IDW/Kriging Interpolation and Spatial Clustering of Average Annual Concentrations of NO2, O3, and PM2.5 within the states of Illinois, Indiana, Wisconsin, Iowa, and Missouri

Abstract:

Air pollution is a rising problem around the world and in many U.S states. Due to this rise in air pollution, monitoring of different air pollutants has become an important task. This study will examine nitrogen dioxide (NO2), ozone (O3), and particulate matter 2.5 (PM2.5) within 5 contiguous U.S states. Illinois, Indiana, Wisconsin, Iowa, and Missouri. Inverse distance weighted (IDW) and kriging interpolation were two different methods used to estimate the air pollution of counties without monitoring sites. Spatial clustering analysis shows regions that contain monitors most similar in air quality. Fit variograms were created for kriging and cross-validation was used for both interpolation methods to determine their reliability. Results show that average annual NO2 concentration is high within counties of Illinois while both Ozone and PM2.5 concentrations are fairly high within parts of Illinois, Indiana, and Missouri. Spatial clustering analysis shows that NO2 is clustered within Iowa, Wisconsin, and part of Illinois and Missouri. Ozone is clustered within parts of Indiana and Illinois, and PM2.5 is clustered within parts of Illinois, Indiana, and Missouri. Conclusions from the study show that Wisconsin and Iowa seem to have the low amounts of all the air pollution concentrations while states like Indiana and Illinois have high amounts of all of the air pollution concentrations. Missouri meanwhile varies between high and low concentrations depending on the county. IDW is a better predictor of Ozone and PM2.5 while kriging is a better predictor of NO2.

1. Introduction

Good air quality is essential in order to maintain human health and prosperity. With factors such as global warming, the rise of dense car filled cities and deforestation, air pollution is on the rise. This has led to increases in the amount of nitrogen dioxide (NO2), ozone (O3), and particulate matter 2.5 (PM2.5) in many areas of the world including the United States. As such, it is important to monitor and analyze these pollutants in order to be aware of the potential hazards that they pose to human health. High concentrations of these pollutants can lead to increased mortality rates and diseases as they are responsible for reduced levels of air quality. The aim of this work will be to study the annual average concentration of these pollutants in 5 contiguous U.S. states during the 2020 year. Illinois, Indiana, Wisconsin, Iowa, and Missouri. In order to determine where there are high annual average concentrations of these pollutants, inverse distance weighted (IDW) and kriging interpolation will be conducted. Both of these methods will then be compared by using the root mean square error to determine which interpolation method is better at predicting which pollutants. Along with the interpolation of these concentrations, a spatial clustering analysis will also be performed in order to identify regions that contain monitors that are most similar in air quality. This will provide a greater understanding of which areas in these states suffer from high concentrations of air pollution and any clusters that exist. By knowing this, action can be taken to reduce the air pollution of these regions and policies can be enacted or improved in order to prevent a further increase of air pollution.

2. Methods

2.1 Study Area

The study area consists of 4 contiguous states that surround the state of Illinois, which are Indiana, Iowa, Missouri and Wisconsin. The geography mainly consists of flat plains and lowlands. The climate of the study area typically consists of a four-season climate, and indicates a more humid, temperate continental climate. A population of ~35 million resided within the study area as of 2020. The area of study was chosen in hopes of observing the concentration levels of nitrogen dioxide, ozone, and particulate matter 2.5 of the state of Illinois and the surrounding areas, which includes the 4 contiguous states previously specified.

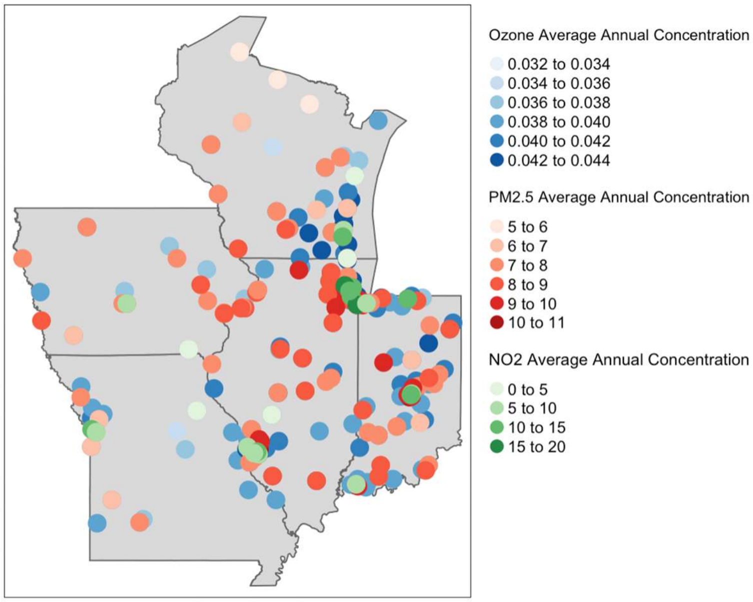

Figure 1: Study Area

Observing the figure, which displays the study area and the three pollutants overlapped on it, we can observe a general pattern of the concentrations of the three pollutants being more concentrated and monitored within areas close to the capital cities of each respective, contingent state.

2.2 Data

Pollutants	# of Monitors
Nitrogen Dioxide	24
Ozone	141
Particulate Matter 2.5	119

Table 1: Pollutant Monitors

The data was obtained from the United States Environmental Protection Agency website, where they provide datasets pertaining to concentrations of air pollution which are captured by outdoor monitors in every state in America.

This data obtained from the U.S. EPA website contains the average annual concentration of a specified pollutant alongside with its geographical coordinates from where this observation was obtained, the datum, the methods for which this data was captured with, the metrics, the pollutant standard, as well as the date and address from where and when this observation was taken.

2.3 Statistical Analysis

2.3.1 IDW Interpolation

In order to conduct IDW interpolation, the CSV data was first filtered to only include average annual concentrations of NO2, Ozone, and PM2.5 in the states of Illinois, Indiana, Wisconsin, Iowa, and Missouri along with their pollutant standards (Appendix A: Lines 33-55). The pollutant standards chosen were NO2 Annual 1971, Ozone 8-hour 2015, and PM25 Annual 2012 as they contained averages based on the longest term data. The average annual concentration of these pollutants was found under the “Arithmetic.Mean” variable. Any duplicate values from the same longitude and latitude were removed (Appendix A: Lines 58-63). Before performing IDW interpolation, the data needed to be converted into a SpatialPointsDataFrame. This was done through using the longitude and latitude coordinates of the pollutant data and assigning a projection system (Appendix A: Lines 67-84). WGS84 was used as most of the pollutant data was using the WGS84 datum with few using NAD83. After making the spatial data frames, a study boundary was added of the counties within our 5 U.S states (Appendix A: Lines 87-95). This counties shapefile was obtained through filtering a 2020 U.S. counties shapefile [5] within ArcGIS Pro. Next, different k values had to be chosen for the interpolation. The k value effects the weighting of distance. The smaller the k, the higher the weights for further points. An initial k parameter of 2 was chosen and leave one out cross-validation was used to assess the prediction of the interpolation for each of the pollutants. This was repeated for different k values (3, 4, and 5) in order to determine the best value to use for the interpolation. After comparing the RMSE values, a k value of 4 for NO2, 2 for Ozone, and 3 for PM2.5 was used as these values had the best fit. The results of the IDW interpolation were then mapped. The RMSE for the different k values can be found in table 2 within the “Results” section.

2.3.2 Kriging

In order to conduct Kriging, variogram models had to be fitted. Before fitting a model, the data had to be checked to make sure that it was normally distributed (Appendix A: Lines 246-255). This was done by viewing the histograms of the average annual concentrations of each of the pollutant as well as conducting a shapiro test. The shapiro test for all of the pollutants displayed p-values greater 0.05 which meant that the data was normally distributed. Next, sample variograms were displayed before fitting in order to view the initial nugget, sill, and range (Appendix A: Lines 259-279). This would show the initial values needed to input into the models. Each of these models were then tested on each of the pollutants (Appendix A: Lines 288-303) and leave out out cross-validation was conducted (Appendix A: Lines 308-326) in order to determine the RMSE of the residuals which would help determine the model that fit best. The results of the model tests are shown in table 3 within the “Results” section. The values of “NA” represent when the model did not converge. The models not shown in the table are those which had no convergence for all of the pollutants (e.g. “Sph”, “Exc”, etc.). After conducting the tests for each of the models on the pollutants, the “Mat” model was the best fit for NO2, the “Lin” model was the best fit for Ozone, and the “Gau” model was the best fit for PM2.5. Different nugget, sill, and range values were then optimized for each of these models to obtain the best fit values. With the best fit models now acquired, the data could be interpolated using kriging and then mapped (Appendix A: Lines 330-366).

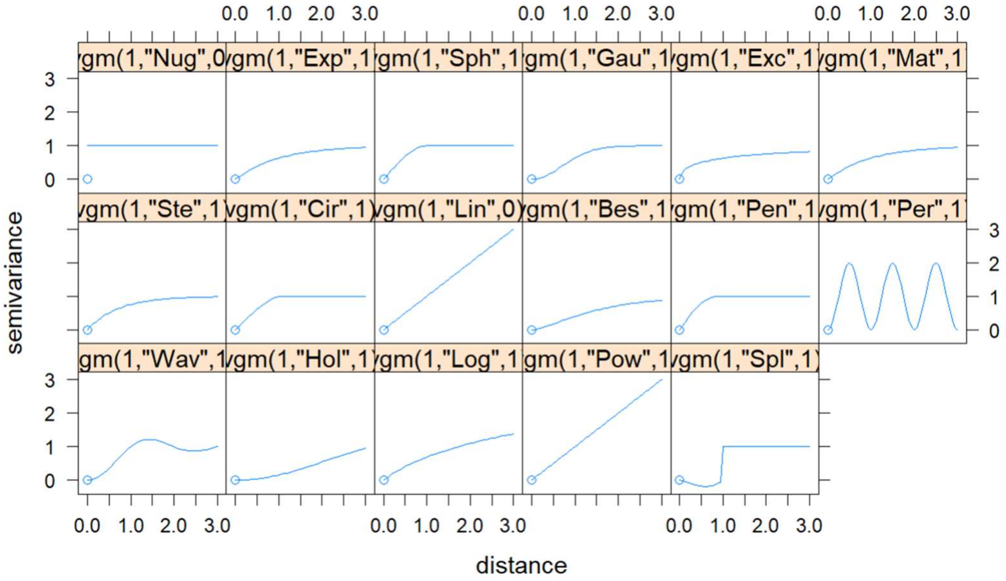

Figure 2: Variogram models tested

2.3.3 Clustering

For Spatial clustering, we process in the following steps: extracting data for clustering, making neighbors lists, making minimum spanning tree and processing contiguity-constrained clustering.

Using extract function with fun=mean to fill area boundaries with mean idw interpolation values in area. Then getting the pollutants idw interpolation values for each area and standardizing the values with scale function. Next, Poly2nb function helps create the neighbor list for study areas. For making a minimum spanning tree, using nbcosts function to create least cost from neighbor list and standard data.Then, nb2listw function creates edge weights based on least function. By weights list, we can have a minimum spanning tree by mstree function. Finally, calculating spatial clustering with k=3 by skater function.

3. Results

3.1 Descriptive Statistics

The NO2 values range from 1.1 to 16.46 with a median of 8.9 and a mean of 8.6.

The Ozone values range from 0.003316 to 0.04385 with a median of 0.03949 and a mean of 0.03947.

The PM2.5 values range from 5.02 to 10.91 with a median of 8.1 and Mean of 8.092.

There were only 24 monitoring sites for NO2, 141 for Ozone, and 119 for PM2.5 (Table 1). With only 24 monitoring sites for NO2, it is more difficult for the interpolation methods to accurately predict the pollutant values across the counties.

3.2 IDW Interpolation Results

Different k values were tested for the IDW interpolation to determine the best model. The smaller the k value, the higher the weights are higher for further points. The larger the k value the lower the weights are for further points. In order to determine the optimal k values, cross validation was conducted for each pollutant with each of the k values. The results of the cross validation provide RMSE values which gives a predictive accuracy of the IDW interpolations. After completing cross validation, the best values to choose for k for each of the pollutants was determined. In this case NO2 should choose k=4, Ozone k=2, and PM2.5 k=3.

Pollutant	RMSE (K=2)	RMSE (K=3)	RMSE (K=4)	RMSE(K=5)
NO2	4.095283	4.096841	4.092321	4.10417
Ozone	0.001886983	0.001945635	0.002014079	0.002062057
PM2.5	0.8325481	0.8128481	0.826702	0.8472032

Table 2: Root Mean Squared Error values for different K values of pollutants

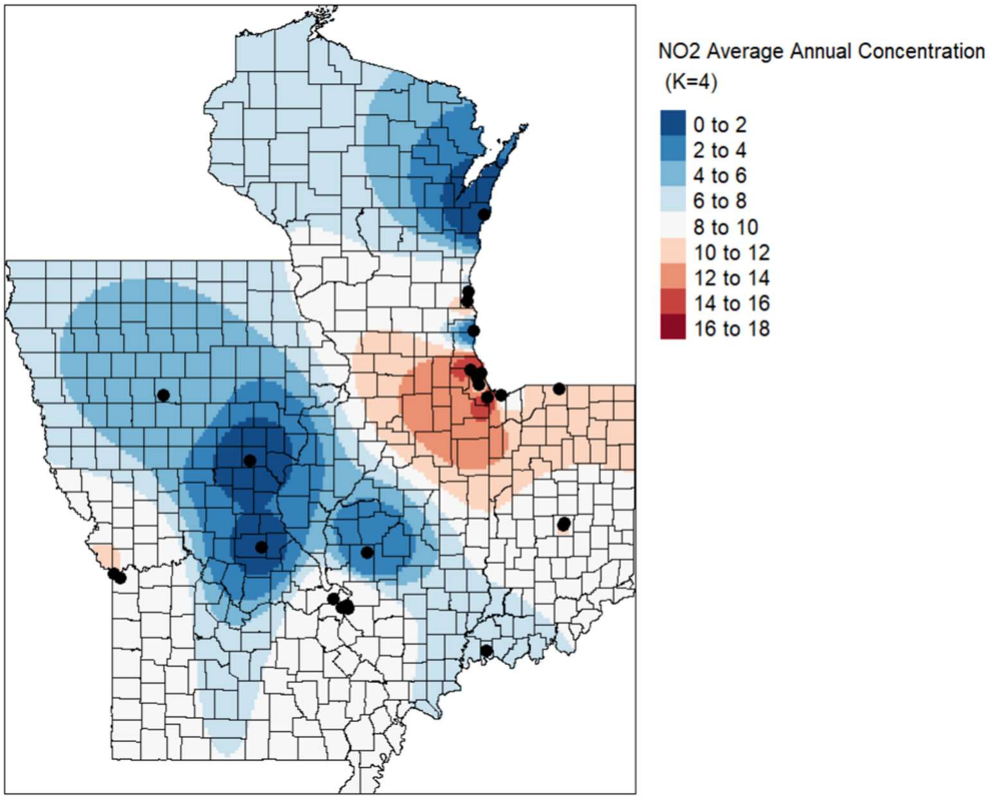

Figure 3: NO2 IDW Interpolation

With an optimal k value now determined, the results of the IDW interpolation were mapped (Figure 3). The NO2 IDW shows low values of NO2 concentration within the state of Wisconsin as well as Iowa. Missouri and Indiana are around the mid range of 8-10 while Illinois is a mix with higher NO2 concentrations in top right of the state and lower NO2 concentrations in the bottom left. With a higher k value, the weights are lower for further points which could be why the map shows much lower concentrations of NO2 near areas with already low NO2 concentrations.

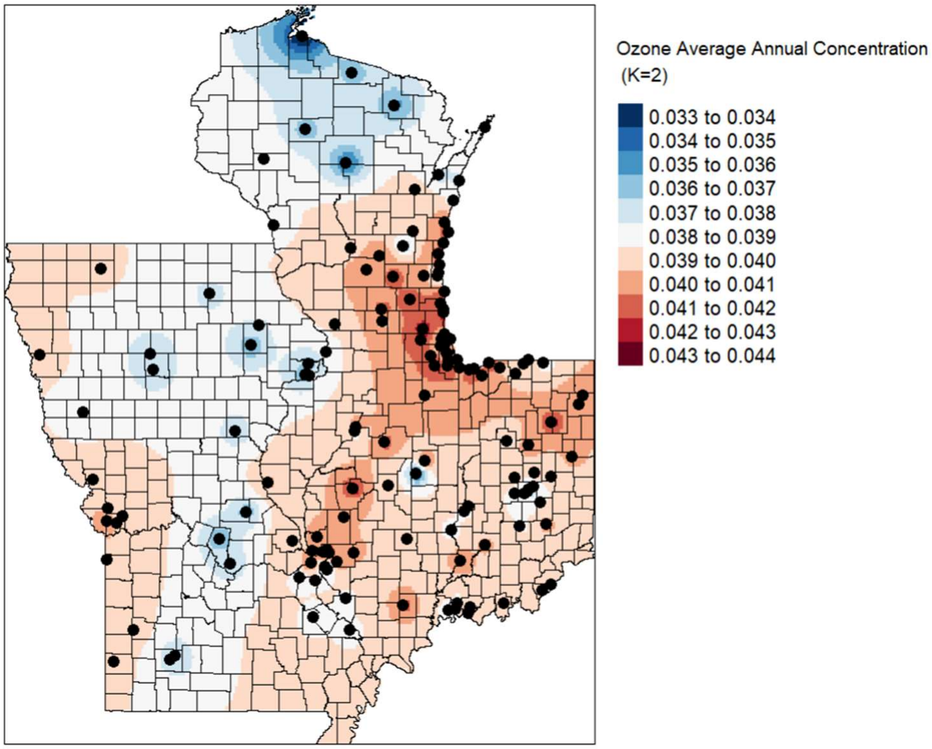

Figure 4: Ozone IDW Interpolation

For Ozone, the cross validation results determine an optimal k value of 2. The Ozone IDW shows low values of Ozone concentration within parts of Wisconsin. However, the general trend of the map seems to show higher concentrations overall. This is likely because the k value is lower which gives higher weights to farther points.

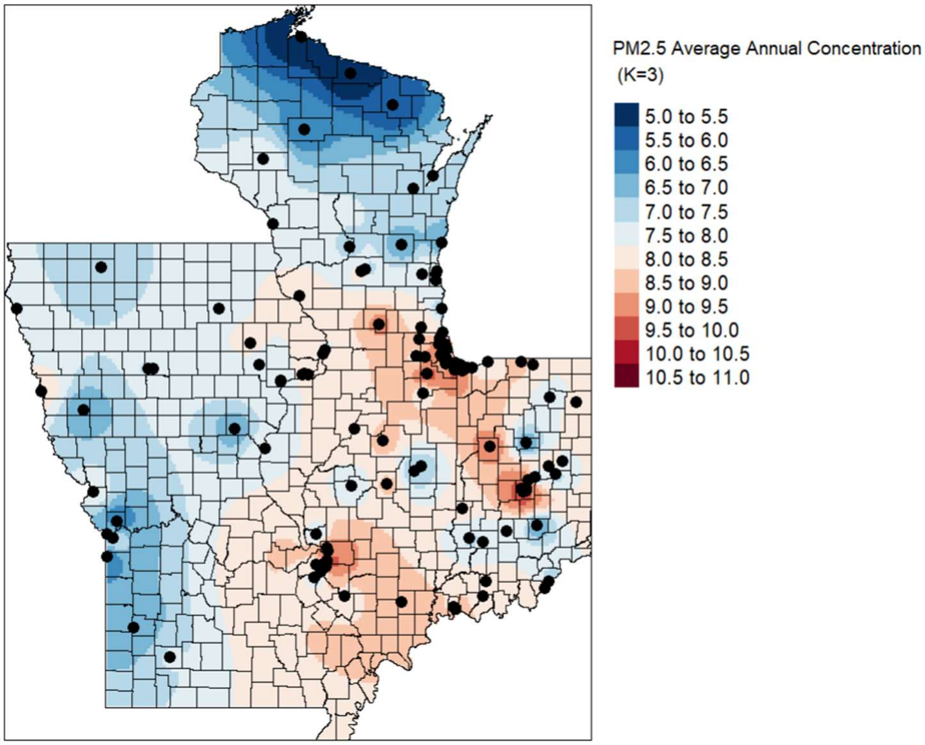

Figure 5: PM2.5 IDW Interpolation

For PM2.5, the cross validation results determine an optimal k value of 3. The PM2.5 IDW shows low values of PM2.5 within Wisconsin. This seems to indicate that the state has the least amount of air pollution as all of the IDW interpolation maps had low values for Wisconsin. Iowa also seems to have low PM2.5 while Missouri is split. Illinois and Indiana both seem to trend towards higher concentrations. With a k value of 3 this is likely why the map shows mixed results compared to the lower k value of the Ozone map which mostly showed high concentrations of air pollution and the higher k value of NO2 which showed lower concentrations of air pollution.

3.3 Kriging Results

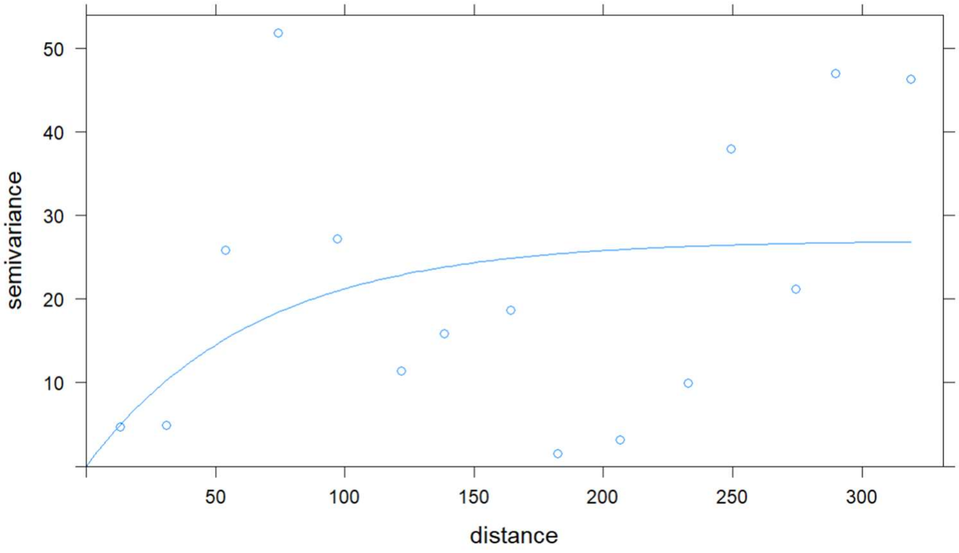

Figure 6: NO2 Fitted Variogram (Mat)

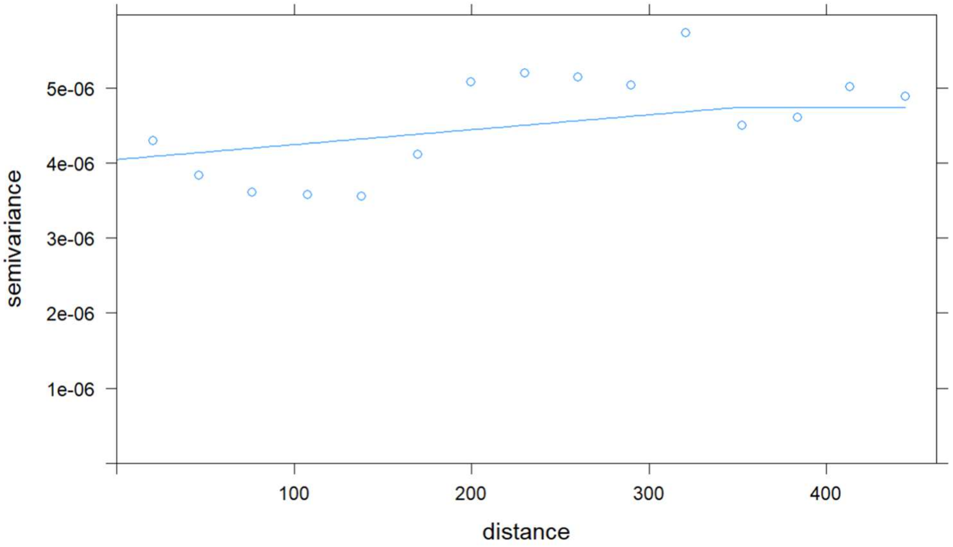

Figure 7: Ozone Fitted Variogram (Lin)

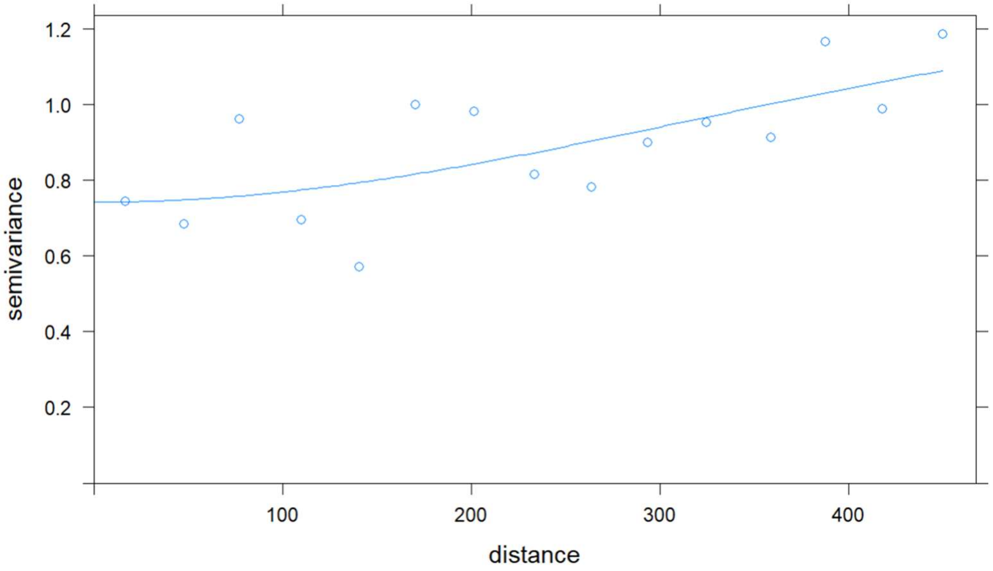

Figure 8: PM2.5 Fitted Variogram (Gau)

Different variogram models were tested in order to determine the best fit for kriging. Before fitting the variogram models, an initial nugget, sill, and range were chosen. Fitting the variograms for Ozone and PM2.5 was a bit easier as they seemed to level off and had a more consistent trend in their values. NO2 however was much harder to fit as the values were much more scattered which made it difficult to determine where it converges. These variogram models were assessed using cross validation which provided RMSE values to help determine the optimal model to use. After completing the cross validation we were able to determine the models to use for each of the pollutants (Table 3). The best model to fit NO2 concentrations was “Mat”. The best model for Ozone was “Lin”, and the best model for PM2.5 was “Gau”. Now that we have the best models that could fit the data, we could optimize the nugget, sill, and range values for these 3 types of models. After changing the parameters for these, we can find the optimal nugget, sill, and range by minimizing the sum of squared errors. The optimized values are as follows. NO2 had a nugget of 5, sill 25, and range of 100. Ozone had a nugget of 4.8e-06, sill of 5.2e-06 and range of 230. PM2.5 had a nugget of 0.7, sill of 0.97, and range of 80. These optimized values provided the lowest RMSE values for the 3 fitted models.

Pollutant	Gau	Mat	Ste	Lin	Bes	Hol
NO2	NA	3.985711	3.98572	4.245885	4.071612	NA
Ozone	NA	NA	NA	0.001902049	NA	0.001933156
PM2.5	0.8534085	NA	NA	NA	0.853739	NA

Table 3: Kriging RMSE values for different models

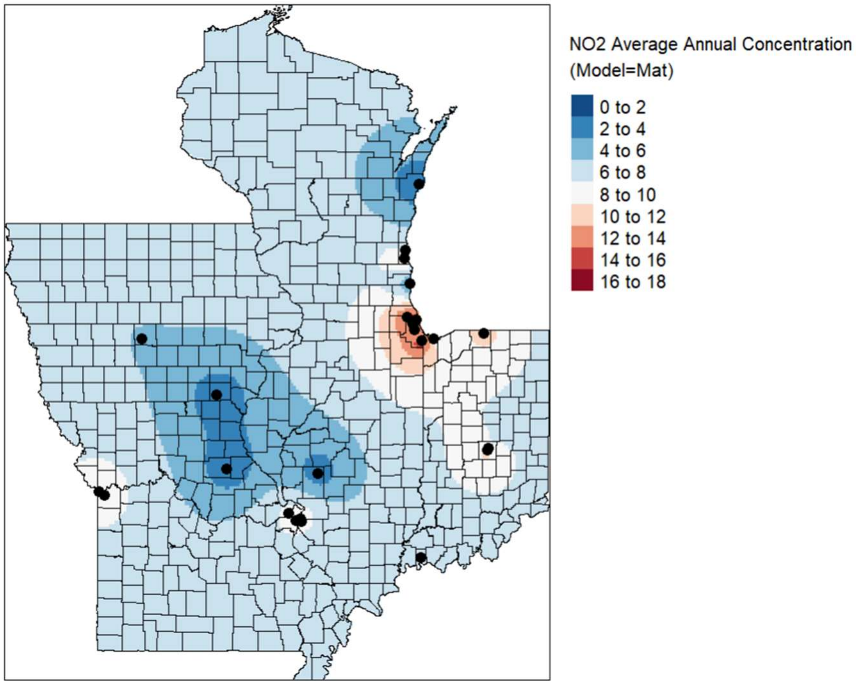

Figure 9: NO2 Kriging Interpolation

The results of the kriging interpolation for NO2 show much lower concentrations of NO2 throughout all of the counties except for a small cluster within Illinois.

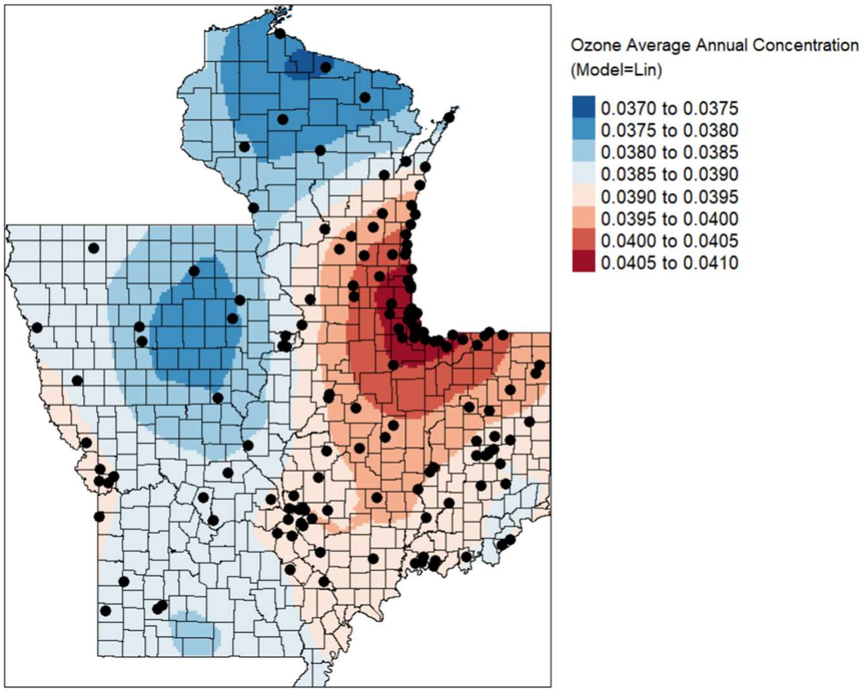

Figure 10: Ozone Kriging Interpolation

The results of the kriging interpolation for Ozone show higher concentrations within Illinois and Indiana with lower concentrations within Wisconsin, Iowa, and Missouri.

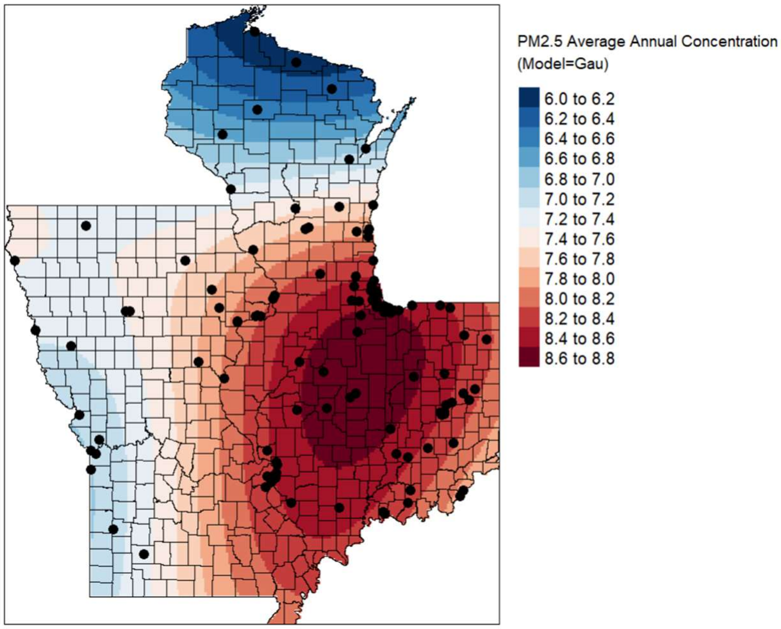

Figure 11: PM2.5 Kriging Interpolation

The results of the kriging interpolation for PM2.5 show low concentrations of NO2 within Wisconsin and parts of Iowa and Missouri with high values of PM2.5 in Illinois, Indiana, and parts of Missouri and Iowa.

3.4 Clustering Results

In the spatial clustering analysis, since many counties do not have pollutants observing stations, there was no data of pollutants. As such, IDW interpolation values were used to fill out area boundaries. Then, pollutant data was standardized to make different pollutants have the same weight standard in calculating edge costs.

Next, a neighbor list for spatial clustering analysis was created as can be seen in Figure 12.

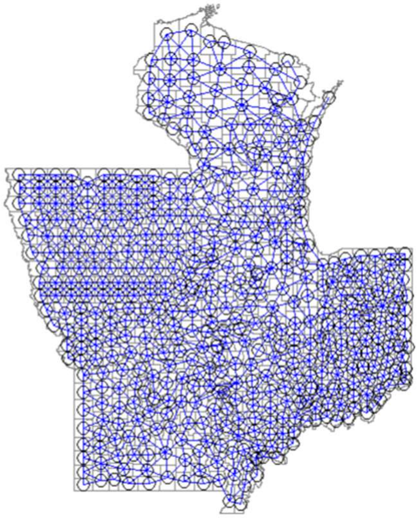

Figure 12: Neighbour List

After creating a Neighbour list, we calculate edge cost and apply weights to edges. Using weighted edges to create a minimum spanning tree. Then Calculate contiguity-constrained clustering with k = 3 based on minimum spanning tree and we get the result in Figure 13. k = 3 was chosen since we have 3 different types of pollutants and we want to have 3 groups of clustering.

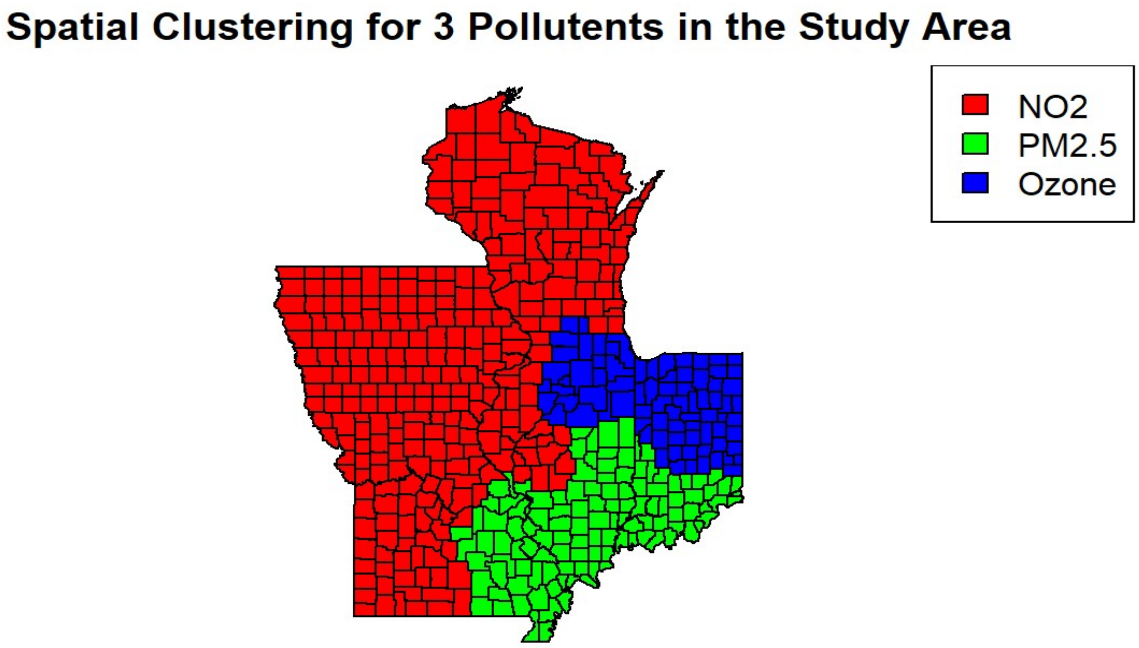

Figure 13: Spatial Clustering

From Figure 13, the clustering groups show spatial features of pollutants in the area. From IDW interpolation, we know that the north of Indiana and northeast of Illinois have heavy Ozone pollution and PM2.5 also mainly affects this area. Meanwhile, the rest of the states seem to mostly be affected by NO2.

4. Discussion

After conducting interpolation using both IDW and kriging, we can determine the optimal methods for predicting air pollution values within our 5 states. For NO2, kriging was a better predictor as it had a lowest RMSE of 3.985711 compared to IDW which had a lowest RMSE of 4.092321. For Ozone, IDW was a better predictor as it had a lowest RMSE of 0.001886983 compared to kriging which had a lowest RMSE of 0.001902049. Lastly, IDW was also a better predictor of PM2.5 as it had a lowest RMSE of 0.8128481 compared to kriging which had a lowest RMSE of 0.8534085. In general, the RMSE values of both methods are rather similar which shows there is not much difference between the two within our study area and both could realistically be used. Spatial clustering showed the counties that contained monitors that are most similar in air quality. NO2, Ozone, and PM2.5 were all high in concentration and clustered in counties of Illinois and Indiana. Meanwhile, all 3 pollutants had low concentrations within Wisconsin. These results somewhat differ from other studies who use newer and improved models to predict air pollution. One study Valencia, Spain [1] found a LIDW (Local Inverse Distance Weighting) to produce better results than IDW or Kriging when taking wind direction into account. Another study [2] tested a new type of kriging model called RIO which is a type of detrended kriging model and found it to be superior to regular kriging for predicting O3, NO2, and PM10. In Seoul, Korea, a model called ConvLSTM was developed which manipulates the spatial and temporal features of data [3] and was found to model PM2.5, PM10, CO, NO2, and SO2 better than models such as CNN and LSTM. However, one study monitoring Ozone and PM10 concentration levels [4] found no significant difference between IDW interpolation and kriging which is similar to the results of this study.

5. Conclusions

In conclusion, we can determine that certain states such as Wisconsin and Iowa have low amounts of air pollution concentrations while states like Indiana and Illinois suffer from higher amounts of air pollution concentrations. States like Missouri vary between low and high concentrations within certain counties. These results are further shown through spatial clustering analysis where we can see Illinois and Indiana have clusters of all types of air pollution while Wisconsin and Iowa are only shown with NO2. We can also conclude that IDW is a better predictor of Ozone and PM2.5 air pollution within this study area while kriging provides better results for NO2. These results can be used to help the governments of these states develop or amend policies in order to reduce air pollution concentrations that are high within their counties.

References

[1] Contreras, L., & Ferri, C. (2016). Wind-sensitive interpolation of urban air pollution forecasts. *Procedia Computer Science*, *80*, 313–323. https://doi.org/10.1016/j.procs.2016.05.343

[2] Janssen, S., Dumont, G., Fierens, F., & Mensink, C. (2008). Spatial interpolation of air pollution measurements using Corine land cover data. *Atmospheric Environment*, *42*(20), 4884–4903. https://doi.org/10.1016/j.atmosenv.2008.02.043

[3] V. Le, T. Bui and S. Cha, "Spatiotemporal Deep Learning Model for Citywide Air Pollution Interpolation and Prediction," *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2020, pp. 55-62, doi: 10.1109/BigComp48618.2020.00-99.

[4] Wong, D. W., Yuan, L., & Perlin, S. A. (2004). Comparison of spatial interpolation methods for the estimation of Air Quality Data. *Journal of Exposure Science & Environmental Epidemiology*, *14*(5), 404–415. https://doi.org/10.1038/sj.jea.7500338

[5] *USA counties*. ArcGIS Hub. (n.d.). https://hub.arcgis.com/datasets/esri::usa-counties/about

Appendix A

Code for analysis is in zipped folder as Appendix_A.Rmd