

# Visualizing Spatial Patterns of Air Pollution in the Eastern United States: Impacts on Human Health

Sanah Sarin<sup>1</sup> Abdullah Shahid<sup>2</sup> and Munim Adil<sup>3</sup>

<sup>1</sup> 1007190941; sana.sarin@mail.utoronto.ca

<sup>2</sup> 1007237624; abdullah.shahid@mail.utoronto.ca

<sup>3</sup> 1007071534; munim.adil@mail.utoronto.ca

Start: March 27, 2024; Due: April 10, 2024

**Abstract:** Air pollution is a significant environmental and public health concern. This study aims to investigate the spatial patterns and clustering of annual average Nitrogen Dioxide ( $\text{NO}_2$ ), ground level Ozone ( $\text{O}_3$ ), and Particulate Matter 2.5 ( $\text{PM}_{2.5}$ ) air pollution concentrations in five contiguous states in the Eastern United States: New York, Pennsylvania, New Jersey, Delaware, and Maryland. We will employ spatial interpolation techniques, including Inverse Distance Weighting (IDW) and Kriging, to estimate air pollution concentrations at unmonitored locations. Additionally, we use a spatial clustering algorithm, known as the skater algorithm for hierarchical clustering, to identify spatial patterns and hotspots of air pollution concentrations in the study area. Our study reveals through Spatial Interpolation, that the concentrations of Nitrogen Dioxide ( $\text{NO}_2$ ), ground level Ozone ( $\text{O}_3$ ), and Particulate Matter ( $\text{PM}_{2.5}$ ) pollutants are consistently within the standards set by the Environmental Protection Agency (EPA) [1]. However, Spatial Clustering analysis reveals that despite overall safe levels of  $\text{NO}_2$  pollutants, New Jersey exhibits relatively higher levels compared to other regions. Additionally, the highest ground level Ozone pollutant levels are concentrated in North Maryland, identifying it as a hotspot for  $\text{O}_3$ . Lastly, Pennsylvania is also approaching the threshold for  $\text{PM}_{2.5}$  concentrations, posing potential health hazards. Overall, other states in the study area remain below the standard threshold and are free from immediate health hazards, however, effective air quality management strategies are important to mitigate the potential upcoming risks.

**Keywords:** Air Pollution; Spatial Interpolation; Spatial Clustering

---

## 1. Introduction

Air pollution is a global health crisis, with 7.6% of global deaths attributed to particulate matter air pollution, and significant mortality rates linked to ground-level ozone, and nitrogen dioxide in the United States [2, 3]. Urgent action is needed to address air quality concerns and mitigate health risks associated with these pollutants, which threaten human health and the environment. Understanding the impacts of nitrogen dioxide ( $\text{NO}_2$ ), ozone ( $\text{O}_3$ ), and particulate matter ( $\text{PM}_{2.5}$ ) is crucial for informing evidence-based policies and interventions to reduce exposure, protect public health, and safeguard the environment, both in the United States and globally.

Exposure to these pollutants is associated with a wide range of adverse health effects, particularly on the respiratory and cardiovascular systems. Short-term exposures to elevated levels of  $\text{NO}_2$  exacerbate respiratory diseases such as asthma, leading to respiratory symptoms and hospitalizations [4]. Ozone, commonly known as smog, also causes respiratory irritation and exacerbates asthma, especially in vulnerable populations [5]. Fine particulate matter ( $\text{PM}_{2.5}$ ), which can penetrate deep into the respiratory tract, has been linked to various respiratory and cardiovascular health impacts, including increased hospital admissions, reduced lung function, and mortality from lung cancer and heart disease [6].

The objective of this research study is to employ advanced spatial interpolation and clustering techniques to estimate and map air pollution concentrations at unmonitored locations within our study area which consists of the selected states of New York, Pennsylvania, New Jersey, Delaware, and Maryland. The study aims to identify spatial patterns and hotspots of high air pollution concentrations in the study area, contributing to a better understanding of the magnitude and distribution of air pollution in the region. By understanding the spatial distribution of air pollution, we aim to gain valuable insights, with the goal to observe patterns and concentrations of air pollution in the study area. This research will be beneficial to policymakers for developing targeted interventions to reduce exposure and mitigate the adverse health outcomes associated with air pollution.

The findings of our study through Spatial Interpolation, indicates that Nitrogen Dioxide ( $\text{NO}_2$ ), ground level Ozone ( $\text{O}_3$ ), and Particulate Matter ( $\text{PM}_{2.5}$ ) concentrations in the study area consistently comply with the safety standards established by the Environmental Protection Agency (EPA) [1].

However, Spatial Clustering analysis unveils distinct findings in the study area. Despite generally safe levels of  $\text{NO}_2$  pollutants, New Jersey stands out with comparatively higher concentrations. Meanwhile, North Maryland emerges as a hotspot for ground level Ozone, with the highest pollutant levels concentrated in that region. Pennsylvania is also approaching the threshold for  $\text{PM}_{2.5}$  concentrations, raising concerns about potential health hazards. Notably, all other states in the study area remain below the standard threshold and are free from immediate health hazards, however, effective air quality management strategies are crucial to proactively address the impending risks.

## 2. Methods

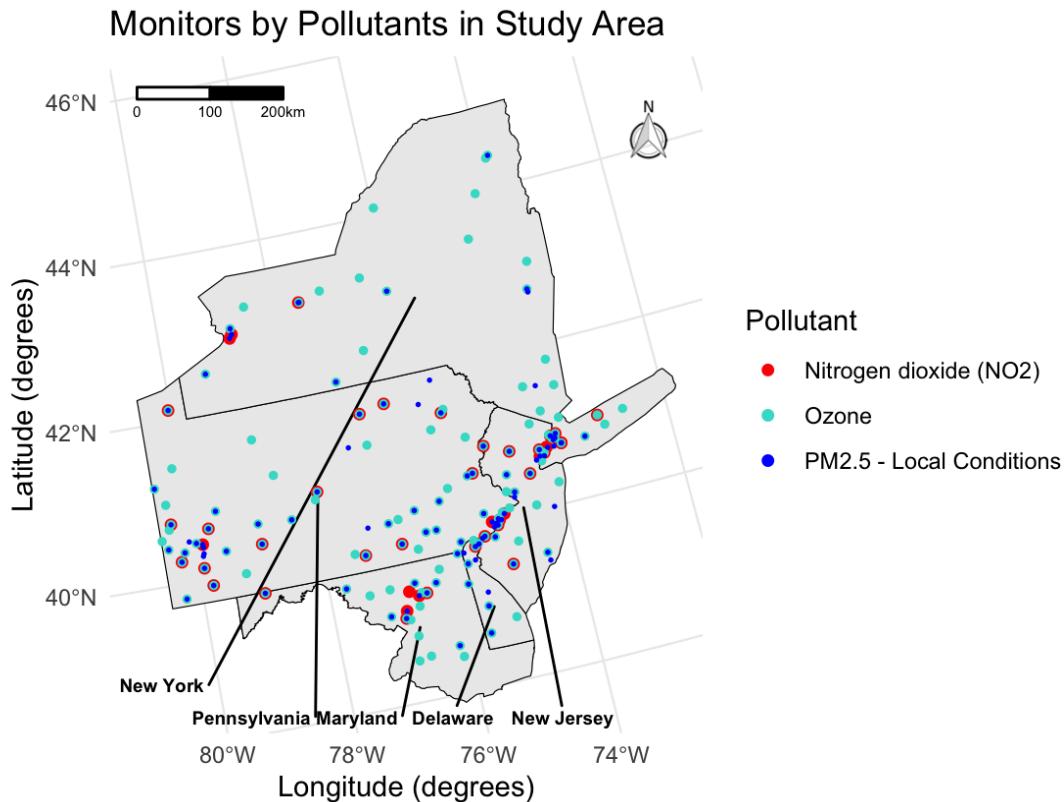
### 2.1 Study Area

The selected states of New York, Pennsylvania, New Jersey, Delaware, and Maryland represent a diverse range of climate, population, and geography, making them ideal for our research study. The northeastern region of the United States, where these states are located, exhibits a variety of climatic conditions, including temperate humid continental, humid subtropical, and oceanic climates. This diversity in climate allows for examining the potential impacts of air pollution on health outcomes under different weather conditions.

In terms of population, these states collectively have a significant population size, including densely populated urban areas, suburban regions, and rural areas. This population diversity allows for investigating the potential disparities in exposure to air pollution and health outcomes among different populations, including vulnerable groups such as children, elderly, and minority communities.

Geographically, the selected states span a range of environmental characteristics, including coastal areas, inland regions, and varying terrain. This diversity in geography enables the exploration of different sources and patterns of air pollution, such as industrial emissions, vehicular emissions, and geographical features that may influence the dispersion of air pollutants.

Overall, the selected states offer a diverse range of climate, population, and geography, making them a compelling choice for our research study to examine the relationships between air pollution and health outcomes in the northeastern region of the United States.



**Figure 1.** A map of monitors in the study area by pollutant.

## 2.2 Data

The data employed for this research comprises the 2021 annual summary of air quality data obtained from the United States Environmental Protection Agency (EPA) [7], which monitors outdoor air quality across the United States. The data pertains to a wide range of parameters, including nitrogen dioxide (NO<sub>2</sub>), ozone (O<sub>3</sub>), and particulate matter 2.5 (PM<sub>2.5</sub>). The EPA collects this data from a network of strategically placed monitors throughout the country, providing robust and up-to-date information on air quality levels.

The standards referenced in this study are safety limits outlined by the US EPA [1]. The standard used for NO<sub>2</sub> is the annual mean of 53 parts per billion (ppb). For O<sub>3</sub>, the standard is 0.070 parts per million (ppm) for the annual fourth-highest daily maximum 8-hour concentration, averaged over 3 years. For PM<sub>2.5</sub>, the standard is 12.0  $\mu\text{g}/\text{m}^3$  for the annual mean, averaged over 3 years. The number of monitors for each pollutant is outlined in Table 1.

**Table 1.** Number of monitors across the study region for each pollutant.

| Pollutant         | # of monitors |
|-------------------|---------------|
| NO <sub>2</sub>   | 44            |
| O <sub>3</sub>    | 127           |
| PM <sub>2.5</sub> | 106           |

Polygons representing the state boundaries of our study area (New York, Pennsylvania, New Jersey, Delaware, and Maryland) were used to define the study area for the interpolation. This data was acquired from the 2021 TIGER/Line Shapefiles provided by the United States Census Bureau, and filtered to the states in the study area using QGIS [8].

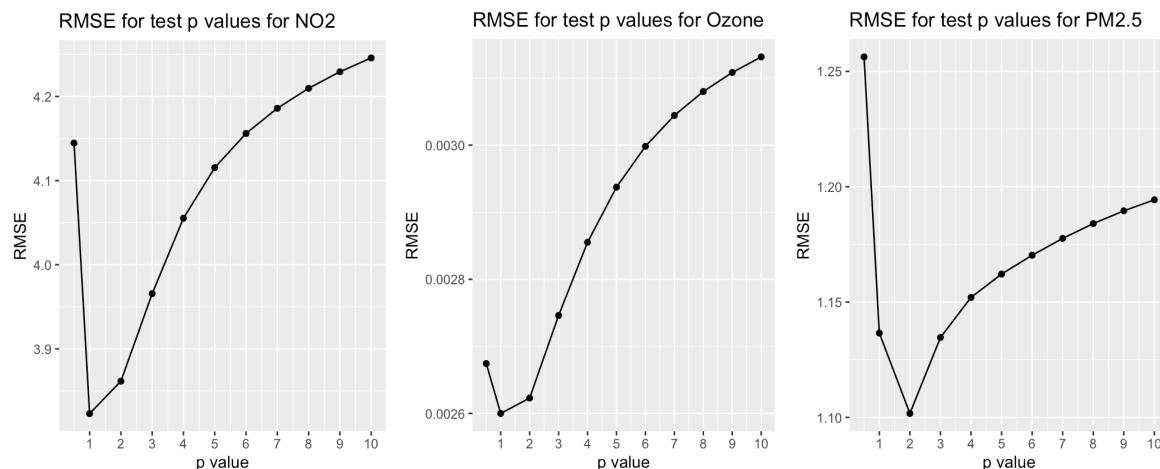
## 2.3 Statistical Analysis

In our research, we utilize spatial interpolation techniques such as Inverse Distance Weighting (IDW) and Kriging to estimate and map air pollution concentrations at unmonitored locations. IDW assigns weights to nearby monitoring stations based on their distances, while Kriging employs statistical methods to model spatial variability. These techniques enable us to fill data gaps and generate continuous maps of pollution concentrations, providing us with a comprehensive understanding of the spatial distribution of pollution.

Additionally, clustering is performed to identify spatial patterns and hotspots of air pollution concentrations. By grouping locations with similar pollution levels together, we can identify areas with high or low pollution concentrations. Clustering aids in understanding the spatial distribution of pollution, identifying potential pollution sources, and informing targeted interventions to reduce exposure and mitigate health risks.

### 2.3.1 IDW Interpolation

The first step in this study involves conducting Inverse Distance Weighting (IDW) spatial interpolation for each pollutant. A grid of equally spaced points is created to cover the study area, which is then limited to points within the study area boundaries. The initial value of the weight or power of distances, "k" (also referred to as "p" or "idp" in R), is set to 2, and IDW is performed. Next, the most suitable value of "k" is determined by looping over multiple values of "k" and performing Leave-One-Out Cross-Validation (LOOCV) for each value of "k" (see Figure 2). The parameter value with the least Root Mean Squared Error (RMSE) between the interpolated and actual values is selected. LOOCV involves sequentially removing one data point at a time from the dataset and interpolating the value at the removed point using the remaining data points, from which the RMSE can be calculated from the residuals. Based on the results of the cross-validation, the best value of "k" is selected for the final IDW interpolation. The chosen parameters are outlined in Table 2.



**Figure 2.** RMSE values for each p (k) value for each pollutant.

**Table 2.** The value of "k" in IDW, representing the weight of distances for each pollutant.

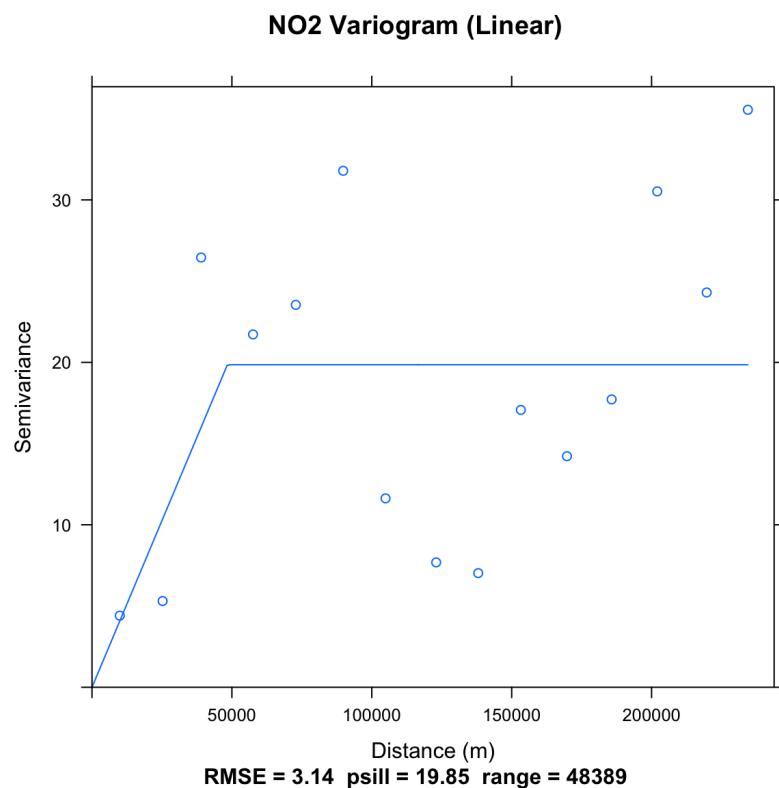
| Pollutant         | k | RMSE   |
|-------------------|---|--------|
| NO <sub>2</sub>   | 1 | 3.8231 |
| O <sub>3</sub>    | 1 | 0.0026 |
| PM <sub>2.5</sub> | 2 | 1.1017 |

### 2.3.2 Kriging

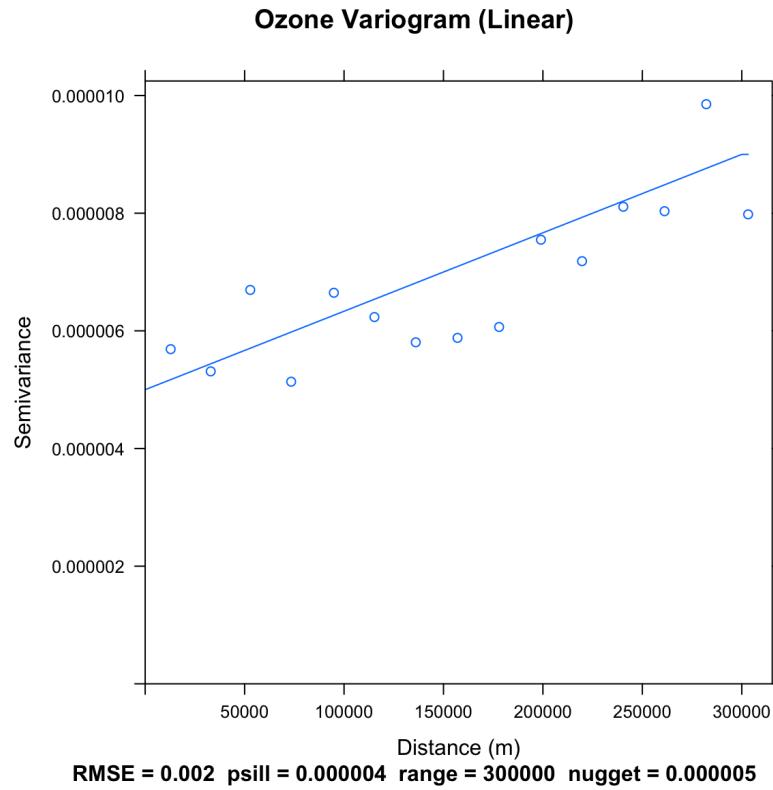
Another effective method of interpolation in this study is ordinary kriging. To perform ordinary kriging, the data is first assessed for meeting the assumptions of a normal distribution,

stationarity, and isotropy. The normal distribution assumption is checked by analyzing the histogram of the data and using the Shapiro test. The stationarity of the data is tested by measuring the mean and variance of the data across sections of the study area and checking for significant variation. For isotropy, the pattern of the IDW interpolation is analyzed for directional trends.

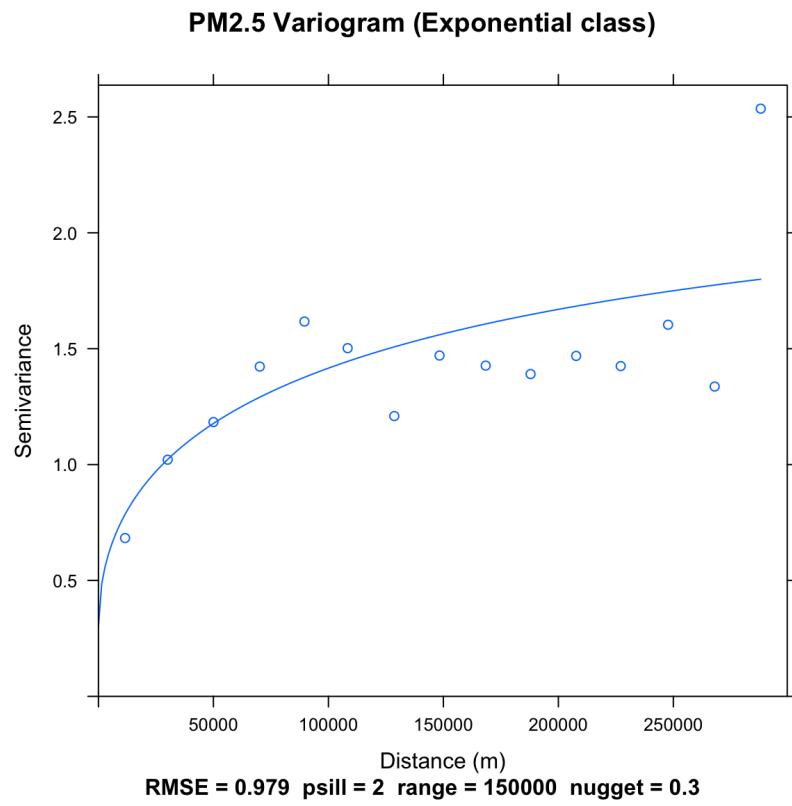
The variogram for kriging is determined using cross-validation similar to the IDW interpolation. For each variogram model type (e.g., "Sph", "Wav", "Lin"), a model is fitted to the data using gstat's fit.variogram function in R, LOOCV is performed with the specified model, and the RMSE is calculated. Building upon this model type, the variogram is manually fitted to the data if necessary, the variogram with the lowest RMSE is determined, and the sum of squared error is calculated to further validate the fit. This variogram is then used to conduct the final kriging interpolation. The variograms for each pollutant can be seen in Figure 3-5.



**Figure 3.** The variogram fitted for the nitrogen dioxide data.



**Figure 4.** The variogram fitted for the ozone data.



**Figure 5.** The variogram fitted for the PM<sub>2.5</sub> data.

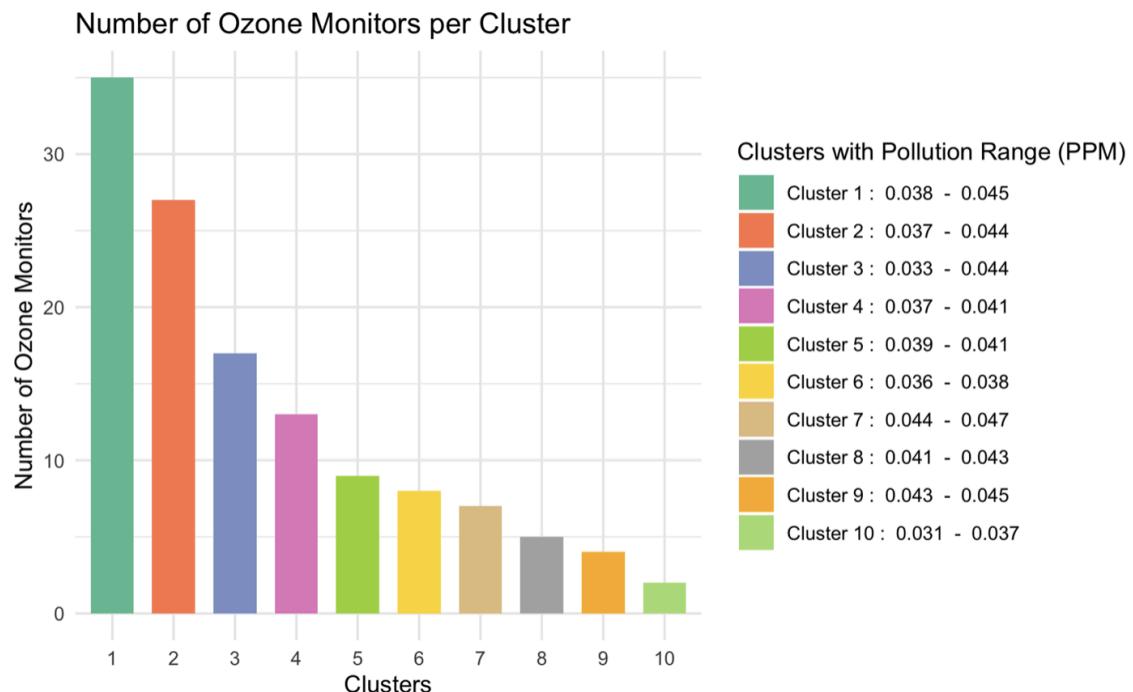
### 2.3.3 Clustering

Lastly, in this study, we employ a clustering analysis using GeoDa software to investigate the spatial patterns and clusters of air pollution concentrations. We begin by filtering the data based on the pollutants of interest, namely NO<sub>2</sub>, Ozone, and PM<sub>2.5</sub>, and export it into their respective CSV files. Next, we import the data into GeoDa software and create Queen contiguity weights using the Weights Manager tool, which allows us to establish spatial relationships between neighboring locations. In our analysis, we use the SKATER algorithm as the chosen clustering method, and through trial and error, we select 10 clusters for Ozone, 4 for NO<sub>2</sub>, and 6 for PM<sub>2.5</sub> to achieve optimal separation. To avoid singletons with only one monitor in a cluster, we set the minimum number of monitors per cluster to 2. We allow the SKATER algorithm to determine the cluster sizes based on this criterion (Figure 6-8).

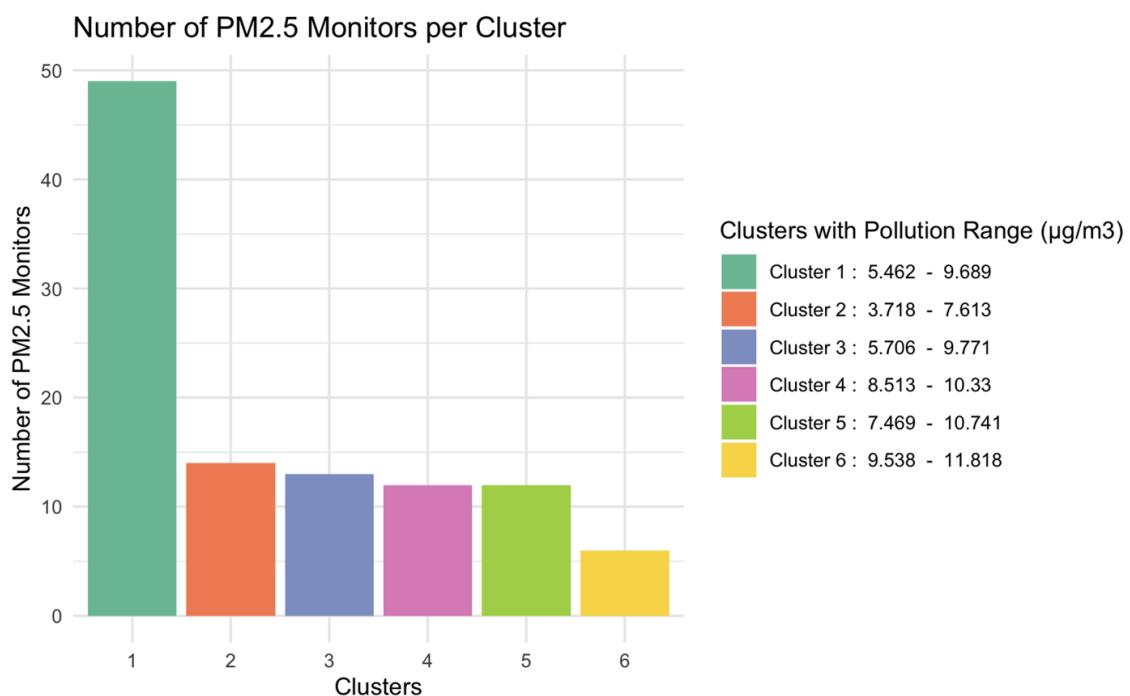
In our study, we include the arithmetic mean as an additional parameter to augment the clustering algorithm with multiple attributes. Specifically, we run the cluster analysis using the "Arithmetic Mean" attribute to enable the algorithm to determine similarity between the monitors based on the average pollutant concentration levels. The results are then exported as a CSV file and imported into R for visualization using GGplot, which allows us to plot the points according to their clusters. Furthermore, the distance function used in our clustering analysis was Euclidean, and no transformation was applied as we only had one non-spatial attribute. This clustering analysis provides valuable insights into the spatial distribution of air pollution in the selected states and aids in the development of targeted interventions to mitigate the adverse health effects associated with air pollution exposure.



**Figure 6.** Number of monitors per cluster for NO<sub>2</sub>



**Figure 7.** Number of monitors per cluster for ground level Ozone.



**Figure 8.** Number of monitors per cluster for PM<sub>2.5</sub>.

### 3. Results

#### 3.1 Descriptive Statistics

##### 3.1.1 Nitrogen dioxide (NO<sub>2</sub>)

Based on the descriptive statistics analysis in Table 3, the study area's air quality, particularly for NO<sub>2</sub> pollution, remained within EPA's safety annual standard of 53 ppb [1]. Both IDW and Kriging methods consistently yielded interpolated NO<sub>2</sub> values below this threshold, indicating safety with recommended limits. Furthermore, the RMSE of 3.82308 for the IDW method with the optimal k value of 1, and RMSE of 3.14 for the Kriging method, shows that these methods performed well in estimating NO<sub>2</sub> values. These findings highlight effective implementation of air quality management measures and acceptable levels of NO<sub>2</sub> pollution in the study area, showcasing no risk to human health, as supported by median, mean, and third quartile values.

**Table 3.** Statistics for interpolated NO<sub>2</sub> annual average (parts per billion).

| Method  | Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max    |
|---------|-------|---------|--------|-------|---------|--------|
| IDW     | 3.280 | 8.001   | 9.068  | 8.925 | 9.710   | 15.668 |
| Kriging | 1.188 | 6.708   | 6.708  | 6.733 | 6.708   | 18.168 |

##### 3.1.2 Ozone

Based on the findings of our study, the estimated ozone pollutant levels in the study area using Inverse Distance Weighting (IDW) and Kriging methods fall within the range of 0.03643 to 0.04372 parts per million (ppm). These values are found to be below the National Ambient Air Quality Standards (NAAQS) limit set by the EPA, which is 0.070 ppm for an 8-hour concentration averaged over 3 years [1]. Furthermore, the low Root Mean Squared Error (RMSE) value of 0.0026 with the optimal k value of 1 for the IDW method demonstrates the accuracy of the calculated Ozone pollutant values. Similarly, the RMSE value of 0.002 obtained for Kriging further validates the accuracy of the estimated Ozone pollutant levels in the study area. Therefore, the study area can be deemed to have normal ozone pollutant levels, indicating that the ozone concentration is within the acceptable limits as per the EPA standards, showcasing no risk to human health.

**Table 4.** Statistics for interpolated ozone annual average (parts per million).

| Method  | Min.    | 1st Qu. | Median  | Mean    | 3rd Qu. | Max     |
|---------|---------|---------|---------|---------|---------|---------|
| IDW     | 0.03643 | 0.04011 | 0.04044 | 0.04054 | 0.04086 | 0.04354 |
| Kriging | 0.03660 | 0.03844 | 0.03965 | 0.03974 | 0.04087 | 0.04372 |

##### 3.1.1 Fine particulate matter (PM<sub>2.5</sub>)

Based on the findings obtained from performing the IDW and Kriging method on the selected states, the PM<sub>2.5</sub> levels were found to be below the revised annual standard of 12.0 µg/m<sup>3</sup> and the retained 24-hour standard of 35 µg/m<sup>3</sup>, indicating good air quality in relation to PM<sub>2.5</sub> pollution [6]. The minimum, first quartile, median, mean, and third quartile values of PM<sub>2.5</sub> concentrations estimated by both IDW and Kriging are all well below the recommended limits, ranging from 3.719 µg/m<sup>3</sup> to 10.909 µg/m<sup>3</sup>. These findings suggest that the estimated PM<sub>2.5</sub> levels using both IDW and Kriging methods are within acceptable levels according to EPA standards, showing minimal risk to human health [1]. Additionally, the IDW method with a chosen k value of 2 resulted in the lowest RMSE of 1.101713, making it a suitable option for spatial interpolation of PM<sub>2.5</sub> levels in the study area. Moreover, the RMSE value of 0.979 in the Kriging method also validates the accuracy of the

$\text{PM}_{2.5}$  pollutant levels in the study area. This suggests that the IDW method, and the Kriging method performed well in estimating  $\text{PM}_{2.5}$  values.

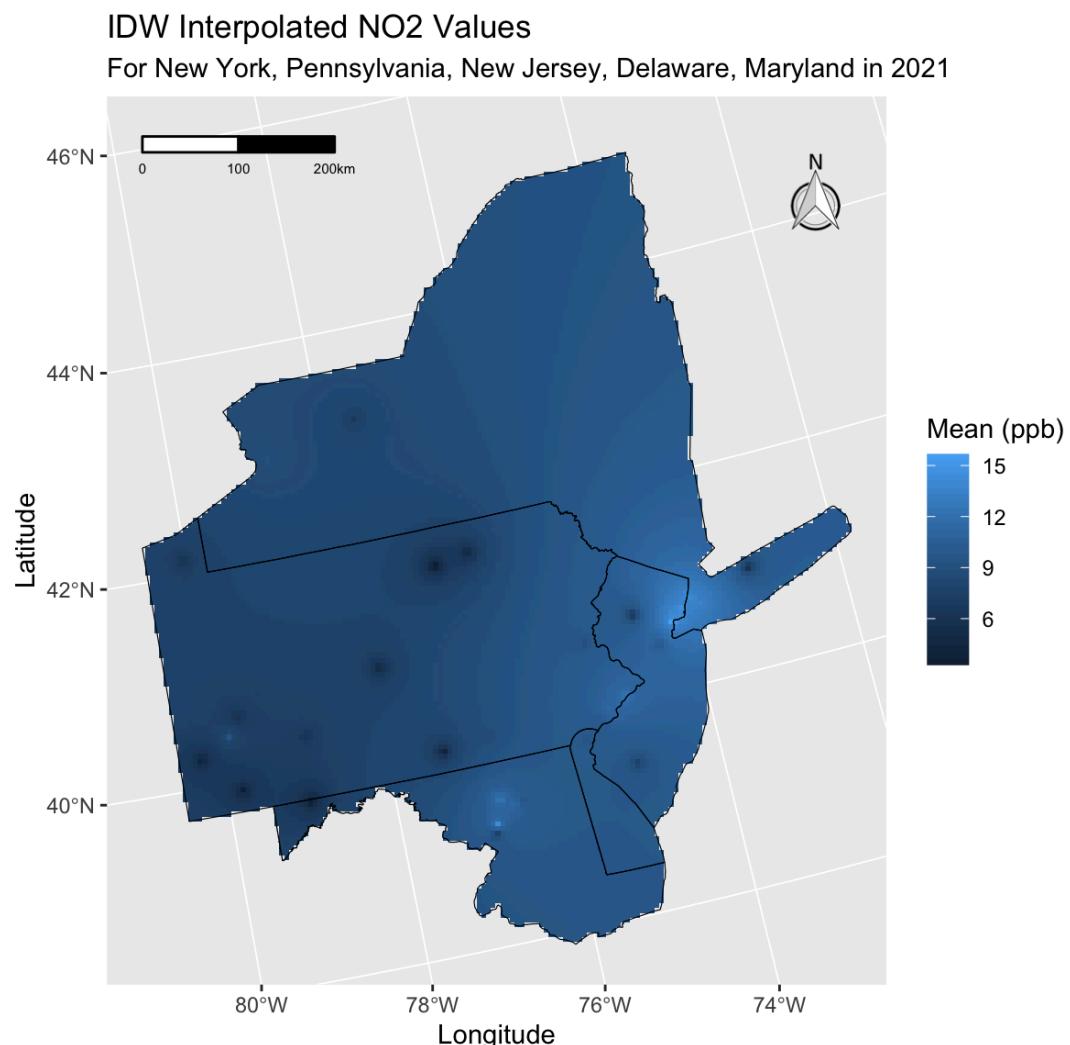
**Table 5.** Statistics for interpolated  $\text{PM}_{2.5}$  annual average (Micrograms/cubic meter)

| Method  | Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max    |
|---------|-------|---------|--------|-------|---------|--------|
| IDW     | 3.719 | 7.254   | 7.690  | 7.688 | 8.253   | 10.909 |
| Kriging | 4.371 | 6.526   | 7.247  | 7.268 | 8.083   | 10.197 |

### 3.2 IDW Interpolation Results

#### 3.2.1 Nitrogen dioxide ( $\text{NO}_2$ )

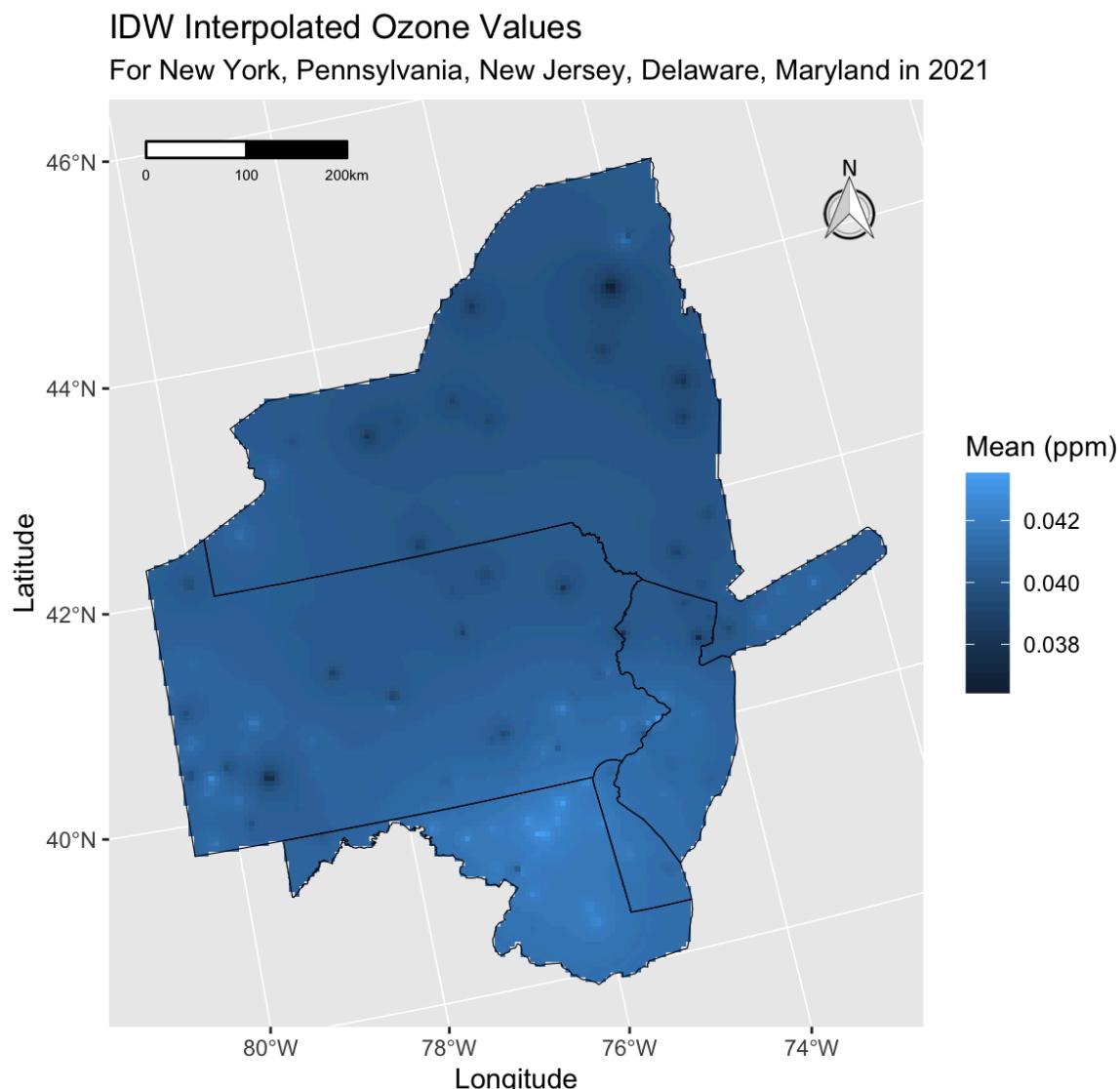
Based on the results obtained from the IDW method, it is concluded that the minimum, first quartile, median, and third quartile values of  $\text{NO}_2$  are all within acceptable limits as per the EPA safety standard of 53 ppb for the annual average [1]. These results suggest that the IDW method yields interpolated  $\text{NO}_2$  values that do not exceed the recommended limits, indicating acceptable air quality levels in the study area for the  $\text{NO}_2$  pollution.



**Figure 9.** The IDW interpolated values for  $\text{NO}_2$  in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

### 3.2.2 Ozone ( $O_3$ )

Based on our study utilizing the Inverse Distance Weighting (IDW) method to estimate ozone pollutant levels in the study area, in accordance with the EPA's NAAQS, it has revealed consistently low values ranging from 0.03643 to 0.04354 ppm [1]. These values are well below the regulatory limit of 0.070 ppm for an 8-hour concentration, averaged over 3 years, indicating that the ozone pollutant levels in the study area are not hazardous to health and do not pose significant risks to human health or the environment.

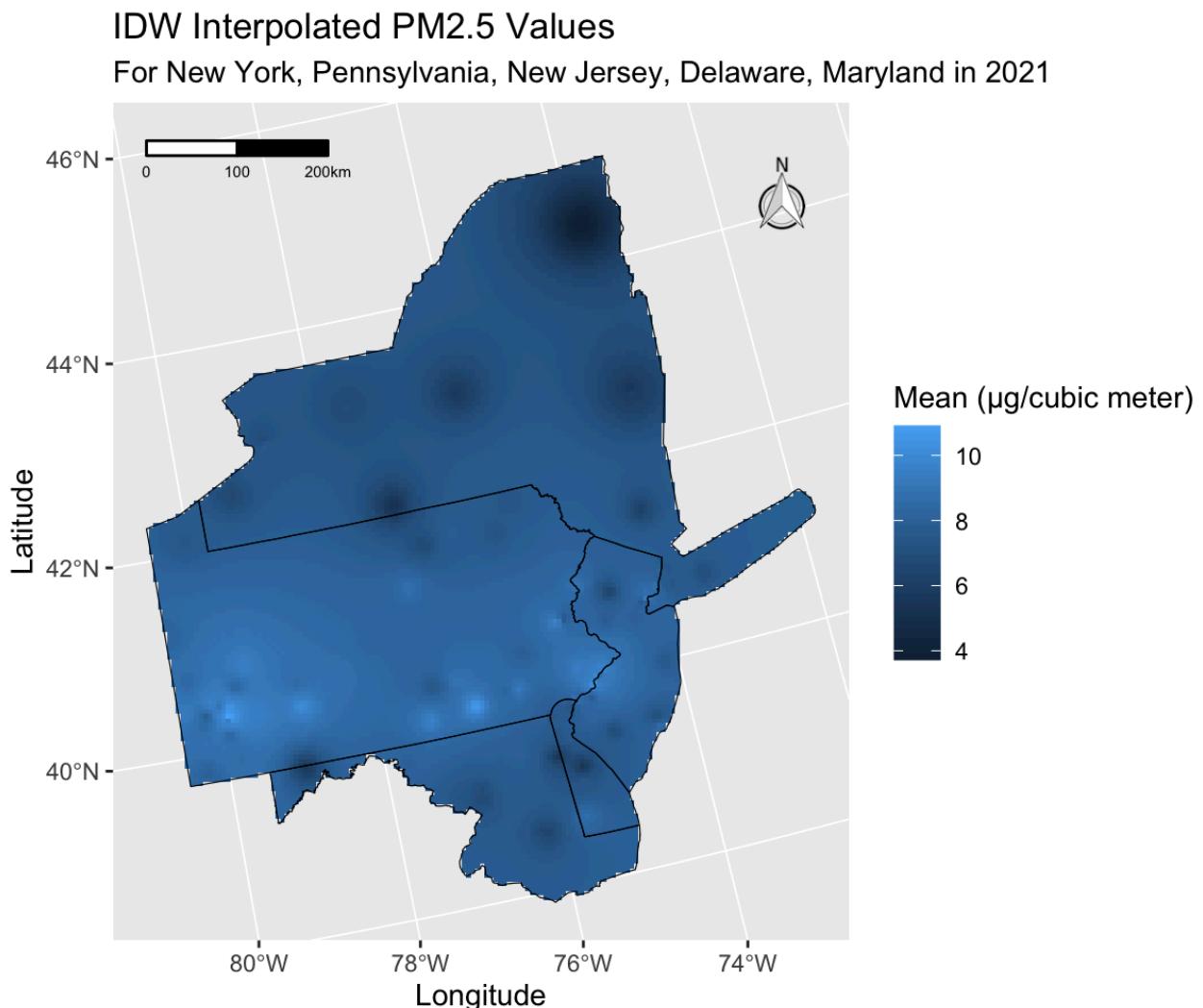


**Figure 10.** The IDW interpolated values for ozone in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

### 3.2.3 Fine particulate matter ( $PM_{2.5}$ )

Based on the findings obtained from the data, the  $PM_{2.5}$  levels, as measured by the IDW method, are below the revised annual standard of 12.0  $\mu\text{g}/\text{m}^3$  and the retained 24-hour standard of 35  $\mu\text{g}/\text{m}^3$  set by the EPA for protecting public health [6]. The minimum, first quartile, median, mean, and third quartile values of  $PM_{2.5}$  concentrations are all well below the recommended limits, ranging from 3.719  $\mu\text{g}/\text{m}^3$  to 10.909  $\mu\text{g}/\text{m}^3$ , and with an annual standard of 7.688  $\mu\text{g}/\text{m}^3$ . These

findings suggest that the PM<sub>2.5</sub> levels in the study area, as estimated by the IDW method, are within acceptable levels according to EPA standards, indicating good air quality in relation to PM<sub>2.5</sub> pollution, posing no risk to human health [1].

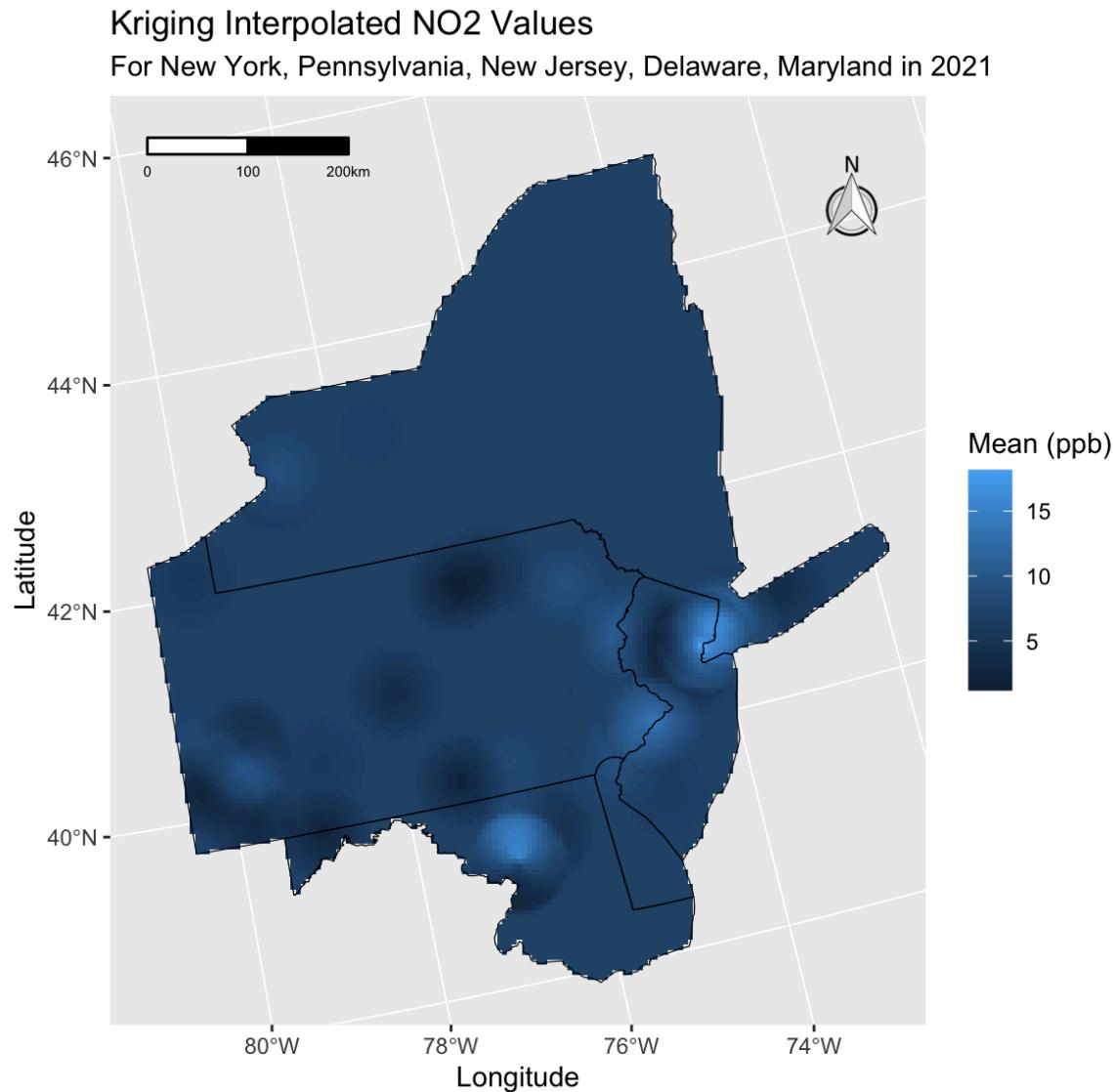


**Figure 11.** The IDW interpolated values for PM<sub>2.5</sub> in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

### 3.3 Kriging Results

#### 3.3.1 Nitrogen dioxide (NO<sub>2</sub>)

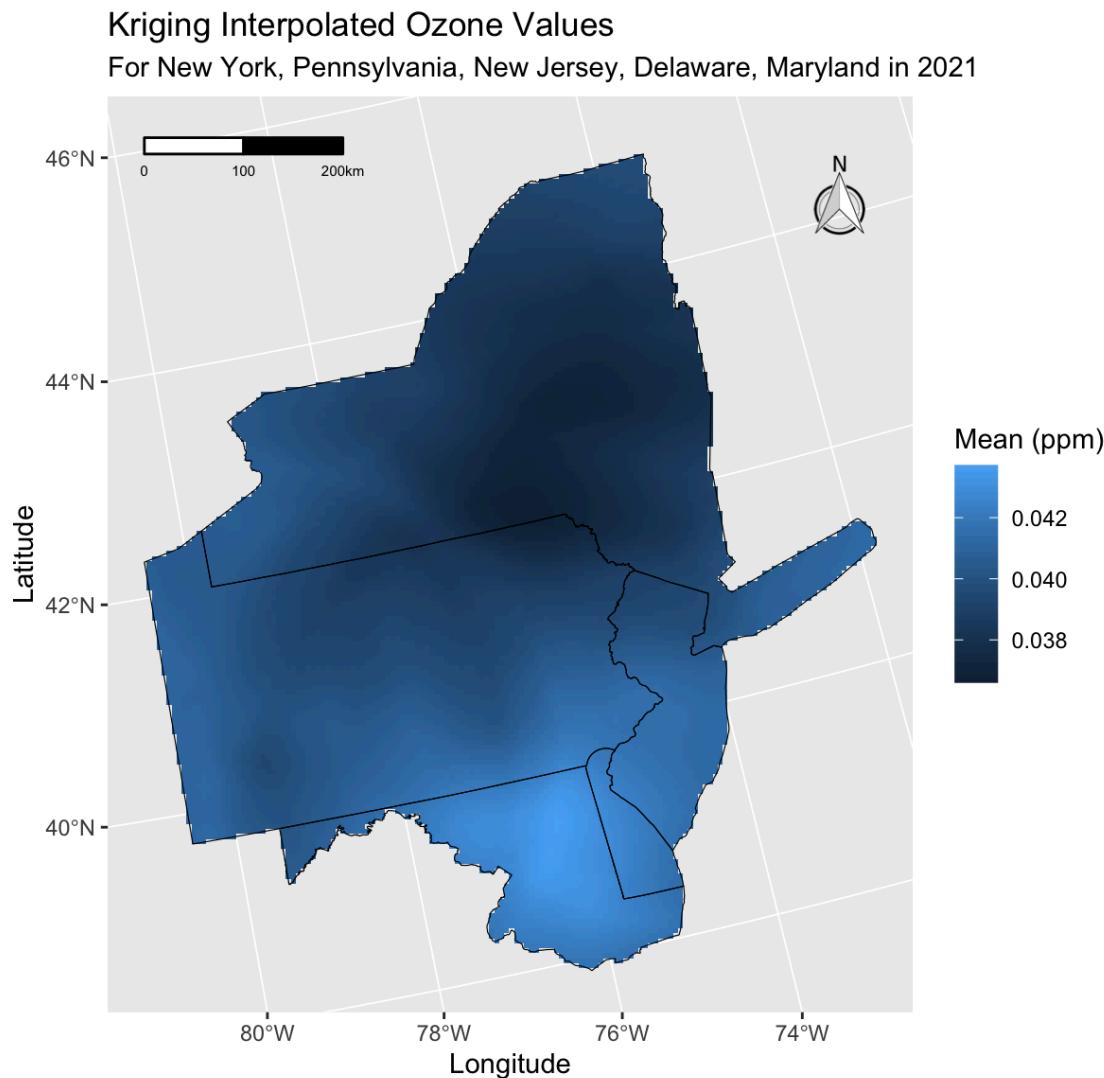
Based on the results obtained from the Kriging method, it can be concluded that the minimum, first quartile, median, and third quartile values of NO<sub>2</sub> were all within acceptable limits as per the EPA safety standard of 53 ppb for the annual average [1]. These findings suggest that the Kriging method resulted in interpolated NO<sub>2</sub> values that did not exceed the recommended limits, indicating acceptable air quality levels in the study area for NO<sub>2</sub> pollution, posing no risk to human health.



**Figure 12.** The Kriging interpolated values for NO<sub>2</sub> in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

### 3.3.2 Ozone

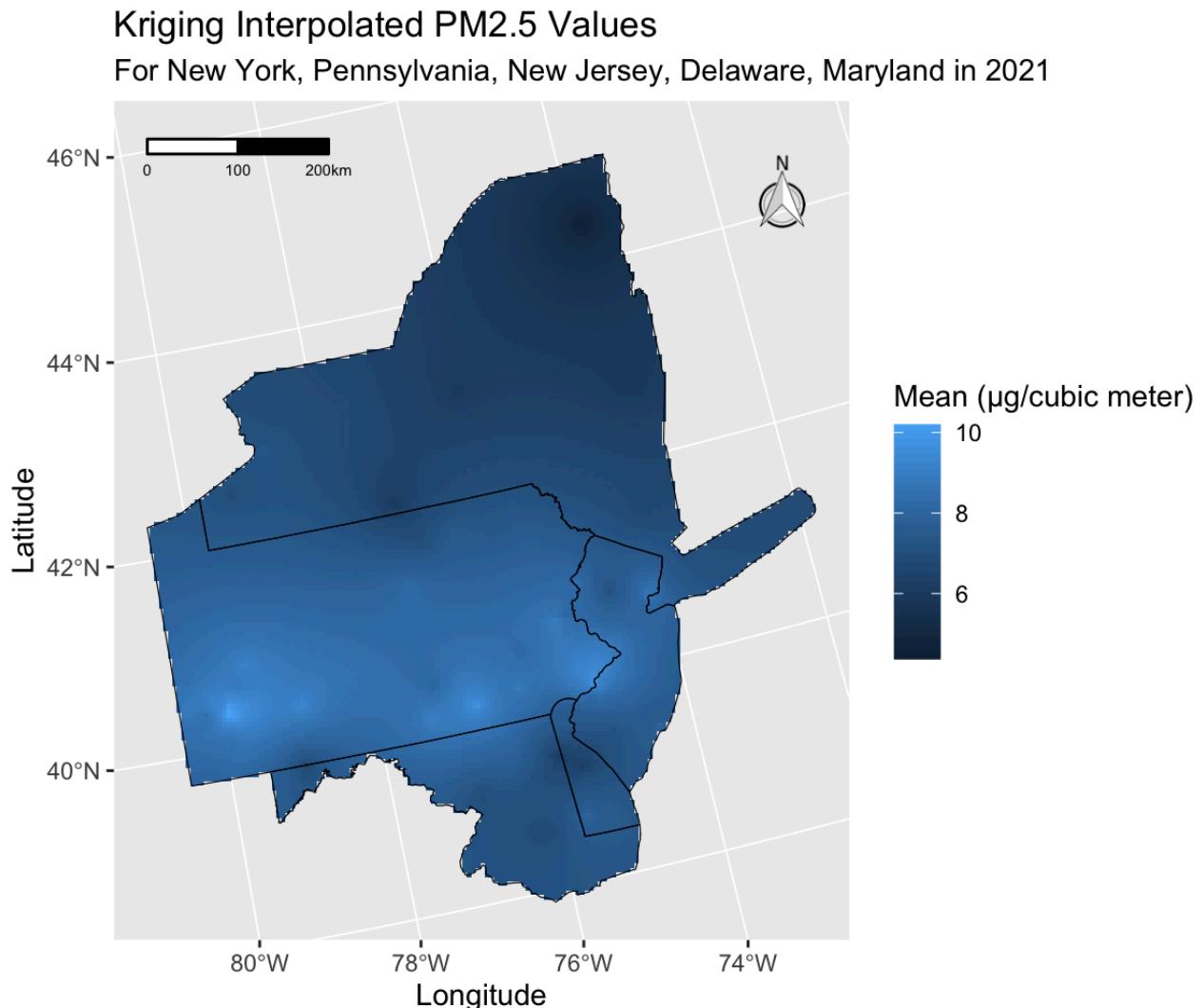
Based on the findings from our study, the Kriging method estimated ozone pollutant levels in the study area to range from 0.03660 to 0.04372 ppm. These values are consistently low and well below the EPA's NAAQS limit of 0.070 ppm for an 8-hour concentration, averaged over 3 years [1]. The Kriging method demonstrates that ozone pollutant levels in the study area are not hazardous to health and are within acceptable limits according to EPA standards. These results further support the conclusion that the ozone pollutant levels in the study area are not posing significant risks to human health or the environment.



**Figure 13.** The Kriging interpolated values for ozone in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

### 3.3.3 Fine particulate matter (PM<sub>2.5</sub>)

The Kriging method, employed in this study, yielded PM<sub>2.5</sub> concentration values ranging from 4.371 µg/m<sup>3</sup> to 10.197 µg/m<sup>3</sup>. These values are well below the revised annual standard of 12.0 µg/m<sup>3</sup> and the retained 24-hour standard of 35 µg/m<sup>3</sup> set by the EPA for protecting public health [1]. This suggests that the PM<sub>2.5</sub> levels in the study area, estimated using the Kriging method, are within acceptable levels according to EPA standards, indicating good air quality in relation to PM<sub>2.5</sub> pollution, and posing no risk to human health.



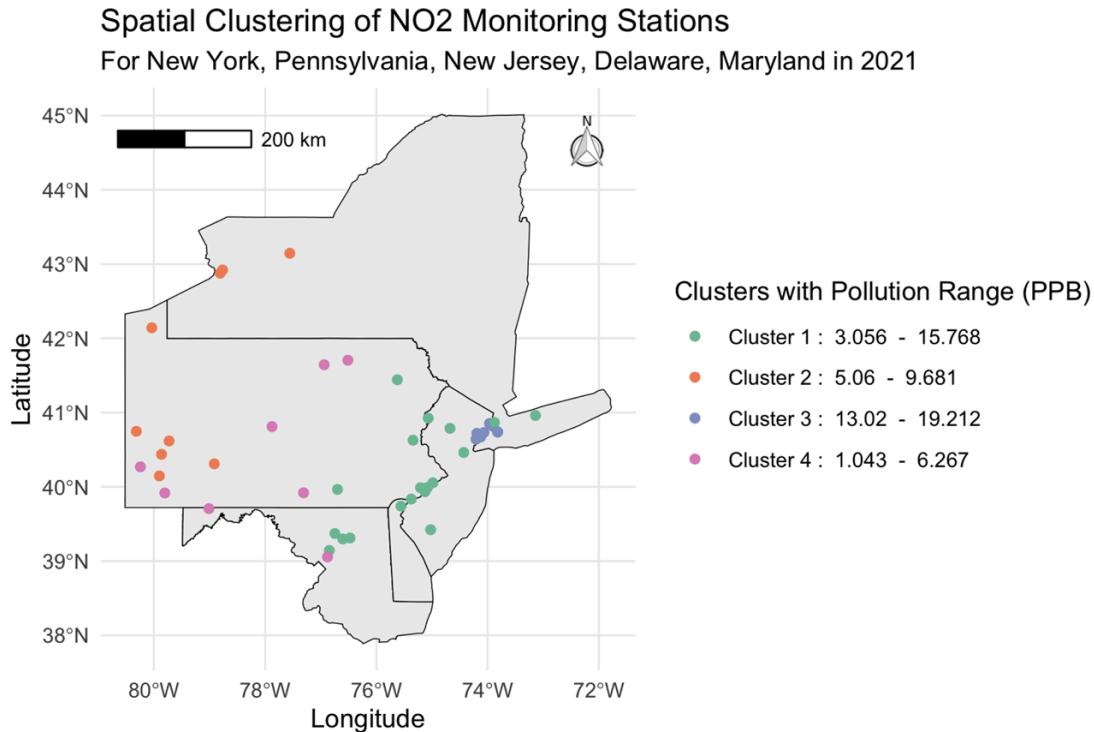
**Figure 14.** The Kriging interpolated values for PM<sub>2.5</sub> in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

### 3.4 Clustering Results

#### 3.4.1 Nitrogen dioxide (NO<sub>2</sub>)

Based on our comprehensive analysis of NO<sub>2</sub> concentrations using Spatial Clustering, the results, as depicted in Figure 15 and Table 6, reveal that Cluster 1 exhibits the widest range, spanning from 3.056 ppb to 15.768 ppb, and is predominantly observed in Eastern Pennsylvania, New Jersey, and the northern part of Maryland.

Cluster 4, on the other hand, shows the lowest NO<sub>2</sub> pollutant levels, ranging from 1.043 ppb to 6.267 ppb, mainly concentrated in Pennsylvania. Moreover, these findings also suggest that Maryland generally experiences low NO<sub>2</sub> pollution levels, as evidenced by the wide range in Cluster 1 and the smaller range in Cluster 4. The highest pollutant levels are evidently found in New Jersey. Despite the overall safe levels of NO<sub>2</sub> pollutants in the study area, New Jersey exhibits relatively higher levels compared to other regions (Figure 15).



**Figure 15.** Spatial Clustering for NO<sub>2</sub> in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

**Table 6.** Statistics for clustered NO<sub>2</sub> annual average (parts per billion)

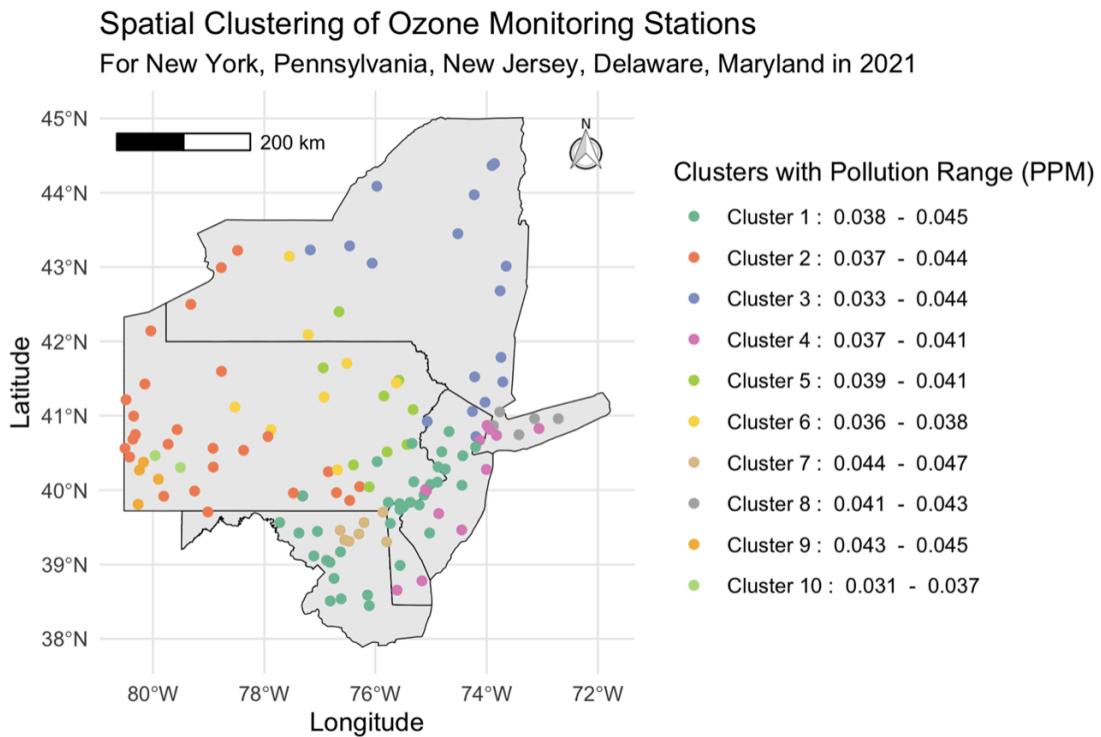
| Cluster   | Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max    |
|-----------|-------|---------|--------|-------|---------|--------|
| Cluster 1 | 3.056 | 8.325   | 9.486  | 9.950 | 12.165  | 15.768 |
| Cluster 2 | 5.060 | 5.770   | 6.626  | 7.184 | 8.243   | 9.681  |
| Cluster 3 | 13.02 | 14.81   | 15.11  | 15.49 | 15.76   | 19.21  |
| Cluster 4 | 1.043 | 2.056   | 2.746  | 2.900 | 3.093   | 6.267  |

### 3.4.2 Ground level Ozone (O<sub>3</sub>)

Through our Spatial Clustering analysis for the ozone pollutant concentrations, we can conclude our findings based on Figure 16 and Table 7. The entire study area exhibits a narrow range, ranging from 0.031 ppm to 0.045 ppm. However, Cluster 3 exhibits the widest range, ranging from 0.003 ppm to 0.044 ppm, indicating a distinct variation in New York. Cluster 10, mainly found in Pennsylvania with a small portion in New York, also shows a wide range of 0.031 ppm to 0.037 ppm.

Conversely, Cluster 6 demonstrates the most average pollutant level, ranging from 0.036 to 0.038, spread over Pennsylvania. The highest pollutant level in our study area is observed in Cluster 7, ranging from 0.044 ppm to 0.047 ppm, covering North Maryland and sharing borders with Delaware and Pennsylvania. Overall, the health risk in the study area is not significant, but through

the Clustering analysis, it reveals that the highest pollutant levels are concentrated in North Maryland (Figure 16).



**Figure 16.** Spatial Clustering for O<sub>3</sub> in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

**Table 7.** Statistics for clustered ozone annual average (parts per million)

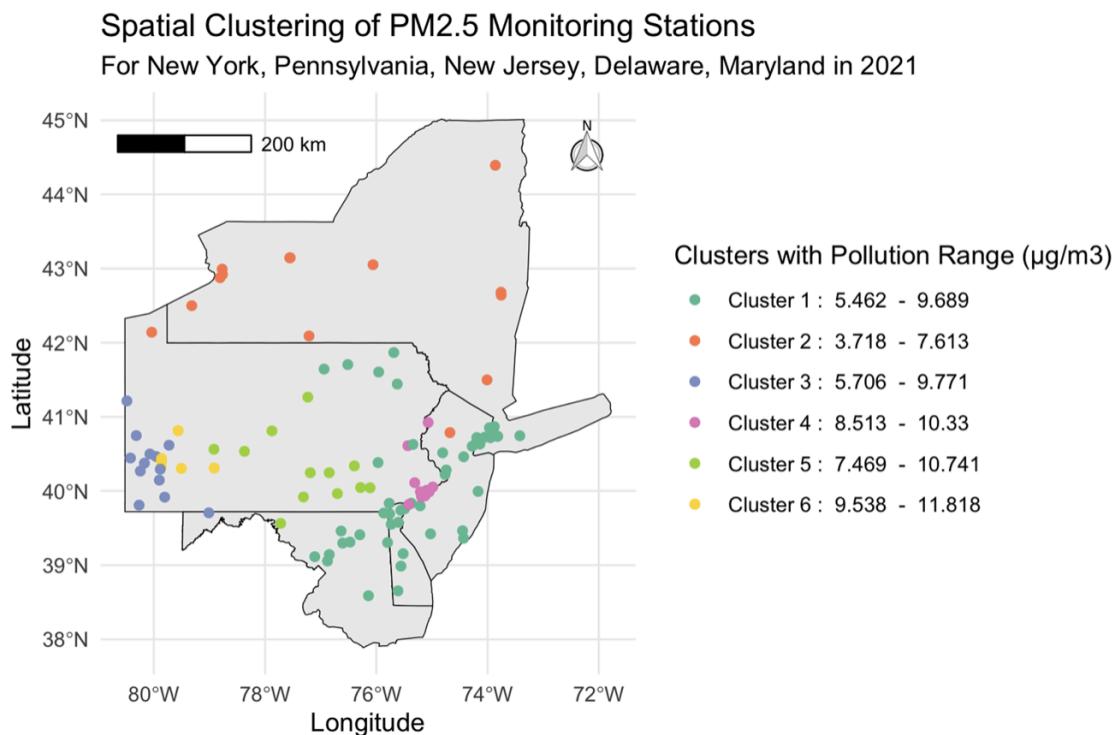
| Cluster   | Min.    | 1st Qu. | Median  | Mean    | 3rd Qu. | Max     |
|-----------|---------|---------|---------|---------|---------|---------|
| Cluster 1 | 0.03774 | 0.04182 | 0.04255 | 0.04251 | 0.04364 | 0.04488 |
| Cluster 2 | 0.03708 | 0.03966 | 0.04158 | 0.04083 | 0.04197 | 0.04369 |
| Cluster 3 | 0.03284 | 0.03715 | 0.03826 | 0.03772 | 0.03851 | 0.04389 |
| Cluster 4 | 0.03714 | 0.03929 | 0.03979 | 0.03958 | 0.04032 | 0.04149 |
| Cluster 5 | 0.03856 | 0.03957 | 0.04012 | 0.03996 | 0.04057 | 0.04088 |
| Cluster 6 | 0.03600 | 0.03640 | 0.03670 | 0.03688 | 0.03739 | 0.03805 |
| Cluster 7 | 0.04392 | 0.04445 | 0.04520 | 0.04526 | 0.04598 | 0.04685 |
| Cluster 8 | 0.04101 | 0.04124 | 0.04259 | 0.04214 | 0.04264 | 0.04321 |

|            |         |         |         |         |         |         |
|------------|---------|---------|---------|---------|---------|---------|
| Cluster 9  | 0.04253 | 0.04260 | 0.04295 | 0.04337 | 0.04372 | 0.04504 |
| Cluster 10 | 0.03148 | 0.03291 | 0.03434 | 0.03434 | 0.03577 | 0.03720 |

### 3.4.3 Fine particulate matter ( $\text{PM}_{2.5}$ )

Based on our results through Figure 17 and Table 8, we can conclude our findings for Spatial Clustering in our study area for fine particulate matter. Clusters 1 and 2 exhibit the widest range of values, ranging from 5.462 to 9.689 for Cluster 1, and from 3.718  $\mu\text{g}/\text{m}^3$  to 7.613  $\mu\text{g}/\text{m}^3$  for Cluster 2. Cluster 1 is spread across New Jersey, Delaware, Northeast Maryland, and Northern Pennsylvania, showcasing distinct spatial variation.

Notably, Cluster 4, 5, and 6 approach the standard safe level of  $\text{PM}_{2.5}$ , which is 12.0  $\mu\text{g}/\text{m}^3$  annually [1]. The states that are on the verge of exceeding this threshold are Pennsylvania and the border region of Pennsylvania and New Jersey. Despite being overall safe, these states are nearing the unhealthy level of this pollutant.



**Figure 17.** Spatial Clustering for  $\text{PM}_{2.5}$  in New York, Pennsylvania, New Jersey, Delaware, and Maryland in 2021.

**Table 8.** Statistics for clustered  $\text{PM}_{2.5}$  annual average ( $\mu\text{g}/\text{m}^3$ )

| Cluster   | Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max   |
|-----------|-------|---------|--------|-------|---------|-------|
| Cluster 1 | 5.462 | 7.239   | 7.638  | 7.613 | 7.963   | 9.689 |
| Cluster 2 | 3.718 | 5.821   | 6.369  | 6.314 | 7.077   | 7.613 |

|           |       |       |        |        |        |        |
|-----------|-------|-------|--------|--------|--------|--------|
| Cluster 3 | 5.706 | 8.082 | 8.366  | 8.334  | 8.742  | 9.771  |
| Cluster 4 | 8.513 | 9.271 | 9.555  | 9.560  | 10.029 | 10.330 |
| Cluster 5 | 7.469 | 8.273 | 8.621  | 8.754  | 9.030  | 10.741 |
| Cluster 6 | 9.538 | 9.683 | 10.222 | 10.340 | 10.610 | 11.818 |

#### 4. Discussion

Our analysis reveals that  $\text{NO}_2$  has higher concentration levels than  $\text{PM}_{2.5}$ , and spatial clustering results show that its clusters are related to urban and industrial presence in the area, such as Pittsburgh. For example, in Figure 15, cluster 3 with the highest mean value is confined to New York, reflecting the primary causes of  $\text{NO}_2$ , which are burning fossil fuels, vehicles, and manufacturing [9].  $\text{PM}_{2.5}$  clusters from Figure 17 are mostly in natural areas, as it is more inclined to natural and agricultural causes [10]. The concentrations of  $\text{O}_3$  are significantly higher than both pollutants.

Similar studies have utilized various methods to estimate air pollution concentrations. Interpretable convolutional neural networks (CNNs) proved to be highly accurate in such estimates [11] (pp. 5-6). The results using this model were similar to the results of the IDW and Kriging techniques, with the annual mean  $\text{PM}_{2.5}$  following the same concentration patterns we observed. However, in southern Pennsylvania, the CNNs model observes higher concentration. This could be due to the CNNs model's use of deep learning techniques to capture complex, non-linear relationships. This model uses multiple input metrics (i.e satellite data, land use etc.) and a larger geographical scale, which may result in different interpolated values and error.

A higher number of monitors in the concerned region would likely produce more accurate results using the methods we used [12]. The low RMSE values implying higher accuracy in the study came from the abundance of monitors in a relatively small region. Moreover, the use of a fixed polygon for the region can lead to variations in interpolation results, as boundaries and other parameters may change over time. To optimize our results, using a smaller network of mobile monitoring could be considered. This also minimizes variation in estimates from different methods due to using a denser network of monitors (as exemplified in the paper for Mexico City). This approach could be implemented in our research to further enhance the accuracy of our results.

Furthermore, we can reinforce this study with an interpolation process that entices diversity and minimizes sample bias. A procedure called Spatial-Temporal Point Interpolation [13] (p. 1) assists with this. The use of machine learning techniques, and mobile monitoring networks can be well reinforced by this utility to produce even more accurate results. It addresses missing datasets, and reduces misrepresentation of regions which is especially important to consider in such a geographical scale. For regions scarce in monitoring stations, sample bias is a prevalent risk as a result of the Kriging method, and this interpolation method fixes that [13] (p. 6).

#### 5. Conclusions

In conclusion, our research utilizing Spatial Interpolation techniques reveals that concentrations of Nitrogen Dioxide ( $\text{NO}_2$ ), ground level Ozone ( $\text{O}_3$ ), and Particulate Matter ( $\text{PM}_{2.5}$ ) in the study area are consistently compliant within the safety standards established by the Environmental Protection Agency (EPA) [1]. However, Spatial Clustering analysis uncovers notable findings, including higher  $\text{NO}_2$  levels in New Jersey, and North Maryland as a hotspot for ground level Ozone. Additionally, Pennsylvania is approaching the threshold for  $\text{PM}_{2.5}$  concentrations,

posing potential health hazards. While other states in the study area are currently free from immediate health hazards, effective air quality management strategies are imperative to proactively address the impending risks.

## References

1. Environmental Protection Agency. (2023, March 15). *NAAQS Table*. NAAQS Table EPA. Retrieved April 9, 2023, from <https://www.epa.gov/criteria-air-pollutants/naaqs-table>
2. Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., ... & Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The lancet*, 389(10082), 1907-1918. [https://doi.org/10.1016/S0140-6736\(17\)30505-6](https://doi.org/10.1016/S0140-6736(17)30505-6)

3. Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., & Kaufman, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental health*, 12(1), 1-16. <https://doi.org/10.1186/1476-069X-12-43>
4. Environmental Protection Agency. (2022, August 2). *Basic Information about NO<sub>2</sub>*. EPA. Retrieved April 9, 2023, from <https://www.epa.gov/no2-pollution/basic-information-about-no2#Effects>
5. Environmental Protection Agency. (2022, June 14). *Ground-level Ozone Basics*. EPA. Retrieved April 9, 2023, from <https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#wwh>
6. New York State. (2018, February). *Department of Health*. Fine Particles (PM 2.5) Questions and Answers. Retrieved April 9, 2023, from [https://www.health.ny.gov/environmental/indoors/air/pmq\\_a.htm](https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm)
7. *United States Environmental Protection Agency Air Data Pre-Generated Data Files*. United States Environmental Protection Agency. (2022, November 14). Retrieved April 9, 2023, from [https://aqs.epa.gov/aqsweb/airdata/download\\_files.html](https://aqs.epa.gov/aqsweb/airdata/download_files.html)
8. *United States Census Bureau TIGER/Line Shapefiles*. United States Census Bureau. (2022, December 5). Retrieved April 9, 2023, from <https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2021.html#list-tab-790442341>
9. Ji, J. S., Liu, L., Zhang, J. (J., Kan, H., Zhao, B., Burkart, K. G., & Zeng, Y. (2022, October 13). *No<sub>2</sub> and PM<sub>2.5</sub> air pollution co-exposure and temperature effect modification on pre-mature mortality in advanced age: A longitudinal cohort study in china - environmental health*. BioMed Central. Retrieved April 10, 2023, from <https://ehjournal.biomedcentral.com/articles/10.1186/s12940-022-00901-8>
10. Government of Ontario, Ministry of the Environment. (n.d.). *Notice: Scheduled network maintenance*. Fine Particulate Matter. Retrieved April 10, 2023, from <http://www.airqualityontario.com/science/pollutants/particulates.php>
11. Yang, Q., Wu, J., Liu, Y., Kloog, I., Hu, X., Bach, S., Chakma, A., Di, Q., Goodfellow, I., Gupta, P., He, K., Hinton, G., & Kingma, D. P. (2019, October 23). *Estimating PM<sub>2.5</sub> concentration of the conterminous United States via interpretable convolutional Neural Networks*. Environmental Pollution. Retrieved April 10, 2023, from <https://www.sciencedirect.com/science/article/abs/pii/S0269749119335341>
12. Rivera-González, L. O., Zhang, Z., Sánchez, B. N., Zhang, K., Brown, D. G., Rojas-Bracho, L., Osornio-Vargas, A., Vadillo-Ortega, F., & O'Neill, M. S. (2015, May). *An assessment of air pollutant exposure methods in Mexico City, Mexico*. Journal of the Air & Waste Management Association (1995). Retrieved April 10, 2023, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4670782/>
13. Xu, C., Wang, J., Hu, M., & Wang, W. (2022, September 21). *A new method for interpolation of missing air quality data at Monitor stations*. Environment International. Retrieved April 10, 2023, from <https://www.sciencedirect.com/science/article/pii/S0160412022004652>