Machine Learning for Finance

Problem Set 1

Bennett, Andrew
Maciel, Rui
Mirabent Rubinat, Guillem

Argimiro Arratia

Barcelona School of Economics

May 9, 2024

0 Introduction

A three question problem set on the analysis of Market Trends using volatility indicators and regression modelling. All authors contributed equally.

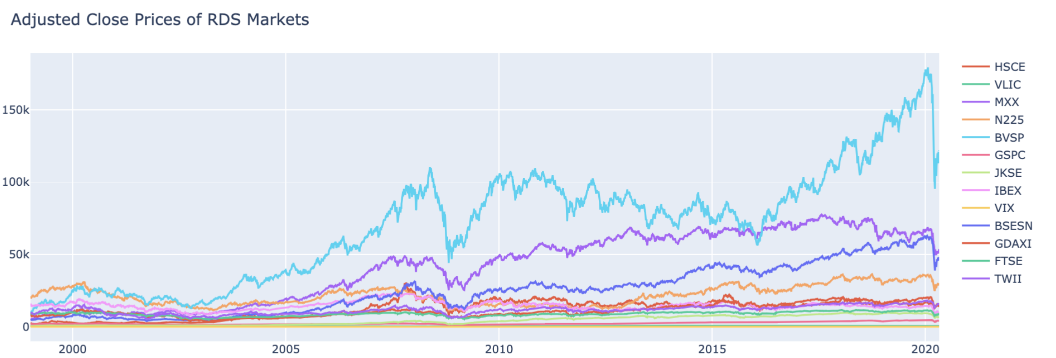
1 Question 1

Show that the EWMA-based variance σ^2 ewma can be obtained also from the following recursion:

$$\sigma^2_{\text{ewma}(t)} = \lambda\sigma^2_{\text{ewma}(t-1)} + (1-\lambda)r^2_{t-1}$$

which is easily derived from the EWMA equation. Use EMA to compute σ^2 ewma(t) with = 0.94 for some market index. Obtain the volatility time series estimation of the market index from this EMA estimation of variance and compare it to a regular (historical) volatility estimation (e.g. by cumulative sum of square returns or Parkinson, or Garman-Klass estimates). Report what you observe.

Solution. We first clean the data removing outliers, then interpolating the data to fill missing values. This method creates two dataframes, one forward filled, one backward filled then averages them together. This process creates an local average for the missing values, while still retaining the consistency of the original values. We plotted and tracked the Adjusted Closing prices across all major markets, seeing high variation among the markets. Notably BVSP is outperforming all other markets, but has a strong downtrend at the of the time period.



We then use the following code to calculate the EWMA recursively:

```
def calculate_ewma_volatility(rets, lam):
    '''
        rets: array of returns
        lam: float lambda value
    '''
    sq_rets = (rets**2).values
    EWMA_var = np.zeros(len(sq_rets))

    for r in range(1, len(sq_rets)):
```

```
        EWMA_var[r] = (1-lam)*sq_rets[r-1] + lam*EWMA_var[r-1]

    EWMA_vol = np.sqrt(EWMA_var)
    return pd.Series(EWMA_vol, index=rets.index, name ="EWMA Vol {}".format(
    lam))[1:]
```

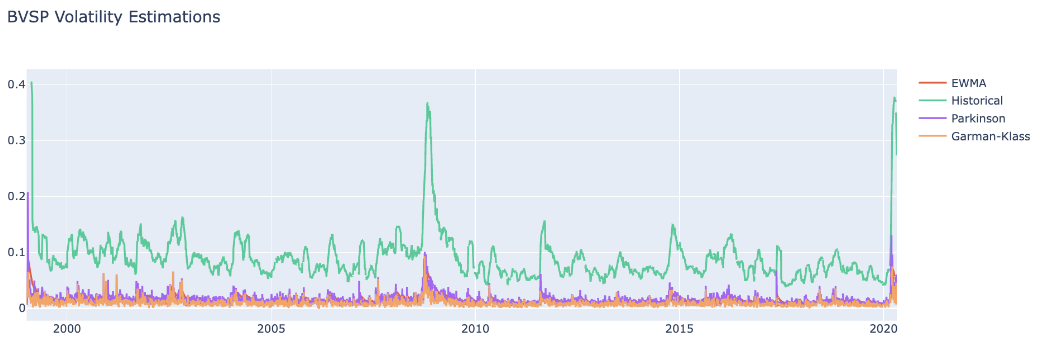
After we tested and completed the EWMA volatility function, we then created function for historical volatility, Parkinson volatility and Garman Klass volatility. The formulas used are defined below

$$HV = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(r_i - \bar{r})^2}$$

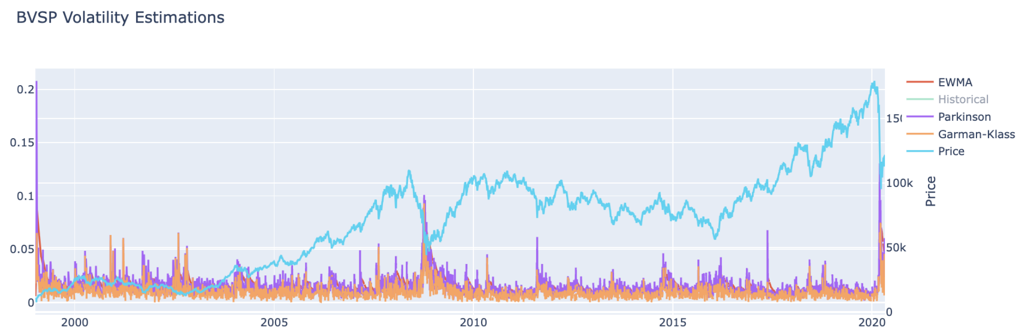
$$GK = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\log\frac{H_i}{L_i}\right)^2 - \frac{2}{N}\sum_{i=1}^{N}\log\frac{C_i}{O_i}\log\frac{H_i}{L_i} + \frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{C_i}{O_i}\right)^2}$$

$$PV = \sqrt{\frac{1}{4N}\sum_{i=1}^{N}\left(\log\frac{H_i}{L_i}\right)^2}$$

We then plotted the formulas on a stacked line chart and observed the proceeding results. For continuity we are viewing the Sao Paulo BVSP market as it is a top performer and has some significant spikes in price.



As we can see in the graph Historical volatility is much larger in magnitude than the other indicators. However, we can still observe that all 4 volatility measures are showing robust activity at similar times throughout the year. To get a better look at the volatility measures, we remove the historical volatility and add a price overlay.



We now can see the volatility estimates more clearly. EWMA is more smooth in its depiction, while the Parkinson and Garman Klass have higher spikes. Additionally, the Parkinson volatility seems to have higher spikes as the price increases, while the Garman Klass has higher spikes at lower prices. Generally, all of the significant price changes are captured by all of the volatility metrics, including the financial crisis in 2008 and the descent during the pandemic.

2 Exercise 2

In our analysis of global market dynamics, the following stock market indexes will be examined:

- Bombay Stock Exchange Sensitive Index (Sensex) (BSES);
- Bovespa Index (Brazil) (BVSP);
- Financial Times Stock Exchange 100 Index (UK) (FTSE);
- Deutscher Aktien Index (Germany) (GDAXI);
- S&P 500 Index (US) (GSPC);
- Hang Seng China Enterprises Index (Hong Kong) (HSCE);
- IBEX 35 Index (Spain) (IBEX);
- Jakarta Composite Index (Indonesia) (JKSE);
- Mexican Stock Exchange Index (Mexico) (MXX);
- Nikkei 225 Index (Japan) (N225);
- Taiwan Stock Exchange Weighted Index (TWII);
- CBOE Volatility Index (US) (VIX);
- For (VLIC), no information could be found.;

The following analysis will be done for the timeframe 2000-01 and 2002-12. The causality analysis done will consider a statistical significance of 5/

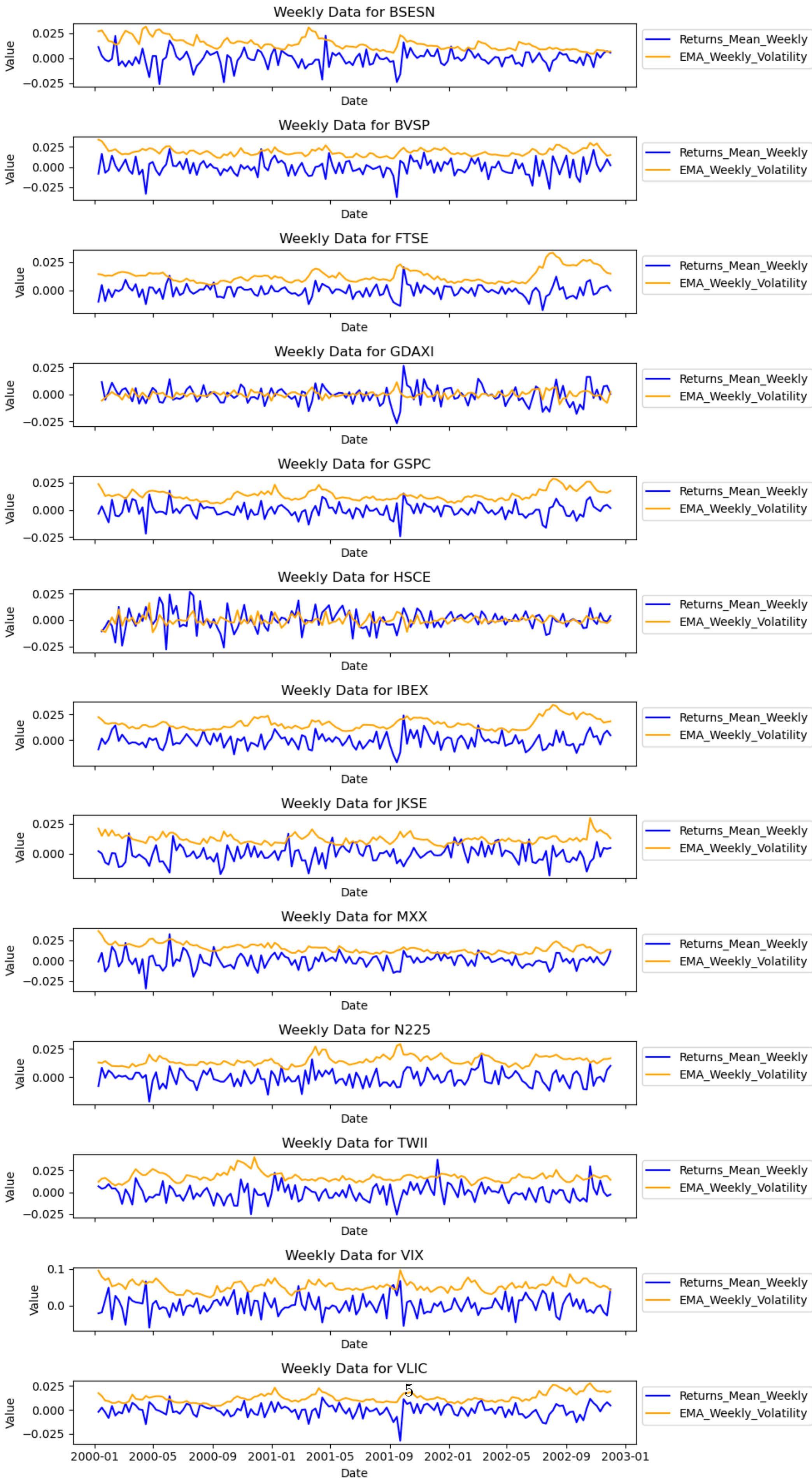
After cleaning, structuring the data and creating the EMA volatility index, we checked the stationarity of the series which we were going to do statistical tests.

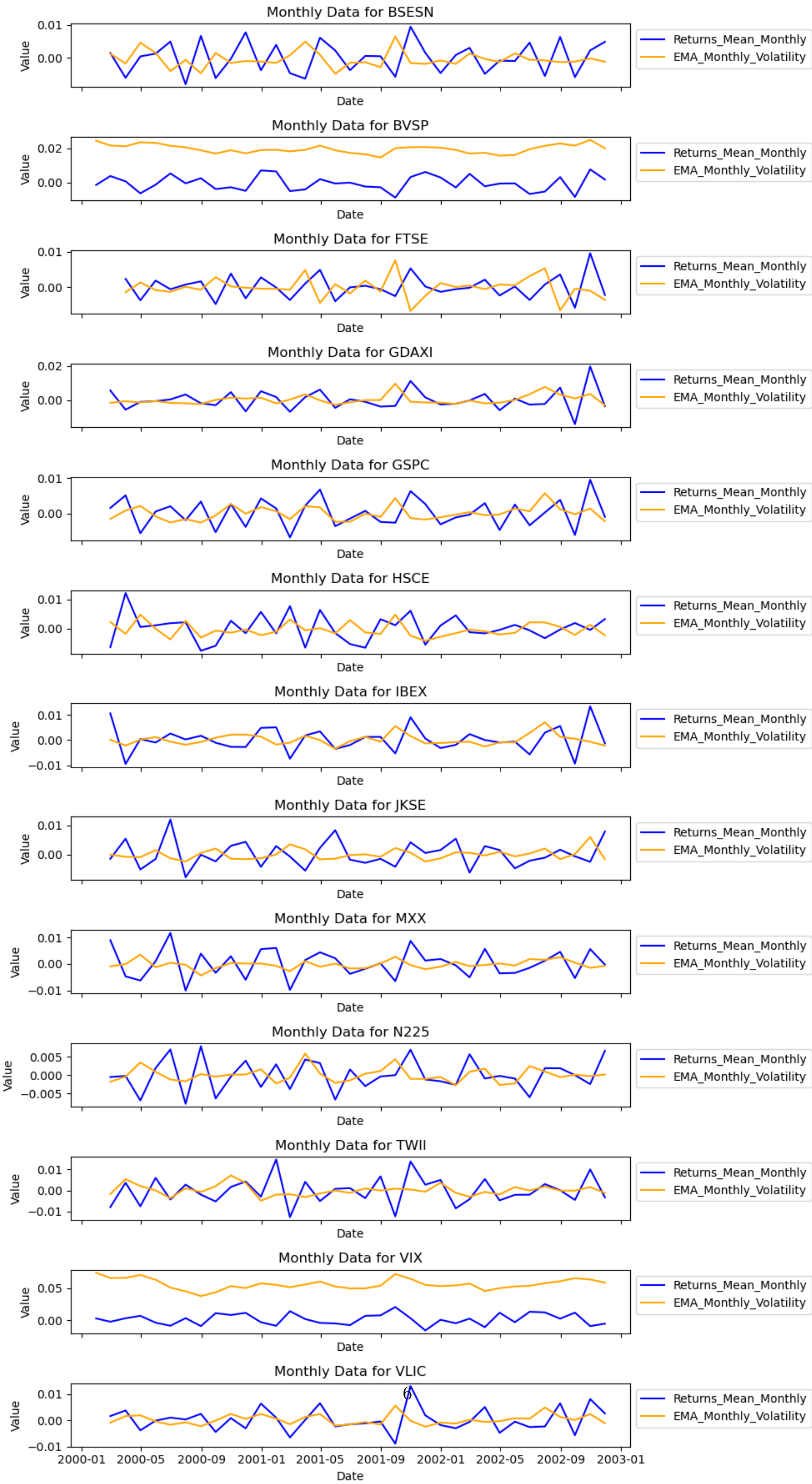
These two series of graphs presents the results of financial data analysis for various stock indexes after ensuring data stationarity. Initially, the Dickey-Fuller test was applied to test for stationarity, indicating that the original data series contained unit roots.

The analysis of the data revealed that most of the time series had achieved stationarity, particularly in the weekly dataset. However, there were a few exceptions where the Exponential Moving Average (EMA) of weekly volatility for the GDAXI and HSCE indexes did not initially exhibit stationarity. These particular series required further transformation through differencing to achieve stationarity.

In the monthly data, the challenge of achieving stationarity was more pronounced, especially with the volatility measures. To address this, differencing was applied to these series. For the FTSE EMA Monthly Volatility, a single round of differencing was insufficient, necessitating a second round of differencing to successfully stabilize the series.

Each graph plots the returns (blue line) and their corresponding Exponential Moving Average (EMA) of volatility (orange line) across several stock indexes. The stationarity of these transformed series is evident as the means and variances appear consistent over time, with no visible trends or seasonal effects, which is a key characteristic of stationary data.


Weekly Data for BSESN
Weekly Data for BVSP
Weekly Data for FTSE
Weekly Data for GDAXI
Weekly Data for GSPC
Weekly Data for HSCE
Weekly Data for IBEX
Weekly Data for JKSE
Weekly Data for MXX
Weekly Data for N225
Weekly Data for TWII
Weekly Data for VIX
Weekly Data for VLIC
Returns_Mean_Weekly
EMA_Weekly_Volatility
Value
Date
0.025
0.000
−0.025
0.1
0.0
2000-01
2000-05
2000-09
2001-01
2001-05
2001-09
2002-01
2002-05
2002-09
2003-01


Monthly Data for BSESN
Monthly Data for BVSP
Monthly Data for FTSE
Monthly Data for GDAXI
Monthly Data for GSPC
Monthly Data for HSCE
Monthly Data for IBEX
Monthly Data for JKSE
Monthly Data for MXX
Monthly Data for N225
Monthly Data for TWII
Monthly Data for VIX
Monthly Data for VLIC
Returns_Mean_Monthly
EMA_Monthly_Volatility
Value
Date
2000-01 2000-05 2000-09 2001-01 2001-05 2001-09 2002-01 2002-05 2002-09 2003-01

As we analyze over 624 data points across various indices (4 matrixes, with 13*13 cells, with each cell with 4 data points, minus the diagonal), our prioritization will focus on the following criteria:

a) **Strong Causality**: This is indicated by the presence of Granger causality across four consecutive lags within the indices. Such a pattern signifies a robust and consistent predictive relationship, where historical values of one index are repeatedly predictive of another's future values.

b) **Bidirectionality**: The existence of bidirectional causality suggests potential endogeneity within the relationship between the indices. This observation warrants a deeper investigation to thoroughly understand the underlying causality dynamics.

2.1 Granger Causality Analysis for Weekly returns

Granger Causality Results for Returns Mean Weekly:

	BSESN	BVSP	FTSE	GDAXI	GSPC	HSCE	IBEX	JKSE	MXX	N225	TWII	VIX	VLIC
BSESN	-	(1, 1, 1, 1)	(1, 1, 1, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 0)
BVSP	(0, 0, 0, 0)	-	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)
FTSE	(0, 1, 0, 0)	(0, 0, 0, 0)	-	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
GDAXI	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
GSPC	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	-	(1, 1, 1, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)
HSCE	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 0, 0)	(1, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 1)	(0, 0, 0, 0)	(1, 1, 1, 0)	(0, 0, 0, 0)
IBEX	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
JKSE	(0, 0, 0, 0)	(0, 0, 1, 1)	(0, 1, 1, 1)	(0, 0, 0, 1)	(1, 1, 1, 1)	(0, 1, 1, 1)	(0, 0, 0, 0)	-	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)
MXX	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 1, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
N225	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 0, 0, 0)	-	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)
TWII	(0, 0, 0, 0)	(1, 1, 1, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	-	(1, 1, 1, 1)	(1, 1, 1, 1)
VIX	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 1, 0, 0)	(1, 1, 1, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)
VLIC	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-

These are the strong relations identified in the Weekly Returns matrix:

- BSESN to BVSP, GDAXI, GSPC, MXX, VIX.
- BVSP to FTSE, GDAXI, GSPC, IBEX, VIX, VLIC.
- FTSE to GDAXI, HSCE.
- GDAXI to None.
- GSPC to FTSE, GDAXI.
- HSCE to None.
- IBEX to GDAXI
- JKSE to GSPC, VIX, VLIC.
- MXX to GDAXI
- N225 to BSVP, FTSE, GDAXI, GSPC, GSPC, HSCE, IBEX, MXX, VIX, VLIC.
- TWII to FTSE, GDAXI, GSPC, HSCE, IBEX, MXX, VIX, VLIC.

- VIX to GDAXI.
- VLIC to FTSE, GDAXI, HSCE.

The most influential indices within the dataset are N225 and TWII, as they exhibit extensive influence across a broad range of other indices. Specifically, N225 influences BVSP, FTSE, GDAXI, GSPC, HSCE, IBEX, MXX, VIX, and VLIC, showcasing its significant role across diverse global markets. Similarly, TWII impacts FTSE, GDAXI, GSPC, HSCE, IBEX, MXX, VIX, and VLIC, indicating its effect on multiple major indices. These indices stand out for their capacity to affect numerous others.

In terms of bidirectional relationships, there are mutual influences between major indices which may indicate intricate economic interdependencies. The relationships between BSESN and GSPC, as well as between FTSE and GDAXI, are particularly significant.

Lastly, the indices GDAXI and HSCE show no strong influence on other indices within the matrix. This absence of influence suggests that these markets might operate under different dynamics or are less connected to the global trends represented by the other indices.

2.2 Granger Causality Analysis for EMA Weekly Volatility

Granger Causality Results for EMA Weekly Volatility:

	BSESN	BVSP	FTSE	GDAXI	GSPC	HSCE	IBEX	JKSE	MXX	N225	TWII	VIX	VLIC
BSESN	-	(0, 0, 0, 0)	(0, 1, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 1)
BVSP	(0, 0, 0, 0)	-	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)
FTSE	(0, 0, 0, 0)	(1, 1, 1, 1)	-	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 1, 0)
GDAXI	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 1, 1, 1)	-	(0, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 0)
GSPC	(0, 0, 0, 1)	(1, 1, 0, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	-	(1, 1, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
HSCE	(1, 0, 0, 0)	(1, 1, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 1)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
IBEX	(0, 0, 0, 0)	(1, 1, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
JKSE	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 0, 0)	(1, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
MXX	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 1)	(1, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
N225	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 1)	(1, 1, 1, 1)	(0, 1, 1, 0)	(1, 1, 1, 1)	(0, 1, 1, 1)	(0, 0, 0, 0)	(0, 1, 1, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
TWII	(0, 0, 1, 1)	(0, 0, 0, 0)	(0, 1, 1, 0)	(0, 1, 0, 0)	(0, 1, 1, 1)	(0, 1, 1, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 1, 1, 1)
VIX	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)
VLIC	(0, 0, 0, 1)	(0, 0, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	-

These are the strong relations identified in the EMA Weekly Volatility matrix:

- BSESN to None
- BVSP to FTSE, GDAXI, GSPC, HSCE, IBEX, N225, VIX, VLIC.
- FTSE to BVSP, GDAXI.
- GDAXI to BSVP.
- GSPC to FTSE, GDAXI, IBEX.
- HSCE to None.
- IBEX to FTSE, GDAXI, N225.
- JKSE to None.

- MXX to BSENS, HSCE.
- N225 to GDAXI, HSCE.
- TWII to None.
- VIX to FTSE, GDAXI.
- VLIC to FTSE, GDAXI, IBEX, N225.

The most influential index in this matrix is BVSP, which exhibits a significant degree of influence over a range of other indices, including FTSE, GDAXI, GSPC, HSCE, IBEX, N225, VIX, and VLIC. This broad spectrum of influence shows statistical influence of the BVSP's role in impacting weekly volatility across various global markets.

In the context of bidirectional relationships within the EMA Weekly Volatility matrix, only BVSP and GDAXI share influences, hinting at closely linked economic or market dynamics.

Lastly, several indices such as BSESN, HSCE, JKSE, and TWII show no strong influence on others within this matrix. The absence of influence from these indices might indicate unique local market conditions or factors that isolate them from the broader trends observed in other global indices.

2.3 Granger Causality Analysis for Returns Mean Monthly

Granger Causality Results for Returns Mean Monthly:

	BSESN	BVSP	FTSE	GDAXI	GSPC	HSCE	IBEX	JKSE	MXX	N225	TWII	VIX	VLIC
BSESN	-	(1, 1, 1, 1)	(0, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 1, 1)	(0, 0, 0, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 1, 0, 0)
BVSP	(1, 0, 1, 1)	-	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 1, 1)	(1, 1, 1, 1)	(0, 0, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 1)	(1, 1, 1, 1)
FTSE	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(1, 0, 0, 0)	(1, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 0, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 0, 1)	(1, 0, 0, 0)
GDAXI	(0, 0, 0, 0)	(1, 0, 0, 0)	(1, 1, 1, 1)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)
GSPC	(0, 0, 0, 0)	(1, 1, 1, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	-	(0, 0, 0, 1)	(0, 0, 0, 1)	(1, 0, 1, 0)	(0, 0, 0, 0)	(0, 1, 1, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)
HSCE	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
IBEX	(0, 1, 0, 0)	(1, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)
JKSE	(0, 0, 0, 0)	(1, 0, 0, 0)	(0, 0, 0, 1)	(1, 1, 1, 0)	(1, 0, 1, 1)	(0, 0, 0, 0)	(1, 0, 0, 0)	-	(1, 1, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 1, 0)
MXX	(0, 1, 1, 0)	(1, 1, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 1, 1)	(0, 0, 0, 0)	(1, 0, 0, 1)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)
N225	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 1, 1, 1)	(0, 1, 1, 0)	(0, 1, 1, 1)	(0, 0, 0, 1)	(0, 0, 0, 1)	(1, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(1, 1, 0, 1)	(0, 1, 1, 0)
TWII	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(1, 1, 1, 1)	(0, 0, 0, 0)
VIX	(0, 0, 1, 1)	(0, 0, 0, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 1, 1)	(1, 1, 1, 1)	(0, 0, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 1)	(1, 1, 1, 1)	-	(1, 1, 1, 1)
VLIC	(0, 0, 0, 0)	(1, 1, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 1)	(0, 0, 0, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	-

These are the strong relations identified in the Returns Mean Monthly:

- BSESN to BSVP, GDAXI, GSPC, VIX.
- BVSP to FTSE, GDAXI, GSPC, IBEX, MXX, TWII, VIX.
- FTSE to None.
- GDAXI to VIX.
- GSPC to FTSE, VIX.
- HSCE to None.

- IBEX to FTSE, VIX.
- JKSE to None
- MXX to FTSE, VIX.
- N225 to BSVP.
- TWII to BSVP, FTSE, VIX.
- VIX to FTSE, GDAXI, GSPC, IBEX, MXX, TWII, VLIC.
- VLIC to VIX.

VIX and BVSP emerge as the most influential indexes. VIX impacts FTSE, GDAXI, GSPC, IBEX, MXX, TWII, and VLIC. BVSP also demonstrates a strong influence, affecting FTSE, GDAXI, GSPC, IBEX, MXX, and TWII, underscoring its importance in both regional and global contexts. Additionally, BSESN has notable influence, affecting BVSP, GDAXI, GSPC, and VIX, reinforcing its significance in the broader market landscape.

Regarding bidirectional relationships, the only mutual influence observed is between VIX and VLIC. This bidirectional relationship suggests a reciprocal influence that could be indicative of underlying economic or financial mechanisms that interlink these two volatility indices.

Several indices show no outgoing influence on others. FTSE, HSCE, and JKSE are among those that do not exert influence on any other indices in the matrix.

2.4 Granger Causality Analysis for EMA Monthly Volatility

Granger Causality Results for EMA Monthly Volatility:

	BSESN	BVSP	FTSE	GDAXI	GSPC	HSCE	IBEX	JKSE	MXX	N225	TWII	VIX	VLIC
BSESN	-	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)
BVSP	(1, 1, 1, 1)	-	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 1)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 0)	(1, 1, 1, 1)
FTSE	(0, 0, 0, 0)	(0, 0, 1, 0)	-	(1, 1, 1, 1)	(0, 0, 1, 1)	(1, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 1, 1)	(1, 0, 1, 1)
GDAXI	(0, 0, 0, 0)	(1, 1, 0, 0)	(1, 1, 1, 1)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)
GSPC	(0, 0, 0, 0)	(1, 1, 1, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 1)	(0, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 1, 1, 0)
HSCE	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
IBEX	(0, 0, 0, 0)	(1, 0, 0, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 0, 1)	(0, 0, 1, 1)	(1, 0, 0, 0)	(0, 0, 0, 0)
JKSE	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
MXX	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 1, 1)	(0, 0, 1, 0)	(0, 1, 0, 1)	(0, 0, 0, 0)	(0, 0, 1, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
N225	(0, 0, 0, 1)	(0, 0, 0, 0)	(1, 0, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)
TWII	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 1, 1)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	-	(0, 1, 1, 1)	(0, 0, 0, 0)
VIX	(1, 1, 1, 1)	(1, 0, 0, 0)	(1, 1, 1, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(1, 0, 1, 0)	(1, 1, 1, 1)	(0, 0, 0, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 0, 0, 0)	-	(1, 1, 1, 1)
VLIC	(0, 0, 0, 1)	(0, 0, 1, 0)	(1, 1, 1, 1)	(1, 1, 1, 1)	(0, 1, 1, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 1, 1, 1)	(0, 1, 1, 1)	(0, 0, 0, 0)	(0, 0, 1, 1)	-

These are the strong relations identified in the EMA Monthly Volatility:

- BSESN to FTSE, GDAXI, GSPC, VIX.
- BVSP to FTSE, DGAXI, GSPC, IBEX, MXX, N225, TWII, VLIC.
- FTSE to GDAXI, GSPC, IBEX, N225.
- GDAXI to None.

- GSPC FTSE, VIX.
- HSCE to FTSE.
- IBEX to FTSE.
- JKSE to None.
- MXX to None.
- N225 to None.
- TWII to None.
- VIX to BSESN, GDAXI,GSPC, IBEX, MXX, N225, VLIC.
- VLIC to FTSE, GSPC.

VIX and BSVP are the most influential index in this matrix. VIX exerts a significant impact on BSESN, GDAXI, GSPC, IBEX, MXX, N225, and VLIC. This reach underscores VIX's role as a critical indicator of market volatility, reflecting its central importance in forecasting market volatility across multiple regions. BVSP also shows a robust influence pattern, affecting FTSE, GDAXI, GSPC, IBEX, MXX, N225, TWII, and VLIC.

In terms of bidirectional relationships, several indices exhibit reciprocal influences. Notably, FTSE and GSPC influence each other, indicating a significant economic interdependence. Additionally, the mutual influence between VIX and several indices, particularly GSPC and VLIC, further emphasizes the interconnected nature of global market volatility.

Several indices show no outgoing influence, indicating isolated behavior within this volatility framework. GDAXI, JKSE, MXX, N225, and TWII are among those that do not exert influence on any other indices in this matrix.

2.5 Comparison between the matrixes

Relation between Monthly and Weekly Returns

To analyze the relationships between Monthly and Weekly Returns matrices, we observe that BVSP consistently impacts other major indices like FTSE, GDAXI, GSPC, IBEX, and VIX on both a monthly and weekly basis. This indicates a significant influence of Brazilian stocks on European and American markets, showcasing its pivotal role in global market dynamics.

Other noteworthy observations include BSESN displaying broader connections in the weekly matrix, suggesting a more immediate reaction to shifts in global markets. Conversely, indices such as FTSE and HSCE show minimal or no relationships consistently, indicating a degree of independence from the global trends captured by other indices. The extensive weekly connectivity of N225 or TWII across a range of indices, compared to its monthly interactions, highlight the indexes' high sensitivity to short-term market changes.

Relation between Monthly and Weekly EMA Volatility

To delve into the relationship between Monthly and Weekly EMA Volatility, we first identify the consistent relationships across both matrices and the indexes that display no relationships in either time frame. Notably, BVSP maintains robust connections with FTSE, GDAXI, GSPC, IBEX, N225, VIX, and VLIC in both monthly and weekly analyses, suggesting strong inter-market volatility correlations. However, both JKSE and TWII show no relationships in either matrix, indicating that their volatility might be independent of the other major global indexes listed.

Additionally, there are significant differences worth highlighting: VIX, which is linked with several indexes on a monthly basis, shows few relationships weekly, suggesting its volatility correlations may be more pronounced over longer periods.

Relation between Weekly variables

Analyzing the relationships between Weekly Returns and Weekly EMA Volatility matrices helps to highlight how market dynamics and volatility are interconnected on a weekly basis. In the Weekly Returns matrix, we observe that BVSP has a significant influence on various indices like FTSE, GDAXI, GSPC, IBEX, VIX, and VLIC, indicating its pivotal role in the global markets.

The EMA Weekly Volatility matrix confirms and extends some of these insights. BVSP's influence remains prominent, impacting the same indices as in the returns matrix, which suggests a correlation between return fluctuations and volatility. However, HSCE shows no relationships in the both matrix, indicating that its returns do not consistently correlate with changes in market volatility or returns on a weekly basis.

The varied connectivity of indices like N225 and VLIC in both matrices underscores their roles as intermediaries in transmitting market shocks and volatility across different markets.

Relation between Monthly variables

Exploring the relationships in the Returns Mean Monthly and EMA Monthly Volatility matrices provides a nuanced view of how return averages and volatility patterns interrelate over a monthly period. In the Returns Mean Monthly matrix, BVSP shows extensive relationships with multiple indices such as FTSE, GDAXI, GSPC, IBEX, MXX, TWII, and VIX, indicating a strong influence on these markets. VIX, as a measure of market volatility, also shows significant connections with many indices, underscoring its central role in capturing market sentiment and fluctuations across various markets.

In the EMA Monthly Volatility matrix, similar patterns emerge, with BVSP again influencing a broad spectrum of indices, and VIX maintaining its connections, further establishing the link between volatility and returns across markets. Interestingly, some indices like HSCE, JKSE, and TWII that show no relationships in the returns matrix have connections in the volatility matrix, suggesting that while their returns may not directly correlate with other markets, their volatility patterns do exhibit some dependencies. This analysis highlights the complexity of financial markets, where not just returns but also the volatility can influence and be influenced by multiple factors across different markets, vital for comprehensive risk management and investment strategy development.

2.6 Sum up

The analysis of relationships across Monthly and Weekly Returns, as well as EMA Volatility, reveals significant influences exerted by certain indexes in the markets studied. The

BVSP (Brazil's Bovespa Index) emerges as a pivotal index, consistently impacting major global indices such as the **FTSE, GDAXI, GSPC, IBEX, VIX**, and **VLIC** across both time frames, returns and volatility measures. This indicates that Brazilian stocks are significant influencers of European and American markets, as well as broader global market dynamics.

The **VIX (Volatility Index)**, known for measuring market risk and investor sentiment, shows extensive connections across various indices in the monthly analysis (both in returns and volatility) demonstrating that it has a volatility spillover over longer spans of time. It links with **FTSE, GDAXI, GSPC, IBEX, MXX**, and **TWII**, underscoring its central role in reflecting and potentially driving market volatility and sentiment.

These findings illustrate the interconnected nature of global financial markets, where indices such as **BVSP, VIX** not only reflect their regional economies but also exert significant influence on international markets. Additionally, understanding these connections helps in predicting potential market movements based on changes observed in these key indices.

2.7 Robustness Check - Bootstrapping

Bootstrapping is a technique to resample the data. In this situation we used Stationarity Bootstrapping. Stationarity bootstrapping is unique because it takes periods of data that retain the sequential features of time series data. This is especially important if there is autocorrelation or other time series specific phenomena. The process for stationarity bootstrapping is as follows:

$$\text{For } i = 1, 2, ..., B:$$

Draw random block indices $\{b_1^{(i)}, b_2^{(i)}, ..., b_{n/b}^{(i)}\}$ with replacement, where $b_j^{(i)} \in \{1, 2, ..., n/b\}$.

$$\text{Define the bootstrap sample } X^{*(i)} = (X[b_1^{(i)}], X[b_2^{(i)}], ..., X[b_{n/b}^{(i)}]).$$

Repeat the process for B iterations to obtain B bootstrap samples $\{X^{*(1)}, X^{*(2)}, ..., X^{*(B)}\}$.

The purpose of this bootstrapping technique is to ensure accurate results of the Granger causality tests with different subsections of data. We first attempted a 12 period sample size and obtained the same results as the standard tests. Then we changed the sample size two 5 and saw some variability. Below are the comparisons of the results. On the x axis are the markets, then on the y-axis there is each time lag for each market. For the weekly returns, the samples were averaged together, and for the monthly returns the samples were rounded to the nearest whole number. This procedure was done to visualize the causality more effectively.

Below we can see there is significantly less causality in the bootstrapped sample across

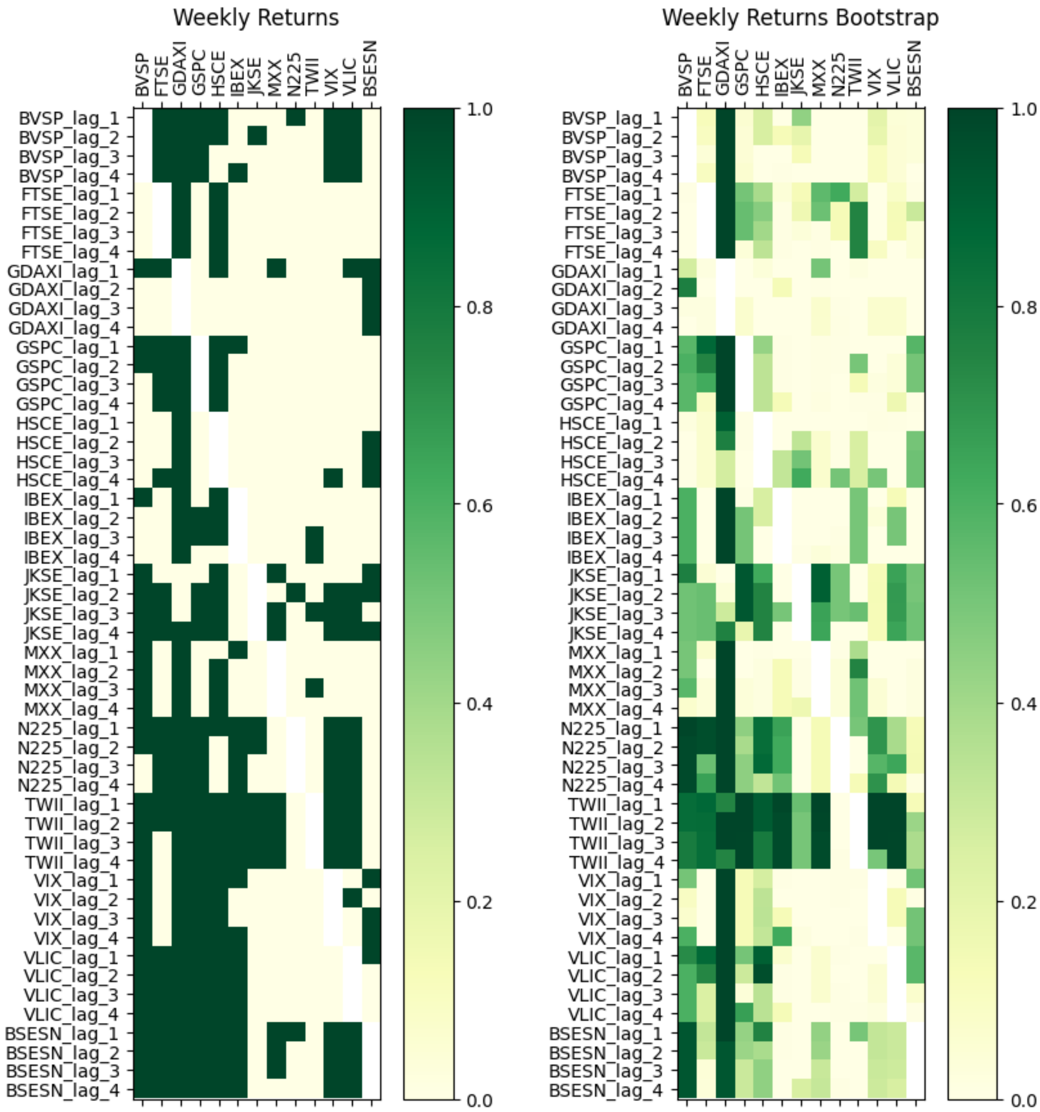


Figure 1: Weekly Bootstrap and Standard Tests Displayed Side by Side

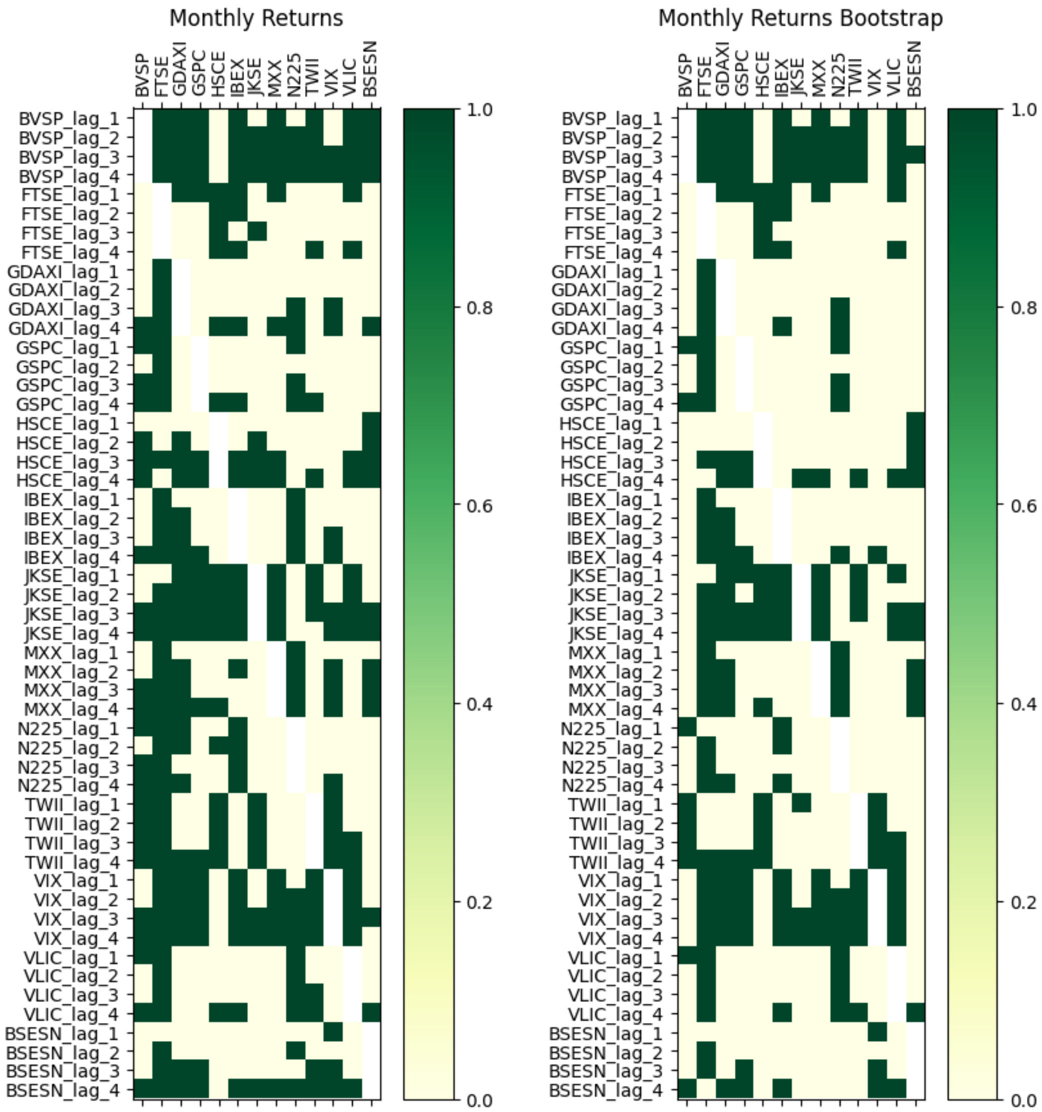


Figure 2: Monthly Bootstrap and Standard Tests Displayed Side by Side

the board. However, it is of note that the TWII index retains much of its causality and in some cases increases its causality compared to the standard approach. This could be due to strong trends in the semi-conductor markets influencing global markets. Additionally, GDAXI also looks to have strong correlation across all markets. These German blue chip companies also retain strong influence over the market at smaller time scales.

At the monthly level, with a 5 period sample size, causation is more strongly retained than in the weekly samples, but still decreases. Most notably is the decrease in causation from BVSP. This is interesting and follows previous volatility analysis that shows BVSP as one of the most volatile markets. With this finding, we can extrapolate that higher volatility has lower causation at larger time frames.

3 Exercise 3

We were assigned the following indicators:

- **Dividend-Price Ratio:** Computed by dividing the Dividends paid between the Price of the index:
$$\frac{Dividend}{Price}$$
or, alternatively:
$$log(Dividend) - log(Price)$$

- **Dividend-Earnings Ratio:** Computed by dividing the Dividends paid between the Earnings of the companies:
$$\frac{Dividend}{Earnings}$$
or, alternatively:
$$log(Dividend) - log(Earnings)$$

- **Earnings-Price Ratio:** Computed by dividing the Earnings of the companies by the Price of the Index:
$$\frac{Earnings}{Price}$$
or, alternatively:
$$log(Earnings) - log(Price)$$

Initially, there was the fatal temptation to use the 0 lag version of those variables, but using them would render us scores near 100% in R^2 and Accuracy (for Log Returns prediction and Direction prediction respectively). This is too good to be true, of course, as the model would then only pick up the fact that DP and EP ratios are computed using the Target variable. As a result, this would lead to data leakage and not to a real demonstration of our variables' predictive capacity. To avoid that, we compute the ratios but then only use them in a lagged fashion, never with 0 lag.

3.1 Pre-processing and Feature Engineering

Continuing on our practice, we load the data into a Python notebook and we check that there are no missing values for our assigned variables. There are none in the whole period from 1927 so we advance to the next step.

Next, we compute the returns using the logarithmic difference :

$$log(Return_t) - log(Returnt - 1)$$

and we also create the columns for the ratios using the logarithmic formula we saw before. We then create the binary variable for the direction of the returns (positive or negative).

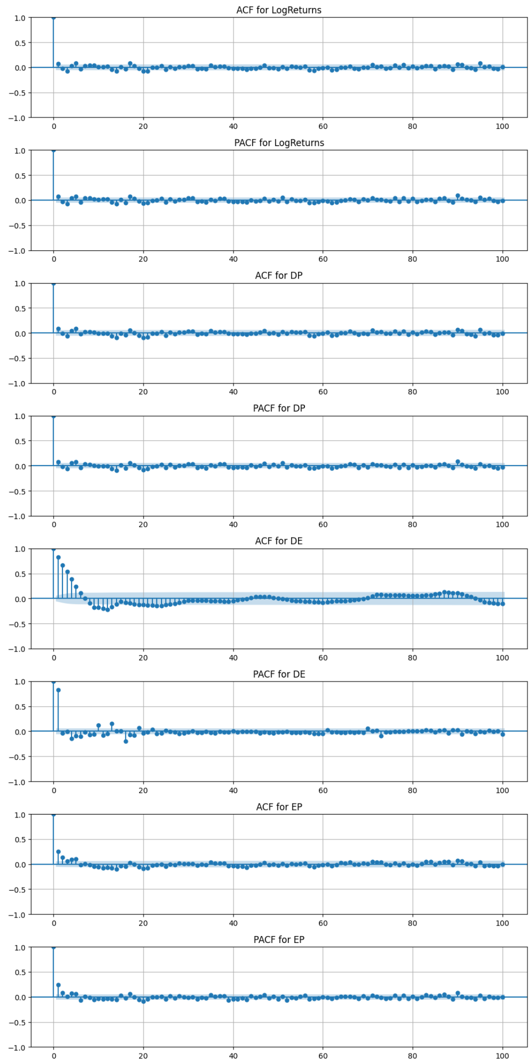
At this point, our Log Returns, DE and EP ratios are already stationary with a high probability, but we have to further difference the DP ratio in order to turn it into a stationary variable. Afterwards, with further analysis, we realize that DE and EP ratios still contain extremely skewed data so we also difference them again in order to reduce this skewness. After this second differentiation, all the variables are mostly centered and completely stationary. We also entertain the choice to further standardize the variables but standardization doesn't really add anything to the data we already have so we decide to not use it and instead use the data as we already have it.

We finally decide to cap some extreme outliers to a limit value to reduce their effect on the overall trends of the data that the models could pick up.

3.2 Lag Creation and Selection

3.2.1 ACF and PACF

We start by computing and plotting the ACF and PACF of the Log Returns and the DP, DE and EP ratios. We can only observe some significant lag auto-correlation for the DE and EP ratios. In DE, we observe relevance up until around lag 10 but for EP the relevance already stops at lag 2.



3.2.2 Granger Causality

Due to the limitation of working with monthly data, we need to take bigger windows of time to perform a rolling window analysis. We take windows of several years, 5 windows in total, which jointly cover from 1927 to 2021.

For each interval and relevant variables (Log Returns, DP, DE and EP), we perform the Granger causality test to up to 5 lags before time t.

We can observe how it appears that the EP ratio tends to be the most causal for LogReturns, especially in some periods more than others. Then we can also see that some lags of the DE ratio also pass the test (although only at a 10% confidence interval). As a result we are going to keep evaluating the lags to use, but we can probably safely try to use more than just 5 lags for DE and EP ratios.

That being said, the results are not too surprising because the Dividends are just one component of the price, but a lot of people value more highly a stock which keeps its value and grows in price more than another stock which returns dividends but doesn't grow as much, because you can always resell the stock if needed.

As a result, a company which is having consistent earnings is more probably higher valued than another with lower earnings and, as such, its price will probably be higher as well, hence this result. If a company does well one month, probably a lot of people will be at least a little bit more interested in buying their stock. The same can be said about an index like the S&P500. It's better to know that they are getting consistent and significant earnings more than knowing whether they paid dividends or not.

```
For lagged EP
 ####################

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=4.3026  , p=0.0418  , df_denom=68, df_num=1
ssr based chi2 test:   chi2=4.4925  , p=0.0340  , df=1
likelihood ratio test: chi2=4.3561  , p=0.0369  , df=1
parameter F test:         F=4.3026  , p=0.0418  , df_denom=68, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=6.0303  , p=0.0040  , df_denom=65, df_num=2
ssr based chi2 test:   chi2=12.9883 , p=0.0015  , df=2
likelihood ratio test: chi2=11.9143 , p=0.0026  , df=2
parameter F test:         F=6.0303  , p=0.0040  , df_denom=65, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=4.3809  , p=0.0073  , df_denom=62, df_num=3
ssr based chi2 test:   chi2=14.6267 , p=0.0022  , df=3
likelihood ratio test: chi2=13.2657 , p=0.0041  , df=3
parameter F test:         F=4.3809  , p=0.0073  , df_denom=62, df_num=3
```

3.2.3 Distance Correlation among Factors

We can observe how there are some relevant correlations among ratios, which is only normal given that those ratios include information about the same variables. That being said, we have to take these into account.

```
 1933-01-01  ,  1970-12-01
 ####################
Distance correlation between DP and DE: 0.13177652507641135
Distance correlation between DP and EP: 0.910709754252578
Distance correlation between DE and EP: 0.2841051912645152


 1971-01-01  ,  1997-12-01
 ####################
Distance correlation between DP and DE: 0.12137788620111253
Distance correlation between DP and EP: 0.9102521650651332
Distance correlation between DE and EP: 0.3976309178201416


 1998-01-01  ,  2005-12-01
 ####################
Distance correlation between DP and DE: 0.191164072368886
Distance correlation between DP and EP: 0.7870057106538956
Distance correlation between DE and EP: 0.5094622595287323


 2006-01-01  ,  2021-11-01
 ####################
Distance correlation between DP and DE: 0.357985684556322
Distance correlation between DP and EP: 0.5309099671380381
Distance correlation between DE and EP: 0.6622304159131994
```

And this is just for Distance Correlations, for normal Correlation we get some of them up to 0.98 for some periods.

	DP	DE	EP
DP	1.00	0.19	0.98
DE	0.19	1.00	-0.00
EP	0.98	-0.00	1.00

	DP	DE	EP
DP	1.00	0.06	0.93
DE	0.06	1.00	-0.32
EP	0.93	-0.32	1.00

	DP	DE	EP
DP	1.00	-0.10	0.95
DE	-0.10	1.00	-0.41
EP	0.95	-0.41	1.00

	DP	DE	EP
DP	1.00	0.08	0.82
DE	0.08	1.00	-0.50
EP	0.82	-0.50	1.00

	DP	DE	EP
DP	1.00	0.34	0.27
DE	0.34	1.00	-0.81
EP	0.27	-0.81	1.00

As a result of our analysis, we decide to create lags of up to 10 periods before t for Log Returns, DP, DE and EP columns. Most of those lags are not going to be used, but the Neural Network is not going to suffer when selecting how many lags to use.

3.2.4 LASSO Analysis

Once we have the lagged columns in place, we perform a LASSO analysis:

```
 1971-01-01  ,  1997-12-01
 ####################
Lasso coefficients: [-0. -0.  0. -0. -0.  0.  0. -0.  0. -0.  0. -0. -0.  0.  0.  0.  0. -0.
 -0. -0. -0.  0. -0.  0. -0.  0.  0.  0. -0.  0.  0. -0. -0.  0.  0.  0.
  0. -0.  0. -0.]
Lasso intercept: 0.007266424654699718


 1998-01-01  ,  2005-12-01
 ####################
Lasso coefficients: [ 0.  0. -0.  0. -0.  0. -0.  0.  0. -0. -0. -0. -0.  0. -0.  0.  0. -0.
  0. -0.  0. -0. -0. -0.  0. -0. -0.  0. -0.  0. -0.  0.  0. -0. -0.  0.
  0. -0. -0.  0.]
Lasso intercept: 0.002622818973870418б


 2006-01-01  ,  2021-11-01
 ####################
Lasso coefficients: [ 0. -0. -0.  0. -0.  0. -0.  0.  0. -0. -0.  0.  0. -0. -0. -0. -0.  0.
  0. -0. -0.  0.  0. -0.  0. -0.  0. -0.  0. -0.  0. -0. -0.  0.  0.  0.
 -0.  0. -0.  0.]
Lasso intercept: 0.0067910048138999915
```

As we can observe, the LASSO model uses none of the lags and this is consistent across all the intervals, which already makes us think that the lags might not be too useful to predict future returns. We tried with different alphas, from higher penalization to lower penalization values. The results shown come from the model with an alpha of 0.01 but the rest of options we tried would also yield similar results.

3.2.5 Teras Virta Test

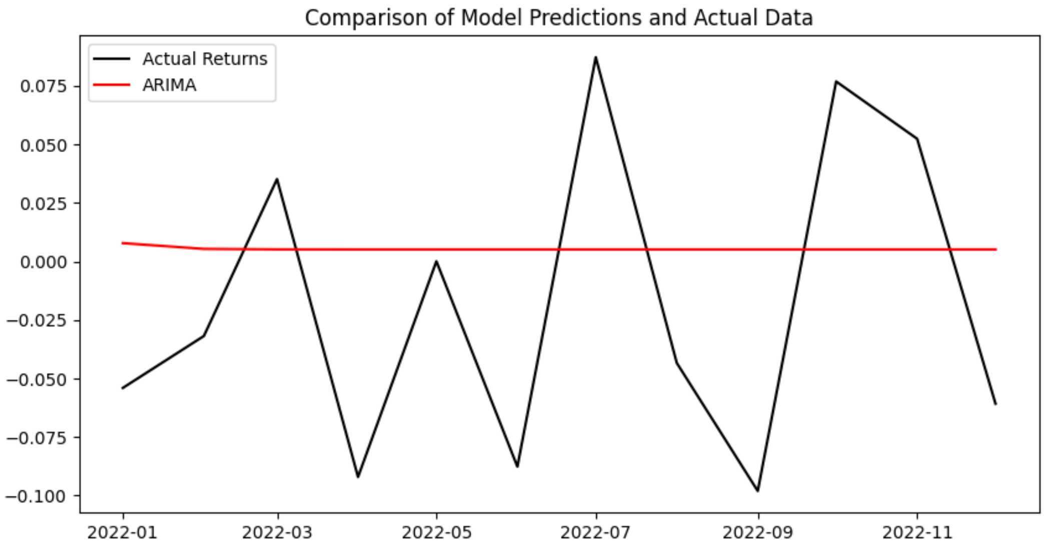
We finally performed a Teras Virta Test in order to detect non-linearities and, in short, none were detected. We performed the test on all the ratios with respect to the Log Returns and none of the tests allowed us to reject the null of linearity. As a result, we are quite certain that the ratios are not affecting the Log Returns in a non-linear way.

3.3 ARMA Baseline Model

We start our analysis by setting a baseline model. We are going to fit a very simple ARMA just using the lagged information about Log Returns to predict the mean of future returns.

We decide to use an ARMA 1,1 because the auto_arima function to use a Moving Average of 1 lag. Instead, we will fit an ARMA, which can still just leave the AR element equal to 0 if it really is the most optimal approach.

We fit the ARMA 1,1 model to our LogReturns data and then we predict on data from 2022. We decided to use the data from 1927 to 2021 as train and validation data, and leave the most recent data, from 2022 (from January to December) as the test data. We predict on the test data with our ARMA model and the results are pretty bad, which is expectable for a simple ARMA 1,1 (which reverts back to the mean in very few periods). We get a negative R^2 Score, which is pretty bad. But there is no need to look at the score to know that the model can be improved greatly:



3.4 Neural Network Model with LSTM features

3.4.1 Predicting Log Returns

We prepare our data by setting aside the last observations of the period 1927-2021 as a validation dataset. We will be using this to tune the neural network more precisely.

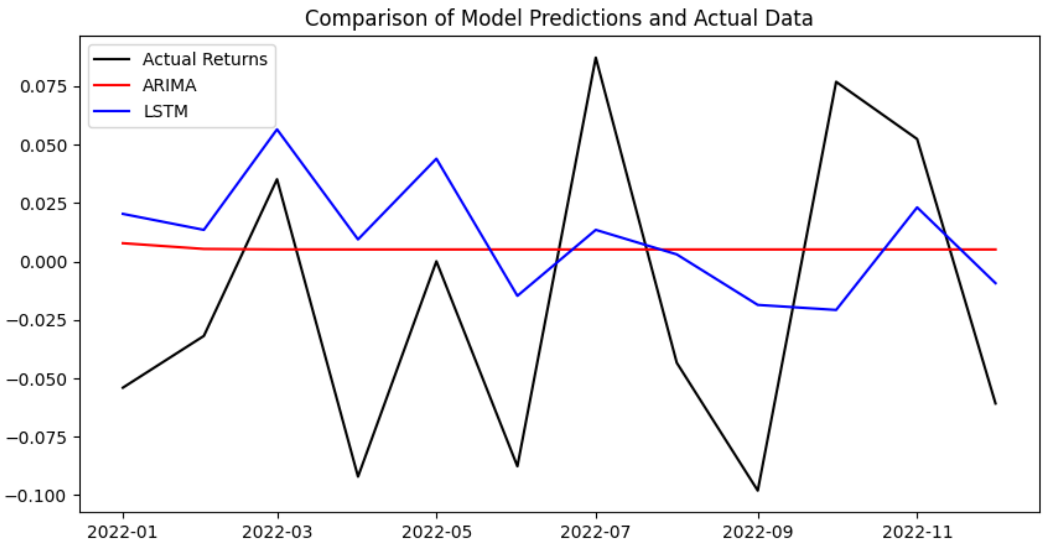
We decide to use an LSTM neural network in order to try to get it to capture the time series characteristics of our stock data.

We tried fitting the network with:

- LSTM layer with: 64, 16, 50 and 128 nodes (64 worked the best some times, although the performance is VERY unstable).
- Dense hidden layer with: 16, 32 and 64 nodes and ReLu activation function (also tried some others but their results were definitely worse).
- Adam optimizer with learning rate of: 0.01, 0.001 and 0.0001 (the best performing one was 0.001).
- MSE as the loss function.
- 50 epochs (tried to use more and also to implement early stopping, but early stopping didn't work properly and more than 50 epochs tended to conduct to huge instances of overfitting).
- Batch Size of 2 (also tried with 64 but the model got more accurate by using smaller batch sizes while not taking too long to run).

We also tried different combinations of using only the lags of the Log Returns themselves or together with the lags of the different ratios. In the end, what worked the best was using both the lags of the returns with the lags of the DE ratio. Some runs we would get R^2 Score up to 0.3 but never more than that. For most of the runs the model would still get a negative R^2 Score.

We can observe on the following comparison how the neural network does a better work than the ARMA model:



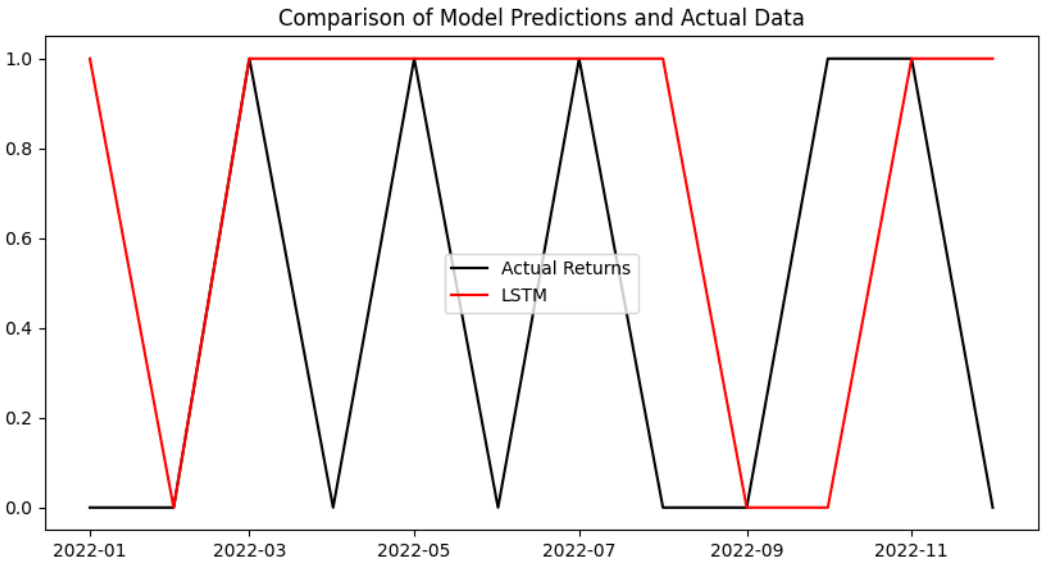
3.4.2 Predicting Direction

We tried whether predicting direction (up or down) of the Log Returns we would get better results.

We used the column that we prepared earlier which converted the returns into a binary variable of 0 if the return was negative and 1 if it was positive.

We applied the same neural network (but changed loss function from MSE to Accuracy because the problem now changed from regression to binary classification) and tuned the parameters in the same way, effectively doubling the amount of different settings we tried, but to no avail. In the case of this second instance, we never got the model to score higher than 0 for the R^2 Score, we only got negative scores.

The performance of the model on the test set can be seen to be very lacking in the following plot:

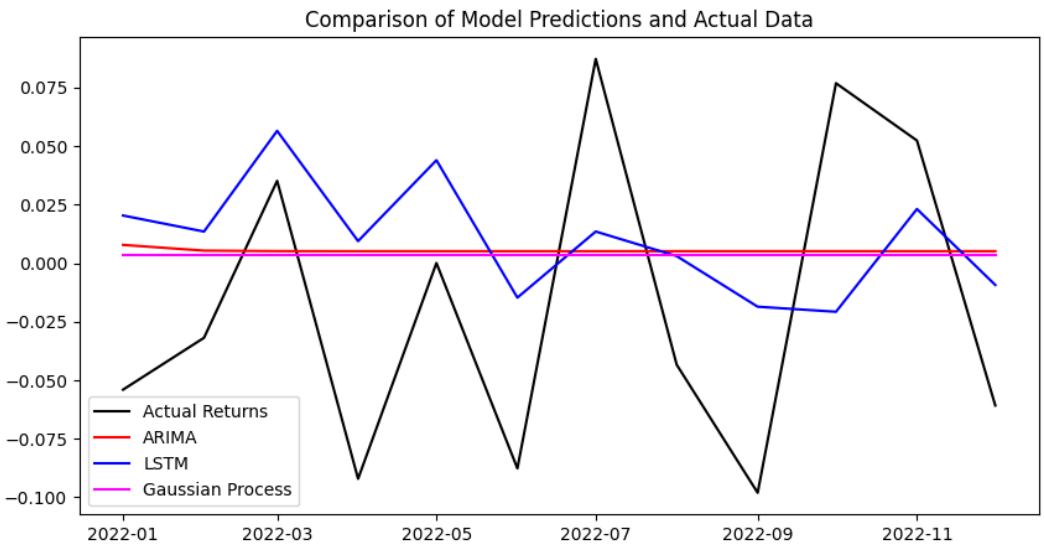


3.5 Gaussian Process Model

3.5.1 Predicting Log Returns

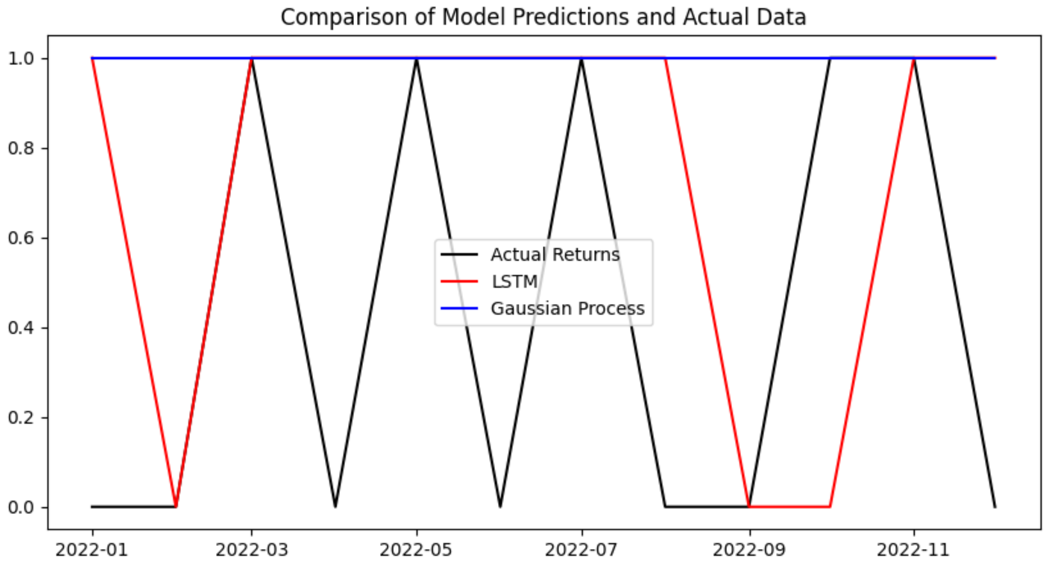
For this model, we didn't need to prepare the data too much, but we also performed the same feature selection as with the neural network. This time, though, none of our variables added anything to the model. Every fit came back with negative R^2 Scores: only lags of the returns, lags of returns and all the ratios, lags of the returns with only lags of one or two ratios. All the predictions were really bad, not even observably better than the predictions from the ARMA model.

We tried with both RBF and C kernels and also tuned the alpha to 0.001, 0.01, 0.1 and 1.0 but none of those alternatives improved the model in any significant way.



3.5.2 Predicting Direction

We also tried to predict direction with the Gaussian Process Model to see if there was any luck but the results were as bad as before. The scores in this case didn't go above 0 either with any of the options. The model always kept only predicting increasing returns all the time:



3.6 Conclusions

We can safely conclude that our ratios are not the most significant when trying to predict the returns of the S&P500. DP, DE and EP are not extremely useful when predicting the returns except if we have them on the same period of the price, of course, but then we are already on the period we try to predict so it renders the prediction moot.

The only ratio that seemed to get some better results (and only with the neural network model) was the DE. The rest of the combinations didn't really improve the best results we got with the neural network applied only to the lags of the log returns.

The direction predictions only made things worse and in that case not even the lags of the log returns were any better than using an MA 1 model.

In the end, we also considered using the stationary bootstrap to strengthen our conclusions, but our ratios were already bad enough that no amount of robustness checks would get them to magically work. The average R^2 Score we got with our variables was negative so we are already very confident in our conclusion that those ratios are not useful to predict the S&P500.

4 Authors Contribution

All authors contributed equally.