Forecasting Time Series

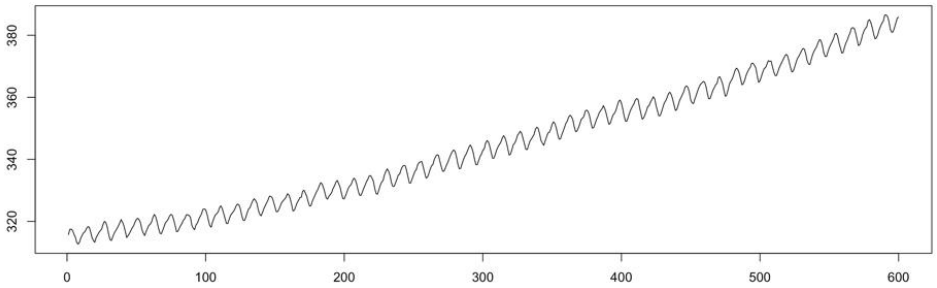
Individual Assignment

ALFONSO VILLEGAS NARVÁEZ
MBD, SECTION 2
Group C

Data Understanding and Model Selection

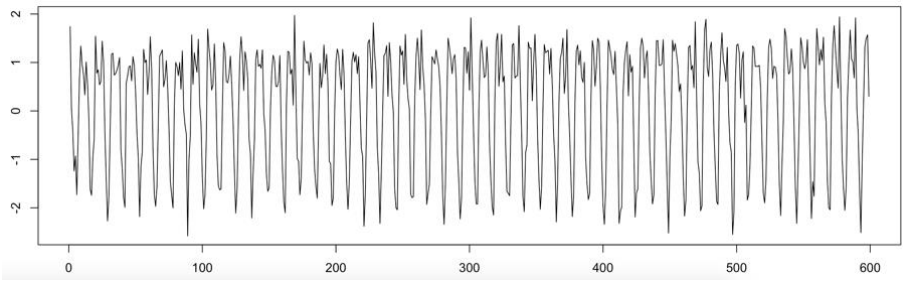
First of all, I will get an overall understanding of the data that I will try to predict. I can see that the data shows monthly mean CO2 values, from 1958 to 2019, with 732 observations.

Plotting the time series, we can see right away that the data is NOT Stationary in the Mean. Most probably Stationary in the Variance

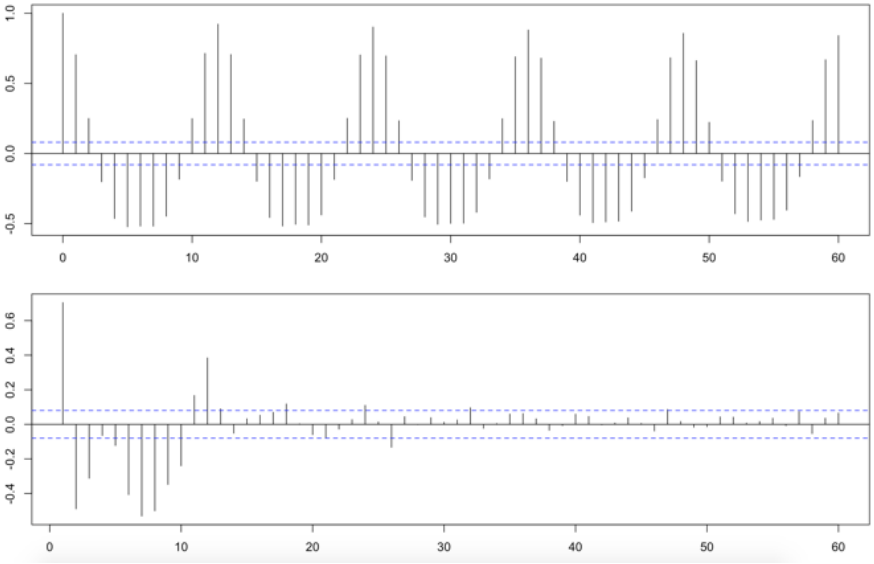


I will choose the **Recursive Scheme**, including as many observations as possible, since there does not seem to be any structural breaks in the data. I will import the first 600 rows to forecast, and use the last 132 to validate my predictions.
We run the Dickey Fuller test for Yt and confirm that we need to do 1 transformation.

After doing the transformation we plot Zt and confirm that it is now Stationary, both with the plot and with the Dickey Fuller test:



We then plot the ACF (top) and PACF (bottom) for Zt to look for linear models to predict Yt:



We can see a very clear seasonality in the ACF, and several significant lags on the PACF (1-3 and 6-12). We will first try to work with the simplest possible model, in the sense that we have to estimate less parameters and use it as a baseline to compare the following more complex models. Due to the found seasonality, we will define S = 12.

First Model

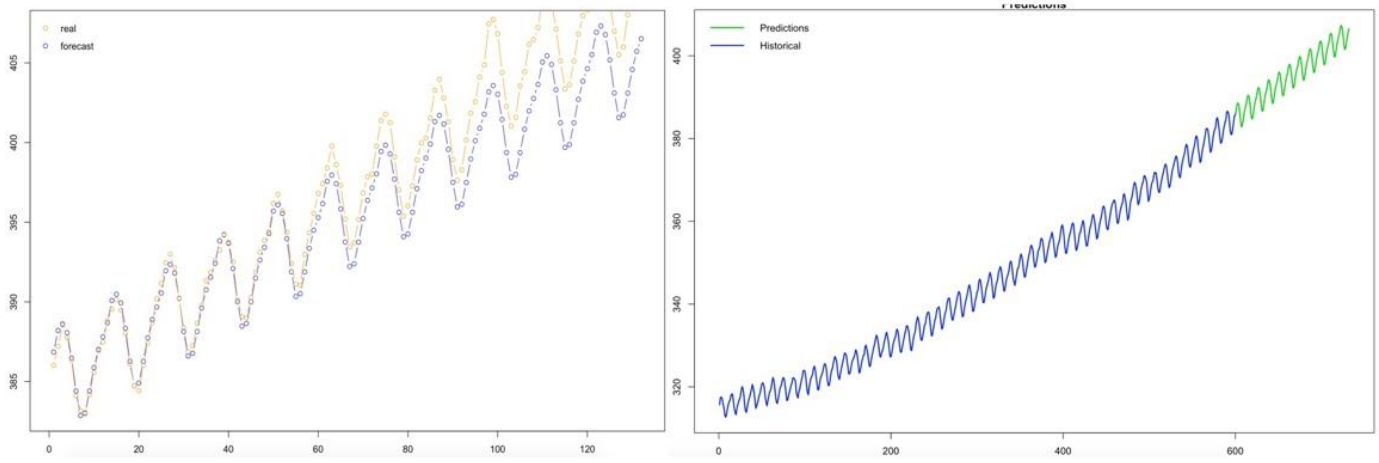
The simplest model that I can think of considering the significant lags, is AR(1), SMA(1). We will therefore do a:
SARIMA (1,1,0) * (0,1,1).

Residuals are White Noise:
I run the model, and first of all plot the ACF and PACF for the residuals to check if they are White Noise. From the perspective of the plots, I can tell that they seem to be WN. I then run the **Ljung-Box Test** and confirm WN, since we got a p-value of 0.08925 and we cannot Reject H0 (uncorrelated data).
We then run the Shapiro Test and plot the Histogram, and confirm that we have Gaussian White Noise.

After that I calculate the Intervals for a 95% confidence and confirm that both our parameters are significant (different from 0):

AR(1) = (-0.3867055, -0.2306895)
SMA(1) = (-0.9191924, -0.8396164)

I am now going to plot predictions compared to the real values:



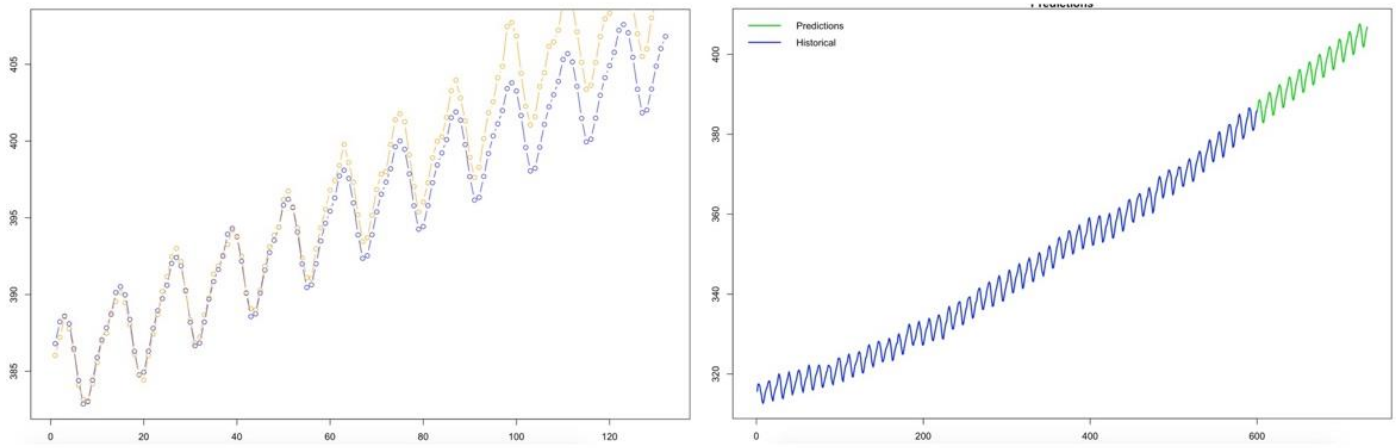
Second Model

After testing a couple of the more complex options, I decided to model a: SARIMA (4,1,0) * (0,1,1).
Residuals are White Noise:
I run the model, and first of all plot the ACF and PACF for the residuals to check if they are White Noise. From the perspective of the plots, I can tell that they seem to be WN. I then run the **Ljung-Box Test** and confirm WN, since we got a p-value of 0.6193 and we cannot Reject H0 (uncorrelated data).
We then run the Shapiro Test and plot the Histogram, and I am unable to confirm with confidence if data is normal or not, which is irrelevant at this point since we will not create a Non-Linear model.

After that I calculate the Intervals for 95% confidence and confirm that all my 5 parameters are significant (different from 0):

AR(1) = (-0.4480283, -0.2833883)
AR(2) = (-0.25106807, -0.07623607)
AR(3) = (-0.19027188, -0.01739988)
AR(4) = (-0.165186307, -0.002898307)
SMA(1) = (-0.9062555, -0.8223675)

I am now going to plot my predictions compared to the real values:



Model Assessment and Selection:

So far I am happy with both model, visually they seem very nice. Let's now assess their forecasting ability using MAPE (Mean Absolut Percentage Error).

MODEL 1, MAPE = 0.42%
MODEL 2, MAPE = 0.3925%

If we take the MAPE metric as an only criteria, then we would select the Model 2, since it's forecasting ability seems to be better. However; we can see that the error for both models is very small and very similar, and considering that Model 1 needs to estimate only 2 parameters, and Model 2 need to estimate 5 parameters, my final choice and recommendation is for Model 1.

This is because it is easier to estimate, and considers only the first normal lag, and the first seasonal lag, which makes a lot of sense: we can capture most of the behavior of Yt using only the previous lag.

Final choice: Model 1: SARIMA (1,1,0) * (0,1,1)