $w = 1000$.
- Considering $w = 1000$, a surprising observation is that the Naive MSE loss of the best model, GARCH$(1, 2)$ with Standard Normal distribution, using the Bootstrap method, table IX, is smaller than the loss obtained using the best model with $h = 1$, EGARCH$(2, 1)$, table VI. This result is surprising given the fact that half of the predicted observations of the conditional volatility are the same, given our choice to fit a model every 2 days and using the given model we predict the future conditional volatilities σ_{t+1} and σ_{t+2}, rather than to fit a model each day and forecast only the future conditional volatility σ_{t+2}. This result would lead to the condition that bootstrapping shocks helps in forecasting the volatility in a more precise manner than the Analytic method that does not consider any shock. An intuition could be the fact that we are better off by bootstrapping, and therefore to draw with replacement from observed shocks rather than trying to estimate the distribution of shocks.
- We can observe that Bootstrap tends to outperform the Simulation method. The intuition behind this result can be taken from the previous point: drawing with replacement from the set of observed residuals performs better rather than imposing a distribution for shocks and simulate a random variable that follows the imposed distribution.
- We observe that GARCH$(1, 1)$ performs quite well using $w = 1000$ and Bootstrap method based on the losses using the hourly realized volatility proxy. This result is not completely surprising, as claimed by Hansen and Lunde [2005]. They stated that this extremely simple GARCH model often outperforms more complex models with additional parameters with the aim to model asymmetry volatility clustering such as EGARCH and GJR-GARCH, in forecasting conditional volatility. A possible explanation could be the fact that more complex models require the estimation of more parameters, which are introduced to describe asymmetric behaviour in volatility. This can be difficult to do systematically well, especially when the fundamentals of the price process tends to change over time. Therefore the GARCH$(1, 1)$ could be able to describe the general behaviour of the conditional volatility, that overall, could be useful to forecast future volatility.

VII. Test sample

In this section we report the results obtained on a test set, BTC returns from $01-01-2022$ to $16-05-2022$, using the models that performed the best on the validation set, shown in the previous section VI. Based on previous result on the validation set we decided to test three models using $h = 1$ and three models using $h = 2$, with the Bootstrap method for forecasting for the first two models and Simulation method for the last one. We reduce the number to test on test set because we do not consider this part as the selection of the best model, but more as a confirmatory analysis, where we desire to see the performance of such models on a data set that have never been used for selection. We decide to test for each horizon one model per GARCH class (EGARCH, GARCH, GJR-GARCH), based on tables VI, IX. Two practical notes: we have used a Standard Normal distribution for all the model tested on the test set, and for $h = 2$ we have selected the EGARCH$(3, 1, 1)$ in order to test an EGARCH model with this horizon. This model is taken from the table of best performing model based on the Simulation method, XII and the performance on the test set is obtained using Simulation, in order to be consistent with the previous results. The models selected on the validation set for $h = 1$ are:

- GARCH$(1, 1)$: best model based on 1h MSE,
- GJR-GARCH$(3, 1, 1)$: best model based on 1h MAE,
- EGARCH$(2, 1, 1)$: best on Naive MSE.

The models selected on the validation set for $h = 1$ are:

- GARCH$(1, 1)$: best model based on 1h MSE and 1h MAE,
- GJR-GARCH$(4, 1, 1)$: best model based on 1h MAE,
- EGARCH$(3, 1, 1)$: best on Naive MSE (Simulation method).

For $h = 1$ we observe that the GARCH$(1, 1)$ outperforms the other models in terms of Naive MSE and Naive MAE, while the best model in terms of 1h MSE and 1h MAE seems to be the GJR-GARCH$(3, 1, 1)$ (consistently to table VI). We visualize the forecasted series in the figure 10. For $h = 2$ we observe that the GARCH$(1, 1)$ seems to be the best model in term of Naive MSE and 1h MSE, while the GJR-GARCH$(4, 1, 1)$ seems to be the best model in terms of Naive MAE and 1h MAE. This result could derive from the different aspects emphasized by the two loss functions, indeed as wrote above, the MSE penalized more large errors, while the MAE tends to choose model that on average performed the better without put a relevant weight on the magnitude of the error. We can visualize the forecasted conditional volatility series in figure 11. A relevant aspect to consider is the fact that the EGARCH models considered did not perform better than the models selected for the other classes. Another aspect that we can note, and that is the confirmation of what shown in the previous section is the fact that the GJR-GARCH models estimated for $h = 2$ seems to be able to reproduce spikes in volatility, even though not precisely at the right time, or in the right direction.

VIII. Conclusion

Our work studied the questions whether it is possible to model the volatility of Bitcoin using GARCH models. We approached this question by finding the best performing GARCH models in-sample and out-of-sample. The best performing models in-sample were the GARCH$(1, 3)$, GJR-GARCH$(1, 3)$ and EGARCH$(1, 1)$ each using the standardized Student's T distribution. It is also worth mentioning that the GJR-GARCH$(1, 3)$ model had a negative, statistically significant asymmetry parameter which can be interpreted that positive shocks in the returns imply a higher next period conditional variance than negative shocks of the same magnitude. Initially we concentrated on the following question: is it possible to use in-sample result to choose a model for forecasting volatility? After having done the above illustrated analysis we can say that some information from the in-sample analysis can be used for forecasting the volatility. Indeed we can observe in table 8 that the best model in term of AIC is the EGARCH$(2, 1, 1)$ with Standardized Student's t distribution. It is possible to note that this model is one of the best performing also out of sample with $h = 2$ and method Bootstrap, but using a different a Standard Normal distribution. The difference in the distribution is not really relevant given the fact that using the Bootstrap method we draw from past shock in order to forecast reliably. The AIC, therefore, seems to be the onlyinformation criteria to follow in order to use some result from the in-sample analysis for predicting the volatility of Bitcoin. Despite these observations, we can say that the out-of-sample analysis generally leads to choosing different models from those that performed better in the in-sample analysis. The relevant results observed in the out-of-sample analysis are that a larger window size tends to lead to better results, and that the bootstrap method tends to outperform the simulation and analytical methods. In fact, it seems to be a winning strategy to use past shocks to create a volatility process that follows realized volatility, in terms of out of sample loss-scores. We can also observe this graphically. We could conclude that Bootstrap method seems to have further potentialities that can be analysed and made explicit with an ad-hoc analysis. Furthermore, we observed that the only model that is able to generate some spikes in volatility similarly to the ones observed in the realized volatility series using intra-day hourly data is the GJR-GARCH$(3, 1, 1)$, with $h = 2$ and Bootstrap method. Although it is able to create these peaks in the predicted series of volatility values, they are not precisely estimated in the sense of position and direction, indeed we can observe this when confronting the forecasted series with realized one. Therefore we can assess the presence of this spikes to the bootstrapping procedure that has drawn some large shocks for that dates. Another important result found is about the GARCH$(1, 1)$: indeed in most of the cases this simple model has shown to be able to beat more complex models which allows for asymmetry in the volatility process, or that includes a larger number of parameters in order to take into account more lagged past volatilities and past squared returns. We think that this result is due to the simplicity of this model that allows to forecast sufficiently well the average trend of the volatility using a small number of parameters. Important is the integration of the Bootstrap method with the GARCH$(1, 1)$ that allows to rely on past shocks to predict a similar forecasted values with a reliable volatility trend. Coming back to our initial research question, we can conclude that it is possible to capture the general trends of the return volatility of Bitcoin but GARCH models are not able to capture the spikes given the extreme jumps in the Bitcoin returns, which is also one of the most significant differences to common currencies.

References

Suliman Zakaria Suliman Abdalla. Modelling exchange rate volatility using garch models: Empirical evidence from arab countries. *International Journal of Economics and Finance*, 4(3):216–229, 2012.

Katja Ahoniemi and Markku Lanne. Overnight stock returns and realized volatility. *International Journal of Forecasting*, 29(4):592–604, 2013.

Serkan Aras. Stacking hybrid garch models for forecasting bitcoin volatility. *Expert Systems with Applications*, 174: 114747, 2021.

ARCH Library. Univariate volatility models, forecasting, 2019.

Chung Baek and Matt Elbeck. Bitcoins as an investment or speculative vehicle? a first look. *Applied Economics Letters*, 22(1):30–34, 2015.

Ole E Barndorff-Nielsen and Neil Shephard. Variation, jumps, market frictions and high frequency data in financial econometrics. 2005.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

Jamal Bouoiyour and Refk Selmi. Bitcoin price: Is it really that new round of volatility can be on way? 2015.

Jamal Bouoiyour, Refk Selmi, et al. Bitcoin: A beginning of a new phase. *Economics Bulletin*, 36(3):1430–1440, 2016.

Elie Bouri, Georges Azzi, and Anne Haubo Dyhrberg. On the return-volatility relationship in the bitcoin market around the price crash of 2013. *Economics*, 11(1), 2017.

Eng-Tuck Cheah and John Fry. Speculative bubbles in bitcoin markets? an empirical investigation into the fundamental value of bitcoin. *Economics letters*, 130:32–36, 2015.

Hao Chen, Jianzhong Zhang, Yubo Tao, and Fenglei Tan. Asymmetric garch type models for asymmetric volatility characteristics analysis and wind power forecasting. *Protection and Control of Modern Power Systems*, 4(1):1–11, 2019.

CoinMarketCap. Historical data for bitcoin. https://coinmarketcap.com/currencies/bitcoin/historical-data/, 2022a. Accessed: 2022-06-16.

CoinMarketCap. Major cryptoassets by percentage of total market capitalization (bitcoin dominance chart). https://coinmarketcap.com/charts/, 2022b. Accessed: 2022-06-15.

Christoffer Dahlvid and Per Granberg. The leverage effect-uncovering the true nature of us asymmetric volatility. 2017.

Stavros Degiannakis and Alexandra Livada. Realized volatility or price range: Evidence from a discrete simulation of the continuous time diffusion process. *Economic Modelling*, 30: 212–216, 2013.

Zhuanxin Ding, Clive WJ Granger, and Robert F Engle. A long memory property of stock market returns and a new model. *Journal of empirical finance*, 1(1):83–106, 1993.

Anne Haubo Dyhrberg. Bitcoin, gold and the dollar–a garch volatility analysis. *Finance Research Letters*, 16:85–92, 2016.

Gemini.

Florian Glaser, Kai Zimmermann, Martin Haferkorn, Moritz Christian Weber, and Michael Siering. Bitcoin-asset or currency? revealing users' hidden intentions. *Revealing Users' Hidden Intentions (April 15, 2014). ECIS*, 2014.

Lawrence R Glosten, Ravi Jagannathan, and David E Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, 48(5):1779–1801, 1993.

Elise Gourier. Financial econometrics: Time series models, lecture notes. 2022.

Marc Gronwald. The economics of bitcoins–market characteristics and price jumps. *Available at SSRN 2548999*, 2014.

Peter R Hansen and Asger Lunde. A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics*, 20(7):873–889, 2005.

Paraskevi Katsiampa. Volatility estimation for bitcoin: A comparison of garch models. *Economics letters*, 158:3–6, 2017.

Jong-Min Kim, Chulhee Jun, and Junyoup Lee. Forecasting the volatility of the cryptocurrency market by garch and stochastic volatility. *Mathematics*, 9(14):1614, 2021.

Viviane Y Naimy and Marianne R Hayek. Modelling and predicting the bitcoin volatility using garch models. *Int. J. Math. Model. Numer. Optimisation*, 8(3):197–215, 2018.

Daniel B Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the econometric society*, pages 347–370, 1991.

Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

Alexandre Zajic. Introduction to aic — akaike information criterion. https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced, 2019. Accessed: 2022-06-15.

APPENDIX

TABLE I
LJUNG-BOX TEST STATISTICS FOR SQUARED LOG RETURNS

Lag	lb statistics	lb p-value
1	42.20	8.23e-11
2	47.43	5.02e-11
3	50.44	6.45e-11
4	64.39	3.45e-11
5	69.59	1.24e-11
6	74.02	6.09e-11

TABLE II
IN SAMPLE RESULTS USING THE WHOLE TRAINING SET

Model	Distribution	AIC	BIC	Naive MSE	Naive MAE	1h MSE	1h MAE
EGARCH(p: 2, o: 1, q: 1)	Standardized Student's t distribution	11494.3	11534.5	14.24	2.96	6.63	2.02
EGARCH(p: 1, o: 1, q: 1)	Standardized Student's t distribution	11500.0	11534.4	14.10	2.97	6.90	2.06
EGARCH(p: 5, o: 1, q: 1)	Normal distribution	12248.6	12300.3	9.89	2.41	4.85	1.56
GJR-GARCH(p: 1, o: 1, q: 3)	Standardized Student's t distribution	11535.8	11581.7	9.90	2.34	4.67	1.48
GARCH(p: 1, q: 3)	Standardized Student's t distribution	11541.2	11581.4	10.06	2.36	4.59	1.48
GARCH(p: 1, q: 3)	Standardized Student's t distribution	11541.2	11581.4	10.06	2.36	4.59	1.48

TABLE III
IN SAMPLE RESULTS WITH 400 DAYS NON-OVERLAPPING BINS

Model	Distribution	AIC	BIC	Naive MSE	Naive MAE	1h MSE	1h MAE
EGARCH(p: 1, o: 1, q: 1)	Standardized Skew Student's t distribution	2023.3	2051.3	15.927	3.119	8.752	2.206
EGARCH(p: 1, o: 1, q: 1)	Standardized Student's t distribution	2023.9	2047.8	16.308	3.168	9.053	2.254
EGARCH(p: 5, o: 1, q: 1)	Normal distribution	2119.0	2154.9	9.332	2.322	5.148	1.534
EGARCH(p: 5, o: 1, q: 1)	Normal distribution	2119.0	2154.9	9.332	2.322	5.148	1.534
GARCH(p: 1, q: 3)	Standardized Student's t distribution	2038.1	2066.1	10.276	2.415	4.906	1.558
EGARCH(p: 5, o: 1, q: 1)	Normal distribution	2119.0	2154.9	9.332	2.322	5.148	1.534

TABLE IV
IN SAMPLE RESULTS WITH 600 DAYS NON-OVERLAPPING BINS

Model	Distribution	AIC	BIC	Naive MSE	Naive MAE	1h MSE	1h MAE
EGARCH(p: 2, o: 1, q: 1)	Standardized Skew Student's t distribution	3017.6	3052.8	17.212	3.355	8.993	2.437
EGARCH(p: 1, o: 1, q: 1)	Standardized Student's t distribution	3018.1	3044.4	20.708	3.671	12.027	2.776
EGARCH(p: 5, o: 1, q: 1)	Normal distribution	3184.7	3224.3	10.164	2.414	5.328	1.602
EGARCH(p: 5, o: 1, q: 1)	Normal distribution	3184.7	3224.3	10.164	2.414	5.328	1.602
GARCH(p: 1, q: 3)	Standardized Student's t distribution	3034.3	3065.0	10.553	2.461	5.19	1.617
EGARCH(p: 4, o: 1, q: 1)	Normal distribution	3188.4	3223.5	10.208	2.422	5.246	1.594

TABLE V
OUT-OF-SAMPLE PERFORMANCE OF BEST IN-SAMPLE MODELS

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GARCH(p: 4, q: 2)	Normal distribution	9.749	2.493	4.138	1.398
GARCH(p: 1, q: 3)	Standardized Student's t distribution	9.86	2.615	3.523	1.389
GARCH(p: 1, q: 1)	Standardized Student's t distribution	9.862	2.616	3.43	1.398
GARCH(p: 1, q: 3)	Standardized Skew Student's t distribution	9.823	2.613	3.529	1.389
GARCH(p: 1, q: 1)	Standardized Skew Student's t distribution	9.834	2.612	3.455	1.397
EGARCH(p: 5, o: 1, q: 1)	Normal distribution	9.551	2.485	4.233	1.474
EGARCH(p: 4, o: 1, q: 1)	Normal distribution	9.364	2.47	4.22	1.467
EGARCH(p: 2, o: 1, q: 1)	Standardized Student's t distribution	11.398	2.867	4.282	1.677
EGARCH(p: 1, o: 1, q: 1)	Standardized Student's t distribution	11.284	2.846	4.262	1.672
EGARCH(p: 2, o: 1, q: 1)	Standardized Skew Student's t distribution	11.328	2.858	4.295	1.675
EGARCH(p: 1, o: 1, q: 1)	Standardized Skew Student's t distribution	11.274	2.848	4.275	1.668
GJR-GARCH(p: 4, o: 1, q: 2)	Normal distribution	8.163	2.217	4.483	1.33
GJR-GARCH(p: 1, o: 1, q: 1)	Normal distribution	8.204	2.131	5.401	1.493
GJR-GARCH(p: 1, o: 1, q: 3)	Standardized Student's t distribution	10.039	2.617	4.964	1.579
GJR-GARCH(p: 1, o: 1, q: 1)	Standardized Student's t distribution	10.346	2.678	4.95	1.608
GJR-GARCH(p: 1, o: 1, q: 3)	Standardized Skew Student's t distribution	9.954	2.598	4.947	1.571
GJR-GARCH(p: 1, o: 1, q: 1)	Standardized Skew Student's t distribution	10.203	2.656	4.934	1.603

TABLE VI
OUT OF SAMPLE RESULTS, W = 1000, H = 1

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
EGARCH(p: 2, o: 1, q: 1)	Normal distribution	9.177	2.507	3.521	1.311
GJR-GARCH(p: 4, o: 1, q: 1)	Normal distribution	9.539	2.473	3.762	1.35
GARCH(p: 1, q: 1)	Normal distribution	9.249	2.501	3.377	1.297
GJR-GARCH(p: 3, o: 1, q: 1)	Normal distribution	9.344	2.48	3.39	1.291

TABLE VII
OUT OF SAMPLE RESULTS, W = 800, H = 1

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	9.383	2.509	3.541	1.326
GARCH(p: 4, q: 1)	Normal distribution	9.669	2.492	3.881	1.365
GARCH(p: 1, q: 2)	Normal distribution	9.492	2.545	3.42	1.33
GARCH(p: 3, q: 4)	Normal distribution	9.546	2.506	3.527	1.321

TABLE VIII
OUT OF SAMPLE RESULTS, W = 600, H = 1

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	9.571	2.506	3.522	1.297
EGARCH(p: 4, o: 1, q: 4)	Normal distribution	9.705	2.49	4.001	1.371
GARCH(p: 4, q: 1)	Standardized Skew Student's t distribution	9.942	2.615	3.430	1.43
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	9.571	2.51	3.522	1.297

TABLE IX
OUT OF SAMPLE RESULTS, W = 1000, H = 2, METHOD = BOOTSTRAP

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GARCH(p: 1, q: 2)	Normal distribution	8.989	2.467	3.364	1.298
GJR-GARCH(p: 4, o: 1, q: 1)	Normal distribution	9.42	2.455	3.771	1.357
GARCH(p: 1, q: 1)	Normal distribution	8.995	2.467	3.361	1.297
GARCH(p: 1, q: 1)	Normal distribution	8.995	2.467	3.361	1.297

TABLE X
OUT OF SAMPLE RESULTS, W = 800, H = 2, METHOD = BOOTSTRAP

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GARCH(p: 1, q: 1)	Normal distribution	9.207	2.506	3.407	1.33
GJR-GARCH(p: 4, o: 1, q: 3)	Normal distribution	9.656	2.469	3.973	1.383
GARCH(p: 1, q: 1)	Normal distribution	9.207	2.506	3.407	1.33
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	9.337	2.5	3.525	1.324

TABLE XI
OUT OF SAMPLE RESULTS, W = 600, H = 2, METHOD = BOOTSTRAP

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GJR-GARCH(p: 1, o: 1, q: 4)	Standardized Skew Student's t distribution	9.594	2.593	3.643	1.421
GJR-GARCH(p: 4, o: 1, q: 3)	Normal distribution	9.959	2.489	3.957	1.39
GARCH(p: 4, q: 1)	Standardized Skew Student's t distribution	9.844	2.603	3.448	1.433
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	9.666	2.515	3.513	1.311

TABLE XII
OUT OF SAMPLE RESULTS, W = 1000, H = 2, METHOD = SIMULATION

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GARCH(p: 1, q: 1)	Normal distribution	9.12	2.491	3.393	1.315
EGARCH(p: 4, o: 1, q: 1)	Normal distribution	9.318	2.456	3.841	1.373
GARCH(p: 3, q: 4)	Normal distribution	9.305	2.497	3.372	1.306
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	9.13	2.472	3.457	1.301

TABLE XIII
OUT OF SAMPLE RESULTS, W = 800, H = 2, METHOD = SIMULATION

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GARCH(p: 1, q: 1)	Normal distribution	9.205	2.506	3.413	1.33
GJR-GARCH(p: 4, o: 1, q: 3)	Normal distribution	9.644	2.468	3.971	1.382
GARCH(p: 1, q: 1)	Normal distribution	9.205	2.506	3.413	1.33
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	9.344	2.5	3.531	1.326

TABLE XIV
OUT OF SAMPLE RESULTS, W = 600, H = 2, METHOD = SIMULATION

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GJR-GARCH(p: 1, o: 1, q: 4)	Standardized Skew Student's t distribution	9.602	2.593	3.638	1.422
GJR-GARCH(p: 4, o: 1, q: 3)	Normal distribution	9.96	2.49	3.966	1.393
GARCH(p: 4, q: 1)	Standardized Skew Student's t distribution	9.841	2.602	3.449	1.433
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	9.667	2.515	3.515	1.311

TABLE XV
TEST SET PERFORMANCE, H = 1, W = 1000

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GARCH(p: 1, q: 1)	Normal distribution	8.003	2.359	2.922	1.343
GJR-GARCH(p: 3, o: 1, q: 1)	Normal distribution	8.349	2.393	2.826	1.34
EGARCH(p: 2, o: 1, q: 1)	Normal distribution	8.489	2.45	2.973	1.394

TABLE XVI
TEST SET PERFORMANCE, H = 2, W = 1000, METHOD = BOOTSTRAP - SIMULATION

Model	Distribution	Naive MSE	Naive MAE	1h MSE	1h MAE
GARCH(p: 1, q: 1)	Normal distribution	8.137	2.384	2.974	1.37
GJR-GARCH(p: 4, o: 1, q: 1)	Normal distribution	8.635	2.377	3.273	1.342
EGARCH(p: 3, o: 1, q: 1)	Normal distribution	8.337	2.419	3.012	1.374

Fig. 10. Test set Performance, h = 1, w = 1000

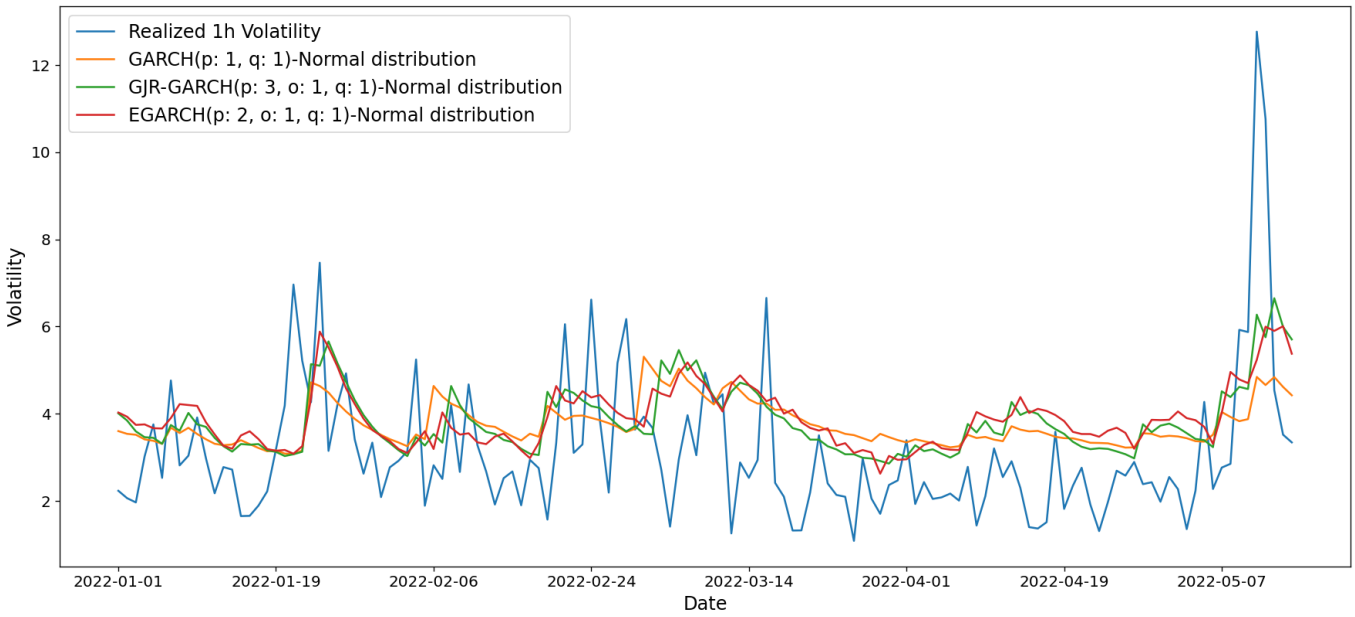


Fig. 11. Test set Performance, h = 2, w = 1000, method = Bootstrap and Simulation

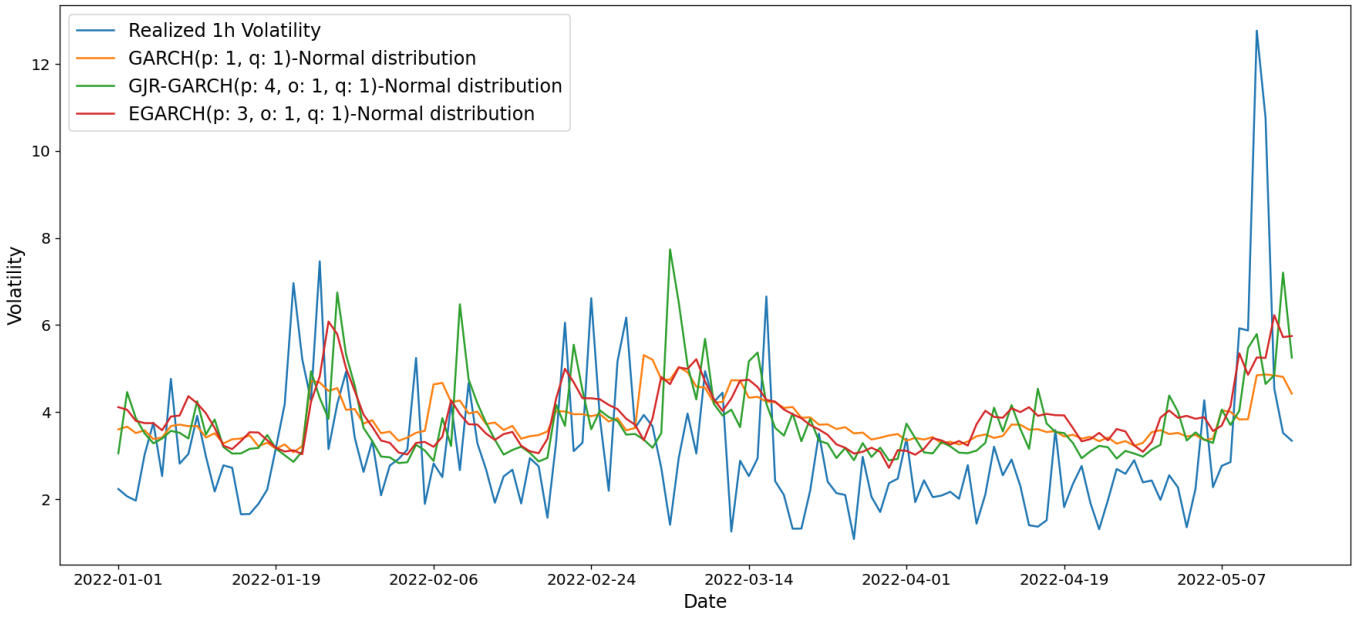


Fig. 12. Minute data manipulation