GIS & SPATIAL ANALYSIS - SOC SCI, 2023

Author: Oguz Alp Eren

Homework 6: Spatial Autocorrelation Diagnostic and Spatial Lag/Error Models

Summary

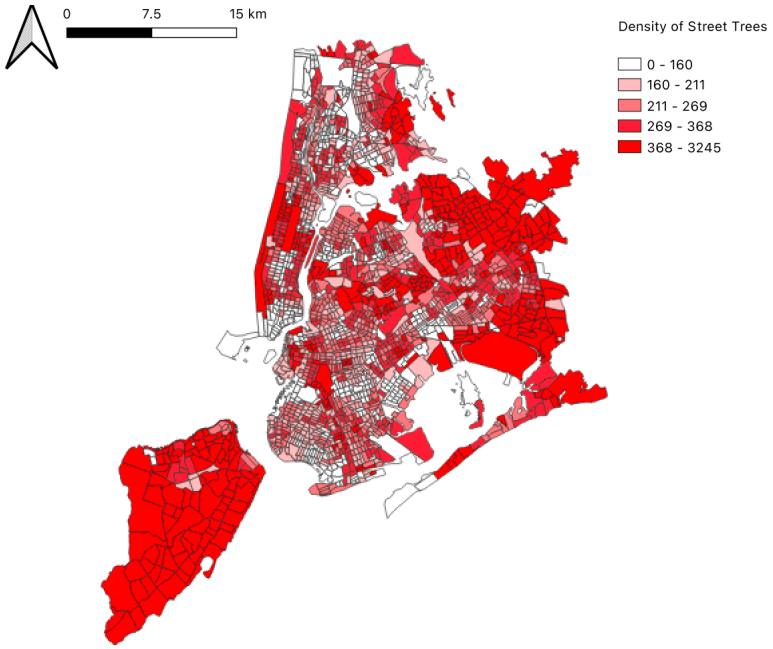
- **Description of main hypothesis:** The hypothesis suggests that areas with higher Black populations in New York City have fewer trees, indicating a disparity in access to green spaces and the benefits they provide. This hypothesis is explored through various statistical models, controlling for factors like household income, population size, and age demographics.
- **Description of the results of final bivariate analysis:** The preliminary analysis reveals a spatial distribution of trees that is dense in residential areas like Queens, Staten Island, and Upper Manhattan, but less in areas like Midtown, parts of Brooklyn, and the Bronx. The correlation coefficients indicate a slight negative correlation between the African American population and the number of trees (-0.1048), while a positive correlation exists between the white population and tree density (0.2506). This suggests a racial division in access to natural resources.
- **Description of the results from the regression analyses:** Three models were employed to investigate the correlation between the Black population and tree density in New York City, each controlling for different variables. The Ordinary Least Squares (OLS) model showed a moderate relationship with an R-squared of 0.260, adjusting for factors like household income and age demographics. The Spatial Lag model revealed significant positive spatial autocorrelation (Rho = 0.755), indicating clustering of areas with higher Black populations. Lastly, the Spatial Error model demonstrated strong spatial autocorrelation (Lambda = 0.846) with a high R-squared value around 0.75, suggesting a substantial explanation of the variance in the distribution of the Black population, emphasizing the clustering and potential environmental disparities in urban tree distribution.

Introduction

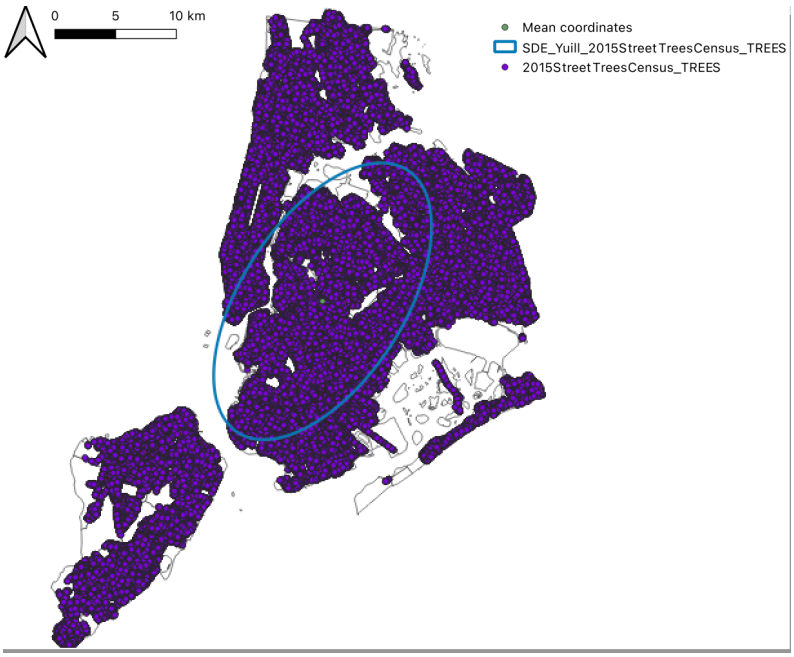
This analysis merges two key datasets: the 2015 Street Tree Census and the 2010 Census Tracts of New York City, to explore the relationship between tree density and demographic factors. The Street Tree Census, with details on species, size, and locations of trees, enables mapping tree distribution trends across boroughs and neighborhoods. Meanwhile, the 2010 Census Tracts dataset provides demographic and socioeconomic contexts, such as income, population density, and living conditions.

Three statistical models investigate the correlation between the Black population and tree numbers in city tracts, adjusting for variables like household income, population size, and the elderly proportion. The Ordinary Least Squares (OLS) model assesses this relationship, factoring in economic and demographic influences and employing Queen Contiguity and Moran's I statistic for spatial correlation, indicating moderate explanatory power and notable spatial autocorrelation. The Spatial Lag model further analyzes the impact of neighboring areas on the Black population, showing significant spatial clustering and high explanatory power. Finally, the Spatial Error model addresses unobserved spatial factors, evident in a substantial spatial error coefficient and a high R-squared value, effectively capturing the complex spatial relationships of demographic patterns in New York City.

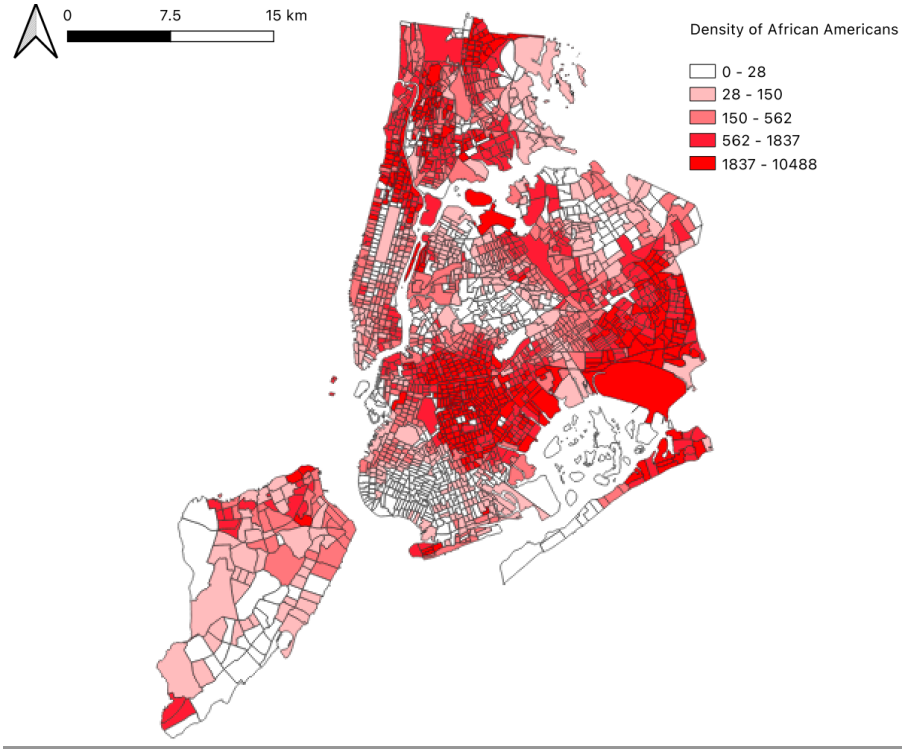
Preliminary Analysis:



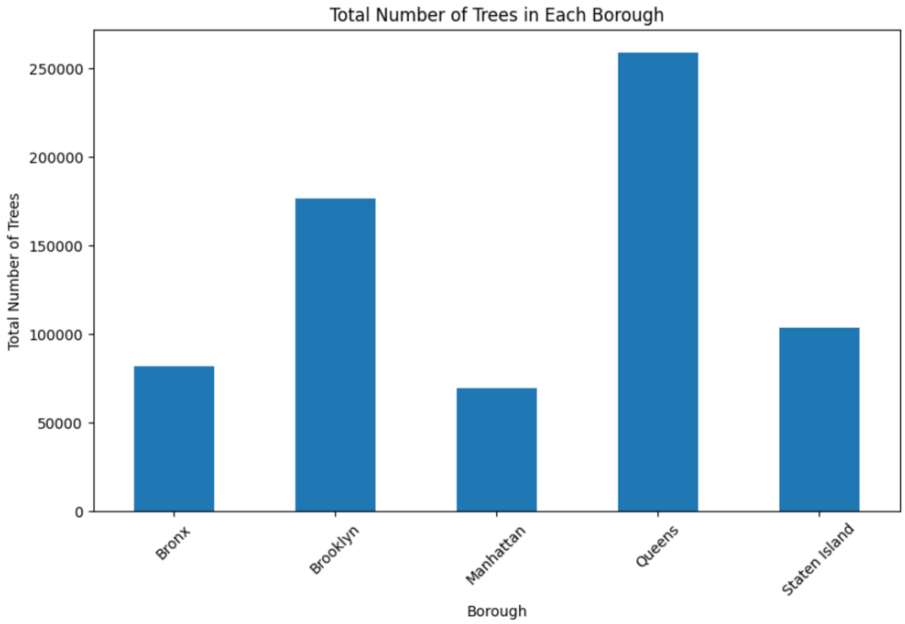
According to the map above, it is evident that almost all the residential areas in NYC are densely populated with trees whereas non-residential places like the financial district and industrial areas are less dense in contrast. We can observe that Queens, Staten Island, Upper Manhattan are very densely populated with trees. However, looking in more depth in the residential areas, we can see that Midtown, some areas in Brooklyn, and the Bronx are less dense.



This map presents the mean coordinates and standard deviational ellipse for individual trees in New York City. It clearly illustrates that NYC has a dense population of trees, with the mean coordinate located centrally in Queens. While this map effectively highlights the general concentration of trees in NYC, it offers limited insight into other parts of tree distribution within New York City as a whole.

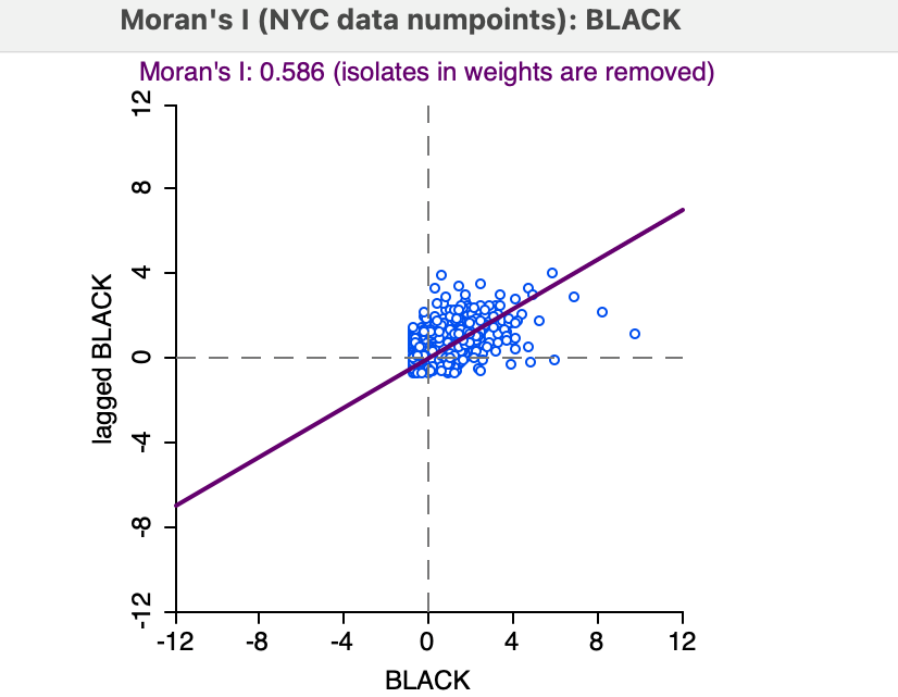


Examining the map that depicts the distribution of African American populations in New York City, it becomes somewhat apparent that there exists an inverse relationship with tree density. This observation could potentially point to a form of environmental inequality affecting the residents of the city.



This is a Histogram showing the number of trees based on Boroughs in NYC. According to this graph, Queens has the most trees among all boroughs and Manhattan has the least trees with Bronx in close second. Although, this map does not show the density of trees in the boroughs, so spatially bigger boroughs have more trees, even if they are less dense per square meter.

In depth analysis reveals that the correlation coefficient is -0.1048 for the African American population and the number of trees in tracts, suggesting a slight negative correlation on average. Conversely, the correlation between the white population and the number of trees in tracts is positively correlated with 0.2506 being the correlation coefficient. This suggests a racial division between the racial groups in NYC. Looking at household income and the number of trees in tracts we can observe a correlation coefficient of 0.1749, again suggesting a division of natural resources among civilians.



Lagged BLACK	Value
Moran's 1	0.586

A Moran's I value close to +1 suggests a strong positive spatial autocorrelation. In your case, 0.59 is a substantial positive value. This means that areas with high concentrations of Black residents are likely to be surrounded by areas with similarly high concentrations, and areas with low concentrations are surrounded by areas with similarly low concentrations. This makes sense as racial segregation patterns are evident in NYC, based on prior analyses.

Models

Model 1, Ordinary Least Squares:

This model test whether there is a significant correlation between the black population and number of trees in tracts in NYC. Household income, population, and age 65 and up population are added as control variables. These control variables serve the purpose of accounting for potential influences on the relationship between the number of trees and the Black population.

Household income helps control for economic disparities that might affect resource allocation, while population size considers the demand for amenities like green spaces. The proportion of the population aged 65 and older acknowledges the potential influence of age-related preferences for outdoor spaces, accounting the effect of more vivid tracts to more peaceful ones.

This output summarizes the results of an Ordinary Least Squares (OLS) regression analysis. The dependent variable "BLACK" has an R-squared of 0.260, indicating that approximately 26% of its variability is explained by the model, which includes four predictors (NUMPOINTS, eSHAVGINC, POP2009, AGE_65_UP). The coefficients for these predictors, along with their p-values, suggest varying levels of statistical significance, with all predictors appearing to be significantly associated with the dependent variable at conventional levels ($p < 0.05$).

Model 1: Coefficients	**Value**
NUMPOINTS	0.0898
eSHAVGINC	0.2060
POP2009	0.6073
AGE_65_UP	0.2569

Model 1: Results	**Value**
R Squared	0.26
Durbin Watson	1.05

Model 2, Spatial Lag:

The model reveals a significant positive spatial autocorrelation in the distribution of the Black population (Rho ≈ 0.755). This spatial pattern indicates that areas with a higher concentration of Black residents tend to be spatially clustered, highlighting a non-random distribution. The model's R-squared value of approximately 0.678 suggests that the included variables collectively explain a substantial portion of the variance in the Black population, indicating a reasonably good fit.

Queen Contiguity is used, which defines spatial relationships between geographic units based on shared edges or vertices. In this case, it considers two geographic units as neighbors if they share

a common boundary or vertex. This weight structure is appropriate for considering both contiguity and adjacency in defining spatial relationships.

Delving into specific coefficients, the negative coefficient for "NUMPOINTS" (-0.255) implies a potential relationship between the number of trees and the Black population. As the number of points increases, the Black population tends to decrease, suggesting that neighborhoods with more trees may have a lower Black population concentration. Additionally, other covariates such as "POP2009," "AGE_65_UP," and "eSHAVGINC" also exhibit significant relationships with the Black population, further underscoring the complexity of factors influencing demographic distribution. Overall, this analysis unveils insights into the spatial distribution of the Black population in New York City, highlighting the presence of spatial autocorrelation and its associations with various socioeconomic and environmental factors.

Model 2: Coefficients	**Value**
W BLACK	0.754741
NUMPOINTS	-0.254936
POP2009	0.245233
AGE 65 _UP	-0.387304
eSHAVGINC	-0.00357377

Model 2: Results	**Value**
R-Squared	0.68
Lag Coeff	0.75

Model 3, Spatial Error:

The model indicates a strong positive spatial autocorrelation, denoted by the spatial error coefficient (Lambda) of approximately 0.846471. This suggests that there is spatial clustering in the distribution of the Black population across New York City. Areas with a high concentration of Black residents tend to be spatially adjacent to other areas with similar characteristics, revealing a non-random pattern in demographic distribution. The R-squared value of around 0.749263 implies that the model explains a significant proportion of the variance in the Black

population, indicating its effectiveness in capturing spatial relationships. The covariates, including "NUMPOINTS," "POP2009," "AGE_65_UP," and "eSHAVGINC," exhibit significant associations with the Black population, underscoring the multifaceted nature of factors influencing demographic patterns. Additionally, the model reports diagnostics for heteroskedasticity, which indicate potential variability in the error terms, and spatial dependence, confirming the presence of spatial autocorrelation.

This spatial error model provides insights into the spatial distribution of the Black population in New York City. The strong spatial autocorrelation suggests that areas with similar demographic characteristics are clustered together, revealing a spatially dependent pattern. The model demonstrates the importance of considering spatial relationships when analyzing demographic distributions, highlighting the significance of various socioeconomic and environmental factors in shaping these patterns.

Model 3: Coefficients	**Value**
NUMPOINTS	-0.25922
POP2009	0.342617
AGE_65_UP	0.5927
eSHAVGINC	0.0045477
LAMBDA	0.846471

Model 3: Results	**Value**
R Squared	0.75
Lag Coefficient	0.85

Conclusion

The fundamental difference between these models lies in how they account for spatial dependence. In the spatial lag model, the emphasis is on the direct influence of neighboring observations on the dependent variable. It assumes that the value of the dependent variable in one location is influenced by the values of the same variable in adjacent areas, quantified by the spatial lag coefficient (Rho). This model is suitable when there is a theoretical basis for direct

spatial interaction, such as in cases where a change in a variable in one location directly impacts nearby locations.

Conversely, the spatial error model takes a different approach by considering spatial autocorrelation as inherent in the error terms of the model. It acknowledges that unobserved spatial factors or omitted variables may lead to spatial correlation in the residuals, and this spatial autocorrelation is captured by the spatial error coefficient (Lambda). The spatial error model is chosen when spatial autocorrelation is attributed to unobserved factors or measurement error, and it accounts for the indirect or unexplained spatial effects that influence the residuals.

Compared to the Ordinary Least Squares (OLS) and spatial lag models, the spatial error model appears to be the most suitable choice based on the outputs. While OLS lacks the consideration of spatial autocorrelation, the lag model primarily focuses on direct spatial interactions. In contrast, the spatial error model accounts for unobserved spatial factors affecting the residuals, which aligns with the strong spatial autocorrelation observed in the Black population distribution. The spatial error model's Lambda coefficient of approximately 0.846 reflects this spatial autocorrelation, while also achieving a high R-squared value of around 0.749263, indicating a good fit. Therefore, the spatial error model provides a more comprehensive understanding of the complex spatial relationships affecting the Black population in New York City, making it the preferred among the tree models.

The hypothesis that the number of trees is correlated with the Black population in New York City is supported by the results. The spatial error model, which accounts for unobserved spatial factors, shows a strong spatial autocorrelation in the Black population distribution, indicating a non-random pattern that is likely influenced by various factors, including the presence of trees. However, this analysis is limited in that it does not consider or test the possibility of a mediating factor that affects both the black population and the density of the trees. So, inferring causal relationship would be very risky.