GIS – Part 3

Introduction

The demographics of Washington, DC over the last 60 years are characterized by great change. In the 1950s, “white flight” combined with an influx of African Americans to shift the city to majority African American by 1960 (Census Bureau 2000). After population declines and poverty increases from 1970 – 2000, the situation changed in the late 1990s. During the 21st century DC’s population has been increasing with the white population growing significantly.

While all city wards experienced positive growth between 2000-2008, it was not all equal (Comey et al 2010). Racial populations within DC are deeply segregated, which in turn means that the racialized aspect of wealth creates geographic disparities (Farrell 2017).

As wealthier white populations return to Washington, DC, areas of the city are beginning to experience gentrification. Gentrification is a process that alters communities, and for better or worse, is a constant throughout the history of cities. There is no agreed upon definition of gentrification as it is often a combination of factors and indicators. However, based on research by Yan Lee, incomes can be a strong indicator as many gentrification indicators, such as housing prices, education attained, and cars per household, are often proxies for income (2010). Although this approach to income as an indicator has been criticized (Papachristos et al 2011), it is useful in this study as the data is accessible.

This study will use regression to investigate the relationship between diversity and household income in census tracts throughout Washington, DC. If a relationship exists it could be of use for city planners as they face the positive, and negative, effects of gentrification. Two examples of regression analysis performed on gentrification factors with racial variables include Freeman and Braconi (2004) in New York City and Hwang and Simpson (2014) in Chicago. This suggests that it is a viable analysis tool to identify this sort of relationship. A further study looked at the effects of gentrification on racial composition in Washington, DC and defined gentrification as “an influx of affluent residents” (Jackson 2014, pg 359). The study did not explicitly look at the relationship between diversity and income per household, though, as gentrification was a combination of factors with income being just one.

Research Question and Hypothesis

The research question posed is the following:

- Is there a relationship between diversity index score and change in the median federal adjusted gross income by census tract?

Due to the historical racial segmentation of Washington, DC into black and white neighborhoods, it is suspected that an increase in diversity score may indicate a positive increase in the change in household income.

As such, our null and alternative hypotheses are the following:

- H_0 (Null hypothesis): There is no relationship between increasing diversity scores and change in median income.

- H_1 (Alternative hypothesis): As diversity score increases we will see a positive change in median income.

Data

The data in this study is from Open Data DC, a robust open database of hundreds of datasets related to Washington, DC. There are several administrative boundaries of DC. At the highest level is Ward data, followed by Neighborhood Clusters, Advisory Neighborhood Commissions (ANCs), and census tracts.

Figure 1: DC Ward boundaries

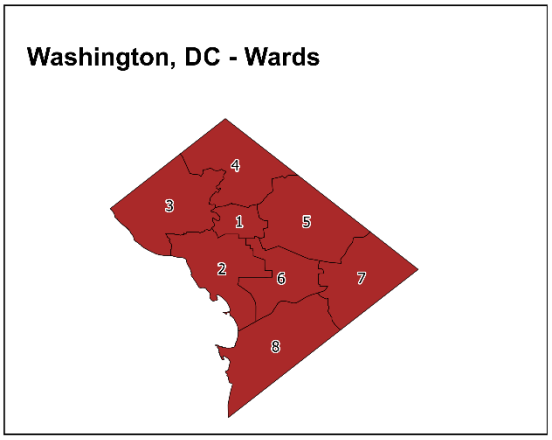


Figure 2: DC Neighborhood Cluster boundaries

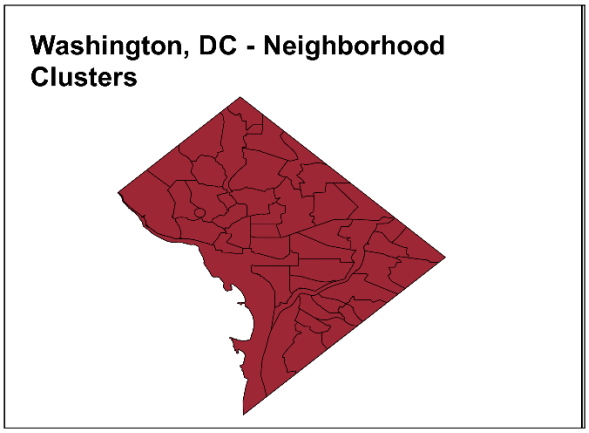


Figure 3: DC ANC boundaries

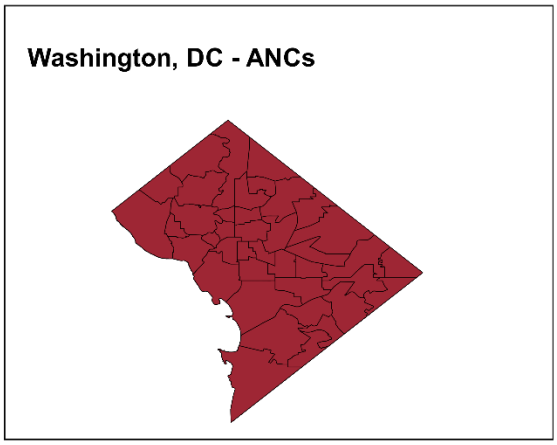
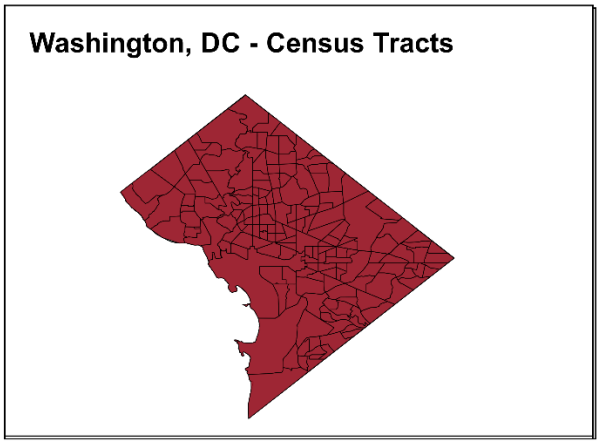


Figure 4: DC Census Tract boundaries



This investigation used census tracts as our study level. Ward, Neighborhood Cluster, and ANC boundaries often change after census' making comparisons difficult and they often have overlapping areas. Further, the census tracts are granular (179 distinct areas) compared to Wards (8), Neighborhood Clusters (39), and ANCs (37).

Data from the 2010 census and American Community Survey (ACS) 2015 5-year estimates was used for this investigation. ACS is an ongoing survey administered by the Census Bureau and provides estimates for the years between census'. Although there is a margin of error in these estimates, the data is sufficient for the purposes of this study. The change in income over time was calculated by using the 2010 census data as a baseline and the ACS 2015 estimates as an endline. The population and race data are from the ACS 2015 dataset.

Two census tracts were removed from our analysis. Tract 6202 was removed because it corresponds to government buildings (including the White House) and National Mall with an extremely low population count that is lower than the ACS margin of error. Tract 6804 was also removed as it corresponds to a tract of federal land that contains a prison and no households.

Methodology

To establish if a result is statistically significant, alpha (α) = 0.05 for this study.

The first step was to transform the data to fit our needs. This involved accumulating the race categories per census tract in 2015 and finding the change in household income from 2010 to 2015. A view of this transformation can be seen in the below tables:

Figure 5: Race population categories

TRACT	Total_Pop	WHITE	BLACK	AMERIND	ASIAN	HAWA	OTHER	MultRace	%White	%Black	%AmerInd	%Asian	%Hawa	%Other	%Multrace
100	5209	4784	66	12	256	0	53	38	91.84%	1.27%	0.23%	4.91%	0.00%	1.02%	0.73%
201	3685	2653	295	0	508	0	33	196	71.99%	8.01%	0.00%	13.79%	0.00%	0.90%	5.32%
202	4817	4082	358	8	221	0	24	124	84.74%	7.43%	0.17%	4.59%	0.00%	0.50%	2.57%
300	6095	5149	363	13	346	0	24	200	84.48%	5.96%	0.21%	5.68%	0.00%	0.39%	3.28%
400	1578	1265	89	0	130	0	21	73	80.16%	5.64%	0.00%	8.24%	0.00%	1.33%	4.63%
501	3502	2855	155	87	149	0	103	153	81.52%	4.43%	2.48%	4.25%	0.00%	2.94%	4.37%
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
11100	5490	660	4737	0	13	0	14	66	12.02%	86.28%	0.00%	0.24%	0.00%	0.26%	1.20%

The seven broad race categories from the ACS data were maintained – White, Black, American Indian, Asian, Hawaiian/Pacific Islander, Other, and Multiracial (two or more races).

Figure 6: Median income by year categories

TRACT	$MedInc_2010	$MedInc_2011	$MedInc_2012	$MedInc_2013	$MedInc_2014	$MedInc_2015	$Change_2010-15
100	$89,357	$88,250	$92,634	$86,102	$108,308	$110,945	$21,588
201	$12,202		$21,740	$23,000	$5,087	$8,785	-$3,417
202	$116,867	$109,461	$106,470	$95,034	$114,630	$115,586	-$1,281
300	$66,915	$58,960	$61,937	$58,822	$73,252	$70,413	$3,499
400	$104,934	$101,766	$102,512	$110,621	$130,924	$121,711	$16,777
501	$65,919	$64,764	$70,218	$68,973	$72,707	$76,950	$11,031
…	$76,777	$76,905	$76,609	$75,829	$90,719	$89,334	$12,557
11100	$87,171	$84,947	$91,538	$81,529	$98,903	$103,485	$16,314

To find the change in median income, we simply subtracted the income from 2015 to 2010. Any negative number indicates a decrease in household income over that time.

The second step to answering the question is calculating the diversity index score of each census tract. The diversity score was calculated using the Diversity Index equation:

$$\text{Diversity index} = 1 - \left(\frac{\sum_j (n_{ij}(n_{ij}-1))}{N_i(N_i-1)} \right)$$

- n_{ij} is the total number of people in area *i* who are classified in race *j*
- N_{ij} is the total number of people across all races *j* in area *i*

The Diversity Index reports "the probability that two randomly selected people in a particular area would be from different ethnic groups" (Stillwell and Ham 2010). The index ranges between 0 and 1. For our study, we multiplied the diversity index by 100 to give us a score between 0 and 100. High values indicate greater racial diversity while low values indicate homogeneity. This index will work with numerous variables, but this study looked explicitly at the ACS data categories of race – White, Black, American Indian, Asian, Hawaiian/Pacific Islander, Other, and Multiracial – to calculate the score.

In addition to finding the diversity score, a variable of "Diversity Score Level" was created. Each tract was thus categorized as "Low", "Medium", or "High" based on the quantile breaks of the diversity scores. Appendix one shows how the diversity score and level was calculated.

Once the diversity score is calculated, a simple linear regression is performed. The dependent variable is the change in median household income. The independent variable is the diversity score. This will show us if a statistical relationship exists between the two variables. Appendix two shows the regression analysis and Moran's I test.

Results

All maps were produced using QGIS. The transformed data tables (Figures 5-6) were joined with the administrative boundary shapefiles taken from Open Data DC by census tract ID.

Figure 7: Diversity by tract

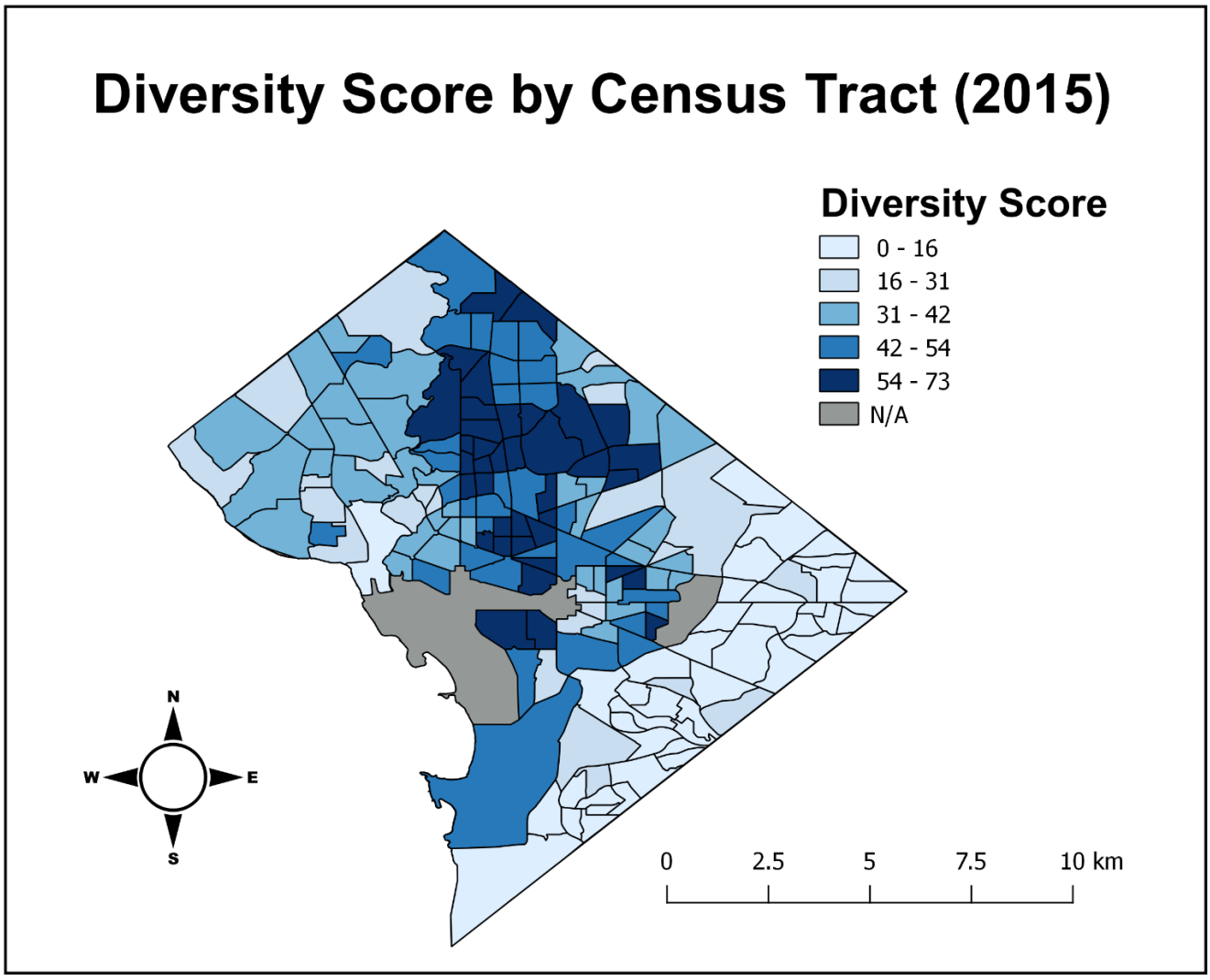


Figure 8: Diversity by tract with Ward layer

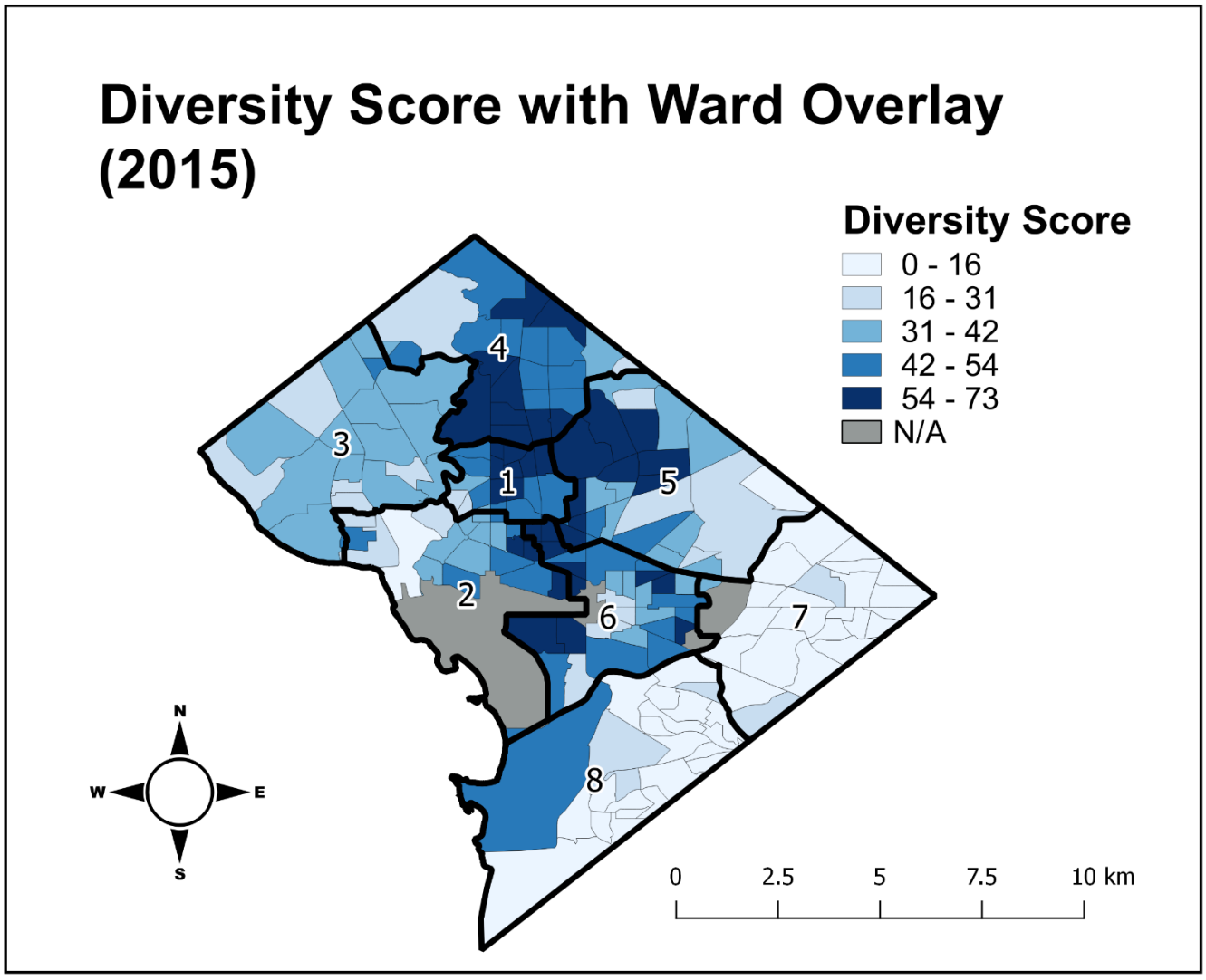


Figure 9: Diversity score level by tract

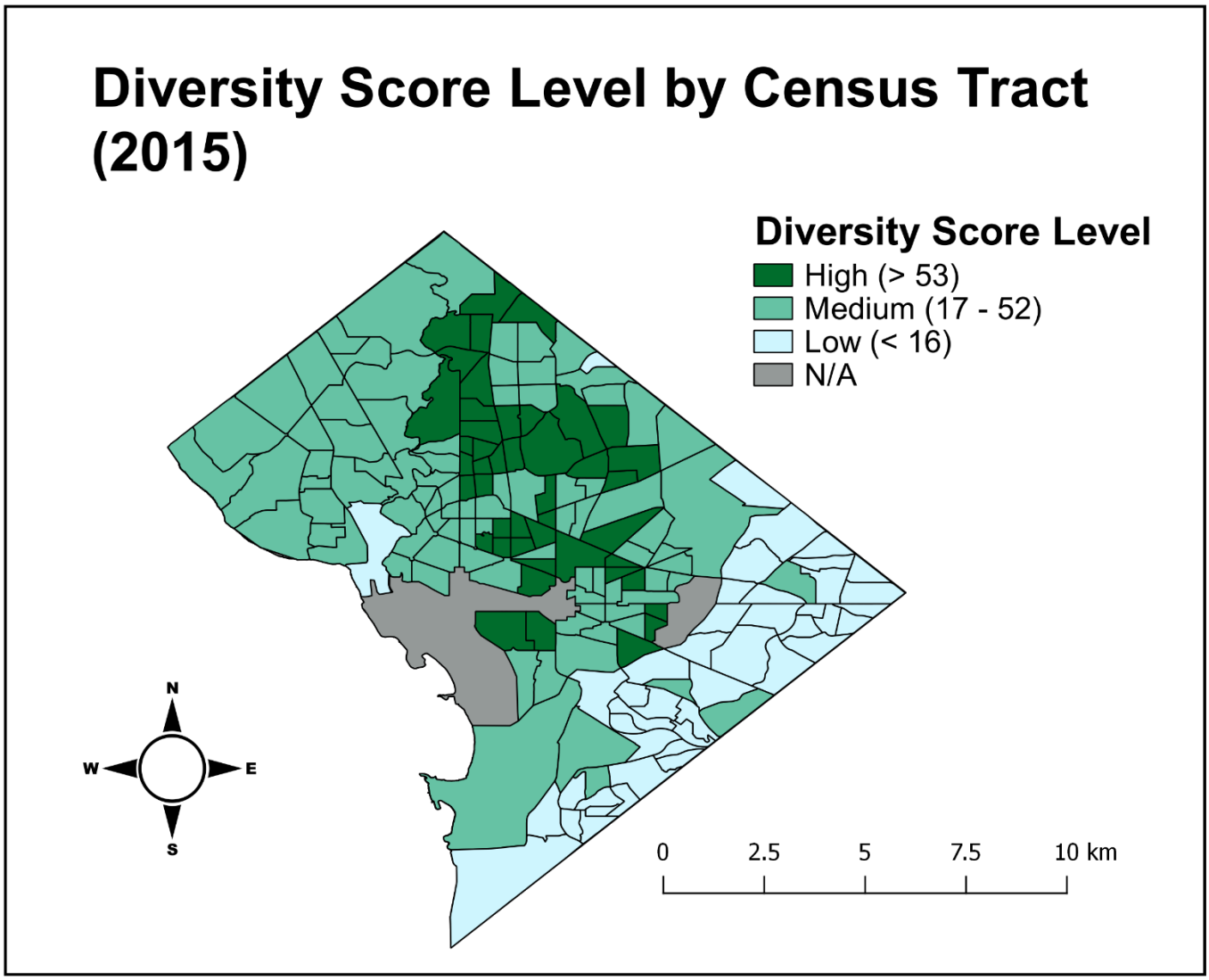
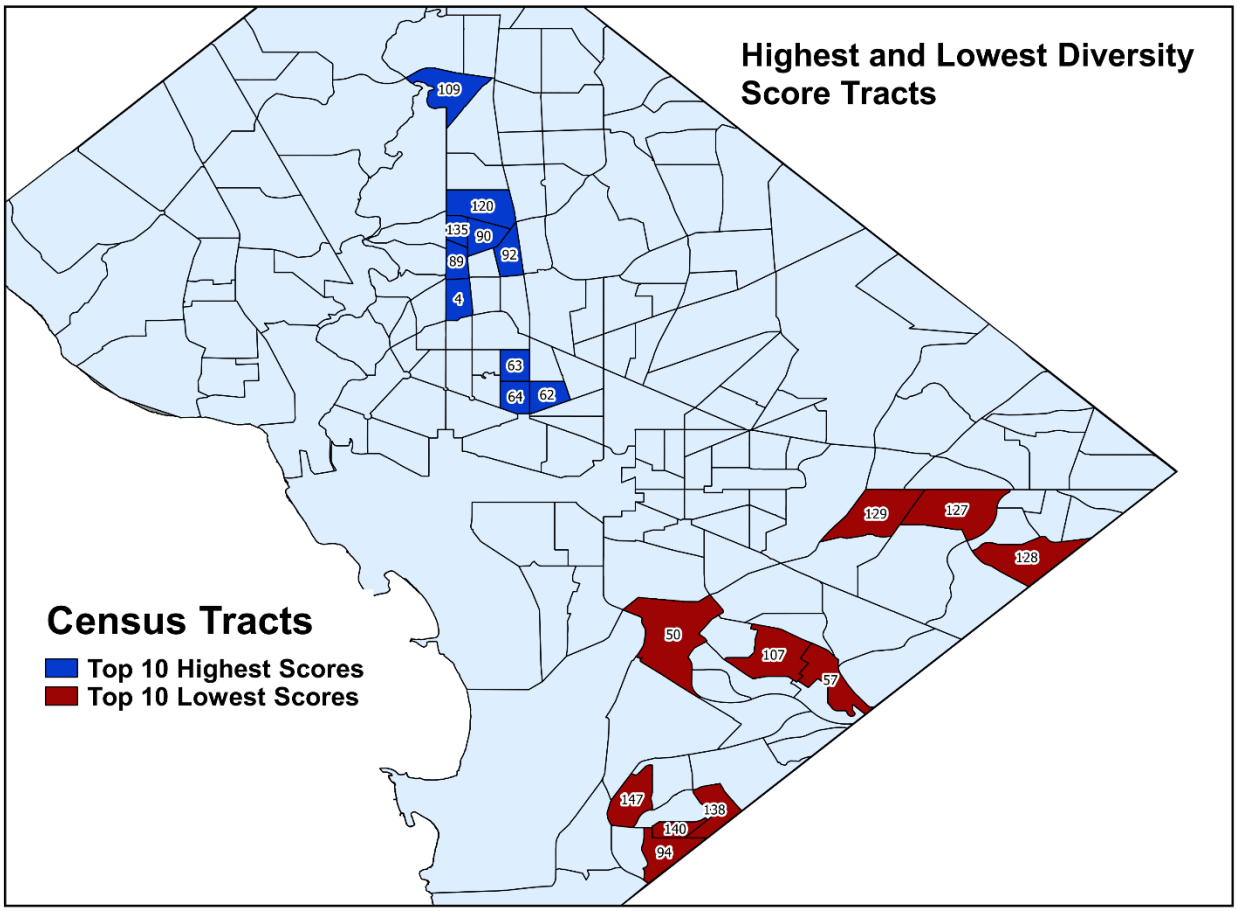


Figure 10: Highest and lowest diversity score tracts



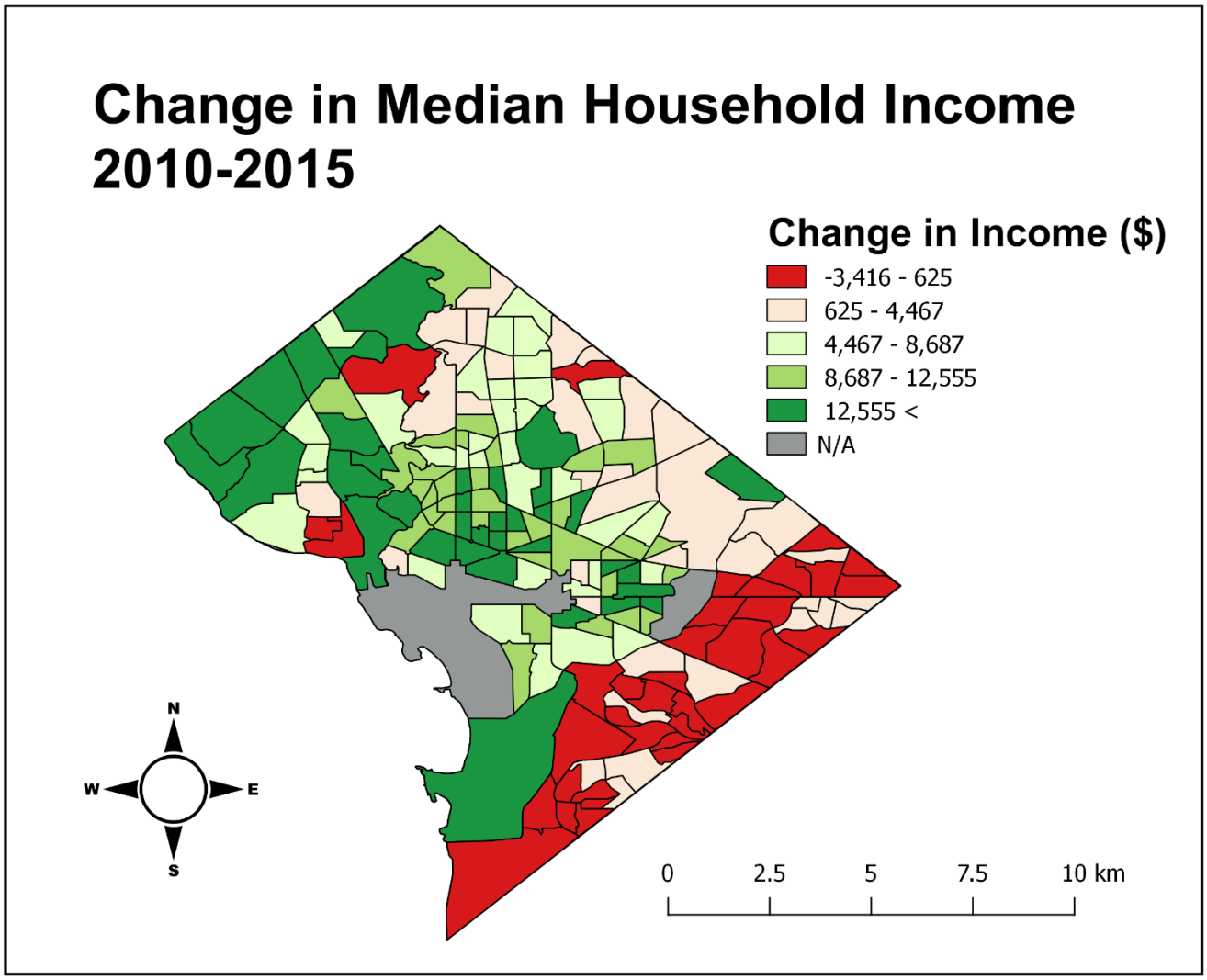
Top 10 Lowest Scores:

Rank	Map ID	TRACT	Total_Pop	%White	%Black	%AmerIn	%Asian	%Hawa	%Other	%Multra	Diversity	Diversity Level
1	129	7708	3070	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0	Low
2	140	9802	1646	0.18%	99.82%	0.00%	0.00%	0.00%	0.00%	0.00%	0.3638557	Low
3	128	7707	4423	0.00%	99.53%	0.00%	0.00%	0.00%	0.00%	0.47%	0.9450732	Low
4	94	9811	5018	0.00%	99.50%	0.00%	0.00%	0.00%	0.34%	0.16%	0.9925289	Low
5	107	7504	2780	0.50%	99.46%	0.04%	0.00%	0.00%	0.00%	0.00%	1.0736763	Low
6	50	7401	2453	0.20%	99.39%	0.00%	0.41%	0.00%	0.00%	0.00%	1.2171756	Low
7	147	9803	2733	0.22%	99.27%	0.29%	0.00%	0.00%	0.00%	0.22%	1.4564171	Low
8	127	7703	5267	0.00%	98.94%	0.00%	0.32%	0.00%	0.74%	0.00%	2.1086187	Low
9	138	9700	3371	0.27%	98.52%	0.00%	0.00%	0.00%	1.22%	0.00%	2.9289732	Low
10	57	7502	5361	1.27%	98.41%	0.00%	0.00%	0.00%	0.00%	0.32%	3.1288168	Low

Top 10 Highest Scores:

Rank	Map ID	TRACT	Total_Pop	%White	%Black	%AmerIn	%Asian	%Hawa	%Other	%Multra	Diversity	Diversity Level
1	92	3100	3350	36.69%	28.39%	6.30%	5.73%	0.00%	21.94%	0.96%	72.934052	High
2	135	2801	4033	29.58%	42.18%	0.00%	4.84%	0.00%	20.16%	3.25%	69.057624	High
3	109	2001	2848	29.60%	43.50%	2.74%	3.44%	0.00%	18.22%	2.49%	68.735945	High
4	89	2802	4750	46.78%	21.18%	0.00%	5.96%	0.00%	23.39%	2.69%	67.733567	High
5	90	2900	4567	48.70%	27.30%	0.15%	4.09%	0.00%	15.09%	4.66%	66.16902	High
6	64	4902	2842	51.65%	21.43%	0.39%	9.96%	0.00%	13.27%	3.31%	65.864897	High
7	120	2502	5852	28.90%	48.19%	0.17%	3.54%	0.00%	17.02%	2.19%	65.358689	High
8	62	4802	3548	32.27%	48.79%	0.00%	11.39%	0.00%	4.96%	2.59%	64.172759	High
9	63	4901	2600	43.27%	41.31%	0.00%	7.00%	0.00%	5.65%	2.77%	63.328136	High
10	4	3700	5860	46.47%	37.80%	0.32%	5.05%	0.00%	9.10%	1.26%	63.020839	High

Figure 11: Change in median income by tract



The following heatmaps of population and diversity were also produced in QGIS. To do so, the geometry tool Polygon Centroid was used on the census tract polygons to create a point layer of centroids. A heatmap fill style was used on the new centroid layer weighted by the variables in each map below. The basemap was layered by using the OSM plugin provided by QGIS.

Figure 12: White populations with Ward overlay

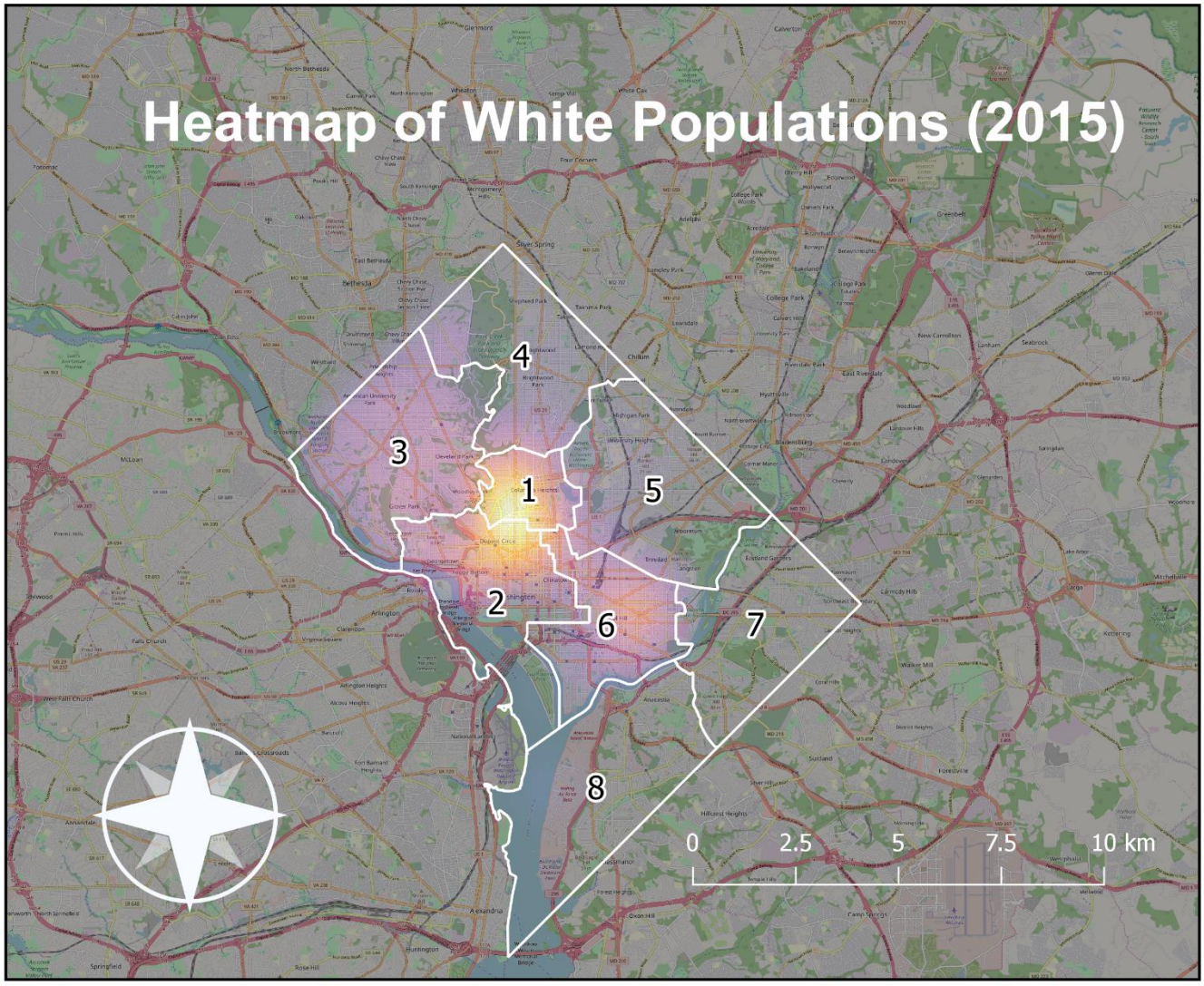


Figure 13: Black populations with Ward overlay

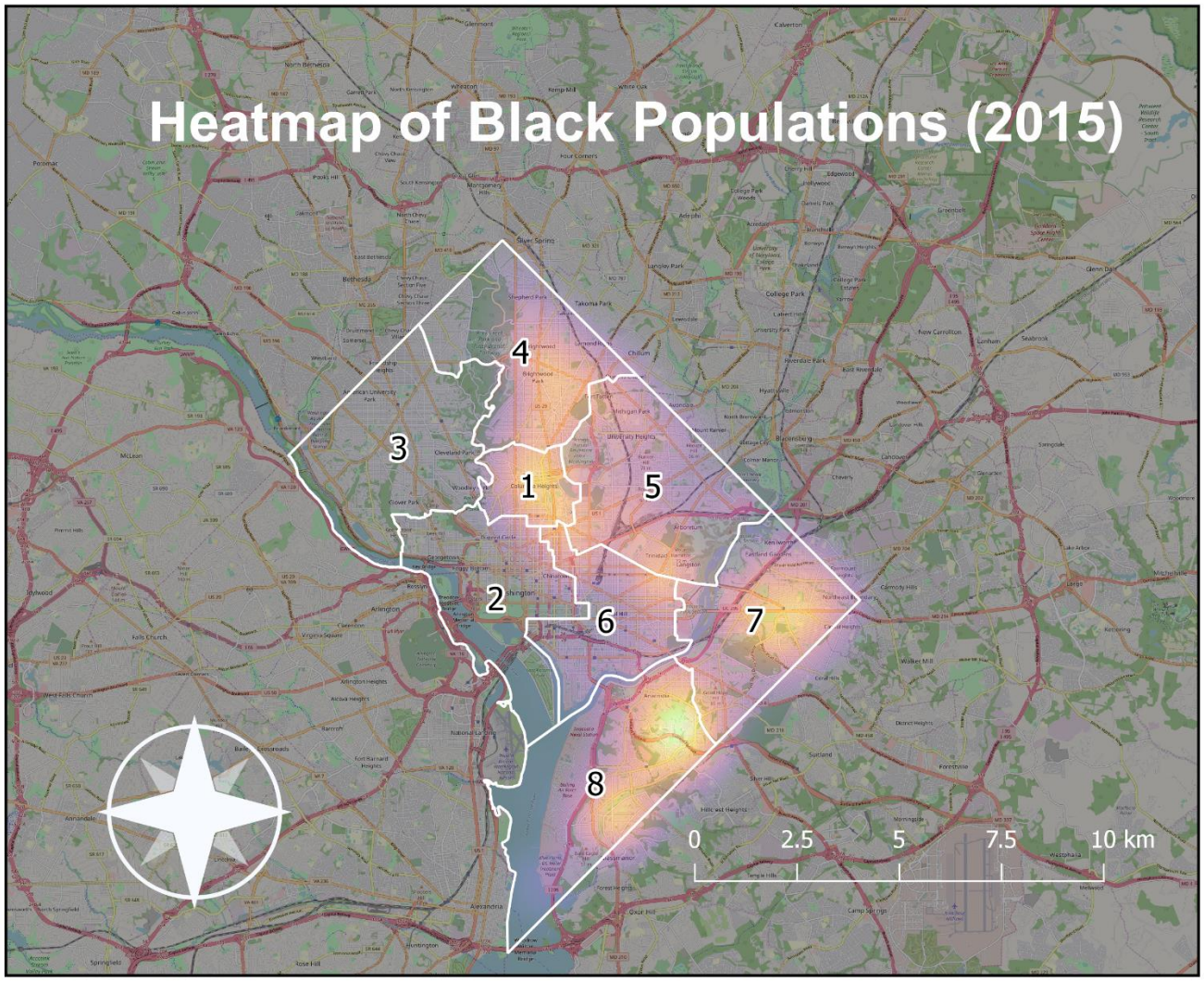
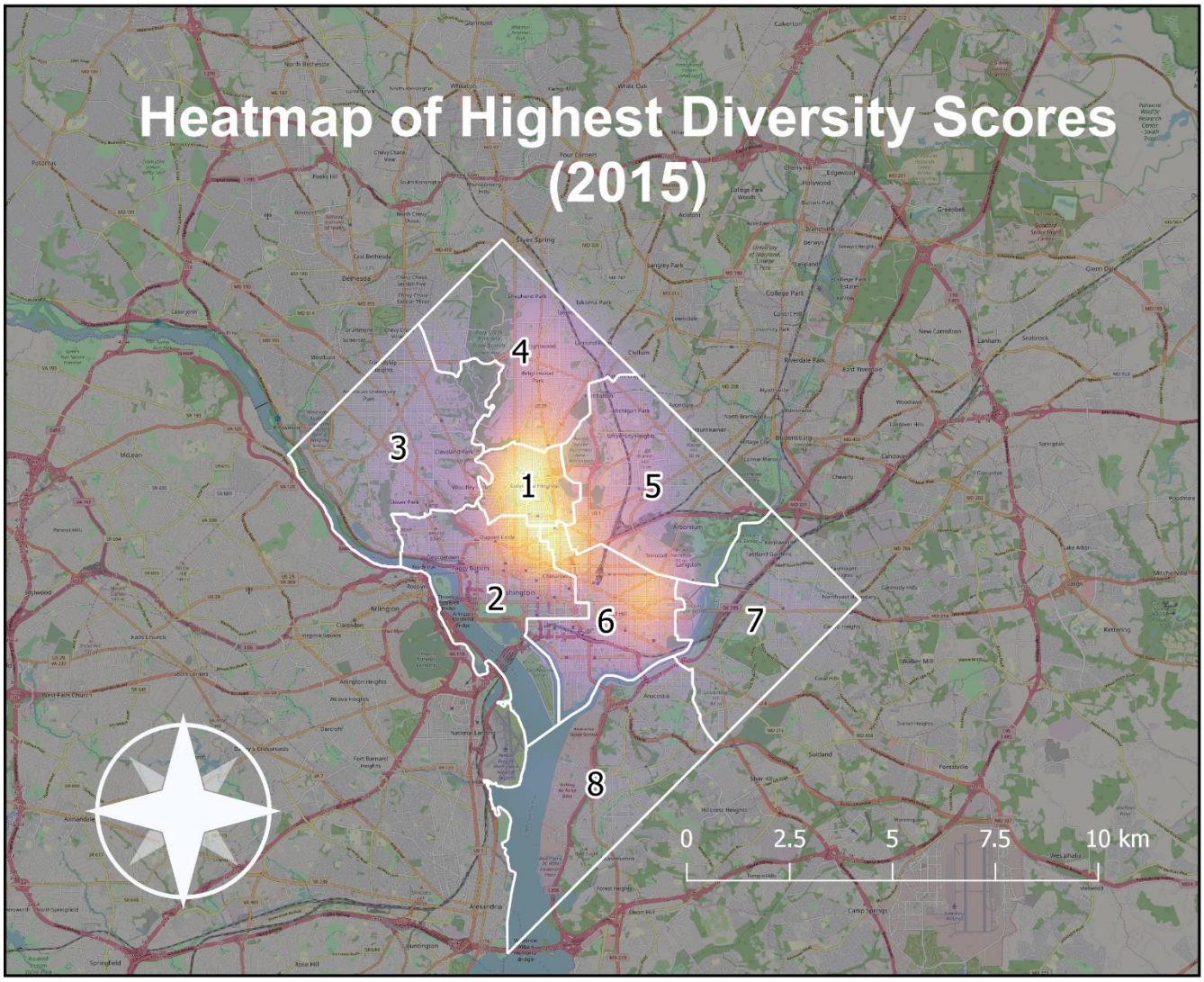


Figure 14: Weighted diversity scores with Ward overlay



Visualizations of the regression results are shown below. The appendix shows the code and calculations of the regression.

Figure 15: Scatterplot with fitted line

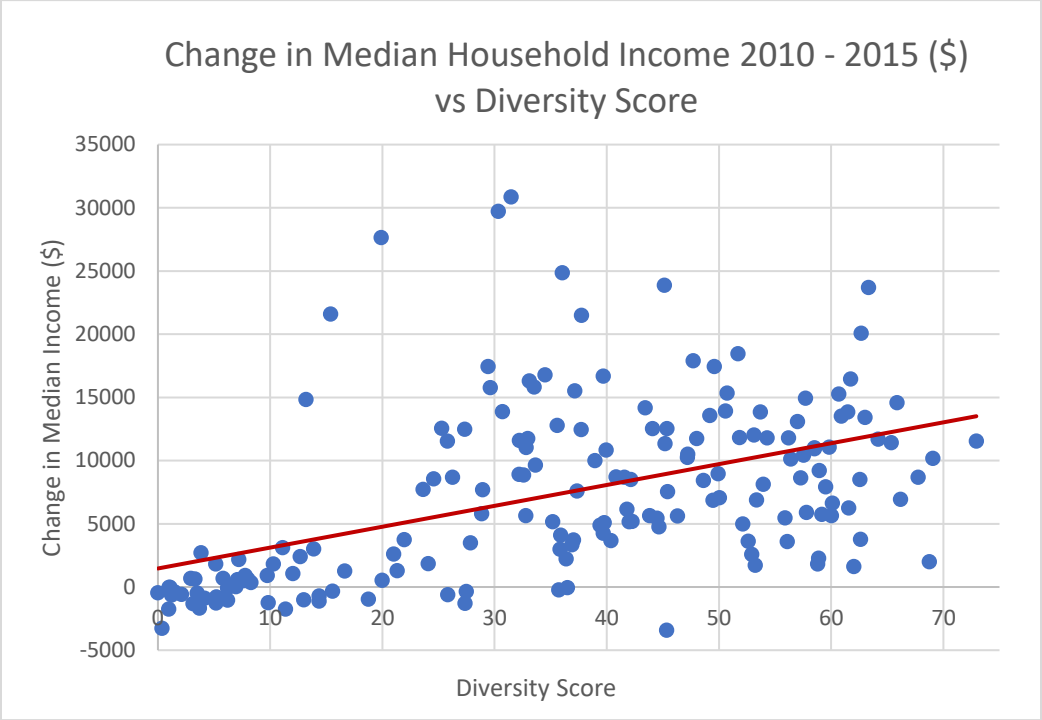


Figure 16: Residuals vs Fits plot of regression

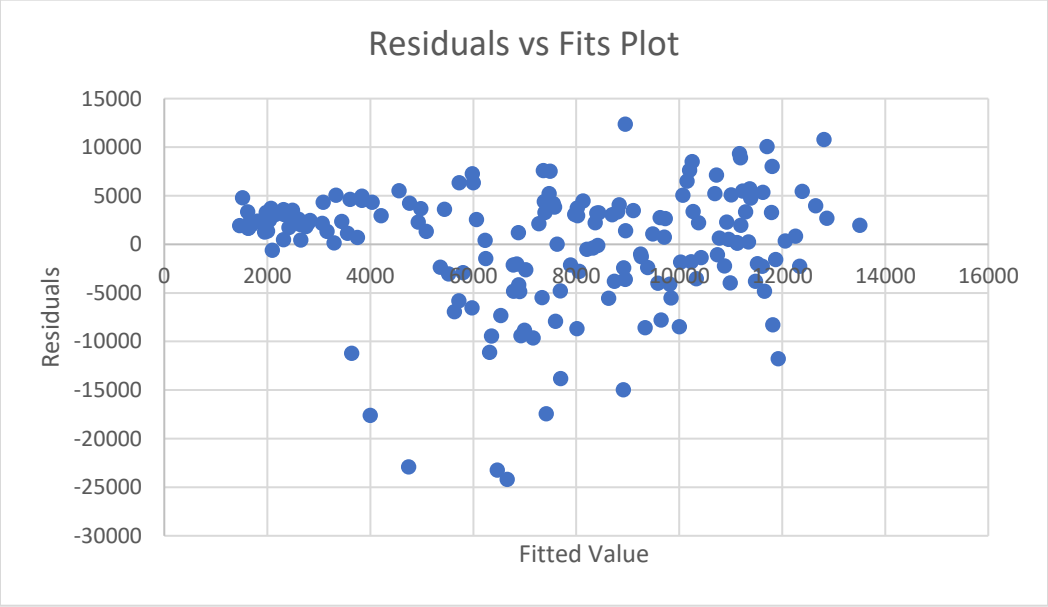


Figure 17: Important results from regression analysis

Gradient	165.12
Intercept	1,465.90
R^2	0.232
MSE	36852868.99
p-value	1.134e-11

Results from the regression provide the fitted equation –

*Estimated Change in Median Income ($\hat{y}$) = 165.12 * x + 1,465.90*

Figure 18: Important results from Moran's I test

Standard Deviation	7.403
Moran I Statistic	0.311
Expectation	-0.006
Variance	0.002
p-value	6.655e-14

An important condition of regression on spatial objects is that residuals should not be autocorrelated. Thus, the Moran's I test was run to identify if the residuals from the model displayed spatial autocorrelation. The results of this test are statistically significant (p-value < 0.05). The value of .31 is a sign that high residual values cluster near each other, and low values cluster near each other. However, with a score of one being perfect spatial autocorrelation and zero being none, we can conclude that the residuals are not spatially autocorrelated and are a weak indication of clustering.

Discussion

Regression of the data suggests that for every 1 point the diversity score increases, we expect about $165.12 positive change in median household income. The R^2 value indicates that 23% of the variation in income change can be explained by the variation in the diversity score. The p-value of $1.134e^{-11}$ suggests a genuine relationship (p-value < 0.05) and is statistically significant. While the result is significant, the low R^2 value is weak. Figure 16 and the Moran's I statistic suggest that regression was an appropriate analysis for these results.

In addition to the regression results, calculating the diversity score revealed some interesting findings.

The tracts with the highest diversity scores are centered on the center and north sections of Washington, DC; Ward 1 in particular. A possible explanation is the strong Central American

community that resides in the Ward 1 neighborhood of Columbia Heights. This community, combined with large numbers of black and white residents, may explain the high diversity score. As shown in Figure 10, the majority of the highest diversity scores are within that area.

The eastern section of Ward 6 also has high diversity scores. This area, Capitol Hill East, is a dense residential district with historic homes. Traditionally a black area of the city (Figure 13), diversity may have increased as affluent residents move back to DC and purchase homes in that area.

The areas of the city that have the lowest diversity scores (Figures 7-8) are the same areas with the greatest negative change in median income (Figure 11) and are in southeast DC. These are also the areas with the greatest black populations (Figure 13).

Figure 11 emphasizes the traditional economically divided classification of Washington, DC – the affluent, white neighborhoods of northwest DC compared to the poor, black neighborhoods of southeast.

Conclusion

Overall, we can reject the null hypothesis and accept the alternative hypothesis that there is a relationship between the diversity score and the change in the median income by census tract. Results of the regression prove this a statistically significant result. However, using regression to predict the amount of change in median household income based on the diversity score is not advisable. The low R^2 value shows it is not a strong relationship and is thus not particularly helpful for prediction purposes.

This study only scratched the surface of potential research avenues that would be helpful for city planners. Based on results from this study the following are suggested areas for additional research:

- Construct a Geographically Weighted Regression model to explore the spatial aspects in more detail.
- Build a multi-linear regression model with the goal of increasing the R^2 value to a level that would be helpful for city planners and researchers for prediction purposes.
- Results suggest that the “white areas” of Washington, DC are actually more diverse than the “black areas” (Figures 12-14). An explanation of this could be fertile ground for further investigations.