Department: Department of Management Science and Technology

M.Sc.: Business Analytics (Full-Time)

Course: Advanced Topics in Statistics

Professor: Dimitris Karlis

Deliverable: Time Series Project

Student Name: Evangelos Lakkas-Pyknis

Registration Number: f2822306

E-mail: eva.lakkaspyknis@aueb.gr

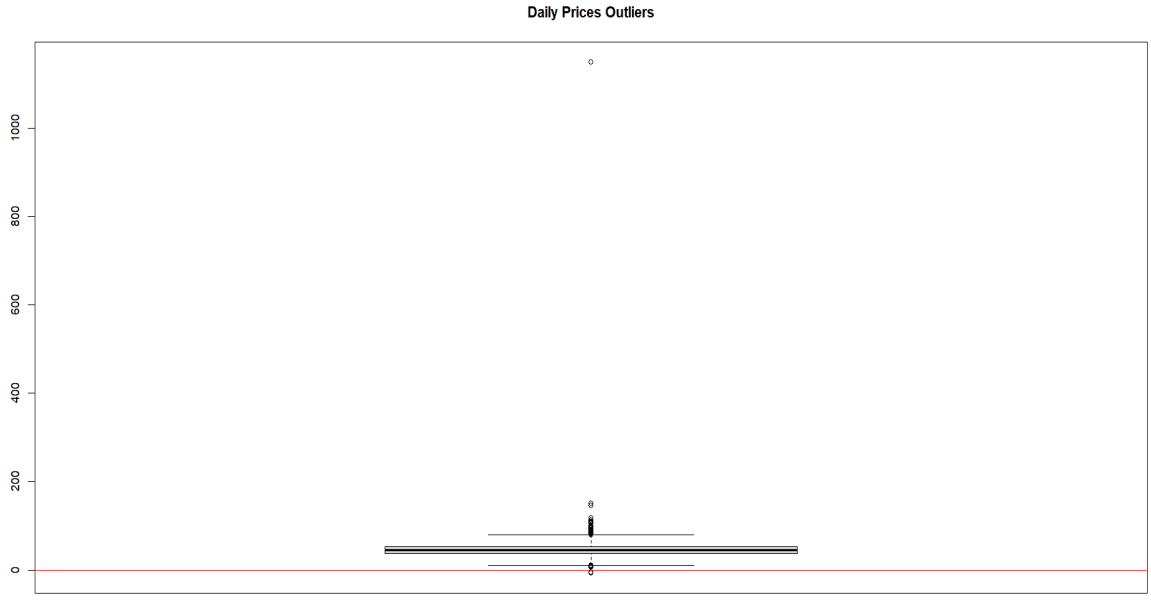
Περιεχόμενα

Introduction

This assignment aims to implement a time series analysis in order to forecast the average monthly prices for the first six months of 2019. The given data refer to the daily kWh price in Europe and consists of 3086 observations which start from the 21st of July 2010 and end on the 31st of December 2018. The process that was followed to find the appropriate forecasting model was checking the stationarity of the time series and applying the necessary transformations on the data in order to achieve this property, observation of autocorrelation, and partial autocorrelation plots to gain an insight on what should be the order of the model, fitting of MA and ARMA models, model comparison and selection based on the smallest AIC score, diagnostic checking the model assumptions of and performance evaluation.

Remarks about the data and data cleaning

As has been said before the available dataset included 3086 observations and generally it had not any severe quality issues. However, I located four daily records that seemed problematic. These records included one unreasonably high price (equal to 1147€, on the 16th of August 2010), two negative prices which I considered that this is due to a possible mistake in the data entry process (obviously prices can't be negative values) on the 25th and 26th of December 2012 and one missing value for the 11th of March 2013. To treat these issues, I replaced them with the mean price of the corresponding month and year (after excluding these outliers from the calculation of each mean). These outliers are also shown in the boxplot below (the negative values are below the red line).

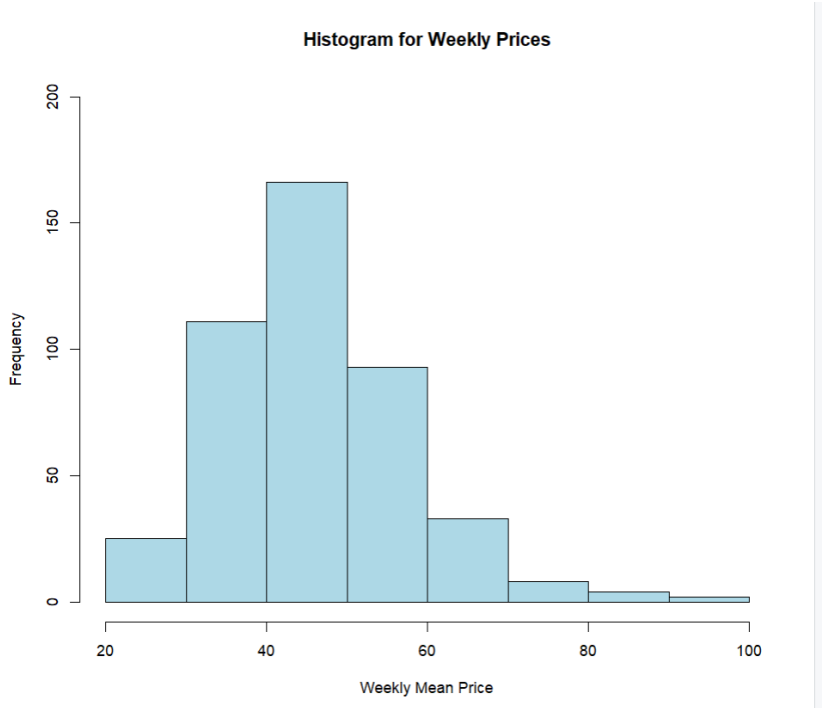


The data have been aggregated on a weekly basis, I preferred to fit the model on the weekly prices in order to lose as few observations as possible in case where taking more than one difference to make the time series stationary was necessary

and hopping that more data points would help to the finding of a model that has good fit and predictive abilities. Also, I used a 75-25 split for the training and testing datasets in the first part of the project. After making the time series stationary (by taking first-order differences) I split the data into training and testing to be able to do an out-of-sample evaluation of the model. The train–test ratio that I used was 92%-8%, I chose this split because 8% of the weekly data corresponds to 36 weeks which is approximately 9 months so in this way I had a slightly larger time period compared to the 6-month prediction that is the final goal of this project for testing the model's performance. I did not use a classic 80%-20% split because the test size that would be derived would correspond approximately to 2 years which is a significantly larger period of prediction compared to one semester so the insight that I would be taking from this testing dataset might not be useful for the task that I want to solve.

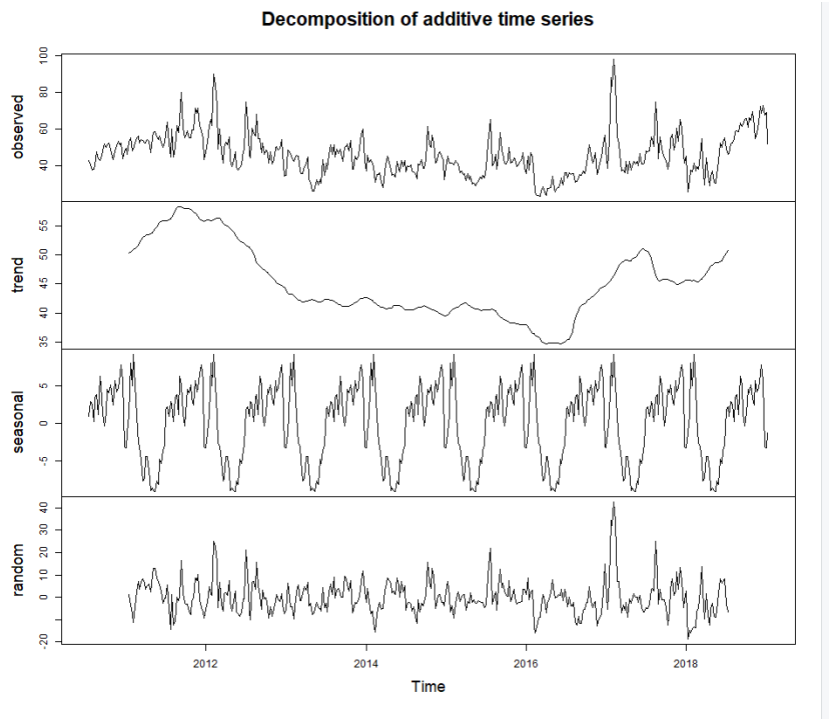
Stationarity and EDA

To gain some knowledge about the weekly electricity price data I implemented some descriptive plots for the time series and the price alone. In the histogram below we see that most prices are within the range [30,60]. Also, the distribution of the weekly prices seems to decline from the normal distribution as there is a right skewness and the histogram is heavy on the right tail. However, the mean and the median weekly price do not seem to have a significant difference (mean=45.99, median=44.67).

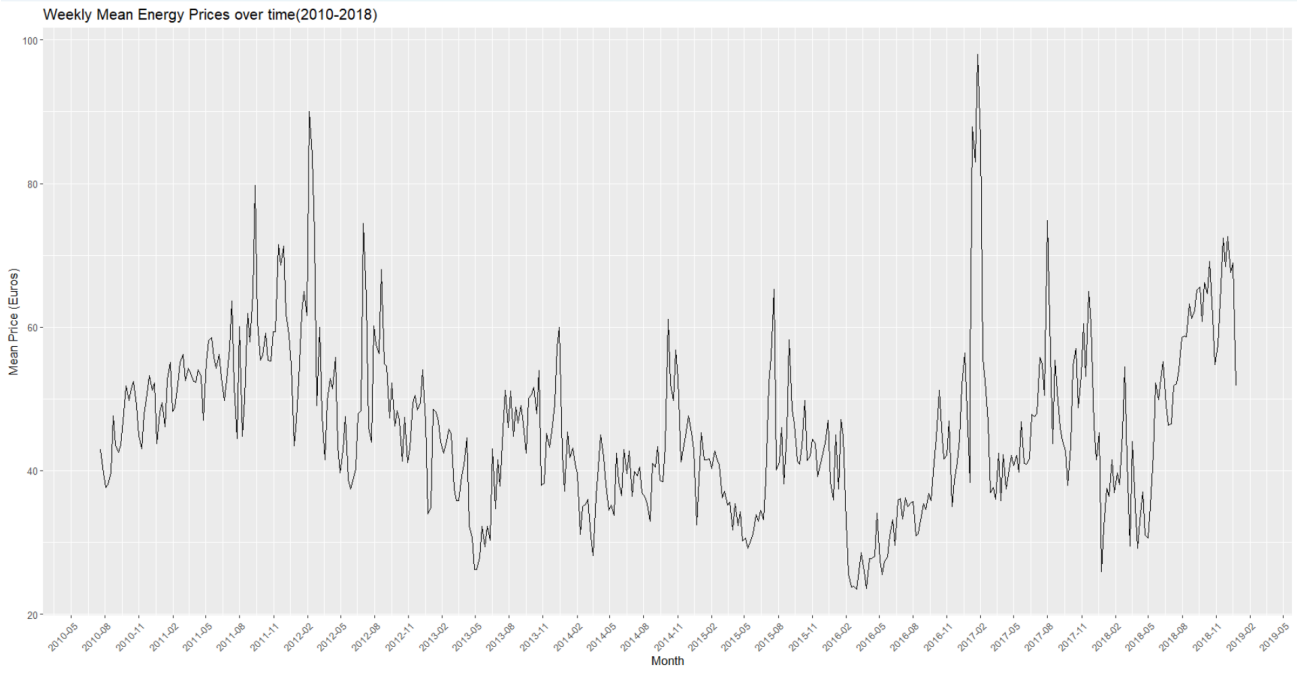


After that in order to search for stationarity and trend I plotted the decomposed time series (using the additive method). As we can see from the plots given below there is a trend that seems to be quadratic at first glance (because of the curve that is being shaped) which makes the price decrease rapidly from 2012 till 2016 and then starting to rise again from 2017 until the end of the data. This testifies to a lack of stationarity.

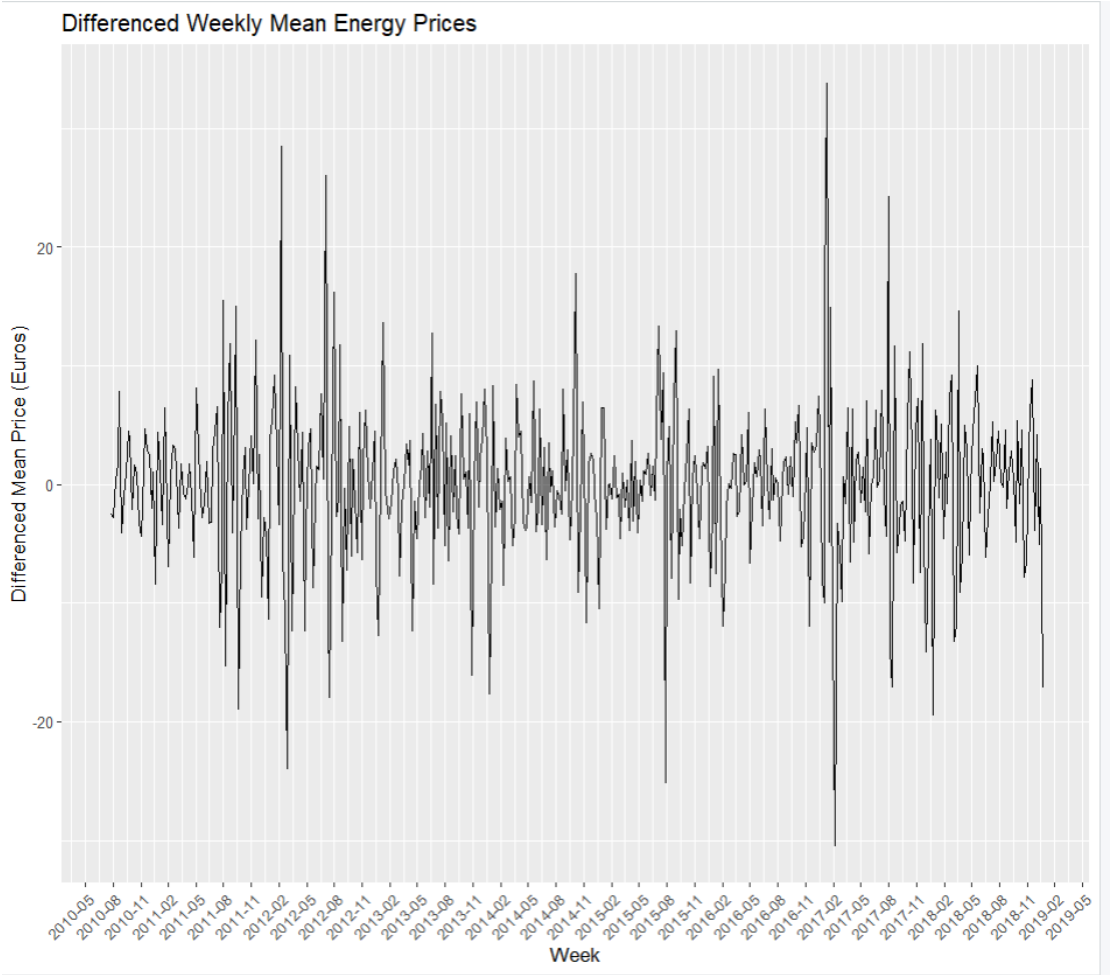
Another thing that shows that weekly data in their original form are not stationary is the volatility that seems to be present at some points in time for both the observed data and the random component (it can be seen approximately around 2012 and 2017).



These indications for lack of stationarity are also shown in the plotted weekly prices over time below. This lack of stable variance is now more apparent, and the trend can also be seen (although not as clearly as in the previous decomposed plot). Finally, from these two plots, there are indications for seasonality because there seems to be an increase in the prices in November, January, and February of each year and a decrease in the prices for March, April, and August.

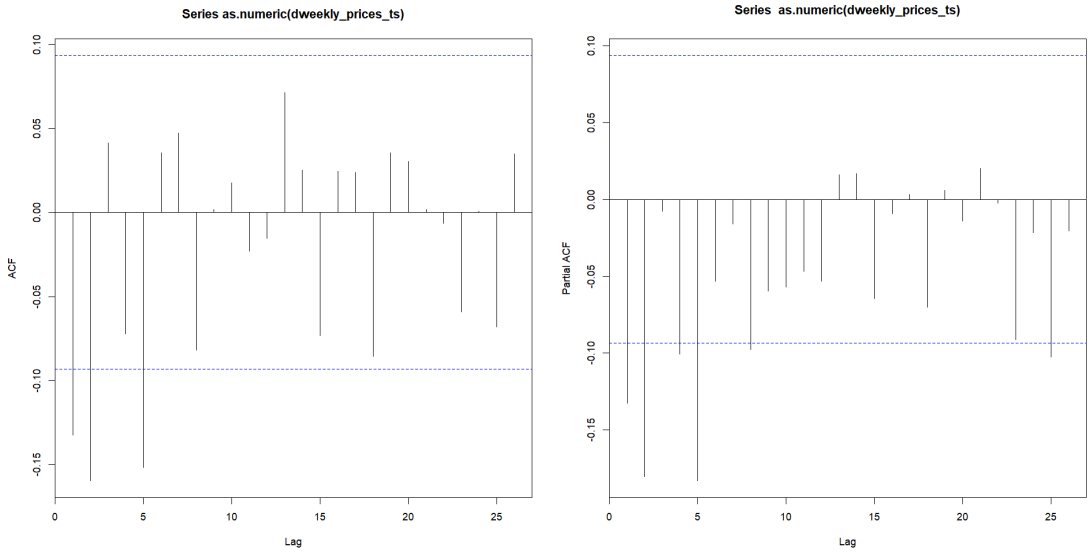


After studying the data visually I also implemented the KPSS statistical test for stationarity. KPSS test rejects the null hypothesis for stationarity (non-presence of unit root) with p-value= 0.01<0.05. So given the above considerations, I concluded that stationarity is not a reasonable assumption for the weekly prices time series so a transformation should be tried in order to treat this problem and be able to fit Box-Jenkins models and search for the best-one. After that, I took first-order differences (i.e. taking the difference of each weekly price with the price of the previous week) on the weekly time series to achieve stationarity. So, I checked once again the statistical tests for stationarity and found out that stationarity is a reasonable assumption for the differenced time series. In this case, the KPSS test gave a p-value=0.1>0.05 and the Augmented Dickey-Fuller test (H0: the presence of unit root, H1: stationarity) gave a p-value=0.01<0.05 so the null hypothesis is being rejected. We now observe that the trend has been removed and the time series is stationary also the variance presents still some volatility but seems to be better now.



Model training and comparison

Having made the time series stationary I began trying to find what model should I use and what should be its order. To do that I generated an autocorrelation and a partial autocorrelation plot that I am showing below to see if I can observe some indications for the model that I should start by locating significant autocorrelations. In the plots below we observe that the autocorrelation in the ACF plot decreases suddenly after lag=5 and then it never surpasses the confidence intervals, in many cases it is close to 0 and even when it starts having an increase in some lags (like 15 or 18) these autocorrelations don't seem to be significant. PACF plot on the other hand seems to decay more gradually and tail off. Although the biggest partial autocorrelations are met in the first 5 lags there are 3-4 cases that partial autocorrelations reach or surpass the confidence intervals (e.g. lag 8, 23,25). With this way of thinking I chose to work with Moving Averages models in the beginning because in cases where ACF plot cuts suddenly at point q and PACF tails off, MA models might be a good approach.



The first model that I fitted in the training data was the MA(5) because at lag=5 the ACF plot presents the last significant autocorrelation. This MA model turned out to have an AIC score equal to 2693.9. The mathematical formulation of this model is the following:

$$y_t = -0.014 - 0.225\varepsilon_{t-1} - 0.216\varepsilon_{t-2} - 0.028\varepsilon_{t-3} - 0.144\varepsilon_{t-4} - 0.160\varepsilon_{\tau-5} + \varepsilon_t$$

After fitting this model I tried to fit two more MA(q) models to see if I can achieve a lower AIC score. So I fitted the MA models for lag=2 (the most significant autocorrelation along with lag=5) and for lag=8 (this autocorrelation is not significant but tends to reach the line of the confidence interval) so I searched if the removal or further addition of error terms will result in a better data fitting. I compared these two models with the MA(5) but it remained the best MA model based on its AIC score. The AIC scores of the three MA(q) models are shown in the table below:

MA(q)	AIC
MA(2)	2709.94
MA(5)	2693.90
MA(8)	2696.05

Seeing that MA(5) was the best MA model that I could find I wanted to search if there was an ARMA model that can fit better in the training data. To do that I implemented a double iterative search in R which searched every combination of p and q where p was taking values in a range [0,8] and q values in a range [0,13] the best model that came from this was an ARMA(3,3). This model had an AIC score of 2688.43 so I concluded that it was a better fit in the data from the MA(5) model, so ARMA(3,3) is my selected model for now.

Out-of-sample predictive ability

To see the out-of-sample performance of the selected model I kept a testing dataset that corresponds approximately to the last 9 months of the time series data. To the accuracy of the prediction I used the Mean Absolute Error and Root Mean Squared Error. These two had values equal to 4.82 and 5.99 respectively. This indicates a mediocre predictive ability. But to get a better view I normalized the MAE based on the range of values that differenced prices take. To do that I used the following formula and got the corresponding result:

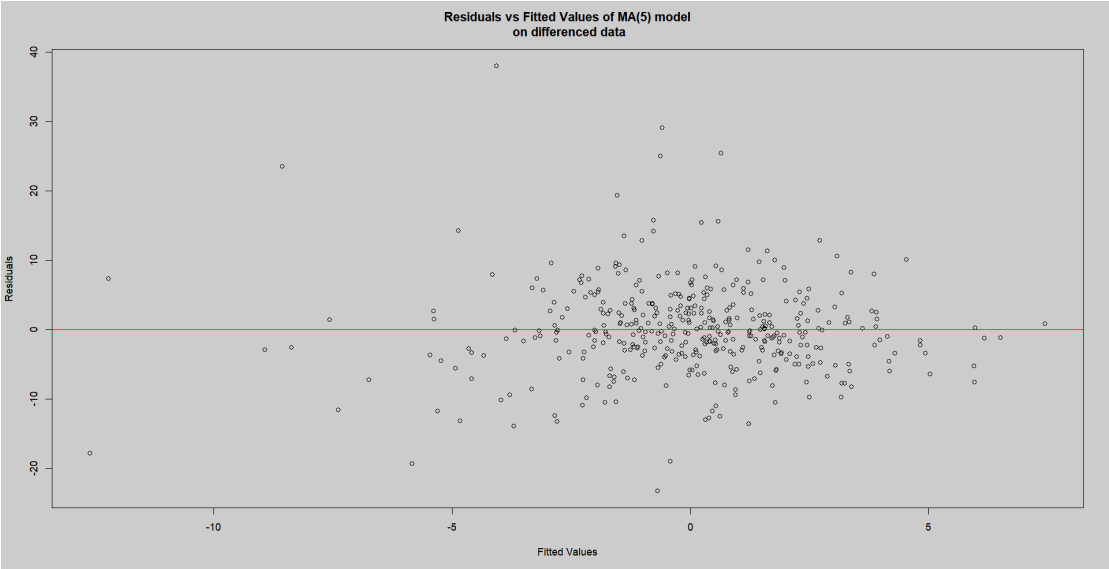
$$\frac{MAE}{\max(price\ in\ the\ test\ data)-\min(price\ in\ the\ test\ data)} = 17.8\%$$

This testifies indeed for a mediocre predictive ability because it means that the average error is 17% of the range of the differenced prices. When I used the same formula to assess the prediction of the MA(5) model the result that I got back was 15.2% which testifies to a slightly (but not significantly) better out-of-sample predictive ability despite the smaller AIC value of the ARMA model in the training set. So in the next part, I will check the assumptions for the residuals of the MA(5) model.

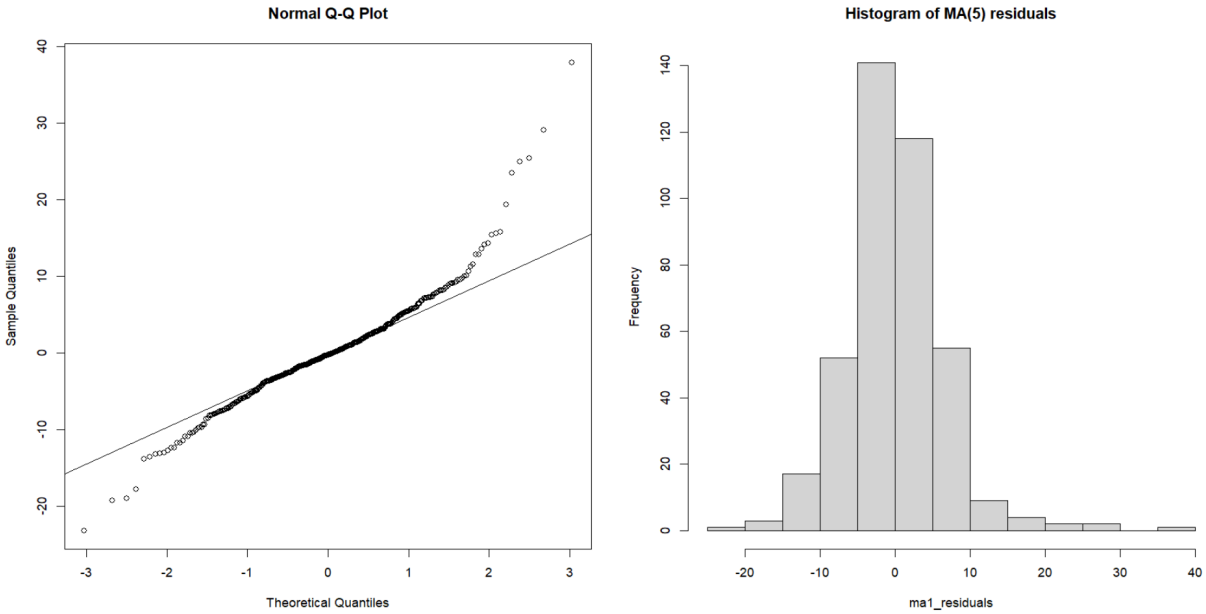
Checking the diagnostics of the residuals

From theory, we know that the residuals of a selected predictive time series model must resemble white noise so we want them to be uncorrelated, homoscedastic, and normally distributed. For the trained model the residuals can be assumed to be uncorrelated using the Ljung-Box I failed to reject the null hypothesis for the lack of correlation between the residuals with p-value=0.998 > 0.05, so this assumption is met. To check the constant variance assumption, I plotted the fitted values of the model against the residuals. In the plot below we can see that the residuals do not present any unusual pattern and do not form clusters, but they are rather scattered randomly in the plot. These characteristics indicate for the presence

of homoscedasticity, also using Arch LM test the null hypothesis for homoscedasticity cannot be rejected (p-value=0.17>0.05).



Although the assumptions for lack of correlation and homoscedasticity of the residuals seem to hold there seems to be an issue with the normality assumption. As we can see from the QQplot and the histogram below the residuals are not far from normality but both these plots are heavy on the tails and present a skewness. Also Kolmogorov-Smirnoff test rejects the null hypothesis for normality (p-value=0.01<0.05). So normally distributed residuals is not a reasonable assumption.

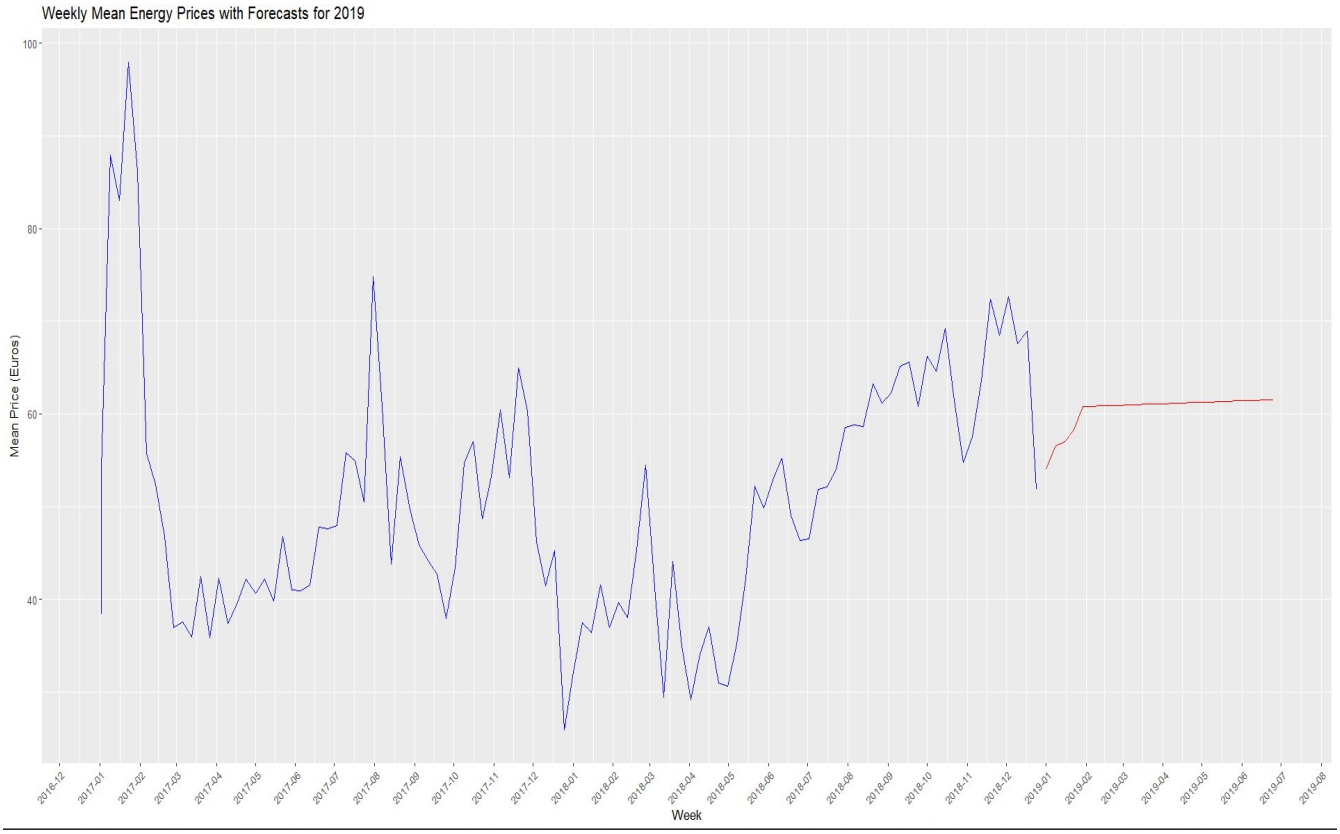


Forecasting

Finally, after using all the time-series data to make the forecasting for the first semester of 2019 the predicted prices are shown in the following table:

Month	Predicted Price
January	57.32
February	60.87
March	61.02
April	61.18
May	61.35
June	61.49

These values were obtained after forecasting 26 weeks ahead and averaging their prices for each month. The evolution of the prices is also shown in the line chart below:



We can see both in the table and in the chart (in which we see the evolution of the historical prices for the years 2017-2018 with the blue line and the predicted values with the red line) that the price of the kWh will rise in the first semester of 2019 more rapidly in January and February and more slightly during the spring months where it will tend to stabilize.