Final Project Report:

Investigating the Relationship between Air Pollution and Educational Attainment in India

GEOG:3520 GIS for Environmental Studies

Professor David Bennett

May 9, 2024

Amira Qidwai

Introduction & Problem Statement

As air pollution worsens worldwide, scholars have become increasingly interested in the relationship between it and the characteristics of the people it affects. Air pollution, like many anthropologically driven environmental phenomena, does not impact all people equally (Jerrett et al., 2001). Several studies have found that high levels of air pollution are correlated with certain socioeconomic factors. For example, high exposure to air pollution is associated with low incomes, low dwelling values, unemployment, social and material deprivation, immigration status, and higher neonatal and infant mortality rates (Hassan et al., 2022; Crouse et al., 2009, Jerrett et al., 2001; Goyal et al., 2019). These investigations show that the burden of air pollution falls unequally on populations. They also demonstrate why further understanding of the relationship between air pollution and socioeconomic factors is necessary. Studies like this have been conducted around the world, in developed and developing countries alike, but few studies have investigated the relationship between air pollution and educational attainment. Educational attainment is one of many factors that impacts an individual's socioeconomic status and could plausibly have a relationship with levels of air pollution.

American analysts have begun to scratch the surface of this issue by finding smaller scale relationships between education and air pollution. For example, one investigation focused on the relationship between standardized test scores and air pollution in Salt Lake City, Utah (Mullen et al., 2020). They found a strong relationship between frequent high levels of air pollution, as measured by $PM_{2.5}$ values, and reduced proficiency in math and ELA amongst third graders (Mullen et al., 2020). Another study focusing on older adults across the U.S. analyzed the relationship between $PM_{2.5}$ values and the incidence of cognitive impairment, specifically investigating how adults' level of education modified their incident of cognitive impairment (Alishire & Walsemann, 2021). They found that the probability of impairment is increased with greater concentrations of $PM_{2.5}$ among those with 8 or fewer years of education, but $PM_{2.5}$ concentrations are unrelated to cognitive impairment among those with 13 or more years of education (Alishire & Walsemann, 2021). This investigation shows how important the relationship between air pollution and different levels of educational attainment is. Based on these studies and the trend of studying air pollution’s relationship with socioeconomic factors, I investigate the correlation between educational attainment and air pollution in India. In recent decades, air pollution has become one of the foremost public health concerns in India. Like most countries, the burden of air pollution does not fall equally over all of India. I theorize that air pollution and educational attainment should be correlated in a country like India, which struggles to limit the impact of air pollution on its citizenry. This investigation does not aim to discover causality or the direction of this relationship, only to understand if there is a correlation between air pollution and educational attainment in the country of India.

Literature Review

Many GIS investigators have focused on mapping the impact of air pollution on different aspects of human life. In this section, the review of human impacts is broken down into health impact studies and socioeconomic factor studies. Although this study is focusing on mapping the correlation between air pollution and a socioeconomic factor, it is important to review health impact studies because they constitute the majority of studies on air pollution in developing countries. The focus of this study is a developing country, therefore, it is important to review the unique facets of GIS investigations in developing countries, including how they source their data and what type of air pollution they focus on. It is also important to review studies investigating the correlation between air pollution and socioeconomic factors other than education. Although most of these studies are conducted in developed countries, they reveal useful information about sources of data and the geographic scale of most socioeconomic factor studies. Lastly, this section will review where this investigation sourced its data from.

Mapping Air Pollution and Health Impacts in Developing Countries

Most studies investigating the human impacts of air pollution in developing countries are focused on the public health threat that air pollution poses. One of the primary examples of this is a study performed by Nihit Goyal and his colleagues analyzing the impacts of increasing $PM_{2.5}$ exposure on births and child mortality in 43 low- and middle- income countries (2019). They pooled 69 nation-wide demographic and health surveys and compared it with high resolution satellite air pollution data using QGIS (Goyal et al., 2019). The results of their analysis showed that there is a strong association between exposure to carbonaceous $PM_{2.5}$ and higher neonatal mortality rates (Goyal et al., 2019). Other studies have taken a more in-depth look at specific health impacts of air pollution in a single developing country. One study investigated the link between $PM_{2.5}$ levels and lung cancer incidence (LCI) in China (Guo et al., 2023). They sourced $PM_{2.5}$ data from NASA's Earth Data collection and combined it with data from the China Cancer Registry Annual Report (Guo et al., 2023). Researchers then used spatial autocorrelation to assess the impact of $PM_{2.5}$ on LCI and found that areas with extreme climate conditions were more affected by $PM_{2.5}$ and struggled with high LCI rates the most (Guo et al., 2023). Another similar study in India used GIS to analyze how the geographic distribution of $PM_{2.5}$ concentrations correlate with the number of emergency room visits for acute respiratory symptoms in India (Kabra et al., 2022). Daily outdoor $PM_{2.5}$ concentrations and air quality index data were compared with the spatial distribution of patients with acute respiratory symptoms (Kabra et al., 2022). Using data from 18,063 patients, researchers found that acute respiratory ER visits were associated with higher $PM_{2.5}$ concentrations, especially during winter when $PM_{2.5}$ levels are at their highest (Kabra et al., 2022). These studies demonstrate major trends in investigations into the impacts of air pollution in developing countries. Most studies utilize national survey or census-like data to document human impacts and satellite derived data to measure $PM_{2.5}$ exposure levels. These studies are also mostly performed at the country level, with some studies taking an international perspective and analyzing an entire class of countries. To further understand these trends, these studies are compared to those investigating the correlation between air pollution and socioeconomic factors to see what data practices and methods and similar and which are different.

Map Air Pollution and Socioeconomic Factors

Few studies have evaluated the relationship between air pollution and education using GIS, but many GIS investigations have analyzed the relationship between air pollution and other socioeconomic factors. One foundational study in this area is Michael Jerrett and his colleagues' environmental justice analysis of particulate matter air pollution in Canada (2001). This study obtained socioeconomic data from the 1991 Census of Canada and used an early version of ArcGIS to process and compare particulate matter levels and socioeconomic factors like income, employment, and housing cost (Jerrett et al., 2001). Jerrett and his colleagues found that low income and unemployment were significant predictors for exposure to high $PM_{2.5}$ levels and dwelling values were negatively associated with $PM_{2.5}$ levels (Jerrett et al., 2001). Similar investigations have been conducted since then, including another in Canada that analyzed the relationship between air pollution, specifically NO_2 exposure levels, and social and material deprivation (Crouse et al., 2009). Social and material deprivation is a measurement of socioeconomic status that also includes low educational attainment (Crouse et al., 2009). This study also sourced its socioeconomic data from the Census of Canada and found that neighborhood tracts characterized by social and material deprivation consistently had higher levels of NO_2 exposure (Crouse et al., 2009). This type of investigation was extended to several other socioeconomic factors in a Danish study examining the relationship between air pollution and marital status, income, country of origin, and number of children in a household (Raaschour-Nielsen et al., 2022). They used GIS software to map the Danish Building Registry, Danish National Patient Registry, and Danish Civil Registration System and compare these databases with air pollution data (Raaschour-Nielsen et al., 2022). This study found that non-Danish origin is linked to higher levels of air pollution exposure, and, at the neighborhood level, high air pollution levels were found in areas with low socioeconomic status (Raaschour-Nielsen et al., 2022). These studies consistently show that air pollution falls unequally across populations worldwide and demonstrates why further investigation into the relationship between socioeconomic factors and air pollution in developing countries is necessary as well.

Most GIS investigations into the correlation between air pollution and socioeconomic factors thus far have been conducted in developed countries. One exception to this is a study done by Shareful Hassan and his colleagues which analyzed the relationship between $PM_{2.5}$ levels and a variety of demographic and population variables in Bangladesh (2022). Bangladesh was one of the first developing countries that socioeconomic factor correlation studies were extended to because it has the most hazardous level of air pollution in the world and numerous methods of study have been applied to the nation in hopes of alleviating the public health threat (Hassan et al., 2022). Hassan and his colleagues sourced spatial $PM_{2.5}$ data from the Socioeconomic Data Applications Center (SEDAC) within NASA (2022). They found that high levels of $PM_{2.5}$ are associated with several demographic and population variables including higher population density, higher poverty rates, and low-income groups in both urban and rural areas (Hassan et al., 2022). This study is an example of how investigations into the correlation between air pollution and socioeconomic factors need to continue to be conducted in developing countries. This collection of studies investigating the correlation between air pollution and socioeconomic factors also echoes some of the data practices and methods from the health impact studies reviewed above. Both sets of studies mostly use national survey or census data to provide health and demographic data. They also mostly use air pollution data obtained from satellite sources and almost entirely focus on $PM_{2.5}$ exposure levels as the primary type of air pollution worth investigating. These studies vary widely in scale with some focusing on individual cities or communities and others focusing on entire groups of countries. To reflect the results of this literature review most accurately, I selected the following data sources and scale for this study.

Data Sources

Based on the data utilized in the studies above, conducting this investigation requires three main areas of data. The first is socioeconomic data concerning educational attainment, which was obtained from the Census of India, mirroring several of the studies above that sourced their data from censuses and national surveys. The most recent Indian Census was conducted in 2011 (India conducts a census every 10 years, but they missed their 2021 census while dealing with the COVID-19 Pandemic) ("C-08," 2023). From the Indian Census data repository, I downloaded a spreadsheet of education data for each state and union territory in India. This totaled to 36 spreadsheets which contained educational statistics broken down by district, age group, and sex. I eliminated any irrelevant data and kept only the total educational statistics for each district. This process is discussed in more detail in the methods section.

The second area of data necessary for this investigation is air pollution data. Almost every study mentioned above focused on $PM_{2.5}$ exposure level data. Hassan and his colleagues cited the SEDAC within NASA as a consistent and reliable source of particulate matter data, and several other studies reviewed above used satellite data like that collected by the SEDAC (2022; Goyal et al., 2019; Guo et al., 2023). For this reason, I use a $PM_{2.5}$ raster dataset obtained from the SEDAC (Hammer et al., 2020). This study uses the global annual dataset from 2011 to logically compare with 2011 census data.

The final area of data required for this investigation is a shapefile of the administrative boundaries of India. India has several levels of administrative boundaries including states, districts, and subdistricts. The investigations reviewed above have a variety of scales. To best emulate them, I chose to conduct a country-wide analysis of India at the district level. Indian Census data is most accurate and obtainable at the state or district level, but the states in India are quite large and would provide few data points for correlation analysis. There are 640 districts in India and their size more accurately match the scale of the investigations reviewed above, especially those investigations conducted in developing countries (Guo et al., 2023; Kabra et al., 2022; Hassan et al., 2022). I obtained a shapefile of the 640 districts of India from an online project called "Community Created Maps of India" run by DataMeet which stores several South Asian administrative boundary datasets for free and easy online usage ("India," n.d.). I downloaded the "District Boundaries" dataset for India and saved the "Census_2011" shapefile for mapping purposes. This shapefile was the best dataset I found because it was derived from Indian Census and listed the districts of India in the same order as the educational attainment census data. This made the process of manually joining the educational attainment data with the "censuscode" field from the shapefile dataset easier and more reliable. This manual joining process is described in the methods section below.

Methods

The research methods of this investigation are broken down into several major steps and conducted across three software: Excel, ArcGIS, and Stata. The software being used is specified in the title of each major step. These steps begin after the relevant data has been downloaded from the sources described above.

I. Compiling and Formatting Indian Census Data in Excel
   a. Delete all irrelevant data: There is a spreadsheet for each state and union territory of India. Each spreadsheet contains educational attainment data for every district within each state and territory of India broken down by age group and sex. Specific data by age group and sex is not relevant to this investigation, therefore only total educational statistics for each district should be taken from each spreadsheet. The following fields for every district were reserved from each spreadsheet.
      i. Total Persons
      ii. Literate Persons
      iii. Below Primary Persons
      iv. Primary Persons
      v. Middle Persons
      vi. Matric/Secondary Persons

vii. Higher Secondary/Intermediate Persons

viii. Graduate & above Persons

b. Create formatted spreadsheet with relevant data fields

   i. Because there is a spreadsheet for every state or union territory in India, each spreadsheet contains a row with all necessary data points for each district in the state or union territory that the spreadsheet focuses on. These are the only rows that need to be copied into the new formatted spreadsheet.

   ii. In a new spreadsheet, create the following fields and copy in all relevant data from the census spreadsheets into the new formatted spreadsheet.

      1. State_NM – This is the name of the state that each district lies in.
      2. Dist_NM – This is the name of the district.
      3. Total_Pop – The total population of the district. This corresponds with the total persons field in the original spreadsheets.
      4. Total_Lit – The total literate population of the district. This corresponds with the literate persons field in the original spreadsheets.
      5. Tot_BPrim – The total number of people in the district at a below a primary level of education. This corresponds with the below primary persons field in the original spreadsheets.
      6. Tot_Prim – The total number of people in the district at a primary education level. This corresponds with the primary persons field in the original spreadsheets.
      7. Tot_Mid – The total number of people in the district at a middle school level of education. This corresponds with the middle persons field in the original spreadsheets.
      8. Tot_Sec – The total number of people in the district at a secondary education level. This corresponds with the secondary persons field in the original spreadsheets.
      9. Tot_HSec – The total number of people in the district at a high secondary or intermediate level of education. This corresponds with the higher secondary/intermediate persons field in the original spreadsheets.
      10. Tot_Grad – The total number of people in the district at a graduate or above education level. This corresponds with the graduate and above persons field in the original spreadsheets.
      11. censuscode – The field that is manually input to make this data joinable with the district shapefile. This will allow the this

spreadsheet of educational attainment data to be mapped on the districts shapefile of India and compared with air pollution values.

c. Upload and open the India districts shapefile in ArcGIS

d. Table to Excel: This tool is used to export an attribute table from ArcGIS as an excel spreadsheet so the data contained in the spreadsheet can be used outside of the ArcGIS application.

    i. Inputs: The input table is the "2011_Dist" attribute table that contains information about all the districts of India. The output table can be named "2011DistMatch".

    ii. Output: This creates an Excel spreadsheet of the exact attributes of the features in the India districts shapefile that the educational attainment data needs to be matched with.

e. Manually match the two spreadsheets: Using the "2011DistMatch" spreadsheet and the formatted educational attainment spreadsheet, manually input the "censuscode" field into the educational attainment spreadsheet so it matches the same district names as the "2011DistMatch" spreadsheet. Both spreadsheets contain a version of the districts name field which are similar enough to use for manually matching the "censuscode" field, but not similar enough to use as a join field themselves. This process sounds tedious, but it is relatively straightforward because within each state the districts are listed in the same order on both spreadsheets. This means users don't need to manually search for each district in input a single value for "censuscode". Instead, there is a set of sequential "censuscode" values for every state and union territory in India and it shouldn't take more than 30 minutes to manually match them.

II. Calculating Educational Attainment Statistics

a. In the original Indian Census Data, educational attainment levels were reported as the levels that individuals were currently at, not the levels that had been completed. For the purpose of comparing these educational statistics with air pollution data, it is important to reflect the total number of persons who have completed each level of education. This is necessary because other research comparing education levels and air pollution data discuss education levels in terms of the total number or percentage of people that have an education at or beyond a certain level. For example, in the cognitive impairment study discussed in the introduction, education levels are reported as 8 or more years of education or 13 or more years of education (Alishire & Walsemann, 2021).

b. To replicate this, I added the educational attainment totals from the census data up backwards so all individuals who have a graduate education are also counted as having a secondary level education, middle school education, and so forth.

The educational statistics listed below were calculated from the educational data already in the formatted spreadsheet using excel spreadsheet functions. The equation used in Excel to calculate each educational attainment statistic is shown below.

i. Literacy Rate (Lit_Rate)
    1. ((Total_Lit) / (Total_Pop)) * 100 = Literacy Rate

ii. Percent with Graduate Education (TPGrad)
    1. ((Tot_GradL) / (Total_Pop)) * 100 = Graduate Percent

iii. Percent with Secondary Education (TPSec)
    1. (((Tot_SecL) + (Tot_HSecL) + (Tot_GradL)) / (Total_Pop)) * 100 = Total Secondary Percent

iv. Percent with Primary Education (TPPrim)
    1. (((Tot_PrimL) + (Tot_MidL) + (Tot_SecL) + (Tot_HSecL) + (Tot_GradL)) / (Total_Pop)) * 100 = Total Primary Percent

c. Note that I did not calculate these statistics for all education levels because some of them are not as reliable groupings as others. For example, the below primary and higher secondary levels aggregate too many different potential educational levels together. Someone with a below primary education could have come very close to completing six years of school or could have only attended for one year ("C-08," 2023). Higher secondary institutions vary greatly in India and could involve a variety of different educational experiences ("C-08," 2023). I also did not calculate a statistic for the middle school level because it does not differ much from the primary school level for the purposes of this analysis.

III. Joining Education Data with Indian Districts

a. Upload and open the formatted educational attainment statistics spreadsheet to ArcGIS as a standalone table.

b. Select by Attribute: The Select by Attributes tool allows users to sort out data based on objects associated values in their attribute table. This tool is valuable at this stage because any districts in India that don't have associated educational attainment statistics or $PM_{2.5}$ values must be deleted from the data. This is true of two districts, the northernmost district in the state of Kashmir and Jammu and the state of Lakshadweep. Both Pakistan and India claim Kashmir and Jammu and the northern most portion is considered "Pakistani-administer" while the southern portion is "Indian-administered". For this reason, India does not collect census data in Pakistani-administered Kashmir and Jammu, meaning there are no educational statistics to compare $PM_{2.5}$ values to in this. The northern most district of Kashmir and Jammu is assigned a "censuscode" of "0" to reflect the fact that it is not a full district of India, and this field can be used to remove the district from the rest of the data. Lakshadweep, on the other hand, is a series of small islands on the western coast of India. The islands are small enough that SEDAC did not calculate the $PM_{2.5}$ values for them meaning there would be nothing to compare the educational statistics to in this area. The "censuscode" for Lakshadweep is "587".

    i. Inputs: The input table is "2011_Dist" and the selection statement "Where 'censuscode' 'is greater than' '1' 'And' 'censuscode' 'is not equal to' '587'" selects out all the districts that have enough data to be relevant to this investigation.
    ii. Output: This creates a selection of all the districts in India other than the northern most district of Kashmir and Jammu and the district of Lakshadweep.

c. Export Features: This tool allows users to create a new layer of only the features selected in the previous step.
    i. Inputs: The input features are the "2011_Dist" layer and the output feature class can be named "IndiaDistricts2011".
    ii. Output: This creates a new feature class, identical to the original imported layer, except that it does not include Kashmir and Jammu or Lakshadweep.

d. Join Field: This tool allows users to combine tables using a common field within both tables. This tool is valuable at this stage for combining the vector layer of the Indian districts and the table of educational attainment data. This is possible because the "censuscode" field from the vector layer of the Indian Districts was included and matched with each district in the educational attainment data.
    i. Inputs: The input table is the previous "IndiaDistricts2011" layer and the input join field in this table is the "censuscode" field. The join table is the India Census education data table named "Form_ICED" and the join table field is the "censuscode" field as well.
    ii. Output: This pulls all the educational attainment data into the feature class that includes the vector layer. This is vital for being able to map and understand the spatial relationship between educational attainment statistics and their relationship with $PM_{2.5}$ values.

IV. Extracting Raster Data
    a. Upload the SEDAC annual $PM_{2.5}$ values raster dataset for the year 2011 to ArcGIS
    b. Extract by Mask: This tool allows users to cut out the shape of one layer in another raster layer. This tool is valuable at this point to eliminate all irrelevant $PM_{2.5}$ values for the rest of the world and preserve just the $PM_{2.5}$ values for the relevant districts of India.

   i. Inputs: The input raster is the original annual $PM_{2.5}$ values raster dataset file that was downloaded from SEDAC. The input raster or feature mask data is the "IndiaDistricts2011" feature class, and the output raster can be named "IndiaPM25". The extraction area should be set to "inside".
   ii. Output: This creates a new raster layer of $PM_{2.5}$ values that has the same extent as the "IndiaDistricts2011" feature class.

c. Polygon to Raster: This tool can turn any polygon or vector dataset into a raster dataset for different types of analysis. This tool is valuable at this point for turning the vector layer of the Indian Districts into a raster dataset in order to use the Zonal Statistics Table tool, which functions best with two input raster datasets instead of one raster dataset and one vector dataset.
   i. Inputs: The input feature is the "IndiaDistricts2011" feature class, and the value field should be a field that is unique for each district which can be "censuscode". The output raster dataset can be named "IndiaDistricts2011_Ras" and the cell assignment type is cell center. The priority field can be "None" because there are few if any overlapping polygons in this dataset. The cell size should be set to the "IndiaPM25" dataset so the cell sizes of the raster being created matches that of the "IndiaPM25" dataset. The build raster attribute table option can also be switched off because this layer is just being created for the function of the Zonal Statistics Table tool.
   ii. Output: This creates a new raster layer named "IndiaDistricts2011_Ras". The raster layer should have values from 1 to 639 defining the 639 districts in India relevant to this investigation if the Polygon to Raster tool was implemented correctly.

d. Zonal Statistics Table: This tool allows users to calculate a variety of zonal statistics using a zonal raster that defines the extent of the zones in which statistics will be calculated and a value raster that contains the values on which statistics will be calculated. This tool is valuable because it can calculate the mean $PM_{2.5}$ values for the individual districts of India for direct comparison to educational attainment statistics.
   i. Inputs: The input raster or feature zone data is the newly created "IndiaDistricts2011_Ras" and the zone field is "Value" because there is only one value in this raster dataset which is equivalent to "censuscode". The input value raster is the "IndiaPM25" raster dataset that contains the air pollution data. The "Ignore No Data in Calculations" option can remain checked. The only relevant statistics type to this investigation is mean and the "Calculate Circular Statistics" and "Process as Multidimensional"

boxes can remain unchecked. Leave the output join layer option blank, inputting a name here will cause the function to fail.

ii. Output: This creates a new standalone table named "MeanPM25Dist" that contains the mean $PM_{2.5}$ values for each district in India. The value field in this table is the same field from the "IndiaDistricts2011Ras" meaning it represents "censuscode".

e. Join Field: This tool's function is described above, but it is useful here to connect the mean $PM_{2.5}$ values for each district with the educational statistics for each district and the vector layer of the districts in order to create maps of the relationships between these variables.

   i. Inputs: The input table is the "IndiaDistricts2011" feature class, and the input field is "censuscode". The join table is "MeanPM25Dist" and the join field is "VALUE".

   ii. Output: This pulls the mean $PM_{2.5}$ values into the "IndiaDistricts2011" attribute table so they can be compared to the educational statistics and associated with their respective district polygons for mapping and statistical purposes.

V. Evaluating Correlation of Statistics with Stata

a. Table to Excel: The function of this tool is described above, but here it is used to export the "IndiaDistricts2011" attribute table with all relevant educational attainment statistics and air pollution data attached.

   i. Inputs: The input table is the "IndiaDistricts2011" table, and the output excel file can be named "IndiaDistricts2011Final"

   ii. Output: The output is an excel spreadsheet version of the attribute table of the "IndiaDistricts2011" feature class.

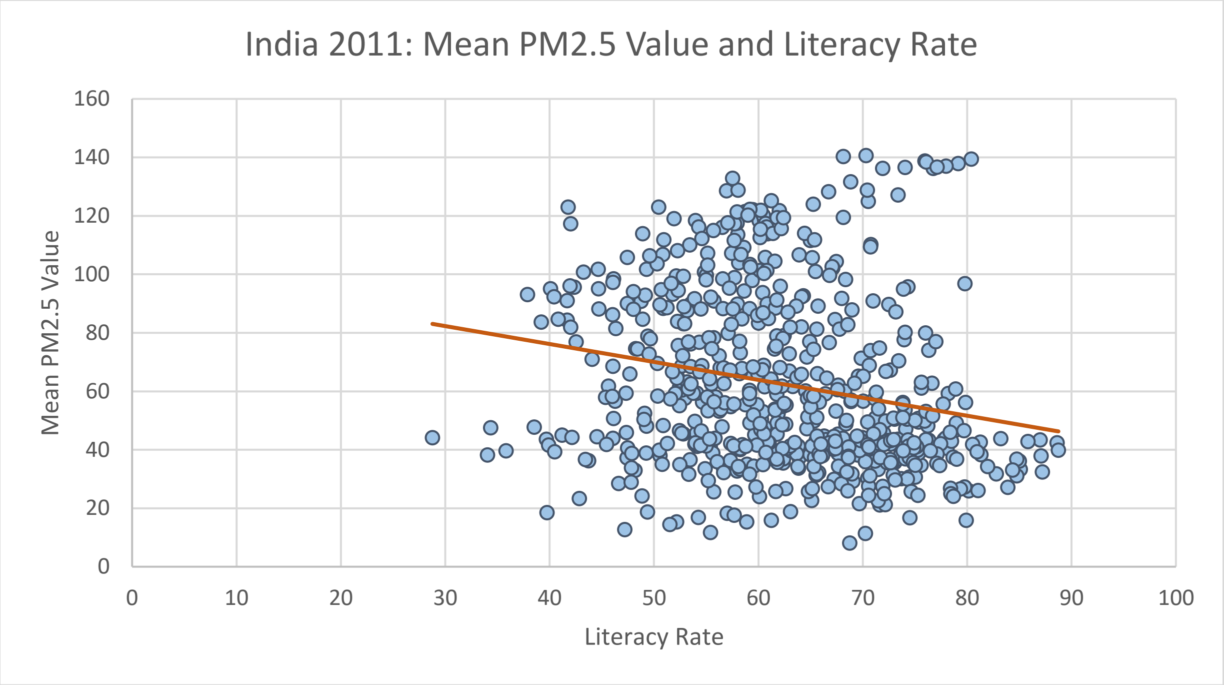
b. Open excel spreadsheet in Stata

   i. You can open an excel spreadsheet in Stata by hovering over the file tab, then import, and then excel spreadsheet. Open the "IndiaDistricts2011Final" spreadsheet and check "import first row as variable names".

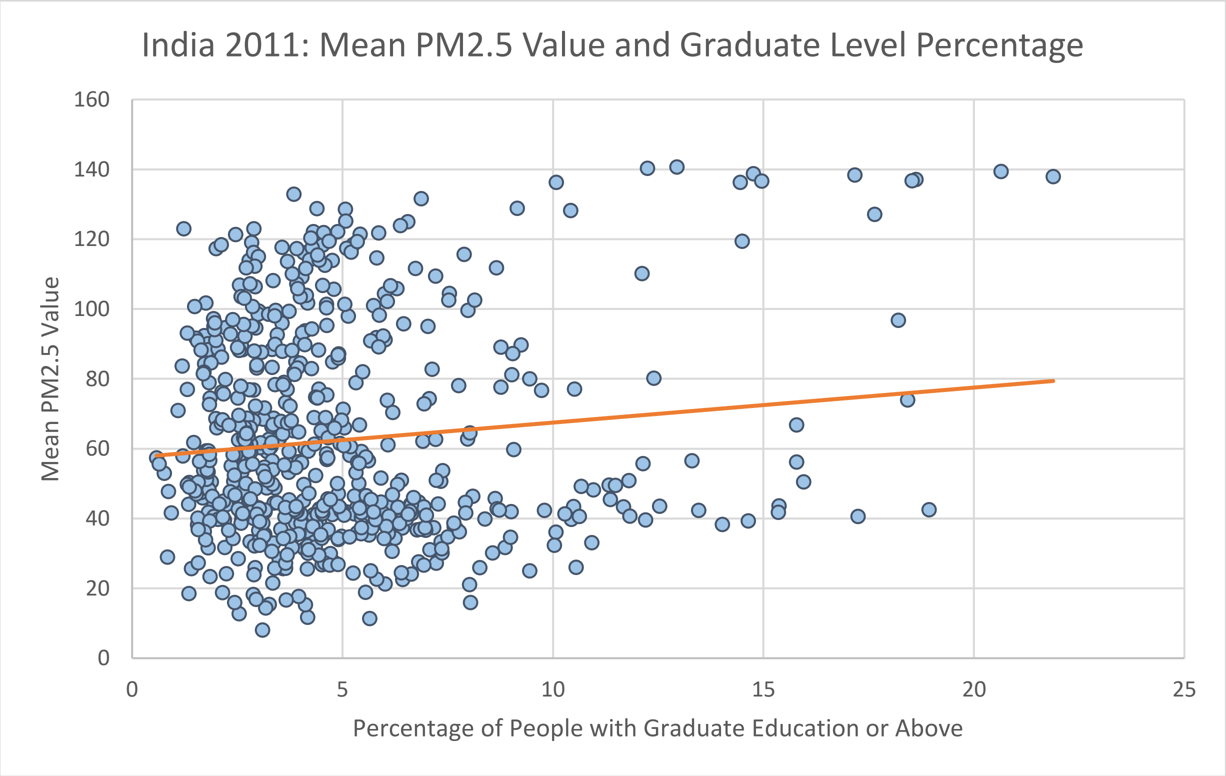
   ii. Once the data is opened in Stata, the following commands can be input. Capture the results of each command as a screenshot for further analysis.

      1. pwcorr MEAN Lit_Rate, sig
      2. pwcorr MEAN TPGrad, sig
      3. pwcorr MEAN TPSec, sig
      4. pwcorr MEAN TPPrim, sig

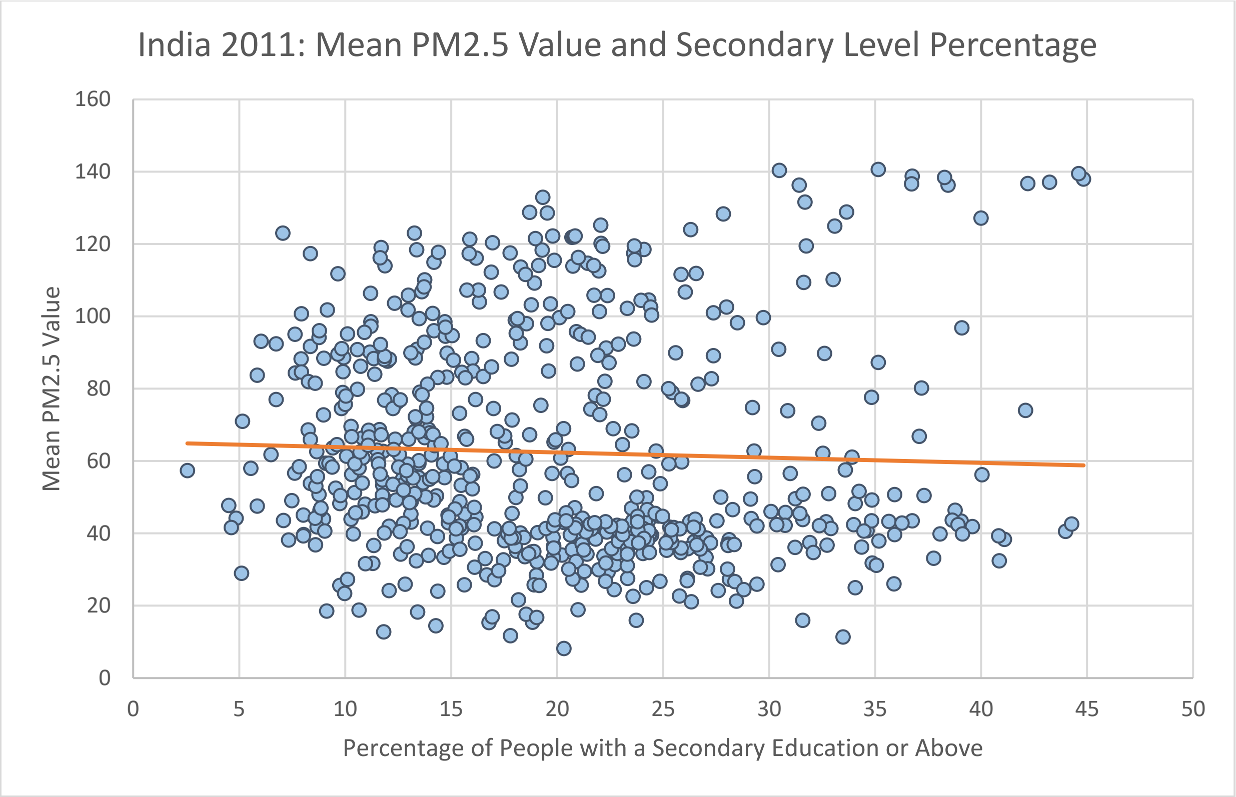
Results



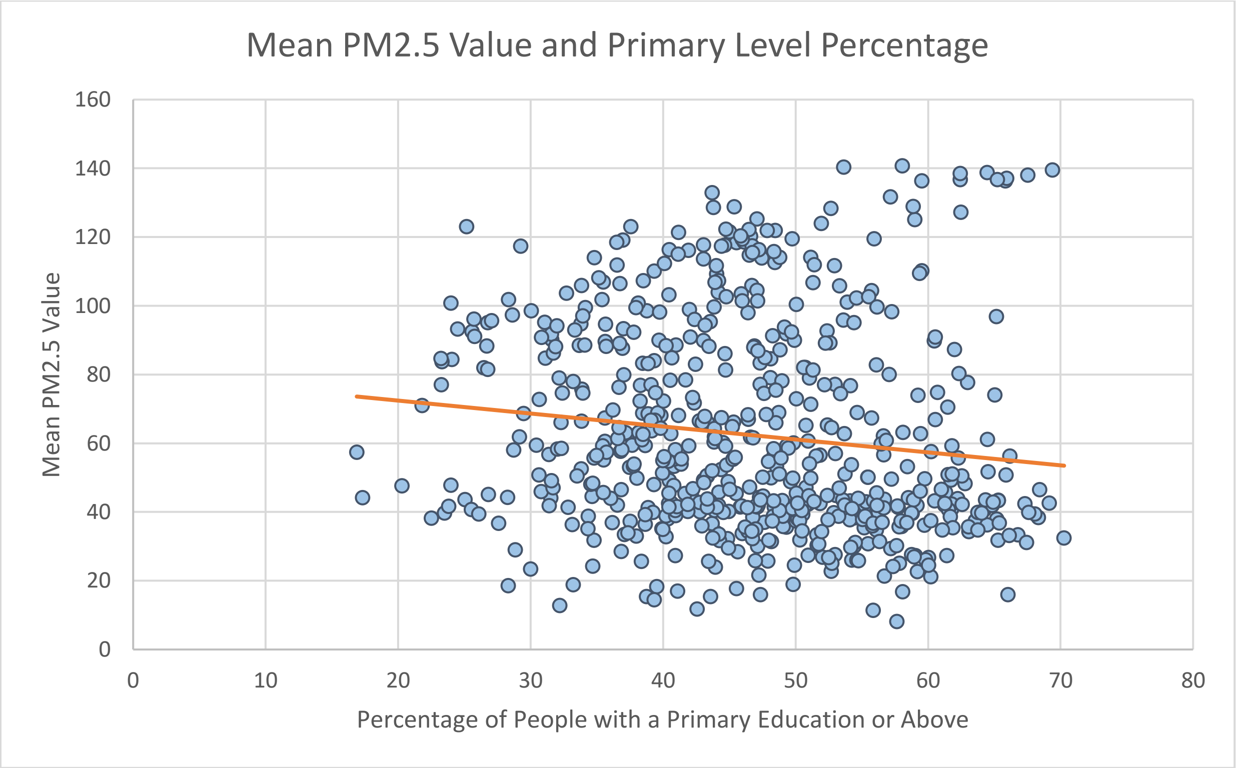
r = -0.2151   p-value = 0.000



r = 0.1084   p-value = 0.0061



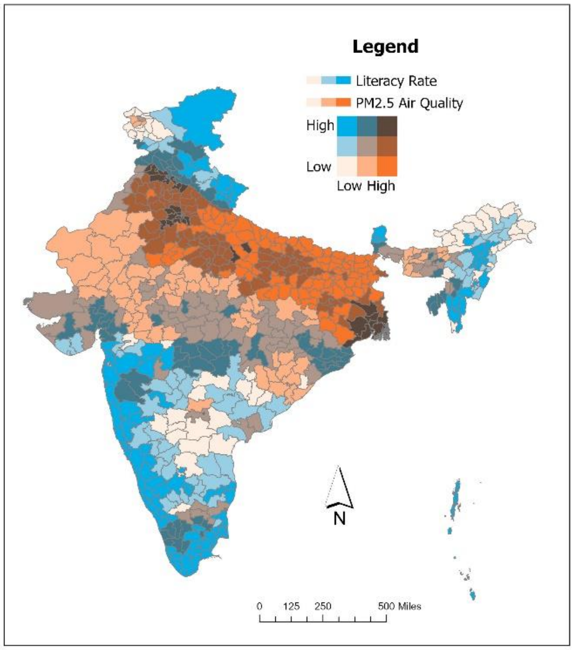
r = -0.0405   p-value = 0.3071



r = -0.1323   p-value = 0.0008

India 2011

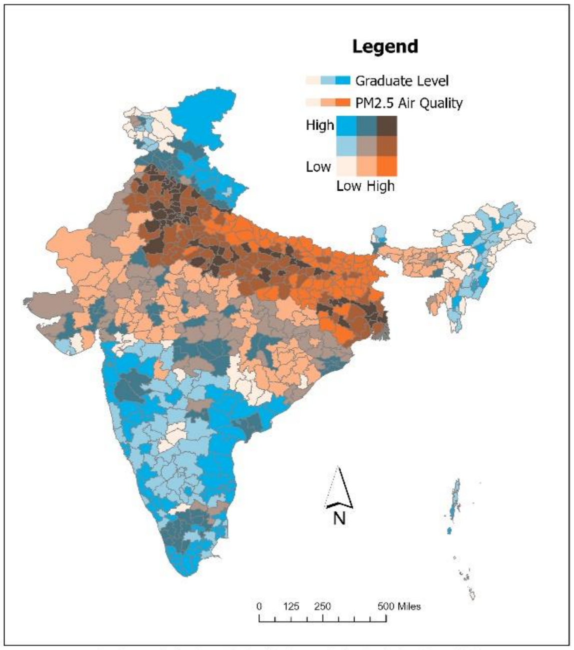
Literacy Rate and PM2.5 Air Pollution



Data Sources: Indian Census 2011 and Socioeconomic Data Applications Center NASA

India 2011

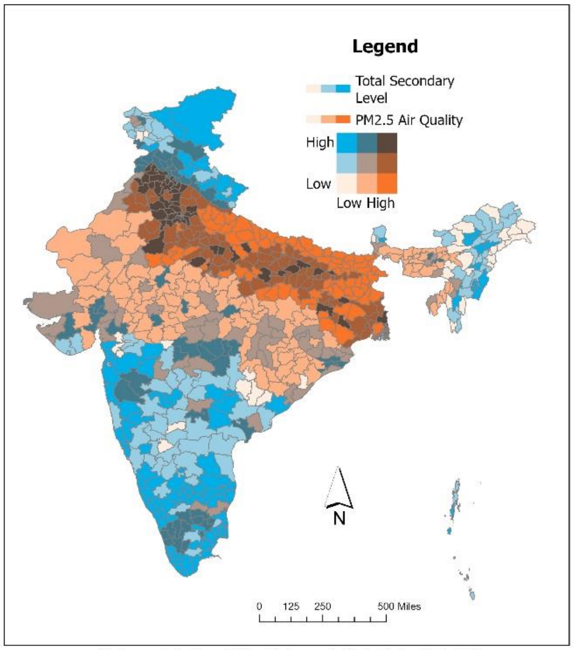
Graduate Level and PM2.5 Air Pollution



Data Sources: Indian Census 2011 and Socioeconomic Data Applications Center NASA

India 2011

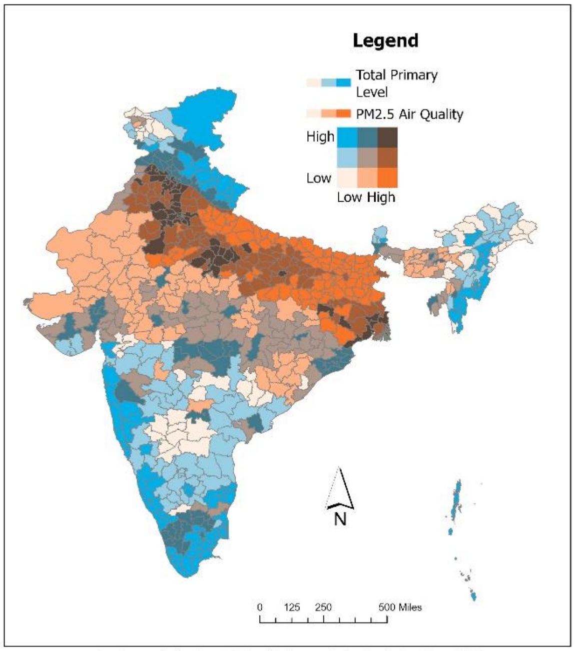
Total Secondary Level and PM2.5 Air Pollution



Data Sources: Indian Census 2011 and Socioeconomic Data Applications Center NASA

India 2011

Total Primary Level and PM2.5 Air Pollution



Data Sources: Indian Census 2011 and Socioeconomic Data Applications Center NASA

The results of this investigation are formatted as both scatterplots and bivariate maps so readers can understand the statistical and spatial relationships of air pollution and educational attainment in India. These results reveal preliminary insights about the correlation between air pollution and educational attainment. The strongest relationship is between $PM_{2.5}$ values and literacy rates. The correlation of these variables has a regression value of -0.2151, meaning a one standard deviation increase in $PM_{2.5}$ values is associated with a 0.2151 standard deviation decrease in literacy rates, on average. This relationship also has a p-value of 0.0000, meaning that the probability that air pollution and the literacy rate have no linear relationship is 0%. This means that the existence of a negative relationship between air $PM_{2.5}$ values and the literacy rate is statistically significant.

The second strongest relationship is between $PM_{2.5}$ values and the percentage of people with a primary level education or above. The correlation of these variables has a regression value of -0.1323, meaning a one standard deviation increase in $PM_{2.5}$ values is associated with a 0.1323 standard deviation decrease in the percentage of people with a primary level education or above, on average. The relationship also has a p-value of 0.0008, meaning there is a 0.08% chance that air pollution and primary education do not have a linear relationship. A p-value of 0.05 or lower is required to claim, with 95% confidence, that a linear relationship does exist. The p-value is lower than 0.05 at 0.0008, which means that, with 95% confidence, this study can claim that there is a statistically significant, negative, linear relationship between $PM_{2.5}$ values and the percentage of people with a primary level education or above.

The third strongest relationship is between $PM_{2.5}$ values and the percentage of people with a graduate education or above. The correlation of these variables has a regression value of 0.1084, meaning a one standard deviation increase in $PM_{2.5}$ values is associated with a 0.1084 standard deviation increase in the percentage of people with a graduate education or above, on average. This relationship has a p-value of 0.0061, which is less than 0.05. This study can claim, with 95% confidence, that there is a statistically significant, positive, linear relationship between $PM_{2.5}$ values and the percentage of people with a graduate education or above.

Finally, the weakest relationship is between $PM_{2.5}$ values and the percentage of people with a secondary education or above. The correlation of these two variables has a regression value of -0.0405 and a p-value of 0.3071. 0.3071 is not less than 0.05 which means that this study cannot draw statistically significant conclusions about the relationship between $PM_{2.5}$ values and the percentage of people with a secondary education or above.

Discussion and Conclusion

This analysis yielded many statistically significant relationships between air pollution and educational attainment, but the substantive significance of these relationships is more debatable. Although the strongest correlation in this research is between $PM_{2.5}$ values and literacy rates, the correlation is weak at only -0.2151. This value is closer to 0 than it is to -1. Therefore it is difficult to claim that increasing literacy rates is associated with a meaningful decrease in air pollution. This relationship is weak, but it is consistent with past findings in socioeconomic factor studies. Almost every study reviewed above found that higher $PM_{2.5}$ values are associated with lower socioeconomic conditions like lower income, lower dwelling values, higher unemployment, and higher neonatal mortality rates (Hassan et al., 2022; Crouse et al., 2009, Jerrett et al., 2001; Goyal et al., 2019). The negative relationship between $PM_{2.5}$ values and literacy rates echoes past findings, lending the results of this study more credibility. This negative relationship is also congruent with studies originally linking air pollution and education in the United States (Mullen et al., 2020). These studies find that higher $PM_{2.5}$ values are related to lower educational attainment, specifically lower standardized test scores, which corresponds with lower literacy rates as well.

On the other hand, the other two statistically significant correlations between $PM_{2.5}$ values and graduate and primary education levels are not completely logical and do not line up with prior research. $PM_{2.5}$ values and primary education levels have a negative relationship while $PM_{2.5}$ values and graduate education levels have a positive relationship. The opposing positive and negative relationships and the weakness of the correlation between these variables calls for further research into the relationship between different educational attainment levels and air pollution to pinpoint what type of relationship exists.

Regardless of these results, the government of India needs to continue to relentlessly pursue increasing educational attainment, especially literacy rates. India, in fact, ratified the International Covenant on Economic, Social, and Cultural Rights in 1979 which states that education is not only a human right but that all ratifying countries are obligated to devote the maximum of their existing resources to realize the rights enumerated in the Covenant ("International Covenant," 2024). Looking at the scatterplots above, readers can see that there are still over 250 districts in India where less than 60% of the population is literate. This alone should be a call to action, and although this research does not provide significant evidence that increased literacy rates will meaningfully decrease $PM_{2.5}$ values, it does find a weak relationship that is worth further investigation in other countries that struggle to improve their air quality.

The results of this research encourage a further understanding of the relationship between air pollution and educational attainment in other developing countries. For example, further research of this relationship in countries like Bangladesh, Pakistan, Chad, Iraq, Tajikistan, and Burkina Faso could reveal a clearer relationship between these variables and help address the dire state of air quality in these countries. This study is part of what is soon to be a long line of investigations into the relationship between air pollution and socioeconomic factors in developing countries. This research, and the continuation of it, is vital to address the ongoing public health crisis of air pollution that threatens the health and wellbeing of so many people.

References

Alishire, J., & Walsemann, K. M. (2021). Education differences in the adverse impact of $PM_{2.5}$ on incident cognitive impairment among U.S. older adults. Journal of Alzheimer's Disease, 79(2), 615-625. https://doi.org/10.3233/JAD-200765

C-08: Education level by age and sex. (2023, March 3). Census of India. Retrieved May 3, 2024, from https://censusindia.gov.in/nada/index.php/catalog/44790

Crouse, D. L., Ross, N. A., & Goldberg, M. S. (2009). Double burden of deprivation and high concentrations of ambient air pollution at the neighborhood scale in Montreal, Canada. Social Science & Medicine, 69(6), 971-981. https://doi.org/10.1016/j.socscimed.2009.07.010

Goyal, N., Karra, M., & Canning, D. (2019). Early-life exposure to ambient fine particulate air pollution and infant mortality: Pooled evidence from 43 low- and middle-income countries. International Journal of Epidemiology, 48(4), 1125-1141. https://doi.org/10.1093/ije/dyz090

Guo, B., Gao, Q., Pei, L., Guo, T., Wang, Y., Wu, H., Zhang, W., & Chen, M. (2023). Exploring the association of $PM_{2.5}$ with lung cancer incidence under different climate zones and socioeconomic conditions from 2006 to 2016 in China. Environmental Science and Pollution Research, 30(60), 126165-126177. https://doi.org/10.1007/s11356-023-31138-8

Hammer, M. S., van Donkelaar, A., Li, C., Lyapustin, A., Sayer, A. M., Hsu, N. C., Levy, R. C., Garay, M. J., Kalashnikova, O. V., Kahn, R. A., Brauer, M., Apte, J. S., Henze, D. K., Zhang, L., Zhang, Q., Ford, B., Pierce, J. R., & Martin, R. V. (2020). Global estimates and long-term trends of fine particulate matter concentrations (1998–2018). Environmental Science & Technology, 54(13), 7879-7890. https://doi.org/10.1021/acs.est.0c01764

Hassan, S., Islam, T., & Bhuiyan, M. A. H. (2022). Effects of economic and environmental factors on particulate matter ($PM_{2.5}$) in the middle parts of Bangladesh. Water Air and Soil Pollution, 233(8). https://doi.org/10.1007/s11270-022-05819-y

India - district boundaries. (n.d.). DataMeet: Community Created Maps of India. Retrieved April 28, 2024, from https://projects.datameet.org/maps/districts/

International Covenant on Economic, Social and Cultural Rights. (2024, May 6). United Nations Treaty Collection. Retrieved May 6, 2024, from https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-3&chapter=4&clang=_en#EndDec

Jerrett, M., Burnett, R. T., Kanaroglou, P., Eyles, J., Finkelstein, N., Giovis, C., & Brook, J. R. (2001). A GIS - environmental justice analysis of particulate air pollution in Hamilton, Canada. Environment and Planning, 33(6), 955-973. https://doi.org/10.1068/a33137

Kabra, S., Yadav, R., Nagori, A., Mukherjee, A., Singh, V., Lodha, R., Yadav, G., Saini, J., Singhal, K., Jat, K., Madan, K., George, M., Mani, K., Mrigpuri, P., Kumar, R., Guleria, R., Pandey, R., Sarin, R., & Dhaliwal, R. (2022). Geographic information system-based mapping of air pollution & Emergency room visits of patients for acute respiratory symptoms in Delhi, India (March 2018-February 2019). Indian Journal of Medical Research, 156(4), 648. https://doi.org/10.4103/ijmr.ijmr_136_21

Mullen, C., Grineski, S. E., Collins, T. W., & Mendoza, D. L. (2020). Effects of $PM_{2.5}$ on third grade students' proficiency in math and English language arts. International Journal of Environmental Research and Public Health, 17(18), 6931. https://doi.org/10.3390/ijerph17186931

Raaschour-Nielsen, O., Taj, T., Poulsen, A. H., Hvidtfeldt, U. A., Ketzel, M., Christensen, J. H., Brandt, J., Frohn, L. M., Geels, C., Valencia, V. H., & Sorensen, M. (2022). Air pollution at the residence of Danish adults, by socio-demographic characteristics, morbidity, and address level characteristics. Environmental Research, 208. https://doi.org/10.1016/j.envres.2022.112714