Time Series Assignment 2

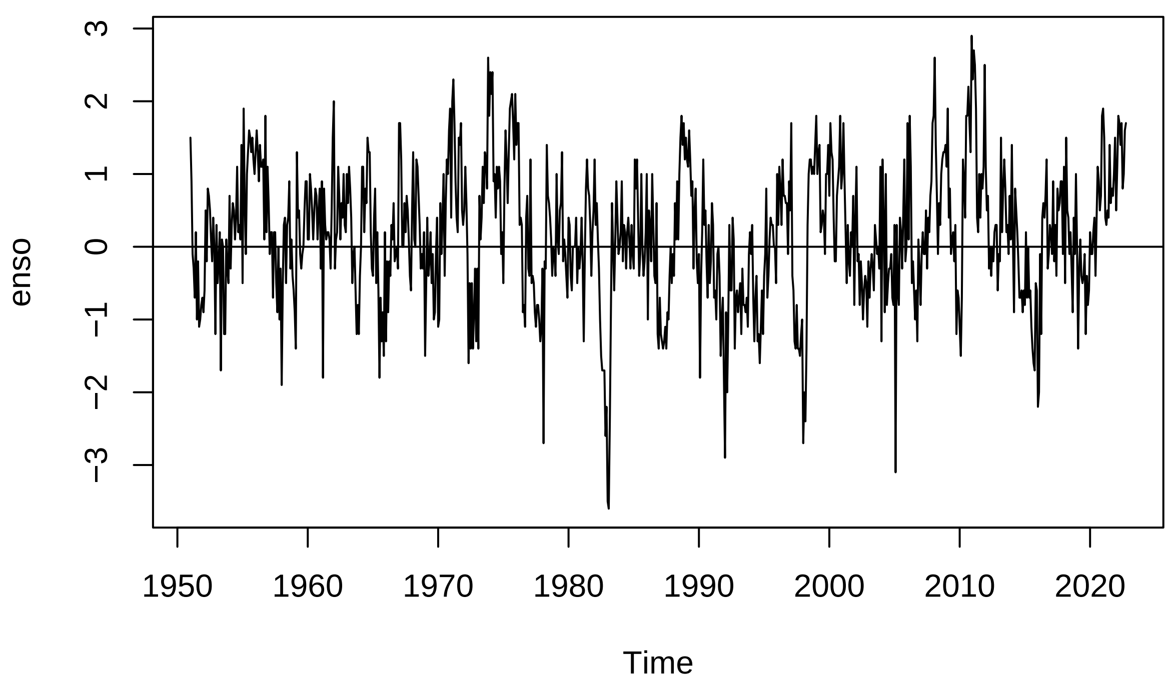
Kane Kesler (200934242)

2023-04-25

Section 1 answers

1. Plotting time series

```r
# plot enso data
enso <- ENSO
plot(enso)
abline(0,0)
```



From visual inspection we see a random oscillation around 0 with a fairly stable variance. In addition isn't any signs of trends nor seasonality which indicates the series is weakly stationary with mean 0. To test for weak-staroinarity, we use the ADF and KPSS tests.

2. ADF and KPSS tests

```
# test for weakly-stationarity
adf.test(enso)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  enso
## Dickey-Fuller = -7.2308, Lag order = 9, p-value = 0.01
## alternative hypothesis: stationary
```

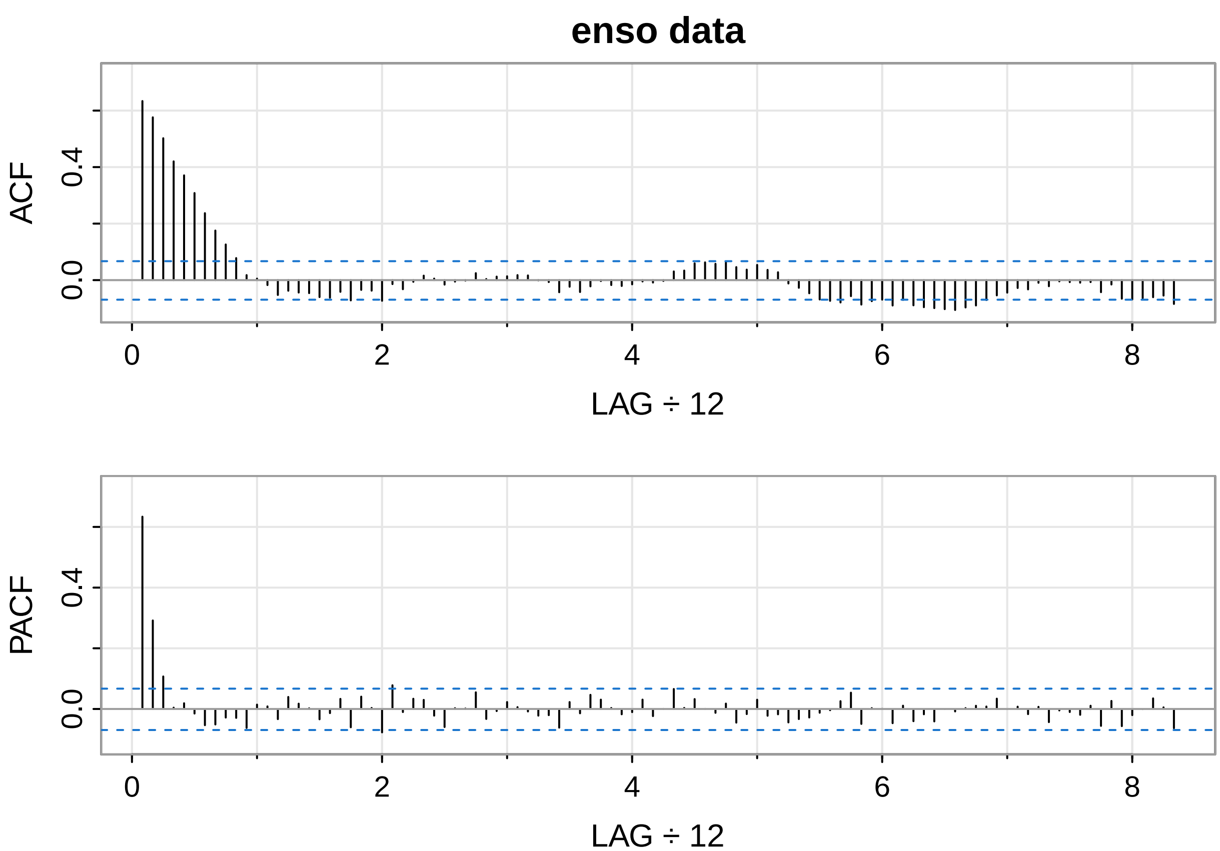
```
unitroot_kpss(enso)
```

```
##   kpss_stat kpss_pvalue
##   0.1714988   0.1000000
```

We see that p-value < 0.05 and kpss-pvalue > 0.05 and hence we conclude that the plot is weakly stationary significance level $\alpha = 0.05$. Now we plot the ACF and PACF to decide on the type of model we need.

3 & 4. Plot ACF and PACF and determine the best model

```
acf2(enso, max.lag=100, main="enso data");
```



We see an exponential decay in the ACF plot and a lag cut-off at lag spike 3, where all three spikes are significant at significance level $\alpha = 0.05$. This indicates a non-seasonal AR(3) model is a suitable choice. Our general structure would be

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)X_t = Z_t,\ Z_t \sim N(0, \sigma^2)$$

where the parameters ϕ_1, ϕ_2, ϕ_3 and σ are to be determined.

5. Estimate your chosen model using maximum likelihood

Since we have seasonality of 12 months, we use the 12-lag seasonal differencing operator to remove seasonality.

```
m <- sarima(enso, 3, 0, 0)
```

```
print(m$ttable)
```

```
##       Estimate     SE t.value p.value
## ar1     0.4181 0.0339 12.3411  0.0000
## ar2     0.2471 0.0358  6.9068  0.0000
## ar3     0.1067 0.0339  3.1421  0.0017
## xmean   0.1654 0.1023  1.6168  0.1063
```

```
print(m$fit)
```

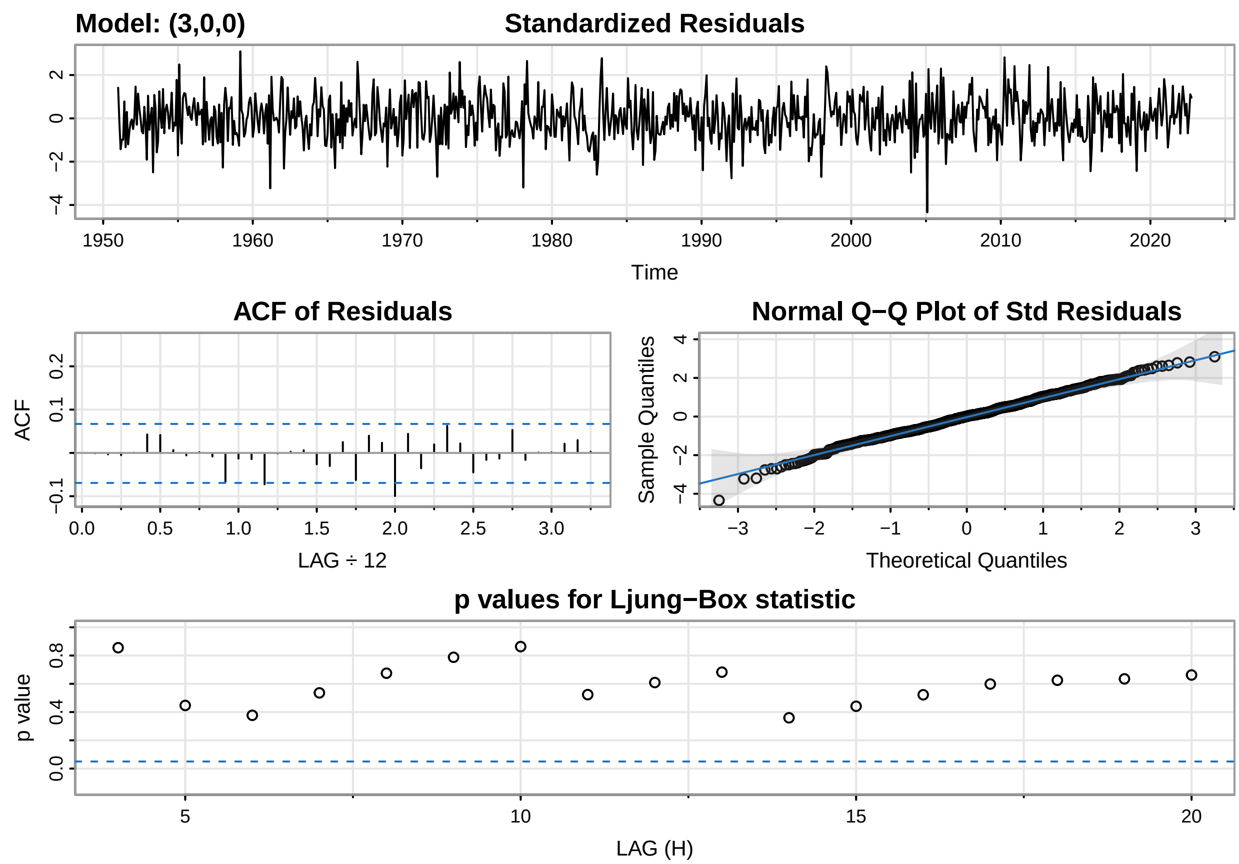
```
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     ar2     ar3   xmean
##       0.4181  0.2471  0.1067  0.1654
## s.e.  0.0339  0.0358  0.0339  0.1023
##
## sigma^2 estimated as 0.4748:  log likelihood = -902.45,  aic = 1814.9
```

The MLE estimates of the AR(3) parameters are $\phi_1 = 0.4181, \phi_2 = 0.2471, \phi_3 = 0.1067$ and $\sigma^2 = 0.4748$, respectively. Each these estimates are significant at level $\alpha = 0.05$, with the exception of the variance, which is not specified in the output. Now that we have our model, we can now perform some diagnostics on the residuals.

6. Perform model diagnostics on the standardised residuals to assess goodness of fit

From model m, we retrieve the following residual plots.

```
m <- sarima(enso, 3, 0, 0)
```



The plot of standardized residuals seems to have random oscillations of constant variation around zero. The ACF plot shows no significant peaks expect at integer lag 2, meaning that there is a slight correlation between every second residual but overall there is little correlation between residuals. Besides the ends of the QQ plot, there is little deviation from the normal line, which gives enough evidence to assume they come from a normal distribution. Finally, the plot of residual p-values are all above the 0.05 significance line so we don't reject the null hypothesis of independence of each lag. To get the p-value for the lag-wise aggregated version for all lags we perform the Ljung-Box Q test:

```
Box.test(resid(m$fit), lag=20, type="Ljung-Box", fitdf=3)
```

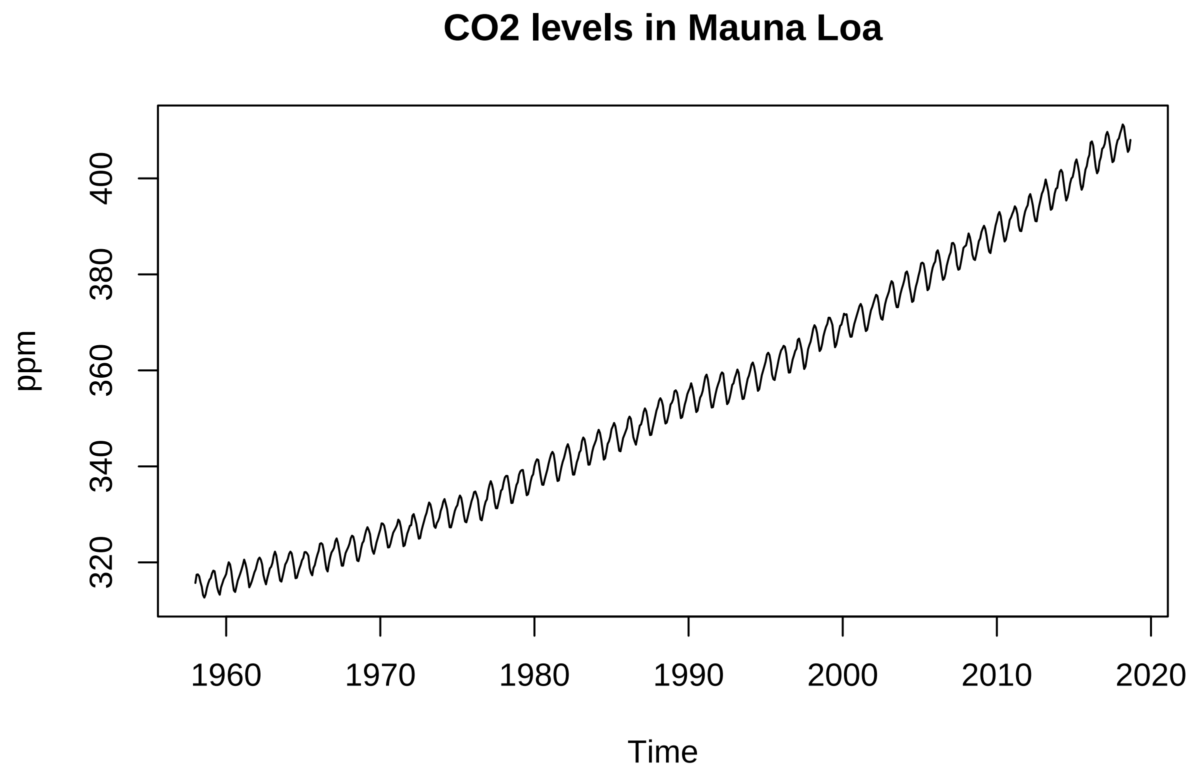
```
##
##  Box-Ljung test
##
## data:  resid(m$fit)
## X-squared = 14.067, df = 17, p-value = 0.6623
```

The p-value is greater than 0.05, so we don't reject the null hypothesis of independence. From this we can assume the residuals come from an i.i.d Normal distribution.

Section 2 answers

1. Plotting time series

```
# plot cardox data
cx <- ts(cardox, start = 1958, frequency = 12)
plot(cx, main="CO2 levels in Mauna Loa", ylab="ppm")
```



The plot indicates an non-linear increasing trend, with seasonality of 12 months, particularly higher in the summer in the winter. Due to the many data points on the plot, determining whether or not the variance is stable is not clear. To ensure variance stability, we apply the Box-Cox transformation:

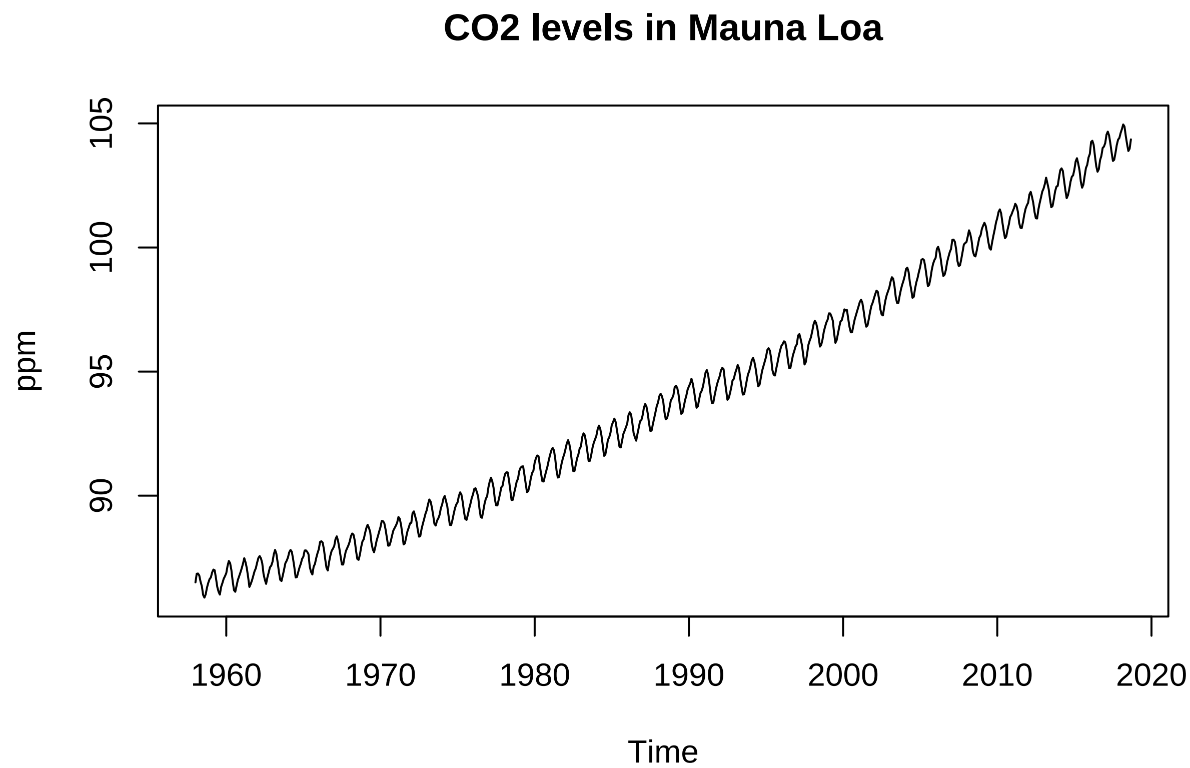
```
lambda <- BoxCox.lambda(cx)
print(lambda)
```

```
## [1] 0.7209449
```

```
cxbox <- BoxCox(cx, lambda)
```

Now we plot the variance-stabalised series:

```
plot(cxbox, main="CO2 levels in Mauna Loa", ylab="ppm")
```

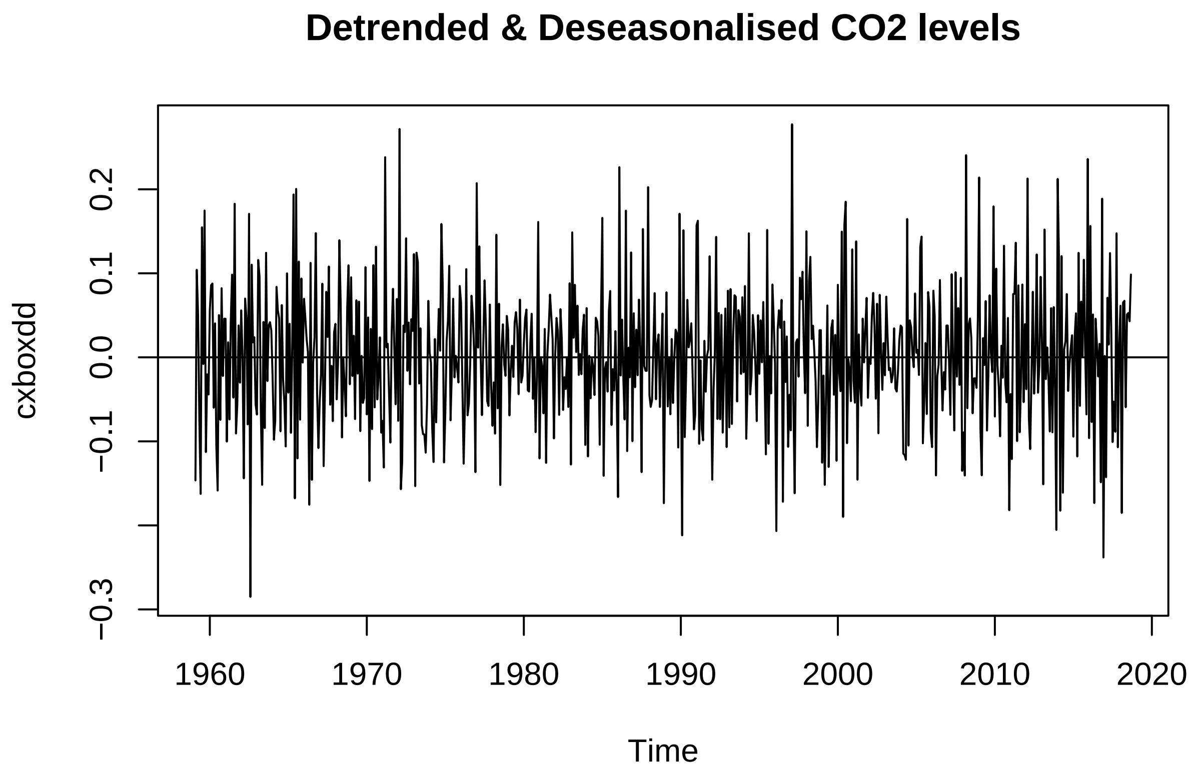


There does seem to be a shift in the position of the data points in the series, making the trend slightly more linear, which shows that the transformation was indeed necessary.

Detrend and Deseasonalise

```
cxboxd <- diff(cxbox)
cxboxdd <- diff(cxboxd, lag=12)

plot(cxboxdd, main="Detrended & Deseasonalised CO2 levels")
abline(0,0)
```



After using the composite operator $\nabla_{12}\nabla$ we get a seemingly random oscillation around 0 with no signs of seasonality or trend. From this, it seems to be weakly-stationary with mean 0 but, we need to perform a few tests to verify this assumption.

```
# test for weakly-stationarity
adf.test(cxboxdd)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  cxboxdd
## Dickey-Fuller = -8.763, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

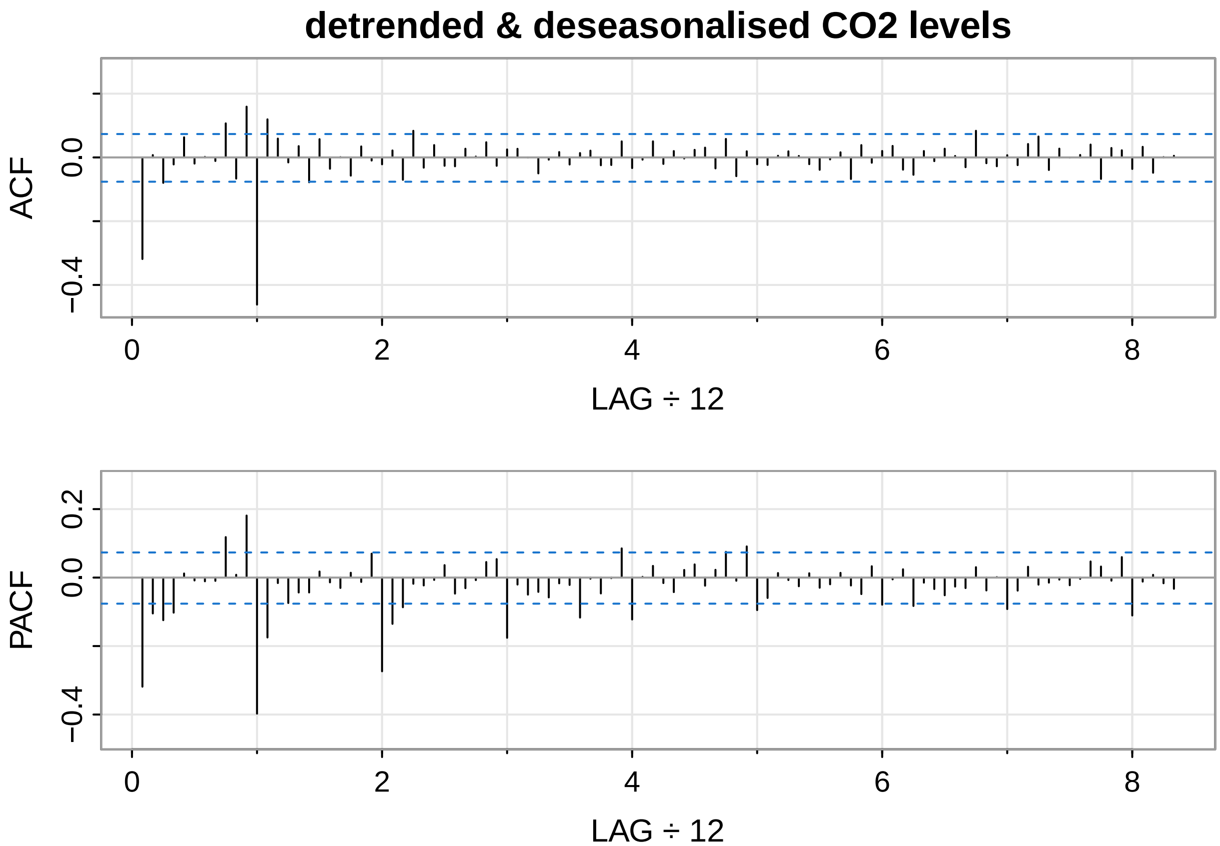
```
unitroot_kpss(cxboxdd)
```

```
##   kpss_stat kpss_pvalue
## 0.009408498 0.100000000
```

We see that p-value < 0.05 and kpss-pvalue > 0.05 and hence we conclude that the plot is weakly stationary significance level $\alpha = 0.05$. Now we plot the ACF and PACF to decide on the type of model we need.

3 & 4. Plot ACF and PACF and determine the best model

```
acf2(cxboxdd, max.lag=100, main="detrended & deseasonalised CO2 levels")
```

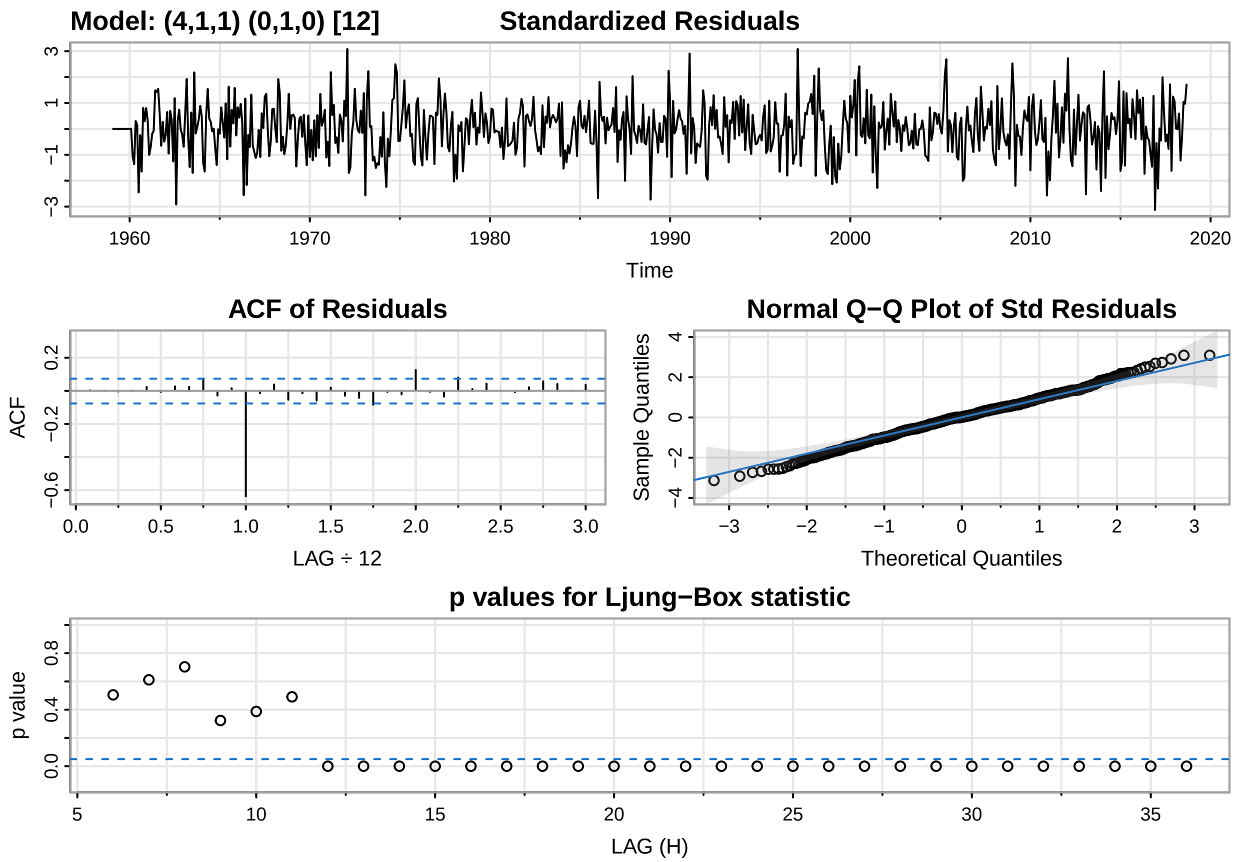


We notice the cut off after lag $1/12$ and $4/12$ on the ACF and PACF plots, respectively with neither plots showing any clear exponential decay. This suggests a non-seasonal MA(1) and AR(1) are present in the series. We have a seasonal period of $s = 12$ months, in addition we detrended with one differencing operator and deseasonalised one period, so we have $d = 1$ and $D = 1$. Hence our initial guess of the model is of the form ARIMA$(4, 1, 1) \times (0, 1, 0)_{12}$. The issue with this model, however, is clear when we look at the residual diagnostics.

```
m_initial <- sarima(cxboxdd, p=4, d=1, q=1, P=0, D=1, Q=0, S=12)
```

```
## initial  value -1.448155
## iter   2 value -1.769405
## iter   3 value -1.828679
## iter   4 value -1.939715
## iter   5 value -1.968020
## iter   6 value -1.974422
## iter   7 value -1.977734
## iter   8 value -1.979223
## iter   9 value -1.980612
## iter  10 value -1.981325
## iter  11 value -1.981367
## iter  12 value -1.981453
## iter  13 value -1.981469
## iter  14 value -1.981471
## iter  15 value -1.981471
## iter  16 value -1.981471
## iter  16 value -1.981471
## iter  16 value -1.981471
```

```
## final  value -1.981471
## converged
## initial  value -1.987561
## iter   2 value -1.993549
## iter   3 value -1.994044
## iter   4 value -1.995191
## iter   5 value -1.995251
## iter   6 value -1.995257
## iter   7 value -1.995257
## iter   7 value -1.995257
## iter   7 value -1.995257
## final  value -1.995257
## converged
```



We see that sometime after lag 10, the p-values are less than the dotted 0.05 significance line, meaning that we have to reject the independence hypothesis after lag 10. The ACF plot indicates a high correlation between neighboring points, with a significant peak at lag 1 and 2 indicating a need for a seasonal MA(2) component. Upon re-inspection of the CO2 ACF plot we see an cut-off at integer peak 1 and exponential decay of integer peaks in the PACF plot, suggesting at least a seasonal MA(1). Combining all of these observations this model needs an additional seasonal MA(2). From this, we have two possible models. We use the model mentioned before with a slight modification, $\text{ARIMA}(4,1,1) \times (0,1,2)_{12}$, or we use a model that counts 1/12 and 9/12 as significant peaks and the rest as white noise in the ACF plot and counts 1/12, 2/12, 3/12, 4/12, 9/12 as significant peaks and the rest as white noise in the PACF plot. With that we get the model second possible model $\text{ARIMA}(9,1,9) \times (0,1,2)_{12}$.

Now we calculate the MLE estimates of the parameters of both models.

```
ma1 <- sarima(cxboxdd, p=4, d=1, q=1, P=0, D=1, Q=2, S=12)
```

```
ma2 <- sarima(cxboxdd, p=9, d=1, q=9, P=0, D=1, Q=2, S=12)
```

```
ma1$fit
```

```
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     include.mean = !no.constant, transform.pars = trans, fixed = fixed, optim.control = list(trace =
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ma1     sma1    sma2
##       -0.3877  -0.1796  -0.1501  -0.1109  -1.0000  -1.9014  0.9085
## s.e.   0.0378   0.0401   0.0400   0.0378   0.0148   0.0262  0.0262
##
## sigma^2 estimated as 0.003551:  log likelihood = 926.23,  aic = -1836.46
```

```
ma2$fit
```

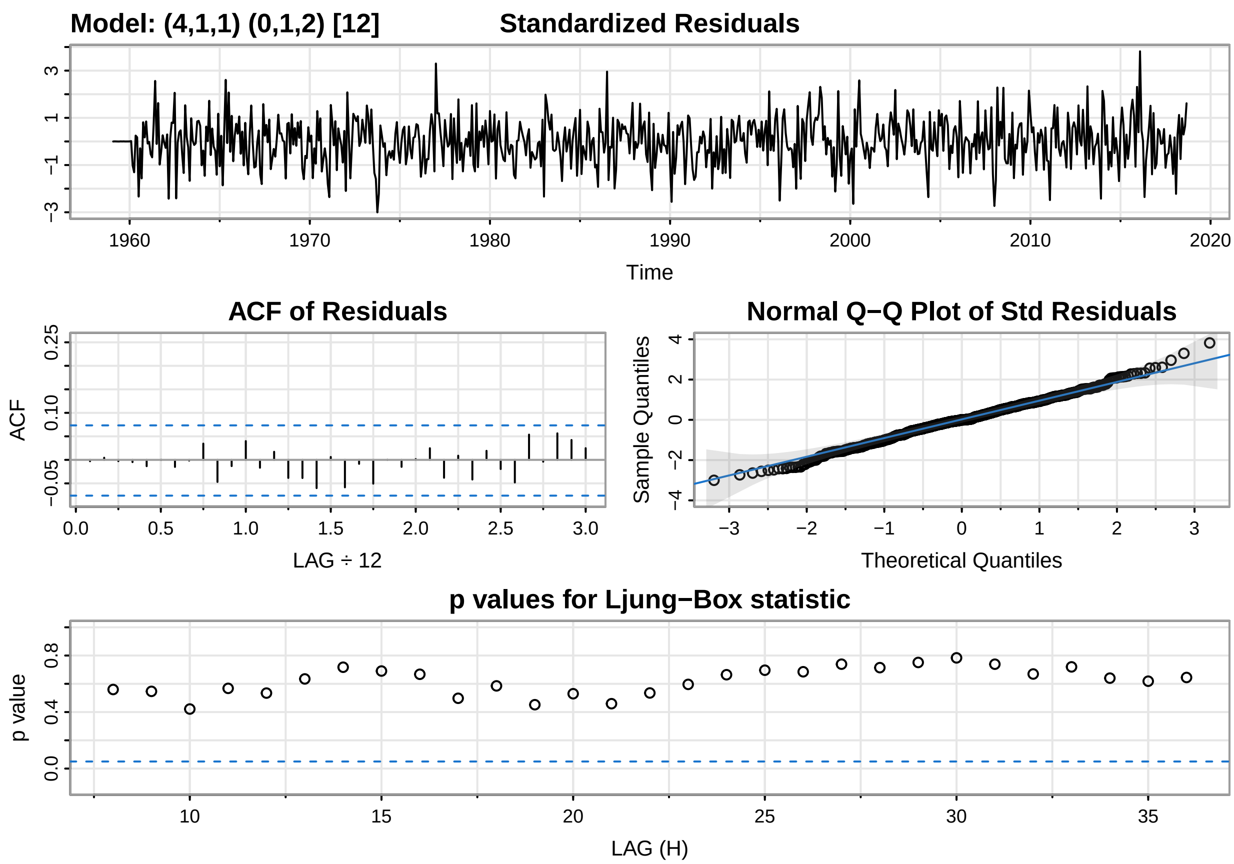
```
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     include.mean = !no.constant, transform.pars = trans, fixed = fixed, optim.control = list(trace =
##         REPORT = 1, reltol = tol))
##
## Coefficients:
```

```
## Warning in sqrt(diag(x$var.coef)): NaNs produced
```

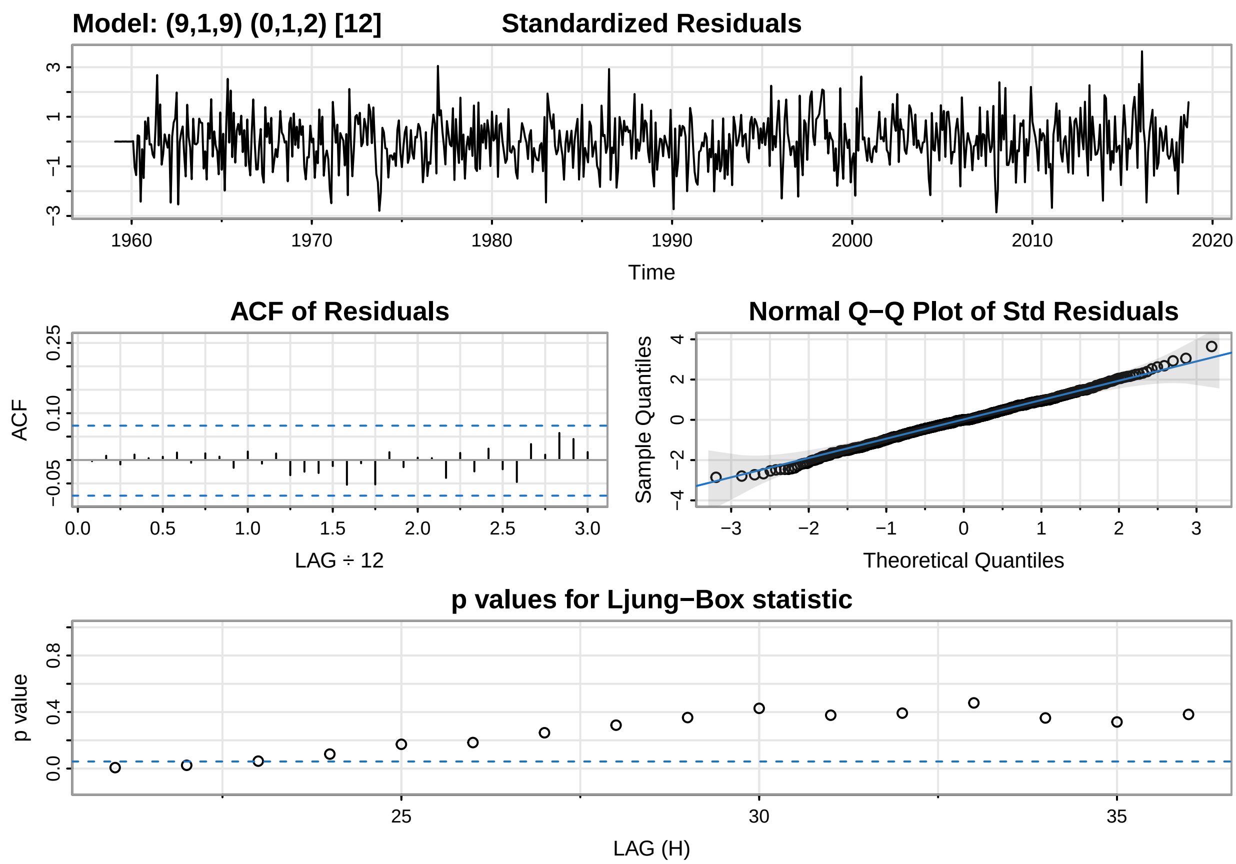
```
##           ar1      ar2     ar3      ar4     ar5     ar6    ar7      ar8     ar9
##       -0.4161  -0.0282  -0.199  -0.1625  0.2791  0.4201  0.688  -0.0364  0.0709
## s.e.      NaN      NaN     NaN   0.0587     NaN     NaN    NaN      NaN  0.0351
##           ma1      ma2     ma3      ma4      ma5     ma6      ma7     ma8
##       -0.9612  -0.2060  0.3011  -0.0967  -0.3622  0.0028  -0.2561  0.9319
## s.e.      NaN   0.1492     NaN      NaN   0.0824     NaN   0.1135     NaN
##           ma9     sma1    sma2
##       -0.3517  -1.9032  0.9055
## s.e.      NaN      NaN  0.0116
##
## sigma^2 estimated as 0.003401:  log likelihood = 932.28,  aic = -1822.55
```

For the sake of brevity, we will not mention the parameters estimates explicity as they are shown in the output above.

```
ma1 <- sarima(cxboxdd, p=4, d=1, q=1, P=0, D=1, Q=2, S=12)
```




```
ma2 <- sarima(cxboxdd, p=9, d=1, q=9, P=0, D=1, Q=2, S=12)
```



Upon inspection, the residuals from both models have seemingly random oscillations around 0, hence a mean of 0. The QQ plot shows no significant deviation from normality with the exception of the end points, so we can assume normality. In addition the ACF plots don't have any significant peak at significance level $\alpha = 0.05$. Where both differ from each other is in the Ljung-Box plot, the ARIMA$(4, 1, 1) \times (0, 1, 2)_{12}$ model has p-values consistently above the 0.05 significance lines hence we can assume independence of the residuals. This can not be said for the ARIMA$(9, 1, 9) \times (0, 1, 2)_{12}$ where a few are below the line, and so the independence assumption is to be rejected. This is reflected in the aggregated Ljung-Box test:

```
Box.test(resid(ma1$fit), lag=20, type="Ljung-Box", fitdf=4+1+2)
```

```
##
##  Box-Ljung test
##
## data:  resid(ma1$fit)
## X-squared = 11.974, df = 13, p-value = 0.5297
```

```
Box.test(resid(ma2$fit), lag=20, type="Ljung-Box", fitdf=9+9+2)
```

```
##
##  Box-Ljung test
##
## data:  resid(ma2$fit)
## X-squared = 5.3055, df = 0, p-value < 2.2e-16
```

We can also postulate that since the ARIMA$(9, 1, 9) \times (0, 1, 2)_{12}$ is more parameterised than the simpler ARIMA$(4, 1, 1) \times (0, 1, 2)_{12}$ model, we would risk over-fitting and so the simpler model should be picked. We can verify this with the information criteria:

5. Use the information criteria to decide between these models.

```
ma1$BIC
```

```
## [1] -2.560476
```

```
ma2$BIC
```

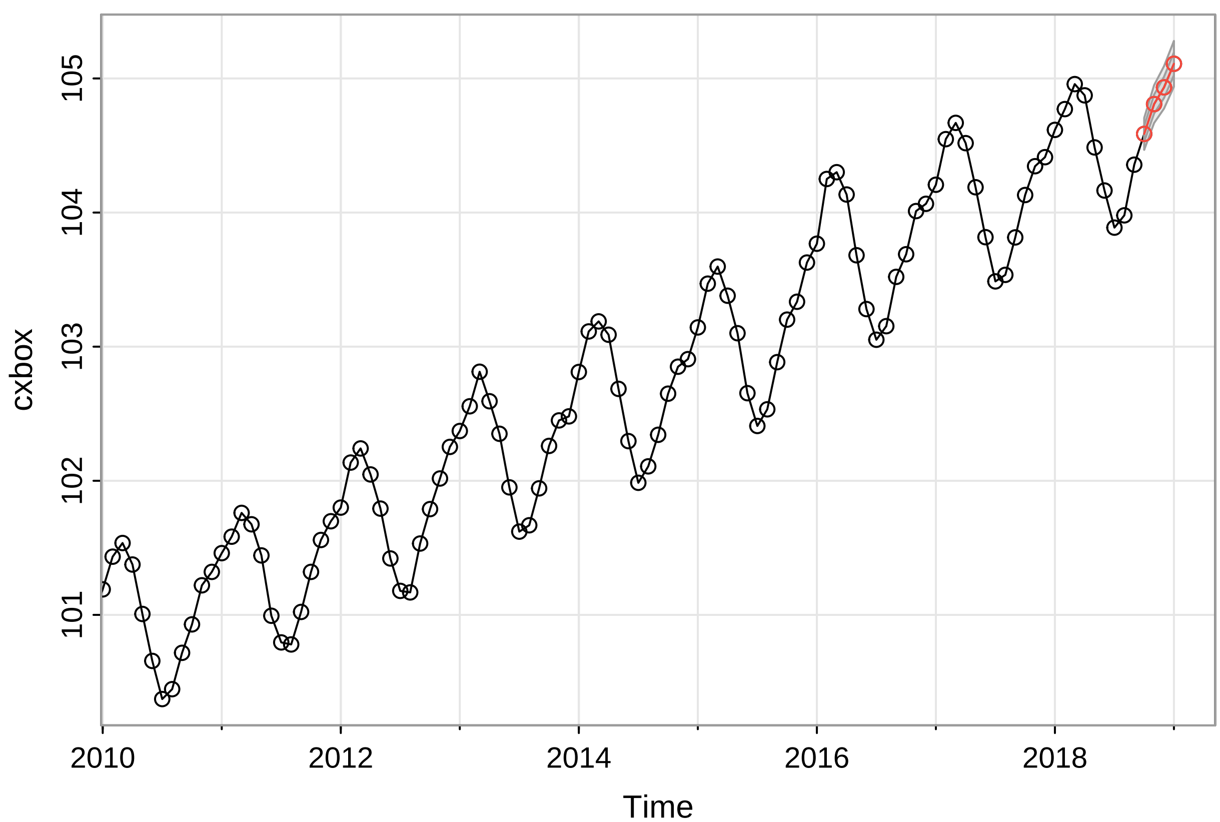
```
## [1] -2.456459
```

The model we pick should have the smaller BIC, in this case we pick the simpler model. (We used BIC since it's used for large models).

6. Forecast the carbon dioxide levels from December 2018 to March 2019.

Now we forecast the CO2 levels 4 months after November using model we choose:

```
sarima.for(cxbox, 4, p=4, d=1, q=1, P=0, D=1, Q=2, S=12)
```




```
## $pred
##           Jan Feb Mar Apr May Jun Jul Aug Sep      Oct      Nov      Dec
## 2018                                          104.5863 104.8087 104.9336
## 2019 105.1094
##
## $se
##             Jan Feb Mar Apr May Jun Jul Aug Sep        Oct        Nov
## 2018                                            0.05962795 0.07049703
## 2019 0.08533174
##             Dec
```

```
## 2018 0.07916906
## 2019
```

In between the January and February predictions, there's a bend that's present in all the actual data shown in the plot as well adhering to the increasing trend (where the CO2 levels in March are at least the same or higher than the levels in May of the year prior). The variance between the predictions don't seem to deviate to the variation already established by the current data. Overall, we can say that the ARIMA$(4, 1, 1) \times (0, 1, 2)_{12}$ is not only a good model for the series $\nabla_{12}\nabla X_t$ but also a good predictor of future of CO2 levels, at least to the extent of 4 months.