Analysis and Modeling of US Pollution Data for 1980-2018 and its implications on the atmosphere

By:

Yash Shailendra Gokhale (ysg)

Course: (19-603) Data Science for Technology,Innovation and Public Policy

Instructor: Professor Alexander Davis

INDEX

Abstract

Pollution has become a word of widespread discussion all over the world, mainly due to its increasing presence in our lives. Be it New Delhi in India, or be it Beijing in China, air pollution has caused rapid disruption in daily activities leading to closure of schools and public places. Apart from harming the social life, pollution has also affected the environment in an adverse manner, which rampant fluctuations in atmospheric parameter. This report considers four major pollutants, Carbon monoxide (CO), Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2) and Ozone (O_3). These pollutants are the key components of any industrial smog or emission which is released into the atmosphere. Moreover, it is easier to measure these concentrations because of readily available techniques. The report aims to correlate the concentrations of these species with two atmospheric parameters, Temperature and %RH (Relative Humidity). Apart from finding direct correlations, the report also takes into account spatial and temporal considerations. For the scope of this report, the area under scrutiny is the United States of America, considering the 50 states and several cities in each state. The report uses different data visualization techniques to establish correlations between the involved parameters. Robust models are developed using data science techniques and are tabulated. Moreover, apart from the concentration factors, the model also aims to correlate the number of samples chosen to arrive at the mean reading to the states in the USA.

Data

Apart from the four pollutants mentioned here, there are several other undetected pollutants, which are present in negligible amounts. For the scope of this project, the concentration of all these pollutants are assumed to have no effect on the proposed models and predictions. They can either be considered as an inclusion in the intercept or a part of the residual error. For the scope of this project, the region in focus is the USA, considering all the 50 states, with a variable number of cities in each state, depending upon the size and feasibility of data collection. The data was obtained from: 'https://aqs.epa.gov/aqsweb/airdata/download_files.html', which is the official data repository for US EPA (Environment Protection Agency). The referred data is in the form of multiple .csv files, with files separated by each year and every pollutant. The data collected for each pollutant follows a general pattern:
(StateCode - CountryCode -SiteCode -ParameterCode -POC -Latitude -Longitude -Datum-SampleDuration -Date- Units -ObservationCount- Percent- Arithmetic Mean-Maxvalue-AQI-Method-State-County-City)
The required features are extracted from the files for further modeling.
The scattered .csv files were combined into a single file for each pollutant, for the years 1980-2018. Lastly, each of those files for individual parameters were combined into a final .csv files, so as to correlate the pollutants to these dependant parameters.
The combined .csv files are made publicly available at:
https://drive.google.com/drive/folders/1yKqv8DsB6o1szznw1qjfyM8zmG-egUOE?usp=sharing .

Data Visualization

Time variance of Pollutants

In order to understand the trend of each individual pollutant over time, the graphs of pollutant concentration was plotted as shown in Figure 1.

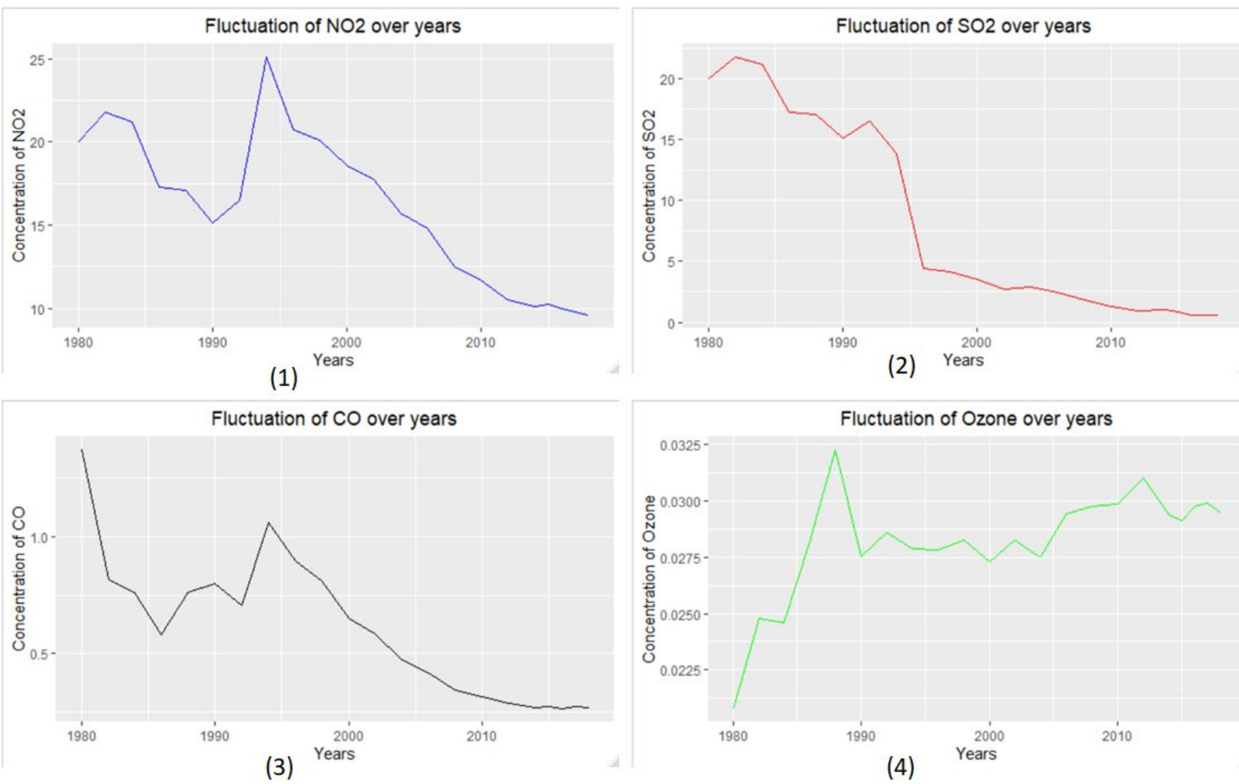

Figure 1: Trend of Pollutants over Time

From the plots, the trend of each pollutant over time can be roughly estimated. The concentration of NO_2, SO_2 and CO have been on the decline for the past 25 years and keep on declining with subsequent years. On the contrary, the concentration of Ozone has a fluctuating pattern, with frequent peaks in the curve. No definite conclusion can be drawn about the Ozone concentration.

For the trend of NO_2, SO_2 and CO, there is a sudden peak observed in the pollutant concentration in the decade of the 1990s, roughly around 1994-96. There seems to be a correlation between each of the individual pollutant's behavior, and thus, these pollutants are correlated later in the regression models.

Time variance of Atmospheric Parameters

In order to understand the trend of Temperature and Relative Humidity over time, the graphs were plotted as shown in Figure 2.

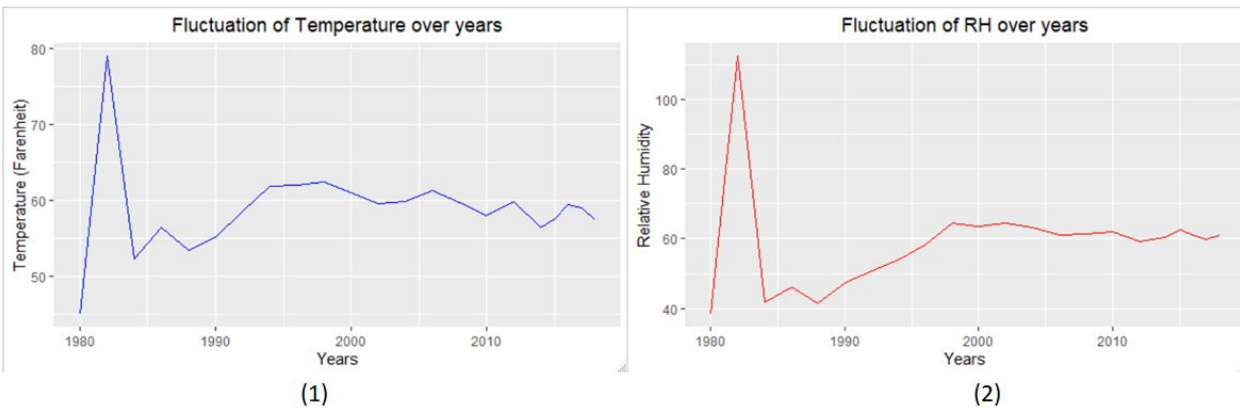

Figure 2: Temperature and %RH fluctuations over Time

The trend observed for Temperature and Relative Humidity over the years is similar to each other, with a significant peak at the exact time. Moreover, the Temperature and RH seems to have minor fluctuations over the past two decades. Also, the sudden drop in Temperature and RH over the third time instant is unexpected and is thus analyzed in the further sections.

Consideration of Sample Size

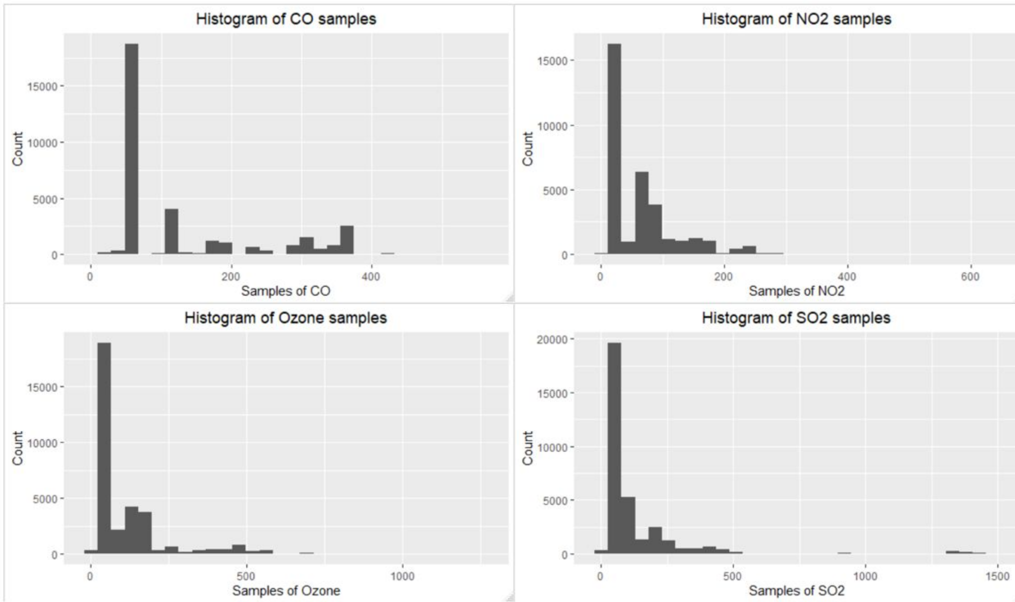

Figure 3: Frequency distribution of sample size

Sample size is an important consideration in the design of experiments. A sample size is the number of times each sample is repeated for better accuracy. In case of areas with tougher atmospheric conditions, multiple samples are collected to ensure higher perfection.. However, here, it is observed that maximum readings have less than 10 samples,while some samples are collected multiple times. Also, the sample size may vary state to state depending upon the ease of setting up measurement instruments. This analysis is done in further part of the report.

Data Modeling

Temperature Model (LM)

Temperature, a dependant parameter was modeled against the following independent parameters:

1) Pollutant concentration
2) Time (Years)
3) State

Thus, the temperature is a spatial-temporal dependant parameter, with State treated as categorical variable whereas the pollutant concentrations and years as continuous.
The model thus fit was:

$$T_i = \beta_0 + \beta_1(SO_2)_i + \beta_2(NO_2)_i + \beta_3(O_3)_i + \beta_4(CO)_i + \sum\beta_i(year)_i + \sum\beta_i(state)_i$$

Months and states were modelled as factors. In order to avoid the dummy variable trap, coefficients are reported except for April (with month as a categorical variable) and Alabama (with state as a categorical variable).
From the tabulated results, it can be inferred that the coefficients of CO and O_3 are satisfactorily high, which implies that these two pollutants have a greater impact on the Temperature in comparison to NO_2 and SO_2, which have sufficiently low coefficients.
The coefficients for months were obtained as follows:

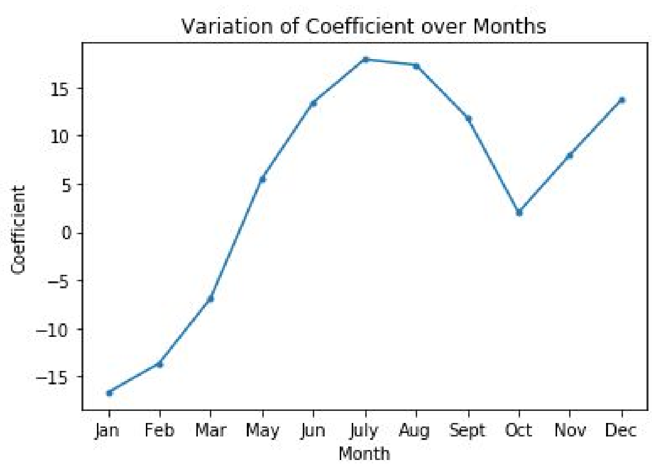

Figure 4: Plot of Regression Coefficients (Monthwise)

As for the states, regression coefficients were tabulated.

term	estimate	term	estimate	term	estimate
(Intercept)	-226.83	Connecticut	-12.5631	Nevada	-5.91858
Avg_CO	4.56914	Delaware	-7.92631	New Hampshire	-14.6006
Avg_NO2	-0.0471	District Of Columbia	-13.7595	New Jersey	-7.67752
Avg_O3	36.0408	Florida	10.73616	New Mexico	1.351426
Avg_SO2	0.00023	Georgia	-3.34047	New York	-10.1976
Year	0.14349	Hawaii	11.7086	North Carolina	-3.65689
Jan	-16.659	Idaho	-16.2028	North Dakota	-21.5627
Feb	-13.627	Illinois	-7.43708	Ohio	-12.0193
Mar	-6.9051	Indiana	-7.60816	Oklahoma	-2.9541
May	5.51411	Iowa	-10.9539	Oregon	-10.2493
Jun	13.4188	Kansas	-6.42664	Pennsylvania	-2.0768
Jul	17.8756	Kentucky	-6.38002	Rhode Island	-8.37961
Aug	17.2962	Louisiana	5.971581	South Carolina	0.103814
Sep	11.864	Maine	-18.5837	South Dakota	-14.5063
Oct	2.01426	Maryland	-8.05173	Tennessee	-8.21806
Nov	-7.9724	Massachusetts	-12.876	Texas	3.83385
Dec	-13.669	Michigan	-13.7403	Utah	-9.72425
Alaska	-34.074	Minnesota	-17.014	Vermont	-16.1446
Arizona	6.1712	Mississippi	2.795587	Virginia	-11.0988
Arkansas	-1.4125	Missouri	-13.315	Washington	-13.1707
California	-2.8004	Montana	-19.1562	Wisconsin	-11.8774
Colorado	-14.213	Nebraska	-20.9554	Wyoming	-21.4451

Table 1: Regression Coefficients for Temperature Model

The adjusted R^2 value for the model was 0.8149. The inclusion of states in the regression model helps improve the prediction and helps in obtaining a better R^2 value.

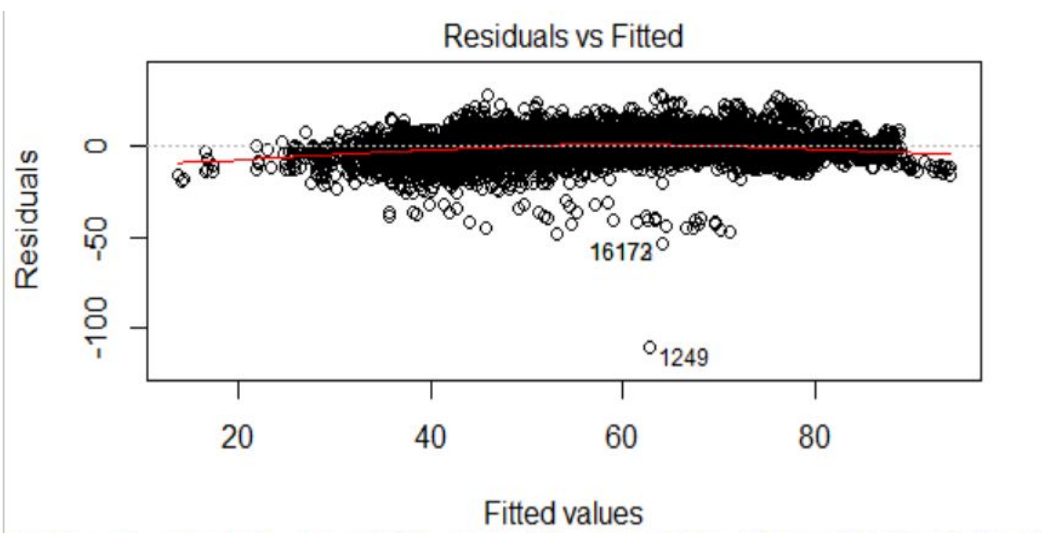

Figure 5: Simple Linear Model for Temperature

Relative Humidity Model (LM)

Being an environmental impact factor as temperature, RH was modeled in a similar manner, with the same number of variables. The equation used to model the RH was:

$$RH_i = \beta_0 + \beta_1(SO_2)_i + \beta_2(NO_2)_i + \beta_3(O_3)_i + \beta_4(CO)_i + \sum\beta_i(year)_i + \sum\beta_i(state)_i$$

The entire modeling for Relative Humidity was identical to that of Temperature. In this case, however, it was assumed that there is no correlation between temperature and Relative Humidity.

Tabulated results are as follows:

term	estimate	term	estimate	term	estimate
(Intercept)	-226.308	Connecticut	-14.7157	Nevada	-23.7249
Avg_CO	-1.74626	Delaware	-6.35304	New Hampshire	4.476569
Avg_NO2	-0.20441	District Of Columbia	6.453698	New Jersey	-0.96659
Avg_O3	-541.845	Florida	2.239833	New Mexico	-15.0416
Avg_SO2	-0.00491	Georgia	7.295611	New York	10.25815
Year	0.15398	Hawaii	-3.70599	North Carolina	-3.01871
Jan	-8.58173	Idaho	-5.06825	North Dakota	0.7829
Feb	-6.05071	Illinois	1.134144	Ohio	-4.9674
Mar	-2.05021	Indiana	-14.1086	Oklahoma	2.074709
May	4.09492	Iowa	3.851344	Oregon	3.478888
June	5.08343	Kansas	1.546353	Pennsylvania	5.634234
Jul	8.54221	Kentucky	1.725784	Rhode Island	12.42894
Aug	7.47955	Louisiana	15.88261	South Carolina	1.282601
Sep	4.277	Maine	5.459046	South Dakota	-0.74717
Oct	0.22345	Maryland	2.169051	Tennessee	9.312006
Nov	-3.6903	Massachusetts	2.392554	Texas	-28.7471
Dec	-0.05086	Michigan	1.088144	Utah	-15.6283
Alaska	-38.6861	Minnesota	-15.3924	Vermont	2.704089
Arizona	-18.7296	Mississippi	-5.84055	Virginia	-7.10243
Arkansas	-27.874	Missouri	-0.69322	Washington	7.256789
California	-4.27962	Montana	-2.07241	Wisconsin	-13.136
Colorado	-15.1254	Nebraska	-40.0889	Wyoming	-1.61903

Table 2: Regression Coefficients for RH Model

In case of the Relative Humidity, it was observed that the coefficient for O_3 was much higher as compared to the other coefficients. Although, inherently, the value of ozone concentration is low (ppm or ppb) and thus, the coefficient is high. While considering the coefficients for months, there was an increasing trend from January to July, thereby again going to smaller values till December.

The adjusted R^2 value for the model was 0.5574. The fit was not satisfactory as compared to the temperature model, however, the trends were similar to that of the temperature fit.
The plot for the residuals indicate a more uniform distribution of the residuals.

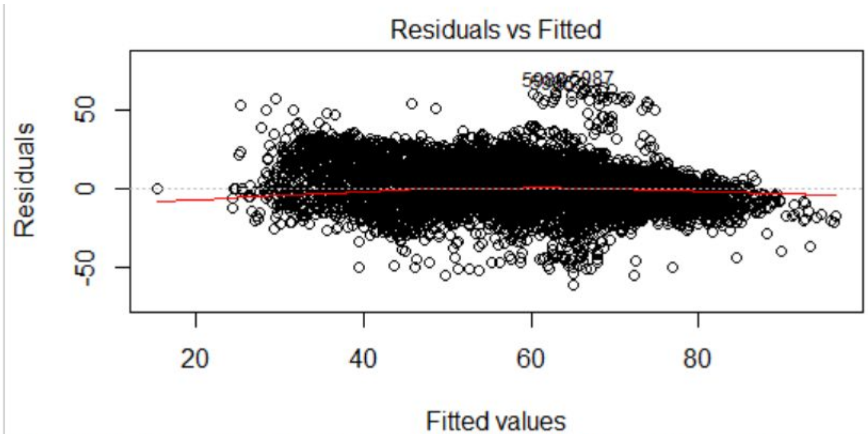

Figure 6: Simple Linear Model for RH

Level Log Transform

Level Log Transform Model was implemented to determine the fit and see its effect on the residuals.
The model implemented was:

$$log(T)_i = \beta_0 + \beta_1(SO_2)_i + \beta_2(NO_2)_i + \beta_3(O_3)_i + \beta_4(CO)_i + \sum\beta_i(year)_i + \sum\beta_i(state)_i$$

$$log(RH)_i = \beta_0 + \beta_1(SO_2)_i + \beta_2(NO_2)_i + \beta_3(O_3)_i + \beta_4(CO)_i + \sum\beta_i(year)_i + \sum\beta_i(state)_i$$

The model gave a lower R^2 value of 0.6916 for Temperature whereas an R^2 value of 0.506 was obtained for the RH Model. Also, the residual errors were higher in both the cases of Temperature and RH.

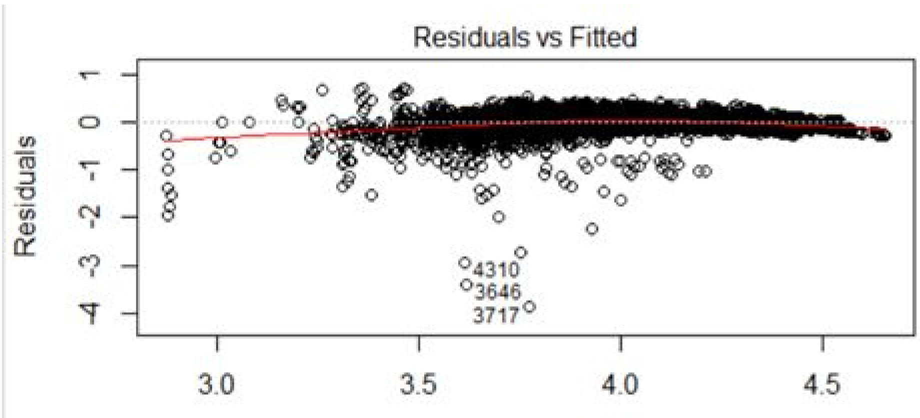

Figure 7: Residual Plot for Temperature Level Log Transform

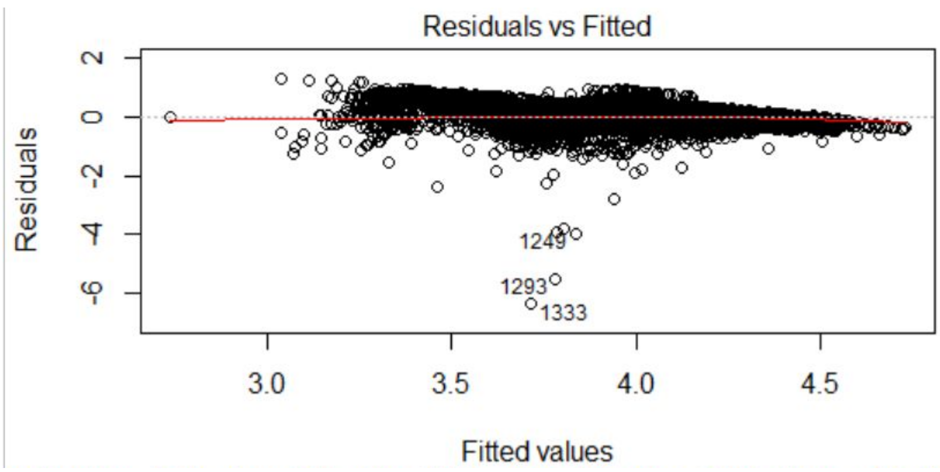

Figure 8: Residual Plot for RH Level Log Transform

The residual plot for Temperature and RH are identical in all sorts. Moreover, the residual values are scattered with an improper fit. The linear model, thus has a better fit as compared to the Level Log Model. An alternate option would be to model a Box-Cox model.

Log Log Model

All the continuous variables (Dependant and Independent Variables) were transformed logarithmically, while keeping the binary variables as same.

$$log(T)_i = \beta_0 + \beta_1 log(SO_2)_i + \beta_2 log(NO_2)_i + \beta_3 log(O_3)_i + \beta_4 log(CO)_i + \sum \beta_i (year)_i + \sum \beta_i (state)_i$$

$$log(RH)_i = \beta_0 + \beta_1 log(SO_2)_i + \beta_2 log(NO_2)_i + \beta_3 log(O_3)_i + \beta_4 log(CO)_i + \sum \beta_i (year)_i + \sum \beta_i (state)_i$$

The adjusted R^2 obtained for the temperature model was 0.6823, whereas that for the RH model was 0.4931. Residual plots obtained were:

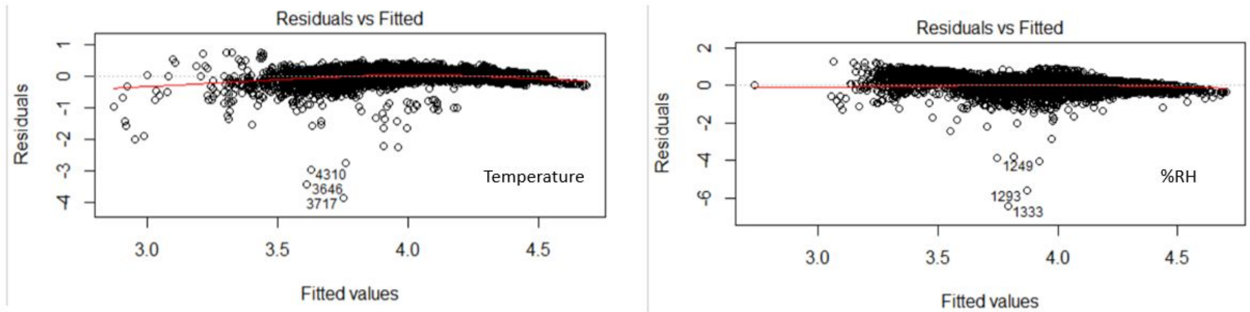

Figure 9: Log-Log Transform model results for the Dependant Variables

The distribution of residuals was uniform for the Log Log model as compared to the Simple Model, in spite of having a lower R^2 fit. From all the models,a general regression model seemed optimum.

Model	Temperature	%RH
Linear Regression Model	0.8149	0.5574
Level Log Model	0.6916	0.506
Log Log Model	0.6823	0.4921

Table 3: Adjusted R^2 values for the parameters

Sample Size Analysis

The number of samples collected is an important parameter influencing the value of the concentration obtained. For instance, if the ease of obtaining a sample is higher, then the number of samples are lesser. In order to assess the states with the maximum number of samples collected, a regression model was fitted for each of the four pollutants.

The regression model followed was:

$$N_i = \beta_0 + \sum \beta_i (state)_i$$

The states with the maximum number of samples collected were as follows:

Sample coefficient	CO	NO_2	SO_2	O_3
Maximum	Arizona	Wyoming	Indiana	Wyoming
2nd Maximum	Newyork	North Dakota	North Dakota	Maine
3rd Maximum	Wisconsin	Montana	Montana	Virginia
Median	Kentucky	New Mexico	Maine	Utah
3rd Least	Rhode Island	Mississippi	Rhode Island	Alaska
2nd Least	Minnesota	Delaware	Maryland	Hawaii
Least	Delaware	South Carolina	Nebraska	South Carolina

Table 4: Sample Size Coefficients for Pollutants

The regression analysis can be found at: https://drive.google.com/drive/u/0/folders/1I1DjKMK0DRxnhsi3S5dY5gntBFBsBW3K.

From the analysis of samples, it can be observed that there is no uniform pattern in location for each of the pollutants. The extent of ease and difficulty of collecting samples varies from pollutant to pollutant. For instance, it is comparatively tough to measure the O_3 concentration in Maine (3rd highest coefficient of regression), whereas it is moderately difficult to collect samples of SO_2 in Maine (Median coefficient of regression). This analysis can be applied in designing new experiments for measurements of pollutants in future. A measurement exercise set up to measure CO concentration in Arizona would take up more effort than setting it up in Delaware or Minnesota.

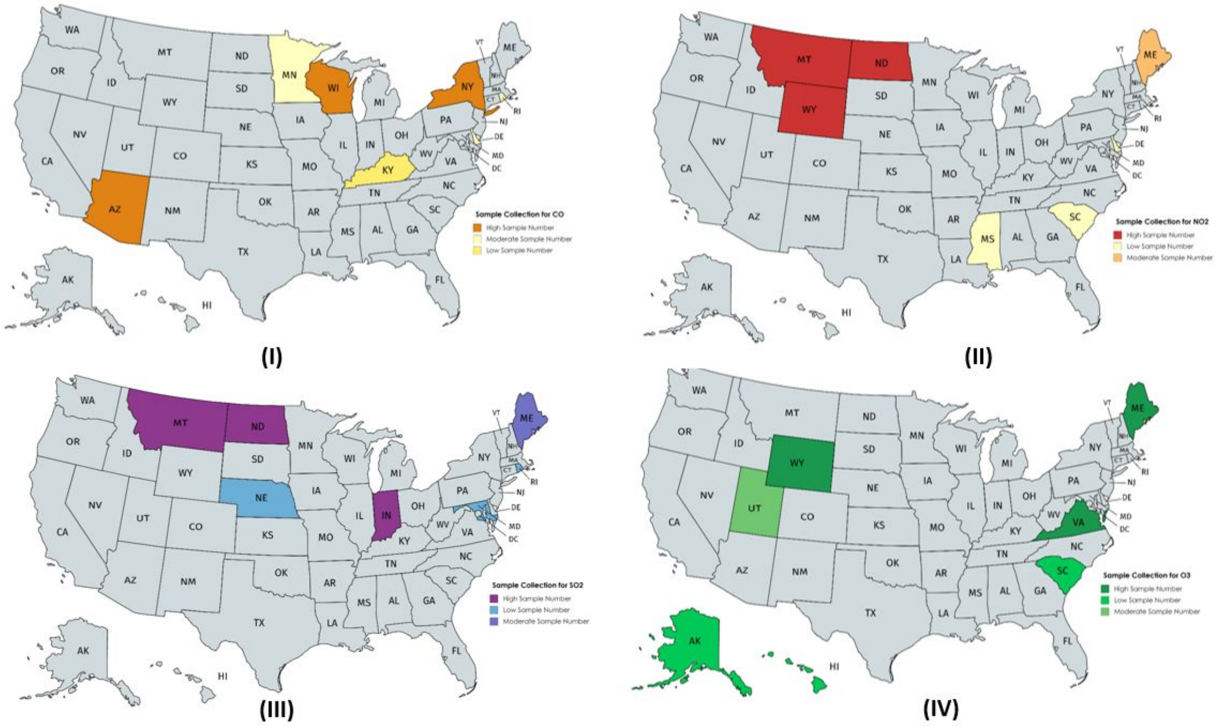

Figure 10: Spatial Distribution Graph for Sample Collections: I-CO, II-NO_2, III-SO_2, IV-O_3

From the spatial map of the sample size, it can be inferred that location plays a part in the number of samples being collected. In the case of NO_2, there seems to be a predilection for the northern states for higher number of samples. A similar observation can be made in case for SO_2. Similarly, lower number of samples are collected for ozone for the islandic states such as Alaska and Hawaii.

Thus, a spatial study of the number of samples collected are useful in determining the future collection strategy for a particular pollutant. For instance, more effective methods for NO_2 sample collection must be used for the northern states so as to reduce the number of samples collected and hence the cost involved.

Gauss Markov Assumptions

In order to implement a regression model, we need to be assured that the Gauss Markov Assumptions are satisfied. For the model implemented in the report, the Gauss Markov Assumptions are satisfied as follows:

1) Linearity in parameters

   In all the three models implemented in the report, it has been modeled by satisfying this assumption. The general structure of any model implemented in the paper is as follows:

   $y = \sum \beta_i * x_i$

   Here, the independent variable x_i is not necessarily a linear value. For instance, in the log-log transformation model, the independent variable has been transformed into a logarithmic value of that variable. However, there is no effect on the nature of parameters in this case. The individual parameters are independent of each other. Each value of β_i is calculated through a standard linear regression model and none of these have interaction amongst each other. Thus, the first condition of Linearity is satisfied.

2) Random Sampling

   The nature of the data is such that the condition of Random Sampling is inherently satisfied. All of the concentrations recorded have been noted at a particular time instant and at a particular location. The concentration of pollutants in the atmosphere are not manually modeled or structured. In case of temperature and Relative Humidity, the data collection is done in a manner similar to that of the pollutants. There is no fixed method in which the data is recorded. Thus, due to the nature of data under consideration, the second condition is satisfied.

3) Zero conditional mean of errors

   The model developed here has a total of 5 independent continuous variables and a large number of independent categorical variables. For every model mentioned here, there are 2 omitted variables, one for the months and one for the States. However, even if certain variables are dropped from the regression coefficients does not imply that they have dropped from the model. Correspondingly, the omitted variables are included in the intercept of the regression. The presence of an intercept ensures that there is indeed a zero condition mean of errors and thus, the third condition is satisfied.

$E(u_i|x_{ik}) = 0, \textit{ for all regressors}$

4) No perfect collinearity

   In the model developed, each of the categorical variable has been converted into a dummy variable. However, in order to avoid the dummy variable trap, a variable of each type has been dropped. For instance, in the case of months, the dummy variable for April has been dropped in the model. Similarly, for the categorical variable ‘State’, the state of Arizona has been dropped. This is done by the linear regressor automatically and the analysis of the regression coefficients confirms this observation.

Thus, all the four Gauss Markov Assumptions have been satisfied, concluding that our model is indeed robust.

Conclusion

The modeling of Temperature and Relative Humidity through several models has been instrumental in analysing the trend of Pollutants and Atmospheric Parameters over the years. From the analysis of the fitted data, it can be proclaimed that the Temperature and Relative Humidity is dependant, not just on the pollutant, but also the spatial and temporal conditions. For instance, even though the concentration might play a visible atmospheric role in influencing the natural parameters, other factors, inherently present in the state and time variables have a significant effect.

As for the trends of Pollutants over time are concerned, there are certain peaks in the graphs, which indicate two possibilities:

1) There were sudden changes in the atmospheric conditions, leading to rather detrimental effects.
2) As the data is averaged over months and state, there might be a case that the values recorded for a particular time period or a particular state were higher or lower, which leads to an unwanted spike in the curve.

In order to have a detailed look at the parameters, the values for each variable (independant or dependant) were split into state and time, in order to obtain a detailed output. From the analysis of the regression coefficients, it can be judged, as to which state has a larger effect on the data. Apart from correlating the atmospheric variables to independent variables, an additional study of the number of samples recorded was carried out, which gave out interesting results.

Assuming that all samples are recorded by the same data collection technique, certain patterns in the ideal number of samples for a state can be understood. In the report, the spatial distribution for number of samples collected was mentioned. It can be concluded that the behavior of the optimal sample number was not constant, and varied from state to state. Certain spatial clusters could be observed upon deeper scrutiny (example of high NO_2 samples in the northern United States). Thus, although the number of samples collected are considered insignificant in the concentration of pollutants, it plays a major role in the spatial analysis of a region.

Through the mentioned project, trends in a particular pollutant over time was observed. Moreover, with the filtering techniques employed in this study, it is possible to dig deeper into

the data and find trends for each pollutant for every state and the corresponding city. In totality, this project stresses upon correlating atmospheric characteristics to the physical particles.
This study can be employed universally to any region in the world, provided it has ample access to valid and consistent data. Although the project is based on analysing the concentrations of a handful of pollutants, it can be further extended to other gaseous intensive studies such as fuel emissions, industrial emissions or natural phenomenon like forest fires. Moreover, additional environmental parameters like Salinity, Turbulence and Productivity can also be studied, with easy access to data.

```
---
title: "Appendix"
title: "Analysis and Modeling of US Pollution Data for '1980-2018'
and its implications on the atmosphere"
title: "R Notebook"
output: pdf_document
---
```

Importing Libraries

```{r}
setwd("C:\\Users\\yashg\\OneDrive\\Desktop\\CMU\\Fall 2019\\Data Science\\Project\\")
library(readxl)
library(ggplot2)
library(tidyverse)
library(lubridate)

data<-read_csv('combined.csv')
```

Plots of individual pollutants are plotted with time using individual files of the concentrations over time.

```{r}
SO2<-read_csv('SO2.csv')
CO<-read_csv('CO.csv')
NO2<-read_csv('NO2.csv')
O3<-read_csv('Ozone.csv')
T<-read_csv('T.csv')
RH<-read_csv('RH.csv')
```

Each file was further grouped over years and plotted. The plots are mentioned in the report.

```{r}

data_f<-data%>%
  group_by(State_Name)%>%
  group_by(City_Name)%>%
  group_by(Year)%>%
  mutate(avgS=mean(Avg_SO2),
         avgCO=mean(Avg_CO),
         avgO3=mean(Avg_O3),
```

    avgNO2=mean(Avg_NO2),
    avgT=mean(Avg_T),
    avgRH=mean(Avg_RH)
)
#1 (Sample Plot)
`ggplot()+
  geom_line(mapping=aes(x=Year,y=avgRH),data=data_f,color='red')`+
  xlab("Years")+
  ylab("Relative Humidity")+
  ggtitle("Fluctuation of RH over years")+
  theme(plot.title = element_text(hjust = 0.5))
```

Regression Models:

Linear Model (As the number of variables was high, the output was exported to a .csv file)
```{r}
linear_T<-lm(Avg_T~Avg_CO+Avg_NO2+Avg_O3+Avg_SO2+Year+Month+State_Name,data=data)
summary(linear_T)
plot(linear_T)

linear_RH<-lm(Avg_RH~Avg_CO+Avg_NO2+Avg_O3+Avg_SO2+Year+Month+State_Name,data=data)
summary(linear_RH)
plot(linear_RH)

```

The Log-level and Log-Log transforms were carried out as follows:
```{r}
levellog_T<-lm(log(Avg_T)~Avg_CO+Avg_NO2+Avg_O3+Avg_SO2+Year+Month+State_Name,data=data)
summary(levellog_T)
write.csv( tidy(levellog_T) , "levloT.csv" )

levellog_RH<-lm(log(Avg_RH)~Avg_CO+Avg_NO2+Avg_O3+Avg_SO2+Year+Month+State_Name,data=data)
summary(levellog_RH)
write.csv( tidy(levellog_RH) , "levellogrh.csv" )

loglog_T<-lm(log(Avg_T+0.0001)~log(Avg_CO+0.0001)+log(Avg_NO2+0.0001)+log(Avg_O3+0.0001)+log(Avg_SO2+0.0001)+log(Year)+Month+State_Name,data=data) summary(loglog_RH)

loglog_RH<-lm(log(Avg_RH+0.0001)~log(Avg_CO+0.0001)+log(Avg_NO2+0.0001)+log(Avg_O3+0.0001)+log(Avg_SO2+0.0001)+log(Year)+Month+State_Name,data=data)
summary(loglog_RH)

```

In order to assess the spatial variance of number of samples:

```{r}
s1<-lm(num_x~State_Name,data=data_f)
s2<-lm(num_y_1~State_Name,data=data_f)
s3<-lm(num_x_1~State_Name,data=data_f)
s4<-lm(num_x_2~State_Name,data=data_f)
write.csv( tidy(s4) , "pollutant4.csv" )
```
```