14-MAY-2019

TIME SERIES ANALYSIS OF POLLUTANT LEVELS IN SAN DIEGO

SNEHA THANASEKARAN
ANUJ MATHUR

EXECUTIVE SUMMARY

Our focus on this project was on the 4 pollutants (Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone) levels in San Diego. The US pollutants data was available on data.world. We pre-processed the data to focus on one city, one pollutant (No2) for univariate Analysis and the interaction of 4 pollutants with each other for multivariate analysis. The data was stationary, and we found seasonality repeating every 12 months. No drastic changes or outliers spotted in the data.

For univariate analysis, we came up with 10 different models, tested the efficiency in prediction through aic and rolling forecasting. After choosing the final model, we predicted the level of pollutants for the next 12 months and compared it to the test data. The actuals were well within the 95% confidence interval. This can be owed to a consistent pattern and seasonality in the data.

For multivariate analysis, we investigated the scatter plot to understand the correlation between the pollutants. There was a significant relationship found between No2 and the other pollutants. S02 remained uncorrelated. We decided to make a linear regression model and build analysis on that. The resulted prediction was well within the range, but time series prediction gave better results. We ventured further into the VAR model. Since the lag was at 4, the model was bigger than anticipated. We decided to finalize the linear regression model for the multivariate analysis.

DATASET

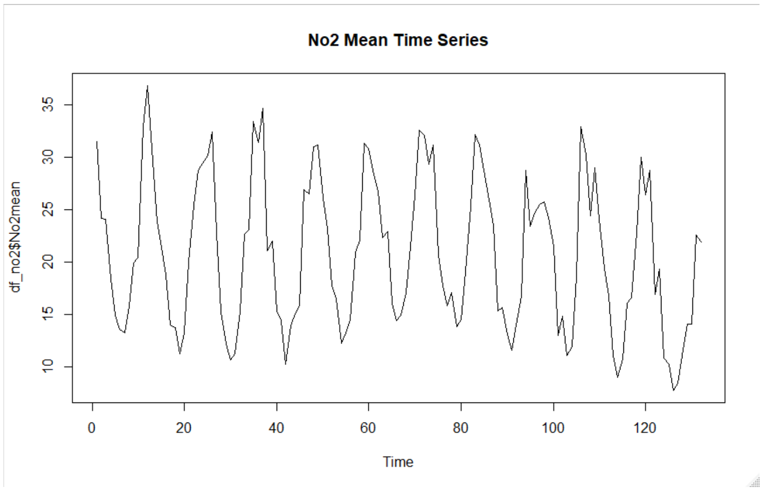
The dataset is from data.world data society (https://data.world/data-society/us-air-pollution-data). The dimension of the dataset is 1.75mx29. The data is recorded daily from 2000 – 2011. It covers every city in the US, but our focus will be only on San Diego: The Data set contains 8 fields about the location and 20 fields on the pollutants - Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone. Each Pollutant has 5 specific columns. Our Focus will be on the City of San Diego and the mean concentration recorded for each gas.

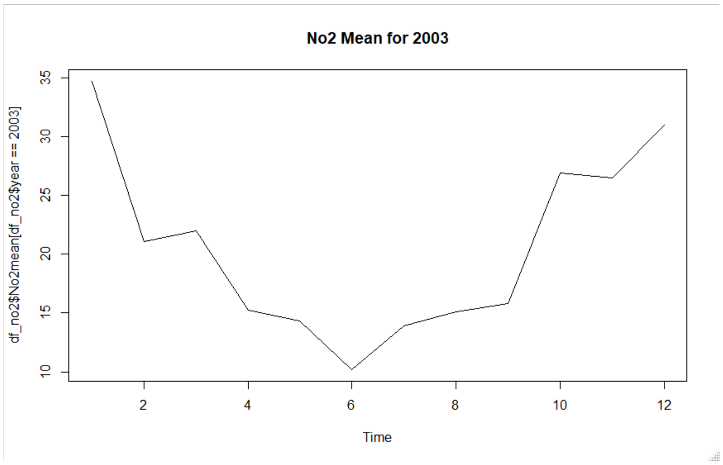
After basic filtering of the dataset to the city of San Diego, there are 29994 rows. The data was recorded daily. We consolidated the mean to give monthly data for our analysis. We kept year 2011 as our test data and made predictions for the same year.

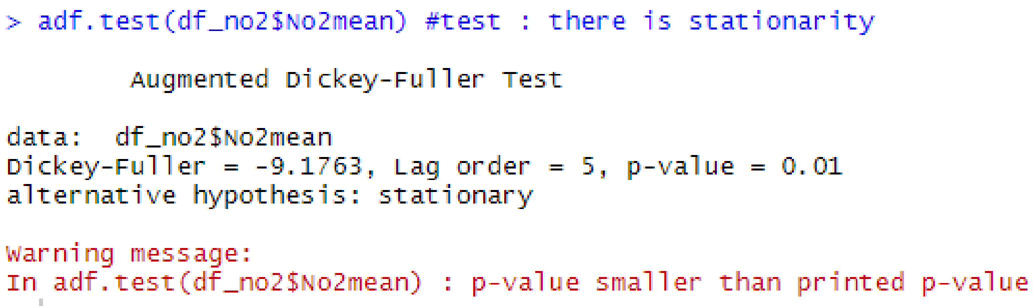
UNIVARIATE ANALYSIS

After preprocessing, we started the analysis by plotting the time series of the No2 mean concentration. The data was stationary, and we confirmed it through Dicky-fuller Test. We also noticed the seasonality repeating every 12 months (From the ACF plot)

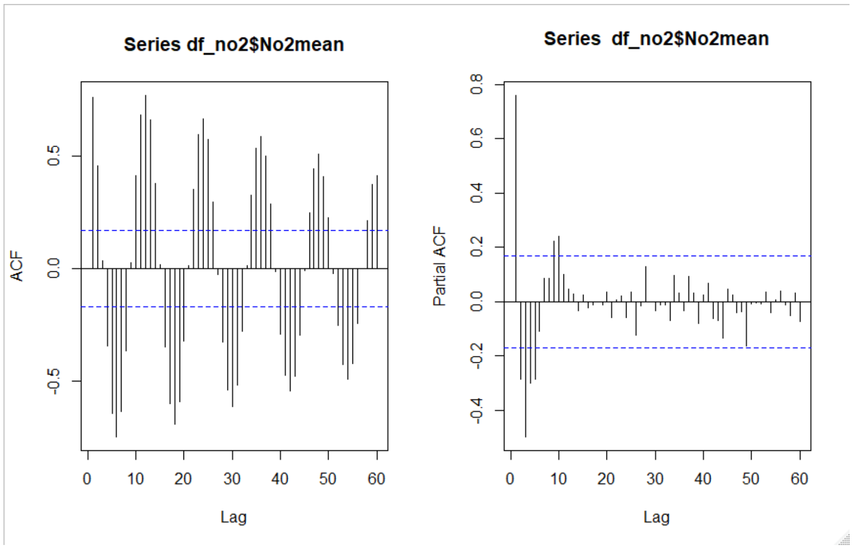
We took a seasonal difference of the data to remove seasonality for modelling.



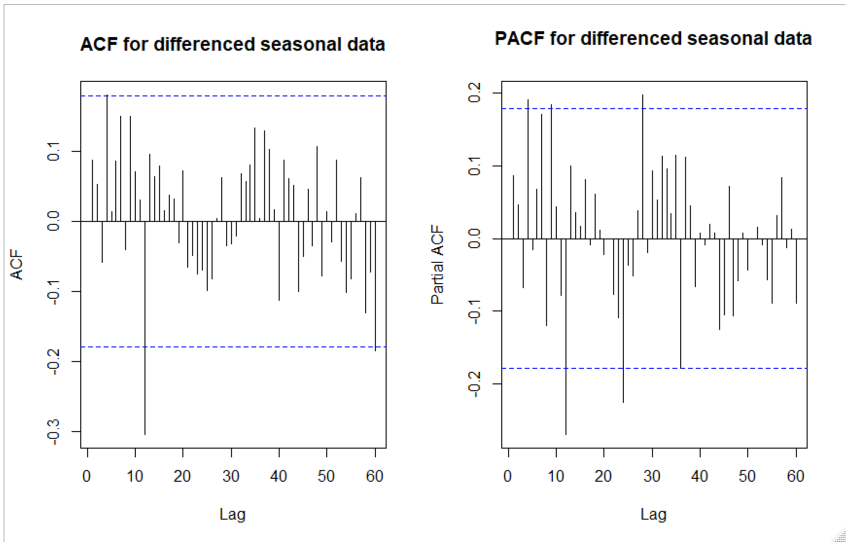




```
> adf.test(df_no2$No2mean) #test : there is stationarity

        Augmented Dickey-Fuller Test

data:  df_no2$No2mean
Dickey-Fuller = -9.1763, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(df_no2$No2mean) : p-value smaller than printed p-value
```



After taking the seasonality, the acf and pacf is plotted to check the pattern. Seasonality was removed and the plot showed an AR lag at 3 and MA lag at 1.



We were able to proceed further and create about 10 models that can have various degrees of significant co-efficients. The acf, pacf and eacf was observed for white noise. When the co-efficients are not significant, they were fixed using manual 0s. Polynomial roots were checked for the efficiency of the system.

We finalized on 5 models and decided to choose the best of the 5.

i. ARMA(c(4,0,0),seasonal=list(order=c(3,0,0))))

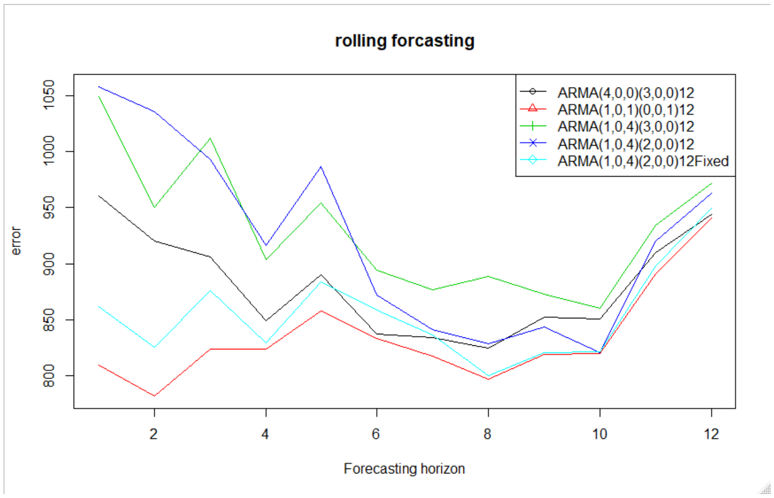
ii. ARMA(c(1,0,1),seasonal=list(order=c(0,0,1))))

iii. ARMA(c(1,0,4),seasonal=list(order=c(3,0,0))))

iv. ARMA(c(1,0,4),seasonal=list(order=c(2,0,0))))

v. ARMA(c(1,0,4),seasonal=list(order=c(2,0,0)), fixed = c(NA,NA,0,0,NA,NA,NA,0)))

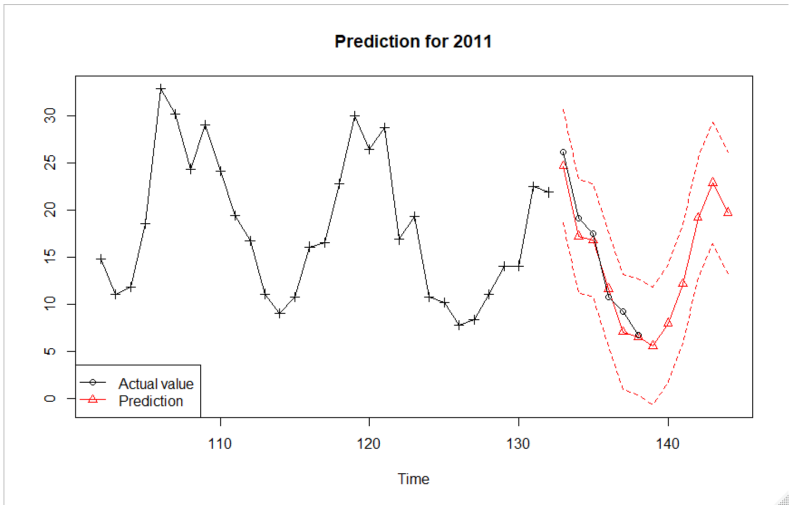
AIC and Rolling Forecasting was performed. From both the results, it was evident that the ARMA(1,0,1)(0,0,1)x12 model outperformed the others by lowest aic and lowest error among others.



The chosen model was used to do the future prediction. We were able to predict the No2 mean concentration for the 12 months of 2011 and compare it with the test data. The predictions were very close to the actuals

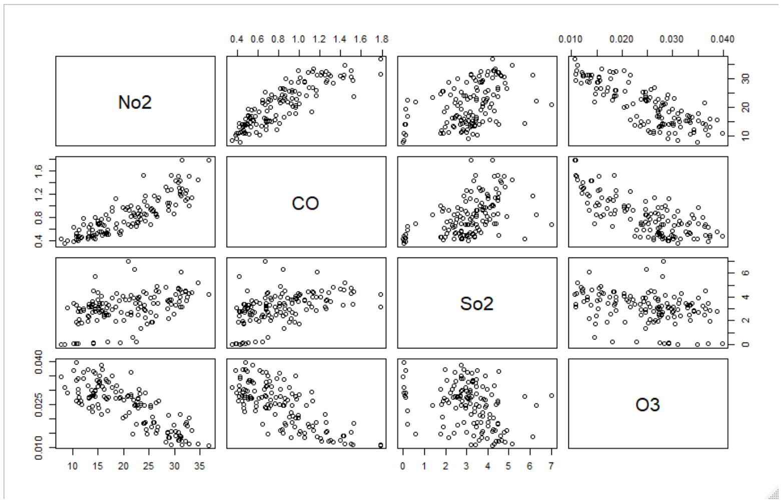
INFERENCE

The power of the prediction can owe to a lot of factor. The data had a consistent seasonality and there was no drastic change in the trend of the data. Unless there is an extra influence, the model can be used to predict the future dates as well.

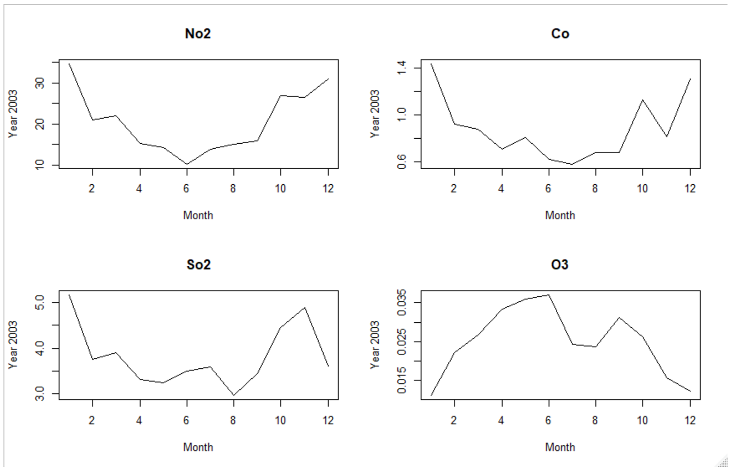


MULTIVARIATE ANALYSIS

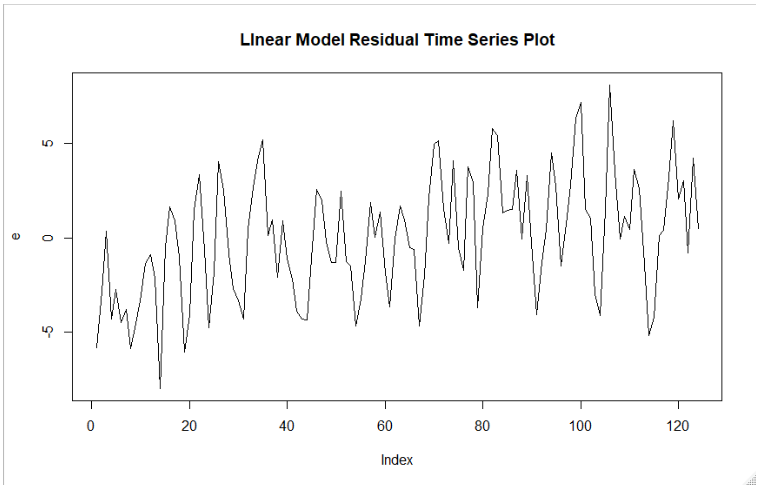
Since we had 4 pollutants in the data, we wanted to proceed further to find the interaction of the pollutants with each other. We found No2 to be significantly interacting with the other gases. It had a positive correlation with Co2 and negative correlation with O3. So2 didn't have any effect with the other gases.

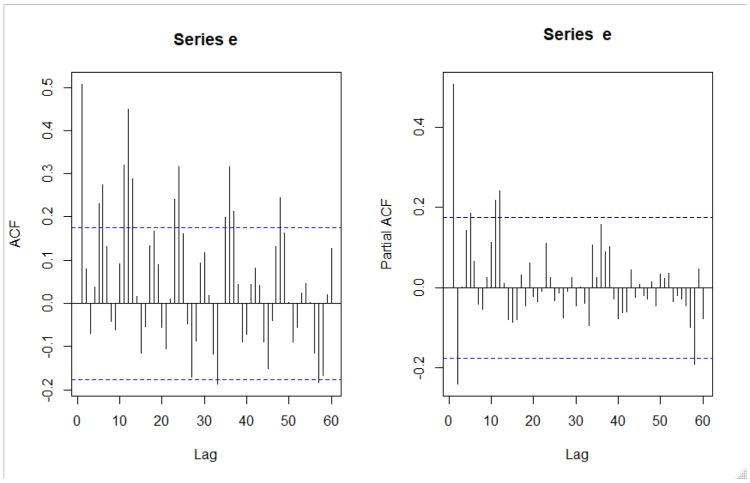


After research and plotting 1-year data of all the 4 gases, we found that N02 increases as Co increases as they both are emitted through heavy traffic and industry pollution. O3 is a result of the interaction of No2 and CO with other different gases. Hence it is inversely proportional to the 2 pollutant levels.



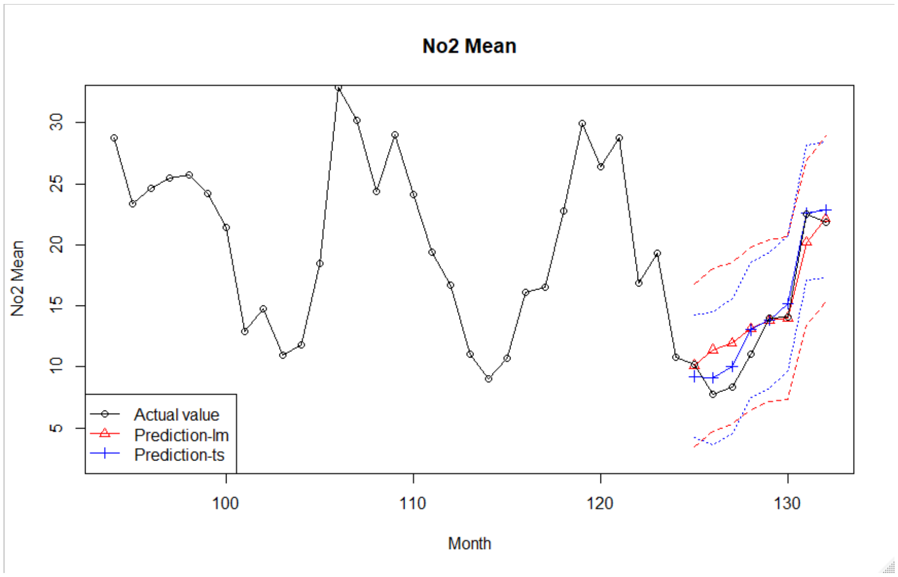
We started by doing a linear regression on the model. We found that the interaction of Co and O3 is significant while So2 is not, as observed in the plot.





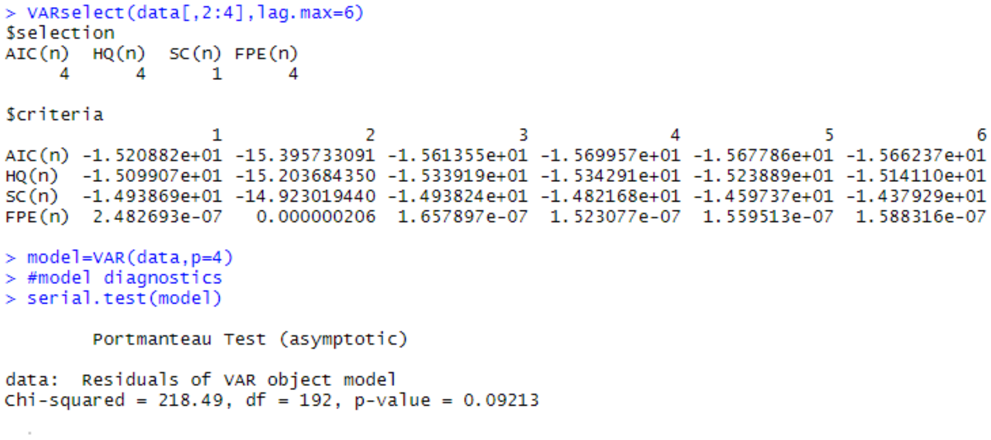
After getting the residual, we were able to do time-series analysis. We did not do seasonal transformation as the seasonality degraded towards 4 lags in ACF. We came up with 3 models. Tested the acf, pacf and eacf for white noise.

Later, aic was performed to finalize the model – ARMA(1,0,0)(1,0,0)x12. Using the finalized model, the prediction was made. The plot contains both the time series and the linear regression time series plots. We can notice that the time series plot out performed the linear regression model.



VAR MODELING

We wanted to venture further into VAR Modelling. When we tested for the lag that can fit the model, the lowest score was for lag 4. The Portmanteau test proved that the test will be a positive one. Since the process will be out of scope of this project, we stopped.



```
> VARselect(data[,2:4],lag.max=6)
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     4      4      1      4

$criteria
                    1            2             3             4             5             6
AIC(n) -1.520882e+01 -15.395733091 -1.561355e+01 -1.569957e+01 -1.567786e+01 -1.566237e+01
HQ(n)  -1.509907e+01 -15.203684350 -1.533919e+01 -1.534291e+01 -1.523889e+01 -1.514110e+01
SC(n)  -1.493869e+01 -14.923019440 -1.493824e+01 -1.482168e+01 -1.459737e+01 -1.437929e+01
FPE(n)  2.482693e-07   0.000000206  1.657897e-07  1.523077e-07  1.559513e-07  1.588316e-07

> model=VAR(data,p=4)
> #model diagnostics
> serial.test(model)

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object model
Chi-squared = 218.49, df = 192, p-value = 0.09213
```

FUTURE PROCESS AND RECOMMENATIONS

This was a small slice of the big process. We were able to understand time-series analysis better and we would like to venture further to

1. predict the output for the other 3 gases,
2. build a multivariate model based on the interaction of the different pollutants,
3. comparing the pollutant levels with the other cities.
4. This could be an interesting topic to venture into.

REFERENCES

1. https://data.world/data-society/us-air-pollution-data
2. Spatiotemporal variations of air pollutants (O3, NO2, SO2, CO, PM10, and VOCs) with land-use types J.-M. Yoo1 , M.-J. Jeong2 , D. Kim3 , W. R. Stockwell4 , J.-H. Yang5 , H.-W. Shin2 , M.-I. Lee6 , C.-K. Song7 , and S.-D. Lee7

THANK YOU