EPA Project

Vasanta Chivukula and Sabrina Ozjan Noor

7/19/2020

Project Background

Anthropogenic activities impact the atmosphere by adding pollutants to the ambient environment and deteriorating the air quality which in turn impacts human health. Additionally, these pollutants pose threat to other ecosystems through acid rains and excessive nitrogen and sulfur depositions (2).

The US passed the Clean Air Act (CAA) in 1963 to mitigate these problems. Sulfur and Nitrogen along with four other components are designated as criteria pollutants by the US Environmental Protection Agency. Thus, CASTNET (Clean Air Status and Trends Network) was established in the US to monitor the air quality for these pollutants.

CASTNET is a national monitoring network established to assess trends in pollutant concentrations, atmospheric deposition, and ecological effects due to changes in air pollutant emissions. Ozone monitoring is one such component of the CASTNET network and the data are submitted near real-time and updated daily. The ozone analyzers are calibrated and checked every night and performance evaluation is done once a year along with a technical system audit every 3 years. The data used for the current research are the concentration of certain pollutants within the ozone monitoring system over the past 29 years, averaged over each year. Ozone data is used to determine if an area meets or exceeds the National Ambient Air Quality Standards.

The variables in this data set are listed below along with a description of each variable -

1. Year: the year the data was measured
2. SO2_CONC: the mean ambient sulfur dioxide concentration in the ozone
3. SO4_CONC: the mean ambient particulate sulfate concentration in the ozone
4. NO3_CONC: the mean ambient particulate nitrate concentration in the ozone
5. HNO3_CONC: the mean ambient particulate nitric acid concentration in the ozone
6. TNO3_CONC: the total ambient nitrate (NO3 + HNO3) concentration in the ozone
7. NH4_CONC: the mean ambient particulate ammonium concentration in the ozone
8. CA_CONC: the mean ambient particulate calcium concentration in the ozone
9. NA_CONC: the mean ambient particulate sodium concentration in the ozone
10. MG_CONC: the mean ambient particulate magnesium concentration in the ozone
11. K_CONC: the mean ambient particulate potassium concentration in the ozone
12. CL_CONC: the mean ambient particulate chloride concentration in the ozone

This data set was chosen since it is a direct measure of human activity and its impact on the environment over the years. The change of pollution levels over the years can be determined using this data set. Additionally, the data set can help identify the most polluting chemical compound thus providing information to potentially reduce the release of that pollutant into the environment.

https://www.epa.gov/castnet/castnet-ozone-monitoring

```r
#install.packages("tidyverse")
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(dplyr)
# Uploaded data set in the Files -> uploads
# Loaded the data into a variable 'CompleteData'
completeData <- read.csv("Concentration - Annual.csv")
```

Data Transformation and Exploratory Data Analyses

Data exploration helps understand any patterns in the data that could help uncover any points of interest

Since carbon, sulfur and nitrogen compounds are the main components affecting the environment (https://gispub.epa.gov/air/trendsreport/2018/#sources), the focus of the current project is to determine their levels in the environment. Thus, only the Sulfur and Nitrogen containing compounds for selected for further analyses.

```r
snData <- completeData[ ,1:10]
```

Removed all the rows that had any cell with no value.

```r
data_notNA <- na.omit(snData)
```

summary of the data_notNA gives the minimum, and maximum values of the pollutant along with other values of central tendency

```r
summary(data_notNA)
```

```
##    SITE_ID               YEAR         DATEON             DATEOFF
##  Length:96          Min.   :1990   Length:96          Length:96
##  Class :character   1st Qu.:2001   Class :character   Class :character
##  Mode  :character   Median :2007   Mode  :character   Mode  :character
##                     Mean   :2007
##                     3rd Qu.:2013
##                     Max.   :2019
##     SO2_CONC          SO4_CONC        NO3_CONC        HNO3_CONC
##  Min.   :0.2860   Min.   :1.000   Min.   :0.2700   Min.   :0.2030
##  1st Qu.:0.6378   1st Qu.:1.925   1st Qu.:0.4170   1st Qu.:0.3600
##  Median :1.1855   Median :2.751   Median :0.4845   Median :0.6425
##  Mean   :1.9128   Mean   :2.978   Mean   :0.7904   Mean   :0.8498
##  3rd Qu.:1.9355   3rd Qu.:3.862   3rd Qu.:1.1660   3rd Qu.:1.0608
##  Max.   :7.5110   Max.   :6.376   Max.   :2.2000   Max.   :2.4450
##    TNO3_CONC        NH4_CONC
##  Min.   :0.755   Min.   :0.1950
##  1st Qu.:1.240   1st Qu.:0.3670
##  Median :1.500   Median :0.5925
##  Mean   :1.626   Mean   :0.6966
##  3rd Qu.:2.050   3rd Qu.:0.8955
##  Max.   :2.844   Max.   :1.7440
```

```r
hist(data_notNA$YEAR) #Histogram showing the frequency of sampling between 1990 and 2019
```

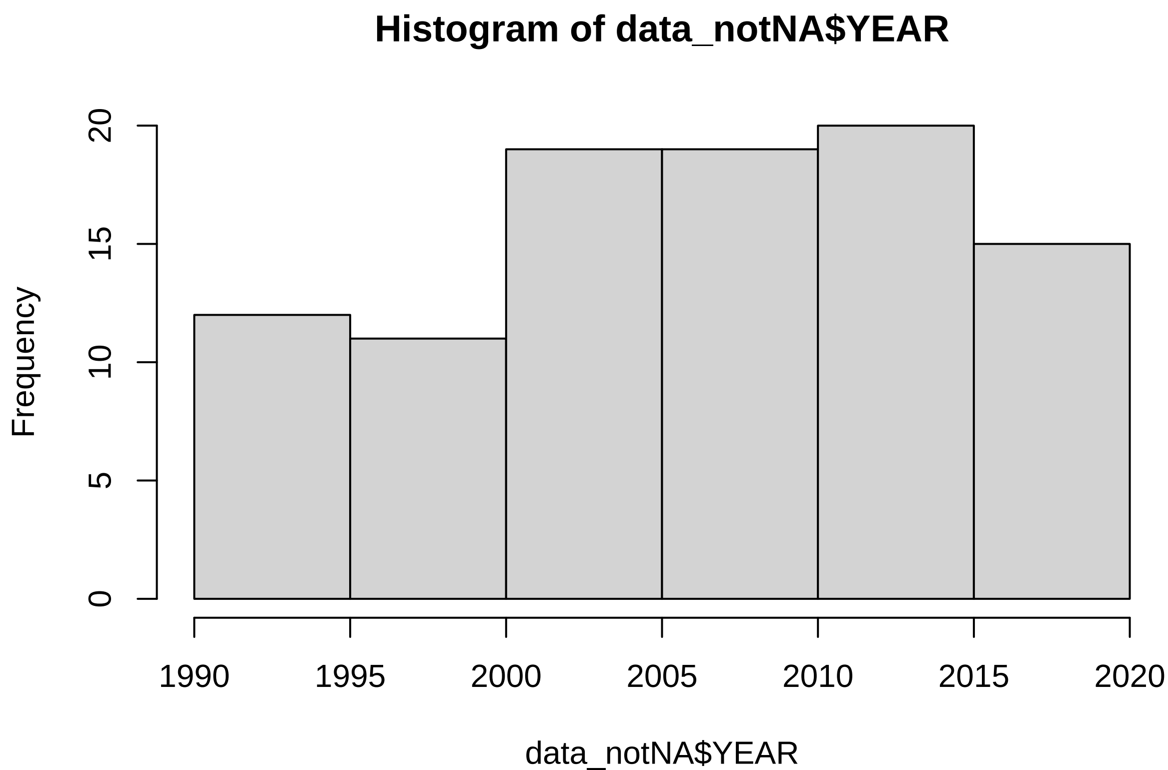


Figure 1: Histogram with sampling frequency The sampling seems more consistent since 2000 and more sites came online since 1999

```r
#sortData_SO2 <- arrange(data_notNA, SO2_CONC) #Sort data for SO2_CONC
#in the increasing order to see the SO2 trends
```

```r
plot(data_notNA$YEAR, data_notNA$NO3_CONC) #Plotting year vs NO3_CONC for any relationship
```

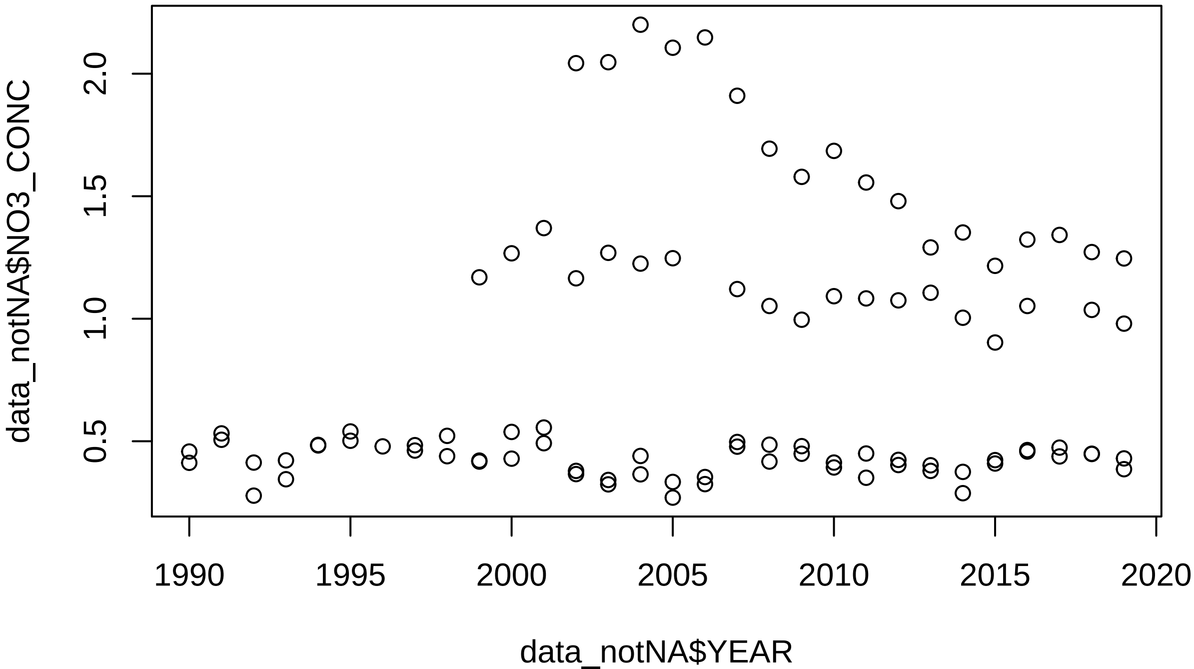
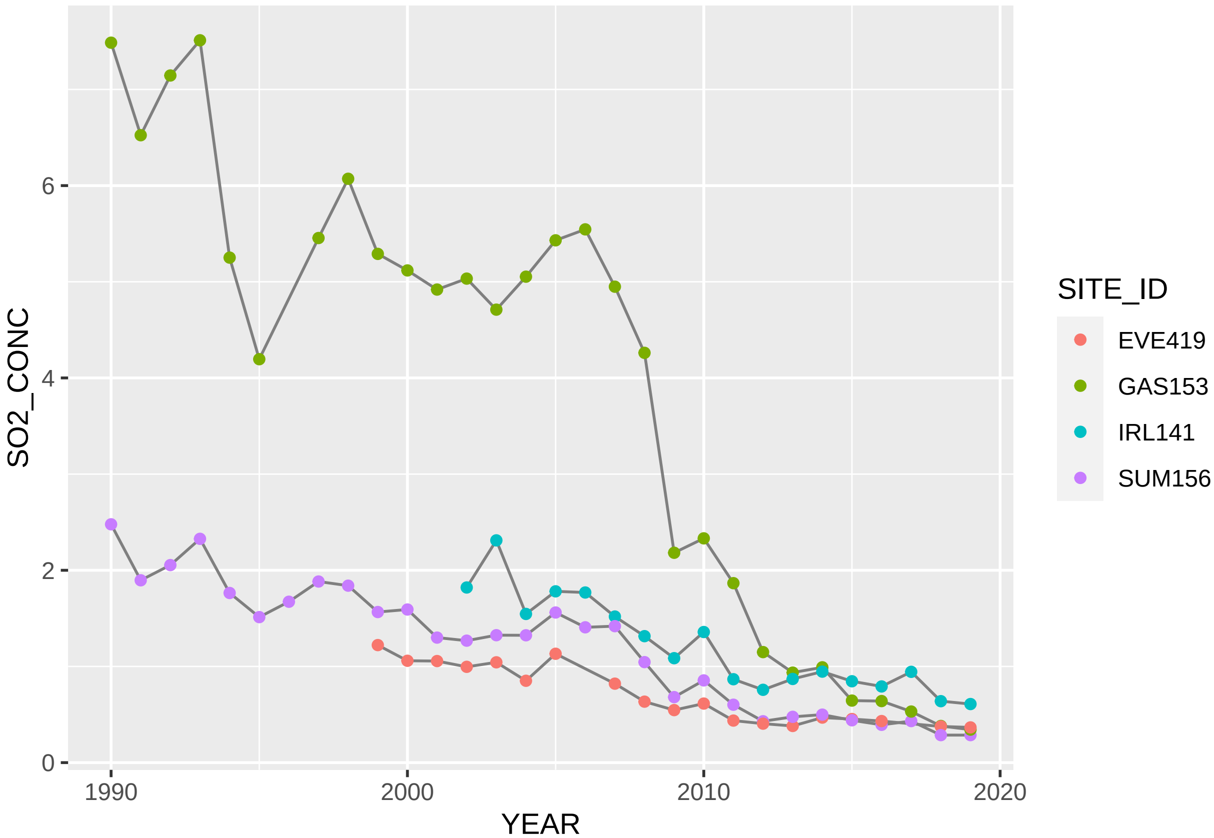


Figure 2: Year vs NO3_CONC

Since 1999, there are three sites that measured the NO3_CONC per year and the fourth one came online since 2002.

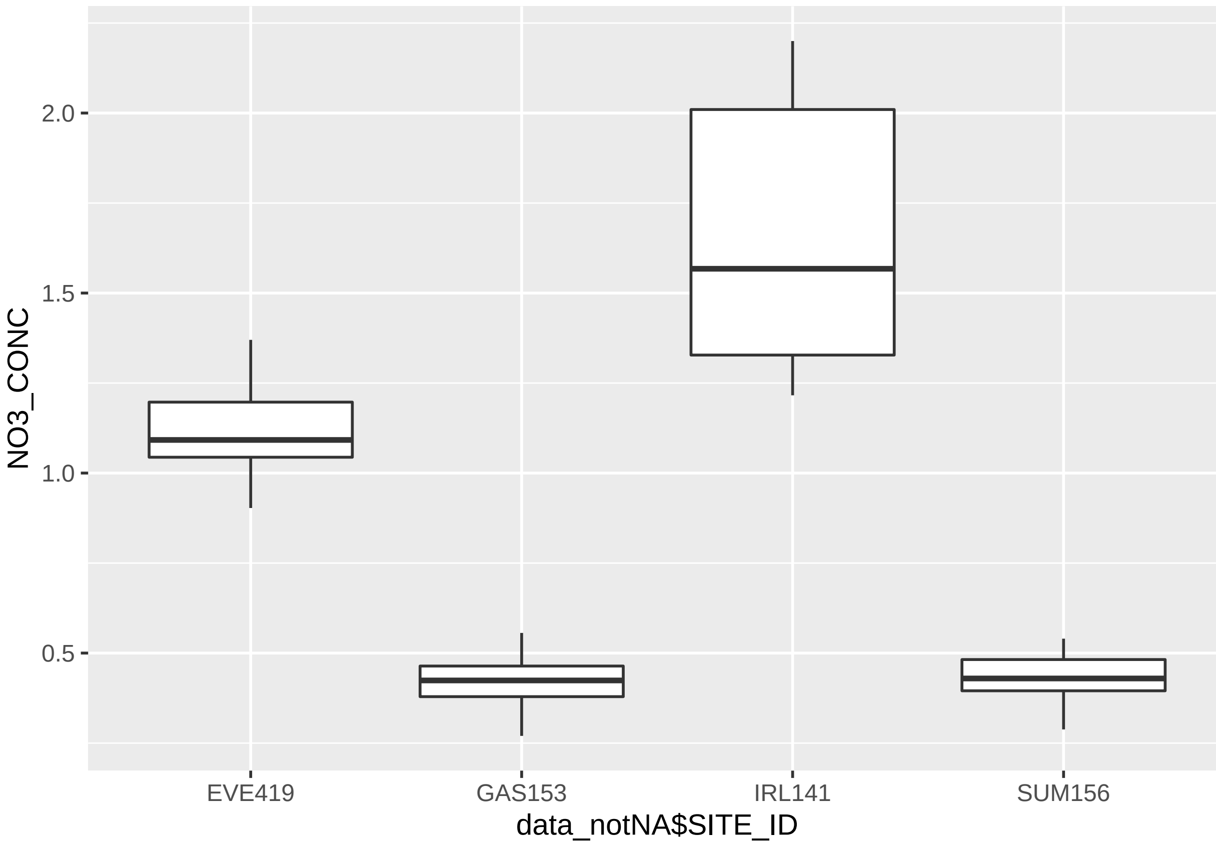
```
ggplot(data_notNA, aes(YEAR, SO2_CONC)) +
 geom_line(aes(group = SITE_ID), color = "grey50") +
 geom_point(aes(color = SITE_ID)) #Plotting year vs SO2_CONC with respect to
```




```
#the four sites in the data set to observe the trends in SO2 in the four sites
```

Figure 3: SO2_CONC per site per year
GAS153 site has higher SO2_CONC when compared to the other three sites. Overall the rates of SO2_CONC have been decreasing since around 2005 at all sites.

```r
data_notNA %>% ggplot(aes(data_notNA$SITE_ID, y=NO3_CONC)) + geom_boxplot()
```




```r
#Boxplot to analyze the NO3 concentrations with respect to the four sites.
```

Figure 4: Boxplot with site vs NO3_CONC
Mean NO3_CONC is highest in IRL141 site. There is less variation with respect to NO3_CONC at SUM156 and GAS153 sites and to a certain extent at EVE419 site. So, the predictability of mean NO3_CONC at these three sites is more dependable.

```r
#ggplot(data_notNA, aes(x=YEAR, y=HNO3_CONC)) + geom_point()
#Geometric point plot to check the relationship between year and HNO3 concentration
```

```r
#install.packages("ggcorrplot")
#install.packages("ggplot2")
#install.packages("RColorBrewer")
library(ggplot2)
library(RColorBrewer)
library(ggcorrplot)
ggplot(data=data_notNA, aes(x=YEAR, y=HNO3_CONC, color=HNO3_CONC))+
  geom_jitter() +
  scale_color_gradientn(colors=topo.colors(15))
```

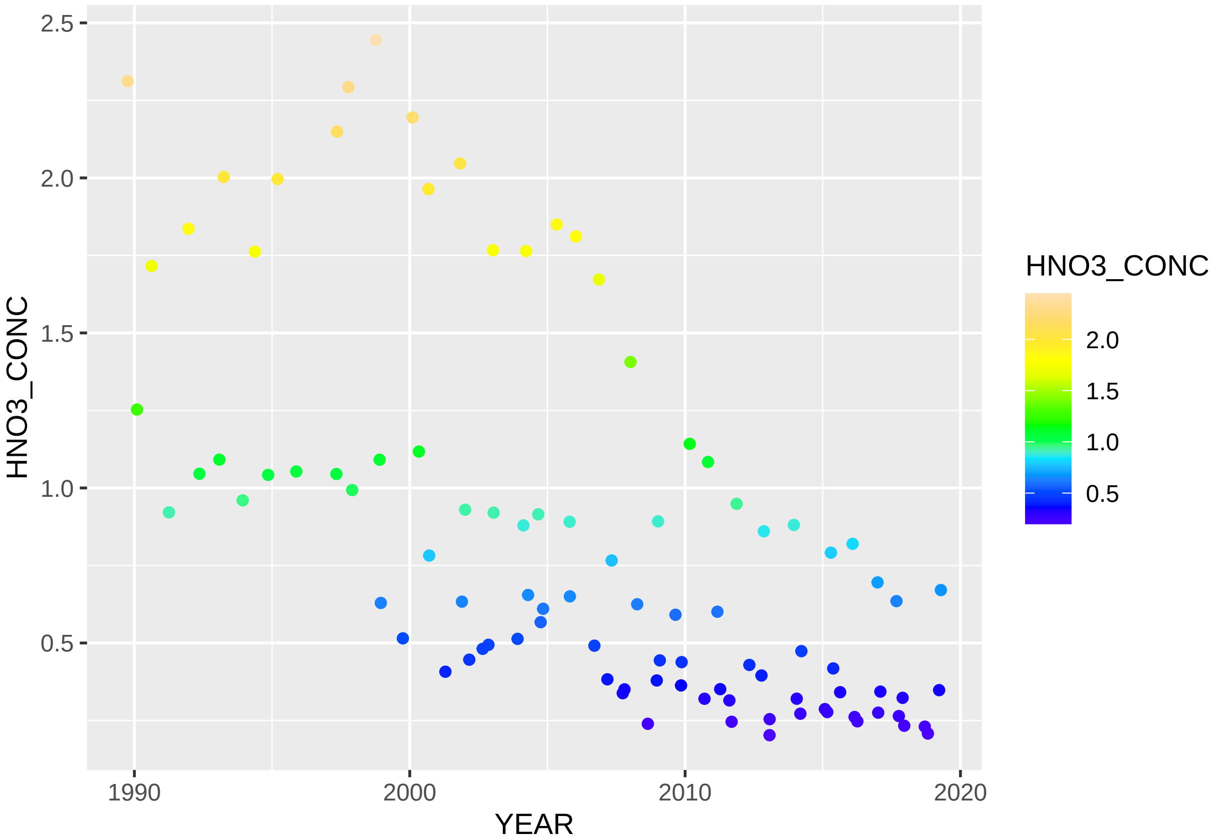
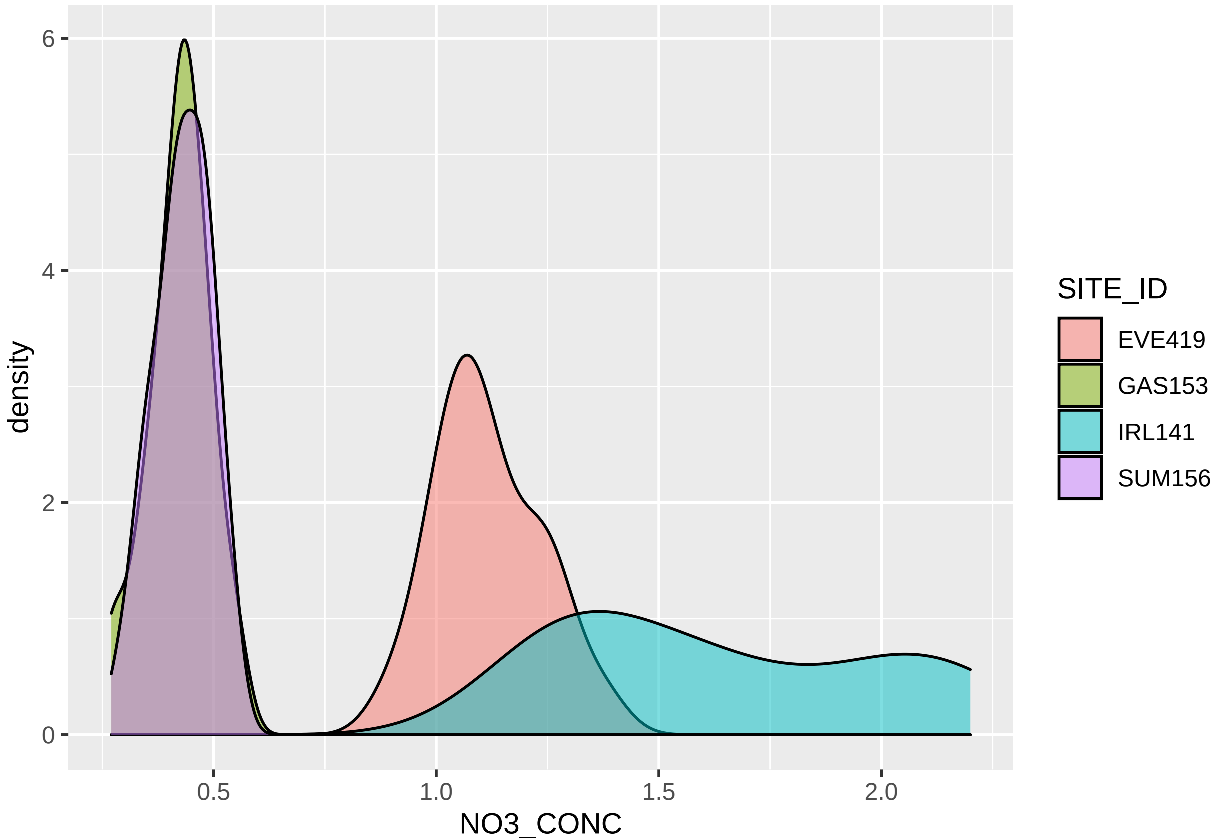


Figure 5: Year vs HNO3_CONC
This plot is similar to the ggplot in fig 3 showing the relationship between SO2_CONC and the year

```r
ggplot(data=data_notNA, aes(x=NO3_CONC, fill=SITE_ID))+
  geom_density(kernal="gaussian", alpha=0.5)
```

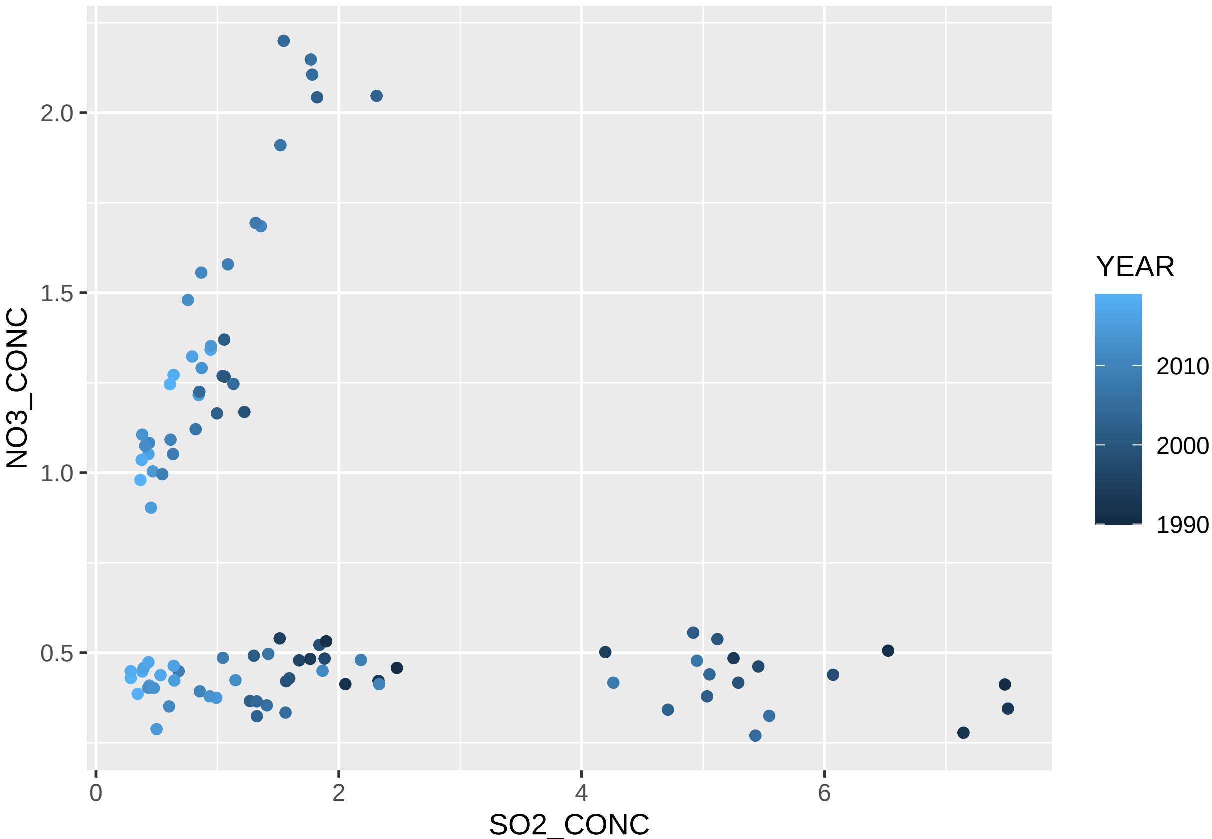
```
## Warning: Ignoring unknown parameters: kernal
```




```r
# Using gaussian distribution to visualize NO3_CONC across the four sites
```

Figure 6: Gaussian Distribution of NO3_CONC at the four sites NO3_CONC at sites GAS153, SUM156 and to an extent at EVE419 follow the normal distribution (bell-shaped curve). Site IRL141 does not fit in the normal distribution which follows the trend seen in figure 4 (box plots)

```r
ggplot(data_notNA, aes(x=SO2_CONC, y=NO3_CONC, color=YEAR)) + geom_point()
```



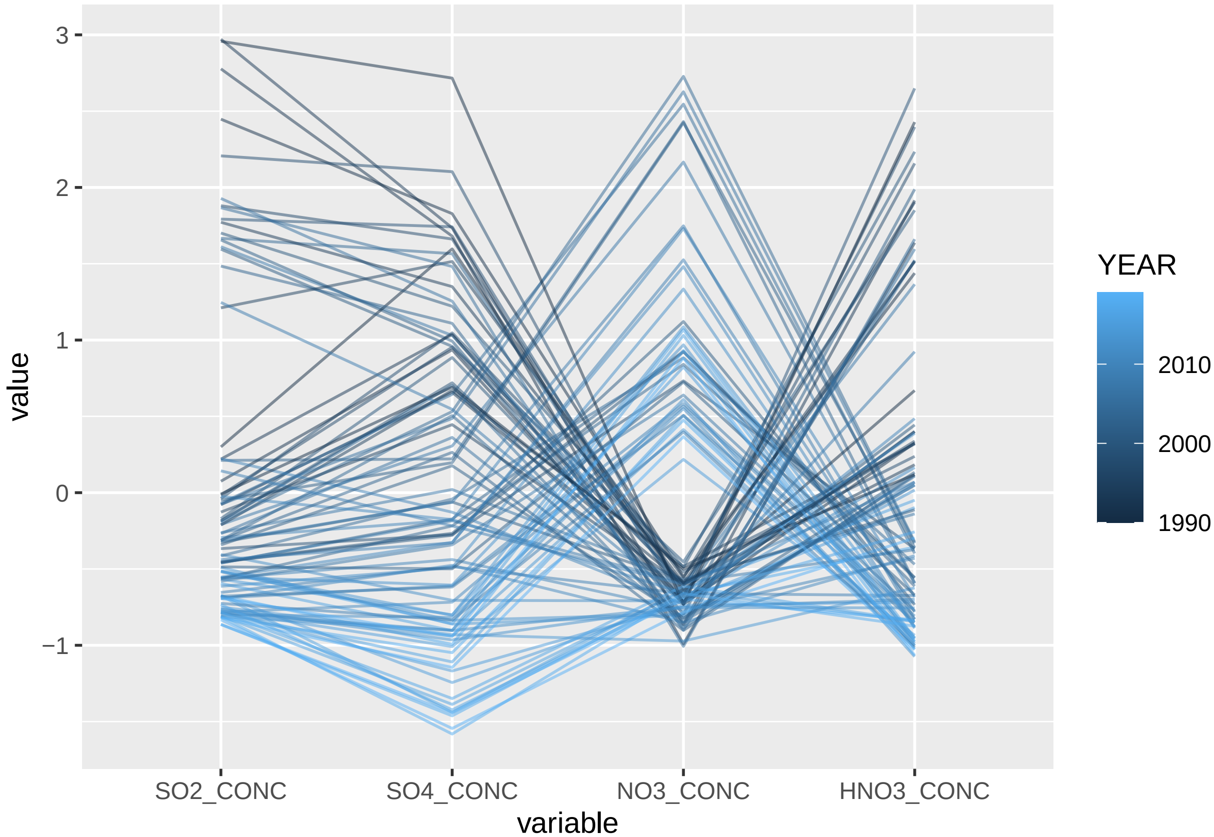

```r
#Both SO2 and NO3 concentrations by year
```

Figiure 6: SO2_CONC and NO3_CONC between 1990 and 2019 The light blue dots congregated mostly near the origin of the graph relate to the decreasing rates of NO3 and SO2 concentrations in the environment over the years

```r
#install.packages("GGally")
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
ggparcoord(data_notNA, columns=5:8, groupColumn = 2, alphaLines = 0.5)
```




```r
#Parallel coordinates plot using data from columns 5-8 in the data set
#and grouping them by year (column 2)
```

Figure 7: Parallel coordinates plot with SO2, SO4, NO3, and HNO3 concentrations

Parallel coordinates plot maps each row in an excel sheet into a line and each value in the row would be a point on the line.So all the 96 observations for the four variables are shown in this plot.

Data Models

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that is used in data mining and machine learning. It groups together points that are close to each other based on a distance measurement. The outliers that result in the end are in the lower density regions of the clustering. DBSCAN finds associations and structures in data that are difficult to find manually (3). DBSCAN is beneficial as it can group all the chemical compounds in the atmosphere that contribute the most to polluting it. The outliers that would result in this method of analysis would be the compounds that have the least effect in contaminating the air.

In this project a smaller subset of the larger data set is used. The DBSCAN approach would be especially useful for the larger data set. Its implementation is fairly simple as there are preexisting packages and libraries available for it. Instead, Principal Component Analysis is used for this project due to the smaller data set used for this project.

Principal Component Analysis

Principal Component Analysis is a feature extraction technique which helps to reduce a multidimensional data into a single dimension. It drops the least important variables from a data set while still retaining the most valuable variables in a data set. By identifying the dimensions that are most important, PCA drops the unimportant dimensions thus making the data simpler for use. Additionally, the new variables obtained are independent of each other which give an added advantage by satisfying the assumptions of a linear model (requires that the variables are independent of each other) (4). In the EPA data set that was selected, there are 6 different pollutants that are measured every year for the past 29 years. By performing PCA, we can identify the pollutants that are mostly present in the environment and are the principle components polluting the environment.

```r
data_numeric <- select(data_notNA, 5:10)
# selecting only the columns with the pollutant concentrations
data_pca <- prcomp(data_numeric, center=TRUE, scale=TRUE)
#Running PCA on the data
summary(data_pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5      PC6
## Standard deviation     2.0879 1.1822 0.39233 0.26729 0.13212 0.008517
## Proportion of Variance 0.7266 0.2329 0.02565 0.01191 0.00291 0.000010
## Cumulative Proportion  0.7266 0.9595 0.98517 0.99708 0.99999 1.000000
```

```r
screeplot(data_pca)
```

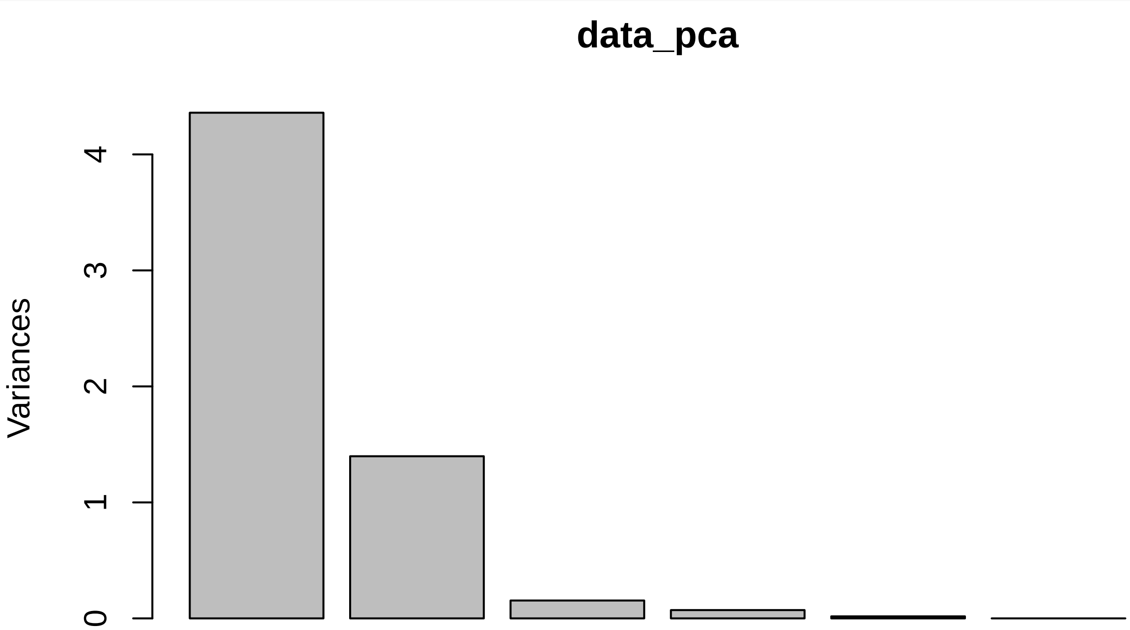


Figure 8: PCA data plot
The first two PCA components explain about 95% of the variation in the data.

```r
data_pca
```

```
## Standard deviations (1, .., p=6):
## [1] 2.087922318 1.182237221 0.392329168 0.267294518 0.132115365 0.008517092
##
## Rotation (n x k) = (6 x 6):
##                  PC1          PC2         PC3         PC4        PC5
## SO2_CONC  -0.4626166  0.001347003  0.37871489 -0.79037664  0.1336077
## SO4_CONC  -0.4491724 -0.083071878 -0.84102999 -0.09387039  0.2742245
## NO3_CONC   0.1547506 -0.799607514 -0.03939867 -0.13432195 -0.1535615
## HNO3_CONC -0.4619404  0.170870447  0.27983452  0.46045462  0.3132167
## TNO3_CONC -0.3496467 -0.565705159  0.25878813  0.35353643  0.1684770
## NH4_CONC  -0.4739625  0.067127012 -0.04911901  0.10697787 -0.8699883
##                    PC6
## SO2_CONC  -0.003581323
## SO4_CONC   0.001013695
## NO3_CONC  -0.541760365
## HNO3_CONC -0.607440429
## TNO3_CONC  0.580845098
## NH4_CONC  -0.010814403
```

```r
biplot(data_pca) #Biplot used to visualize the PCA data
```

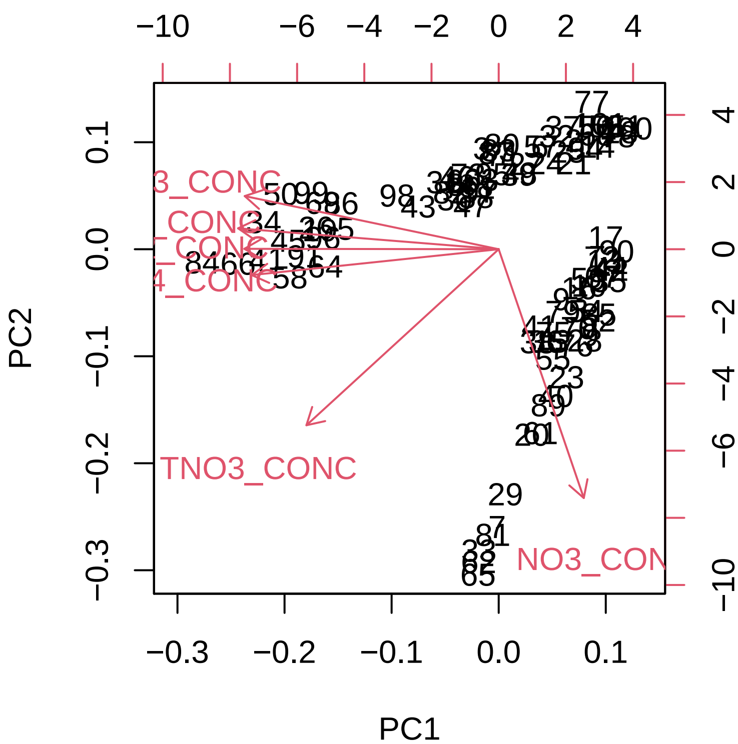


Figure 9a: Biplot of the PCA data

```r
biplot(data_pca, expand=10, xlim=c(-0.30, 0.0), ylim=c(-0.1, 0.1))
```

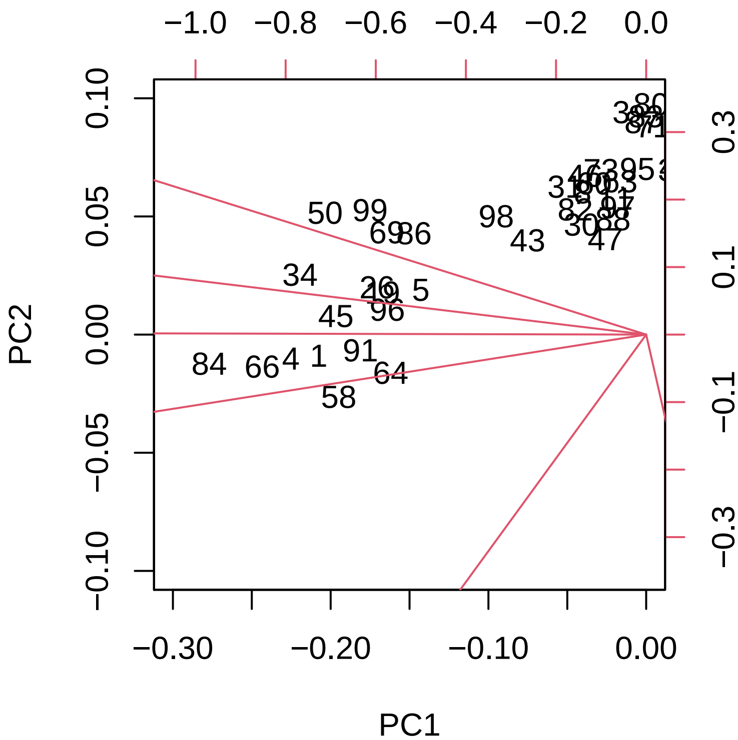


Figure 9b: Biplot of the PCA data using options for better visualization of the data

Discussion

Figure 2 illustrates the reduction in the concentration of Nitrate (NO3) over the years. This result is consistent with multiple studies (5,6). The Sulfur dioxide (SO2) concentrations also show a drop at all 4 sites (figure 3). It is interesting that the GAS153 site reports an initial SO2 concentration that is significantly higher than the other 3 sites. This may be due to the airspace in which they carried out the measurements that could have had a higher concentration. Furthermore, while the remaining 3 sites (IRL141, EVE419, and SUM156) show a relatively steady decline in concentration levels, GAS153 drops substantially and was also prone to more aggressive fluctuations than the others. Nitric Acid (HNO3) concentrations are similar to the ones observed for SO2. They decline over the 20-year period with more fluctuations and steeper declines observed from GAS153 as shown in figure 5. Figure 6 depicts the changes in the SO2 concentrations over the years on X-axis while the the changes for NO3 concentrations are depicted on the Y-axis. The darker color represents the later years and it is evident that as time progresses, both the SO2 and NO3 concentrations declined.

The Principal Component Analysis provided six principal components that explain the total variation in the dataset. PC1 explains 73% of the total variance, which means that nearly three-fourths of the information in the dataset (6 variables) can be encapsulated by just that one Principal Component. PC2 explains 23% of the variance. So, by knowing the position of a sample in relation to just PC1 and PC2 one can explain 96% of the variance.

Bibliography

1. https://www.epa.gov/castnet/castnet-ozone-monitoring
2. Feng et al. Air quality in the eastern United States and Eastern Canada for 1990–2015: 25 years of change in response to emission reductions of SO2 and NOx in the region. Atmos. Chem. Phys., 20, 3107–3134, 2020
3. http://www.cs.fsu.edu/~ackerman/CIS5930/notes/DBSCAN.pdf
4. Zhang, Nina et al. "Data-Driven Analysis of Antimicrobial Resistance in Foodborne Pathogens from Six States within the US." International journal of environmental research and public health vol. 16,10 1811. 22 May. 2019, doi:10.3390/ijerph16101811 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6572035/
5. https://www.epa.gov/air-trends/nitrogen-dioxide-trends
6. Li, Yi et al. The importance of reduced nitrogen deposition. Proceedings of the National Academy of Sciences May 2016, 113 (21) 5874-5879; DOI: 10.1073/pnas.1525736113