Athens University of Economics and Business
Department of Management Science and Technology

Advanced Topics in Statistics
Professor: D. Karlis

Project I
Time Series Assignment
Electricity Price

Marianna Konstantopoulou
A.M: P2822122

M.Sc. Business Analytics
Part Time 2021-2023

Athens, 02/07/2023

Table of Contents

1. Introduction

The dataset at hand contains daily price records of the kilowatt hour (kWh) spanning from 2010 to 2018. The primary objective of this study is to develop a reliable forecasting model to predict the average monthly prices for the first half of 2019 and assess whether prices are likely to increase during that period.

To achieve this goal, we will adopt a data-driven approach and explore various forecasting methods while considering different data frequencies, such as daily, weekly, or monthly data. Our methodology will be presented in a clear and concise manner, detailing the steps taken to build the forecasting model. By leveraging advanced statistical techniques and employing suitable modeling strategies, we aim to provide valuable insights into the future trends of electricity prices during the specified timeframe.

The outcome of this study will serve as a valuable resource for decision-makers in diverse industries, empowering them to make informed choices in response to potential fluctuations in electricity prices. This assignment is meant to contribute meaningfully to understanding energy price dynamics and their implications on various sectors.

2. Exploratory Analysis and Model Development

In this initial chapter, we will understand the dataset and prepare the groundwork for developing our forecasting models. The chapter begins with an Exploratory Data Analysis (EDA), where we dive into the daily price records of the kilowatt hour (kWh) for the period 2010-2018, examining patterns, trends, and potential outliers.

The initial step involves providing a concise overview of the dataset, outlining its structure and content. To prepare the data for analysis, a cleaning process was conducted to ensure it is in the appropriate format. This included merging the date and price columns into singular ones. Due to variations in column lengths, they were handled separately and later combined uniformly. Subsequently, a check for missing values was performed, revealing one instance on the date 2013-03-11. This data point was excluded from the dataset to focus solely on forecasting prices from 2016 onwards.

Furthermore, two prices were identified that were recorded as below zero, and as a result, they were excluded from the dataset. Including such values would have led to non-applicable (N/A) results when fitting the logarithmic time series. Additionally, one extremely high outlier with a price of 1147.95 was excluded, as it disrupted the scaling of the time series during the fitting process. Removing these problematic data points ensures the accuracy and reliability of the time series analysis.

Following that it was time to create the time series with frequency 365 as we have daily data from year 2010 to 2018 so we can quickly check how our data looks. (i.e. Figure 1)

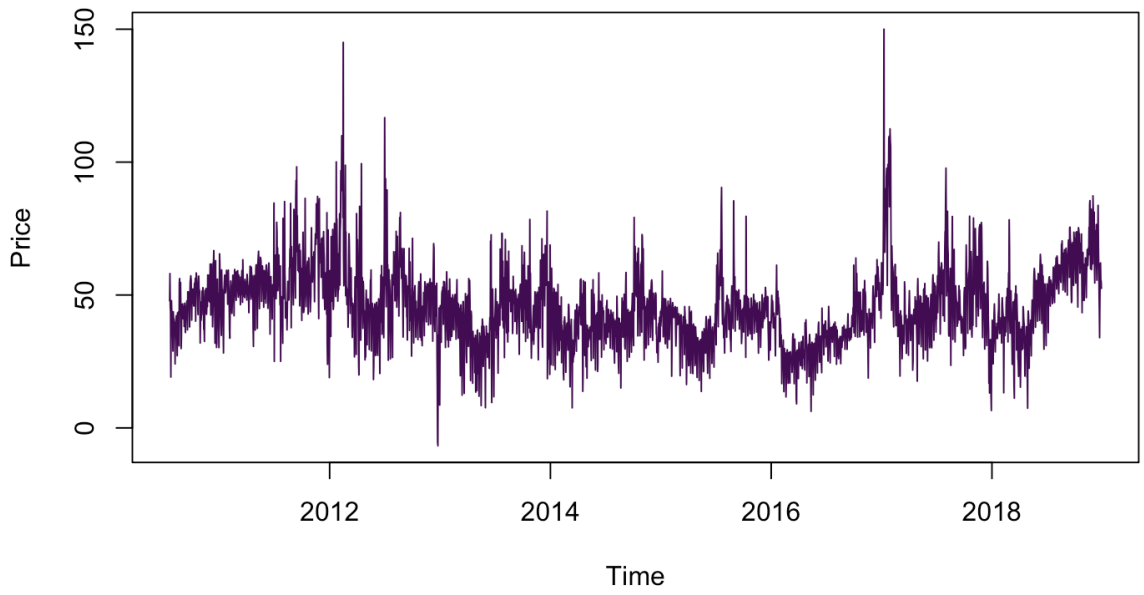

Figure 1: Times series with electricity price daily data from 2010-2018

An Augmented Dickey-Fuller test was conducted so we could check the stationarity. Based on the Augmented Dickey-Fuller test results, with a p-value of 0.01, we can reject the null hypothesis of non-stationarity. The alternative hypothesis is that the time series is stationary. Therefore, with a significance level of 0.05, we have enough evidence to conclude that the time series data is stationary.

Checking for normality a Shapiro-Wilk normality test was conducted for our data and the result showed that the p-value is less than 2.2e-16, which indicates strong evidence against the null hypothesis that our time series is normally distributed. Therefore, based on this test, we can conclude that it does not follow a normal distribution. However, upon inspecting the histogram and the QQ plot of our data (i.e., Figure 2), it is evident that the normality assumption does not appear to be severely violated.

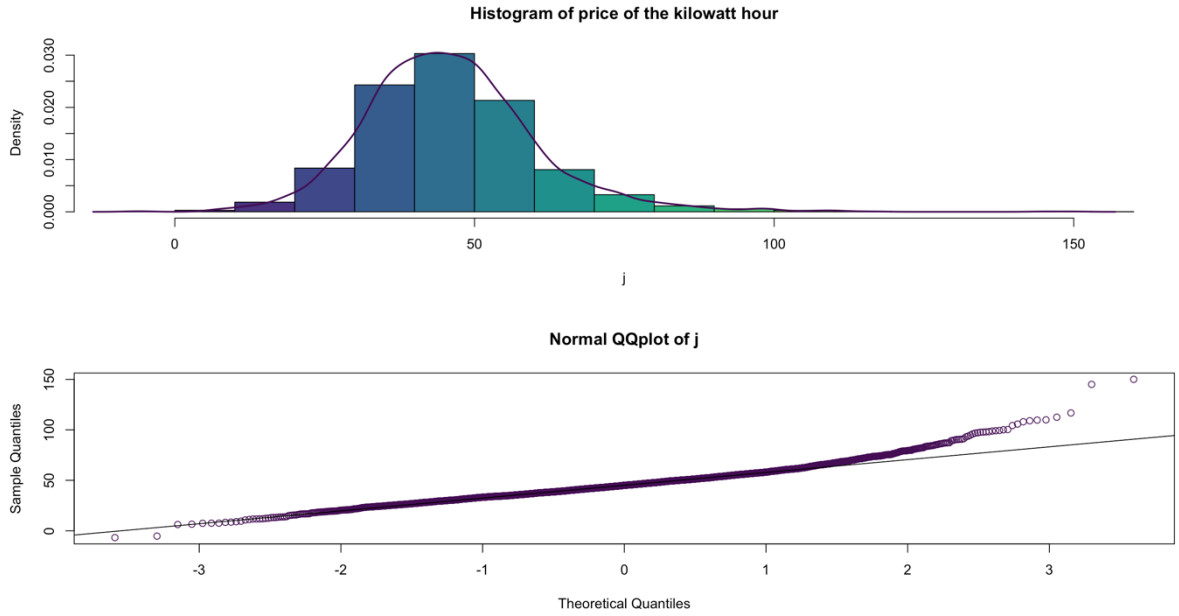

Figure 2: Times series qqplot and histogram for the daily electricity data

While drawing the autocorrelation and partial autocorrelation plots, we can see patterns that repeat every seven days, so 30 lags were included to see that the behavior starts at 7 and repeats at multiples of 7 (e.g., 14, 21) in the autocorrelation and partial autocorrelation plots. This allowed us to capture the autocorrelation and partial autocorrelation at the weekly level. (i.e Figure 3). In order to capture and analyze the weekly autocorrelation and partial autocorrelation patterns in the data, the time series data is converted into a seasonal time series with a frequency of 7 (representing a weekly pattern).

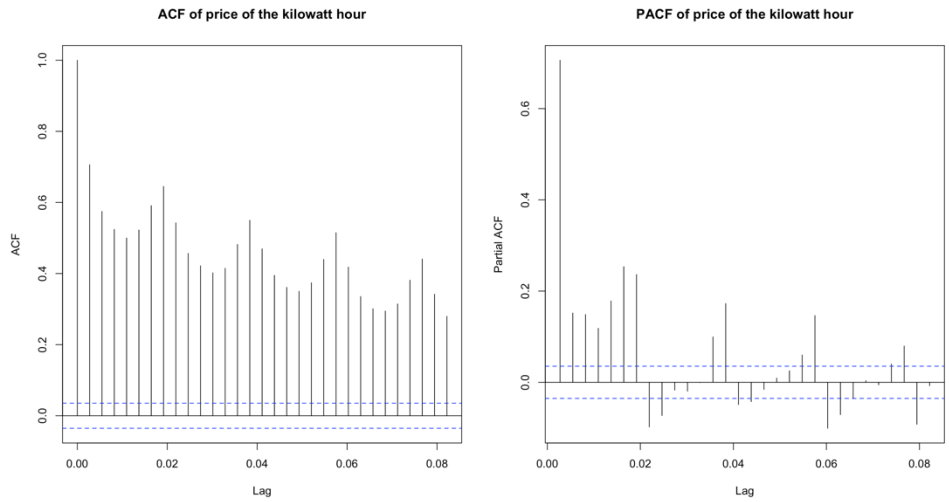

Figure 3: ACF and PACF plots of price of kilowatt hour

To prepare the dataset for time series forecasting by dividing it into training and testing sets, the data is initially split into an 80% training set and a 20% testing set. Next, the training data is processed to calculate the average weekly price. This is achieved by grouping the data by weeks and computing the mean price for each week, resulting in a dataset containing the weekly average prices for the training period. It's worth mentioning that when choosing the weekly frequency the variance will be affected and actually it will be smaller. To ensure consistency in the analysis, only weeks starting from "2016-01-03" and onwards are retained in the training dataset. This step is essential for the forecasting task as it focuses on data from the year 2016 onwards. Similarly, the testing data is also processed to calculate the average weekly price for each week in the test period. After completing these steps, two separate datasets are obtained: one with the weekly average prices for the training period and another with the weekly average prices for the test period. These datasets are now suitable for further analysis and time series forecasting.

Then we create a new time series object with now frequency 52 as we are referring to weekly data. (i.e Figure 4)

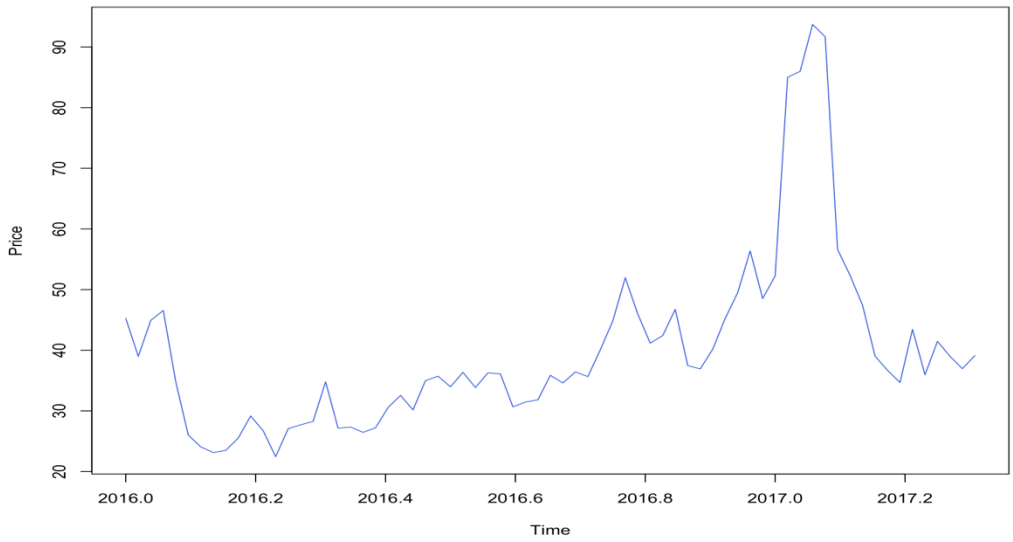

Figure 4: Times series with electricity price weekly data from 2016-2018

After conducting an Augmented Dickey-Fuller test for stationarity, the Augmented Dickey-Fuller test results, with a p-value of 03367, showed we can not reject the null hypothesis of non-stationarity. The alternative hypothesis is that the time series is stationary. Therefore, with a level of 0.05, we have do not enough evidence to conclude that the time series data is stationary.

This led to computing the logarithmic time series and the first differences of the logarithmic series. We performed an Augmented Dickey-Fuller test on both series. Interestingly, only the second series showed a p-value lower than the significance level of 0.05, indicating that it is stationary. We can visually validate this result while drawing the plots (i.e Figure 5)

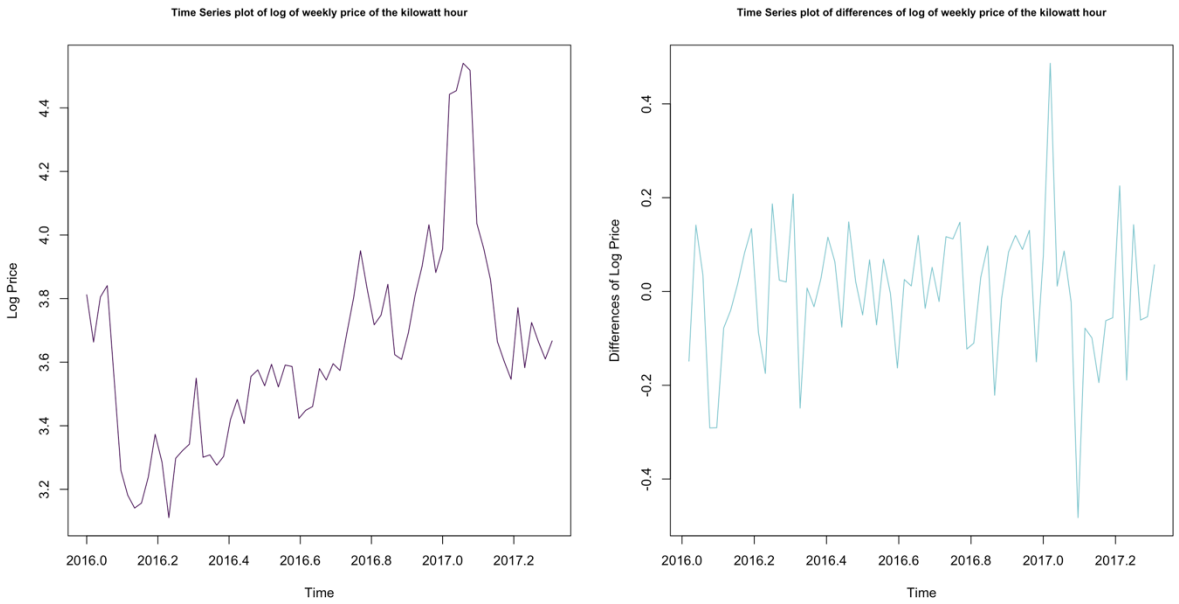

Figure 5: Times series log of weekly prices and the first differences of log

We can also check the autocorrelation and partial correlation plot of the weekly time series and the first differences logarithmic time series. (i.e Figure 6) The autocorrelation plot with a lag of 52 showed the correlation between each observation and its lagged value with a one-year time difference (52 weeks). This can reveal if there are significant seasonal patterns repeating every year that might influence the current week's price based on the previous year's price for the same week. For our stationary first differences logarithmic time series we can see that our weeks look uncorrelated and don't show any strong patterns. The partial autocorrelation function plot for the first differences of logarithmic time series with a lag of 52 shows the correlation between each observation and its lagged difference, excluding the effects of intermediate lags. Significant partial autocorrelation at lag 52 would indicate a strong relationship between the price changes in the current week and the price changes exactly one year ago. Again, for our stationary first differences logarithmic time series we do not have any strong relationships between price changes.

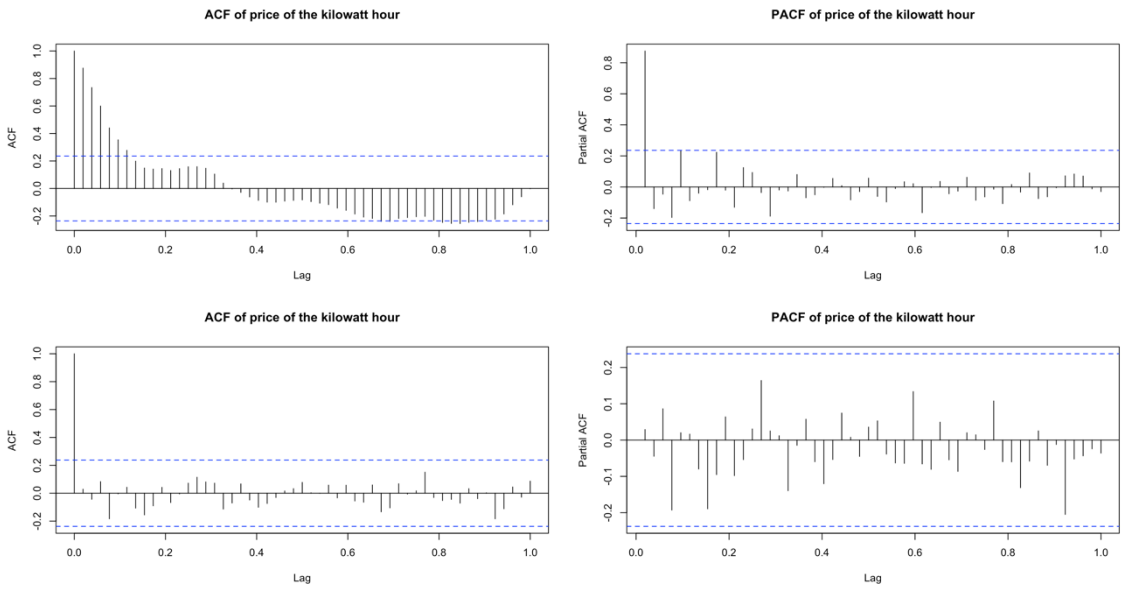

Figure 6: ACF and PACF plots for the time series and the first differences of log of the time series

We will use grid search to find the best ARIMA model for the given time series data. It tries different combinations of model parameters (AR and MA orders) using nested loops. For each combination, it fits an ARIMA model and calculates the Akaike Information Criterion (AIC) value, which helps assess the model's performance and complexity. The combination that yields the lowest AIC value is considered the best model. After checking 81 different models the model with the lowest AIC is an ARIMA model with no autoregressive component, differenced once to make it stationary and no moving average component, it was also was seasonally differenced once (period = 52) and the AIC is -1.1.

After finding the best model to use, we will conduct some diagnostic tests for the residuals to check the heteroscedasticity and the normality. (i.e Figure 7)

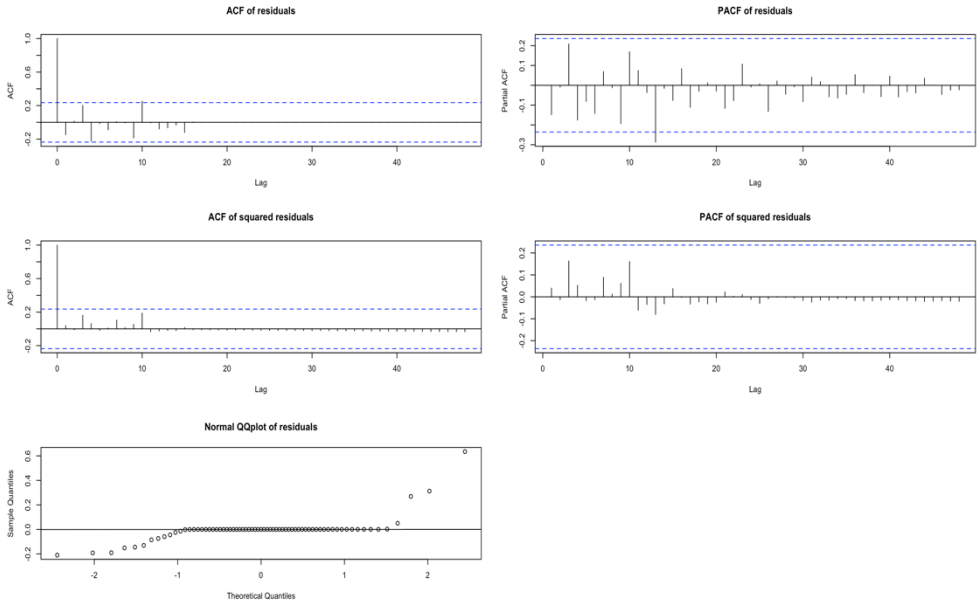

Figure 7: Diagnostics for residuals (ACF, PACF and QQ-plot)

The residuals are uncorrelated, as seen in the autocorrelation and partial autocorrelation plots, they display homoscedasticity, evident in the autocorrelation and partial autocorrelation plots of squared residuals, and they follow a normal distribution, as observed in the QQ-plot.

3. Forecasting

After successfully estimating an identified model and since the residuals resemble a white noise process, we are ready to test our model and forecast the weekly prices from week 2017-04-23 till the end of 2018 and validate our forecasting with the test dataset.

We will forecast 89 weeks using the drift method and we will also calculate the upper and lower limits. (i.e Figure 8)

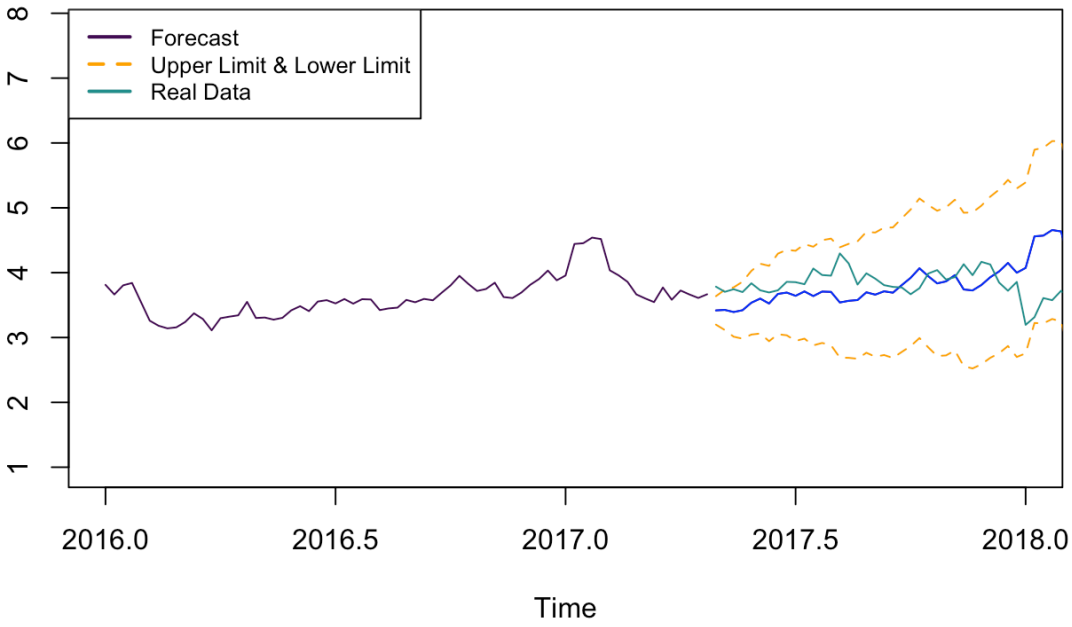

Figure 8: Forecasting the time period 2017-04-23 to 2018-12-30 with the identified ARIMA model

To assess the forecast's quality, we examine various accuracy measures and compare them with the summary statistics of the test dataset. The Mean Error (ME) of 46.18 indicates that, on average, the forecasted values are 46.18 units higher than the actual values. The Root Mean Squared Error (RMSE) of 47.58 implies an average error of 47.58 units between the forecast and actual values. The Mean Absolute Error (MAE) of 46.18 reflects an average absolute error of 46.18 units in the forecasted values compared to the actual values. The Mean Percentage Error (MPE) of 91.85% suggests that, on average, the forecasted values are 91.85% higher than the actual values. Additionally, the Mean Absolute Percentage Error (MAPE) of 91.85% indicates an average absolute percentage error of 91.85% in the forecasted values compared to the actual values.

Overall, the positive ME and relatively high RMSE, MAE, MPE, and MAPE reveal significant deviations between the forecasted and actual values in the test set. The forecast tends to consistently overestimate the actual values, suggesting that the model may not accurately capture the underlying patterns and variations in the data. As a result, improvements to the forecasting model or exploration of alternative forecasting approaches may be necessary to achieve higher accuracy in predictions.

Using the same model and now using all the data from 2016-2018 we will try to forecast the monthly electricity prices for the first half of 2019. (i.e Figure 9)

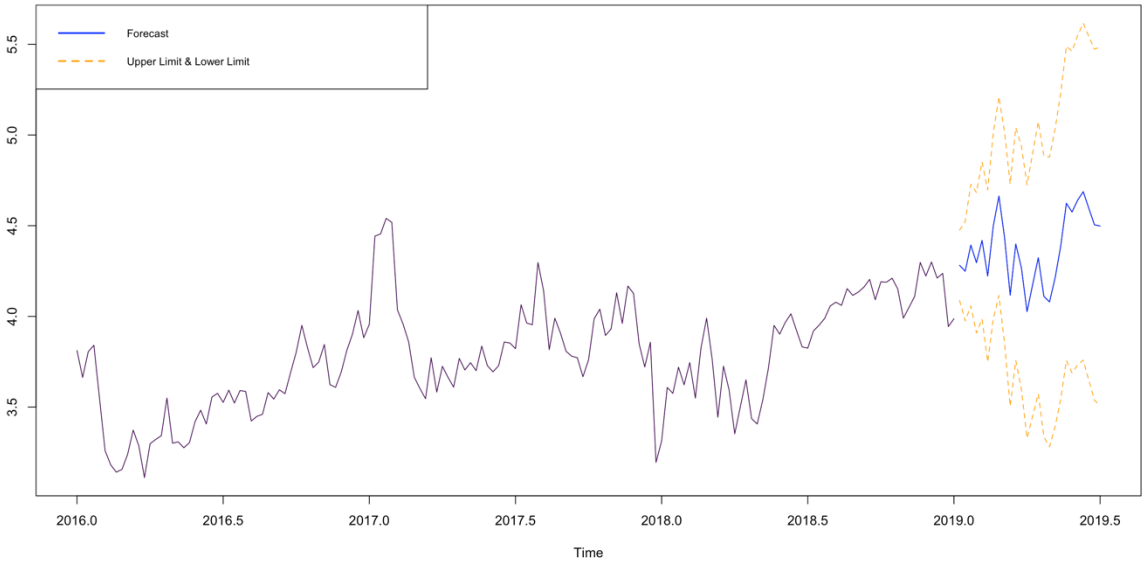

Figure 9: Forecasting the first half of 2019 with the identified ARIMA model

We will also calculate the monthly average prices which are the following:

Jan-19	74.05
Feb-19	85.71
Mar-19	74.03
Apr-19	63.98
May-19	75.66
Jun-19	101.95

For the 2019 forecast, it is important to be mindful that there may be deviations between the predicted and actual prices. The analysis reveals that the model tends to consistently overestimate the actual values, which indicates the need for caution when interpreting the forecasted results. While the forecast provides valuable insights, it is essential to consider the possibility of discrepancies from the real prices. Therefore, decision-making should take into account the potential variability and uncertainties associated with the forecast, and adjustments or alternative approaches might be required to enhance the accuracy of the predictions.