Universidad Carlos III de Madrid

MSc in Statistics for Data Science

Simulation Project

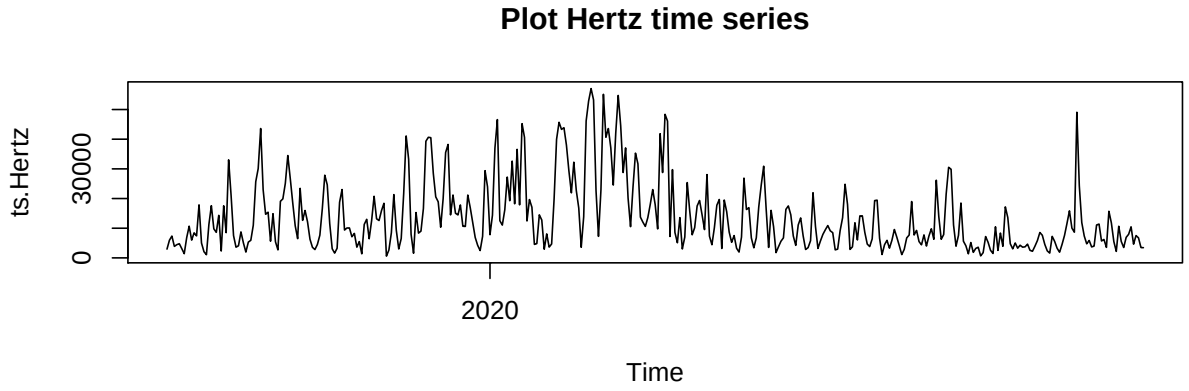
Simulation and resampling

Authors:
David de la Fuente López
Diego Perán Vacas

10th March 2021

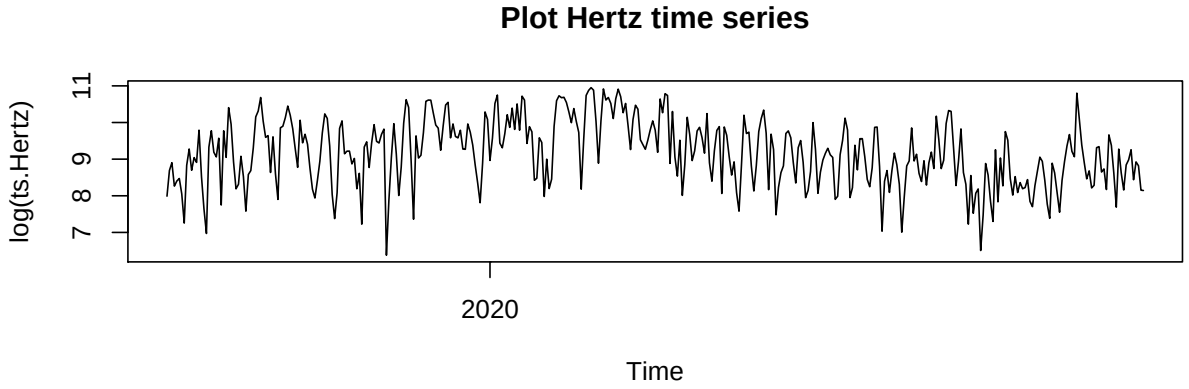
Hertz time series

Time series without any transformation



Time series after logarithm transformation

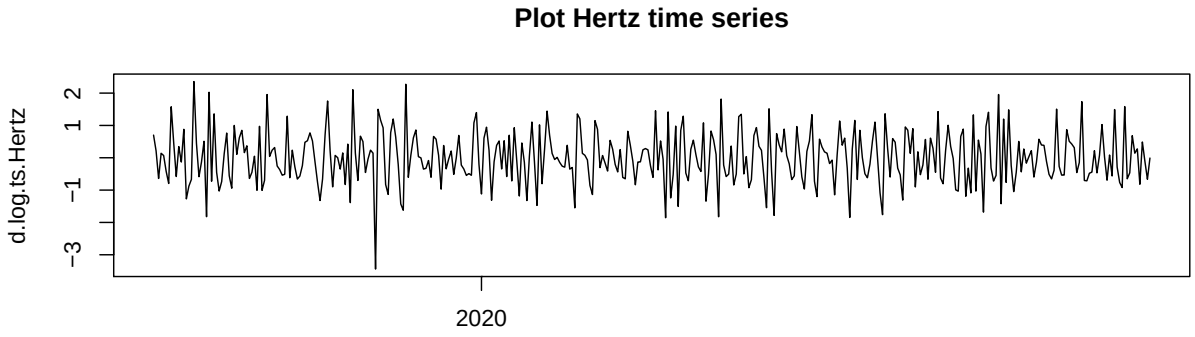
We take the logarithm transformation.

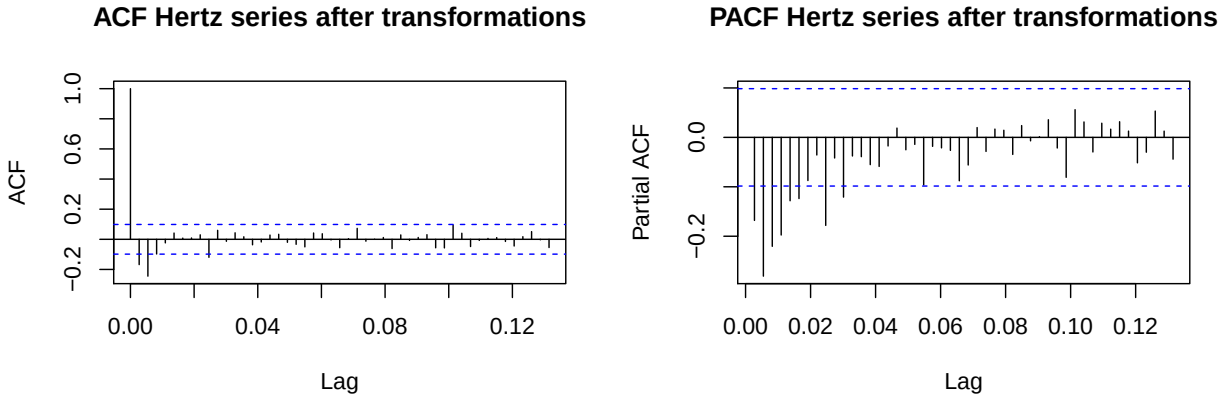


We see some trend in the first part of the time series. In this time interval, it seems the unconditional mean of the series tends to increase. Then, after the second month of 2020 (in particular, the maximum of the series is reached at the day 173. Taking into account we start on the day 235 of 2019, this corresponds to the twelfth of February), it starts a new trend in the time series. In this second time interval, the time series has a decreasing trend until it finds a new peak at time 370 (which corresponds to 27th of August 2020). In fact, we consider this peak as a candidate to be treated as outlier.

Due to this trendy behavior (although the trend changes over the time periods), we consider it is justified to take the first difference on the series (after having applying logarithm transformation).

Time series after log and differences





First, note the point which was a candidate of outlier does not appear anymore after considering the first order difference in the regular part.

Second, at this point, we can get some insights from the ACF and PACF. On the one hand, the PACF presents clearly a slow decay pattern towards zero from the second lag. On the other hand, the ACF shows only two significant peaks (apart from the first one), at the first and second lag. This might indicate we are facing a MA(2) process. If we consider the series before applying the difference (i.e. just applying the logarithm transformation), we will be facing an ARIMA(0,1,2) process.

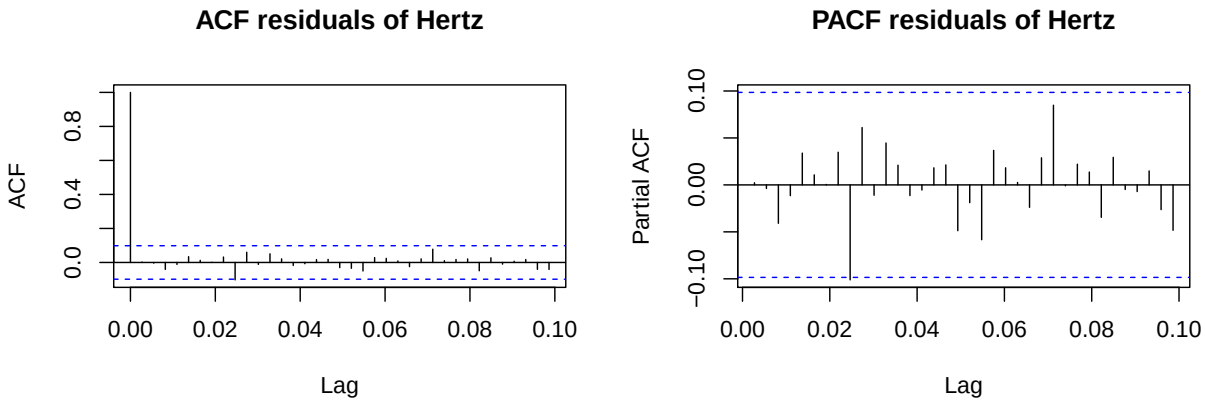
At this point, we will run auto.arima() to verify our considerations (first, we do not consider outliers).

```
## Series: d.log.ts.Hertz
## ARIMA(1,0,2) with zero mean
##
## Coefficients:
##          ar1      ma1      ma2
##       0.2020  -0.6399  -0.2868
## s.e.  0.1073   0.1033   0.0908
##
## sigma^2 estimated as 0.5014:  log likelihood=-424.48
## AIC=856.96   AICc=857.07   BIC=872.89
```

The function detects an ARIMA(1,0,2) model. The AR(1) part may be justified in the previous plots by the fact the first peak in the PACF is smaller than the second one. This is not an expected result for a simple

MA(2) model. In the case the model was simply an AR(1), we should find one significant peak on the PACF (the first one) and then none of the rest significant; and a decay pattern in the ACF from the first peak. Since here we have an ARMA(1,2) model, the PACF is contaminated with the slow decay pattern from the second peak, characteristic of the MA(2) part. This is the reason why it is complicated to detect an ARMA(1,2) model by visual inspection.

Diagnosis of the model

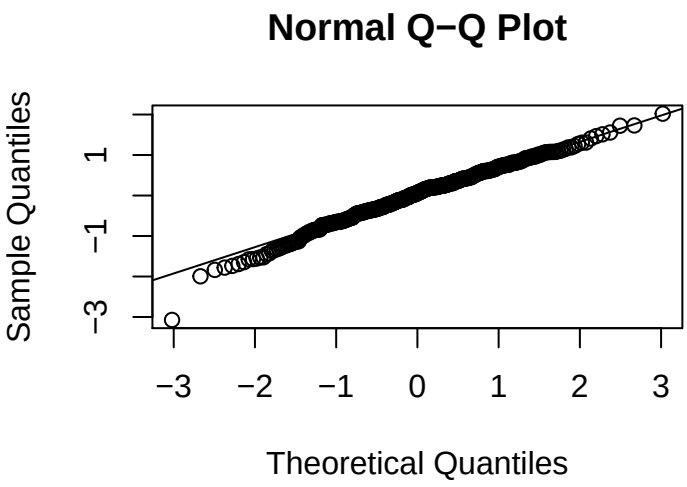



```
##
##  Box-Ljung test
##
## data:  res.Hertz
## X-squared = 0.0020104, df = 1, p-value = 0.9642
```

The ACF and PACF of the residuals show there is no correlation remaining to be explained. The PACF presents just one significant peak, but with a quite small value. In the case of the ACF, we show the same situation, just one peak that may be considered as significant, but right in the limit to do so.

The Ljung-Box test yields to a very high p-value, which does not support the rejection of the null hypothesis. Therefore, we can consider there is no correlation left to be explained.

Let's check know the normality of the residuals.




```
## [[1]]
##
##  Jarque Bera Test
##
## data:  res.Hertz
## X-squared = 17.197, df = 2, p-value = 0.0001844
```

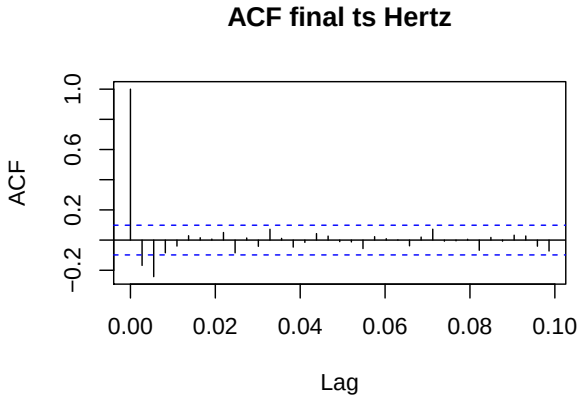
However, in the Jarque-Bera test for normality, we get a small p-value, which supports the rejection of the null hypothesis. Therefore, the residuals are not normal, and the ARIMA(1,0,2) model is not enough to explain the time series.

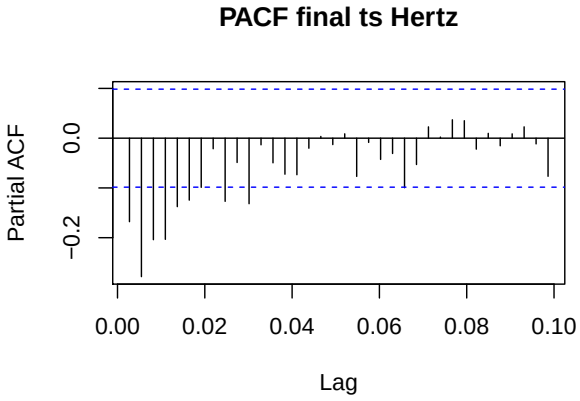
Considering outliers

```
## Series: d.log.ts.Hertz
## Regression with ARIMA(0,0,2) errors
##
## Coefficients:
##           ma1      ma2     IO89
##       -0.4607  -0.4183  -3.0559
## s.e.   0.0433   0.0431   0.6972
##
## sigma^2 estimated as 0.4824:  log likelihood=-416.74
## AIC=841.47   AICc=841.58   BIC=857.4
##
## Outliers:
##   type ind     time coefhat  tstat
## 1   IO  89 2019:324  -3.056 -4.383
```

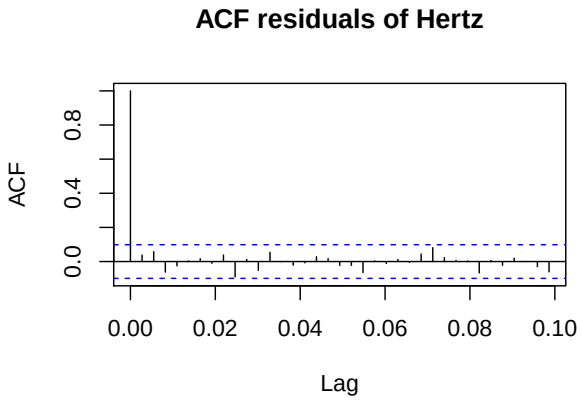
At this point, we have considered the presence of outliers in our time series. We accept four types of outliers: additive, innovative, level shift and transitory change. The method identifies an outlier at time 89. Since the time series begins in the day 235 of 2019, this outlier appears in the day 324 of 2019. This corresponds to the twentieth of November. Note in the plot of the time series after having applied logarithms and the first order regular difference, this outlier clearly appears.

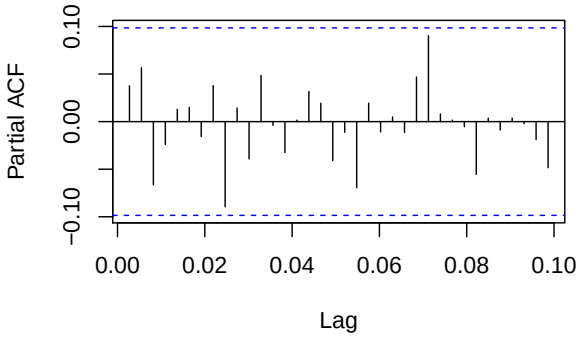
We present the plots of the resulting time series after: taking logarithm, doing the first regular difference, and replacing the outlier.





Second model diagnosis

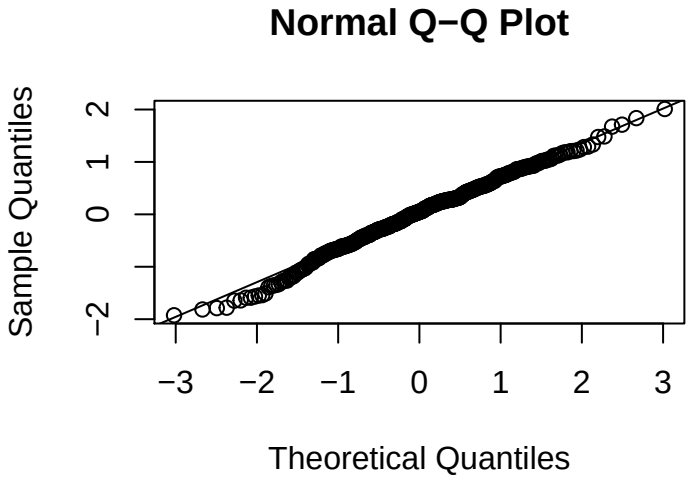





```
##
##  Box-Ljung test
##
## data:  mod.out.Hertz$fit$residuals
## X-squared = 0.56531, df = 1, p-value = 0.4521
```

First, we check the ACF and PACF of the residuals and it is clear there is no correlation left to explain in

our final time series of the Hertz company. The Ljung-Box test gives us the same insights.

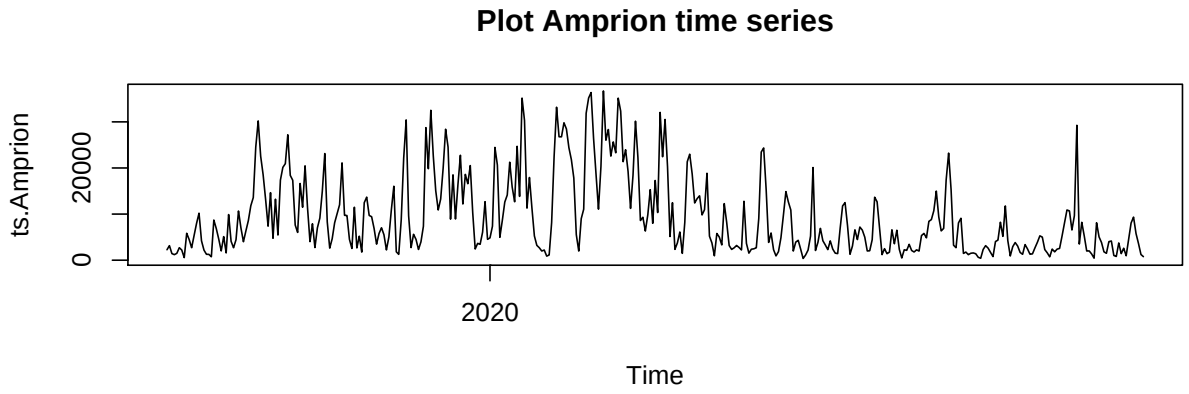



```
## [[1]]
##
##  Jarque Bera Test
##
## data:  mod.out.Hertz$fit$residuals
## X-squared = 3.0868, df = 2, p-value = 0.2137
```

Second, with respect to the normality Jarque-Bera test, we get a high p-value, indicating the residuals are normally distributed. In addition, in the qq-plot we see an improvement due to the removal of the worse residual (with certainty related to the outlier which has been removed). This implies our time series can be described as an ARIMA(0,0,2) after replacing one outlier and doing the two transformation that we have considered.

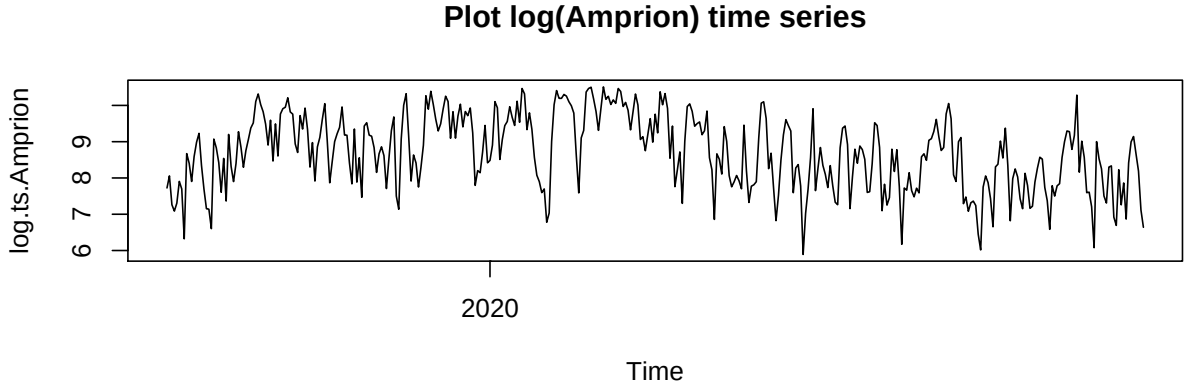
Amprion time series

It is now the turn to do the study for the energy company Amprion. To do this, we begin by analyzing the behavior of this time series through its plot without any transformation in it.



Time series after logarithm transformation

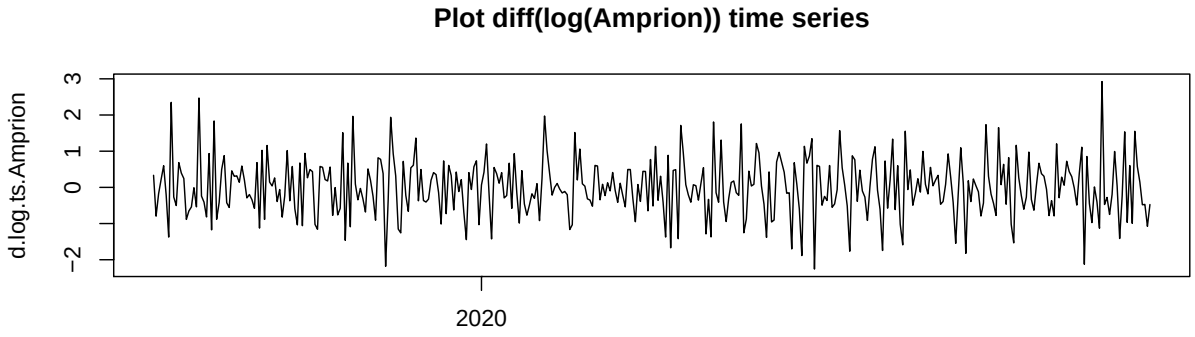
Now, as we have done in the previous cases, we apply the logarithm in order to decrease the variance of the time series , since observing the previous plot we verify that the variance varies over time.

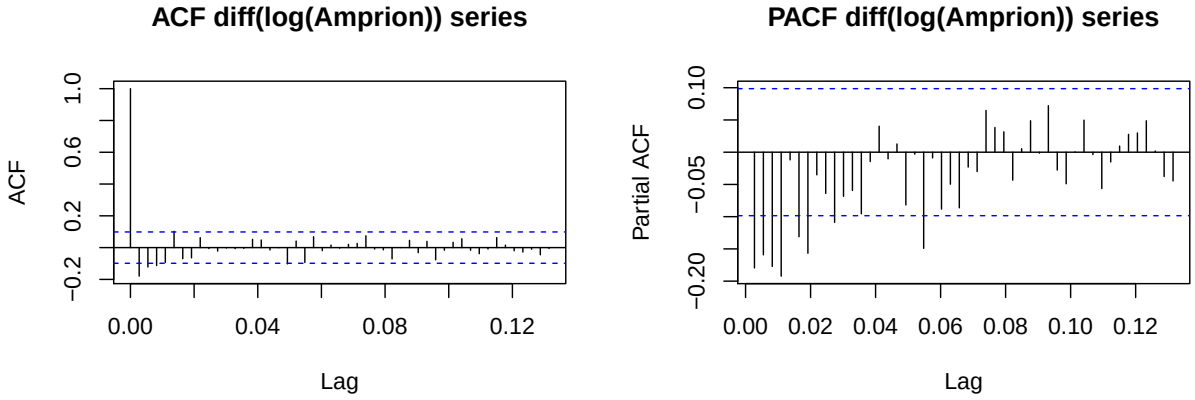


If you look closely, the plot shows various trend changes over time. At the beginning of the time series it seems clear that there is an upward trend until it stabilizes at values higher than the initial ones. Then, in 2020 it seems that there is a downward trend until it returns to values similar to the initial ones.

It is also because of these trend changes that we apply the first difference to the time series as in the previous cases.

Time series after log and differences





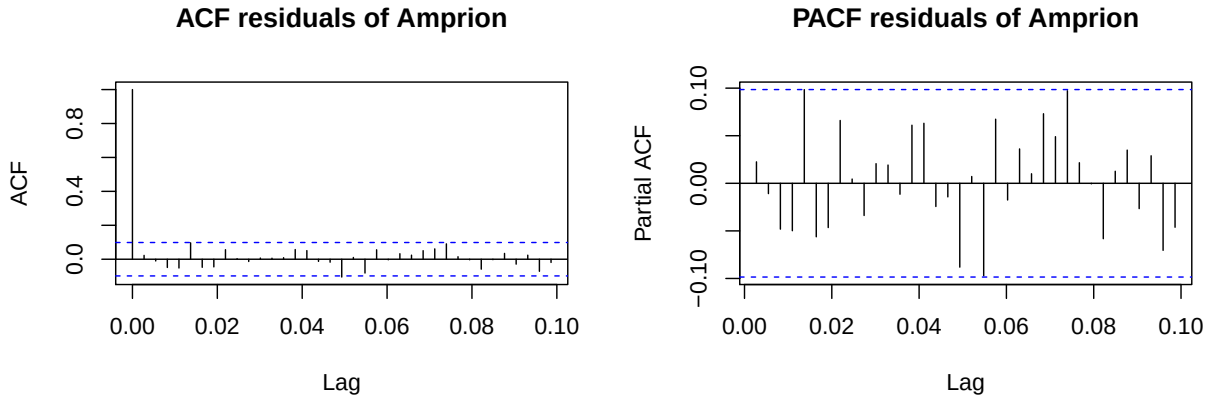
After these two transformations of this time series we can already draw some conclusions from the correlograms. In the first place, there are three significant peaks in the ACF, although the first is more significant, and perhaps the other two should not be taken into account (since they are at the limit). Second, the partical correlogram clearly shows a slow decay. Due to these two characteristics, it seems clear that we are dealing with an MA, although the value of q is not clear (it seems to be 1).

As we are not yet clear about the value of q, we perform the auto.arima() to verify what we have predicted. It must be remembered that this initial study is without taking into account the presence of outliers.

```
## Series: d.log.ts.Amprion
## ARIMA(1,0,1) with zero mean
##
## Coefficients:
##          ar1      ma1
##       0.5735  -0.9548
## s.e.  0.0491   0.0187
##
## sigma^2 estimated as 0.5244:  log likelihood=-433.69
## AIC=873.39   AICc=873.45   BIC=885.33
```

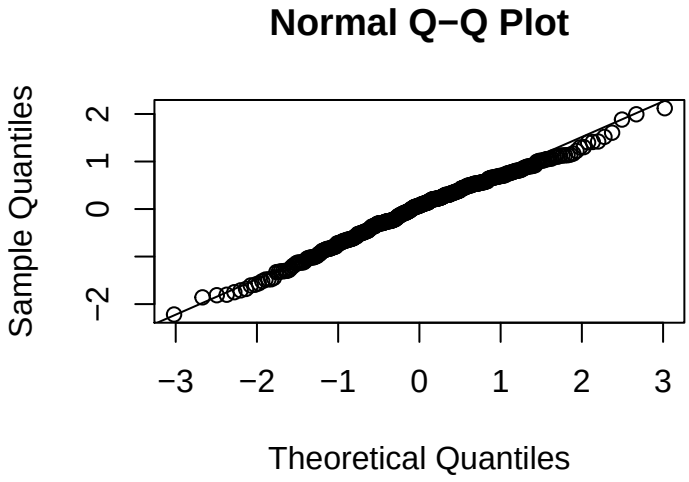
This function implemented by R detects an ARIMA(1,0,1) (taking into account the application of the first difference).It is possible that the presence of AR(1) in the correlograms can be predicted by the slight slow decay that exists in the ACF. Said slow decay does not allow us to guess the q value of the MA, since if it is a pure MA there would only be q significant peaks in the ACF.

Diagnosis of the model




```
##
##  Box-Ljung test
##
## data:  res.Amprion
## X-squared = 0.20292, df = 1, p-value = 0.6524
```

From what we see in the two correlograms of the residuals, no correlation appears to remain. The ACF may have a significant peak, but it appears to be very small. On the other hand, the PACF does not show any significant peaks. The Ljung-Box test presented a very high p-value, so we do not reject the null hypothesis and we can consider that there is no correlation to be explained. Let us check the normality of these residues.




```
## [[1]]
##
##  Jarque Bera Test
##
## data:  res.Amprion
## X-squared = 4.2011, df = 2, p-value = 0.1224
```

In this case, using the Jarque-Bera test to test normality, we find a p-value large enough not to reject the

null hypothesis. Therefore, in this case the residuals are normal.

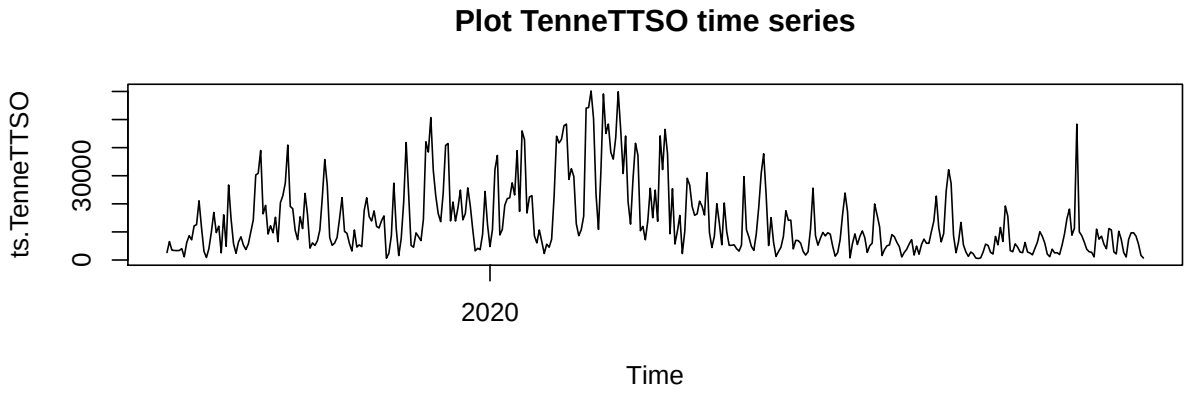
Considering outliers

```
## Series:
## ARIMA(1,0,1) with zero mean
##
## Coefficients:
##          ar1      ma1
##       0.5735  -0.9548
## s.e.  0.0491   0.0187
##
## sigma^2 estimated as 0.5244:  log likelihood=-433.69
## AIC=873.39   AICc=873.45   BIC=885.33
##
## No outliers were detected.
```

In this case, of the four types of outliers that we have chosen as possible, the method has not detected any. When visualizing the plot of the time series, outliers did not seem to exist at first glance and the method has corraborated this, therefore it is not necessary to make a separate diagnosis.

TenneTTSO time series

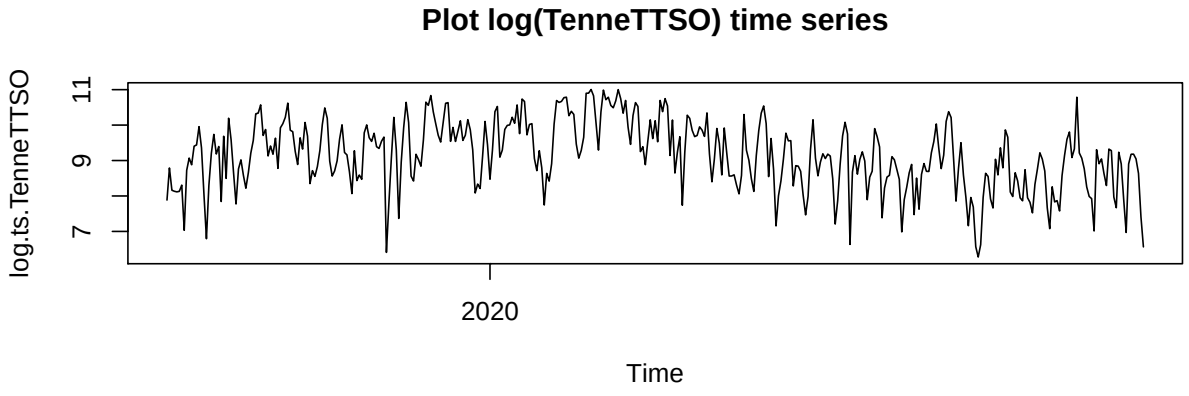
Let start with the analysis of the time series corresponding to the energy company TenneTTSO. To do this we will follow the same steps as in the previous examples.



The plot clearly shows abrupt changes in the variance of the series over time. The time series begins with lower values but in the final months of 2019 and the first of 2020 it experiences a change in its variance and begins to take higher values. After a while the time series seem to return to their initial values. To avoid this change in variance we apply the logarithm.

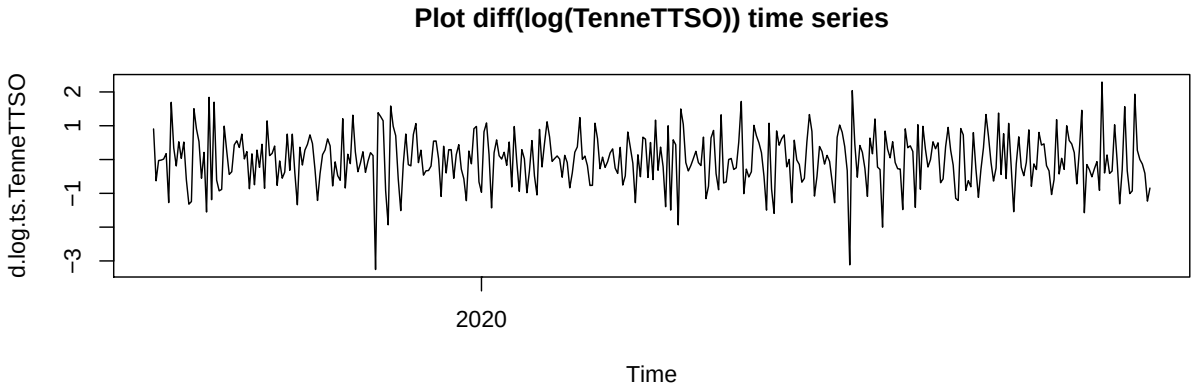
Time series after logarithm transformation

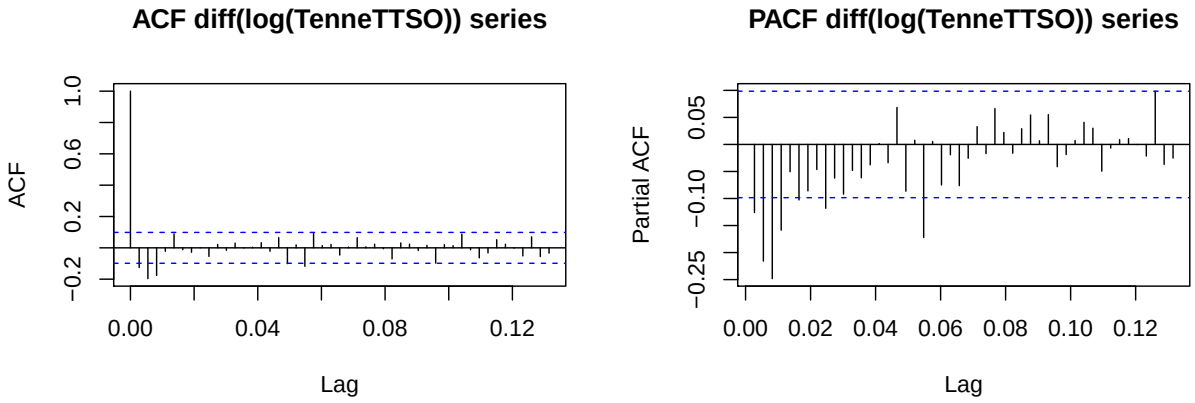
To avoid this change in variance we apply the logarithm to the time series.



Time series after log and differences

Then, since there are obvious changes in the trend in the previous plot, it is justified to use the first difference in order to eliminate that trend.





Now, since we have performed all the necessary transformations to the time series, the corresponding correlograms closely resemble those of the previous case. Therefore, due to the apparent slow decay in the partial correlogram and the three significant peaks in the ACF, it seems that we are facing an MA (3). However, due to the similarity to the previous time series, it is possible that it is a ARMA.

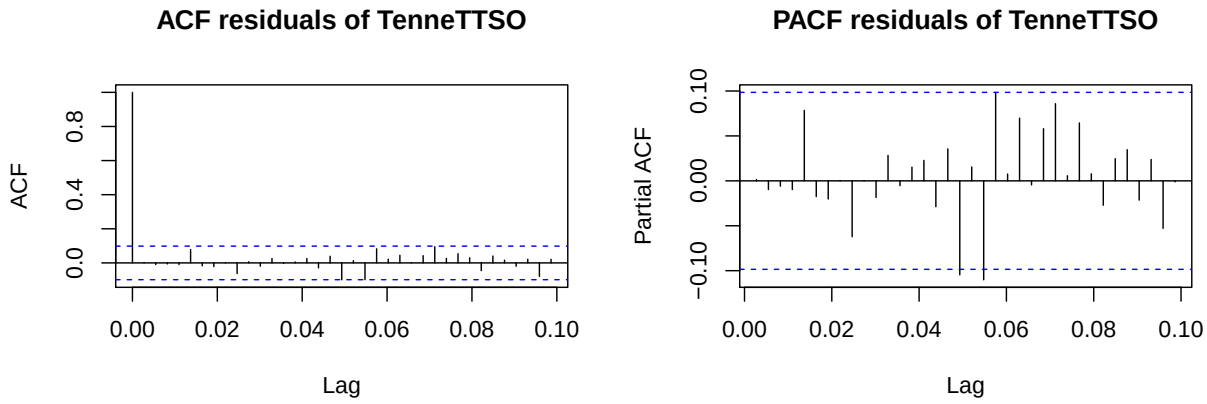
Otherwise, looking now at the plot of the time series we transform we find two candidate outlier points. Later, when we do the outlier analysis we will see if they actually turn out to be so.

```
## Series: d.log.ts.TenneTTSO
## ARIMA(1,0,3) with zero mean
##
## Coefficients:
##          ar1      ma1      ma2      ma3
##       0.0863  -0.4335  -0.2822  -0.1909
## s.e.  0.1856   0.1798   0.0897   0.0957
##
## sigma^2 estimated as 0.4867:  log likelihood=-418.02
## AIC=846.05   AICc=846.2   BIC=865.95
```

Indeed, as we suspected, we find an ARIMA (1,0,3). As in the previous example, it is not a pure MA. We had predicted this time the existence of the MA and the value of q but the model estimated by the method also predicts an AR (1).

Diagnosis of the model

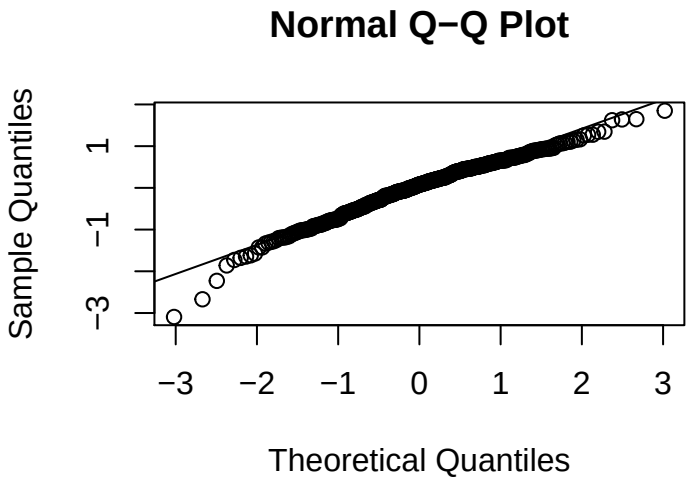
Let's do the diagnosis of this model:




```
##
##  Box-Ljung test
##
## data:  res.TenneTTSO
## X-squared = 0.00081949, df = 1, p-value = 0.9772
```

In the first place, the ACF has a small significant peak, but in principle it could be accidental because it is very small. However the PACF also has a significant peak and this time of considerable size. Therefore, we can conclude that can be some correlation to be explained in the model and that the estimate is not enough.

However, the Box-Ljung test, as in the previous examples, presents a very high pvalue, which allows us not to reject the null hypothesis. Therefore, we can consider that there is no correlation to be explained in the estimated model. Let's now check the normality of the residuals.




```
## [[1]]
##
##  Jarque Bera Test
##
```

```
## data:  res.TenneTTSO
## X-squared = 41.378, df = 2, p-value = 1.035e-09
```

We once again performed the Jarque-Bera test, which provides us with a remarkably low pvalue, which allows us to reject the null hypothesis. Therefore, we conclude that the residuals are not normal.

Considering outlier

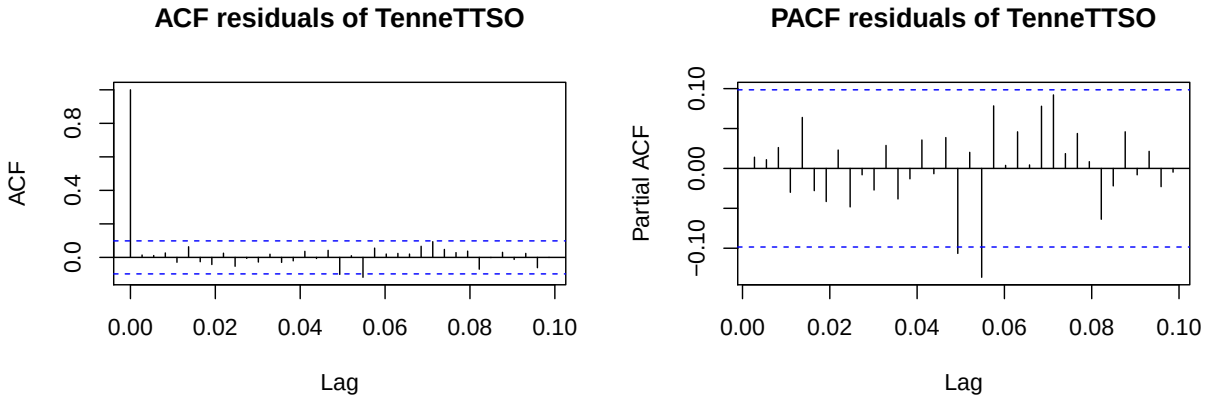
Before applying the R function called tso() in order to estimate the model, this time considering the possible outliers, it is worth remembering the two candidates for outliers detected in the previous analysis of the series.

```
## Series: d.log.ts.TenneTTSO
## Regression with ARIMA(0,0,3) errors
##
## Coefficients:
##           ma1      ma2      ma3     IO89    IO277
##       -0.3262  -0.3120  -0.2448  -3.0977  -2.6626
## s.e.   0.0478   0.0547   0.0516   0.6669   0.6705
##
## sigma^2 estimated as 0.4462:  log likelihood=-400.29
## AIC=812.57   AICc=812.79   BIC=836.46
##
## Outliers:
##   type ind     time coefhat  tstat
## 1   IO  89 2019:324  -3.098 -4.645
## 2   IO 277 2020:147  -2.663 -3.971
```

In this case, the method finds two outliers, just the ones we had predicted. In the first place, it should be noted that considering the outliers the method now estimates an ARIMA (0,0,3), that is, the part of the AR disappears and becomes an MA (3).

Second, outliers are where we suspected, exactly on the days 20th of November and 26th of May. The type of both outliers is an IO (innovation outliers) and their estimated coefficients are considerably large. Due to the existence of these large outliers, it is necessary to repeat a diagnosis of the model, this time taking it into account.

Second model diagnosis

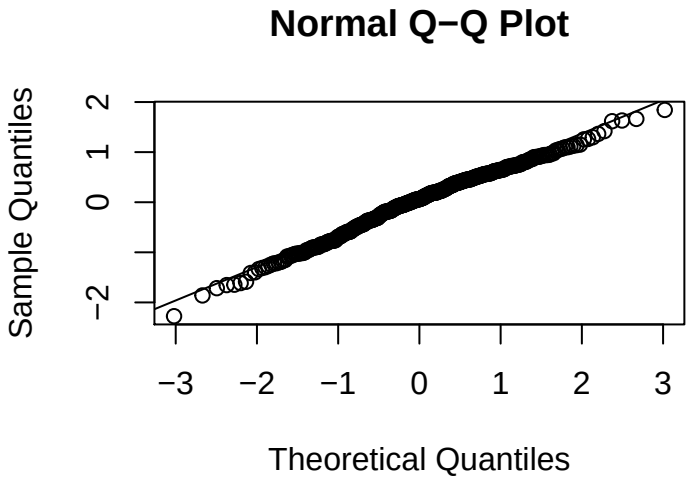



```
##
```

```
##  Box-Ljung test
##
## data:  mod.out.TenneTTSO$fit$residuals
## X-squared = 0.078876, df = 1, p-value = 0.7788
```

First, in the ACF there seem to be no significant peaks to take into account, but in the PACF we find two significant peaks. However, when performing the Box-Ljung test, the p value resulting from the test is very high, so we do not reject the null hypothesis. In this way, we can consider that the model does not leave any correlation unexplained.

Let's check the normality of the residuals.




```
## [[1]]
##
##  Jarque Bera Test
##
## data:  mod.out.TenneTTSO$fit$residuals
## X-squared = 6.7674, df = 2, p-value = 0.03392
```

Focusing on the results of the Jarque-Bera test, we verify that the p-value has increased compared to the test with the model without considering outliers. Despite this, if we consider a $\alpha = 0.05$, the pvalue is still low enough to reject the null hypothesis, and therefore the residuals would not be normal.

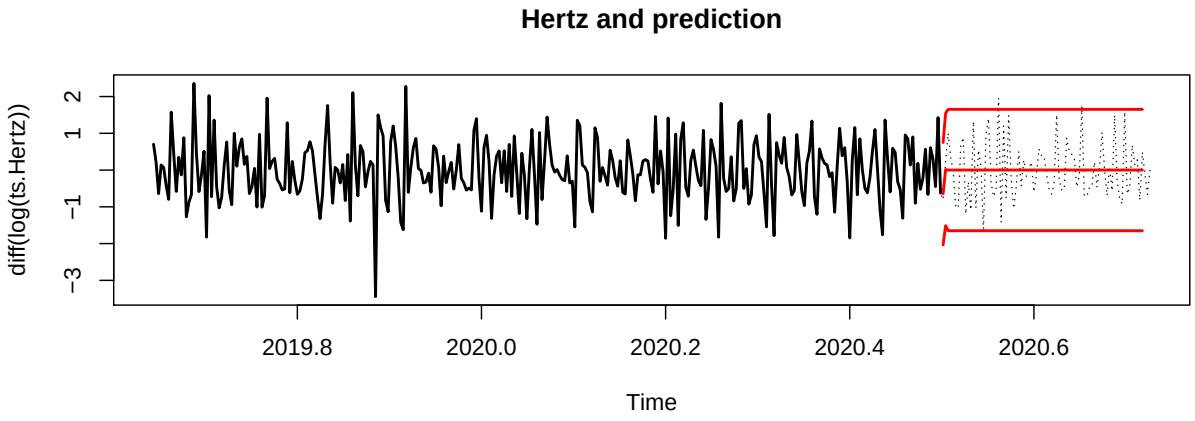
Forecasting

We continue the study of this set of time series with their forecasting. In this section we are only going to forecast models that have turned out to be good enough, that is, in which the estimated model is sufficient to estimate the time series.

As we have been explaining throughout the project, we consider that a model is sufficient to explain a time series when, after performing the model diagnostic, we do not reject either of the two null hypotheses of the tests performed. This condition is fulfilled in two cases: the estimation of the Hertz time series considering outliers and the estimation of the model for Amprion.

Forecasting of Hertz

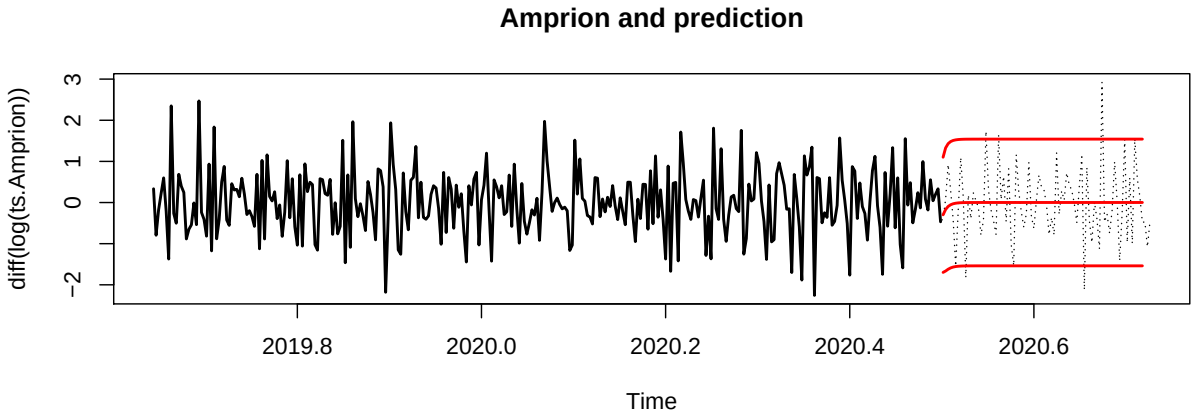
For the time series corresponding to the Hertz company, we use the estimated model considering the previously detected outlier. This method estimated an ARIMA (0,0,2), that is, an MA (2) (after applying logarithms and the first difference).



As we can see in the plot, the prediction of the time series has the characteristics of a forecasting on an MA. There seems to be 2 different periods to 0 and from there the best prediction always has the value 0. On the other hand, there are only two peaks that are outside the confidence intervals, otherwise it seems to be a fairly reliable estimate.

Forecasting of Amprion

This example lacked detected outliers, so the model used will be the one estimated with the auto.arima (). In this case the estimated model for the Amprion time series was an ARMA (1,0,1). So let's make the forecast and analyze its behavior.



Being an ARMA process, the behavior of the predictions has part of the characteristics of an MA and an AR. If it were a pure RA, the best estimate would take more periods to stabilize at 0. However, the presence of the MA causes that after a few periods both the best estimate and the confidence intervals stabilize.

Conclusions

The project is based on an in-depth study on three time series corresponding to the production of 3 energy companies. In all three cases, to begin with, the transformation of the initial time series was necessary, in order to eliminate the changes in the variance of the series and the trends. For this we have applied both the logarithm and the first difference. For these previous steps, it was only necessary to observe the plots of the transformed series. It should be noted that already in these plots, certain points that were already candidates for outliers stood out.

Once we had the time series with the necessary transformations, the next step was to try to predict a model that fits well with each series. For this we made use of the correlograms, both the ACF and the PACF. In all three cases we found a correlogram similar to those of an MA. To check if our predictions were correct or not, we made use of the auto.arima () function. We choose the model resulting from this function for the next step.

Later, a model diagnosis was necessary to check to what extent the estimated model explained the time series sufficiently well. To do this, we applied two tests to check if the residuals of the estimated model still maintained some unexplained correlation and if they were normal. If the two null hypotheses of the two tests were not rejects, we could consider that this estimate of the time series was good.

Next, it was necessary to carry out a parallel study, considering the possible outliers of the time series. And as we had done previously, carry out a model diagnosis. In two of the time series outliers were detected, and the estimation of the model taking these outliers into account improved compared to the previous one. These outlers detected by the tso() function and previously displayed in the plot may be due to simple errors when transmitting the information of the companies. However, some of them may be due to days when the company has not been operational and therefore production has fallen to very low or even zero levels.

To finish the study of the three time series, we complete it with a forecasting of the estimated models that fit the time series well. For example, in the case of the Hertz time series, the estimation of the model with an outlier was taken into account, since the prediction of the model with the presence of this outlier would have been less efficient.

Finally, it should be noted the existence of the same trend in all the time series discussed. As these are time series that show the production of an energy company over a year, it is clear that their production increases considerably in winter. That is, the months prior to the beginning of 2020 (November and December 2019)

present higher values than the previous ones, remaining a few months after the beginning of 2020 (until approximately March). The existence of this difference in values taken by the time series could have been detected by the tso () function as a Level Shift, but since it is not an excessively abrupt change, the method has not considered it. It would be interesting to see to what extent the estimation of these time series could be improved considering a LS during the winter months.