Measuring the Exposure to Air Pollution for Pedestrians in Boston Metropolitan Area and its Relation to Walkability

Patrali Ghosh and Vishwesh Srinivasan

11 May 2023

Abstract

Measuring air pollution has been a rising topic of interest among researchers as climate change is gaining popularity on a global scale. The exposure to air pollution changes based on one's mode of transportation. The method of transport significantly impacts a person's travel time and thus their exposure to air pollution. This paper focuses on pedestrians' exposure to air pollution in the Greater Boston Region. We found places in the center of Boston with high concentrations of NOx. Based on a literature review, we hypothesized that the places with high walkability index are the places where the exposure to air pollution is also high. In this paper, we performed multiple spatial regression analyses to find if there is any correlation between walkability and the variables used to calculate it, and the concentration of NOx at a census block group level in the Greater Boston Region. We found evidence of a positive correlation between these two sets of variables.

Introduction

Air pollution is an environmental issue globally, leading to long-term adverse effects like climate change due to increased CO2 and other greenhouse gases. Air pollution has also been linked to various hostile health conditions humans face. People are exposed to air pollutants through occupation, residence, or daily travel patterns. In urban areas, transportation is one of the primary sources of air pollution to which people are exposed.

Measuring air pollution exposure based on the mode of transportation is critical to help people make data-driven decisions related to their travel. This air pollution measurement can also help policymakers design urban systems, especially public transportation systems, to minimize exposure to air pollution while traveling.

Walking is an integral part of commuting as one will walk to the nearest transit stop or bus stop to avail public transport. Calculating the air pollution exposure while walking to the nearest public transport stop or job location can allow policies to be created in order to safeguard citizens. An Internal Medicine paper written by Jeroen de Bont et al. 2022, there found a substantial increase in the effects of ambient air pollution on cardiovascular diseases. Higher long-term exposure to air pollution was strongly evident in connections to all-cause cardiovascular mortality and morbidity, stroke, blood pressure, and ischemic heart diseases. Short-term exposures to $PM_{2.5}$, $PM_{10,}$ and Nitrogen Oxides (NO_x) were consistently associated with an increased risk of myocardial infarction and both fatal and non-fatal strokes. Long-term exposure to $PM_{2.5}$ is associated with an increased risk of incident myocardial infarction, hypertension, incident stroke,

and stroke mortality. Having an insight into how much exposure to air pollution people have on a daily basis while commuting may allow better public policies.

The National Walkability Index is a measure or tool used to assess the pedestrian-friendliness or walkability of a given area, such as a city or neighborhood, in the United States. It is based on measures of the built environment that affect the probability of whether people walk as a mode of transportation: street intersection density, proximity to transit stops, and diversity of land uses. Although numerous factors influence walking, these measures were chosen for the National Walkability Index because they can be measured using variables in the Smart Location Database (SLD), which has nationwide data availability and consistency at the block group level. The selected variables from the SLD are:

(i) Intersection density: Higher intersection density is correlated with more walk trips.

(ii) Proximity to transit stops: Distance from the population center to the nearest transit stop in meters. Shorter distances correlate with more walking trips.

(iii) Employment mix: The mix of employment types in a block group (such as retail, office, or industrial). Higher values correlate with more walk trips.

(iv) Employment and household mix: The mix of employment types and occupied housing. A block group with a diverse set of employment types (such as office, retail, and service) plus many occupied housing units will have a relatively high value. Higher values correlate with more walk trips.

The Community Line Source Modeling System (C-LINE) is a computer software tool developed by the U.S. Environmental Protection Agency (EPA) for modeling air emissions from various types of

line sources, such as roadways, railways, and other linear emission sources. It is specifically designed for use in assessing air quality impacts from mobile sources, and it is commonly used for estimating emissions from transportation-related activities.

C-Line uses mathematical algorithms and emission factors to estimate pollutant emissions from line sources based on input data such as vehicle activity data, vehicle characteristics, and emission factors for different pollutants. It can provide estimates of emissions for a variety of pollutants, including criteria pollutants such as particulate matter (PM), nitrogen oxides (NO_x), volatile organic compounds (VOCs), and others.

A paper by Marshall et al. looks into walkability and air pollution. Their findings are that neighborhoods with high walkability tend to have high levels of traffic-related pollution, such as NOx, but a lower level of ozone pollution. High walkability and high traffic pollution tend to be in and around lower-income neighborhoods.

Research Questions

This work aims to answer the following research questions:

(1) What is the exposure to air pollution, specifically NOx, for pedestrians in the Greater Boston Region?
(2) What is the walkability for the areas in the Greater Boston Region?
(3) Is there any correlation between the NOx concentrations and the walkability for the areas in Greater Boston Region?
(4) Does this correlation determined vary spatially?

Data

We used the following three datasets:

(i) Census 2010 Census Block Groups (CBGs) boundary shapefile provided by MassGIS[1].

(ii) National Walkability Index dataset, part of Smart Location Mapping provided by the United States Environmental Protection Agency[2].

(iii) NOx Concentration in ppb downloaded for Greater Boston Area (10m x 10m grids) from the C-LINE tool, jointly provided by Chapel Hill's Institute for the Environment and the United States Environmental Protection Agency[3].

We downloaded the census block groups boundary shapefile since the basic unit of analysis in our project is the census block group. The National Walkability Index dataset provides the walkability score for each census block group in the United States, along with the values of the variables using which this score is calculated.

The C-LINE dataset was downloaded in 10m x 10m grids and had many overlaps. We took all the shapefiles having the NOx concentration (ppb) at the census block group level and combined them. We removed the duplicate CBGs and combined partial CBGs. Then, we merged this dataset with the National Walkability Index dataset and the CBGs boundary shapefile on the GOEID10 column, which is a unique identifier used to identify a census block group. The geometry value from the CBGs boundary file was retained for visualization purposes.

[1] MassGIS Data: 2010 U.S. Census | Mass.gov
[2] Smart Location Mapping | US EPA
[3] Community-LINE Source Model (C-LINE) to estimate roadway emissions | US EPA, This dataset was downloaded and provided by Hongkun Huang (Hongkun.Huang@tufts.edu)

Methods

For visualization purposes, we had the following layers for the National Walkability Index:

(i) The first layer is the base map of the Greater Boston Region obtained from the Contextily package of Python.

(ii) The second layer is the walkability index derived from the National Walkability Index dataset at a census block group level. This value ranges from 1 – 20.

For visualization purposes, we had the following layers for the NOx concentrations:

(i) The first layer is the base map of the Greater Boston Region obtained from the Contextily package of Python.

(ii) The second layer is the NOx concentrations obtained from the C-LINE dataset. This dataset gives the concentrations of NO_x at a census block group level after data preprocessing.

We plot all the visuals using the geopandas plot function, part of the geopandas package, and the matplotlib package of Python. According to the user guide provided by EPA for the National Walkability Index, the scores have been divided into four categories which are 'Least Walkable,' 'Below Average Walkable,' 'Above Average Walkable,' and 'Most Walkable.' Each category has a color indicated to it. They are orange, yellow, light green, and dark green respectively. In order to maintain uniformity, the walkability index column was split into four categories with scores 0-5.75 being least walkable, 5.75 - 10.25 being below average walkable, and 10.25 - 15.25 being above average walkability.

Similarly, in order to maintain uniformity, we split the pollution levels into four categories 'Very Low,' 'Low,' 'High,' 'Very High,' with ranges from 0 – 18.75, 18.75 - 37.5, 37.5 - 56.25, 56.25 - 75 respectively colored light blue, salmon, red, and brown. Both the walkability and pollution exposure maps have been plotted for the Greater Boston Region and for Boston City.

For statistical analysis, we examine the spatial distribution of NOx concentrations across different locations in the Greater Boston Region and the walkability scores to identify if there is any correlation between the walkability and air quality measured at the census block group level.

First, we focus on two variables – the national walkability index as the independent variable and NOx concentration (ppb) as the dependent variable. Both variables are operationalized as continuous values. We use the latitude and longitude of the centroid of the census block groups to compute the spatial weight matrix that connects every observation to its k nearest neighbors. We fit the following regression models:

(i) OLS regression with weights matrix

(ii) Spatially lagged exogenous regressors with the national walkability index as the spatially lagged variable

(iii) Spatially lagged endogenous regressors

Then, we modify our inputs and use the following variables as input variables:

(i) Intersection Density (D3B)

(ii) Proximity to transit stops (D4A)

(iii) Employment Mix (D2B_E8MIXA)

(iv) Employment and household mix (D2A_EPHHM)

The dependent variable and the spatial weights matrix remain unchanged in the analysis. All the variables are operationalized as continuous values. Again, we fit the following regression models:

(i) OLS regression with weights matrix

(ii) Spatially lagged exogenous regressors with the D4A as the spatially lagged variable

(iii) Spatially lagged endogenous regressors

Results

We obtain the following maps representing the national walkability index and NOx concentration for the Greater Boston Region:

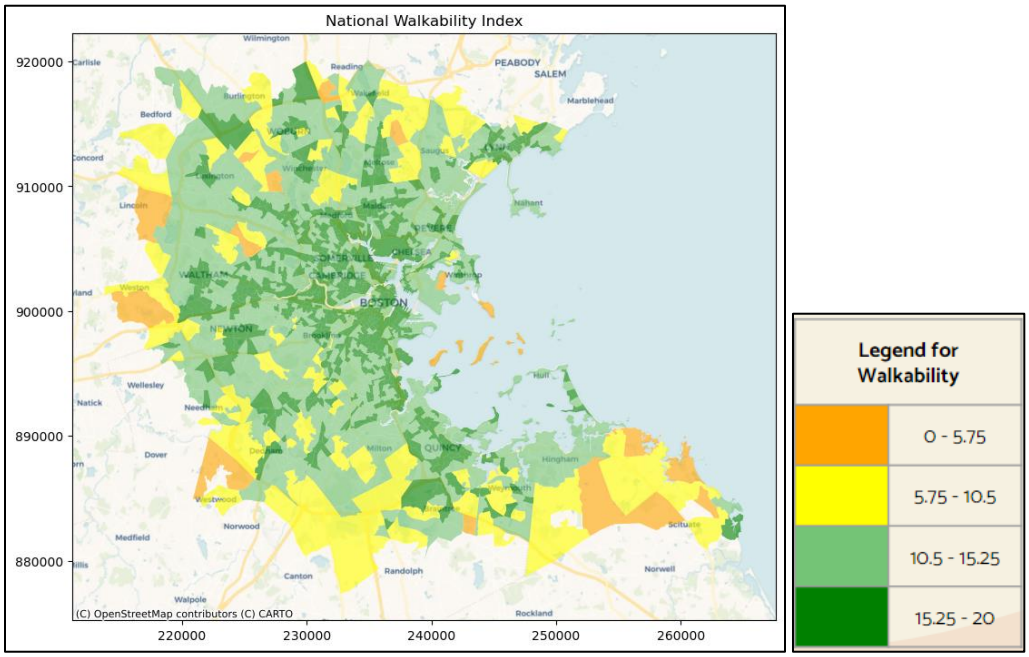

Figure 1: National Walkability Index for Greater Boston Region

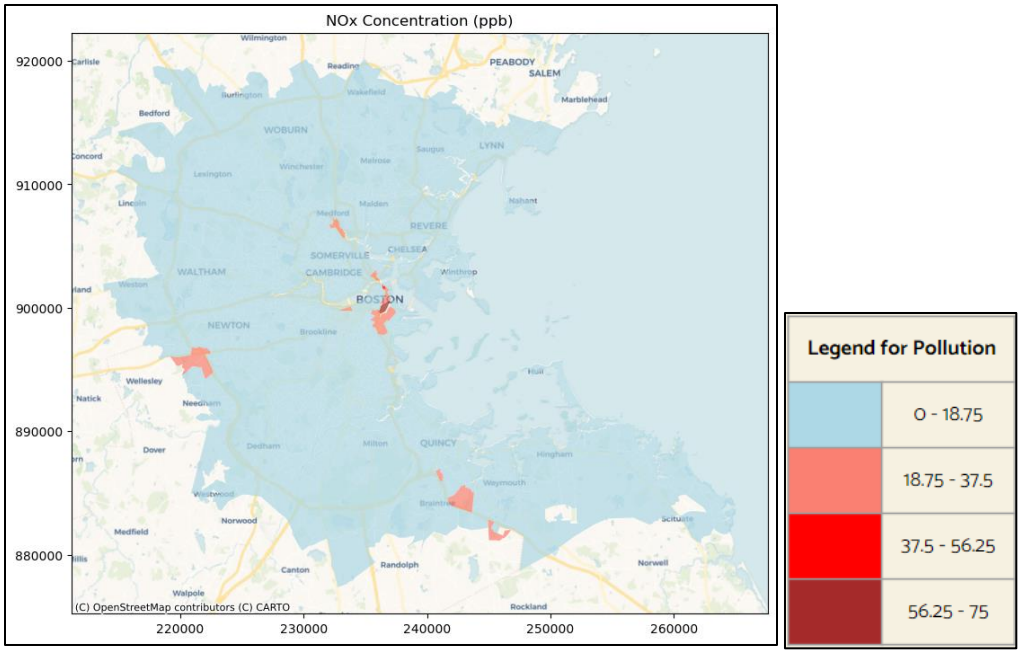

Figure 2: NOx Concentration (ppb) for Greater Boston Region

We obtain the following maps representing the national walkability index and NOx concentration for Boston City:

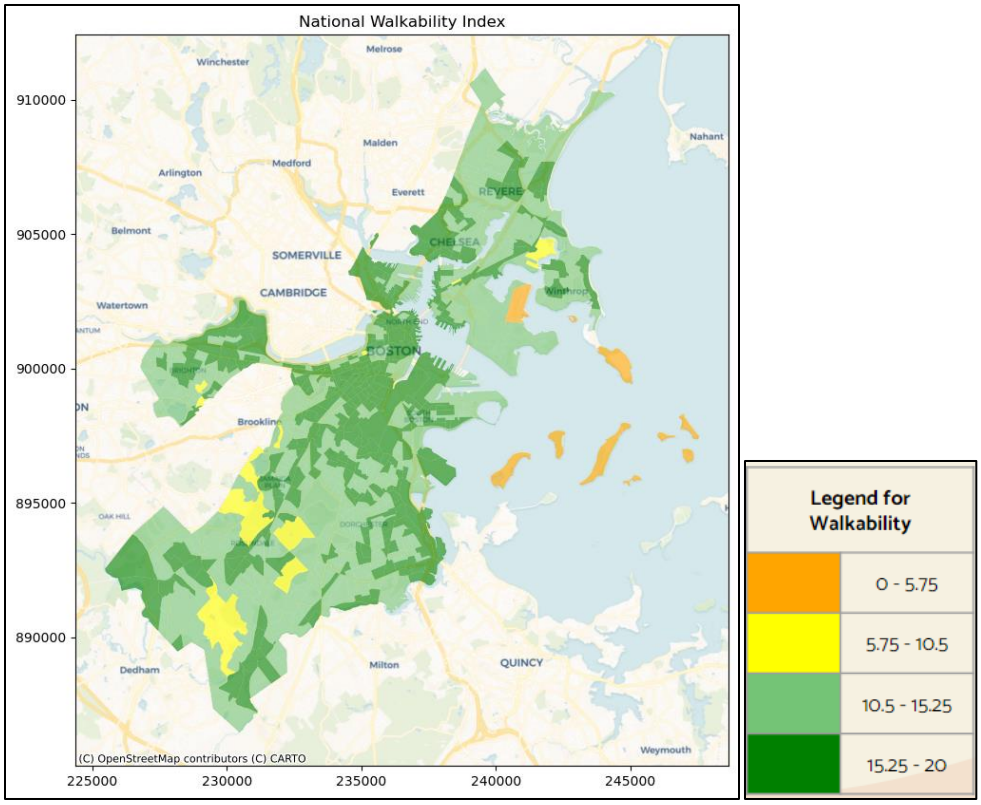

Figure 3: National Walkability Index for Boston City

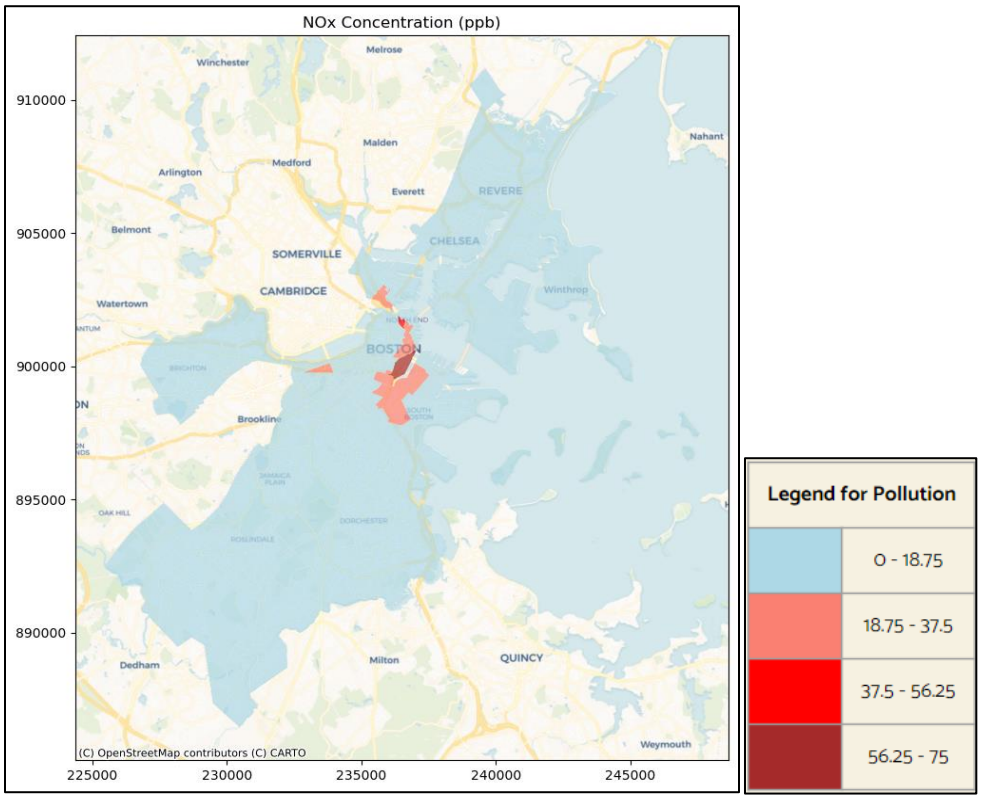

Figure 4: NOx Concentration (ppb) for Boston City

With the maps focusing on the Greater Boston Region, we can see that high levels of pollution are mostly focused in the Boston Metro Area, specifically around Downtown Boston, which is also a high walkability area. The suburbs in the Greater Boston Area show low levels of pollution but are mostly walkable. The suburbs on the edge of the region show low levels of pollution and low levels of walkability, which mostly follows the general hypothesis that high walkability and high pollution levels are related positively.

The maps focusing on Boston City, we can see a majority of the city has low levels of pollution while being highly walkable. This is not following the general hypothesis. This may be because of various reasons, primarily being green cover. Boston has a lot of parks and green areas, like the Boston Public Garden, the Boston Common, the Commonwealth Ave Mall, the Charles River Esplanade, the Arnold Arboretum, and so on.

Below is the summary statistics for all the variables of interest mentioned in the previous section.

	TOTAL_CONC	D3B	D4A	D2B_E8MIXA	D2A_EPHHM	NatWalkInd
count	1677	1677	1677	1677	1677	1677
mean	2.68	173.93	-5042.83	0.55	0.48	14.42
std	4.19	116.59	22620.71	0.19	0.22	2.75
min	0.03	0.00	-99999.00	0.00	0.00	1.00
25%	0.78	100.61	174.35	0.43	0.30	13.00
50%	1.38	154.76	273.59	0.58	0.48	14.67
75%	2.73	220.35	419.77	0.70	0.66	16.33
max	73.66	1355.44	1190.91	0.89	1.00	19.83

Table 1: Summary Statistics for the following variables: NOx Concentration (TOTAL_CONC), Intersection Density (D3B), Proximity to transit stops (D4A), Employment Mix (D2B_E8MIXA), employment and household mix (D2A_EPHHM) and National Walkability Index (NatWalkInd).

We observed that some of the entries in 'D4A' had -99999 values if the distance from the population-weighted centroids to transit stops is more than three-quarter miles. These rows are dropped since they are invalid entries before performing statistical analysis.

For our initial analysis with National Walkability Index as the independent variable and NOx concentration as the dependent variable, we obtained the below results summarized in tables along with inferences derived from them.

OLS regression with weights matrix – OLS + w

Variable	Coefficient	Probability
National Walkability Index	**0.3006929**	**0**

Table 2: Variable name, coefficient, and the p-value for the independent variables in the OLS + w regression.

We find the national walkability index variable statistically significant since the p-value is less than 0.05. It has a positive coefficient of 0.3, indicating that when the value of the national walkability index is increased by one unit and keeping everything else constant, the NOx concentration will increase by 0.3 units. The spatial dependence tests like Lagrange Multiplier (lag) and others have p-value less than 0.05 indicating the spatial dependence between the variables.

Spatially lagged exogenous regressors with the national walkability index as the spatially lagged variable – spreg (WX)

Variable	Coefficient	Probability
National Walkability Index	**0.2222103**	**0.0000155**
National Walkability Index lagged	**0.2987857**	**0.0000206**

Table 3: Variable name, coefficient, and the p-value for the independent variables in the spatially lagged exogenous regressors model with national walkability as the lagged variable.

The national walkability index is still significant. In addition, the spatially lagged variable is also significant. Both have positive coefficients.

Spatially lagged endogenous regressors – spreg (WY)

Variable	Coefficient	Probability
National Walkability Index	**0.1098082**	**0.0317483**
Total Concentration lagged	**0.8469829**	**0.0000000**

Table 4: Variable name, coefficient, and the p-value for the independent variables in the spatially lagged endogenous regressors model with total concentration as the lagged variable.

The national walkability is still significant, and also the NOx concentration, when used as an independent variable, is also significant. We can infer from this that there is evidence of pollution at a particular place being affected by the pollution in nearby areas.

For further analysis, when we considered all four variables – D3B, D4A, D2B_E8MIXA, and D2A_EPHHM as independent variables and NOx concentration as the dependent variable, we obtained the following results summarized in tables along with inferences derived from them.

OLS regression with weights matrix – OLS + w

Variable	Coefficient	Probability
D3B	**0.0040663**	**0.0000106**
D4A	-0.0008531	0.0693542
D2B_E8MIXA	-0.3519191	0.5365889
D2A_EPHHM	**4.5951272**	**0.0000000**

Table 5: Variable name, coefficient, and the p-value for the independent variables in the OLS + w regression.

We find the intersection density and employment and household mix as statistically significant since the p-value is less than 0.05. Both have positive coefficients. This points to the positive correlation between the places with more walk trips and high concentrations of NOx. The spatial

dependence tests like Lagrange Multiplier (lag) and others have p-value less than 0.05 indicating the spatial dependence between the variables.

Spatially lagged exogenous regressors with the D4A as the spatially lagged variable - spreg (WX)

Variable	Coefficient	Probability
D3B	**0.0037797**	**0.0000499**
D4A	-0.0003539	0.5009030
D4A lagged	**-0.0014395**	**0.0359316**
D2B_E8MIXA	-0.2443625	0.6687602
D2A_EPHHM	**4.6417723**	**0.0000000**

Table 6: Variable name, coefficient, and the p-value for the independent variables in the spatially lagged exogenous regressors model with D4A as the spatially lagged variable.

The intersection density and employment and household mix are still significant and have positive coefficients. In addition, the spatially lagged variable is also significant. We observe a negative coefficient here, which is reasonable. We infer that when proximity to transit stops is less, people will tend to use more public transportation, and hence this reduces emissions from vehicles leading to a reduction in the NOx concentrations.

Spatially lagged endogenous regressors – spreg (WY)

Variable	Coefficient	Probability
D3B	-0.0006573	0.4557038
D4A	-0.0006146	0.0975157
D2B_E8MIXA	0.2118613	0.6398337
D2A_EPHHM	**2.0334248**	**0.0000222**
Total Concentration lagged	**0.8698729**	**0.0000000**

Table 7: Variable name, coefficient, and the p-value for the independent variables in the spatially lagged endogenous regressors model with total concentration as the spatially lagged variable.

Now, the employment and household mix variable is significant, and also the NOx concentration, when used as the independent variable, is also significant. We can infer from this that there is evidence of pollution at a particular place being affected by the pollution in nearby areas.

Below is the table summarizing all the regression models with their corresponding R-squared values. We find reasonable R-squared values for both the two settings in the case of spatially lagged endogenous regressors. This makes us infer that the exposure to pollution in an area is largely dependent on the air pollution levels in the nearby areas.

Dependent Variable	Independent Variable	Type	R-squared
NOx Concentration	National Walkability Index	OLS + w	0.0241
NOx Concentration	National Walkability Index	spreg (WX)	0.0352
NOx Concentration	National Walkability Index	spreg (WY)	**0.4105**
NOx Concentration	D3B, D4A, D2B_E8MIXA and D2A_EPHHM	OLS + w	0.0714
NOx Concentration	D3B, D4A, D2B_E8MIXA and D2A_EPHHM	spreg (WX)	0.074
NOx Concentration	D3B, D4A, D2B_E8MIXA and D2A_EPHHM	spreg (WY)	**0.4217**

Table 8: All the different types of regression models with the R-squared values.

Policy Implications

Since we see a positive correlation between walkability and NOx concentrations, policies implemented to encourage people to use public transportation, such as buses and trains, instead of driving personal vehicles in high walkability areas can help reduce pollution levels in those areas. Additionally, policies developed to increase green spaces, such as parks and urban forests, in areas with high employment and household mix can again help reduce air pollution exposure. The presence of green spaces will absorb pollutants and improve air quality, creating more livable environments for residents. Moreover, promoting transportation options, like bicycles, through policies is crucial. This approach will effectively reduce the emissions of pollutants and encourage sustainable transportation methods that are healthier for both people and the environment. By combining these initiatives, policymakers can create vibrant, sustainable communities that prioritize public transportation, green spaces, and clean transportation options for the benefit of all.

References

Agency, U. S. (2021). National Walkability Index Methodology and User Guide. EPA.

Jeroen de Bont, S. J. (2022). Ambient air pollution and cardiovascular diseases: An umbrella review of systematic reviews and meta-analyses. Wiley Journal of Internal Medicine.

Julian D. Marshall, M. B. (2009). Healthy Neighborhoods: Walkability and Air Pollution. Environmental Health Perspectives.

Timothy M. Barzyk, V. I. (2015). A near-road modeling system for community-scale assessments of traffic-related air pollution in the United States. Elsevier.

Appendix

Team members' task allocation:

(i) **Study of Literature and Formulation of Problem Statement** – Patrali and Vishwesh

(ii) **Data Collection and Preparation** – Patrali and Vishwesh

(iii) **Visualizations** – Patrali

(iv) **Statistical Analysis** – Vishwesh

(v) **Presentation and Report** – Patrali and Vishwesh

Code and Data path in R drive folder:

All our datasets and code are present in the **group_air_pollution** folder. We have three data sources, and below are the listed folders for each dataset:

(i) **MA_census_block_groups** folder has the Census 2010 Census Block Groups (CBGs) boundary shapefile provided by MassGIS.

(ii) **WalkabilityIndex** folder has the National Walkability Index dataset, part of Smart Location Mapping provided by the United States Environmental Protection Agency.

(iii) **Air quality files** folder has the NOx Concentration in ppb downloaded for Greater Boston Area (10m x 10m grids) from the C-LINE tool, jointly provided by Chapel Hill's Institute for the Environment and the United States Environmental Protection Agency.

Code.ipynb is the jupyter notebook that reads the data from the above three folders at different stages and merges them for visualization and statistical analysis.