MSc Business Analytics

Time Series Analysis

STAMATOULA-GERASIMOULA MESOLORA

Registration number: f2822308

Advanced Topics in Statistics

July 2024

Contents

Introduction

This study involves time series analysis, using a dataset of the daily price of kilowatt-hour (kWh) for the period 2010-2018. The aim of the analysis is to use the data to forecast for the 1st half of 2019 the average monthly prices and whether the price will rise or not. The initial phase of the analysis focuses on the time series decomposition in order to find out whether it follows a tr and and if it has seasonality. The second phase of the analysis involves time series' transformation for making it stationary, removing tr and and seasonality. After that, follows the creation of the appropriate model, which is going to be used for forecasting and a Diagnostic checking step for examining if the chosen (estimated) model fits the data reasonably well, to test if the residuals of the estimated model are uncorrelated, homoscedastic and normal (white noise). Finally, there is the forecasting of the average monthly prices and the possible rise or fall of the price for the 1st half of 2019.

Methodology

The methodology of this study involves several steps to analyze the daily price of kilowatt-hour (kWh) electricity for the period 2010-2018 and forecast the average monthly prices for the first half of 2019 and determine whether the price will rise.

The initial dataset is consolidated into a single time series and formatted removing any missing values. Monthly average prices are calculated by grouping the data by month. This step ensures the data is ready for time series analysis by smoothing out daily fluctuations and highlighting monthly tr ands.

To understand the underlying patterns in the data, time Loess (STL) decomposition is performed. After decomposition, the series is transformed to achieve stationarity. Stationarity is a crucial property for time series modeling, indicating that the statistical properties of the series do not change over time. This is achieved by differencing and logarithmic transformation to remove tr ands and seasonality. Stationarity is then tested using the Augmented Dickey-Fuller test (ADF).

An appropriate Seasonal ARIMA (SARIMA) model is selected to fit the stationary time series. Specifically, a SARIMA(1, 0, 1)(0, 0, 1)[12] model is chosen, which accounts for both non-seasonal and seasonal components of the time series. The SARIMA model parameters are estimated and the model's adequacy is evaluated through diagnostic checks. These checks include examining the residuals of the model to ensure they exhibit properties of white noise: uncorrelated, homoscedastic and normally distributed. The residuals are further analyzed using plots and statistical tests to confirm these properties.

After confirming the stationarity of the selected model, we then proceed to forecast future prices and tr and, providing insights for predicting the electricity market's behavior in the upcoming months.

Time Series Decomposition

STL decomposition is robust to outliers and uses a smoothing technique (LOESS) to estimate tr and. This process separates the series into tr and, seasonal and random components. Identifying these components helps in understanding the behavior of data over time. The decomposition of the time series data into its tr and, seasonal and remainder components provides insights into the patterns of electricity prices.

In Figure 1 we can see the electricity prices over time and it is obvious there is both tr and seasonality.

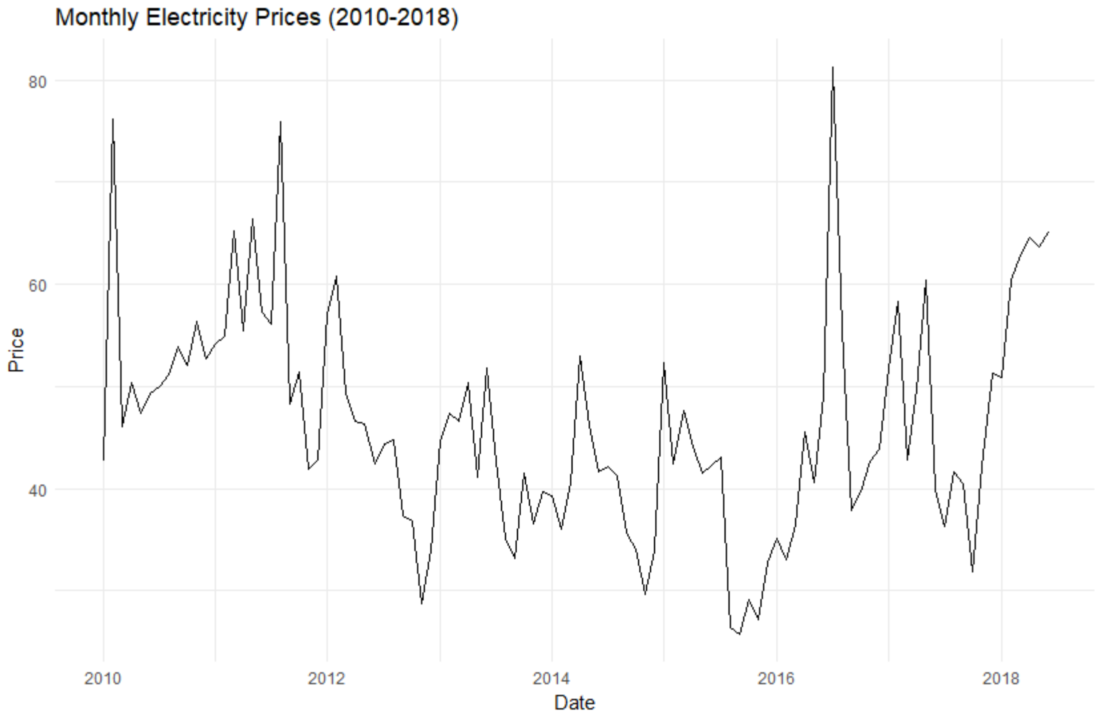

Figure 1 – Monthly Electricity Prices (2010 - 2018)

In order to gain a better understanding we have to analyze each component separately. As illustrated in Figure 2, the STL decomposition allows us to separately analyze the seasonal patterns, long-term tr ands and residual variations in the data.

The seasonal component, shown in the first panel of Figure 2, reveals regular, repeating patterns within each year. The strong, consistent peaks and troughs in the seasonal component are evident of the presence of seasonality in electricity prices. The middle panel of Figure 2 depicts the tr and component, which captures the long-term movement in the electricity prices. The tr and line shows a general decrease in prices from 2010 to around 2016, followed by a noticeable upward tr and starting in 2017. The remainder or residual component, shown in the bottom panel of Figure 2, represents the irregular, random variations that are not explained by the seasonal or tr and components. This component captures the noise and unpredictable fluctuations in the data. The residuals appear relatively stationary and have lower magnitudes compared to the seasonal and tr and components, indicating that the STL decomposition effectively isolates the systematic patterns in the time series.

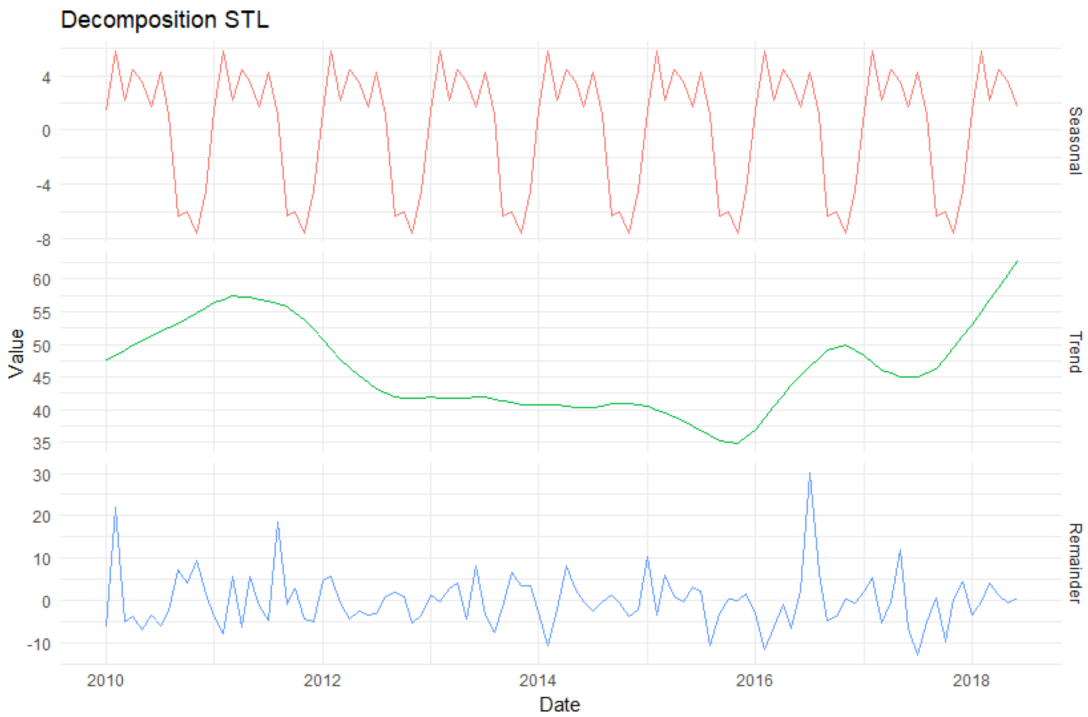

Figure 2 - Decomposition STL

Stationarity

Stationarity is a crucial property for time series modeling, indicating that the statistical properties of the series do not change over time. This is achieved by differencing and logarithmic transformation to remove tr ands and seasonality. Stationarity is tested using the Augmented Dickey-Fuller test (ADF) and the p – value of the initial time series is approximately 0,39, as shown in Table 1, which indicates that the time series is not stationary.

In order to remove the seasonality, there has been applied a logarithmic transformation, depicted in Figure 3, which compressed the scale of the data and reduced the impact of outliers on the overall scale of the data, making them less prominent. This reduces the effect of extreme values on the results and contributes to stabilizing the variance of the data.

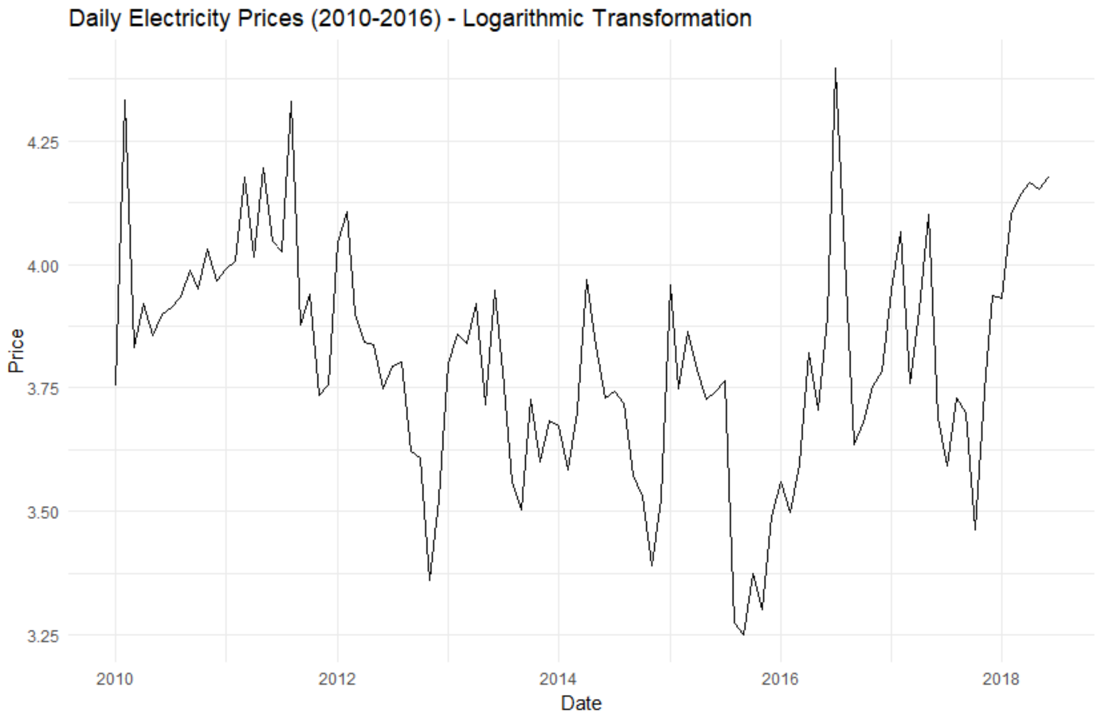

Figure 3 - Daily Electricity Prices (2010-2016) - Logarithmic Transformation

For the removal of the tr and component there have been applied first differences, depicted in Figure 4. This transformation helps stabilize the variance and makes the data more stationary.

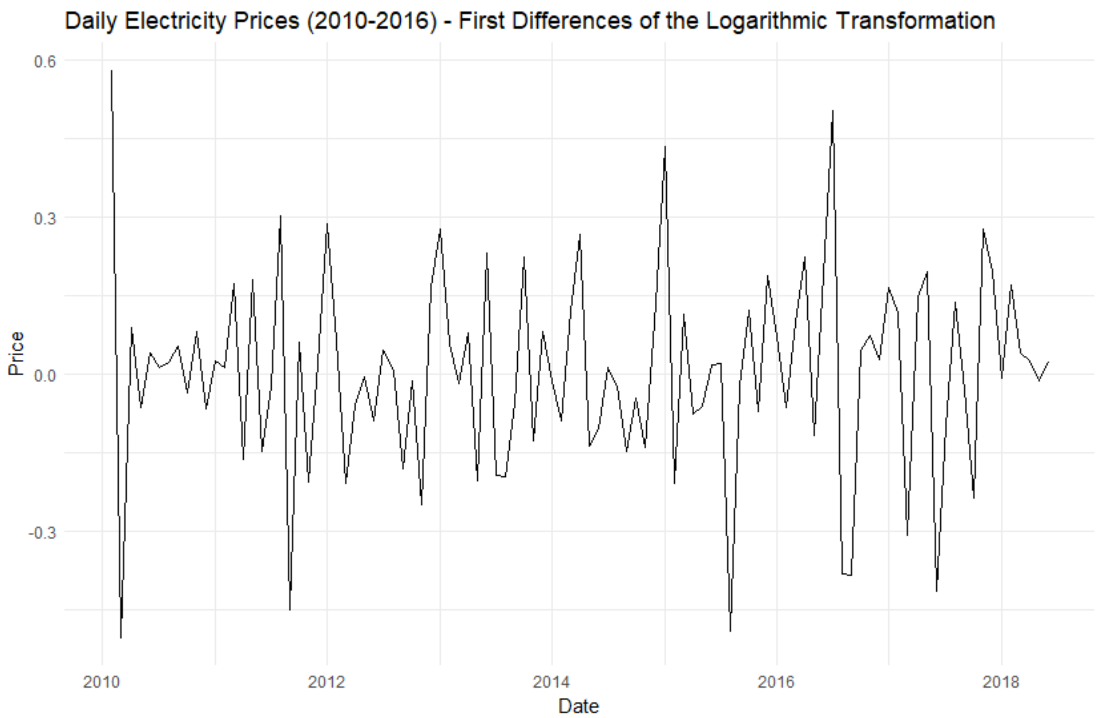

Figure 4 - Daily Electricity Prices (2010-2016) - First Differences of the Logarithmic Transformation

After the complete transformation, the p – value of the Augmented Dickey-Fuller test (ADF) decreases to 0.01, which indicates that stationarity has been achieved.

Time Series	**P-Value**
Initial Time Series	0.39
Transformed Time Series	0.01

Table 1 - Augmented Dickey-Fuller test (ADF)

Model Selection

Since we have made the time series stationary, we aim to identify the most suitable model for the data. To achieve this, we employ two essential diagnostic tools: the Autocorrelation Function (ACF) plot and the Partial Autocorrelation Function (PACF) plot. These plots provide insights into the autocorrelation structure of data, which is crucial for selecting an appropriate model. The ACF plot helps us understand the overall autocorrelation pattern, while the PACF plot allows us to identify the direct relationships between the time series and its lagged values. By examining these plots, we can determine the order of the autoregressive (AR) and moving average (MA) components, ultimately guiding our choice of model.

The Autocorrelation Function (ACF) plot provides insights into the correlation between observations of a time series at different lags. The presence of significant autocorrelations at specific lags informs the choice of the AR (AutoRegressive) terms in the model. The ACF plot shown in Figure 5 displays the autocorrelation values of the time series at various lags and each bar represents the correlation between the time series and its lagged values. The height of the bars indicates the strength and direction of the correlation. Positive values indicate a positive correlation, while negative values indicate a negative correlation. The dashed red lines in the ACF plot represent the approximate 95% confidence intervals for the autocorrelations. These confidence intervals are calculated based on the assumption that the time series follows a white noise process, where each observation is uncorrelated with all others. If an autocorrelation value exceeds these bounds, it is statistically significant at the 5% level, suggesting a deviation from the white noise assumption. In the below ACF plot, we observe that most of the autocorrelation values fall within the confidence intervals, indicating that the differenced and log-transformed series exhibits characteristics of a stationary process. However, the plot shows that the time series has autocorrelation at lag 1 and lag 8, while autocorrelation decays rapidly as the lag increases between those lags and after lag 8. This means that the time series exhibits a strong dependence on its past values.

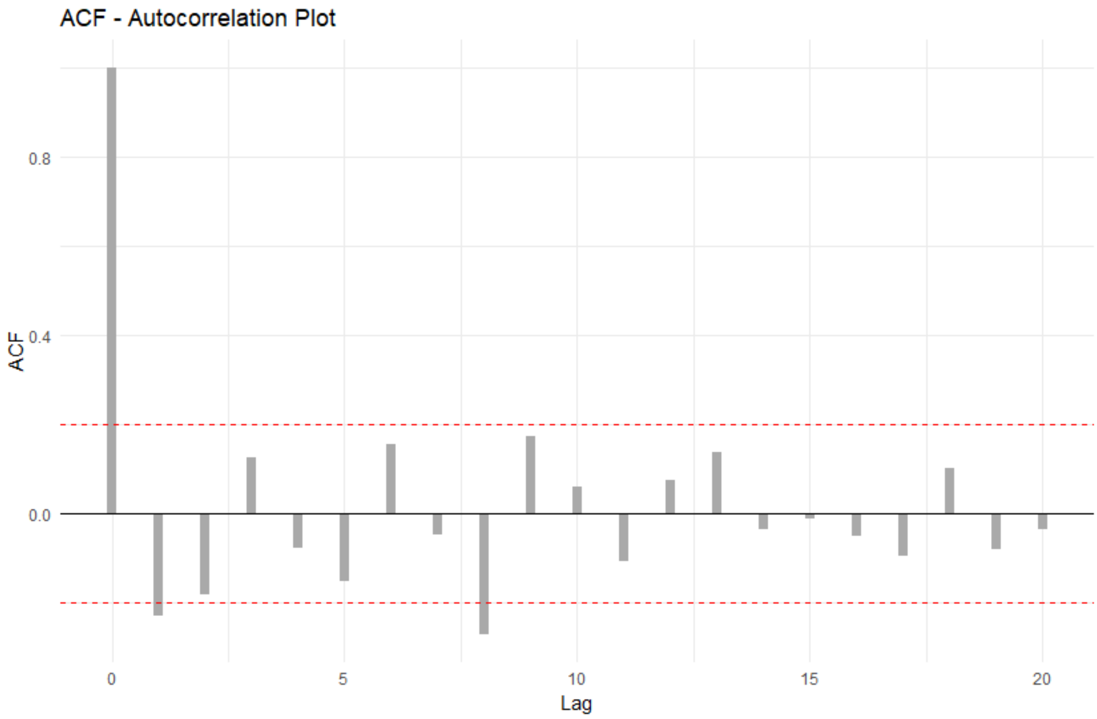

Figure 5 – ACF Autocorrelation Plot

The PACF plot shows the partial autocorrelation of the time series. The partial autocorrelation at lag k is the correlation between the time series at time t and the time series at time t-k, after removing the effects of the time series at all lags between 1 and k-1. The ACF plot in Figure 6 displays the autocorrelation values of the time series at various lags. Each bar represents the correlation between the time series and its lagged values. The confidence intervals, depicted as dashed lines, represent the range within which autocorrelations are considered statistically insignificant at the 5% significance level.

We can see that there is a significant partial autocorrelation at lag 1, lag 2 and lag 8, indicating that there is a direct relationship between the time series at time t and the time series at time t-1.

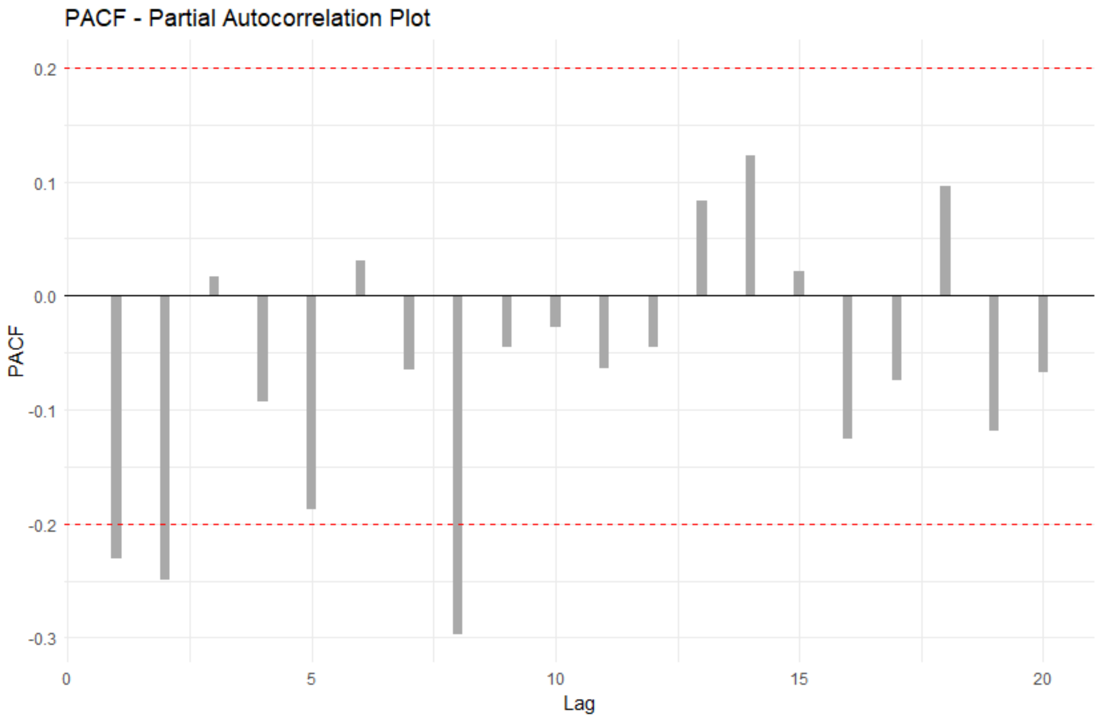

Figure 6 - PACF Partial Autocorrelation Plot

The selected seasonal AutoRegressive Integrated Moving Average (SARIMA) model is specified as ARIMA(1,0,1)(1,0,1), with period, as ACF showed a 1-month and 8-month lagged correlation. The model includes one autoregressive term, one moving average term, one seasonal autoregressive term and one seasonal moving average term. This model was chosen based on the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, which indicated the presence of autocorrelation at lag 1 and lag 8.

The non-seasonal part of the model:

- p = 1: The model includes one autoregressive (AR) term, which means that the current value of the time series is a function of the previous value.
- d = 0: The model does not include any differencing terms, since the data has already been differenced.
- q = 1: The model includes one moving average (MA) term, which means that the current value of the time series is a function of the previous error term.

The seasonal part of the model:

- P = 1: The model includes one seasonal autoregressive (SAR) term, which means that the current value of the time series is a function of the value from the same season in the previous year.
- D = 0: The model does not include any seasonal differencing terms.
- Q = 1: The model includes one seasonal moving average (SMA) term, which means that the current value of the time series is a function of the error term from the same season in the previous year.

The model selection process involved evaluating different models based on their AIC values and selecting the model that best balances model complexity and goodness of fit. The AIC absolute value of the model is equal to 62,07.

Diagnostic Plots & Tests

After fitting the SARIMA model, diagnostic checks are performed to ensure that the residuals exhibit properties of white noise—uncorrelated, homoscedastic and normally distributed. These checks validate the adequacy of the model and its suitability for forecasting.

Figure 7 shows the residuals of the model. The residuals are centered around zero, indicating that the model is a good fit for the data. The residuals are also randomly distributed, which is another good indicator of a well-fitted model.

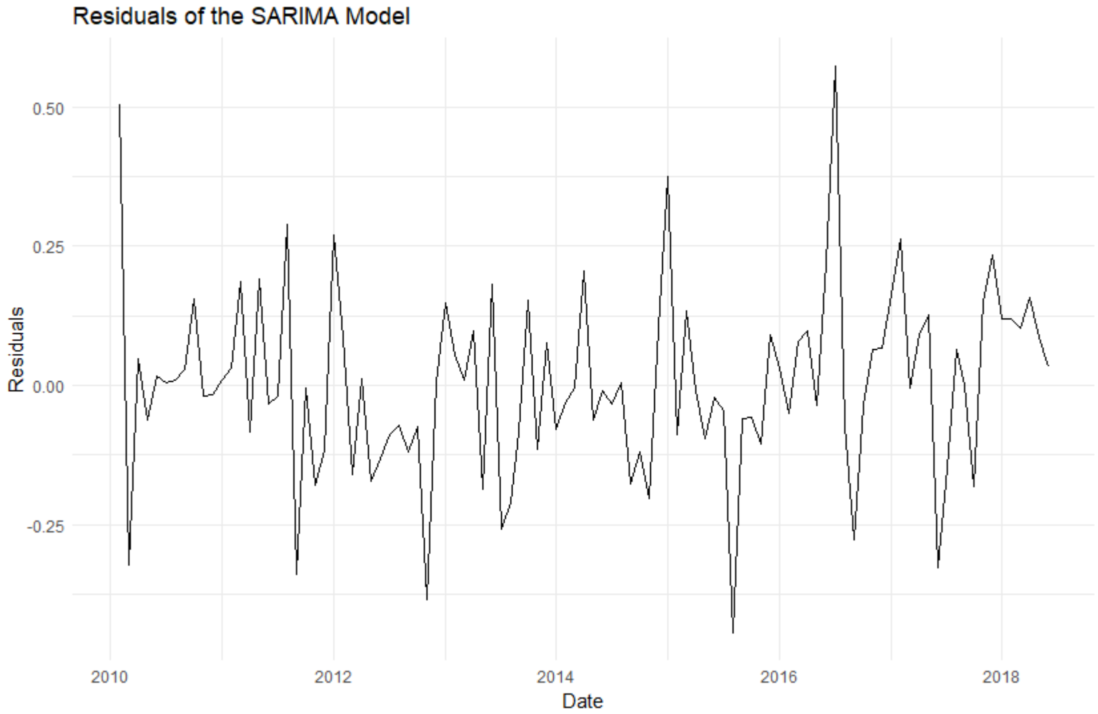

Figure 7 – SARIMA Residuals

The normality of the residuals was assessed using the Lilliefors and Shapiro-Wilk tests. The results of both tests indicated a p-value greater than 0.05, which suggests that the null hypothesis of normality cannot be rejected. This implies that the residuals are approximately normally distributed, with a mean close to zero.

Normality Tests	P - Value
Lilliefors Test	0.38
Shapiro-Wilk test	0.05

Table 2 – Normality Tests

In Figures 8 & 9 we have visualized the distribution of the model’s residuals with a histogram and a q-q plot respectively. Both plots show a roughly linear pattern, indicating that the residuals are approximately normally distributed as there are a few outliers on the tails of the plots. Although, the model is reasonably well-fitted.

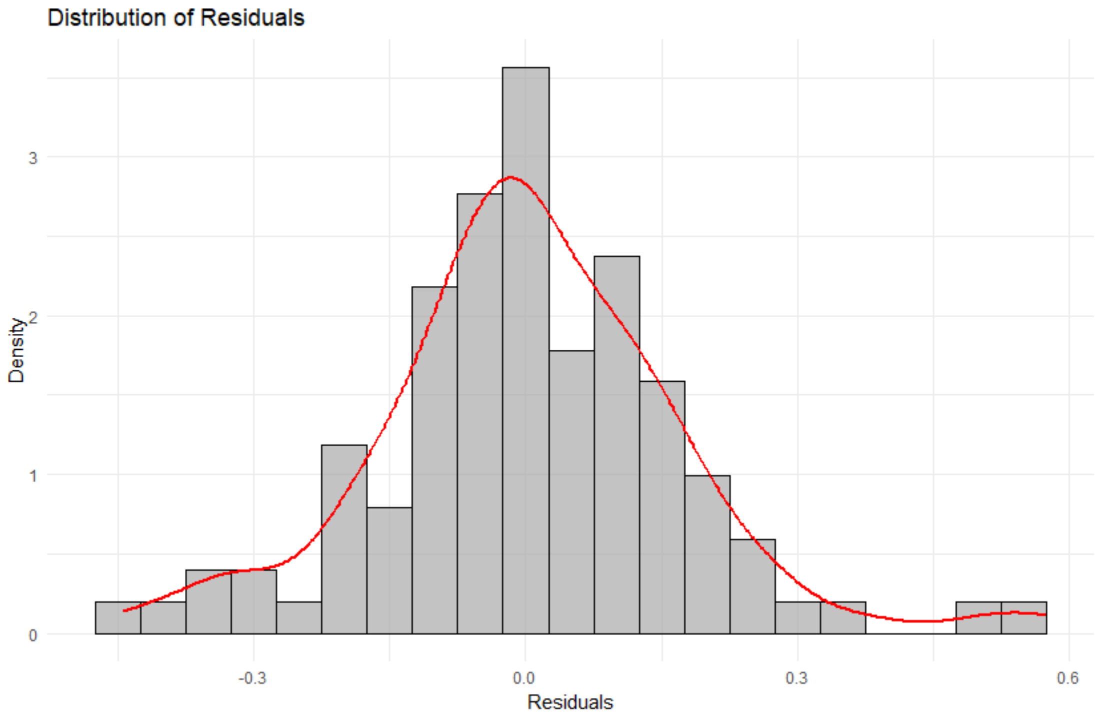

Figure 8 - Distribution of Residuals

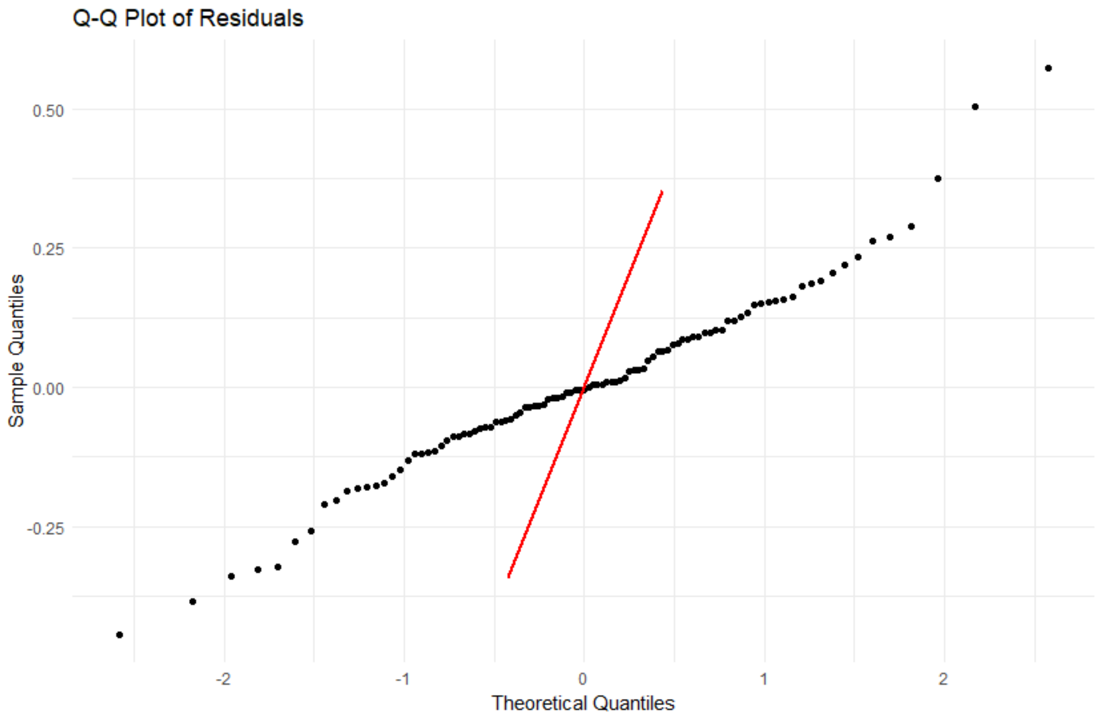

Figure 9 - Q-Q Plot of Residuals

The Autocorrelation of the residuals was assessed using the Box-Pierce and Ljung-Box tests and the Durbin-Watson and Breusch-Godfrey tests, as well as Autocorrelation and Partial Autocorrelation Plots. The results of all tests indicated a p-value greater than 0.05, which suggests that the null hypothesis that there is no serial correlation of any order up to 1 cannot be rejected.

Autocorrelation Tests	**P - Value**
Box-Pierce Test	0.87
Ljung-Box Test	0.71
Durbin-Watson Test	0.29
Breusch-Godfrey test	0.93

Table 3 – Autocorrelation Tests

As we can see in Figures 10 & 11 below, the model has successfully captured the underlying patterns in the data and the residuals are white noise. This is a desirable outcome, as it indicates that the model is well-specified and that there are no additional autoregressive terms that need to be included.

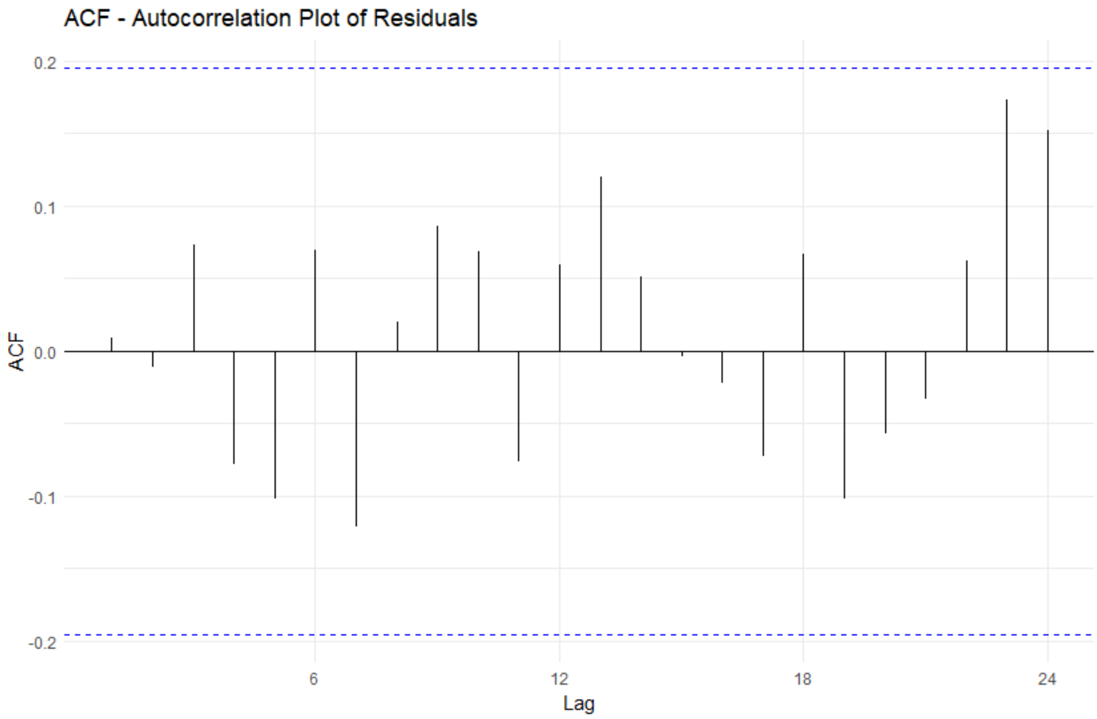

Figure 10 – ACF Autocorrelation Plot of Residuals

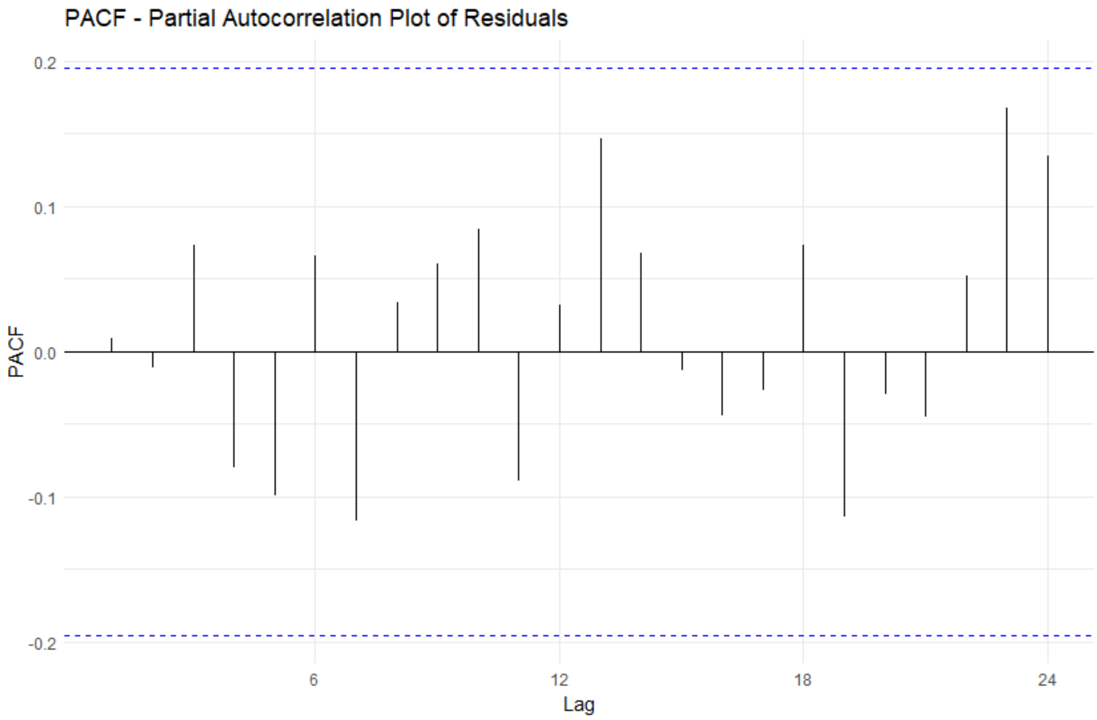

Figure 11 - PACF Partial Autocorrelation Plot of Residuals

The Heteroskedasticity of the residuals was assessed using the Ljung-Box - Autocorrelation Test of Squared Residuals, Goldfeld-Quandt Test, Breusch-Pagan Test, White Test and ARCH LM Test of Heteroscedasticity, as well as Autocorrelation and Partial Autocorrelation Plots of squared residuals. The results of all tests indicated p-values greater than 0.05, suggesting that we cannot reject the null hypothesis of homoscedasticity, meaning there is no significant evidence of heteroskedasticity in the residuals. Additionally, the autocorrelation plot shows that there is no significant autocorrelation whereas partial autocorrelation plot of squared residuals shows a small amount of autocorrelation, further supporting the conclusion that the problem of heteroscedasticity is very small.

Heteroskedasticity Tests	**P - Value**
Autocorrelation Test of Squared Residuals	0.67
Goldfeld-Quandt Test	0.39
Breusch-Pagan Test	0.53
White Test	0.69

Table 4 - Heteroskedasticity Tests

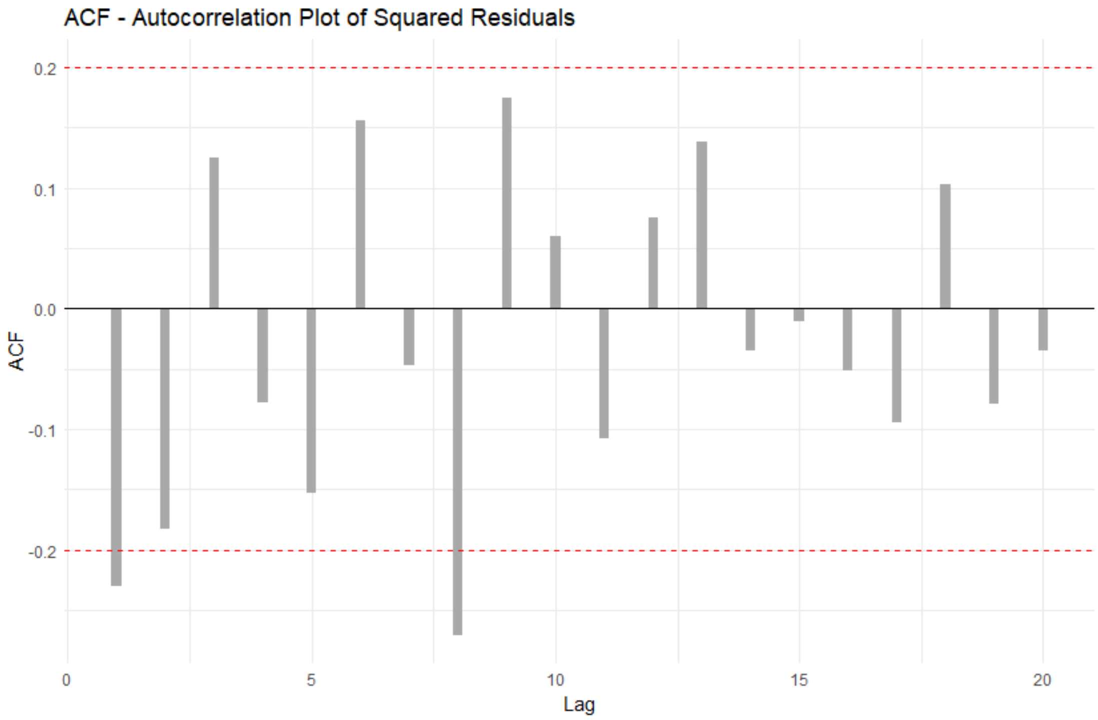

Figure 12 - ACF - Autocorrelation Plot of Squared Residuals

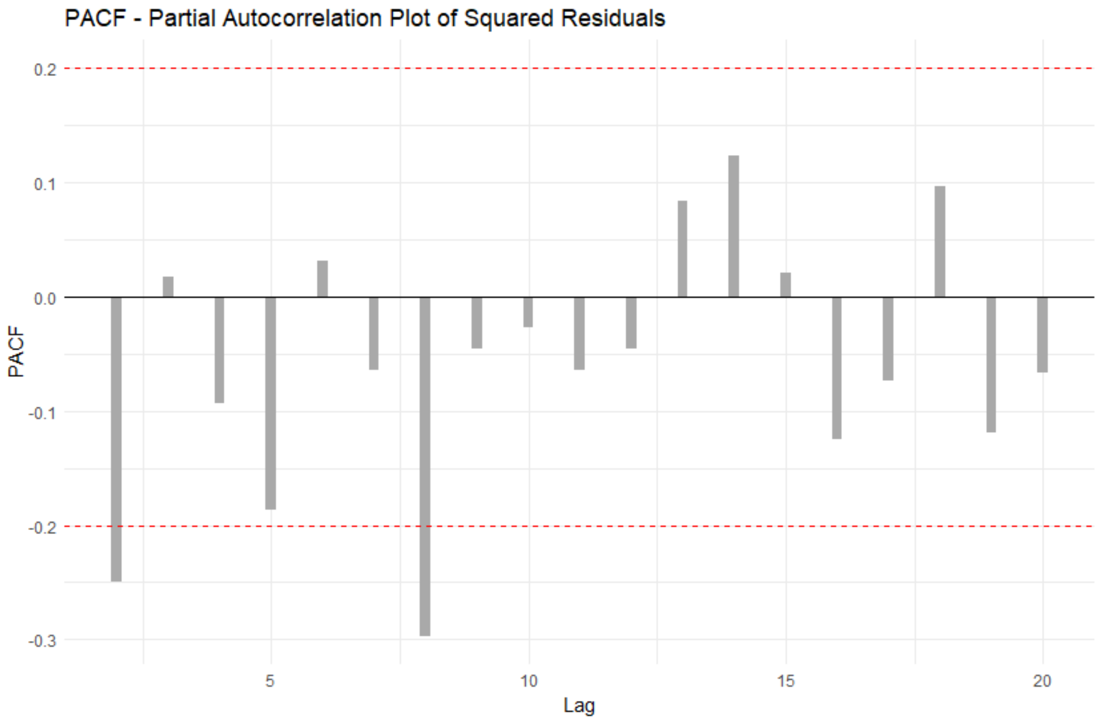

Figure 13 - PACF – Partial Autocorrelation Plot of Squared Residuals

The Stationarity parameter of the residuals was assessed using the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) and Phillips-Perron Unit Root Tests. The KPSS test, where the null hypothesis is that the data is stationary, has a p-value greater than 0.05 to confirm stationarity. The Phillips-Perron test, on the other hand, tests for the presence of a unit root with a null hypothesis of non-stationarity; a p-value less than 0.05 indicates that we reject the null hypothesis, supporting stationarity. So, the results of both tests indicate that the residuals of the model are stationary.

Stationarity Tests	**P - Value**
KPSS Test	0.10
Phillips-Perron Unit Root Test	0.01

Table 5 - Stationarity Tests

Forecast

The insights gained from the STL decomposition and the fitted SARIMA model enable accurate forecasting of future electricity prices. The model's ability to capture the seasonal and trend components allows for more reliable predictions of monthly average prices.

After we inverse the logarithmic transformation and the first differences, we are able to proceed to the forecast of the kwh price of the rest months of 2018 and the first half of 2019.

The forecasted monthly average electricity prices for the first half of 2019 show an increasing trend, with prices ranging from $65,94 the first month to $76,83 in June. The average forecasted price for the first half of 2019 is $71,32, which is higher than the average price of $61,26 in 2018. This suggests that the price is expected to increase in 2019 compared to 2018.

The forecasted monthly average electricity prices for the first half of 2019 are as follows:

Date	**Price**
January 2019	$71.78
February 2019	$72.79
March 2019	$73.81
April 2019	$74.82
May 2019	$75.83
June 2019	$76.83

Table 6 - Forecasted Monthly Average Electricity Prices for the First Half of 2019

The forecast also includes confidence intervals to represent the uncertainty in the predictions. The upper and lower limits for the 95% confidence intervals were transformed back to the original scale and are presented alongside the forecasts. The confidence interval for the forecasted prices is between $65,69 and $82,04, indicating a range of possible values within which the true prices are likely to fall. The upper bound of the confidence interval is $82,04, while the lower bound is $65,69. This interval provides a measure of uncertainty associated with the forecasted prices, highlighting the potential variability in the actual prices that may occur.

Figure 14 below illustrates the forecasted prices with the confidence intervals:

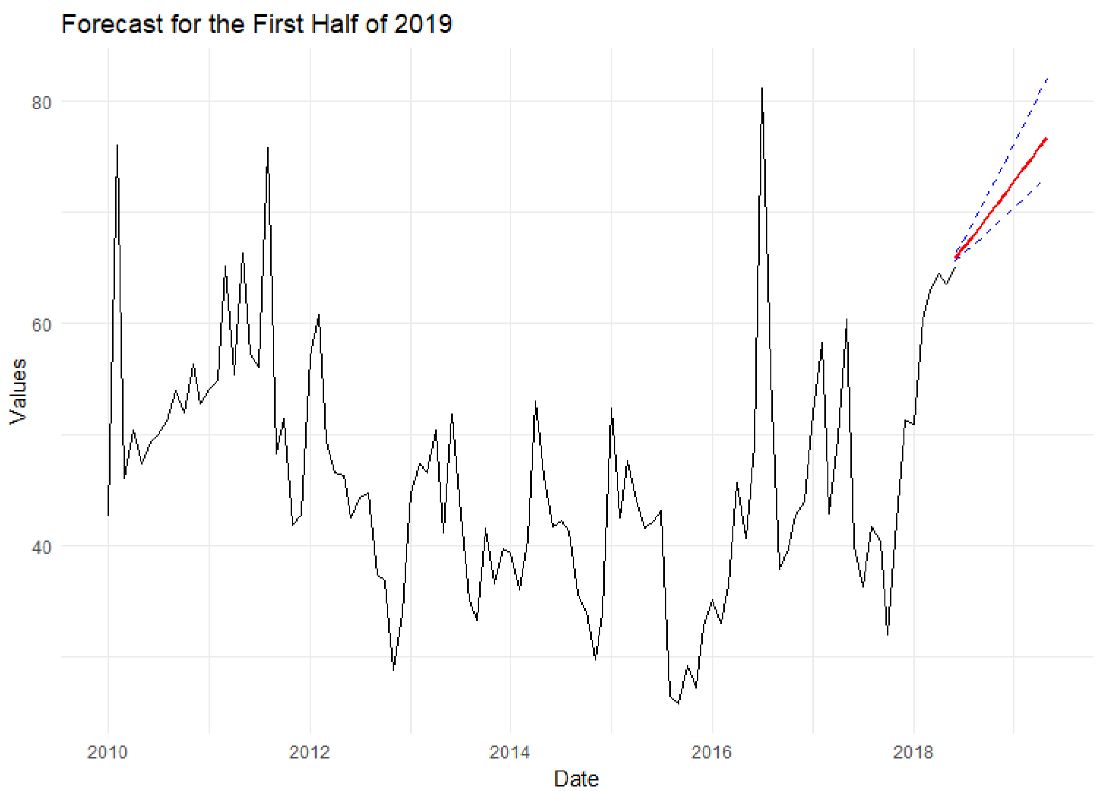

Figure 14 - Forecast for the First Half of 2019

The average price for 2018 is $61,26 and the average forecasted price for the first half of 2019 is $68,34, therefore, the price is expected to increase in 2019 compared to 2018.

Diagnostic Plots & Tests of the Forecast

In order to be sure about the results of the forecast we have to make sure that the residuals is white noise. Thus, there have been conducted several tests to assess Normality, Autocorrelation and Heteroscedasticity. The p – value of Shapiro-Wilk normality test is equal to 0,05, so the residuals follow an approximately normal distribution, as shown in Figure 15. The p – value of the Durbin-Watson test is 0,29 which suggests that there is no serial correlation, as shown in Figures 16 & 17. Finally, the p – value studentized Breusch-Pagan test is 0,53 meaning that there is no significant evidence of heteroskedasticity in the residuals and they are White Noise.

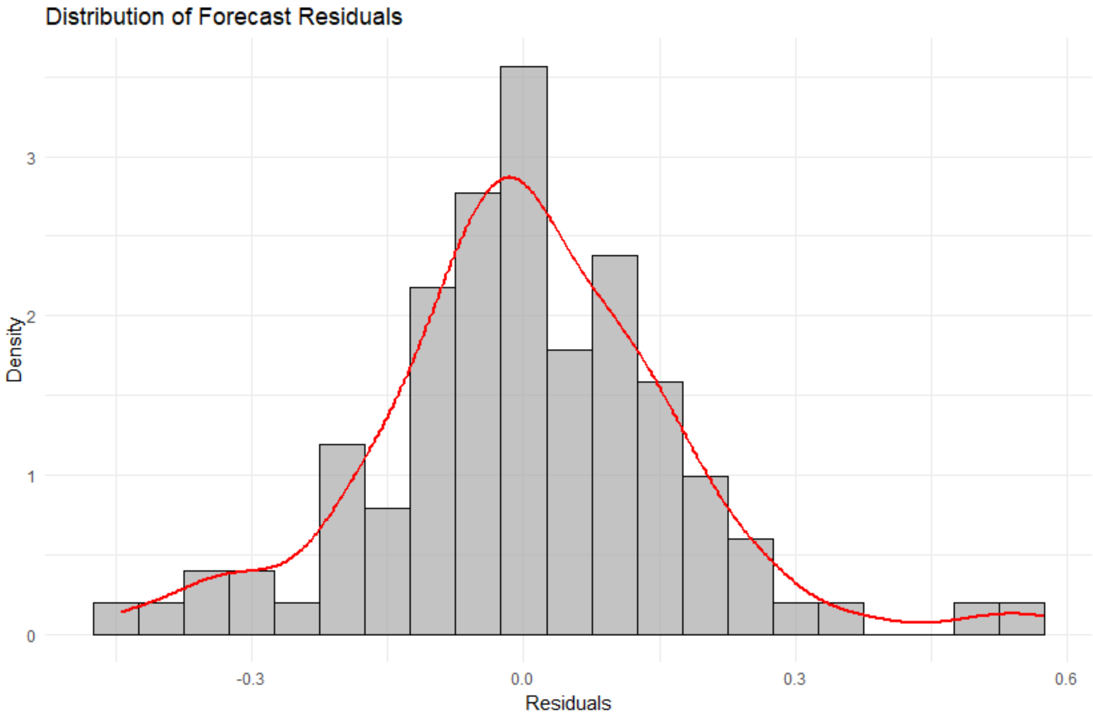

Figure 15 – Distribution of Forecast Residuals

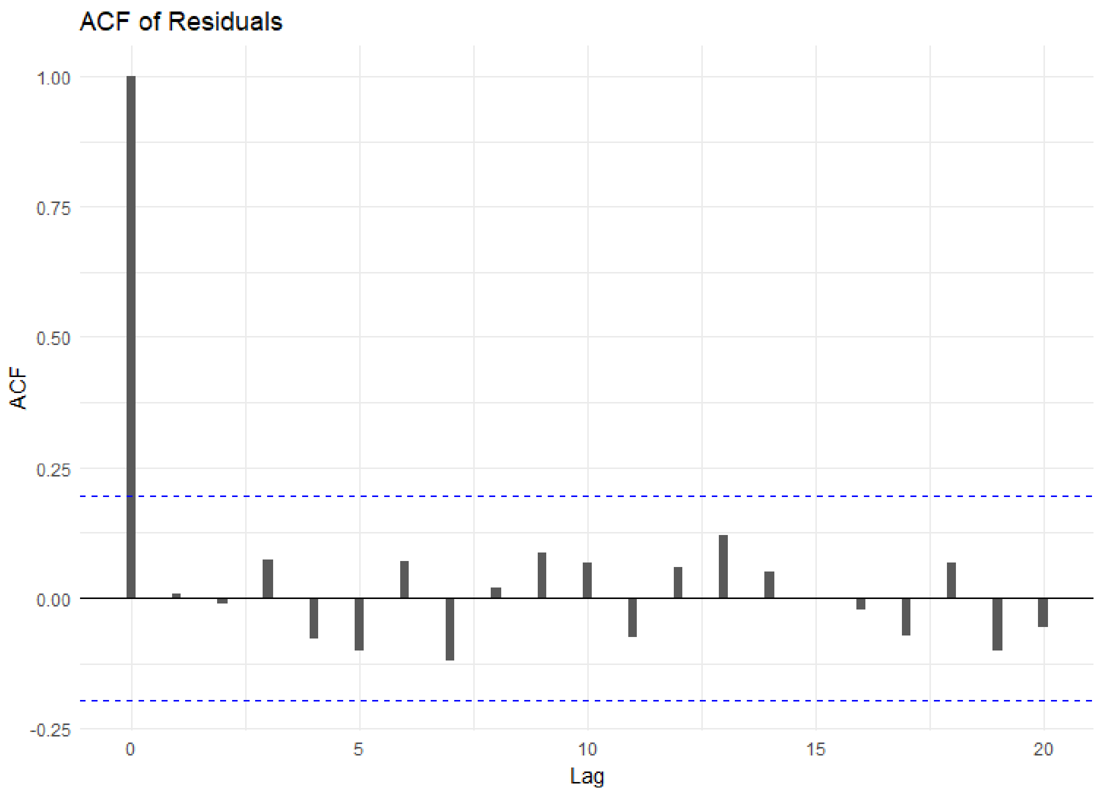

Figure 16 - ACF of Forecast Residuals

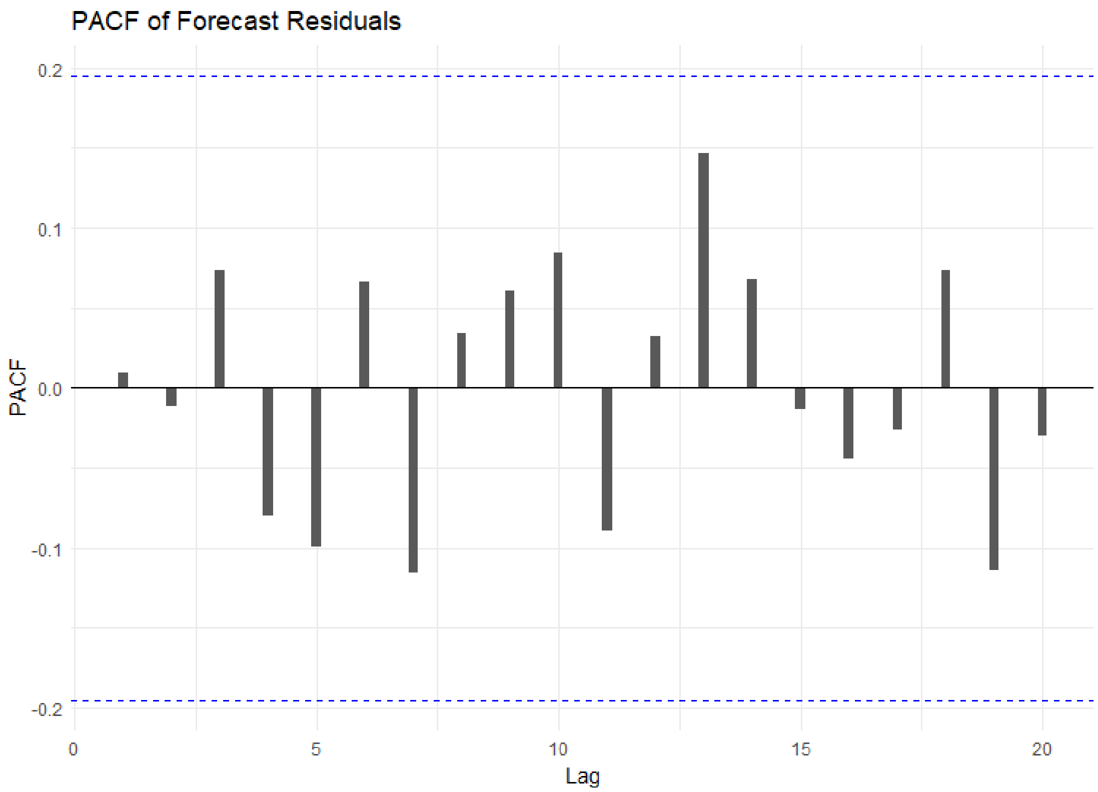

Figure 17 - PACF of Forecast Residuals

Figures

Tables