

# ECU33143 Introduction to Big Data in Economics (Updated) Research Proposal

Raj Aryan Upadhyaya (Student number 23375852)

## Research Question:

My research question develops in the field of Quantitative Finance, specifically in Cryptocurrencies.

I have been investigating the price movement of cryptocurrencies in comparison to macroeconomic indicators, and also with regards to the relation with sentiment analysis.

**Main aim:** I wish to see whether or not cryptocurrencies have integrated more into the macroeconomy versus before, where it was primarily driven by social sentiment, and potentially modelling to see what comes next.

## Literature Review:

Alongside my research project, I have recently concluded a Literature Review (for 2 main reasons - 1) to really get a head start on my dissertation, and 2) to understand what methods to look at, and why they would be useful), where I took 2 papers from each of the following topics (all sourced from Google Scholar search).

- 1) Cryptocurrency market efficiency
  - a. [Sahoo and Sethi \(2022\)](#)
  - b. [Souza and Carvalho \(2022\)](#)

Findings:

- Both papers found cryptocurrencies to be informationally efficient (i.e. the efficiency of the cryptocurrency market depended on the level of information present for the price to move), with some exceptions (as we see with Sahoo & Sethi).
- Souza and Carvalho (2022) validates the need for utilising higher-frequency data in order to explain the cause of more abrupt shifts, as well as emphasising on looking at this from a multivariate perspective, as cryptocurrency exchanges show causality when compared amongst one another in the Granger sense and also allows to gain a broad understanding of the inherent risks associated with cryptocurrencies.

- Sahoo & Sethi (2022) found exceptions for information efficiency in Ripple (one of my tested cryptocurrencies), DASH, and XMR, but the majority follow an efficient market hypothesis.

## 2) Cryptocurrency links to Macroeconomic Data

- [Sakemoto and Nakagawa \(2021\)](#)
- [Baranovskyi et al \(2021\)](#)

### Findings:

- Both papers found a correlation between Macro and Cryptocurrencies.
- Sakemoto and Nakagawa (2021) found strong correlations through the use of a factor model (using Output, Labour, housing, bonds prices, amongst other factors) previously discussed in S&W, where Inflation and Money supply don't tend to affect it too much. Primarily used regression to test their theory.
- Baranovskyi et al (2021) found there is correlation, and that multi-regression paired with correlation analysis has shown the market's impact on modern investment vehicles such as cryptocurrencies. Their main methods used were multiple regression as well as machine learning techniques.

## 3) Cryptocurrency links to Social Media via Sentiment Analysis

- [Colianni et al \(2015\)](#)
- [Wooley et al \(2019\)](#)

### Findings:

- Both tended to find a link between cryptocurrency price shifts with regards to sentiment analysis.
- Colianni et al (2015) looked at this from a sentiment perspective on Twitter, where they found similar, well-correlated results, and had looked at Logistic Regression, with Bernoulli Naïve Bayes methods.
- Wooley et al (2019) were looking at this from a sentiment perspective on Reddit, where they had looked at sentiment time series, amongst other forms of time series analysis in order to highlight where there is correlation, which was proven via Reddit. They had used VADER[21] as their Sentiment model, and had also utilised a rolling z-score for analysis, GARCH for volatility, and a Self-Organising Fuzzy Neural Network to look at non-linear complex relationships.

4) Existing work that combines both Cryptocurrency-Macro and Cryptocurrency-Sentiment analyses.

a. [Kabo et al \(2025\)](#)

b. [Frendo \(2025\)](#)

Findings:

- Both of these papers found that there was a need to further explore studies of this kind, and both of these papers approached the area in different ways, which carried out their own results. Both of these papers being new also indicates that there is indeed progression being made on amalgamating both cryptocurrency-macro and cryptocurrency-sentiment in analysis of cryptocurrencies.
- Kabo et al (2025) were looking purely from a predictive standpoint, investigated primarily on bitcoin, and only tracked GDP, but otherwise showed that there is indeed feasibility for my work, as they had built a model by integrating ARIMA and LSTM, which had improved by combining both sentiment and GDP into their model.
- Frendo (2025) looked at this primarily from the perspective of cryptocurrency-macro, but had integrated cryptocurrency-sentiment, allowing them to find meaningful insight into cryptocurrency price movement, but had suggested to consider regulations for sentiment, as well as more robust methodology in order to fully capture cryptocurrency price movements.

Overall, we found that a lot of the studies had utilised Granger Causality, allowing us to use that to pinpoint the relationships between the different factors involved, and similarly, had given insight towards how to better structure this project for methodology going forward in ways I may not have considered before (such as a self-organising Fuzzy Neural Network, should I choose to utilise one)

## Data Sources:

For all my data extracted,

1. I used the [FRED API](#) and the python yfinance library to extract macroeconomic data for the following variables: CPI, Fed Funds Rate, Money Supply, 10 year bond yields, and Crude Oil Prices, as well as other indexes on yfinance.
2. I used the [Binance API](#) to look at hourly historical cryptocurrency pricing through querying for market data for 10 popular cryptocurrencies, including Bitcoin, Ripple, Ethereum, and DOGE amongst others.
3. I would like to use Reddit and X APIs to extract post data for Sentiment Analysis.
4. Additional data may be incorporated if relevant.

## Data Cleaning:

For all data involved, I intend to clean any missing data primarily through forward fills, with some backward fills where NaN entries still exist since my data is time series, and I am using different period changes for some metrics (i.e. Macro data tends to update much less frequently vs Cryptocurrency price data), and for this, I have utilised forward fills in order to match the data to the appropriate timing for the Cryptocurrency price data, which allows me to match cryptocurrency price data where needed.

## Methods (and implementation pipeline):

**Stage 1:** ~~Conduct regression analysis (time-series OLS, Vector Autoregression (VaR)) to establish baseline relationships between macroeconomic indicators and Cryptocurrency prices.~~ (Completed)

**Stage 2:** ~~Introduce Granger Causality and IRF visualisation to VaR, allowing us to highlight Causal relationships between variables and coins.~~ (Completed)

**Stage 3:** Integrate sentiment analysis from social media platforms (such as X (formerly Twitter) or Reddit) to assess whether market sentiment provides additional explanatory power.

**Stage 4:** Introduce Volatility Modelling/ML-based prediction to explore potential nonlinear/complex relations I may have missed.

**Stage 5 (optional):** Utilise ML to try and find further factors which might cause price shifts otherwise/in the future for cryptocurrencies.

**As I work through this, it's a continuous process:**

Compare and evaluate everything found so far to determine robustness and draw conclusions on the key drivers of Cryptocurrency pricing.

## Findings so far:

**Stage 1:**

- After preliminary static Linear Regression, I had realised that while the  $R^2$  ranged from 0.63 (AVA) to 0.96 (BTC), the other factors within the regression analyses, such as the Durbin-Watson statistics being low on average ( $<0.1$ ) and high condition numbers ( $>1.5e6$ ) indicate potential autocorrelation and multicollinearity issues, warranting further time-series correction (e.g., via differencing or VAR models).

- This led towards utilising a VAR model, which showed more plausible results, which you can see by referring to Figures 1-10

**Stage 2:**

- After introducing Granger Causality and visualisation, we were able to find that for approximately half of the variables involved, there has been a link that's developed (namely via USD Index, the S&P 500, VIX, and to an extent, gold prices), whereas there still are some parts of the macro infrastructure that have also been mixed (such as the M2 Money Supply (liquidity)), but otherwise, the rest tend to show that while cryptocurrencies have integrated as a risk asset, where the motions between how it moves and how certain indexes are moving as well show a progressing cryptocurrency-macro link from 2016, there is still a long way to go.

## Figures:

Figure 1

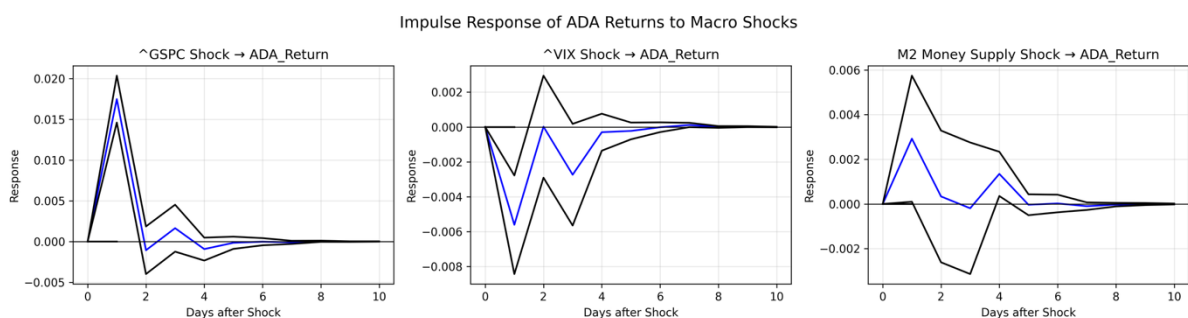


Figure 2

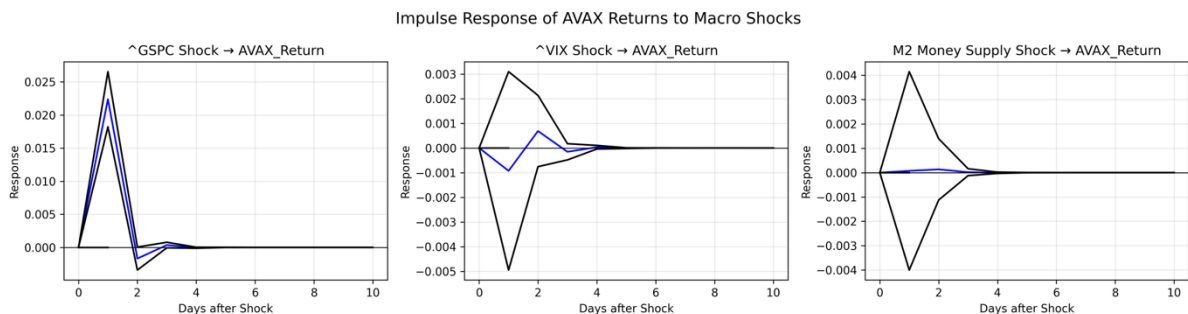


Figure 3

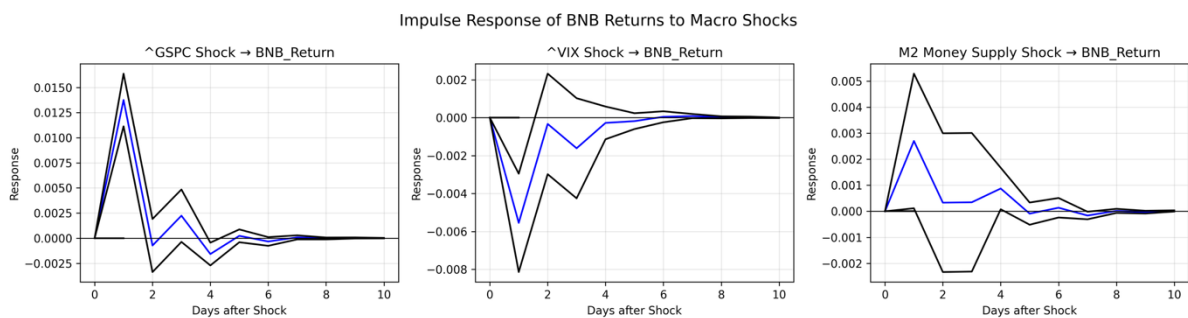


Figure 4

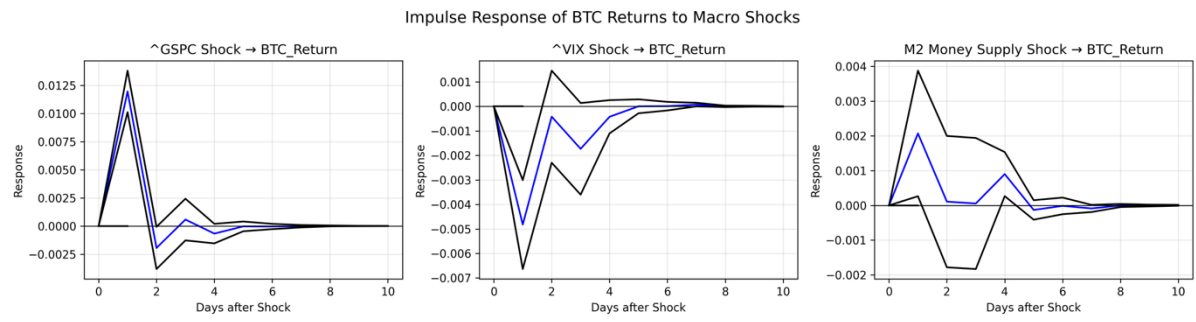


Figure 5

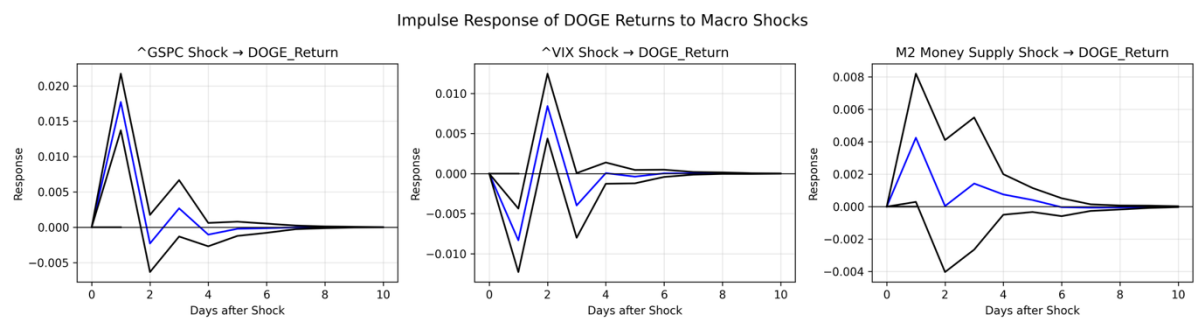


Figure 6

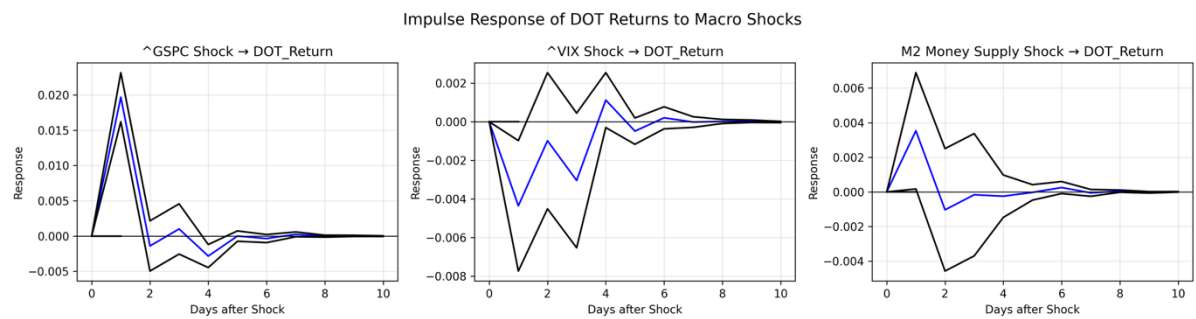


Figure 7

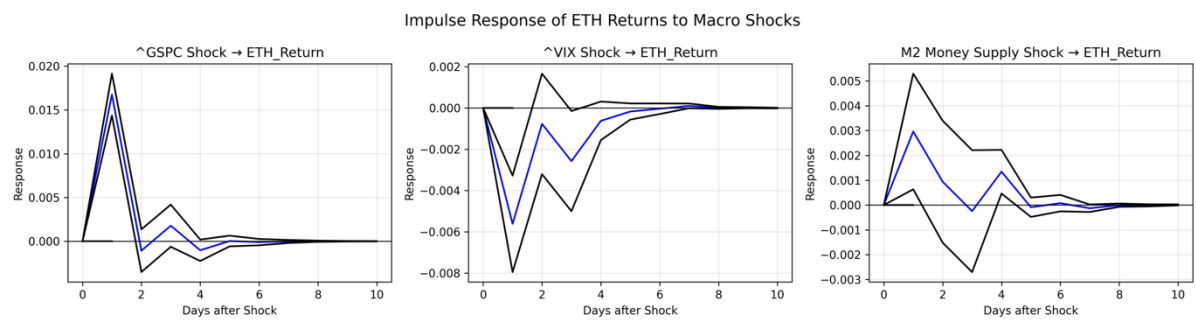


Figure 8

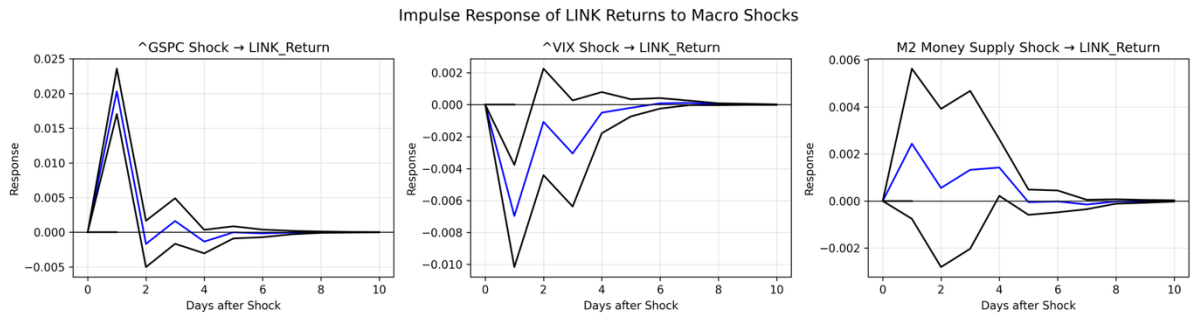


Figure 9

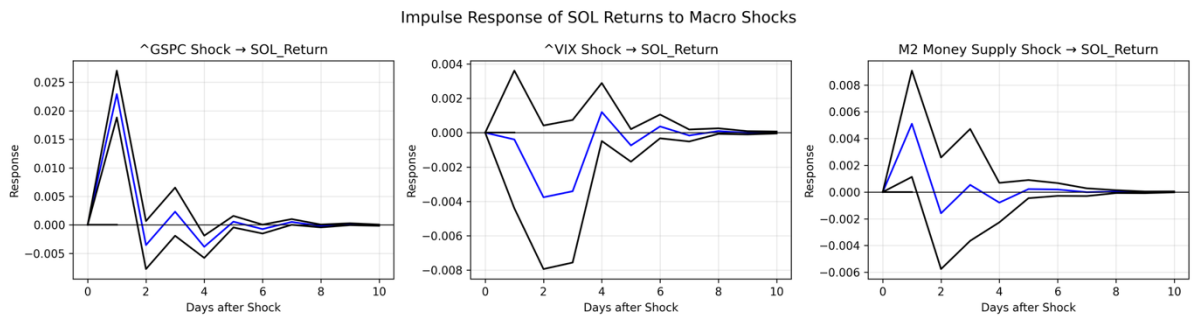


Figure 10

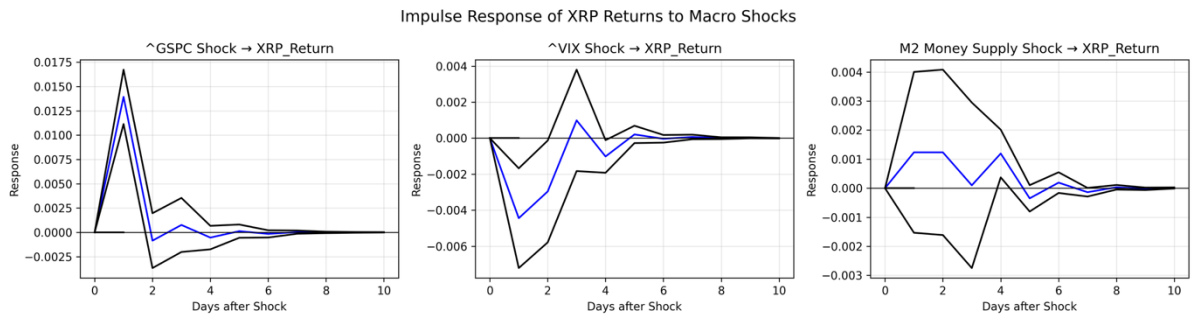


Figure 11

