Exploratory Data Analysis of US Air Quality

Xuemao Zhang*

Abstract

What are the patterns of U.S. air quality over years? The data set Air Quality Measures on the National Environmental Health Tracking Network available on `https://www.data.gov/` is analyzed in this paper. This data set about ozone and particulate matter (PM2.5) contains data from approximately 4,000 monitoring stations around the country from 1999 to 2013. The tools of map visualizations, cluster analysis and longitudinal data analysis are applied. It is found that Eastern United States and California have highest PM2.5 levels, while the Central United States and Hawaii have lowest PM2.5 levels. California has largest, over 20, DOZ (Number of days with maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard), while all other areas in US have average DOZ 10 or less. Moreover, California has largest, over 3, PRPM (Percent of days with PM2.5 levels over the National Ambient Air Quality Standard), while other areas in US have average DOZ 2 or less. Furthermore, it can be seen that the overall air quality of US has been improved over the 15 years that the data were collected.

Key Words: Air quality, cluster analysis, longitudinal data analysis, map data visualization

1. Introduction

The "Air Quality Measures on the National Environmental Health Tracking Network" dataset (`https://catalog.data.gov/dataset/`) compiles various measures of air pollution collected by approximately 4,000 monitoring stations around the United States. The EPA (the Environmental Protection Agency) maintains a database called the Air Quality System (AQS) which contains data from these monitoring stations and the data from the AQS is considered the gold standard for determining outdoor air pollution. The air pollution data set is about ozone and particulate matter (PM2.5) provided to Centers for Disease Control and Prevention (CDC) for the Tracking Network by EPA. A Downscaler statistical model(Holland, D., n.d.)was used to predict air pollutant levels in rural areas due to low coverage of the monitoring systems.

The Downscaler model combines output from the Community Multi-Scale Air Qaulity Model (CMAQ), a gridded atmospheric model developed by the EPA, and point air pollution measurements. CMAQ estimates are subject to calibration error and monitoring data have both missing and sparsely collected data, but fusion of the two sets of data accounts for the resulting bias. Therefore, the Downscaler model provides better fine-scale predictions of levels of air pollutants at both local and community scales. The variables measured in the data set are listed in the following.

DOZ, Number of days with maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard. The National Ambient Air Quality Standard (NAAQS) set by the EPA for Ozone concentration averaged over 8 hours is 0.070 parts per million (ppm) (United States Environmental Protection Agency, 2016).

*East Stroudsburg University, 200 Prospect Street, East Stroudsburg, PA 18301

$PRPM$, Percent of days with PM2.5 levels over the National Ambient Air Quality Standard. The NAAQS set by the EPA for PM2.5 averaged over one year is 12.0 micrograms per cubic meter of air ($\mu g/m^3$) (United States Environmental Protection Agency, 2016).

$PDOZ$, Number of person-days with maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard. As opposed to a 24 hour day, a person-day is generally speaking an 8 hour day, reflecting the work time of one person during a day.

$PDPM$, Person-days with PM2.5 over the National Ambient Air Quality Standard.

$PMAV$, Annual average ambient concentrations of PM2.5 in micrograms per cubic meter. Rather than a discrete count of days (1, 2, 3, etc.) or a proportion derived from such a count, this variable is continuous (ex. 6.320 μg/m^3).

The air quality data set from 1999 to 2013, including monitor only data and monitor & modeled data from over 1,106 counties in the 50 states and the District of Columbia. The R package dplyr was used to conduct data manipulations. In section 2, the distributions of monitor only and monitor & modeled data are compared and the two types of data are merged for subsequent analysis due to their similarity.

In section 3, the air quality data are visualized by maps which are colored by states. Furthermore, the states are grouped according to relative average pollutant levels using cluster analysis. Longitudinal analysis by genearalzied estimating equations in section 4 shows that the pattern of air quality had been improved over the 15 report years. Conclusions follow in section 5.

2. Distributions of monitor only data and monitor & modeled data

In general, data were collected from monitor only and monitor & modeled as can be seen from Figure 1. But for the report year 1999, 2000, 2012 and 2013, the Downscaler statistical model was not used and data were collected from the monitoring stations only. For example, Figure 2 shows the distribution of the monitoring stations in 2013. Furthermore, missing values for each variable are removed in the following data analysis. The R packageg gplot2 (Wickham, 2010) and usmap(Lorenzo, 2019) are used to generate the graphs in this section.

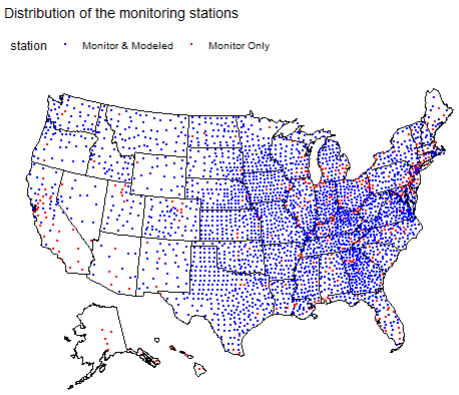

Figure 1: Distribution of the all monitoring stations, Year 1999-2013

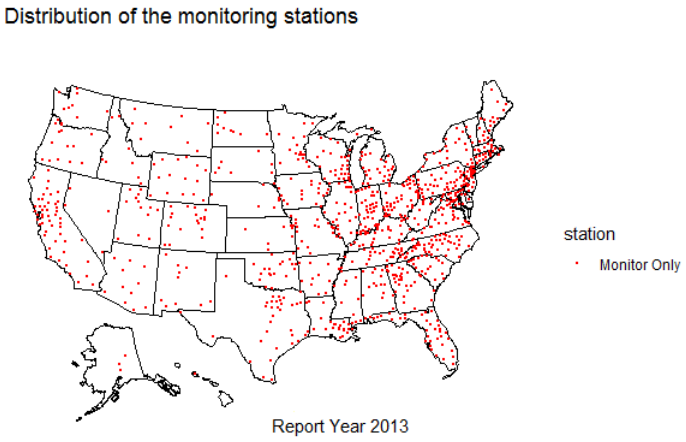

Figure 2: Distribution of the monitoring stations in 2013

Monitor only data is lacking in representation of rural areas. Therefore,representing air quality data using monitor only data could be misleading. It is of interest to combine monitor data and monitor & modeled data for visualization and analysis of air quality between states and over time. For this, we compare the distribution of the five variables for monitor data and monitor & modeled data. It can be seen from Figure 3 to Figure 5 that the distributions of monitor only & monitor and modeled data for all five variables are very similar. Therefore, all data analysis in this paper is based on the complete data where monitor only & monitor and modeled data are combined

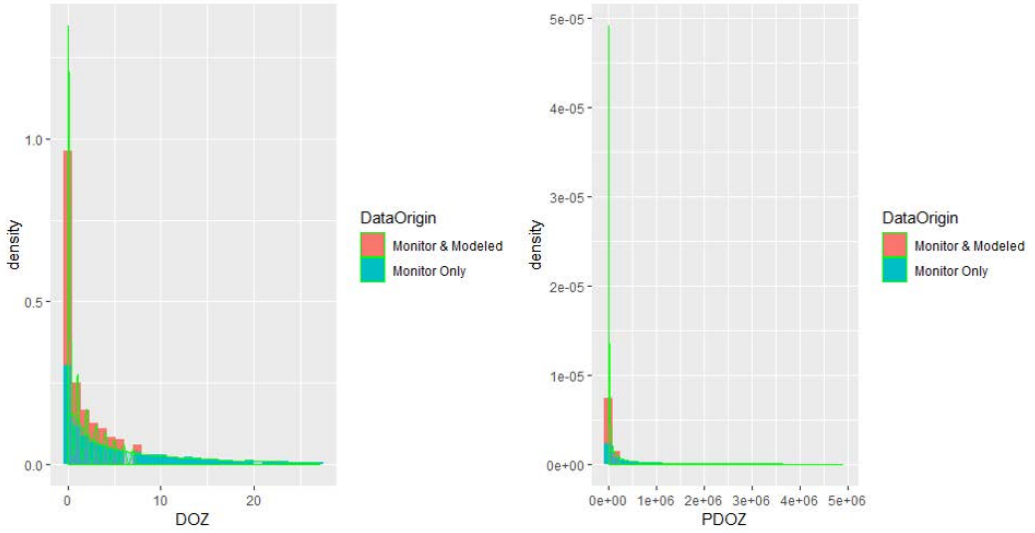

Figure 3: Distribution of monitor only & monitor and modeled data for variables DOZ and PDOZ

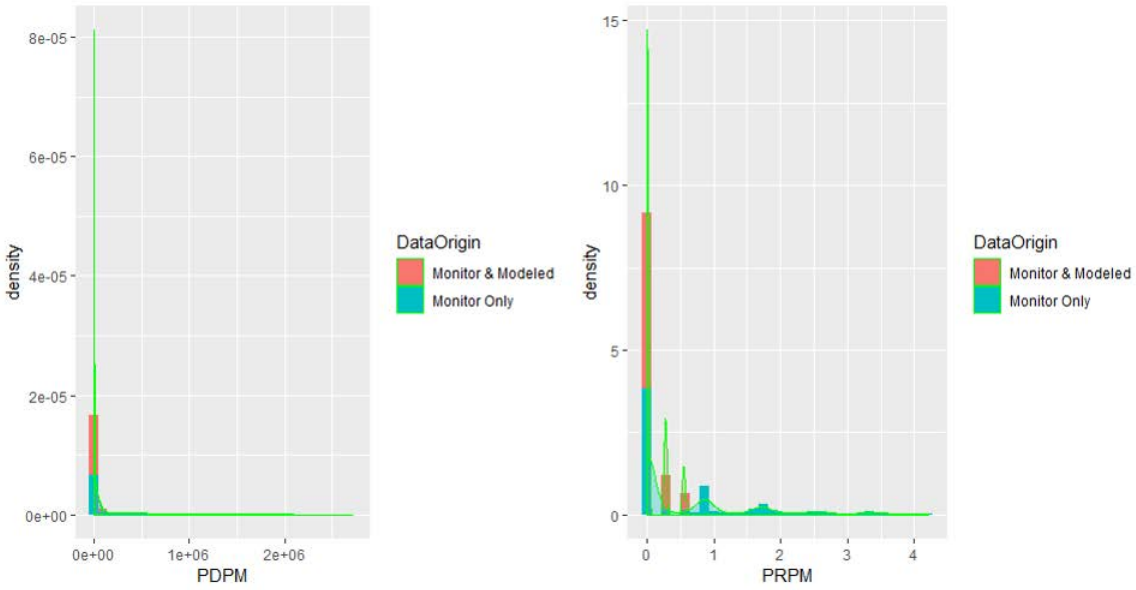

Figure 4: Distribution of monitor only & monitor and modeled data for variables PDPM and PRPM

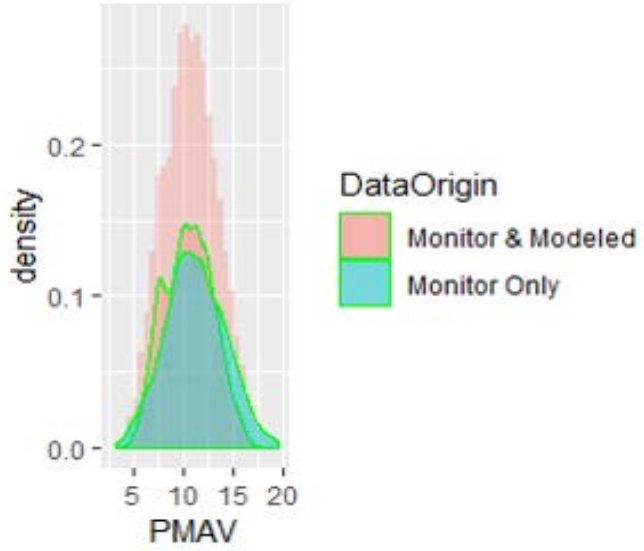

Figure 5: Distribution of monitor only & monitor and modeled data for variable PMAV

The distribution of the variable PMAV is unimodal and about symmetric. But the distributions of the first four variables, DOZ, PDOZ, PDPM and PRPM, are highly skewed to the right. Furthermore, the similarity of the distributions of the variables can be seen from their summary statistics as well. The summary statistics of the monitor only and monitor & modeled data are analyzed and compared, as shown in Table 1, rounded to two decimal places.

Table 1: Comparison of Monitor Only and Monitor and Modeled data

variable	data source	min	max	Q_1	median	Q_3	mean	s.d.
DOZ	Monitor only	0	146	0	3	11	8.21	13.45
	Monitor & modeled	0	144	0	1	5	4.11	8.13
PDOZ	Monitor only	0	33.23	0	0	1.68	1.27	2.47
	Monitor & modeled	0	32.06	0	0	0.55	0.48	1.12
PDPM	Monitor only	0	1.08×10^{9}	0	3.82×10^{5}	1.96×10^{6}	5.00×10^{6}	3.45×10^{7}
	Monitor & modeled	0	1.08×10^{9}	0	1.49×10^{4}	2.15×10^{5}	1.28×10^{6}	1.71×10^{7}
PRPM	Monitor only	0	1.03×10^{9}	0	0	1.08×10^{6}	2.86×10^{6}	6.02×10^{5}
	Monitor & modeled	0	8.18×10^{8}	0	0	6.76×10^{4}	5.86×10^{5}	3.44×10^{4}
PMAV	Monitor only	0	51.20	9.07	11.06	13.30	11.23	3.23
	Monitor & modeled	3.67	30.35	8.75	10.66	12.43	10.64	2.57

3. Visualization of the air quality variables

3.1 Map visualization

Choropleth maps in this section are used to display the five air quality variables by state and year. The 50 states are colored in proportion to the statistical mean of each variable. The choropleth maps provide an easy way to visualize how each measurement varies across the united states. The center of measurement mean instead of median is used so that the outliers of each variable are counted. The differences among the 50 states for each variable can not be detected from the maps otherwise. The R package fiftystater (Murphy, 2016) is used to generate the maps in this section.

The maps for PMAV, shown in Figure 6, represent the distribution of PM2.5 pollution. It can be seen that in general the east US and California generally have higher PM2.5 pollutant. The PM2.5 pollution was higher for the year 1999 and 2000 compared to the maps in other report years. Alaska was mostly polluted in 2004, 2012 and 2013 due to serious wildfire in those years. This pattern in general is true for the distribution of other variables as can be seen from Figures 7 - 10.

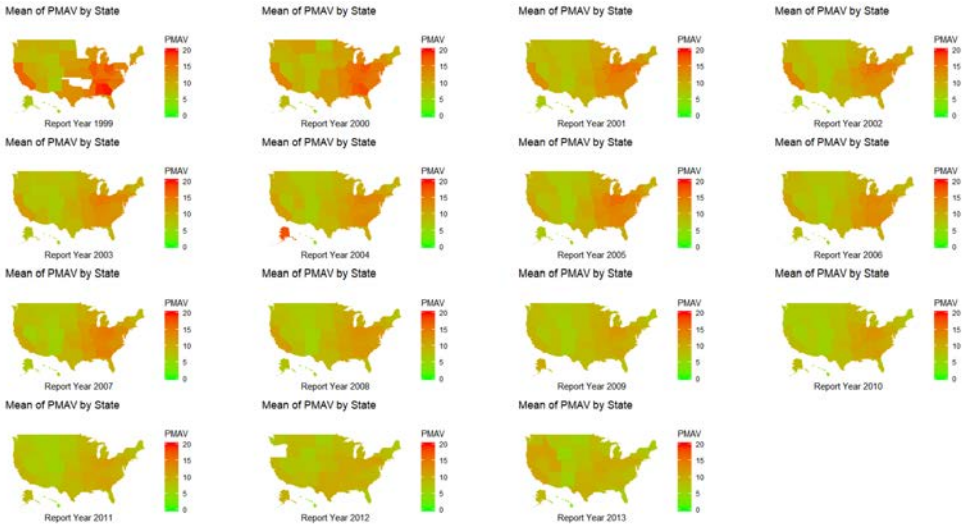

Figure 6: Distribution of PMAV by state and year

The two variables DOZ and PDOZ are counts of days above the NAAQS standard by any amount. It can be seen from Figure 7 and Figure 8 that California has more days and person-days of with maximum 8-hour average ozone concentration over the NAAQS standard per year in most years. The measures from other states are much lower. The percentage differences generally are about 50% which is significant.

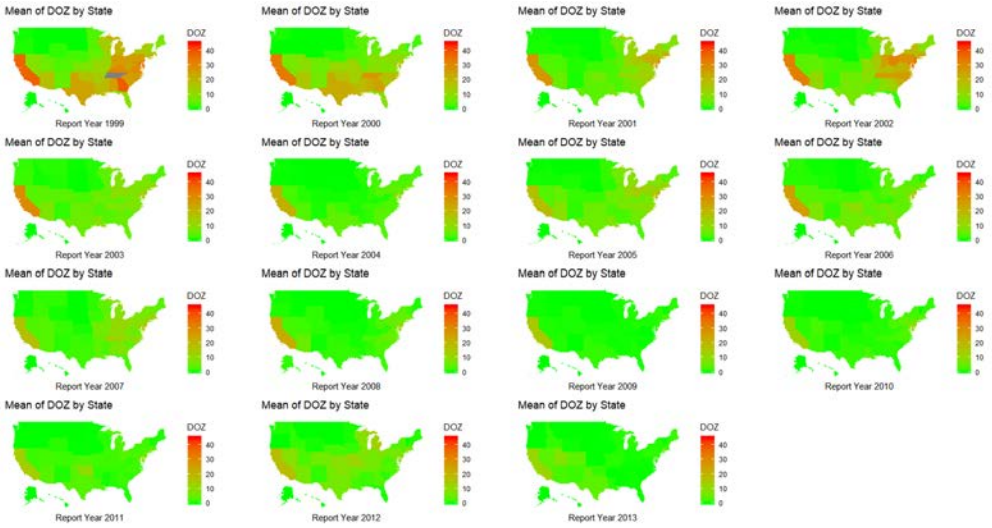

Figure 7: Distribution of DOZ by state and year

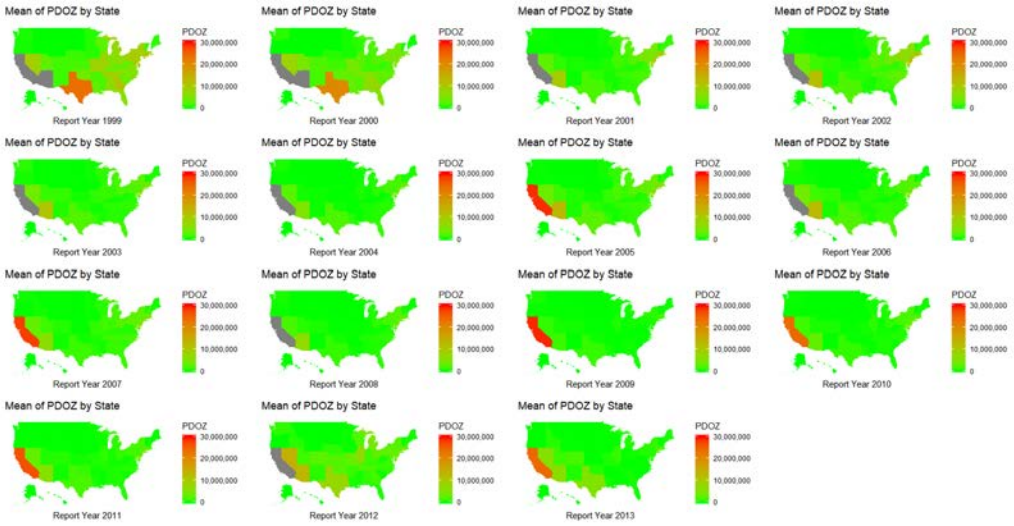

Figure 8: Distribution of PDOZ by state and year

PDPM and PRPM are percent of days/person-days with PM2.5 levels over the NAAQS standard. Again, California stood out in most years which can be seen from Figure 9 and Figure 10.

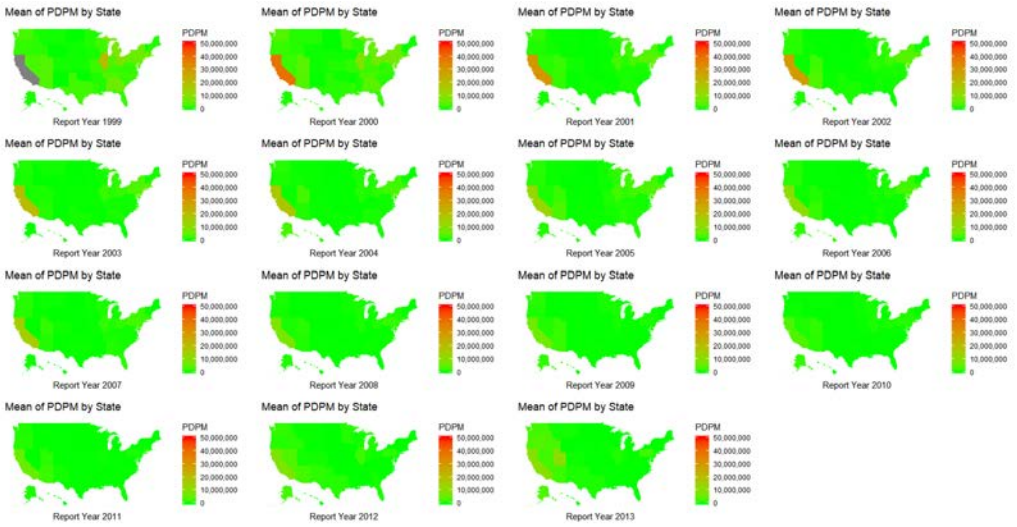

Figure 9: Distribution of PDPM by state and year

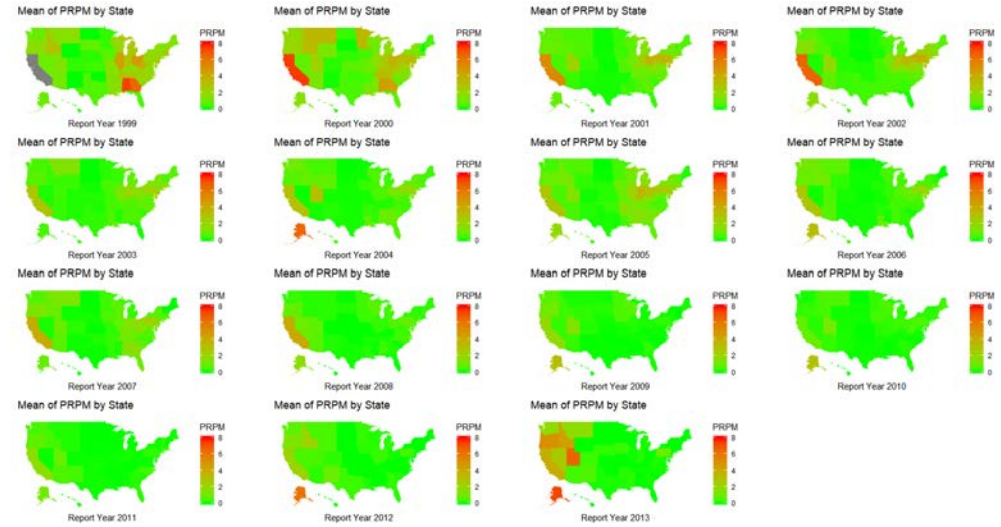

Figure 10: Distribution of PRPM by state and year

3.2 Cluster analysis of air quality

Cluster analysis is a unsupervised learning method to group together subsets of observations in a data set based on the similarity or distance between those observations (Ramachandran and Tsokos, 2009). The air quality combining all the five variables for the 50 states are grouped into three(low, medium and high pollution) clusters using the K-means method(Rencher and Christensen, 2012, p. 532).

The cluster analysis was conducted for each year with results not shown here. In most years, California stood out, the east US was more polluted than the middle US due to more developed industries. The cluster analysis combing all 15 years' data is summarized in Figure 11.

Figure 11: cluster analysis

Furthermore, Figure 12 shows the pairwise scatter plot and correlation among the five air quality variables. It is not surprising that PDPM, PRPM and PMAV are highly correlated and DOZ and PMAV are highly correlated. The R package GGally (Schloerke etc., 2018) is used to generate the scatter plot matrix.

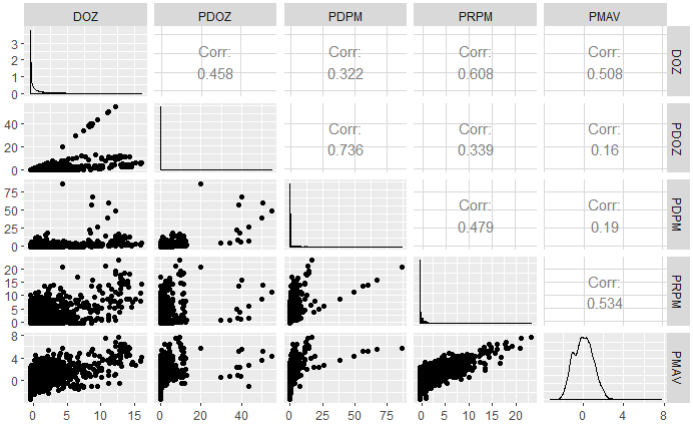

Figure 12: pairwise scatter plots and correlations among the air quality variables

4. Longitudinal data analysis

The trend of air quality can not be easily seen by map visualization. In this section, the method of Generalized estimating equations (GEE) (Liang and Zeger 1986, Zeger and Liang 1986, Horton and Lipsitz, 1999) is used to analyze the longitudinal clustered data for the 50 states air quality over 15 years. In the analysis, each state is regarded as a subject such that the air quality measurements for each state are correlated while measurements among the states are assumed to be independent. Moreover, since the working correlation structure in GEE does affect the consistency of the parameter estimates, a working exchangeable correlation structure is used.

It can be found from Figure 13 - 15 that there is a negative trend over time for all five variables as shown by the linear regression fit. That is, the US air quality had been improving over the 15 year period.

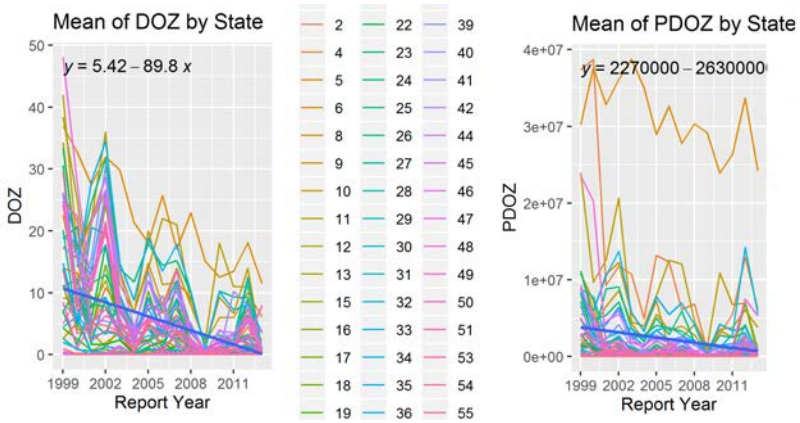

Figure 13: Visualization of longitudinal mean of DOZ and PDOZ by state

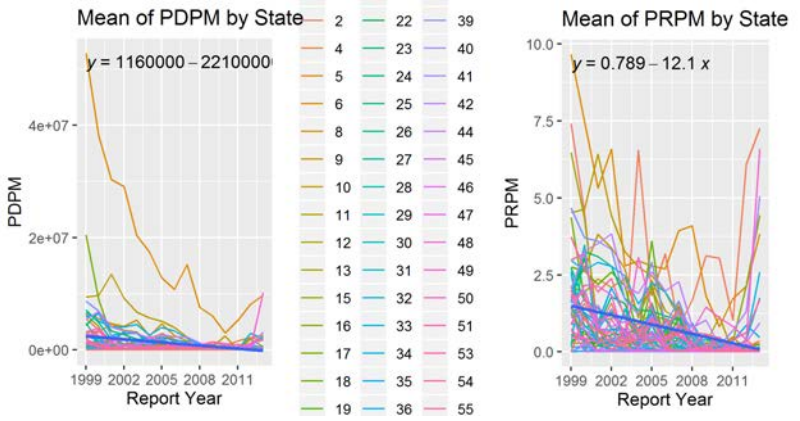

Figure 14: Visualization of longitudinal mean of PDPM and PRPM by state

Furthermore, it is easy to see that some states generally stood out due to low air quality over years. In Figure 13 and Figure 14, the top orange curve is for California. The spike in 2004 of both PRPM and PMAV is for Alaska. It is hard to tell the exact StateFips for the other top curves. The summary statistics for the 50 states over 15 years will help one identify all the time series curves in these graphs. The tables are too long to be listed here.

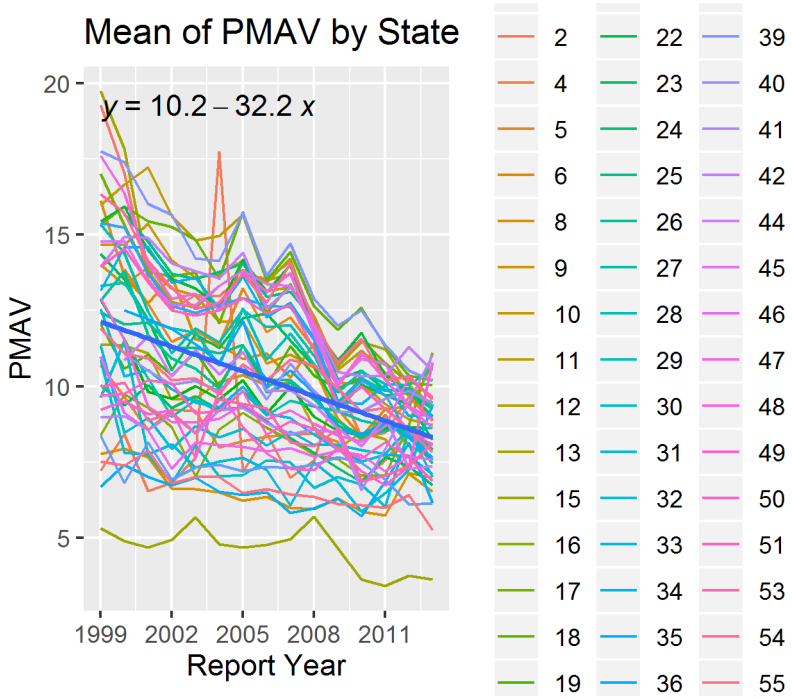

Figure 15: Visualization of longitudinal mean of PMAV by state

5. Conclusions

In this paper, I conduct exploratory data analysis of levels of air pollution in the United States from 1999 to 2013. Five air quality variables are analysed. Data visualization is the main tool to detect the pattern of 50 states over 15 years. Map visualizations and cluster analysis show that Eastern United States and California have higher PM2.5 levels, while the Central United States and Hawaii have lowest PM2.5 levels. California has largest, over 20, DOZ (Number of days with maximum 8-hour average ozone concentration over the NAAQS Standard), while all other areas in US have average DOZ 10 or less. Moreover, California has largest, over 3, PRPM , while other areas in US have average DOZ 2 or less. Overall, the central US had highest air quality, then Eastern US and last California stood out. Furthermore, The air quality was unusually low for some year (e.g. 1999). Longitudinal data analysis shows that the overall air quality of US had been improved over the 15 years. The correlations between some variables are strong (e.g. PDPM, PRPM and PMAV, and DOZ and PMAV).

References

[1] Holland, D. *Downscaler Modeol for predicting daily air pollution.* Retrieved from https://www.epa.gov/air-research/downscaler-model-predicting-daily-air-pollution

[2] Horton, N.J. and Lipsitz, S.R. (1999). *Review of Software to Fit Generalized Estimating Equation Regression Models*, The American Statistician, **53** (2), 160–169.

[3] Lorenzo, P.D. (2019). *US Maps Including Alaska and Hawaii.* https://github.com/cran/usmap.

[4] Liang, K.Y. and Zeger, S. L. (1995). *Inference based on estimating functions in the presence of nuisance parameters. Statistical Science*, **10**, 158–173.

[5] Murphy, W. (2016). *fiftystater: Easy 50 state maps for R & ggplot2.* https://github.com/wmurphyrd/fiftystater.

[6] Ramachandran, K.M. and Tsokos, C. P. (2009). *Mathematical Statistics with Applications.* Burlington, Massachusetts: Elsevier Academic Press.

[7] Rencher, A.C. and Christensen, W.F. (2012). *Methods of Multivariate Analysis, Third Edition*, John Wiley & Sons, Inc.

[8] Schloerke, B., Crowley, J., Cook, Di, Hofmann, H., Wickham, H., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Larmarange, J. (2018). *GGally: Extension to 'ggplot2'. R package version 1.4.0.* https://cran.r-project.org/web/packages/GGally/.

[9] Wickham, H. (2010). *ggplot2: Elegant Graphics for Data Analysis.* Journal of Statistical Software. **35** (1), 1–3.