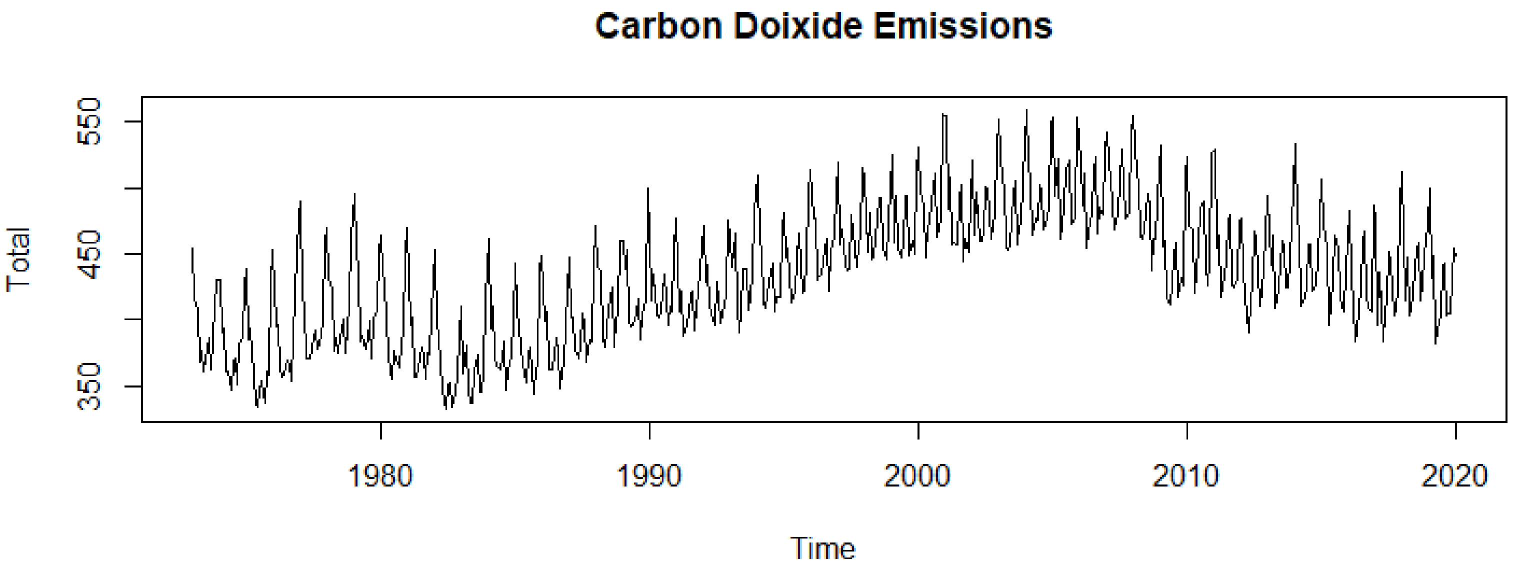
# Study of the Carbon Dioxide Emissions From Energy Consumption in USA time series

Victor Fernandes Malheiro

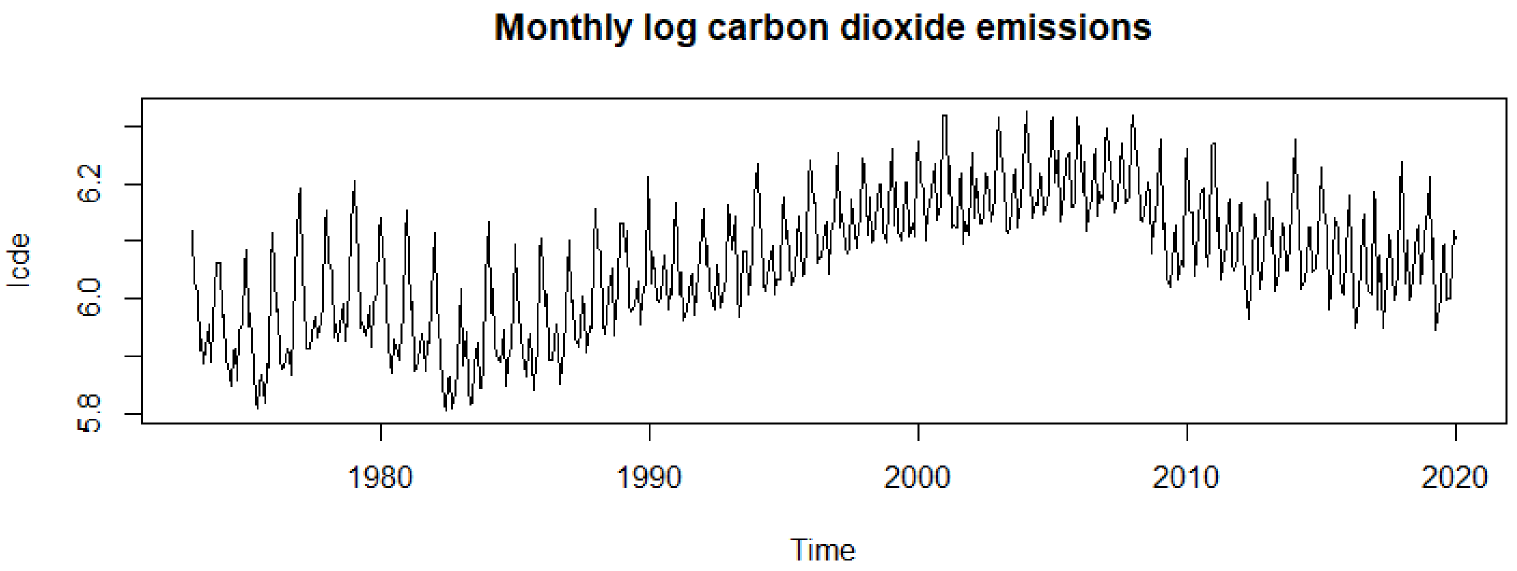
# Introduction

This dataset represents the monthly Total U.S energy related CO2 emissions from January 1973 to January 2020 in million metric tons.
The objective of this assignment was to be able to make reliable forecasts of the last 12 months using different techniques and models.
First, I will analyze the time series checking for seasonality, trend and stationarity. After that, I will apply smoothing methods and decomposition methods to make forecasts.
In the end, I will try to fit the time series into SARIMA models that will allow me to make good forecasts.

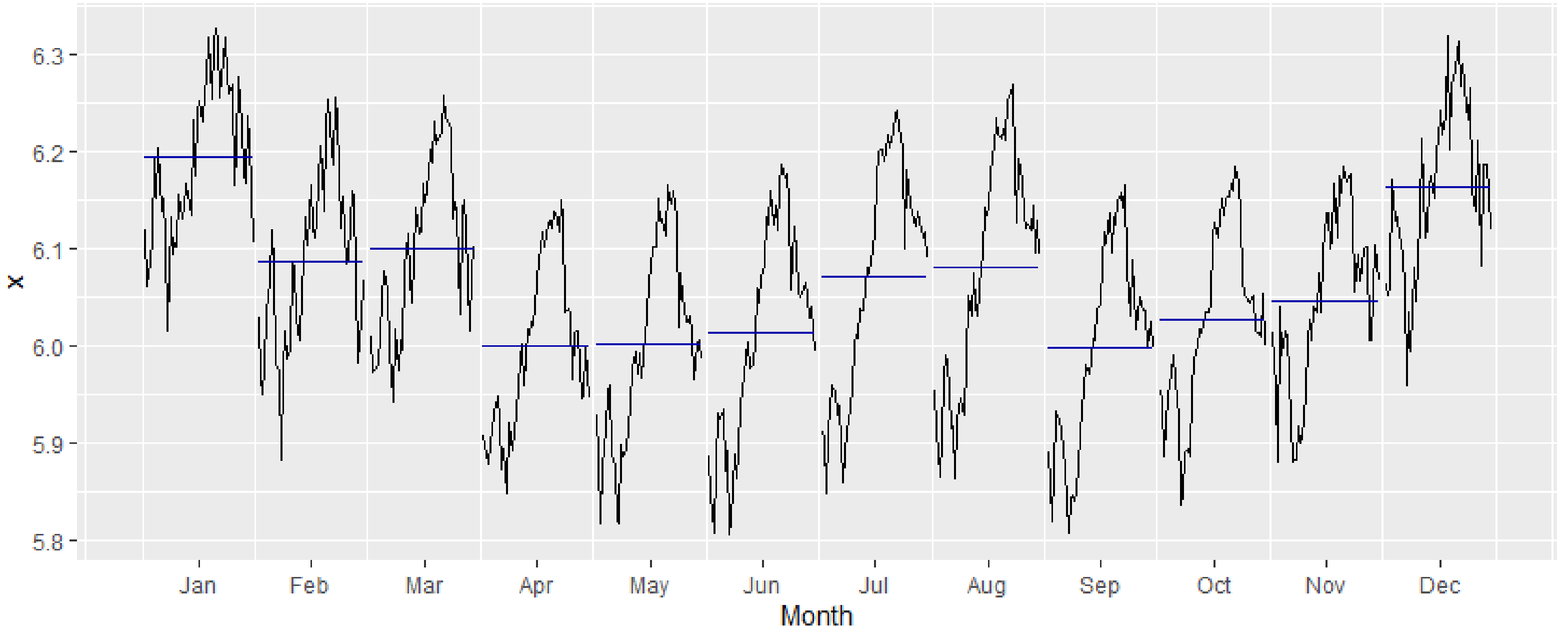
# Time series description

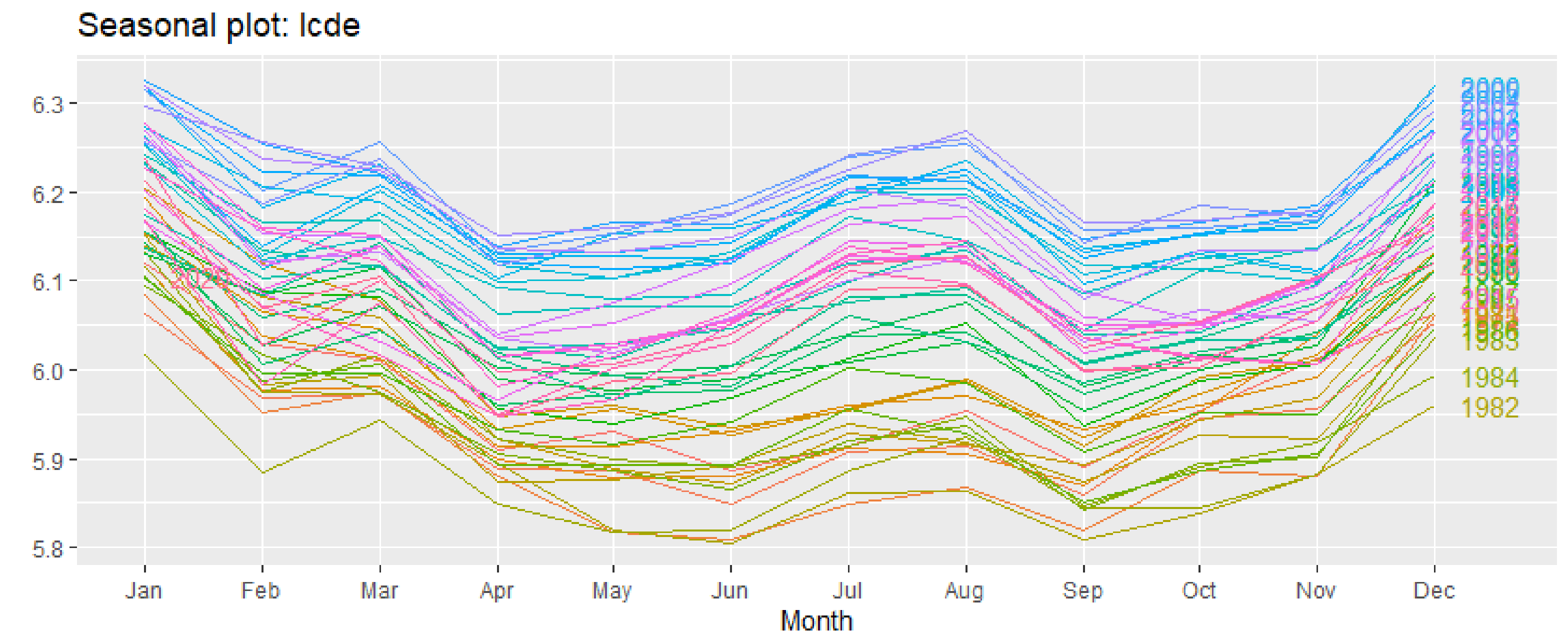



This dataset represents the monthly Total U.S energy related CO2 emissions from January 1973 to January 2021 in million metric tons.
We can see different trends along the time, being the biggest one between 1983 and 2008, on the beginning of 2020 we can see a big drop that must be due to the beginning of the Covid-19 pandemic, so I decided to make the analysis without the last 12 months of observation. With the analysis, I can conclude that this time series is not stationary.
Being the ACF not meaningful because the time series is not stationary there is no need to present the plot.
The variance of the data seems to vary in some points of the time series, to stabilize it I used a Box-Cox transformation taking the log of the data that can also help to remove some non-linearity on the trend that exists. Below can be seen the time series with the log transformation and it does not appear to have a big impact on the time series. Using the BoxCox function in R to see if it gave a better result than the log transformed data, I could see that no major difference was found.

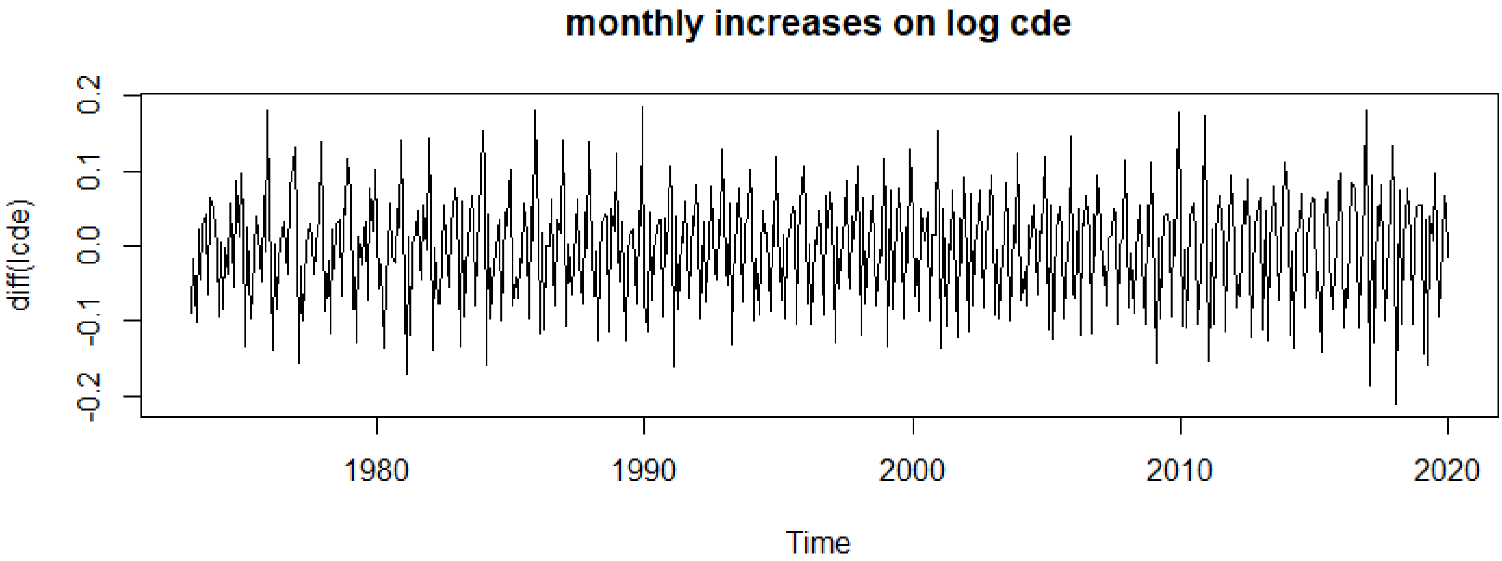


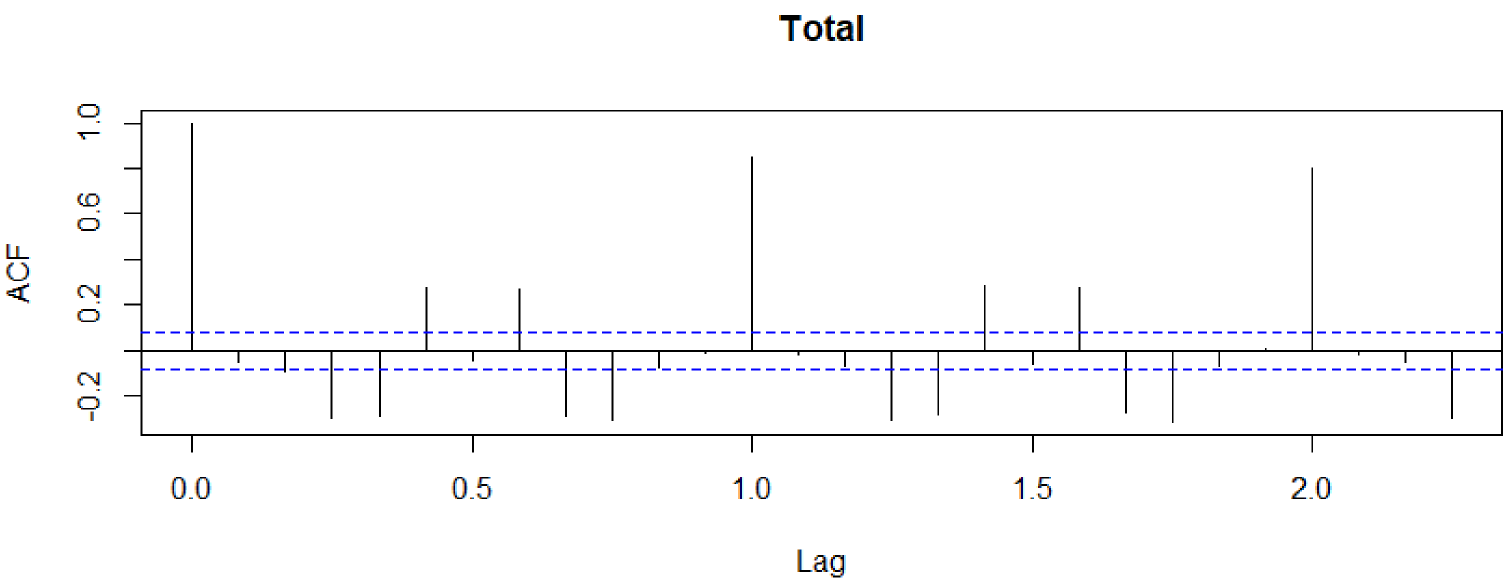
In addition, to see the seasonality that is present in the data I computed the below two plots.



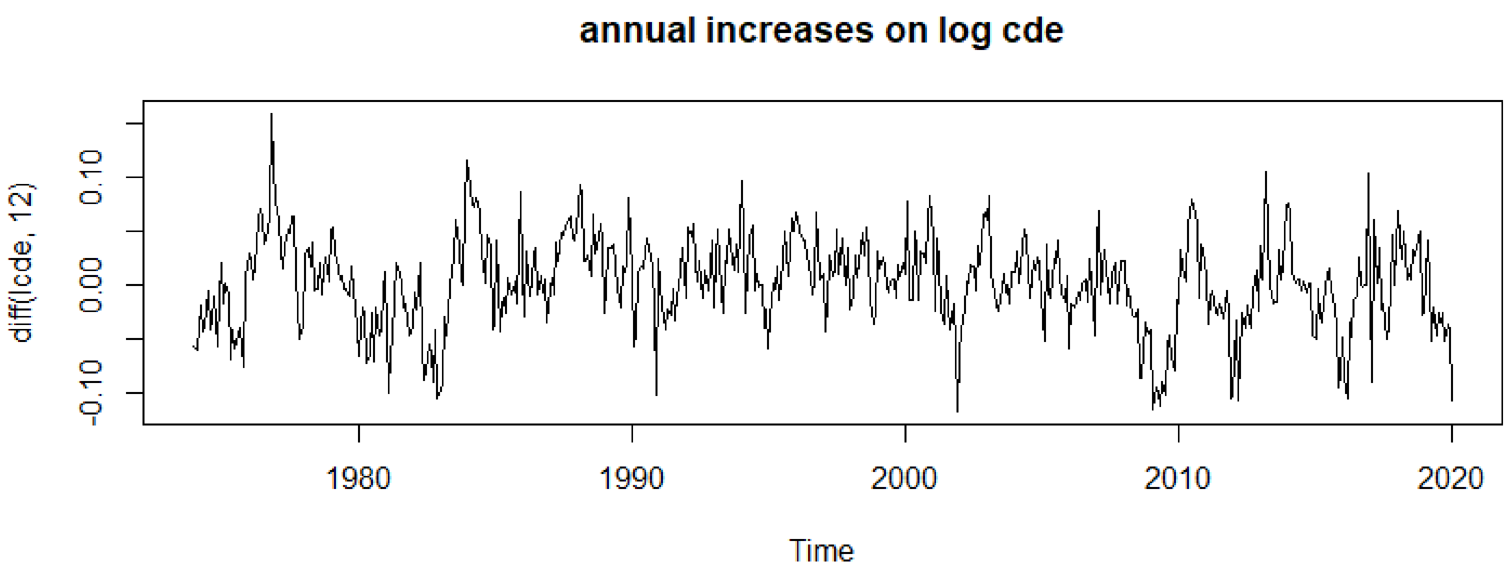



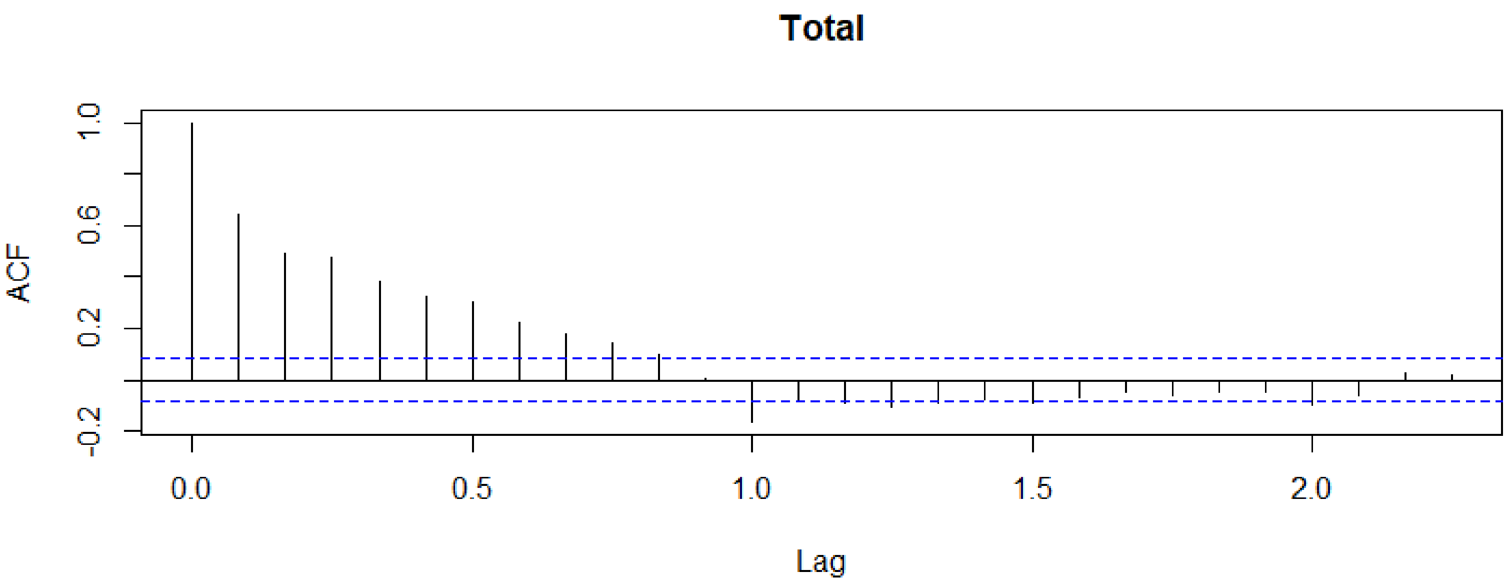
Studying the monthly increases in the log transformed data and its ACF plot I can state that the monthly increases are stationary around zero and seasonal.



The annual increases look approximately stationary around a mean that is zero



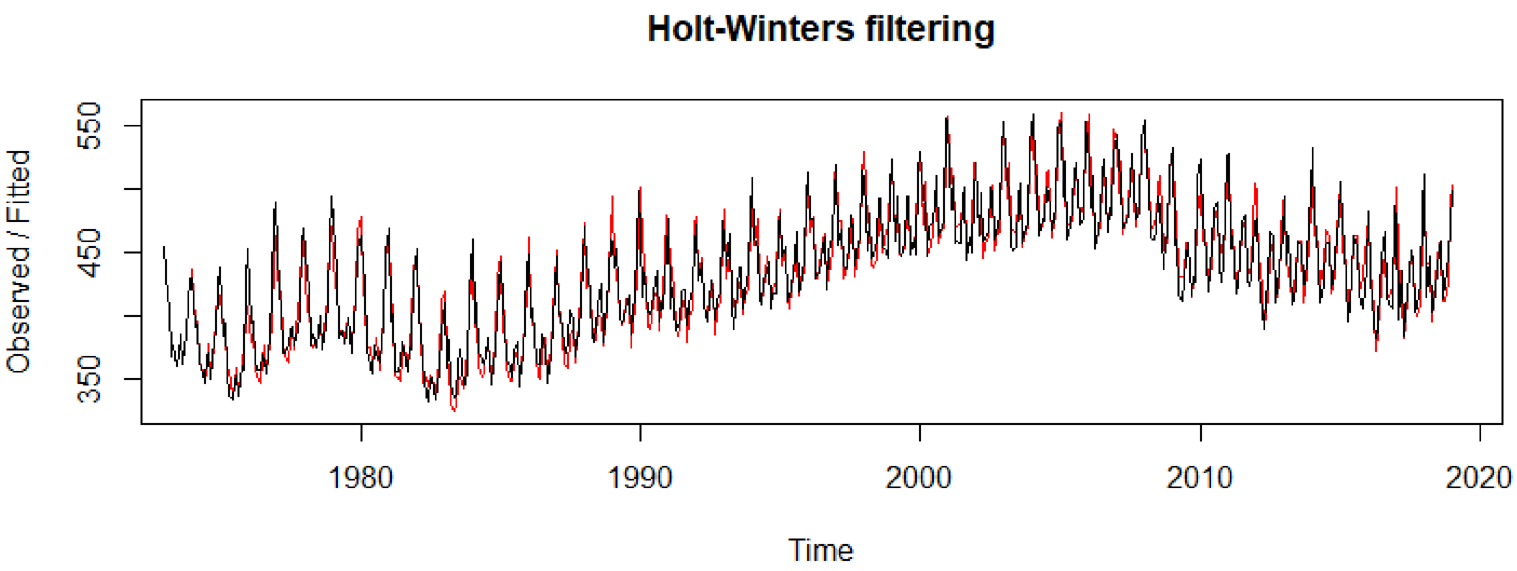



# Smoothing Method

Because the seasonality does not have the same dimension every year and it seems to depend on the increasing or decreasing trend, I think the most adequate model is the Winter's multiplicative model. Although the observed differences are not significant I think the multiplicative model will be a bit better than the additive model.

Below we can see the results of the application of the model and after that we can check the plot of the predictions with the observed data and we can see that the fit is generally good. The results of the forecast are below too.

```
> xhw
Holt-Winters exponential smoothing with trend and multiplicative seasonal component.

Call:
HoltWinters(x = tscde_pred, seasonal = "m")

Smoothing parameters:
 alpha: 0.3476143
 beta : 0.007938737
 gamma: 0.4750947

Coefficients:
            [,1]
a    432.8074512
b     -0.0795642
s1     0.9919145
s2     1.0370782
s3     0.9274471
s4     0.9502017
s5     0.9904900
s6     1.0697056
s7     1.0693879
s8     0.9672105
s9     0.9741305
s10    1.0020427
s11    1.1029855
s12    1.1557635
> |
```



```
> xhw.f
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Feb 2019       429.2291 418.6491 439.8090 413.0485 445.4096
Mar 2019       448.6901 436.5792 460.8011 430.1681 467.2122
Apr 2019       401.1846 388.3171 414.0522 381.5055 420.8638
May 2019       410.9520 396.7223 425.1816 389.1896 432.7144
Jun 2019       428.2974 412.6002 443.9946 404.2907 452.3042
Jul 2019       462.4659 444.9077 480.0240 435.6130 489.3187
Aug 2019       462.2435 443.7235 480.7634 433.9196 490.5673
Sep 2019       418.0002 399.8379 436.1626 390.2233 445.7772
Oct 2019       420.9134 401.7131 440.1136 391.5491 450.2776
Nov 2019       432.8943 412.3668 453.4217 401.5002 464.2883
Dec 2019       476.4150 453.3692 499.4609 441.1694 511.6606
Jan 2020       499.1196 343.8089 654.4303 261.5924 736.6468
>
```

Checking the accuracy of the forecast we can see that for the MAPE measure the accuracy of the test set is a bit more than 4,06%, which is a good value. Comparing with the accuracy of the application of the additive model, that is 3,99%, I get a slightly better forecast with the multiplicative model like I was predicting on the beginning.

```
> xhw.f.acc
                      ME     RMSE      MAE        MPE     MAPE      MASE      ACF1 Theil's U
Training set   0.5755817 12.55641  9.72145  0.1012626 2.232079 0.6827452 0.2229354        NA
Test set     -16.7625905 21.57611 17.23558 -3.9547271 4.064199 1.2104687 0.1642636 0.7682283
>
```

# Decomposition Method

## Classical Method

As I concluded before, the best model for this time series is the multiplicative one, so I'm going to make the decomposition of this time series. Below I show the graphical representation of the observed values, the trend (that have the cycle included), the seasonal component and the errors The values for each component for the last 5 years and the seasonally adjusted data are below, for the trend and error components, we can see that for the first six and last six months no value is available because they were computed through a 12 terms centered MA.

Seasonal component:

```
2015 1.1383210 1.0198309 1.0331434 0.9349454 0.9356141 0.9487132 1.0039485 1.0124880 0.9319075 0.9605509
2016 1.1383210 1.0198309 1.0331434 0.9349454 0.9356141 0.9487132 1.0039485 1.0124880 0.9319075 0.9605509
2017 1.1383210 1.0198309 1.0331434 0.9349454 0.9356141 0.9487132 1.0039485 1.0124880 0.9319075 0.9605509
2018 1.1383210 1.0198309 1.0331434 0.9349454 0.9356141 0.9487132 1.0039485 1.0124880 0.9319075 0.9605509
2019 1.1383210 1.0198309 1.0331434 0.9349454 0.9356141 0.9487132 1.0039485 1.0124880 0.9319075 0.9605509
2020 1.1383210
```

```
2015 0.9791176 1.1014196
2016 0.9791176 1.1014196
2017 0.9791176 1.1014196
2018 0.9791176 1.1014196
2019 0.9791176 1.1014196
2020
```

Trend component:

```
2015 446.5854 446.7708 446.5219 445.7293 443.3815 440.0806 437.4974 434.9148 431.4081 428.9543 427.5997
2016 426.0897 426.3939 426.8777 426.8379 426.7972 428.7869 430.5471 428.7640 428.2852 429.3732 429.9780
2017 429.4582 428.1857 426.5936 425.8869 426.6551 427.4700 428.9060 431.0999 432.3140 433.5898 434.4946
2018 435.0746 435.7895 436.8763 438.0978 439.7702 440.1338 439.0330 439.2180 439.9934 439.2000 438.0263
2019 435.6382 434.5017 433.3988 432.0424 430.4778 429.0641 426.2075       NA       NA       NA       NA
2020       NA
```

```
2015 426.5357
2016 430.1761
2017 434.7804
2018 436.9095
2019       NA
2020
```

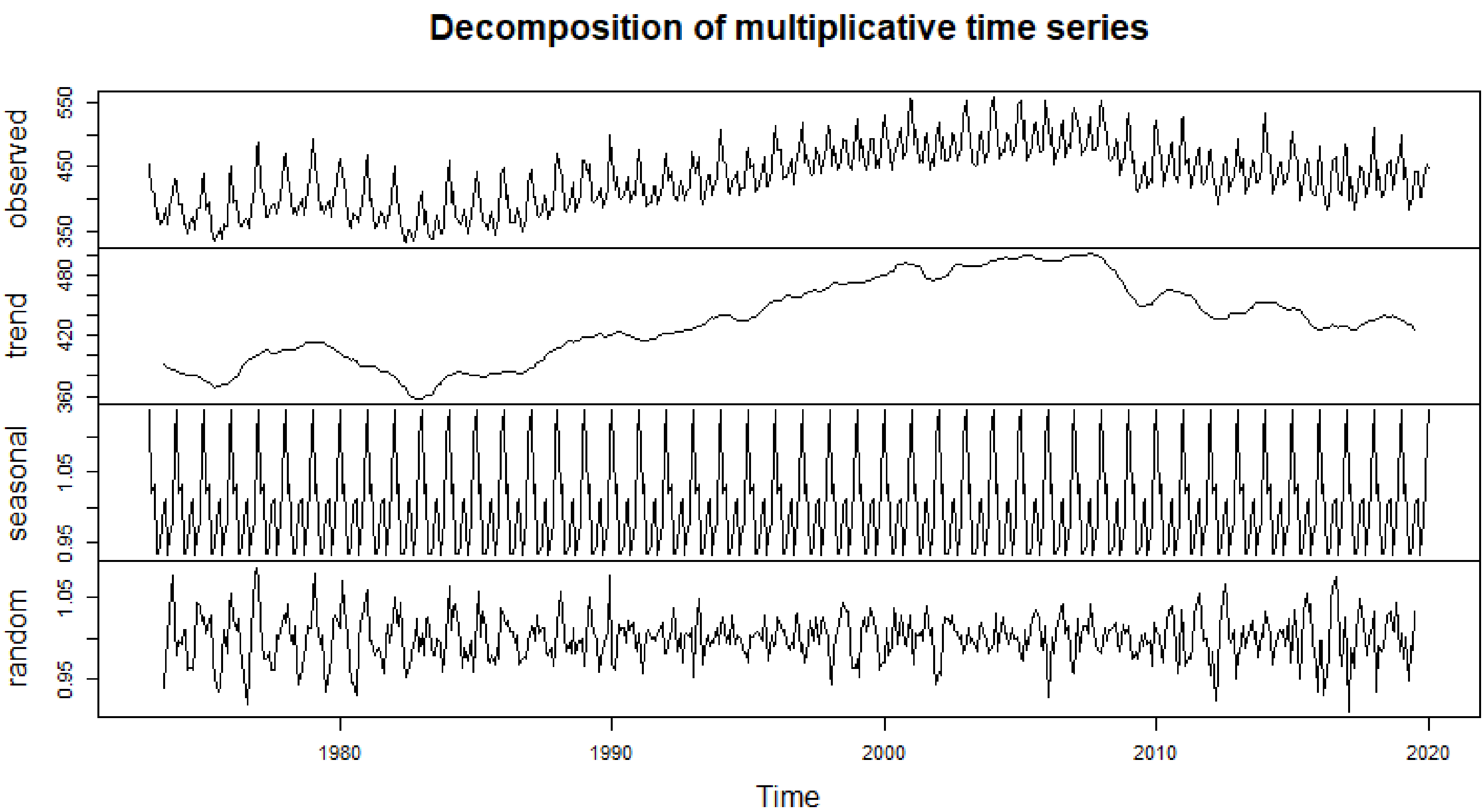
Random component:

```
2015 0.9964404 1.0361199 0.9878627 0.9499161 0.9871538 1.0322668 1.0556006 1.0327788 1.0409308 0.9943368
2016 0.9957985 0.9972758 0.9296165 0.9589943 0.9770445 1.0442538 1.0621840 1.0748242 1.0479998 0.9915531
2017 0.9762606 0.9082319 0.9882506 0.9624970 1.0123790 1.0247336 1.0476464 1.0173857 0.9993083 0.9788394
2018 1.0335675 0.9329557 0.9896004 0.9833043 0.9875461 1.0063949 1.0289394 1.0317539 1.0095037 1.0089624
2019 1.0060333 0.9750603 1.0001717 0.9470311 0.9891247 0.9860395 1.0326641        NA        NA        NA
2020        NA
```

```
2015 0.9701856 0.9328351
2016 0.9642777 1.0262040
2017 1.0005987 1.0150046
2018 1.0441889 0.9821524
2019        NA        NA
2020
```

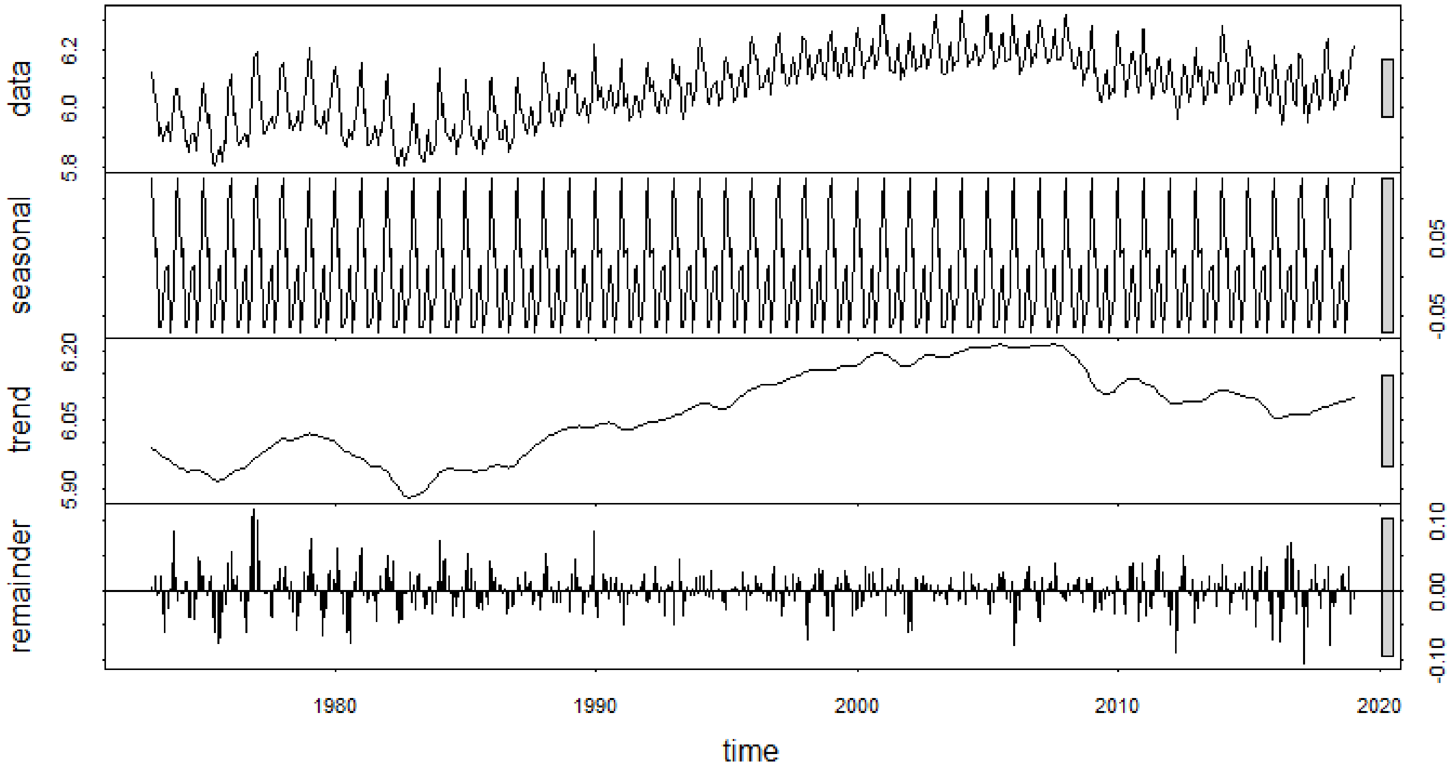
Seasonally adjusted data:

```
2015 505.4097 471.0682 454.6889 394.9261 408.5694 430.0333 462.6421 453.7675 417.5561 408.7384 405.2089 437.1396
2016 481.8507 432.6452 408.9519 381.7711 389.2154 423.8493 458.1221 465.5885 417.3481 407.9904 404.9809 485.1186
2017 476.1177 395.5842 434.5209 382.3131 403.1904 414.6283 450.1121 443.0595 401.6661 406.7114 424.6969 484.9596
2018 510.7407 413.6152 445.6289 401.8241 405.3954 419.2823 452.5181 457.8115 412.9981 424.6944 446.8519 471.5306
2019 497.7497 431.0472 446.8069 381.6051 397.4454 400.4273 440.8631 442.1895 402.1511 402.9874 431.0099 453.4426
2020 447.2787
>
```



## Seasonal and Trend decomposition using Loess - STL method

To compute this method in R I used the "stl" function with the log of the data to use the multiplicative model, with outer set to 20 and inner to 2. Below can be seen the plot of the entire components and the data.




After making the exp of the data, because it was logged, I can get the values of the three components of the decomposition for the last 12 months.

| | seasonal | trend | remainder |
|---|---|---|---|
| 542 | 1.0267338 | 437.0382 | 0.9240356 |
| 543 | 1.0370105 | 437.9005 | 0.9836042 |
| 544 | 0.9387314 | 438.9063 | 0.9775345 |
| 545 | 0.9383411 | 439.9144 | 0.9843534 |
| 546 | 0.9513936 | 440.4634 | 1.0028085 |
| 547 | 1.0075099 | 441.0132 | 1.0206985 |
| 548 | 1.0159481 | 441.5858 | 1.0227264 |
| 549 | 0.9331529 | 442.1592 | 1.0032182 |
| 550 | 0.9617860 | 442.7691 | 0.9995441 |
| 551 | 0.9764501 | 443.3799 | 1.0343991 |
| 552 | 1.0995753 | 443.9472 | 0.9682040 |
| 553 | 1.1362385 | 444.5152 | 0.9877498 |

With this information, I can compute the forecast for the data and then check the accuracy of the predictions. Below I have the values of the forecast that are the exp of the results of the forecast.

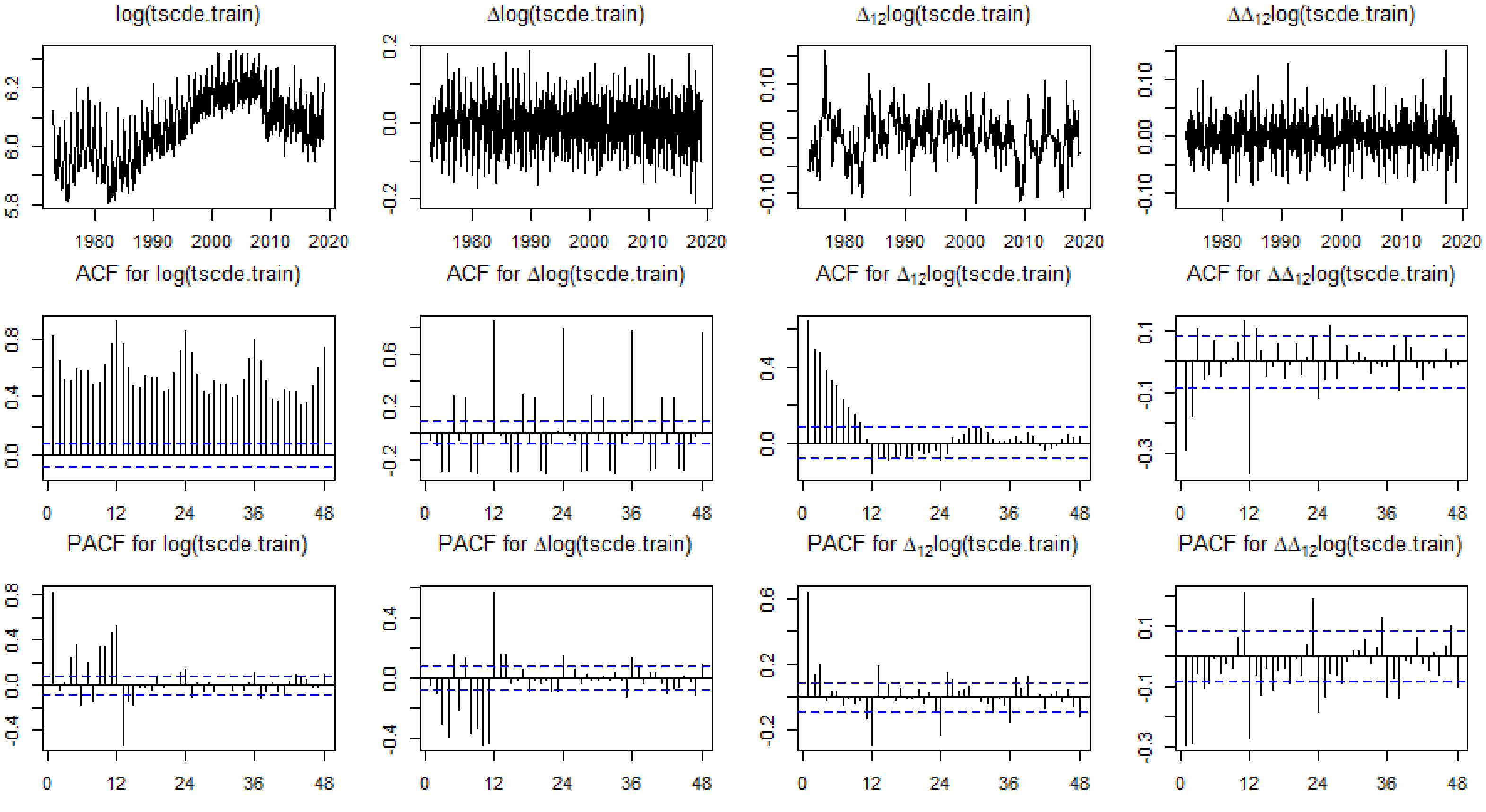
```
> v2 <- exp(stltscde.f[[2]])
> v2
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov
2019          449.9386 454.4421 411.3739 411.2029 416.9228 441.5143 445.2121 408.9293 421.4770 427.9031
2020 497.9261
          Dec
2019 481.8595
2020
```

And to compute the accuracy I used these values against the original time series and got good results with 3,71% for the MAPE measure.

```
> stltscde.f.acc
                ME     RMSE      MAE       MPE     MAPE      ACF1 Theil's U
Test set -14.95397 20.73658 15.69374 -3.542321 3.713264 0.1030058 0.7048346
```

# Modelling with a SARIMA model

Because the time series have seasonality and trend, I had to consider a seasonal AR and/or MA component besides the regular AR and/or MA component. I divided the dataset into a train set and test set, with only the last 12 months of observations, and because the variance seems to change over time, I will consider the log of the data that is more stable.



To know the best SARIMA model, I will study the data, the ACF and PACF plots of the logged data, the differenced data, the annual differenced data and the annual difference of the differenced data.

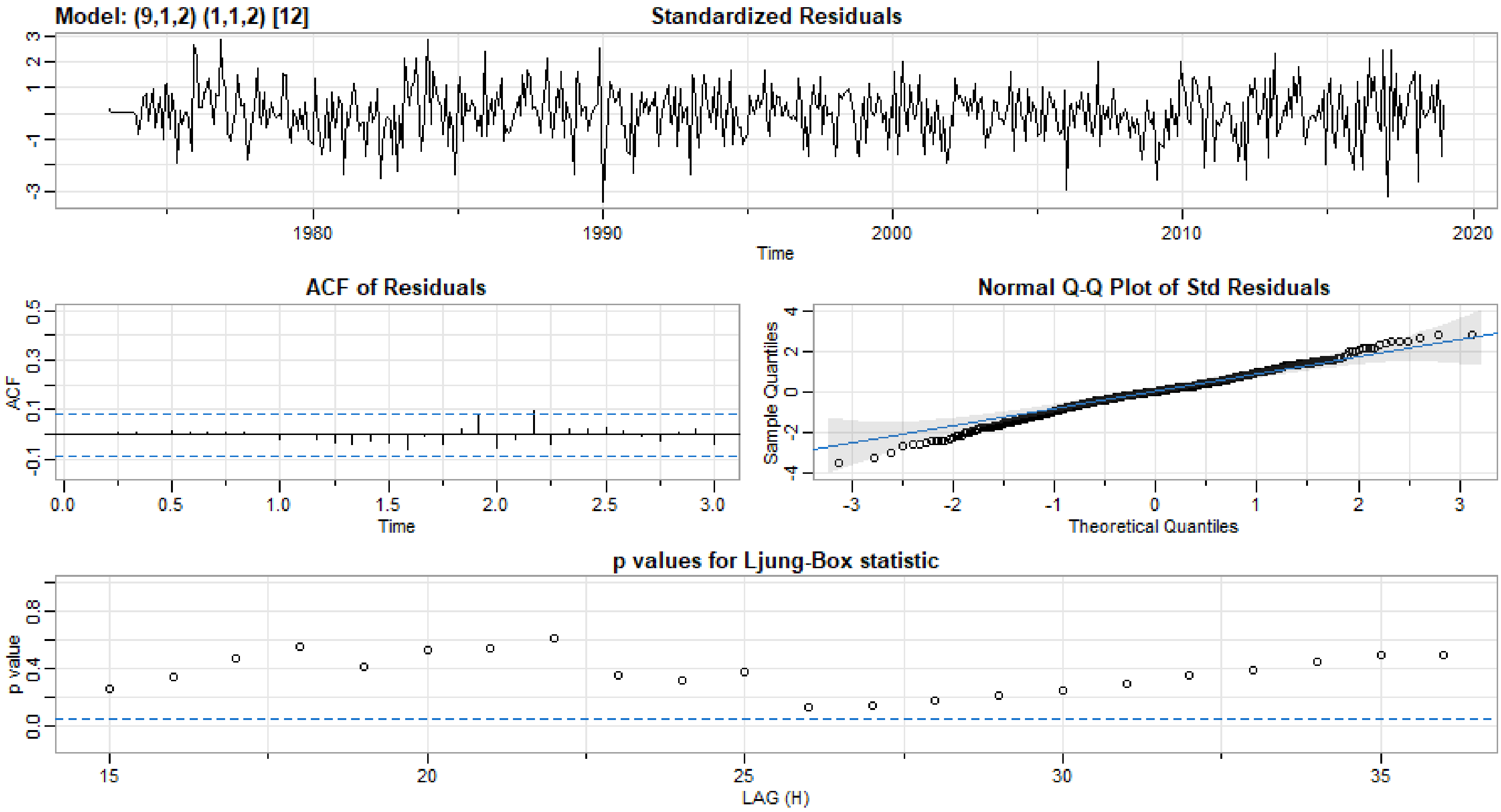
```
> nsdiffs(ltrain)
[1] 1
> ndiffs(ltrain)
[1] 1
> ndiffs(diff(ltrain,12))
[1] 0
> ndiffs(diff(diff(ltrain,12)))
[1] 0
> 
```

Analyzing the plots and the output of functions ndiffs() and nsdiffs(), i could see that can be d=0 or d=1 and D=1.

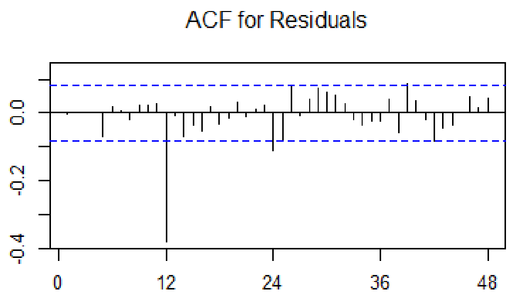
Although the PACF and ACF plots indicate an AR model, it is not easy to clearly indicate a model for this time series.

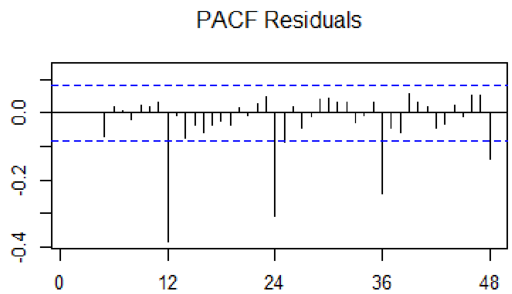
Starting with a large p and a small P, the first model that i called model1 is sarima(lcde,9,1,2,1,1,2,12). Despite the no correlation of the residuals, I could see that with this model a few parameters are not statistically significant, so I made them equal to zero and re-estimated the model.

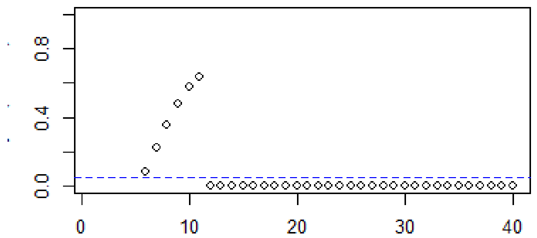
```
sigma^2 estimated as 0.0006576:  log likelihood = 1204.68,  aic = -2379.35

$degrees_of_freedom
[1] 526

$ttable
     Estimate     SE t.value p.value
ar1    0.5426 0.3587  1.5128  0.1309
ar2   -0.5473 0.1347 -4.0617  0.0001
ar3   -0.0632 0.0895 -0.7064  0.4803
ar4   -0.2337 0.0668 -3.4956  0.0005
ar5   -0.1089 0.0732 -1.4883  0.1373
ar6   -0.0109 0.0674 -0.1614  0.8719
ar7   -0.2142 0.0599 -3.5728  0.0004
ar8   -0.0029 0.0621 -0.0469  0.9626
ar9   -0.1128 0.0678 -1.6637  0.0968
ma1   -0.9207 0.3585 -2.5684  0.0105
ma2    0.5544 0.2004  2.7667  0.0059
sar1  -0.4129 0.3051 -1.3534  0.1765
sma1  -0.3757 0.2910 -1.2909  0.1973
sma2  -0.4180 0.2457 -1.7016  0.0894

$AIC
[1] -4.318245
```




After the re-estimation, I had to recalculate the Lung-Box statistic and I could see that almost all the residuals are correlated.



After the analysis of model1 and model2, I can have the conclusion that model1 can explain the serial correlation very well, although, some parameters are not statistically significant. However, model2 is not a very good option to explain the serial correlation.

```
> model1
Series: ltrain
ARIMA(9,1,2)(1,1,2)[12]

Coefficients:
         ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8      ar9      ma1     ma2
      0.5426  -0.5473  -0.0632  -0.2337  -0.1089  -0.0109  -0.2142  -0.0029  -0.1128  -0.9207  0.5544
s.e.  0.3587   0.1347   0.0895   0.0668   0.0732   0.0674   0.0599   0.0621   0.0678   0.3585  0.2004
         sar1     sma1     sma2
      -0.4129  -0.3757  -0.4180
s.e.   0.3051   0.2910   0.2457

sigma^2 estimated as 0.000676:  log likelihood=1204.68
AIC=-2379.35   AICc=-2378.44   BIC=-2314.98
> |
```

```
> model2
Series: ltrain
ARIMA(9,1,2)(1,1,2)[12]

Coefficients:
      ar1      ar2  ar3      ar4  ar5  ar6      ar7  ar8  ar9      ma1     ma2  sar1  sma1  sma2
        0  -0.2151    0  -0.1060    0    0  -0.0421    0    0  -0.4108  0.0208     0     0     0
s.e.    0   0.1196    0   0.0575    0    0   0.0423    0    0   0.0430  0.1165     0     0     0

sigma^2 estimated as 0.001047:  log likelihood=1088.9
AIC=-2165.8   AICc=-2165.64   BIC=-2140.05
> |
```

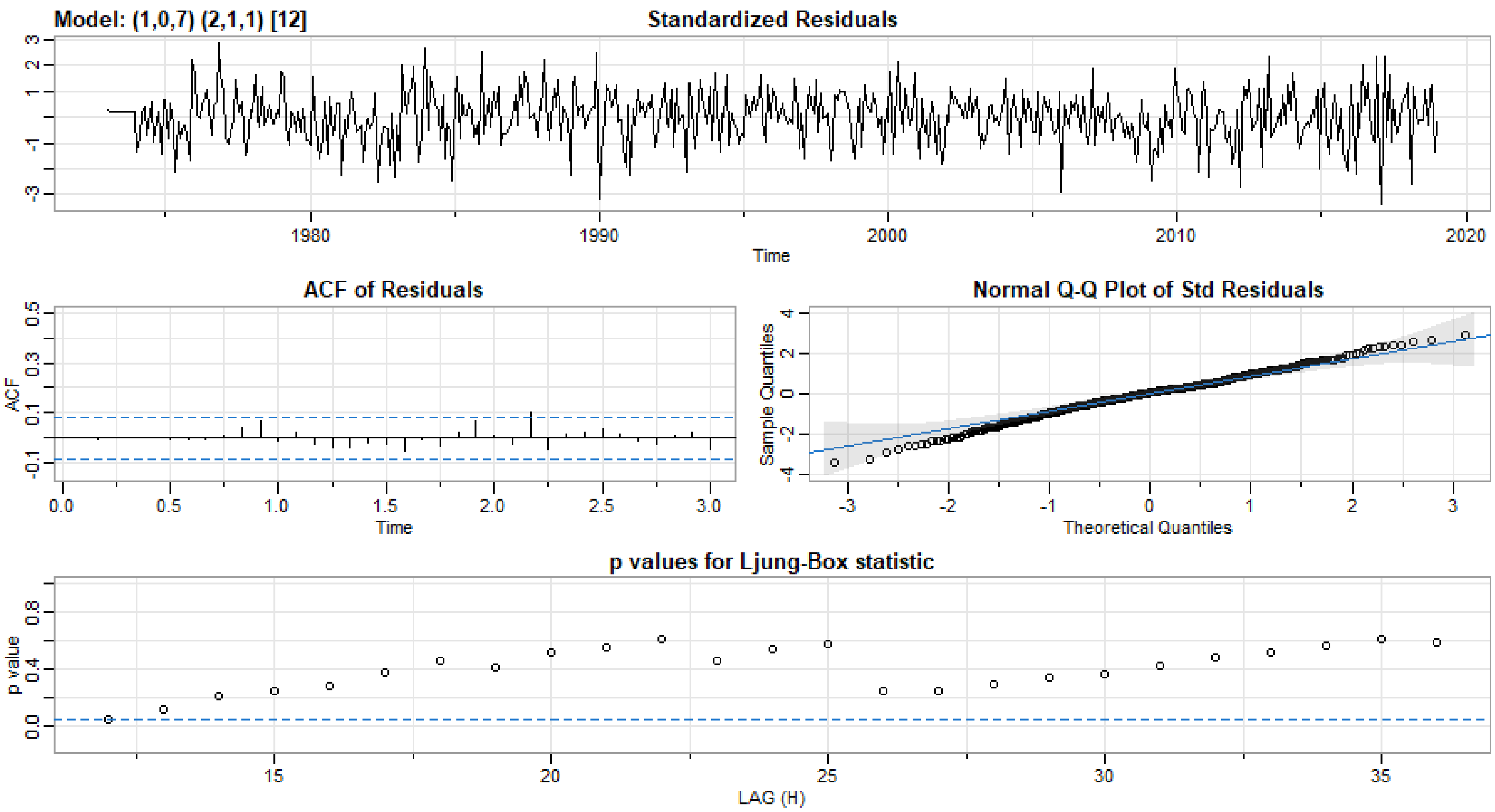
Comparing the information criteria of both of them, we can see that model1 has lower information criteria and residuals are uncorrelated so I will keep Model1 to produce forecasts.

| Model | Order | NPar | AIC | AICc | BIC |
|---|---|---|---|---|---|
| **model1** | (9,1,2)x(1,1,2)[12] | 14 | -2379.35 | -2378.44 | -2314.98 |
| **model2** | (9,1,2)x(1,1,2)[12] | 14-9 | -2165.8 | -2165.64 | -2140.05 |

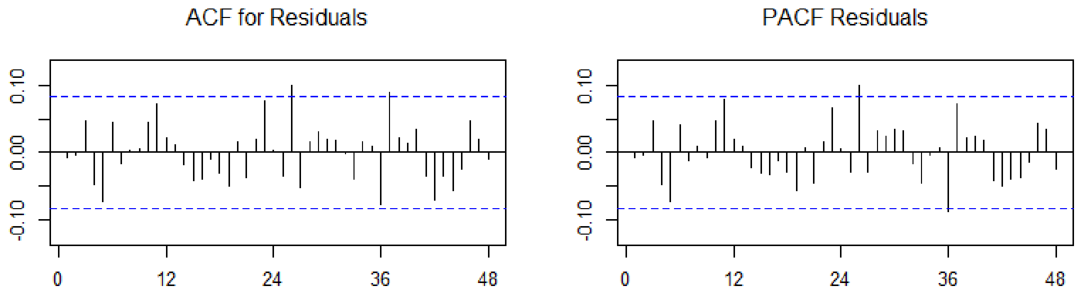
I tried other values of the orders of the seasonal components of the model but it did not produced better results. However, increasing the order of the MA component, lowering the order of the AR component and removing the differencing of the regular components I was able to get a good model that I wish can produce good forecasts.

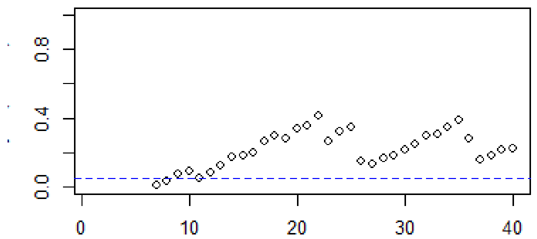
Using model3 as sarima(lcde,1,0,7,2,1,1,12) I could get good results, with almost all the residuals uncorrelated but with some parameters being statistically not significant.

```
sigma^2 estimated as 0.0006522:  log likelihood = 1210,  aic = -2394

$degrees_of_freedom
[1] 529

$ttable
         Estimate     SE  t.value p.value
ar1        0.9845 0.0096 103.0481  0.0000
ma1       -0.3795 0.0442  -8.5849  0.0000
ma2       -0.1880 0.0463  -4.0600  0.0001
ma3        0.0444 0.0469   0.9481  0.3435
ma4       -0.0676 0.0469  -1.4409  0.1502
ma5       -0.0698 0.0461  -1.5149  0.1304
ma6        0.0978 0.0499   1.9605  0.0505
ma7       -0.0947 0.0462  -2.0478  0.0411
sar1       0.0741 0.0555   1.3347  0.1826
sar2      -0.1544 0.0509  -3.0307  0.0026
sma1      -0.7863 0.0395 -19.9018  0.0000
constant   0.0002 0.0004   0.5104  0.6100

$AIC
[1] -4.336951
```




After the re-estimation, I had to recalculate the Lung-Box statistic, as before, and I could see that the residuals are almost all uncorrelated and all the parameters are statistically significant.



After analyzing these two models, I could see that model4 has a better AIC, so I will keep the model to produce forecasts.

```
> model3
Series: ltrain
ARIMA(1,0,7)(2,1,1)[12]

Coefficients:
         ar1      ma1      ma2     ma3      ma4      ma5     ma6      ma7    sar1     sar2     sma1
      0.9858  -0.3805  -0.1883  0.0441  -0.0681  -0.0701  0.0973  -0.0955  0.0739  -0.1548  -0.7863
s.e.  0.0088   0.0441   0.0463  0.0468   0.0469   0.0461  0.0498   0.0462  0.0555   0.0508   0.0394

sigma^2 estimated as 0.0006668:  log likelihood=1209.88
AIC=-2395.76   AICc=-2395.16   BIC=-2344.23
> |
```

```
> model4
Series: ltrain
ARIMA(1,0,7)(2,1,1)[12]

Coefficients:
         ar1      ma1      ma2  ma3  ma4  ma5  ma6      ma7  sar1     sar2     sma1
      0.9848  -0.3799  -0.1947    0    0    0    0  -0.0772     0  -0.1783  -0.7534
s.e.  0.0087   0.0446   0.0458    0    0    0    0   0.0430     0   0.0481   0.0344

sigma^2 estimated as 0.0006711:  log likelihood=1205.63
AIC=-2397.25   AICc=-2397.04   BIC=-2367.2
> |
```

| **Model** | **Order** | **NPar** | **AIC** | **AICc** | **BIC** |
|---|---|---|---|---|---|
| **model3** | (1,0,7)x(2,1,1)[12] | 11 | -2395.76 | -2395.16 | -2344.23 |
| **model4** | (1,0,7)x(2,1,1)[12] | 11-5 | -2397.25 | -2397.04 | -2367.2 |

After trying to tune the parameters of model3, I noticed that this one is the better one. Therefore, I will use model1, model3 and model4 to produce forecasts for this time series, with model4 being the one with lower information criteria.

The results for the forecast accuracy for the training set and the test set can be obtained, as can be seen below.

```
> model1.f12.acc
                        ME       RMSE        MAE          MPE      MAPE      MASE          ACF1
Training set  0.0003505047 0.02535803 0.01927134  0.004881536 0.3174356 0.5868769 -0.0001236171
Test set     -0.0365960825 0.04553148 0.03816599 -0.605356432 0.6310743 1.1622820  0.0233747205
             Theil's U
Training set        NA
Test set     0.6779146
> model3.f12.acc
                       ME       RMSE        MAE          MPE      MAPE      MASE         ACF1 Theil's U
Training set  0.000616192 0.02527933 0.01952090  0.008930613 0.3216950 0.5944768 -0.002059282        NA
Test set     -0.035098945 0.04400316 0.03742953 -0.580594039 0.6188225 1.1398544 -0.014522117 0.6520406
> model4.f12.acc
                        ME       RMSE        MAE          MPE      MAPE      MASE         ACF1 Theil's U
Training set  0.0005995737 0.02548034 0.01974398  0.008630388 0.3253736 0.6012702 -0.007446257        NA
Test set     -0.0346949987 0.04371624 0.03714162 -0.573826174 0.6140143 1.1310865 -0.028975430 0.6473619
```

For the training set, I got the following accuracy measures,

| **Model** | **RMSE** | **MAE** | **MAPE** | **MASE** |
|---|---|---|---|---|
| **model1** | 0.0254 | 0.0193 | 0.3174 | 0.5869 |
| **model3** | 0.0253 | 0.0195 | 0.3217 | 0.5945 |
| **model4** | 0.0255 | 0.0197 | 0.3254 | 0.6013 |

All the models have a similar performance, with model4 being worse than the others on all measures, because information criteria penalizes models with the smallest number of parameters.

For the test set, the measures are the one below,

| Model | RMSE | MAE | MAPE | MASE |
|---|---|---|---|---|
| **model1** | 0.0455 | 0.0382 | 0.6311 | 1,1623 |
| **model3** | 0.0440 | 0.0374 | 0.6188 | 1.1399 |
| **model4** | 0.0437 | 0.0371 | 0.6140 | 1.1311 |

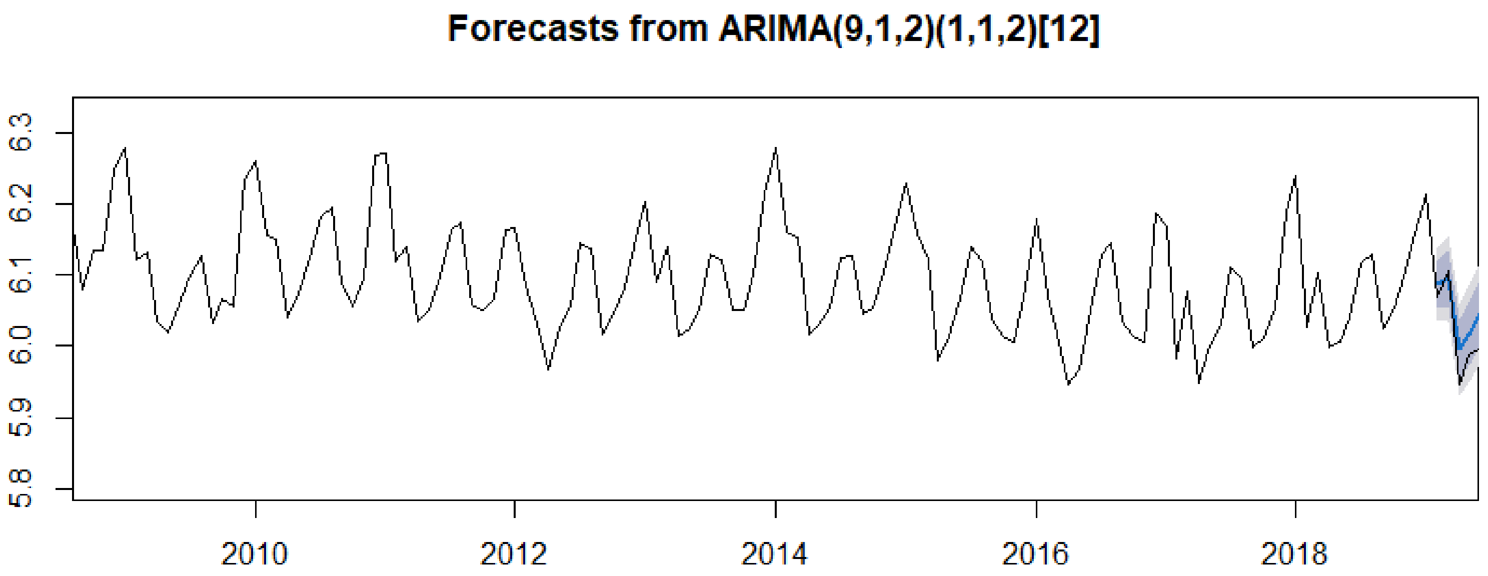
However, in the test set is model4, that was the worst of the 3 models on the training set, to have the best accuracy measures with MAPE of 0.614%.

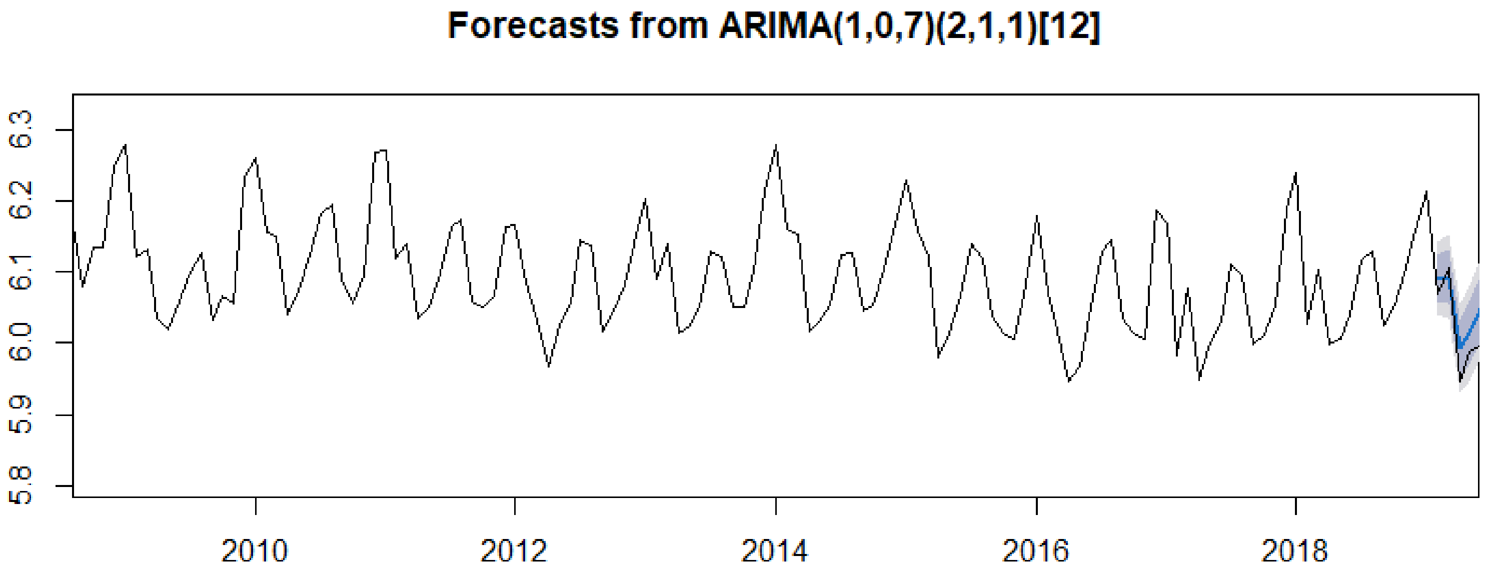
Below I show the 95% confidence intervals in the forecasts for the best performing model in the test set that is model4.

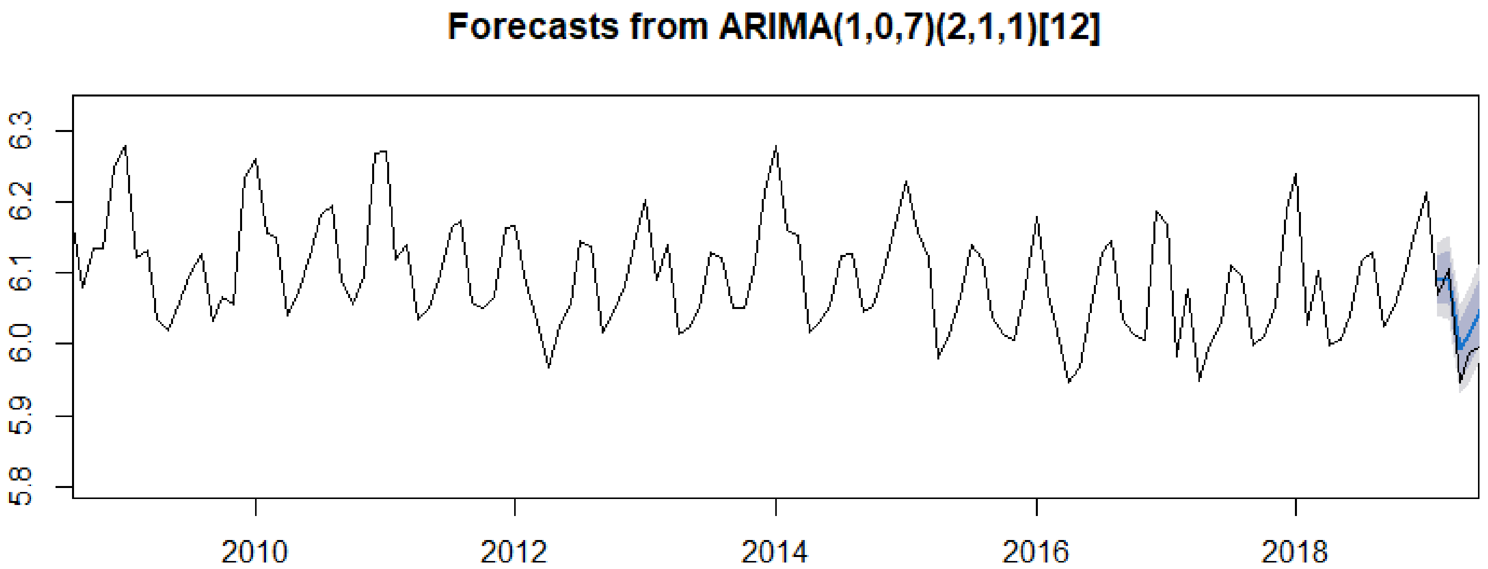
```
> model4.f12$lower[1:12,2]
 [1] 6.040670 6.037129 5.929707 5.943649 5.981672 6.054890 6.057355 5.958393 5.960996 5.983441 6.082560
[12] 6.123095
> model4.f12$upper[,2]
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov
2019          6.142218 6.155811 6.055181 6.075375 6.119191 6.197802 6.205311 6.109336 6.114781 6.139932
2020 6.284632
          Dec
2019 6.241632
2020
```

Analyzing the plots of the forecasts with the confidence intervals and the observed values, I noticed that all the models performed similarly. With forecasts that overestimate the observed values on the last months, but in other months they achieved very good forecasts.

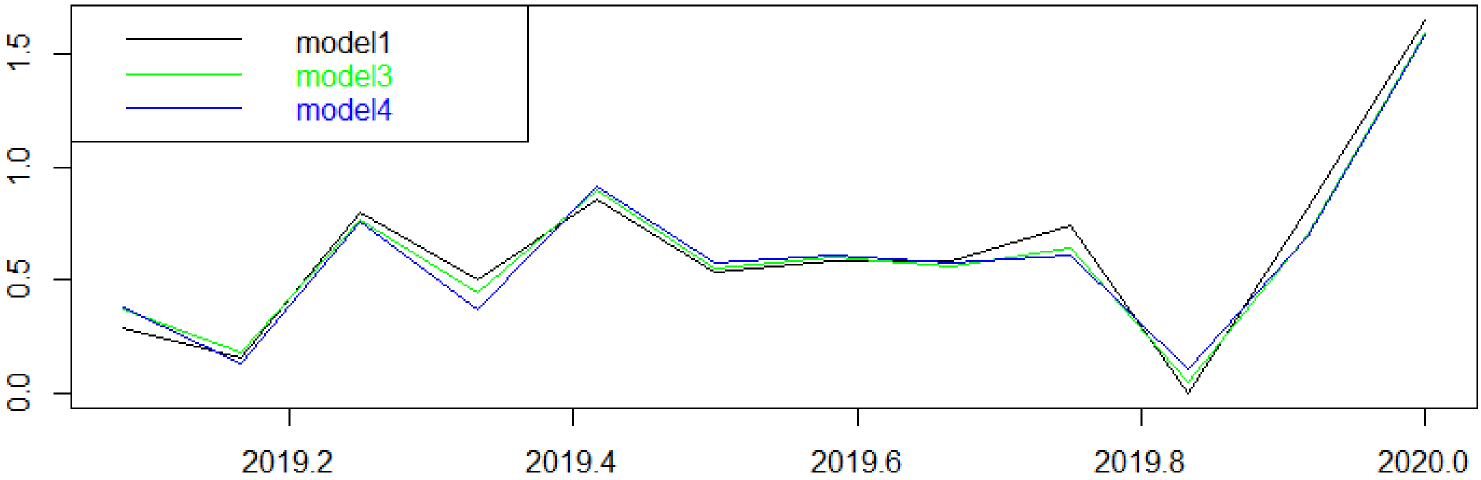
For a better visualization, I only set the plots for the period after 2019.



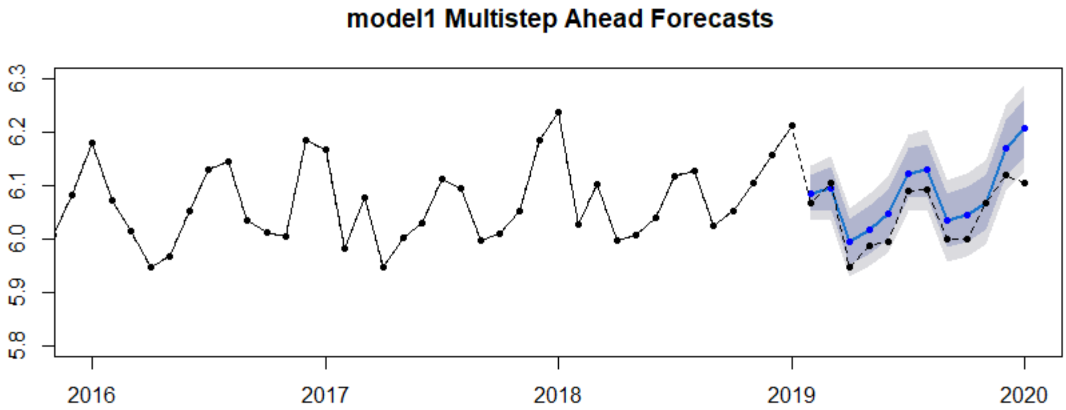


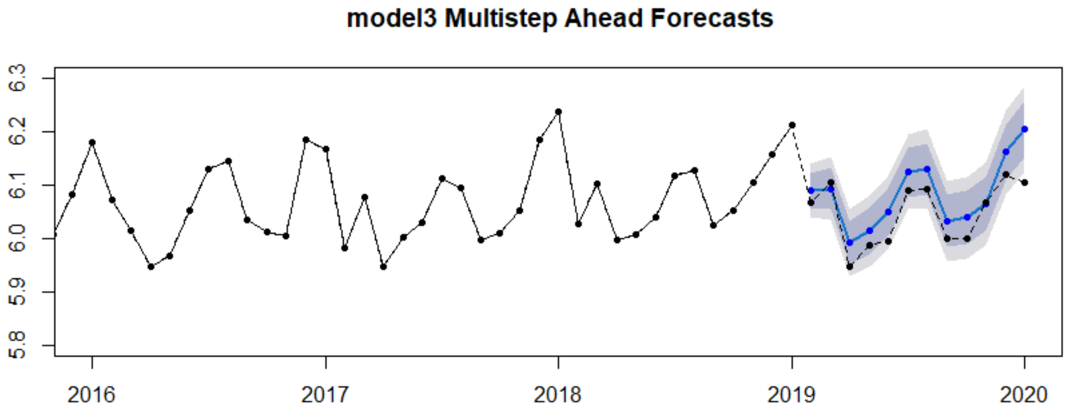



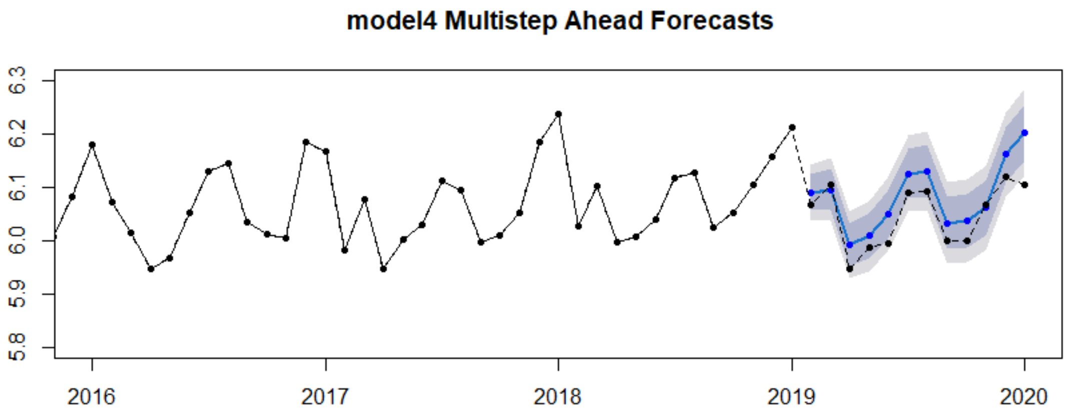

In the Graphic below, we can see that the forecast was very good until after August, where the forecasting errors increased a lot.




Finally, we can have the multistep ahead forecasts for the 3 models, as can be seen below.




model4 Multistep Ahead Forecasts
6.3
6.2
6.1
6.0
5.9
5.8
2016
2017
2018
2019
2020

# Conclusion

After doing the assignment I can conclude that the SARIMA models were very accurate on forecasting and that the most accurate forecasts can be modeled by several different models and finding the best one is very difficult.

As I stated before model4 is the best from all the models that I used but the other ones can make good predictions too. Despite being the model with less parameters.

After doing this assignment, I became aware that this is just a small part of time series analysis and forecasting and on this assignment I would like to have used multivariate time series analysis to the decomposed data that I used but I did not had time for that, although, that is a thing that I would like to explore in the future.

The references that I used were manly the class materials and some internet searches.
The dataset was downloaded from https://www.eia.gov/todayinenergy/detail.php?id=44837