Data Science Coursework

Ektor Avlonitis, SN: 23165499, COMP0047, Group: 1

ABSTRACT

This report focuses on forecasting steel and crude oil prices by analyzing daily trade data. It examines how economic and geopolitical influences affect these key commodities and uses advanced statistical techniques to select the best forecasting models. The findings offer insights into the future of commodity prices and the dynamics of international trade networks.

I. INTRODUCTION

The research contributes to a larger group discussion by addressing a pivotal question: which countries are central to global trade? In answering this, the project highlights key influencers in the international market, thus offering a well-rounded view of the strategic economic landscapes that drive commodity prices. It examines the trade dynamics of steel and crude oil, commodities fundamental to global commerce and economic indicators. Then, the study explores the historical trends and of the prices and log returns of these commodities in order to delve into the main objective of this research. Individually, the project zeroes in on predicting the future prices of these vital commodities. The study focuses on daily trade data for steel within the period 2005 to 2020 to uncover trends and evaluate how changes in the global economy influence this commodity. It also analyzes daily crude oil trade data in the same period, paying close attention to how geopolitical events and economic policies affect market dynamics. Through detailed statistical analysis ensuring data stationarity, and the deployment of sophisticated forecasting models like ARIMA for returns and GARCH for volatility, the study provides insightful conclusions on commodity price movements.

II. METHODOLOGY & DATA

A. Data used

The data used in this study is sourced from daily steel and crude oil price figures provided by Yahoo Finance.[1]

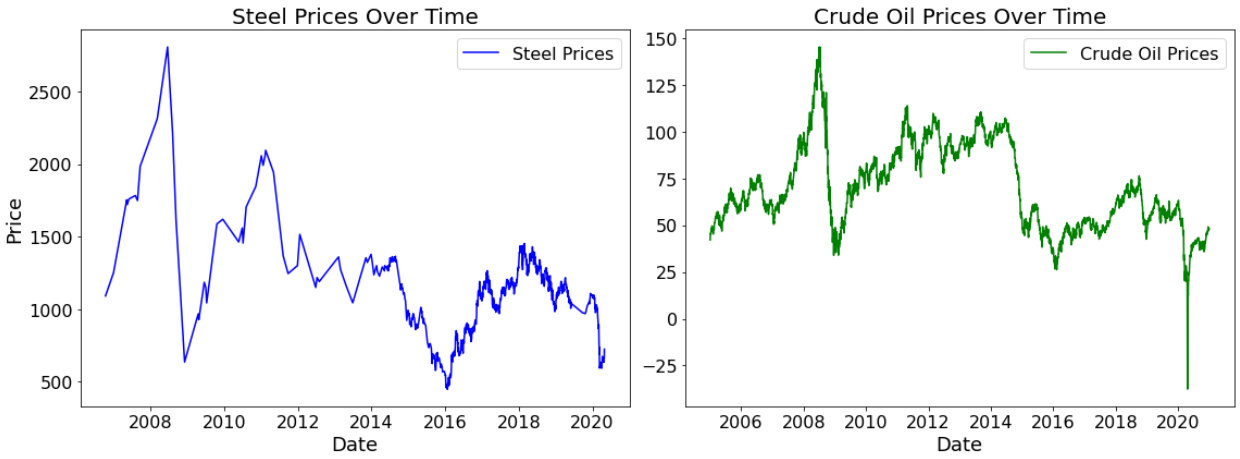

FIG. 1. Steel and Crude Oil Prices Over Time

The plots for steel and crude oil prices show that both markets have experienced significant volatility over the period from 2005 to 2020. Just like steel, crude oil prices have been influenced by global economic shifts, with notable price movements. Peaks and troughs in these commodities often correlate with global demand, supply constraints, and geopolitical tensions that affect market sentiment and trading decisions. A more detailed analysis of steel and crude oil price trends over time will be presented in the *Historical Trends and Log Return Analysis* subsection below.

The Augmented Dickey-Fuller (ADF) test indicate non-stationarity in both steel and crude oil prices, suggesting the importance of analyzing log returns to achieve stationarity for effective time series forecasting. The ADF test results reveal a statistic of -2.814 for steel prices and -2.091 for oil prices, with corresponding p-values of 0.056 and 0.248, which confirm the initial non-stationarity. However, the log returns for both steel and oil display significant negative ADF statistics (-6.429 and -10.981, respectively) with p-values very close to zero, suggesting that the log return series are stationary and suitable for reliable forecasting models. This enables the application of time series forecasting models to deliver

more reliable predictions of future steel and crude oil price trends.

B. Trade dynamics

Firstly, the trade dynamics of steel and oil were examined. The analysis utilized as metrics the Degree of Countries, Degree Centrality, Betweenness Centrality, and Closeness Centrality to identify the countries with the most significant roles in the global marketplace.

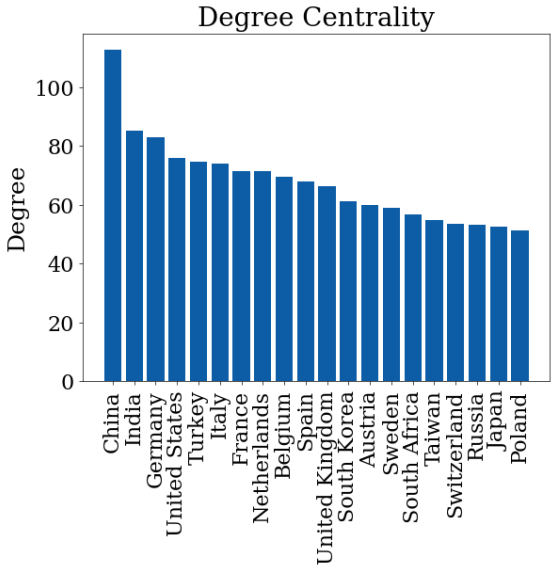

FIG. 2. Degree Centrality

As observed in Figure 2, China leads significantly in degree centrality, indicating it has the highest number of direct trade connections with other countries. This suggests China's crucial role in global trade networks, followed by India, Germany, and the United States. The high degree centrality of these countries reflects their extensive involvement in global trade, acting as major centers for the import and export of goods.

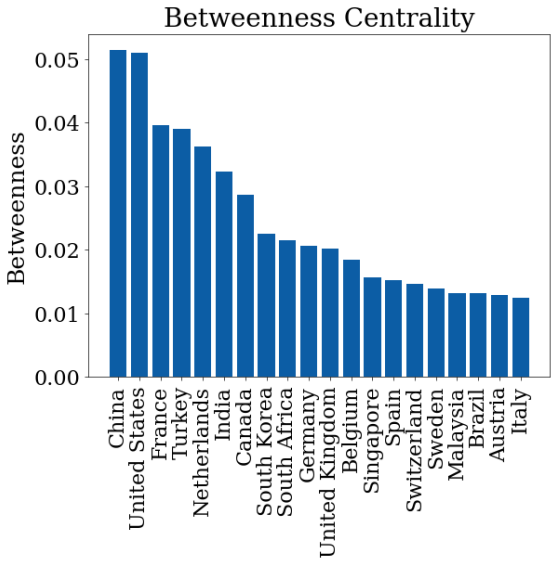

FIG. 3. Betweenness Centrality

In Figure 3, China again tops the chart, highlighting its role also as a key intermediary in global trade routes. This means that trade flows between many countries are likely to pass through China. The United States, France, and India also show high betweenness centrality, meaning they have strategically positioned within the trade network.

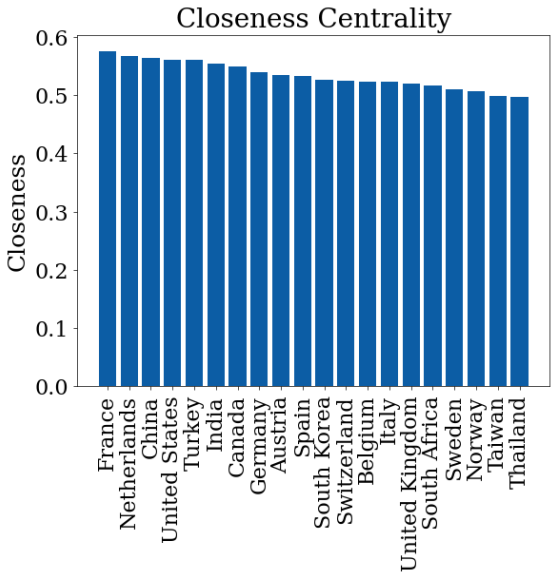

FIG. 4. Closeness Centrality

France leads in closeness centrality as seen in Figure 4, followed by the Netherlands,

China, and the United States. High closeness centrality indicates these countries can quickly interact with any other country in the network, suggesting efficient access to global markets.

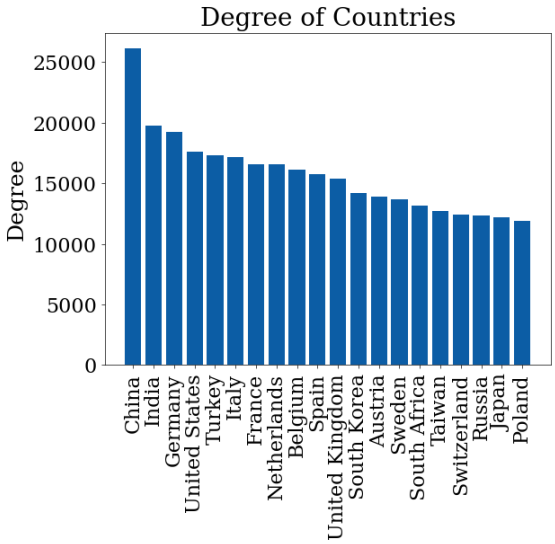

FIG. 5. Degree of Countries

Figure 5 showcases the "Degree of Countries" and shows China's unmatched participation into the global trade network, with a significantly higher number of direct trade connections than any other nation, exhibiting its status as a global trading powerhouse. Following China are India, Germany, and the United States, all of which play vital roles in the international exchange of goods, especially in the steel and oil sectors.

C. Historical Trends and Log Return Analysis

The trajectory of steel prices over time has been formed by a combination of supply and demand dynamics, technological progress, and global economic policies. Economic expansion and industrial growth typically lead to increased demand for steel, resulting in higher prices, as observed in the mid-2000s when emerging economies, particularly in Asia, invested heavily in infrastructure and construction, thus driving up steel demand significantly [2]. Thus, the global financial crisis of 2008 led to a sharp downturn in industrial activity and construction, as seen in Figure 1, causing steel prices to plummet as demand contracted rapidly [2].

In the period from around 2016 to 2020, steel prices displayed a more stable yet declining trend. This period likely reflects market responses to trade tensions and policy changes,

such as the United States' introduction of Section 232 tariffs, which induced adjustments in global steel trade flows [2]. Additionally, the industry's transition toward green steel, along with sustainability initiatives, began to impact the market. The push for decarbonization in the steel sector, along with technological advancements in production, hinted at a potential restructuring within the market that could influence prices across various regions depending on their pace of decarbonization and energy costs [3].

Towards the latter part of this period, a noticeable downward trend in steel prices can be observed, which could be attributed to several factors, including the European Union's introduction of the Carbon Border Adjustment Mechanism (CBAM) and other environmental and trade policies, indicating a shift in the industry toward cleaner production methods that could potentially impact global supply and demand dynamics [4]. The overall trend suggests that the steel market was contending with these emerging forces while also managing the typical economic cycles that influence commodity prices [2].

Crude oil prices have also undergone significant fluctuations influenced by various geopolitical, economic, and sector-specific factors. The early 2000s saw a sudden rise in crude oil prices, driven by heightened global demand, geopolitical tensions, and supply constraints. The following decline in prices post-2008 reflected the economic downturn, which drastically reduced demand for energy and affected commodity prices across the board.

Crude oil prices have seen dramatic changes, often due to OPEC's output choices, new energy policies, and the rise of shale oil production, especially in the USA [5]. These factors have made oil prices fluctuate considerably, as the Figure 1 shows. Similar to steel, the oil industry is gradually focusing more on renewable energy, which might affect future oil prices as global energy priorities shift towards sustainability [6].

The figure below (Figure 6), showing daily log returns of steel prices and crude oil prices over time, reflects series that closely resemble random walks, characterized by its apparent lack of a pattern or trend.

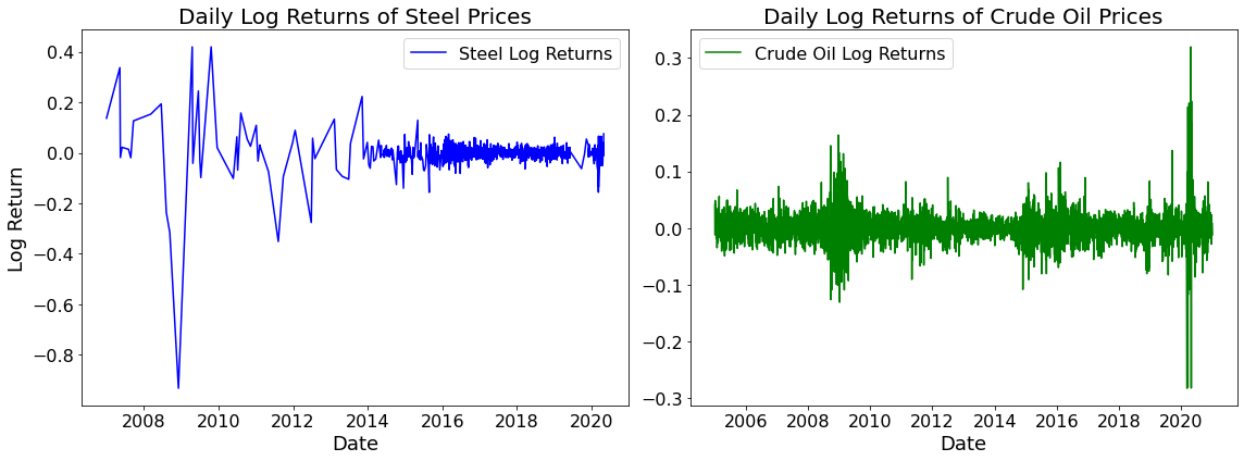

FIG. 6. Daily Log Returns of Steel and Crude Oil Prices

This fluctuating nature of the returns, rapidly changing from one period to the next, indicates the difficulty in forecasting such a series. Therefore, the project adopts a strategy of aggregation, focusing on monthly log returns (Figure 7). This approach aims to reduce noise and capture more substantial, long-term movements.

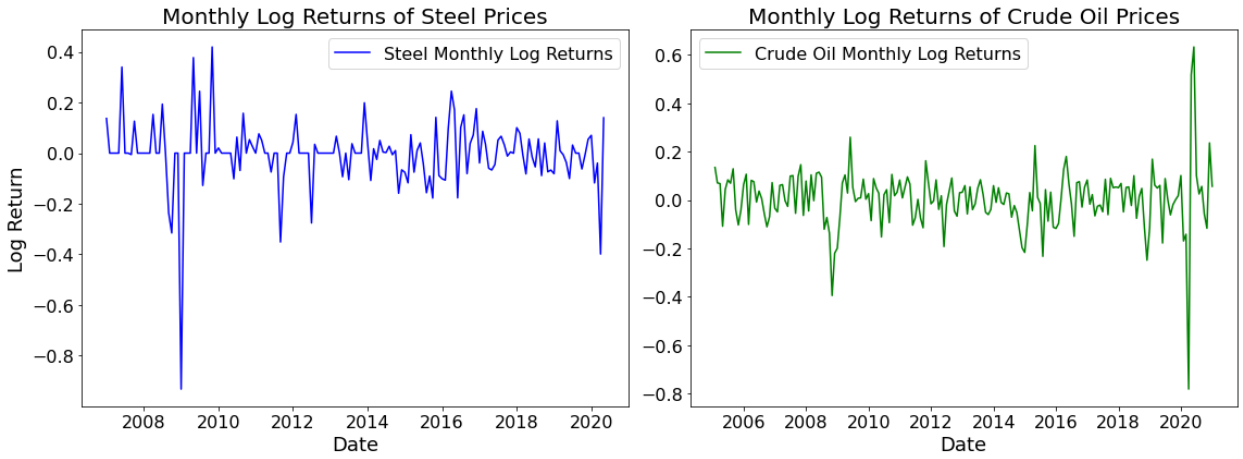

FIG. 7. Monthly Log Returns of Steel and Crude Oil Prices

D. Distribution Analysis

In the report's forecasting section, the first task was to understand the monthly returns distribution by studying the monthly log returns of steel and crude oil prices. An initial statistical analysis of the log returns was conducted to understand the fundamental properties

of the data:

TABLE I. Statistical Analysis of Log Returns

Statistic	Steel	Crude Oil
Mean	-0.0025	0.0038
Standard Deviation	0.1290	0.1239
Skewness	-2.1286	-0.5572
Kurtosis	19.7648	14.8599

Three theoretical distributions, the normal, log-normal, and Student's t, were then fitted to the log returns data and displayed in Figure 15. These distributions were chosen to model log returns due to their suitability for capturing key financial data characteristics: symmetry and mild deviations in the normal, positivity in log-normal, and heavy tails in Student's t.

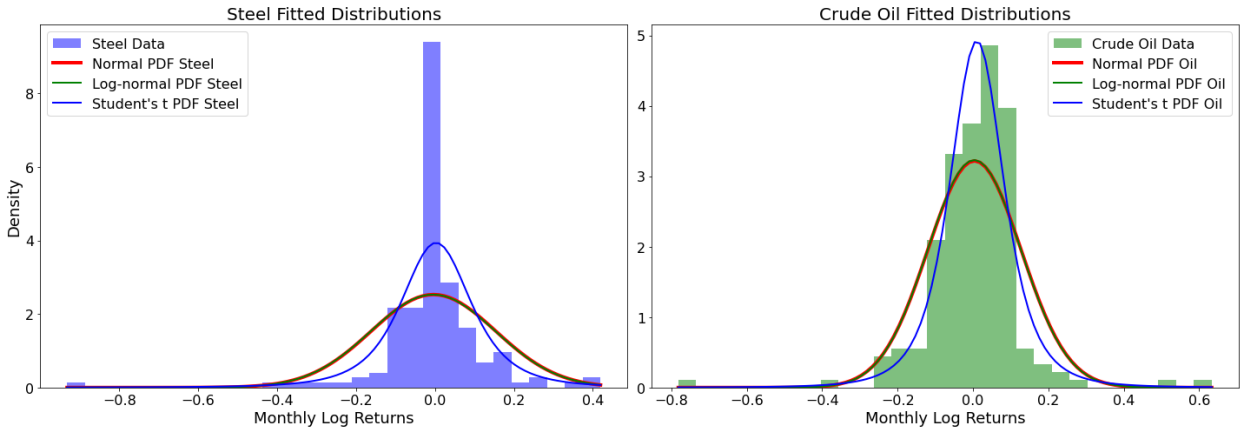

FIG. 8. Log returns and fitted distributions for Steel and Crude Oil

Following this, QQ plots for each theoretical distribution were compared, providing a visual analysis of how each distribution aligns with the empirical distribution of log returns.

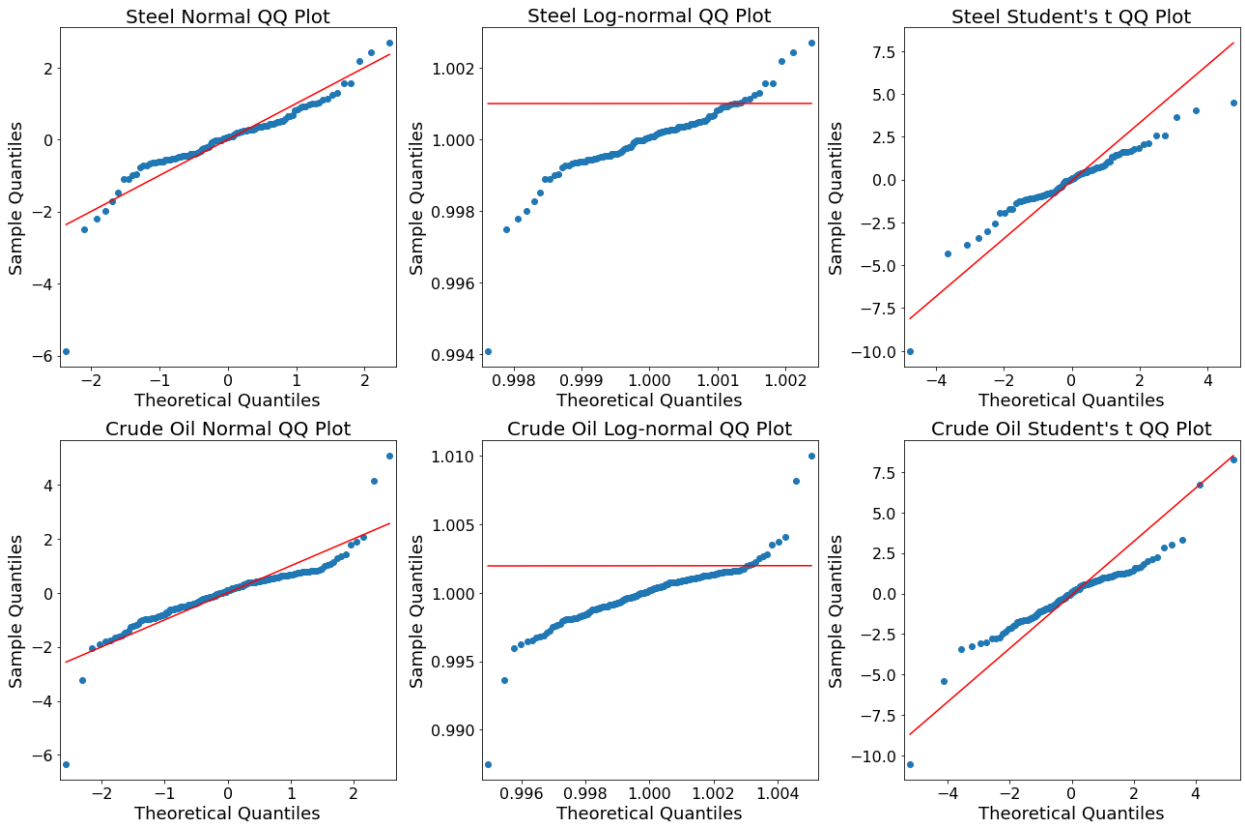

FIG. 9. Different distributions Q-Q plots for Steel and Crude Oil

The next step was to use Kolmogorov-Smirnov (K-S) statistics to compare theoretical and empirical log return distributions, and focus on the 95% Value at Risk (VaR) and Expected Shortfall (ES) for extreme market risk assessment.

TABLE II. Kolmogorov-Smirnov Test and Risk Metrics

Statistic	Steel	Crude Oil
Normal K-S Statistic	0.2112	0.1301
Normal p-value	7.344×10^{-7}	0.0027
Log-normal K-S Statistic	0.2135	0.1287
Log-normal p-value	5.304×10^{-7}	0.0031
Student's t K-S Statistic	0.1821	0.0678
Student's t p-value	3.359×10^{-5}	0.3264
95% VaR	-0.2127	-0.1730
95% ES	-0.4183	-0.2861

The examination of monthly log returns for steel and crude oil prices revealed significant market dynamics, highlighting heavy tails and data asymmetry. The statistics, mean close to zero, notable standard deviation, negative skewness, and high kurtosis, indicate distributions with frequent large deviations from the mean for both commodities. Kolmogorov-Smirnov (K-S) tests and QQ plots indicate that, unlike normal and log-normal distributions, the Student's t distribution closely matches the empirical data for monthly log returns, particularly for crude oil with a p-value over 0.05, suggesting a statistically significant fit. This highlights the importance of using models that can effectively handle the irregularities and extreme values common in financial data. Tail risk assessments through Value at Risk (VaR) and Expected Shortfall (ES) for both commodities suggest that forecasting models accounting for the heavy-tailed nature of distributions could provide more accurate risk evaluations. This indicated that employing forecasting models that account for the heavy-tailed nature of the distribution would likely yield more accurate risk assessments. This suggests that it would be beneficial to employ forecasting models that can better handle these "irregularities" of the data.

III. RESULTS & DISCUSSION

A. Forecasting Analysis

1. Forecasting model for monthly log returns

In the analysis of monthly log returns, the data was decomposed into its components, trend, seasonality, and residuals, using the seasonal decomposition technique over a specified period to capture annual seasonality. Each component was forecasted separately to understand distinct patterns and behaviors within the data. The trend component captures long-term movements, the seasonal component captures recurring patterns within a year, and the residual component accounts for irregularities not explained by trend or seasonality. The forecasts from these individual components were then aggregated to construct the complete time series forecast. Below Figure 10 shows the decomposed components over time for both steel and crude oil prices.

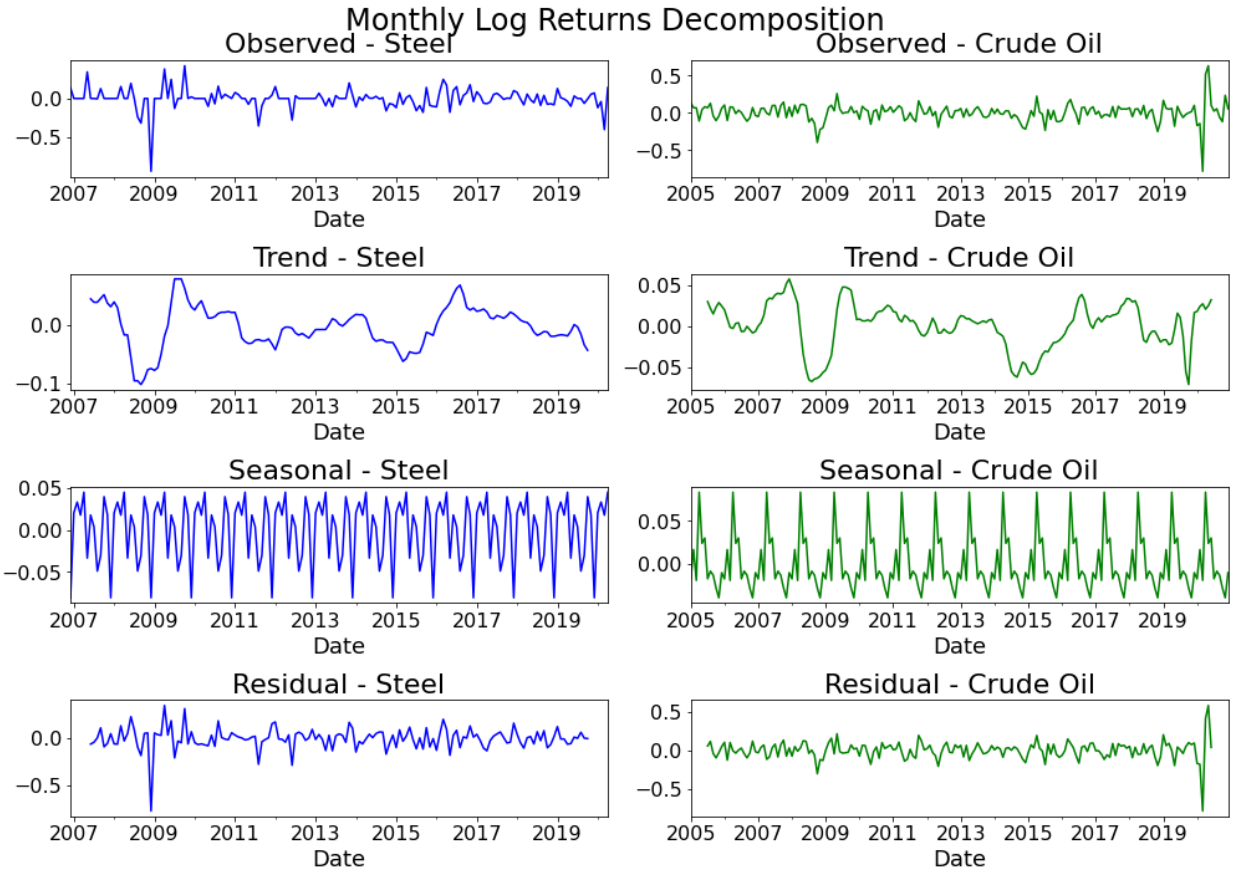

FIG. 10. Monthly Log Returns of Steel and Crude Oil - Decomposition

For the trend component of the monthly log returns, the project employed four different predictive models to capture the underlying patterns in the data as shown in Table III. The training set was 70% of the available data and the remaining 30% served as a test set to evaluate their performance.

TABLE III. Forecasting Models for Trend Component

Model	**Prediction Capability**
ARIMA	Captures autocorrelations
Linear Regression	Models linear trends
ETS	Smoothens trends and adjusts for seasonality
GBM	Handles complex, non-linear patterns

The Autoregressive Integrated Moving Average (ARIMA) model selection was facilitated by the "auto_arima" function. For the first time series regarding steel prices, this function

determined the optimal fit with parameters: p = 2, d = 0, q = 3, based on minimizing the Akaike Information Criterion (AIC), which resulted in an AIC value of -1016.57. Similarly, for the crude oil time series, the best fitting ARIMA model identified by "auto_arima" had parameters: p = 2, d = 0, q = 2, optimizing for the lowest AIC, which yielded an AIC of -1290.66. The Exponential Smoothing (ETS) model utilized an 'additive' trend without seasonality, ideal for data with a more consistent trend over time. The GBM model implemented had 100 estimators, a learning rate of 0.1, and a tree depth of 3. It's suited for capturing complex patterns in the data by building multiple decision trees.

TABLE IV. MSE Comparison of Forecasting Models for Trend Component for Steel and Crude Oil

Model	Steel - MSE	Oil - MSE
ARIMA	0.0005	0.0005
Linear Regression	0.0015	0.0017
ETS	0.1052	0.0188
Gradient Boosting Model	0.0008	0.0007

Based on Table IV comparing MSE values, the ARIMA model, which gave the lowest MSE for both time series, was chosen for its superior forecasting accuracy of the trend component.

For the seasonal component, the analysis revealed that the seasonal pattern remained consistent and unchanged over the observed period. This consistency means that forecasting for the seasonal component might not be necessary, as its future behavior could be accurately anticipated based on its past patterns. Therefore, future values are predicted with confidence without the need for complex forecasting models.

For the residual component of the monthly log returns, the project employed three different predictive models to capture the underlying patterns in the data as shown in Table V. Again the set was split for train and test the same way.

TABLE V. Forecasting Models for Residual Component

Model	**Prediction Capability**
ARIMA	Captures residual autocorrelations
Linear Regression	Models potential linear patterns in residuals
Random Forest	Handles complex, non-linear relationships in residuals

In the analysis of the residuals component, the Random Forest model was included for its ability to capture complex, non-linear relationships. When the ARIMA model was applied to the steel price time series, it identified the optimal parameters as p = 0, d = 0, and q = 0, achieving the lowest Akaike Information Criterion (AIC) value compared to other configurations. This (0,0,0) parameter set indicates that the model does not incorporate autoregressive terms, differencing, or moving average terms, suggesting that the best fit, according to the AIC, is essentially a model that predicts using the series' mean. This finding implies that, for the residuals, employing a model that uses the overall mean for predictions is statistically more favorable. For the crude oil time series, on the other hand, the ARIMA model's optimal parameters were found to be p = 0, d = 0, q = 1.

TABLE VI. MSE Comparison for Residual Component Forecasting Models

Model	Steel MSE	Oil MSE
ARIMA	0.0045	0.0268
Linear Regression	0.0045	0.0269
Random Forest	0.0260	0.0278

Based on Table VI which compares MSE values, the ARIMA model was selected for its superior forecasting accuracy of the residuals, as it yielded the lowest MSE for both time series. It's worth noting that the Linear Regression model performed quite closely, but ultimately, ARIMA was chosen for its slight edge in forecasting precision.

After aggregating the forecasts for the trend, seasonal, and residual components using the best performing models, a comprehensive forecast for the monthly log returns was created. The comparison between the actual data and the combined forecast, is represented in the plot in Figure 11, which showcases the model's effectiveness. The accuracy of the combined forecast for the steel price time series was evaluated using three metrics: it achieved a Mean

Squared Error (MSE) of 0.0129, a Root Mean Squared Error (RMSE) of 0.1136, and a Mean Absolute Error (MAE) of 0.0836. In the case of crude oil, the combined forecast attained an MSE of 0.0306, an RMSE of 0.175, and an MAE of 0.1065, reflecting the precision of the model in capturing the movements of the monthly log returns.

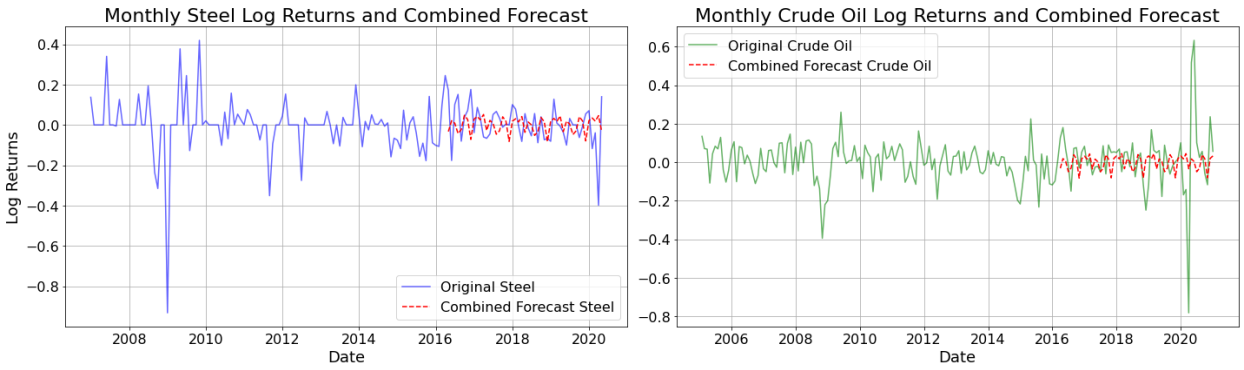

FIG. 11. Monthly Log Returns and Combined Forecast for Steel and Crude Oil

From the graph, it is observed that for both commodities the original monthly log returns are tracking closely with the combined forecast, particularly in capturing the central trend. While there are discrepancies, particularly with larger spikes in the actual data, the forecast follows the general pattern well.

Next, causality was examined to determine if it could offer additional insights into forecasting accuracy.

2. Causality

Following the analysis of monthly log returns for steel and crude oil prices, the study also explored the causal relationships between these two commodities. The investigation was primarily focused on determining if changes in crude oil prices could predict variations in steel prices and vice versa. The Granger causality test results revealed significant p-values at various lags, indicating a statistically significant predictive power of oil prices on steel prices. This statistical significance suggests that historical values of oil prices contain useful information that could potentially forecast future movements in steel prices.

TABLE VII. Summary of Granger Causality Test Results

Lag	**P-value**
1	0.0043
2	9.18×10^{-6}
...	...
12	0.0034

Incorporating this information regarding causality, an ARIMAX model was developed by including the lagged oil prices as an exogenous variable in the forecasting models for the trend and residual components of steel prices. Despite the statistical evidence from the Granger causality tests, the incorporation of lagged oil prices as an exogenous variable in the ARIMAX models did not result in an improvement in forecasting accuracy, as measured by the Mean Squared Error (MSE), compared to the optimum ARIMA model without exogenous variables. Thus, the addition of oil price data did not significantly enhance the model's predictive capability for steel prices. This outcome led to the use of the ARIMA model as the optimal choice for forecasting the trend and residual components of monthly log returns for steel prices. Considering the complexities and unpredictable movements in returns, the next step was to focus towards predicting volatility, which should provide a clearer prediction and understanding of market risk and price variability over time.

3. Forecasting model for volatility

Distribution analysis of steel and crude oil price returns revealed data irregularities, such as asymmetry in the distribution of returns and the presence of fat tails, suggesting that extreme observations are more common. To address these complexities and improve forecast accuracy, the GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model was employed. The GARCH model was particularly suited for this task due to its capacity to model volatility as a function of past errors and conditional variances, thus capturing the persistence of shocks and the clustering of volatility observed.

The analysis was conducted utilizing the derived log returns. To identify the optimal representation of the observed volatility patterns, various GARCH models were considered. The best model was the one with the lowest AIC (Akaike Information Criterion) and

BIC (Bayesian Information Criterion). For the steel price time series the best choice was the GARCH(4,4) model (p = 4 is is the order of the GARCH terms, indicating the number of lagged variance terms included in the model and q =4 is the order of the ARCH terms, representing the number of lagged squared-error terms). However, recognizing that AIC and BIC values alone do not mean comprehensive model evaluation, and thus the GARCH(4,4) model's performance was further examined by comparing it with the simple GARCH(1,1) model. Also, out-of-sample forecasting revealed that the GARCH(4,4) model achieved a marginally lower Mean Squared Error (MSE) of 0.00026, compared to the simpler GARCH(1,1) model's MSE of 0.00031. For the crude oil time series the GARCH(2,5) model gave the best results with an MSE of 0.0012 which is identical to the simpler model (1,1) with an MSE of 0.00112 too.

In the analysis of volatility forecasting models, ARIMA models for both commodities were also put to test alongside the GARCH model in order to compare performances. For the steel commodity, the best ARIMA model, based on the lowest AIC metric, had as optimal parameters p = 2, d = 2 and q = 0. Despite the ARIMA model's utility in various forecasting scenarios, it exhibited a higher Mean Squared Error (MSE) of 0.00045, compared to the GARCH (4,4) model's MSE of 0.00026. This result underlines the GARCH model's superior capability in capturing and predicting the dynamic nature of volatility. For the crude oil commodity, in the analysis of volatility forecasting models, the ARIMA(2,1,2) was also tested alongside the GARCH(2,5), with an MSE of 0.0018.

The efficacy of the GARCH(4,4) model for steel and GARCH(2,5) model for crude oil are clearly demonstrated below in Figure 12. The figure displays the out-of-sample (20% test set) volatility forecast using the two GARCH model in comparison to the actual volatility values. The model's forecasts closely mirror the actual volatility trends.

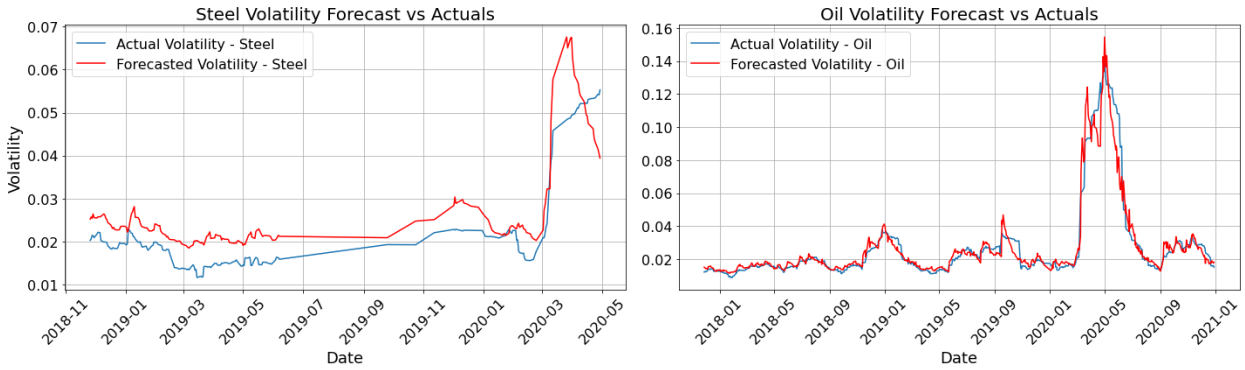

FIG. 12. GARCH Model Volatility Forecast vs Actuals

Next, standardized residuals were examined for the GARCH(4,4) model for steel and for the GARCH(2,5) model for crude oil and revealed no systematic patterns, indicating that they effectively modeled the data's volatility (Figure 13).

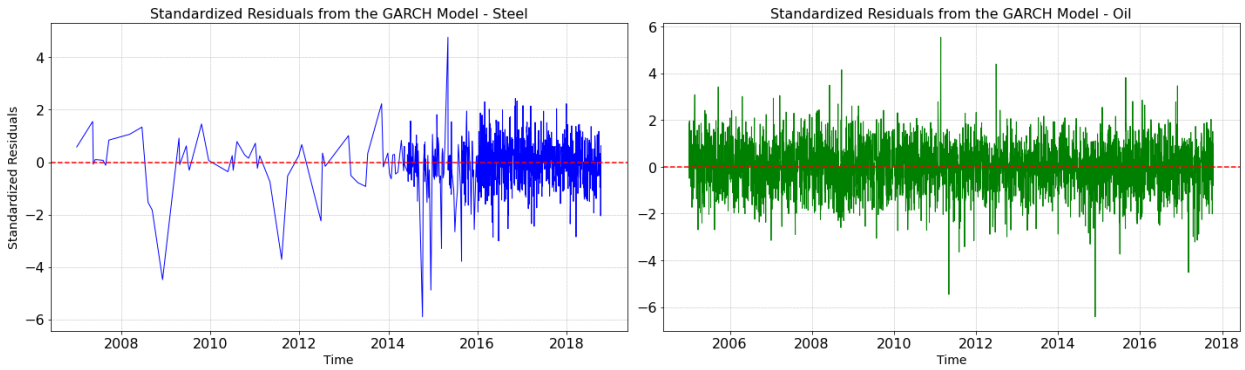

FIG. 13. Standardized Residuals from the GARCH Model

The ACF plots (Figure 14) for squared standardized residuals showed no significant autocorrelations, showing that the models captured volatility clustering successfully. This was confirmed by high p-values from the Ljung-Box test, which verified the lack of autocorrelation in the residuals at lag 10.

Lag	**Steel**		**Crude Oil**	
	Ljung-Box Statistic	**P-value**	**Ljung-Box Statistic**	**P-value**
10	4.178	0.939	8.355	0.594

TABLE VIII. Ljung-Box Test Results for Steel and Crude Oil

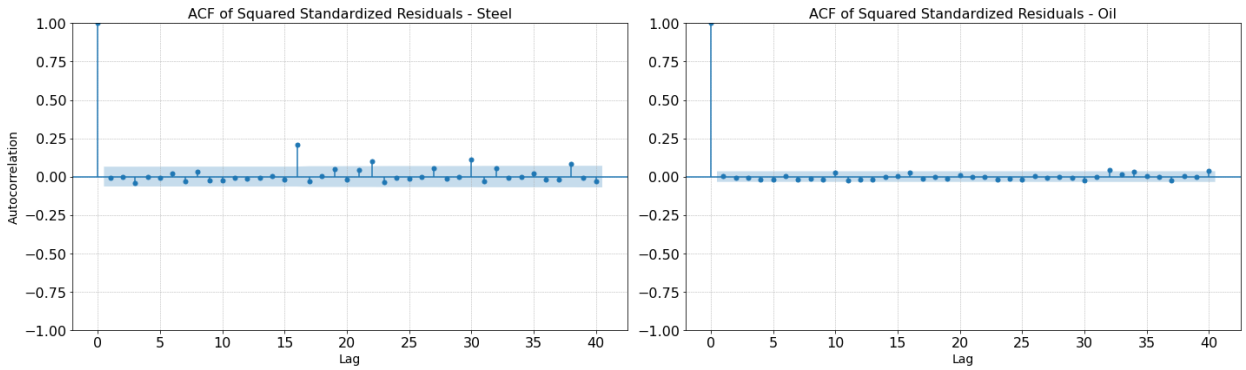

FIG. 14. ACF of Squared Standardized Residuals

Following these tests, by plotting the conditional volatility it was observed that the GARCH models efficiently captured periods of high and low volatility, demonstrating their capability to model volatility over time accurately.

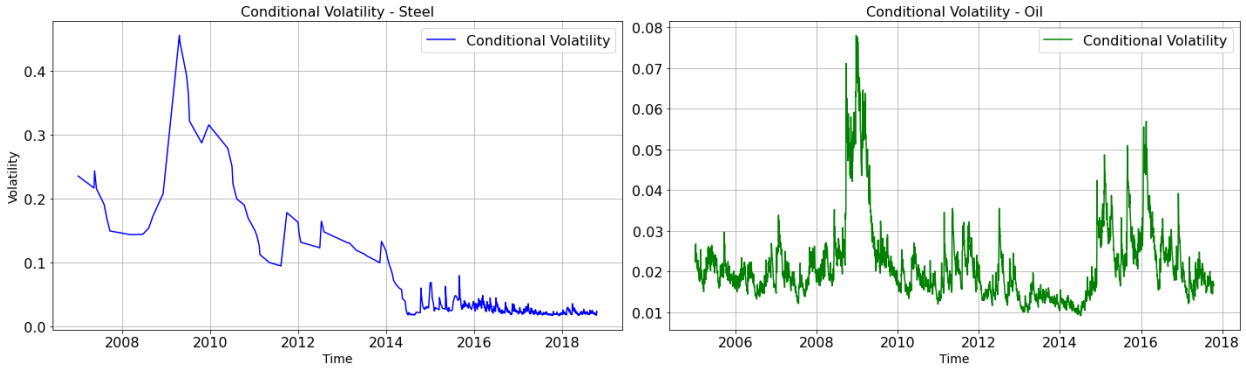

FIG. 15. Conditional Volatility from GARCH Model

IV. CONCLUSION & OUTLOOK

The analysis demonstrated that the GARCH(4,4) model adeptly captured the volatility patterns in steel prices, while the GARCH(2,5) model performed equivalently well for crude oil prices. These conclusions were supported by minimal autocorrelation in the residuals, as confirmed by the Ljung-Box test results, and the models' capability to reflect market volatility accurately in their conditional volatility forecasts. The results showed the GARCH models' proficiency in modeling volatility and responding well to market dynamics for both commodities. Forecasting volatility is more reliable than predicting returns, benefits from

the ability of the GARCH models to handle irregularities, particularly in the context of commodities like steel and crude oil, where market conditions can rapidly change.

Therefore, for forecasting purposes, particularly concerning the complex behavior of commodity prices, the GARCH models were optimal choices. Nevertheless, it is crucial to take into account the economic factors that may impact the performance of these models. As financial markets grow more complex, incorporating economic reasoning into the model selection process is crucial for achieving forecasts that are not just statistically accurate, but also economically intuitive.

---

[1] Yahoo Finance, Daily steel and crude oil prices (2023).

[2] McKinsey & Company, The resilience of steel: Navigating the crossroads (2022).

[3] McKinsey & Company, Safeguarding green steel in europe: Facing the natural-gas challenge (2022).

[4] European Commission, Carbon border adjustment mechanism: Questions and answers (2021).

[5] U. E. I. Administration, Short-term energy outlook (2020), accessed: 2024-04-03.

[6] I. E. Agency, World energy outlook (2020), accessed: 2024-04-03.