EC4304 Economic and Financial Forecasting

AY2019/20 Semester 1

Group Project Report

Arima Otsuka (A0147095Y)

Heng Jian Shun (A0166871U)

Kew Yu Jing (A0171351N)

Serena Lum (A0171623L)

Yu Xinyao (A0173372E)

Table of Contents

Abstract

The objective of this paper is to examine whether asymmetric (leverage) effects common in many asset classes are applicable in forecasting Bitcoin volatility, as well as to inspect the forecasting power of the models across various forecasting horizons. Specifically, we forecasted the daily realised variance of returns in one particular Bitcoin exchange, GEMINI with STATA. We did so by using two specified benchmark models, GARCH(1,1) and HAR, and the extension models from these two benchmarks. We then executed a pseudo out-of-sample forecast for both daily (1-step-ahead), weekly (5-steps-ahead) and monthly (22-steps-ahead) frequencies. Lastly, we produced combined forecasts and evaluated the performances of both combined forecasts and selected models. From this analysis, we are able to conclude from the DM-tests that asymmetric effects do not facilitate the forecast of Bitcoin volatility, and that both benchmark models provide limited insight into volatility when the forecast horizon increases to 22 days and beyond.

Section 1: Introduction

Background

Bitcoin is an expanding market which has one of the highest trading volumes when compared against other cryptocurrencies like Ethereum and Litecoin. Volatility has been rife throughout Bitcoin's short history; in recent times, speculation of an increase in volatility has gained traction following a prolonged period of stable trading prices. Meanwhile, the attractiveness of Bitcoin has this year resurfaced amongst investors; this contrasts with other currencies in the digital currency market which face headwinds to their development. For instance, in October'19, high-profile Multinational Corporations like Visa Inc., Mastercard Inc. and even the US Federal Reserve withdrew support for Facebook Inc.'s plan to launch the global cryptocurrency, Libra, over national security concerns (Hajric, 2019).

Despite Bitcoin being an increasingly popular platform for transaction, research regarding Bitcoin volatility forecasting to date has been lacklustre. Moreover, the high degree of transparency achieved through Bitcoin's publicly available transactions facilitates our foray into Bitcoin forecasting. Katsiampa encapsulates Bitcoin's attractiveness as an intriguing research focus in a nutshell:

> "*This can be attributed to its innovative features, simplicity, transparency and its increasing popularity, while since its introduction it has posed great challenges and opportunities for policy makers, economists, entrepreneurs, and consumers.*" (2017)

We now proceed to lay out the groundwork for volatility forecasting commonly used by practitioners in this area.

Volatility Forecasting

GARCH models are the traditional go-to models in volatility forecasting for conventional financial assets to date. Poon and Granger (2003) highlight in their review of financial asset volatility forecasting that GARCH is more parsimonious than ARCH and is the most popular model for many financial forecasts. For typical financial assets like exchange rates, Hansen and Lunde (2005) have not been able to identify evidence that suggests that more sophisticated models outperform the GARCH(1,1) model in forecasting exchange rates. As a result, early studies of Bitcoin forecasting have focused on GARCH-type models (Katsiampa, 2017).

In addition, with the increasing prevalence of high-frequency intraday data in today's era of big data, the relatively newer HAR model has become an important contender in volatility forecasting. As such, we used intra-daily squared returns as a proxy for volatility and extended our analysis to cover more models in the HAR category as well.

Asymmetric responses, which is a notable property in volatility forecasting, have been exploited by researchers in the domain of forecasting financial market variables (Corsi, 2012). In fact, one of our candidate models, the modified asymmetric HAR model, expands on existing literature which examined the persistence in leverage effects, with past negative returns correlated with current volatility (Corsi, 2012). The same asymmetric model is commonly used in equity volatility forecasting such as for stock indices like the Dow Jones Industrial Average, where negative returns are associated with higher volatility innovations than positive returns of the same magnitude (Diebold, 2001).

Comparison with other financial assets

Broadly, it has been acknowledged that Bitcoin's volatility largely surpasses that of other asset classes (Baur, 2017) - some reasons cited include its susceptibility to cyber-attacks, asymmetric information and absence of regulation. Bitcoin is known to have exhibited 30 times the volatility of foreign exchange markets and is hence deemed as having "excess volatility", according to Baur.

Compared to safe haven assets like gold and USD, Bitcoin's sensitivity and reaction render it highly susceptible to market movements. In terms of volatility distribution, Bitcoin displays more extreme observations than the foreign exchange market - its maximum and minimum values are registered as 10-fold higher than that of the EUR or JPY exchange rates (Baur, 2017). In addition, there are spikes in volatility that are only registered for Bitcoin but not for exchange rate and stock markets, owing to Bitcoin's lower degree of regulation (Baur, 2017).

Intended Outcomes

Moving forward, this paper acknowledges Bitcoin's unprecedented volatility levels and examines the robustness of the GARCH(1,1) and HAR models, which are staple benchmark models for volatility forecasting in other asset classes, in forecasting Bitcoin volatility. Our approach is twofold; we first examine if asymmetric (leverage) effects, which are unique to other asset classes, are significant to Bitcoin and thereafter proceed to inspect forecasts of increasing horizons to validate or disprove the predictive prowess of such models in day-, week- and month-ahead forecasting for Bitcoin data.

Section 2: Data & Methodology

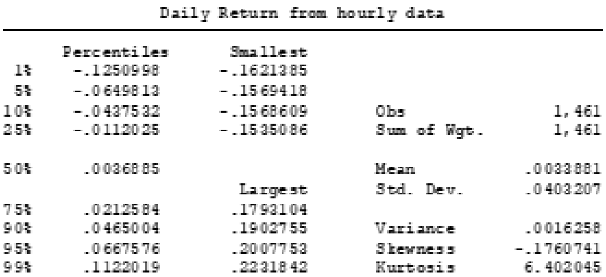
Data Source

We selected GEMINI from numerous other Bitcoin exchange platforms for two reasons - it not only has one of the widest databases publicly available and a sizeable dataset that houses intraday returns but is also a regulated financial services entity that has obtained a license from the New York Department of Financial Services (Roberts, 2015). This, as opposed to some other cryptocurrency digital platforms like Bitmex which are not regulated, provides GEMINI with an edge; the institutionalisation of Bitcoin has positioned it nearer to the forefront of trading today, attracting more mainstream investors and conferring a stronger reputation unto GEMINI. Henceforth, all subsequent mentions of Bitcoin in this paper refer solely to the GEMINI exchange.

Data Collection

To forecast Bitcoin's daily realised variance, we first extracted the hourly data of GEMINI via CryptoDataDownload, a free database established in 2017 for aggregated research purposes. Using the data obtained in hourly intervals, we worked with the hourly closing prices in USD between 8 October 2015 and 9 October 2019, with an estimated 35,000 observations. After transforming and consolidating the data using Microsoft Excel, we created the necessary variables used in our analysis by calculating hourly returns via computing the log difference of hourly prices. Next, we summed the 24 log differences (24 hours in a day) in closing prices to obtain the *daily returns (tr)*. We also computed the *daily realised variance (rv)* by summing the squared returns across 24 hours. These calculated values were then used in our forecasting analysis. For the 1-step-ahead forecasts, we reserved 300 observations (1162-1461) to form our pseudo-out-of-sample forecasts for comparison, while for the 5-step-ahead forecasts, we reserved 296 observations (1166-1461) instead. 279 observations (1183-1461) were reserved for the 22-step-ahead forecast.

Statistical Properties of Bitcoin



```
                  Daily Return from hourly data
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.1250998      -.1621385
 5%    -.0649813      -.1569418
10%    -.0437532      -.1568609       Obs               1,461
25%    -.0112025      -.1535086       Sum of Wgt.       1,461

50%     .0036885                      Mean            .0033881
                        Largest       Std. Dev.       .0403207
75%     .0212584       .1793104
90%     .0465004       .1902755       Variance        .0016258
95%     .0667576       .2007753       Skewness       -.1760741
99%     .1122019       .2231842       Kurtosis        6.402045
```

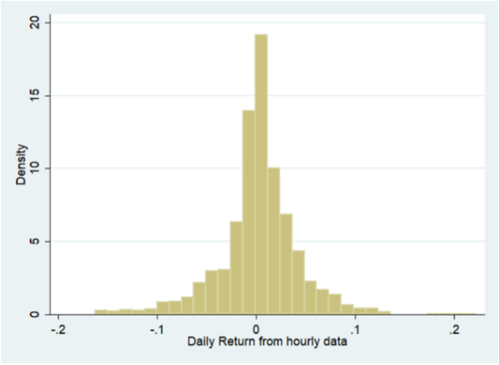


Figure 1: Properties of Bitcoin's daily returns

To have a better understanding of the data we obtained, we analysed the details of daily returns. From Figure 1, it is no surprise that Bitcoin is leptokurtic, having a kurtosis coefficient of more than 3. A high value of kurtosis and skewness could also be used to explain the distribution of realised variance, which does not follow a normal distribution (Baur, 2017) in our case.

Following this, we plotted the time series of realised variance and compared it against that of the S&P 500 index, widely recognised as the most common market proxy in stock markets. This data was obtained from the Oxford-Man Institute of Quantitative Finance's Realised Library.

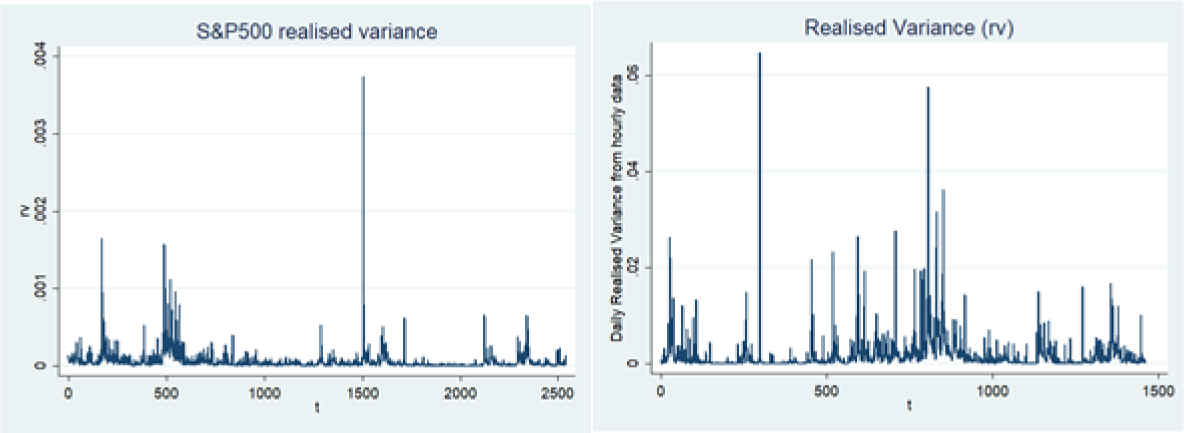


Figure 2: Comparison of realised variance between the S&P 500 (Left) and Bitcoin (Right)

From Figure 2, we see significantly more spikes for the realised variance of Bitcoin than the conventional S&P500 diagrams. Also, the magnitude of realised variance seems to be about 10 times higher than that of the S&P500, reinforcing the validity of our findings from the aforementioned research papers.

Methodology

Given the benchmark GARCH(1,1) and HAR models decided prior, we then evaluated the suitability of extension models for these benchmarks. For the GARCH family, we evaluated GARCH models up to order 2 - Threshold-GARCH (T-GARCH) and GARCH-in-mean (GARCH-M) models and selected the GARCH-M(1,1) model based on AIC values. For the HAR family, we expanded on asymmetric aspects and the use of RAV variables, generating asymmetric HAR, RAV-HAR, and asymmetric RAV-HAR models.

We then proceeded to evaluate the in-sample fit of all respective models for observations 1-1161, by observing the residuals and standardised residuals for each model. After generating our 1-step-ahead forecast using a Rolling Window Estimation, we proceeded to assess forecast optimality by checking for appropriate errors and systematic bias.

Next, we carried out forecast combination methods – namely Simple Averaging, Granger-Ramanathan, and Bates Granger – to improve the general forecasts for both 1-step ahead and 5-step ahead forecasts. To achieve this, we created a holdout sample of 100 observations (1062 to 1161) for the 1-step ahead forecasts, another containing 96 observations (1066 to 1161) for the 5-step ahead forecasts and one more containing 79 observations (1083 to 1161) for the 22-step-ahead forecasts. We then evaluated the forecast outputs by comparing the Root Mean Squared Errors across all forecasts made. Lastly, we evaluated the models via their loss functions.

We considered 2 loss functions - squared and QLIKE loss and ran the Diebold-Mariano tests to evaluate the ability of models to minimise risks. These loss functions were selected since they are robust to measurement errors. We then repeat the above procedure for 5 & 22-step ahead forecasts.

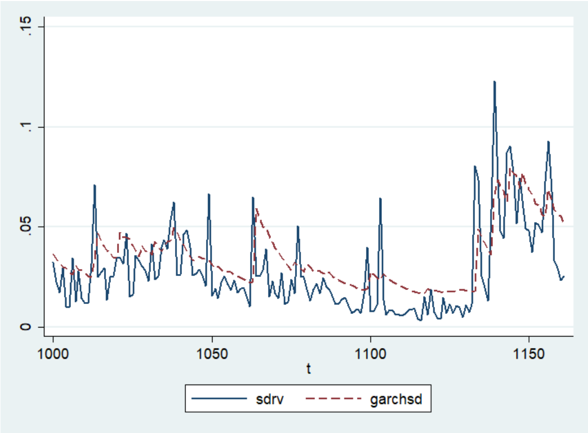
Section 3: Models

Benchmark (Vanilla) Models

In this section, we will be examining two categories of benchmark models, namely, GARCH (1,1) and the HAR model. Performance of the in-sample fitting will be examined, and h-step ahead forecasts will then be executed.

Benchmark GARCH(1,1) Model

GARCH(1,1), one of the most popular models used in financial time series modelling, suits financial data; like Bitcoin, data is generally collected at discrete intervals.



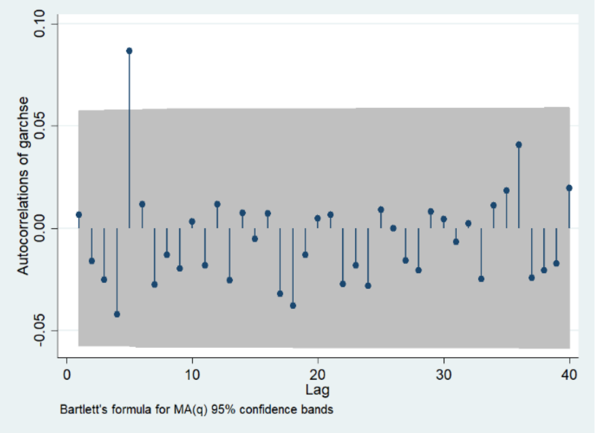


Figure 3 (left): Selected capture of bitcoin volatility December 2018 - October 2019 (in-sample fit using GARCH(1,1))

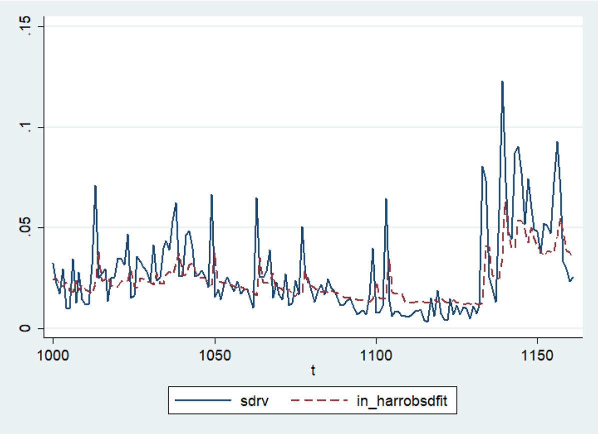
Figure 4 (right): ACF of GARCH (1,1) squared standardised residual

We fitted the in-sample data with GARCH (1,1) in Figure 3 and examined the fit. Focusing on in-sample data for periods 1000 (December 2018) to 1161 (October 2019), the data is fitted relatively well though it tends to overestimate periods with lower volatility and underestimate periods with higher volatility. Figure 4 indicates insignificant serial correlation in the residuals. From the Ljung-Box Q-test at lag 34, the null hypothesis that the residuals are white noise (p-value = 0.9339) cannot be rejected. From this, we infer that the GARCH (1,1) model has captured the ARCH effects nicely.

Benchmark HAR Model

The HAR model became increasingly popular among researchers, owing to its desirable properties of computational simplicity using fundamentals from the Ordinary Least Squares(OLS) regression and its out-of-sample performance (Xie, 2019). Robust regression is chosen for our models since it is the conventional method and robust to outliers, which are crucial in forecasting; according to Trucios, robust procedures outperform non-robust ones (2019). From Figure 5, we observed that the HAR (robust) fits the in-sample data relatively better than the GARCH (1,1) model, aside from underestimates of the spikes in volatility

.



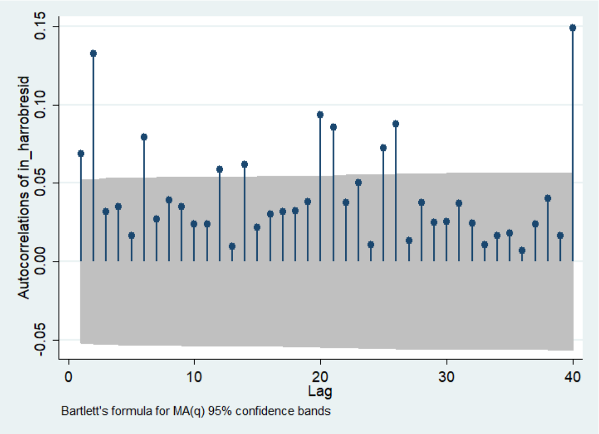


Figure 5 (Left): Bitcoin Volatility December 2018 - October 2019 (in-sample fit using HAR robust)

Figure 6 (Right): ACF of HAR robust residual

Figure 6 indicates significant serial correlation in the residuals. From the Ljung-Box Q-test at lag 33, the null hypothesis that the residuals are white noise (p-value= 0.0000) is rejected. We then tried fitting in-sample data with a HAR robust model using 7-day weeks and 30-day months to investigate any significant difference in performance.

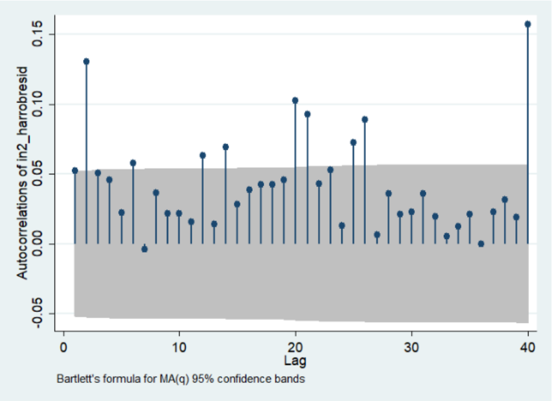
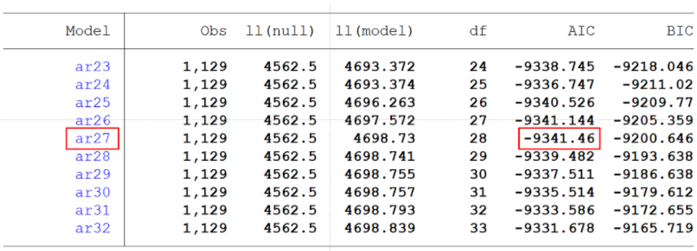


Figure 7: ACF of HAR robust (7-day weeks & 30-day months) residual

Figure 7 shows there is still a significant serial correlation in residuals. The Ljung-Box Q-test conducted at lag 33 again rejected the null hypothesis that the residuals are white noise (p-value= 0.0000). It appears that neither the 7-day nor 30-day months could provide a more well-behaved HAR model. Hence, the conventional 5-day weeks and 22-day months were chosen instead. The intuition behind this could be that Bitcoin trading activity on weekends were registered as relatively low in our working dataset and hence we could safely ignore these effects for the purpose of our research.



```
       Model │       Obs  ll(null)  ll(model)      df         AIC        BIC
─────────────┼──────────────────────────────────────────────────────────────
        ar23 │     1,129    4562.5   4693.372      24   -9338.745  -9218.046
        ar24 │     1,129    4562.5   4693.374      25   -9336.747   -9211.02
        ar25 │     1,129    4562.5   4696.263      26   -9340.526   -9209.77
        ar26 │     1,129    4562.5   4697.572      27   -9341.144  -9205.359
        ar27 │     1,129    4562.5    4698.73      28    -9341.46  -9200.646
        ar28 │     1,129    4562.5   4698.741      29   -9339.482  -9193.638
        ar29 │     1,129    4562.5   4698.755      30   -9337.511  -9186.638
        ar30 │     1,129    4562.5   4698.757      31   -9335.514  -9179.612
        ar31 │     1,129    4562.5   4698.793      32   -9333.586  -9172.655
        ar32 │     1,129    4562.5   4698.839      33   -9331.678  -9165.719
─────────────┴──────────────────────────────────────────────────────────────
```

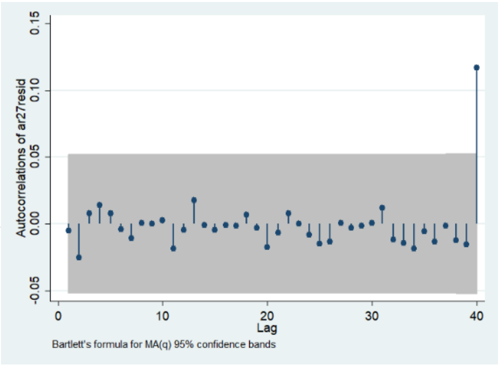
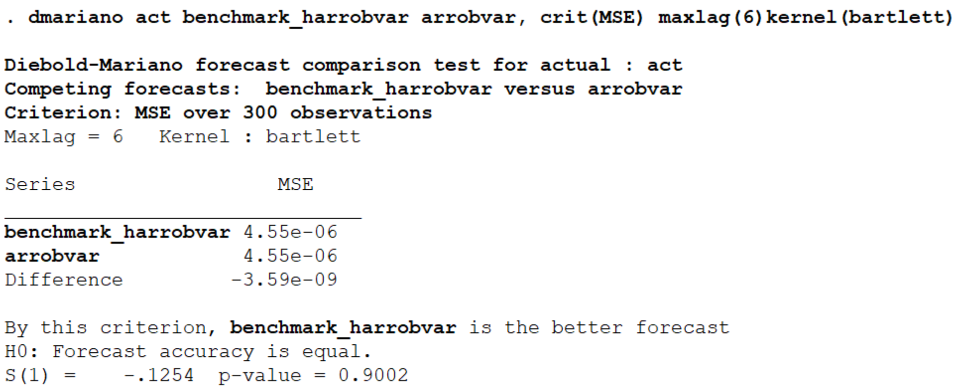


Figure 8 (Left): AIC values of AR models

Figure 9 (Right): ACF of AR(27) residual

The HAR benchmark model was selected on the basis that it is more parsimonious than an AR model. The comparison of AIC values across AR models with different lags in Figure 8 revealed that AR(27) minimises AIC (-9,341.46). Figure 9 reveals there is no significant serial correlation in the residuals of the AR(27) model. Hence, we conclude the model is well-fitted. We proceeded to compare the HAR (robust) model against the AR(27) model by creating a series of 1-step ahead forecasts and performing a Diebold-Mariano (DM) test.



```
. dmariano act benchmark_harrobvar arrobvar, crit(MSE) maxlag(6)kernel(bartlett)

Diebold-Mariano forecast comparison test for actual : act
Competing forecasts:  benchmark_harrobvar versus arrobvar
Criterion: MSE over 300 observations
Maxlag = 6   Kernel : bartlett

Series                   MSE
______________________________
benchmark_harrobvar 4.55e-06
arrobvar            4.55e-06
Difference         -3.59e-09

By this criterion, benchmark_harrobvar is the better forecast
H0: Forecast accuracy is equal.
S(1) =    -.1254  p-value = 0.9002
```

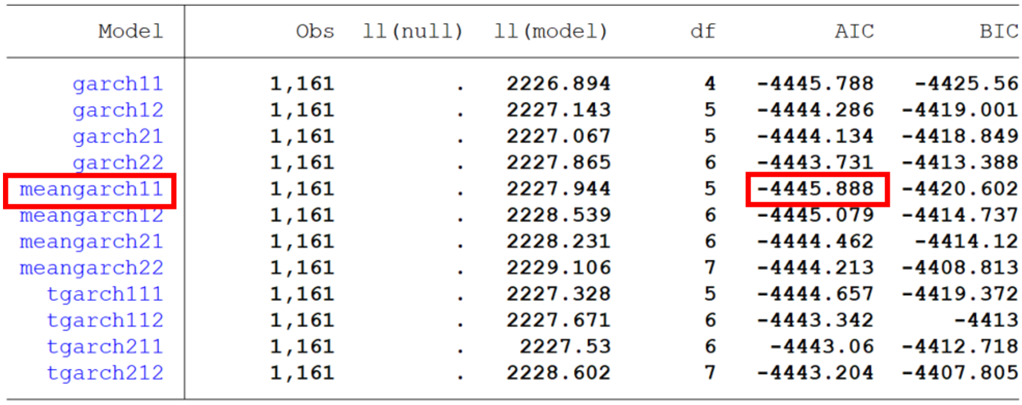
Figure 10: DM test for HAR (robust) and AR (27)

We are unable to reject the null hypothesis based on the DM test results (p-value= 0.9002) in Figure 10, indicating that the predictive ability of the two models are similar. Since the HAR model involves fewer parameters, it justifies being selected as a benchmark model. We then constructed pseudo out of sample 1-step ahead forecasts for the HAR (robust) model.

Extension Models

GARCH Extension Models (AIC Selection)

While the benchmark GARCH(1,1) model is simple and able to capture volatility clustering, it is unable to capture asymmetric behaviour perfectly (Kreinovich and Sriboonchitta, 2019). This could explain why our in-sample plot using GARCH (1,1) obtained above showed some slight deviations from the actual volatility. We then shifted our attention to GARCH models that account for asymmetric effects. Since there are numerous such models to choose from, we used the Akaike Information Criteria (AIC) for our model selection.



```
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
     garch11 |      1,161         .   2226.894       4   -4445.788   -4425.56
     garch12 |      1,161         .   2227.143       5   -4444.286  -4419.001
     garch21 |      1,161         .   2227.067       5   -4444.134  -4418.849
     garch22 |      1,161         .   2227.865       6   -4443.731  -4413.388
 meangarch11 |      1,161         .   2227.944       5   -4445.888  -4420.602
 meangarch12 |      1,161         .   2228.539       6   -4445.079  -4414.737
 meangarch21 |      1,161         .   2228.231       6   -4444.462   -4414.12
 meangarch22 |      1,161         .   2229.106       7   -4444.213  -4408.813
   tgarch111 |      1,161         .   2227.328       5   -4444.657  -4419.372
   tgarch112 |      1,161         .   2227.671       6   -4443.342      -4413
   tgarch211 |      1,161         .    2227.53       6    -4443.06  -4412.718
   tgarch212 |      1,161         .   2228.602       7   -4443.204  -4407.805
```

Figure 11: AIC statistics of models in the GARCH family

The extension models we considered from the GARCH family were conventional GARCH models up to order 2 and included asymmetric models such as GARCH-M and T-GARCH. The ARCH and GARCH effects above order 2 were not considered as lower-order GARCH models are typically well-equipped to sufficiently deal with variance dynamics and are hard to beat in forecasting. According to Figure 11 above, after comparing AIC values, GARCH-M(1,1) emerges with the lowest AIC (-4445.888) among the family of ARCH models and was thus selected.

HAR Extension Models

With the baseline HAR model as a new benchmark to beat in volatility forecasting, we have considered 3 other HAR extension models that include not only a different proxy but also asymmetric effects; these are the Asymmetric HAR, RAV-HAR and Asymmetric RAV-HAR. Of the three, the Asymmetric RAV-HAR is simply a combination of the former 2 models.

RAV-HAR:

$$rv_t = \alpha + \beta rav_{t-1} + \beta^{(5)} rav_{t-1}^{(5)} + \beta^{(22)} rav_{t-1}^{(22)} + \varepsilon_t$$

We substitute the realized variance proxy for another proxy known as the realised absolute variance (RAV), which is denoted by rav; studies have shown that it exhibits more persistent dynamics and greater robustness to noise and jumps than realised variance (Corsi & Reno, 2012). We compute daily RAV by summing up the absolute returns within a day, dividing the sum by square root of the number of intradaily observations (24 given hourly data) and multiplying the normalizing constant $\sqrt{\frac{\pi}{2}}$.

Asymmetric HAR:

$$rv_t = \alpha + \beta rv_{t-1} + \beta^{(5)} rv_{t-1}^{(5)} + \beta^{(22)} rv_{t-1}^{(22)} + \gamma tr_{t-1}^{-} + \gamma^{(5)} tr_{t-1}^{(5)-} + \gamma^{(22)} tr_{t-1}^{(22)-} + \varepsilon_t$$

In addition, studies have shown that volatility increases more with negative shocks compared to positive ones of equal magnitudes (Corsi & Reno, 2012). We account for this asymmetric effect in the models by including lagged negative returns at different frequencies as additional regressors (tr^{-}). We compute negative returns by replacing positive returns with zeros and keeping negative ones.

Asymmetric RAV-HAR:

$$rv_t = \alpha + \beta rav_{t-1} + \beta^{(5)} rav_{t-1}^{(5)} + \beta^{(22)} rav_{t-1}^{(22)} + \gamma tr_{t-1}^{-} + \gamma^{(5)} tr_{t-1}^{(5)-} + \gamma^{(22)} tr_{t-1}^{(22)-} + \varepsilon_t$$

By incorporating the asymmetric responses and realised absolute values, the models can then be better accounted for in high frequency data in continuous time (Corsi & Reno, 2012). Ergo, deploying the use of these extension models as a parsimonious approach allows us to avoid the long lags of the classic AR model while still summarising the high persistence in volatility.

Section 4: Results – 1-step-ahead forecasts

Forecast Optimality

We examine if the 1-step-ahead forecasts are optimal. Table 1 below details our findings.

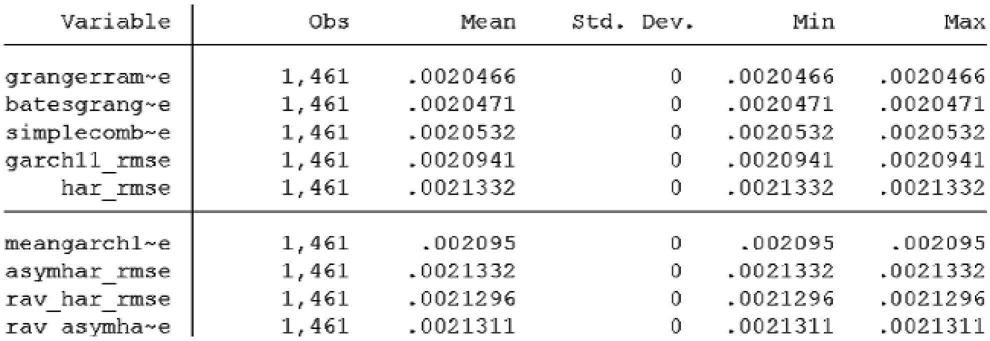
Model	**Forecast optimality** **(forecast error Q test p-value)**	**Check for systematic bias** **(MZ test p-value)**
GARCH (1, 1)	Forecast errors are white noise (0.1068)	No systematic bias (0.3051)
HAR	Forecast errors are not white noise (0.0432)	Systematic bias (0.0000)
GARCH-M (1,1)	Forecast errors are white noise (0.1233)	No systematic bias (0.2792)
RAV-HAR (robust)	Forecast errors are not white noise (0.0026)	Systematic bias (0.0000)
Asymmetric HAR (robust)	Forecast errors are not white noise (0.0341)	Systematic bias (0.0000)
Asymmetric RAV-HAR (robust)	Forecast errors are not white noise (0.0033)	Systematic bias (0.0000)

Table 1: Forecast Optimality across models

A clear distinction is seen; both GARCH models return optimal forecasts while all HAR models are sub-optimal.

Forecast Combination

Forecast combinations have been shown to improve forecasts significantly. Hence, we decided to generate a series of forecast combination methods to improve the 1-step-ahead forecasts. We have chosen simple averaging, Granger Ramanathan and Bates Granger forecast combinations as they generally work well in various settings. For the latter two methods, we first generated a hold-out sample of 100 observations before the pseudo-out-of-sample observations (observations 1061 to 1160) to generate the subsequent forecast combination weights; this optimises our combination approach; simply using actual pseudo-out-of-sample observations generated to compute weights would create bias towards forecasts' errors already produced.



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
grangerram~e |      1,461    .0020466           0   .0020466   .0020466
batesgrang~e |      1,461    .0020471           0   .0020471   .0020471
simplecomb~e |      1,461    .0020532           0   .0020532   .0020532
garch11_rmse |      1,461    .0020941           0   .0020941   .0020941
    har_rmse |      1,461    .0021332           0   .0021332   .0021332
-------------+---------------------------------------------------------
meangarch1~e |      1,461     .002095           0    .002095    .002095
asymhar_rmse |      1,461    .0021332           0   .0021332   .0021332
rav_har_rmse |      1,461    .0021296           0   .0021296   .0021296
rav asymha~e |      1,461    .0021311           0   .0021311   .0021311
```

Figure 12: RMSE comparison

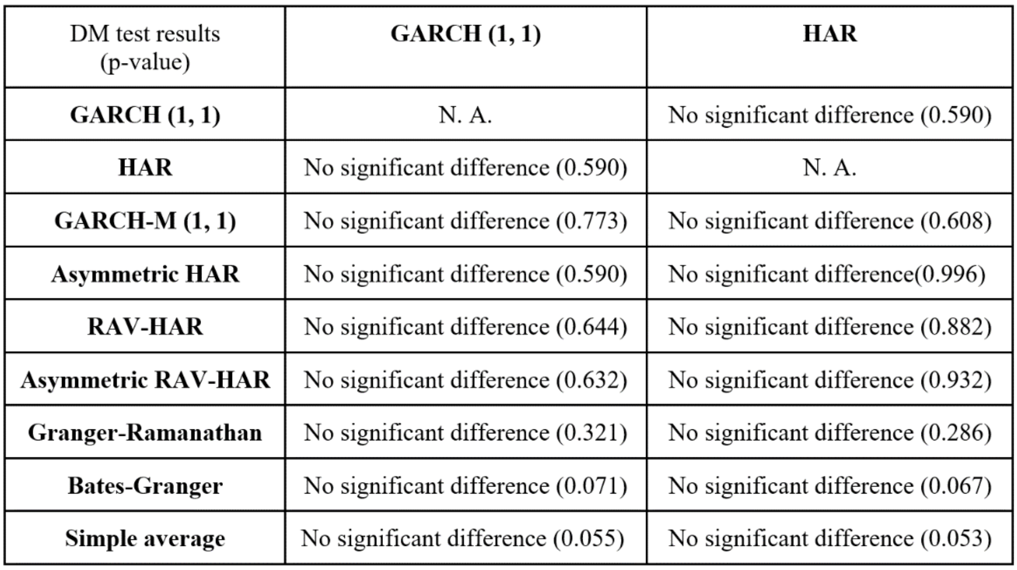
From Figure 12 above, all 3 combined forecasts return lower RMSEs than forecasts from the individual models. The Granger-Ramanathan forecast has the lowest RMSE followed by the Bates-Granger forecast and finally the Simple Average. To examine the extent to which such differences in RMSE are significant, we proceed to evaluate our forecasts.

Forecast Evaluation

In this section, models are evaluated under 2 loss functions which are more relevant to the models we have selected - Squared Loss and QLIKE Loss. Thereafter, 3 separate comparisons of forecast accuracy are made, one between the 2 benchmark models, next between the 4 selected models against both benchmark models and last pitting the 3 combined forecasts against both benchmark models. This is done through the Diebold-Mariano test.

Under Squared Loss

We first compared the forecast accuracy between both benchmark models. Since we are unable to reject the null hypothesis that there is no disparity in forecast accuracy, we conclude both benchmark models have similar forecast accuracy. We then proceed to compare the individual and combined models against the benchmark GARCH(1,1) model. We are unable to reject the null hypothesis that there is no disparity in forecast accuracy, implying that all 4 individual models and 3 combined models have equivalent forecast accuracy as the benchmark GARCH(1,1) model. Next, we compared the selected models against the benchmark HAR model. Again, the p-values reveal that we are unable to reject the null hypothesis that both the individual and combined models are better than the benchmark HAR model. Table 2 below shows the results.

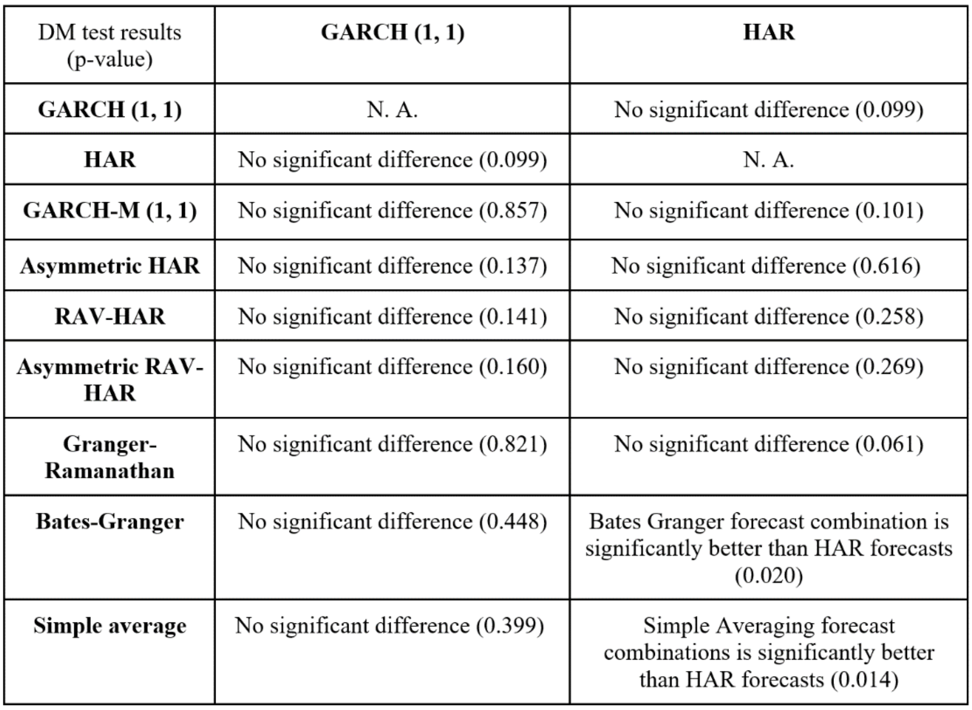


DM test results (p-value)	**GARCH (1, 1)**	**HAR**
GARCH (1, 1)	N. A.	No significant difference (0.590)
HAR	No significant difference (0.590)	N. A.
GARCH-M (1, 1)	No significant difference (0.773)	No significant difference (0.608)
Asymmetric HAR	No significant difference (0.590)	No significant difference(0.996)
RAV-HAR	No significant difference (0.644)	No significant difference (0.882)
Asymmetric RAV-HAR	No significant difference (0.632)	No significant difference (0.932)
Granger-Ramanathan	No significant difference (0.321)	No significant difference (0.286)
Bates-Granger	No significant difference (0.071)	No significant difference (0.067)
Simple average	No significant difference (0.055)	No significant difference (0.053)

Table 2: Comparing forecasting power across models

Under QLIKE Loss

First comparing between both benchmark models, we are still unable to reject the null hypothesis that there is no disparity between forecast accuracy. Comparing the selected models against the benchmark GARCH(1,1) models, similar to squared loss, all 4 models and forecast combinations have the same forecast accuracy as the GARCH(1,1) model as we are unable to reject the null hypothesis that there exists a disparity in forecast. Comparing the models against the benchmark HAR model, all 4 models are shown to have the same forecast accuracy as the benchmark HAR model. However, both the Bates-Granger and Simple Average combined forecasts return significantly lower RMSE than the benchmark HAR model. Table 3 below shows the results.



DM test results (p-value)	**GARCH (1, 1)**	**HAR**
GARCH (1, 1)	N. A.	No significant difference (0.099)
HAR	No significant difference (0.099)	N. A.
GARCH-M (1, 1)	No significant difference (0.857)	No significant difference (0.101)
Asymmetric HAR	No significant difference (0.137)	No significant difference (0.616)
RAV-HAR	No significant difference (0.141)	No significant difference (0.258)
Asymmetric RAV-HAR	No significant difference (0.160)	No significant difference (0.269)
Granger-Ramanathan	No significant difference (0.821)	No significant difference (0.061)
Bates-Granger	No significant difference (0.448)	Bates Granger forecast combination is significantly better than HAR forecasts (0.020)
Simple average	No significant difference (0.399)	Simple Averaging forecast combinations is significantly better than HAR forecasts (0.014)

Table 3: Comparing forecasting power across models

Forecast Plots

Based on our evaluation, we gather that not only do the different model extensions considered fail to perform significantly better than the benchmark models, they are also more complicated. As such, it seems that the benchmark models provide reasonable forecasts and we now compare their forecast plots against actual pseudo out of sample data. We also decided to examine the Granger-Ramanathan combined forecast plot since it returned the lowest RMSE. When constructing the forecast plot, we have opted to compare the forecast standard deviations against the actual Bitcoin standard deviation. Figure 13 below shows the plot chart.

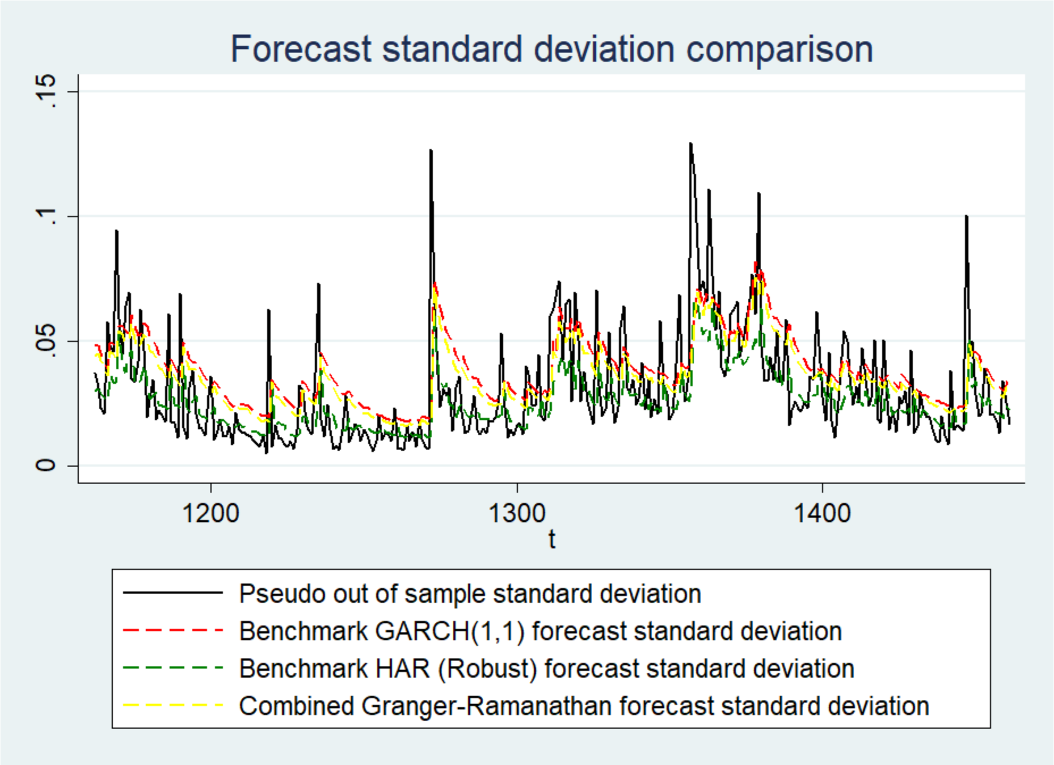


Figure 13: 1-step-ahead forecast standard deviation comparison

From Figure 13, three observations surface. Firstly, while all models' forecasts are able to capture the large overall dynamics of bitcoin volatility movement, they fail to capture finer fluctuations. In other words, all models seek to smooth out the jaggedness of bitcoin volatility data; however, it seems like the HAR model is able to retain a larger portion of the minute fluctuations. Secondly, for both the benchmark GARCH(1,1) and HAR models, a trade-off exists between forecasting different levels of volatility; GARCH(1,1) seems to better account for high volatility values but overestimates low volatility values while HAR underestimates high volatility values to a larger extent than GARCH models but more accurately reflects low volatility values. The Granger-Ramanathan combined forecast plot opts for a safer in-between value between both models; this is to be expected since the weights were derived from one model in the GARCH family and another in the HAR family. Finally, we noticed that all models' forecasts are unable to fit large spikes well, thereby underestimating large fluctuations in volatility. This seems reasonable since our chosen models were not adequately equipped to handle large spikes; in our limitations section below, we discuss the possibility of introducing jumps into our models, which could potentially mitigate this shortcoming.

<u>Section 5: Results – 5 & 22-step-ahead forecasts</u>

5-step-ahead Forecasts

Forecast Optimality

We examine if the 5 step-ahead forecasts are optimal. Table 4 below details our findings.

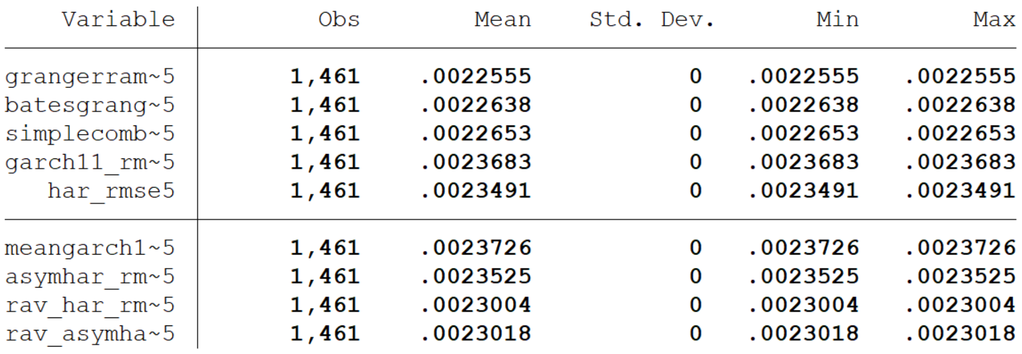
Model	Forecast optimality (forecast error Q test p-value)	Check for systematic bias (MZ test p-value)
GARCH(1,1)	Optimal – an MA(4) process fits the forecast errors nicely	Systematic bias (0.0055)
HAR	Optimal – an MA(4) process fits the forecast errors nicely	Systematic bias (0.0000)
GARCH-M(1,1)	Optimal – an MA(4) process fits the forecast errors nicely	Systematic bias (0.0047)
RAV-HAR (robust)	Optimal – an MA(4) process fits the forecast errors nicely	Systematic bias (0.0010)
Asymmetric HAR (robust)	Optimal – an MA(4) process fits the forecast errors nicely	Systematic bias (0.0007)
Asymmetric RAV-HAR (robust)	Optimal – an MA(4) process fits the forecast errors nicely	Systematic bias (0.0011)

Table 4: Checking for forecast optimality

We see that although all forecast errors are appropriate, all forecasts also contain systematic bias.

Forecast Combination

Similarly, as we have done in 1-step-ahead forecasts, we conducted various forecast combination methods - simple averaging, Granger Ramanathan and Bates Granger forecast combinations. The hold-out sample is 96, ranging from observations 1066 to 1161.



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
grangerram~5 |      1,461    .0022555           0    .0022555   .0022555
batesgrang~5 |      1,461    .0022638           0    .0022638   .0022638
simplecomb~5 |      1,461    .0022653           0    .0022653   .0022653
garch11_rm~5 |      1,461    .0023683           0    .0023683   .0023683
   har_rmse5 |      1,461    .0023491           0    .0023491   .0023491
-------------+---------------------------------------------------------
meangarch1~5 |      1,461    .0023726           0    .0023726   .0023726
asymhar_rm~5 |      1,461    .0023525           0    .0023525   .0023525
rav_har_rm~5 |      1,461    .0023004           0    .0023004   .0023004
rav_asymha~5 |      1,461    .0023018           0    .0023018   .0023018
```

Figure 14: 5-step-ahead forecast RMSE comparison

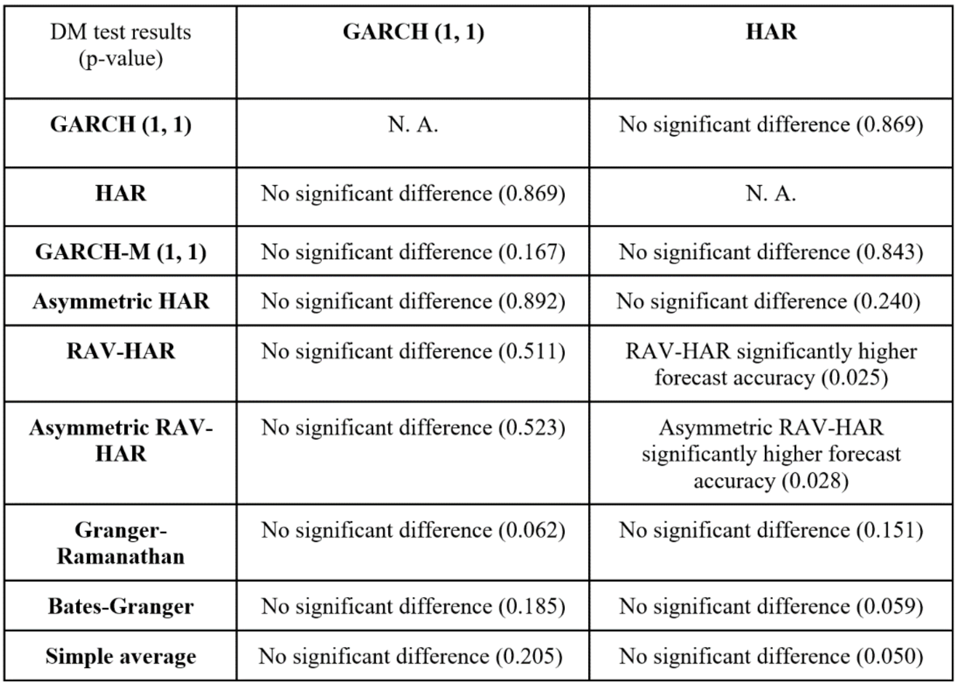
From Figure 14 above, all 3 combined forecasts return lower RMSEs than the individual models. The Granger-Ramanathan forecast has the lowest RMSE, followed by the Bates-Granger forecast, and finally the simple averaging. To examine the extent to which such differences in RMSE are significant, we proceed to evaluate our forecasts.

Forecast Evaluation

As in 1-step-ahead forecast evaluations, we compare the forecasting abilities of the models under squared loss and QLIKE loss. We compare the benchmark models against each other, and the selected models against the respective benchmark models.

Under Squared Loss

We first compared the forecast accuracy between both benchmark models. Since we are unable to reject the null hypothesis that there is no disparity in forecast accuracy, we conclude that both benchmark models have similar forecast accuracy. We then proceed to compare the selected models against the benchmark GARCH(1,1) model. We are unable to reject the null hypothesis that there is no disparity in forecast accuracy, implying that all selected models have equivalent forecast accuracy as the benchmark GARCH(1,1) model. Next, we compared the selected models against the benchmark HAR model. At the 5% level, the RAV-HAR and Asymmetric RAV-HAR forecasts tested to be significantly different from the Benchmark HAR forecast, with the former models having higher forecast accuracy. The remaining models are not significantly different from the benchmark HAR model. Table 5 below shows the results.

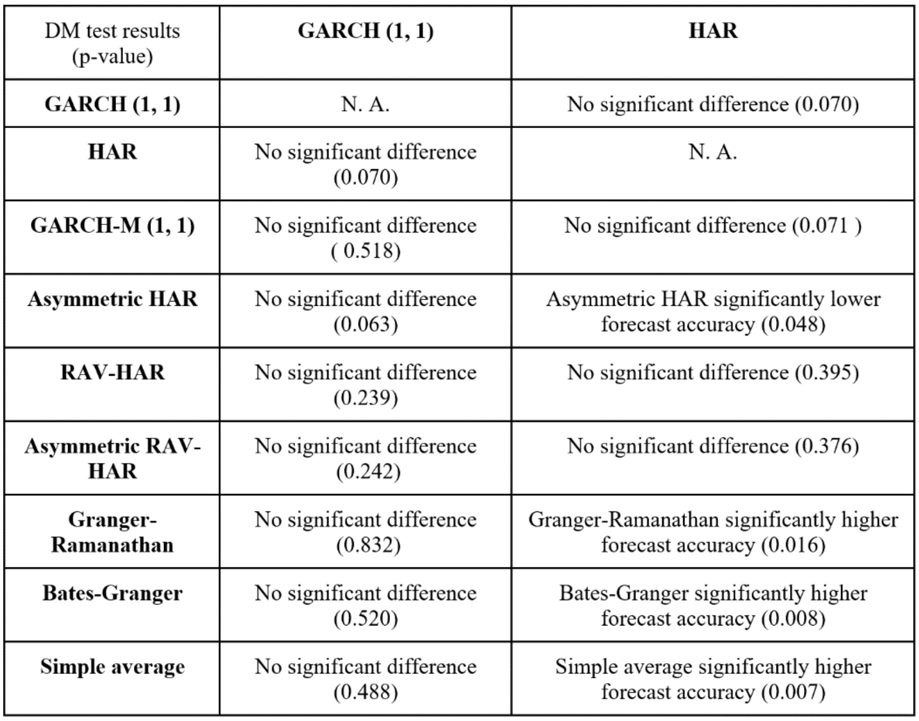


DM test results (p-value)	**GARCH (1, 1)**	**HAR**
GARCH (1, 1)	N. A.	No significant difference (0.869)
HAR	No significant difference (0.869)	N. A.
GARCH-M (1, 1)	No significant difference (0.167)	No significant difference (0.843)
Asymmetric HAR	No significant difference (0.892)	No significant difference (0.240)
RAV-HAR	No significant difference (0.511)	RAV-HAR significantly higher forecast accuracy (0.025)
Asymmetric RAV-HAR	No significant difference (0.523)	Asymmetric RAV-HAR significantly higher forecast accuracy (0.028)
Granger-Ramanathan	No significant difference (0.062)	No significant difference (0.151)
Bates-Granger	No significant difference (0.185)	No significant difference (0.059)
Simple average	No significant difference (0.205)	No significant difference (0.050)

Table 5: DM test results under squared loss

Under QLIKE Loss

We first compared the forecast accuracy between both benchmark models. Since we are unable to reject the null hypothesis that there is no disparity in forecast accuracy, we conclude that both benchmark models have similar forecast accuracy. We then proceed to compare the individual and combined models against the benchmark GARCH(1,1) model. Again, we are unable to reject the null hypothesis that there is no disparity in forecast accuracy, implying that all selected models have equivalent forecast accuracy as the benchmark GARCH(1,1) model. Next, we compared the selected models against the benchmark HAR model. At the 5% level, Asymmetric HAR, Granger Ramanathan, Bates-Granger and simple averaging are significantly different from the Benchmark HAR forecast. Asymmetric HAR does worse while the forecasts combinations outputs do better in terms of forecast accuracy. All other models are not significantly different from the benchmark HAR model. Table 6 below shows the results.



DM test results (p-value)	**GARCH (1, 1)**	**HAR**
GARCH (1, 1)	N. A.	No significant difference (0.070)
HAR	No significant difference (0.070)	N. A.
GARCH-M (1, 1)	No significant difference (0.518)	No significant difference (0.071)
Asymmetric HAR	No significant difference (0.063)	Asymmetric HAR significantly lower forecast accuracy (0.048)
RAV-HAR	No significant difference (0.239)	No significant difference (0.395)
Asymmetric RAV-HAR	No significant difference (0.242)	No significant difference (0.376)
Granger-Ramanathan	No significant difference (0.832)	Granger-Ramanathan significantly higher forecast accuracy (0.016)
Bates-Granger	No significant difference (0.520)	Bates-Granger significantly higher forecast accuracy (0.008)
Simple average	No significant difference (0.488)	Simple average significantly higher forecast accuracy (0.007)

Table 6: DM test results under QLIKE loss

Forecast Plots

Based on our evaluation, we gather that no extension or combined models have outperformed GARCH (1, 1) under either squared loss or QLIKE loss so we compare the GARCH (1, 1) forecast plot against the actual pseudo out of sample data. We decided to examine the simple average combined forecast plot since it produced the most significantly better forecast than the benchmark HARnmodel under QLIKE loss. We also decided to examine the RAV-HAR forecast plot since it produced the most significantly better forecast than benchmark HAR model under squared loss. When plotting the forecast plot, we have opted to compare the forecast standard deviations against the actual Bitcoin standard deviation. Figure 15 below shows the plot chart.

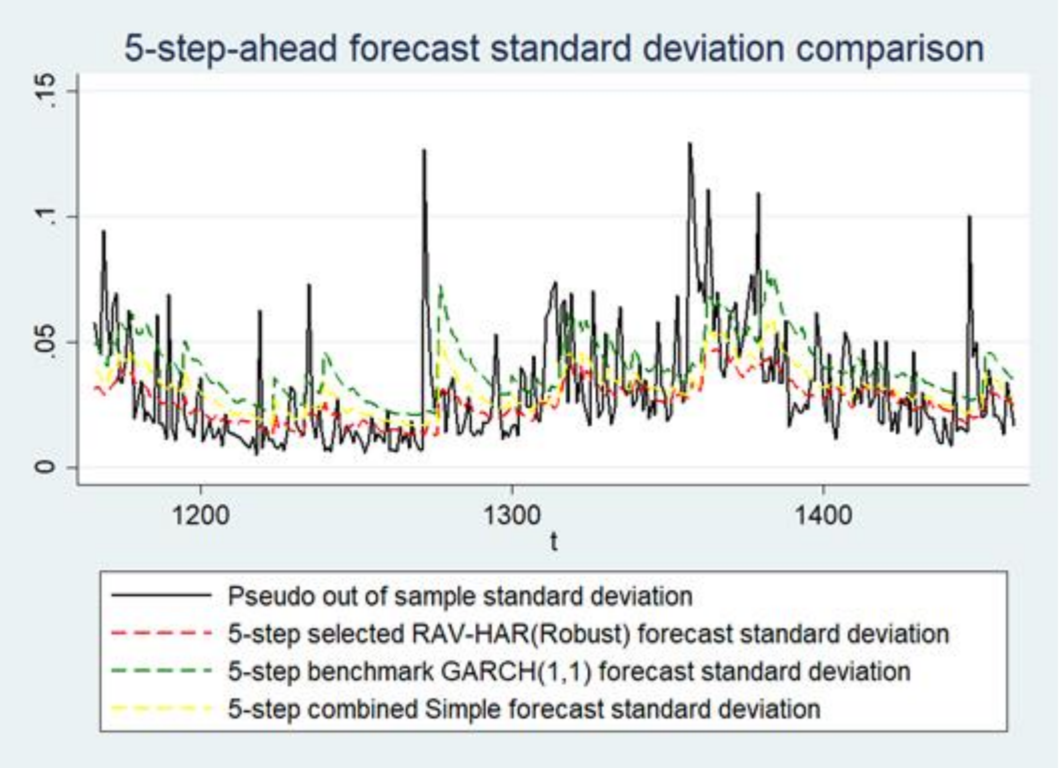


Figure 15: 5-step-ahead forecast standard deviation comparison

Similar to the performance of the 1-step-ahead forecasts, all 5-step-ahead forecasts, while somewhat able to capture the dynamics of low values of Bitcoin volatility, once again smooth out the original jaggedness of bitcoin volatility but only to a larger extent. Hence, the 5-step-ahead forecasts are less able to reflect the day-to-day changes in the actual volatility values compared to 1-step-ahead forecast. Meanwhile, we also observe a slight lag of the forecasts behind the actual values.

22-step-ahead Forecast

We constructed 22-step-ahead forecasts for both benchmark models. Based on the analysis detailed below, we believe that the 22-step-ahead forecasts offer limited insights about volatility. Hence, we do not detail discussions regarding forecast optimality or forecast evaluation for 22-step-ahead forecasts in this section.

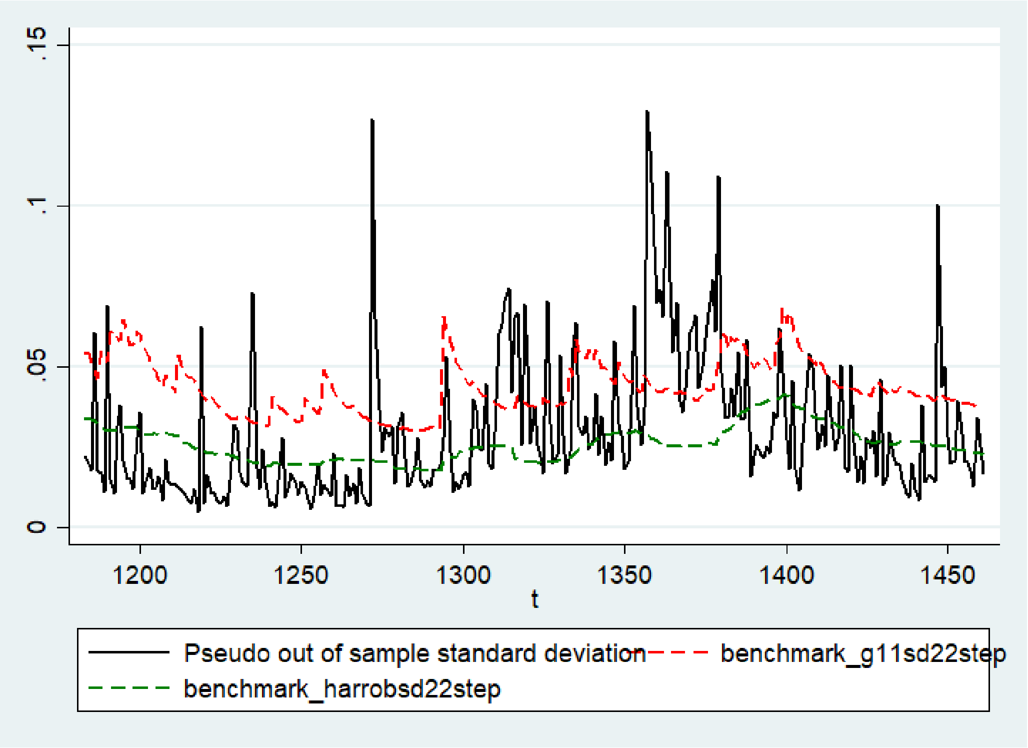


Figure 16: 22-step-ahead standard deviation comparison

We constructed the 22-step-ahead forecasts for both benchmark models and examined the plots with respect to the actual values. We plotted the two benchmarks as a basis to gauge the accuracy of our 22-step-ahead forecasts. The benchmark HAR forecast suggests a relatively constant level of volatility hovering around 0.025 throughout the pseudo out of sample period even in periods of significant volatility fluctuations. For example, the actual value of volatility decreased from 0.07 to 0.04 between t= 1350 and t=1450 yet the HAR forecast suggests a rather constant level of volatility around 0.04 during the same period. GARCH (1, 1)'s forecast similarly suggests a relatively constant level of volatility around 0.05 and reflects negligible day-to-day changes in volatility. While GARCH (1, 1) did indicate day-to-day changes, the direction of change is opposite to the actual change. For example, GARCH (1, 1) forecasts indicate a fall in volatility between t=1300 and t= 1325 when actual values increased. Hence, both benchmark models provide limited insight into changes in volatility as forecast horizon reaches 22 days. This leads us to believe that volatility forecasting of periods longer than 22 days will add little to no value to the analysis of Bitcoin's volatility dynamics.

Limitations

GEMINI has fewer volumes of Bitcoin exchanges (market capitalisation of USD7.7mn) than major cryptocurrency exchanges like BINANCE (market capitalisation of USD861mn). As a result, the fluctuations of Bitcoin in GEMINI may not reflect the actual volatility of Bicoin in general. However, given that GEMINI provides higher frequency data as compared to major cryptocurrency exchanges, in conjunction with the fact that major cryptocurrency exchanges like BINANCE and OKex only provides access to only 2 years of data, our team feels that this is an inevitable trade-off to ensure the optimal quality of data obtained and processed.

We obtained sub-optimal results, whereby the one-step ahead forecast error for HAR models are not MA(h-1), contrary to our theoretical understanding for all covariance stationary processes. Hence, we suspect that this is due to the inherent nature of Bitcoin in which the models are unable to capture properly. Perhaps, Corsi's proposed Leverage Heterogeneous Auto-Regressive with Continuous Volatility and Jumps (LHAR-CJ) model which exhibited a good forecasting performance for the S&P 500 stocks might have modelled and captured Bitcoin's volatility more closely, given Bitcoin's greater susceptibility to large spikes in volatility.

In addition, if we had access to data of higher-frequency intervals, say 5-minute data intervals, the forecast performance of our respective models would be strengthened. This could improve the in-sample fit of our HAR models.

Conclusion:

In summary, our research paper discussed the forecast of Bitcoin daily returns with the goal of examining the significance of asymmetry in forecast models and examining the forecasting ability of models across various forecast horizons. Employing models from both the GARCH and HAR families to examine the pseudo out of sample performance for 1-step, 5-steps and 22-steps ahead forecasts, our results reveal that as forecast horizons reach 22 days, the forecasts contribute little to understanding changes in volatility. DM test results also reveal that models containing asymmetric components generally do no better than the benchmark HAR or GARCH(1,1) models, revealing that asymmetric effects are not as relevant for Bitcoin as they are in other conventional assets.

On this note, we are aware that there is potential room for improvement and contribution to further research. Some researchers like (Peng, 2017) have combined the tried-and-tested ARCH related models with novel approaches like Machine Learning and Support Vector Regression (SVR), capitalising on 21st century technological advancements to enhance existing models in order to better capture the volatility of financial assets. The short history of Bitcoin provides a relatively limited amount of data to ascertain its dynamics compared to traditional financial assets. An increased amount of data could potentially offer a better understanding of the dynamics and in turn lead to the development of more accurate forecasting models.

Bibliography

Corsi, F & Reno, R. (2012). Discrete-Time Volatility Forecasting With Persistent Leverage Effect and the Link With Continuous-Time Volatility Modeling, Retrieved from *Journal of Business & Economic Statistics*, 30:3, 368-380.

Cryptodatadownload (2019). *US & UK Exchange Data.* **https://www.cryptodatadownload.com/data/northamerican/**

Dyhrberg, A. H. (2016a). Bitcoin, gold and the dollar–a garch volatility analysis. Retrieved from Finance Research Letters, 16, 85–92.

Forsberg, L., and Ghysels, E. (2007), "Why Do Absolute Returns Predict Volatil-ity so Well?" Retrieved from *Journal of Financial Econometrics*, 5, 31–67.

Hajric, V. (2019). *Bitcoin Volatility Set for Comeback After Trading Turned Boring.* Retrieved from https://www.bloomberg.com/news/articles/2019-10-21/bitcoin-volatility-set-for-comeback-after-trading-turned-boring.

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? Retrieved from *Journal of Applied Econometrics*, *20*(7), 873–889.

Katsiampa, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. Retrieved from *Economics Letters*, *158*, 3–6. doi: 10.1016/j.econlet.2017.06.023

Kreinovich, V., & Sriboonchitta, S. (2019). *Structural changes and their econometric modeling*. doi: https://doi.org/10.1007/978-3-030-04263-9_23.

Peng, Y. (2018) Forecasting High Frequency Volatility: A study of the Bitcoin Market using Support Vector Regression. Retrieved from *Expert Systems With Applications.* 97 177–192.

Poon, S., & Clive W. J. Granger. (2003). Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature,41*(2), 478-539. Retrieved from http://www.jstor.org.libproxy1.nus.edu.sg/stable/3216966.

Roberts, D. (2015). *With Gemini, Winklevoss brothers seek respect in bitcoin.* Retrieved from https://fortune.com/2015/10/05/gemini-winklevoss-bitcoin/

Trucois, C. (2019). Forecasting Bitcoin risk measures: A robust approach. Retrieved from *International Journal of Forecasting*,35, 836–847.

Williams, B. (2011, July 15). GARCH(1,1) models. Retrieved November 5, 2019, from https://math.berkeley.edu/~btw/thesis4.pdf.

Xie, Tian. "Forecast Bitcoin Volatility with Least Squares Model Averaging." *Econometrics*, vol. 7, no. 3, 2019, doi:10.3390/econometrics7030040.