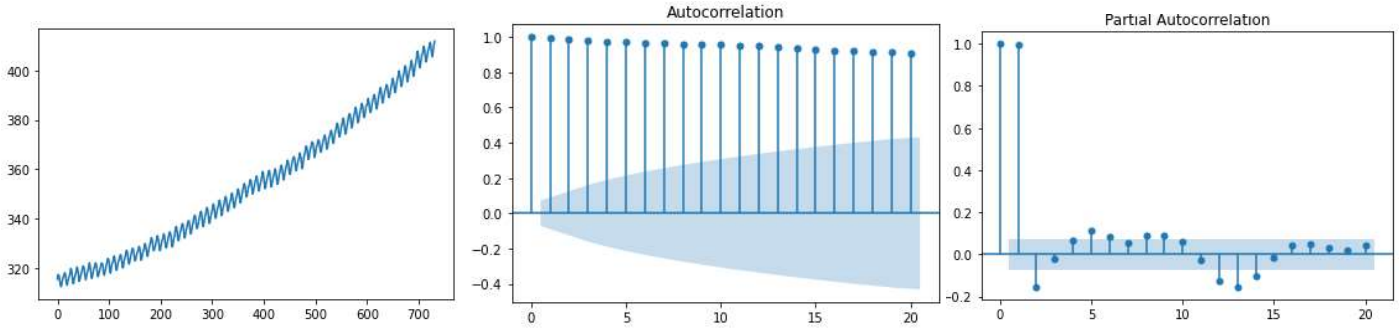
Ramón Denia
deniachavarri@student.ie.edu

1) Find at most two-time series models, using the Box-Jenkins methodology, for the monthly mean CO2 mole fraction at Mauna Loa Observatory, Hawaii, from March 1958 to February 2019. The mole fraction of CO2, expressed as parts per million (ppm), is the number of molecules of CO2 in one million molecules of dried air (water vapor removed)

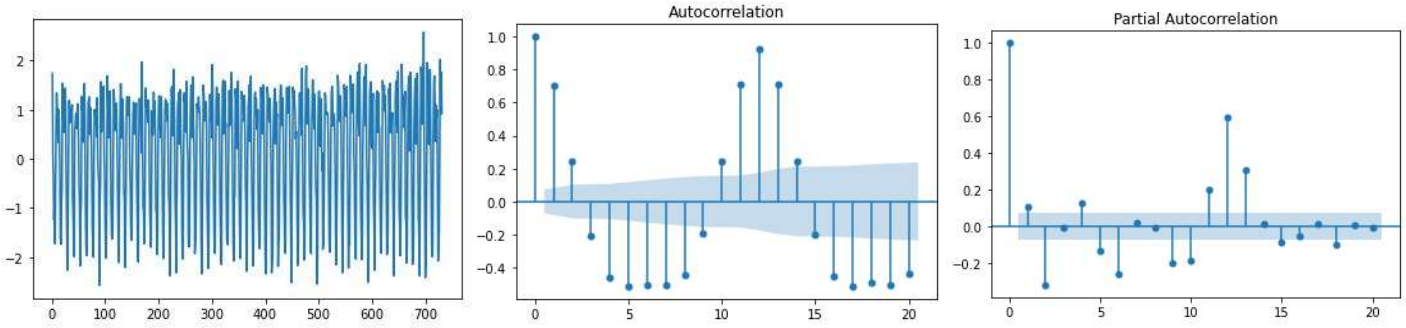
First, we need to see if the data it's stationary or not. We observe that the data it's not stationary and it has a strong seasonal pattern. We can look as well to the ACF to detect the descending trend which means the non-stationary model. Also, we ran as well the ADF test to double-check our data.

```
ADF test for the original series
Statistic Value: 12.039534898542087
p-value: 1.0
```

p_value > 0.05 --> This means that we accept H(0) (not stationary)



We need to transform our data into stationary. Due to this, we need to apply the difference to have our data stationary.




```
ADF test for the original series
Statistic Value: -15.738116094526454
p-value: 1.265496015501439e-28
```

p_value < 0.05 --> This means that we accept H(1) (stationary)

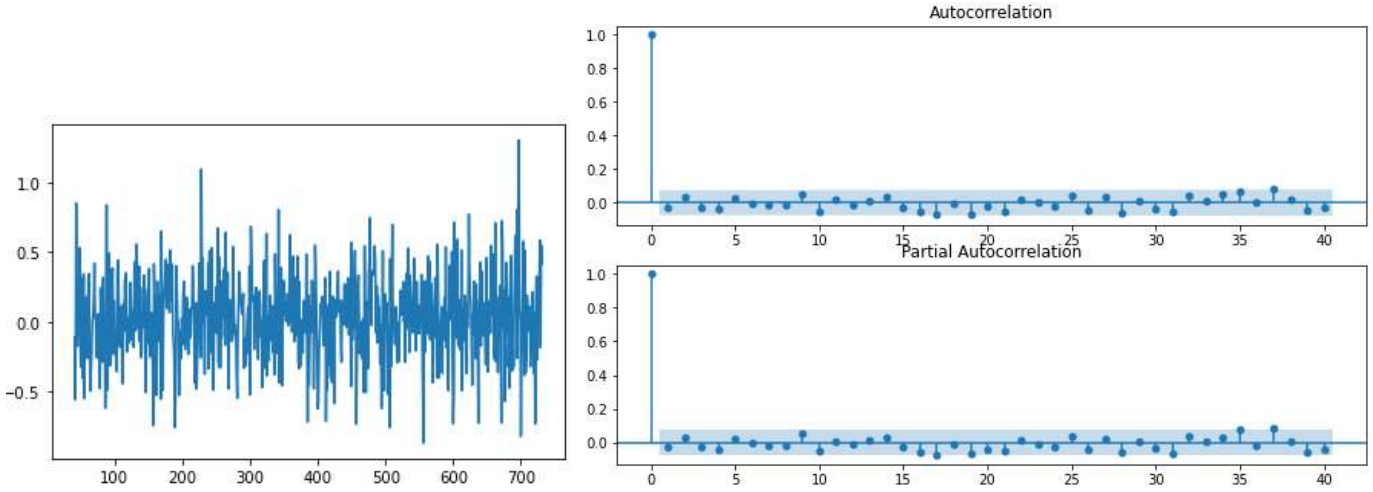
Now we have stationary data (ADF test above). We can advise that we have seasonal data and, we see points out of bounds for the ACF and the PACF. So, we don't have White Noise because the PACF and the ACF are not zero. Now, we want to obtain the residuals of the estimated model and check if they are White Noise or not. Our estimated model is the following:

SARIMAX Results

Dep. Variable:	CO2	No. Observations:	732
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 12)	Log Likelihood	-209.189
Date:	Fri, 26 Nov 2021	AIC	428.377
Time:	20:26:25	BIC	451.350
Sample:	0	HQIC	437.240
	- 732		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2105	0.036	5.793	0.000	0.139	0.282
ma.L1	-0.5582	0.039	-14.262	0.000	-0.635	-0.482
ar.S.L12	0.9996	0.000	3068.944	0.000	0.999	1.000
ma.S.L12	-0.8645	0.021	-40.337	0.000	-0.907	-0.822
sigma2	0.0962	0.005	20.421	0.000	0.087	0.105

Ljung-Box (L1) (Q):	0.26	Jarque-Bera (JB):	4.28
Prob(Q):	0.61	Prob(JB):	0.12
Heteroskedasticity (H):	1.14	Skew:	-0.00
Prob(H) (two-sided):	0.31	Kurtosis:	3.37



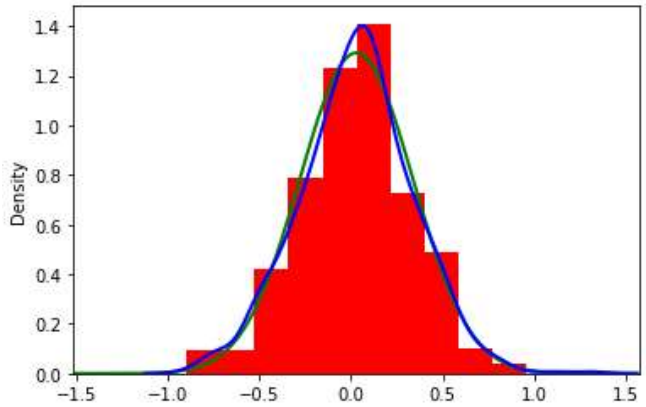
Analyzing our residuals for our estimated model we can see that we have:

```
Mean 0.02901069732336451
The mean is nearly zero.
```

- Zero mean -->
- Constant Variance
- PACF and ACF are equals to zero --> Uncorrelated data.

We have White Noise

Also, we can say as well that we have **Strict White Noise** and **Gaussian White Noise** because the variables are independent and identically distributed with a normal distribution:




```
Saphiro ShapiroResult(statistic=0.9959223866462708, pvalue=0.06892590969800949)

It is NORMALLY distributed
```

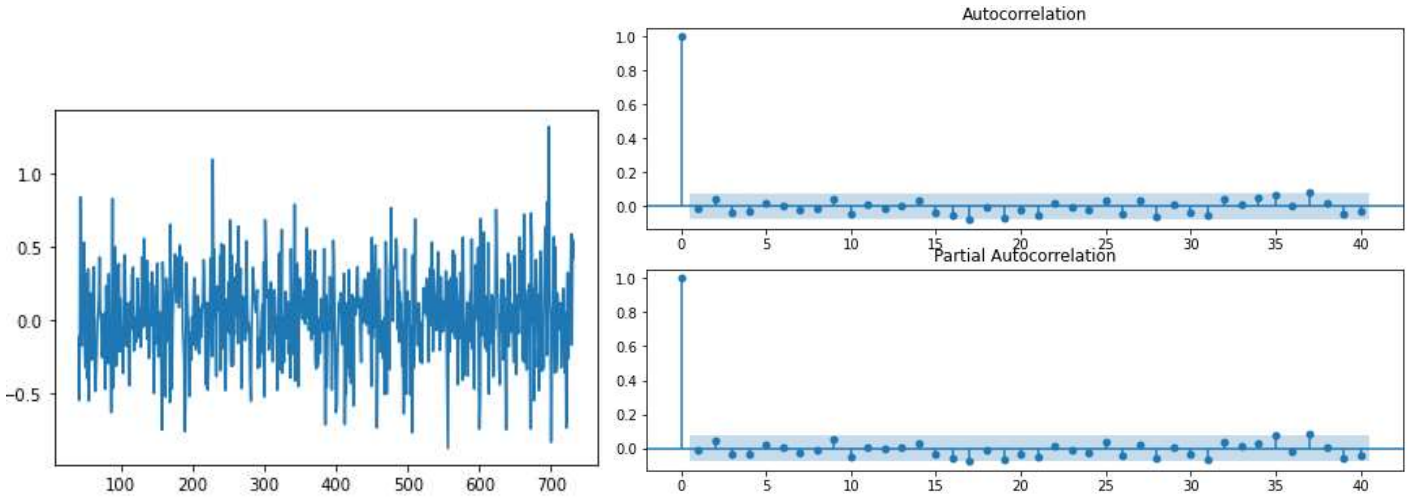
We have estimated another model that follows the previous one. The second model is the following:

SARIMAX Results

Dep. Variable:	CO2	No. Observations:	732
Model:	SARIMAX(2, 1, 2)x(1, 0, [1], 12)	Log Likelihood	-208.874
Date:	Fri, 26 Nov 2021	AIC	431.748
Time:	20:46:50	BIC	463.909
Sample:	0	HQIC	444.155
	- 732		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7704	0.098	-7.864	0.000	-0.962	-0.578
ar.L2	0.1814	0.082	2.202	0.028	0.020	0.343
ma.L1	0.4062	0.089	4.558	0.000	0.232	0.581
ma.L2	-0.5451	0.081	-6.738	0.000	-0.704	-0.387
ar.S.L12	0.9997	0.000	3878.420	0.000	0.999	1.000
ma.S.L12	-0.8638	0.021	-41.045	0.000	-0.905	-0.823
sigma2	0.0939	0.005	20.589	0.000	0.085	0.103

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	4.77
Prob(Q):	0.93	Prob(JB):	0.09
Heteroskedasticity (H):	1.15	Skew:	-0.01
Prob(H) (two-sided):	0.27	Kurtosis:	3.39

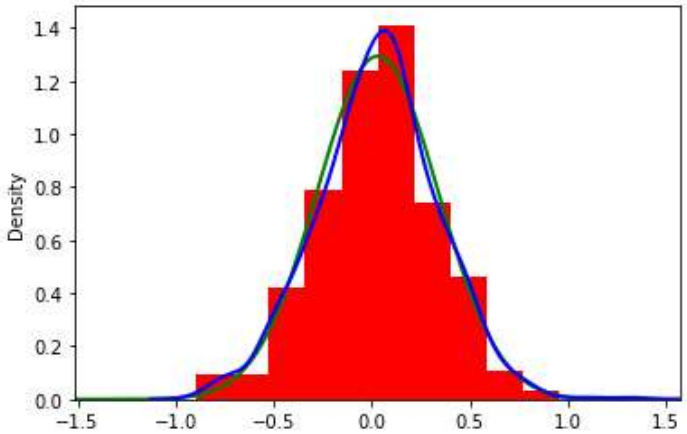


Analyzing our residuals for our second estimated model we can see that we have:

```
Mean 0.02965096778019756
The mean is nearly zero.
```

- Zero mean -->
- Constant Variance
- PACF and ACF are equals to zero --> Uncorrelated data.

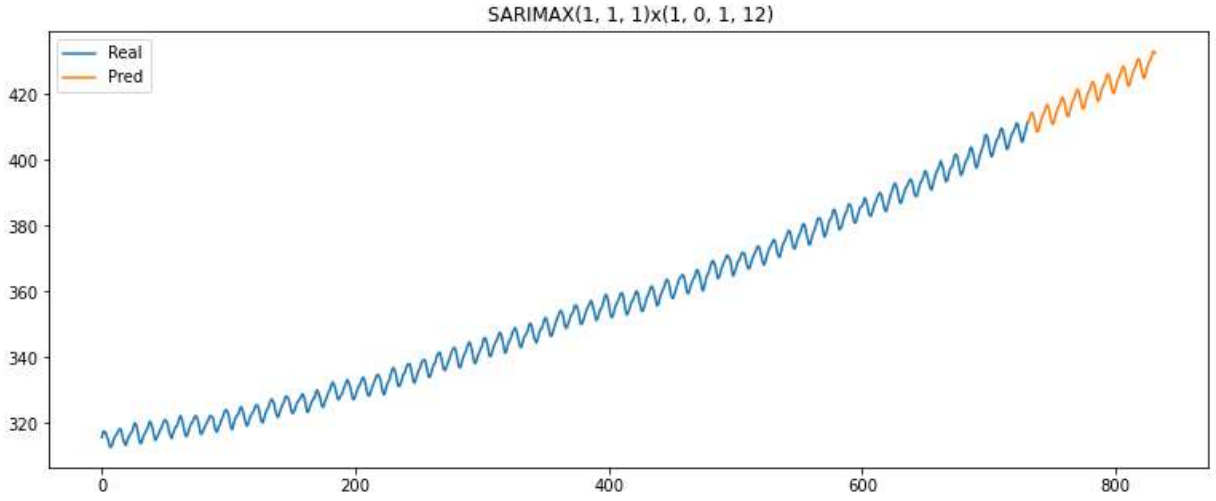
We have White Noise




```
Saphiro ShapiroResult(statistic=0.9959092140197754, pvalue=0.0678914487361908)


It is NORMALLY distributed
```

Now, we can calculate our point predictions and the confidence intervals in a recursive way for the first model I have shown. Let's see how our prediction is:



The confidence intervals are:

```
95% confidence intervals using quantiles
Lower limit 2.5%: -0.5679334368064133 Upper limit 97.5%: 0.6218635057421409
95% confidence intervals assuming normality
Normal Percentile 2.5%: -0.605252480125199 Normal Percentile 97.5%: 0.605252480125199
```

The first 10 predicted values are:

index	predicted_mean
731	411.750000
732	412.480752
733	413.908047
734	414.556242
735	413.820130
736	412.072267
737	410.024952
738	408.568414
739	408.872521
740	410.472388

The best model to be studied is the SARIMA (1,1,1)x(1,0,1,12) because the variance for this model is less than the other. We take into consideration that both are White Noise, SWN and GWN so we compare both with the variance:

VARIANCE FOR SARIMA (1,1,1)x(1,0,1,12) --> `5.288053521219361` --> **BEST MODEL**

VARIANCE FOR SARIMA (2,1,2)x(1,0,1,12) --> `5.2948440062680735`

Ramón Denia
deniachavarri@student.ie.edu

2) For the model, or models, identified in the previous step, leave the last 120 real values to run the usual out-of-sample forecasting exercise. If you only found one model, which is the forecasting performance of your model? In case you found two models, which one is the best model and why is this your final choice?

What we want it's to compare how our model behaves and we need to compare the last 120 values of our real data with the prediction of our model for 120 predicted values. Having 2 models, we want to compare the MAPE and the MSFE. The best model will be the one with smallest MAPE or MSFE.

The last 120 real values from our model are:

```
611    387.48
612    388.82
613    389.55
614    390.14
615    389.48
        ...
727    406.00
728    408.02
729    409.07
730    410.83
731    411.75
```

SARIMA(1,1,1)x(1,0,1,12) has the following errors: `MSFE [[7.39329594]  MAPE % [[0.53962166]`

SARIMA(2,1,2)x(1,0,1,12) has the following errors: `MSFE [[7.16838933] MAPE % [[0.53060696]`

We can see that in this case, the best model for our out of sample forecast is the **SARIMA(2,1,2)x(1,0,1,12)** because it has the lowest values for both errors.

The prediction then that better fits with the real values is:

```
611    387.480000
612    388.285954
613    389.713491
614    390.207465
615    389.650609
          ...
727    401.375121
728    402.702305
729    404.169518
730    405.339110
731    406.090204
```