ARIMAX Model for CO2 Ind USA

Universitat Politècnica de Catalunya

Raúl López Martinez, Jan Leyva and Andreu Meca

6 de juny, 2021

Contents

1 Introduction **3**

2 Identification **4**

2.1 Determine the needed transformations to make the series stationary. Justify the transformations carried out using graphical and numerical results. . . . 4

2.2 Analyze the ACF and PACF of the stationary series to identify at least two plausible models. Reason about what features of the correlograms you use to identify these models. . . . 12

3 Estimation **13**

3.1 Use R to estimate the identified models. . . . 13

4 Validation **16**

4.1 Perform the complete analysis of residuals, justifying all assumptions made. Use the corresponding tests and graphical results. . . . 16

4.2 Include analysis of the expressions of the AR and MA infinite models, discuss if they are causal and/or invertible and report some adequacy measures. . . . 25

4.3 Check the stability of the proposed models and evaluate their capability of prediction, reserving the last 12 observations. . . . 29

4.4 Select the best model for forecasting. . . . 31

5 Prediction **32**

5.1 Obtain long term forecasts for the twelve months following the last observation available; provide also confidence intervals. . . . 32

6 Outlier Treatment: **34**

6.1 Analyze of the Calendar Effects are significant. . . . 34

6.2 For the last selected model, apply the automatic detection of outliers and its treatment. Try to give the interpretation of detected outliers . . . 47

6.3 Once the series has been linearized, free of calendar and outliers' effects, perform forecasting. Compare forecasts results for the original series: classical ARIMA vs ARIMA extension (by using the linearized models). 48

6.4 Identification of the model 49

6.5 Estimation of the linearized model 49

6.6 Validation of the linearized model 51

6.7 Forecasting linearized serie 54

7 References **57**

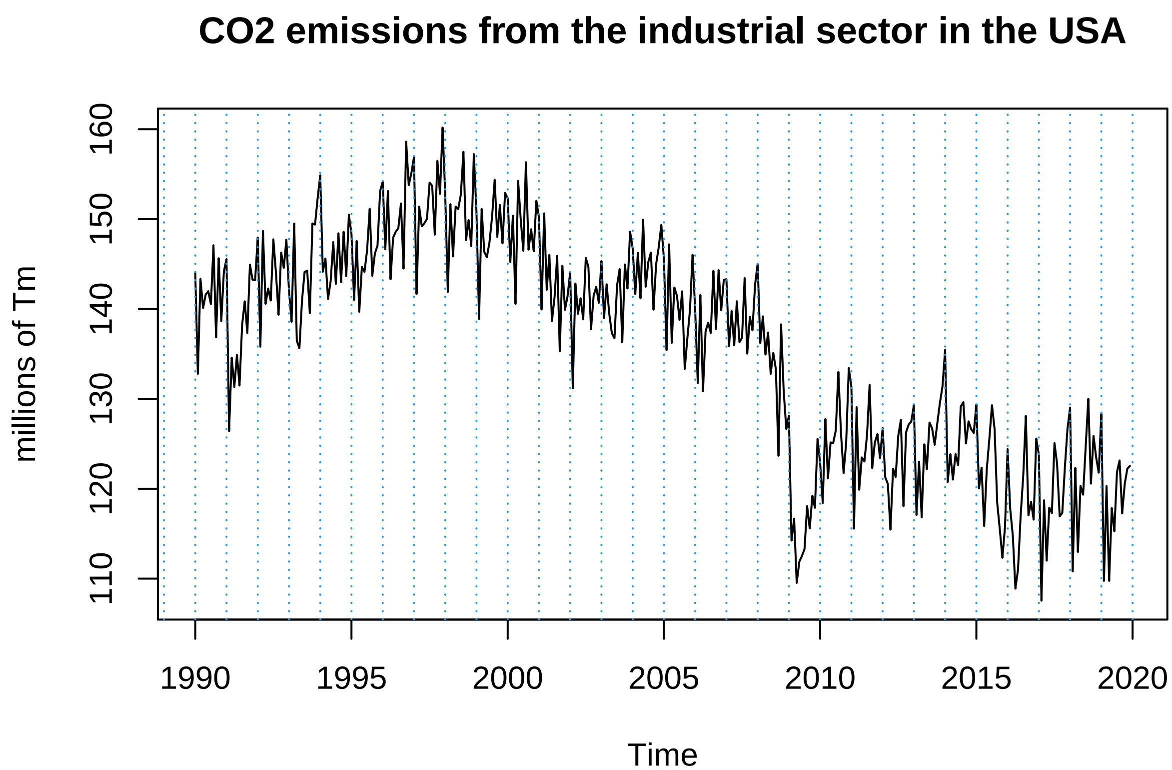
1 Introduction

The aim of this project is to apply the Box-Jenkins ARIMA methodology, following the 4 steps this method comprises which are identification, estimation, validation and finally predicting. This project also includes the outlier treatment and calendar effect to the series.

The data used for the project is based on the CO2 emissions from the industrial sector in the USA from 1990 to 2020.

Source: US Energy Information Administration. Source

The data is collected in millions of tonnes in a 20 year period, which includes the recession at the beginning of the 1990's, the recession of the 2000, caused by the dotcom and the 11S attacks, and the great recession of 2008 caused by the subprime mortgage crisis. To have a better idea, let's plot the data.



The code used for the project can be found at github

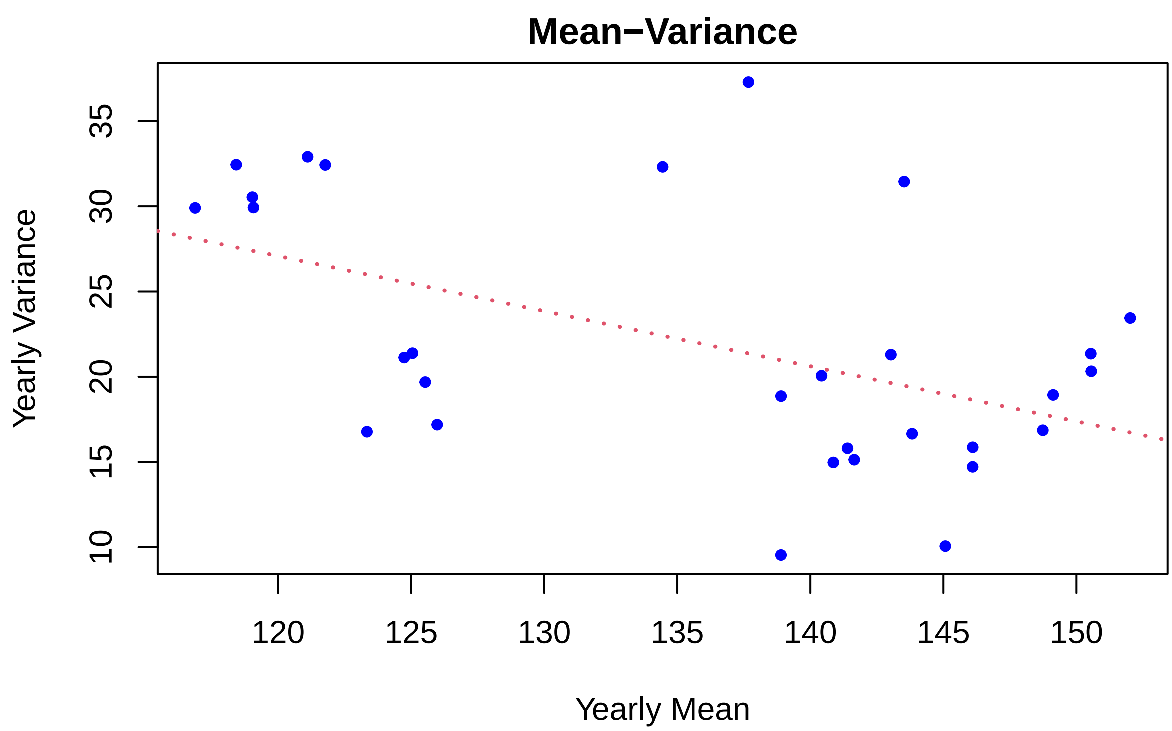
2 Identification

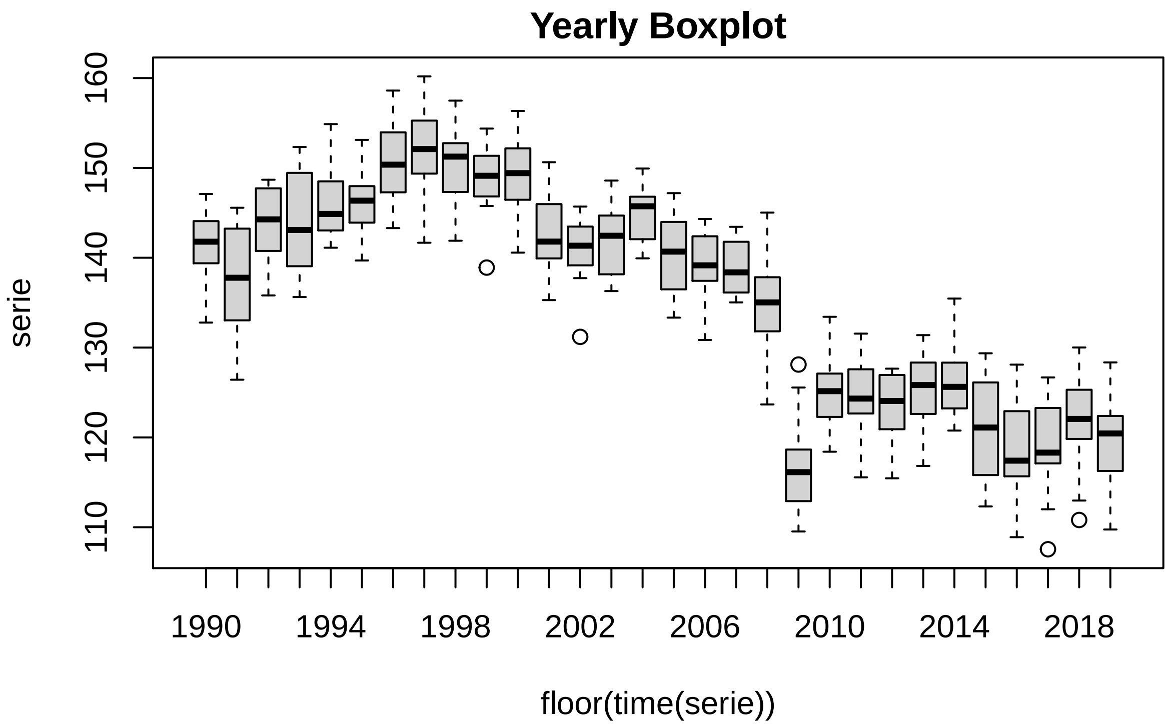
The first step of the Bok-Jenkins methodology is to identify the time series. That means to check if it is stationary, if not apply a series of transformations until it reaches stationarity, and then identify if it has an autoregressive and/or moving average component/s.

2.1 Determine the needed transformations to make the series stationary. Justify the transformations carried out using graphical and numerical results.

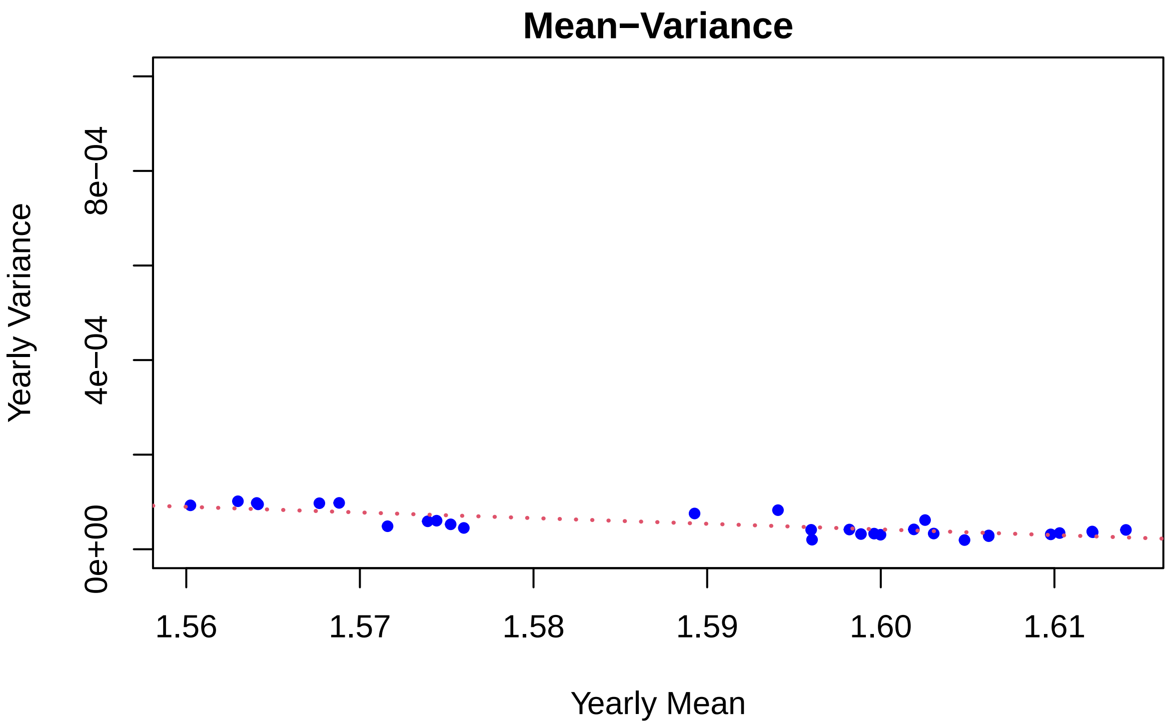
To check if the series needs to be differenced we have to check 3 characteristics: Variance, Seasonality and Mean. For every transformation we apply to the series we'll have to check again the 3 characteristics.

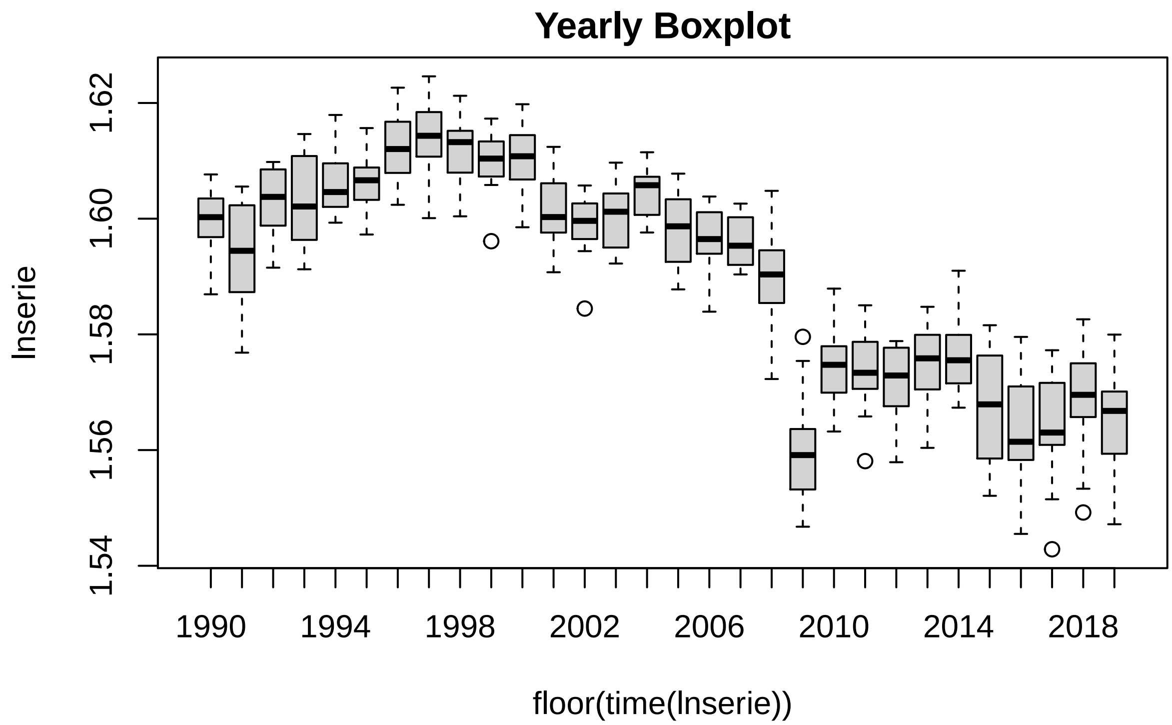
The analysis starts by checking if the variance is constant, which is done with a mean-variance plot and a yearly boxplot of the series:



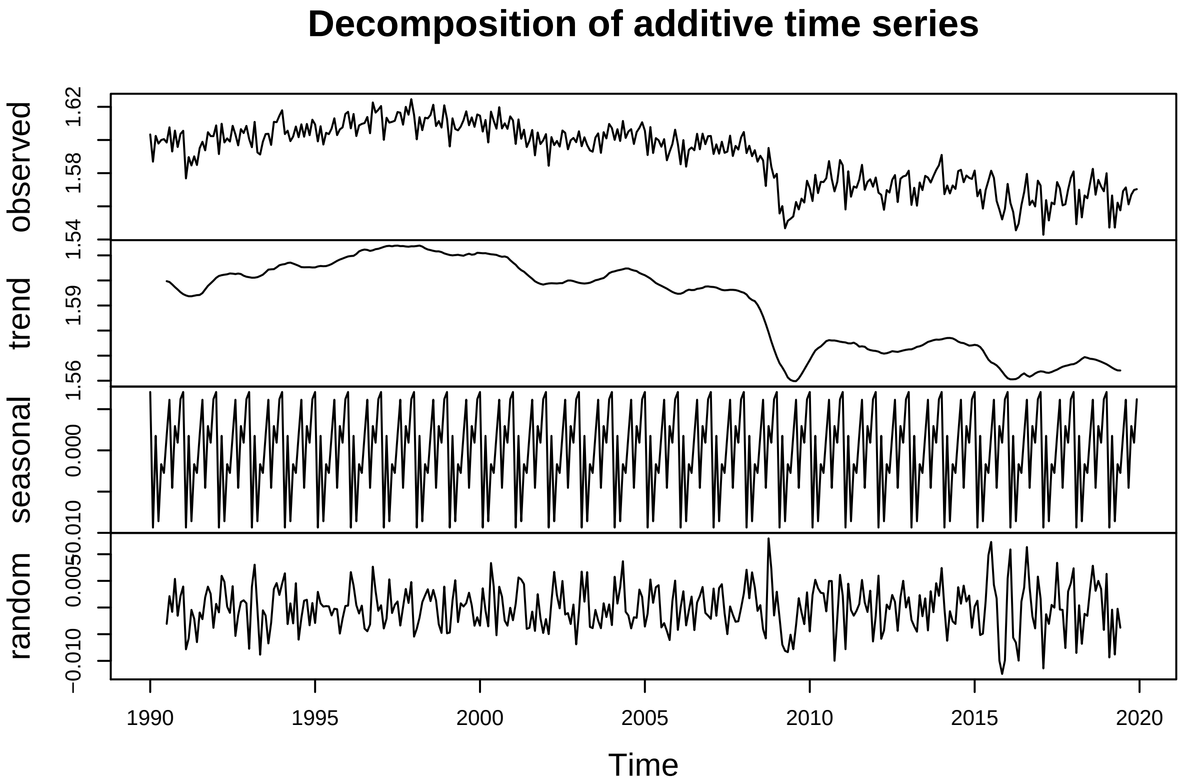


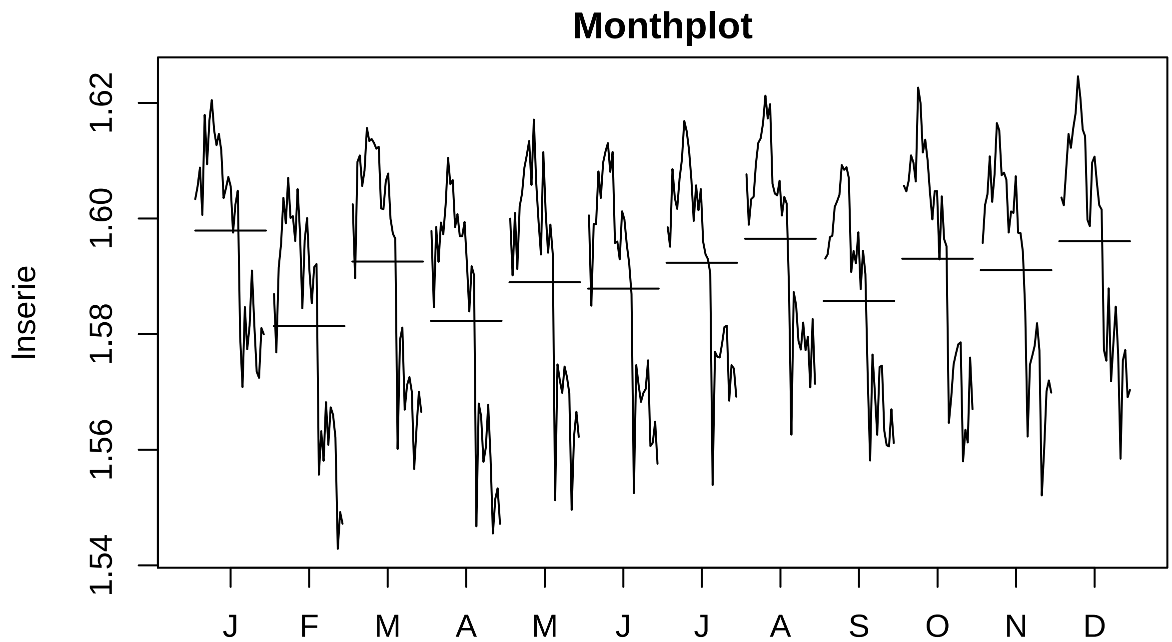
It clearly isn't constant as in the mean-variance plot the fitted line is not straight, indicating that the variance is not constant throughout the series, so we'll apply the logarithm to make it constant. Now, if we check again, we can see that it is a straight line and all the dimensions of the boxes in the boxplot are quite similar, as the magnitude in the y-axis is much lower than before, indicating that the difference in the variance is much more constant than it was before.


Mean−Variance
Yearly Variance
0e+00
4e−04
8e−04
1.56
1.57
1.58
1.59
1.60
1.61
Yearly Mean

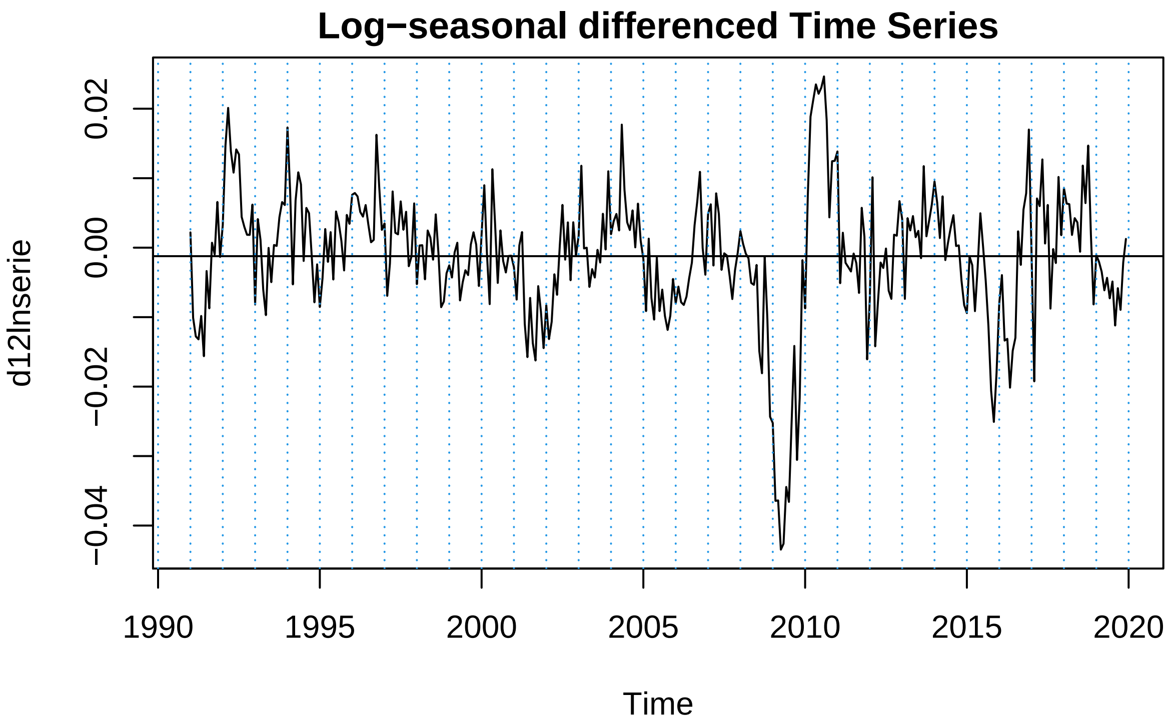


Secondly, it is checked if there exists a seasonal component by plotting the Decomposed series into 3 components: trend, seasonal and random and the Monthplot.


Decomposition of additive time series
observed
1.54
1.58
1.62
trend
1.56
1.59
seasonal
-0.010
0.000
random
-0.010
0.000
0.005
1990
1995
2000
2005
2010
2015
2020
Time



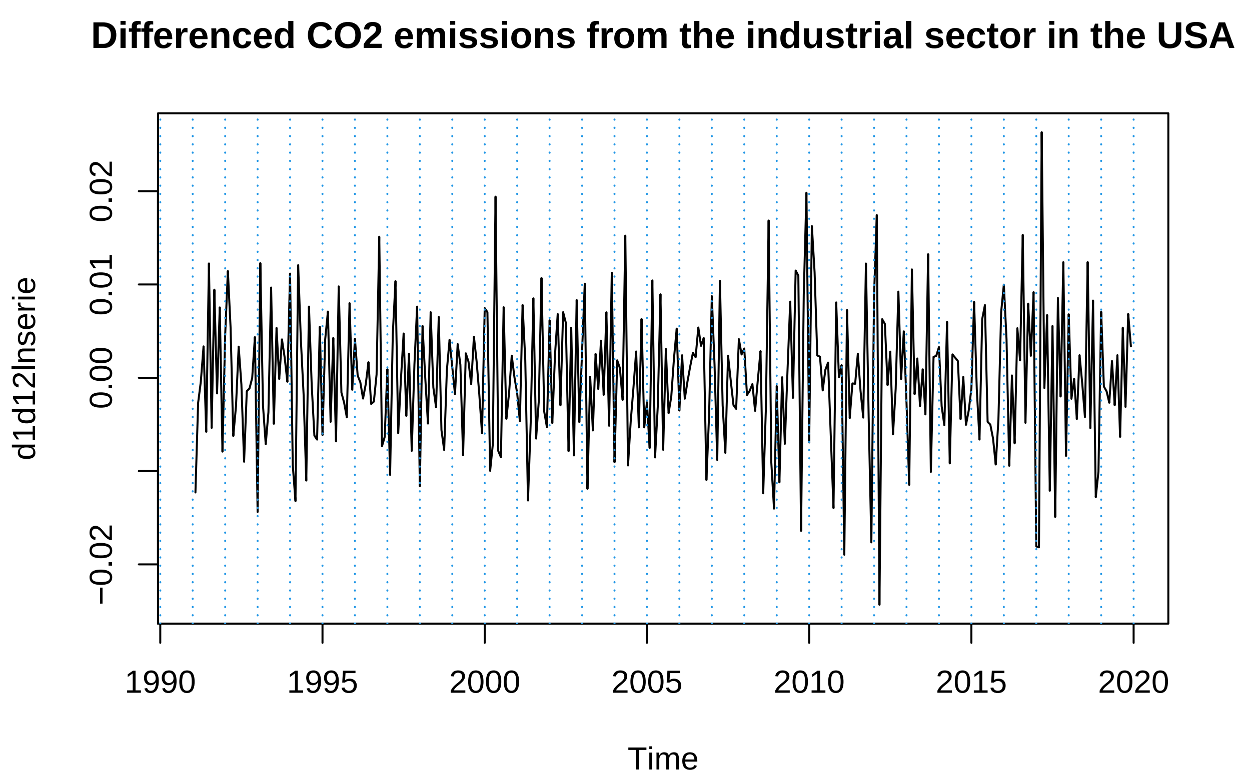
It is possible to see a seasonal component as, for example, January and August have a constant higher value whereas February, April and September have a lower value. In order to deseasonalize the time series, a difference is applied with a 12 month frequency


Log−seasonal differenced Time Series
d12lnserie
0.02
0.00
−0.02
−0.04
1990
1995
2000
2005
2010
2015
2020
Time

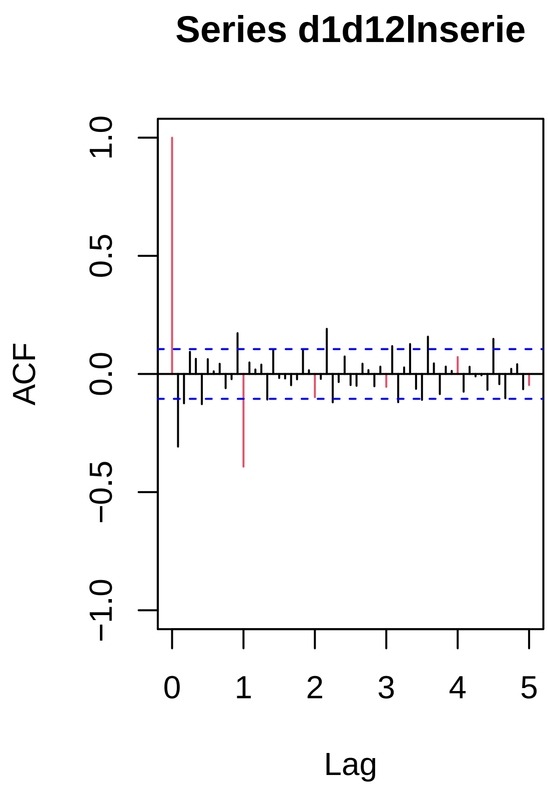
Lastly, in a stationary series the mean is equal to 0 and, although it seems to be close to it, to be on the safe side we'll run a test to see if applying two more differences is useful or not. To check that, the variance of the series is calculated and it will be over-differenced once the variance is higher than the last differenced series.

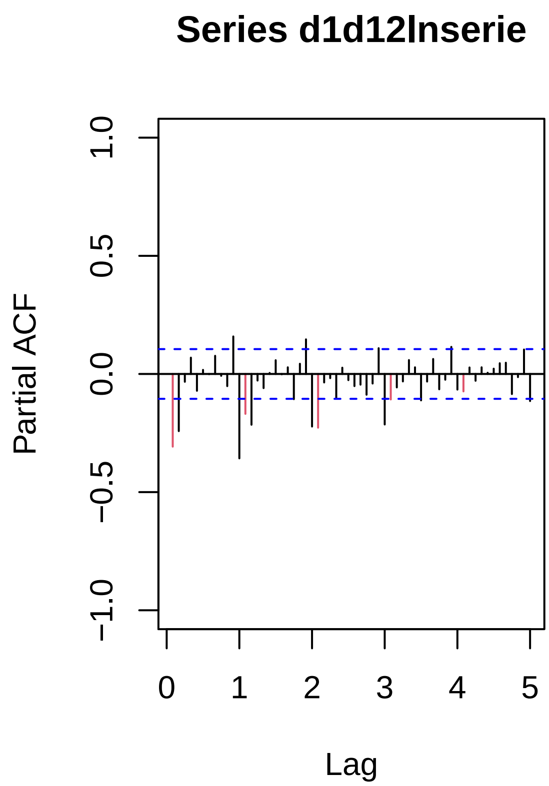
	Variance
Var serie	150.0063098
Var-lnserie	0.0003554
Var-d12lnserie	0.0000944
Var-d12d1lnserie	0.0000519
Var-d12d1d1lnserie	0.0001356

As it could be hypothesized, one regular difference is enough, two is over-differencing. So, finally, we've applied the logarithm, one difference in the seasonal 12-month component and one regular difference to reach stationarity of the series. Finally, a look at the current time series we have with said transformations.



2.2 Analyze the ACF and PACF of the stationary series to identify at least two plausible models. Reason about what features of the correlograms you use to identify these models.





First, we'll look at the regular part of the series. It clearly has a first significant lag in both the ACF and PACF which, if we think they rapidly decrease to 0, leads to think of a p=1 and q=1 . If we take the confidence intervals as a very strict measure then it is possible to see a q=5. As for the seasonal part, it clearly has a 1 significant lag in the ACF and it quickly decreases to 0 in the PACF, which leads to a p=0, q=1.

So, the two chosen models will be, in one hand an $ARIMA(1,0,1)(0,0,1)_{12}$ and on the other hand an $ARIMA(1,0,5)(0,0,1)_{12}$

3 Estimation

To proceed with the estimation we'll take as a starting point the two models chosen in the last section, we'll compute their estimate using the arima function in R from the stats package. Once they are estimated, the p-value of the parameters will be computed and if they are significant they'll remain in the model and, if not, they'll be deleted.

3.1 Use R to estimate the identified models.

$ARMA(1,0,1)(0,0,1)_{12}$

```
##
## Call:
## arima(x = d1d12lnserie, order = c(1, 0, 1), seasonal = list(order = c(0, 0,
##     1), period = 12))
##
## Coefficients:
##          ar1      ma1     sma1  intercept
##       0.1003  -0.5573  -0.8507      0e+00
## s.e.  0.1020   0.0830   0.0350      1e-04
##
## sigma^2 estimated as 2.646e-05:  log likelihood = 1328.44,  aic = -2646.89
```

```
##        ar1        ma1       sma1  intercept
##  0.9837049  6.7166530 24.3229994  0.1460198
```

The non-significant parameters are the first auto-regressive and the intercept, which will be taken out of the model. As for the intercept, now the estimated model will be based on the logarithmic series and the differences will be applied in the computation done by the function itself.

```
##
## Call:
## arima(x = lnserie, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##          ar1      ma1     sma1
##       0.0996  -0.5563  -0.8499
## s.e.  0.1019   0.0829   0.0349
##
## sigma^2 estimated as 2.647e-05:  log likelihood = 1328.39,  aic = -2648.77
```

The coefficient `ar1` is still not significant in this model without intercept so we definitely take it out.

```
##
## Call:
## arima(x = lnserie, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ma1     sma1
##       -0.4844  -0.8476
## s.e.   0.0481   0.0347
##
## sigma^2 estimated as 2.656e-05:  log likelihood = 1327.9,  aic = -2649.8
```

Note that the AIC has decreased in each step and now we have a model ($ARIMA(0,1,1)(0,1,1)_{12}$) where all parameters are significant.

$ARMA(1,0,5)(0,0,1)_{12}$

```
##
## Call:
## arima(x = d1d12lnserie, order = c(1, 0, 5), seasonal = list(order = c(0, 0,
##     1), period = 12))
##
## Coefficients:
##           ar1     ma1      ma2      ma3     ma4      ma5     sma1  intercept
##       -0.7938  0.3586  -0.4627  -0.0212  0.1080  -0.0578  -0.8499      0e+00
## s.e.   0.1560  0.1599   0.0851   0.0631  0.0653   0.0648   0.0370      1e-04
##
## sigma^2 estimated as 2.591e-05:  log likelihood = 1332.18,  aic = -2646.36
```

```
##        ar1        ma1        ma2        ma3        ma4        ma5       sma1
##  5.0872378  2.2435081  5.4343280  0.3355899  1.6546746  0.8916944 22.9536562
##  intercept
##  0.1390556
```

The intercept and some of the `ma` coefficients are non-significant. The same procedure will be applied it was done in the first model.

```
##
## Call:
## arima(x = lnserie, order = c(1, 1, 5), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ar1     ma1      ma2      ma3     ma4      ma5     sma1
##       -0.7927  0.3578  -0.4622  -0.0208  0.1086  -0.0575  -0.8490
## s.e.   0.1561  0.1600   0.0852   0.0631  0.0651   0.0647   0.0369
##
## sigma^2 estimated as 2.592e-05:  log likelihood = 1332.13,  aic = -2648.26
```

Coefficients `ma3` to `ma5` are non-significant, we take them out one by one.

```
##
## Call:
## arima(x = lnserie, order = c(1, 1, 4), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ar1     ma1      ma2      ma3     ma4     sma1
##       -0.8626  0.4277  -0.4964  -0.0442  0.1330  -0.8515
## s.e.   0.1266  0.1313   0.0745   0.0591  0.0586   0.0372
##
## sigma^2 estimated as 2.597e-05:  log likelihood = 1331.76,  aic = -2649.52
```

Now coefficient `ma4` is significant, but we still want to check if we can get a model with a lower AIC by taking it out.

```
##
## Call:
## arima(x = lnserie, order = c(1, 1, 3), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##          ar1      ma1     ma2     ma3     sma1
##       0.2171  -0.6697  0.0035  0.0894  -0.8393
## s.e.  0.6259   0.6231  0.2842  0.0682   0.0359
##
## sigma^2 estimated as 2.635e-05:  log likelihood = 1329.56,  aic = -2647.12
```

Note that the AIC has increased and the coefficients `ar1` and `ma1` are not significant now. We won't go further on this analysis and we will stay with the previous model, $ARIMA(1,1,3)(0,1,1)_{12}$.

So finally we propose two seasonal ARIMA models, $ARIMA(0,1,1)(0,1,1)_{12}$ and $ARIMA(1,1,4)(0,1,1)_{12}$, for fitting the `logseries` and they both have similar AIC.

4 Validation

In order to validate the models the residuals will be analyzed, a look at the AR and MA infinite models will be taken to see if the models are invertible and/or causal and the stability of the model will also be checked.

4.1 Perform the complete analysis of residuals, justifying all assumptions made. Use the corresponding tests and graphical results.

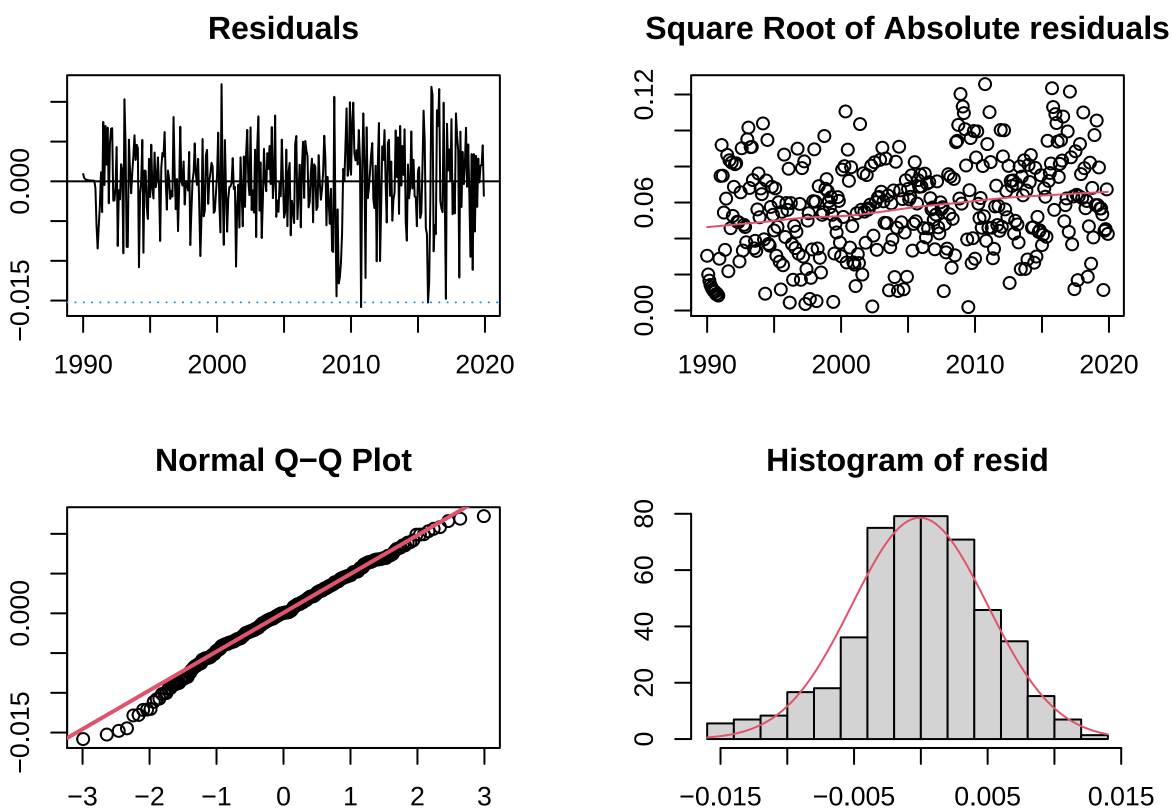
When checking the residuals, 3 aspects are analyzed:

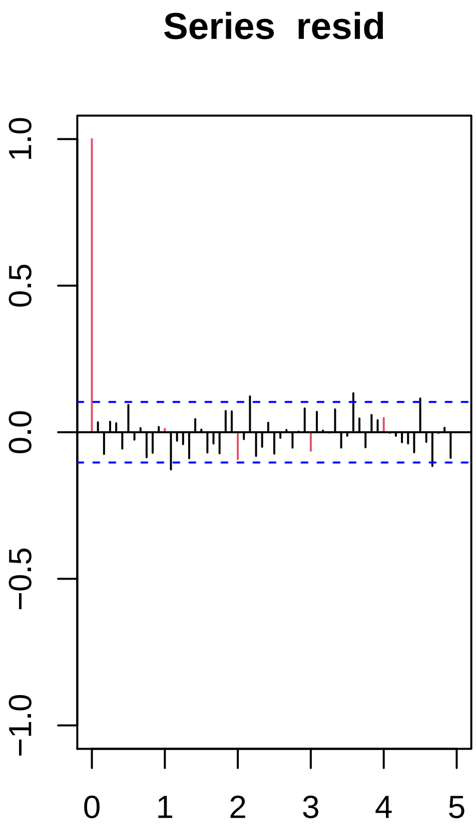
1. Homogeneity of variance, for which the residuals, the square root of absolute values of the residuals with smooth fit and the ACF and PACF of square residuals are plotted.

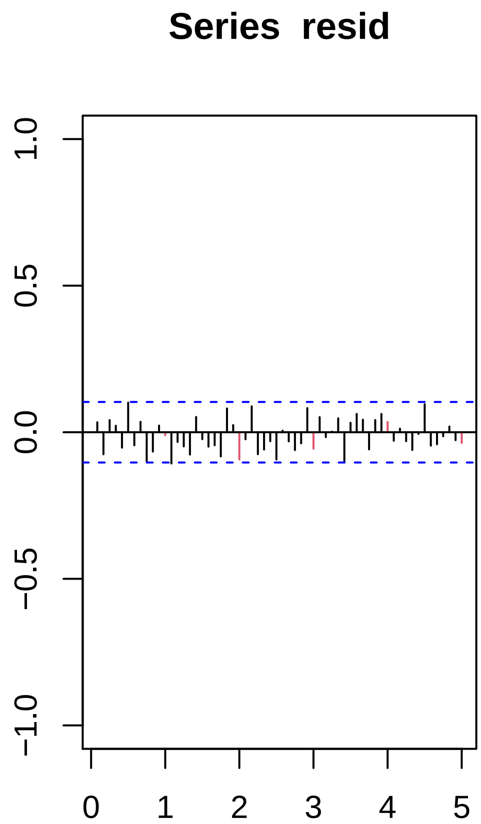
2. Normality, for which the Quantile-Quantile and the histogram with theoretical density overlapped are plotted.

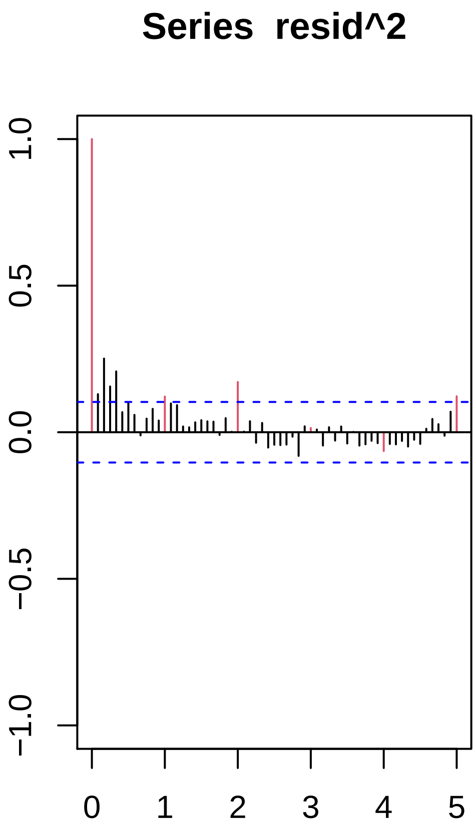
3. Independence, for which the ACF and PACF of residuals are plotted and LJung-Box test is run.

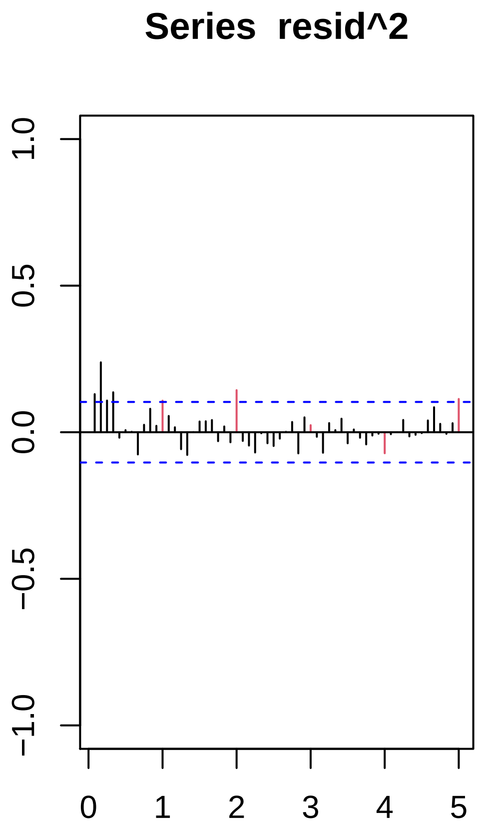
$ARIMA(0,1,1)(0,1,1)_{12}$

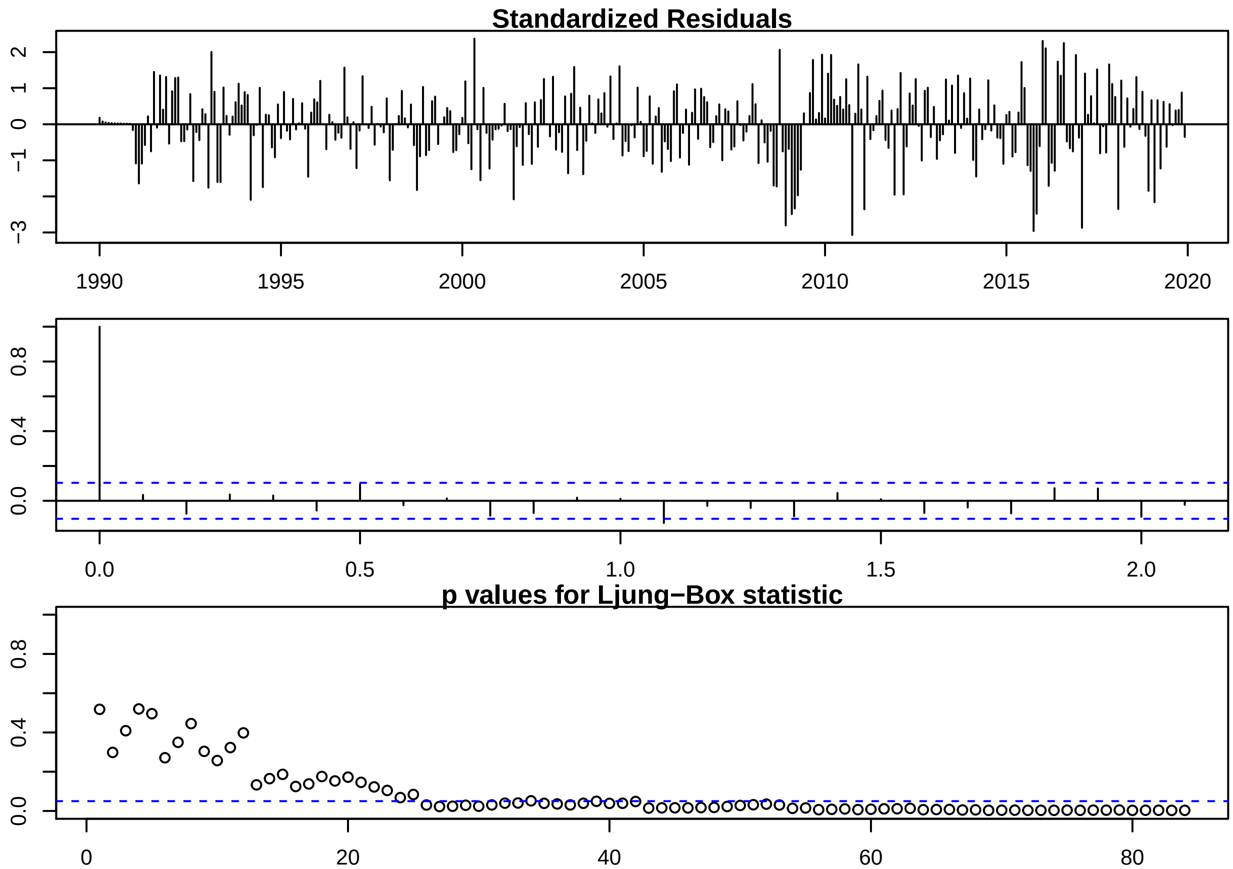



Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5

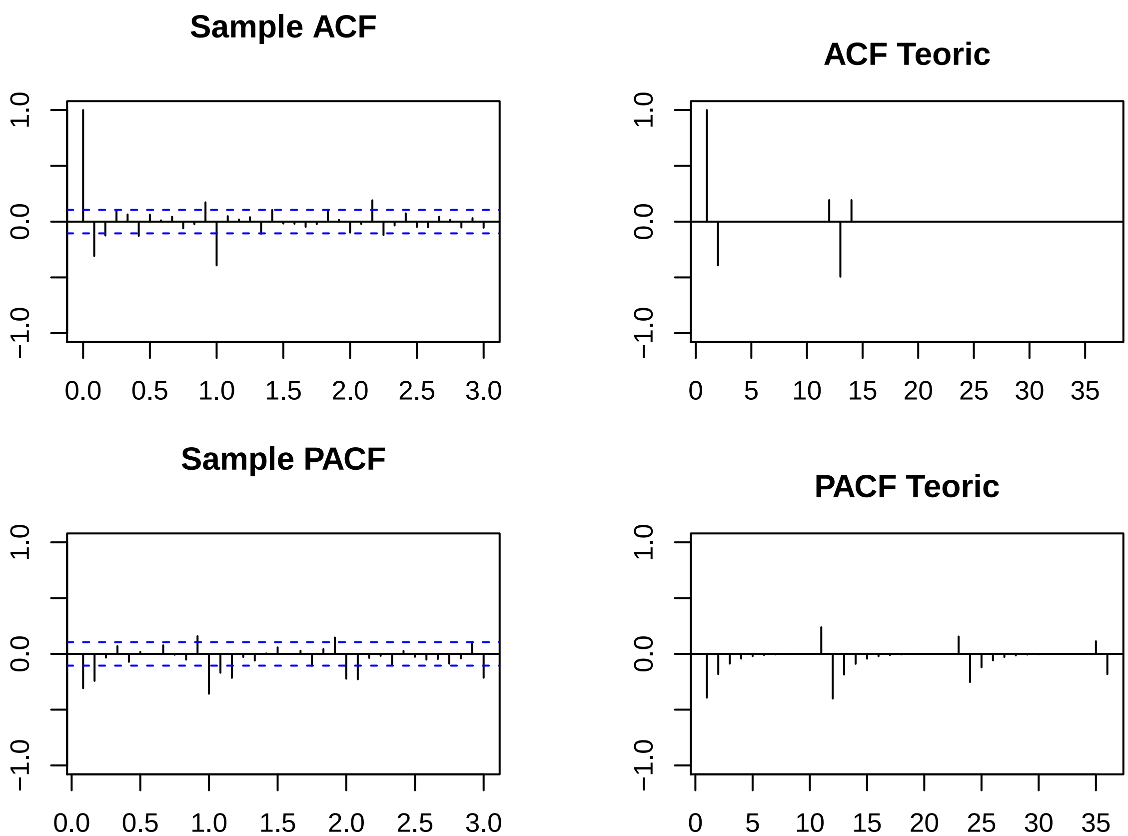

Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid^2
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid^2
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5

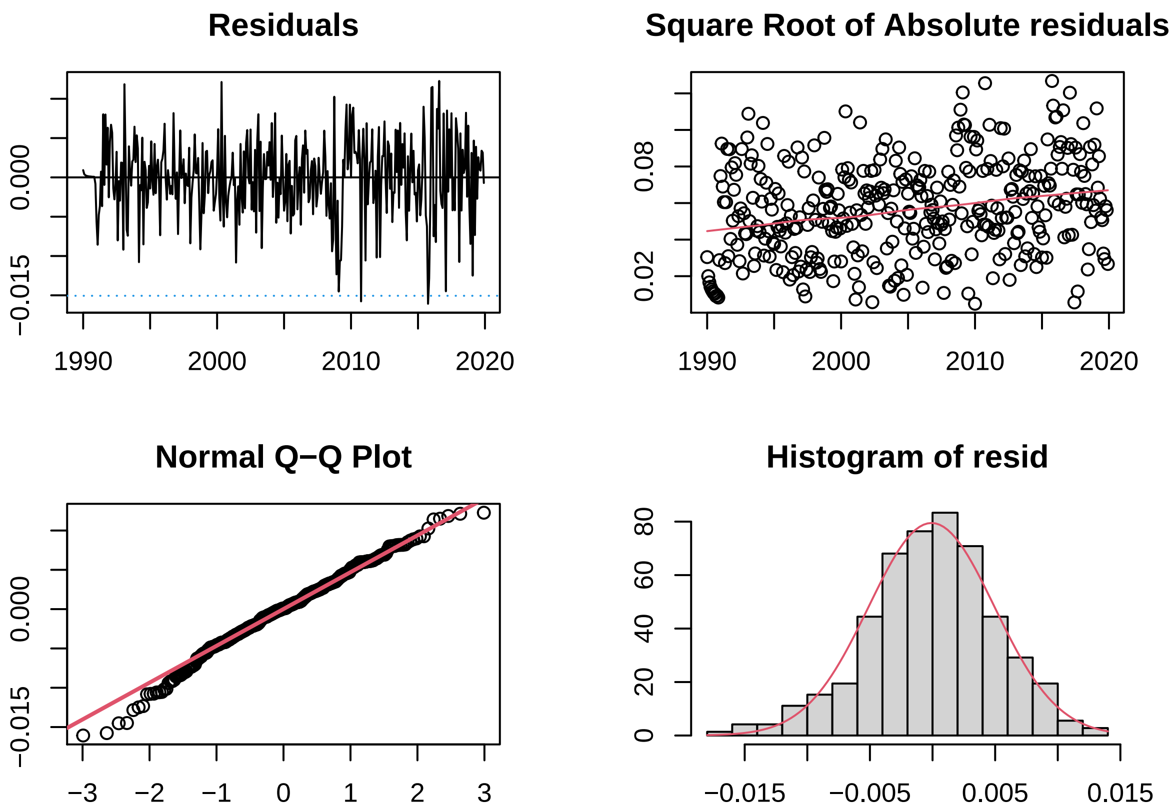


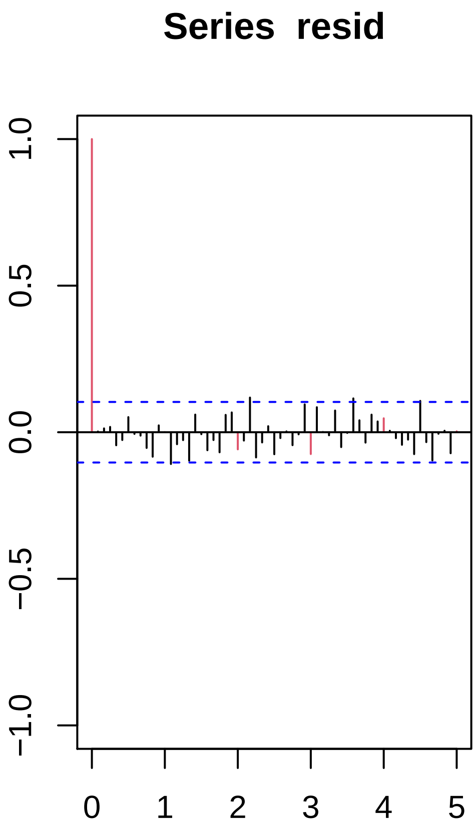

```
##
## ----------------------------------------------------------------------
##
## Call:
## arima(x = lnserie, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ma1     sma1
##       -0.4844  -0.8476
## s.e.   0.0481   0.0347
##
## sigma^2 estimated as 2.656e-05:  log likelihood = 1327.9,  aic = -2649.8
##
## Ljung-Box test
##      lag.df  statistic    p.value
## [1,]      1  0.4182792 0.51779696
## [2,]      2  2.4195221 0.29826854
## [3,]      3  2.8903181 0.40884693
## [4,]      4  3.2315818 0.51984509
## [5,]     12 12.6128999 0.39779284
## [6,]     24 35.0305870 0.06794858
## [7,]     36 52.6955937 0.03577061
## [8,]     48 70.7527662 0.01796118
```

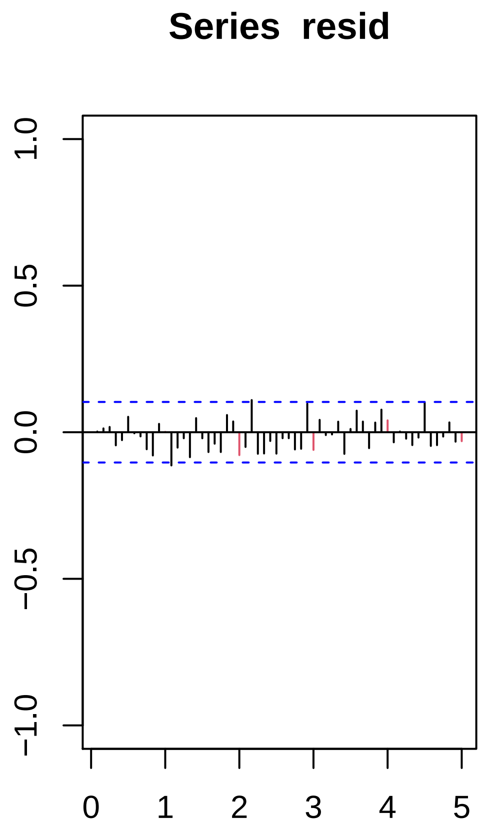


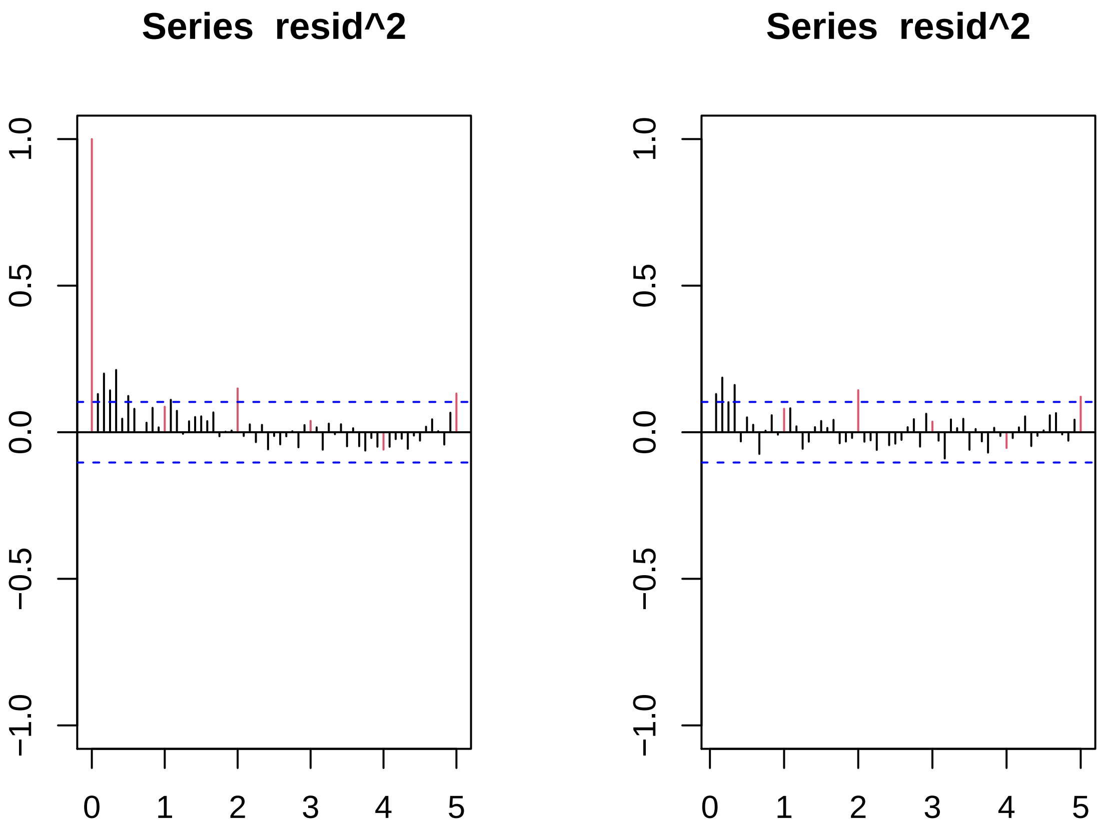
It seems that the main concern we may have is about the LJung-Box test for the independence property as p-values fall into the rejection band pretty early. About the normality of the residuals, it is close to be fulfilled but we noted that the distribution of the residuals is not symmetric with respect to zero, having more big negative residuals than positive ones. It is one of the kind of issues that one would expect to solve by applying ARIMA extensions. The variance of the residuals can also be considered non-constant, since it increases from approximately 2008.

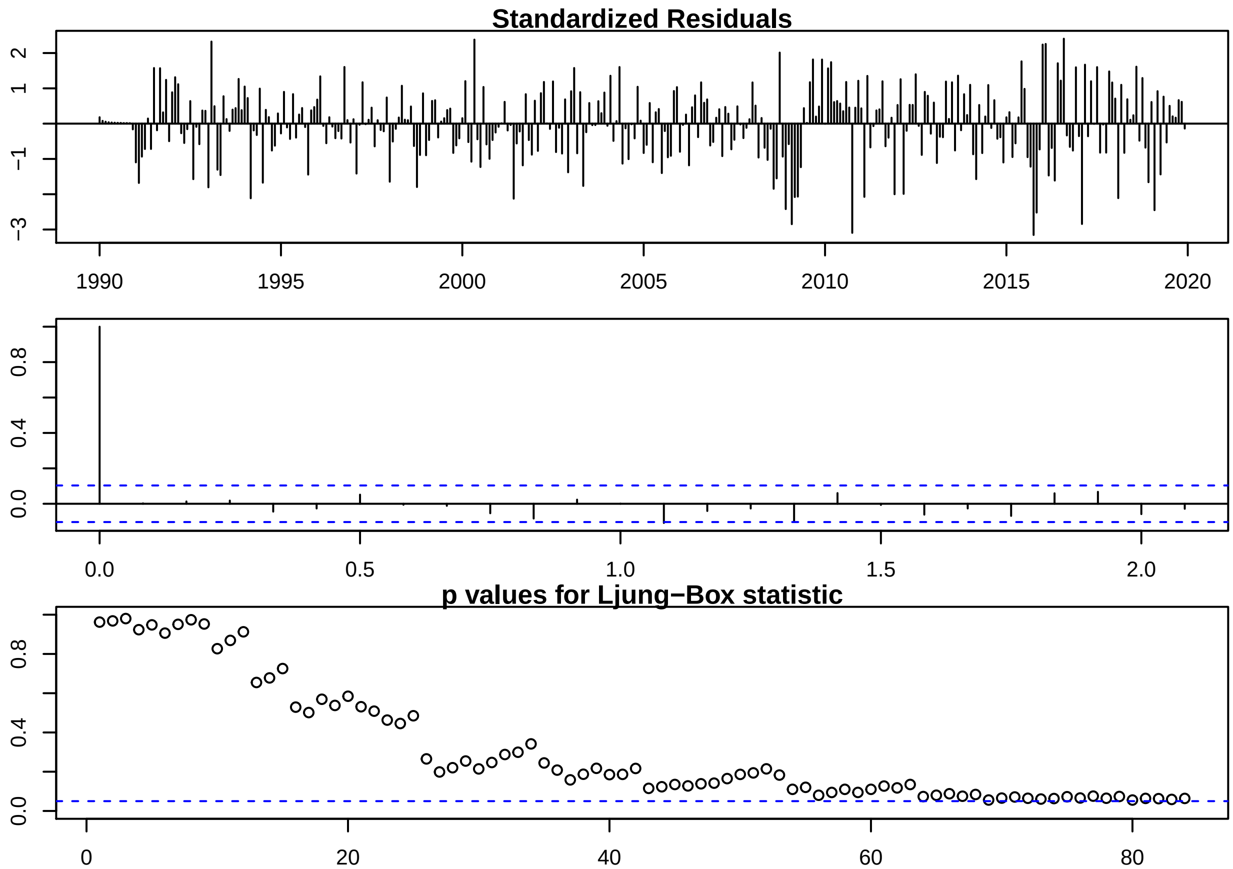
$ARIMA(1,1,4)(0,1,1)_{12}$


Residuals
Square Root of Absolute residuals
Normal Q−Q Plot
Histogram of resid
−0.015
0.000
0.02
0.08
1990
2000
2010
2020
−3
−2
−1
0
1
2
3
0
20
40
60
80
−0.015
−0.005
0.005
0.015

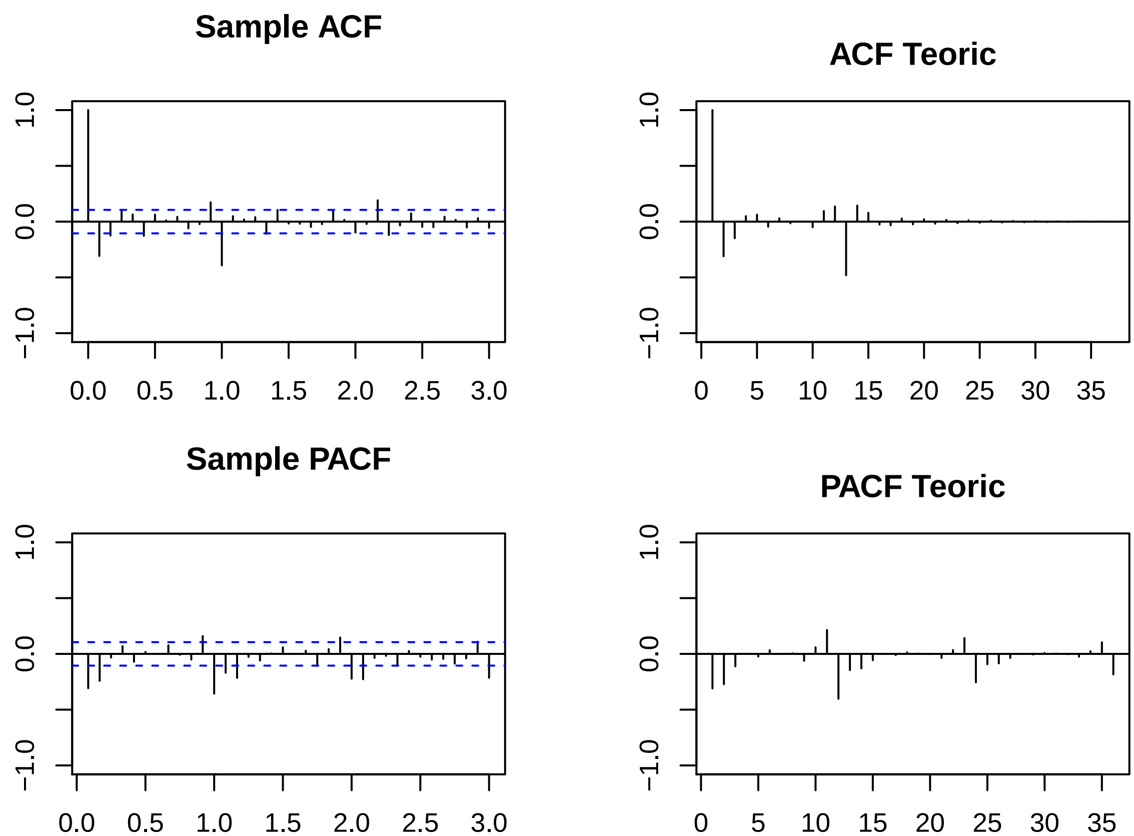

Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid^2
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5
Series resid^2
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5




```
##
## --------------------------------------------------------------------
##
## Call:
## arima(x = lnserie, order = c(1, 1, 4), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ar1     ma1      ma2      ma3     ma4      sma1
##       -0.8626  0.4277  -0.4964  -0.0442  0.1330  -0.8515
## s.e.   0.1266  0.1313   0.0745   0.0591  0.0586   0.0372
##
## sigma^2 estimated as 2.597e-05:  log likelihood = 1331.76,  aic = -2649.52
##
## Ljung-Box test
##      lag.df    statistic   p.value
## [1,]      1  0.002245209 0.9622075
## [2,]      2  0.064376945 0.9683241
## [3,]      3  0.183366691 0.9802287
## [4,]      4  0.906610864 0.9236105
## [5,]     12  6.065240688 0.9127576
## [6,]     24 24.282191770 0.4455590
## [7,]     36 42.606429035 0.2081049
## [8,]     48 58.540220859 0.1416970
```



It seems that the independence property is better fulfilled now, but p-values fall end up getting near the rejection band. About the normality of the residuals, it is once again close to be fulfilled but the distribution of the residuals is not symmetric with respect to zero (it is even more asymmetric than before). The variance is also increasing from 2008 when using this second model.

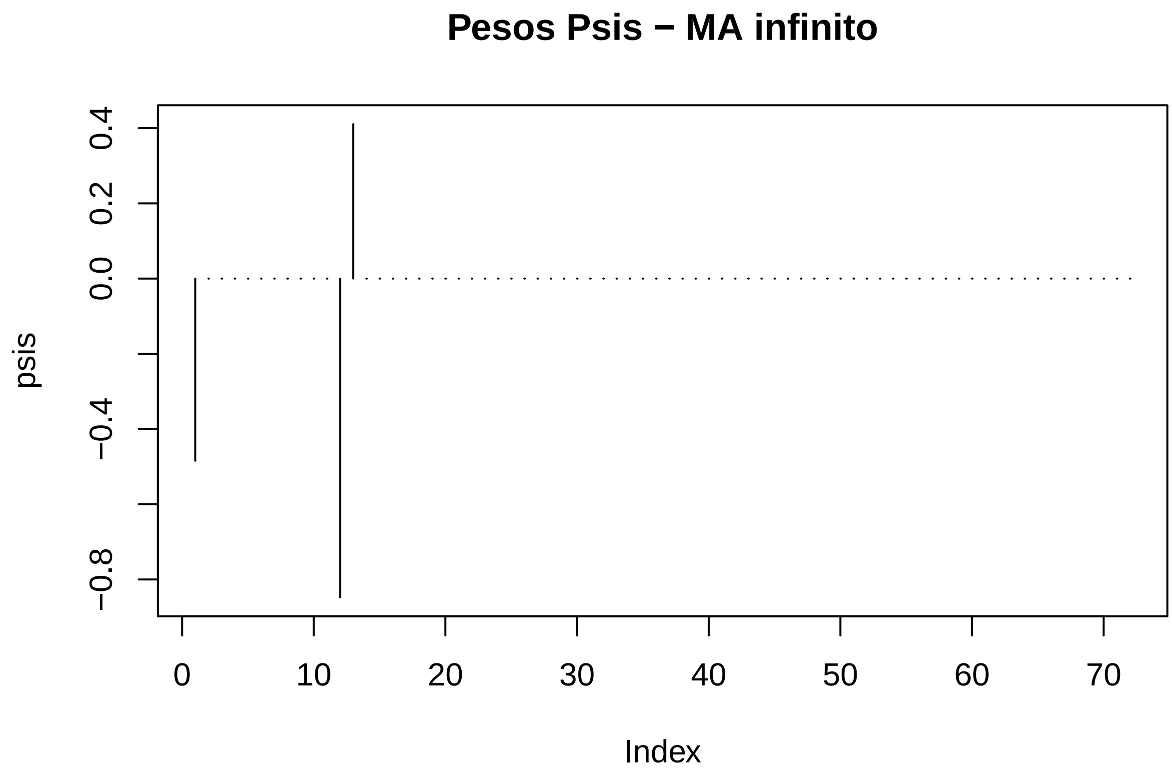
4.2 Include analysis of the expressions of the AR and MA infinite models, discuss if they are causal and/or invertible and report some adequacy measures.

To check if the models are causal and/or invertible the modul of their coefficients will be computed and if they are outside the unit root then they are causal and/or invertible.

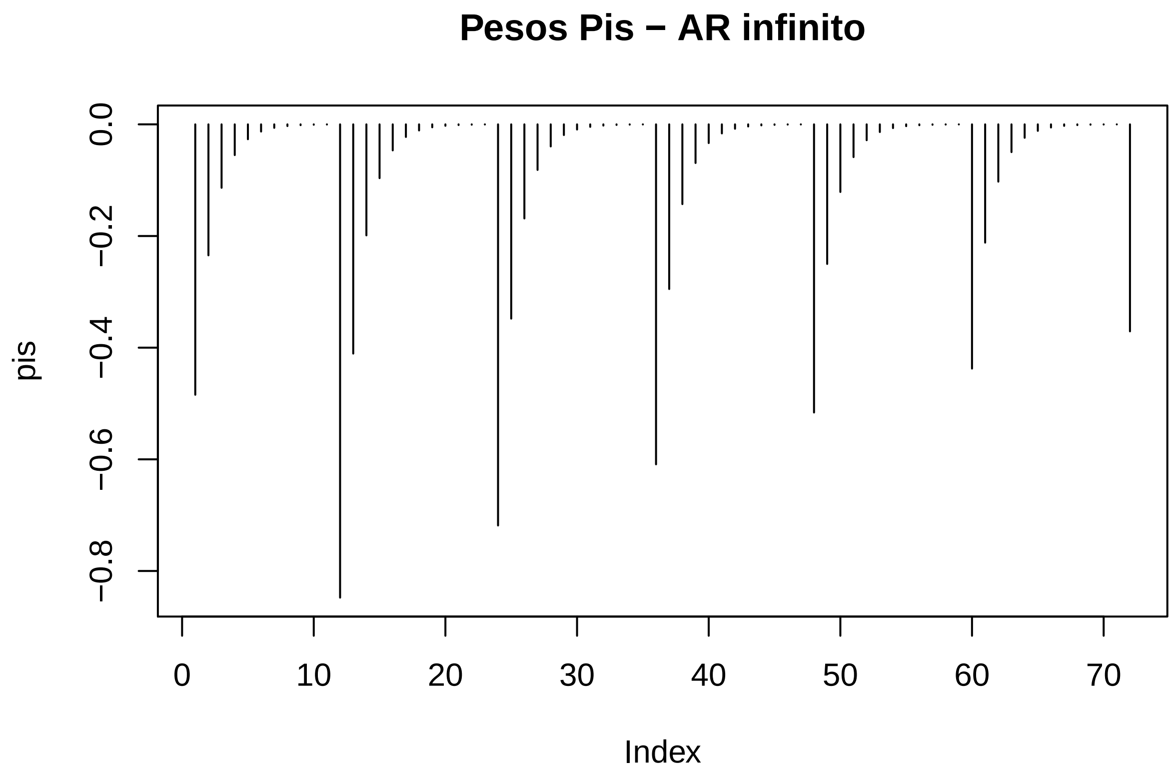
$ARIMA(0,1,1)(0,1,1)_{12}$

```
##
## Modul of AR Characteristic polynomial Roots:
##
## Modul of MA Characteristic polynomial Roots:  1.013878 1.013878 1.013878 1.013878 1.013878 1.013878 1
##
## Psi-weights (MA(inf))
##
## --------------------
##       psi 1      psi 2      psi 3      psi 4      psi 5      psi 6      psi 7
## -0.4843532  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
##       psi 8      psi 9     psi 10     psi 11     psi 12     psi 13     psi 14
##  0.0000000  0.0000000  0.0000000  0.0000000 -0.8475584  0.4105176  0.0000000
##      psi 15     psi 16     psi 17     psi 18     psi 19     psi 20
```

```
##  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
```



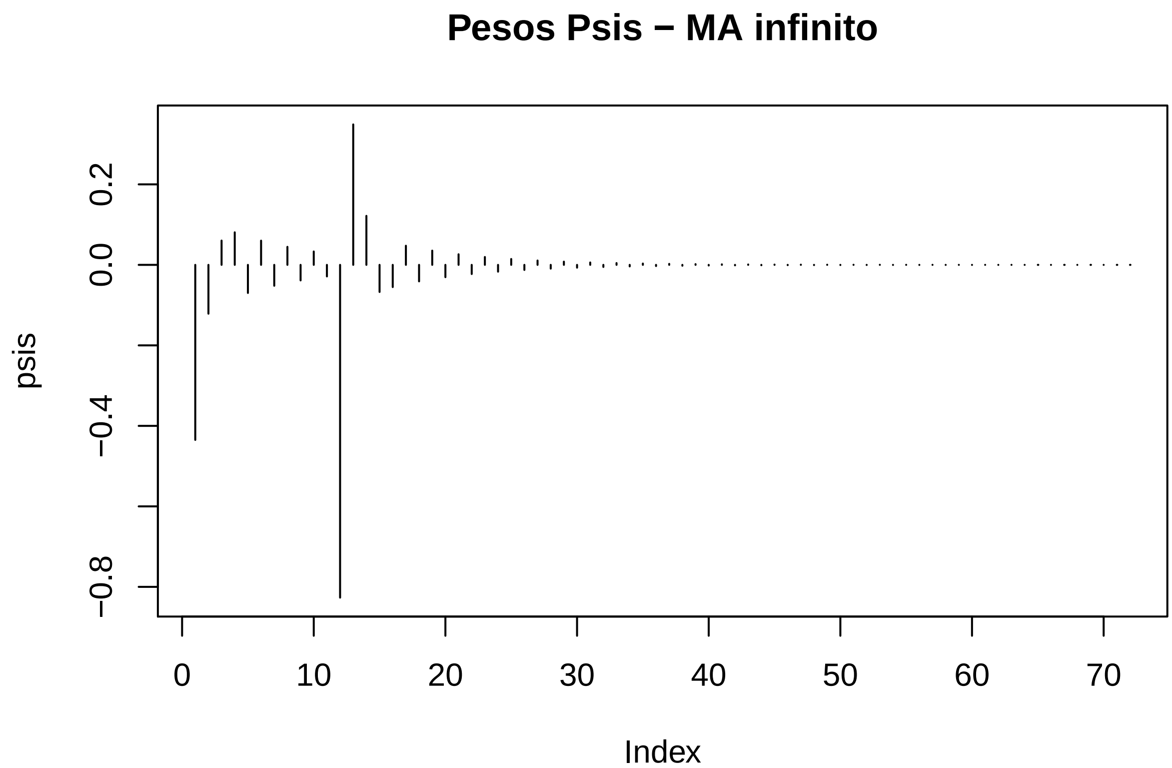

```
##
## Pi-weights (AR(inf))
##
## --------------------
##          pi 1          pi 2          pi 3          pi 4          pi 5
## -0.4843532013 -0.2345980236 -0.1136283038 -0.0550362327 -0.0266569755
##          pi 6          pi 7          pi 8          pi 9         pi 10
## -0.0129113914 -0.0062536738 -0.0030289869 -0.0014670995 -0.0007105943
##         pi 11         pi 12         pi 13         pi 14         pi 15
## -0.0003441786 -0.8477251139 -0.4105983728 -0.1988746363 -0.0963255668
##         pi 16         pi 17         pi 18         pi 19         pi 20
## -0.0466555966 -0.0225977876 -0.0109453108 -0.0053013963 -0.0025677483
```



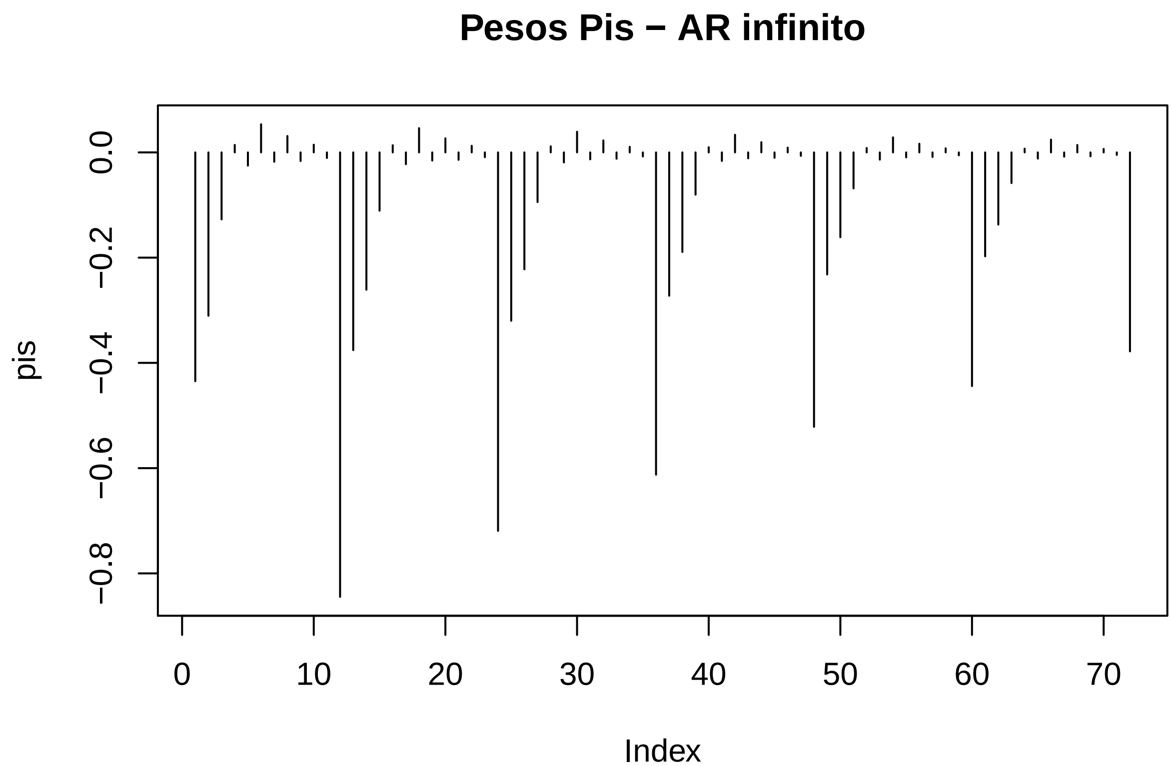
This model has no autoregressive part, therefore it is causal/stationary. The modulo of the roots of the MA Characteristic polynomial fall outside the unit root, so we could consider it as invertible. However, we should be careful since the modulo all the roots of the MA Characteristic polynomial except one are pretty near to one (`1.0138..`).

$ARIMA(1,1,4)(0,1,1)_{12}$

```
##
## Modul of AR Characteristic polynomial Roots:  1.15933
##
## Modul of MA Characteristic polynomial Roots:  1.013484 1.013484 1.013484 1.013484 1.013484 1.013484 1
##
## Psi-weights (MA(inf))
##
## --------------------
##        psi 1       psi 2       psi 3       psi 4       psi 5       psi 6
## -0.43485900 -0.12132698  0.06042427  0.08090064 -0.06978224  0.06019187
##        psi 7       psi 8       psi 9      psi 10      psi 11      psi 12
## -0.05191954  0.04478409 -0.03862929  0.03332036 -0.02874105 -0.82673310
##       psi 13      psi 14      psi 15      psi 16      psi 17      psi 18
##  0.34890898  0.12175797 -0.06736288 -0.05516528  0.04758376 -0.04104419
##       psi 19      psi 20
##  0.03540337 -0.03053779
```




```
##
## Pi-weights (AR(inf))
##
## --------------------
##        pi 1        pi 2        pi 3        pi 4        pi 5        pi 6
## -0.43485900 -0.31042933 -0.12732885  0.01414313 -0.02514256  0.05343660
##        pi 7        pi 8        pi 9       pi 10       pi 11       pi 12
## -0.01777370  0.03113572 -0.01643237  0.01459045 -0.01065649 -0.84459176
##       pi 13       pi 14       pi 15       pi 16       pi 17       pi 18
## -0.37571697 -0.26098893 -0.11082451  0.01357065 -0.02238503  0.04612635
##       pi 19       pi 20
## -0.01549889  0.02673191
```



The modulo of the only root of the AR Characteristic polynomial is `1.1589` (outside the unit circle) so the model is invertible. The modulo of the roots of the MA Characteristic polynomial again fall outside the unit root, but the modulo of most of the roots is pretty near 1, so we can say the model is causal but we should still take into account that fact.

4.3 Check the stability of the proposed models and evaluate their capability of prediction, reserving the last 12 observations.

To check the stability of the proposed models and evaluate their capability of prediction, what we'll do is to estimate each model two times, one with the whole series and one leaving out the last 12 observations. Once both are estimated we'll compare the significance, sign and magnitude of the parameters. If they are the same, then it is stable and if not, it is not stable.

$ARIMA(0,1,1)(0,1,1)_{12}$

Estimation without 12 last observations

```
##
## Call:
## arima(x = lnserie2, order = pdq, seasonal = list(order = PDQ, period = 12))
##
## Coefficients:
##           ma1     sma1
##       -0.4816  -0.8700
## s.e.   0.0504   0.0356
##
## sigma^2 estimated as 0.0006305:  log likelihood = 750.34,  aic = -1494.68
```

```
##       ma1      sma1
##  9.557673 24.452071
```

Estimation with the complete series

```
##
## Call:
## arima(x = lnserie1, order = pdq, seasonal = list(order = PDQ, period = 12))
##
## Coefficients:
##           ma1     sma1
##       -0.4880  -0.8533
## s.e.   0.0479   0.0347
##
## sigma^2 estimated as 0.0006277:  log likelihood = 778.97,  aic = -1551.95
```

```
##      ma1     sma1
## 10.19127 24.57959
```

Both estimations are very close in significance and magnitude and have the same sign, so we'll conclude that this model is stable.

$ARIMA(1,1,4)(0,1,1)_{12}$

Estimation without 12 last observations

```
##
## Call:
## arima(x = lnserie2, order = pdq, seasonal = list(order = PDQ, period = 12))
##
## Coefficients:
##           ar1     ma1      ma2      ma3     ma4     sma1
##       -0.7906  0.3857  -0.5009  -0.0647  0.1666  -0.8745
## s.e.   0.1083  0.1156   0.0691   0.0609  0.0598   0.0372
##
## sigma^2 estimated as 0.00061:  log likelihood = 755.71,  aic = -1497.41
```

```
##       ar1       ma1       ma2       ma3       ma4      sma1
##  7.298597  3.337610  7.244818  1.062570  2.784285 23.490850
```

Estimation with the complete series

```
##
## Call:
## arima(x = lnserie1, order = pdq, seasonal = list(order = PDQ, period = 12))
##
## Coefficients:
##           ar1     ma1      ma2      ma3     ma4     sma1
##       -0.8617  0.4247  -0.5018  -0.0433  0.1372  -0.8573
## s.e.   0.1232  0.1280   0.0738   0.0591  0.0582   0.0371
##
## sigma^2 estimated as 0.0006133:  log likelihood = 782.96,  aic = -1551.92
```

```
##        ar1        ma1        ma2        ma3        ma4       sma1
##  6.9955239  3.3192335  6.8043224  0.7336495  2.3590658 23.0792545
```

As in the first model, both estimations are very close in significance, sign and magnitude, so this model is also stable.

4.4 Select the best model for forecasting.

Once both models are validated and knowing both are stable, choosing a model for the forecasting step will be based on the simplicity of the model and the different criterion offered by the estimations.

Both models have passed through all the validation steps with similar outcomes, so we will keep the model with the lowest AIC which, in our case, is $ARIMA(0,1,1)(0,1,1)_{12}$. It turns out that it is also the simplest one between both.

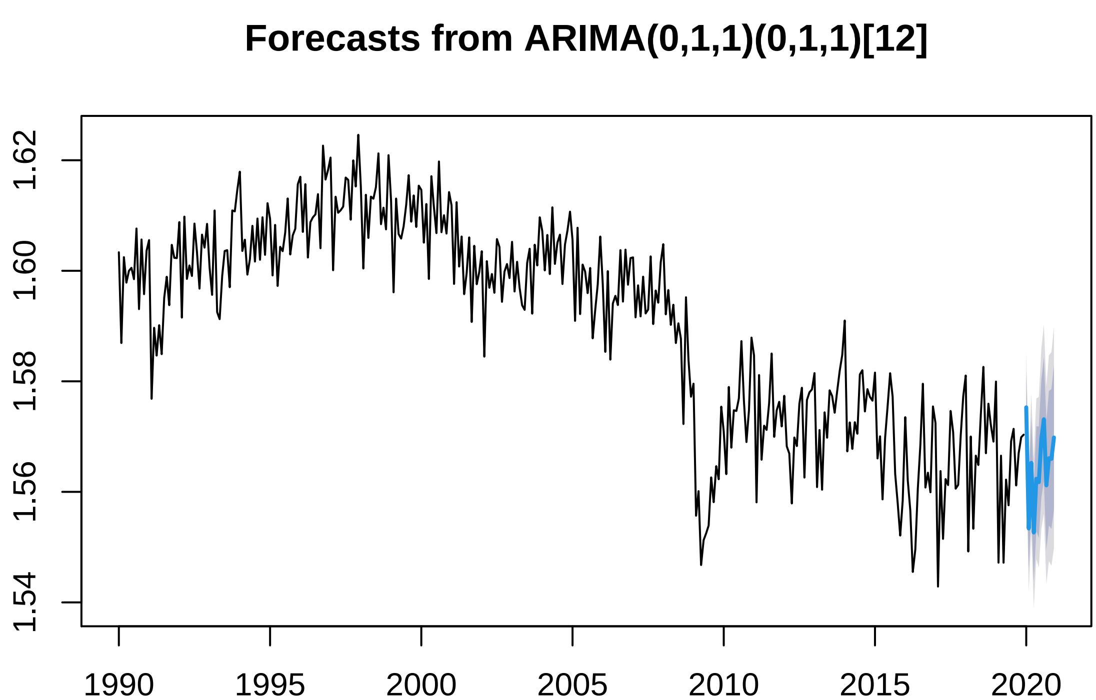
5 Prediction

In this section, the function forecast of the package forecast will be used to perform the predictions. The parameters it needs are two, the estimated model and how many periods we want it to predict. Then, it computes the predictions and calculates a 80% and 95% confidence intervals.

5.1 Obtain long term forecasts for the twelve months following the last observation available; provide also confidence intervals.

$ARIMA(0,1,1)(0,1,1)_{12}$

```
##          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## Jan 2020       1.575292 1.568687 1.581896 1.565190 1.585393
## Feb 2020       1.553370 1.545939 1.560801 1.542005 1.564735
## Mar 2020       1.565248 1.557074 1.573422 1.552746 1.577750
## Apr 2020       1.552644 1.543788 1.561499 1.539100 1.566187
## May 2020       1.562368 1.552880 1.571856 1.547858 1.576878
## Jun 2020       1.561761 1.551680 1.571841 1.546344 1.577178
## Jul 2020       1.569625 1.558985 1.580266 1.553352 1.585898
## Aug 2020       1.573136 1.561964 1.584308 1.556050 1.590222
## Sep 2020       1.561174 1.549494 1.572854 1.543311 1.579037
## Oct 2020       1.566085 1.553919 1.578252 1.547479 1.584692
## Nov 2020       1.565967 1.553333 1.578601 1.546645 1.585288
## Dec 2020       1.569830 1.556745 1.582915 1.549818 1.589841
```



- Accuracy measurements

```
##                         ME       RMSE         MAE          MPE     MAPE
## Training set -8.464947e-05 0.00506891 0.003953969 -0.005963917 0.249339
##                   MASE       ACF1
## Training set 0.5027677 0.03394492
```

6 Outlier Treatment:

6.1 Analyze of the Calendar Effects are significant.

In the original series, we can note a fall in mean CO2 emissions around 2009. In this webpage, they say that "The economic downturn, combined with natural gas displacing some coal as a source of electricity generation, is projected to lead to a 5 percent decline in fossil-fuel based (carbon dioxide) emissions in 2009". About the industry, they say that "Fuel switching by electricity generators and declines in industrial use were projected to lead to a 7.9 percent decline in carbon emissions from coal in 2009".

We're going to take that fact into account when doing the calendar effects analysis, so we will be creating an auxiliary variable for data before/from 2009 and also variables for Easter and trading days configurations of the corresponding month. Then, we will fit all pertinent models and check their AIC and coefficients levels of significance to choose one.

```
##
## Call:
## arima(x = lnserie, order = pdq, seasonal = list(order = PDQ, period = 12), xreg = wTradDays)
##
## Coefficients:
##           ma1     sma1  wTradDays
##       -0.4446  -0.8402      4e-04
## s.e.   0.0514   0.0342      1e-04
##
## sigma^2 estimated as 2.449e-05:  log likelihood = 1342.27,  aic = -2676.54
```

```
##
## Call:
## arima(x = lnserie, order = pdq, seasonal = list(order = PDQ, period = 12), xreg = wEast)
##
## Coefficients:
##           ma1     sma1    wEast
##       -0.4829  -0.8464  -0.0012
## s.e.   0.0482   0.0347   0.0011
##
## sigma^2 estimated as 2.648e-05:  log likelihood = 1328.48,  aic = -2648.97
```

```
##
## Call:
## arima(x = lnserie, order = pdq, seasonal = list(order = PDQ, period = 12), xreg = from2009)
##
## Coefficients:
##           ma1     sma1  from2009
##       -0.5542  -0.8440   -0.0135
## s.e.   0.0506   0.0349    0.0044
##
## sigma^2 estimated as 2.593e-05:  log likelihood = 1332.14,  aic = -2656.28
```

```
##
## Call:
## arima(x = lnserie, order = pdq, seasonal = list(order = PDQ, period = 12), xreg = data.frame(wTradDay
##     wEast))
##
```

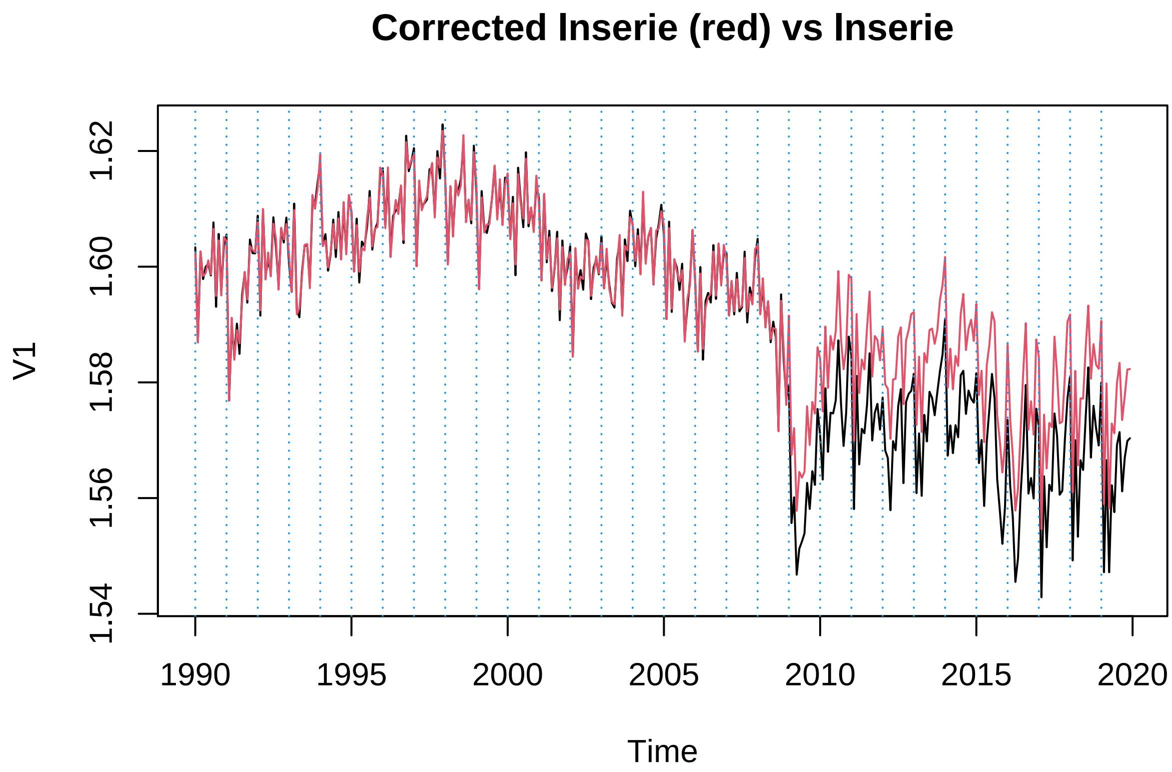
```
## Coefficients:
##           ma1     sma1  wTradDays   wEast
##       -0.4440  -0.8396      4e-04  -7e-04
## s.e.   0.0514   0.0342      1e-04   1e-03
##
## sigma^2 estimated as 2.446e-05:  log likelihood = 1342.48,  aic = -2674.97
```

```
##
## Call:
## arima(x = lnserie, order = pdq, seasonal = list(order = PDQ, period = 12), xreg = data.frame(from2009
##     wEast))
##
## Coefficients:
##           ma1     sma1  from2009    wEast
##       -0.5528  -0.8429   -0.0135  -0.0012
## s.e.   0.0507   0.0349    0.0044   0.0011
##
## sigma^2 estimated as 2.585e-05:  log likelihood = 1332.72,  aic = -2655.44
```

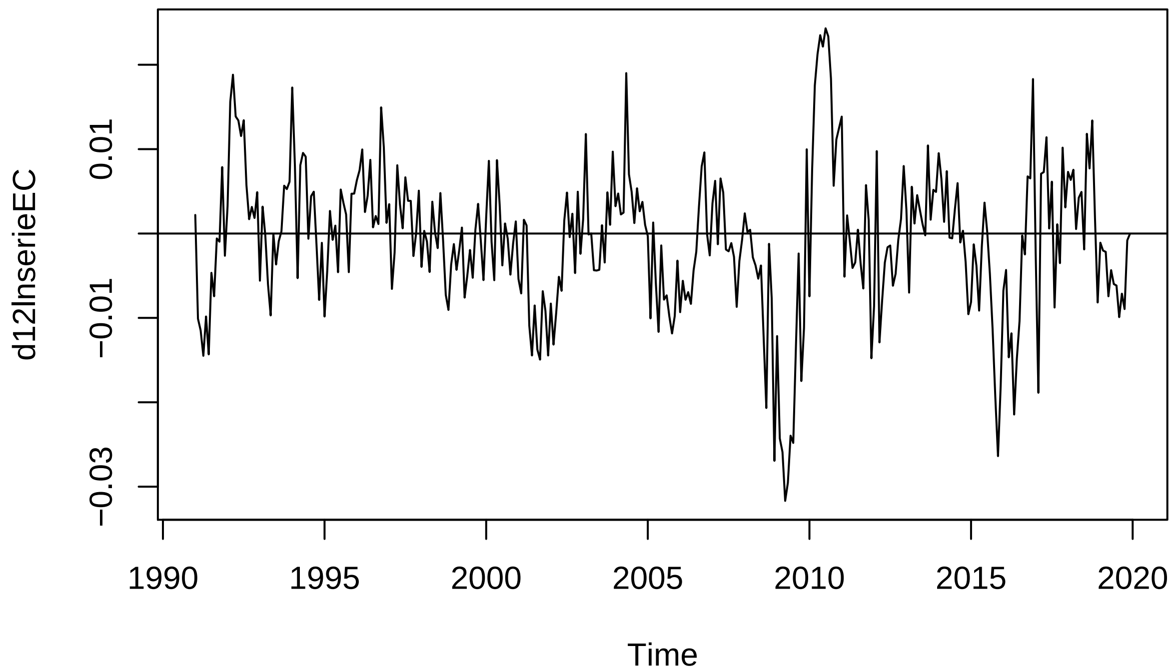
```
##
## Call:
## arima(x = lnserie, order = pdq, seasonal = list(order = PDQ, period = 12), xreg = data.frame(wTradDay
##     from2009))
##
## Coefficients:
##           ma1     sma1  wTradDays  from2009
##       -0.5124  -0.8374      4e-04   -0.0118
## s.e.   0.0546   0.0343      1e-04    0.0044
##
## sigma^2 estimated as 2.404e-05:  log likelihood = 1345.5,  aic = -2681
```

```
##
## Call:
## arima(x = lnserie, order = pdq, seasonal = list(order = PDQ, period = 12), xreg = data.frame(wTradDay
##     wEast, from2009))
##
## Coefficients:
##           ma1     sma1  wTradDays   wEast  from2009
##       -0.5120  -0.8368      4e-04  -7e-04   -0.0118
## s.e.   0.0547   0.0344      1e-04   1e-03    0.0044
##
## sigma^2 estimated as 2.402e-05:  log likelihood = 1345.72,  aic = -2679.45
```

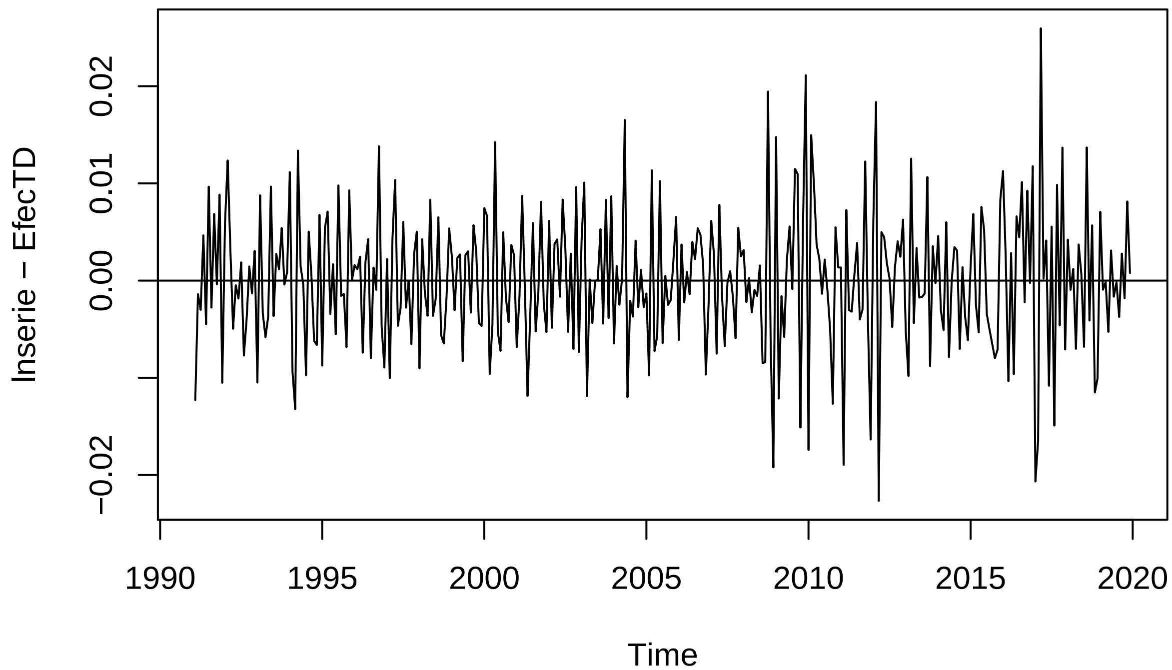
Note that the `wEast` coefficient is never significant. Besides, the found model with lowest AIC is the one that includes correction only for trading days and "2009 effect". Now we are going to estimate the calendar effects and get the corrected series.



Next, we see which transformations are needed to make this new series stationary. First, we eliminate the seasonal component by taking an order 12 seasonal difference.



Note that the mean is not constant, we take a regular difference.



Now the mean seems to be constant equal zero. We check if an extra regular difference is needed:

```
## var(lnserieEC) = 0.0002003915
```

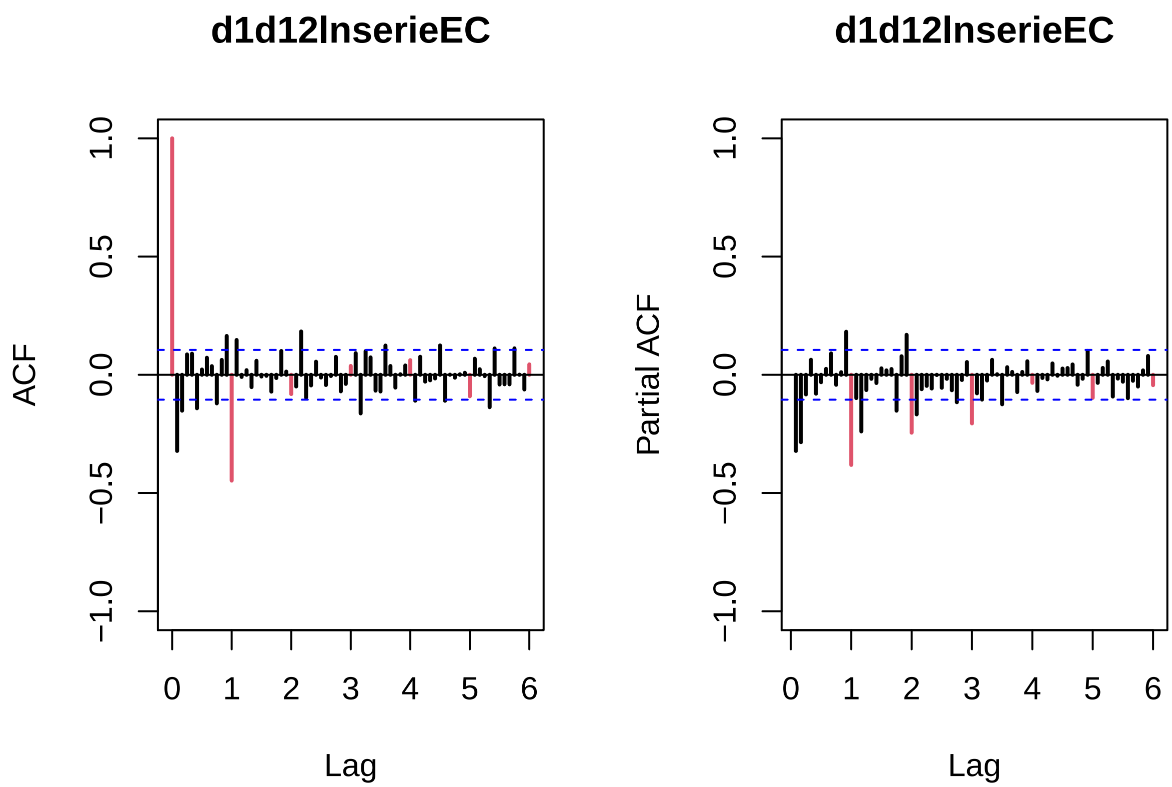
```
## var(d12lnserieEC) = 7.537243e-05
```

```
## var(d1d12lnserieEC) = 5.022209e-05
```

```
## var(diff(d1d12lnserieEC)) = 0.0001327281
```

An extra regular difference artificially increases the variance.

Now let's identify some plausible model for this data and see if we should select it instead of the non-extended ARIMA.



We propose AR(2)/ARMA(1,1) for the regular part and MA(1) for the seasonal part.

```
##
## Call:
## arima(x = lnserie, order = c(2, 1, 0), seasonal = list(order = c(0, 1, 1), period = 12),
##     xreg = data.frame(wTradDays, from2009))
##
## Coefficients:
##           ar1      ar2     sma1  wTradDays  from2009
##       -0.4043  -0.2643  -0.8209      4e-04   -0.0098
## s.e.   0.0536   0.0537   0.0349      1e-04    0.0044
##
## sigma^2 estimated as 2.424e-05:  log likelihood = 1344.68,  aic = -2677.35
```

```
##
## Call:
## arima(x = lnserie, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12),
##     xreg = data.frame(wTradDays, from2009))
##
## Coefficients:
##          ar1      ma1     sma1  wTradDays  from2009
##       0.1712  -0.6359  -0.8424      4e-04   -0.0124
## s.e.  0.0983   0.0791   0.0345      1e-04    0.0045
##
## sigma^2 estimated as 2.381e-05:  log likelihood = 1347.05,  aic = -2682.09
```

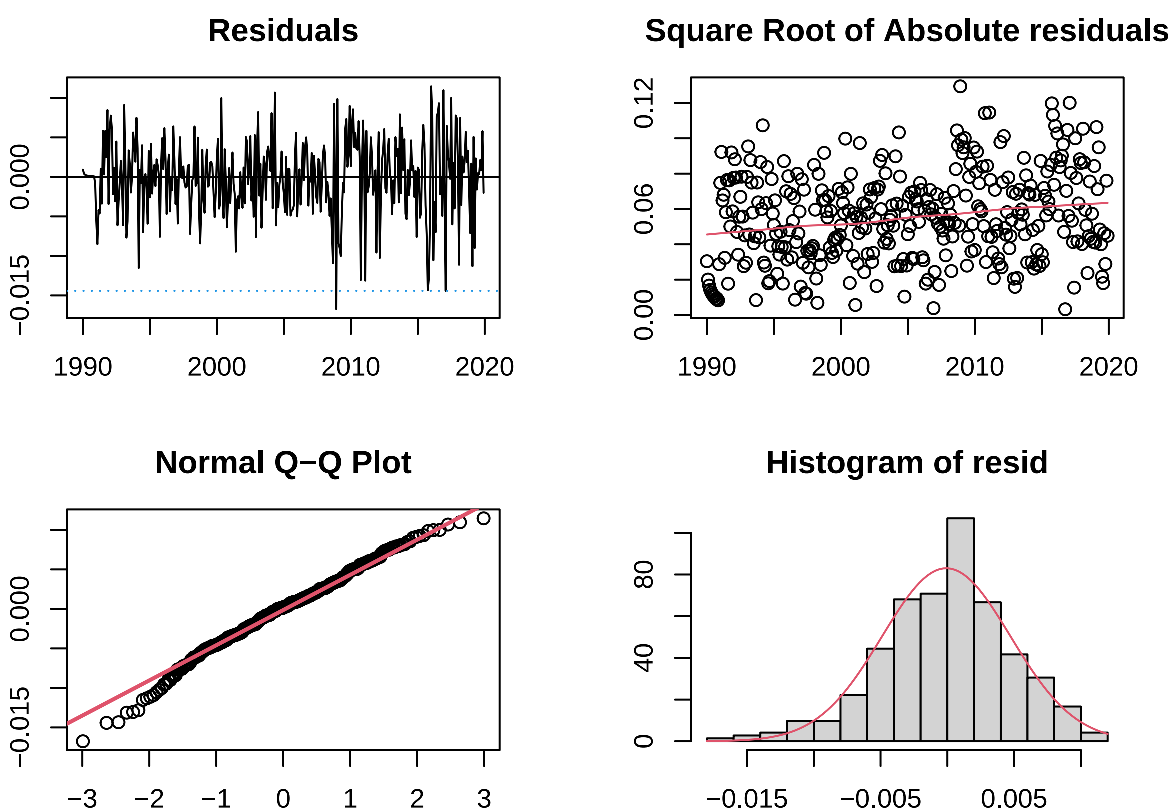
Note that the second model has the lowest AIC seen until now, but the coefficient `ar1` is not significant. Let's see if AIC improves by removing it:

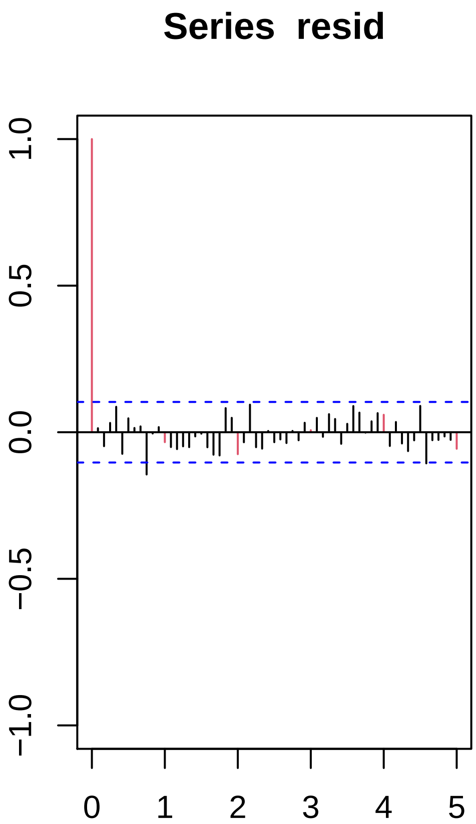
```
##
## Call:
## arima(x = lnserie, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12),
##     xreg = data.frame(wTradDays, from2009))
##
## Coefficients:
##           ma1     sma1  wTradDays  from2009
##       -0.5124  -0.8374      4e-04   -0.0118
## s.e.   0.0546   0.0343      1e-04    0.0044
##
## sigma^2 estimated as 2.404e-05:  log likelihood = 1345.5,  aic = -2681
```

AIC does not decrease, so we stay with $ARIMA(1,1,1)(0,1,1)_{12}$ for the corrected logseries.

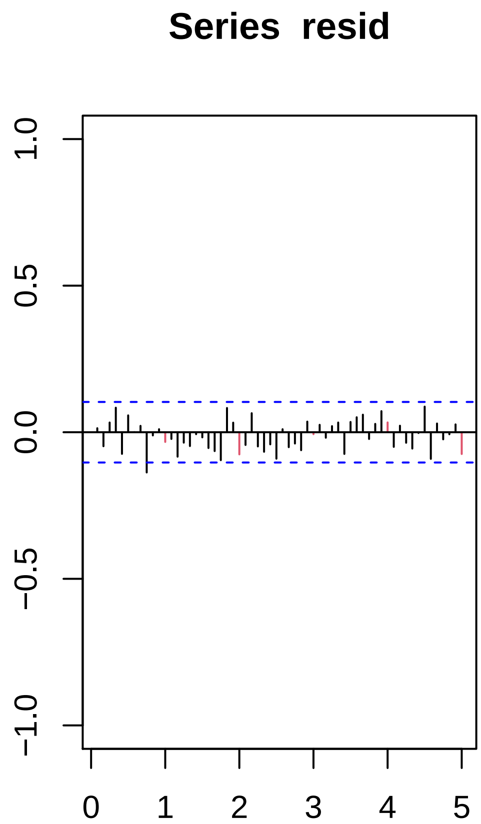
Let's validate this model now:

Residuals analysis

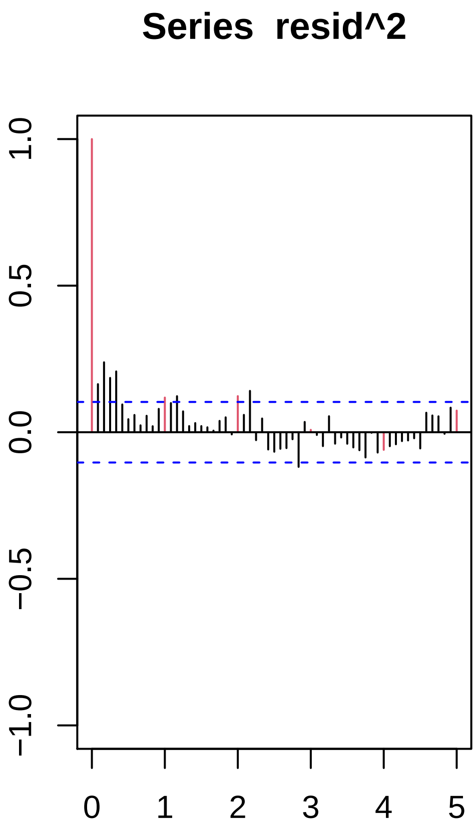


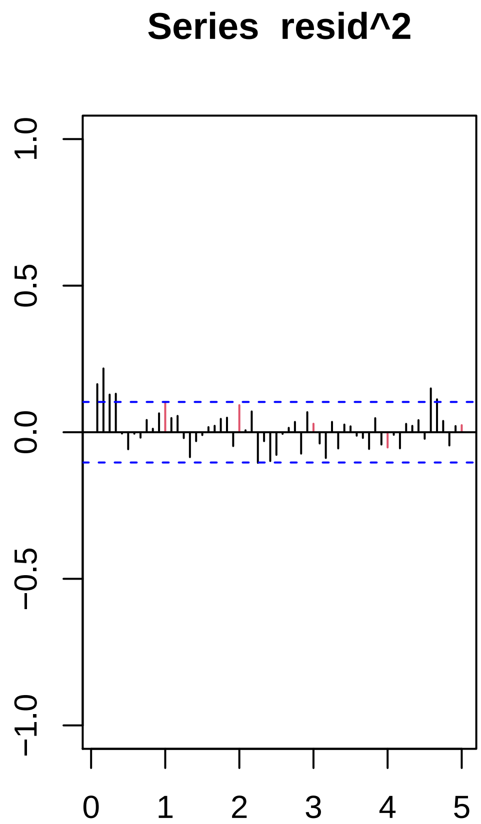


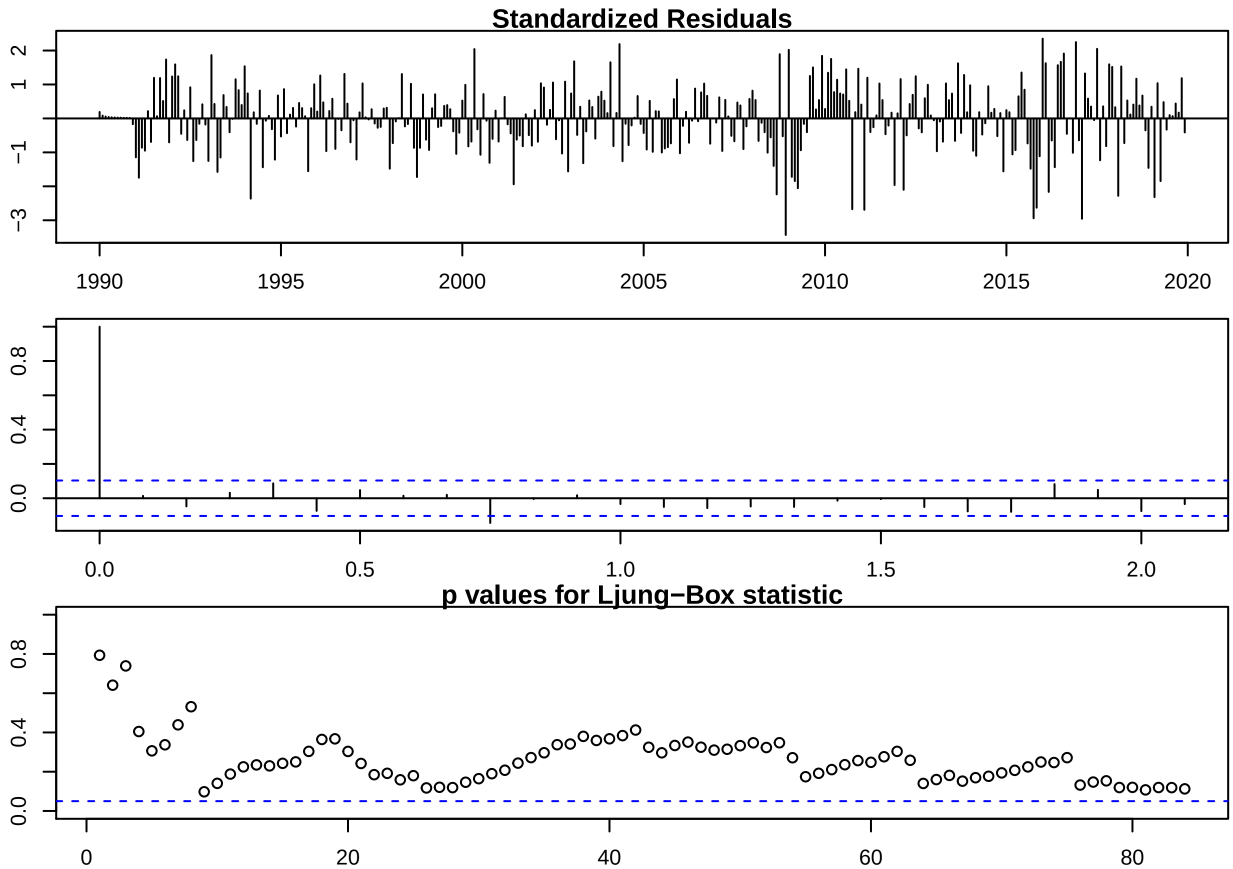
Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5



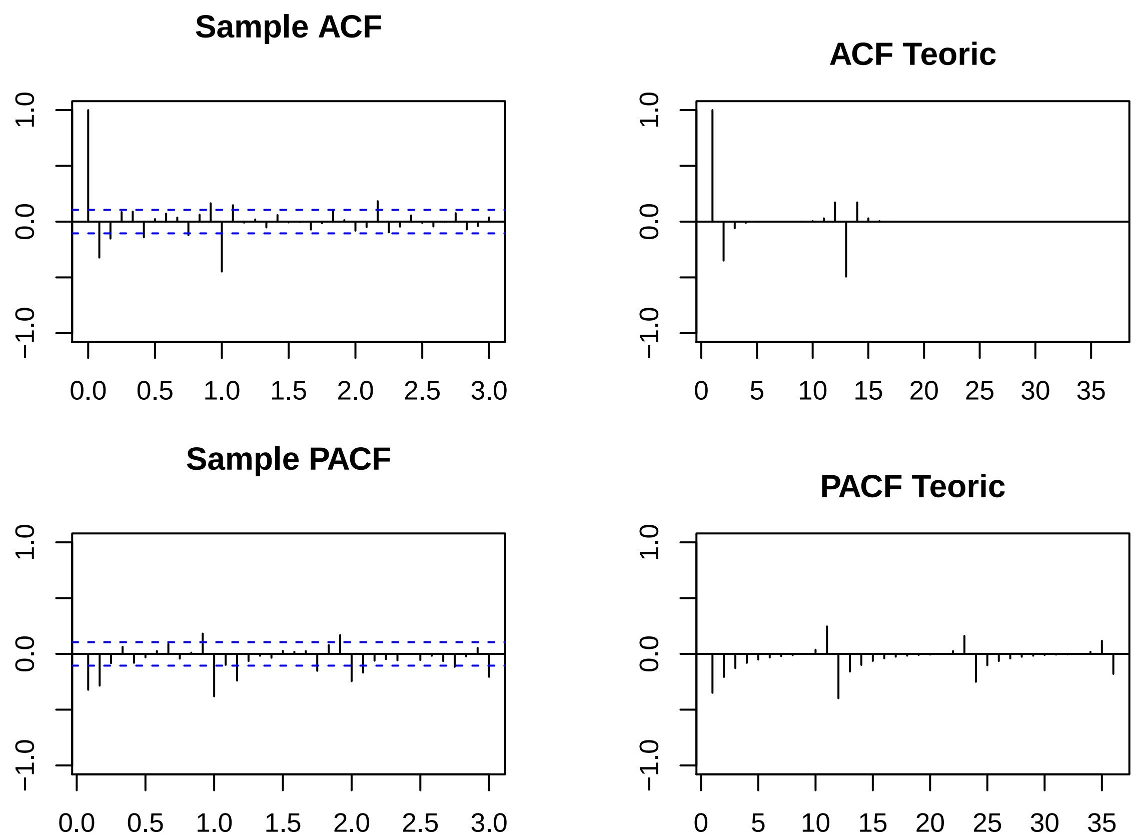
Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid^2
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid^2
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5

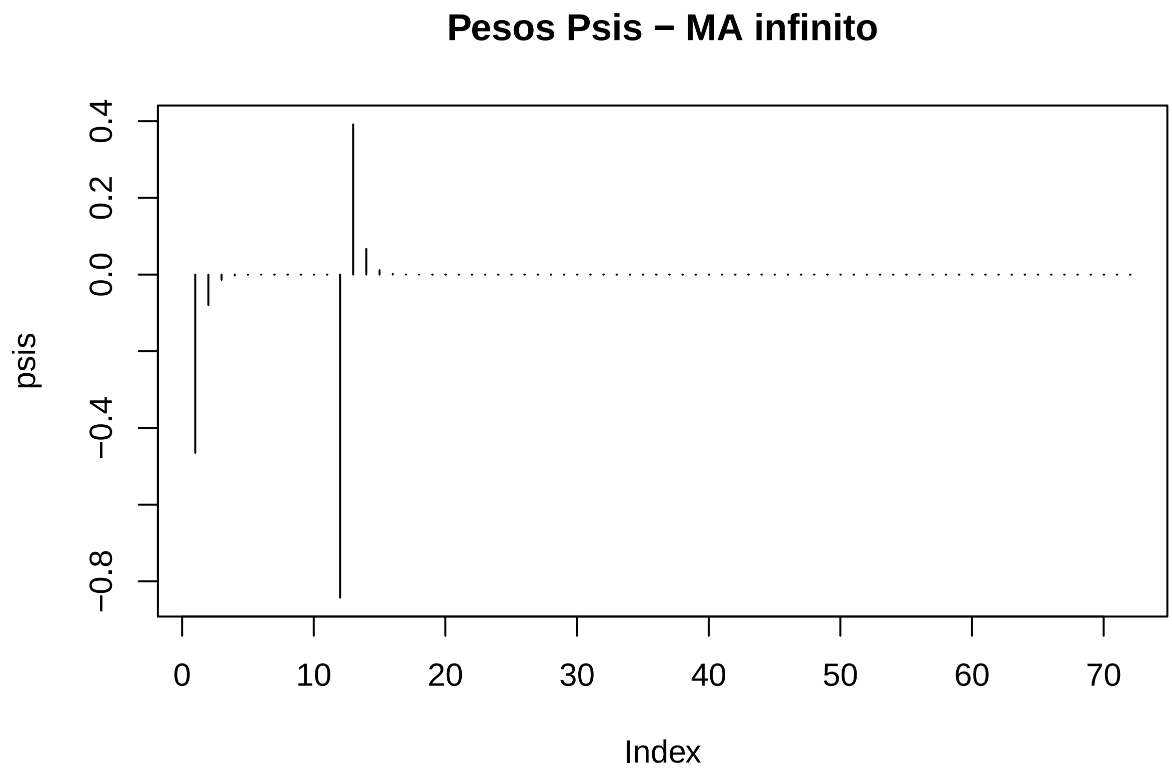



```
##
## --------------------------------------------------------------------
##
## Call:
## arima(x = lnserie, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12),
##     xreg = data.frame(wTradDays, from2009))
##
## Coefficients:
##          ar1      ma1     sma1  wTradDays  from2009
##       0.1712  -0.6359  -0.8424      4e-04   -0.0124
## s.e.  0.0983   0.0791   0.0345      1e-04    0.0045
##
## sigma^2 estimated as 2.381e-05:  log likelihood = 1347.05,  aic = -2682.09
##
## Ljung-Box test
##      lag.df   statistic   p.value
## [1,]      1  0.06867857 0.7932706
## [2,]      2  0.89002408 0.6408166
## [3,]      3  1.25850567 0.7390091
## [4,]      4  4.00921358 0.4047604
## [5,]     12 15.31364563 0.2247306
## [6,]     24 30.83204373 0.1586847
## [7,]     36 38.96940317 0.3376409
## [8,]     48 52.32031369 0.3099547
```

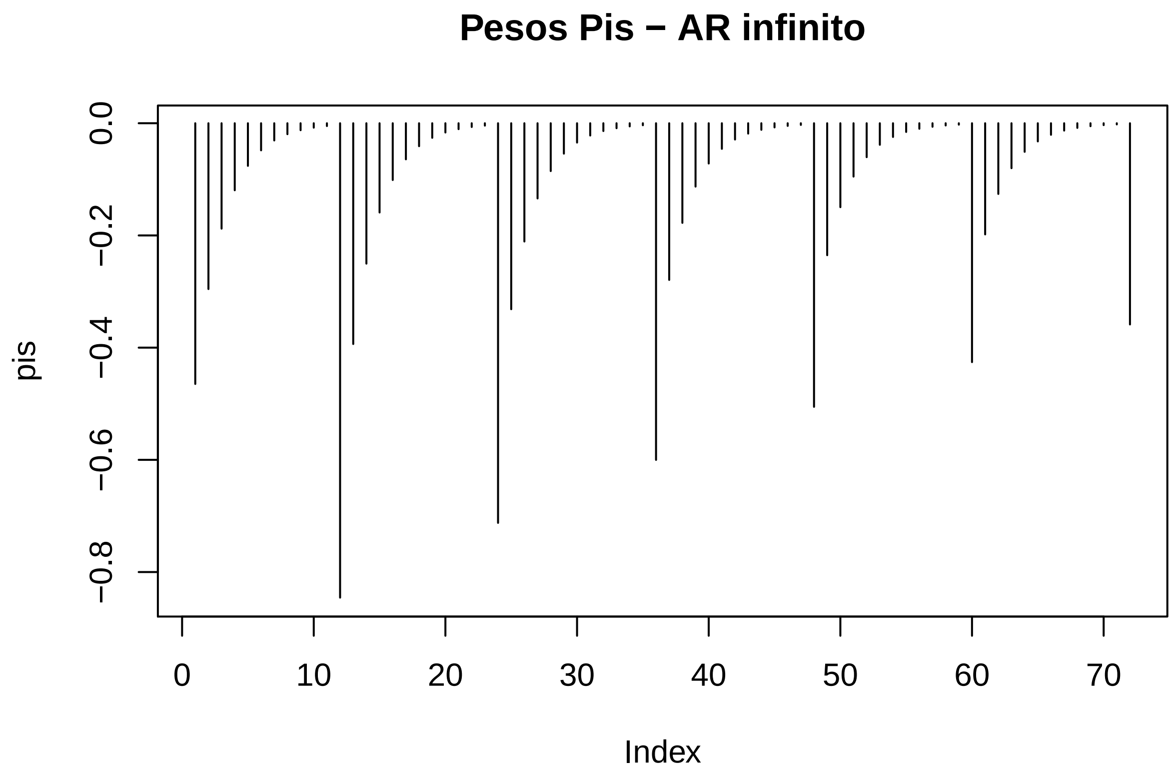


Infinite models: causality and invertibility

```
##
## Modul of AR Characteristic polynomial Roots:  5.841997
##
## Modul of MA Characteristic polynomial Roots:  1.014394 1.014394 1.014394 1.014394 1.014394 1.014394 1
##
## Psi-weights (MA(inf))
##
## --------------------
##         psi 1         psi 2         psi 3         psi 4         psi 5
## -4.647347e-01 -7.955066e-02 -1.361703e-02 -2.330887e-03 -3.989880e-04
##         psi 6         psi 7         psi 8         psi 9        psi 10
## -6.829651e-05 -1.169061e-05 -2.001133e-06 -3.425426e-07 -5.863450e-08
##        psi 11        psi 12        psi 13        psi 14        psi 15
## -1.003672e-08 -8.424014e-01  3.914932e-01  6.701359e-02  1.147101e-02
##        psi 16        psi 17        psi 18        psi 19        psi 20
##  1.963542e-03  3.361080e-04  5.753308e-05  9.848187e-06  1.685757e-06
```




```
##
## Pi-weights (AR(inf))
##
## --------------------
##         pi 1         pi 2         pi 3         pi 4         pi 5         pi 6
## -0.464734683 -0.295528981 -0.187929548 -0.119506097 -0.075995006 -0.048325911
##         pi 7         pi 8         pi 9        pi 10        pi 11        pi 12
## -0.030730883 -0.019542046 -0.012426964 -0.007902418 -0.005025219 -0.845597015
##        pi 13        pi 14        pi 15        pi 16        pi 17        pi 18
## -0.393525262 -0.250246268 -0.159133862 -0.101194660 -0.064350598 -0.040921126
##        pi 19        pi 20
## -0.026022114 -0.016547697
```



Stability

Estimation without last 12 observations:

```
##
## Call:
## arima(x = lnserie2, order = pdq, seasonal = list(order = PDQ, period = 12),
##     xreg = data.frame(wTradDays2, from20092))
##
## Coefficients:
##           ma1     sma1  wTradDays2  from20092
##       -0.5090  -0.8564      0.0019     -0.056
## s.e.   0.0576   0.0351      0.0003      0.022
##
## sigma^2 estimated as 0.000569:  log likelihood = 768.08,  aic = -1526.16
```

```
##        ma1       sma1 wTradDays2  from20092
##   8.831314  24.378742   5.357969   2.548525
```

Estimation with the complete series:

```
##
## Call:
## arima(x = lnserie1, order = pdq, seasonal = list(order = PDQ, period = 12),
##     xreg = data.frame(wTradDays1, from20091))
##
```

```
## Coefficients:
##           ma1     sma1  wTradDays1  from20091
##       -0.5142  -0.8428      0.0018    -0.0567
## s.e.   0.0544   0.0344      0.0003     0.0216
##
## sigma^2 estimated as 0.0005677:  log likelihood = 796.78,  aic = -1583.55
```

```
##        ma1       sma1 wTradDays1  from20091
##   9.456815  24.530505   5.326536   2.626939
```

ote that we still have some issues with the normality of the residuals (residuals histogram/q-qplot). Also note that the variance of the residuals is still higher for the latest observations. The model is causal by a very small margin (roots with modulo approx `1.01`) and invertible. The model is also stable as the sign, order of magnitude and significance of the coefficients doesn't change drastically when fitting the incomplete series.

6.2 For the last selected model, apply the automatic detection of outliers and its treatment. Try to give the interpretation of detected outliers

```
## Estimated residual variance after outliers detection and treatment: 1.667091e-05
```

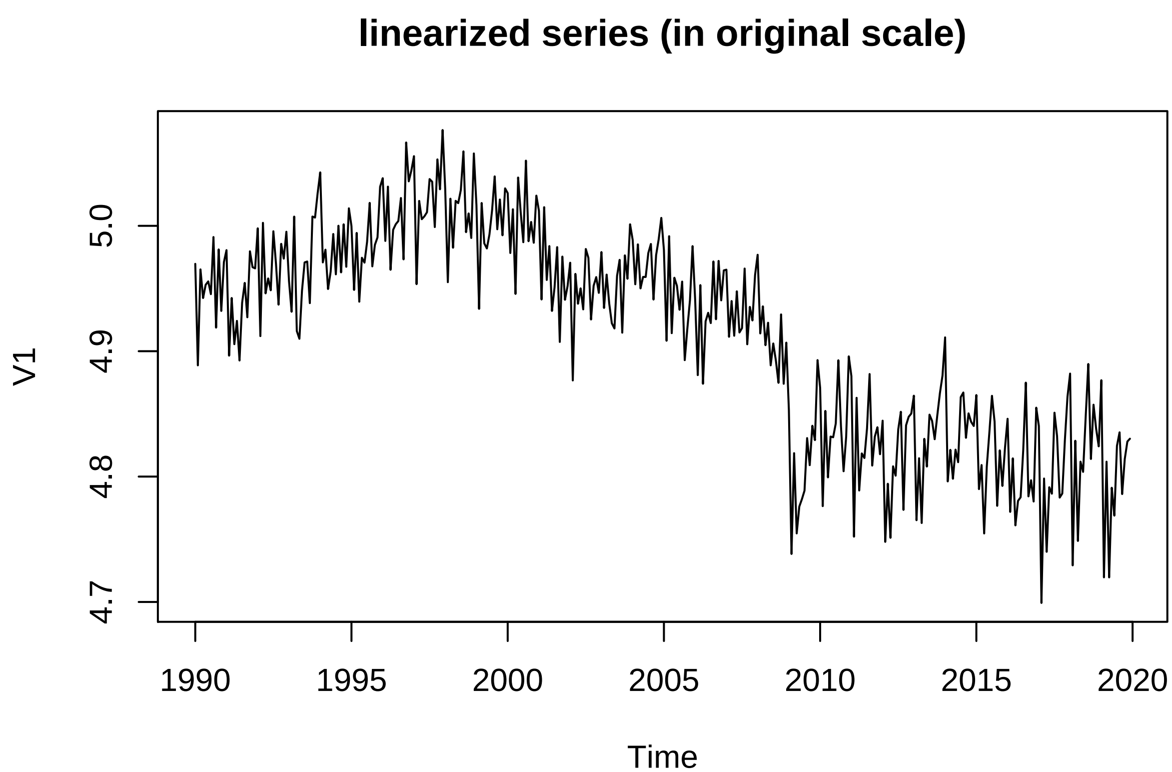
Table with detected outliers, their types, magnitud, statistic values and cronology

```
##    Obs type_detected     W_coeff ABS_L_Ratio    Fecha  perc.Obs
## 8   14            TC -0.01167651    3.100440 Feb 1991  98.83914
## 9  173            AO  0.01022374    2.939404 May 2004 101.02762
## 6  225            AO -0.01179250    3.261804 Sep 2008  98.82768
## 3  228            AO -0.01341843    3.549249 Dic 2008  98.66712
## 4  231            LS -0.01237681    3.357216 Mar 2009  98.76995
## 5  242            LS  0.01188932    3.270822 Feb 2010 101.19603
## 7  266            AO  0.01100434    3.075302 Feb 2012 101.10651
## 2  310            LS -0.01448445    3.768487 Oct 2015  98.56199
## 10 313            AO  0.01030627    2.957137 Ene 2016 101.03596
## 1  314            AO  0.01433763    3.608153 Feb 2016 101.44409
## 11 318            LS  0.01042880    3.061124 Jun 2016 101.04834
```

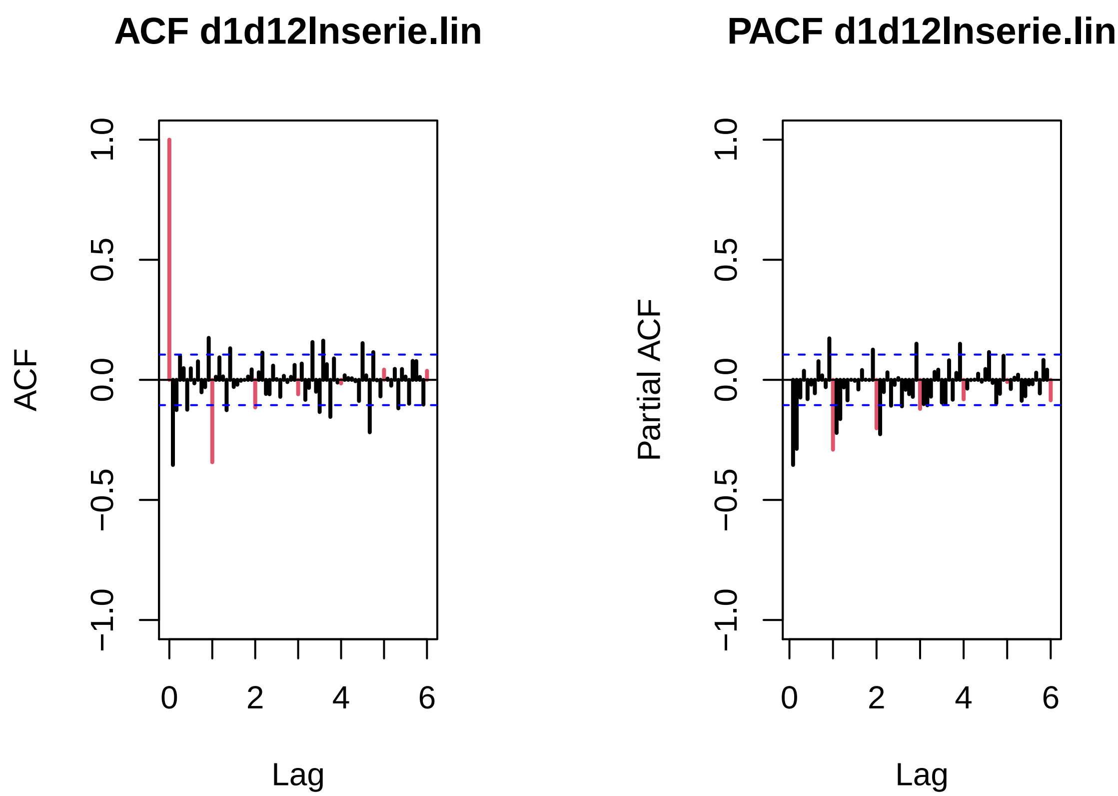
On the table we can observe the outliers, their type and his magnitude. For example: * In Feb 1991 we found a transitory change (TC) type of outlier with a significant statistic's value $|3.059| > 2$. Its magnitud is given by Wcoeff $= -0.0568$ in the log scale (our series was log-transformed), which means that a decrease in the CO2 emissions with respect to what would have happened if this atypical had not taken place.

- The second is an additive outlier (AO) that occurs in May 2004. As learned in theory, its effect is only noticed at that specific date.
- In Mar 2009 a level shift (LS) type of outlier is detected. Its effect takes place from that moment on. Also another one in Feb 2010, Oct 2015 and Jan 2016.

6.3 Once the series has been linearized, free of calendar and outliers' effects, perform forecasting. Compare forecasts results for the original series: classical ARIMA vs ARIMA extension (by using the linearized models).



6.4 Identification of the model

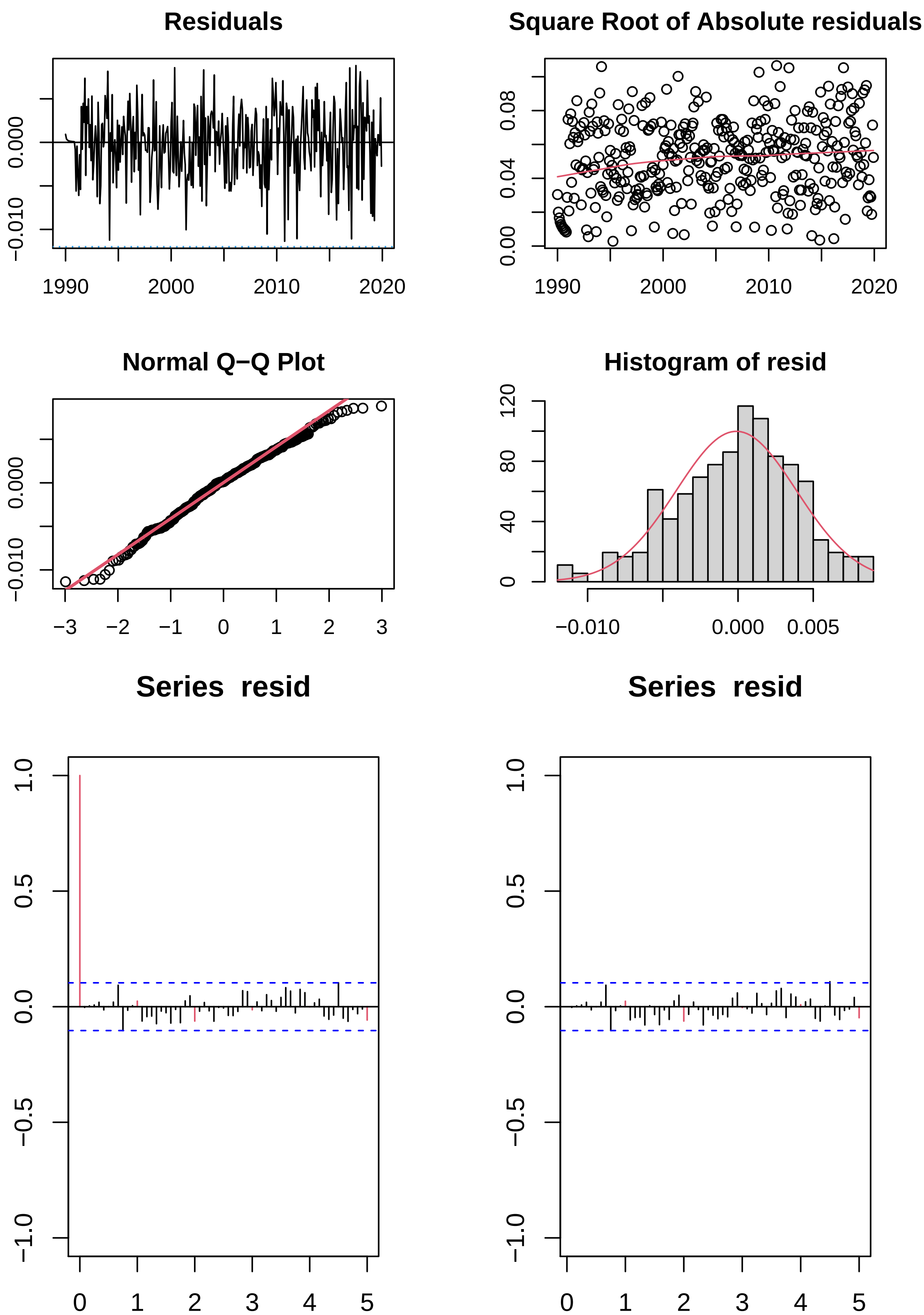


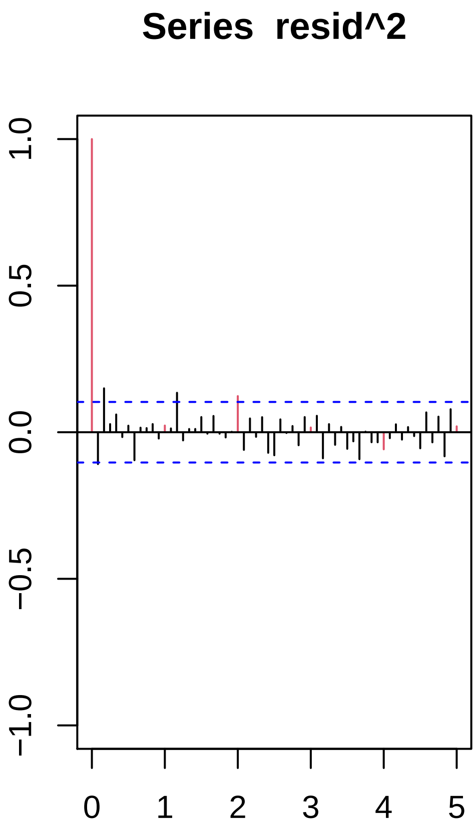
6.5 Estimation of the linearized model

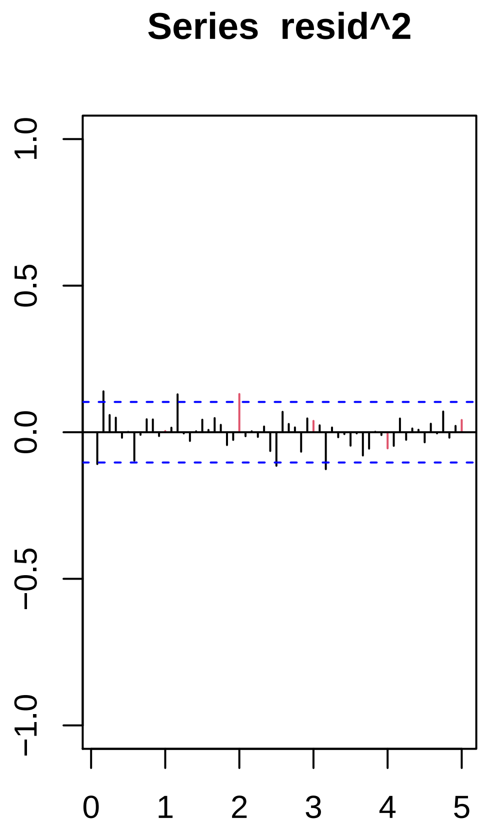
After some search, this is the model with lower AIC we've found: $ARIMA(0,1,5)(0,1,1)_{12}$

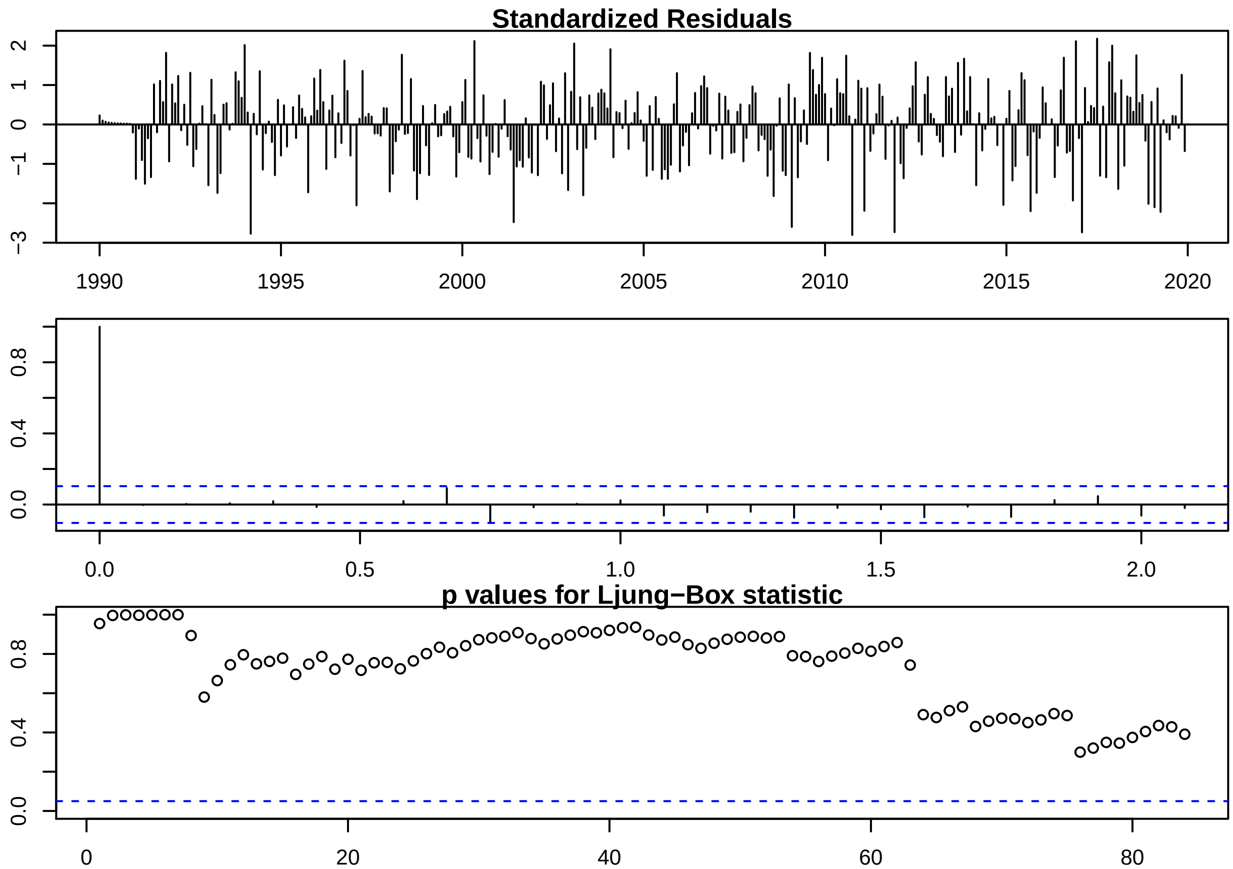
```
##
## Call:
## arima(x = lnserie.lin, order = c(0, 1, 5), seasonal = list(order = c(0, 1, 1),
##     period = 12), xreg = data.frame(wTradDays, from2009))
##
## Coefficients:
##            ma1      ma2     ma3     ma4      ma5     sma1  wTradDays  from2009
##        -0.5456  -0.1165  0.0168  0.0809  -0.1441  -0.8043      4e-04   -0.0164
## s.e.    0.0537   0.0616  0.0617  0.0596   0.0529   0.0370      1e-04    0.0030
##
## sigma^2 estimated as 1.642e-05:  log likelihood = 1412.42,  aic = -2806.83
```

6.6 Validation of the linearized model

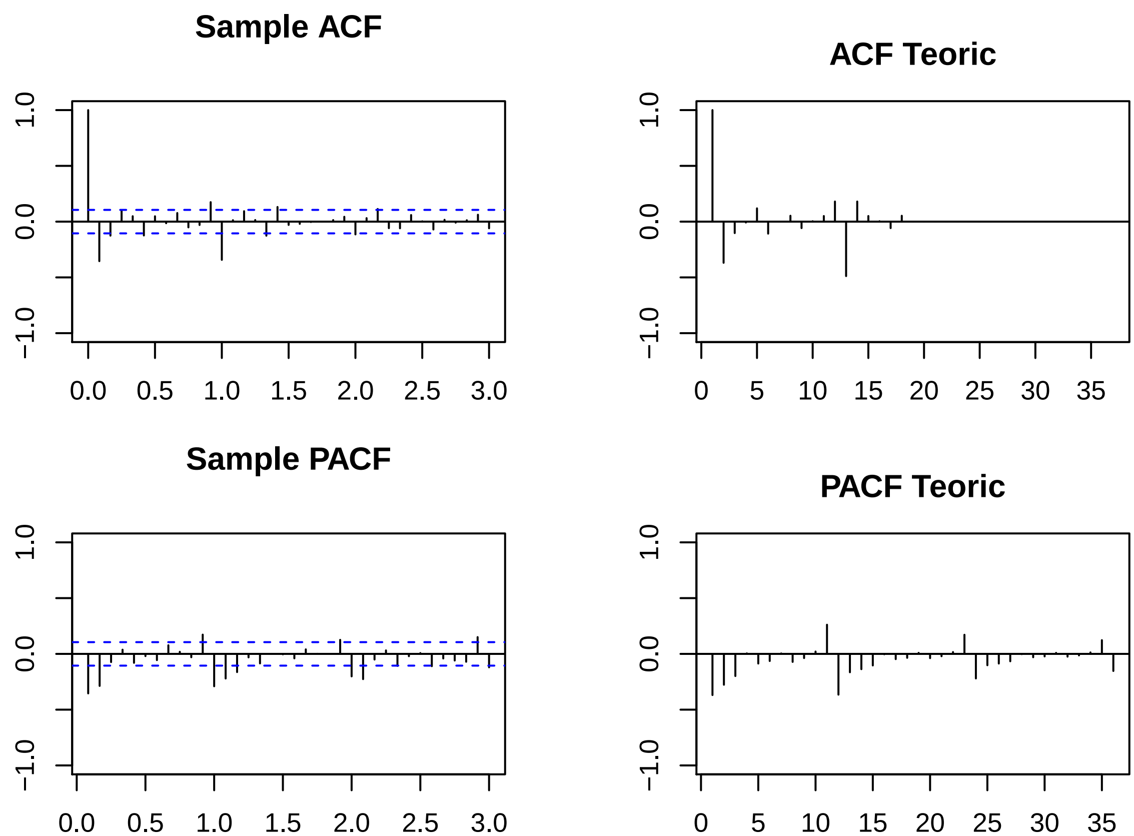



Series resid^2
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid^2
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5




```
##
## --------------------------------------------------------------------
##
## Call:
## arima(x = lnserie.lin, order = c(0, 1, 5), seasonal = list(order = c(0, 1, 1),
##     period = 12), xreg = data.frame(wTradDays, from2009))
##
## Coefficients:
##           ma1      ma2     ma3     ma4      ma5     sma1  wTradDays  from2009
##       -0.5456  -0.1165  0.0168  0.0809  -0.1441  -0.8043      4e-04   -0.0164
## s.e.   0.0537   0.0616  0.0617  0.0596   0.0529   0.0370      1e-04    0.0030
##
## sigma^2 estimated as 1.642e-05:  log likelihood = 1412.42,  aic = -2806.83
##
## Ljung-Box test
##      lag.df    statistic   p.value
## [1,]      1  0.003219395 0.9547525
## [2,]      2  0.007603176 0.9962056
## [3,]      3  0.026815268 0.9988415
## [4,]      4  0.159678554 0.9969775
## [5,]     12  7.860769788 0.7959139
## [6,]     24 19.512729018 0.7241216
## [7,]     36 26.490535847 0.8764746
## [8,]     48 37.798664090 0.8545827
```



Again we have to verify the following hypotesis:

1. Homogeneity of variance, for which the residuals, the square root of absolute values of the residuals with smooth fit and the ACF and PACF of square residuals are plotted.

2. Normality, for which the Quantile-Quantile and the histogram with theoretical density overlapped are plotted.

3. Independence, for which the ACF and PACF of residuals are plotted and LJung-Box test is run.

In the validation of this last model, it seems that all the hypotesis are accomplish except for maybe the normality of the residuals (check residuals q-qplot and histogram).

6.7 Forecasting linearized serie

```
##
## Call:
## arima(x = lnserie1.lin, order = pdq, seasonal = seas, xreg = reg)
##
## Coefficients:
##           ma1      ma2     ma3     ma4      ma5     sma1  wTradDays  from2009
##       -0.5456  -0.1165  0.0168  0.0809  -0.1441  -0.8043      4e-04   -0.0164
## s.e.   0.0537   0.0616  0.0617  0.0596   0.0529   0.0370      1e-04    0.0030
##
## sigma^2 estimated as 1.642e-05:  log likelihood = 1412.42,  aic = -2806.83
```

Fitted the model to the subset series (without 2018 data): lnserie2

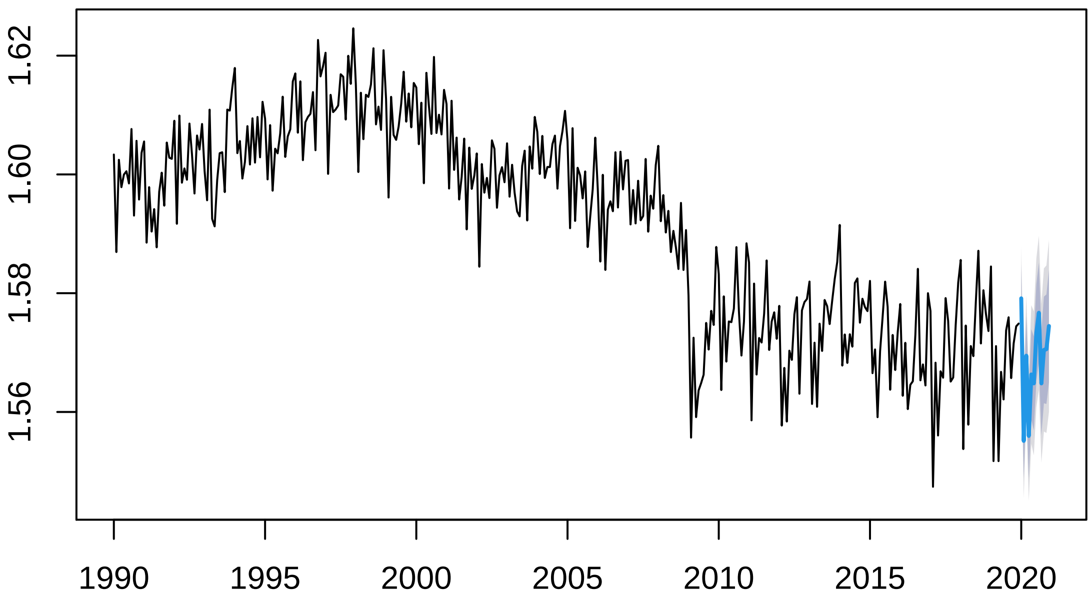
```
##
## Call:
## arima(x = lnserie2.lin, order = pdq, seasonal = seas, xreg = reg2)
##
## Coefficients:
##           ma1      ma2     ma3     ma4      ma5     sma1  wTradDays2  from20092
##       -0.5176  -0.1624  0.0286  0.0875  -0.1359  -0.8252       4e-04    -0.0164
## s.e.   0.0552   0.0627  0.0610  0.0596   0.0551   0.0373       1e-04     0.0031
##
## sigma^2 estimated as 1.628e-05:  log likelihood = 1364.2,  aic = -2710.4
```

The model is stable it accomplish the three hypotesis in significance, sign and magnitude.

6.7.1 Predictions

```
##          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## Jan 2020       1.579131 1.573391 1.584871 1.570352 1.587909
## Feb 2020       1.555131 1.548828 1.561433 1.545492 1.564770
## Mar 2020       1.569477 1.562830 1.576124 1.559311 1.579643
## Apr 2020       1.555973 1.548823 1.563122 1.545039 1.566907
## May 2020       1.566317 1.558690 1.573944 1.554652 1.577982
## Jun 2020       1.564817 1.556901 1.572732 1.552711 1.576923
## Jul 2020       1.573319 1.565125 1.581513 1.560788 1.585851
## Aug 2020       1.576703 1.568240 1.585166 1.563760 1.589646
## Sep 2020       1.564814 1.556090 1.573538 1.551472 1.578156
## Oct 2020       1.570443 1.561466 1.579420 1.556714 1.584172
## Nov 2020       1.570573 1.561350 1.579797 1.556467 1.584679
## Dec 2020       1.574472 1.565009 1.583935 1.559999 1.588945
```

Forecasts from ARIMA(0,1,5)(0,1,1)[12]



- Accuracy measurements

```
##                         ME        RMSE         MAE          MPE      MAPE
## Training set -0.0001094045 0.004407751 0.003452217 -0.007384824 0.2172253
##                   MASE         ACF1
## Training set 0.4413399 -0.007398264
```

- Table ARIMA vs ARIMA extension

```
##                              par      Sigma2Z       AIC       BIC     RMSE
## ARIMA(0,1,1)(0,1,1)_12         2 2.656083e-05 -2649.802 -2638.254 0.005069
## ARIMA(0,1,5)(0,1,1)_12+Atip   19 1.642223e-05 -2784.832 -2707.846 0.004408
##                                   MAE       MPE     MAPE
## ARIMA(0,1,1)(0,1,1)_12      0.003954 -0.005964 0.249339
## ARIMA(0,1,5)(0,1,1)_12+Atip 0.003452 -0.007385 0.217225
```

As expected the final model without outliers and with calendar effects have a better performance in the predictions. Also having less values for AIC and BIC, for this reason for forecast this serie we will choose the second model $ARIMA(0,1,5)(0,1,1)_{12} + Atip$.

To sum up, is important take into account all the steps that we perform in this project. First of all make the serie stationary, then identify the model, predict values and check out the outliers and calendar effects that the serie could have in order to obtain better predictions.

7 References

Josep A. Sanchez and Lesly Acosta, Time serie [class notes], 2021