1 2 UNIVERSIDADE D COIMBRA 9 0

Francisco André Tomé Pereira

FORECASTING BITCOIN REALIZED VOLATILITY: THE ROLE OF BLOCKCHAIN INFORMATION

Dissertação no âmbito do Mestrado em Métodos Quantitativos em Finanças, orientada pelo Professor Doutor Helder Miguel Correia Virtuoso Sebastião e pelo Professor Doutor Pedro Miguel Avelino Bação, apresentada ao Departamento de Matemática da Faculdade de Ciências e Tecnologia e à Faculdade de Economia.

Março de 2022

Forecasting Bitcoin Realized Volatility: The Role of Blockchain Information

Francisco André Tomé Pereira

1 2 9 0

UNIVERSIDADE D COIMBRA

Master in Quantitative Methods in Finance

Mestrado em Métodos Quantitativos em Finanças

MSc Dissertation | Dissertação de Mestrado

Março 2022

Acknowledgements

To all my teachers who accompanied me in these four and a half years. And in particular to my advisors, Professor Helder Sebastião and Professor Pedro Bação, who throughout this last semester, despite the difficulties arising from the pandemic, always supported me and were always available to help me.

To all my colleagues in the bachelor's and master's degrees that, in one way or another, contributed to my academic success and with whom I created memories that I am sure I will never forget.

Finally, to my family and friends that since the day I first entered the University of Coimbra until today have never stopped being by my side in good and bad times. And who always gave me the best advice and didn't just say what I wanted to hear.

Abstract

Predicting the realized volatility of Bitcoin has become an increasingly and recurrent subject in the literature. This dissertation, as in most of the papers that adress this issue, uses HAR-type models. Using data from nine exchanges from January 1, 2015, until October 19, 2021, several models were implemented in order to find out which variables are the most important in predicting 1-day ahead volatility. One of the main objectives of this work is to find out if Blockchain and other market information are relevant to predict future volatility. The results point out that the models where Blockchain information is introduced do not present more accurate results, and that the HAR-J-LN is the best model, meaning that log transformation of realized volatility and including jumps are important aspects when forecasting the realized volatility of Bitcoin.

Resumo

Prever a volatilidade realizada da Bitcoin tornou-se um assunto cada vez mais recorrente na literatura. Esta dissertação, tal como outros trabalhos que estudam este tema, utiliza modelos do tipo HAR. Utilizando dados de nove bolsas, de 1 de janeiro de 2015 até 19 de outubro de 2021, foram implementados vários modelos a fim de descobrir quais as variáveis mais importantes na previsão da volatilidade. Um dos principais objectivos deste trabalho é descobrir se a informação da Blockchain e outra informação de mercado são relevantes na previsão da volatilidade futura. Os resultados permitem concluir que os modelos onde a informação da Blockchain é introduzida não apresentam resultados mais precisos, e que o modelo HAR-J-LN é o melhor modelo, o que significa que a transformação logarítmica da volatilidade realizada e a inclusão da variável *jumps* são aspetos importantes, na previsão da volatilidade realizada da Bitcoin.

Table of Contents

List of Acronyms **xi**

List of Figures **xiii**

List of Tables **xv**

1 Introduction **1**

2 Literature Review **3**

3 Data and Preliminary Analysis **7**

3.1 Data 7

3.2 Preliminary Analysis 9

4 Methodology **11**

4.1 Variables 11

4.1.1 Realized Volatility 11

4.1.2 Trading Volume 12

4.1.3 Blockchain and other Market Information 12

4.1.4 Jumps 12

4.2 HAR-type models 13

5 Forecasting Performance **19**

6 Conclusion **23**

References **25**

Appendix A Description of market and blockchain variables **27**

List of Acronyms

- AR: AutoRegressive
- GARCH: Generalized AutoRegressive Conditional Heteroskedasticity
- GARCH-MIDAS: Generalized AutoRegressive Conditional Heteroskedasticity-Mixed Data Sampling
- HAR: Heterogeneous AutoRegressive
- HAR-CJ: Heterogeneous AutoRegressive-Continuous volatility and Jumps
- HAR-CJ-EPU: Heterogeneous AutoRegressive-Continuous volatility and Jumps-Economic Policy Uncertainty
- HAR-CJ-L/LHAR-CJ: Heterogeneous AutoRegressive-Continuous volatility and Jumps-Leverage/ Leverage-Heterogeneous AutoRegressive-Continuous volatility and Jumps
- HAR-DUJ: Heterogeneous AutoRegressive-Downside and Upside Jumps
- HAR-DUJ-NT: Heterogeneous AutoRegressive-Downside and Upside Jumps-Number of Transactions
- HAR-Full: Heterogeneous AutoRegressive with l=[1,2,...,30]
- HAR-J: Heterogeneous AutoRegressive-Jumps
- HAR-J-LN: Heterogeneous AutoRegressive-Jumps-Logarithm
- HAR-L/LHAR: Heterogeneous AutoRegressive-Leverage/Leverage-Heterogeneous AutoRegressive
- HAR-LN: Heterogeneous AutoRegressive-Logarithm
- HAR-NT: Heterogeneous AutoRegressive-Number of Transactions
- HAR-PC: Heterogeneous AutoRegressive-Principal Components
- HAR-PC-J: Heterogeneous AutoRegressive-Principal Components-Jumps
- HAR-PC-J-LN: Heterogeneous AutoRegressive-Principal Components-Jumps-Logarithm

- HAR-PC-LN: Heterogeneous AutoRegressive-Pricipal Components-Logarithm
- HARQ: Heterogeneous AutoRegressive Quarticity
- HARQ^F: Heterogeneous AutoRegressive Quarticity Fully Adjusted
- HARQ-J: Heterogeneous AutoRegressive Quarticity-Jumps
- HARQ-J^{MAC}: Heterogeneous AutoRegressive Quarticity-Jumps-Model Averaging Coefficient
- HARQ^{MAC}: Heterogeneous AutoRegressive Quarticity-Model Averaging Coefficient
- HARQ-RS: Heterogeneous AutoRegressive Quarticity-Realized Semi-Variances
- HARQ-RS^{MAC}: Heterogeneous AutoRegressive Quarticity-Realized Semi-Variances-Model Averaging Coefficient
- HAR-RS: Heterogeneous AutoRegressive-Realized Semi-Variances
- HAR-RS-I: Heterogeneous AutoRegressive-Realized Semi-Variances (Type I)
- HAR-RS-II: Heterogeneous AutoRegressive-Realized Semi-Variances (Type II)
- HAR-V: Heterogeneous AutoRegressive-Volume
- HAR-V-J: Heterogeneous AutoRegressive-Volume-Jumps
- HAR-V-J-LN: Heterogeneous AutoRegressive-Volume-Jumps-Logarithm
- HAR-V-LN: Heterogeneous AutoRegressive-Volume-Logarithm
- HAR-V-PC: Heterogeneous AutoRegressive-Volume-Principal Components
- HAR-V-PC-J: Heterogeneous AutoRegressive-Volume-Principal Components-Jumps
- HAR-V-PC-J-LN: Heterogeneous AutoRegressive-Volume-Principal Components-Jumps-Logarithm
- HAR-V-PC-LN: Heterogeneous AutoRegressive-Volume-Principal Components-Logarithm
- H-MAHAR: Heteroskedasticity-robust Model Averaging Heterogeneous AutoRegressive
- HRCP: Heteroskedasticity-robust Mallows' C_p
- JMA: Jackknife Model Averaging
- LASSO: Least Absolute Shrinkage and Selection Operator
- MAHAR: Model Averaging Heterogeneous AutoRegressive
- MSE: Mean Squared Error
- TGARCH: Treshold Generalized AutoRegressive Conditional Heteroskedasticity
- VIX: Chicago Board Options Exchange Volatility Index

List of Figures

3.1 Training Sample, Validation Sample and Test Sample 7
3.2 Volume Weighted Bitcoin Daily Prices in USD 8
3.3 Realized Volatility of Bitcoin 9

List of Tables

2.1 List of Studies on Forecasting Realized Volatility of Bitcoin (in alphabetical order) . 5

3.1 Descriptive Statistics of Realized Volatility 9

5.1 Result for 1-day Ahead Regressions by OLS 20
5.2 Forecasting Performance of each Model 21
5.3 Diebold-Mariano Tests 22

A.1 Description of the Variables Collected from Coinmetrics 27

Chapter 1

Introduction

Satoshi Nakamoto, founder of Bitcoin, describes it as a peer-to-peer version of electronic cash that allows online payments to be sent directly from one party to another without going through a financial institution, leading to lower costs and a more streamlined process (Nakamoto [21]).

Unlike regular currencies such as the Euro or the Dollar, Bitcoin's activity happens in a public decentralized ledger, known as Blockchain, which means that all users collectively retain control, and changes in rules only occur when the majority approves. This ledger provides pseudo-anonymity (users cannot be easily identified unless they reveal their identity) and avoids double-spending, i.e., a user cannot use the same cryptocurrency more than once (Lansky [18]). The security of Blockchain is guaranteed by "miners" (users who generate new cryptocurrency units by solving cryptographic problems on their computers), who in exchange for Bitcoin units, ensure the integrity of the ledger.

As time went by and Bitcoin became more popular, more cryptocurrencies started to appear. According to CoinMarketCap, assessed on February 1, 2022, there are more than 17250 cryptocurrencies with a total market capitalization of approximately $1.760.150.348.599. Bitcoin, the number one crypto in terms of market capitalization and price, represents 41.3% with a market capitalization of $726.5B, followed by Ethereum ($328.17B) and Binance Coin ($63.31B).

The main objective of this work is to find out if Blockchain and other market information can help predict the realized volatility of Bitcoin (BTC/USD), something that has not been much explored in the literature. To this end, we will implement several HAR models on the realized volatility of Bitcoin using also several exogenous variables, such as volume, jumps, and Blockchain information. In addition to the 16 models applied, we also created two other prediction schemes (from those generated by the models): the arithmetic mean and a weighted mean. Another differentiating aspect of this work is the fact that we collect data from nine different exchanges, which in our view, leads to a better representation of the overall Bitcoin market.

The remainder of this paper is organized as follows. Chapter 2 presents a literature review focusing on forecasting the realized volatility of Bitcoin. Chapter 3 describes the dataset and performs a preliminary analysis. Chapter 4 explains the methodology used in this work. Chapter 5 presents the results and chapter 6 concludes.

Chapter 2

Literature Review

Since its inception in 2008, when Satoshi Nakamoto created Bitcoin, cryptocurrencies have been the subject of numerous studies, the most important ones for this work being those conducted from a financial perspective. Some of these studies focus on the speculative nature of Bitcoin. For example, Fry and Cheah [13] concludes that Ripple and Bitcoin have experienced negative bubbles (from 2014 onwards), while Cheah and Fry [8] concludes that Bitcoin fundamental value is zero and speculative bubbles exist, and Chaim and Laurini [7] provides evidence that a bubble existed from early 2013 to mid-2014.

Other studies, for example, Kyriazis [16], Fang et al. [12], and Matkovskyy et al. [20], focus on Bitcoin's relationship with other markets. All conclude that Bitcoin may act as a hedge against uncertainty in traditional markets (just like gold). The latter also concludes that monetary policy shocks increase Bitcoin volatility. On the other hand, Klein et al. [15] concludes that Bitcoin is not the new gold as it shows a positive coupling effect and declines when markets are in a downward trend.

Another of the most explored themes in the literature is modeling and forecasting cryptocurrencies volatility using GARCH-type models. Baur and Dimpfl [3] and Bouri et al. [4] study the existence of asymmetry in volatility by employing a TGARCH and an Asymmetric-GARCH, respectively. Baur and Dimpfl [3] find that volatility increases more in response to positive shocks than in response to negative shocks, implying an asymmetric effect in most cryptocurrencies, except in Bitcoin and Ethereum. Bouri et al. [4] explore Bitcoin volatility and conclude that the previous phenomenon only happened before the 2013 price crash.

To test which model best forecasts one-step-ahead volatility and Value-at-Risk (VaR) in the Bitcoin market, Trucíos [24] implements a wide range of GARCH-type models. His paper shows that robust procedures outperform non-robust ones and highlights the importance of outliers while modeling and forecasting Bitcoin volatility measures. Also, Ardia et al. [2] conclude that Markov-switching GARCH models outperform standard GARCH models when forecasting the one-day-ahead VaR and find evidence of regime changes in Bitcoin volatility dynamics. Köchling et al. [17] conclude that it's not an easy task to choose one model that outperforms the others.

Conrad et al. [9] and Walther et al. [25] apply the GARCH-MIDAS model to determine the potential drivers of Bitcoin volatility. The first considers the VIX, risk in the US stock market, and a measure of global economic activity. Their findings support the evidence that Bitcoin volatility is pro-cyclical, i.e., increases with higher levels of economic activity. Walther et al. [25] applies the model to five cryptocurrencies (Bitcoin, Ethereum, Litecoin, Ripple, and Stellar) and concludes that Global Real Economic Activity is the most significant exogenous driver of Bitcoin volatility.

Similar to GARCH models, HAR models are also widely explored in the literature (e.g., Qiu et al. [22], Bouri et al. [5] and Aalborg et al. [1]). The standard in these models is to use realized volatility (square root of the sum of squared returns within a fixed period) as the dependent variable, following Corsi [10]. Returns are calculated using, most of the time, 5-minute data as it seems to be the best option, according to Liu et al. [19]. This model builds on the assumption of three different types of investors creating three different types of volatility: short-term, medium-term, and long term, hence the use of daily, weekly, and monthly realized volatility.

Qiu et al. [22] implements the HAR model along with variations such as HAR-J (adding jumps), HAR-RS (decomposing the variance into two signed semi-variances), and the HARQ-type models (adding realized quarticity). They also apply the MAC estimator that allows for model specification uncertainty and aims to minimize the MSE of the coefficients, concluding that $HARQ^{MAC}$ models provide the best forecasts.

Other approaches can be used to improve the accuracy of forecasts. Bouri et al. [5] uses Machine-Learning techniques, such as Random Forests, to analyze the role of the US-China trade war in forecasting out-of-sample daily realized volatility of Bitcoin returns. The authors extend the HAR model to include a metric of US-China trade tensions (based on Google Trends), jumps, realized skewness, and realized kurtosis. They concluded that US-China trade uncertainty improves the accuracy of volatility forecasts. Gkillas et al. [14] also uses Random Forests alongside with the inclusion of jumps, which improves out-of-sample forecast accuracy, according to the authors.

Xie [26] implements a wide range of methods to predict Bitcoin volatility. The findings support the evidence that there is excessive model uncertainty when modeling Bitcoin volatility by conventional regression methods and that H-MAHAR performs significantly better than conventional regressions at a 5% level.

Aalborg et al. [1] base their volatility models on the HAR model, including additional variables to see if any of them can improve Bitcoin daily forecastability: Google Trends, transaction volume (Bitcoins exchanged for goods or services), trading volume, unique addresses, changes in the VIX Index, and returns. They concluded that only trading volume is significantly correlated with Bitcoin daily volatility and can help predict it. On a similar approach, Yu [27] investigates the impacts of leverage effect and economic policy uncertainty on Bitcoin volatility, using HAR-type models. The author finds that adding those variables to the benchmark model can improve predictions.

Table 2.1 presents some studies on forecasting the realized volatility of Bitcoin, highlighting the sample period, data frequency, data source, models used and main conclusions.

Table 2.1 List of Studies on Forecasting Realized Volatility of Bitcoin (in alphabetical order)

Authors	Sample	Data Frequency	Data Source	Models	Main Conclusions
Aalborg et al. [1]	March 1, 2012 - March 19, 2017	10 min	Bitcoincharts	HAR	Daily volatility is correlated with and can be predicted by the trading volume of Bitcoin
Bouri et al. [5]	July 1, 2017 - June 30, 2016	60 min	CryptoCompare	HAR, HAR-J, HAR-RS	US-China trade war improves forecast accuracy
Catania and Sandholdt [6]	September 13, 2021 - March 18, 2021	5 min	Bitstamp, Coinbase	HAR, HAR-L, HAR-J, HAR-CJ, HAR-CJ-L	HAR-L and HAR-CJ-L are the best performers. Predictability of Bitcoin realized variance is increased over time, and predictability is higher for lower forecast horizons
Gkillas et al. [14]	January 1, 2014 - March 7, 2020	60 min	Bittrex	HAR, HAR-NT, HAR-DUJ, HAR-DUJ-NT	Random Forests based on HAR, with inclusion of transaction activity and jumps improves forecast accuracy
Qiu et al. [22]	October 10, 2017 - October 10, 2018	5 min	Binance	HAR, HARQ, $HARQ^{F}$, HAR-J, HARQ-J, $HARQ\text{-}J^{F}$, HAR-RS, HARQ-RS, $HARQ\text{-}RS^{F}$, $HARQ^{MAC}$, $HARQ\text{-}J^{MAC}$, $HARQ\text{-}RS^{MAC}$	$HARQ^{MAC}$ models demonstrate superior forecasting performance

Continues on the next page

Authors	Sample	Data Frequency	Data Source	Models	Main Conclusions
Xie [26]	January 1, 2018 - December 20, 2018	5 min	Binance	AR(1), HAR-Full, HAR, HAR-J, HAR-CJ, HAR-RS-I, HAR-RS-II, HAR-SJ-I, HAR-SJ-II, LASSO, MAHAR, HRCP, JMA, H-MAHAR	Model averaging methods outperform conventional regressions. H-MAHAR performs significantly better than conventional regressions at the 5% level
Yu [27]	March 1, 2013 - December 30, 2018	5 min	Bitcoincharts	HAR, HAR-CJ, LHAR-CJ, HAR-CJ-EPU, LHAR	Leverage effect has significant impact on future bitcoin volatility. Adding leverage effect and economic policy uncertainty can improve predictive ability

Chapter 3

Data and Preliminary Analysis

3.1 Data

In this dissertation, the object of study is Bitcoin (BTC) volatility in the period from January 1, 2015, to October 19, 2021. The sample was divided into three sub-samples. The first, "Training-sample”, corresponds to the period from January 1, 2015, to May 26, 2018. The second, “Validation-sample”, covers the period from May 27, 2018, to February 6, 2020. The third sub-sample, "Test-sample" covers the remaining period. This data partition implies a 50/25/25 split, as it is common in Machine-Learning applications.[1]

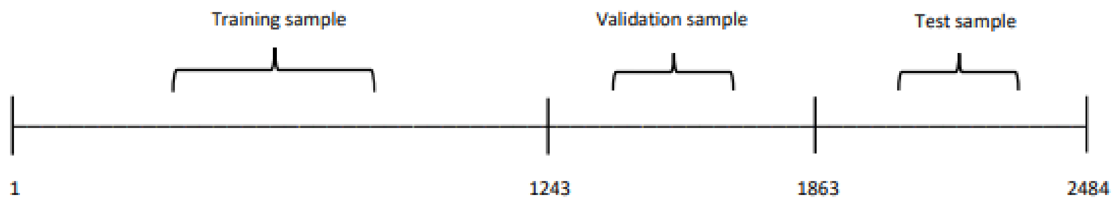

Fig. 3.1 Training Sample, Validation Sample and Test Sample

The data was collected from three sources: Cryptodatadownload (https://cryptodatadownload.com), Bitcoincharts (https://bitcoincharts.com/), and Coinmetrics (https://coinmetrics.io/).

Cryptodatadownload provides intraday data (1-minute) on OHLC (Open/High/Low/Close) prices in USD and trading volume from two exchanges: Gemini and Ftx. Bitcoincharts provides tick-by-tick trade prices in USD and number of Bitcoins traded (volume) recorded at several exchanges, from which only seven have uninterrupted data for the period under scrutiny: Bitbay, Bitfinex, Bitflyer, Bitkonan, Bitstamp, HitBTC, and Kraken. The data was then filtered to 5 minute closing prices. These are the last recorded prices before the sampling moment. Next, the log-returns were calculated, $r_t = ln(\frac{P_t}{P_{t-1}})$, which were later used to calculate the realized volatility using Formula 4.2

Figure 3.2 presents the path of volume-weighted Bitcoin daily prices, using data from the nine

[1] https://online.stat.psu.edu/stat508/lesson/2/2.2

exchanges referred to above, using the formula:

$$P_t = \frac{\sum_{i=1}^{n} v_{t,i} P_{t,i}}{\sum_{i=1}^{n} v_{t,i}}, \tag{3.1}$$

where i refers to exchange i, n is the number of exchanges , $P_{t,i}$, and $v_{t,i}$, are the price and trading volume in exchange i at day t, respectively.

From 2015 until mid-2017, Bitcoin prices experience an almost linear growth. After that, Bitcoin behavior changes drastically, presenting an exponential growth until hitting roughly 20,000 USD in late 2017. This trend ended and gave place to a sharp decline in prices, reaching almost 3,000 in 2019. Bitcoin prices then remained between 4,000 and 14,000, until late 2020, when there was an explosive price behaviour that sent Bitcoin to all-time highs at around 57,600. In 2021, there is a new retraction followed by a new episode of explosive prices. The recent history of Bitcoin prices is sintomatic of its speculative nature and its susceptibility to bubble-like events (like those described in Fry and Cheah [13], Cheah and Fry [8], and Chaim and Laurini [7]).

Figure 3.3 presents the realized volatility of Bitcoin. The main feature is the existence of periods of extremely high volatility throughout the overall sample.

The last source, Coinmetrics, provides daily Blockchain and other market information. This dataset contains 139 variables and three were excluded because they were null. Appendix A.1 provides a detailed description of these variables.

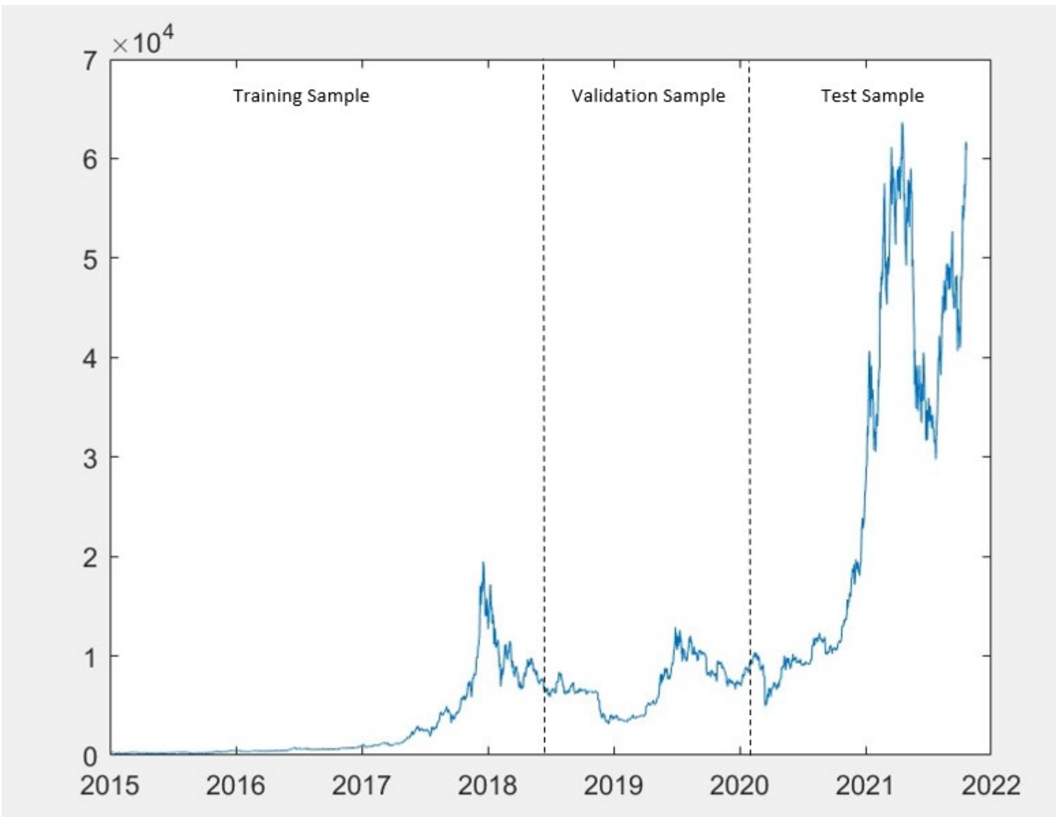

Fig. 3.2 Volume Weighted Bitcoin Daily Prices in USD

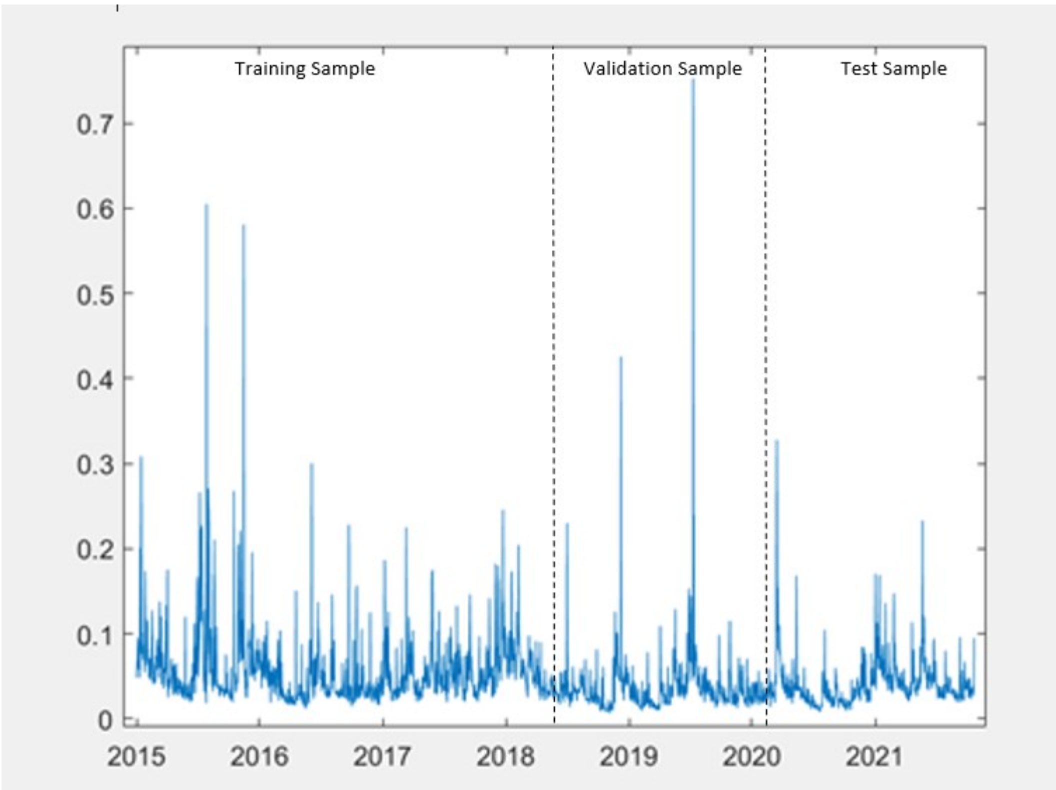

Fig. 3.3 Realized Volatility of Bitcoin

3.2 Preliminary Analysis

Table 3.1 presents the descriptive statistics of Bitcoin realized volatility. As can be seen, there is no notable difference in the statistics presented for the various samples. In all subsamples, Bitcoin realized volatility presents a kurtosis greater than 3, so its distribution is leptokurtic. Regarding skewness, one can conclude that the distributions are right-skewed, which is expected because realized volatility is always non-negative. The Jarque-Bera test rejects, at a 1% significance level, the normality of the distributions.

Table 3.1 Descriptive Statistics of Realized Volatility

	Full Sample	Training Sample	Validation Sample	Test Sample
Observations	2484	1242	621	621
Mean	0.0478	0.0567	0.0367	0.0409
Median	0.0376	0.0457	0.0286	0.0409
Std. Dev.	0.0403	0.0434	0.0402	0.0281
Minimum	0.0067	0.0122	0.0067	0.009
Maximum	0.7528	0.605	0.7528	0.3280
Kurtosis	75.5654	44.5537	179.1123	29.3812
Skewness	6.1458	4.692	11.176	3.9138
Jarque-Bera	560641	93914	815454	19530
JB p-value	<0.001	<0.001	<0.001	<0.001

Chapter 4

Methodology

The essence of the methodology used in this work is based on Corsi [10] which gives an in-depth analysis of the HAR model. Other papers such as Aalborg et al. [1] and Bouri et al. [5] are also relevant because they show how one can modify the basic HAR model to include exogenous variables.

4.1 Variables

4.1.1 Realized Volatility

Corsi [10] calculates realized volatility as the square root of the sum of squared intraday returns. Given that we are working with data from several exchanges, we use the following formula to calculate volume-weighted squared returns at the 5-minute interval j in day t:

$$r_{t,j}^2 = \frac{\sum_{i=1}^{n} v_{t,j,i} r_{t,j,i}^2}{\sum_{i=1}^{n} v_{t,j,i}} \tag{4.1}$$

where i refers to exchange i, n is the number of exchanges , $r_{t,j,i}$, and $v_{t,j,i}$, are the log-return and trading volume in exchange i at the 5-min interval j in day t, respectively. Subsequently, we calculate realized volatility in day t as:

$$RV_t = \sqrt{\sum_{j=1}^{T} r_{t,j}^2} \tag{4.2}$$

Since Bitcoin is available for trading 24/7, we have 288 (5-minute) intervals in a day, hence in Equation (4.2).

$$T = 288 \tag{4.3}$$

4.1.2 Trading Volume

The trading volume data were obtained from each exchange. The daily trading volume was computed by adding the trading volume from all exchanges ($Volume_t = \sum_j \sum_i v_{t,j,i}$). The volume variable ($V_t$) was then standardized according to Equation (4.4)

$$V_t = \frac{Volume_t - \overline{Volume}}{\sigma(Volume)}, \tag{4.4}$$

where $Volume_t$ represents the total volume at day t, $\overline{Volume}$ stands for the volume average, and $\sigma(Volume)$ represents its standard deviation.

4.1.3 Blockchain and other Market Information

The principal focus of this paper is to find out if Blockchain information plays an important role when forecasting Bitcoin volatility.

The dataset obtained from Coinmetrics contains 139 variables, of which three are excluded because they are null, leaving us with 136 variables. Most of these variables are related to Blockchain, while others are related to the Bitcoin transaction market (e.g., market capitalization). It would be almost impossible to incorporate all these variables into the HAR models. To solve this dimensionality we normalize the data set and then reduce the size of the exogenous variables space by applying Principal Component Analysis (PCA). PCA is a mathematical technique that transforms (possibly) correlated variables into a set of variables called principal components (see Richardson [23] for a more in-depth review on PCA). We use the Matlab built-in function *pca*[1] that uses the singular value decomposition (SVD) algorithm to calculate the principal components coefficients and also returns the percentage of the total variance explained by each principal component. The variables in the initial data set are linear functions of the principal components. In this work, we select the principal components that account for most (at least 75%) of the variation in the Blockchain variables. By using those principal components instead of the initial variables, we significantly decrease the number of variables to be included in the regression models.

4.1.4 Jumps

Another variable often used in HAR-type models (e.g., Qiu et al. [22] and Bouri et al. [5]) is the daily jump component (J_t). A price jump can be defined as an abrupt price change that is considerably larger when compared with the current market situation, usually computed as:

$$J_t = max(RV_t - BPV_t, 0) \tag{4.5}$$

where BiPower Variation (BPV) is given by:

$$BPV_t = \sqrt{\frac{\pi}{2} \sum_{j=2}^{288} |r_{t,j-1}||r_{t,j}|} \tag{4.6}$$

[1] https://www.mathworks.com/help/stats/pca.html

4.2 HAR-type models

The standard HAR model that postulates the h-step ahead realized volatility, proposed by Corsi [10], is the following:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(5)} + \beta_m RV_t^{(22)} + e_{t+h}, \tag{4.7}$$

where $RV_t^{(l)} = l^{-1} \sum_{s=0}^{l-1} RV_{t-s}$, is the l period averages of lagged RV, the βs are the coefficients and e_{t+h} is a zero mean innovation process.

The model described in 4.7 is usually applied to traditional financial assets, therefore the use of lag indexes $l = [1,5,22]$, representing the daily, weekly and monthly volatility of these markets, as they are closed on weekends. But since Bitcoin is available for trading 24/7, the suitable HAR model is:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(7)} + \beta_m RV_t^{(30)} + e_{t+h}. \tag{4.8}$$

As mentioned above, the HAR model can be modified to include more variables. The HAR-V model is one example of that and includes the variable volume:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(7)} + \beta_m RV_t^{(30)} + \beta_v V_t + e_{t+h}. \tag{4.9}$$

The HAR-PC model incorporates the principal components of Blockchain and other market information and it is defined as follows:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(7)} + \beta_m RV_t^{(30)} + \beta_1 PC_t^1 + \beta_2 PC_t^2 + \beta_3 PC_t^3 + e_{t+h}, \tag{4.10}$$

where PCs represents the principal components (in this case three PCs are included in the model, which is enough to explain 75% of the variability of the 136 variables in the Coinmetrics database).

The last variable used in this study is the daily jump component, J_t, and gives rise to the HAR-J model:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(7)} + \beta_m RV_t^{(30)} + \beta_j J_t + e_{t+h}. \tag{4.11}$$

The following four models result from the combination of the variables used in previous models.

HAR-V-PC:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(7)} + \beta_m RV_t^{(30)} + \beta_v V_t + \beta_1 PC_t^1 + \beta_2 PC_t^2 + \beta_3 PC_t^3 + e_{t+h}. \tag{4.12}$$

HAR-V-J:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(7)} + \beta_m RV_t^{(30)} + \beta_v V_t + \beta_j J_t + e_{t+h}. \tag{4.13}$$

HAR-PC-J:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(7)} + \beta_m RV_t^{(30)} + \beta_1 PC_t^1 + \beta_2 PC_t^2 + \beta_3 PC_t^3 + \beta_j J_t + e_{t+h}. \tag{4.14}$$

HAR-V-PC-J:

$$RV_{t+h} = \beta_0 + \beta_d RV_t^{(1)} + \beta_w RV_t^{(7)} + \beta_m RV_t^{(30)} + \beta_1 PC_t^1 + \beta_2 PC_t^2 + \beta_3 PC_t^3 + \beta_v V_t + \ldots \\ \beta_j J_t + e_{t+h}. \quad (4.15)$$

Applying logarithms to the previous models (similar to Catania and Sandholdt [6]), we thus have more eight models.

HAR-LN:

$$ln(RV_{t+h}) = \beta_0 + \beta_d ln(RV_t^{(1)}) + \beta_w ln(RV_t^{(7)}) + \beta_m ln(RV_t^{(30)}) + e_{t+h}, \quad (4.16)$$

where $ln(RV_t^{(l)}) = l^{-1} \sum_{s=0}^{l-1} ln(RV_{t-s})$.

HAR-V-LN:

$$ln(RV_{t+h}) = \beta_0 + \beta_d ln(RV_t^{(1)}) + \beta_w ln(RV_t^{(7)}) + \beta_m ln(RV_t^{(30)}) + \beta_v V_t + e_{t+h}. \quad (4.17)$$

HAR-PC-LN:

$$ln(RV_{t+h}) = \beta_0 + \beta_d ln(RV_t^{(1)}) + \beta_w ln(RV_t^{(7)}) + \beta_m ln(RV_t^{(30)}) + \beta_1 PC_t^1 + \beta_2 PC_t^2 + \ldots \\ \beta_3 PC_t^3 + e_{t+h}. \quad (4.18)$$

HAR-J-LN:

$$ln(RV_{t+h}) = \beta_0 + \beta_d ln(RV_t^{(1)}) + \beta_w ln(RV_t^{(7)}) + \beta_m ln(RV_t^{(30)}) + \beta_j ln(J_t) + e_{t+h}, \quad (4.19)$$

where $ln(J_t) = max(ln(R_t) - ln(BPV_t), 0)$

HAR-PC-V-LN:

$$ln(RV_{t+h}) = \beta_0 + \beta_d ln(RV_t^{(1)}) + \beta_w ln(RV_t^{(7)}) + \beta_m ln(RV_t^{(30)}) + \beta_1 PC_t^1 + \beta_2 PC_t^2 + \beta_3 PC_t^3 + \ldots \\ \beta_v V_t + e_{t+h}. \quad (4.20)$$

HAR-V-J-LN:

$$ln(RV_{t+h}) = \beta_0 + \beta_d ln(RV_t^{(1)}) + \beta_w ln(RV_t^{(7)}) + \beta_m ln(RV_t^{(30)}) + \beta_v V_t + \beta_j ln(J_t) + e_{t+h}. \quad (4.21)$$

HAR-PC-J-LN:

$$ln(RV_{t+h}) = \beta_0 + \beta_d ln(RV_t^{(1)}) + \beta_w ln(RV_t^{(7)}) + \beta_m ln(RV_t^{(30)}) + \beta_1 PC_t^1 + \beta_2 PC_t^2 + \beta_3 PC_t^3 + \ldots \\ \beta_j ln(J_t) + e_{t+h}. \quad (4.22)$$

HAR-V-PC-J-LN:

$$ln(RV_{t+h}) = \beta_0 + \beta_d ln(RV_t^{(1)}) + \beta_w ln(RV_t^{(7)}) + \beta_m ln(RV_t^{(30)}) + \beta_1 PC_t^1 + \beta_2 PC_t^2 + \beta_3 PC_t^3 + \ldots \\ \beta_v V_t + \beta_j ln(J_t) + e_{t+h}. \quad (4.23)$$

The next step is to find out which window length best fits each model. We start by dividing the sample into three sub-samples in a 50/25/25 split, as indicated in Figure 3.1.

The procedure for each model is as follows:

1. The model is estimated using observations 30 to 1242 (from the Training sample), i.e., with a window length of 1212. It is important to note that it's not possible to use a window length equal to 1241 (length of Training sample until observation 1241, the last observation used in the first rolling window) because the first 30 observations are needed to compute the first realization of variable $RV_t^{(30)}$. Using this information, we compute the forecast for observation 1243 (first observation of Validation sample). Then we re-estimate the model using observations 31 to 1243 and compute the forecast for observation 1244 (second observation of Validation sample) and so on until we forecast every value for the Validation sample. So the forecasts in the Validation Sample are obtained using a rolling window with fixed length Afterwards we calculate the RMSE statistic (see Equation 4.35).

2. We repeat this process, but this time using a window length of 1211, that is, we estimate the model using observations 31 to 1242 from the Training sample to then calculate the forecast for observation 1243. Then we proceed analogously to what was explained above.

3. This process is executed 1183 times until the window length equals 30. The window length that results in the smallest RMSE value is the one selected.

4. After ascertaining the best window length for each model, the forecasts for the Test sample are calculated. It's important to note that, in this paper, we only calculate 1-day ahead forecasts, i.e., h=1.

Equations 4.24 and 4.25 illustrate the computations performed in step 1, while Equations 4.26 and 4.27 illustrate the computations performed in step 2 for the HAR model.

$$\begin{bmatrix} RV_{31} \\ . \\ . \\ RV_{1242}^{(30)} \end{bmatrix} = \begin{bmatrix} 1 & RV_{30}^{(1)} & RV_{30}^{(7)} & RV_{30}^{(30)} \\ . & . & . & . \\ . & . & . & . \\ 1 & RV_{1241}^{(1)} & RV_{1241}^{(7)} & RV_{1241}^{(30)} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_d \\ \beta_w \\ \beta_m \end{bmatrix} \tag{4.24}$$

$$RV_{1243} = \beta_0 + \beta_d RV_{1242}^{(1)} + \beta_w RV_{1242}^{(7)} + \beta_m RV_{1242}^{(30)} \tag{4.25}$$

$$\begin{bmatrix} RV_{32} \\ . \\ . \\ RV_{1242}^{(30)} \end{bmatrix} = \begin{bmatrix} 1 & RV_{31}^{(1)} & RV_{31}^{(7)} & RV_{31}^{(30)} \\ . & . & . & . \\ . & . & . & . \\ 1 & RV_{1241}^{(1)} & RV_{1241}^{(7)} & RV_{1241}^{(30)} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_d \\ \beta_w \\ \beta_m \end{bmatrix} \tag{4.26}$$

$$RV_{1243} = \beta_0 + \beta_d RV_{1242}^{(1)} + \beta_w RV_{1242}^{(7)} + \beta_m RV_{1242}^{(30)} \tag{4.27}$$

At the end of this process we have computed 16 sets of predictions, one for each model. We also calculate two additional forecasts (using the previous ones): the arithmetic mean and a weighted mean (WM). The weighted mean forecast for day $t+1$ is computed as:

$$WM_{t+1} = \sum_{i=1}^{16} w_i \hat{y}_{i,t+1} \tag{4.28}$$

where $\hat{y}_{i,t+1}$ is the forecast of model i at day $t+1$, and w_i is the weight defined by:

$$w_i = \frac{\phi_i^{-1}}{\sum_{j=1}^{16} \phi_j^{-1}} \tag{4.29}$$

where ϕ_i is given by:

$$\phi_i = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} e_{i,t+1}^2 \tag{4.30}$$

where t_0 is the first observation and T is the last observation in the window used to obtain $\hat{y}_{i,t+1}$ and $e_{i,t+1}$ is the error.

The choice of these weights implicit assume that the errors of the different models are not correlated with each other. The weights decrease with the variance of the errors, hence the more accurate are the models the higher are the weights.

The forecasts are then compared with the actual values to measure the forecasting performance of the different models. The statistics used are the Mean Error (ME), the Mean Absolute Error (MAE), the Mean Percentage Error (MPE), the Mean Absolute Percentage Error (MAPE), the Root Mean Square Error (RMSE), and Theil's U (U), computed as follows:

$$ME = \frac{1}{n} \sum_{t=0}^{n-1} (y_{t+1} - \hat{y}_{t+1}) \tag{4.31}$$

$$MAE = \frac{1}{n}\sum_{t=0}^{n-1}|y_{t+1} - \hat{y}_{t+1}| \tag{4.32}$$

$$MPE = \frac{100\%}{n}\sum_{t=0}^{n-1}\left(\frac{y_{t+1} - \hat{y}_{t+1}}{y_{t+1}}\right) \tag{4.33}$$

$$MAPE = \frac{100\%}{n}\sum_{t=0}^{n-1}\left|\frac{(y_{t+1} - \hat{y}_{t+1})}{y_{t+1}}\right| \tag{4.34}$$

$$RMSE = \sqrt{\frac{\sum_{t=0}^{n-1}(y_{t+1} - \hat{y}_{t+1})^2}{n}} \tag{4.35}$$

$$U = \sqrt{\frac{\sum_{t=0}^{n-2}\left(\frac{\hat{y}_{t+2} - y_{t+2}}{y_{t+1}}\right)^2}{\sum_{i=0}^{n-2}\left(\frac{y_{t+2} - y_{t+1}}{y_{t+1}}\right)^2}} \tag{4.36}$$

Where n is the number of forecasts, y_{t+1} is the actual value of the forecasted variable (RV_{t+1} or $ln(RV_{t+1})$) at observation t+1.

We also apply the test proposed by Diebold and Mariano [11] with the null hypothesis of no difference in the accuracy of two competing forecasts.

Chapter 5

Forecasting Performance

For illlustration purposes, Table 5.1 shows the initial OLS regression results for all 16 models. The coefficients presented where then used to forecast the first observation in the test-sample. The models presented in the table can be divided into two groups: HAR (1-8) and HAR-LN (9-16)

In the first group, the results allow us to conclude that lagged RV has a positive relation with future RV (except for $RV^{(30)}$ in model HAR-V-PC), and that $RV^{(1)}$ and $RV^{(7)}$ are always significant at 1% level. Similar to RV, the PC coefficients are always positive (except for PC^3 in model HAR-V-PC-J). On the other hand, the statistical significance of those coefficients varies across models. The volume coefficient is significant at the 1% level, except for model HAR-V-J. The 1-day lagged jump has a negative relation with future volatility and it is always significant at the 1% level.

Moving on to the second group of models, we see that the coefficients of $ln(RV^{(1)})$, $ln(RV^{(7)})$, $ln(RV^{(30)})$, $ln(J)$ are similar to the coefficients of $RV^{(1)}$, $RV^{(7)}$, $RV^{(30)}$, J from the first group of models, however in this group of models the coefficients of $ln(RV^{(7)})$ aren't always significant. PC^1 and PC^3 have a negative relation with future volatility, and PC^2 has a positive relation. Regarding V_t, it is only significant in model HAR-V-LN.

Table 5.1 Result for 1-day Ahead Regressions by OLS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
$RV_t^{(1)}$	0.195*** (0.034)	0.138*** (0.035)	0.1715*** (0.0334)	0.566*** (0.058)	0.140*** (0.034)	0.578*** (0.077)	0.516*** (0.057)	0.580*** (0.063)								
$RV_t^{(7)}$	0.492*** (0.064)	0.451*** (0.064)	0.357*** (0.067)	0.279*** (0.074)	0.370*** (0.065)	0.278*** (0.075)	0.224*** (0.072)	0.215*** (0.066)								
$RV_t^{(30)}$	0.007 (0.034)	0.006 (0.034)	0.0014 (0.0332)	0.035 (0.036)	-0.001 (0.032)	0.037 (0.037)	0.018 (0.031)	0.018 (0.031)								
PC_t^1			0.000** (0.000)		0.001** (0.000)		0.000 (0.000)	0.000 (0.000)			-0.040 (0.063)		-0.046 (0.064)		-0.047 (0.062)	-0.442 (0.063)
PC_t^2			0.002*** (0.000)		0.001*** (0.000)		0.001 (0.000)	0.001*** (0.000)			0.065** (0.029)		0.062** (0.030)		0.054* (0.029)	0.056* (0.030)
PC_t^3			0.000 (0.000)		0.000 (0.000)		0.001 (0.001)	-0.003 (0.002)			-0.021** (0.033)		-0.027 (0.034)		-0.039 (0.033)	-0.037 (0.033)
V_t		0.006*** (0.001)			0.004*** (0.001)	-0.000 (0.001)		-0.003** (0.002)		0.054*** (0.014)			0.136 (0.136)	-0.052 (0.053)		-0.077 (0.155)
J_t				-0.601*** (0.067)		-0.614*** (0.086)	-0.532*** (0.067)	-0.604*** (0.073)								
$ln(RV_t^{(1)})$									0.285*** (0.079)	0.361*** (0.036)	0.230*** (0.082)	0.451*** (0.048)	0.193** (0.091)	0.476*** (0.061)	0.386*** (0.097)	0.425*** (0.123)
$ln(RV_t^{(7)})$									0.371*** (0.143)	0.475*** (0.057)	0.235*** (0.153)	0.365*** (0.081)	0.267* (0.157)	0.339*** (0.093)	0.110 (0.156)	0.089 (0.165)
$ln(RV_t^{(30)})$									0.086 (0.078)	-0.028 (0.033)	0.067 (0.076)	0.007 (0.452)	0.074 (0.078)	0.020 (0.050)	0.102 (0.076)	0.107 (0.077)
$ln(J_t)$												-0.493*** (0.090)		-0.550*** (0.106)	-0.446*** (0.155)	-0.493*** (0.181)
R^2	0.217	0.237	0.245	0.243	0.228	0.277	0.277	0.288	0.303	0.549	0.549	0.494	0.326	0.511	0.372	0.353

Table reports the coefficients for the various models when forecasting the first observation in test-sample. Numbers in parentheses are the associated standard errors. *, ** and *** denote significance at the 10%, 5% and 1% level, respectively. The models are the following: (1) - HAR, (2) - HAR-V, (3) - HAR-PC, (4) HAR-J, (5) HAR-V-PC, (6) HAR-V-J, (7) HAR-PC-J, (8) HAR-V-PC-J, (9) HAR-LN, (10) HAR-V-LN, (11) HAR-PC-LN, (12) HAR-J-LN, (13) HAR-V-PC-LN, (14) HAR-V-J-LN, (15) HAR-PC-J-LN, (16) HAR-V-PC-J-LN

Most notably, Table 5.2 presents the forecasting statistics of each model. From these statistics, it's hard to draw any concise conclusion, as the values are quite similar. Although there are no marked differences, it can be seen that the HAR-V-J model presents some of the best values in four statistics (RMSE, ME, MAE, MAPE, Theil's U). The Mean and Weighted Mean also show interesting results in those statistics. The MAPE statistic exhibits a higher variability when compared to the RMSE and MAE statistics. It also stands out that HAR-LN models have, on average, a smaller window length (413 vs 1082).

Table 5.2 Forecasting Performance of each Model

Models		RMSE	ME	MAE	MPE	MAPE	Theil´s U	Window length
HAR	(1)	0.022104	0.0000	0.0118	-16.81 %	30.80%	1.0146	1131
HAR-V	(2)	0.022308	-0.0033	0.0131	-16.88%	29.39%	0.9580	1131
HAR-PC	(3)	0.022392	-0.0008	0.0125	-4.01%	23.71%	0.9642	1149
HAR-J	(4)	0.021387	-0.0017	0.0119	-6.57%	24.31%	0.9543	917
HAR-V-PC	(5)	0.022371	-0.0035	0.0132	-5.72%	23.95%	0.9539	1206
HAR-V-J	(6)	0.021413	0.0008	0.0114	-6.18%	23.97%	0.9504	917
HAR-PC-J	(7)	0.021653	-0.0019	0.0122	-19.63%	33.72%	1.0534	997
HAR-V-PC-J	(8)	0.021611	0.0007	0.0132	-17.84%	30.59%	0.9794	1206
HAR-LN	(9)	0.022657	0.0032	0.0109	-7.95%	25.95%	0.9647	214
HAR-V-LN	(10)	0.021774	0.0020	0.0109	-7.81%	25.81%	0.9653	1131
HAR-PC-LN	(11)	0.022963	0.0011	0.0117	-24.83%	35.72%	1.0491	219
HAR-J-LN	(12)	0.021576	0.0022	0.0106	-10.13%	26.62%	0.9586	585
HAR-V-PC-LN	(13)	0.021910	0.0000	0.0117	-26.78%	37.38%	1.0932	214
HAR-V-J-LN	(14)	0.021808	0.0024	0.0107	-11.65%	28.26%	0.9790	509
HAR-PC-J-LN	(15)	0.022875	0.0011	0.0117	-9.56%	26.57%	0.9645	219
HAR-V-PC-J-LN	(16)	0.021681	0.0030	0.0117	-9.11%	26.44%	0.9653	216
Mean	(17)	0.021275	0.0002	0.0112	-12.56 %	26.80%	0.9542	——
WM	(18)	0.021259	0.0002	0.0112	-12.55%	26.80%	0.9541	——

As mentioned before, the differences in the forcasting performance statistics are minimal, so it is not possible to conclude which models produce the best results. To better adress this issue the Diebold-Mariano test is used (Table 5.3).

In most cases (137 out of 153), the test value is in the range [-1.96, 1.96], which means that there are no significant differences at the 5% level in the compared forecasts. It is therefore important to analyze what happens in the other 16 tests (marked in bold in the table). The HAR-J-LN (12) model immediately stands out because it presents superior results in relation to seven models, which leads us to conclude that the introduction of the jumps variable together with the logarithm produces more accurate forecasts. Other models such as HAR-V-LN, HAR-V-J-LN, arithmetic mean, and the weighted mean also show good results. These results put into perspective that volume may help forecasting Bitcoin realized volatility and that the combination of forecasts from several models also may provide marginal forecasting benefits. Clearly, Blockchain information and other market information do not increase forecasting accuracy.

Table 5.3 Diebold-Mariano Tests

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
HAR	(1)		-0.59	-1.60	0.65	-0.95	0.62	0.48	0.53	-1.10	1.22	-1.59	**2.67**	0.08	1.70	-1.52	0.22	1.39	1.37
HAR-V	(2)			-0.20	1.05	-0.29	0.97	0.90	0.91	-0.48	**2.69**	-0.89	**2.64**	0.19	1.47	-0.86	0.37	**2.51**	**2.45**
HAR-PC	(3)				0.87	0.08	0.85	0.76	0.81	-0.57	1.78	-1.26	**3.30**	0.21	**2.58**	-1.14	0.37	1.76	1.73
HAR-J	(4)					-1.03	-0.14	-1.15	-0.75	-0.84	-0.42	-1.08	-0.18	-0.39	-0.38	-1.11	-0.32	0.19	0.23
HAR-V-PC	(5)						0.97	0.93	0.95	-0.44	**2.47**	-0.98	**3.44**	0.21	1.95	-0.94	0.40	**2.42**	**-2.36**
HAR-V-J	(6)							-0.81	-0.85	-0.82	-0.38	-1.06	-0.16	-0.38	-0.35	-1.09	-0.31	0.23	0.27
HAR-PC-J	(7)								0.20	-0.75	-0.16	-1.03	0.09	-0.17	-0.16	-1.07	-0.026	0.91	0.99
HAR-V-PC-J	(8)									-0.78	-0.21	-1.06	0.04	-0203	-0.21	-1.09	-0.07	0.82	0.91
HAR-LN	(9)										1.36	-0.89	**1.99**	0.28	1.88	-0.52	0.43	1.36	1.35
HAR-V-LN	(10)											-1.78	1.14	-0.06	-0.14	-1.83	0.05	1.18	1.17
HAR-PC-LN	(11)												**2.41**	0.41	**2.22**	0.32	0.59	1.75	1.73
HAR-J-LN	(12)													-0.15	-1.84	**-2.64**	-0.06	0.60	0.60
HAR-V-PC-LN	(13)														0.044	-0.390	0.48	0.35	0.36
HARV-V-J-LN	(14)															**-2.36**	0.07	0.87	0.87
HAR-PC-J-LN	(15)																0.58	1.85	1.83
HAR-V-PC-J-LN	(16)																	0.29	0.31
Mean	(17)																		0.73
WM	(18)																		

The Diebold-Mariano test statistic is the difference between the sum of the squared errors of two models. If the difference is positive, the errors are larger in the 1st model. Otherwise, they are larger in the 2nd model. Then if the statistic is greater than 1.96, the difference is statistically significant (at a 5% level) and the second model is better. If the statistic is less than -1.96, the difference is statistically significant, and the first model is better (Diebold and Mariano [11]). For each cell in the table, the first model is the one indicated in the corresponding row, and the second model is the one indicated in the corresponding column.

Chapter 6

Conclusion

This paper aims to contribute to the existing literature related to forecasting Bitcoin realized volatility by adding two aspects that have, so far, been scarcely explored. Firstly, realized volatility is computed from several exchanges and not from only one, which gives a better picture on the overall Bitcoin market. Secondly, it additionally considers Blockchain and other market information into HAR models as exogenous variables. Other papers have already used some information from the Blockchain in their models however, by using PCA we reduced the size of the dataset (from 139 to 3) while taking into account most of its information.

Using 5-minute data covering the period from January 1, 2015, through October 19, 2021, we implemented several HAR-type models to forecast 1-day ahead volatility. One of the main conclusions to be drawn from the results is that none of the eight models where Blockchain information was introduced, produces superior results compared to the other models. Another interesting result is that 3 of the eight models, where logarithms are used, produce the best predictions (conclusions drawn by analyzing the Diebold-Mariano tests), with the HAR-J-LN model demonstrating superior results, which leads us to conclude that the introduction of logarithms can improve predictions. The fact that the average forecasts (Mean and Weighted Mean) show interesting results, both in the forecasting statistics and in the Diebold-Mariano test (in terms of the RMSE statistic, these predictions present the best values), highlights that there is no model significantly superior to the others. If one forecast is more accurate than the others, one would expect their combination to produce worse results, however, this is not the case.

The main conclusions of this work could be further sustained (or refuted) by exploring some aspects such as: extending the forecast horizon beyond 1-day (forecasting for 7, 14, or 30 days following what has been usual in other papers) or applying other models like the HAR-RS which uses signed semi-variances, or HARQ-type models that use realized quarticity.

References

[1] Aalborg, H. A., Molnár, P., and de Vries, J. E. (2019). What can explain the price, volatility and trading volume of bitcoin? *Finance Research Letters*, 29:255–265.

[2] Ardia, D., Bluteau, K., and Rüede, M. (2019). Regime changes in bitcoin garch volatility dynamics. *Finance Research Letters*, 29:266–271.

[3] Baur, D. G. and Dimpfl, T. (2018). Asymmetric volatility in cryptocurrencies. *Economics Letters*, 173:148–151.

[4] Bouri, E., Azzi, G., and Dyhrberg, A. H. (2017). On the return-volatility relationship in the bitcoin market around the price crash of 2013. *Economics*, 11(1).

[5] Bouri, E., Gkillas, K., Gupta, R., and Pierdzioch, C. (2021). Forecasting realized volatility of bitcoin: The role of the trade war. *Computational Economics*, 57:29–53.

[6] Catania, L. and Sandholdt, M. (2019). Bitcoin at high frequency. *Journal of Risk and Financial Management*, 12(1):36.

[7] Chaim, P. and Laurini, M. P. (2019). Is bitcoin a bubble? *Physica A*, 517:222–232.

[8] Cheah, E.-T. and Fry, J. (2015). Speculative bubbles in bitcoin markets? an empirical investigation into the fundamental value of bitcoin. *Economics Letters*, 130:32–36.

[9] Conrad, C., Custovic, A., and Ghysels, E. (2018). Long- and short-term cryptocurrency volatility components: A garch-midas analysis. *Journal of Risk and Financial Management*, 11(2):23.

[10] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7:174–196.

[11] Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuuracy. *Journal of Business and Economic Statistics*, 13(3):253–263.

[12] Fang, T., Su, Z., and Yin, L. (2020). Economic fundamentals or investor perceptions? the role of uncertainty in predicting long-term cryptocurrency volatility. *International Review of Financial Analysis*, 71:101566.

[13] Fry, J. and Cheah, E.-T. (2016). Negative bubbles and shocks in cryptocurrency markets. *International Review of Financial Analysis*, 47:343–352.

[14] Gkillas, K., Tantoula, M., and Tzagarakis, M. (2021). Transaction activity and bitcoin realized volatility. *Operations Research Letters*, 49:715–719.

[15] Klein, T., Thu, H. P., and Walther, T. (2018). Bitcoin is not the new gold – a comparison of volatility, correlation, and portfolio performance. *International Review of Financial Analysis*, 59:105–116.

[16] Kyriazis, N. A. (2021). A survey on volatility fluctuations in the decentralized cryptocurrency financial assets. *Journal of Risk and Financial Management*, 14(7):293.

[17] Köchling, G., Schmidtke, P., and Posch, P. N. (2020). Volatility forecasting accuracy for bitcoin. *Economics Letters*, 191:108836.

[18] Lansky, J. (2018). Possible state approaches to cryptocurrencies. *Journal of Systems integration*, 9(1):19.

[19] Liu, L. Y., Patton, A. J., and Sheppard, K. (2015). Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187:293–311.

[20] Matkovskyy, R., Jalan, A., and Dowling, M. (2020). Effects of economic policy uncertainty shocks on the interdependence between bitcoin and traditional financial markets. *The Quaterly Review of Economics and Finance*, 77:150–155.

[21] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260.

[22] Qiu, Y., Wang, Z., Xie, T., and Zhang, X. (2021). Forecasting bitcoin realized volatility by exploiting measurement error under model uncertainty. *Journal of Empirical Finance*, 62:179–201.

[23] Richardson, M. (2009). Principal component analysis. *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si*, 6:16.

[24] Trucíos, C. (2019). Forecasting bitcoin risk measures: A robust approach. *International Journal of Forecasting*, 35:836–847.

[25] Walther, T., Klein, T., and Bouri, E. (2019). Exogenous drivers of bitcoin and cryptocurrency volatility – a mixed data sampling approach to forecasting. *Journal of International Financial Markets, Institutions Money*, 35:836–847.

[26] Xie, T. (2019). Forecast bitcoin volatility with least squares model averaging. *Econometrics*, 7(3):40.

[27] Yu, M. (2019). Forecasting bitcoin volatility: The role of leverage effect and uncertainty. *Physica A: Statistical Mechanics and Its Applications*, 533:120707.

Appendix A

Description of market and blockchain variables

Table A.1 Description of the Variables Collected from Coinmetrics

ID	Name	Description
AdrActCnt	Active Adresses	The sum count of unique addresses that were active in the network (either as a destination or source of a ledger change) that day. All parties in a ledger change action (source and destination) are counted. Individual addresses are not double-counted if previously active.
AdrBal1in100KCnt	Addr Cnt with $\geq$ 0.001% Supply	The sum count of unique addresses holding at least 0.001% of the current supply of native units as of the end of that day. Only native units are considered (e.g., an address with less than one hundred-thousandth ETH but with ERC-20 tokens would not be considered).
AdrBal1in100MCnt	Addr Cnt with $\geq$ 0.000001% Supply	... at least 0.000001%...
AdrBal1in10BCnt	Addr Cnt with $\geq$ 0.00000001% Supply	... at least 0.00000001%...
AdrBal1in10KCnt	Addr Cnt with $\geq$ 0.01% Supply	... at least 0.01%...

Continues on the next page

ID	Name	Description
AdrBal1in10MCnt	Addr Cnt with $\geq$ 0.00001% Supply	... at least 0.00001%...
AdrBal1in1BCnt	Addr Cnt with $\geq$ 0.0000001% Supply	... at least 0.0000001%...
AdrBal1in1KCnt	Addr Cnt with $\geq$ 0.1% Supply	... at least 0.1% ...
AdrBal1in1MCnt	Addr Cnt with $\geq$ 0.0001% Supply ... at least 0.0001% ...	
AdrBalCnt	Address Cnt Bal $>$ 0	The sum count of unique addresses holding any amount of native units as of the end of that day. Only native units are considered.
AdrBalNtv0.001Cnt	Addr Cnt of Bal $\geq$ 0.001 (native units)	The sum count of unique addresses holding at least 0.001 native units as of the end of that day. Only native units are considered.
AdrBalNtv0.01Cnt	Addr Cnt of Bal $\geq$ 0.01 (native units)	... at least 0.01 ...
AdrBalNtv0.1Cnt	Addr Cnt of Bal $\geq$ 0.1 (native units)	... at least 0.1 ...
AdrBalNtv100Cnt	Addr Cnt of Bal $\geq$ 100 (native units)	... at least 100 ...
AdrBalNtv100KCnt	Addr Cnt of Bal $\geq$ 100K (native units)	... at least 100k ...
AdrBalNtv10Cnt	Addr Cnt of Bal $\geq$ 10 (native units)	... at least 10 ...
AdrBalNtv10KCnt	Addr Cnt of Bal $\geq$ 10K (native units)	... at least 10k ...

Continues on the next page

ID	Name	Description
AdrBalNtv1Cnt	Addr Cnt of Bal ≥ 1 (native units)	... at least 1 ...
AdrBalNtv1KCnt	Addr Cnt of Bal ≥ 1K (native units)	... at least 1k ...
AdrBalNtv1MCnt	Addr Cnt of Bal ≥ 1M (native units)	... at least 1M ...
AdrBalUSD100Cnt	Addr Cnt of Bal ≥ $100	The sum count of unique addresses holding at least $100 as of the end of that day. Only native units are considered (e.g., an address with less than $100 but with more than $100 in ERC-20 tokens would not be considered).
AdrBalUSD100KCnt	Addr Cnt of Bal ≥ $100K	... at least $100K ...
AdrBalUSD10Cnt	Addr Cnt of Bal ≥ $10	... at least $10 ...
AdrBalUSD10KCnt	Addr Cnt of Bal ≥ $10K	... at least $10K ...
AdrBalUSD10MCnt	Addr Cnt of Bal ≥ $10M	... at least $10M ...
AdrBalUSD1Cnt	Addr Cnt of Bal ≥ $1	... at least $1 ...
AdrBalUSD1KCnt	Addr Cnt of Bal ≥ $1K	... at least frm[o]–K...
AdrBalUSD1MCnt	Addr Cnt of Bal ≥ $1M	... at least $1M ...
AssetEODCompletionTime	Completion Time	The time that the last metric for the asset was calculated, indicating that all metrics for that asset have been calculated
BlkCnt	Block Cnt	The sum count of blocks created that interval that were included in the main (base) chain.
BlkSizeMeanByte	Mean Block Size (in bytes)	The mean size (in bytes) of all blocks created that interval.

Continues on the next page

ID	Name	Description
BlkWghtMean	Mean Block Weight	The mean weight of all blocks created that interval. Weight is a dimensionless measure of a block's “size”. It is only applicable for chains that use SegWit (segregated witness).
BlkWghtTot	Sum Block Weight	The sum weight of all blocks created that interval. Weight is a dimensionless measure of a block's “size”. It is only applicable for chains that use SegWit (segregated witness).
CapAct1yrUSD	Active Market Cap (1yr) (USD)	The sum USD value of all active native units in the last year. Native units that transacted more than once are only counted once.
CapMVRVCur	MVRV (Market Cap / Realized Market Cap)	The ratio of the sum USD value of the current supply to the sum "realized" USD value of the current supply.
CapMVRVFF	Free Float MVRV (Free Float Market Cap / Realized Market Cap)	The ratio of the free float market capitalization (CapMrktFFUSD) to the sum "realized" USD value of the current supply (CapRealUSD).
CapMrktCurUSD	Market Cap (USD)	The sum USD value of the current supply. Also referred to as network value or market capitalization.
CapMrktFFUSD	Free Float Market Cap (USD)	The sum USD value of the free float supply. Also referred to as free float network value or free float market capitalization.
CapRealUSD	Realized Market Cap (USD)	The sum USD value based on the USD closing price on the day that a native unit last transacted for all native units.
DiffLast	Difficulty	The difficulty of the last block in the considered time period. Difficulty represents how hard it is to find a hash that meets the protocol-designated requirement (i.e., the difficulty of finding a new block) that day. The requirement is unique to each applicable cryptocurrency protocol. Difficulty is adjusted periodically by the protocol as a function of how much hashing power is being deployed by miners.

Continues on the next page

ID	Name	Description
DiffMean		
FeeByteMeanNtv	Mean Tx Fee per Byte (native units)	The mean transaction fee per byte of all blocks that interval in native units.
FeeMeanNtv	Mean Tx Fee (native units)	The mean fee per transaction in native units that interval.
FeeMeanUSD	Mean Tx Fee (USD)	The sum USD value of the mean fee per transaction that interval.
FeeMedNtv	Median Tx Fee (native units)	The median fee per transaction in native units that interval.
FeeMedUSD	Median Tx Fee (USD)	The sum USD value of the median fee per transaction that day.
FeeTotNtv	Total Fees (native units)	The sum native units value of all fees paid to miners, transaction validators, stakers and/or block producers that interval. In certain cryptonetworks, fees might be burned (destroyed), but they are still accounted for in this metric.
FeeTotUSD	Total Fees (USD)	The sum USD value of all fees paid to miners, transaction validators, stakers and/or block producers that interval. In certain cryptonetworks, fees might be burned (destroyed), but they are still accounted for in this metric.
FlowInExNtv	Exchange Deposits (native units)	The sum number of native units sent to exchanges that interval, excluding exchange to exchange activity
FlowInExUSD	Exchange Deposits (USD)	The sum USD value sent to exchanges that interval, excluding exchange to exchange activity.
FlowOutExNtv	Exchange Withdrawals (native units)	The sum in native units withdrawn from exchanges that day, excluding exchange to exchange activity.
FlowOutExUSD	Exchange Withdrawals (USD)	The sum USD value withdrawn from exchanges that day, excluding exchange to exchange activity

Continues on the next page

ID	Name	Description
FlowTfrFromExCnt	Exchange Withdrawal Cnt	The sum count of transfers from any address belonging to an exchange in that interval. Transfers between exchanges are not counted.
HashRate	Mean Hash Rate	The mean rate at which miners are solving hashes that day. Hash rate is the speed at which computations are being completed across all miners in the network. The unit of measurement varies depending on the protocol.
HashRate30d	Mean Hash Rate, 30 Day	The mean rate at which miners are solving hashes over the last 30 days.
IssContNtv	Coinbase Issuance (native units)	The sum of native units issued that day. Only those native units that are issued by a protocol-mandated continuous emission schedule are included.
IssContPctAnn	Annual Inflation Rate	The percentage of new native units (continuous) issued on that day, extrapolated to one year (i.e., multiplied by 365), and divided by the current supply on that day. Sometimes referred to as the annual inflation rate.
IssContPctDay	Daily Inflation Rate	The percentage of new native units (continuous) issued on that day divided by the current supply on that day. Also referred to as the daily inflation rate.
IssContUSD	Coinbase Issuance (USD)	The sum of USD value issued that day. Only those native units that are issued by a protocol-mandated continuous emission schedule are included.
IssTotNtv	Total Issuance (native units)	The sum of all new native units issued that day.
IssTotUSD	Total Issuance (USD)	The sum USD value of all new native units issued that day.
NDF	NDF (Network Distribution Factor)	The ratio of supply held by addresses with at least one ten-thousandth of the current supply of native units to the current supply.

Continues on the next page

ID	Name	Description
NVTAdj	NVT	The ratio of the network value (or market capitalization, current supply) divided by the adjusted transfer value. Also referred to as NVT.
NVTAdj90	NVT 90-day Moving Avg	The ratio of the network value (or market capitalization, current supply) to the 90-day moving average of the adjusted transfer value. Also referred to as NVT.
NVTAdjFF	Free Float NVT	The ratio of the free float network value (or market capitalization, free float) divided by the adjusted transfer value. Also referred to as FFNVT.
NVTAdjFF90	Free Float NVT 90-day Moving Avg	The ratio of the free float network value (or market capitalization, free float) to the 90-day moving average of the adjusted transfer value. Also referred to as FFNVT.
PriceBTC	BTC Denominated Price	The fixed closing price of the asset as of 00:00 UTC the following day (i.e., midnight UTC of the current day) denominated in BTC.
PriceUSD	USD Denominated Price	The fixed closing price of the asset as of 00:00 UTC the following day (i.e., midnight UTC of the current day) denominated in USD.
ROI1yr	ROI, 1 Year	The return on investment for the asset assuming a purchase 12 months prior.
ROI30d	ROI, 30 Days	The return on investment for the asset assuming a purchase 30 days prior.
RevAllTimeUSD	All Time Miner Revenue (USD)	The sum USD value of all miner revenue (fees plus newly issued native units) for all time.
RevHashNtv	Miner Revenue per Hash (native units)	The mean miner reward per estimated hash unit performed during the period, in native units. The unit of hashpower measurement depends on the protocol.

Continues on the next page

ID	Name	Description
RevHashRateNtv	Miner Revenue per Hash per Sec (native units)	The mean daily miner reward per estimated hash unit per second performed during the period, in native units.
RevHashRateUSD	Miner Revenue per Hash per Sec (USD)	The USD value of the mean daily miner reward per estimated hash unit per second performed during the period, also known as hashprice.
RevHashUSD	Miner Revenue per Hash (USD)	The mean miner reward per estimated hash unit performed during the period, in USD.
RevNtv	Miner Revenue (native units)	The sum native units of miner revenue (fees plus newly issued native units) that interval.
RevUSD	Miner Revenue (USD)	The sum USD value of all miner revenue (fees plus newly issued native units) that day.
SER	SER (Supply Equality Ratio)	The ratio of supply held by addresses with less than one ten-millionth of the current supply of native units to the supply held by the top one percent of addresses.
SplyAct10yr	10 Year Active Supply	The sum of unique native units that transacted at least once in the trailing 10 Years up to that interval. Native units that transacted more than once are only counted once.
SplyAct180d	180 Day Active Supply	... trailing 180 days ...
SplyAct1d	1 Day Active Supply	... trailing 1 day ...
SplyAct1yr	1 Year Active Supply	... trailing 1 year ...
SplyAct2yr	2 Year Active Supply	... trailing 2 years ...
SplyAct30d	30 Day Active Supply	... trailing 30 days ...

Continues on the next page

ID	Name	Description
SplyAct3yr	3 Year Active Supply	... trailing 3 years ...
SplyAct4yr	4 Year Active Supply	... trailing 4 years ...
SplyAct5yr	5 Year Active Supply	... trailing 5 years ...
SplyAct7d	7 Day Active Supply	... trailing 7 days ...
SplyAct90d	90 Day Active Supply	... trailing 90 days ...
SplyActEver	Active Supply (transacted at least once)	The sum of unique native units held by accounts that transacted at least once up to that interval. Native units that transacted more than once are only counted once.
SplyActPct1yr	1 Year Active Supply %	The percentage of the current supply that has been active in the trailing 1 year up to that day.
SplyAdrBal1in100K	Val in Addrs w/ Bal $\geq$ 0.001% of Current Supply	The sum of all native units being held in addresses whose balance was at least 0.001% of the current supply of native units as the end of that day. Only native units are considered.
SplyAdrBal1in100M	Val in Addrs w/ Bal $\geq$ 0.000001% of Current Supply	... at least 0.000001% ...
SplyAdrBal1in10B	Val in Addrs w/ Bal $\geq$ 0.00000001% of Current Supply	... at least 0.00000001% ...
SplyAdrBal1in10K	Val in Addrs w/ Bal $\geq$ 0.01% of Current Supply	... at least 0.01% ...

Continues on the next page

ID	Name	Description
SplyAdrBal1in10M	Val in Addrs w/ Bal $\geq$ 0.00001% of Current Supply	... at least 0.00001% ...
SplyAdrBal1in1B	Val in Addrs w/ Bal $\geq$ 0.0000001% of Current Supply	... at least 0.0000001% ...
SplyAdrBal1in1K	Val in Addrs w/ Bal $\geq$ 0.1% of Current Supply	... at least 0.1% ...
SplyAdrBal1in1M	Val in Addrs w/ Bal $\geq$ 0.0001% of Current Supply	... at least 0.0001% ...
SplyAdrBalNtv0.001	Val in Addrs w/ Bal $\geq$ 0.001 (native units)	The sum of all native units being held in addresses whose balance was at least 0.001 native units at the end of that day. Only native units are considered.
SplyAdrBalNtv0.01	Val in Addrs w/ Bal $\geq$ 0.01 (native units)	... at least 0.01 ...
SplyAdrBalNtv0.1	Val in Addrs w/ Bal $\geq$ 0.1 (native units)	... at least 0.1 ...
SplyAdrBalNtv1	Val in Addrs w/ Bal $\geq$ 1 (native units)	... at least 1 ...
SplyAdrBalNtv10	Val in Addrs w/ Bal $\geq$ 10 (native units)	... at least 10 ...
SplyAdrBalNtv100	Val in Addrs w/ Bal $\geq$ 100 (native units)	... at least 100 ...
SplyAdrBalNtv100K	Val in Addrs w/ Bal $\geq$ 100K (native units)	... at least 100K ...

Continues on the next page

ID	Name	Description
SplyAdrBalNtv10K	Val in Addrs w/ Bal ≥ 10K (native units)	... at least 10K ...
SplyAdrBalNtv1K	Val in Addrs w/ Bal ≥ 1K (native units)	... at least 1K ...
SplyAdrBalNtv1M	Val in Addrs w/ Bal ≥ 1M (native units)	... at least 1M ...
SplyAdrBalUSD1	Val in Addrs w/ Bal ≥ $1 USD	The sum of all native units being held in addresses whose balance was at least $1 at the end of that day. Only native units are considered.
SplyAdrBalUSD10	Val in Addrs w/ Bal ≥ $10 USD	... at least $10 ...
SplyAdrBalUSD100	Val in Addrs w/ Bal ≥ $100 USD	... at least $100 ...
SplyAdrBalUSD100K	Val in Addrs w/ Bal ≥ $100k USD	... at least $100K ...
SplyAdrBalUSD10K	Val in Addrs w/ Bal ≥ $10k USD	... at least $10K ...
SplyAdrBalUSD10M	Val in Addrs w/ Bal ≥ $10M USD	... at least $10M ...
SplyAdrBalUSD1K	Val in Addrs w/ Bal ≥ $1K USD	... at least $1K ...
SplyAdrBalUSD1M	Val in Addrs w/ Bal ≥ $1M USD	... at least $1M ...
SplyAdrTop100	Value in Top 100 Addrs (native units)	The sum of all native units held by the richest 100 addresses at the end of that time interval.

Continues on the next page

ID	Name	Description
SplyAdrTop10Pct	Value in Top 10% of Addrs (native units)	The sum of all native units held by the richest 10% of addresses at the end of that interval.
SplyAdrTop1Pct	Value in Top 1% of Addrs (native units)	The sum of all native units held by the richest 1% of addresses at the end of that interval.
SplyCur		
SplyExpFut10yr	10 Year Expected Supply (native units)	The sum of all native units counting current supply and including all those expected to be issued over the next 10 years from that day if the current known continuous issuance schedule is followed. Future expected hard-forks that will change the continuous issuance are not considered until the day they are activated/enforced.
SplyFF		
SplyMiner0HopAllNtv	Miner Supply (native units)	The sum of the balances of all mining entities. A mining entity is defined as an address that has been credited from a transaction debiting the 'FEES' or 'ISSUANCE' accounts in accordance with Coin Metric's Universal Blockchain Data Model (UBDM).
SplyMiner0HopAllUSD		
SplyMiner1HopAllNtv	Supply One Hop from Miners (native units)	The sum of the balances of all addresses within one hop of a mining entity. An address within one hop of a mining entity is defined as an address that has been credited from a transaction debiting the 'FEES' or 'ISSUANCE' accounts in accordance with Coin Metric's Universal Blockchain Data Model (UBDM), or any address that has been credited in a transaction sent by such an address.
SplyMiner1HopAllUSD	Supply One Hop from Miners (USD)	

Continues on the next page

ID	Name	Description
TxCnt	Tx Cnt	The sum count of transactions that day. Transactions represent a bundle of intended actions to alter the ledger initiated by a user (human or machine). On certain occasions, transactions are counted regardless of whether they result in the transfer of native units or not. As long as such transactions are recorded on the chain, they will be included in the calculation of this metric. Changes to the ledger algorithmically mandated by the protocol, such as coinbase transactions or post-launch new issuance, are not included here.
TxCntSec	Tx per Second Cnt	The sum count of transactions divided by the number of seconds that day.
TxTfrCnt	Xfer Cnt	The sum count of transfers that interval.
TxTfrValAdjNtv	Xfer'd Val, Adj (native units)	The sum of native units transferred that interval removing noise and certain artifacts. Also known as Adjusted Transfer Value (native units).
TxTfrValAdjUSD	Xfer'd Val, Adj (USD)	The USD value of the sum of native units transferred that interval removing noise and certain artifacts. Also known as Adjusted Transfer Value (USD).
TxTfrValMeanNtv	Mean Tx Size (native units)	The sum value of native units transferred divided by the count of transfers (i.e., the mean size of a transfer) between distinct addresses that interval.
TxTfrValMeanUSD	Mean Tx Size (USD)	The sum USD value of native units transferred divided by the count of transfers (i.e., the mean "size" in USD of a transfer) that interval.
TxTfrValMedNtv	Median Tx Size (native units)	The median count of native units transferred per transfer (i.e., the median "size" of a transfer) that interval.
TxTfrValMedUSD	Median Tx Size (USD)	The median USD value transferred per transfer (i.e., the median "size" in USD of a transfer) that interval.

Continues on the next page

ID	Name	Description
VelCur1yr	1 Year Current Supply Velocity	The ratio of the value transferred (i.e., the aggregate "size" of all transfers) in the trailing 1 year divided by the current supply on that day. It can be thought of as a rate of turnover – the number of times that an average native unit has been transferred in the past 1 year.
VtyDayRet180d	180 Day Volatility	The 180 days volatility, measured as the deviation of log returns
VtyDayRet30d	30 Day Volatility	The 30 days volatility, measured as the deviation of log returns