Air Pollution in Seoul

Daniel Chen

4/15/2020

Abstract

Air pollution is a growing problem in many countries. Many countries, especially in fast growing countries in Asia, have increasing air pollution problems due to the increasing urbanization and modernization of their societies. South Korea, for example, has one of the largest GDP in Eastern Asia and has have issues due to air quality in the region due to various factors. The Seoul Metroplitical Government in South Korea has several measurement stations to collect pollution data to help with their management of air quality. Using data collected over 24 hours for 3 years, we have granularity into the pollution patterns in Seoul. We are interested in building a forecasting model which might be able to predict pollution levels of various chemicals which could potentally be used to help with air pollution policies.

Introduction

One of the major causes of death in the world is due to air pollution. Air pollution comes from many sources as a mixture of solid particles and chemicals coming from source such as dust or car exhaust. These particles and chemicals can do long term damage to those that inhale them. Countries run ambient air quality monitoring in order to understand the extent of the pollution and to provide input into emission control strategies. Asia for example, has massive issues with air quality due to the population numbers and rapid industrialization. Seoul in South Korea for example, has some of the worst air quality in the region potentially due to many factors including geographical proximity to some countries and older coal power plants. As approximately 4.2 million people a year die due to these outdoor air pollution, understanding the trends of air pollution can be powerful in reducing pollution and lowering comorbidities due to pollution.

The data from these monitoring stations are time dependent measurements of various particles and chemicals. Carbon Monoxide(CO) typically comes from sources such as exhaust and can cause a variety of symptoms including headaches. Sulfur Dioxide(SO2) is gaseous and is generally formed from energy generation and is an irritant to the lungs. Nitrogen Dioxide(NO2) is also gaseous and comes from the burning of fuels such as from cars. This chemical is an irritant to the lungs and can potentially cause acid rain. Ozone(O3) is typically atmospeheric but can be generated on the ground through chemical plants and gasoline pumps. Prolonged exposure can cause skin cancer and many respiratory problems. Particulate matter 10 micrometer(PM10) or less and particulate matter 2.5 micrometer or less(PM2.5) are small particles which may be absorbed and cause various issues including fungal infection and cancer. These are all potententially harmful ambient air quality issues and need to be monitored.

Time series modeling is similar to traditional ordinary least squares linear modeling however, with differences stemming from the introduction of time. Time series data can move seasonally and in cycles. The data must show some pattern like with regular data otherwise the time series without a pattern will just be white noise. Vector autoregression(VAR) are a unique time series based model used to capture the linear interdependencies between predictors among multiple time series. The model is built upon the univariate autoregressive model in which it fits a linear model against two different time points of the same variable. The equation for the model is the following:

y_t =

$$\sum_{i=1}^{\infty} a_i * y_{t-i} + e_t$$

y_t = Current value of variable a_i = Parameter coefficient y_{t-i} = Value at lagged time period e_t = Error term

As these collected measurements should be correlated to each other, any time series model built upon this data should be built in the assumption that all of the variables interact with each other in some way is why the VAR model was used.

As air pollution is a massive global problem, building a forecasting model would be helpful in an advanced warning system and being able to better model public policy changes to pollution.

Results

First, we load all necessary packages for this analysis.

Dataset Summary

There are four datasets in this analysis. One is the summary and the others are elements used to help build the summary. The variable keys for the datasets can be seen below.

Variable Keys

Measurement Summary

Variables	Explaination
Measurement Date	Date of Measurement
Station code	Station code
Address	Address of monitoring station
Latitude	Latitude of station
Longitude	Longitude of station
SO2	Average Sulfur Dioxide(ppm)
NO2	Average Nitrogen Dioxide(ppm)
O3	Average Ozone(ppm)
CO	Average Carbon Monoxide(ppm)
PM10	Average Particulate Matter 10µg or less (µg/m^3)
PM2.5	Average Particulate Matter 2.5µg or less(µg/m^3)

Measurement Info

Variables	Explanation
Measurement Date	Measurement Date
Station Code	Station Code Primary key
Item Code	Item Code Primary Key
Average Value	Average value for given item code
Instrument Status	Status of instrument

Measurement Item Info

Variables	Explaination
Item code	Item code primary key
Item name	Measured item name
Unit of measurement	Unit of measurement
Good(blue)	Good value
Normal(green)	Normal value
Bad(yellow)	Bad value
Very bad(Red)	Very bad value

Measurement Station Info

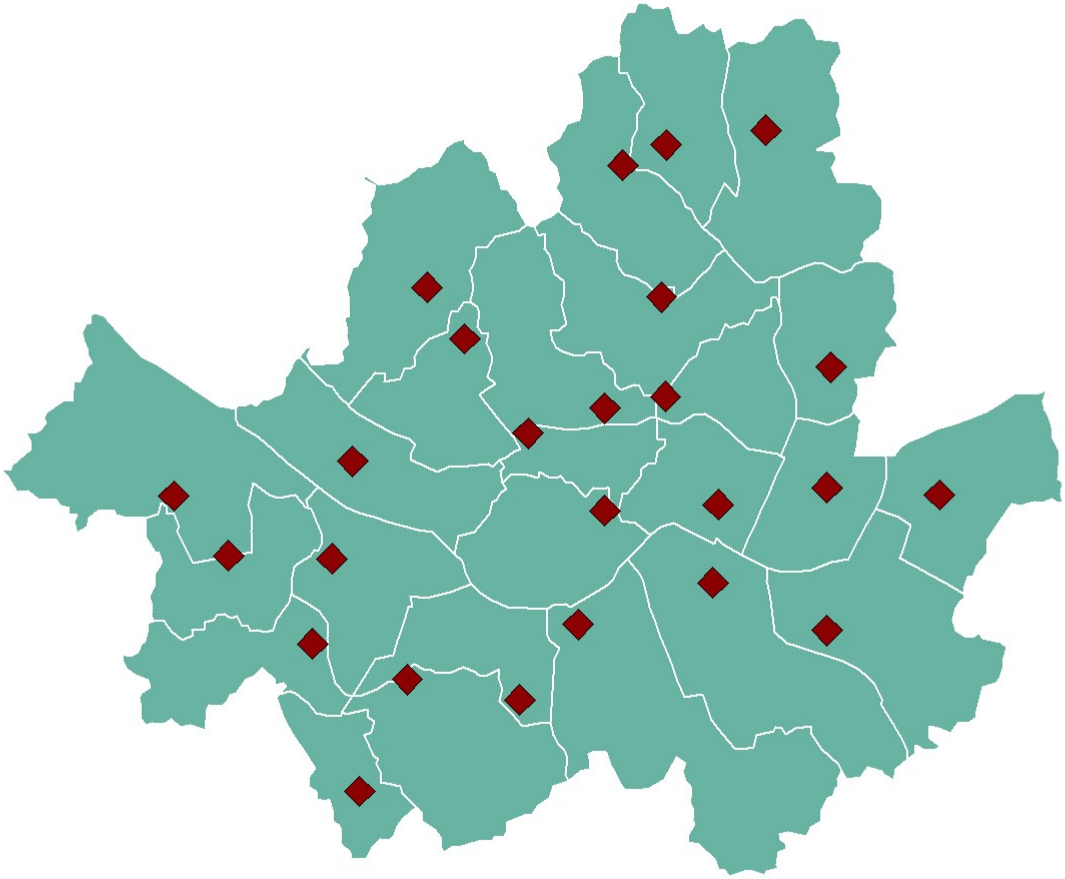
Variables	Explaination
Station code	Station code primary key
Station name(district)	District name for station
Address	Address of station
Latitude	Latitude
Longitude	Longitude

Exploratory Data Analysis

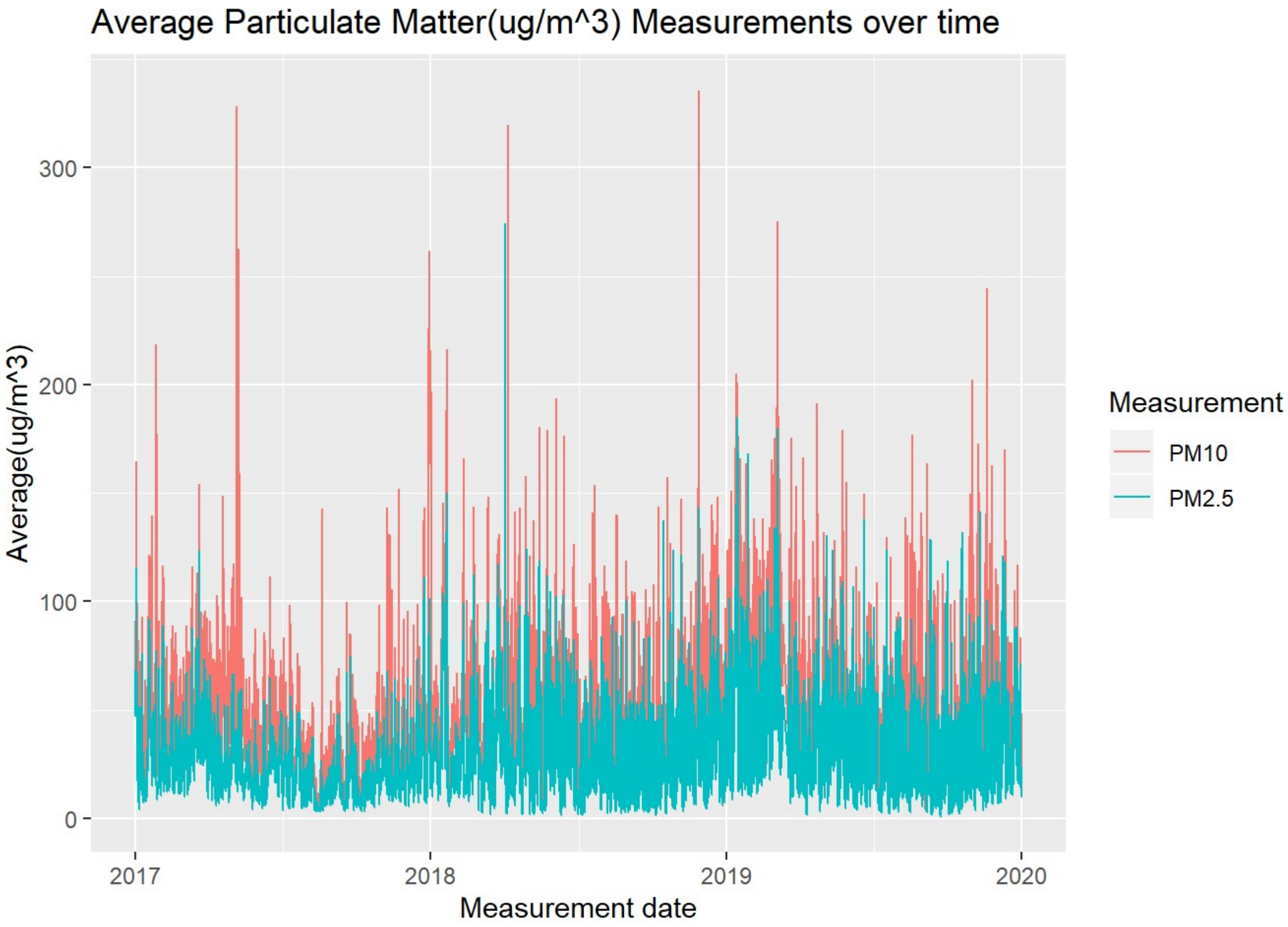
```
##  Measurement date               Station code
##  Min.   :2017-01-01 00:00:00   110    : 25906
##  1st Qu.:2017-09-27 19:00:00   116    : 25906
##  Median :2018-06-24 14:00:00   101    : 25905
##  Mean   :2018-06-27 21:39:49   102    : 25905
##  3rd Qu.:2019-03-30 15:30:00   106    : 25905
##  Max.   :2019-12-31 23:00:00   111    : 25905
##                                (Other):492079
##                                                            Address
##  369, Yongmasan-ro, Jungnang-gu, Seoul, Republic of Korea     : 25906
##  71, Gangseo-ro 45da-gil, Gangseo-gu, Seoul, Republic of Korea: 25906
##  10, Poeun-ro 6-gil, Mapo-gu, Seoul, Republic of Korea        : 25905
##  14, Sillimdong-gil, Gwanak-gu, Seoul, Republic of Korea      : 25905
##  15, Deoksugung-gil, Jung-gu, Seoul, Republic of Korea        : 25905
##  16, Sinbanpo-ro 15-gil, Seocho-gu, Seoul, Republic of Korea  : 25905
##  (Other)                                                      :492079
##     Latitude       Longitude           SO2                 NO2
##  Min.   :37.45   Min.   :126.8   Min.   :-1.000000   Min.   :-1.00000
##  1st Qu.:37.52   1st Qu.:126.9   1st Qu.: 0.003000   1st Qu.: 0.01600
##  Median :37.54   Median :127.0   Median : 0.004000   Median : 0.02500
##  Mean   :37.55   Mean   :127.0   Mean   :-0.001795   Mean   : 0.02252
##  3rd Qu.:37.58   3rd Qu.:127.0   3rd Qu.: 0.005000   3rd Qu.: 0.03800
##  Max.   :37.66   Max.   :127.1   Max.   : 3.736000   Max.   :38.44500
##
##        O3                 CO               PM10             PM2.5
##  Min.   :-1.00000   Min.   :-1.0000   Min.   :  -1.00   Min.   :  -1.00
##  1st Qu.: 0.00800   1st Qu.: 0.3000   1st Qu.:  22.00   1st Qu.:  11.00
##  Median : 0.02100   Median : 0.5000   Median :  35.00   Median :  19.00
##  Mean   : 0.01798   Mean   : 0.5092   Mean   :  43.71   Mean   :  25.41
##  3rd Qu.: 0.03400   3rd Qu.: 0.6000   3rd Qu.:  53.00   3rd Qu.:  31.00
##  Max.   :33.60000   Max.   :71.7000   Max.   :3586.00   Max.   :6256.00
##
```

We take a quick look at the data to understand how the summary is laid out. The data looks like a wide data format of various measurement values across 3 years from multiple in Korea. The data ranges from 2017 to the end of 2019.

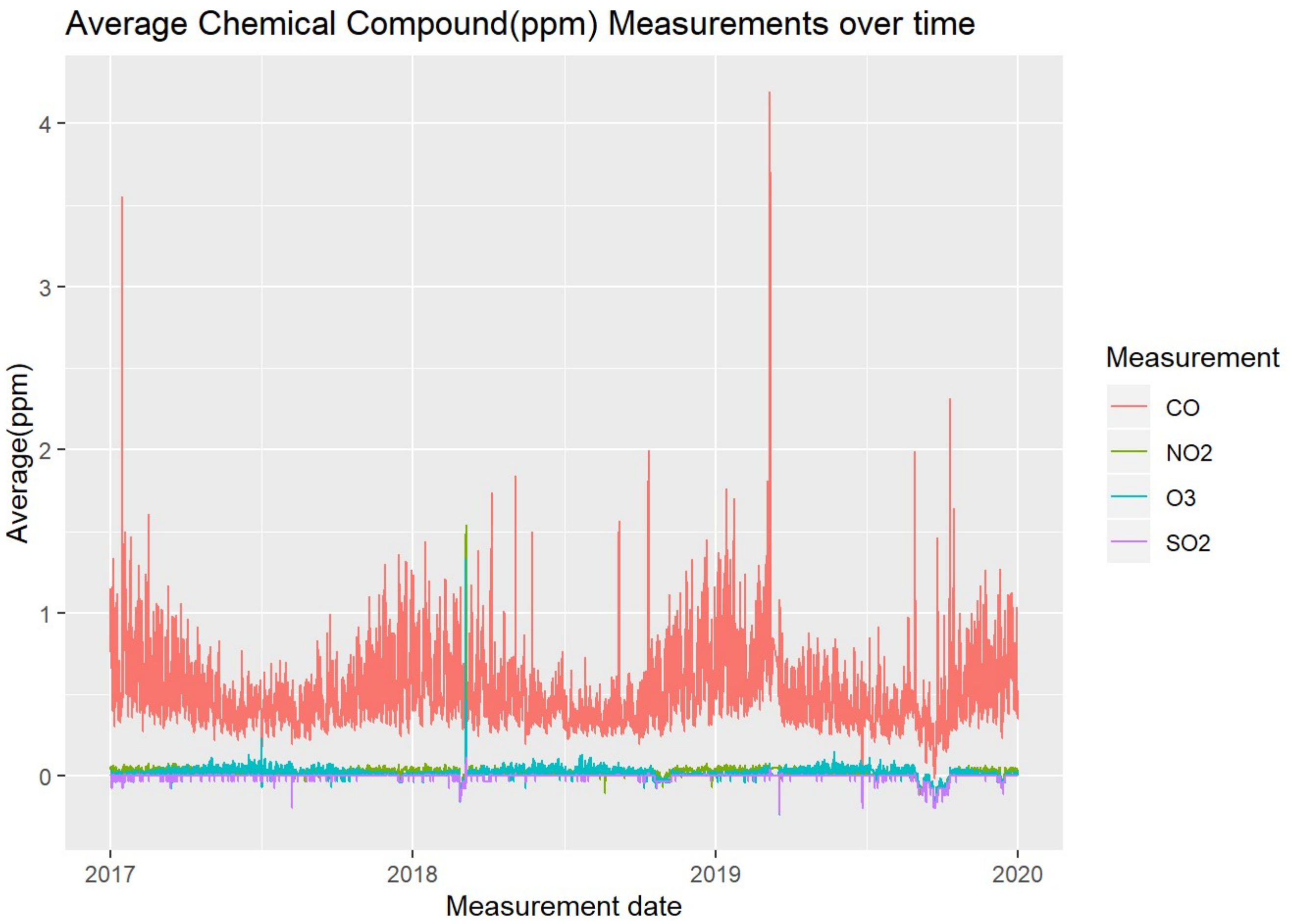
```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```



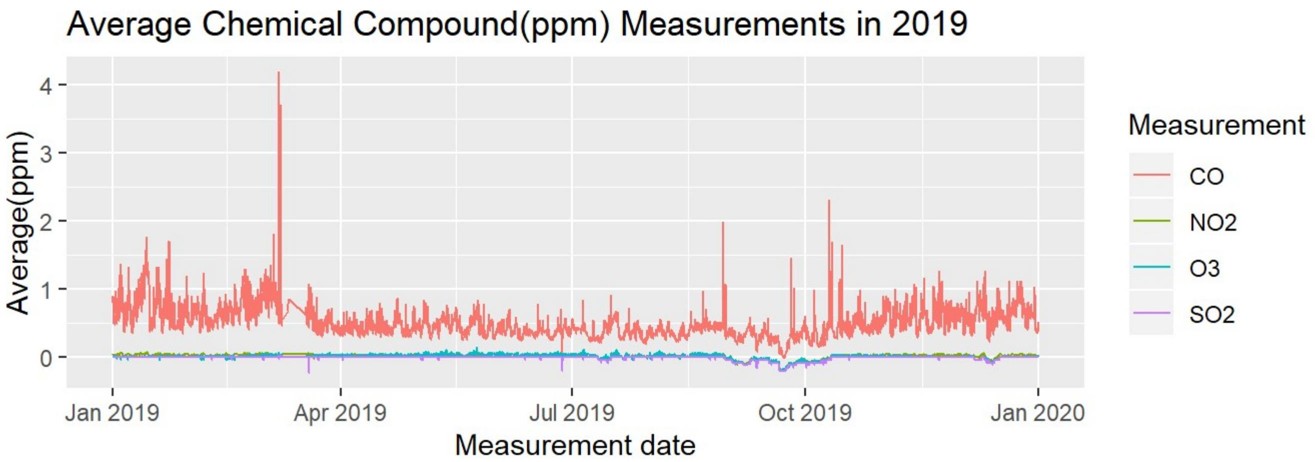
The figure above is a plot of Seoul with the lines representing the divisions between different municipalities. The red diamonds show the location of the 25 monitoring stations in this study.

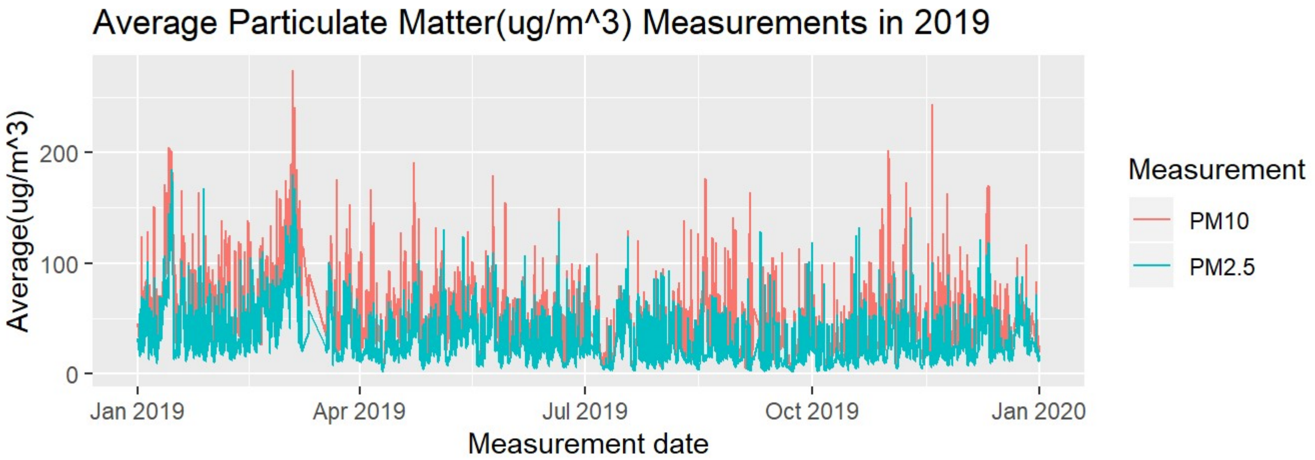


The plot above shows the average particulate matter value over the course of about three years. In general, it looks like PM10 has higher measured values which would make sense as these are larger particles which would be correlated with PM2.5 as it includes those measurements. .



We can see in this plot above that the highest measured values out of the chemicals is Carbon Monoxide. The other measurements have lower average ppm and deviations can be seen including what appear to be measurement errors as they are below 0. There seemed to have been some sort of sensor malfunction around Fall of 2019 and some major spikes especially in 2019.

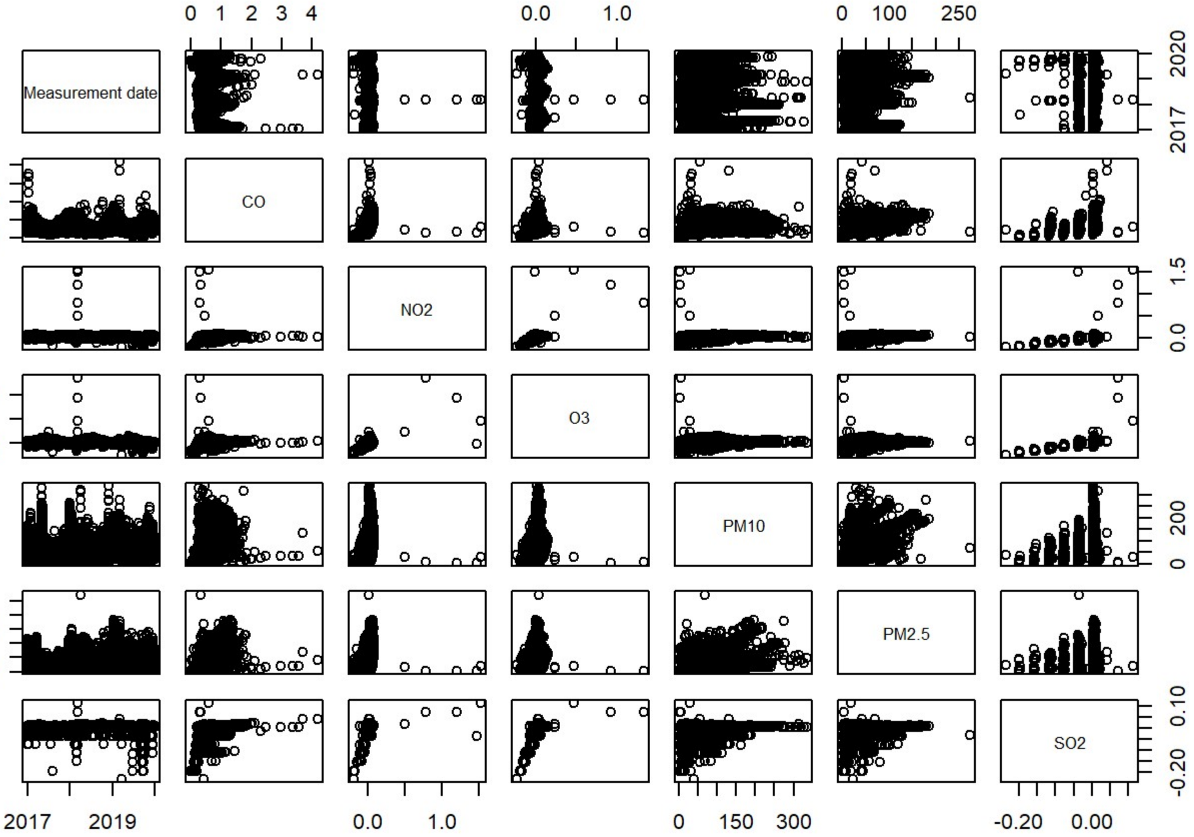




We try to examine 2019 a bit more closer. There was a cyclical event at around early March which caused a spike in polution levels in both the chemical level and particle level. The pollution levels look somewhat seasonal as it looks like CO levels are higher during the mid fall to early spring.

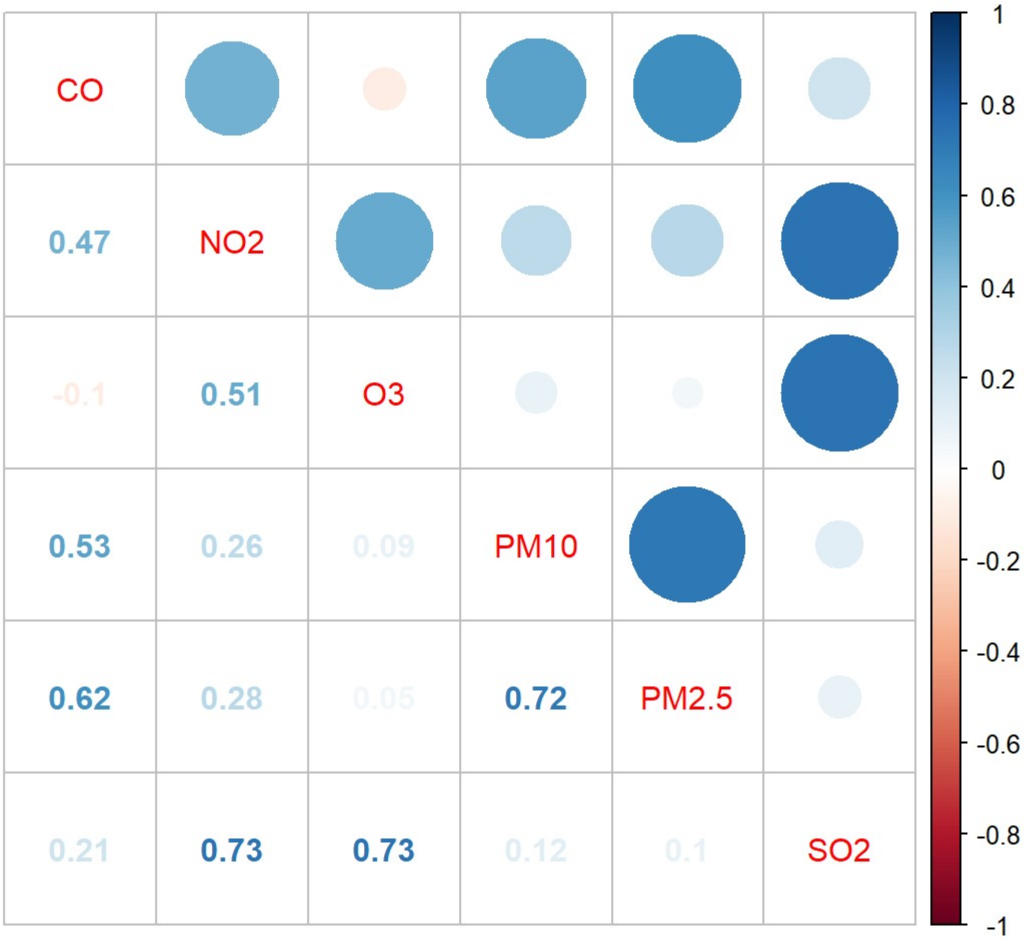
```
measurement_summary_wide <- spread(measurement_summary_long, Measurement, Average)
pairs(measurement_summary_wide %>% dplyr::select(-`Measurement date`))
```

```
## Adding missing grouping variables: `Measurement date`
```



We look at the pairwise plot fo the average measurements. It looks like in general that everything is positively correlated although some pairwise correlations do look more nonlinear in nature like with CO and NO2.

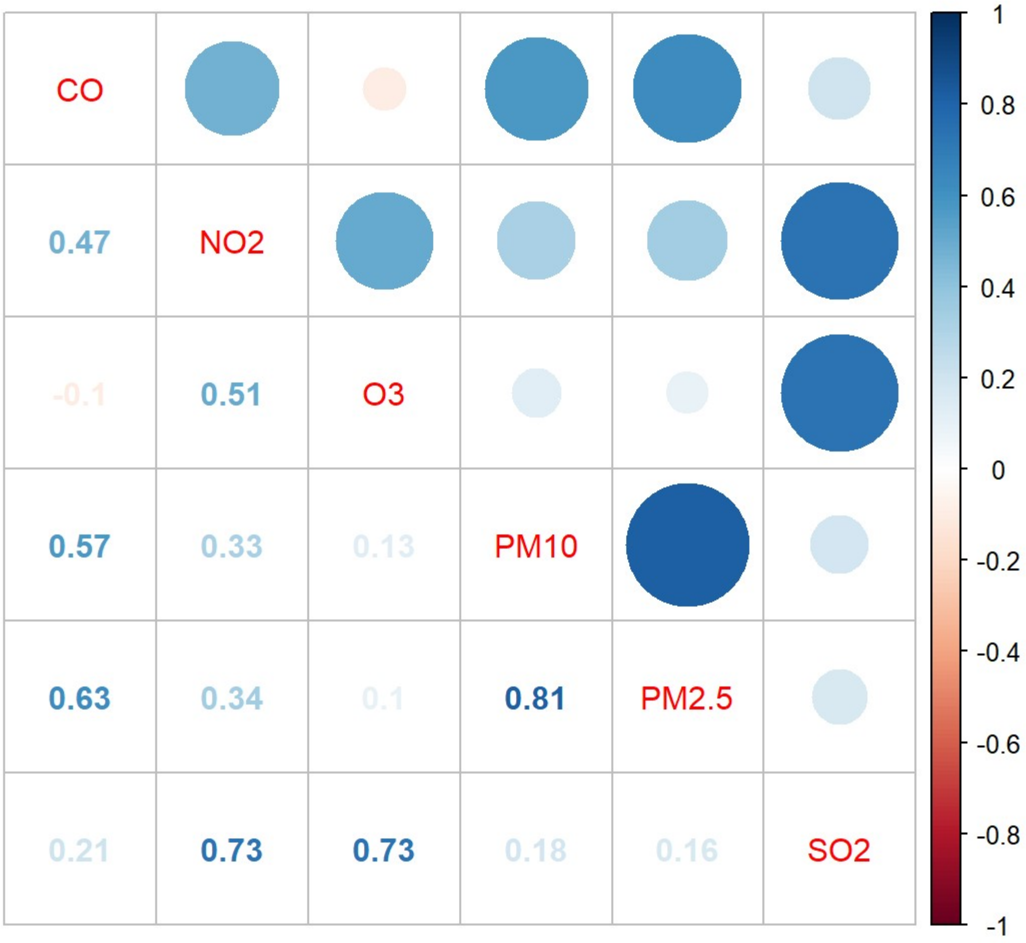
```
corrplot.mixed(measurement_summary_wide %>% ungroup %>% dplyr::select(-`Measurement d
ate`) %>% as.matrix %>% cor)
```



As can be seen in the aggregated correlation plot above, it looks like PM10, PM2.5, and CO are highly correlated with each other. With NO2, O3, and SO2, we can see that they are correlated together. There is a good chance of multicollinearity with this dataset especially as these models autoregress upon itself.

As seen in the plots above, the scale of the dataset distribution is not balanced especially evident by the PM values compared to the others. Log 10 transformation of the PM columns

```
measurement_summary_wide_log <- measurement_summary_wide %>% ungroup %>% dplyr::selec
t(-`Measurement date`) %>% mutate(PM2.5 = log10(PM2.5), PM10 = log10(PM10))
corrplot.mixed(measurement_summary_wide_log %>% as.matrix %>% cor)
```



With this log 10 transformation, we find that the correlation between the variables is even higher than before. Due to risk from even higher colinear effects, the untransformed data will be used.

We would like to double check now that this dataset is not stationary. What that means is that we are interested in finding out whether or not the data follows seasonal or cyclical trends. We are checking a lag time of 24 or 24 hours in this case. The data is tested using a Ljung-Box Q Test which checks to see if the time series has a non zero autocorrelation at each lag point.

```
for(i in c("CO", "NO2", "O3", "PM10", "PM2.5", "SO2")){
  print(Box.test(measurement_summary_wide[i] %>% pull, lag=24, type="Ljung-Box"))
}
```

```
## 
##  Box-Ljung test
## 
## data:  measurement_summary_wide[i] %>% pull
## X-squared = 251049, df = 24, p-value < 2.2e-16
## 
## 
##  Box-Ljung test
## 
## data:  measurement_summary_wide[i] %>% pull
## X-squared = 208835, df = 24, p-value < 2.2e-16
## 
## 
##  Box-Ljung test
## 
## data:  measurement_summary_wide[i] %>% pull
## X-squared = 236937, df = 24, p-value < 2.2e-16
## 
## 
##  Box-Ljung test
## 
## data:  measurement_summary_wide[i] %>% pull
## X-squared = 258130, df = 24, p-value < 2.2e-16
## 
## 
##  Box-Ljung test
## 
## data:  measurement_summary_wide[i] %>% pull
## X-squared = 235454, df = 24, p-value < 2.2e-16
## 
## 
##  Box-Ljung test
## 
## data:  measurement_summary_wide[i] %>% pull
## X-squared = 477877, df = 24, p-value < 2.2e-16
```

All of the measurement columns under a lag period of 24 hours is significant so all of the data columns do show some level of seasonaliity and cyclical behaviors.

Now that we understand the data, we would like to see what the forecasted values for all pollution metrics in the 10 months after the end of 2019. We first convert the data into a time series object rolling by months to better capture large scale changes.

```
# Quarterly
measurement_summary_ts <- ts(measurement_summary_wide %>% ungroup %>% dplyr::select(-
`Measurement date`), start = c(2017, 1, 1), end = c(2019, 12, 31), frequency = 12)
```

In order to optimze for the best lag time, we run a grid search to find the optimal time lag.

```
VARselect(measurement_summary_ts, type = "none", lag.max = 5)
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      5      5      5      5
##
## $criteria
##                    1             2             3             4
## AIC(n) -3.946501e+01 -4.117552e+01 -4.175289e+01 -4.882301e+01
## HQ(n)  -3.892218e+01 -4.008985e+01 -4.012437e+01 -4.665166e+01
## SC(n)  -3.779974e+01 -3.784497e+01 -3.675706e+01 -4.216191e+01
## FPE(n)  7.472199e-18  1.692849e-18  1.987995e-18  1.359217e-20
##                    5
## AIC(n) -2.261386e+02
## HQ(n)  -2.234244e+02
## SC(n)  -2.178122e+02
## FPE(n)  2.869666e-93
```

According to the criteria, we find that the optimal lag period is 1 month in this case.

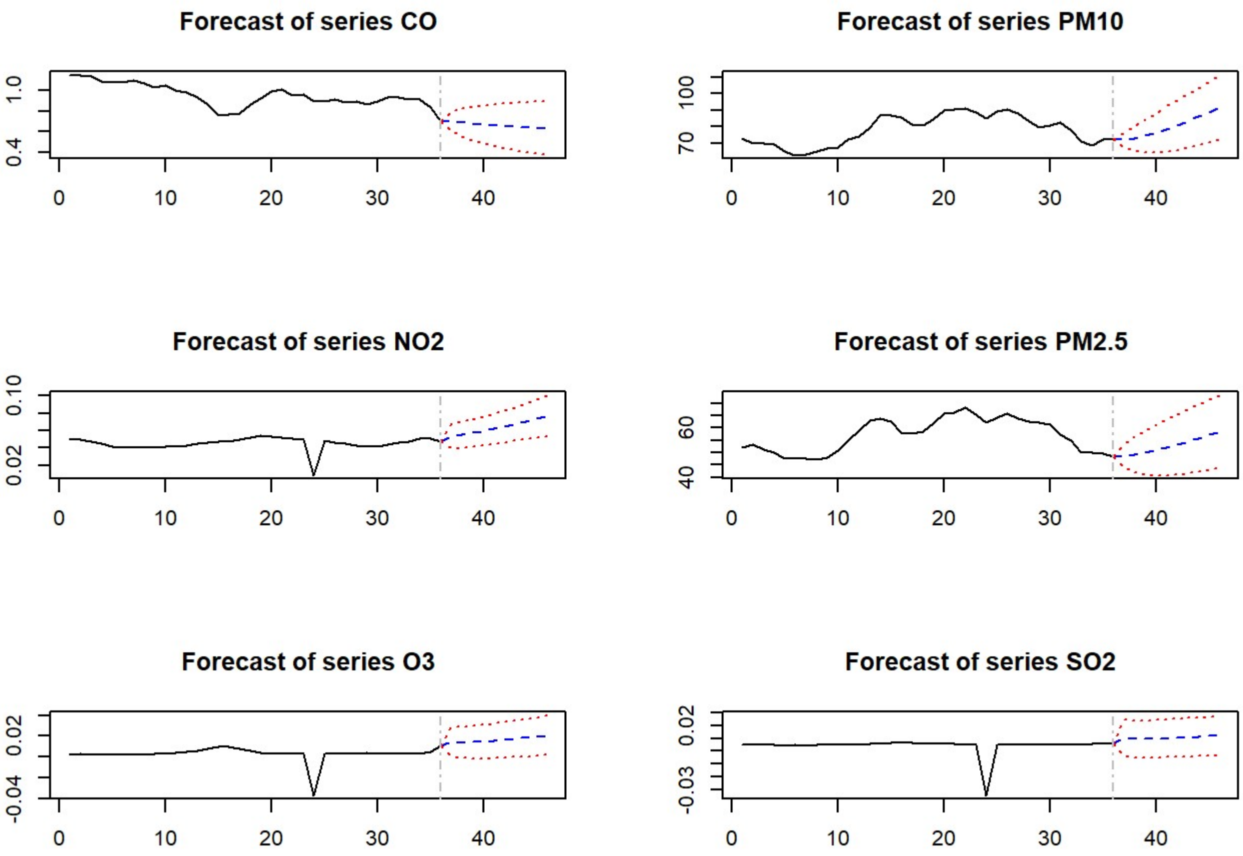
We then run the Vector Autoregression model with a lag of 1 month.

```
# Get the estimated coefficients
VAR_est <- VAR(measurement_summary_ts, p = 1, type = "none")
summary(VAR_est)
```

```
##
## VAR Estimation Results:
## =========================
## Endogenous variables: CO, NO2, O3, PM10, PM2.5, SO2
## Deterministic variables: none
## Sample size: 35
## Log Likelihood: 432.251
## Roots of the characteristic polynomial:
## 1.074 0.9894 0.8906 0.7689 0.3854 0.08213
## Call:
## VAR(y = measurement_summary_ts, p = 1, type = "none")
##
##
## Estimation results for equation CO:
## ===================================
## CO = CO.l1 + NO2.l1 + O3.l1 + PM10.l1 + PM2.5.l1 + SO2.l1
##
##           Estimate Std. Error t value Pr(>|t|)
## CO.l1     0.965675   0.072952  13.237 8.06e-14 ***
## NO2.l1   -0.694044   2.747199  -0.253    0.802
## O3.l1    -0.315907   6.647447  -0.048    0.962
## PM10.l1   0.001557   0.003711   0.420    0.678
## PM2.5.l1 -0.001267   0.004552  -0.278    0.783
## SO2.l1    0.696399   7.627269   0.091    0.928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 0.05005 on 29 degrees of freedom
## Multiple R-Squared: 0.9977,  Adjusted R-squared: 0.9972
## F-statistic:  2090 on 6 and 29 DF,  p-value: < 2.2e-16
##
##
## Estimation results for equation NO2:
## ====================================
## NO2 = CO.l1 + NO2.l1 + O3.l1 + PM10.l1 + PM2.5.l1 + SO2.l1
##
##            Estimate Std. Error t value Pr(>|t|)
## CO.l1     0.0138671  0.0110794   1.252   0.2207
## NO2.l1    0.9060597  0.4172260   2.172   0.0382 *
## O3.l1     1.8299242  1.0095692   1.813   0.0803 .
## PM10.l1   0.0001756  0.0005635   0.312   0.7576
## PM2.5.l1 -0.0002822  0.0006913  -0.408   0.6861
## SO2.l1   -2.8428295  1.1583780  -2.454   0.0204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 0.007601 on 29 degrees of freedom
## Multiple R-Squared: 0.977,   Adjusted R-squared: 0.9722
## F-statistic: 205.4 on 6 and 29 DF,  p-value: < 2.2e-16
##
##
## Estimation results for equation O3:
## ===================================
## O3 = CO.l1 + NO2.l1 + O3.l1 + PM10.l1 + PM2.5.l1 + SO2.l1
##
##            Estimate Std. Error t value Pr(>|t|)
## CO.l1     0.0115403  0.0107694   1.072   0.2927
## NO2.l1    0.0362098  0.4055485   0.089   0.9295
## O3.l1     2.2215035  0.9813129   2.264   0.0312 *
## PM10.l1   0.0001142  0.0005478   0.208   0.8363
## PM2.5.l1 -0.0002636  0.0006720  -0.392   0.6977
## SO2.l1   -2.3387490  1.1259568  -2.077   0.0468 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 0.007388 on 29 degrees of freedom
## Multiple R-Squared: 0.2648,  Adjusted R-squared: 0.1127
## F-statistic: 1.741 on 6 and 29 DF,  p-value: 0.1469
##
##
## Estimation results for equation PM10:
## =====================================
## PM10 = CO.l1 + NO2.l1 + O3.l1 + PM10.l1 + PM2.5.l1 + SO2.l1
##
##            Estimate Std. Error t value Pr(>|t|)
## CO.l1    -2.183e+00  3.489e+00  -0.626  0.53645
## NO2.l1    3.762e+02  1.314e+02   2.863  0.00771 **
## O3.l1     6.914e+02  3.179e+02   2.175  0.03792 *
## PM10.l1   2.907e-03  1.774e-01   0.016  0.98704
## PM2.5.l1  1.156e+00  2.177e-01   5.310 1.07e-05 ***
## SO2.l1   -1.206e+03  3.647e+02  -3.307  0.00252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 2.393 on 29 degrees of freedom
## Multiple R-Squared: 0.9992,  Adjusted R-squared: 0.9991
## F-statistic:  6307 on 6 and 29 DF,  p-value: < 2.2e-16
##
##
## Estimation results for equation PM2.5:
## ======================================
## PM2.5 = CO.l1 + NO2.l1 + O3.l1 + PM10.l1 + PM2.5.l1 + SO2.l1
##
##           Estimate Std. Error t value Pr(>|t|)
## CO.l1       1.5639     3.5528   0.440    0.663
## NO2.l1    146.3080   133.7891   1.094    0.283
## O3.l1     316.8386   323.7318   0.979    0.336
## PM10.l1    -0.2969     0.1807  -1.643    0.111
## PM2.5.l1    1.2868     0.2217   5.805 2.72e-06 ***
## SO2.l1   -546.6657   371.4493  -1.472    0.152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 2.437 on 29 degrees of freedom
## Multiple R-Squared: 0.9985,  Adjusted R-squared: 0.9982
## F-statistic:  3247 on 6 and 29 DF,  p-value: < 2.2e-16
##
##
## Estimation results for equation SO2:
## ====================================
## SO2 = CO.l1 + NO2.l1 + O3.l1 + PM10.l1 + PM2.5.l1 + SO2.l1
##
##            Estimate Std. Error t value Pr(>|t|)
## CO.l1     0.0121015  0.0104578   1.157    0.257
## NO2.l1   -0.0293895  0.3938180  -0.075    0.941
## O3.l1     1.3015873  0.9529283   1.366    0.182
## PM10.l1   0.0001509  0.0005319   0.284    0.779
## PM2.5.l1 -0.0002774  0.0006525  -0.425    0.674
## SO2.l1   -1.3572982  1.0933883  -1.241    0.224
##
##
## Residual standard error: 0.007174 on 29 degrees of freedom
## Multiple R-Squared: 0.3149,  Adjusted R-squared: 0.1732
## F-statistic: 2.222 on 6 and 29 DF,  p-value: 0.06937
##
##
##
## Covariance matrix of residuals:
##              CO       NO2        O3     PM10    PM2.5       SO2
## CO    2.505e-03 1.046e-04 2.007e-05 0.002100 0.023912 6.719e-05
## NO2   1.046e-04 5.777e-05 5.307e-05 0.005552 0.004514 5.327e-05
## O3    2.007e-05 5.307e-05 5.458e-05 0.004957 0.003478 5.225e-05
## PM10  2.100e-03 5.552e-03 4.957e-03 5.727541 3.617519 4.702e-03
## PM2.5 2.391e-02 4.514e-03 3.478e-03 3.617519 5.940474 3.755e-03
## SO2   6.719e-05 5.327e-05 5.225e-05 0.004702 0.003755 5.147e-05
##
## Correlation matrix of residuals:
##            CO    NO2      O3    PM10  PM2.5    SO2
## CO    1.00000 0.2749 0.05429 0.01753 0.1960 0.1871
## NO2   0.27487 1.0000 0.94514 0.30523 0.2437 0.9770
## O3    0.05429 0.9451 1.00000 0.28037 0.1932 0.9859
## PM10  0.01753 0.3052 0.28037 1.00000 0.6202 0.2739
## PM2.5 0.19603 0.2437 0.19316 0.62018 1.0000 0.2148
## SO2   0.18712 0.9770 0.98591 0.27387 0.2148 1.0000
```

We get back a variety of linear models for each each predictor. We use a lag of 1 to get the difference between seasons and see some terms are significant in various predictors. For instance, PM10 has some relationship with PM2.5 which is expected. It looks like some of the models have overfitting issues as their adjusted r^2 is greater than 95%. SO2 and O3 have very poor performance which makes sense due to how highly correlated they are. Furthermore, their F statistic tells us that it is not useful in forecasting with all of the other variables in this case. It is also interesting to see that much of the significance in each model is due to auto regression with the exception of PM10 which is has significance relationship with many other factors. This makes sense as you would expect that larger particles would have highest association with a derived variable(PM2.5) and other chemicals on heavy pollution days. Now, we plot the forecasted pollution values up to 10 months away from the end of 2019.

```
plot(predict(VAR_est))
```



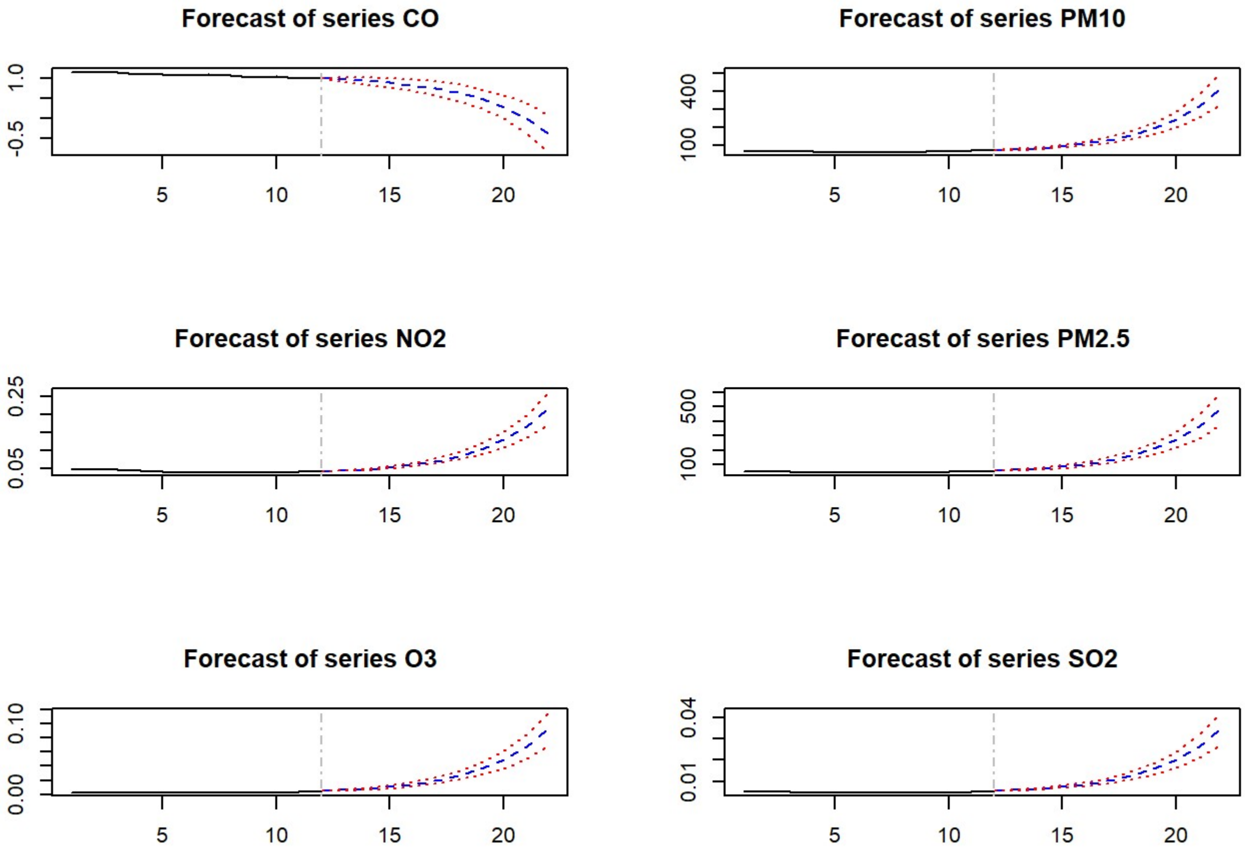
In the series of forecasted plots, we can see the predicted trend in blue and the 95% confidence interval in red. We can see two trends from the forecasts. PM10 and PM2.5 both are gradually increasing in each month since 2016 and is expected to increase again in the future. It makes sense as most of the data shows it increasing and then fluctuating around 60. Most of the other chemicals measured will increase again in the future. For instance, NO2, O3, and SO2. The sensor seeming to malfunction can be seen in the sharp drop in the NO2 and O3 and SO3 forecast plots. For some of the plots, we can see the confidence interval widening over time as it becomes more unconfident in long term predictions. We can also try quarterly to see if there are any differences.

```
measurement_summary_quarter_ts <- ts(measurement_summary_wide %>% ungroup %>% dplyr::
select(-`Measurement date`), start = c(2017, 1, 1), end = c(2017, 12), frequency = 4)
var.aic <- VARselect(measurement_summary_quarter_ts, type = "none", lag.max = 3)
var.aic
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n) 
##      2      2      2      2 
## 
## $criteria
##                     1    2    3
## AIC(n) -1.607762e+02 -Inf -Inf
## HQ(n)  -1.624787e+02 -Inf -Inf
## SC(n)  -1.599873e+02 -Inf -Inf
## FPE(n)  7.856685e-70    0    0
```

Doing this optimization does indeed to show that a lag of 1 is the best model using all criterions.

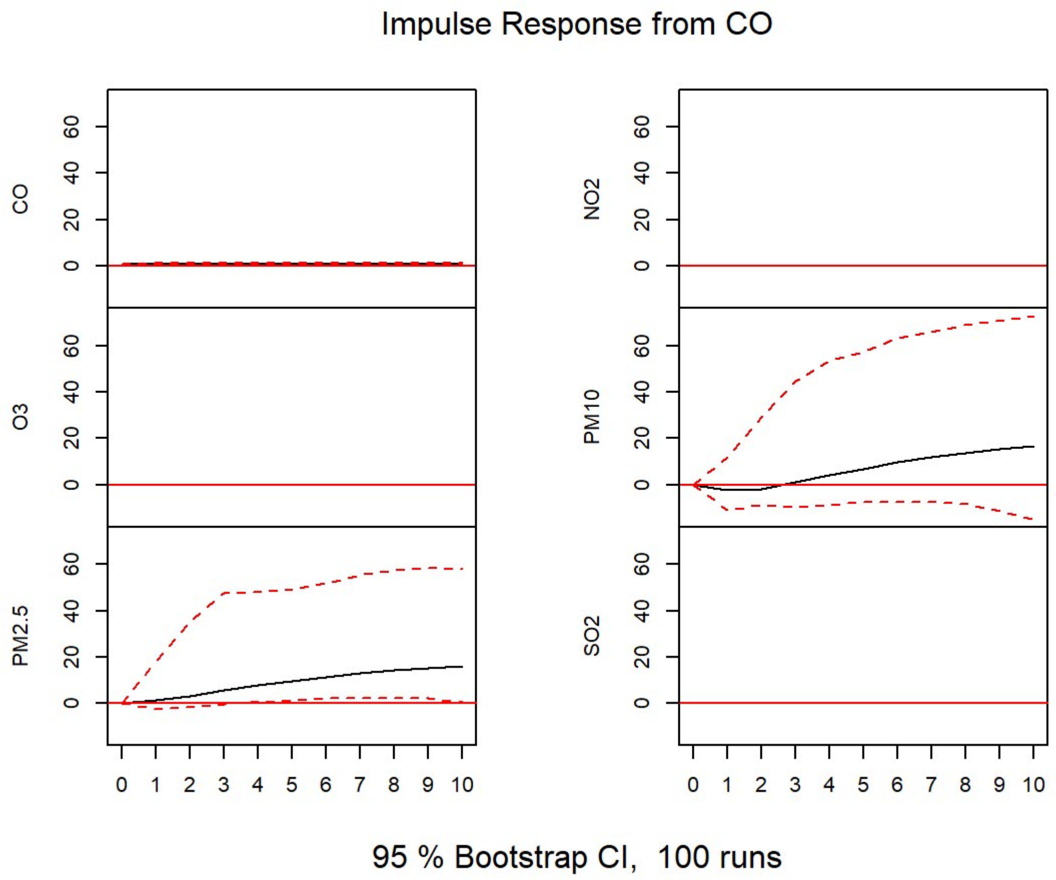
```
VAR_est2 <- VAR(measurement_summary_quarter_ts, p = 1, type = "none")
plot(predict(VAR_est2))
```

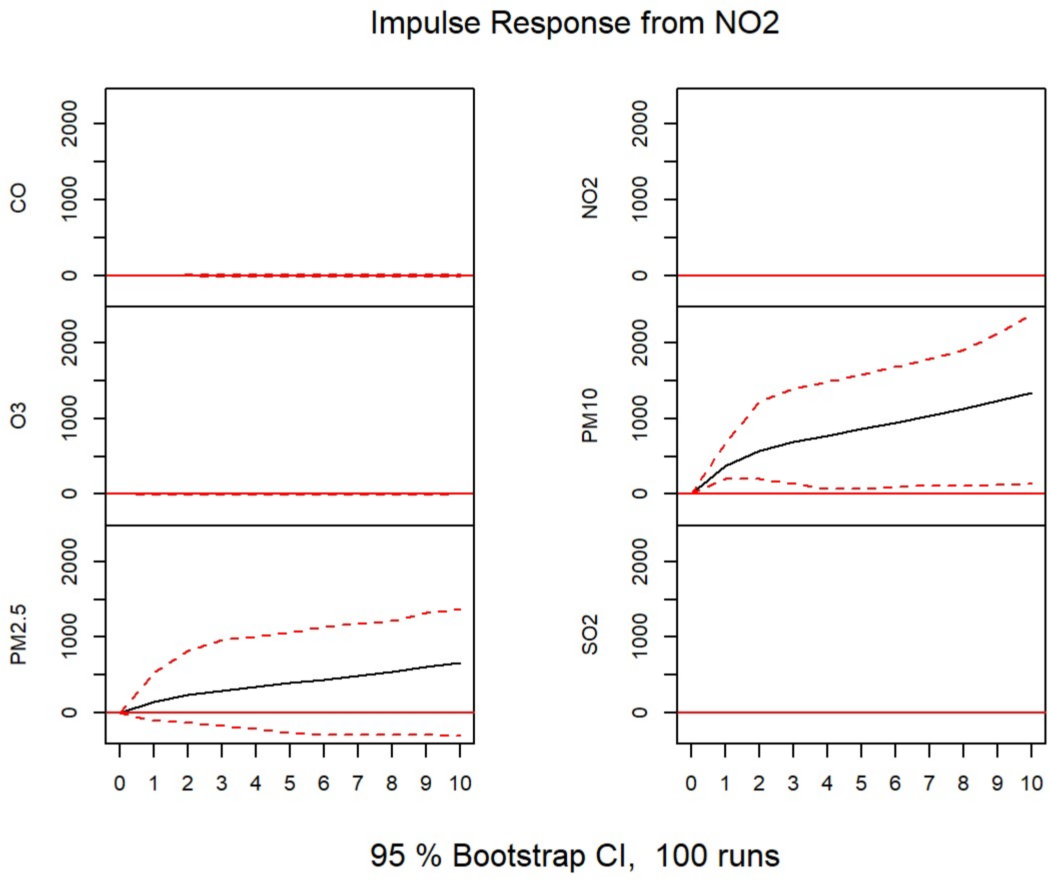


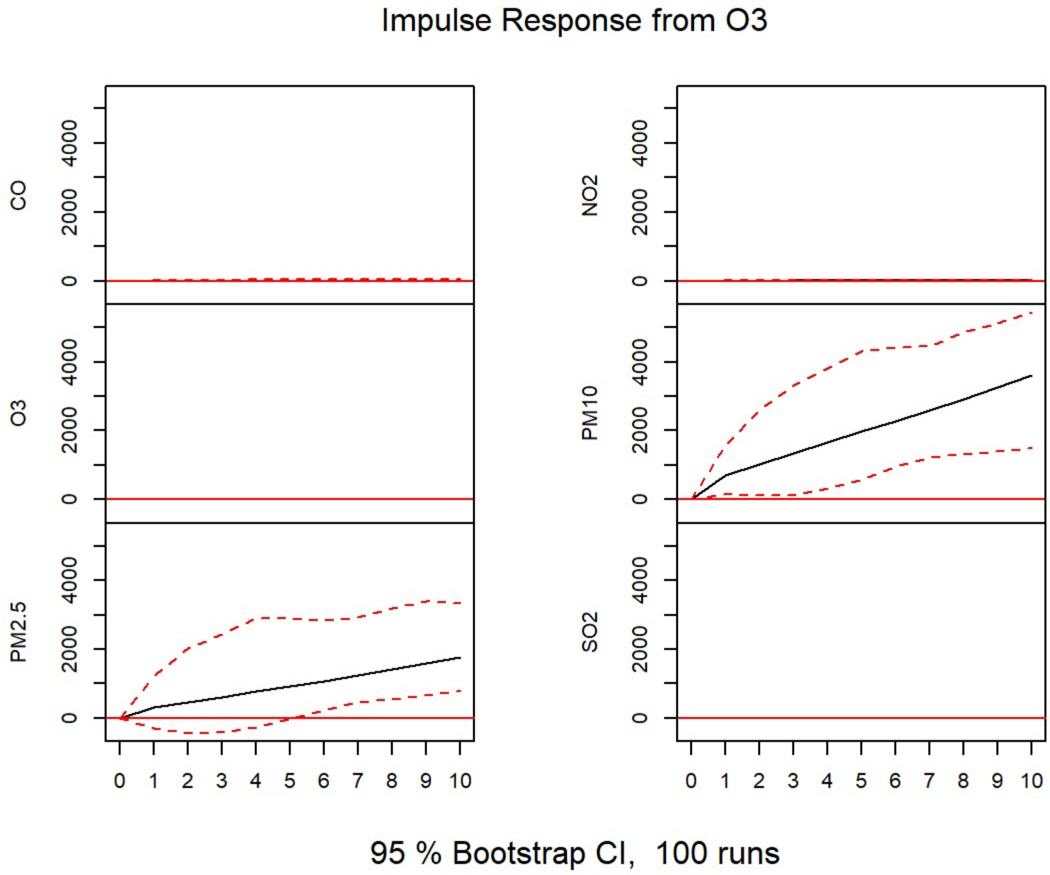
When we try doing the same thing with quartering averages, we can see major issues with this version of the model due to the statistical power of the samples due to 12 points. Because all of the trends in the 12 points are either moving up or down, the forecasted value expotentially moves in one direction or another.

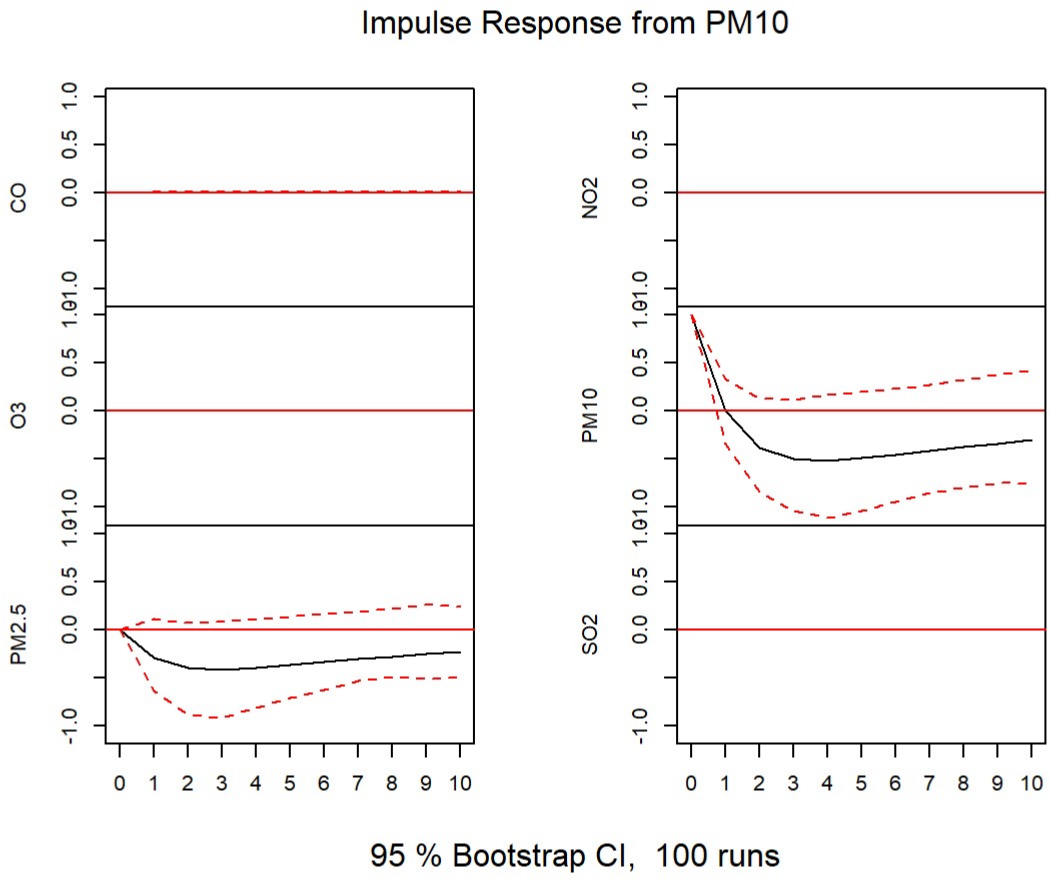
The prior models have little insight into the actual relationship between variables. In order to understand more about it, we can run a error impulse reponse in which we model the response in relationship to a given variable by giving it shocks via stoichastic error. We can look using the monthly moving average and look 10 months ahead.

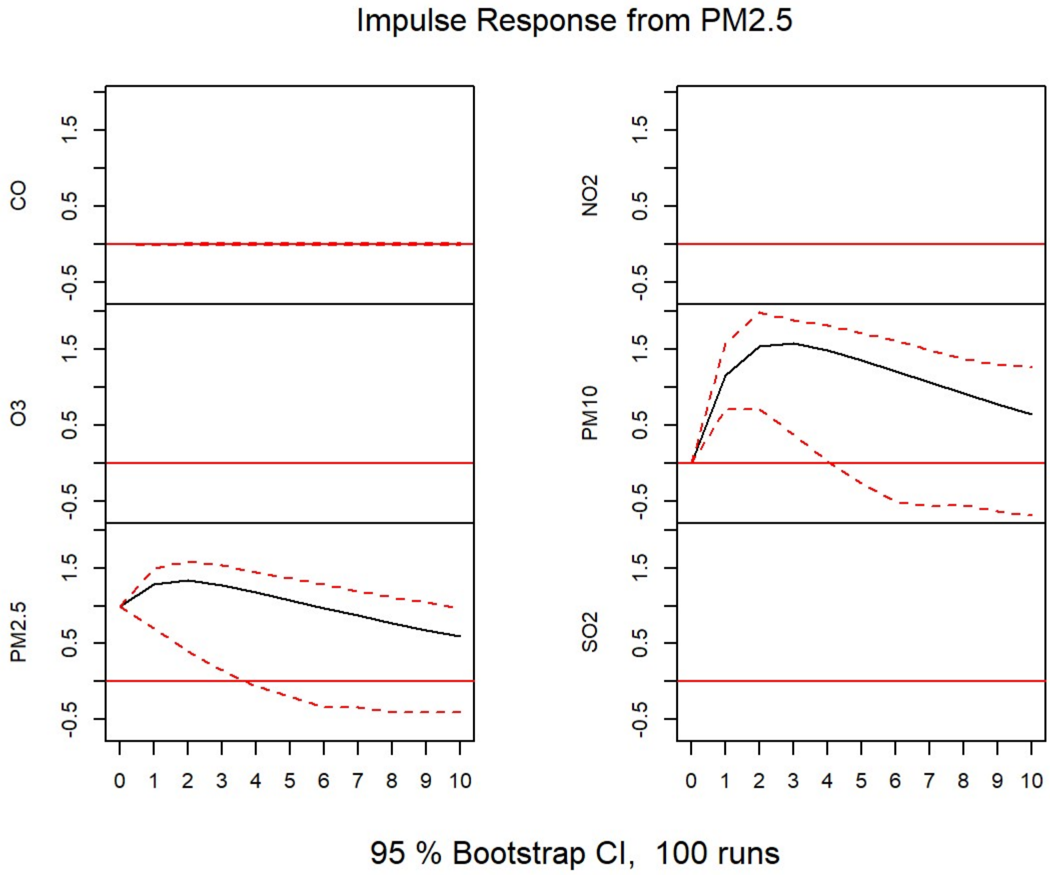
```
for(i in c(names(measurement_summary_wide)[-1])){
  plot(irf(VAR_est, impulse = i, n.ahead = 10, ortho = FALSE))
}
```

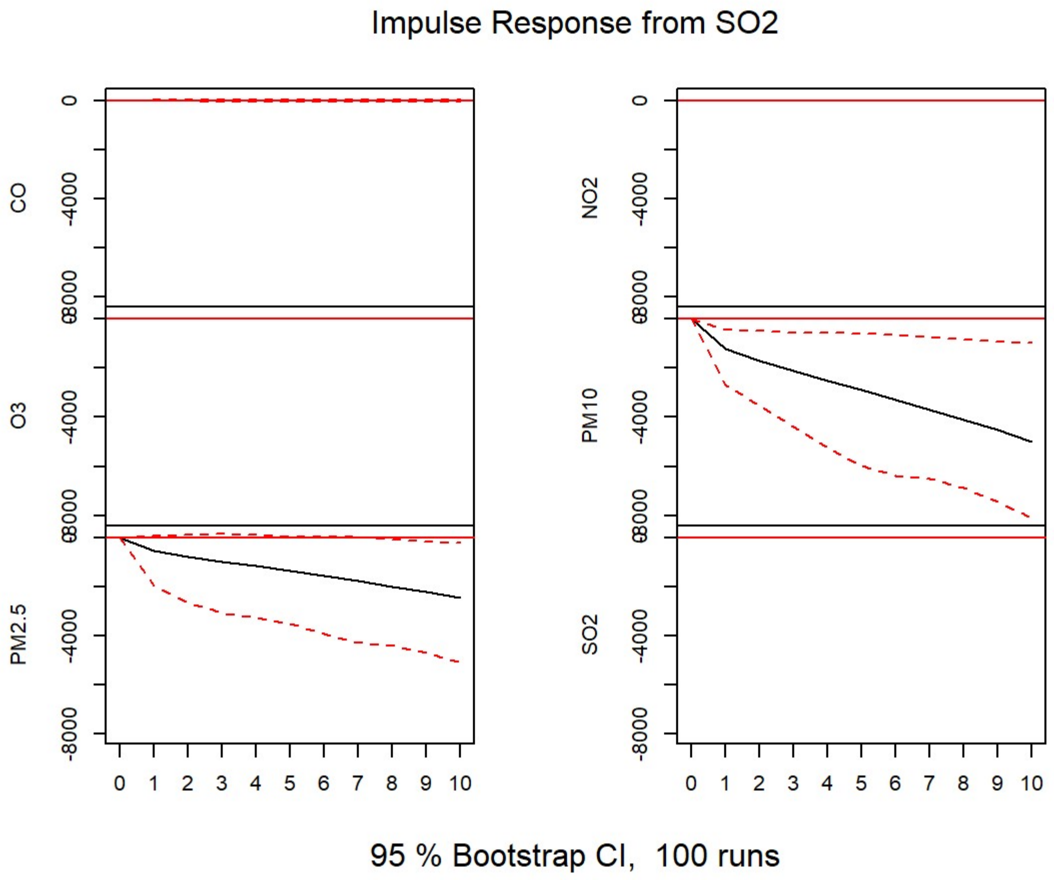



Impulse Response from NO2
CO
NO2
O3
PM10
PM2.5
SO2
0
1000
2000
0 1 2 3 4 5 6 7 8 9 10
95 % Bootstrap CI, 100 runs


Impulse Response from O3
CO
NO2
O3
PM10
PM2.5
SO2
0
2000
4000
0 1 2 3 4 5 6 7 8 9 10
95 % Bootstrap CI, 100 runs


Impulse Response from PM10
CO
O3
PM2.5
NO2
PM10
SO2
1.0
0.5
0.0
-1.0
0 1 2 3 4 5 6 7 8 9 10
95 % Bootstrap CI, 100 runs


Impulse Response from PM2.5
CO
O3
PM2.5
NO2
PM10
SO2
1.5
0.5
-0.5
0 1 2 3 4 5 6 7 8 9 10
95 % Bootstrap CI, 100 runs



There are mulitple plots above showing the effects of variation from a given variable will have on other variables over the course of 10 months. The black line shows the predicted values and the red are the 95% confidence interval. The x axis represents the number of predicted quarters and the y axis represent standard unit changes. Note that the confidence interval bands in every plot overlap 0 so it's hard to say where the true effect will actually be.

In the response to changes in CO, we can see that every chemical and gas remains constant but the particulate matter will gradually increase. For NO2, we get a similar pattern with shocks to NO2 however, particulate matter quickly increases and then gradually increase. With shocks to O3, we see almost linear increase. With shocks to PM10, we will expect both PM10 and PM2.5 to decrease exponentially until stabilizing. Inversely, shocks to PM2.5 will increase PM10 and PM2.5 expotentially until steady decrease. Finally, shocks to SO2 significantly decrease PM10 and PM2.5.

Discussion

VAR models were chosen because all of these variables are related to one another. They are all pollutants that are commonly measured and should be considered as a whole instead of using a univariate model like ARIMA. The does not appear stationary according to the Box Ljung test which is statistically significant and tells us that this time seris data shows some cyclical or seasonal trend.

We fit two different VAR models in order to forecast future pollution levels. The first model we fit used monthly moving averages. The second model we fit using the quartly moving averages. In both models, we optimize for the best time lag using a grid search using AIC. We find in that model that some of the estimated terms to have very poor fits. However, when we try doing the same thing but with quarterly data, we run into statistical under power issues due to sample size. Although this data does appear to show collinearity, the data was not normalized due to even higher correlation between variables.

Based on the forecasted data using the monthly model, we see increasing trends amongst all pollutants except for carbon monoxide. One of the limitations of the forecasting is that it does not show causation. Although there are signs of collinearity, correlation is still not causation. We can try to show some response and relationship by using the Impulse Reponse to see the effects that adding noise to a variable will have in relationship to the other variables. When we shock most of the chemicals, we see that it tended to have an increase in particle emissions. Major changes to the particle emissions themselves tended to actually have future decrease in particle emissions.

It is always important to note that these are predicted values based on prior data. Although this model could undergoing cross validation to find test and training error, it is also important to note that this 10 month prediction has potentially been impacted due to virus issues. It would be interesting to see if any of these trends would hold given the COVID-19 effect on travel.

Limitations

This analysis is limited only to forcasting trends in pollution metrics that were collected. This has nothing to do with causal relationships and is simply forecasting future pollution levels given prior changes. The time periods collected also make analysis potentially very noisy as forecasting using hourly measures is possible but computationally taxing while higher levels such as quarterly or biannually is too sparse. Exploring day average and especially hourly average would be interesting models as one could expect higher pollutants when people are typically awake. There is also potentially location based biases which would be better to explore as Seoul is separated by the Han river and the districts can highly vary from high rises to smaller older homes. Although VAR is similar to a multivariate ARIMA model, it might be better to separate out the poor performing endogenous variables and forecast them using ARIMA separately. Furthermore, the analysis can change completely depending on how the time series is sliced. Finally, this forecast is based on prior results. So when black swan events like pandemics happen, forecasts will be completely wrong.

Conclusion

A vector auto regressing model was built in order to forecast future levels of pollution in the city of Seoul. We find that the data from the monitoring station to show non-stationary patterns and are able to use the data. Our monthly forecasting model finds that most polluants are expected to increase from January 2020 to October 2020. When we Impulse Response to see the effect of noise to the data, we find that it generally leads to an increase in particular matter.

Acknowledgements

I would like to acknowledge Prof. Parzen and the teaching assitants of Stat 109 for teaching the course. Additionally, I would like to acknowledge bappe, Kaggle, and the Korean government for releasing this dataset.

References

16.1 Vector Autoregressions. https://www.econometrics-with-r.org/16-1-vector-autoregressions.html (https://www.econometrics-with-r.org/16-1-vector-autoregressions.html), May 2020
An Introduction to Vector Autoregression (VAR). https://www.r-econometrics.com/timeseries/varintro/

(https://www.r-econometrics.com/timeseries/varintro/), May 2020
Hu, Elise. Korea’s Air is Dirty, But It’s Not All Close-Neighbor China’s Fault. https://www.npr.org/sections/parallels/2016/06/03/478796463/koreas-air-is-dirty-but-its-not-all-close-neighbor-chinas-fault (https://www.npr.org/sections/parallels/2016/06/03/478796463/koreas-air-is-dirty-but-its-not-all-close-neighbor-chinas-fault), May 2020.
Air Pollution in Seoul. https://www.kaggle.com/bappekim/air-pollution-in-seoul (https://www.kaggle.com/bappekim/air-pollution-in-seoul), May 2020.
Managing Air Quality - Ambient Air Monitoring. https://www.epa.gov/air-quality-management-process/managing-air-quality-ambient-air-monitoring (https://www.epa.gov/air-quality-management-process/managing-air-quality-ambient-air-monitoring), May 2020.
Huzar, Timothy. Air pollution may be a leading global cause of death. https://www.medicalnewstoday.com/articles/air-pollution-may-be-a-leading-global-cause-of-death (https://www.medicalnewstoday.com/articles/air-pollution-may-be-a-leading-global-cause-of-death), May 2020.
Carbon Monoxide Posioning. https://www.health.harvard.edu/a_to_z/carbon-monoxide-poisoning-a-to-z (https://www.health.harvard.edu/a_to_z/carbon-monoxide-poisoning-a-to-z), May 2020.
Sulfure Dioxide(SO2). https://www.environment.gov.au/protection/publications/factsheet-sulfur-dioxide-so2 (https://www.environment.gov.au/protection/publications/factsheet-sulfur-dioxide-so2), May 2020.
Basic Information about NO2. https://www.epa.gov/no2-pollution/basic-information-about-no2 (https://www.epa.gov/no2-pollution/basic-information-about-no2), May 2020.
What is Ozone. https://www.epa.gov/ozone-pollution-and-your-patients-health/what-ozone (https://www.epa.gov/ozone-pollution-and-your-patients-health/what-ozone), May 2020.
Particulate Matter (PM10 and PM2.5). http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25 (http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25), May 2020.
An Introduction to Impulse Reponse Analysis of VAR Models. https://www.r-econometrics.com/timeseries/irf/ (https://www.r-econometrics.com/timeseries/irf/), May 2020.
Floyd, John. Vector Autogression Analysis: Estimation and Interpretation, September 19 2005.
VAR forecasting methodology. https://stats.stackexchange.com/questions/191851/var-forecasting-methodology (https://stats.stackexchange.com/questions/191851/var-forecasting-methodology), May 2020.