CRP3850 Urban Data Science
Prof. Zhang
Final Project Report
Edward Guo, Tim Wang, Yucheng Zhang

Research Question: What are the current patterns of environmental inequalities in New York City, and do trees and green spaces help address such environmental inequality?

Introduction and Literature Review

Environmental inequality in the U.S. has a long history closely linked to physical community segregation and unequal distribution of pollution sources and environmental amenities. In terms of physical segregation, the creation of redlining maps by the Home Owners' Loan Corporation and Federal Housing Administration in the 1930s, for example, acted as a powerful tool to skew resource distribution based on a community's racial composition (Hoffman, Shandas, and Pendleton 2020). Although redlining was outlawed in the 1970s, it continues to have profound impacts on urban environments. Research has shown that redlined neighborhoods are on average 5 degrees Fahrenheit warmer in the summer compared to non-redlined neighborhoods (Hoffman et al. 2020). The historic lack of public investment resulted in a lower number of trees and urban parks in redlined neighborhoods, exposing them to greater risks of extreme heat. In terms of the unequal distribution of pollution and environmental amenities, research has consistently found that low-income and minority households face disproportionate environmental burdens (Boyce, Zwickl, and Ash 2016). For example, areas with higher-than-average minority or low-income populations are exposed to significantly higher levels of PM 2.5, reflecting the concentration of pollution sources like vehicle traffic and polluting industries in these areas (Jbaily et al. 2022).

Given that environmental inequality is a widespread issue in U.S. cities, we want to investigate potential solutions that may help mitigate environmental hazards like high urban temperature and air pollution. After reviewing relevant literature, we discovered that urban greeneries may be effective in reducing environmental inequality. Urban greenery has been shown to be effective in improving public health. In Greensboro, NC, greater availability of tree canopy coverage and open spaces are correlated with increased physical activities and improved community health (Bruton & Floyd, 2014). Greater greenness in schoolyards in Chicago is associated with lower stress levels, better concentration, increased motivation, and higher predicted math performance among students (Kuo et al., 2018). Urban greenery can also bring economic benefits. In New York City, researchers matched Green View Indexes (GVI) developed using Google Street View with commercial property prices and found that properties with a higher level of surrounding greenery have significantly higher rents and transaction prices (Yang et al., 2021). Proximity to amenities like parks is also correlated to increased financial value (Yang et al., 2021). These health and market trends indicate that access to nature improves the quality of life by making a neighborhood more desirable to live, work, and visit, making urban greeneries promising in reducing environmental inequality.

Knowing that environmental inequalities exist and that urban greenery may be able to reduce such inequality, we decided to use New York City as a case study and investigate the two in detail. Thus, in the first part of this research, we examine the patterns of urban inequalities through the lens of environmental amenities, indicated by urban greenery distribution, and environmental burdens, as reflected by land surface temperature and PM 2.5 concentration. In the second part of this research, we explore the potential for cities to use urban greenery as a tool to reduce environmental inequality by analyzing the impact urban greenery has on pollution and temperature. Overall, our results suggest that environmental inequality persists in New York City and that urban greenery is effective at reducing land surface temperature, but its effect on pollution is still ambiguous.

Data / Methods

Pollution Data

We obtained pollution data for nitrogen dioxide (NO_2), fine particulate matter (PM2.5), nitric oxide (NO), and ozone (O_3) from the NYC Open Data portal (New York City Department of Health and Mental Hygiene 2022). The data, in raster format with a spatial resolution of 300 meters, shows the annual average predicted value for each of the pollution types based on a model derived from the New York City Community Air Survey (New York City Department of Health and Mental Hygiene n.d.). We derived the average pollution level for each census block group using the Zonal Statistics tool in ArcGIS. Unfortunately, due to the low spatial resolution of the pollution data, we were only able to obtain averages for 4,065 block groups out of the 6,287 land census block groups in the City. However, most missing data were for the same block groups and these block groups tend to be in the more densely populated areas, meaning there remain enough data points for all types of areas across the City.

Tree Density Data

We derived the tree density data based on the 2015 New York City Street Tree Census, which is accessed via the NYC Open Data portal (NYC Department of Parks and Recreation 2020). The spreadsheet contained the longitude and latitude for 683,788 trees, dead and alive, along the City's streets. We removed the dead trees as they do not contribute to reducing surface temperature or pollution, leaving us with 584,036 trees. The living trees were then mapped in

ArcGIS using the XY to Points function, and we utilized the Summarize Within feature to count the number of trees within each block group. The tree density was derived by dividing the number of trees in each block group by its size. This dataset did not include trees inside New York City's numerous parks, and therefore all parks are shown to have very low tree density.

Green View Index (GVI)

The Green View Index (GVI) is calculated from the Google Street View (GSV) images using Semantic Image Segmentation. GVI is a representation of the residents' eye-level perception of urban greenery. Traditional methods of greenery measurement like tree density data and remote sensing images have limitations compared to GVI. First, traditional methods only measure greenery at a two-dimensional level while GVI includes three-dimensional features like walls covered with vines, lawns, and bushes (Ki and Lee 2021). Secondly, by using the GSV Images, GVI takes into account the urban greeneries' accessibility. Traditional methods based on an overhead view can differ from pedestrians' actual perceived greenness and accessible public greenery by including inaccessible areas like mountains and enclosed private gardens.

We used the road network and boundary shapefile of the City provided by NYC Open Data as the input to determine which points to sample GVI from. Using the Python libraries Fiona and Shapely, we created 140,000 points along the street network with a sampling distance of 200 meters. To make the input points more accurately reflect ordinary residents' perceived greenness, and to meet our constraints in computational power and budget, we filtered out the points that are less accessible: points on highways or in tunnels were removed, for example.

In order to fetch and download GSV images through the GSV API, we need to specify the resolution, location (in latitude and longitude), field of view, heading, and pitch for each of our sampled points. Overall, we chose to obtain six static street view images to create a panorama for future calculations for each sampled point. Each image is at a resolution of 400 by 400 pixels, a field of view of 60 degrees, headings ranging from 0 to 360 with an interval of 60 degrees, and a pitch of 0 to represent eye-level perception. We also took into account the seasonal difference in perceived greenness, so we only collected images from May to September. Xiaojiang Li has developed a batch processing method using python, and using the parameters mentioned above as the input, we retrieved 496,158 images from 82,693 sampling points (Li et al. 2015). Due to GSV availability, the number of points we were able to retrieve images from is lower than the number of points we planned to sample.

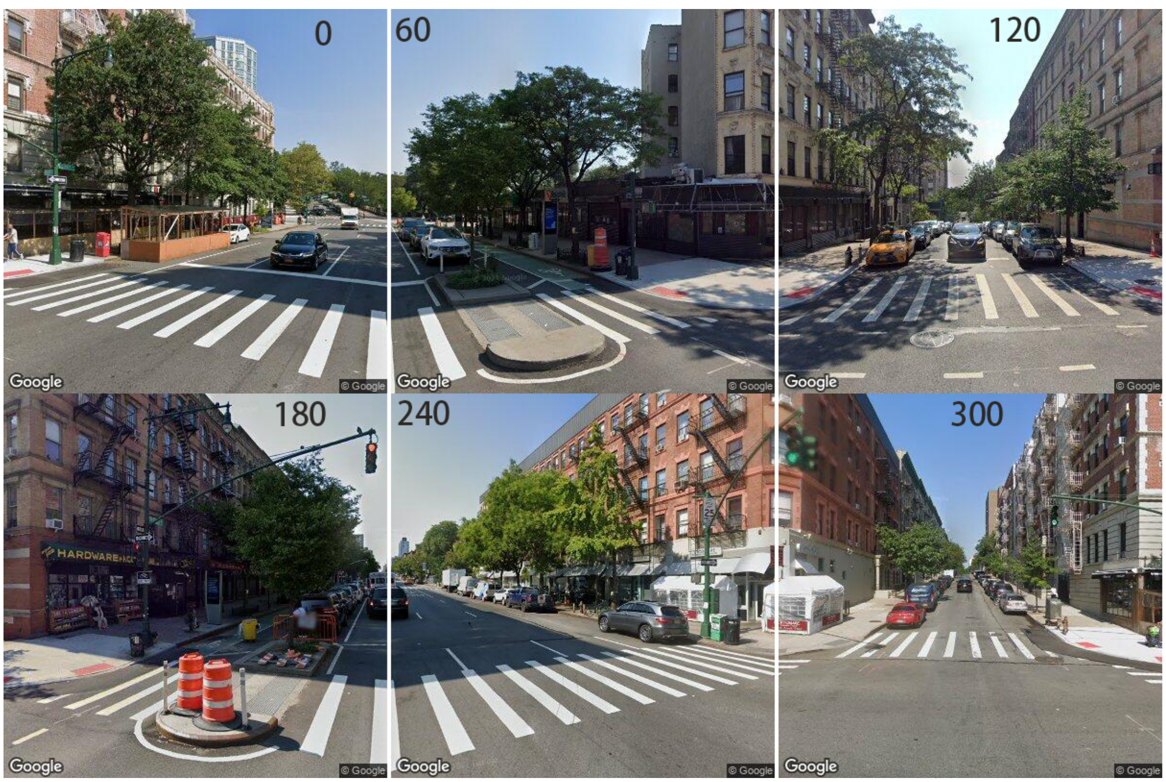

Figure 1: Sample GSV images fetched. The numbers on the image stand for the heading of the image. The six images combined provide a panorama of a given position.

After collecting the GSV images, we calculated the GVI from the percentage of vegetation within a given image. We used PyTorch and the MIT ADE20K scene parsing dataset to recognize vegetation pixels within the GSV images. The Semantic Segmentation Algorithm categorizes every pixel in the scene into 150 types of objects (Zhou et al. 2016, 2017). We summed up the number of pixels recognized as “tree”, “grass”, and “palm” and divided them by 1600 to get the percentage of vegetation, which is the GVI for the given point.

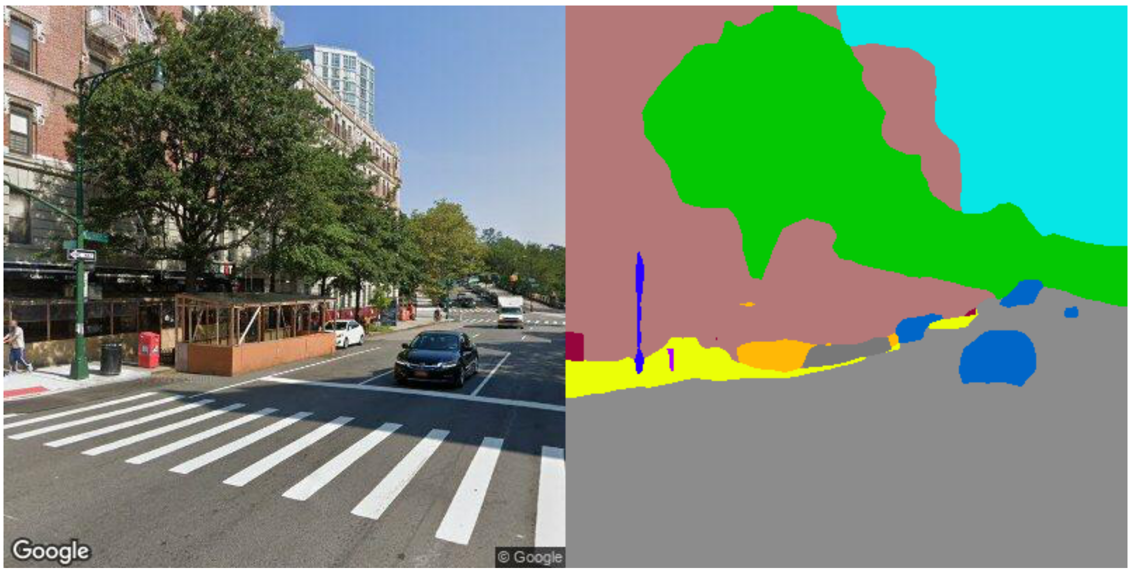

Figure 2.1: Predicted vegetation result created from the MIT ADE20K scene parsing dataset.

Figure 2.2: Pixels categorized as “trees”. Notice that the model successfully distinguished the green street light from the street trees.

The 82,693 calculated GVI points were imported into ArcGIS. Using the Summarize Within function, we calculated the average GVI for the given block group using an average of all the GVI points within the block group.

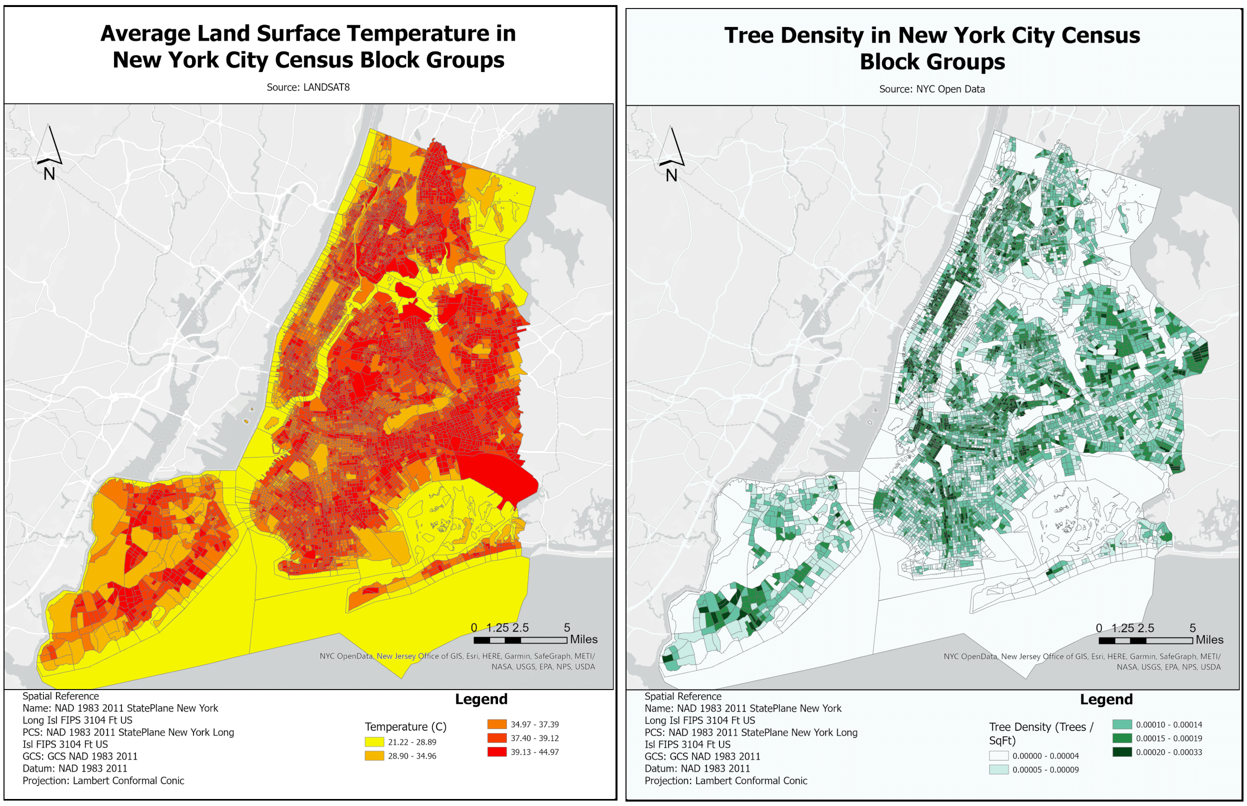

Figure 3: Visualization of Average Land Surface Temperature and Tree Density Data in New York City by Block Groups

Land Surface Temperature

We obtained land surface temperature data via Google Earth Engine. First, we filtered Landsat 8 Collection 2 Level 2 remote sensing infrared data imagery between June and August for each year from 2015 to 2021. We excluded the 2020 data due to visible data errors in the satellite image. We then applied scaling factors to the Band 10 infrared data to obtain the land-surface temperature in Kelvin, then converted the numbers to Celsius. The raster image of the average summer land surface temperature at a spatial resolution of 30 meters was exported to

ArcGIS, in which we used the Zonal Statistic tool and calculated the mean temperature for all 6,287 block groups.

Demographics Data

Demographics data was obtained from the 2019 American Community Survey 5-year Estimates using Social Explorer. We downloaded race, education, employment, household income, housing value, and rent burden status for every block group in the City.

Data Processing and Regression Analysis Methods

All data processed and generated using ArcGIS (pollution, tree density, GVI, and land surface temperature) were joined together in one table using the FIPS code associated with each block group. The table was exported as an Excel file, which was later combined with the demographics data by FIPS code into one Pandas DataFrame via Python.

We first attempted to conduct regression analysis using linear regression, but spatial autocorrelation issues in the dataset produced confusing and often contradictory results. For example, the OLS Regression model from Table 1 shows that land surface temperature is positively correlated with median household income while negatively correlated with median housing value. We expected the two independent variables to have the same sign since households with a higher income often choose to spend more on housing. Realizing that we need to correct for spatial autocorrelation, we decided to conduct our regression analysis using spatial error and spatial lag models via GeoDa. The decision of whether to use a spatial error or a spatial lag model is based on the results of the Robust Lagrange Multiplier test on the OLS regression model for both the spatial error and spatial lag. We chose either the spatial error or spatial lag model depending on which test has the smallest p-value for the given model specification.

Variable	Coefficient	Standard Error	p-value
Mean Green View Index	-0.0584	0.004	0.000
% White Population	-0.0057	0.001	0.000
% Bachelor or Higher Degree	-0.0183	0.002	0.000
Unemployment Rate	0.0005	0.005	0.922
Median Household Income	0.1938	0.076	0.011
Median Housing Value	-0.1779	0.049	0.000
% Residents with Severe Rent Burden	0.0038	0.001	0.011
Population Density	-12.3417	1.603	0.000
Constant	40.3929	0.982	0.000

$R^2 = 0.186$

Table 1: OLS Regression Model, using the surface temperature as the dependent variable

Results

```
R-squared           :    0.560292  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-6821.241283
Sigma-square        :      1.2309  Akaike info criterion :     13658.5
S.E of regression   :     1.10946  Schwarz criterion     :     13709.4

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT       40.7439      0.832501       48.9415      0.00000
       * Pct_White    -0.0053013    0.00120106      -4.41383      0.00001
      * Pct_Bachel    -0.0110929    0.00164941      -6.72537      0.00000
        Pct_Unempl    0.00327352    0.00366256      0.893778      0.37144
        Pct_Rent_5    0.00043168     0.0010424      0.414121      0.67879
         Ln_Income   3.24927e-06     0.0611413   5.31437e-05      0.99996
      * Ln_Housing      -0.13431      0.045677      -2.94044      0.00328
      * Populati_1      -4.59469       1.50009      -3.06294      0.00219
            LAMBDA      0.721552     0.0114143       63.2148      0.00000
-----------------------------------------------------------------------------
```

Table 2: Results of the spatial error regression exploring the distribution of land surface temperature[1] [2]

Most socio-economic indicators correlate with temperature to show that neighborhoods with more disadvantaged populations tend to have hotter summers. Both the percentage of white population and those with a Bachelor's or higher degrees have significant, negative correlations with temperature, showing that racial minorities and non-college graduates tend to live in places with higher temperatures. Similarly, neighborhoods with high average housing values also have lower temperatures. This relationship can possibly be explained by people's residential choices: places that have lower temperatures tend to be more inhabitable and have higher market values.

Interestingly, population density also has a significant negative correlation with summer temperature. This is likely because we utilized residential density for the regression, which decreased the measured density in city-center neighborhoods that only have high activity levels during the day.

R-squared	: 0.450258	R-squared (BUSE)	: -
Sq. Correlation	: -	Log likelihood	:37192.429540
Sigma-square	: 1.44936e-09	Akaike info criterion	: -74368.9
S.E of regression	: 3.80705e-05	Schwarz criterion	: -74318

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	-0.000132064	2.80646e-05	-4.70571	0.00000
* Pct_White	1.62057e-07	3.75024e-08	4.32124	0.00002
Pct_Bachel	5.89027e-08	5.39003e-08	1.09281	0.27448
Pct_Unempl	-5.37885e-08	1.27387e-07	-0.422246	0.67285
Pct_Rent_5	-3.5254e-08	3.62066e-08	-0.97369	0.33021
* Ln_Income	6.06161e-06	2.09078e-06	2.89922	0.00374
* Ln_Housing	1.1687e-05	1.51963e-06	7.69068	0.00000
* Populati_1	0.000959622	4.96114e-05	19.3428	0.00000
LAMBDA	0.600008	0.0144715	41.4612	0.00000

[1] Pct_White is the % of white population, Pct_Bachel is the % of residents with a Bachelor's degree or higher, Pct_Unempl denotes the % unemployment, Pct_Rent_5 denotes % of renters who are severely rent burdened, Ln_Income is the median household income, Ln_Housing is the median housing value, population_1 is a control variable for population density in the census block group.

[2] * denotes the variable is significant at the 5% significance level.

Table 3: Results of the spatial error regression exploring the distribution of urban greenery (measured by tree density)

R-squared	: 0.366890	R-squared (BUSE)	: -
Sq. Correlation	: -	Log likelihood	:-13250.647511
Sigma-square	: 26.6043	Akaike info criterion	: 26517.3
S.E of regression	: 5.15793	Schwarz criterion	: 26568.2

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	0.719014	3.79068	0.189679	0.84956
* Pct_White	0.0124032	0.0050076	2.47688	0.01325
Pct_Bachel	-0.0028229	0.00724728	-0.389511	0.69690
Pct_Unempl	0.014578	0.0172875	0.843268	0.39908
Pct_Rent_5	-0.00224479	0.00491285	-0.456923	0.64773
* Ln_Income	1.71646	0.282932	6.06669	0.00000
* Ln_Housing	-0.462669	0.204786	-2.25929	0.02387
* Populati_1	-40.19	6.68072	-6.01583	0.00000
LAMBDA	0.582302	0.0148708	39.1575	0.00000

Table 4: Results of the spatial error regression exploring the distribution of urban greenery (measured by GVI)

The regressions between social-demographic indicators and urban greenery measurements yield intriguing results. The percentage of white population, median household income, housing value, and residential density are all significant, but only the former two indicators yield consistent results between tree density and GVI. A higher percentage of white population and higher income both lead to more urban greenery. This relationship confirms the impact of historic redlining practices that denied investment to minority neighborhoods and resulted in the lack of public amenities in these areas.

The correlation between residential density and urban greenery measurements is an interesting one, as places with higher densities tend to have more street trees but lower GVI values. Comparing GIS maps for the individual indicators, we conclude that this correlation occurs because low-density neighborhoods tend to have more trees, shrubs, and lawns, but only a limited portion of those are street trees. In contrast, neighborhoods such as Upper East Side, Manhattan, and Park Slope, Brooklyn have more designated street trees but not much greenery otherwise because buildings tend to be directly adjacent to streets.

Similarly, the correlation between housing value and urban greenery measures also depends on whether a neighborhood has more street trees or greenery away from the roadways. We reason that because housing values tend to be the highest for city-center neighborhoods with more street trees than suburbs, high housing value is correlated with higher street tree density. In contrast, suburban neighborhoods have overall lower housing values but more lawn space and natural greenery that can be effectively captured by the GVI.

R-squared	: 0.149850	R-squared (BUSE)	: -
Sq. Correlation	: -	Log likelihood	:-10170.332466
Sigma-square	: 6.8309	Akaike info criterion	: 20356.7
S.E of regression	: 2.6136	Schwarz criterion	: 20407.5

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	2.11704	1.58082	1.3392	0.18051
* Pct_White	-0.00407406	0.00144219	-2.82491	0.00473
Pct_Bachel	-0.00509913	0.00268457	-1.89942	0.05751
* Pct_Unempl	-0.0189972	0.00870348	-2.18272	0.02906
Pct_Rent_5	-0.000406509	0.00249237	-0.163102	0.87044
* Ln_Income	0.51883	0.123204	4.21115	0.00003
Ln_Housing	-0.1342	0.0752783	-1.78272	0.07463
* Populati_1	-61.6673	2.48028	-24.863	0.00000
LAMBDA	-0.145533	0.0237083	-6.13849	0.00000

Table 5: Results of the spatial error regression exploring the distribution of pollution (measured by PM2.5 concentration)

The regression of social demographics on pollution gives very different results for each of the independent variables. First, the impact of percentage of white population follows an expected relationship, as increases in the proportion of white population leads to a significant decrease in pollution. This correlation further demonstrates the effect of systemic racism on urbanization patterns, as minority groups are more likely to live close to pollution sources like industries and power plants.

The regression of median household income on pollution shows a significantly positive correlation, which contradicts the regression for housing value. The special geography of New York City does mean that many neighborhoods in the lower part of Manhattan have both the highest traffic density and concentration of expensive apartments. The results from these wealthy neighborhoods with poor air quality may have skewed the correlation. We also recognize that there could be variables not included in our model that impact pollution levels.

Density also has a significant, negative correlation with pollution, which is similar to the regression on temperature. We believe that the bias of residential density data toward areas outside the city center is the main factor that led to the negative correlation.

R-squared	: 0.154804	R-squared (BUSE)	: -
Sq. Correlation	: -	Log likelihood	:-10159.320198
Sigma-square	: 6.7911	Akaike info criterion	: 20336.6
S.E of regression	: 2.60597	Schwarz criterion	: 20393.9

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	1.22929	1.58131	0.777384	0.43693
* Pct_White	-0.00361501	0.00143031	-2.52743	0.01149
Pct_Bachel	-0.00504516	0.00266427	-1.89364	0.05827
* Pct_Unempl	-0.0191547	0.00866732	-2.20999	0.02711
Pct_Rent_5	-0.000564061	0.00248303	-0.227167	0.82029
* Ln_Income	0.553477	0.122608	4.51421	0.00001
Ln_Housing	-0.0702078	0.075932	-0.924614	0.35517
* Populati_1	-57.7935	2.59576	-22.2645	0.00000
* Tree_densi	-3823.74	809.573	-4.72315	0.00000
LAMBDA	-0.156844	0.0237369	-6.60758	0.00000

Table 6: Results of the spatial error regression exploring the correlation between pollution (as measured by PM2.5 concentration) and urban greenery (as measured by tree density)

R-squared	: 0.150438	R-squared (BUSE)	: -
Sq. Correlation	: -	Log likelihood	:-10168.931040
Sigma-square	: 6.82618	Akaike info criterion	: 20355.9
S.E of regression	: 2.6127	Schwarz criterion	: 20413.1

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	1.92144	1.58429	1.21281	0.22520
* Pct_White	-0.0043359	0.00144968	-2.99094	0.00278
Pct_Bachel	-0.00517492	0.00268335	-1.92853	0.05379
* Pct_Unempl	-0.0188535	0.00870038	-2.16697	0.03024
Pct_Rent_5	-0.000297666	0.00249222	-0.119438	0.90493
* Ln_Income	0.495896	0.123907	4.00216	0.00006
Ln_Housing	-0.111004	0.0765087	-1.45086	0.14682
* Populati_1	-60.8205	2.53004	-24.0394	0.00000
Mean_GVI	0.0102431	0.0061173	1.67445	0.09404
LAMBDA	-0.146126	0.023708	-6.16358	0.00000

Table 7: Results of the spatial error regression exploring the correlation between pollution (as measured by PM2.5 concentration) and urban greenery (as measured by GVI)

The results of the regressions examining the effect of trees and urban greeneries on PM2.5 are contradictory to intuition. After controlling for demographic characteristics, tree density is negatively correlated with PM2.5 concentration while GVI is positively but not significantly correlated with PM2.5 concentration. This difference in signs between the two measures of urban greenery is not entirely unreasonable. Research in England has found that the aerodynamic dispersive effect of trees reduces PM2.5 concentration at greater levels than the deposition effect from trees and grass (Jeanjean, Monks, and Leigh 2016). Given that tree density measures solely street trees whereas GVI, due to the calculations used to derive it, tend to measure more types of urban greenery, it is possible that our tree density measure captured more of the dispersive effect and our GVI measure captured more of the deposition effect and therefore producing contradicting results in the regression. However, we do not know precisely why the sign and significance differ.

In this regression, some previously significant controls become statistically insignificant. Race, for example, is no longer statistically correlated with PM2.5 levels, contradicting earlier findings. Additionally, population density remains negatively correlated with PM2.5 concentration, which is contradictory to the fact that PM2.5 concentration is higher in the City's dense downtown and midtown areas. There may be other variables that we failed to account for that are affecting the results, or urban greenery simply does not have a significant effect on reducing PM2.5 pollution in New York City. Unfortunately, our research cannot provide a definitive conclusion.

R-squared	: 0.566758	R-squared (BUSE)	: -	
Sq. Correlation	: -	Log likelihood	:-6794.074507	
Sigma-square	: 1.2128	Akaike info criterion	: 13606.1	
S.E of regression	: 1.10127	Schwarz criterion	: 13663.4	

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	41.1222	0.828368	49.6424	0.00000
* Pct_White	-0.00581532	0.00119767	-4.85554	0.00000
* Pct_Bachel	-0.0111947	0.00163965	-6.82746	0.00000
Pct_Unempl	0.00347128	0.0036339	0.955248	0.33945
Pct_Rent_5	0.000532271	0.00103438	0.514578	0.60685
Ln_Income	-0.019512	0.0607448	-0.321214	0.74805
* Ln_Housing	-0.170109	0.0456361	-3.72751	0.00019
* Populati_1	-7.71269	1.55038	-4.9747	0.00000
* Tree_densi	3337.81	450.952	7.40171	0.00000
LAMBDA	0.725326	0.0113095	64.1341	0.00000

Table 8: Results of the spatial error regression exploring the correlation between land surface temperature and urban greenery (as measured by tree density)

R-squared	: 0.576811	R-squared (BUSE)	: -	
Sq. Correlation	: -	Log likelihood	:-6738.205212	
Sigma-square	: 1.18466	Akaike info criterion	: 13494.4	
S.E of regression	: 1.08842	Schwarz criterion	: 13551.6	

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	40.6363	0.81665	49.7598	0.00000
* Pct_White	-0.00502794	0.00117756	-4.2698	0.00002
* Pct_Bachel	-0.0112058	0.0016175	-6.92787	0.00000
Pct_Unempl	0.00399954	0.00359404	1.11283	0.26578
Pct_Rent_5	0.000376795	0.00102277	0.368406	0.71257
Ln_Income	0.0674821	0.0602056	1.12086	0.26235
* Ln_Housing	-0.138578	0.0448017	-3.09313	0.00198
* Populati_1	-5.87411	1.47444	-3.98395	0.00007
* Mean_GVI	-0.0428017	0.00328898	-13.0137	0.00000
LAMBDA	0.720457	0.0114446	62.9518	0.00000

Table 9: Results of the spatial error regression exploring the correlation between land surface temperature and urban greenery (as measured by GVI)

The temperature regression with tree density is done using the spatial error model and generates interesting results. The most surprising result is that tree density is positively and significantly correlated with temperature in this model. This means that controlling for all other values, an increase in tree density will lead to an increase in temperature, the exact opposite of our hypothesis. The other surprising result is that population density is negatively correlated with temperature according to our regression. The percentage of white population, residents with at least a Bachelor's degree, and median housing value are negatively and significantly correlated with temperature, fitting the reasonable assumption that those financially better off are more likely to live in more comfortable neighborhoods. Income, unemployment, and rent-burdened status are insignificant in this regression.

The temperature regression with GVI is similarly conducted using the spatial error model. In this model, GVI is significantly and negatively correlated with temperature. Percent white population, percent having Bachelor's or higher, and median housing value remain negatively and significantly correlated with temperature, and income, unemployment, and rent-burdened status remain insignificant. Population density continues to show negative correlation with temperature.

Comparing the two results, it is evident that GVI serves as a better indicator of urban greenery. The street tree density measure failed to account for trees in the City's numerous parks which are areas of low land surface temperature. Additionally, our tree density measure does not account for the size of trees, which has been proven to reduce urban land surface temperature (Wang and Akbari 2016). The GVI does better in accounting for the size of trees as it counts the number of vegetation pixels in a given image, and we sampled points inside parks to ensure full coverage. The quality of the measures is reflected in the regression results: tree density is positively correlated with temperature while GVI is negatively correlated. Using the GVI measure, we therefore conclude that trees and urban greenery have a significant and negative effect on urban land surface temperature. The negative correlation between population density and temperature, however, is difficult to explain in theory and in practice, and would require further research to discern the precise effect of density alone on urban temperature.

Conclusion and Limitation

Based on our research and analysis, we conclude that environmental inequalities continue to exist across race, income, and education lines, while urban greeneries are effective at reducing urban temperature. We based the first part of our conclusion on the evidence that the patterns of surface temperature correlate significantly with neighborhood racial composition, median income, and percentage of residents with a Bachelor's or higher degree. We also found PM 2.5 levels to be consistently higher in neighborhoods with a larger non-white population. On the other hand, minority and low-income neighborhoods also have consistently lower levels of urban greenery. The second part of our conclusion is informed by the clear negative correlation between mean GVI and surface temperature. Although we were unable to arrive at a definitive conclusion on the effect of urban greenery on pollution, the significant relationship between urban greenery and surface temperature suggests that trees can be considered a tool to combat existing environmental inequality at least in New York City.

However, there are some limitations to our research. First, while using spatial-weighted regression improved our regression results and significance, the regression methods we used could be perfected. We decided between using either the spatial error or the spatial lag model as GeoDa does not allow running a model that accounts for both spatial lag and error. However, it is possible for the error terms as well as the dependent variables in the regressions to be correlated with errors and the independent variables from nearby areas, therefore requiring the regression model to account for both spatial error and lag. Without accounting for both, there may be bias in our estimates.

Second, there may be unknown or unobserved variables that influenced the estimates of the regression. While we controlled for demographic differences, a number of variables could have potentially influenced the measures used in the data. The angle of the sun when temperature images were taken could have influenced the land surface temperature as skyscrapers cast long shadows in the City. The construction of highways nearby, the age of the building as well as the inequality within a census block group could all influence the regression results, but we were unable to find satisfactory measures to be able to control for these variables. These omitted and unobserved variables may have caused bias in our regression estimates.

Additionally, the accuracy of data may reduce the accuracy of our results as well. The tree density data did not account for the size of the tree while the GVI score failed to account for the effects of scale — a tree nearby occupies a greater number of pixels than a tree further away, but both may generate the same effect on pollution and temperature. Additionally, the machine learning model we used to extract urban greenery may have identified other objects of similar color and shape as greenery, but it was impossible to verify the tens of thousands of images we collected effectively. These inaccuracies in the measures of urban greenery may have caused imprecise estimates of their effect.

Bibliography

Boyce, James K., Klara Zwickl, and Michael Ash. 2016. “Measuring Environmental Inequality.” *Ecological Economics* 124:114–23. doi: 10.1016/j.ecolecon.2016.01.014.

Hoffman, Jeremy S., Vivek Shandas, and Nicholas Pendleton. 2020. “The Effects of Historical Housing Policies on Resident Exposure to Intra-Urban Heat: A Study of 108 US Urban Areas.” *Climate* 8(1):12. doi: 10.3390/cli8010012.

Jbaily, Abdulrahman, Xiaodan Zhou, Jie Liu, Ting-Hwan Lee, Leila Kamareddine, Stéphane Verguet, and Francesca Dominici. 2022. “Air Pollution Exposure Disparities across US Population and Income Groups.” *Nature* 601(7892):228–33. doi: 10.1038/s41586-021-04190-y.

Jeanjean, A. P. R., P. S. Monks, and R. J. Leigh. 2016. “Modeling the Effectiveness of Urban Trees and Grass on PM2.5 Reduction via Dispersion and Deposition at a City Scale.” *Atmospheric Environment* 147:1–10. doi: 10.1016/j.atmosenv.2016.09.033.

Ki, Donghwan, and Sugie Lee. 2021. “Analyzing the Effects of Green View Index of Neighborhood Streets on Walking Time Using Google Street View and Deep Learning.” *Landscape and Urban Planning* 205:103920. doi: 10.1016/j.landurbplan.2020.103920.

Li, Xiaojiang, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. 2015. “Assessing Street-Level Urban Greenery Using Google Street View and a Modified Green View Index.” *Urban Forestry & Urban Greening* 14(3):675–85. doi: 10.1016/j.ufug.2015.06.006.

New York City Department of Health and Mental Hygiene. 2022. “NYCCAS Air Pollution Rasters.” *NYC Open Data*. Retrieved April 25, 2022 (https://data.cityofnewyork.us/Environment/NYCCAS-Air-Pollution-Rasters/q68s-8qxv).

New York City Department of Health and Mental Hygiene. n.d. “The New York City Community Air Survey: Neighborhood Air Quality 2008 – 2020.” *NYC Health*. Retrieved April 25, 2022 (https://nyccas.cityofnewyork.us/nyccas2022/report/3).

NYC Department of Parks and Recreation. 2020. “2015 Street Tree Census - Tree Data.” *NYC Open Data*. Retrieved April 25, 2022 (https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh).

Wang, Yupeng, and Hashem Akbari. 2016. “The Effects of Street Tree Planting on Urban Heat Island Mitigation in Montreal.” *Sustainable Cities and Society* 27:122–28. doi: 10.1016/j.scs.2016.04.013.

Zhou, Bolei, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2016. “Semantic Understanding of Scenes through the Ade20k Dataset.” *ArXiv Preprint*

ArXiv:1608.05442.

Zhou, Bolei, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. "Scene Parsing through ADE20K Dataset." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.