**Air Pollution Interpolation and Clustering of New York and Surrounding States for 2013**

**Christian Geofroy 1,Jong Su Kim 2, Tausif Reshad Ahsan3  and** 4 **Ziran Jeffrey Zhou**

1 1003193723; christian.geofroy@mail.utoronto.ca

2 1004410648; jongsu.kim@mail.utoronto.ca

3 1001863487; tausif.ahsan@mail.utoronto.ca

4 11002425477; jeffrey.zhou@mail.utoronto.ca

Start: March 11, 2020; Due: April 7, 2020

**Abstract:** Air quality is known to be a key factor in affecting the wellbeing and quality of life of the general populace. Utilizing the data available from EPA this study investigates the spatial clustering and interpolation of the air pollution data from New York, Vermont, Massachusetts, Pennsylvania, and Connecticut. Inverse distance weighted interpolation was done and cross-validated with the LOOCV method. Kriging was done using a spherical variogram. Clustering was done using a k-means of 4 by using the elbow method. The interpolation resulted in the higher levels of pollutants across the state of Pennsylvania and hotspots in areas of high population such as New York City.

**Keywords:** AirPollution, Interpolation, Clustering

**1. Introduction**

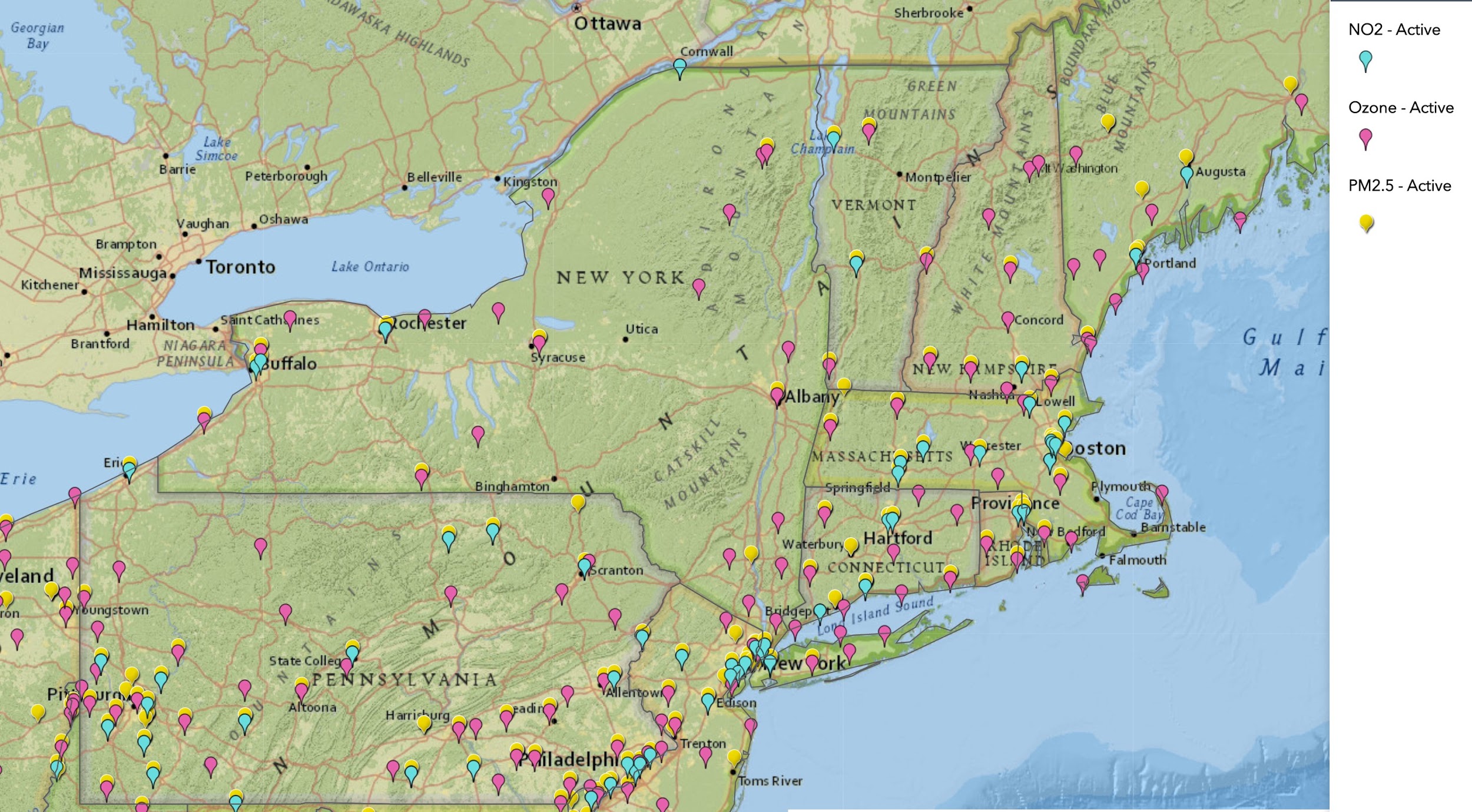
Air pollution is a global issue which has exposed a significant proportion of the world’s population to adverse health effects and continues to get worse every year. Despite having relatively less harmful levels of air pollutants than countries with the most critical levels of pollution, such as India and China, the issue is still significant in the US. Air pollution has been linked to chronic obstructive pulmonary disease, cardiovascular disease, and lung cancer, and is the leading cause of worsening asthma conditions among other health risks[1].

This study aims to analyze spatial patterns of air pollution using statistical methods. Inverse distance weighting and kriging interpolation of air pollutant data to predict levels of pollution throughout the states from data provided by the EPA, collected at specified monitoring stations. Then the aggregated data points were analyzed to determine clusters of air pollutant groups within the study area. Utilizing these methods, we hope to model a spatial pattern of pollution in the state of New York for the year 2013. This can further add to existing research on air pollution throughout the US and beyond, and contribute to a wider model of air pollution patterns that can be applied on a broader scale.

**2. Methods**

*2.1 Study Area*

For our study region we were assigned New York State as the base for our investigation. We then chose Pennsylvania, Connecticut, Massachusetts and Vermont as the other four surrounding states to measure along with New York. The climate of this region is a mix between the Koppen Climate zones of; Dfa, Dfb, Cfa and Cfb, which means that the more inland places have warm or hot summer continental climates whereas the places close to the coastline have oceanic or humid subtropical climates. The population of the combined states is roughly 43.3 million as of 2019, with the biggest centers of population being the cities of New York City, Philadelphia and Boston. The geography of the region is dominated by the northernmost part of the Appalachian Mountain Range, which is mostly located in Pennsylvania and New York. Along the coast the geography changes to be coastal forests and there are many islands along the coast. We selected these states because we wanted to have the states with the highest population that surround New York for more accurate testing of urban vs. rural pollutant levels.



*Figure 1:* Map of the states included in our study area, along with the locations of the monitoring stations (Source: EPA AirData Air Quality Monitors ESRI Map)

*2.2 Data*

The data used for our research came from the United States’ Environmental Protection Agency’s Air Quality Datasets for Annual Concentrations by Monitors for 2013. This data includes the data from every monitor location for the entire year. The dataset was then sorted to suit our needs in focusing on the specific three parameters that we needed to study (ozone, nitrogen dioxide and PM 2.5). The data collected in this region is highly skewed for the urban areas closer to the coastline with larger population centers so this could lead to an observation bias, where the clustering would occur in the higher populated areas. The data is also much more comprehensive for PM2.5 with more than five times the number of monitoring stations compared to nitrogen dioxide stations. Vermont also has a lack of stations, most likely due to the population being much less than the other surrounding states.

| Type of Pollutant: | Ozone | Nitrogen Dioxide | PM 2.5 |
| --- | --- | --- | --- |
| Monitor Station in NY | 32 | 5 | 46 |
| Monitor Station in Pennsylvania | 58 | 22 | 136 |
| Monitor Station in Massachusetts | 17 | 12 | 40 |
| Monitor Station in Connecticut | 12 | 5 | 24 |
| Monitor Station in Vermont | 2 | 2 | 14 |
| Total number of monitor stations | 121 | 46 | 260 |

*Table 1:* Table of the distribution of monitoring locations for the three pollutant types chosen.

*2.3 Statistical Analysis*

2.3.1 IDW Interpolation

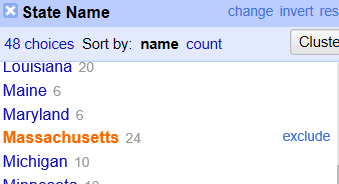
IDW was done by first filtering the raw data acquired from the EPA. We first filtered by year, along with the specific variables we needed. Then the specific states in our study area were filtered out (Pennsylvania, Massachusetts, New York, Vermont, Connecticut). The specific metrics of the pollutant variables were filtered out. Finally three separate files were created for each of the three pollutants that were being studied. Spatial points files were created from the latitude/longitude and then spatial points data frames were created from that. The three variables needed to be standardized using a coordinate reference system, and then the IDW was done for each variable.

2.3.2 Kriging

Kriging was first done by taking the data acquired from the EPA and filtering it like how the IDW was done. A spatial points data frame was then created and the sample variogram was fitted. We chose a spherical base variogram to start with. The grid of points was then created and the ordinary kriging could be done.

2.3.3 Clustering

Clustering was done by first filtering the data that was acquired from EPA. It was important to filter by the state of interest to select Pennsylvania, Massachusetts, New York, Vermont, Connecticut; parameter names as Nitrogen Dioxide, Ozone, and PM2.5; and metric used. This was done using open source software Open Refine and was exported as a csv file.



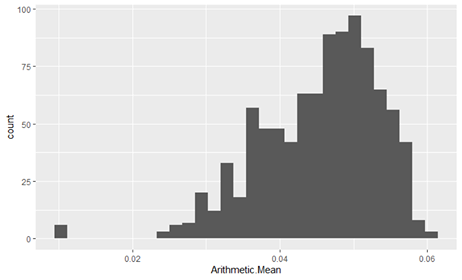
*Figure 2:* Example of the filtering methods used

Using the filtered data, Geoda was used for spatial clustering. The data were plotted using the longitude and latitude and k means clustering was conducted using the the longitude, latitude and arithmetic mean data. Experiments were conducted to find the optimal k value. We experimented with k values from 1-10. In this case, we conducted the elbow method and determined optimal k to be four.

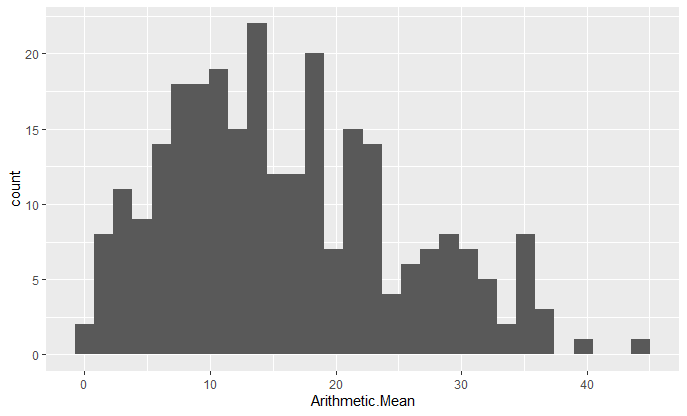
**3. Results**

3.1 Descriptive Statistics

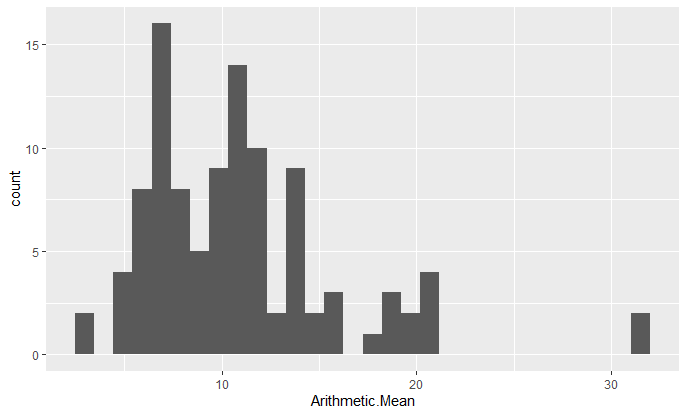
We checked the normality of each variable we are testing and the histogram is below. There is almost normal distribution for ozone, except for the outlier close to 0. The distributions of NO2 and PM2.5 are similar, with both having a positive right-skewed distribution whereas the ozone is left-skewed.



*Figure 3:* Ozone data distribution



*Figure 4*: NO2 Histogram data



*Figure 5:* PM2.5 Distribution

3.2 IDW Interpolation Results

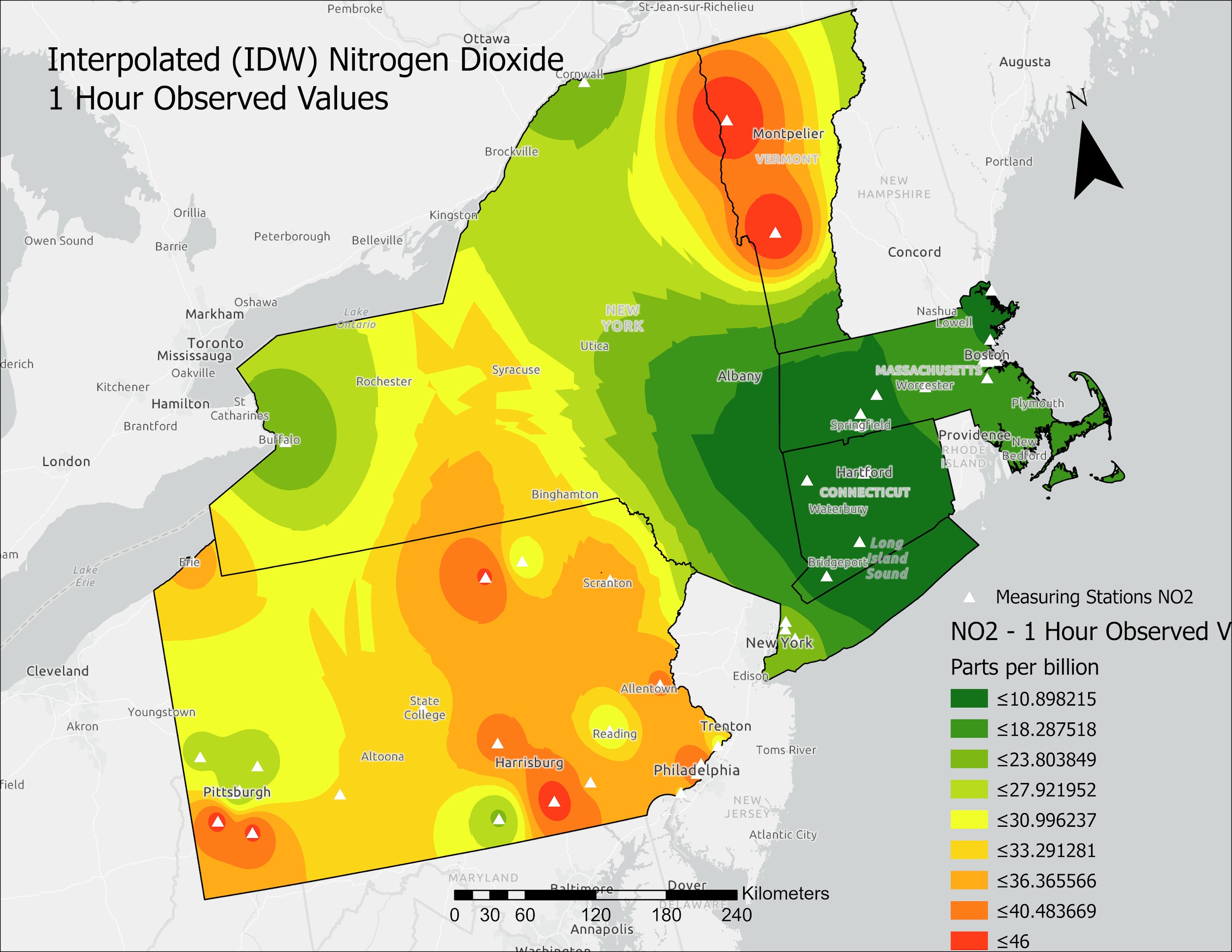
The following maps were created in ArcGIS Pro because of the superior graphics capabilities.



*Figure 6:* Ozone IDW



*Figure 7:* IDW PM2.5

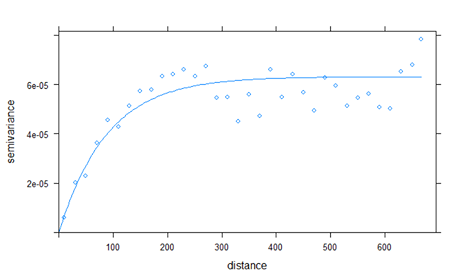


*Figure 8:* IDW of Nitrogen Dioxide

These maps show there is a large concentration of pollutants in the state of Pennsylvania, most likely due to the high amount of manufacturing in the state and also concentrations in the New York City area due to the large population. On the ozone and nitrogen dioxide maps there are also large concentrations in the areas of Montpelier and Bennington in Vermont. These high concentrations are not present on the PM2.5 map.

3.3 Kriging Results

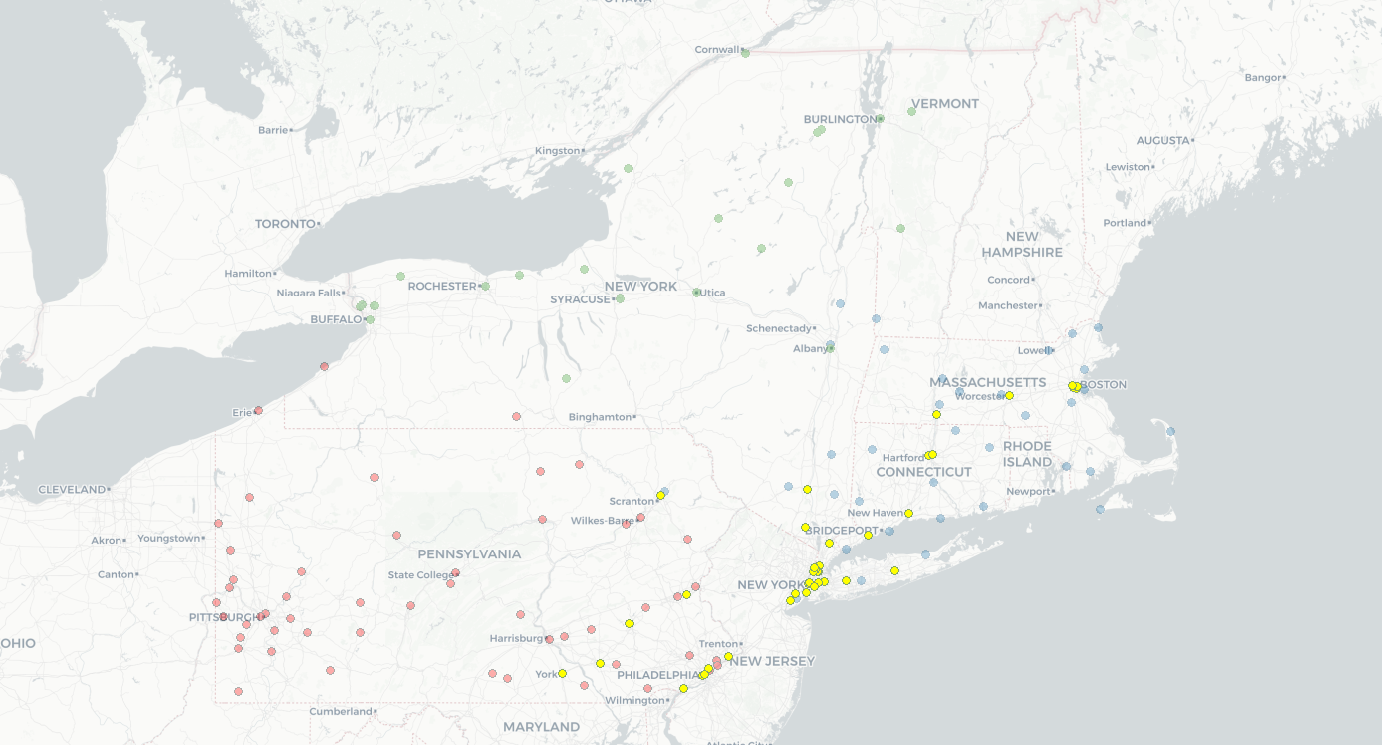
The semivariogram was generated using a spherical model to best fit the data. The range of 6e-05 represents that there is no spatial autocorrelation past this value. The sill of approximately 350-400 indicates that there is lots of variance in the distances of



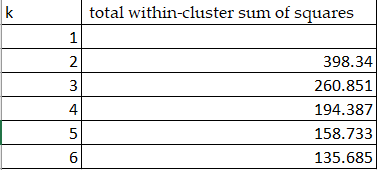
*Figure 9*: Fit Spherical Variogram on all data

3.4 Clustering Results

The clustering results show four clusters as determined by the elbow method and graphing the total within-cluster sum of squares.



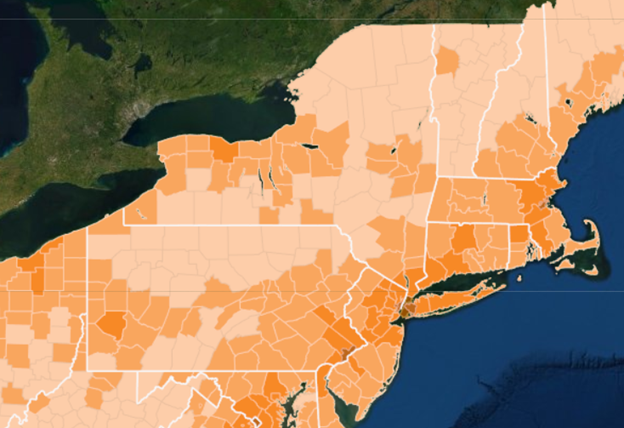
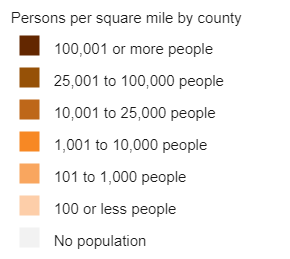
*Figure 10:* Spatial Cluster results by colour.



*Table 2:* Total within-cluster of square value by k value tested.

There are four clusters you can see in the map in green, yellow, red and blue. It is interesting to note the highly clustered amounts of yellow in the New York City Area and the overlapping of blue and red points in the yellow cluster.

This data aligns well with the population density of the United States as displayed in the figure.



*Figure 11:* Population Density Map

It is easy to observe the highest density in the New York City Area and coincides with the yellow clusters and other yellow points in Connecticut, Massachusetts, and Pennsylvania which coincides with major cities in the states like Philadelphia, Pittsburgh, Boston and Hartford. Areas in blue also show a similar pattern and seem to be clustered in areas that have high density as well along the eastern seaboard in areas that have at least 100 persons per square mile. Areas that have a high population density seem to have similar air quality data in surrounding areas. This can clearly be seen in the high density of yellow plot in the Manhattan area and continue to have similar values in the outskirts of the cities like Long Island and New Haven all within a commutable distance from the city.

**4. Discussion**

LOOCV cross-validation should be used as an “exploratory step for assessing the potential benefits of performing a more comprehensive residual correction” such as more stages of IDW or Kriging. We only used one stage of IDW and one stage of cross-validation so that could mean that our data isn't the most accurate it could be [2]. The main areas of clustering of pollutants were in the highest population areas in our study area (New York City, Boston Pittsburgh) and this was to be expected. The higher the population the more pollution that is created from things like manufacturing in Pittsburgh or traffic in New York City. There is also a high amount of clustering in the rural areas of Pennsylvania, most likely due to old manufacturing jobs. This clustering adds to the notion that air pollution affects the poor and uneducated disproportionally more than people with higher economic status [3].

**5. Conclusions**

Intuitively, we can expect clustering of pollutants in high population, urban areas which are likely to have relatively high levels of emission due. For example, the high population density region around New York City and the state of New Jersey indicates a clustering. Our clustering centres can be compared to a 2013 study which looked at pollutant distribution on a nationwide level using the same inputs, which had similar density clusters around the state of New York and neighboring states [4].

The IDW regions for ozone and PM2.5 seem to suggest that highest levels of air pollutants in the New York City metropolitan region, but relatively lower values throughout the rest of New York state. However, this may be due to having a large number of monitoring stations collecting data in the areas showing the highest values for pollutants. Our PM2.5 IDW map was compared to a 2016 study which used a similar IDW method for generating a nationwide pollutant map, which seemed to show similar concentration values of PM2.5 in the state of Pennsylvania and New York City region [5].

**Author Contributions:**

Christian contributed the study area, data sections of methods, helped with IDW and Kriging from the statistical analysis section and wrote the discussion section. Ziran contributed with the IDW interpolation maps. Jong Su contributed with the clustering and kriging computations in the statistical analysis and results sections. Tausif contributed with acquiring and formatting references , introduction and conclusion.

**References**

1. Peden, D. B. (2018). The Unexpected Health Effects of Air Pollution. *North Carolina Medical Journal*, 309-311.
2. Yuval, Levy, I., & Broday, D. M. (2017). Improving modeled air pollution concentration maps by residual interpolation. *Science of The Total Environment*, *598*, 780–788. doi: 10.1016/j.scitotenv.2017.04.117
3. Zou, B., Peng, F., Wan, N., Mamady, K., & Wilson, G. J. (2014). Spatial Cluster Detection of Air Pollution Exposure Inequities across the United States. *PLoS ONE*, *9*(3). doi: 10.1371/journal.pone.0091917
4. Austin, E., Coull, B. A., Zanobetti, A., & Koutrakis, P. (2013). A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. *Environment International* .
5. Lixin, L., Zhou, X., Kalo, M., & Piltner, R. (2016). Spatiotemporal Interpolation Methods for the Application of Estimating Population Exposure to Fine Particulate Matter in the Contiguous U.S. and a Real-Time Web Application. *International Journal of Environmental Research and Public Health*.

**Appendix A**

Appendix\_A.R

#r code to filter code

library(tidyverse)

data2013 <- data %>%

dplyr::select(State.Name,City.Name,County.Name,Latitude,

Longitude, Parameter.Name, Units.of.Measure,

Metric.Used, Sample.Duration, Arithmetic.Mean)

#filter out required states

states <- data2013 %>%

filter(State.Name == "New York"| State.Name == "Pennsylvania"| State.Name == "Connecticut"| State.Name == "Massachusetts"| State.Name == "Vermont" &

Parameter.Name %in% c("Nitrogen dioxide (NO2)", "Ozone",

"PM2.5 - Local Conditions"))

#filter out metric used

state\_data2013 <- states %>%

filter(Metric.Used %in% c("Observed values", "Daily maximum of 8-hour running average", "Daily Mean"))

state\_data2013 <- states %>%

filter(Parameter.Name %in% c("Nitrogen dioxide (NO2)", "Ozone","PM2.5 - Local Conditions"))

summary(state\_data2013)

#histogram plot

ggplot(ozone\_data)+

geom\_histogram(mapping = aes(x =Arithmetic.Mean) )

ggplot(no2\_data)+

geom\_histogram(mapping = aes(x =Arithmetic.Mean) )

ggplot(pm2.5\_data)+

geom\_histogram(mapping = aes(x =Arithmetic.Mean) )

write.csv(state\_data2013, 'C:\\Users\\jsk\\Documents\\ggr376\\state\_data1.csv' )

#dataframe

spoints\_ozone <- SpatialPoints(ozone\_data[7:6], proj4string=CRS("+proj=longlat +datum=NAD83"))

spointsdf\_ozone <- SpatialPointsDataFrame(spoints\_ozone, ozone\_data)

spointsdf\_ozone <- as.data.frame(spointsdf\_ozone, stringsAsFactors = FALSE)

summary(spointsdf\_ozone)

#variogram

install.packages("gstat")

library(gstat)

gs <- gstat(formula=Arithmetic.Mean~1, data = spointsdf\_ozone)

v <- variogram(gs, width=20)

##variogram fit

fvs <- fit.variogram(v, vgm(85, "Sph", 75, 20))

fvs

plot(variogramLine(fvs, 200), type='l', ylim=c(0,4.5e-05) ,col='blue', lwd=2)

points(v[,2:3], pch=20, col='red')

**Appendix B**

Partial IDW code

#import library

install.packages('dplyr')

install.packages('sp')

install.packages('rspatial')

install.packages('maps')

install.packages('ggmap')

install.packages('mapdata')

install.packages('dismo')

install.packages('gstat')

install.packages('rgeos')

library(dplyr)

library(sp)

library(rgdal)

library(ggplot2)

library(maps)

library(ggmap)

library(mapdata)

library(dismo)

library(gstat)

library(rgeos)

#read csv

setwd("D:/GGR376/GP")

data<-(read.csv('annual\_conc\_by\_monitor\_2013/air.csv'))

#filter out necessary data

data2013 <- data %>%

dplyr::select(State.Name,City.Name,County.Name,Latitude,

Longitude, Parameter.Name, Units.of.Measure,

Metric.Used, Sample.Duration, Arithmetic.Mean)

#filter out required states

states <- data2013 %>%

filter(State.Name == "New York"| State.Name == "Pennsylvania"| State.Name == "Connecticut"| State.Name == "Massachusetts"| State.Name == "Vermont" &

Parameter.Name %in% c("Nitrogen dioxide (NO2)", "Ozone",

"PM2.5 - Local Conditions"))

#filter out metric used

state\_data2013 <- states %>%

filter(Metric.Used %in% c("Observed values", "Daily maximum of 8-hour running average", "Daily Mean"))

summary(state\_data2013)

#separate by air type

ozone <- state\_data2013 %>%

filter(Parameter.Name == "Ozone")

nitrogen\_dioxide <- state\_data2013 %>%

filter(Parameter.Name == "Nitrogen dioxide (NO2)")

pm2.5 <- state\_data2013 %>%

filter(Parameter.Name == "PM2.5 - Local Conditions") %>%

filter(Units.of.Measure == "Micrograms/cubic meter (LC)")

#converting data to spatial points data frame

#EPSG 2163 is the US National Atlas Equal Area projection

ozone\_spdf <- SpatialPointsDataFrame(ozone[5:4], ozone, proj4string = CRS('+init=epsg:2163'))

ozone\_spdf <- as.data.frame(ozone\_spdf, stringsAsFactors = FALSE)

no2\_spdf <- SpatialPointsDataFrame(nitrogen\_dioxide[5:4], nitrogen\_dioxide, proj4string = CRS('+init=epsg:2163'))

no2\_spdf <- as.data.frame(no2\_spdf, stringsAsFactors = FALSE)

pm2.5\_spdf <- SpatialPointsDataFrame(pm2.5[5:4], data= pm2.5, proj4string = CRS('+init=epsg:2163'))

pm2.5\_spdf <- as.data.frame(pm2.5\_spdf, stringsAsFactors = FALSE)

#create boundary for contiguous states

us\_state <- map\_data('state')

us\_east <- subset(us\_state, region %in% c('new york', 'pennsylvania', 'connecticut', 'massachusetts', 'vermont'))

#check if boundary is correct

study\_region <- ggplot(data=us\_east, mapping = aes(x = long, y = lat, group=group))+

geom\_polygon(color = 'blue', fill ='gray')

study\_region

#converting map boundary to spatial points data frame

east\_bound\_spdf <- SpatialPointsDataFrame(us\_east[1:2], us\_east, proj4string = CRS('+init=epsg:2163'))

#NAEA = National Atlas Equal Area

NAEA <- CRS('+init=epsg:2163')

#transforming to identical CRS

oz\_spdf <- SpatialPointsDataFrame(ozone[5:4], ozone, proj4string = CRS('+init=epsg:2163'))

ozone\_NAEA <- spTransform(oz\_spdf, NAEA)

no2\_NAEA <- spTransform(no2\_spdf,CRSobj = NAEA)

pm2.5\_NAEA <- spTransform(pm2.5\_spdf,CRSobj = NAEA)

east\_NAEA <- spTransform(east\_bound\_spdf, CRSobj = NAEA)

#IDW for Ozone

ozone\_voro <- voronoi(ozone\_NAEA)

plot(ozone\_voro)

east\_coast <- aggregate(east\_NAEA, FUN = mean, na.rm = TRUE)

plot(east\_coast)

ozone\_boundary <- intersect(ozone\_voro, east\_coast)

spplot(ozone\_boundary, 'Arithmetric.Mean', col.regions=rev(get\_col\_regions()))

spplot(ozone\_boundary)

ozone\_raster <- raster(east\_coast, res=)

vr <- rasterize(ozone\_boundary, ozone\_raster, 'Observation.Count')

plot(vr)