Forecasting the Sweet Spot: A Time Series Analysis on Cocoa Prices

Ethan Wu [1008507661], Rebecca Li [1008885949]

2025-04-04

Abstract

The present study investigates the application of time series analysis techniques to model and forecast cocoa futures prices. Commodity price forecasting is a crucial task in fields such as finance, agriculture, and trade, as price volatility can significantly impact economic decision-making. Cocoa prices are subject to numerous external factors, including weather conditions, geopolitical events, supply chain disruptions, and market speculation. The purpose of this study is to develop and evaluate predictive models for cocoa futures prices using historical price data and external variables such as climate factors. The datasets used in this study include daily cocoa futures prices from the International Cocoa Organization (ICCO) and climate data from Ghana, sourced from the National Centers for Environmental Information (NCEI). A range of statistical and machine learning models will be explored, including classical time series models (ARIMA, GARCH) and more advanced approaches (e.g. LSTM neural networks). The results of this study will help to assess the effectiveness of various forecasting models and provide insights into the key factors driving cocoa price movements. Future research could expand on this work by incorporating additional macroeconomic indicators and alternative modeling techniques to improve forecasting accuracy.

Introduction

The ability to predict commodity prices is critical for financial markets, agricultural stakeholders, and policymakers. Cocoa, a vital ingredient in chocolate production, experiences price fluctuations due to various supply and demand factors, including climate conditions, geopolitical instability, and speculative trading. The objective of this study is to develop a robust forecasting model for cocoa futures prices by leveraging time series analysis techniques. Two key research questions arise from this study: whether time series models effectively predict cocoa futures prices based on historical data, and how external factors (such as climate variables) influence cocoa price movements. It is hypothesized that time series models will be effective in predicting cocoa futures prices based on historical data, and that relevant climate variables will have a significant effect on the movement of cocoa prices.

The data section of this report will describe the datasets used in this study and the preprocessing steps applied. Next, the literature review section will summarize and analyze existing research on time series forecasting and commodity price modelling, highlighting common approaches and limitations. The methodology section will then outline the statistical models chosen for forecasting, including the rationale for each approach. Next, the results section will present the findings of the analysis, comparing model performance using appropriate evaluation metrics. Finally, the discussion and conclusion section will interpret these findings, address main study limitations, and propose future research directions.

Data

Data Collection

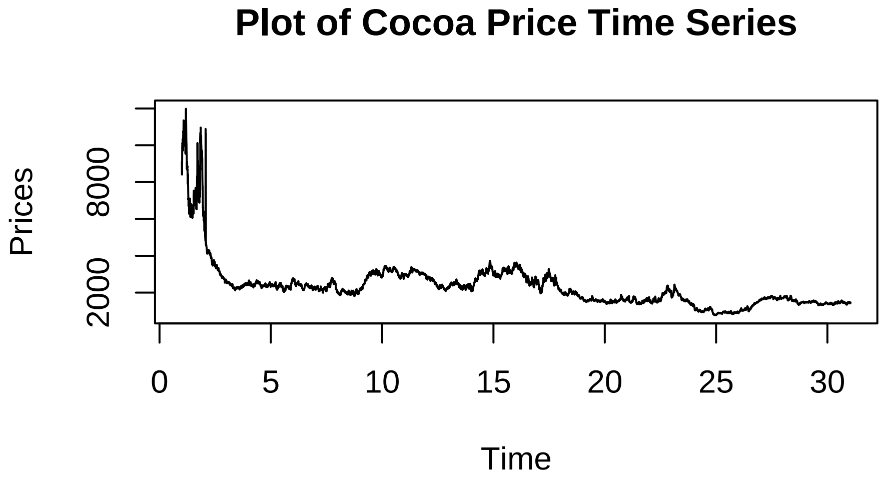
The present study utilizes two primary datasets: the cocoa futures price data, and the Ghana climate data. The cocoa futures prices data consists of daily closing prices for cocoa futures contracts from ICCO, covering the period from March 10, 1994 to February 27, 2025. The Ghana climate data consists of daily records of temperature and precipitation from the NCEI, providing climate insights for one of the largest cocoa-producing countries.

Data Cleaning and Preparation

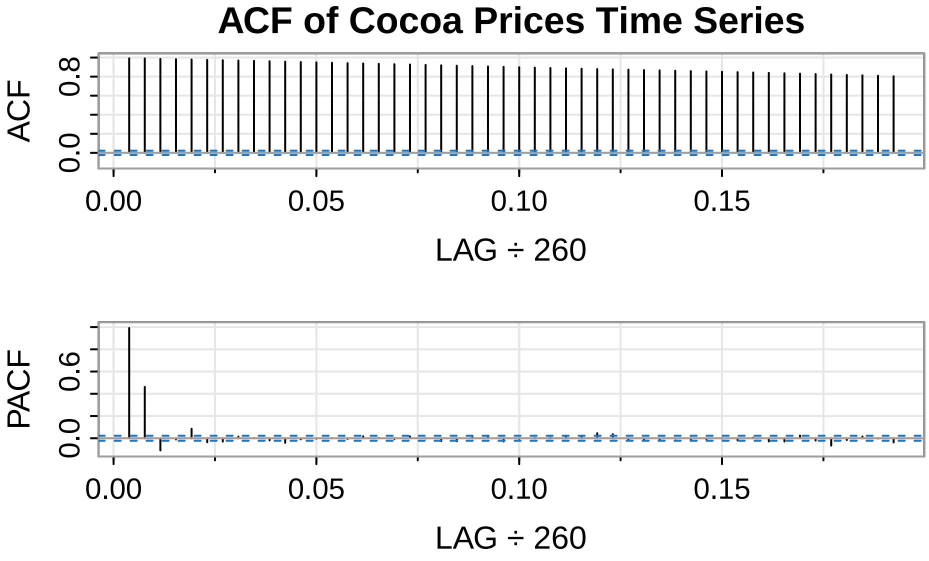
To prepare these datasets for analysis, several preprocessing steps were applied. Missing values in the climate dataset were handled by interpolation where appropriate, and extreme outliers in the price dataset were examined to determine if they were due to reporting errors, or if they should remain in the dataset. Additionally, all variables were standardized to ensure comparability (across measurements and datasets).

Initial visualizations of the cocoa price data reveal strong seasonal patterns, suggesting that certain times of the year exhibit recurring price trends. A decomposition analysis further highlights the presence of long-term trends and cyclical behaviour in the data. Correlation analyses between cocoa prices and climate variables suggest that temperature and precipitation fluctuations may play a role in cocoa price movements, which will be further investigated in the modeling section.

Descriptive Statistics and Visualizations



The time series plot of cocoa prices shows volatility early on, with prices spiking before sharply declining. However, the rest of the series fluctuates within a more stable range, with visible trends but no obvious seasonality. The variance of this time series appears to change over time, which indicates that models such as GARCH may be appropriate. Differencing may also be needed to achieve stationarity.



```
	Augmented Dickey-Fuller Test

data:  cocoa_ts
Dickey-Fuller = -6.6276, Lag order = 19, p-value = 0.01
alternative hypothesis: stationary
```

The ACF plot shows a very gradual decay, with autocorrelations remaining positive and significant over a large number of lags. This indicates clear non-stationarity in the series – the mean of this series is not constant over time, and differencing will likely be required. The PACF plot shows a sharp cutoff after lag 1, which is consistent with the presence of an autoregressive component in the data. Together, these patterns suggest that an ARIMA model may be appropriate, specifically the ARIMA(1,1,0) model.

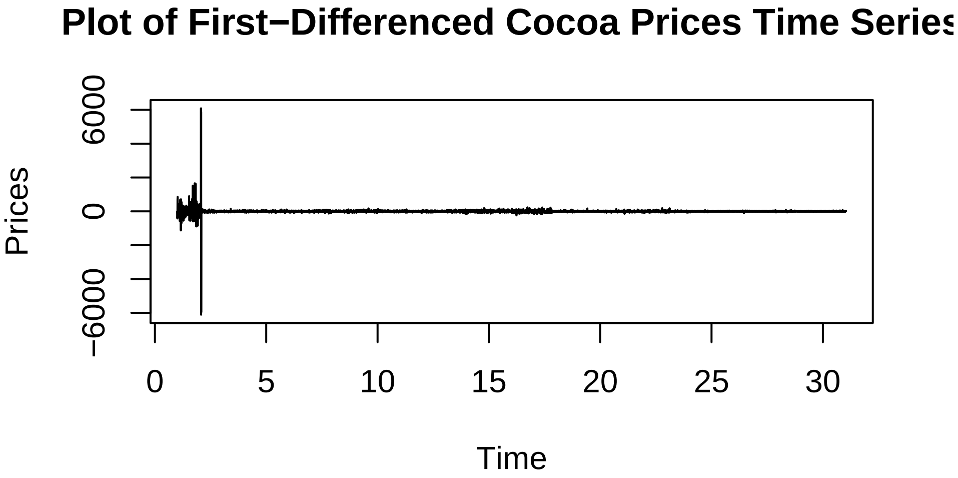
The ADF test result ($p = 0.01$) suggests that the cocoa price series is stationary, as the null hypothesis is rejected. However, this contrasts with the ACF plot, which shows a slow decay indicative of non-stationarity. This may be due to the high persistent autocorrelationsor structural shifts in the data that the ADF test does not fully capture. Despite the test result, the visual evidence supports differencing to ensure robustness in modeling.

```
[1] 2359043
```

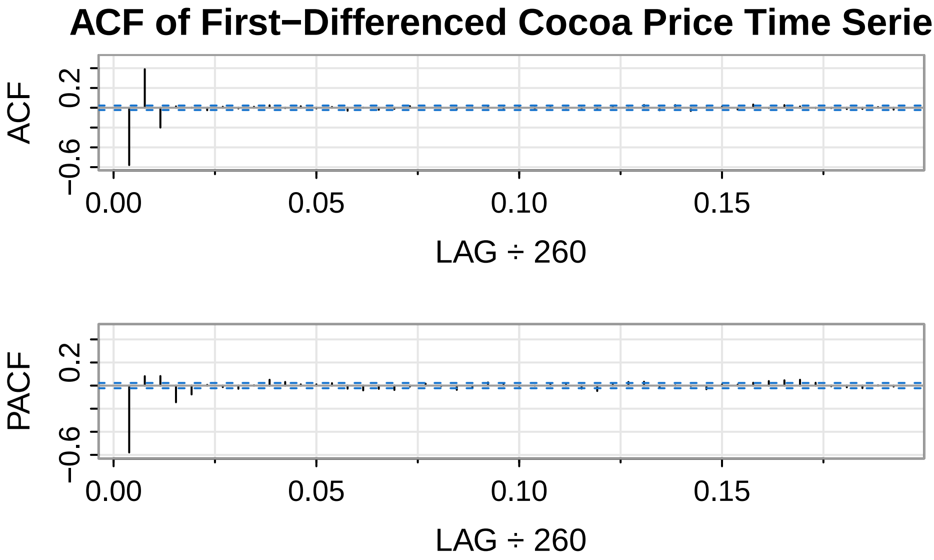
```
[1] 278149.1
```

˜˜˜ will make this pretty :P

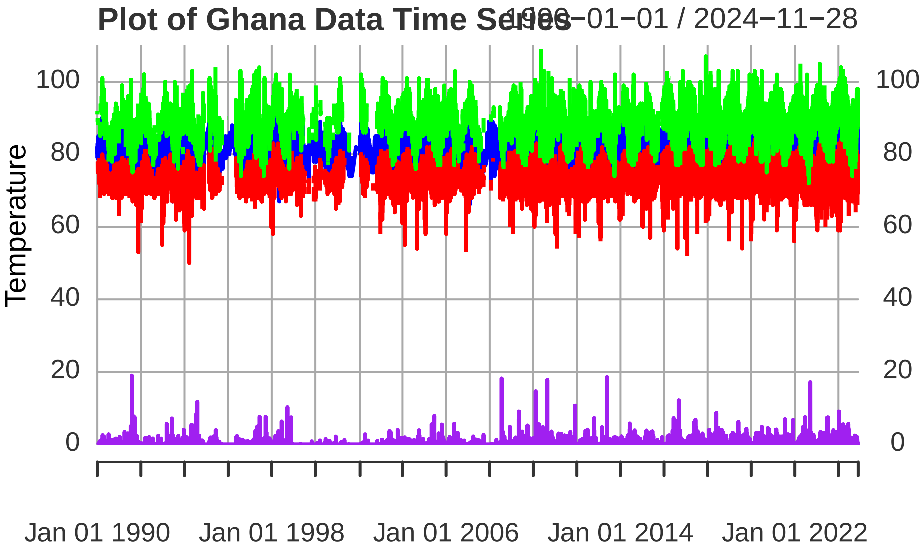
The variance in the first half of the cocoa price series is substantially larger (2,359,043) than in the second half (278,878.2) of the series, indicating a significant decrease in variability over time. This strong difference in variance confirms the presence of heteroskedasticity in the series, suggesting non-constant volatility and the potential need for techniques such as GARCH to properly represent the data.



This plot shows that differencing has successfully removed the long-term trend, with the series now fluctuating around a stable mean. However, high volatility still remains in the early portion of the series, with large spikes in both directions. This indicates that while first differencing may have helped to address non-stationarity in the mean, the issue of heteroskedasticity still remains and may impact model performance. The explicit modeling of volatility through a GARCH model thus may be necessary.

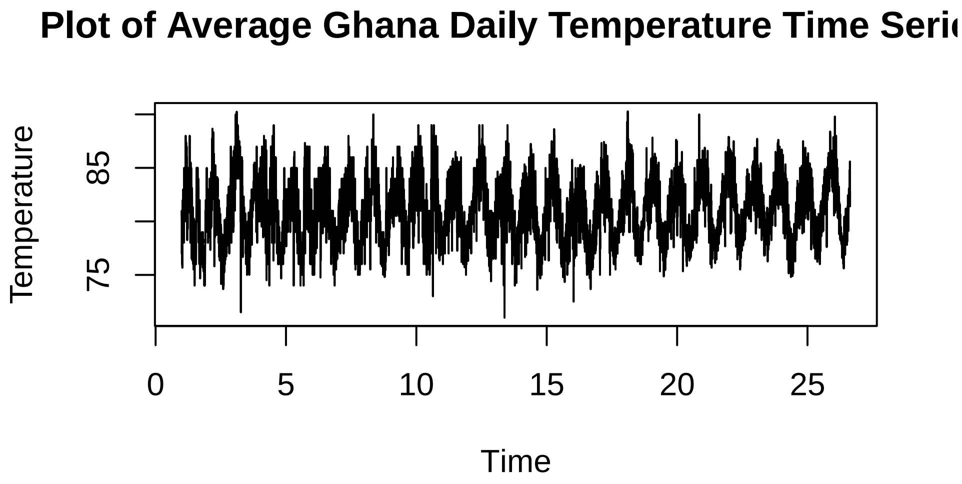


The ACF and PACF plots for the first-differenced cocoa price time series show that autocorrelations drop off quickly and remain within the significance bounds after the first few lags – this indicates that differencing has removed the non-stationarity present in the time series. Both the ACF and PACF plots show a significant spike at lag 1, suggesting that an ARIMA(1,1,0) model may be appropriate for capturing the trends and patterns present in the data.

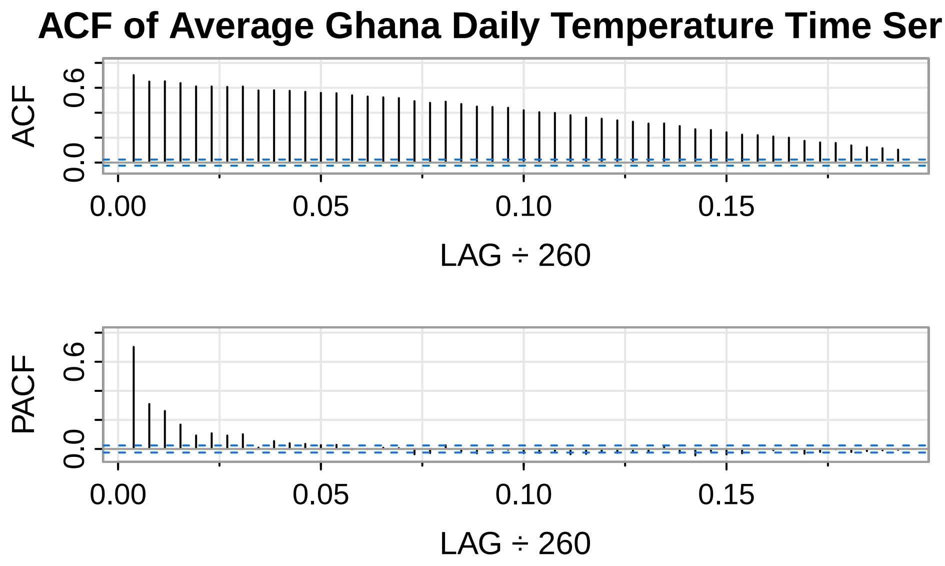


This time series plot shows strong and consistent seasonal patterns in the temperature variables (TMIN, TAVG, and TMAX), with temperatures peaking and dipping in regular yearly cycles from 1990 to 2024. The amplitude and shape of these cycles remain relatively stable over time, suggesting a predictable seasonal structure to the data. The precipitation series (purple) appears more irregular, with sporadic spikes and no clear seasonal trend – this indicates greater variability. From this plot, temperature data could be useful as seasonal predictors in modeling cocoa prices.

Note that the plot shows large chunks of missing data from the TMAX and TMIN variables, which may be difficult to extrapolate. Therefore, the best course of action here is to proceed with only the TAVG variable as a predictor for modeling cocoa prices. From the nature of this dataset, there are multiple observations every day from different stations – therefore the overall average of the average temperatures from each date are taken to obtain one final average nationwide temperature for each day.



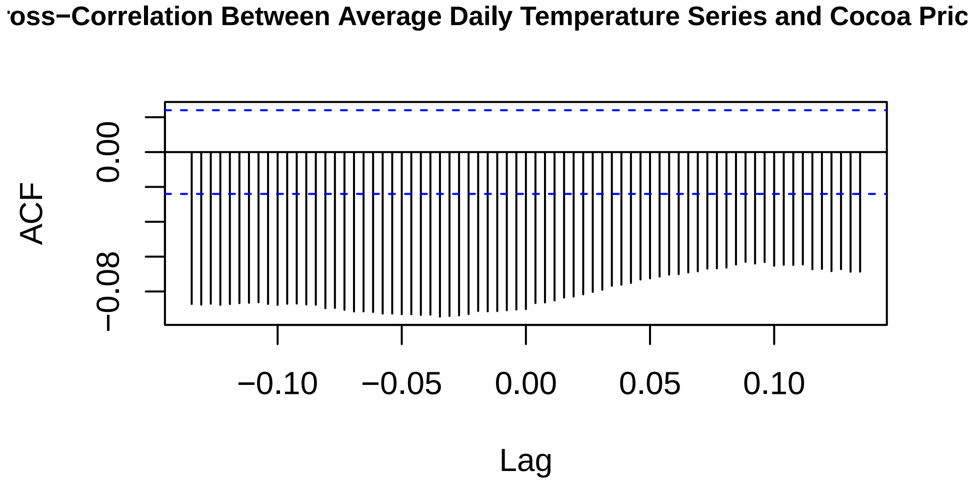
This plot shows a clear and consistent seasonal pattern, with temperatures oscillating regularly throughout the year. This strong seasonality, as well as relatively stable variance and no evident long-term trend, suggests that temperature follows a stationary seasonal process. These characteristics make TAVG a suitable and stable predictor for forecasting cocoa prices.



The ACF plot of the average daily temperature time series shows strong autocorrelation at early lags and a slow, gradual decay, which is characteristic of a seasonal and potentially non-stationary process. The PACF plot shows a series of spikes that taper off, suggesting the presence of both autoregressive and seasonal components. These patterns support the conclusion that temperature follows a highly persistent seasonal process.

```
	Augmented Dickey-Fuller Test

data:  ghana_ts_final
Dickey-Fuller = -7.1671, Lag order = 18, p-value = 0.01
alternative hypothesis: stationary
```

The ADF test result for this time series, $p = 0.01$, indicates that the series is stationary. Combined with the ACF and PACF plots, this confirms that while the series is stationary, it does contain strong seasonal autocorrelation to be accounted for in modeling.



The cross-correlation plot shows statistically significant negative correlations at many lags, indicating that past values of average temperature are inversely related to future cocoa prices. This suggests a lagged effect, where higher temperatures may lead to lower cocoa prices after a delay. This is potentially due to climate impacts on production or harvesting cycles.

Literature Review

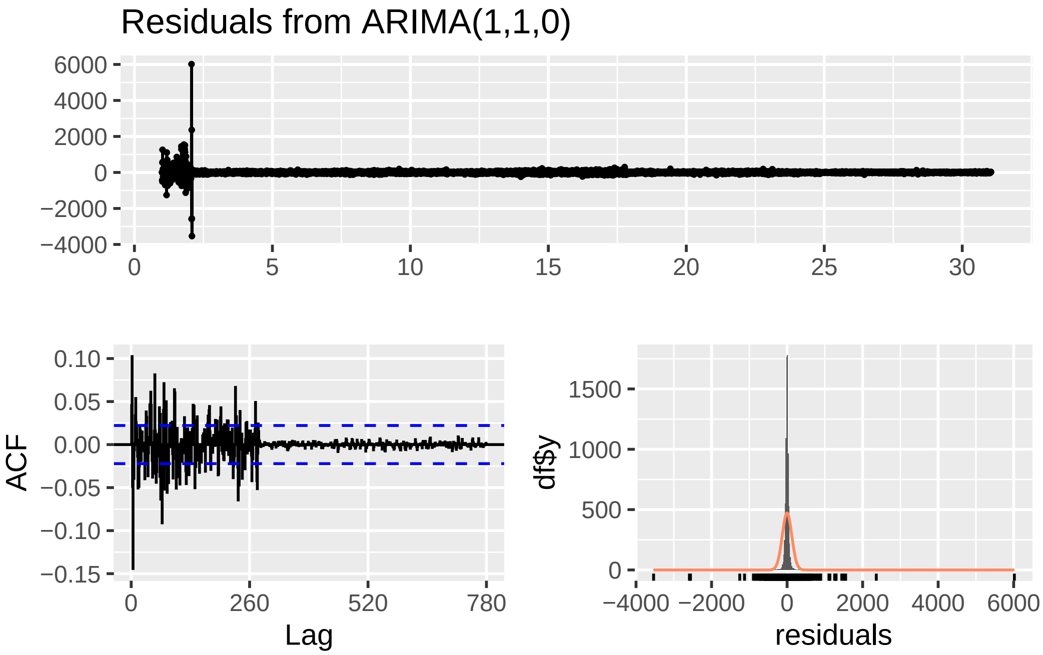
to be done...

Methodology and Assumptions

ARIMA(1,1,0)

The ARIMA(1,1,0) model is a classical linear time series model that combines autoregressive (AR) terms with differencing to achieve stationarity of the series. The structure includes one autoregressive term, first-order differencing, and no moving average component. This model captures short-term dependencies in the data while controlling for long-term trends.

EDA results showed that the original cocoa price time series was non-stationary in the mean but became stationary after first differencing. The ACF of the differenced series quickly tapers off, and the PACF shows a significant spike at lag 1 – this suggests an autoregressive component of order 1. An ARIMA(1,1,0) model is thus appropriate in this context, as it aligns with the observed autocorrelation structure of the data and captures both short-term persistence and long-term trend removal.


Residuals from ARIMA(1,1,0)

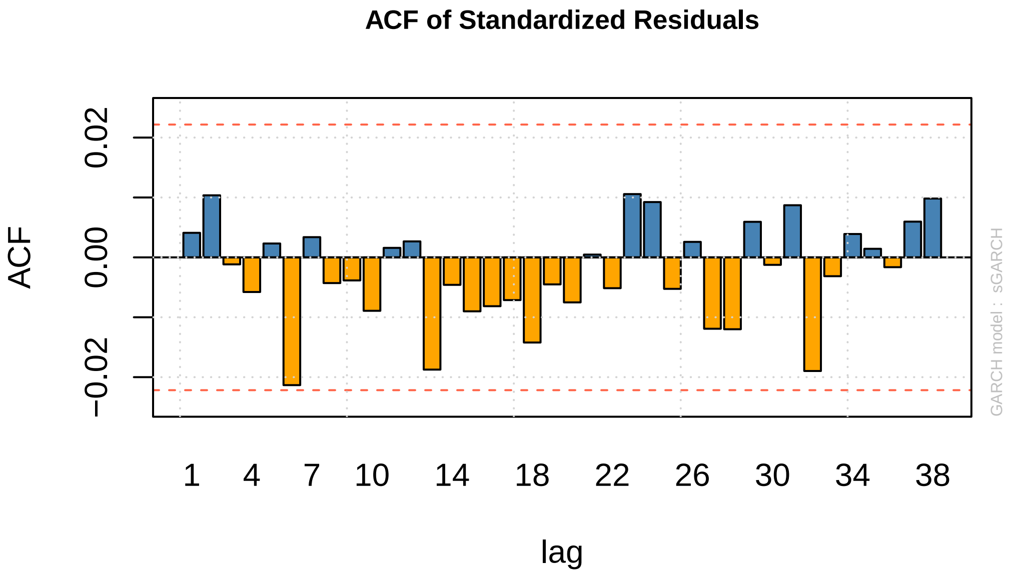
```
    Ljung-Box test

data:  Residuals from ARIMA(1,1,0)
Q* = 1951, df = 519, p-value < 2.2e-16

Model df: 1.   Total lags used: 520
```

Interpretation of assumption checking...

GARCH(1,1)

The GARCH(1,1) model is a volatility model that accounts for time-varying variance, commonly used in financial time series modeling. It consists of a constant, a lagged squared error term (ARCH), and a lagged conditional variance term (GARCH). This allows the model to capture clustering in volatility and heteroskedasticity over time.

The cocoa price time series exhibited strong heteroskedasticity, shown by the large discrepancy in variance between the first and second halves of the dataset. This pattern is typical in commodities markets, where volatility tends to cluster during periods of market shocks. A GARCH(1,1) model is therefore appropriate to model the variance process and improve forecast accuracy by addressing volatility dynamics.

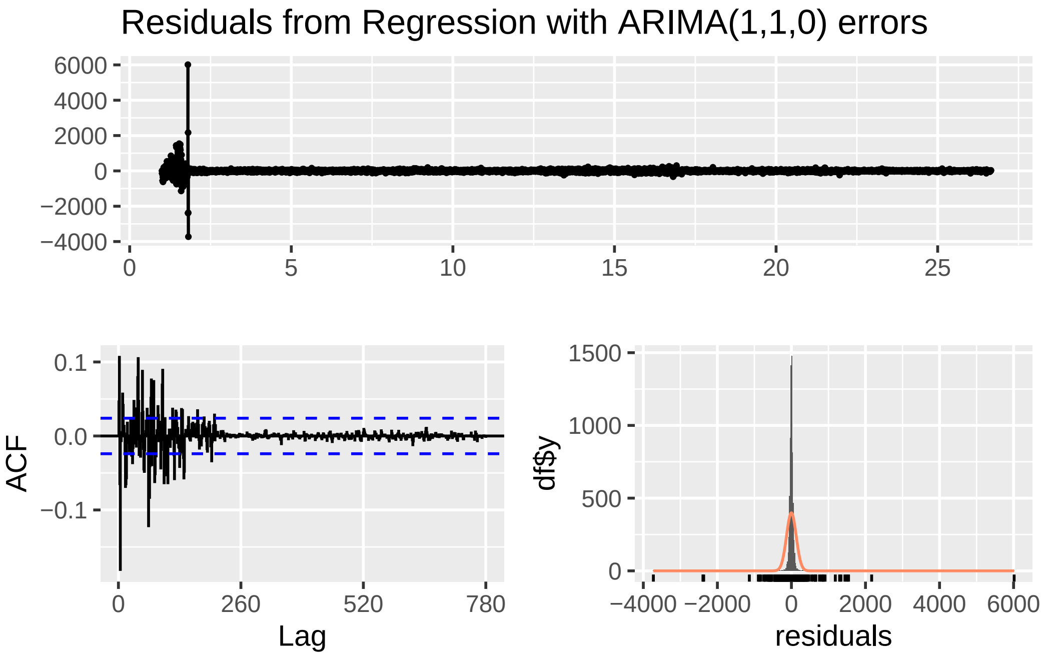


Interpretation of assumption checking...

ARIMA(1,1,0) With Ghana Climate Data (ARIMA-X)

The ARIMA-X model extends the ARIMA model framework by incorporating an exogenous regressor. In this model, we include the average daily nationwide temperature (TAVG variable) from the Ghana climate dataset as an external predictor to explain variation in cocoa futures prices beyond their own past values. This structure allows us to capture climate-driven influences on market prices.

Earlier cross-correlation analysis showed statistically significant negative correlations between lagged temperature values and cocoa futures prices, indicating that temperature may play a meaningful role in cocoa price dynamics. By including the TAVG variable from the Ghana climate dataset as a regressor, the ARIMA-X model may better capture potential cocoa price effects linked to climate variability.



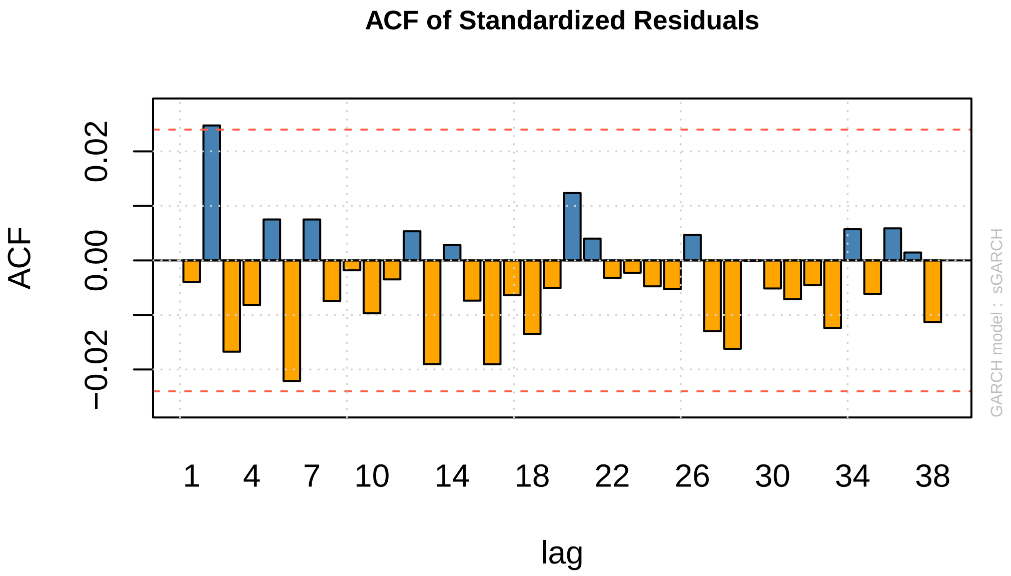
```
	Ljung-Box test

data:  Residuals from Regression with ARIMA(1,1,0) errors
Q* = 1698.4, df = 519, p-value < 2.2e-16

Model df: 1.   Total lags used: 520
```

Interpretation of assumption checking...

GARCH(1,1) With Ghana Climate Data (GARCH-X)

The GARCH-X model extends the standard GARCH framework by including an exogenous regressor. In this case, we incorporate the TAVG variable from the Ghana climate dataset into the mean equation to account for its influence on cocoa futures prices. This model allows for simultaneous modeling of both climate impacts and volatility.

The average daily nationwide temperature was shown to be significantly and negatively correlated with cocoa prices at several lags – therefore including this regressor in a GARCH model allows us to explore its influence while still capturing the dynamic nature of price volatility. This dual-focus structure is well-suited for commodity modeling, and can potentially better represent this dataset than the GARCH itself is able to.



Interpretation of assumption checking…

Model Comparison

To determine which candidate model best represents the dataset here, we perform two layers of model comparison. First, an in-sample comparison is performed with AIC and BIC, to choose between the ARIMA and ARIMA-X models, as well as the GARCH and GARCH-X models. Then, once the best two candidate models have been obtained, the RMSE is computed for each and the final model is selected as the one with the lowest RMSE.

```
    Model         AIC         BIC
1   ARIMA 68205.97700 68219.14109
2 ARIMA-X 68207.40637 68227.15250
3   GARCH    10.30576    10.31193
4 GARCH-X    10.30613    10.31353
```

From the table above, it can be seen that when comparing the ARIMA and ARIMA-X model, the better model is the ARIMA model. similarly, when comparing the GARCH and GARCH-X model, the better model is the GARCH model. This result may be due to the fact that including the extra predictor of average daily temperature does not significantly improve model fit, and therefore we should proceed with the simple ARIMA and GARCH models for comparison.

```
  Model     RMSE
1 ARIMA 1044.964
2 GARCH 1059.909
```

The RMSE comparison shows that the ARIMA(1,1,0) model achieves a slightly lower forecast error (RMSE = 1044.964) compared to the GARCH(1,1) model (RMSE = 1059.909). This indicates that, based on out-of-sample forecast accuracy, the ARIMA model performs marginally better in predicting cocoa futures prices. Although the GARCH model accounts for volatility, its added complexity does not translate into significantly improved forecast performance in this context. Therefore, the ARIMA(1,1,0) model is the better-performing model and will be selected as the final model to move forward with analysis.

Forecasting and Results

Discussion and Conclusion

This study aims to assess the viability of using time series models to predict cocoa futures prices. By incorporating both financial and climate data, we seek to understand the key drivers of cocoa price movements and evaluate the effectiveness of different forecasting approaches. A key limitation of this study is the reliance on historical data, which may not fully capture future economic and geopolitical shocks that affect cocoa prices. Additionally, while climate factors are included, other external factors such as currency exchange rates and global demand fluctuations could further improve predictive accuracy. Future research could explore alternating modeling techniques, such as deep learning architectures, to enhance forecast precision and accuracy.

Understanding cocoa price trends is crucial for stakeholders across the supply chain, from farmers to multinational corporations. This study contributes to the growing body of research on agricultural commodity forecasting by demonstrating the potential of data-driven approaches in predicting future market movements.

Appendix

… include all code and supplementary materials …

… include a references page for literature review and anything else necessary …