Predicting NO2, O3, PM2.5 and Clustering Regions with Similar Traits in New York and Contiguous States (2016)

Andrew Sud [1], Hardeep Dhatt [2], Antonio LaPlaca [3] and Iman Kewalramani [4]

[1] 1000848782; a.sud@mail.utoronto.ca
[2] 1003069519; hardeep.dhatt@mail.utoronto.ca;
[3] 1003944869; Antonio.laplaca@mail.utoronto.ca
[4] 1003053143; Iman.kumarkewalrama@mail.utoronto.ca

Start: March 12, 2020; Due: April 7, 2020

Abstract:

This study focuses on the air pollutants NO_2, O_3, and $PM_{2.5}$, with a goal of predicting the levels of each of the air pollutants in our study area where there are a lack of monitoring stations to collect data. This study will be achieving this through Inverse Distance Weighted Interpolation and Kriging Interpolation, then contrasting these two methods to determine which method better predicts each pollutant individually. This study will then use a clustering method, that method of which is determined using statistical analysis, then clustering regions with similar traits of NO_2, O_3, and $PM_{2.5}$ within them. This study will then use previously known, physical real-world sources of air pollutants, and determine whether our results accurately represent these sources, then concluding that our work. Which for NO_2, IDW was found to be a better predictor, while O_3 and $PM_{2.5}$ was more accurate using the kriging method. We found three main cluster regions in our study area. We found these cluster regions to be accurate representations of real-world phenomena, and correlate with other research in this area of study.

Keywords: 1;"Spatial Clustering" 2;"Interpolation" 3;"Air Pollution"

1. Introduction

In this study we will be looking the three different types of air pollutant concentrations, those of which include nitrogen dioxide (NO_2), ozone (O_3), and particulate matter 2.5 ($PM_{2.5}$), within New York State, and four contiguous states (Massachusetts, Vermont, New Jersey, and Pennsylvania). The data collected will be from the year 2016. In this study, we will spatially visualize the concentrations of these air pollutants, and use interpolation techniques to predict values for areas without monitoring stations, so that it may be used to make inferences and connections between sources of pollutants and possible affects on humans, wildlife, and plant life in these areas more accurately. To achieve this we will be creating interpolated air pollution

surfaces for each of the pollutants in five states (New York, New Jersey, Pennsylvania, Vermont and Massachusetts) using various methods. We will also use spatially constrained clustering to identify regions that contain monitors to correlate similarities in air quality using all three pollutants.

2. Methods

2.1 Study Area

The study area will be the five contiguous US states of New York, New Jersey, Pennsylvania, Vermont, and Massachusetts. The population of this area totals more than 48.7 million people, representing 14.9% of the entire population of the United states. Of these states, most of the population is concentrated in New York and Pennsylvania, with a total share of 66.3% of the population of the 5 states combined [1]. The climate of the study area varies greatly. On the high end, coastal parts of Massachusetts and New York experience oceanic and humid subtropical climates, to warm-summer humid continental climates in the northern parts of New York, Pennsylvania, and Vermont on the low end [2]. This variation in climate is reflective of the diverse characteristics of the physical geography of this contiguous region. Three of these states, being New York, New Jersey, and Massachusetts all have significant coastal frontage on the Atlantic, contributing to the humid subtropical climate experienced in those regions. Conversely, the inland regions of upstate New York, much of Pennsylvania and the entire state of Vermont are forested, mountainous areas that experience a milder climate as a result of the Appalachian Mountains winding their way through the area, and generally higher altitude compared to the coastal regions in question. This area was selected because of their contiguity with the State of New York, as well as the climate variation outlined above.

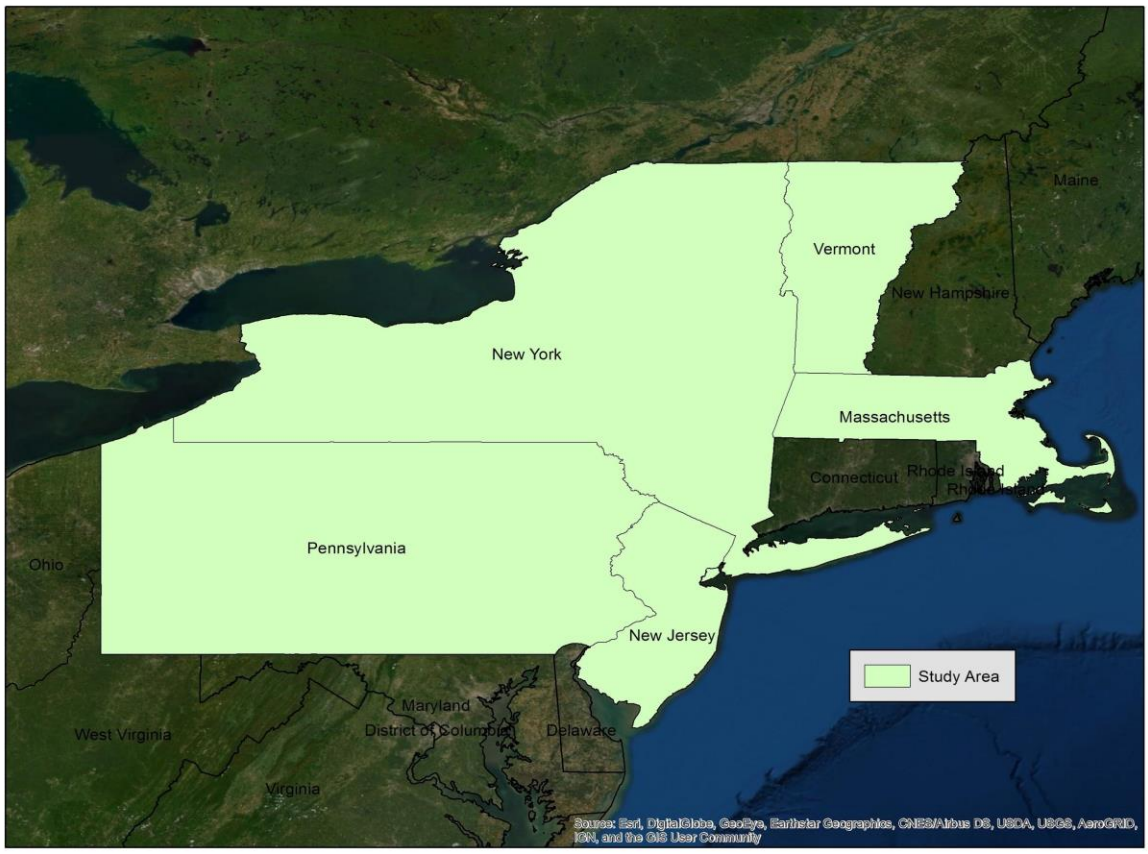


Figure 1. Map of the contiguous study area, comprised of the states of Pennsylvania, New York, New Jersey, Vermont, and Massachusetts

2.2 Data

The EPA dataset used contains observations of concentrations of multiple air pollutants from monitoring stations, as well as atmospheric conditions across the United States, although only those pertaining to the selected pollutants (O_3, NO_2, $PM_{2.5}$) in states included in the study area were retained. The dataset also contains multiple sample durations for pollutant measurements, as well as multiple units of measure used to represent concentrations of pollutants in the atmosphere. The initial dataset also included multiple readings per pollutant per monitoring station (i.e. many readings in a single location). This mix of metrics and sample durations means that the dataset needed to be narrowed down to the specified standards of the study, which are mean observed parts per billion over 8-hour durations for ozone, mean observed part per million over 1-hour durations for nitrogen oxide, and mean observed micrograms per cubic meter over 24-hour

durations for particulate matter. The resulting dataset after filtering for the specific pollution standards contained unique readings per monitoring station per pollutant, which was more suitable for the nature of this analysis. For analysis purposes, sample durations were averaged annually. See table 1 below for monitoring station counts.

Table 1. Number of monitoring stations for each pollutant

Pollutant	Monitor Count
Ozone	120
Particulate Matter	60
Nitrogen Dioxide	48

[1] United States Environmental Protection Agency.

2.3 Statistical Analysis

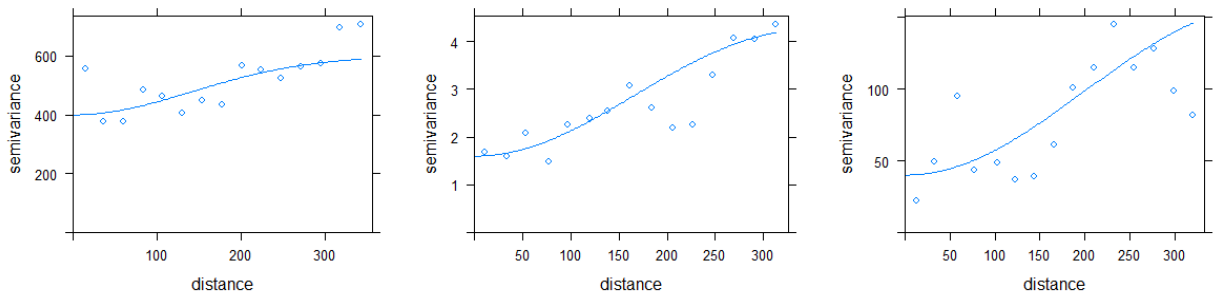
2.3.1 IDW Interpolation

In order to perform the Inverse Distance Weighted (IDW) Interpolation for the different pollution concentrations for this study, a shapefile of the selected five states was sourced from the US Census Bureau. A spatial transform was performed on this spatial data to ensure it aligns with the air pollution data, as some coordinates needed to be transformed from NAD83 to WGS84. The air pollution data also needed to be converted to spatial points data frame in order to conduct this analysis. To perform the IDW interpolation a spatial grid had to be created such that it encapsulated the extent of the aforementioned spatial points data frame. This grid was then passed as a parameter to the IDW function which is part of the *gstat* library. The interpolated surface that was generated

by this function was then converted into a raster format and was applied to the states data frame using the mask function. The only parameter selected for this method was the k-value, an initial value of 2 was chosen. This selection then had to be cross validated using the leave one out cross validation method (LOOCV), which iterates over the dataset removing one value to compare to the rest. This was accomplished using a 'for' loop to iterate n times, where n represents the number of values in the air pollution data set. On each iteration of the loop one value was removed and was used to perform the IDW analysis and stored in a vector, once the loop exited the sum of that vector was used to compute the root mean squared error (RMSE) of the model. This process was applied to each air pollution concentration. In all cases a k-value of 2 produced the lowest value of RMSE so that was used for the final prediction value.

2.3.2 Kriging

When performing the Kriging interpolation, the same dataset containing the spatial points data frame with air pollution data, and the spatial polygons for the selected states was used. An ordinary kriging process was performed by first creating a variogram for the annual average of pollution concentration using the variogram function from the *gstat* library. This variogram was then fit with the *fit.variogram* function. The parameters that needed to be selected for this function were the model, sill, range and nugget. By plotting the variogram against the fit (Figure 1), the fit was assessed and ensured for accuracy. The same aforementioned LOOCV method was used to cross validate the fit and the model. Once the fit of the variogram was optimized with the final values for the parameters (Table 1), the kriged surface could be generated. The chosen model was passed as a parameter to the krige function which generated the kriged surface of spatially interpolated values, which was then converted to a raster and applied to the states using the mask function. This process was then repeated for each air pollution concentration.



O_3 $PM_{2.5}$ NO_2

Figure 2. Final variogram plots for Ozone (Left), Particulate Matter (Centre) and Nitrogen Dioxide (Right)

Table 2. Variogram Model fit parameters

Pollutant	Model	Nugget	Sill	Range
Ozone	Gau	400	600	200
Particulate Matter	Hol	1.6	3.8	80
Nitrogen Dioxide	Wav	40	140	300

2.3.3 Clustering

To perform the spatial clustering, we were required to find unique monitoring stations that measure all 3 variables (NO_2, $PM_{2.5}$ & O_3). Only 21 such stations are monitoring for all 3 variables, out of a total of 143 monitoring stations. Combining the results of all 143 stations produces a data frame like the table shown below, with *NA* representing the missing readings for certain variables for the 132 stations.

Table 3. Attribute Table of Cluttering Dataset Before Using Interpolation Results

StationID	Longitude	Latitude	MeanPM	MeanOzone	MeanN02
81	75.34111	40.62806	10.801639	0.044505	19.750789
82	75.37250	39.83556	10.999118	0.044280	20.510324
83	75.41397	39.81871	9.258154	NA	NA
84	75.43250	40.61194	10.568116	0.042691	NA
85	75.57819	41.47912	NA	0.044680	NA

Using the results of our interpolation, we successfully determined the potential readings across the 143 monitoring stations by extracting values from the rasters produced by the interpolation. A snippet of the new data frame, in the same context as the one above, is displayed below.

Table 4. Attribute Table of Cluttering Dataset After Using Interpolation Results

StationID	Longitude	Latitude	MeanPM	MeanOzone	MeanN02
81	75.34111	40.62806	10.801639	0.04450500	19.750789
82	75.37250	39.83556	10.999118	0.04428000	20.510324
83	75.41397	39.81871	9.258154	0.04421575	20.588484
84	75.43250	40.61194	10.568116	0.04269100	20.461533
85	75.57819	41.47912	8.480992	0.04468000	16.038783

With all the 143 stations containing all the readings, we were able to proceed with the clustering. To account for the distribution and nature of the 3 variables used, we needed to standardize them by transforming each variable to have a mean of 0 and a standard deviation of 1. This would allow the clustering algorithm to equally weight all 3 variables.

The clustering algorithms that are being considered are K-Means, Hierarchical and Skater. All of which require K, the number of clusters. To deduce our value for K, we generated a Dendrogram, which included all the attributes we intend on using to cluster, excluding *StationID*, to find branching of the data that form natural clusters.

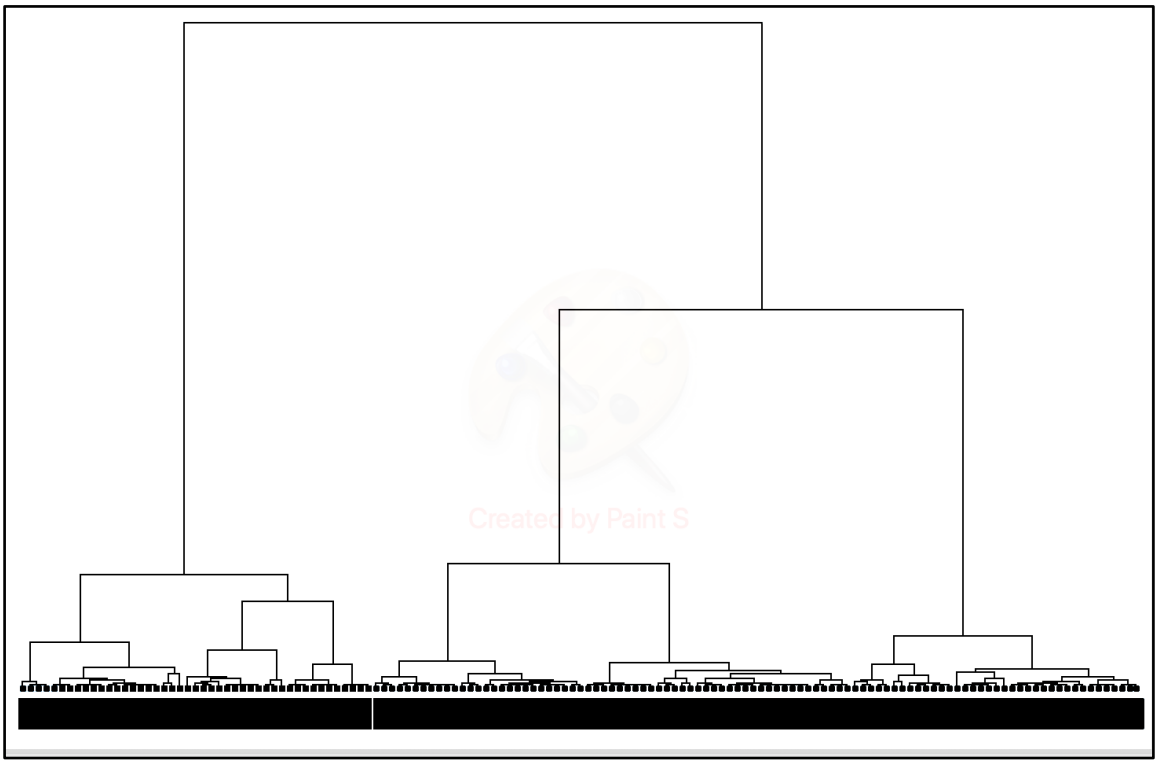


Figure 3. Dendrogram Produced in GeoDa giving equal weights to all variables (Excluding stationID)

The Dendrogram shows that the data is naturally partitioned into 2 categories, with the first category roughly consisting of ⅓ of the total stations and the second category consisting of ⅔ of the total stations. The second category is then partitioned again, around halfway through the clustering process, signaling that there exist distinct differences within the second category. The second category is roughly split into equal portions. At this stage the 3 clusters are roughly the same size. As a result, we decided to run our algorithms based on 3 clusters.

The main goal is to minimize the *"Total within-cluster sum of squares"*, which represents the difference between variables within the same cluster, and to maximize *"Between cluster sum*

of squares", which represents the differences between the clusters as a whole. This relationship is captured using the *"Ratio of between to total sum of squares."* Our goal is to maximize this ratio.

Initially running the K-Means algorithm without any spatial constraints yields a ratio of 0.55. This is realistically the highest attainable ratio score since the algorithm was free to cluster the three variables (NO_2, $PM_{2.5}$, and O_3) without having to take their spatial attributes into account. A consequence of using K-Means without spatial constraints means that the clusters' members are not contiguous. The map does however show a level of localization of clusters, even though it was given no spatial attributes. One cluster is concentrated on the northeast, the second southwest and the third southeast. This is a great indicator that the variables form natural clusters.

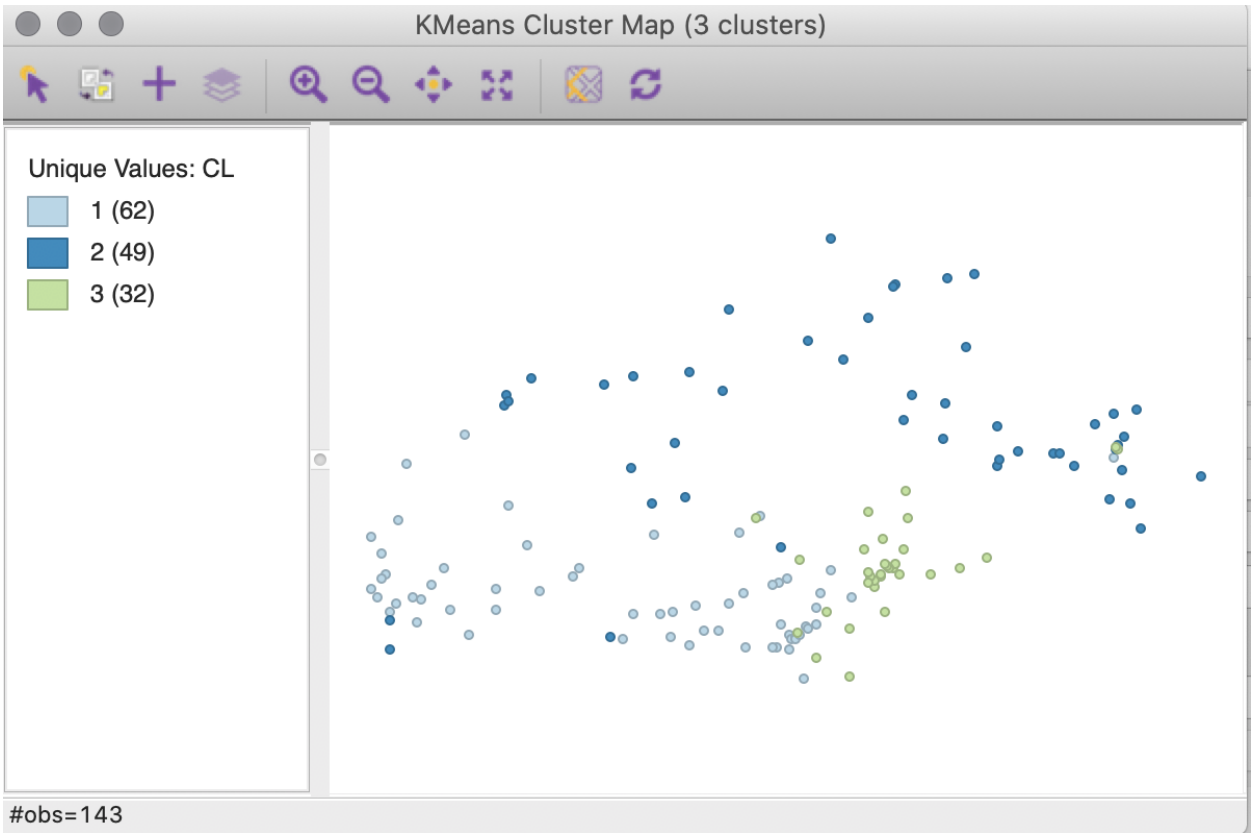


Figure 4. Clustering results after running K-Means clustering without spatial constraints.

We proceeded by running K-means again, now weighting spatial attributes equal to the other variables, this case also yielding noncontiguous with a very high ratio score. The fact that the

results are not contiguous indicates that further weights need to be given to spatial attributes to make the results contiguous. After finding the minimum weight needed for the result to be contiguous, which was 0.625, we obtained our first realistic cluster.

We then ran a hierarchical algorithm with spatial constraint, with the spatial attributes weight at the minimum to ensure contiguity, using ward's linkage (the linkage method that resulted in the highest ratio score). Finally, a graph-based algorithm named *Skater* was also used. Skater ensures contiguity by using a connectivity graph of the monitoring stations. Below is the connect graph used in the Skater algorithm.

Figure 5. Connectivity Graph of Monitoring Stations Used by the Skater Algorithm

3. Results

3.1 Descriptive Statistics

For all three pollutants analyzed in this study LOOCV was performed in order to assess the effectiveness of the models created. Both kriging and inverse distance weighted techniques were used to calculate RMSE values for all pollutants, although for each pollutant only one is used as they provide different results with different degrees of accuracy depending on the pollutant. Based on the results of the LOOCV which provided the root mean squared error of each model. For nitrogen dioxide, the IDW calculation was determined to be the more appropriate error range of the two, and provided a RMSE of 21.13544, and the kriging model was omitted as its RMSE was higher at 21.12636. For ozone, the inverse distance weighted method provided an RMSE of 0.4222651, and 0.4222673 for the kriging method. Unlike nitrogen dioxide however, the method selected for ozone was kriging. This was the same for particulate matter, which resulted in a RMSE of 8.439834 for kriging, and 8.440167 for IDW.

Table 5. Root mean squared error for the IDW and Kriging models for each pollutant concentration

Pollutant	IDW	Kriging
Ozone	0.4222651	0.4222673
Particulate Matter	8.440167	8.439834
Nitrogen Dioxide	21.13544	21.12636

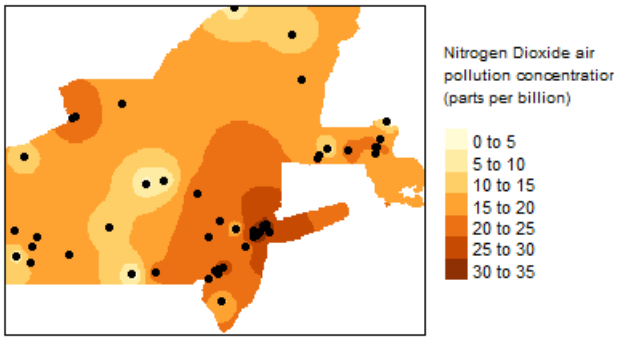
For clustering, the calculations performed provided average values of each pollutant within each cluster. These pollutant values are analyzed in greater detail in section 3.4 but the most

notable results were as follows: For particulate matter, cluster 2 gave a much lower value than the other two at 5.98. For nitrogen dioxide, cluster 3 provided a value that was significantly higher than the other two at 28.81. In contrast to these two significant results, ozone was fairly consistent among all three clusters with values ranging between 0.040 and 0.044.

3.2 IDW Interpolation Results

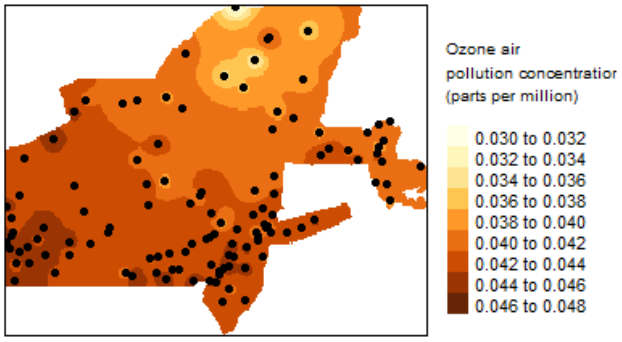
For NO_2, we will be using the IDW Interpolation results (explained in detail in 3.3) for our interpretation. Although, we did receive an Inverse Direction Weighted Interpolated Root Mean Square Error using LOOCV of 21.13544. Which, although not significantly different from the kriging method, will be our range of error for interpretation. To the right of this text is the spatial plot of the Inverse Direction Weighted Interpolation for NO_2.

NO_2 IDW Spatial Plot



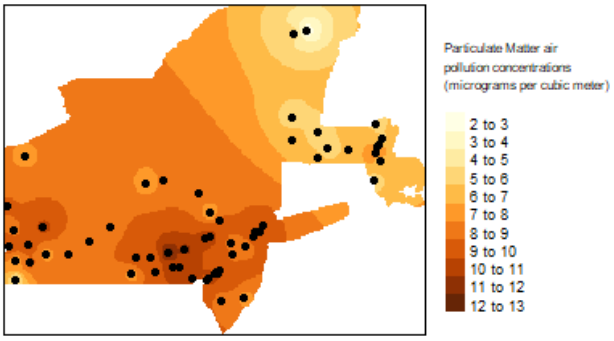
For O_3, we will not be using the IDW Interpolation results for our interpretation, as kriging is more accurate. Although, we did receive an Inverse Direction Weighted Interpolated Root Mean Square Error using LOOCV of 3.94094. Which, although lower than kriging, it is not significantly so, thus this will not be our range of error for interpretation. To the right of this text is the spatial plot of the Inverse Direction Weighted Interpolation for O_3.

O_3 IDW Spatial Plot



For $PM_{2.5}$, we will not be using the IDW Interpolation results for our interpretation, as kriging is more accurate. Although, we did receive an Inverse Direction Weighted Interpolated Root Mean Square Error using LOOCV of 8.440167. Which, was higher than kriging, paired with the fact kriging is more accurate than IDW, this will not be our range of error for interpretation. To the right of this text is the spatial plot of the Inverse Direction Weighted Interpolation for $PM_{2.5}$.

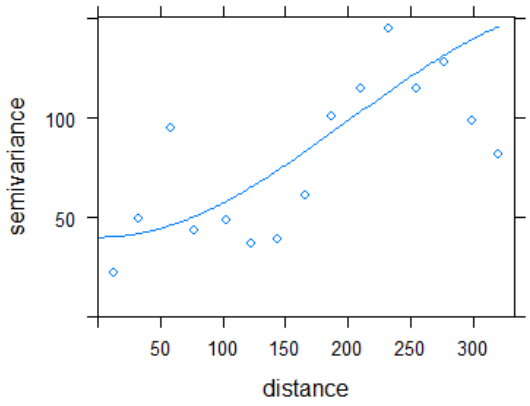
$PM_{2.5}$ IDW Spatial Plot



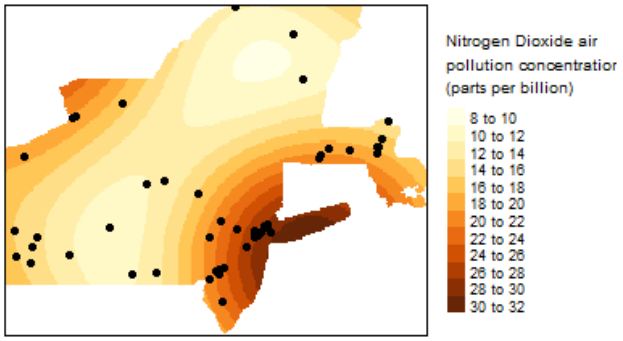
3.3 Kriging Results

For NO_2 the kriging results proved to be insufficient. The first obvious reason is the fact that the data we are using (levels of NO_2) is not stationery and subject to change as the environment changes. The stationarity is not what was found to be un-useful, but rather the Variogram that was produced itself. We did not observe a plateau off in the plot graph, but rather a great fluctuation of changes (With a sill of around 150 at around the 3rd quarter, before immediately dropping in values). A log transformation was conducted, to which it was found not to be different than the original NO_2 variogram. Thus, the variogram was not constant, and broke the second main assumption, rendering this data not reliable. It may be noted that the Root Mean Square Error of the Kriging Interpolation using LOOCV for NO_2 was 21.12636. The following are the spatial visualization of the kriging and the statistical visualization of the variogram for NO_2;

NO_2 Kriging

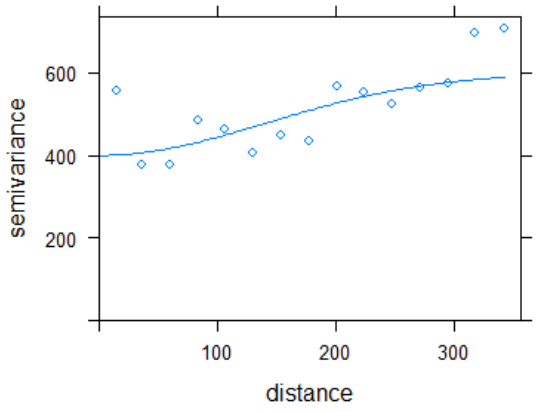


NO_2 Kriging Spatial Plot

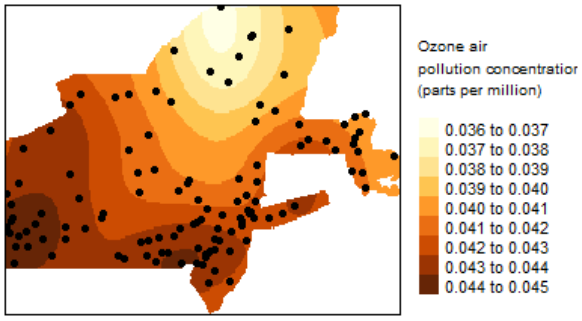


For O_3 the kriging results came out to be much more reliable. The variogram was very uniform, with a very low slope, plateauing off at the end. Thus we are able to say that this is a constant variogram, and will be able to be used in our interpretation of our results. Also, the Root Mean Square Error of the Kriging Interpolation using LOOCV for O_3 was significantly lower than NO_2, which was 3.975054, making it the smallest of the RMSE's for all three pollutants (for Kriging). The following are the spatial visualization of the kriging and the statistical visualization of the variogram for O_3;

O_3 Kriging



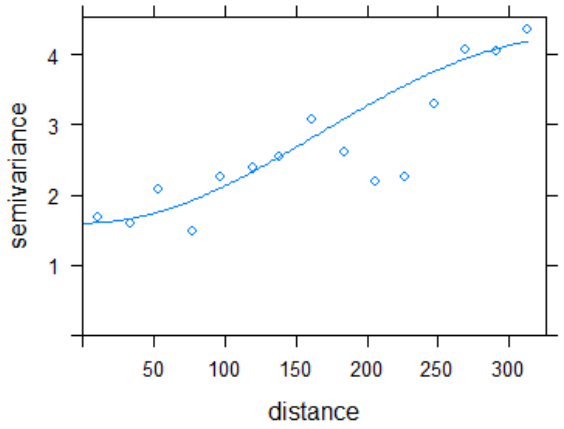
O_3 Kriging Spatial Plot



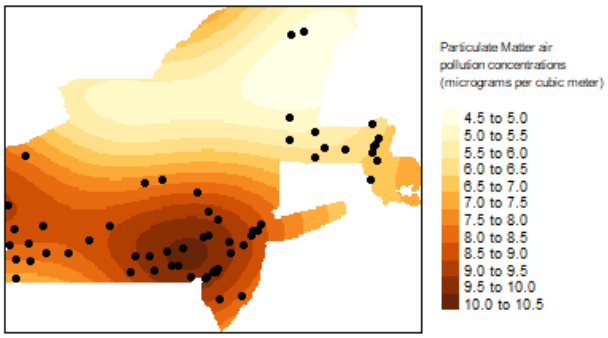
For $PM_{2.5}$, the kriging results were again, much more reliable than NO_2, but slightly less reliable than O_3. Although, we are still able to say that this produces a constant variogram, and will be able to be used in our interpretation of our results. While at the end, we can see a sill, it appears that from a distance of around 200 to 260 there is an exponential increase in semi-

variance, but then it appears to plateau off at the very end. This exponential anomaly must also be considered when using the data for interpolation. Otherwise, we have another Root Mean Square Error of the Kriging Interpolation using LOOCV for $PM_{2.5}$ that's significantly lower than NO_2, but not as low as O_3, which was 8.439834. The following are the spatial visualization of the kriging and the statistical visualization of the variogram for $PM_{2.5}$;

$PM_{2.5}$ Kriging



$PM_{2.5}$ Kriging Spatial Plot



3.4 Clustering Results

Table 6. Summary of Results of Clustering Algorithms

Algorithm	Total within-cluster sum of squares	Between cluster sum of squares	Ratio of between to total sum of squares	Contiguous (Y/N)
K Means with no spatial attributes	195.756	230.244	0.55048	N

K-Means with spatial attributes	315.883	394.117	0.5551	N
K-means with spatial weighted: 0.625	231.545	194.455	0.456	Y
Hierarchical	252.37	173.63	0.407581	Y
Skater	230.025	195.975	0.4600	Y

Although the Ratio Score is the main assessment criteria of which clustering algorithm to use, contiguity is an important consideration in the algorithm selection because spatial distribution is a defining characteristic of spatial clustering. K-Means with no spatial attributes resulted in the highest ratio score, it is non contiguous. The best performing algorithm that maintains contiguity is the graph based, Skater Algorithm.

Table 7. Pollutant Cluster Mean Centers for Clustering Using Skater Algorithm

Pollutant	**Cluster 1**	**Cluster 2**	**Cluster 3**
Mean Ozone	0.0437909	9.07866	16.8995
Mean Particulate Matter	0.0402402	5.9764	16.4412
Mean Nitrogen Dioxide	0.0420174	8.58484	28.8146

As can be seen in the table, ozone levels are fairly equal across all three clusters, as ozone values normally range between 0.02 to 0.07. The fact that the mean ozone levels for the clusters are very close, indicates that ozone levels have little spatial significance. For particulate matter and nitrogen dioxide this is not the case, as each pollutant is more heavily concentrated in certain regions than others. For Particulate matter, the value for cluster 2 is much lower than clusters 1 and 3, with a value of 5.98, compared to 9.08 in cluster 1 and 8.58 in cluster 3. The results for nitrogen dioxide are even more polarized, showing a value of 28.81 on the high end in cluster 3, with 16.9 and 16.44 for clusters 1 and 2 respectively.

Table 8. Size of clusters from the Skater Algorithm

Cluster	Size
Cluster 1	71
Cluster 2	49
Cluster 3	23

The skater algorithm produced clusters with largely different sizes. Their sizes, however, are not an indication of the area occupied by each of the clusters as the monitoring stations are randomly placed, not uniformly distributed, around the 5 selected states. To provide an accurate image of the spatial boundaries of the clusters, Thiesson polygon algorithm was performed to each monitoring station. Polygons with the same cluster ID were dissolved and what remains were the boundaries of the clusters, shown in figure 4.

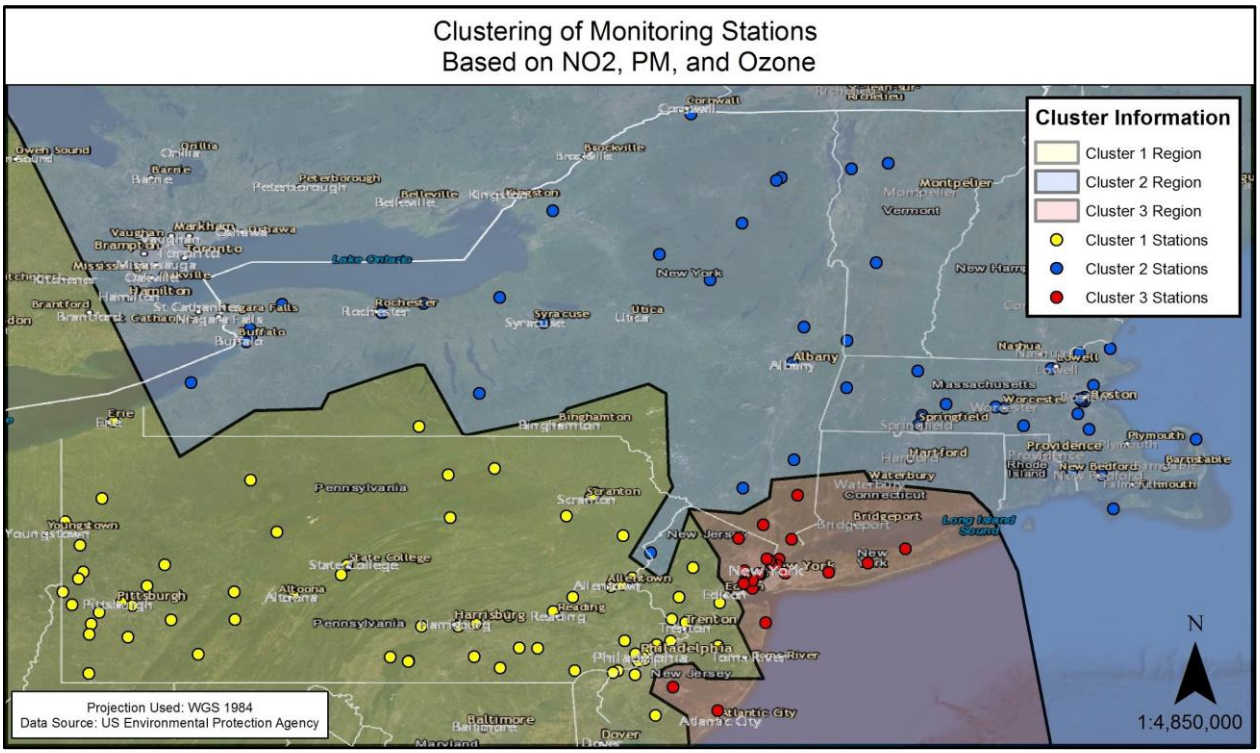


Figure 6. Complete map representing the results of the Skater clustering algorithm

4. Discussion

We can see large amounts of $PM_{2.5}$ North East of Philadelphia, around Conshohocken where there appears to be a large cluster of industrial activities such as automotive and energy plants in that area. So our kriging spatial modelled appeared to accurately predict the locations of high concentration $PM_{2.5}$. We can believe to a degree of certainty a correct correlation as vehicles contribute largely to $PM_{2.5}$, thus we can see large clusters around Philadelphia and New York City [3]. We may also notice that there are almost a moderate (according to AQI) amount of pollution in the south west of our study area, which we may conclude accurate because the Coal Power Plant in the Springdale & Cheswick boroughs in Philadelphia [4]. The rest of the study area, outside of these large urban areas and coal power plant, have good levels (4.5-9 $\mu g/m^3$) of $PM_{2.5}$.

O_3 is similar to $PM_{2.5}$ in that it is also emitted by cars, power plants, but also chemical plants, and other sources that react chemically in the presence of sunlight. Thus, our kriging

model accurately predicted high amounts of clustering around Philadelphia and the coal power plant in Springdale & Cheswick boroughs.

The clustering of our monitoring stations represents our findings to be accurate as per the locations. In the *Cluster 2 Region* we notice that across all our pollutants, this area is the lowest of all regions, significantly in O_3 and $PM_{2.5}$, which makes sense, since in this region there are no major urban areas or plants significantly affecting the region. In the Cluster 1 region we see the highest amounts of O_3 and $PM_{2.5}$, this is because of Philadelphia and the coal power plants, and other plants in the area. We notice a similar amount of NO_2 the Cluster 2 Region, most likely due to the fact that the area receives most of it's pollutants from the industrial sector rather than the urban centers [5]. The Cluster 3 Region is the densest out of all regions, that being it is located around New York City itself. While the O_3 in the Cluster 3 Region is between C1R and C2R, the $PM_{2.5}$ is about the same as C1R. The most significant change is NO_2, where it is almost double that of C1R and C2R, and this can be attributed to the fact that New York City is a high densely populated city, with lots of traffic, creating much larger (but still good according to EPA) amounts of NO_2 in that region, especially in the high traffic & wide road areas such as Queens [6].

5. Conclusions

Through spatial analysis techniques, and through confirmation of physical sources of pollutants, we were able to confirm that the following be true for our study area;

This study is confident that for NO_2, Inverse Distance Weighted Interpolation gave us the most accurate predictions, due to the issue of a great lack of monitoring stations within our study area.

This study is confident that for O_3 and $PM_{2.5}$, Kriging Interpolation gave us the most accurate predictions, due the much greater number of monitoring stations, and good distribution of those monitoring stations across our study area.

This study found that K-Means Clustering produced the most accurate air pollutant cluster regions for our study area, from our statistical analysis, and confirmation using physical sources of pollutants, that being different urban areas (New York is different from Philadelphia) and large coal power plants, contrasted with fairly less urban areas in the north of our study area.

Author Contributions:

Antonio LaPlaca: Sections 1, 2.1, 3.2, 3.3, 4, 5

Iman Kewalramani: Sections 2.3.3, 3.4, Clustering R Code, Data Cleaning

Hardeep Dhatt: Sections 2.3.1, 2.3.2, Kriging R Code, IDW R Code, Data Cleaning

Andrew Sud: Sections 2.1, 2.2, 3.1, 3.4, ArcMap Map Figures

References

1. "State Population Totals and Components of Change: 2010-2019." United States Census Bureau, 30 December 2019, https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html
2. "Time Series Values for Individual Locations." Oregon State University, N.d., http://www.prism.oregonstate.edu/explorer/
3. Hsu, Wan-Hsiang, et al. "Seasonal and Temperature Modifications of the Association between Fine Particulate Air Pollution and Cardiovascular Hospitalization in New York State." *Science of The Total Environment*, vol. 578, 2017, pp. 626–632., doi:10.1016/j.scitotenv.2016.11.008.
4. Bray *, Casey D., et al. "Characterization of Particulate Matter (PM2.5 and PM10) Relating to a Coal Power Plant in the Boroughs of Springdale and Cheswick, PA." *Atmosphere*, vol. 8, no. 12, 2017, p. 186., doi:10.3390/atmos8100186.

5. “Nitrogen Dioxide (NO2) Pollution.” *EPA*, Environmental Protection Agency, 13 June 2019, www.epa.gov/no2-pollution.
6. Masiol, M., et al. “Analysis of Major Air Pollutants and Submicron Particles in New York City and Long Island.” *Atmospheric Environment*, vol. 148, 2017, pp. 203–214., doi:10.1016/j.atmosenv.2016.10.043.

Appendix A

R Code submitted as: Appendix_A.R