technische universität
dortmund

Case Studies

Summer 2023

Project 1
Forecasting Bitcoin

Author : ÖCAL KAPTAN
May 16, 2023

Group 4
In collaboration with MD JAMAL HOSSAIN,
HIEN THI THANH NGUYEN AND
THANH QUYEN NGUYEN

Lecturers: M.Sc. KARSTEN REICHOLD AND
M.Sc. SVEN PAPPERT

Contents

Figures

Tables

1.*Introduction*

Bitcoin is a digital currency that has seen a extreme rise in popularity over the last few years. It is increasingly attracting interest among researchers and investors, both from a theoretical and an empirical perspective. (Garratt & Wallace(2018), Hendrickson et al. (2016)) Its value has been the subject of much speculation, and many people are interested in understanding the factors that can influence its growth.

One of the key factors that can impact the growth rate of Bitcoin is the economic situation in the United States. Specifically, the Unemployment Rate, Inflation Rate, Federal Funds Rate, and growth rate of the S&P500 can all have an impact on Bitcoin's value. We see that ,there is typically less economic activity when the unemployment rate is high, which can result in reduced consumer spending and a decreased demand for Bitcoin. On the other hand, when the unemployment rate is low, there is often more economic activity, leading to increased consumer spending and a greater demand for Bitcoin. Similarly, high inflation rates can lead to decreased purchasing power for consumers and businesses, resulting in a lower demand for Bitcoin. However, Bitcoin has been considered a hedge against inflation, and its value may increase during periods of high inflation. Changes in the federal funds rate can also impact the value of the US dollar, which can in turn impact the value of Bitcoin. Finally, the growth rate of the S&P 500, which is a benchmark for the performance of the US stock market, can also impact Bitcoin's value. During times of economic stability, investors may be less likely to invest in alternative assets like Bitcoin. However, during times of economic uncertainty or market volatility, investors may turn to Bitcoin as a safe investment asset.

Overall, it's important to note that Bitcoin is a relatively new and volatile asset, and its value can be influenced by a range of unpredictable factors beyond just these economic indicators.

*Against this background, this paper relies on data sources by Coin Market Capital and by Fred-Database collecting the information of Bitcoin*prices and four additional macroeconomic variables to capture the dynamics of BTC[1]. `Geometric*

[1] In this paper, BTC is abbreviation for Bitcoin Growth Rate.

mean` transformation has been applied to obtain Bitcoin prices for monthly and quarterly datasets and growth rate transformation has been applied to Bitcoin, SP500. Also Inflation Rate has been derived from CPI[2] by using same growth rate formula which will be explained in Chapter 2.

In this paper, mostly focused on specific methods like AR (Autoregressive Process), VAR Vector Autoregressive Process) and U-MIDAS (Unrestricted Mixed-Data Sampling) processes that explained in the Chapter 3. To find the optimal lag, AIC (Akaike Information Criteria) have been used. Granger Causality has been tested in VAR(1) model. Finally, the methods have been evaluated by using RMFSE (Root Mean Forecast Squared Errors) and compared to each other.*

In the end it's been found that, AR(1) model has outperformed VAR models in quarterly datasets. This is due to the macroeconomic indicators that used in the study are not good in the quarterly observed time series for Bitcoin growth rate. Although, the first lag is not significant in AR(1) model , but still has the better RMSFE than VAR models. When U-MIDAS medals are included we have seen that U-MIDAS outperformed AR(1) model due to usage of high frequency variables in the model. Finally, VAR regression have been applied to monthly time series. With lag 3, VAR model have outperformed to AR ,VAR U-MIDAS models. This is again due to frequency of the time series , which includes more information about the variables. Additionally, we have tested the stationary of the variables in each dataset and found that UNEMP, INF, FEDF variables in monthly times series are not stationary.

The remaining paper is structured as follows. The following section presents a detailed data description will be given. In Chapter 3, describes the overall methodology. Chapter 4 provides empirical findings with discussion and an evaluation of the methods. Chapter 5 concludes the paper.

[2] Consumer Price Index

2. Data

2.1 Data Collection and Transformation

Data for the four variables which are S&P500, FEDF, UNEMP and INF[3] have been collected from ***Fred-Database***[4]*. Historical data starts from 1960 to 2023. Inflation rate transformation has been applied to* ***CPI***[5] *by using the growth rate formula below:*

$$100 * (y_t - y_{t-1})/y_{t-1}$$

After that, the growth rate transformation has been applied to S&P500 as well. When it comes to collect the data for Bitcoin, ***Coin Market Capital***[6] *is used to get historical data from 2010 to 2023. Next, the collected data for Bitcoin has been transformed to quarterly and monthly frequencies by using geometric mean. After that growth rate transformation has been applied as well. Transformed datasets start form 2010:Q4 to 2022:Q4, with 5 columns and 49 observations for each variables in quarterly dataset and 150 observations in monthly dataset which is start 2010 August to 2022 December. Both datasets have been checked for the missing values but not detected any.*

2.2 Statistical Properties

After the collection and the transformation processes, the datasets show the following statistical properties in ***Table 1*** *(Quarterly Dataset). When comparing BTC[7] to other economic indicators such as Unemployment Rate (UNEMP), the Federal*

[3] S&P500 : Stock market index that tracks 500 publicly traded domestic companies
FEDF: Federal Funds Rate in United States
UNEMP : Unemployment Rate in United States
INF : Inflation Rate in United States

[4] Source : https://research.stlouisfed.org/econ/mccracken/fred-databases

[5] CPI is an abbreviation for Consumer Price Index

[6] Source :https://coinmarketcap.com/currencies/bitcoin/

[7] In this paper ,mostly BTC is used abbreviation for Bitcoin growth rate

Funds Rate (FEDF), Inflation Rate(INFR), and the S&P 500 (SP500), it's clear that Bitcoin growth rates have been much more extreme in comparison. UNEMP has a mean of 5.27%, FEDF has a mean of 0.7855%, INF has a mean of 2.4791%, and SP500 has a mean of 2.580%. BTC have been much more volatile than these economic indicators. These extreme movements in BTC can be seen also in ***Figure 1.***

	BTC	UNEMP	SP500	FEDF	INF
Minimum	-67.26	3.567	-7.98787	0.0600	-0.9612
1st Quantile	-11.48	4.067	-0.09619	0.0933	0.3586
Median	19.59	5.283	3.51040	0.1550	0.5538
Mean	53.84	5.956	2.58010	0.6552	0.6391
3rd Quantile	78.40	7.683	5.69906	1.1025	0.8073
Maximum	613.45	12.967	13.41401	3.6533	2.3314

Table 1 : Statistical Properties of Quarterly Dataset

Figure 1 *shows the graphical representation of the quarterly (left) and monthly (right) datasets. It's hard to comment about other variables by looking at Figure 1 because of different scales and large movements in Bitcoin growth rate. That's why the scaled version of graphs have been added.* ***Figure 2*** *shows the scaled version of 4 macroeconomic variables in different plots with Bitcoin growth rate in the quarterly frequency. Monthly plots have not been displayed because it follows the same pattern visually. Figure 1 shows that, BTC has many extremely high peaks in both datasets. In the time period between 2010 and 2012, Bitcoin climbed by 600.0% (in quarterly plots). The reasons for these peaks before 2012 are still unknown, as Bitcoin didn't gain popularity until after 2013. This suggests that there weren't many exchange markets trading Bitcoin at that time.*

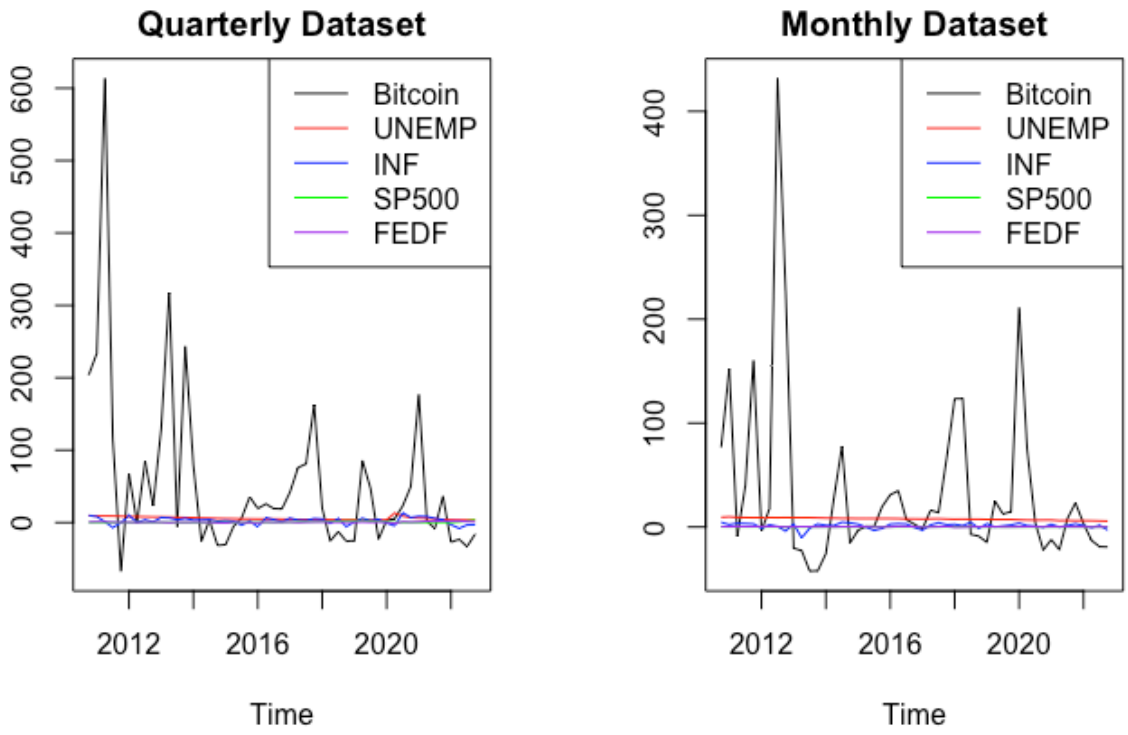


Figure 1: Graphs of Quarterly and Monthly Datasets

When it comes to commenting on the scaled versions, it's unclear whether there is a clear relationship between Bitcoin and other macroeconomic variables. One explanation could be that the relationship between these macroeconomic variables changes over time. For example, in ***Figure 2*** *Bitcoin and the S&P 500 appeared to be positively correlated from 2014 to 2020*[8]*, but it's hard to say something about their correlation after that time period.*

[8] The plot of the scaled monthly dataset shows the same patterns, that's why it's not displayed here.

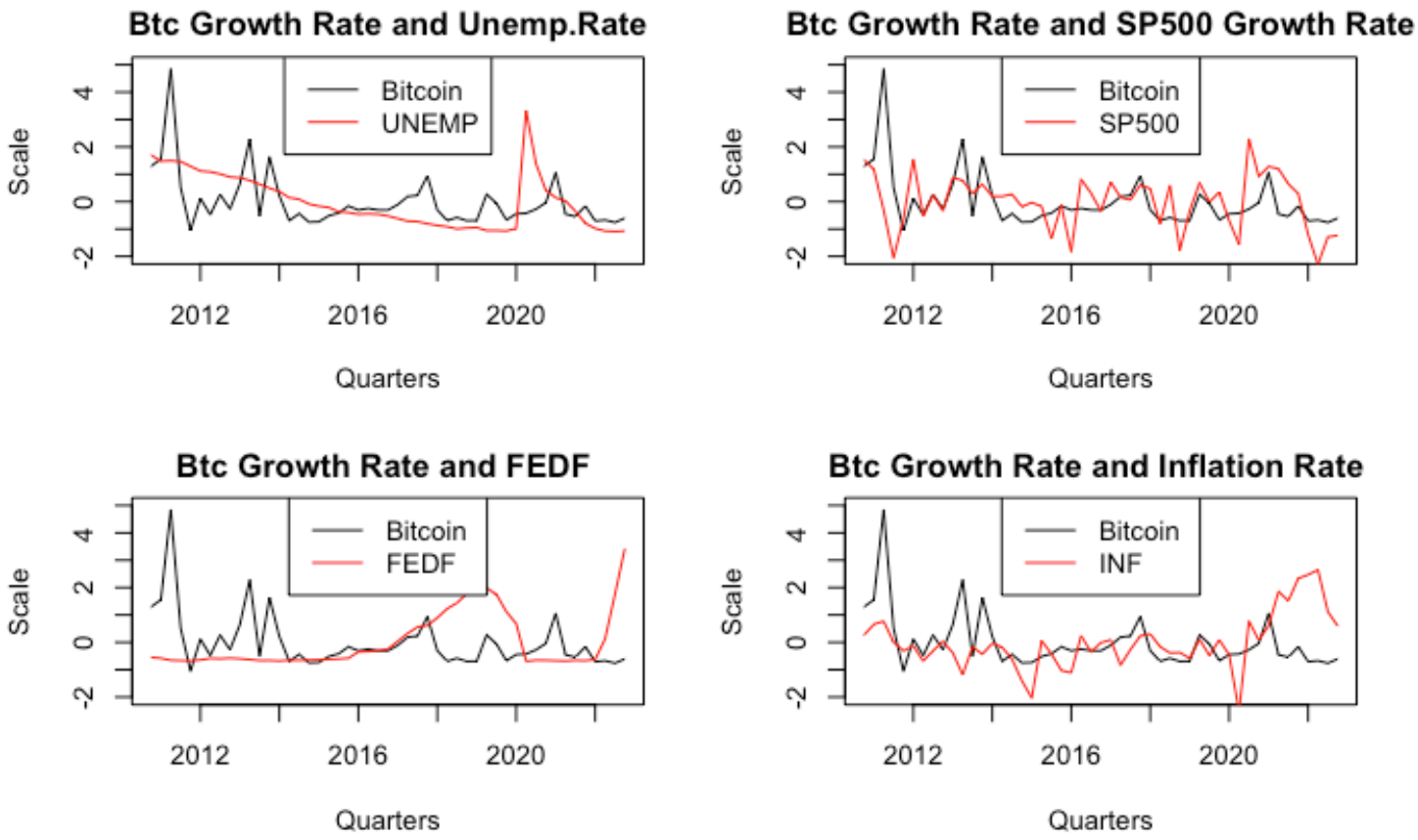


Figure 2 : Plots of the Scaled Dataset (Quarterly)

Overall, Bitcoin has experienced significant growth rates over the period 2010Q4 to 2022Q4, but it has also been highly volatile. The extreme fluctuations in BTC suggest that investing in BTC is a high-risk investment.

3. Methodology

Forecasting the future value of a time series variable is a common task in statistics. The goal is to predict the value of the variable at the next time point, based on historical data up to the current time point. This is known as ***one-step ahead*** *forecast. A more complicated task is to make multi-step ahead forecasts, predicting the value of the variable for several time points into the future.*

Some methods require some property assumption before using it. An example when it comes to use some statistical methods to model the data, the time series should be stationary. If it is not stationary time series has to make stationary by using some methods like exponential smoothing or first difference. After this transformation the specific models like ***Autoregressive*** *or* ***Vector Autoregressive*** *models can be used to model the time series. If a time series is non-stationary, its statistical properties can change over time, making it difficult to make accurate forecasts. For example, if a time series has a trend, meaning that it is increasing or decreasing over time, the mean value of the series will change over time, making it difficult to forecast future values based on past data. The stationary of a time series can be checked by looking at the plot or by using some statistical methods an example ADF[9] test. Here, all the variables are assumed as stationary time series.*

3.1 Autoregressive Process

As explained above, by assuming that a time series is stationary, that some statistical methods can be uses to model the data and make forecasts.The autoregression is a common method for making these forecasts. It expresses the conditional mean of a time series variable as a linear function of its own lagged values. In other words, it uses the variable's past values to predict its future values.

[9] ADF is the commonly used abbreviation for the Augmented Dickey-Fuller test, which is a statistical test used to determine whether a time series has a unit root (i.e., a trend) and is stationary or non-stationary.

Autoregressive model **AR(p)** *process can be modelled as :*

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + e_t \qquad (1)$$

The autoregressive coefficients, ϕ_1 *,.....,*ϕ_p *represent the weights given to the previous values of the time series in predicting the next value,* ***c*** *is a constant , the order of the autoregressive model,* ***p****, represents the number of previous values (lags) used in the model, and* e_t *is the error term at time t. Here****,*** *the error terms are supposed to be normally disturbed. In this project, one of the tasks requires working with an AR(1) process. This is an autoregressive model of order 1, where the next value in the series is modelled as a linear function of the previous value only. In other words, the current value at time* ***t*** *is highly correlated with the value at time* ***t-1.***

AR(1) process defined as below :

$$y_t = c + \phi y_{t-1} + e_t \qquad (2)$$

Once model has decided to use then coefficients can be estimated by using OLS (or any other method ,etc. Yule Walker). Assuming that the coefficients are estimated ,then the estimated coefficient can be used to predict the values for the desired time periods. Of course every method comes with advantages and disadvantages.

AR(p) and specially AR(1) models, in particular, are useful when the time series exhibits a strong autoregressive structure, meaning that the current value is heavily influenced by the previous values. This can be the case in many natural systems, where current conditions are largely determined by recent history. For example, the growth rate of the population mostly influenced by the size of the population in the

previous year. Therefore, AR(p) models ,might give inaccurate prediction when there is certain market conditions, such as financial crises or periods of rapid technological change.

3.2 Vector Autoregressive Process

Vector Autoregressive Model (VAR) is a multivariate extension of univariate autoregressive (AR) process where the lagged values of the series included in the model as regressors. In other words, past (lagged) variables contain information that are sufficient to capture the dynamics or relationship between variables of interest.

Here is a very simple example to VAR(2) model with ***K=3*** *endogenous variables* X_t, Y_t, Z_t *then the model equations ;*

$$Y_t = B_{10} + B_{11}Y_{t-1} + B_{12}Y_{t-2} + \phi_{11}X_{t-1} + \phi_{12}X_{t-2} + \theta_{11}X_{t-1} + \theta_{12}Xt - 2 + e_{1t} \quad (3)$$

$$X_t = B_{20} + B_{21}Y_{t-1} + B_{22}Y_{t-2} + \phi_{21}X_{t-1} + \phi_{22}X_{t-2} + \theta_{21}X_{t-1} + \theta_{22}X_{t-2} + e_{2t} \quad (4)$$

$$Z_t = B_{30} + B_{31}Y_{t-1} + B_{32}Y_{t-2} + \phi_{31}X_{t-1} + \phi_{32}X_{t-2} + \theta_{31}X_{t-1} + \theta * 32X_{t-2} + e_{3t} \quad (5)$$

VAR models differ from univariate autoregressive models because they allow feedback to occur between the variables in the model. It can capture the complex real-world behaviours, and also the dynamics of the time series. After construction the model, OLS can be use to estimate coefficients.
Once VAR method explained, then another important topic can be discussed. How to choose optimal lag for `VAR` process?

3.3 Akaike Information Criterion (AIC)

Finding the optimal lag important for the model. Because too many lag might over fit the model or too few lags might omit the some important information.

Generally these methods balance the benefits of including more lags against the costs of additional estimation error, allowing to select the optimal number of lags for

the regression. Akaike Information Criteria (AIC) is a commonly used lag selection criteria that used to evaluate the efficiency of a model by using the methods, goodness of fit and penalizes for higher number of parameters. According to Kilian, L., Lütkepohl, H. (2017) (page.55) , AIC can be calculated as below:

$$AIC = log(det(\tilde{\Sigma})) + (2/T) * (p * K^2 + K) \quad (6)$$

where p is the lag, K equations in the VAR model ,c_T = 2/T which c_T is a sequence of weights and $\tilde{\Sigma}$ is the residual covariance matrix estimator.

3.4 Granger Causality

Granger causality is a statistical concept that is often used in econometrics to test whether one time series is useful for forecasting another time series. In the context of vector autoregressive (VAR) models, Granger causality tests are typically conducted by estimating a VAR model for a set of variables, and then testing whether the lagged values. The null hypothesis of the `Granger` causality test is that the lagged values of the other variable do not help to predict the current and future values of the variable. If the null hypothesis is rejected, then we conclude that the other variable Granger-causes the variable of interest.

Kilian and Lütkepohl (2017)[10] explain that[11] the `Granger causality` concept to a multivariate setting and show how to test for `Granger causality` between multiple variables. For example, we could test whether the past values of the unemployment rate help to predict the current value of BTC If there is an evidence of Granger causality between Unemployment Rate and BTC, we would conclude that Unemployment Rate Grange causes BTC.

Discusses the concept of Granger causality in Chapters 2.5 and 7

[11] In Chapter 2.5 of Kilian and Lütkepohl (2017)] a variable `X` `Granger causes` `Y` if the inclusion of past values of `X` in the forecasting model of `Y` improves the accuracy of the forecasts. They extend ^[In Chapter 7 of Kilian and Lütkepohl

Granger Causality should be used when :

-Focus is forecasting performance ,not the theoretical model behind the forecast.

-To use it, data is assumed to be stationary.

As conclusion Granger causality is useful to improve forecast accuracy, model reliability and it can reduce time running on the invalid models.

3.5 Unrestricted Mixed Data Sampling (U-MIDAS)

*Working with the time series data in different frequencies can be challenging. The unrestricted mixed data sampling **(U-MIDAS)** method can be used to overcome this issue. U-MIDAS is an econometric model that incorporates high-frequency data as proxy variables for low-frequency data, providing more accurate and timely forecasts. This approach also allows for more reliable estimates of dynamic relationships between variables. In other words, it rely on low-frequency data by integrating the high frequency data.*

*However, U-MIDAS models can also be more complex to estimate, as they require the use of mixed frequency regressions **(MFR)** and the selection of appropriate proxy variables. An example to U-MIDAS model:*

$$g_{y,3s} = \mu_0 + \mu_1 g_{y,3(s-1)} + \Phi_0 X_{3s-1} + \Phi_1 X_{3s-2} + \ldots + \Phi_K X_{3s-(K+1)} + \epsilon_{3s} \quad (7)$$

*$g_{y,3s}$ represent target variable in quarterly frequency, $g_{y,3(s-1)}$ quarter before **3s, X** is vector of collection of the variables, **K** is the lag and error terms ϵ_{3s}. The goal is estimating the coefficients and predict the target variable in given time. So, this can be done by OLS.*

3.6 Root Mean Squared Forecast Error (RMSFE)

*When it comes to discussing the **RMSFE**, let's first briefly talk about the error terms for each of the methods that we explained earlier. Then we can move on to discussing the root mean squared forecast error. Let's start with **AR** models.*

AR models assume that the error terms of a stationary time series are normally disturbited with constant mean , finite variance and uncorrelated over time. VAR models also assume that error terms are normally disturbed and i.i.d. The difference between VAR and AR models is that VAR models allow for multiple dependent variables to be modelled simultaneously. The assumption of i.i.d. error terms in VAR models is necessary for the estimation of the model parameters using maximum likelihood or other statistical methods (etc. OLS).

In UMIDAS models, the error term assumptions are similar to those of AR and VAR models. U-MIDAS models assume that the errors are normally distributed with constant variance and that the proxy variables used to represent the low-frequency data are valid and have a stable relationship with the target variable over time.

Overall, these assumptions are necessary for the models to accurately capture the temporal dependence in the data and provide reliable estimates of the model parameters. However, it is important to note that these assumptions may not hold in practice, and it is important to test the validity of the assumptions using diagnostic tests such as the ***Ljung-Box Test*** *or the* ***Breusch-Godfrey Test*** *for autocorrelation, and* ***Jarque-Bera Test*** *can be used to check normality of the error terms.*

To evaluate the models, ***RMFSE*** *has been used which is stands for Root Mean Squared Forecast Error. RMFSE measures the square root of the average squared differences between the predicted values and the actual values. It gives an idea of how much the predicted values change from the actual values on average.*
RMFSE can be calculated as :

$$RMFSE = \sqrt{1/T \sum_{t=1}^{T} (y_{t+1} - \hat{y}_{t+1|t})^2} \quad (8)$$

4. Empirical Results

In this chapter, all the methods that have explained before will be used to model the data and make predictions. At the end of the chapter, the evaluation of each methods will be made by using RMFSE.

4.1 Autoregressive Model AR(1)

AR(1) *model uses the first lag of the variable, assuming that the future or current price is highly correlated with the previous price. By using the simple* ***AR(1),*** *model the model has been constructed as described in chapter 3 and the predictions have been made accordingly. The coefficients were estimated using the* ***OLS****, which stands for Ordinary Least Squares, and the predictions were made based on those coefficients. The predictions start from the 3rd observation because the* ***OLS*** *method requires at least two observations to estimate the coefficients.* ***Figure 3*** *shows the predictions of BTC using the* ***AR(1)*** *model. The red points represent the predictions, while the black points represent the actual BTC values for each quarter.*

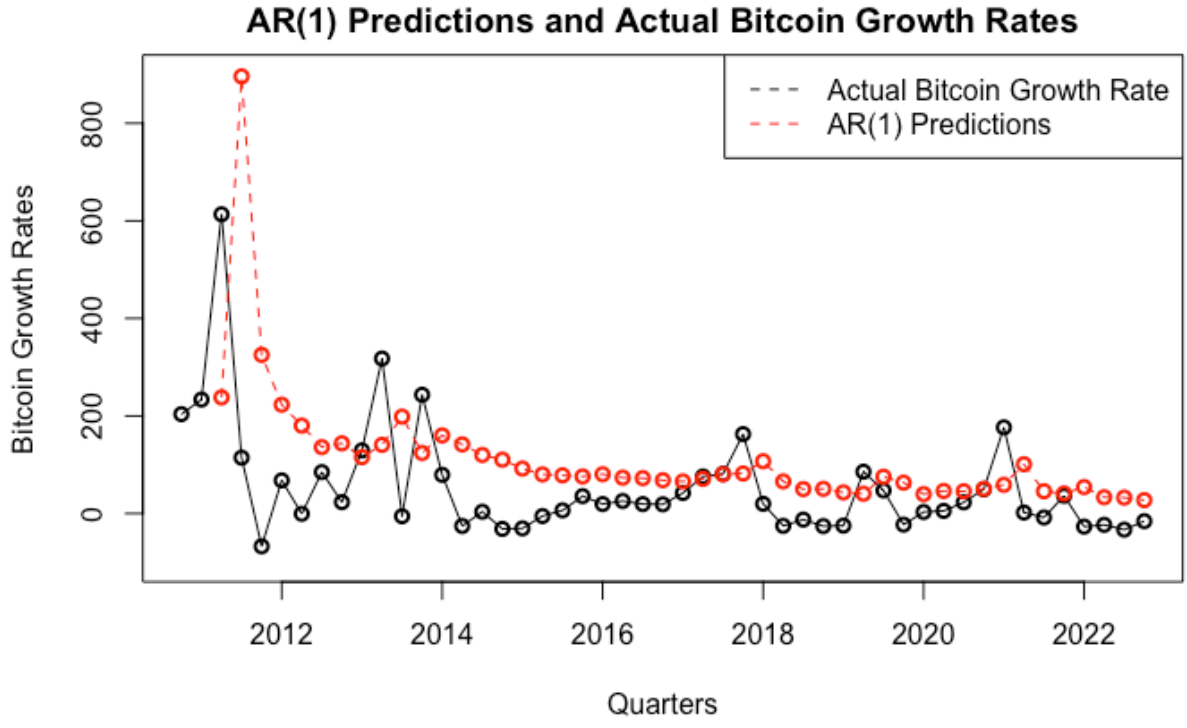


Figure 3: AR(1) predictions from 2011Q2 to 2022Q4

The predictions and residuals have obtained, lets check some assumptions of the AR(1) process.

It is assumed that the time series is stationary, and this can be shown by examining the coefficient ϕ of the estimated model. For a stationary time series in AR(1) setting, ϕ is expected to be $|\phi| \leq 1$. In this AR(1) setting, ϕ has found as 0.31, that shows the time series is stationary. This can be also checked running ***ADF*** *test.* ***p-value*** *of the ADF test has found as* ***0.01*** *which means that H_0[12] can be rejected since* ***p-value*** *is less than* ***0.05*** *percent level.*

Secondly, as explained before, AR(1) model assumes that first lag is correlated with the current value. But there is no correlation when PACF of BTC visualised . This can be confirmed also by running a Breusch-Godfrey Test [13] on BTC. So, ***p-value*** *for the test comes bigger than 0.05 level which shows that there is no autocorrelation. In addition, we can examine the normality of residuals using the Jarque-Bera Test or by inspecting a plot of the residuals on a linear line. In this case, the Jarque-Bera test indicates that the residuals do not follow the normality assumption (p-value of the corresponding test statistic is* ***2.22e-16****).*

4.2 Vector Autoregressive Model

4.2.1 VAR(1)

Figure 4 *Shows VAR(1) with AR(1)*and actual BTC. It obvious that VAR(1) does not capture dynamics of Actual BTC but some periods it fits well.(etc, 2022Q1 to 2022Q4). Mostly, VAR(1) overfits the data. The reason that ,VAR(1) includes the first lag of all the variables in the model which are not significant. Before discussing the significancy of the coefficients, lets check the stationarity of the variables. This can be check again by ADF test. Regarding ADF Test H_0 is failed to reject. p-value of UNEMP which is* ***0.44****, it is higher than 0.05 which means that UNEMP rate is non-stationary. Other variables has been obtained as stationary.*

---

[12] Null hypothesis claims that time series is non-stationary

[13] Breusch-Godfrey Test says that under Null hypothesis there is no autocorrelation.

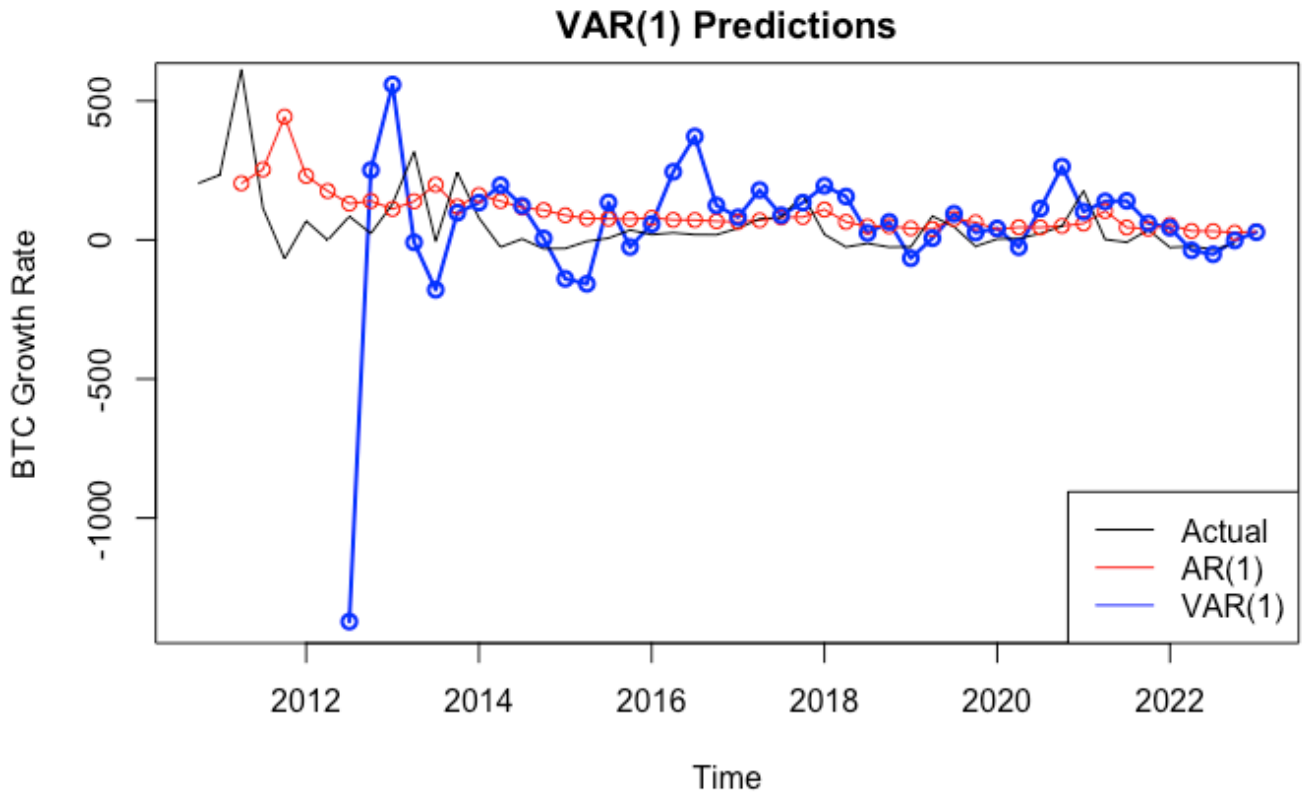


Figure 4: VAR(1) and AR(1) predictions with Actual BTC.

By checking t-values of the coefficients from OLS estimation, lag 1 of the UNEMP is the only significant regressor. This can be confirmed by running Granger Causality Test which uses the F-test. According to F-test, H_0 *for the variables INF , SP500 and FEDF have been failed to reject but* H_0 *for UNEMP has been rejected . Which means that UNEMP rate Granger cause to BTC. Respective p-value is* ***0.048*** *which is lower than 0.05 significance level. But as showed above , UNEMP is not stationary. For the Granger Causality Test , the stationarity of the time series is required.*

4.2.2 VAR(3)

Before proceeding to the predictions of VAR(p) model, it is necessary to determine the optimal lag. As defined in the chapter 3 , the formula which computes the AIC (Akaike Information Criterion) with the maximum lag restricted to 3 to utilize the entire sample, as suggested by Kilian and Lütkepohl (2017, p.55). After

computing the AIC for the entire sample, the optimal lag was found to be 3, and this value was set as the optimal lag for the model.

As shown in the ***Figure 5****, VAR(3) did not capture the dynamics of the actual data. Due to insignificant lags, the model failed to capture the true dynamics of the data and overfit the model.*

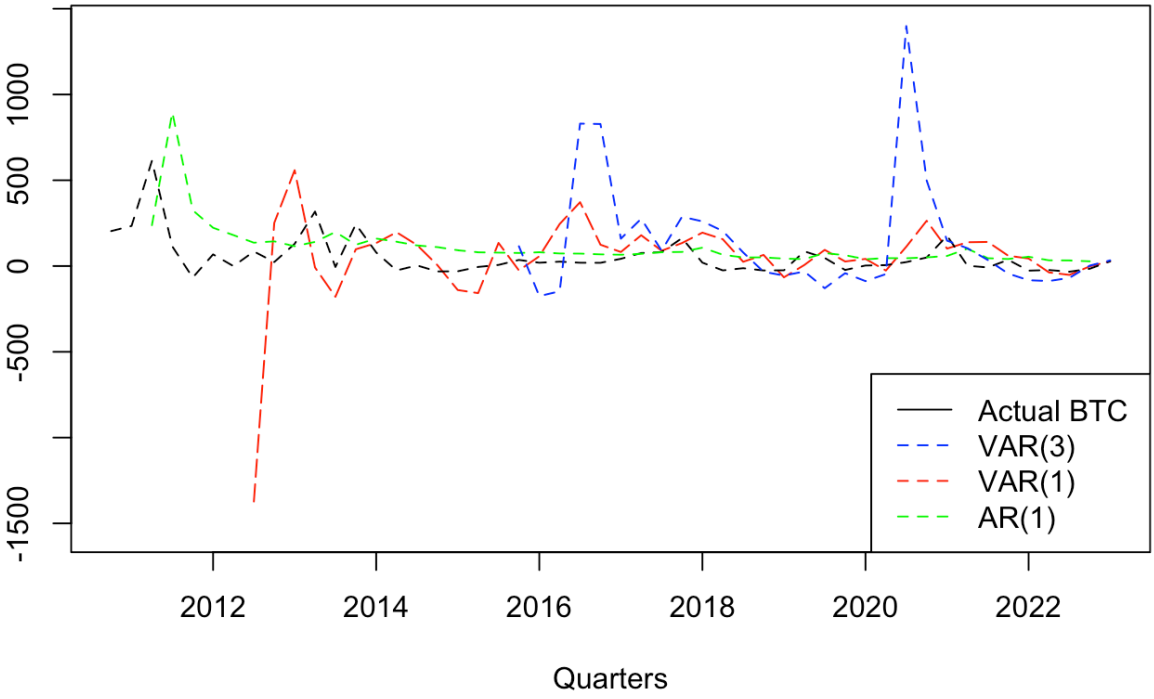


Figure 5: Predictions to VAR and AR models with Actual BTC

4.2.3 Evaluation of the VAR and AR model

The evaluation of the models, both RMFSE and adjusted R^2 *have been used.* ***Table 2*** *shows the RMFSE of each model on the X axis and the RMFSE of each method on the Y axis. It is clear that the AR(1) model has the lowest RMFSE among them. The reason for this is that the AR(1) model only uses the first lag of the BTC as a regressor, which makes it less complicated than the others. Including additional variables and their lags did not improve the models. VAR(3) model, has the lowest Adj.*R^2 *and higher RMSFE than AR(1) model. Similarly, VAR(1) model has the*

highest Adj.R^2, but its RMSFE is very high compared to the others. Even though there is no autocorrelation at the first lag of BTC, it still makes sense to use the AR(1) model instead of the VAR(1) and VAR(3) models. More discussion on this will be held in the last chapter.

Models	Adj.R^2	RMSFE
AR(1)	0,10	164.36
VAR(1)	0,178	268.82
VAR(3)	0,079	353.43

Table 2: Comparison the Models with RMSFE and Adjusted R^2

4.3 Unrestricted Mixed-Data Sampling (U-MIDAS)

*U-MIDAS model was constructed for different values of $K+1$ where $1 \leq K \leq 3$. To find the optimal lag, the AIC criterion has been used, however the AIC formula from (**6**) cannot be applied here. Since the model U-MIDAS, has only one equation, not like the whole system in VAR. That's why the formula (9) is used instead of (6) in the chapter 3.*

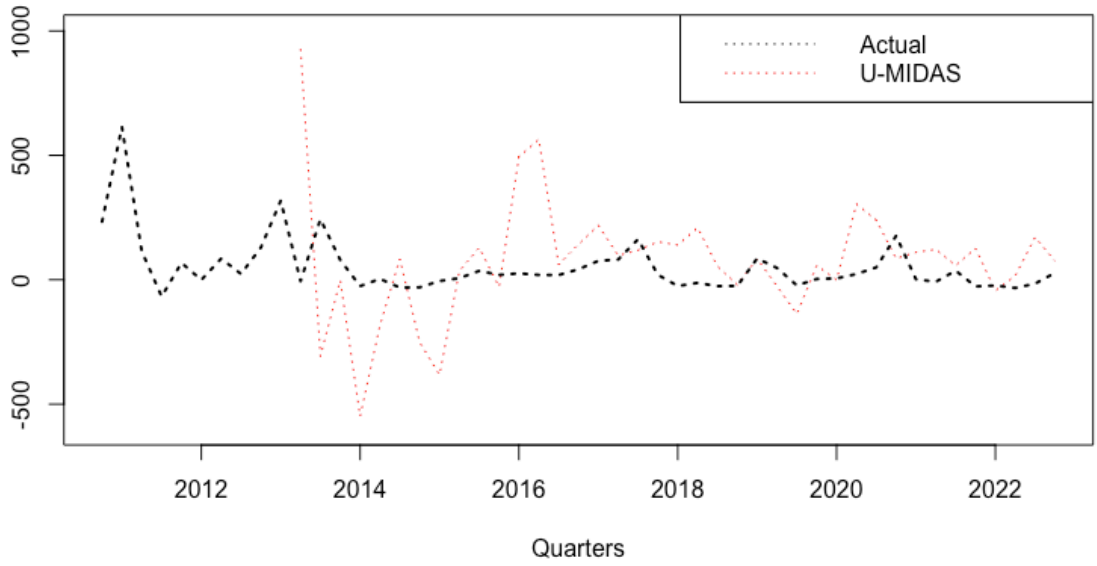
$$AIC = log(SSR/T) + (p+1) * T/2 \qquad (9)$$

The method was applied for $K = 1,2,3$ using the entire sample, and the optimal lag was found to be $K = 1$. Then U-MIDAS with $K = 1$ model can be defined as:

$$g_{y,3s} = u_0 + u_1 g_{y,3(s-1)} + \phi_0 UN_{3s-1} + \theta_0 FE_{3s-1} + \omega_0 IN_{3s-1} + \delta_0 SP_{3s-1} +$$
$$\phi_1 UN_{3s-2} + \theta_1 FE_{3s-2} + \omega_1 IN_{3s-2} + \delta_1 SP_{3s-2} + e_{3s} \qquad (10)$$

*where the model includes the first lag of the BTC, first and second lag of the other four variables. The predictions of U-MIDAS are represented in **Figure 6**.*

Figure 6: *U-MIDAS Full Model Predictions*



U-MIDAS captures dynamics of the model but it is still not good as AR(1). This is due to insignificant variables that included in the model. The significance of the coefficients from U-MIDAS model can be checked by using t-test. T-values and respective probabilities of the coefficients[14] obtained by ***summary()*** *function in* ***R.*** *According to* ***Table 3*** *, first lag of UNEMP at 0.10 % significance level is different than zero. Excluding other coefficients[15], reduced U-MIDAS model has been performed with lag K=1.*

Coefficients	Estimate	Std. Error	T-value	Pr(>|t|)
ϕ_0	109.8536	58.0021	1.86	0.0798 .

Table 3 : T-values and respective probabilities of the estimated coefficients

[14] The coefficients are not significantly different from zero , are not displayed in the table.

[15] The coefficients that are not different from zero.

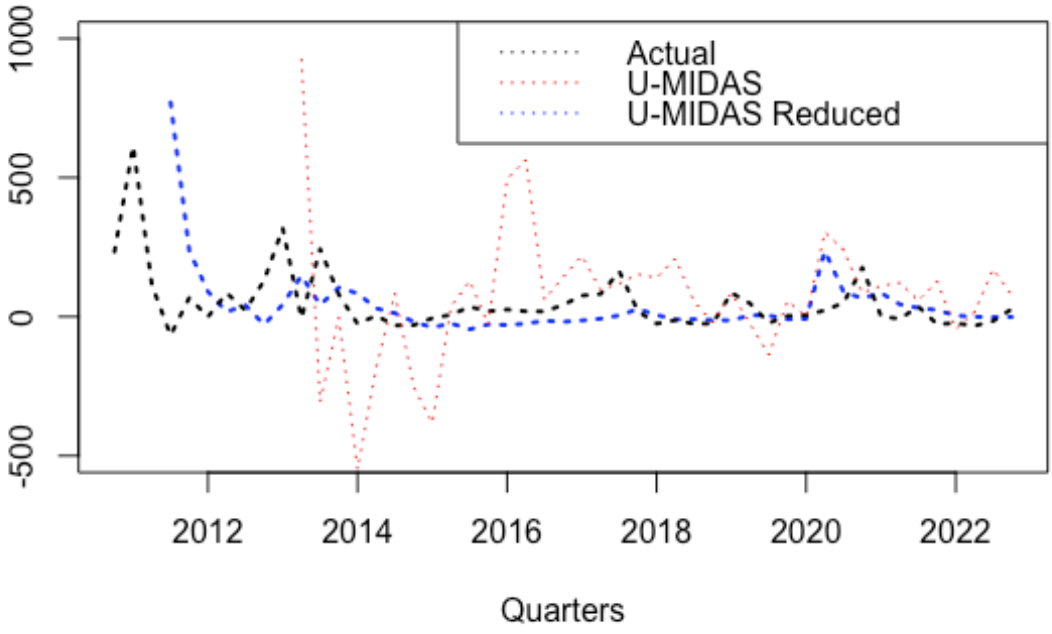


Figure 7 : *Predictions of U-MIDAS Models with Actual BTC*

Figure 7 *shows the predictions of the reduced U-MIDAS where insignificant regressors (lags) are excluded in the model. As shown on the* ***Table 4****, reduced model has been increased the model accuracy by decreasing the RMSFE and increasing the Adj.* R^2***.***

Models	Adj.R^2	RMSFE
U-MIDAS	94	239.21
U-MIDAS Reduced	0.1336	150.64

Table 4: Comparison of the U-MIDAS Models with RMSFE and Adjusted R^2

Additionally, the normality assumption and autocorrelation of the residuals have been checked by running Jarque-Bera Test and Partial Autocorrelation Function. According to PACF , there is no autocorrelation between the lags of residuals, and Jarque-Bera Test have confirmed the normality assumption don't hold for the residuals of U-MIDAS and Reduced U-MIDAS models.

4.4 VAR(p) Monthly Data

As discussed before in the Chapter 4.2.2, same method has been used to model monthly observed time series. By restricting the maximum lag to 12, the optimal lag has been obtained as 3. ***Figure 8****. Shows the predictions from VAR(3).*

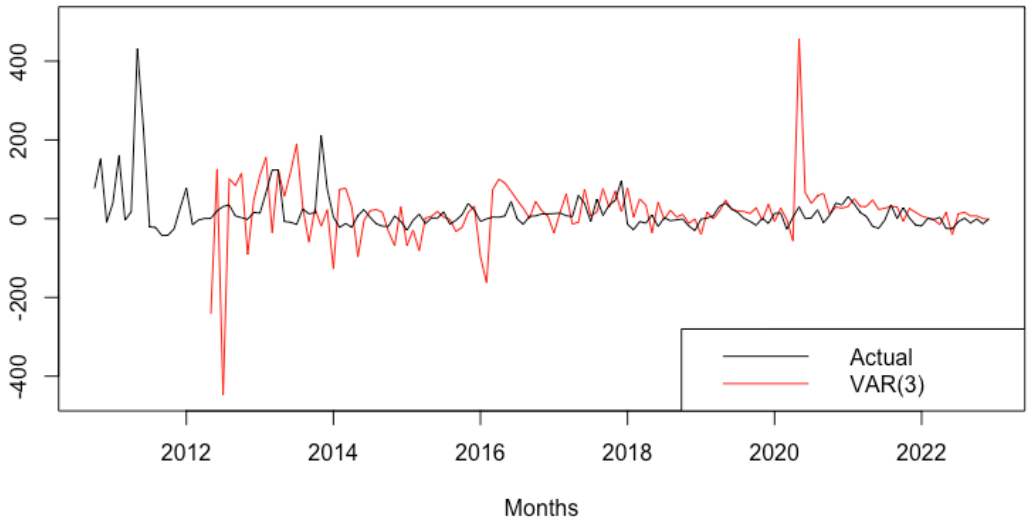


Figure 8 :*VAR(3) monthly predictions with Actual BTC*

The predictions from the VAR(3) method fit the data well and capture the dynamics of the model mostly. The RMSFE and adjusted R-squared have been reported as ***83.92*** *and* ***19.91%****, respectively, which is better than expected. Although the method uses 3 lags of the variables as regressors in the model, It still achieves higher accuracy compared to other models. It seems that the accuracy of the models increases when high-frequency data has been included in the model.*

Additionally, the residuals have been checked for the normality and autocorrelation. It has been found that the residuals doesn't hold the normality assumption for the VAR(3). This might be due of insignificant and non-stationary variables that used in the model or the model is not true data generating process for BTC. ***Table 5*** *shows the ADF test for the variables in VAR() model. As seen on the table, Except BTC and SP500, the p-values of the variables are bigger than at*

0.01,0.05, 0.1 *level of significance. Which tells that they are non-stationary. Using a simple exponential smoothing or taking first difference of the variables would fix the non-stationary problem and would give higher accuracy. VAR(3)[16] model with the stationary times series has performed again and RMSFE has been obtained as* ***61.79*** *which is higher than before.*

Variables	Level	P-values
BTC	-4.60	0.01
UNEMP	-2.74	0.26
FEDF	-2.76	0.2582
SP500	-5.23	0.01
INF	-3.29	0.07484

Table 5 : ADF Test of the Variables

5. Conclusion

Considering the performance of all models, we can conclude that the AR(1) model outperforms VAR(1) and VAR(3) for low-frequency time series predictions. So, AR and VAR models are very sensitive to economic events like economic crises etc. . As shown in Figure 5, the 2020 pandemic heavily affected the global economy, especially in the United States. Excluding these periods and making predictions accordingly would increase the accuracy of the model. When U-MIDAS models are included in the analysis, incorporating high-frequency data into the model have

[16]First difference of the non stationary variables are used

improved the accuracy. Reduced U-MIDAS model showed that it has been performed much better than other models in the previous sections. The model could be improved by excluding some periods where there are extreme peaks (Covid.19) in the model. The monthly time series in VAR(3) settings showed that even with non-stationary variables, the model outperformed other models. Because high-frequency time series have more information and the models can capture the dynamics the times series better than low frequency time series. We can conclude that that the accuracy of the models increases when high-frequency data is included in the model.

In the end, considering the predictions of all the models that, the most relevant macroeconomic indicator for BTC is Unemployment Rate[17]. The unemployment rate is closely related to a country's economic growth. (Mukit, Abdel-Razzaq, Islam, 2020) Since , the unemployment rate is a indicator of the economic growth in the country. If there is low unemployment rate then there is less economic activity. Which means that, if there is a economic growth this allowing people to invest more, therefore the demand for Bitcoin , Gold and S&P500 are increasing as well.

We should also consider that the unemployment rate (also the Inflation rate) of other countries like China as an indicator for Bitcoin growth rate. Which means that macroeconomic indicators of China might increase the accuracy of the models as well.

[17] However, the unemployment rate is not significant for monthly VAR(p) settings due to non-stationarity. Taking the first difference of the unemployment rate would increase the accuracy of the model.

6. References

1. *Ghysels, E., Santa-Clara, P., & Valkanov, R. (2002). The MIDAS touch: Mixed data sampling regression models*
2. *Forni, M., Gambetti, L. (2016). The dynamic effects of monetary policy: A structural factor model approach with mixed frequency data.*
3. *Lütkepohl, H. (2006). New Introduction to Multiple Time Series Analysis. Springer Science & Business Media.*
4. *Foroni, I., Guérin, P., & Marcellino, M. (2017). Markov-switching MIDAS models. Journal of Applied Econometrics, (Page 32)*
5. *Foroni, C., Marcellino, M., Schumacher C., U-MIDAS: MIDAS regressions with unrestricted lag polynomials.*
6. *Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic control 19, 716 – 723.*
7. *Kilian, L., Lütkepohl, H. (2017). Structural Vector Autoregressive Analysis. Cambridge University Press, Cambridge. Chapter 2*
8. *Hanck, C., Arnold, M., Gerber, A., and Schmelzer A.,(2023-5-5) -Introduction to Econometrics with R, Chapters 14,15,16*
9. *Stock, J. H., and M. W. Watson. (2015). Introduction to Econometrics, Fourth Update, Global Edition. Pearson Education Limited, Chapter 14,15*
10. *MUKIT, M., ABDEL-RAZZAQ, A., ISLAM, M.,(2020) Relationship between Unemployment and Macroeconomics Aggregates: Evidence from Bangladesh*
11. *Mario Arturo Ruiz Estrada, How Inflation and Unemployment can affect the Cryptocurrencies' Performance: The Case of Bitcoin*