**Unveiling Behavioural Finance Through Market Indices: Predictive Insights from Google Trends and Advanced Machine Learning Models**

**1. Introduction:**

Amongst the finance industry, many competitors and key players hold a steadfast interest in behavioural finance, particularly concerning key macroeconomic events and the nature of the four principal market indices in the United States. This extensive analysis adopts an analytical approach to investigate the interplay between these pivotal financial indices, the Volatility Index (VIX), S&P 500, Dow Jones, and NASDAQ, with public search behaviour as captured by Google Trends.

By integrating traditional market metrics with alternative data sources, the analysis explores the potential for behavioural factors to influence financial outcomes. The methodology is structured as follows:

* **Data Preprocessing:** Aligning and synchronising datasets to ensure temporal consistency for robust analysis.
* **Descriptive Statistics:** Examining summary statistics to gain a foundational understanding of trends across indices and search interest.
* **Data Transformation:** Converting data into logarithmic scales to more accurately represent rates of growth and underlying trends.
* **Exploratory Data Analysis (EDA):** Using visualisations to uncover patterns, correlations, and anomalies within and between datasets.
* **Time Series Analysis:** Applying ARIMA models to each market index to model and forecast financial time series data.
* **Residual Diagnostics:** Analysing residual plots and ACF plots to identify spikes and autocorrelation, indicating unmodeled macroeconomic shocks.
* **Ljung-Box Test:** Using test statistics and p-values to assess the adequacy of the ARIMA models and identify patterns not captured by the models.

By combining these foundational steps with advanced methodologies, this study integrates traditional time series analysis and machine learning models to examine the predictive relationships between search interest and market metrics.

**2. Literature Review:**

Behavioural finance explores how psychological factors, and cognitive biases influence investor behaviour and market outcomes. The Efficient Market Hypothesis (EMH) has been challenged by researchers (Ţiţan, 2015) who demonstrate that market movements are not solely driven by fundamental valuations but also by investor sentiment and psychological responses to macroeconomic events (Malkiel, 1989). Studies like Kahneman and Tversky's Prospect Theory (1979) (Kahneman, 2013) laid the groundwork for understanding risk aversion and irrational decision-making in volatile markets.

The role of indices like the S&P 500, Dow Jones, NASDAQ, and VIX in reflecting market behaviour has been widely discussed in the literature. The VIX, often dubbed the "fear index," has received significant attention for its ability to predict market volatility (Whaley, 2000).

Mainly, it challenged the traditional Expected Utility Theory, which assumes that individuals act rationally and maximise utility based on probabilities of outcomes (Grant, 2007). Instead, Prospect Theory posits that human behaviour often deviates from rational decision-making due to cognitive biases and emotional factors, especially in volatile and uncertain market conditions (Kahneman, 2013).

Similarly, Vlastakis and Markellos explored the role of information demand, as reflected in search behaviour, on the volatility of financial assets. They found that higher search volume was often associated with greater uncertainty and increased market volatility (Vlastakis, 2012). These studies have therefore underlined the potential of public sentiment data to act as leading indicators of market trends.

Traditional time series techniques such as ARIMA (Box, 2013) and GARCH (Bollerslev, 1987) have been widely used to model and forecast financial time series data (Chen, 2011). These methods have been highly effective in capturing trends, seasonality, and volatility in markets (Xing, 2024). However, the ability of these methods to incorporate external or alternative data (e.g., search interest) is limited. More recent studies have proposed augmenting these traditional methods with external predictors, such as search interest, to improve the accuracy of time series forecasting (Choi, 2012).

**3. Keywords:**

Selecting keywords for Google Trends analysis required a strategic approach to pinpoint terms that encapsulate public reactions during times of market flux. The following keywords were meticulously chosen for their direct correlation to market sentiment and volatility:

* **"Market Crash"** - This term directly relates to significant market downturns, effectively capturing public fear and heightened anxiety about the possibility of substantial financial losses. Its usage aligns with the principles of Prospect Theory, which highlights the outsized emotional impact of potential losses over equivalent gains.
* **"Economic Uncertainty"** - A broader term that mirrors public concerns regarding the stability of economic conditions. Often preceding or coinciding with heightened market volatility, it encapsulates general ambiguity about future macroeconomic trends, which, according to behavioural finance, can lead to risk-averse decision-making and increased market fluctuations.
* **"Financial Crisis"** - This term invokes memories of past economic downturns, such as the 2019 Covid Pandemic, signalling public anxiety about either ongoing or potential crises. It is particularly relevant in environments where fear and loss aversion dominate investor behaviour, often triggering sharp shifts in market sentiment and volatility.
* **"Stock Market Volatility"** - A precise term that embodies the concept of market fluctuations, making it particularly relevant to the VIX ("fear index"). This keyword is vital for linking public sentiment to actual volatility metrics, as it underscores the market's sensitivity to perceived investor uncertainty.

These keywords were selected not only for their immediate relevance to volatility and market sentiment but also for their basis in behavioural finance theories. Specifically, they align with the idea that public fear, anxiety, and emotional responses can influence market dynamics, often leading to overreactions or underreactions during macroeconomic events.

The keyword selection process was justified through two complementary elements:

1. **Behavioural Insights**: Recognising that emotional responses, such as fear and anxiety, play a key role in driving investor behaviour and, by extension, market movements. Words like "market crash" and "financial crisis" are particularly reflective of these reactions, tapping into the psychological dimensions outlined in Prospect Theory (Kahneman, 2013), where loss aversion and risk sensitivity dominate decision-making under uncertainty.
2. **Economic Context**: Acknowledging the role of macroeconomic indicators in shaping investor sentiment. Terms such as "economic uncertainty" capture systemic, large-scale concerns that often precede or amplify market volatility. These terms bridge individual behavioural responses with broader financial and economic environments, offering a more holistic perspective.

**4. Data (Methodology, Summary)**

**Figure 1. VIX, Nasdaq, S&P500, DJI Indices**

A group of graphs showing different colored lines

Description automatically generated

Figure 1: Comparison of financial indices (VIX, Nasdaq, S&P 500, Dow Jones) on a logarithmic scale with Google Trends data for keywords like 'Inflation,' 'Recession,' and 'Stock Crash' (2014–2024), highlighting correlations between market volatility and public interest during major events such as the 2020 pandemic.

For this analysis, financial data for four key market indices were collected over a ten-year period, spanning from **October 15, 2014**, to **October 15, 2024**. These indices include:

* **VIX** **(^VIX):** Often referred to as the "fear index," used to measure expected market volatility.
* **NASDAQ (^IXIC)**: A tech-heavy market index representing the performance of numerous technology companies and growth stocks.
* **S&P 500 (^GSPC)**: An index representing 500 of the most prominent publicly traded companies in the U.S.
* **Dow Jones Industrial Average (^DJI)**: A price-weighted index of 30 significant, large publicly traded U.S. companies.

Financial data for VIX, NASDAQ, S&P 500, and Dow Jones indices were collected from Yahoo Finance over a ten-year period (2014-2024). To ensure a valid and coherent analysis, missing values were removed to ensure data consistency and integrity.

**4.1 Logarithmic Returns**

To perform a complete comparison, the study encompasses all possible combinations of the selected market indices with the Google Trends search interest data. Since financial indices operate on numerical scales representing dollar values, while Google Trends data reflects normalised search interest on a scale of 0 to 100, these datasets are inherently on vastly different magnitudes. Without addressing this disparity, direct comparisons could lead to inaccurate or biased interpretations.

To standardise the data and make comparisons meaningful, logarithmic transformations were applied to both the market index and search interest data. Logarithmic transformations are well-suited for analysing datasets with varying magnitudes because they scale down large values, focusing on proportional changes rather than absolute differences. This transformation is particularly common in economic and financial analyses due to its interpretive simplicity.

The formula for applying a natural logarithmic transformation is as follows:

Where:

* ( ) is the transformed (logarithmic) value,
* ( ) is the original value of the variable (e.g., the raw closing prices/values for each market) and,
* ( ) referring to a logarithm to base ‘e’, where (e roughly 2.718)).

So therefore, for an adequate and suitable comparison to be made, each of the closing returns present for the market indices (VIX, Nasdaq, S&P500, Dow Jones) were transformed (see Figure 1) using the following equation:

**4.2 Summary Statistical Analysis:**

**VIX:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mean** | **Median** | **St. Dev** | **Min** | **Max** | **Q1.25%** | **Q3.75%** | **Kurtosis** | **Skewness** |
| 18.243 | 16.275 | 7.2904 | 9.14 | 82.69 | 13.3075 | 21.32 | 12.888 | 2.600 |

The **VIX** demonstrates high kurtosis (12.88) and significant positive skewness (2.6), implying a leptokurtic or "fat-tailed" distribution with occasional extreme spikes in volatility, often aligning with market crises. Figure 3 thus presents its ACF and PACF graphs.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **ARIMA (1,0,0)** | **ARIMA (2,0,0)** | **Comparison** |
| ME | 0.00009 | 0.0011 | Both values are valid, as they near zero (show no bias) |
| RMSE | 1.965 | 1.943 | Lower values show better predictability, ARIMA (2,0,0) is slightly better |
| MAE | 1.155 | 1.152 | Lower values show better predictability, ARIMA (2,0,0) is slightly better |
| MAPE | 5.41% | 5.41% | Essentially the same. |
| MASE | 0.995 | 0.993 | Both models perform similarly to naïve forecasts. |
| ACF1 | -0.148 | -0.005 | ARIMA (2,0,0) has negligible residual autocorrelation |

Table 1 – Error Metrics Comparison for VIX Time-Series Data. Both models confirm stationarity as the data does not require differencing as well as the value for ACF1 being -0.05, suggesting that residuals aren’t autocorrelated. Overall, ARIMA(2,0,0) is a more improved model.

Throughout the process of analysing the summary statistics for each market index (VIX, Nasdaq, S&P500, Dow Jones), similar process as featured in Table 1 will be implemented to locate the optimal settings in order to obtain the most accurate forecasting predictions obtainable by the ARIMA model. For the case of VIX, the error metrics for both models (1,0,0) and (2,0,0) were obtainable, which proves non-stationarity, a feature that’s ideal for valid comparisons between data.

In general, ARIMA (2,0,0) displayed more promising error metric values as they near towards recommended values compared to that of (1,0,0), such as the lower values of RMSE and MAE which illustrate an enhanced capability to obtain more enhanced predictions.

**Nasdaq Summary:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mean** | **Median** | **St. Dev** | **Min** | **Max** | **Q1.25%** | **Q3.75%** | **Kurtosis** | **Skewness** |
| 9543.73 | 8147.60 | 3947.95 | 4215.320 | 18647.449 | 5903.41 | 13132.15 | -1.1154 | 0.4111 |

**A screenshot of a graph

Description automatically generated**

Figure – Residual Graphs for NASDAQ (stationarity) closing values regarding the ARIMA model with values P = 3, D = 1, Q = 0

The Nasdaq Index analysis reveals intriguing dynamics shaped by both market behaviour and macroeconomic events. Descriptive statistics highlight notable volatility, with a mean of 9,543.73 and a substantial standard deviation of 3,947.95.

The mildly right-skewed distribution (skewness = 0.4111) and platykurtic nature (kurtosis = -1.1154) suggest fewer extreme events than a normal distribution, though the presence of outliers, such as a minimum of 4,215.32 and a maximum of 18,647.45, points to the influence of unusual macroeconomic shocks, such as the COVID-19 pandemic or the Russia-Ukraine War in 2022.

The residual diagnostics from the ARIMA(3,1,0) model, shown in the provided screenshot, indicate clear challenges in capturing these dynamics. The residual plot reveals spikes, likely corresponding to these unmodelled macroeconomic shocks, while the ACF plot suggests some lingering autocorrelation at specific lags. The histogram of residuals approximates normality but includes some deviations, further supporting the hypothesis that external events drive these outliers.

**S&P500 Summary:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mean** | **Median** | **St. Dev** | **Min** | **Max** | **Q1.25%** | **Q3.75%** | **Kurtosis** | **Skewness** |
| 3272.180 | 2942.395 | 1030.994 | 1829.080 | 5859.850 | 2371.888 | 4151.980 | -0.9023 | 0.4856 |

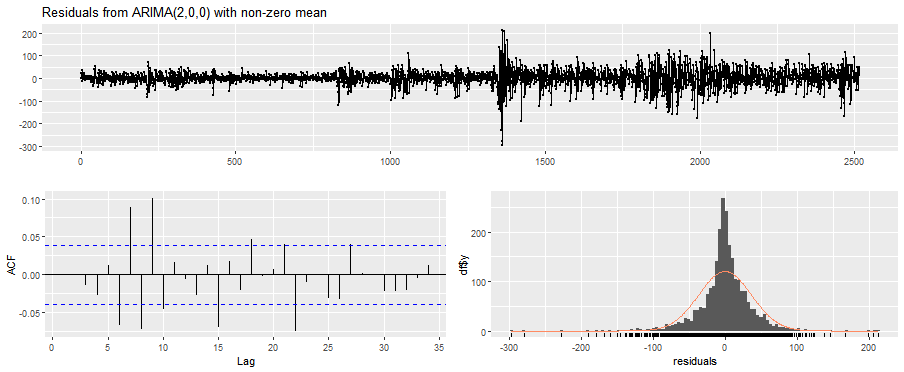


Figure – Residual Graphs for S&P500 (stationarity) closing values regarding the ARIMA model with values P = 2, D = 0, Q = 0

As for the S&P500, the residuals plot indicates that while the model captures much of the time series' variability, there are notable spikes in volatility, particularly around the 1500 mark, suggesting periods of market instability not fully accounted for by the model.

The ‘Auto Correlation Function’ (ACF) plot shows that most lags fall within the confidence intervals, indicating that the residuals are not significantly autocorrelated, which supports the adequacy of the ARIMA(2,0,0) model in removing autocorrelation from the original series.

Furthermore, the histogram with the overlaid density plot of the residuals approximates a normal distribution, although with some deviations at the tails, suggesting that while the model's assumptions are largely met, there might be some outliers or non-normality in the data. This analysis underscores the model's reasonable fit to the data but also highlights potential areas for further refinement or the inclusion of external factors to better capture the dynamics of the S&P 500 over the specified period.

**Dow Jones Summary:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mean** | **Median** | **St. Dev** | **Min** | **Max** | **Q1.25%** | **Q3.75%** | **Kurtosis** | **Skewness** |
| 27137.31 | 26385.02 | 7067.033 | 15660.18 | 43065.22 | 20833.47 | 33679.09 | -1.1366 | 0.1194 |

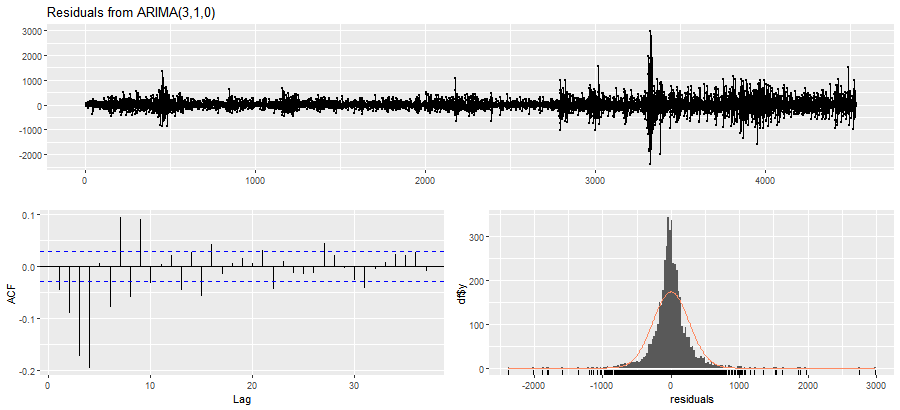


Figure - Residual Graphs for ARIMA (stationarity) closing values regarding the ARIMA model with values P = 3, D = 1, Q = 0

The residuals plot for the Dow Jones shows fluctuations around zero, which is indicative of the model capturing some level of mean reversion, similar to what was observed with the S&P 500 and Nasdaq.

However, there are pronounced spikes, particularly around the 3000 and 4000 (arguably) mark, suggesting periods of (yet again) heightened volatility or significant macroeconomic events affecting the Dow Jones, a pattern that echoes the findings from the other indices where macroeconomic events like pandemics or geopolitical conflicts might have introduced unexpected market movements.

The AutoCorrelation Function (ACF) plot for the Dow Jones indicates that there is some residual autocorrelation, especially at lag 1, where values exceed the confidence intervals. This suggests that the ARIMA(3,1,0) model has not fully accounted for all the time series dependencies within the Dow Jones data. This persistent autocorrelation across all three indices might imply that the chosen ARIMA models are not capturing all the nuances of market behavior or that there are external influences not included in the models.

Lastly, the histogram of residuals for the Dow Jones, when compared to the S&P 500 and Nasdaq, shows a distribution that approximates normality with the density plot closely following the bell curve. However, similar to the other indices, there are deviations at the tails, hinting at the presence of outliers or slight non-normality. This pattern is consistent with the findings from the S&P 500 and Nasdaq, where external events might have led to residuals that deviate from a perfect normal distribution.

Overall, while the models provide a reasonable fit, the comparison across these indices underscores the challenge of modeling financial time series due to the complex relationships of various economic, political, and global factors influencing market indices.

**5. Conclusion**

The analysis of the ARIMA models for the VIX, Nasdaq, S&P 500, and Dow Jones indices reveals significant patterns and influences that were not captured by the models. The high test statistics and low p-values from the Ljung-Box tests across all indices indicate that the models missed key components or external factors, such as macroeconomic events and investor sentiment, which play a crucial role in market dynamics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Market Index** | **Test Statistic** | **Degrees of Freedom** | **P-Value** | **Total Lags Used** | **Arima Model** |
| VIX | 73.468 | 7 |  | 10 | (3,0,0) |
| Nasdaq | 450.01 | 7 |  | 10 | (3,1,0) |
| S&P500 | 78.791 | 8 |  | 10 | (2,0,0) |
| Dow Jones | 483.39 | 7 |  | 10 | (3,1,0) |

Here's how it breaks down:

* **VIX**: With an ARIMA(3,0,0) model, the test gave us a statistic of 73.468. This high number means there are still patterns in the VIX's behavior that our model didn't catch. The p-value reveals that there's virtually no chance this result happened by accident; the model definitely missed something.
* **Nasdaq**: Using an ARIMA(3,1,0) model, the test statistic was even higher at 450.01. This indicates we missed even more patterns in the Nasdaq's movements. The tiny p-value again confirms our model isn't capturing everything.
* **S&P 500**: For this index, with an ARIMA(2,0,0) model, we got a test statistic of 78.791. It's telling us there are significant patterns our model didn't account for, with a p-value showing we're very confident in this conclusion.
* **Dow Jones**: Here, with an ARIMA(3,1,0) model, the test statistic was the highest at 483.39, suggesting our model missed a lot of the Dow Jones's behavior. The p-value, again almost zero, confirming the model's inadequacy.

These conclusions therefore emphasise the complication of financial markets and limitations of traditional time series models. While ARIMA models seem to provide an introductory approach to being able to utilise Google Trends values to predict future market index values, they require heavy refinement and an inclusion of additional explanatory variables to better capture the nuances of market behavior. A few examples include GARCH models and ARIMAX (allowing for exogenous regressors).

This ultimately suggests a need for more sophisticated modeling techniques that integrate both traditional financial metrics and alternative data sources, such as Google Trends, to enhance predictive accuracy and provide deeper insights into market trends.

This heavily ties into a plethora of other attempts to utilise similar methodologies with ARIMA, stating that while ARIMA models serve as a good starting point, they often require enhancement through the integration of other statistical or machine learning models and the inclusion of external data to better capture the complexities of financial markets (Ade, 2023).

# References

Ade, M. (2023, June). The role of exogenous variables in time series forecasting of economic indicators. Retrieved from http://bit.ly/40rhRFp

Bollerslev, T. (1987). A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *The review of economics and statistics*, 542-547. doi:https://doi.org/10.2307/1925546

Box, G. (2013). Box and Jenkins: Time Series Analysis, Forecasting and Control. *A Very British Affair: Six Britons and the Development of Time Series Analysis During the 20th Century*, 161-215. doi:https://doi.org/10.1057/9781137291264\_6

Chen, C. H. (2011, June). Short-time traffic flow prediction with ARIMA-GARCH model. *2011 IEEE Intelligent Vehicles Symposium (IV)*, 607-612. doi:10.1109/IVS.2011.5940418

Choi, H. &. (2012). Predicting the present with Google Trends. *Economic record*, 2-9. doi:https://doi.org/10.1111/j.1475-4932.2012.00809.x

Grant, S. H. (2007). Expected Utility Theory. *INSEAD Business School Research Paper*. doi:https://dx.doi.org/10.2139/ssrn.1033982

Kahneman, D. &. (2013). Prospect theory: An analysis of decision under risk. *Handbook of the fundamentals of financial decision making: Part I*, 99-127. doi:https://doi.org/10.1142/9789814417358\_0006

Malkiel, B. G. (1989). Efficient market hypothesis. *Palgrave Macmillan UK*, 127-134. Retrieved from https://link.springer.com/chapter/10.1007/978-1-349-20213-3\_13

Ţiţan, A. G. (2015). The efficient market hypothesis: Review of specialized literature and empirical research. *Procedia Economics and Finance*, 442-449.

Vlastakis, N. &. (2012). Information demand and stock market volatility. *Journal of Banking & Finance*, 1808-1821. doi:https://doi.org/10.1016/j.jbankfin.2012.02.007

Xing, Y. Y. (2024). Predicting NVIDIA's Next-Day Stock Price: A Comparative Analysis of LSTM, MLP, ARIMA, and ARIMA-GARCH Models. *arXiv preprint*. doi:https://doi.org/10.48550/arXiv.2405.08284