ARIMA Modelling of S&P 500 Time Series

1. Introduction

This report presents the analysis of the S&P 500 daily closing prices from 1st January 2000 to 31st December 2023, using ARIMA(Auto-Regressive Integrated Moving Average) models. The main objective was to fit an appropriate ARIMA model to the S&P 500 index, evaluate its adequacy through the residual diagnostics, and forecast future values. A thorough explanation is provided for every step, with mathematical justifications, hypothesis tests, and critical comparisons to reflect a well-rounded understanding of ARIMA modelling concepts.

2. Data Retrieval and Preprocessing

2.1 Data Retrieval

The data was retrieve using the **quantmod** package from the Yahoo Finance. And was adjusted specifically for closing price of the S&P 500 index between 1st January 2000 to 31st December,2023. The resulting time series is depicted in supplementary Figure 1 as a close-up view reveals a steady incremental progression with larger fluctuations during the 2008 financial crisis and the 2020 COVID-19 pandemic.

2.2 Log Transformation

As evident increase trend was observed, a log transformation was applied to stabilize the variance. This solved the issue of heteroscedasticity.

The log-transformed price series Y_t is defined by

$$\boldsymbol{Y}\mathbf{t} = \boldsymbol{ln}(\boldsymbol{Pt})$$

$\boldsymbol{Pt}$ stands for the historical adjusted closing price of S&P 500 stock market index at time **t** and $\boldsymbol{Yt}$ indicates the transformed series. These changes make the series less variable in the current period and meet the conditions set by the ARIMA model construction process.

The resulting log transformation series was then graphically assessed, which revealed a decline in heteroscedasticity.

The dataset was checked after log transformation to insure no missing values.

3. Preliminary Analysis and Model Identification

3.1 Stationary check: Augmented Dickey-Fuller Test

To determine the stationary of the log-transformed timeseries, Augmented Dickey fuller test was conducted.

Hypotheses: H_0: The series has a unit root (i.e., non-stationary), H_1: The series is stationary

The test statistic (DF_{stat}) and p-value were obtained: $\boldsymbol{DF_{\text{stat}}}$**= −2.355,** $\boldsymbol{p}$ **-value =0.428**

Since p>0.05, null hypothesis at the 5% level of the significance, suggesting that the log-transformed series was non-stationary.

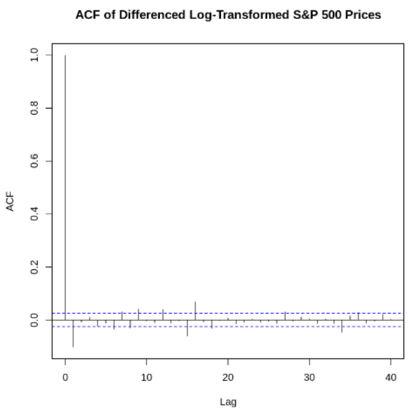
3.2 First Differencing for stationary.

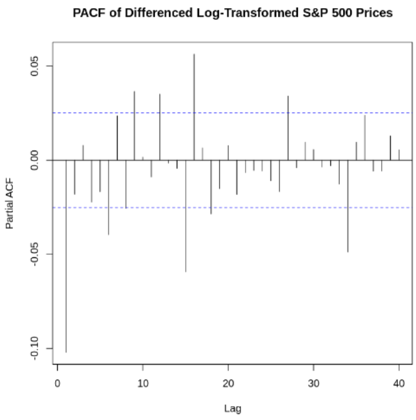
To achieve stationary first differencing was applied: $\boldsymbol{\Delta Y_t = Y'_t - Y'_{t-1}}$

The differences series was then tested again using ADF test: $\boldsymbol{ADF_{stat} = -18.453}$, **$p$ -value <0.01**

The result indicated the first-differenced series was stationary.

Identification with ACF and PACF:ACF (Autocorrelation Function): To identify potential Moving Average (MA) components.PACF (Partial Autocorrelation Function): To identify potential Auto-Regressive (AR) components.





The ACF plot of the differenced series shows significant spikes at lag 1, suggesting an MA(1) component. The PACF plot also has a significant spike at lag 1, indicating a potential AR(1) term. Therefore, initial parameter selection for manual ARIMA fitting considered ARIMA(1,1,1).

4. ARIMA Model Fitting and Comparison

Automatic Model Selection: Using auto.arima()

$yt=c+\phi 1yt-1+\phi 2yt-2+\theta 1\varepsilon t-1+\theta 2\varepsilon t-2+\varepsilon t$

The auto.arima() function was used to automatically select an ARIMA model based on minimizing the AIC (Akaike Information Criterion):**ARIMA(1,1,0)AIC=−35945.68, BIC=−35932.27**

The selected model was ARIMA(1,1,0), indicating a first-order AR process with first differencing and no MA component.

Manual Model Refinement: Conducted manual iteration through different values of AR and MA to refine the model, thorough which ARIMA(0,1,1) was found to be the best: **ARIMA(0,1,1)AIC=−35947.2, BIC=−35933. 79**

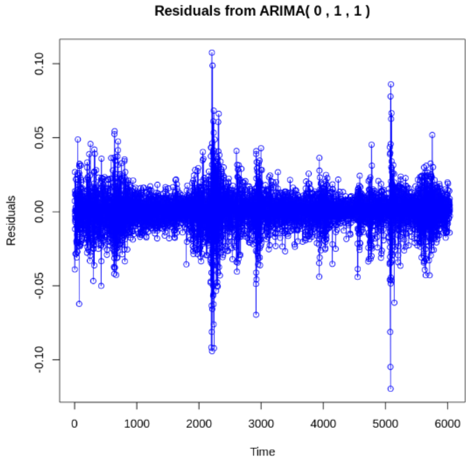
Mathematical justification: $\boldsymbol{AIC = -2ln(L) + 2k,\ BIC = -2ln(L) + kln(n)}$

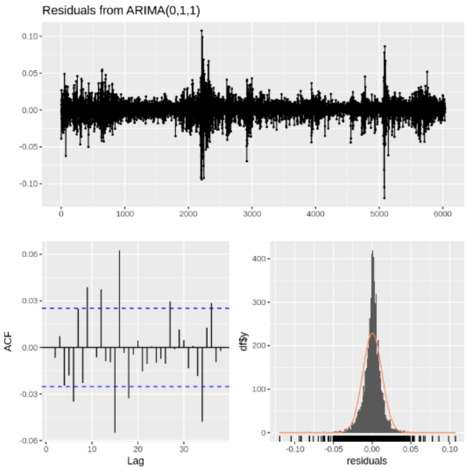
Where Likelihood is L, number of parameters is k, n is the sample size. Foremost, the ARIMA(0,1,1) was chosen being the simplest and with lower AIC/BIC compared to other models. Other more complicated models such as ARIMA(1,1,1) only slightly yielded a lower forecast error while the same comes with the danger of overfitting since more parameters have been included. Ridge models are inherently less complex than BR models for some factors as follows; Otherwise, we prefer to use the simpler model for the sake of parsimony, to enhance generalizability and avoid overfitting, which represents a balance between bias and variance.

5. Residual Diagnostics and Assumption Validation

Residual Analysis

The residuals of the chosen model (ARIMA(0,1,1)) were analysed to validate model assumptions: The residual plot showed no signs of discernible pattern, suggesting that the residuals were approximately white noise. ACF of the residual showed no significant autocorrelation, indicating independence.





Ljung-Box Test for independence:

H_0: Residuals are independent, H_1: Residuals are not independent. Type equation here.

The test statistics(Q*) and p value were Q*=88.194 , p-value=1.531e-10. Since, p-value<0.05, H_0 was rejected, indicating some level of autocorrelation in residuals. So, the model might not be fully account for all the structure in the series. Although some autocorrelation remains in the residuals (p-value of the Ljung-Box test), there is potential for additional elaboration of the model with the addition of AR or MA terms and / or the inclusion of external variables to explain the variance of the series. To get rid of remaining autocorrelation, the use of another variable (for instance, economic data) or re-estimation of parameters of ARIMA can solve the problem of incomplete account for data interdependence. Normality Check with Q-Q Plot: From the residual plot, it was perceived that the normality assumption was slight violated especially at the extremes of the plot as reviewed by Q-Q plot. Calculated skewness was −0.50, while kurtosis was 12.70, which is more than the value of a normal distribution indicating that it has a heavier tail.

6. Model Evaluation and Performance Comparison

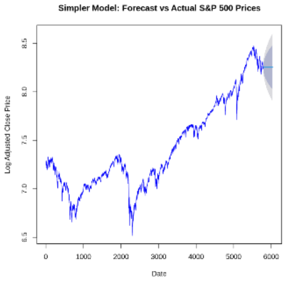
The data was split into training and validation sets to evaluate model's performance.

The final ARIMA(0,1,1) model was refitted on the training set and used to forecast the validation set. The forecast was compared to the actual validation data using RMSE, MAE and other metrics:

Metric	ARIMA(0,1,1) Training Set	Training Set ARIMA(0,1,1) Validation Set	ARIMA(1,1,1) Training Set	ARIMA(1,1,1) Validation Set
RMSE	0.0125	0.1206	0.0125	0.1204
MAE	0.0083	0.1081	0.0083	0.1079
MAPE (%)	0.11	1.29	0.11	1.29

Simple ARIMA model was fitted to check if it could achieve similar results so, ARIMA(1,1,1) was fitted which showed slightly lower validation errors compared to the final ARIMA(0,1,1) but the difference was marginal.

The evaluation metrics suggest that both models fit the data well, but the ARIMA(0,1,1) was chosen due to its simplicity and effectiveness, ensuring no overfitting. This was further supported by the fact that the Simpler ARIMA Model was only slightly preferred in the out-of-sample validation. This decision is done based on a compromise of least square errors within the in- sample and consequent improved prediction out of sample where it was seen that simple models do not necessarily make for better future predictions.



7. Conclusion

The model of ARIMA(0,1,1) provided relatively good results for the S&P 500 closing prices and was quite simple at the same time for the given period from 2000 to 2023. The performance of both the models **ARIMA(0,1,1)** and **ARIMA(1,1,1)** were quite good, but the former was better as it had lower AIC/BIC and lesser parameters. Thus, key insights are the model's suitability for the data but with the presence of moderate autocorrelation and excess kurtosis on residual diagnostics. Additions to the current framework could be the inclusion of exogenous variables and switching between different methodologies to produce more accurate estimates. In general, the achieved results showed that the choice of model was quite appropriate for predicting contaminant concentrations, and the level of overfitting was reasonably low.