Modeling Urban Heat Islands: Investigating Variables Influencing Urban Heat Island Intensity in the United States

Sarah Kwon

MGIS, Pennsylvania State University

Geog 870

Dr. Fritz Connor Kessler

April 20, 2024

1. Introduction

The Urban Heat Island effect (UHI) is a phenomenon where cities and metropolitan areas, on average, experience higher land and surface air temperatures compared to their surrounding rural areas (1). According to the U.S. Environmental Protection Agency (EPA), urban heat islands exhibit daytime temperatures between 1 and 7 degrees Fahrenheit warmer, while nighttime temperatures are 2 to 5 degrees warmer in the city than in their surrounding areas (2). Despite growing recognition of UHI significance, our understanding of its spatial and temporal dynamics across different cities remains limited. Existing research has primarily focused on individual cities or regions, often overlooking the broader regional variability of UHI intensity and its underlying drivers. To address these gaps in knowledge, this project seeks to investigate the regional variability of UHI intensity across different cities in the U.S. For this project, UHI intensity is characterized by increased surface air temperatures, land temperatures, or temperature anomalies. The purpose is to identify and quantify the key variables associated with UHI intensity and assess how these variables vary spatially and temporally within cities in the U.S using regression analyses.

2. Literature Review

UHI is a large environmental and human health issue as the impacts of a warming city results in increased energy consumption, elevated emissions of air pollution, increased instances of heat-related illnesses and death, and impaired water quality (2). It is essential to understand the current variables and indicators associated with this phenomenon. While delving into the literature, the following variables were notably associated with UHI intensity: impervious surfaces, vegetation, demography, and climate.

Impervious Surfaces

One frequented indicator of UHI is the percentage of impervious surfaces within a city (3). Impervious surfaces consist of pavements, roads, and other indications of human settlements. There is strong evidence that impervious surfaces impact the surrounding environment (2, 3, 5, 6). Studies have shown that impervious surfaces increase UHI intensity by trapping heat within a city, blunting evaporative cooling, and reducing albedo (5, 6, 7). In addition to increasing urban heat intensity, impervious surfaces degrade their surrounding environment by increasing stormwater run-off, minimizing ground-water recharge, and reducing greenspace (2, 4). Shi et al. (2023) and Hua et al. (2020) suggested that using impervious surface area, as a proxy for urbanization, is a reasonably effective method for predicting and understanding the spatial and temporal patterns of UHI at a global scale (3, 5). This is also further supported by mitigation strategies that target these surfaces. Multiple studies found that modifying such surfaces yield significant results in reducing urban heat through energy dissipation or reflection (2, 7, 8, 9, 10).

Vegetation

Another extensively researched variable associated with UHI is vegetative coverage. Terms such as vegetation or green space refer to land composed of natural elements and plant cover. A systematic review investigating the impact of vegetative coverage on urban areas concluded that they have a cooling effect, with the caveat that the effectiveness of vegetation can vary based on characteristics such as tree species, distribution, and overall urban design (11). The review also indicated that vegetation reduced surface air temperature through actions such as reflection of UV radiation, evapotranspiration by plants, diminished heat absorption, and filtration of atmospheric pollutants, including greenhouse gases (11). Canopy coverage stood out as a key influencer on urban heat intensity due to its role in providing shade and reflecting solar radiation (11).

Beyond the previously mentioned systematic review, consistent findings across different research also support these claims (2, 7, 11, 12). However, it seems there are more studies examining the cooling effect of vegetation than it being a good indicator of UHIs. One study estimated that a 10% increase in vegetative cover can potentially lead to a 7% reduction in heat-related deaths in Washington, D.C. (12). The Environmental Protection Agency also recognizes this impact and documents canopy coverage as an effective strategy to reduce urban heat islands (2). Based on previous research, vegetation stands as a key variable associated with UHIs, warranting its inclusion in the analytical framework of this project.

Demography

Multiple studies examining the relationship between demographic factors and urban heat islands (UHI) provided evidence of thermal disparities. The overall findings include the significant correlation of population-related parameters, such as city size, population density, race, and income, on UHI intensities. Recent literature showcased the role of population density and size, in determining temperature variations within the city (13). This aligns with previous studies that emphasized the interaction between UHI intensity and city size (1, 14). This overarching trend suggests that both larger and denser cities tend to exhibit heightened urban heat intensity.

Other demographic studies reveal an unsettling trend across various U.S. cities. Nation-wide research reported that communities of color and low-income are positively correlated with increased UHI intensity (15, 16). Moreover, Saverino et al. (2021) study in Richmond, VA, adds a historical dimension to the narrative, providing evidence for past discriminatory practices in urban development contributing to the present spatial distribution of extreme heat (17). There is a clear association between UHI intensity and communities facing socio-economic challenges. Incorporating demographic variables into the regression analysis can reveal socio-economic influences on UHI intensity, aiding in identifying vulnerable populations and guiding targeted mitigation strategies.

Climate

Finally, research has also demonstrated the influence of climate types and weather on UHI intensity. According to research, factors like precipitation, humidity, and spatial/temporal patterns significantly affect urban heat (18, 19, 20, 21). Two studies specifically discuss the impact of precipitation on heat retention within a city (18, 21). Chen & Wang (1995) noted that while rain can slow daytime temperature rises, it also reduces nighttime cooling (18). Similarly, Yang et al. (2019) found that prolonged daytime precipitation in general leads to lower urban heat (21).

Moreover, other studies highlight the spatial and temporal drivers of UHI across various cities, revealing that geographic location does have a significant impact on UHI (19, 22). Varquez & Kanda (2018) found that cities within a drier climate regime, as categorized by the Köppen -Geiger climate sub-classes, have a higher nighttime UHI, while tropical cities had the lowest UHI (22). These global findings suggest that similar dynamics may occur in urban areas of the U.S., warranting further investigation into how climate factors influence heat intensity in different U.S. cities.

3. Data

This project will address the question of if impervious surfaces, tree canopy coverage, median income, and population density are associated with UHI intensity in selected cities across the U.S. The project will also address if these variables play a larger or smaller role in determining UHI intensity. To investigate this question, the following data and methods will be used:

NOAA Urban Heat Mapping Campaign

The National Oceanic and Atmospheric Administration's (NOAA) Urban Heat Mapping Campaign will be used for the majority of the data (23). This is an ongoing nationwide dataset that contains surface air temperature readings (°F) coupled with Sentinel-2 satellite data across well-known cities in the United States. Temperature readings for this dataset were collected via volunteers using car and bike-mounted sensors over the summer months for the last four years. Surface air temperature recordings were collected in the morning (6 a.m.), afternoon (3 p.m.), and evening (7 p.m.) on a single day. Dataset variables for each city include temperature and temperature anomalies, impervious surface percentage, canopy coverage percentage, and various demographics data (i.e., total population, population <5, population >65, minority, median income, and poverty) for each neighborhood. Temperature anomalies are defined as neighborhood temperature compared to the citywide average based on the Climate Adaptation Planning and Analytics (CAPA) data. CAPA is a NOAA funded initiative aimed at climate-focused data analytics.

City Selection

Additionally, cities selected for statistical analyses were chosen considering their climate type. This approach is justified by the well-established literature on the significant impact of climate on UHIs (18, 19, 20, 21). As highlighted in the literature review, climatic conditions play a crucial role in shaping the intensity and characteristics of UHI. The classification of climate type adhered to the Köppen-Geiger climate classification system, which provides an overview based on temperature and precipitation patterns (24). Cities were selected based on a variety of climate types available within the U.S., such that the model can provide predictions across different environmental contexts. Four cities were selected: 1) District of Columbia, 2) Detroit, Michigan, 3) El Paso, Texas, and 4) Miami, Florida. These four cities were selected to represent the climate types of C - Temperate, D - Continental, B – Arid, and A – Tropical, respectively. At the time of this project, only a handful of cities were mapped in NOAA’s Urban Heat Mapping dataset. Therefore, the final selection of cities was chosen based on meeting the criteria of 1) having a diverse range of climate types and 2) having a sample size of at least 30.

4. Methodology

Statistical Analyses

To create a comprehensive model for Urban Heat Islands, a regression analysis was used. To determine a model fit, each city dataset went through a series of exploratory data analyses. These include normality tests on the data and the OLS residuals. Outliers were removed and transformations were performed in attempts to normalize each dataset. Outliers were detected using the Interquartile Range (IQR), where observations outside of the lower and upper bounds were removed. Any non-normal datasets were transformed based on skewness. For data with skewness greater than 1, a log transformation was applied with a small constant added to handle zero values. For data with skewness less than -1, an exponential transformation was applied. For data with skewness between -1 and 1 and containing negative values, a cube root transformation was applied. For data with skewness between -1 and 1 and containing only positive values, a square root transformation was applied. After, a spatial autocorrelation test was performed to check for clustering in the OLS residuals of each variable. Spatial weights were calculated using the K-Nearest Neighbor (KNN) algorithm due to its basis in a distance metric. The number of neighbors (k) was calculated using the formula: $k = \sqrt{(N)}/2$, where N represents the sample size [25].

If clustering was present, a Spatial Error Regression (SER) model or a Spatial Lag Regression (SLR) model was used. In the analysis, four predictor variables were utilized across the four chosen U.S. cities. The predictor variables for the analysis include: 1) Impervious Surface Percentage, 2) Canopy Coverage Percentage, 3) Total Population Density, and 4) Median Income. These four variables represent heat-absorbing urban surfaces, vegetative land cover, city size, and socio-economic status. The dependent variable for this analysis is UHI intensity, represented by afternoon temperature anomalies. Afternoon temperature anomalies consist of neighborhood temperatures (°F) during peak heating hours relative to the citywide average. These variables were chosen to represent UHI factors as outlined by NOAA and the EPA (23, 26). Python was used to carry out the regression analyses. Python packages and modules used include SciPy, ArcPy, spreg, numpy, and pandas. Model summary and static maps will be used to communicate the results of the projects. All maps were created using ArcGIS Pro and Esri's World Imagery basemap (27).

5. Results

Normality Tests and Data Transformations

Exploratory Data Analysis (EDA) was used to identify which regression modeling approach yielded better outcomes for predicting afternoon temperature anomalies. A Shapiro-Wilks Normality test was performed on individual datasets from each city to assess its distribution. The test was performed on the raw data, after removal of outliers, and after applying transformations. A normally distributed dataset would align with one of the assumptions of an Ordinary Least Squares model, while a non-normally distributed dataset violates the assumptions of OLS. For this analysis, p-values > 0.05 fails to reject the null hypothesis of a normally distributed data. The final selection of datasets was based on those yielding the highest resulting p-values. There were some cases where the transformation led to lower p-values. In those instances, the datasets with only the outliers removed were used. Outliers were removed for each dataset even when the raw datasets were normally distributed.

Some variables did have p-values > 0.05, but none of the four cities had all five variables normally distributed after outliers were removed and/or transformations were applied.

Table 1: Shapiro-Wilks Normality Test Results:

** Indicates a normally distributed dataset*

District of Columbia		
Variable Name	**Raw data p-value**	**Final data p-value**
Afternoon Temperature Anomalies	0.065*	0.208*
Median Income	0.001	0.001
Tree Canopy	2.473e-15	2.412e-13
Impervious Surfaces	0.002	0.000
Population Density	1.31e-08	0.277*

Detroit		
Variable Name	**Raw data p-value**	**Final data p-value**
Afternoon Temperature Anomalies	0.017	0.017
Median Income	5.828e-17	0.506*
Tree Canopy	7.148e-20	0.468*
Impervious Surfaces	0.000	0.147*
Population Density	3.011e-07	0.018

El Paso		
Variable Name	**Raw data p-value**	**Final data p-value**
Afternoon Temperature Anomalies	1.874e-11	0.066*
Median Income	0.006	0.107*
Tree Canopy	1.488e-21	1.675e-08
Impervious Surfaces	0.007	0.298*
Population Density	0.360*	0.820*

Miami		
Variable Name	**Raw data p-value**	**Final data p-value**
Afternoon Temperature Anomalies	1.812e-06	0.071*
Median Income	4.916e-09	0.083*
Tree Canopy	9.628e-14	0.473*
Impervious Surfaces	0.035	0.092*
Population Density	7.029e-11	0.015

Exploratory Data Analysis

Both visual and analytical methods were used to determine Ordinary Least Squares (OLS) overall fit of the regression line for each city. An example of a visual method is included below (Figure 1, 2). The scatter plots were used in conjunction with their R-squared values to determine linearity. R-squared values closer to 1.0 indicate a strong linear relationship. Furthermore, the residuals of each city were plotted and tested for normality. Randomness in the residual plots can be indicative of error term normality (homoscedasticity), while non-random patterns can be indicative of a mean error other than 0 (heteroscedasticity). The Shapiro-Wilks Normality test (Table 2) was employed on the residuals to provide a clear indication of normality. P-values > 0.05 failed to reject the null hypothesis of a normal distribution while p-values ≤ 0.05 is statistically significant in rejecting the null hypothesis of a normal distribution.

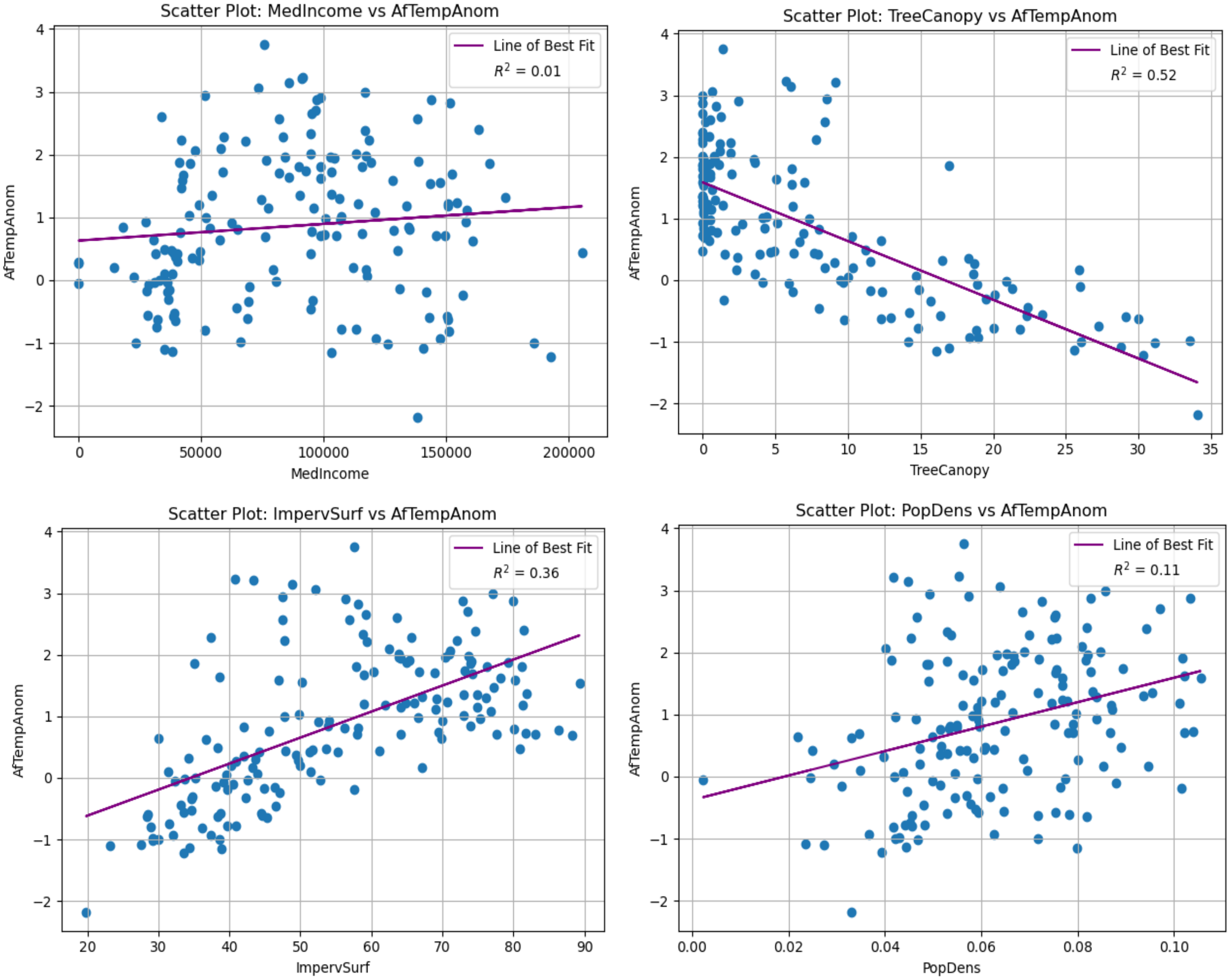


Figure 1. Scatter plot for the District of Columbia: Scatter plots showing the dependent variable (Afternoon Temperature Anomalies) against the independent variables (Median Income, Tree Canopy, Impervious Surfaces, and Population Density). The R-squared values and the line of best fit are included at the top to indicate the strength of linearity.

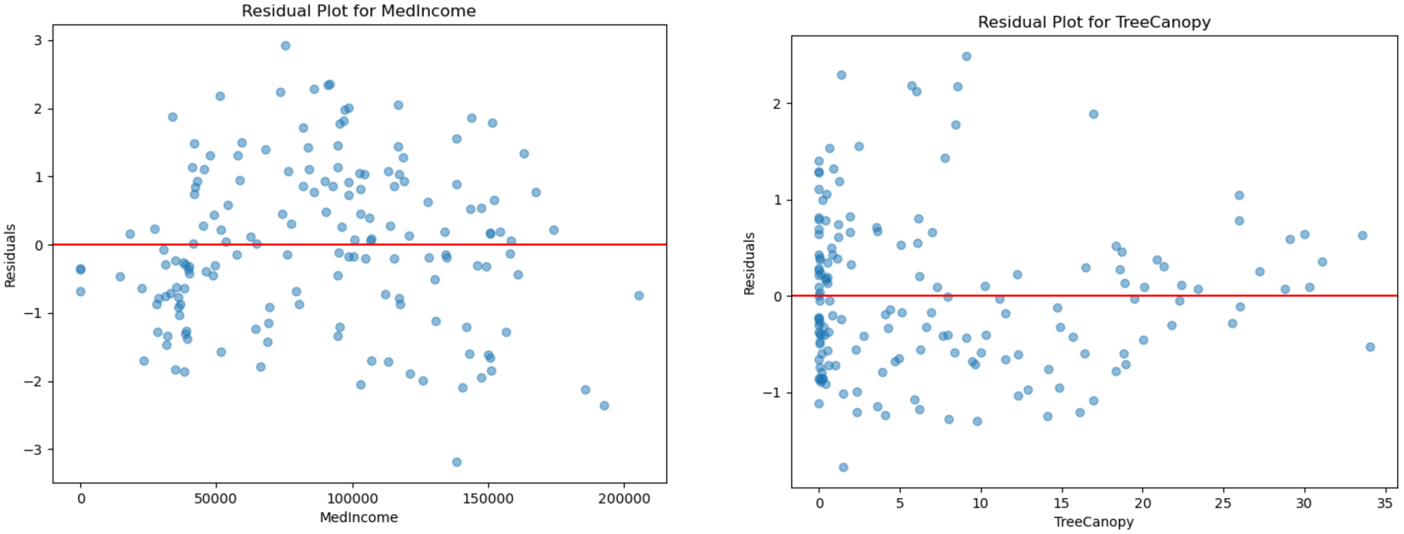


Figure 2. Residual plots for the District of Columbia: Residual plots showing the error terms against Median Income and Tree Canopy. The residual plot for Median Income (left) shows a normally distributed error variance. The residual plot for Tree Canopy (right) shows a cone-shaped pattern, indicative of heteroscedasticity.

Table 2. Normality Test on the Residuals:

** Indicates a normally distributed dataset*

District of Columbia		
Variable Name	**R-Squared**	**Residuals (p-values)**
Median Income	0.010	0.545*
Tree Canopy	0.525	0.000
Impervious Surfaces	0.363	5.940e-07
Population Density	0.106	0.156*

Detroit		
Variable Name	**R-Squared**	**Residuals (p-values)**
Median Income	0.012	0.010
Tree Canopy	0.046	0.005
Impervious Surfaces	0.081	0.002
Population Density	0.030	0.034

El Paso		
Variable Name	**R-Squared**	**Residuals (p-values)**
Median Income	0.020	0.119*
Tree Canopy	0.006	0.107*
Impervious Surfaces	0.003	0.062*
Population Density	0.001	0.056*

Miami		
Variable Name	**R-Squared**	**Residuals (p-values)**
Median Income	0.080	0.060*
Tree Canopy	0.011	0.041
Impervious Surfaces	0.016	0.192*
Population Density	0.046	0.107*

Based on the R-Squared values, none of the cities had a strong degree of linear correlation. Tree Canopy % in the District of Columbia had the strongest linear relationship (0.535) compared to the remaining predictor variables.

In terms of looking at the residual plots, D.C. had two out of four variable residuals normally distributed, while Detroit had no variable residuals normally distributed. The city of El Paso had all four predictor variable residuals normally distributed, and Miami had three out of four variable residuals normally distributed.

Finally, a Spatial Autocorrelation test with performed on the regression model's residuals of each city. Spatial Autocorrelation test analyzes if there is spatial clustering within the data. A p-value < 0.05 indicates that there is significant clustering present within the data. The Moran's I provide insight into the direction of clustering. A positive Moran's I suggest that neighboring observations tend to have similar residual values, indicating spatial autocorrelation, while a negative Moran's I implies dissimilarity among neighboring residuals.

Table 3. Spatial Autocorrelation Test Results:

** Indicates significant Moran's I*

District of Columbia		
Variable Name	**Moran's I**	**p-values**
Median Income	0.611	0.000*
Tree Canopy	0.527	0.000*
Impervious Surfaces	0.638	0.000*
Population Density	0.631	0.000*

Detroit		
Variable Name	**Moran's I**	**p-values**
Median Income	0.606	0.000*
Tree Canopy	0.640	0.000*
Impervious Surfaces	0.659	0.000*
Population Density	0.585	0.000*

El Paso		
Variable Name	**Moran's I**	**p-values**
Median Income	0.357	0.000*
Tree Canopy	0.314	0.000*
Impervious Surfaces	0.342	0.000*
Population Density	0.329	0.000*

Miami		
Variable Name	**Moran's I**	**p-values**
Median Income	0.745	0.000*
Tree Canopy	0.733	0.000*
Impervious Surfaces	0.716	0.000*

Population Density	0.749	0.000*

All four cities had highly significantly positive spatial autocorrelation present. El Paso had the least positive Moran’s index (0.314 - 0.357), while Miami had the highest positive Moran’s index (0.719 – 0.749).

For D.C., Impervious Surface had the most positive Moran’s index (0.638), and Tree Canopy had the least positive Moran’s index (0.527). For Detroit, Impervious Surface had the most positive Moran’s index (0.659) while Population Density had the least positive Moran’s index (0.585). For El Paso, Median Income had the most positive Moran’s index (0.357), while Tree Canopy had the least positive Moran’s index (0.314). Finally, for Miami, Population Density had the most positive Moran’s index (0.749), while Impervious Surfaces had the least positive Moran’s index (0.716).

Exploratory Data Analysis Conclusions

Since each regression model exhibited spatial autocorrelation in the residuals, the OLS model may not be the best fit for the data. The presence of highly positive spatial autocorrelation in the residuals, violates the assumption of independent errors. The R-squared values for each city were also relatively low, indicating that predictors in the model are not explaining a large portion of the variance in Afternoon Temperature Anomalies. Due to the presence of highly significant spatial autocorrelation, a SER or a SLR model may be more suitable. Both models account for spatial autocorrelation, but in different ways. A SER accounts for spatial dependence in the model by modeling spatial autocorrelation in the residuals. On the other hand, SLR addresses spatial autocorrelation in the dependent variable.

Model Comparisons

Based on the EDA, a SER and a SLR was performed on each city. A comparison of the OLS, SER, and SLR models for each city are shown below. Significant Correlation Coefficients are indicated by a p-value ≤ 0.05 and are denoted by an asterisk (*). Two indicators of model fit were included: Akaike information criterion (AIC) and Log likeliness. Both provide a measure of goodness of fit; however, AIC considers model complexity. The AIC value will be used as the main metric of model fit in the comparison. As AIC and log likeliness values move closer to 0, a better model fit is concluded. The lambda value reports similar metrics as Moran's I where lambda reports the direction of spatial autocorrelation. Positive lambda values suggest that closer values are more similar, whereas negative lambda values indicate that closer values are less similar. P-values ≤ 0.05 indicate significant spatial autocorrelation in the residuals. Similarly, the spatially lagged dependent variable (W_y) captures how the dependent variable of a particular observation is influenced by its own independent variables and by the values of the dependent variable in neighboring observations based on the spatial weight matrix. The KNN weighs neighboring observations higher compared to observations that are further away.

With OLS, R-squared values are used to determine how much of the variation are in the dependent variable is accounted for by the predictor variables. Pseudo R-squared values are reported as diagnostics for the spatial error and spatial lag models where traditional R-squared values cannot be calculated. Pseudo R-squared values should be limited to comparisons between the same models, as they can be affected by model complexity. Finally, Spatial Pseudo R-squared values are like traditional R-squared values, but they are adapted to account for spatial dependence in the data.

Table 4. Model Summary for District of Columbia:

** Indicates significant Correlation Coefficients*

Ordinary Least Squares Model					
Variable Name	**Coefficient**	**p-value**	**R-squared**	**AIC****	**Log likelihood**
Median Income	2.66e-06	0.194	0.010	518.1	-257.05
Tree Canopy	-0.051	0.000*	0.525	398.6	-197.30
Impervious Surfaces	0.021	0.000*	0.363	446.4	-221.18
Population Density	-0.466	0.000*	0.106	501.6	-248.79

Spatial Error Regression Model		
Variable Name	**Coefficient**	**p-value**
Median Income	0.000	0.968
Tree Canopy	-0.051	0.000*
Impervious Surfaces	0.021	0.003*
Population Density	-0.466	0.890
lambda	0.773	0.000*
AIC**	305.656	
Log likelihood	-147.828	
Pseudo R-squared	0.495	

Spatial Lag Regression Model		
Variable Name	**Coefficient**	**p-value**
Median Income	-0.000	0.862
Tree Canopy	-0.058	0.000*
Impervious Surfaces	-0.001	0.934
Population Density	3.083	0.318
W_y	0.606	0.000*
AIC**	322.950	
Log likelihood	-155.475	
Pseudo R-squared	0.734	
Spatial Pseudo R-squared	0.539	

Table 5. Model Summary for Detroit:

** Indicates significant Correlation Coefficients*

Ordinary Least Squares Model					
Variable Name	**Coefficient**	**p-value**	**R-squared**	**AIC****	**Log likelihood**
Median Income	5.005e-06	0.109	0.012	173.8	-84.903
Tree Canopy	-0.112	0.002*	0.046	166.3	-81.131
Impervious Surfaces	0.015	0.000*	0.081	158.2	-77.112
Population Density	7.316	0.012*	0.030	169.9	-82.975

Spatial Error Regression Model		
Variable Name	**Coefficient**	**p-value**
Median Income	-0.000	0.332
Tree Canopy	-0.048	0.149
Impervious Surfaces	0.013	0.000*
Population Density	1.3634	0.488
lambda	0.850	0.000*
AIC**	-29.229	
Log likelihood	19.614	
Pseudo R-squared	0.081	

Spatial Lag Regression Model		
Variable Name	**Coefficient**	**p-value**
Median Income	-0.000	0.998
Tree Canopy	0.006	0.840
Impervious Surfaces	0.013	0.000*
Population Density	1.823	0.309
W_y	0.822	0.000*
AIC**	-18.826	
Log likelihood	15.413	
Pseudo R-squared	0.6719	
Spatial Pseudo R-squared	0.0831	

Table 6. Model Summary for El Paso:

** Indicates significant Correlation Coefficients*

Ordinary Least Squares Model					
Variable Name	**Coefficient**	**p-value**	**R-squared**	**AIC****	**Log likelihood**
Median Income	-0.001	0.192	0.020	44.43	-20.215
Tree Canopy	0.026	0.484	0.006	45.68	-20.841
Impervious Surfaces	0.002	0.613	0.003	45.92	-20.961
Population Density	18.993	0.765	0.001	46.09	-21.047

Spatial Error Regression Model		
Variable Name	**Coefficient**	**p-value**
Median Income	-0.003	0.005*
Tree Canopy	0.060	0.069
Impervious Surfaces	0.014	0.002*
Population Density	76.674	0.188
lambda	0.698	0.000*
AIC**	15.598	
Log likelihood	-2.799	
Pseudo R-squared	0.030	

Spatial Lag Regression Model		
Variable Name	**Coefficient**	**p-value**
Median Income	-0.002	0.027*
Tree Canopy	0.042	0.213
Impervious Surfaces	0.007	0.040*
Population Density	78.112	0.175
W_y	0.648	0.000*
AIC**	25.829	
Log likelihood	-6.915	
Pseudo R-squared	0.381	
Spatial Pseudo R-squared	0.000	

Table 7. Model Summary for Miami:

** Indicates significant Correlation Coefficients*

Ordinary Least Squares Model					
Variable Name	**Coefficient**	**p-value**	**R-squared**	**AIC****	**Log likelihood**
Median Income	-1.087	0.012*	0.080	171.5	-83.773
Tree Canopy	0.130	0.347	0.011	177.2	-86.596
Impervious Surfaces	-0.150	0.267	0.016	176.8	-86.417
Population Density	-0.786	0.058	0.046	174.4	-85.200

Spatial Error Regression Model		
Variable Name	**Coefficient**	**p-value**
Median Income	-0.200	0.467
Tree Canopy	-0.076	0.507
Impervious Surfaces	0.227	0.054*
Population Density	-0.397	0.073
lambda	0.887	0.000*
AIC**	77.027	
Log likelihood	-33.514	
Pseudo R-squared	0.001	

Spatial Lag Regression Model		
Variable Name	**Coefficient**	**p-value**
Median Income	-0.260	0.245
Tree Canopy	-0.102	0.380
Impervious Surfaces	0.051	0.648
Population Density	-0.316	0.151
W_y	0.876	0.000*
AIC**	85.433	
Log likelihood	-36.717	
Pseudo R-squared	0.796	
Spatial Pseudo R-squared	0.001	

Ordinary Least Squares Model

The OLS model exhibited the highest AIC value among all four cities (DC: 398.6 – 518.1, Detroit: 158.2 – 173.8, El Paso: 44.4 – 46.0, Miami: 171.5 – 177.2). Additionally, the log likelihood values were the lowest in the model (DC: -257.0 – -248.8, Detroit: -84.9 – -77.1, El Paso: -21.0 - -20.2, Miami: -85.2 – -83.8). The R-squared values were also relatively low, suggesting poor linearity or non-linear datasets. This observation is further supported by the spatial autocorrelation test, which revealed spatial dependence in the residuals. Overall, these diagnostic tests provide evidence that the OLS model is a poorer fit compared to both the SER and the SLR.

The SER model exhibited the lowest AIC value among the three (D.C: 305.656, El Paso: 15.598, Miami: 77.027) out of four cities. This is also reflected in their log likelihood values (DC: -147.828, El Paso: -6.91, Miami: -36.717). The lambda value for all four cities came back significantly positive, reinforcing a better fit of a model which considers spatial dependence. The pseudo-R-squared values range from approximately <0.01 – 0.49.

The SLR model exhibited the lowest AIC value in one out of four cities (Detroit: -18.826). This is also paralleled by its log likelihood value (15.413). The pseudo-R-squared values for all four cities ranged from 0.381 – 0.796, whereas the spatial pseudo-R-squared values ranged from < 0.01 – 0.539. The W_y is highly significantly positive across all four cities, indicating the presence of spatial spillover effects, where the value of the dependent variable at one location is influenced by the values of neighboring locations. This again provides evidence for the need of a model that considers spatial dependence.

Spatial Error Model and Spatial Lag Model

For the District of Columbia (Table 4), the SER model shows Tree Canopy and Impervious Surfaces as the only two variables that were significant predictors of afternoon temperature anomalies for the climate type of C - Temperate. Tree Canopy was slightly negatively correlated with dependent variable (coefficient: -0.051, p-value: 0.000), while Impervious Surfaces was slightly positively correlated with the dependent variable (coefficient: 0.021, p-value: 0.002). Tree Canopy showed a stronger significance than Impervious Surfaces. Median Income (coefficient: 0.000, p-value: 0.968) and Population Density (coefficient: -0.466, p-value: 0.890) fail to attain statistical significance as predictors of afternoon temperature anomalies in the SER model for the District of Columbia. The pseudo- R-squared values (0.495) report that both Tree Canopy and Impervious Surfaces accounted for 49.5% of the variation in afternoon temperature anomalies, leaving 50.5% of the variation unaccounted for.

For Detroit (Table 5), the SLR model indicates that Impervious Surfaces is the only variable that is a significant predictor of afternoon temperature anomalies for the climate type of D – Continental. Impervious Surfaces showed a slightly positive correlation (coefficient: 0.013, p-value: 0.000) with afternoon temperature anomalies. The remaining variables, Median Income (coefficient: -0.000, p-value: 0.998), Tree Canopy (coefficient: 0.006, p-value: 0.840), and Population Density (coefficient: 1.823, p-value: 0.309), are not statistically significant. Furthermore, the spatially lagged dependent variable (W_y) showed a significantly positive correlation (coefficient: 0.822, p-value: 0.000) suggesting the presence of spatial dependence from neighboring locations. The pseudo-R-squared value indicates that Impervious Surfaces, explains 69.17% of the variation in afternoon temperature anomalies. However, the Spatial pseudo-R-squared value (0.0831) suggests that Impervious Surfaces only explain for 8.31% of the variation in afternoon temperature anomalies once the spillover effect is accounted for.

For El Paso (Table 6), the SER shows that Median Income and Impervious Surfaces as the only two variables that were significant predictors of afternoon temperature anomalies for the climate type of B – Arid. Median Income was slightly negatively correlated (coefficient: -0.002, p-value: 0.040) with the dependent variable while Impervious Surfaces with slightly positively correlated (coefficient: 0.007, p-value: 0.027) with the dependent variable. Impervious surfaces had a more significant correlation with the dependent variable compared to Median Income. The pseudo- R-squared value (0.030) indicates that, collectively, Median Income and Impervious Surfaces explain 3.00% of the variation in afternoon temperature anomalies, leaving 97% unexplained.

For Miami (Table 7), the SER shows Impervious Surfaces as the only variable that was a significant predictor of afternoon temperature anomalies for the climate type of A - Tropical. Impervious Surfaces had a slightly positive correlation (coefficient: 0.227, p-value: 0.054) with the dependent variable. The Pseudo-R-squared value (<0.01) indicates that Impervious Surfaces explained < 1% of the variation in the data, leaving approximately 99.99 % unaccounted for.

All four cities, with their respective models, had highly significant positive spatial autocorrelation present (DC – lambda: 0.773, p-value: 0.000; Detroit – W_y: 0.822, p-value: 0.000; El Paso – lambda: 0.698, p-value: 0.000; Miami – lambda: 0.887, p-value: 0.000). This indicates that the SER and the SL models are capturing a degree of the spatial dependence.

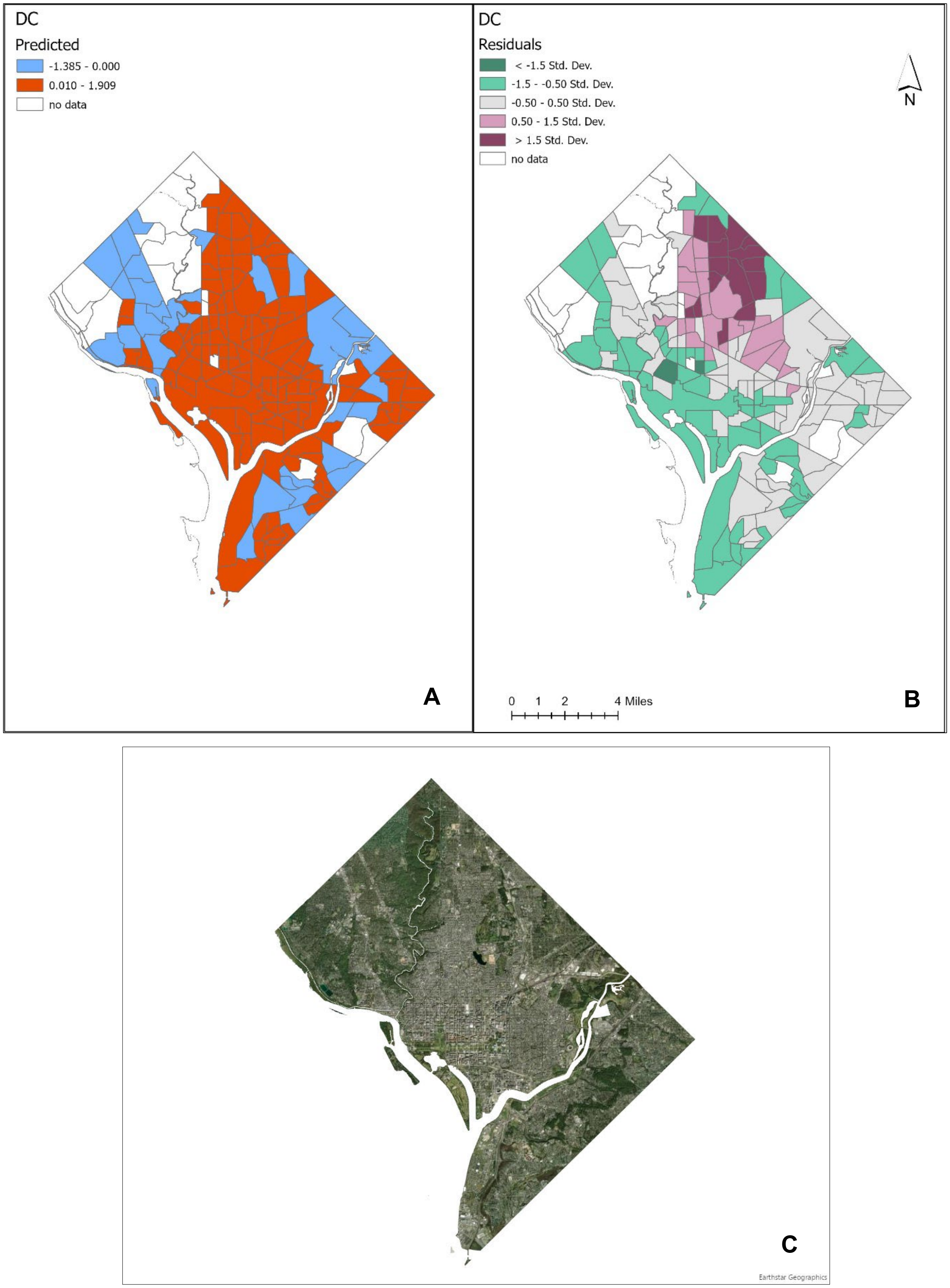


Figure 3. Predicted vs Residual Maps for the District of Columbia: Visual representations include (A) predicted neighborhood afternoon temperature anomalies (°F), (B) differences between predicted and observed values, and (C) satellite imagery of the area for reference. Satellite imagery retrieved from Esri's World Imagery basemap (27).

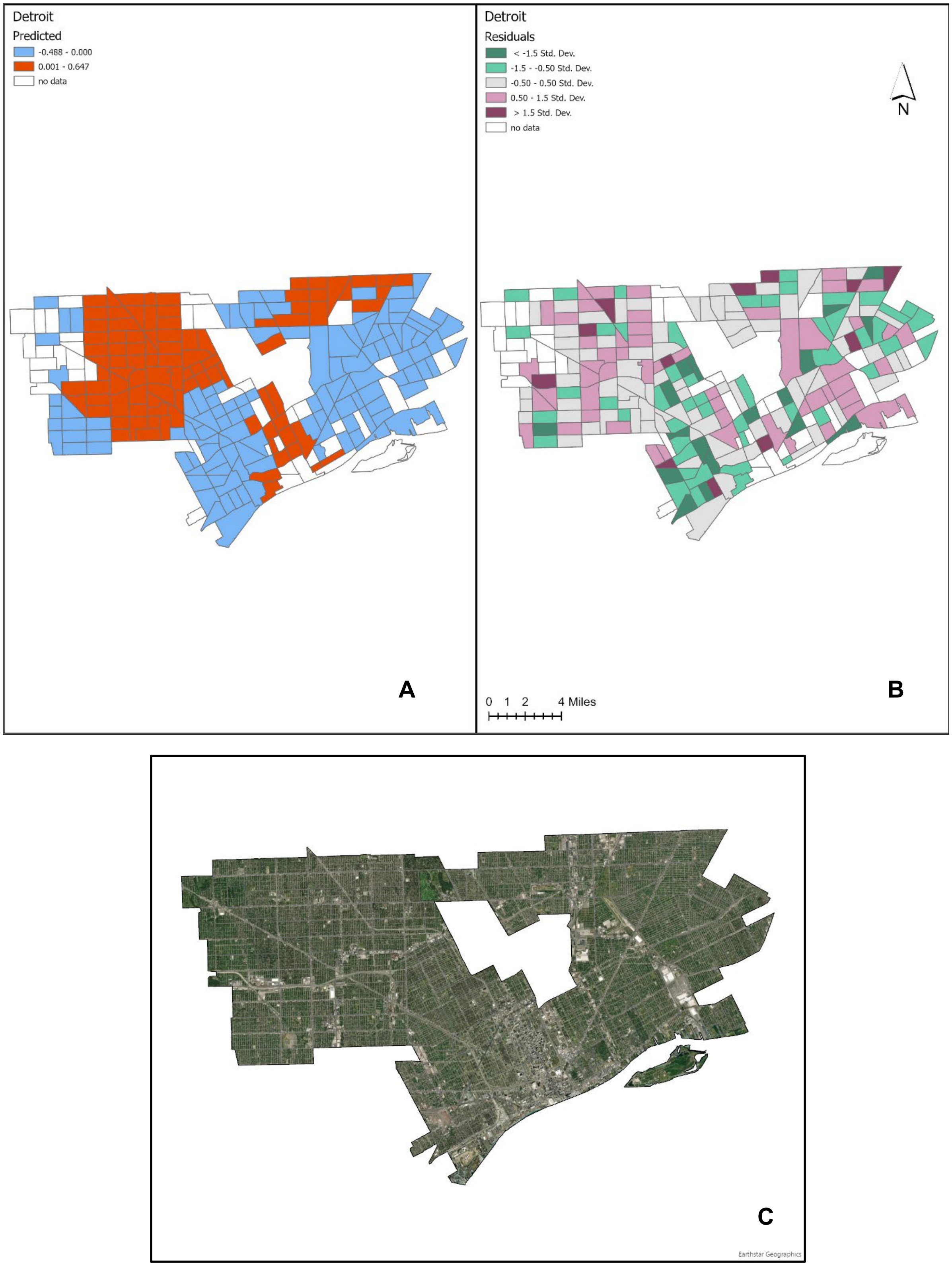


Figure 4. Predicted vs Residual Maps for the Detroit, Michigan: Visual representations include (A) predicted neighborhood afternoon temperature anomalies (°F), (B) differences between predicted and observed values, and (C) satellite imagery of the area for reference. Satellite imagery retrieved from Esri's World Imagery basemap (27).

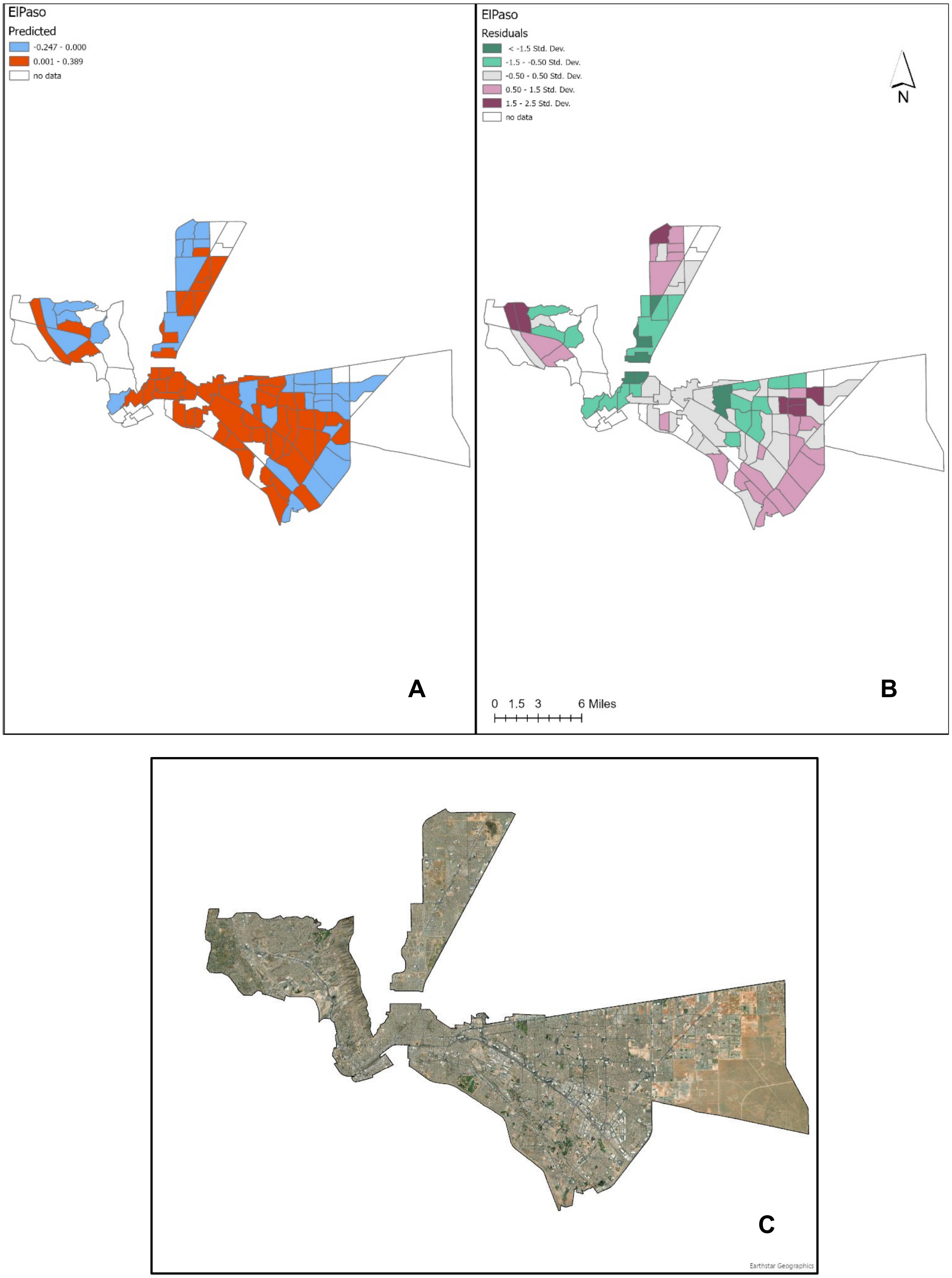
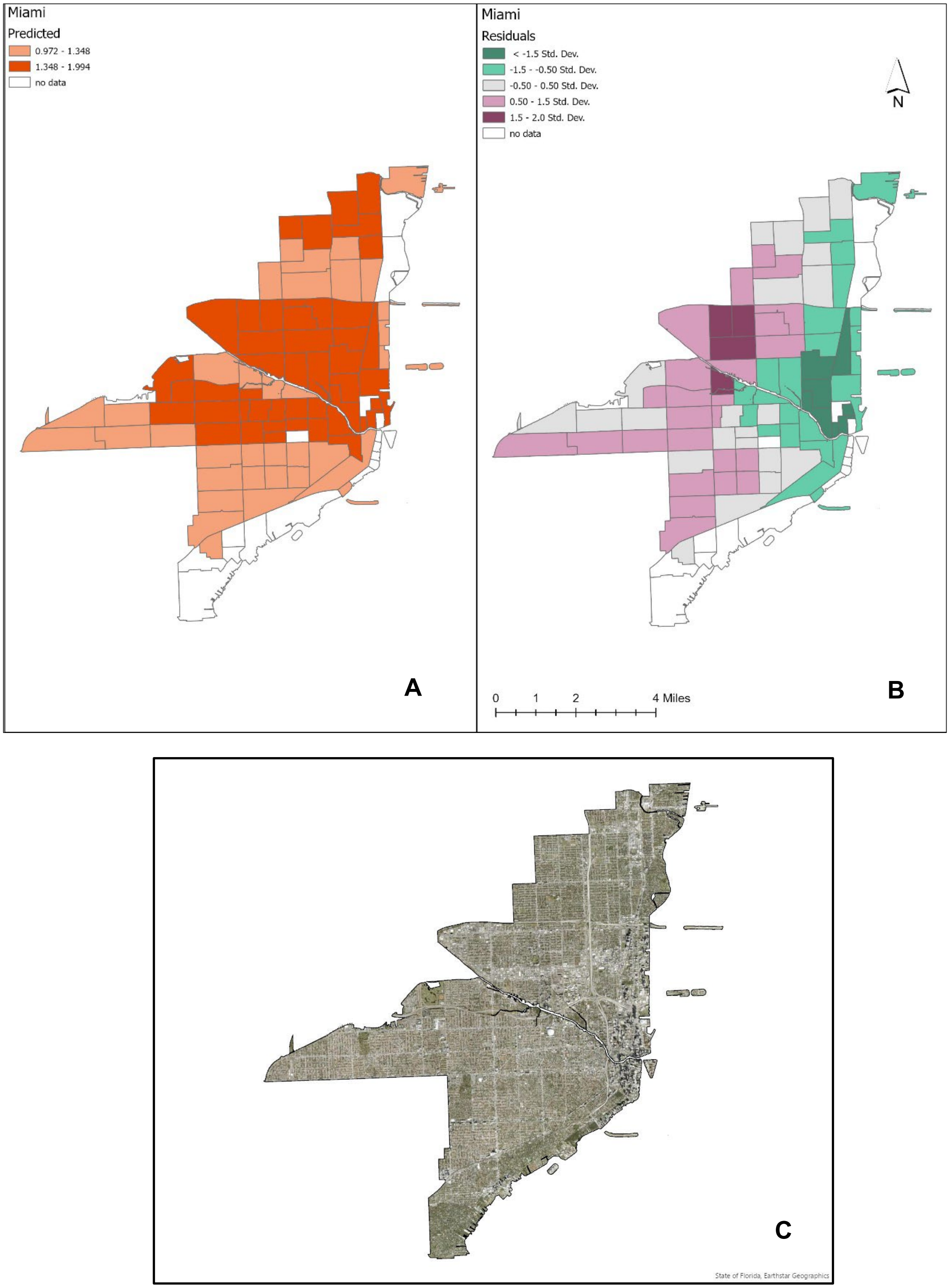


Figure 5. Predicted vs Residual Maps for the El Paso, Texas: Visual representations include (A) predicted neighborhood afternoon temperature anomalies (°F), (B) differences between predicted and observed values, and (C) satellite imagery of the area for reference. Satellite imagery retrieved from Esri's World Imagery basemap (27).



State of Florida, Earthstar Geographics

Figure 6. Predicted vs Residual Maps for the Miami, Florida: Visual representations include (A) predicted neighborhood afternoon temperature anomalies (°F), (B) differences between predicted and observed values, and (C) satellite imagery of the area for reference. Satellite imagery retrieved from Esri's World Imagery basemap (27).

Predicted vs. Residuals

According to the predicted map of Washington, D.C. (Figure 3), the SER model anticipates higher temperature anomalies (depicted by red shades) concentrated in the central downtown area, near Lincoln Heights and near the Anacostia Naval Station and Airforce Base. Lower temperature anomalies (depicted by blue shades) are observed in outlying regions, particularly around notable parks such as the National Arboretum, Rock Creek Park, and Dupont Park. Analyzing the residual maps reveals clusters of over-predictions (depicted by dark burgundy shades) in the northern neighborhoods near Fort Totten and Manor Park. Conversely, there are under-predictions (depicted by dark green shades) in the eastern side of the city near the neighborhood of Sheridan-Kalorama. Furthermore, there are clusters of slightly underpredicted areas (depicted by light green shades) on the eastern and southeastern side of the city, near downtown and in proximity to the Anacostia Naval Station and Air Force Base.

Based on the predicted map of Detroit (Figure 4), the SLR model indicates a concentration of elevated temperature anomalies in the upper west side of the city, particularly around districts 6 and 7. Conversely, clusters of lower temperature anomalies are observed on the east side of the city. Both the high and low anomaly clusters are situated outside the primary downtown district. Analyzing the residual map of Detroit, there are small clusters of slightly overpredicted areas (depicted by light burgundy shades) on the western and southeastern side of the city, while there are small clusters of underpredictions on the southwestern and eastern side of the city.

Based on the predicted map of El Paso (Figure 5), the SER model estimates a cluster of elevated temperature anomalies around central El Paso, near the Mexican border. There are also smaller clusters of lower temperature anomalies on the northern side of the city near Franklin Mountains State Park and on the western side of the city. Areas of lower predicted temperature anomalies are on the outskirts of the city and/or at the edge of the state park. Analyzing the residual maps reveals clusters of overpredictions away from central El Paso on the edges of the city boundary. Conversely, clusters of underpredictions are seen toward the middle of the city, near Franklin Mountain State Park, and next to El Paso's International Airport.

According to the predicted map of Miami (Figure 6), the SER model only estimates elevated temperature anomalies. The model estimates clusters of elevated temperature anomalies near the neighborhoods of downtown Miami, Little Havana, Civic Center, Allapattah, and Lemon City. The model also estimates clusters of slightly elevated temperature anomalies (depicted in orange shades) near the neighborhoods of Coral Way and Liberty City. Furthermore, there are areas of slightly elevated temperature anomalies on islands, such as Biscayne and San Marco islands. Analyzing the residual maps reveals clusters of overpredictions away from downtown Miami, while there are clusters of underpredictions toward downtown Miami, near the beaches.

All four residual maps have some kind of spatial pattern (clustering) present, as the lambda values all returned highly significant (p-value: 0.000).

6. Discussion

This project aimed to identify and explore key variables associated with UHI intensity across four distinct cities in the U.S using regression analyses. These four cities were chosen based on their climate types, each representing the climate of C – Temperate, D – Continental, B – Arid, or A – Tropical. EDA was performed to determine an appropriate statistical model. Overall, the SER and SL models proved to be better model fits compared to the OLS model. Impervious Surfaces was significant in three out of the four cities, while Tree Canopy and Median Income were significant in only one city. Population Density was not a significant predictor of UHI intensity in any of the cities. Finally, the presence of spatial clustering in the residuals indicates that the model is capturing a degree of the spatial dependence but suggests that there are additional spatial dependencies not fully accounted for by the models.

In Washington, D.C., the analysis suggests that increasing greenspace and reducing impervious surfaces may be an effective way to mitigate Urban Heat Island (UHI) intensity. This finding aligns with the understanding that green spaces provide natural cooling through shade and evapotranspiration, while impervious surfaces absorb and radiate heat, exacerbating UHI effects (3, 11).

However, tree canopy coverage did not significantly predict UHI intensity in El Paso, Detroit, or Miami. This lack of significance could be due to various factors, such as the local climate type or the type of tree species available. For example, in the arid climate of El Paso, tree canopy coverage might not provide as much cooling effect as in more temperate regions. The native cacti species thrive in the desert-like climate, and canopy coverage from these plants may not be as effective at cooling the surrounding air compared to deciduous tree species found on the East Coast. Yet, in Detroit and Miami, where deciduous trees are more common, other factors may have a more significant impact on UHI mitigation.

In El Paso, Median Income came out as a significant factor associated with lower temperature anomalies. This finding suggests that gentrification may mitigate the effects of UHI in this region, or it may align with current studies that found communities of low income and of color are disproportionately exposed to UHI (15, 16, 17). The impacts of past redlining practices may still be evident in El Paso.

Although some variables were significant predictors of afternoon temperature anomalies, they did not account for most of the observations. There is considerable unexplained noise in the dataset, indicating that afternoon temperature anomalies may be unpredictable or that key predictor variables may be missing from the model. It's likely that variables such as tourism, land elevation, proximity to large bodies of water, and others could serve as additional indicators of UHI intensity. For example, El Paso, Texas, is situated next to Ciudad Juárez, touching the border of the United States and Mexico. Thus, factors such as the movement of people across the border or conditions in the neighboring city could impact surrounding air temperatures. To reduce the data noise, temperature readings over a longer period of time should be used for more accurate predictions. Further research into potential predictor variables should be conducted for each city before performing additional regression analyses.

Bibliography

[1] Oke, T. R. (1973). City size and the urban heat island. *Atmospheric Environment (1967)*, 7(8), 769–779. https://doi.org/10.1016/0004-6981(73)90140-6

[2] U.S. Environmental Protection Agency. (2008). Reducing urban heat islands: Compendium of strategies. https://www.epa.gov/heat-islands/heat-island-compendium

[3] Shi, Z., Li, X., Hu, T., Yuan, B., Yin, P., & Jiang, D. (2023). Modeling the intensity of surface urban heat island based on the impervious surface area. *Urban Climate, 49*, 101529. https://doi.org/10.1016/j.uclim.2023.101529

[4] Chithra, S. V., Nair, M. H., Amarnath, A., & Anjana, N. S. (2015). Impacts of impervious surfaces on the environment. *International Journal of Engineering Science Invention, 4*(5), 27-31.

[5] Hua, L., Zhang, X., Nie, Q., Sun, F., & Tang, L. (2020). The impacts of the expansion of urban impervious surfaces on urban heat islands in a coastal city in china. *Sustainability, 12*(2), 475. https://doi.org/10.3390/su12020475

[6] Vujovic, S., Haddad, B., Karaky, H., Sebaibi, N., & Boutouil, M. (2021). Urban heat island: Causes, consequences, and mitigation measures with emphasis on reflective and permeable pavements. *CivilEng, 2*(2), 459–484. https://doi.org/10.3390/civileng2020026

[7] Ziter, C. D., Pedersen, E. J., Kucharik, C. J., & Turner, M. G. (2019). Scale-dependent interactions between tree canopy cover and impervious surfaces reduce daytime urban heat during summer. *Proceedings of the National Academy of Sciences, 116*(15), 7575–7580. https://doi.org/10.1073/pnas.1817561116

[8] Ibrahim, S. H., Ibrahim, N. I. A., Wahid, J., Goh, N. A., Koesmeri, D. R. A., & Nawi, M. N. M. (2018). The impact of road pavement on Urban Heat Island (UHI) phenomenon. *International Journal of Technology, 9*(8), 1597-1608.

[9] Yang, C.-C., Siao, J.-H., Yeh, W.-C., & Wang, Y.-M. (2021). A study on heat storage and dissipation efficiency at permeable road pavements. *Materials, 14*(12), 3431. https://doi.org/10.3390/ma14123431

[10] Shin, M. H., Kim, H. Y., Gu, D., & Kim, H. (2017). Leed, its efficacy and fallacy in a regional context—An urban heat island case in California. *Sustainability, 9*(9), 1674. https://doi.org/10.3390/su9091674

[11] Lai, D. Y., Leung, T. M., & Cheung, S. H. (2021). The cooling effect of vegetation on microclimate in a high-density urban environment: A review of the modeling and experimental studies. *Environmental Evidence, 10*(1), 18. https://doi.org/10.1186/s13750-021-00226-y

[12] District Department of the Environment. (2013). Assessing the Health Impacts of Urban Heat Island Reduction Strategies in the District of Columbia. *Department of Energy and Environment*. https://ddoe.dc.gov/sites/default/files/dc/sites/ddoe/publication/attachments/20131021_Urban%20Heat%20Island%20Study_FINAL.pdf

[13] Ramírez-Aguilar, E. A., & Lucas Souza, L. C. (2019). Urban form and population density: Influences on urban heat island intensities in Bogotá, Colombia. *Urban Climate, 29*, 100497. https://doi.org/10.1016/j.uclim.2019.100497

[14] Zhou, B., Rybski, D., & Kropp, J. P. (2017). The role of city size and urban form in the surface urban heat island. *Scientific Reports, 7*(1), 4791. https://doi.org/10.1038/s41598-017-04242-2

[15] Hsu, A., Sheriff, G., Chakraborty, T., & Manya, D. (2021). Disproportionate exposure to urban heat island intensity across major US cities. *Nature Communications, 12*(1), 2721. https://doi.org/10.1038/s41467-021-22799-5

[16] Johnson, D. P. (2022). Population-based disparities in u. S. Urban heat exposure from 2003 to 2018. *International Journal of Environmental Research and Public Health, 19*(19), 12314. https://doi.org/10.3390/ijerph191912314

[17] Saverino, K. C., Routman, E., Lookingbill, T. R., Eanes, A. M., Hoffman, J. S., & Bao, R. (2021). Thermal inequity in Richmond, VA: The effect of an unjust evolution of the urban landscape on urban heat islands. *Sustainability, 13*(3), 1511. https://doi.org/10.3390/su13031511

[18] Chen, Y.-L., & Wang, J.-J. (1995). The effects of precipitation on the surface temperature and airflow over the island of hawaii. *Monthly Weather Review, 123*(3), 681–694. https://doi.org/10.1175/1520-0493(1995)123<0681:TEOPOT>2.0.CO;2

[19] Zhou, D., Zhao, S., Liu, S., Zhang, L., & Zhu, C. (2014). Surface urban heat island in China's 32 major cities: Spatial patterns and drivers. *Remote Sensing of Environment, 152*, 51–61. https://doi.org/10.1016/j.rse.2014.05.017

[20] Hardin, A. W., Liu, Y., Cao, G., & Vanos, J. K. (2018). Urban heat island intensity and spatial variability by synoptic weather type in the northeast U.S. *Urban Climate, 24*, 747–762. https://doi.org/10.1016/j.uclim.2017.09.001

[21] Yang, P., Ren, G., & Hou, W. (2019). Impact of daytime precipitation duration on urban heat island intensity over Beijing city. *Urban Climate, 28*, 100463. https://doi.org/10.1016/j.uclim.2019.100463

[22] Varquez, A. C. G., & Kanda, M. (2018). Global urban climatology: A meta-analysis of air temperature trends (1960–2009). *Npj Climate and Atmospheric Science, 1*(1), 1–8. https://doi.org/10.1038/s41612-018-0042-8

[23] Urban Heat Island mapping campaign cities. (2021, August 25). *ArcGIS StoryMaps.* https://local.storymapsdev.arcgis.com:3443/stories/c57435110d264e1f8ccd75e35731ed5f

[24] Geiger, R. (1961). Überarbeitete Neuausgabe von Geiger, R.: Köppen-Geiger / Klima der Erde. (Wandkarte 1:16 Mill.) – Klett-Perthes, Gotha.

[25] Ranga Vamsi. (2020, July 8). *K-nearest neighbors algorithm in Python. Medium.* *https://medium.com/@rangavamsi5/k-nearest-neighbors-algorithm-in-python-64f38792193*

[26] United States Environmental Protection Agency. (2014, June 17). *Learn about heat islands*. Retrieved from https://www.epa.gov/heatislands/learn-about-heat-islands

[27] Esri. (2009, December 12). *World Imagery* [Basemap]. Retrieved April 20, 2024, from https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9.