Varsha Rajasekar, Kush Aswani
Tejeshwar Reddy Nayani

Final Project
12/05/2019

*Analysis of Air Quality*

## I. Introduction

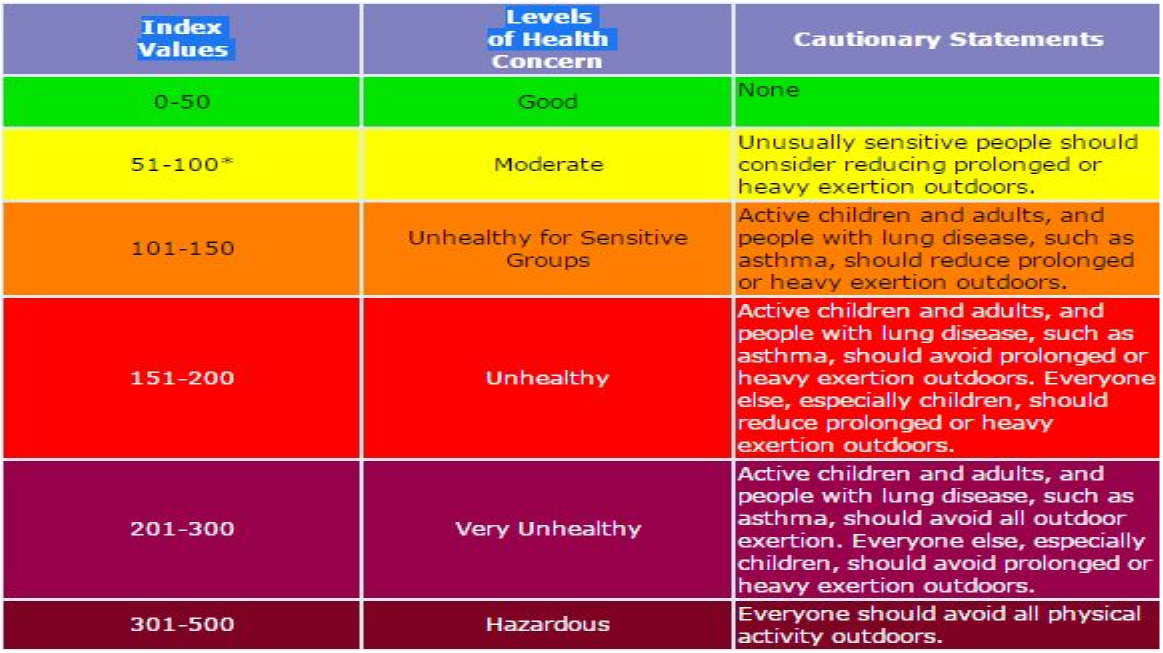
One of the major environmental threats that today's urban population needs to be concerned about is air pollution. Air pollution is more common in cities and poses serious health threats to the large population living in such big cities. Exposures to major air pollutants like sulphur dioxide and nitrogen dioxide adversely affect infants, seniors and people with lung and heart conditions. Even the healthy population is not immune to health outcomes owing to common air pollutants like eye and skin irritation, respiratory diseases, decreased lung and liver function. Air pollution is a consequence of inefficient burning of fuel, emissions from power plants and factories, smoke from car and truck exhausts, etc. Even though we don't realize, our daily routines can have a huge impact on worsening or improving air pollution. Climate change is another factor contributing to the increase in allergic air pollutants like pollen and mold.

In the 1950s and 1960s, New York City experienced severe air pollution. This raised concerns for the public health which was threatened by the harmful gases released by unregulated burning of coal and other fossil fuels. This lead to the involvement of state and eventually, the federal authorities in regulating laws like the Clean Air Act that have improved the air quality sufficiently but not enough. The New York-Newark-Jersey City metropolitan region was the most populated area in the United States in 2018 with about 19.97 million residents(Erin Duffin,2019). At current levels, the air quality in this region can become hazardous if awareness is not raised among the public. Since the population is directly correlated to pollution, therefore, it is important to keep track of the pollutant levels and identify any deterioration in the air quality.

This report will analyze the air quality data obtained from the US EPA website by employing time series algorithms and forecast the future values of the series. The forecasted values can be useful to air quality regulatory bodies like the United States Environmental Protection Agency(US EPA) in formulating laws and standards to reduce and control the emission of air pollutants. It can also help them examine the potential impact of new regulatory requirements on the present population.

## II. Dataset

We got the dataset from Google Big Query Public EPA datasets(https://console.cloud.google.com/marketplace/details/epa/historical-air-quality?filter=solution-type:dataset&q=epa&id=198c2178-3986-4182-a7c7-4c9ae81dfc5d) for the New York-Newark-Jersey City region. The datasets on air quality from 1990-2018 were obtained. We performed time-series analysis on the air quality index (AQI) of three major pollutants: O3 (Ground-level Ozone), SO2 (Sulphur Dioxide) and NO2 (Nitrogen Dioxide). The reason for choosing these three pollutants was that they all had hourly data from 1990-2018. Other pollutant’s dataset was not this consistent.



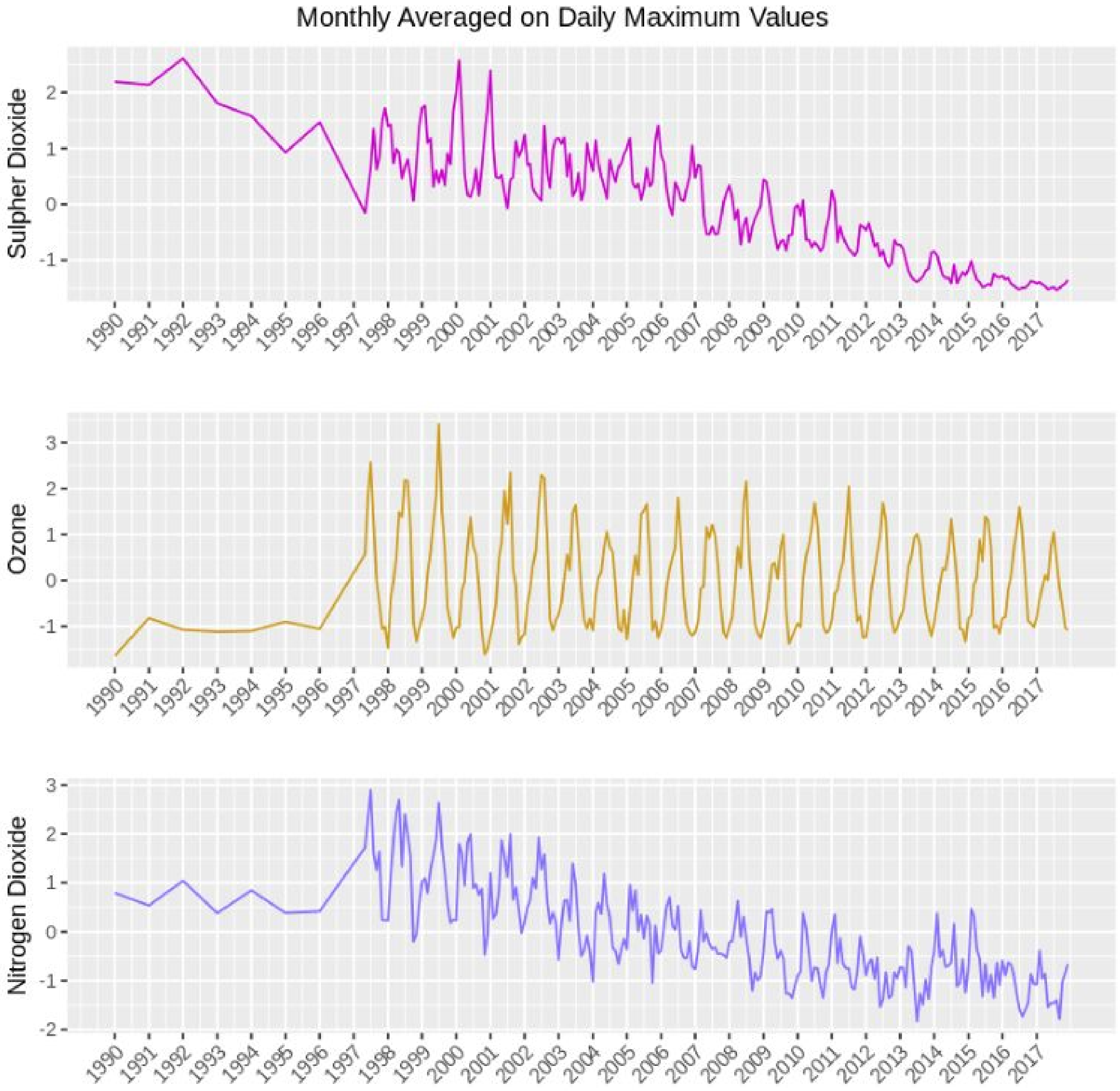
| Index Values | Levels of Health Concern | Cautionary Statements |
|---|---|---|
| 0-50 | Good | None |
| 51-100* | Moderate | Unusually sensitive people should consider reducing prolonged or heavy exertion outdoors. |
| 101-150 | Unhealthy for Sensitive Groups | Active children and adults, and people with lung disease, such as asthma, should reduce prolonged or heavy exertion outdoors. |
| 151-200 | Unhealthy | Active children and adults, and people with lung disease, such as asthma, should avoid prolonged or heavy exertion outdoors. Everyone else, especially children, should reduce prolonged or heavy exertion outdoors. |
| 201-300 | Very Unhealthy | Active children and adults, and people with lung disease, such as asthma, should avoid all outdoor exertion. Everyone else, especially children, should avoid prolonged or heavy exertion outdoors. |
| 301-500 | Hazardous | Everyone should avoid all physical activity outdoors. |

AQI reports the hourly air quality and tells us the health effects associated with a particular index. It is measured on a scale of 0 to 500 where 0-50 interval means the air quality conditions are good and 301-500 interval means the conditions are hazardous. The Clean Air Act provides regulations for five major pollutants and their AQI is calculated by EPA which has also established air quality standards for the country.

## III. Methodology

We first checked for missing values in the data that may have to be removed from the dataset for analysis. After getting the summary of the observations we visualized the data to observe it's distribution and checked for outliers. Visualization of the distribution was also done on the monthly average of daily maximum values. The distribution had lesser outliers after aggregation. Then, we observed the time series flow of the three pollutants to see how their data has changed over time. For this, we considered the monthly average of daily maximum values, monthly maximum values, and hourly maximum values. The stationarity of the data is also checked by plotting the mean of each hour. We wanted to know the maximum harm done by a pollutant in a day so we decided to predict the monthly average of daily maximum values of the pollutants. The reason for this being if we considered the average values of pollutants it won't show the full extent of harm being done during peak hours.

Next, we performed decomposition of the time series plot of each pollutant(monthly averaged over daily maximum values) to realize the seasonality and trend in the time series. We used ARIMA and Exponential Smoothing State Space(ETS) Model models for forecasting. The ARIMA model is determined by using the auto.arima() function and the ETS model is determined by using the ets() function. To validate the model we split the data into training and testing sets where we used the data from 1998 to 2015 as training sets and the data from 2016-2017 were used as for testing. After getting the training vs testing sets plot, we use the MAPE(Mean Absolute Percentage Error) and the Ljung-Box test to measure the prediction accuracy of the ARIMA model and test for the absence of autocorrelation at a particular number of lags respectively. In the training vs testing plots, the green line indicates the actual values and the red line represents the forecasted values.

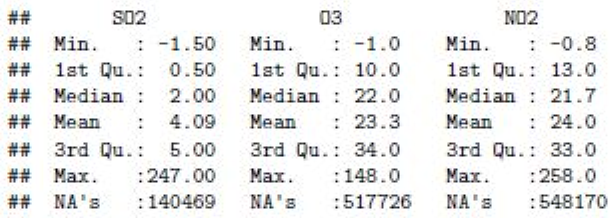


As we could see data from 1990 to 1998 and 1998-2017 varies significantly due to unknown reasons. Hence we considered data only from 1998 for the analysis.

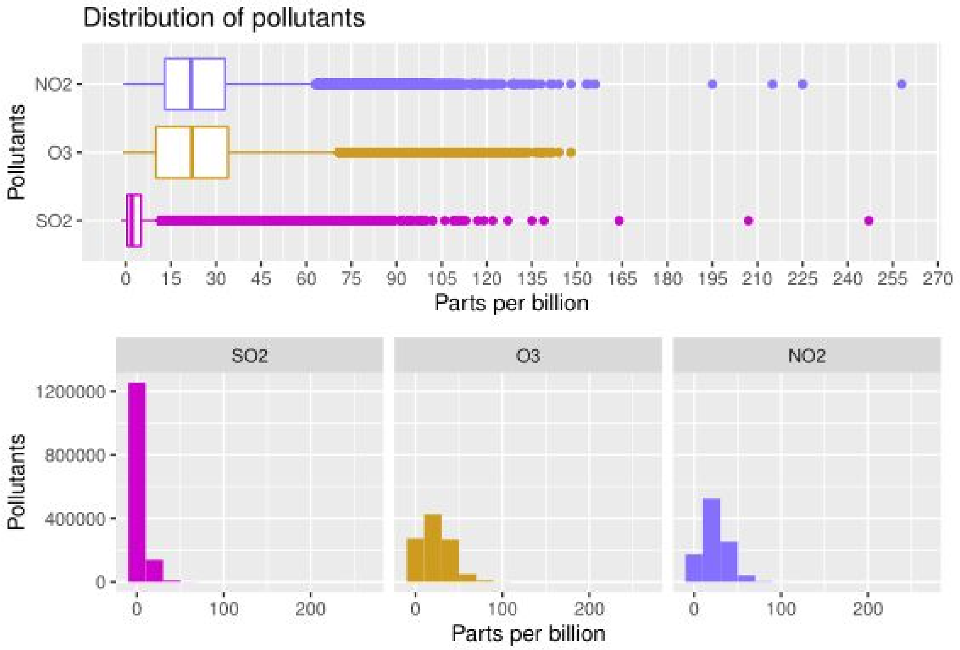
We have also plotted the ACF plot and the residuals plot to check the ARIMA model before forecasting. Then we used the data from 1998-2017 as the training set and forecasted data for 2018-2019.

## IV. Results

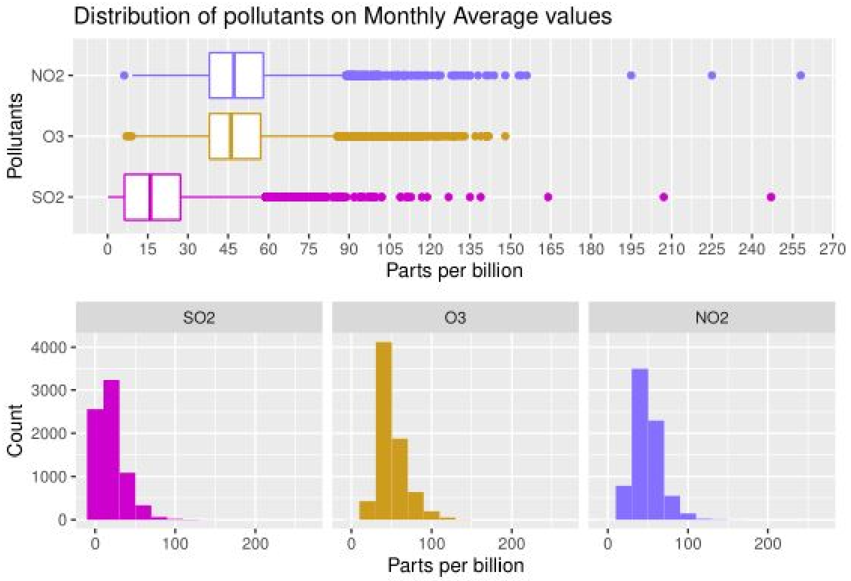
Summary statistics of pollutants Suplur dioxide, Ozone, and Nitrogen Dioxide.



```
##       SO2              O3              NO2
##  Min.   : -1.50   Min.   : -1.0   Min.   : -0.8
##  1st Qu.:  0.50   1st Qu.: 10.0   1st Qu.: 13.0
##  Median :  2.00   Median : 22.0   Median : 21.7
##  Mean   :  4.09   Mean   : 23.3   Mean   : 24.0
##  3rd Qu.:  5.00   3rd Qu.: 34.0   3rd Qu.: 33.0
##  Max.   :247.00   Max.   :148.0   Max.   :258.0
##  NA's   :140469   NA's   :517726  NA's   :548170
```

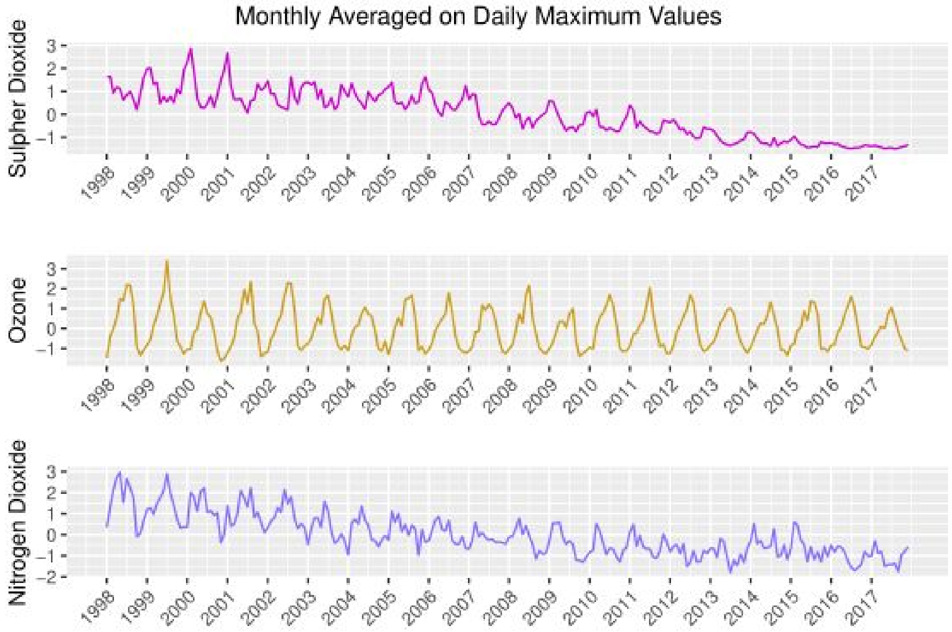


We can see that there are many outliers in each of the pollutants. One reason for this can be an event causing heavy traffic. Because of this, we cannot disregard these values completely. We took the monthly average of daily maximum values and looked at the distributions to see if we can get better distributions.



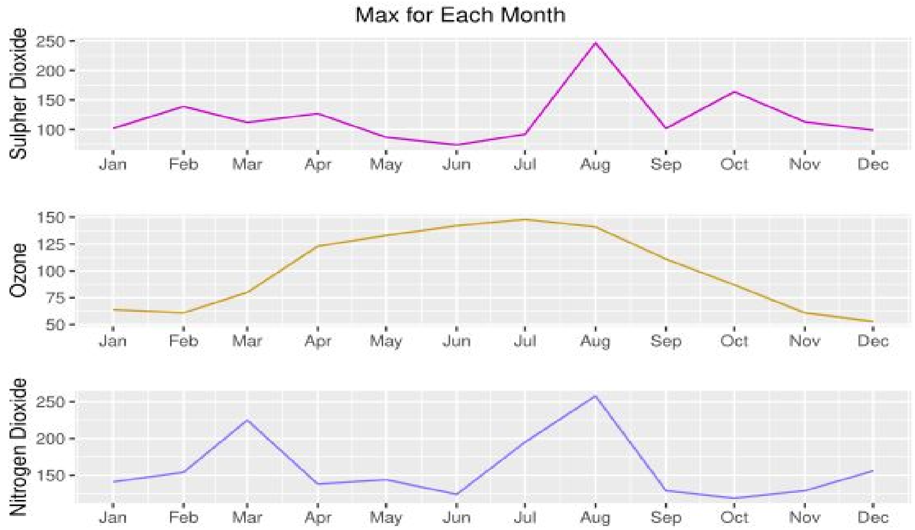
After aggregation, the distribution shows considerably fewer outliers and better distributions.

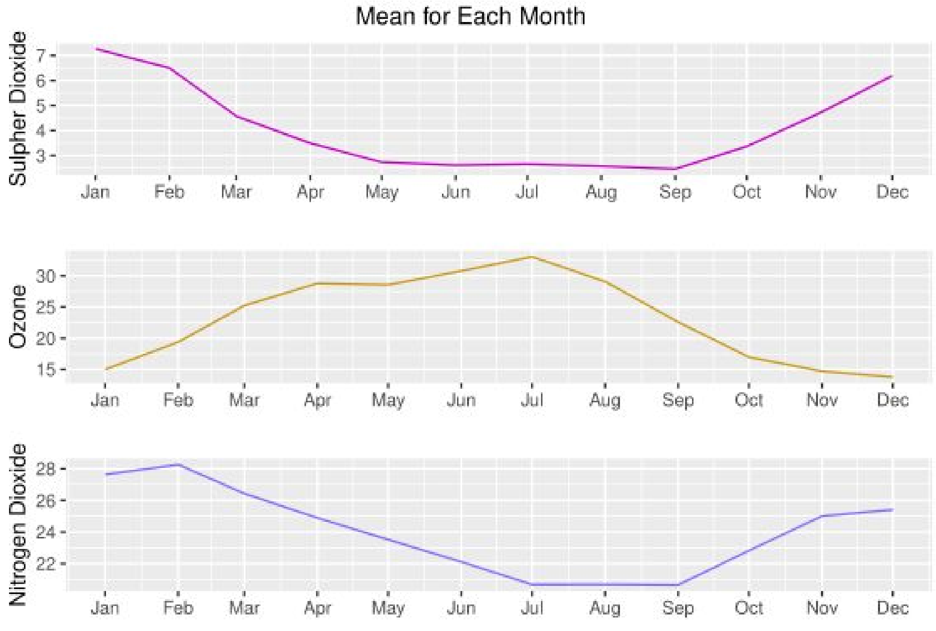
Below is the graph of monthly averaged on daily maximum values considering data from 1998 to 2017.



This shows us a better trend than when we considered the data from 1990-2017. Both SO2 and NO2 values decrease overtime while Ozone remained almost constant.

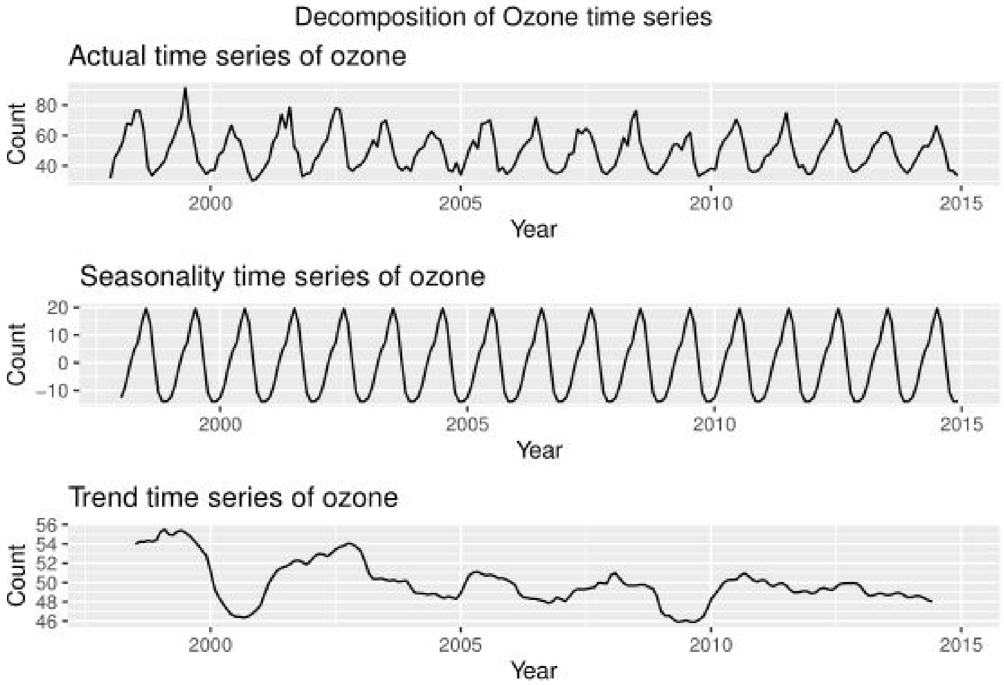
Below we have graphs for max and mean values for each month.





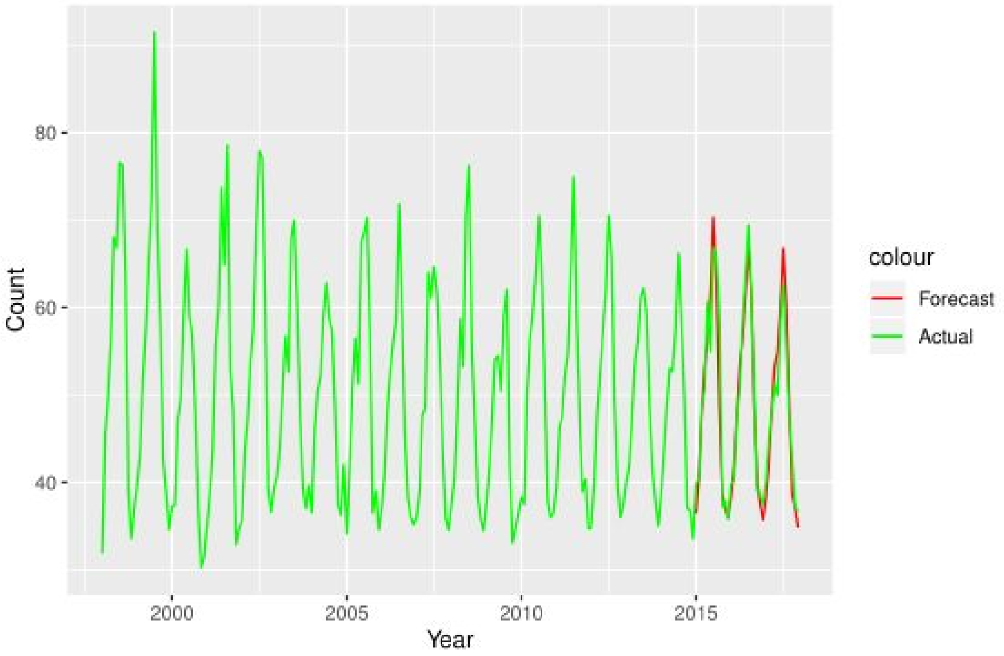
From the mean values for each month we can notice that Ozone has high values during summer and lower values during winter. SO2 and NO2 follow reverse trends compared to Ozone.

### OZONE

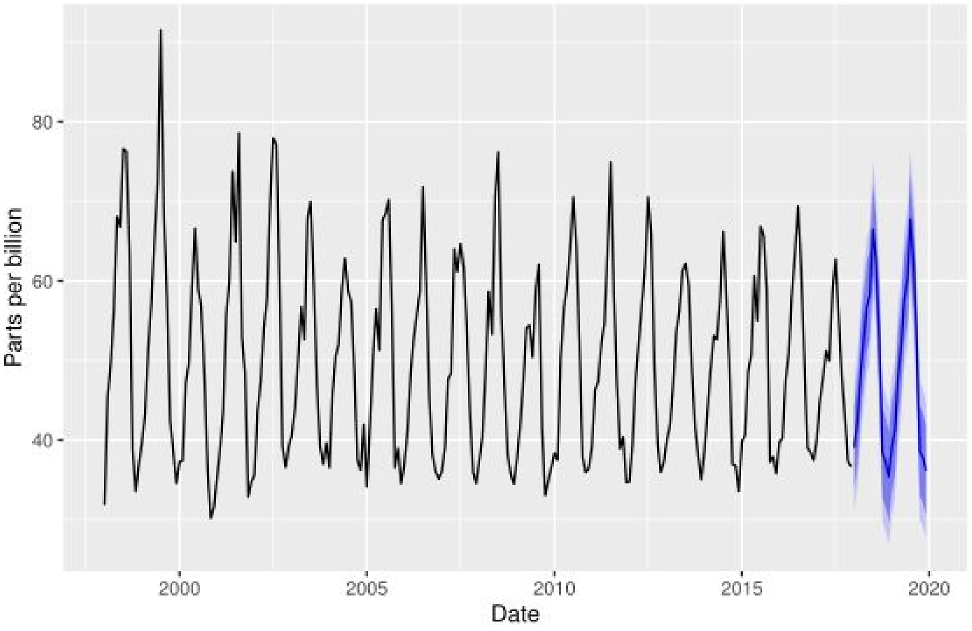


We could recognize that there is little seasonality effect and there is a slightly decreasing trend over time.

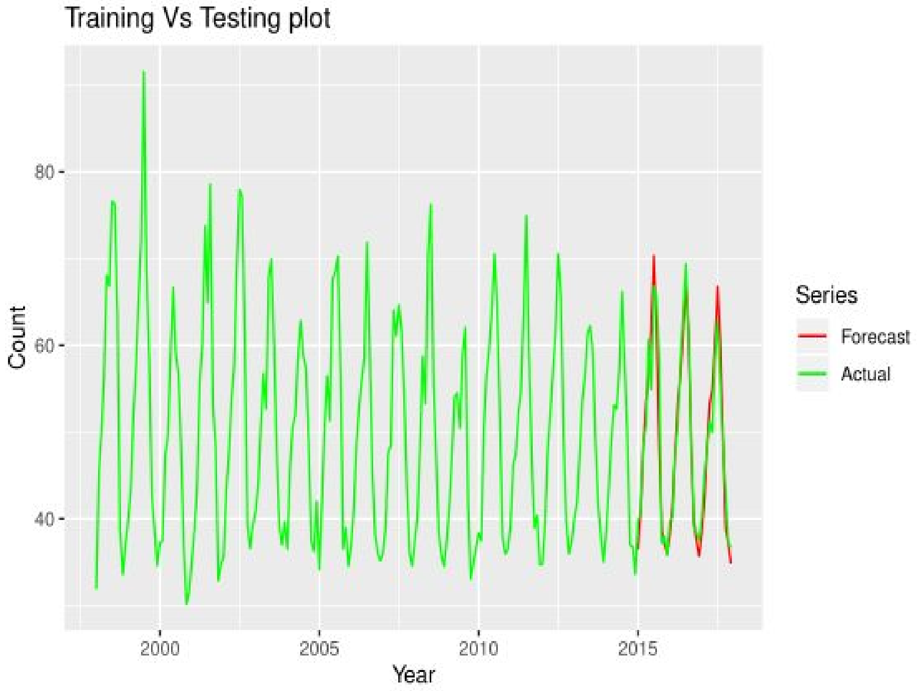
We trained both the ARIMA and ETS models for Ozone. Below we have training vs testing and forecast model plots for both models.

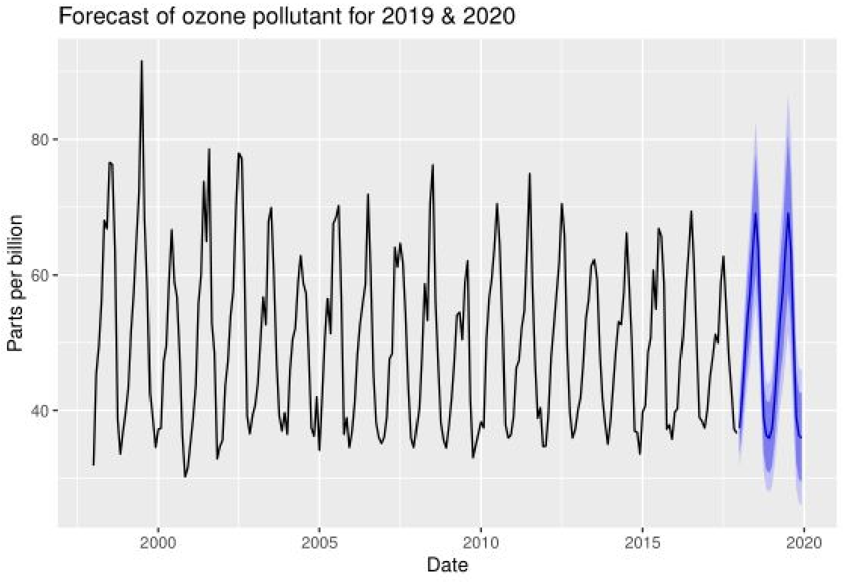


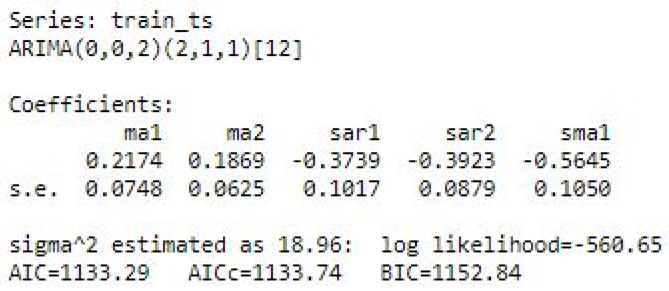
Training Vs Testing plot



Forecast of ozone pollutant for 2018 & 2019


Training Vs Testing plot
Count
80
60
40
2000
2005
2010
2015
Year
Series
Forecast
Actual

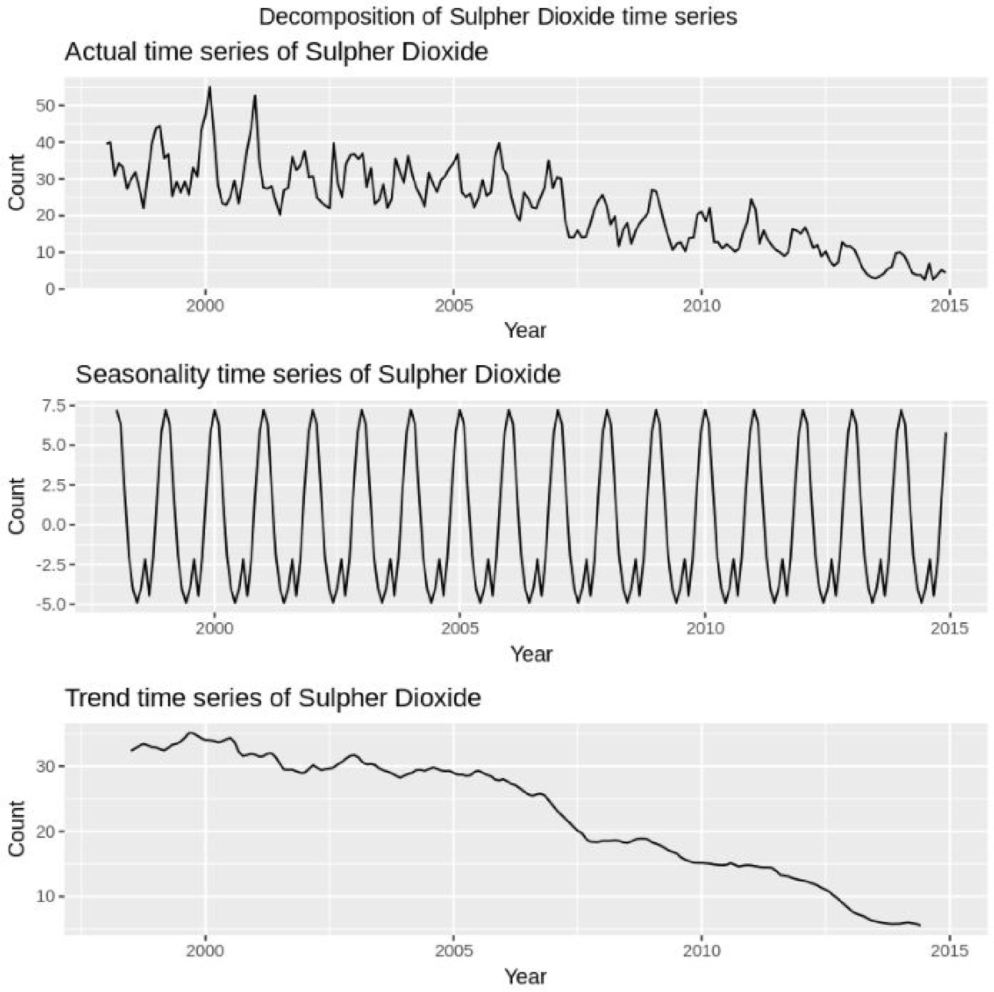

Forecast of ozone pollutant for 2019 & 2020
Parts per billion
80
60
40
2000
2005
2010
2015
2020
Date



```
Series: train_ts
ARIMA(0,0,2)(2,1,1)[12]

Coefficients:
         ma1     ma2     sar1     sar2     sma1
      0.2174  0.1869  -0.3739  -0.3923  -0.5645
s.e.  0.0748  0.0625   0.1017   0.0879   0.1050

sigma^2 estimated as 18.96:  log likelihood=-560.65
AIC=1133.29   AICc=1133.74   BIC=1152.84
```

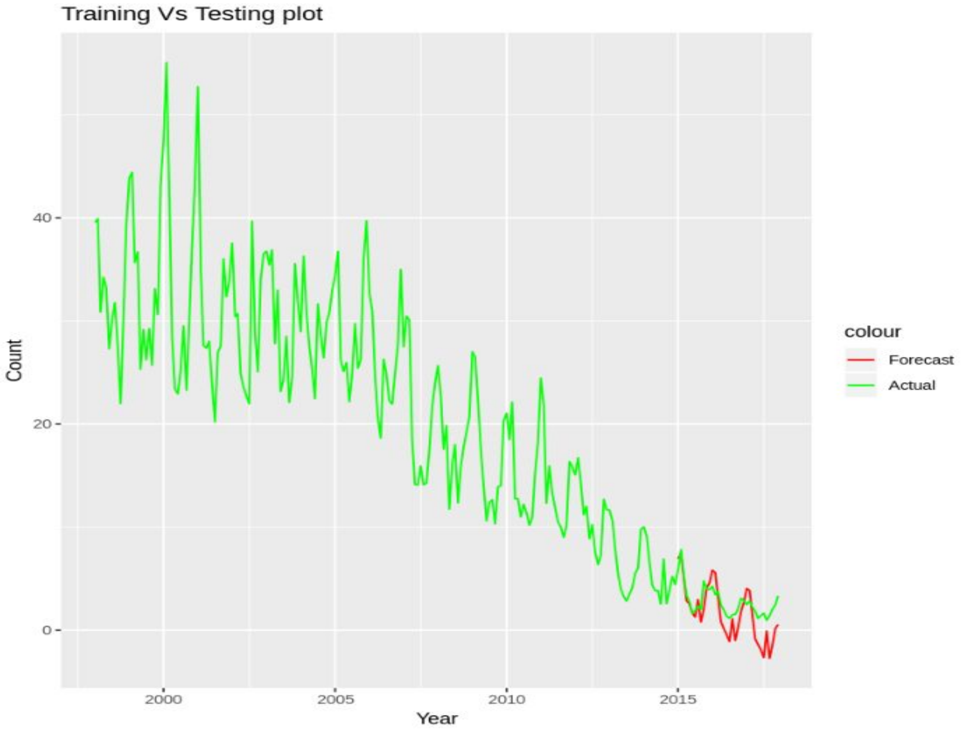
Comparing both ARIMA and ETS models based on the MAPE results, ARIMA(0,0,2) is chosen.

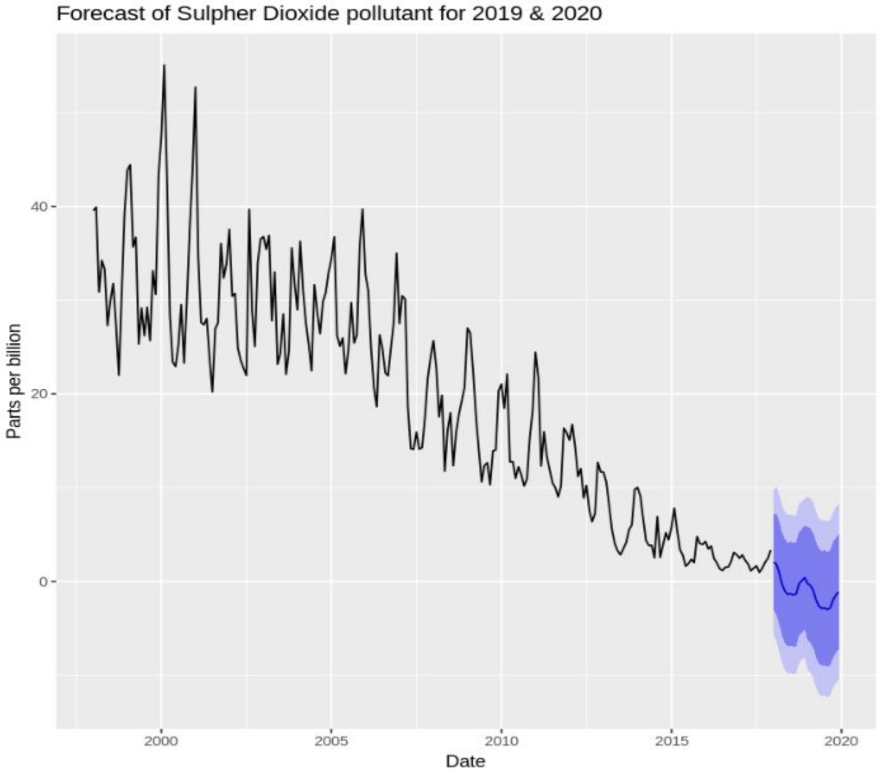
## SULPHUR DIOXIDE:

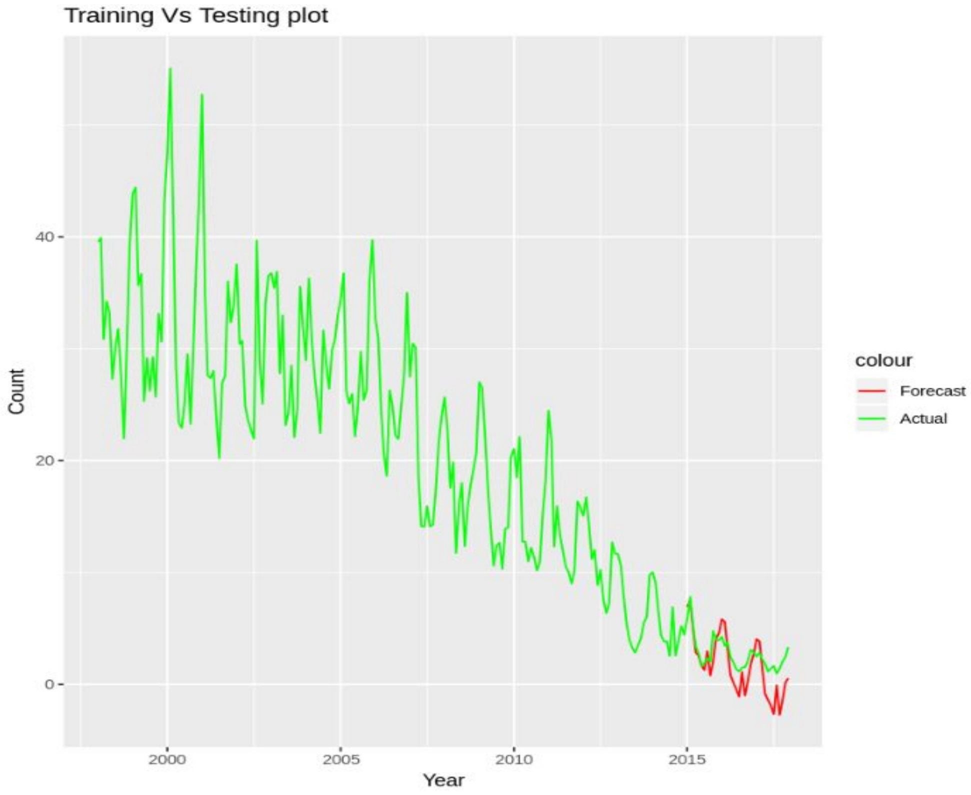


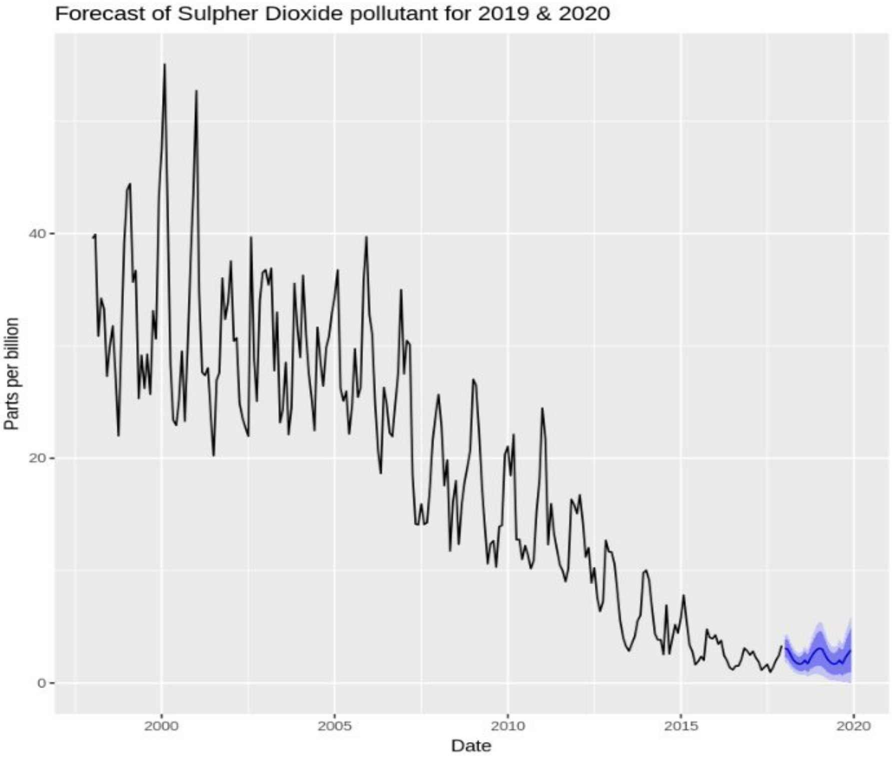
We could recognize that there is a clear decreasing trend over time.

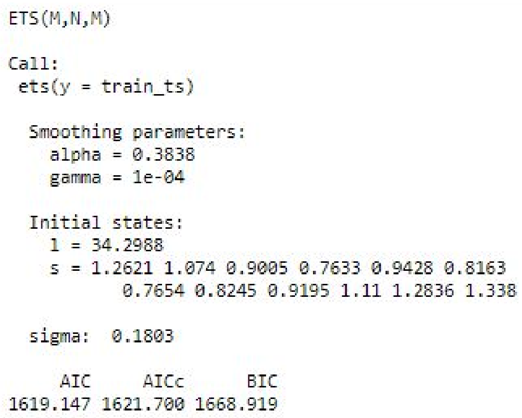
We trained both the ARIMA and ETS models for SO2. Below we have training vs testing and forecast model plots for both models.






Training Vs Testing plot
Count
40
20
0
2000
2005
2010
2015
Year
colour
Forecast
Actual

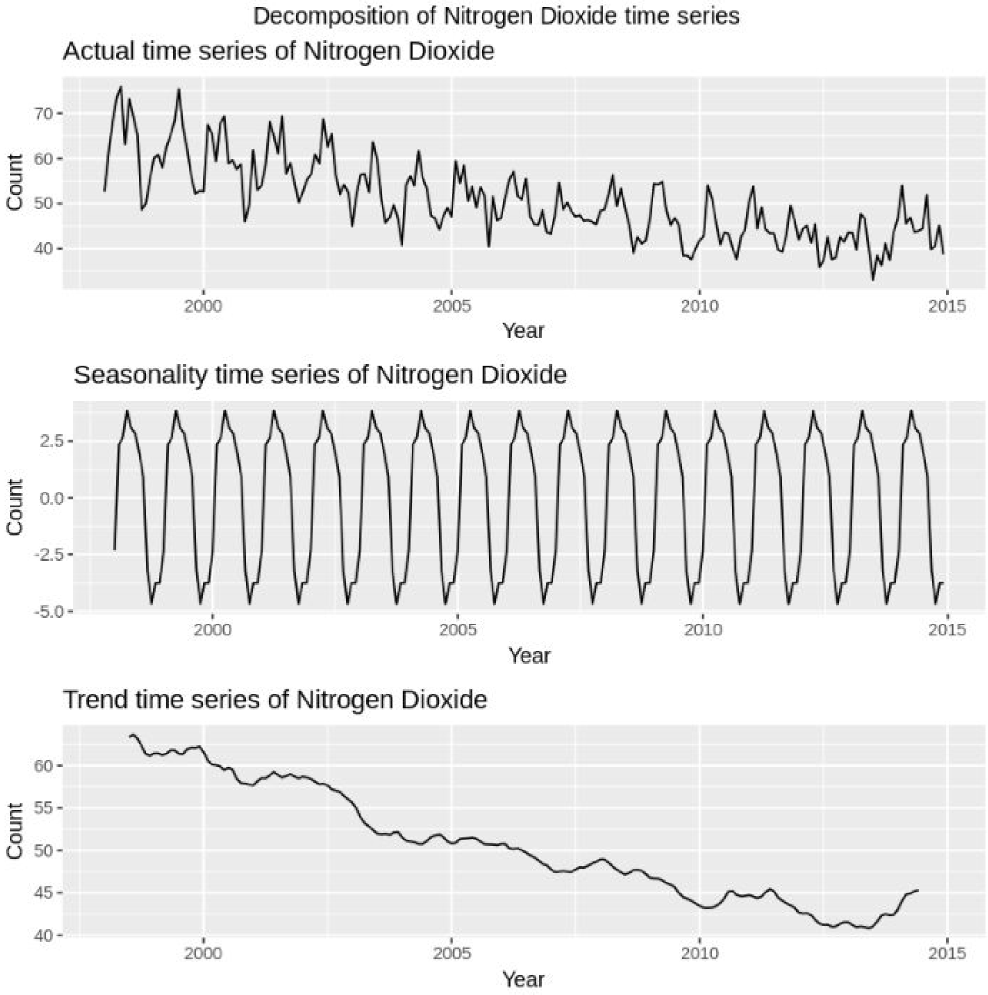

Forecast of Sulpher Dioxide pollutant for 2019 & 2020
Parts per billion
40
20
0
2000
2005
2010
2015
2020
Date



```
ETS(M,N,M)

Call:
 ets(y = train_ts)

  Smoothing parameters:
    alpha = 0.3838
    gamma = 1e-04

  Initial states:
    l = 34.2988
    s = 1.2621 1.074 0.9005 0.7633 0.9428 0.8163
           0.7654 0.8245 0.9195 1.11 1.2836 1.338

  sigma:  0.1803

     AIC     AICc      BIC
1619.147 1621.700 1668.919
```

Comparing both ARIMA and ETS models based on the MAPE results, ETS is chosen.

## NITROGEN DIOXIDE:



Decomposition of Nitrogen Dioxide time series

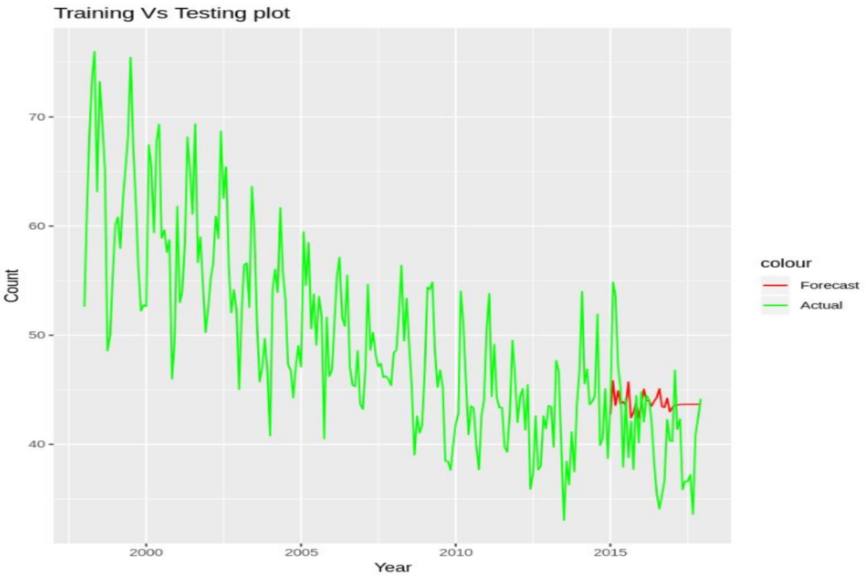
Actual time series of Nitrogen Dioxide

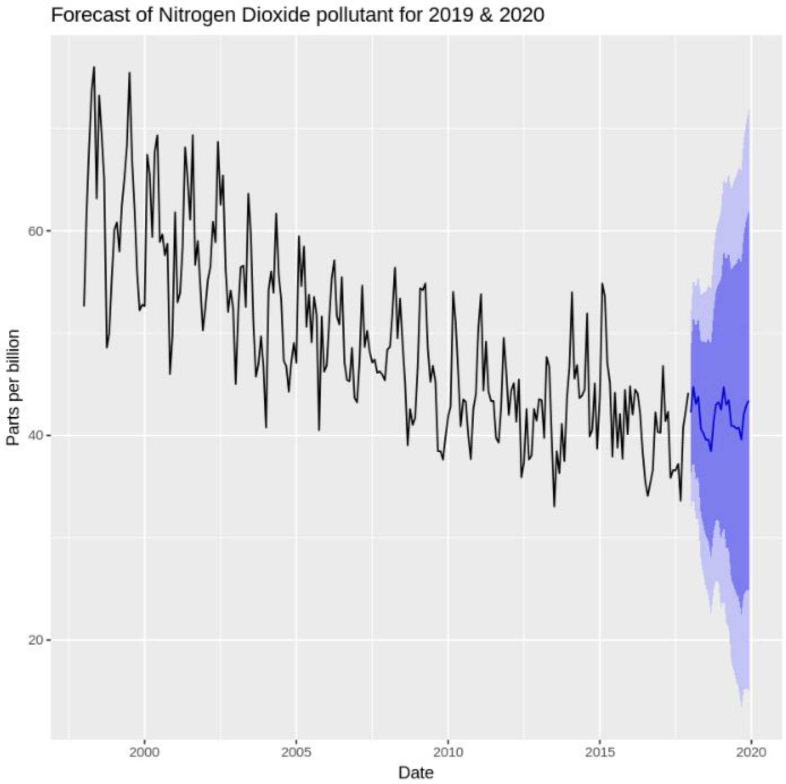
Seasonality time series of Nitrogen Dioxide

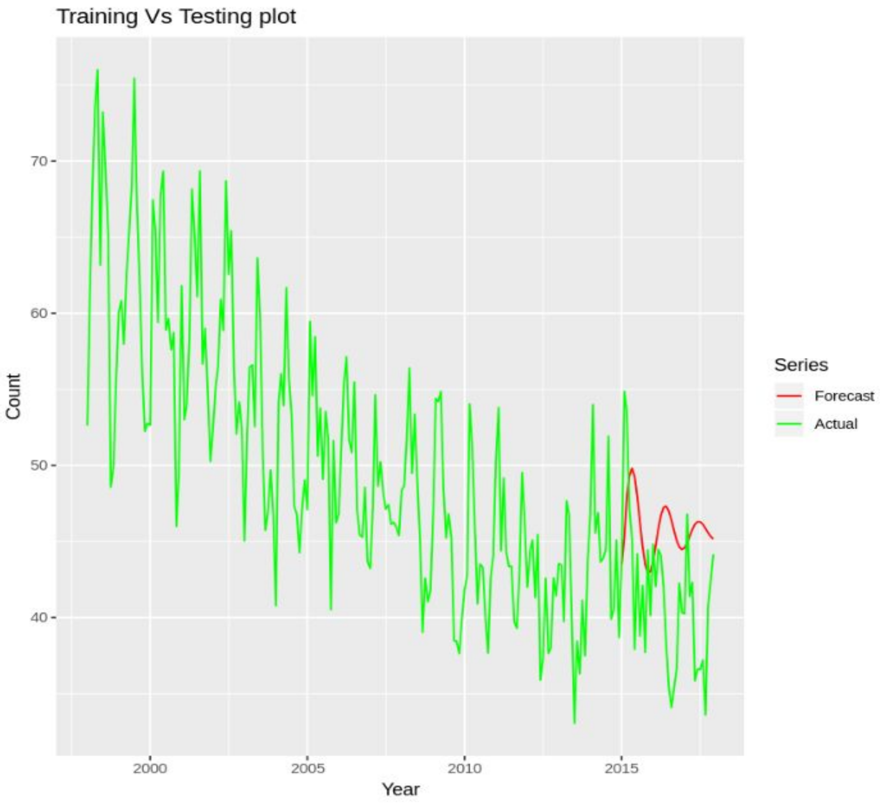
Trend time series of Nitrogen Dioxide

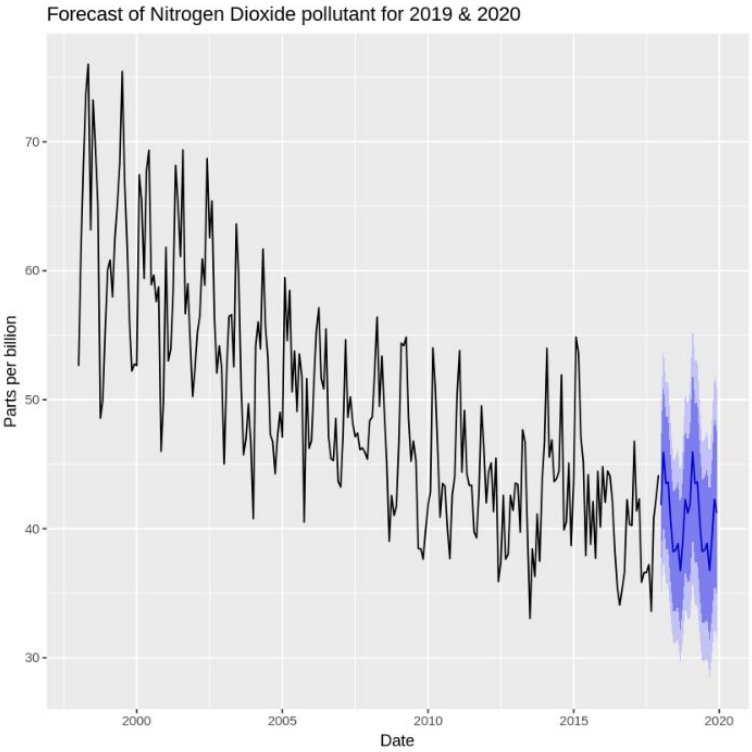
We could recognize that there is little seasonality effect and there is a clear decreasing trend over time.

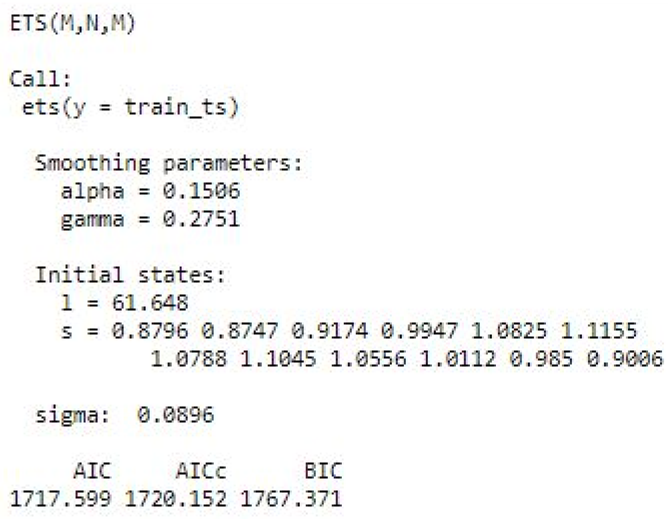
We trained both the ARIMA and ETS models for NO2. Below we have training vs testing and forecast model plots for both models






Training Vs Testing plot
Count
70
60
50
40
2000
2005
2010
2015
Year
Series
Forecast
Actual


Forecast of Nitrogen Dioxide pollutant for 2019 & 2020
Parts per billion
70
60
50
40
30
2000
2005
2010
2015
2020
Date

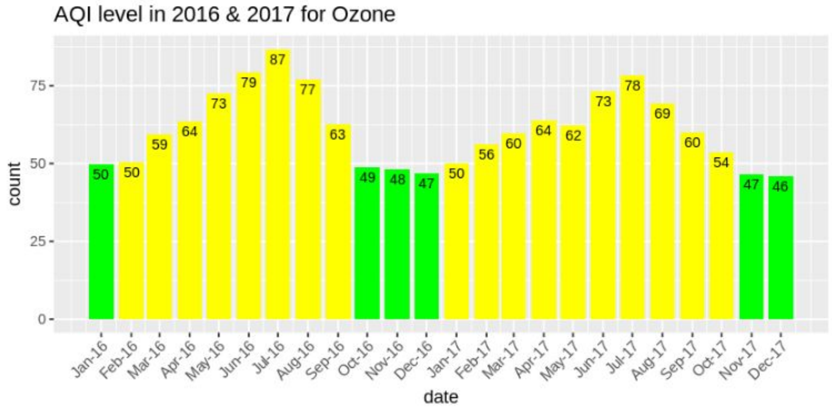


```
ETS(M,N,M)

Call:
 ets(y = train_ts)

  Smoothing parameters:
    alpha = 0.1506
    gamma = 0.2751

  Initial states:
    l = 61.648
    s = 0.8796 0.8747 0.9174 0.9947 1.0825 1.1155
           1.0788 1.1045 1.0556 1.0112 0.985 0.9006

  sigma:  0.0896

     AIC     AICc      BIC
1717.599 1720.152 1767.371
```

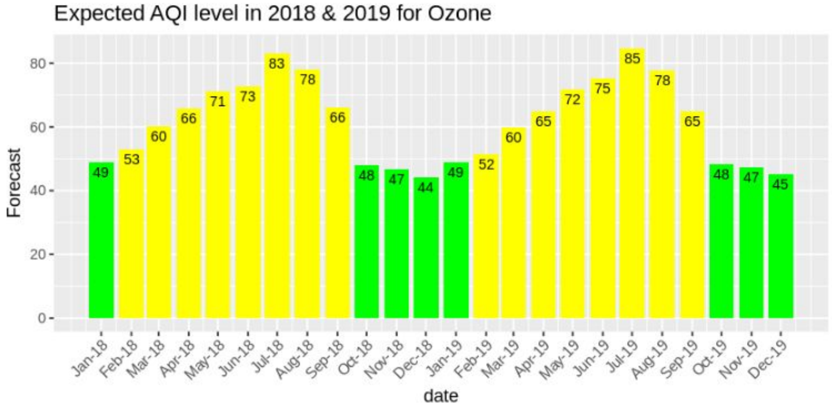
Comparing both ARIMA and ETS models based on the MAPE results, ETS is chosen.

## V. Conclusion

The units of values from data are parts per billion(ppb). For each pollutant, a relevant formula is used to convert it to AQI value.

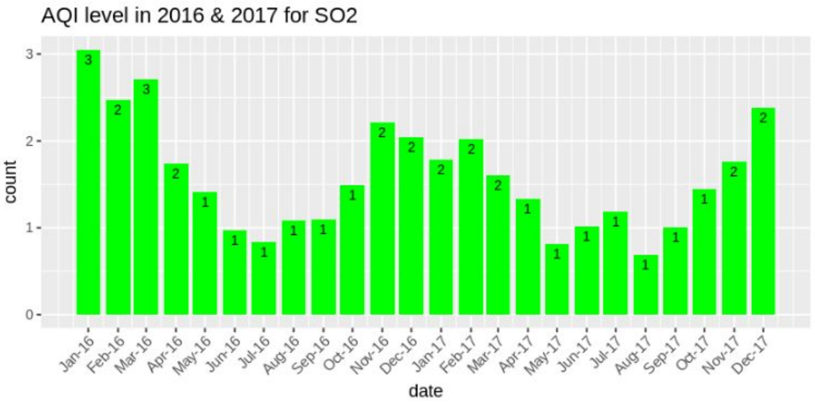


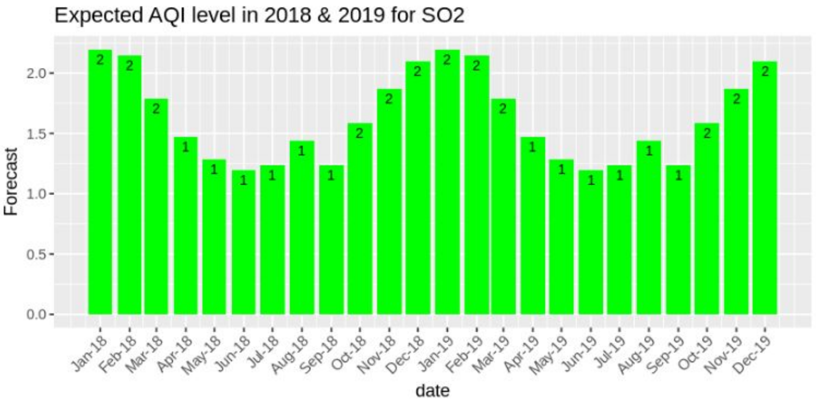
AQI level in 2016 & 2017 for Ozone



Expected AQI level in 2018 & 2019 for Ozone

We could notice the similarity in trends and values. It is the same as what we observed from data(1998-2017).  The image shows that in the summer effect of Ozone on health is moderate.

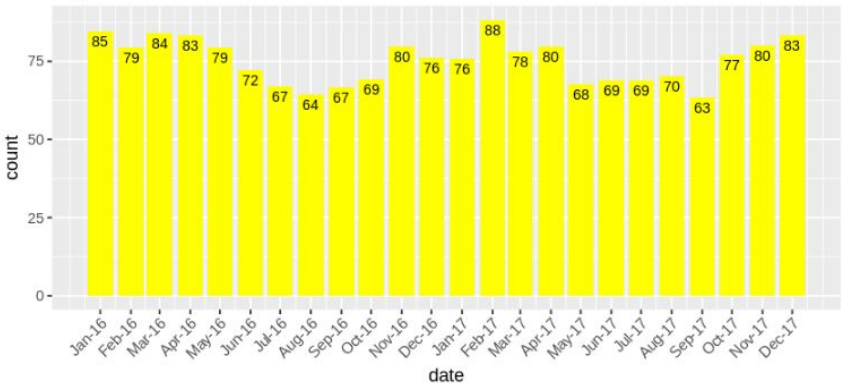




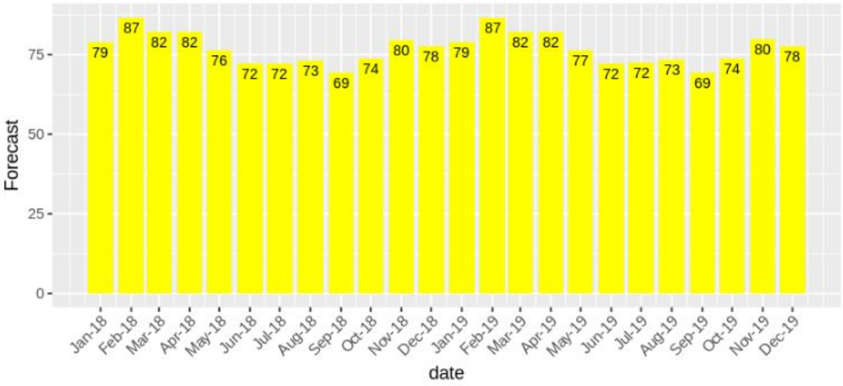
We could notice the trend of low values in summer and high values in winter. And also the maximum value over the year has decreased from 3 to 2.

Below we could see AQI values for NO2. We expect the values to decrease over time but over forecasted weren’t less compared to previous year values. This might due to the effect of the increase in values in the year 2013-14 had on the model.
The image shows that the effect of NO2 on health is moderate all over the year.

AQI level in 2016 & 2017 for NO2



Expected AQI level in 2018 & 2019 for NO2



Overall we had a better model for Ozone compared to NO2 and SO2.

## VI. References

[1]https://otexts.com/fpp2/arima-ets.html

[2]https://airnow.gov/index.cfm?action=aqibasics.aqi