Effects of Air Pollution Variables on Health Disparities in the Philadelphia Region

Karan Sampath* and Blanca E. Himes, PhD

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA

**Correspondence can be sent to ksampath@seas.upenn.edu*

Introduction

Air Pollution is known to be a pervasive and extremely harmful phenomenon and is a major cause of health care and morbidity costs in the United States[1]. The usage of Census and Environmental Protection Agency (EPA) data is valuable to study the effects of such relationships in current tract areas, and due to the comprehensive nature of the data, it is used here to model the effects of air pollution variables on health disparities in the Philadelphia Region.

Previous literature has shown that PM 2.5 pollutants significantly and disproportionately affect people of color every year[2]. We now expand this to every criteria air pollutant, focusing on the Philadelphia region to look at whether any specific region irregularities were not reflected in wider US data.

Linking EPA Data to Census-Derived Data

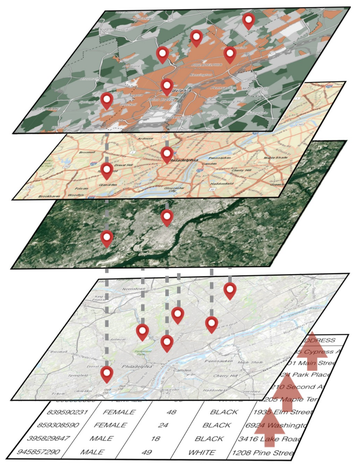


Figure 1. Geospatial variables derived from EPA data can be used to spatially link EPA data to census data on Philadelphia tract regions.

EPA data for 2020 from recording stations in the vicinity of the Philadelphia region was linked to Census data through the following steps:

1. **Read and wrangle the EPA data.** *This step requires cleaning of data to ensure that the data is readable and parsed into an appropriate format.*
2. **Read Census data and shapefiles.** *Census data and shapefiles should be read in using appropriate American Community Survey (ACS) data, adjusting for state and county constraints.*
3. **Fit EPA data to census polygons.** *EPA data should be converted and rasterized, with the means of the data used while fitting to polygons. This is then converted to a data frame to allow for effective correlation.*
4. **Assess associations between pollution data from the EPA and demographic ACS data.** *Correlations can be modelled by using individual univariate analyses, while accounting for significant p-values, with a standard p-value of 0.01 being used here.*

Methods

Data Sources Data was collected from the Environmental Protection Agency (EPA) and the US Census Bureau. The EPA data consisted of daily recordings of criteria air pollutant levels in the year 2020 for the Philadelphia region. The Census Bureau data was American Community Survey estimates for the year 2019, the most recent iteration of the 5-year ACS. Means of the EPA data were calculated and then used with the tract level data from the ACS.

Exploratory analysis Initial exploratory analysis was conducted to understand the salient features of the datasets, including the size, mean and standard deviation. Furthermore, the census data was cleaned to find percentages of various ethnicities in overall population. Pollutant data was plotted out as heatmaps to only use pollutants with more than one recording station in the region, which led to PM 10 data being discarded for this project.

Statistical analysis A correlation matrix using Pearson's correlation was first computed for the various pollutants with individual ethnicities. A full pairwise matrix was plotted to see if there were any other interesting correlations. Following this, a p-value of 0.01 was used to disregard insignificant values in the matrix. The final correlation matrix was plotted as an upper triangular matrix.

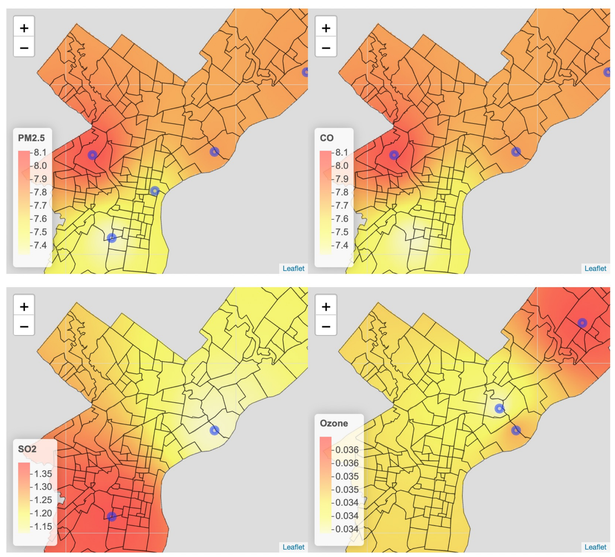


Figure 2. Heatmaps of the prevalence of various EPA criteria air pollutants in the Philadelphia region. Recording stations are highlighted in blue, with rasters used to create the heatmap. (Clockwise from top left: PM 2.5, Carbon Monoxide, Ozone, and Sulphur Dioxide).

Results

Figure 3. Hispanic or Latino demographic concentrations in the Philadelphia region.

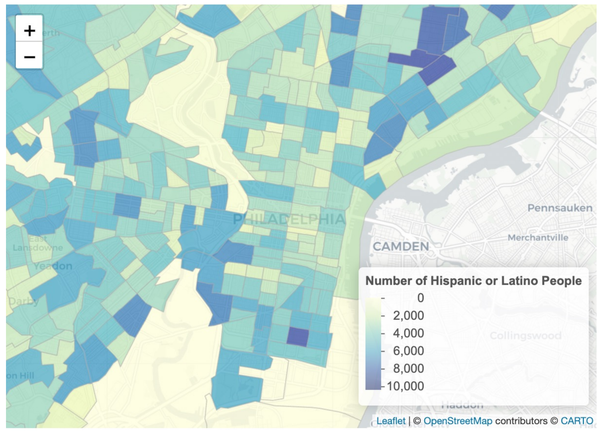
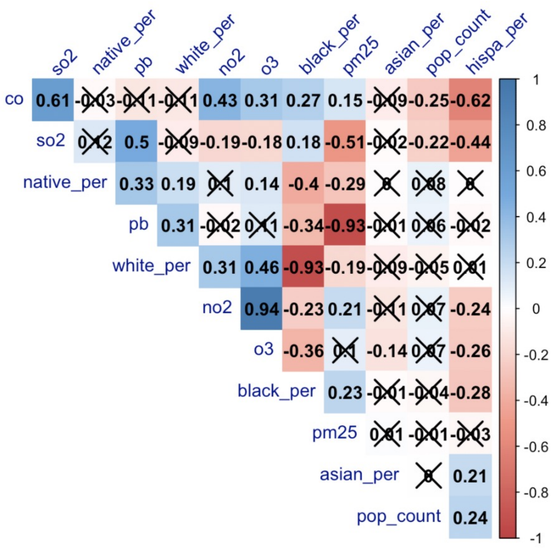


Figure 4. Upper triangle correlation matrix plots visualized with correlations represented and insignificant results removed from the visualization.



Key: *co - Carbon Monoxide prevalence, so2 – Sulphur Dioxide prevalence, native_per – Native American percentage of population, pb – Lead prevalence, White_per – White percentage of population, no2 – Nitrogen Dioxide prevalence, o3 – Ozone prevalence, black_per – Black percentage of population, pm25 – PM 2.5 prevalence, asian_per – Asian American percentage of population, pop_count - Population Count of Census Tract, hispa_per - Hispanic American percentage of population.*

There is a significant negative correlation between areas with higher Hispanic population percentages and the presence of Carbon Monoxide, Sulphur Dioxide, Nitrogen Dioxide and Ozone. Similarly, there is a significant negative correlation between areas with higher Black population percentages and the presence of Lead, Nitrogen Dioxide and Ozone, while there is a positive correlation with Carbon Monoxide, Sulphur Dioxide and PM 2.5. White population percentages are positively correlated with Lead, Nitrogen Dioxide and Ozone, with only PM 2.5 showing a slight negative correlation. Asian American population percentages are mostly uncorrelated with all pollutants.

The results also indicate strong negative correlations between the black and white population percentages, reflecting the presence of ethnic enclaves in the Philadelphia region. The presence of pollutants such as PM 2.5 are also significantly negatively correlated with pollutants like Lead and Sulphur Dioxide. This suggests the presence of a sparse distribution of polluting structures for PM 2.5 across the region. Similarly, the presence of Carbon Monoxide and Ozone is very strongly correlated with other criteria air pollutants, suggesting that those polluting structures are densely clustered.

Conclusion

There are multiple conclusions to draw from the data. While previous literature concluded that pollution disproportionately affected communities of color, this only holds true for some pollutants in the Philadelphia region. Moreover, the most disproportionately affected community in the Philadelphia region was White Americans, showing strong positive correlations for their population. This could be due to the income levels of White Americans in the Philadelphia region, thereby masking a stronger correlation.

Further Discussion

To further improve this work, correlations with income could also be looked at as a future direction. Moreover, the relationship between the various variables could also be modelled using a regression framework rather than only an exploratory analysis.

These conclusions also raise a few important questions for further study. Primarily, what factors determine the placement of structures such as factories most at fault for emitting criteria air pollutants? What policies can be laid out for places affected by high prevalence of criteria air pollutants?

Acknowledgements

There are a few individuals without whom this research would not have been possible. First, I would like to thank my mentor, Dr. Blanca E. Himes for her constant support and mentorship throughout the project. I would also like to thank Colin Christie, Dr. Sherrie Xie, Avantika Diwadkar and Alexandra Rizaldi for their help in this project. Finally, I would like to thank the Center for Undergraduate Research and Fellowships for giving me the ability to conduct independent research by sponsoring it under the Penn Undergraduate Research Mentoring Program.

Bibliography 1. Bernstein, Jonathan A., et al. Journal of allergy and clinical immunology. 2004 2. Tessum et al. *Science Advances, 2021.*