Modeling and Forecasting of Carbon Dioxide Emission from Monthly Mean Data

KARISHMA CHAND

Roll no - DST-19/20-014

Motivation

Climate change is one of the significant challenges of the 21st century. The increasing levels of carbon dioxide in the atmosphere traps additional heat, which causes global warming and the calamities associated with it, such as extreme weather, forest fires, temperature rise, heavy precipitation, acid rain, ice caps melting, sea-level rise and ocean acidification. Too much carbon dioxide gas in the air causes air pollution and smog, which can impact human health. The carbon dioxide gas can have severe consequences on the wildlife and plants if it crosses the certain threshold value. Scientists have been measuring the carbon dioxide levels in parts per million since 1958 at the carbon dioxide monitoring station Mauna Loa, Hawaii.

The dataset downloaded from NOAA site contains the monthly mean carbon dioxide as a mole fraction. As the carbon dioxide level is measured with respect to time, univariate time series analysis can be applied to this dataset to fit a model. Then, the model can be used to forecast the carbon dioxide level for future time points. By predicting the carbon dioxide level for the future, we can get an idea about the severity of this issue. The data of the past 20 years will be used to build the model.

Data Structure

Variable - Monthly mean CO2 in parts per million collected at Mauna Loa, Hawaii

Start date - Jan 1990

End date - Jun 2020

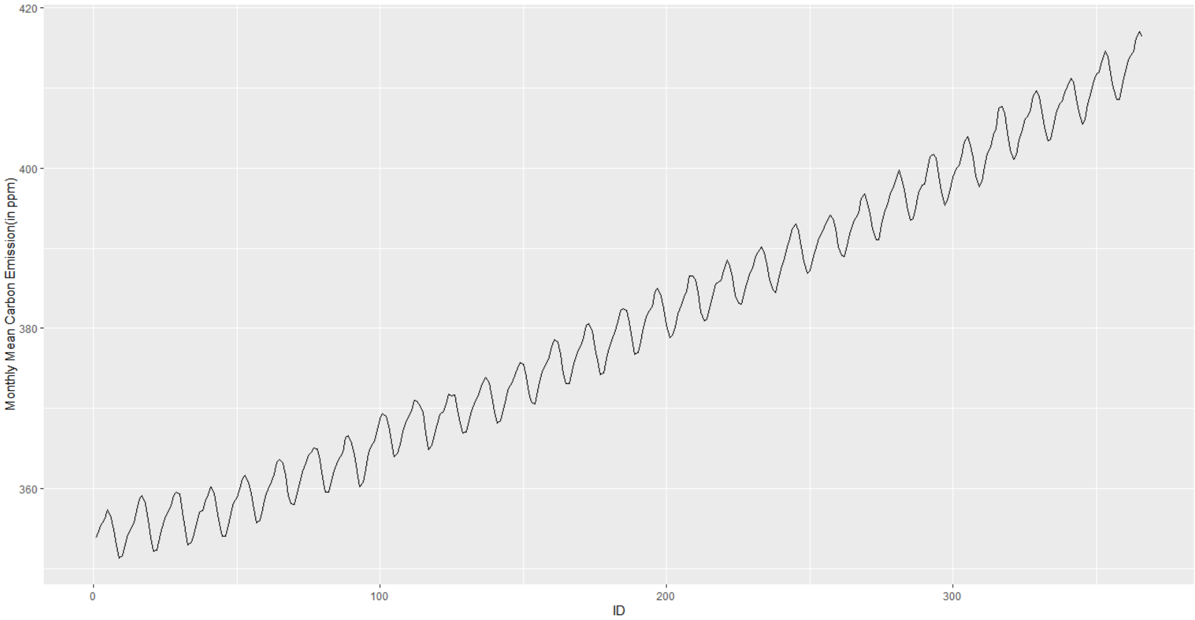
Frequency - Monthly data

Observations - 366

Source - https://www.esrl.noaa.gov/gmd/ccgg/trends/gl_data.html

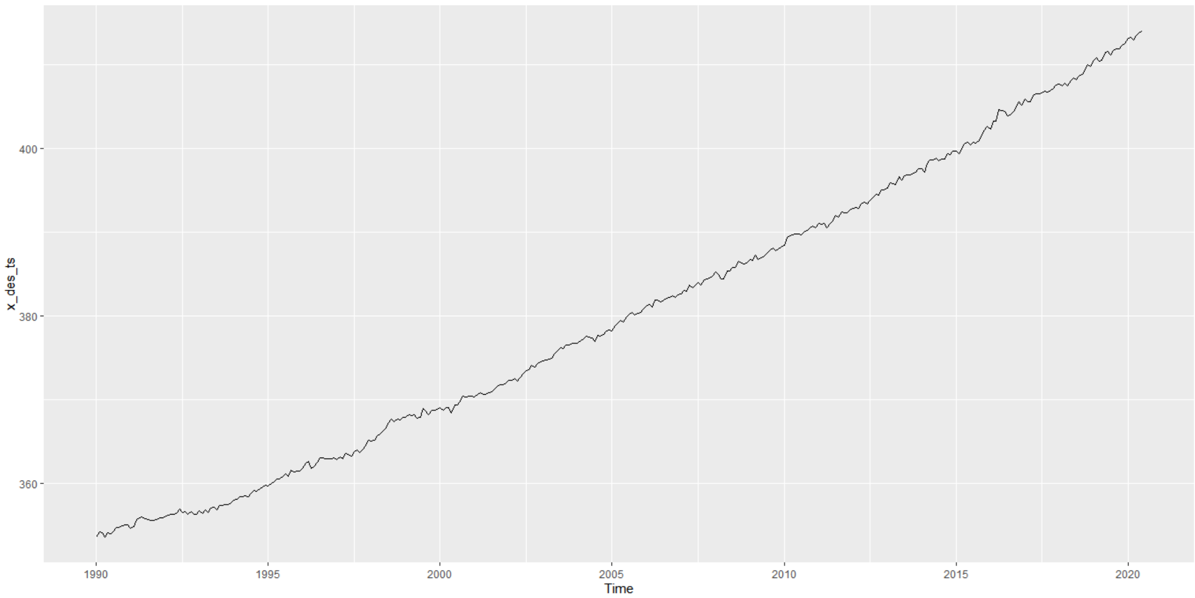
Step 1

By analyzing the plot and the data, we see that the mean monthly carbon dioxide emission is maximum in the month of May for most of the years. Hence, it can be concluded that the data may contain seasonality. By looking at the plot, we also see an upward trend, which indicates that the time series data may have a trend.



Step 2

Collect the deseasonalized time series.



Step 3

Perform the ADF test on the deseasonalized time series.

As the observed test statistic of the ADF test (-2.025) is greater than the critical value at 5% (-3.42), we can not reject the null hypothesis, and the series contains a unit root. So, we will apply differencing to create a new time series $y_t = x_t - x_{t-1}$, where $\{x_t\}$ is the deseasonalized time series. The observed test statistic for the ADF test of y_t is -18.091, which is less than the critical value at 5% (-3.42). Hence, the differenced series y_t does not contain unit root.

```
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.87504 -0.20110  0.00968  0.20216  1.16693 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.159348   3.963173   2.059   0.0402 *  
z.lag.1     -0.022956   0.011335  -2.025   0.0436 *  
tt           0.004258   0.001895   2.247   0.0252 *  
z.diff.lag  -0.279576   0.050249  -5.564 5.16e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3137 on 360 degrees of freedom
Multiple R-squared:  0.1059,	Adjusted R-squared:  0.09846 
F-statistic: 14.21 on 3 and 360 DF,  p-value: 8.95e-09


Value of test-statistic is: -2.0253 47.4723 5.8604 

Critical values for test statistics: 
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

```
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.94891 -0.20903  0.00763  0.19992  1.25673 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.1565967  0.0340200   4.603 5.78e-06 ***
z.lag.1     -1.5098029  0.0834548 -18.091  < 2e-16 ***
tt           0.0005045  0.0001583   3.187  0.00156 ** 
z.diff.lag   0.1696746  0.0518593   3.272  0.00117 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3113 on 359 degrees of freedom
Multiple R-squared:  0.6556,	Adjusted R-squared:  0.6528 
F-statistic: 227.8 on 3 and 359 DF,  p-value: < 2.2e-16


Value of test-statistic is: -18.0913 109.0991 163.6486 

Critical values for test statistics: 
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Step 4

To check the deterministic trend of y_t, lm function is used. As the p-value for the slope is not statistically significant for 5% significant level, we can not reject the null hypothesis. Thus, the differenced time series does not contain any trend. Hence, the differenced time series y_t is a stationary series.

```
Call:
lm(formula = y_ts ~ time(y_ts))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90815 -0.21558  0.00546  0.20732  1.19575 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -7.464386   3.939330  -1.895   0.0589 .
time(y_ts)   0.003805   0.001964   1.937   0.0535 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3295 on 363 degrees of freedom
Multiple R-squared:  0.01023,	Adjusted R-squared:  0.007503 
F-statistic: 3.752 on 1 and 363 DF,  p-value: 0.05353
```

Step 5

Calculate the sample ACFs and prepare the corresponding table by applying Ljung Box test. Based on the conclusion, we can say that the sample ACF is diminishing. So, the model follows either AR or ARMA process.

```
    k   sample_acf   Q_stat      p_value conclusion
1   1 -0.276372473 28.10911 1.146648e-07     reject
2   2 -0.059699428 29.42431 4.079354e-07     reject
3   3 -0.034922238 29.87560 1.465758e-06     reject
4   4 -0.019656601 30.01898 4.851093e-06     reject
5   5  0.034290364 30.45650 1.199038e-05     reject
6   6  0.043153778 31.15137 2.371687e-05     reject
7   7  0.029656399 31.48045 5.068565e-05     reject
8   8  0.042386849 32.15460 8.738585e-05     reject
9   9 -0.024790865 32.38585 1.707426e-04     reject
10 10 -0.096454851 35.89643 8.769607e-05     reject
11 11  0.073231623 37.92577 8.052554e-05     reject
12 12  0.048883651 38.83257 1.121306e-04     reject
13 13 -0.031133414 39.20143 1.856642e-04     reject
14 14  0.048170709 40.08699 2.472736e-04     reject
15 15 -0.049023571 41.00681 3.190296e-04     reject
16 16  0.009449147 41.04108 5.475803e-04     reject
17 17 -0.022204464 41.23086 8.651446e-04     reject
18 18  0.039861450 41.84425 1.162641e-03     reject
19 19 -0.050569065 42.83429 1.365643e-03     reject
20 20  0.079642047 45.29707 1.005556e-03     reject
21 21 -0.064890628 46.93677 9.576971e-04     reject
22 22 -0.016092919 47.03791 1.451024e-03     reject
23 23  0.017268792 47.15471 2.142963e-03     reject
24 24  0.010034746 47.19427 3.171674e-03     reject
25 25  0.081548697 49.81435 2.246317e-03     reject
> |
```

Step 6

Calculate the sample PACFs and prepare the corresponding table. As the sample ACF is diminishing and the sample PACF cuts off at lag 4, the model will be an AR model of order 3.

	k	sample_pacf	test_statistic	conclusion
1	1	-0.276372473	-5.28731674	reject
2	2	-0.147334865	-2.81868194	reject
3	3	-0.104664682	-2.00235327	reject
4	4	-0.079416581	-1.51932865	accept
5	5	-0.010120658	-0.19361959	accept
6	6	0.043495666	0.83212109	accept
7	7	0.065657989	1.25611128	accept
8	8	0.098024082	1.87531111	accept
9	9	0.045694648	0.87419009	accept
10	10	-0.076006383	-1.45408772	accept
11	11	0.025629966	0.49033011	accept
12	12	0.060954016	1.16611899	accept
13	13	-0.007414715	-0.14185184	accept
14	14	0.046708804	0.89359204	accept
15	15	-0.015026993	-0.28748330	accept
16	16	0.004209372	0.08053003	accept
17	17	-0.024882785	-0.47603571	accept
18	18	0.025728244	0.49221030	accept
19	19	-0.060248577	-1.15262314	accept
20	20	0.041597815	0.79581305	accept
21	21	-0.028239519	-0.54025381	accept
22	22	-0.033538063	-0.64162092	accept
23	23	-0.009469835	-0.18116861	accept
24	24	0.013424466	0.25682515	accept
25	25	0.091038164	1.74166263	accept

Step 7

The AR(3) model for the above stationary series is estimated using the ar function in R. The model equation is given by

$$y_t = -0.332y_{t-1} - 0.1803y_{t-2} - 0.0987y_{t-3} + e_t$$

Where e_t is the error term.

Step 8

After estimating the model, Ljung Box test is used to check whether the residuals are white noise or not. As all the conclusions are accepted in the below table, there is no correlation between the error terms.

```
    k  sample_acf       Q_stat   p_value conclusion
1   1 -0.004101517  0.006140331 0.9375415     accept
2   2 -0.011393638  0.053655502 0.9735289     accept
3   3 -0.026851919  0.318302006 0.9565504     accept
4   4 -0.046932805  1.129038910 0.8896379     accept
5   5  0.050096818  2.055361008 0.8414330     accept
6   6  0.082919604  4.600282559 0.5960014     accept
7   7  0.069772211  6.407234353 0.4930823     accept
8   8  0.051644640  7.400023900 0.4941507     accept
9   9 -0.037849907  7.934791246 0.5407309     accept
10 10 -0.083953833 10.573238100 0.3917177     accept
11 11  0.090495576 13.647618007 0.2531089     accept
12 12  0.072714450 15.638215046 0.2083739     accept
13 13 -0.004843502 15.647072390 0.2687141     accept
14 14  0.034548700 16.099026288 0.3073643     accept
15 15 -0.050062768 17.050748488 0.3158397     accept
16 16 -0.013271107 17.117821558 0.3780128     accept
17 17 -0.017338403 17.232639309 0.4387156     accept
18 18  0.026744596 17.506622302 0.4885711     accept
19 19 -0.031088163 17.877906064 0.5306073     accept
20 20  0.054065696 19.004136988 0.5215571     accept
21 21 -0.060478421 20.417508936 0.4949710     accept
22 22 -0.015624046 20.512114805 0.5510389     accept
23 23  0.022147450 20.702773982 0.5992332     accept
24 24  0.026846247 20.983744290 0.6396840     accept
25 25  0.071758190 22.997111565 0.5777320     accept
>
```

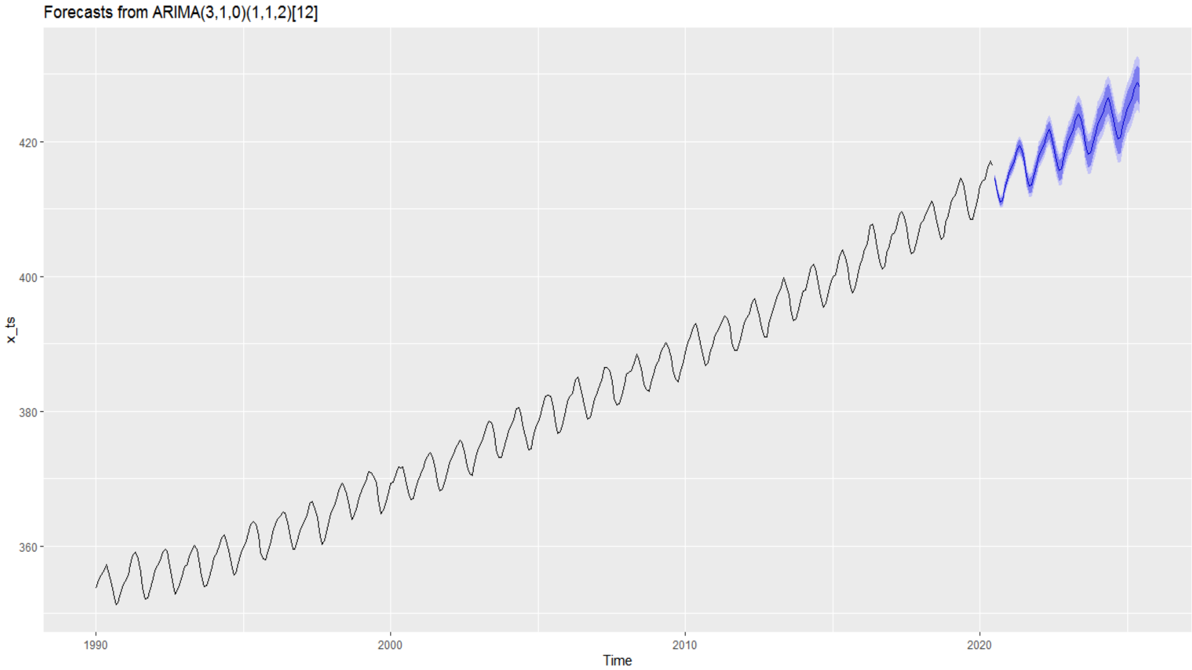
Step 8

For forecasting, we will use the Arima function on our original non-stationary series with the appropriate arguments. Below are the results of the forecasting.

```
Series: x_ts
ARIMA(3,1,0)(0,1,1)[12]

Coefficients:
          ar1      ar2      ar3     sma1
      -0.3596  -0.1855  -0.1039  -0.8656
s.e.   0.0534   0.0555   0.0533   0.0358

sigma^2 estimated as 0.1107:  log likelihood=-111.21
AIC=232.42   AICc=232.6   BIC=251.76
>
```



Conclusion

The above plot indicates the successful forecasting of monthly mean carbon emission in ppm for the next five years. From the plot, it can be concluded that the carbon emission level will increase in the upcoming years, which is not a good sign for the earth.