US Gasoline Prices: Linear Regression Model and ARIMA Model Forecast

Xinyu Wang*

IDepartment of Engineering, University of Toronto, Toronto, Canada

* Corresponding Author Email: thea.wang@mail.utoronto.ca

Abstract. The price of gasoline has always been a concern in consumer's daily lives. As the price increases, consumers spend more on driving. According to the 2022 BP Statistical Review of World Energy, the United States contributed more to the growth of the oil and gasoline consumption in 2021. The study focuses on analyzing past US monthly gasoline price changes and forecasting the prices for the next year. The combination of the linear regression model and Auto Regressive Integrated Moving Average (ARIMA) model is performed to predict the trend and display the forecasted price ranges. The study reveals that there will be an increasing trend for US monthly gasoline prices, starting from $0.964 in August 2023 to $1.108 in July 2024. The lower prices and higher prices from the interval show the same trend of moving upwards. For future research, crude oil demand and supply, and prices are considered as important factors in predicting the gasoline prices because gasoline is a refined product from crude oil.

Keywords: Linear Regression model, ARIMA model, US gasoline prices, Forecast.

1. Introduction

Gas is a crucial component of transportation. It has an influence on the households as they drive and the logistic enterprises. The rises or drops in gasoline prices are certainly noticed by consumers. The attention for the rise often sparks arguments and blame because it is related to the cost. Gasoline also varies by grade. The three main grades are regular, midgrade and premium. Over the past few years, the gasoline price has experienced wild swings as the oil market fluctuates. Gasoline is refined from the crude oil and the petroleum refining industry in the United States is the largest and most advanced in the world in 2019 [1]. With oil demand and crude prices falling precipitously due to the Covid-19 pandemic in 2020, gasoline prices also experienced the drop. In 2021, the temporary reduction was reversed and most of the oil consumption growth came from gasoline and diesel taking place in the US [2]. An interesting research found out that after the COVID-19 pandemic has been controlled, a unit increase in global search interest of the pandemic caused a reduction in New York Harbor Conventional Gasoline Regular spot price by 0.104% after one day [3]. However, the recent event of Russia-Ukraine crisis has affected the oil supply chain and pushed the oil and gasoline prices even higher [4]. Gasoline price is volatile and higher prices may also impact the economy through consumer spending to the automobile industry [5]. According to Trading Economics and their macro models, the United States gasoline price is expected to trend around $1.06 USD/Liter in 2024 [6].

The study uses the different models to forecast the price. It first analyzes the past changes in US gasoline prices and applies the linear regression models for the two periods in the last 9 years. Then, based on the initial analysis and the second period's data, the ARIMA model for the residuals is elaborated. Forecast results and the discussion are presented at the end. The study aims to provide a reference on potential changes of US gasoline prices in the monthly basis to the relevant households and businesses.

2. Methodology

2.1. Data

The study uses monthly end-user total prices for gasoline in the United States from the IEA (International Energy Agency) Energy Prices database [7]. United States is a member country in the

agency and the database sources are from US Energy Information Administration and the US Bureau of Labour Statistics. The time series data extracted is from January 2015 to July 2023 to see overall changes in the US gasoline market within the 9-year period and the forecasting in this study considers recent data starting from January 2019. The unit for gasoline prices is US dollars per liter.

2.1.1 Exploratory analysis

The time series plot for US monthly gasoline prices is illustrated in Fig. 1. There is a blue vertical line in the plot to separate the time periods before the year 2019 (48 observations) and after the year 2019 (55 observations).

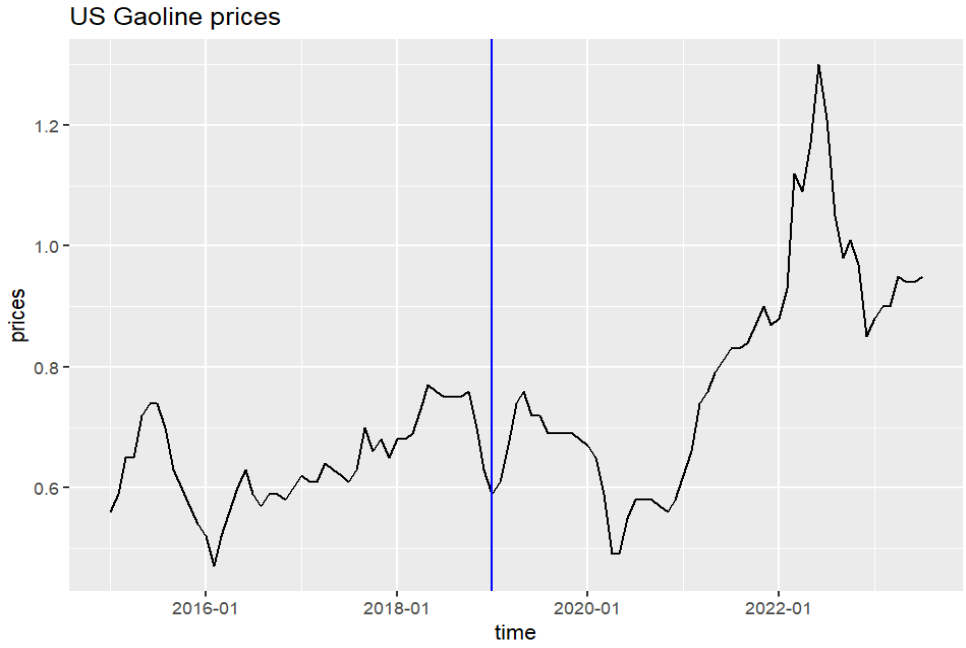

Fig. 1 Plot for US monthly gasoline prices from January 2015 to July 2023

It can be seen from Fig. 1, for the prices before the year 2019, the figure shows no clear seasonality and strong trend, and it fluctuates around 0.65 with the maximum price of 0.77 achieved in May 2018. However, there is a clear upward movement for the prices after 2019. To be specific, the gasoline prices rose by around four-fifths from \$0.49 in May 2020 to \$0.88 in January 2022 and increased 150% to \$1.21 in July 2022, as the maximum price in the whole period. The COVID-19 pandemic and middle east conflict are important factors in influencing the demand and supply for the oil market and causing the jumps or drops in gasoline prices [3, 8].

The study further estimates the linear regression model of gasoline prices for two periods to provide a better understanding of the changes. The model is written as:

$$price_t = \alpha + \beta * t + e_t \quad (1)$$

The slope term beta can capture and trend and the rising speed.

The results are displayed in Table 1. The monthly price change rate is 0.0026 before 2019 and a larger number of 0.0087 after 2019. The two positive coefficients are both significant to the model as two p-values are below 0.05, indicating the increasing trend for both periods. However, the linear model does not fit the pre-2019 data well as it only captures 25% of the price variations. The second model for data after the year 2019 captures a considerable part of the price change.

Table 1. Linear regression models

Period	Model
Before 2019	$price_t = 0.57844 + 0.002598 * t + e_t$
After 2019	$price_t = 0.55 + 0.008721 * t + e_t$

2.2. Model Selection

To forecast the future US gasoline prices, the study uses the monthly prices from January 2019 to July 2023 for selecting the model. It has been discovered that there is a clear growing trend, and the linear regression model is able to explain 54.2% of the variation within prices. The residual plot is displayed in Fig. 2.

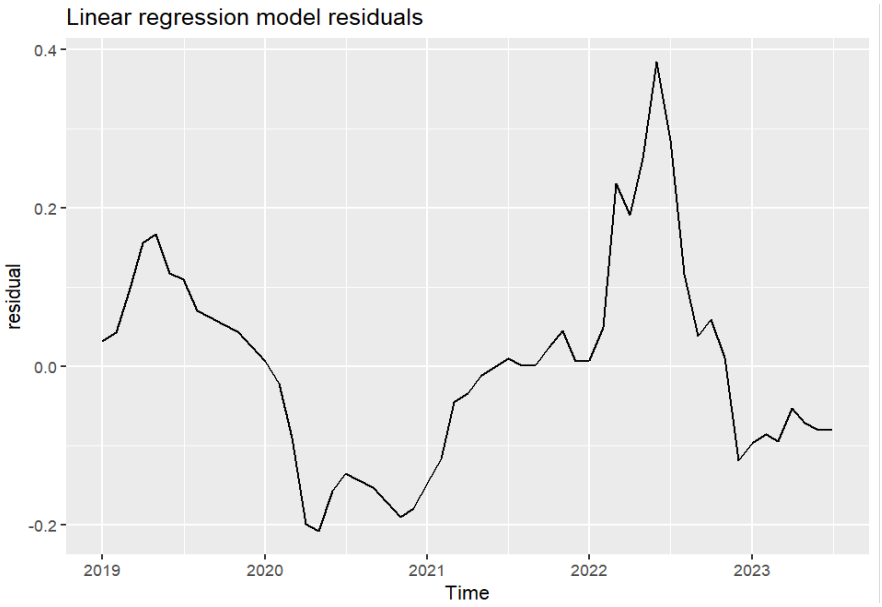

Fig. 2 Residual plot for the linear model

The Fig. 2 shows no apparent trend or seasonality, but some cyclic behavior with a period of about 1-2 years. Through Breusch-Godfrey Test, the correlation between the residuals is assessed and the result of p-values less than 0.05 indicates there exists autocorrelation at some order less than 3. Therefore, the study considers information left in the residuals in the modelling process. ARIMA Model is applied to model and forecast the errors. ARIMA stands for Autoregressive Integrated Moving Average. It makes use of lagged values, differencing and moving averages to smooth the time series data, as referring to three parameters (p, d, q) in the model [9]. Before selecting the parameters, the variance stability is checked through boxcox() function in R. However, it is noticeable that there are negative residual values, and the function could be used for only positive values. A positive value of 1 is added to all the residuals to form a positive series because the maximum negative value does not exceed -1. The study will reverse this transformation at the end of the forecast. The variance stability of the new data series is reassessed. Fig. 3 shows the 95% confidence interval does not contain lambda value of 1 and the transformation was needed to stabilize the data variance.

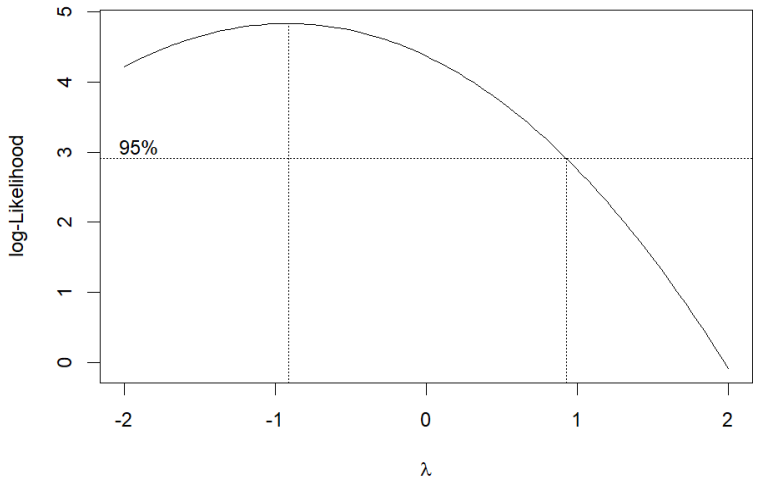

Fig. 3 Result plot from the function “boxcox()” in R

2.2.1 Parameter Selection

The selection of parameters is based on the new data series of all positive values. Firstly, a KPSS unit root test was used to test whether the time series is stationary and determine the differencing values. The test statistic was within the 10 percent significant level so there was no need to do the differencing and the parameter d value would be 0.

2.2.2 Train and Test

Secondly, the left two parameters p and q are determined in the train and test process. The training set comprises 90% of gasoline price data, and the accuracy is evaluated by root mean square error and mean absolute error for the left 10% of test set data. The function "auto. arima()" in R is employed to choose model parameters. The arguments of "stepwise=false" and "approximation=false" inside the function explore a wider variety of possible parameter combinations to minimize the AIC value. The argument of "lambda=auto" considers the Box-Cox transformation. The outcome displays the model ARIMA (1,0,1) with a lambda value of -0.8999268, AIC value of -153.28 and BIC value of -147.61 for the training set. The Ljung-Box test is applied to test if residuals have autocorrelation or not. The result shows the values are within the limits in the ACF diagram and p-value is larger than 0.05, indicating the model is suitable for forecasting (Figure 4).

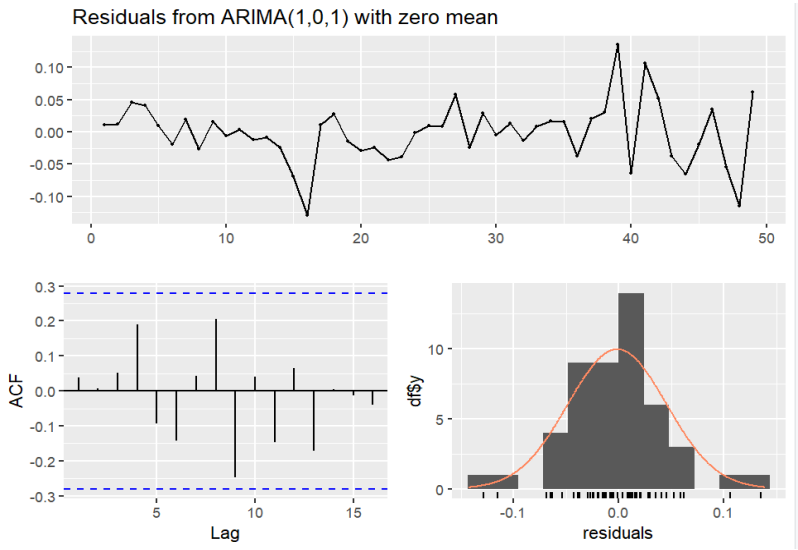

Fig. 4 Ljung-Box test results for the ARIMA model

Then the model is used to forecast the length of the test set. The forecasted values in blue are slightly higher than the actual values in red as shown in Fig. 5. The computed root mean square error is 0.0390 and the mean absolute error is 0.0365.

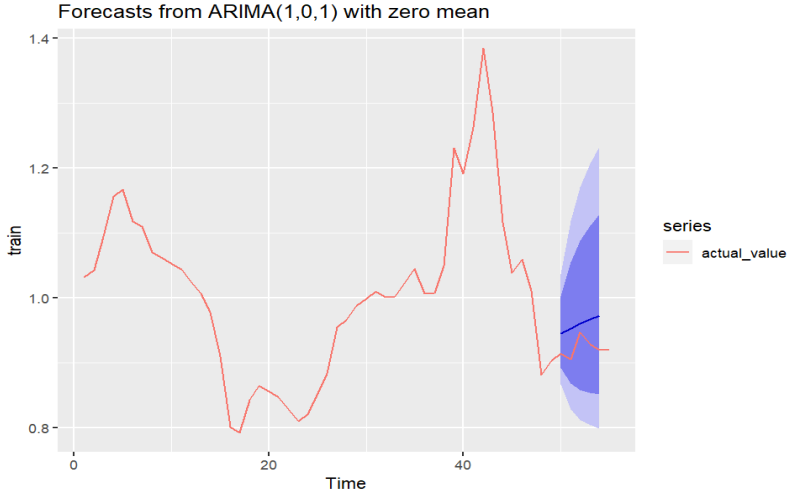

Fig. 5 Plot for the transformed residuals and forecasted data of test set

Finally, the ARIMA model parameter values are 1,0 and 1 separately for p,d and q. The transformation lambda value is -0.8999268.

3. Forecast Gasoline Prices

The study predicts US monthly gasoline prices for the next one year using a combination of the linear regression model and ARIMA model. There are 55 observations from January 2019 to July 2023, so the starting forecasting time point is 56. Firstly, the linear model can be written as:

$$price_t = 0.57844 + 0.002598 * t \quad (2)$$

The forecasting results from the linear model are shown in Table 2. Secondly, the ARIMA model for residuals is defined by the function "Arima ()" with arguments of residual series, the parameters and the transformation lambda value. The outcomes are in the third column of Table 2. Figure 6 displays the next year's forecasts and prediction intervals in blue. There is an increasing trend with possibilities of both upward and downward trends indicating by intervals, however, the range for the upper 95% prediction intervals is larger than that for the lower 95% prediction intervals.

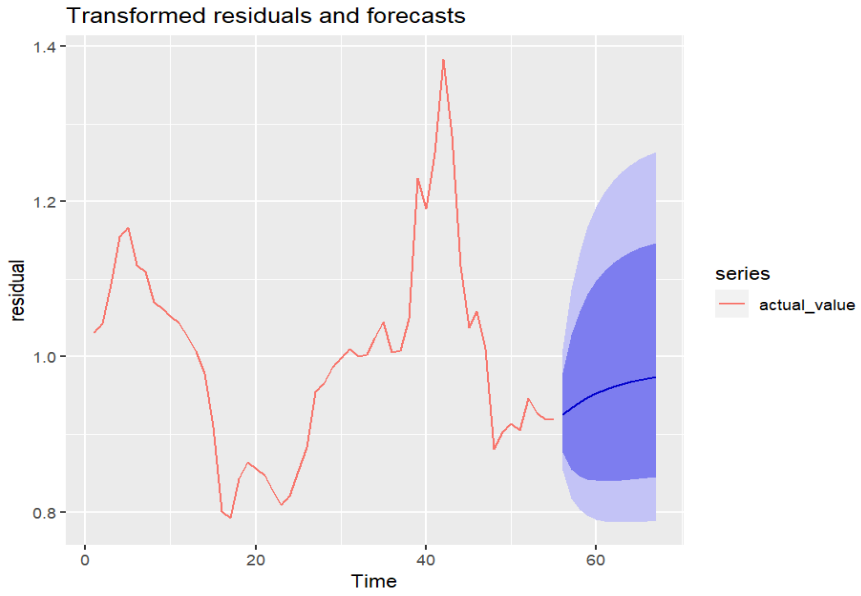

Fig. 6 Plot for the transformed residuals and forecasted data of next year

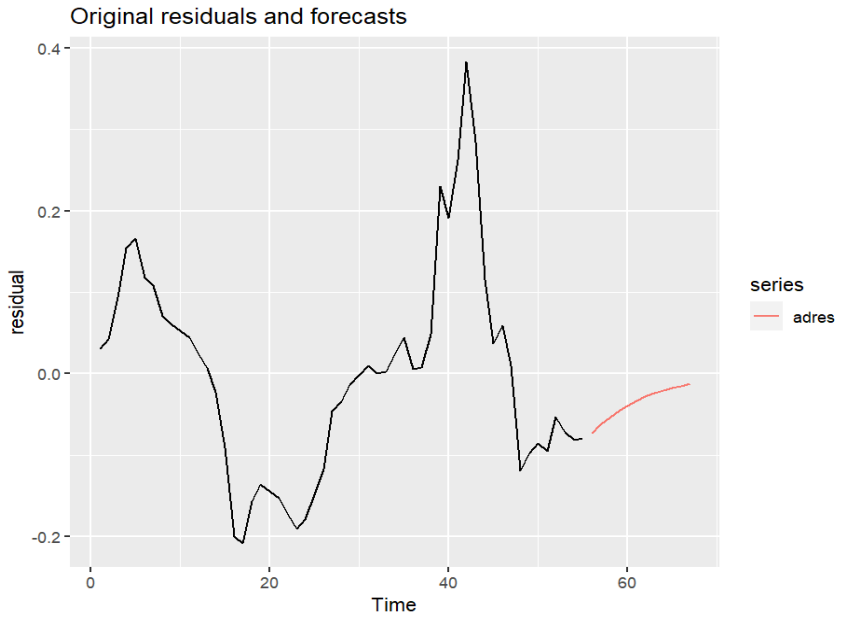

Fig. 7 Plot for the original residuals and forecasted data of next year

Table 2. Original forecasted values from two models and adjusted results

Points/Results	Linear model	ARIMA model	Adjusted	Final
56	1.038364	0.925308	-0.07469	0.963672
57	1.047084	0.933861	-0.06614	0.980945
58	1.055805	0.941204	-0.0588	0.997009
59	1.064526	0.947493	-0.05251	1.012019
60	1.073247	0.952867	-0.04713	1.026114
61	1.081968	0.957452	-0.04255	1.03942
62	1.090688	0.961357	-0.03864	1.052045
63	1.099409	0.96468	-0.03532	1.064089
64	1.10813	0.967503	-0.0325	1.075633
65	1.116851	0.9699	-0.0301	1.086751
66	1.125571	0.971934	-0.02807	1.097505
67	1.134292	0.973658	-0.02634	1.10795

As the data used in this model is transformed by adding one to the original residuals, the forecasted values are adjusted by subtracting one and shown in the fourth column. Figure 7 shows the original residuals and the adjusted values with the same trend as the transformed plot. The final forecasting results are calculated by adding the linear model values and adjusted ARIMA model values. In Figure 8, the black line represents the past prices, and the rest three lines are for forecasting prices (red), upper 95% prediction values (green) and lower 95% prediction values. It shows the US gasoline prices would generally increase every month in the next year, however, there is still possibility for the price to be decreased for the next three months. The lowest price in the forecast interval is $0.858 in October 2023. The highest price would reach $1.399 in July 2024 based on the interval, which is greater than the maximum price in the past prices of $1.30.

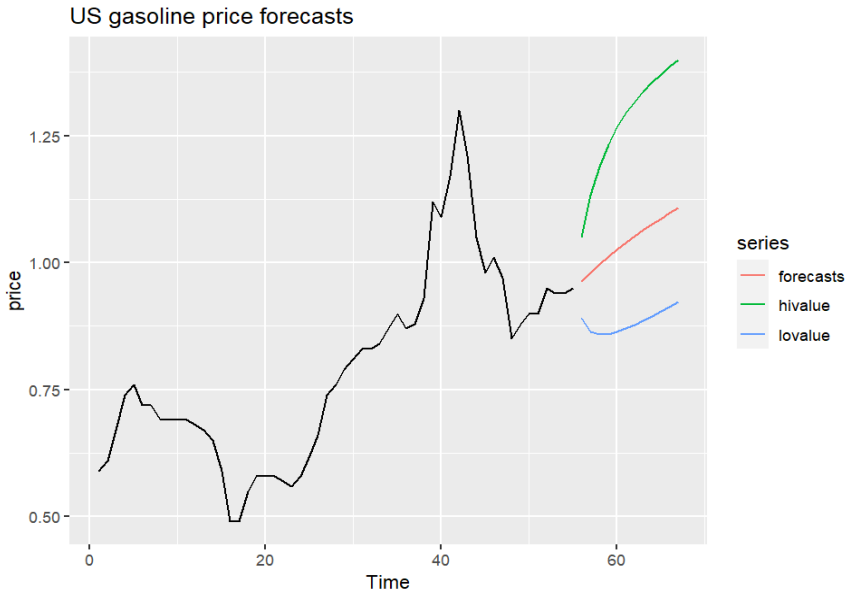

Fig. 8 Plot for the original residuals and forecasted intervals of next year

4. Discussion

The study shows a high potential for growing US gasoline prices in the next year based on the patterns of past prices. As the Covid-induced historic decline in gasoline demand, the prices stayed low. However, the global oil demand has rebounded from 2021 [10], and it is reasonable that the prices would increase and fluctuate at around $1 in the future. Nevertheless, the study uses the univariate time series data. Gasoline is the fuel refined from crude oil. The spot crude oil affects the spot gasoline price through the channel of oil-specific demand shocks [11]. In the longer-term outlook,

the oil demand is uncertain because of the increasing share of electric cars in the US vehicle fleet and alternative energy resources. For future research, the multivariate time series data including factors such as crude oil prices and oil demand and supply might be able to increase accuracy of the US gasoline price forecast.

5. Conclusion

This study forecasts the US gasoline price for the next year by applying and combining the linear regression model and the ARIMA model. Specifically, the linear model and model accounted for residuals is ARIMA(1,0,1) with a lambda value of -0.8999268. The source data is monthly end-user total prices for gasoline in the United from January 2019 to July 2023. The results show that the price will be trending upwards and reach $1.108 USD per liter in July 2024. It would be helpful for the individual consumers and the businesses to have notice and prepare for the rise of the gasoline price. The analysis is based on the univariate time series data, therefore it is better to consider more factors such as the crude oil prices and supply in the future research.

References

[1] Isabella Ruble. The U.S. crude oil refining industry: Recent developments, upcoming challenges and prospects for exports. The Journal of Economic Asymmetries, Volume 20, 2019, Article e00132.

[2] BP p.l.c. Bp statistical review of world energy 2022. June 28, 2023. Retrieved on August 16, 2023. Retrieved from: http://bp.com/statisticalreview.

[3] Behzod B. Ahundjanov, Sherzod B. Akhundjanov, Botir B. Okhunjanov. Risk perception and oil and gasoline markets under COVID-19. Journal of Economics and Business, Volume 115, 2021, Article 105979.

[4] Pippa Stevens. Rising fuel costs are a massive problem for business and consumers — Here's why they're so high. May 19, 2022. Retrieved on August 19,2023. Retrieved from: https://www.cnbc.com/2022/05/19/fuel-is-a-problem-for-business-and-consumers-why-prices-are-so-high.html.

[5] Jean Folger. How gas prices affect the economy. September 01, 2021. Retrieved on August 23, 2023. Retrieved from: https://www.investopedia.com/financial-edge/0511/how-gas-prices-affect-the-economy.aspx.

[6] Trading Economics. United states gasoline prices. August 19, 2023. Retrieved on August 19,2023. Retrieved from: https://tradingeconomics.com/united-states/gasoline-prices.

[7] International Energy Agency. Monthly oil price statistics. August 2023. Retrieved on August 10,2023. Retrieved from: https://www.iea.org/data-and-statistics/data-product/monthly-oil-price-statistics-2#monthly-oil-product-prices.

[8] Michael Jefferson. A crude future? COVID-19s challenges for oil demand, supply and prices. Energy Research & Social Science, Volume 68, 2020, Article 101669.

[9] Adam Hayes. Autoregressive integrated moving average (ARIMA) prediction model. Retrieved on August 21, 2023. Retrieved from: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp.

[10] International Energy Agency. Oil – Why is oil important. Retrieved on August 27, 2023. Retrieved from: https://www.iea.org/energy-system/fossil-fuels/oil.

[11] Louis H. Ederington, Chitru S. Fernando, Thomas K. Lee, Scott C. Linn, Huiming Zhang. The relation between petroleum product prices and crude oil prices. Energy Economics, Volume 94, 2021, Article 105079.