Salem State University

The Graduate School

Department of Geography

Which Urban Residents Vote and Why?
A Geospatial Analysis of Voting Behavior in Worcester, MA

A Thesis in Geo-Information Science

by

John D. Holbrook

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

August 2018

ACKNOWLEDGEMENTS

I would like to thank Keith Ratner and Marcos Luna for developing an accessible, high quality GIS graduate program at Salem State University and providing valuable feedback during this study. I would also like to thank William Hansen and Thomas Conroy, of Worcester State University, for encouraging my pursuit of this topic. Special thanks also to Niko Vangjeli, Assistant City Clerk for the City of Worcester, for his responsiveness to communication and interest in voter participation in the city.

ABSTRACT

WHICH URBAN RESIDENTS VOTE AND WHY?
A GEOSPATIAL ANALYSIS OF VOTING BEHAVIOR IN WORCESTER, MA

AUGUST 2018

JOHN D. HOLBROOK, B.S., THE UNIVERSITY OF MAINE, ORONO

M.S., SALEM STATE UNIVERSITY

Directed by: Keith Ratner

This study investigates the relationship between voter travel distance to polling places in Worcester, MA and voter turnout. Linear and geographically-weighted regression are used to evaluate the significance of travel distance and demographic control variables. Worcester appears to be unique when compared to previous studies investigating travel distance and voter turnout. Travel distance to polling place does not reliably predict voter turnout in Worcester, but vehicle ownership, race, and age do.

ABBREVIATIONS

CSV: comma-separated value; a file format for large data sets

GIS: geographic information system

GUI: graphical user interface

GWR: geographically-weighted regression

IDE: integrated development environment

OLS: ordinary least squares

TABLE OF CONTENTS

INTRODUCTION

This study uses GIS to examine the relationship between voter turnout and the personal cost of traveling to the polling place to vote in Worcester, Massachusetts. It seeks to understand whether distance from the polls, as a metric for personal travel cost, affects the likelihood of voting. It also seeks to identify demographics that may be associated with higher or lower voter turnout and whether they relate to distance from polling place. Linear regression and geographically-weighted regression are used to identify trends between the variables.

The Geography of Counting Votes

Individual participation is the cornerstone of democracy. Like other rights guaranteed to individuals by their government's foundational documents, the right to vote is only useful if exercised. An engaged electorate influences change it wants, while an apathetic electorate may find itself disenfranchised. In the United States, 55.7% of eligible voters cast ballots in the 2016 presidential election (DeSilver, 2017). As a democracy which attempts to serve as an example to the rest of the world, the voter turnout in the United States should be higher. There are many possible reasons why turnout is low: apathy, the perception of social norms, difficulty in accessing polling places, lack of education or information about elections, or the perception of a lack of choice in candidates or policies (Brady and McNulty, 2011) (Blais, 2006) (Gerber, Green, and Larimer, 2008).

The number of people who vote in a given election is known as "turnout." While the number of people who cast a vote is easily summed, it is usually reported as a percentage of a larger number. The choice of which larger number is used to calculate this percentage changes the outcome. Defining which people are eligible to vote is an evolving process because voting has not always been a right in the United States; the definition of who is eligible to vote directly affects reported voter turnout. The most common calculation of turnout percentage is the number of registered voters who voted out of the voting-age population (DeSilver, 2017). The method varies in sholarly articles, but number of votes out of voting-age population and number of votes out of number of registered voters are the most common (Geys, 2006).

Every ten years, the United States federal government executes its constitutional mandate to conduct a census of the population (U.S. Census Bureau, 2016). The census asks detailed demographic questions of individuals, including their race or ethnicity, age, gender, housing status, occupation, income level, education level, family size, and more. After each census, electoral votes for each state are assessed to match the new population count, and state legislatures use the new data to re-draw senate, congressional, and state representative districts. The legislature also draws voting districts based on census data, with the goal of creating districts with similar population counts (U.S. Census Bureau, 2018). During this process, the geography of counting votes has also been used as a tool to suppress them. The process now known as gerrymandering is that of redrawing electoral districts in such a way that benefits or harms a candidate, political party, or specific group of people (Encyclopedia Britannica, 2004).

The census uses geographic units of varying scale to aggregate data collected on individuals. In doing so, the privacy of an individual is protected and comparisons between equivalently-sized geographic units can be made. The primary geographic census units, by increasing size, are blocks, block groups, and tracts. Blocks are often only one part of a city street, while block groups encompass a neighborhood or two and are generally defined as between 600 and 3,000 people. Tracts contain at least one block group and are generally between 2,500 and 8,000 people. For representative election purposes, census blocks are used at the local level to draw municipal wards and precincts. Wards are areas for which city or town representatives are elected, and precincts are divisions of wards whose residents vote at a single polling place (U.S. Census Bureau, 1994).

Worcester as a Case Study

Worcester is the second most populous city in both Massachusetts and New England, with a population of 181,000 at the 2010 Census (MassGIS, 2017). Worcester is divided into ten wards and fifty precincts, or five precincts per ward. Each precinct contains about 3,600 people, which reflects the effort to balance the number of people in each of the fifty precincts. The wards and precincts are defined using census blocks, and in most cases closely follow the boundaries of census block groups. As of 2018, the city had 42 polling places, with some precincts sharing. Not every polling places is within the boundaries of its respective precinct, even where precincts are not shared (City of Worcester, 2017). An initial study using Worcester's voting precincts

revealed that there was a disparity in voter participation across Worcester for the 2016 presidential election (Conroy, Hansen, and Holbrook, 2017). There are demographic trends accompanying this disparity, particularly that the predominately white northwest side of the city votes in higher numbers than the Hispanic communities. It also illustrated the difference in perceived voter turnout, using both registered voters and potentially eligible voters as the denominator.

While Worcester residents represent a cross-section of the average American city, there are aspects which also make it unique. Worcester's topography is typical of New England mill towns: It is hilly and crossed by many small streams and rivers. The era in which the city was settled and expanded is reflected by how its road network was built to accommodate the topography (Worcester Historical Museum, 2018). Streets are narrow, twisting, and in no way reflect the "grid" of planned cities. They are typical of the road networks in the region, most of which were established prior to the widespread introduction of the automobile. The confusing and inefficient road network in the city means that transportation can be a hassle. Getting from one point to another is never simple, whether traveling by foot, bicycle, bus, or car. When considering the reasons why a Worcester resident will or won't vote, the time and energy spent getting to the polls cannot be ignored.

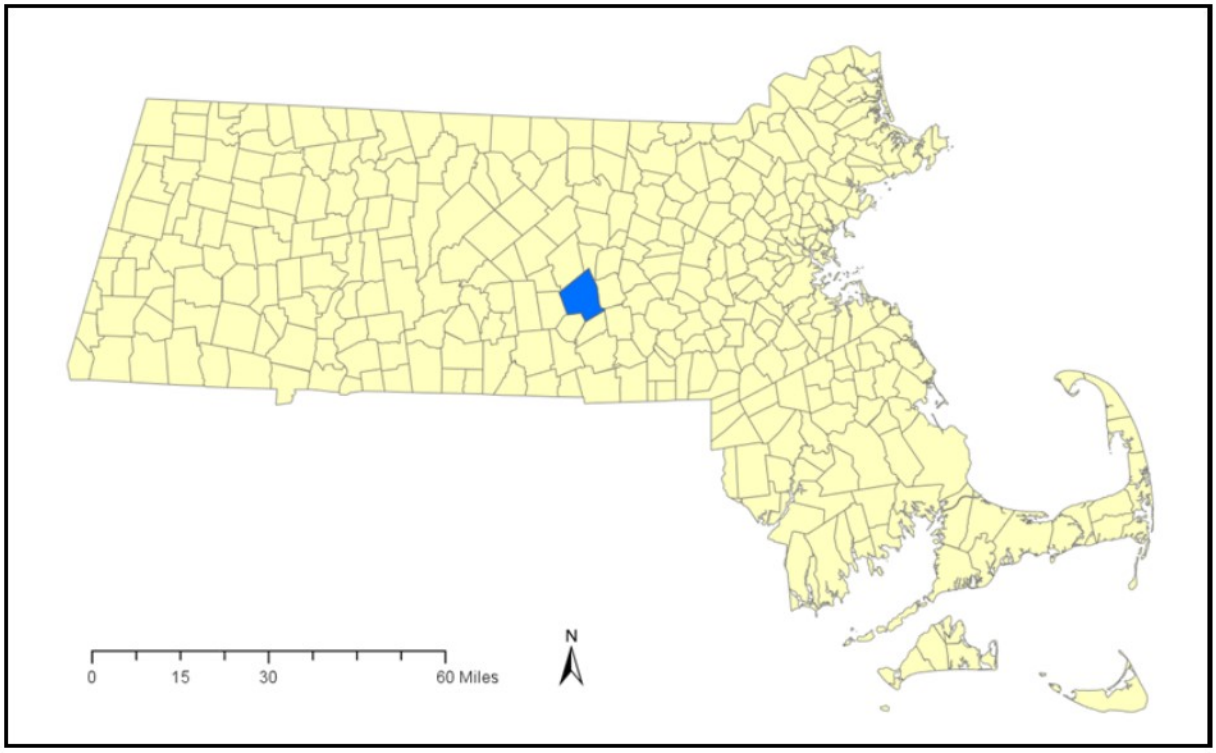

Figure 1. Location of Worcester within Massachusetts

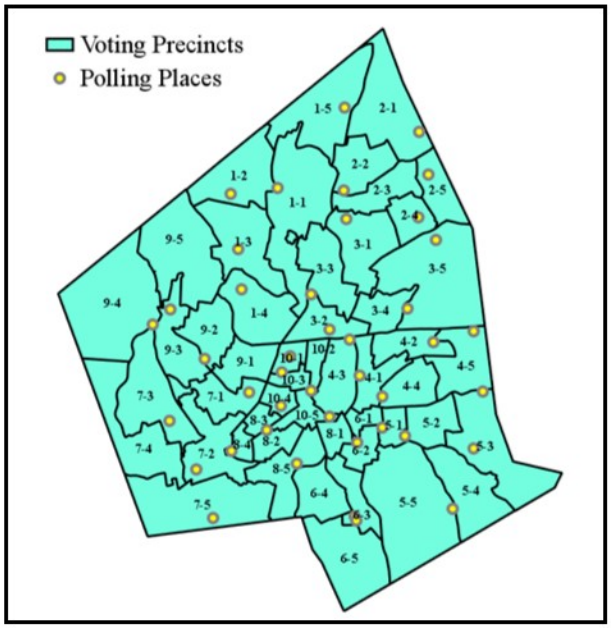

Figure 2. Voting precincts and polling places in Worcester

Figure 3. Worcester road network

LITERATURE REVIEW

The Geography of Casting a Vote

Estimating the perceived personal "costs" of voting has been studied for decades. The decision to vote or not can be thought of as an economic cost-benefit analysis, in that voting takes one's time, effort, and money for transit, and an individual's perceived benefits from voting are compared against these costs. If an individual perceives the benefit to outweigh the cost, they will vote, and vice versa (Haspel and Knotts, 2005). However, only more recently has GIS been used to quantify distance from a polling place as one of these costs. As such, the first to approach the issue used different methods of spatial modeling to varying degrees of success. The common ground among these studies is the use of the same independent and dependent variables: voter distance from the polling place and voter turnout, respectively.

Previous studies utilizing GIS to analyze these variables quantified voter distance to polling place using different methods. An early use of statistical GIS showed a strong geospatial trend about where people who voted lived relative to their polling place (Sui and Hugill, 2002). This study was concerned more with examining the Neighborhood Effect, that is the effect higher voting communities have on the voting behavior of neighboring communities, rather than the influence of distance to the polls on turnout. It is significant to note nonetheless, as it provides a window into the process for preparing a geospatial study of voters and voting.

The first study to use GIS to measure the effect of voter's travel distance to the polls on turnout hypothesized that voter turnout will be higher in precincts where polling locations are more easily accessible, measured as shorter travel distance (Gimpel and Schuknecht, 2003). The authors created a GIS model which took the geographic centroids of precinct polygons, accounted for uninhabited land, and calculated the distance from the centroid of the polygons to the precinct's polling location using the Manhattan Block method. This was the basic measure for ease of access to polling location, combined with impedance. However, in the context of more recent GIS technology, this assumption indeed seems overly simplistic. The distribution of the population is likely not even across a precinct, which means that the centroid of the precinct does not necessarily indicate the average distance a resident will travel from home to the polls.

A more precise measure of travel costs can be accomplished by using voters' home addresses, rather than simply the centroid of a voting precinct to calculate distance to polling place (Haspel and Knotts, 2005) (Gimpel, Dyke, and Shaw, 2006). The addresses of registered voters can be geocoded using address databases in a GIS, and the Manhattan Block method can then be used to calculate the distance from each address to the voter's respective polling place. Both of these studies found there was a non-linear relationship between distance to polling place and voter turnout. Multiple studies adopted the practice of using the geocoded addresses of registered voters to calculate distance to polling place, with varying methods of distance calculation such as straight-line, Manhattan Block, and road network analysis. A thorough comparative analysis of these methods showed that selecting one over another did not yield significant changes in the measurement (McNulty, Dowling, and Ariotti, 2009).

The most complex consideration when measuring the impact of travel cost on voter turnout is accounting for other variables which influence turnout. Gimpel and Schuknecht (2003) identified several "ecological variables" which they incorporated into their model, including percent of population with a four-year college degree (or more), percent new to the area, percent of female headed households with children, and percentage of youth. The authors used geospatial statistics to measure the relationships between travel distance to polls and voter turnout. They tested for spatial dependency of turnout between precincts and established using OLS and Global Moran's I that there was a significant spatial relationship, and thus had to adjust for this spatial dependency in evaluating independent variables. Despite their confidence in these adjustments, they admit "some important explanatory variables may still be missing." Nonetheless, the study's findings reinforced most assumptions about their control variables, showing that better educated populations voted in higher numbers, populations new to an area voted in higher numbers, female-headed households with children voted in lower numbers, and younger voters voted in fewer numbers. The authors found that distance to polling place retained statistical significance even when control variables were removed from the model, although it reduced the magnitude of significance by about half. This suggests that if distance is a significant variable that explains voter turnout, the relationship should be evident even when other demographic variables are missing.

The best fit predictive model between distance from polling place and likelihood of voting was identified as logarithmic (Haspel and Knotts, 2005). The authors explain that people living close to polling place who walk or do not drive have a low initial travel cost, but the cost increases rapidly as the distance increases. The rapid increase in travel cost plateaus once the distance is far enough that most people would consider driving. At distances beyond this point, it is assumed that vehicular travel is the preferred mode, and likelihood of voting gradually decreases as distance increases. The distinction was also made between a vehicle being available or not, and the two distinct trends were plotted separately. The authors still found the logarithmic model to fit best, but the trend was more dramatic when a vehicle was available, presumably because voters living close to the polling place with a car available would find driving and parking inconvenient enough to avoid the hassle altogether.

Haspel and Knotts (2005) also found that turnout increased when polling places move, which could be for the same reasons that caused newer populations to vote more. A related finding that could work into future site selection and drawing precinct boundaries, is that turnout increased when precincts were split, outweighing the confusion over getting to a new polling place (Haspel and Knotts, 2005). Later studies in different geographic regions found that consolidating precincts, and thus relocating polling places, reduced voter turnout (Brady and McNulty, 2011), (McNulty, Dowling, and Ariotti, 2009) because of the information cost associated with adapting to a new polling place. Unfortunately, the comparison between consolidated precincts and split precincts is collectively discussed in ambiguous terms, e.g. “relocated polling place,” making it difficult to discern the trend.

In the analysis that follows, the addresses of registered voters will be used to calculate an average distance to polling place by voting precinct. Linear regression and geographically-weighted regression will be used to examine the relationship between distance to polling place and voter turnout. The models will incorporate other demographic control variables to account for additional factors which contribute to voter turnout.

METHODS AND DATA

Data Acquisition

The data sets used in this study include geographic data from MassGIS, demographics from the U.S. Census Bureau, precincts and polling places in Worcester, voter registration lists for Worcester, and vehicle ownership data from MAPC Boston. MassGIS data is publicly available for download from the MassGIS website and was retrieved in georeferenced formats suitable for use in popular GIS, such as shapefile and file geodatabase (MassGIS, 2017). Specific geographies used include the boundary for the city of Worcester, census blocks and census block groups in Worcester County, and voting precincts and polling places in the city of Worcester. Census data is publicly available through the American FactFinder website and was retrieved as either Excel or csv files (U.S. Census Bureau, 2018). Precincts and polling places in Worcester are available on the City of Worcester's open data website and were retrieved as georeferenced shapefiles. Voter registration lists are not immediately available to the public through a similar online data store, but were provided at request by the Worcester City Clerk's office as Excel workbooks (City of Worcester, 2017). This study used the voter registration list and the list of voters who cast a vote in the 2016 November election. Vehicle ownership data was obtained from Boston's Metropolitan Area Planning Council as Excel sheets summarized by census block group (MAPC Boston, 2018).

Census block groups were used as the primary geographic unit in this study. Certain sensitive demographic information is only made available in census block groups, such as income, which prevents sensitive information from potentially being associated with specific individuals. Analysis of demographics therefore requires a summary of the data at varying geographic units. Although race and age data are available at the block level, income data are only available at the block group level. Since block level data can simply be summarized to the block group level, census block groups were the logical choice of geographic unit for this study.

Data Cleaning and Address Geocoding

Data importation, cleaning, and address geocoding were performed using Microsoft Excel 2016, ArcGIS Pro 2.1, ArcMap 10.5, and Python 3. Initial preparation for examining Worcester geographies involved common GIS functions which extracted the boundary for the

city of Worcester from all towns in Massachusetts and the census block groups from the block groups in Worcester County.

The voter list includes a record for each person who is registered to vote, and each record includes a unique Voter ID number, the individual's name, address, gender, ward-precinct, and their party affiliation. Before importing the list into the GIS, a cleaned version was created which removed names for privacy and redundant or incomplete attributes (columns). The list was imported into ArcGIS Pro using the Excel to Table tool. This table was processed in the ESRI World Geocoder, which searches a global database of addresses, each of which has a global position in latitude and longitude. Through this process, a simple table of addresses becomes a georeferenced entity (or "feature class" in ESRI parlance) which displays each address as a point on a map in a GIS. It is worth noting that this process repeatedly failed in ArcGIS Pro 2.1, but ArcMap 10.5 processed all 109,000 records in about 20 minutes with a greater than 99% confidence in the results.

Although the address geocoding tool reported 100% accuracy, a manual inspection of the geocoding results revealed a few inconsistencies. First, the points derived from three addresses actually fell outside the city limits shown by the MassGIS shapefile of towns. Second, a comparison of the address points with the population totals of the census blocks revealed blocks which had zero population, but with several active voters. Essentially, because census blocks are so small (often only a portion of a city block), address points which were within 100 feet of a house sometimes fell into an adjacent census block. These inconsistencies were not significant enough to skew the overall analysis when summarized by the larger geographic unit of census block groups.

Distance Calculations

Calculations were performed in ArcGIS Pro 2.1 to determine the distances to each of the 42 unique polling places in Worcester from each registered voter's address. ArcGIS includes a "Near" tool which calculates the distance from one location to another. When presented with two separate point feature classes, the tool calculates for one feature class a single distance per record to the nearest location in the other feature class. All feature classes used in the Near tool were converted to or confirmed to be in the "NAD 1983 (2011) State Plane Massachusetts Mainland

FIPS 2001" projected coordinate reference system, which is necessary for the tool to calculate distances accurately. The mean of voter address distances to polling places were summarized by census block group.

Distances from voter addresses to their polling place were calculated using the Near tool in ArcGIS. Given the feature class of voter addresses and the feature class of polling places as inputs, the Near tool calculated the distance from each address to its closest polling place, rather than the polling place at which an individual might be required to vote. In an initial attempt to bypass this problem, the single feature class of all 42 polling places was split into 42 different feature classes, each containing one point representing an individual polling place. The voter addresses and each polling place were run through the Near tool with the same result. In order to ascertain the distance for the correct polling place, the Near tool had to be run once for each precinct and polling place. This required splitting the 109,000 voter addresses into 50 different feature classes based on the precinct to which each voter was assigned.

In order to automate this process, the Python 3 programming language and the ArcGIS Python library, arcpy, were used. Utilizing the Spyder IDE running on the ArcGIS Pro Python 3 kernel, a script was written which applied the Near tool to each precinct's voter addresses and the appropriate polling location. The Near tool appended an attribute column to each of the 50 feature classes showing the distance from an address to the correct polling place. These 50 feature classes were then merged to recreate a single feature class containing all 109,000 records and the distance to polling place for each. This new point feature class containing the distance each registered voter must travel to the polls was then summarized by census block group as an average (mean) distance to polling place.

Regression Analysis

Voter turnout of registered voters was used as the dependent variable to assess the significance of several explanatory variables. Although the contrast between registered voter turnout and potentially eligible voters was shown to be significant in Worcester, for the purposes of this study, registered voter turnout was used because the data available was more recent and granular than 2010 census data. The primary explanatory variable of interest was distance from

polling place as a measure of travel time/cost. Demographic explanatory variables tested in combination with distance from polling place were:

- Median household income
- Passenger vehicles per person over 18 (as of 2016)
- Percent Hispanic
- Percent non-Hispanic Asian
- Percent non-Hispanic Black
- Percent non-Hispanic White
- Percent college enrolled
- Percent ages 18 to 30
- Percent ages 56 and older

Statistical significance of explanatory variables was tested using Exploratory Regression in ArcGIS Pro. Explanatory variables were removed or added based on changes in the various indicators of model performance. The highest-performing combination of variables were modeled using Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR). Both OLS and GWR models were tested for Spatial Autocorrelation using global Moran's I (ESRI, 2018).

Initial Exploratory Regression included explanatory variables of distance to polls, vehicles per person, median income, percent Hispanic, percent non-Hispanic black, percent non-Hispanic Asian, and percent non-Hispanic white. Vehicles per person and percent non-Hispanic white showed significant positive linear trends. Percent Hispanic showed a significant negative linear trend. Although median income had a positive linear trend, it was only 60% significant. Other racial demographics were less significant and did not show strong linear trends. Distance to polling place was only 36% significant and split between positive and negative linear trends. Variables used in this model showed promising R-squared values, indicating that at least some of the independent variables tested explained a linear relationship with voter turnout. Some variables showed significant coefficient p-values indicating they had strong linear relationships with voter turnout. Most had low VIF values, indicating that some variables were likely exhibiting multicollinearity, but most were not. Jacque-Bera p-values were consistently low, which typically indicates the distribution of residuals in the model may not be normally

Percentage of Search Criteria Passed

Search Criterion	Cutoff	Trials	# Passed	% Passed
Min Adjusted R-Squared	> 0.50	119	104	87.39
Max Coefficient p-value	< 0.05	119	25	21.01
Max VIF Value	< 7.50	119	93	78.15
Min Jarque-Bera p-value	> 0.10	119	4	3.36
Min Spatial Autocorrelation p-value	> 0.10	18	0	0.00

Summary of Variable Significance

Variable	% Significant	% Negative	% Positive
VEH_PP	100.00	0.00	100.00
NH_WHT_PCT	100.00	0.00	100.00
INCOME	68.42	0.00	100.00
HISP_PCT	59.65	75.44	24.56
NH_BLK_PCT	56.14	63.16	36.84
DISTANCE	36.84	73.68	26.32
NH_ASN_PCT	21.05	56.14	43.86

Summary of Multicollinearity

Variable	VIF	Violations	Covariates
INCOME	1.96	0	--------
DISTANCE	1.37	0	--------
VEH_PP	1.81	0	--------
HISP_PCT	163.29	26	NH_WHT_PCT (96.30), NH_ASN_PCT (14.81), NH_BLK_PCT (14.81)
NH_ASN_PCT	10.00	4	HISP_PCT (14.81), NH_BLK_PCT (14.81), NH_WHT_PCT (14.81)
NH_BLK_PCT	19.87	4	HISP_PCT (14.81), NH_ASN_PCT (14.81), NH_WHT_PCT (14.81)
NH_WHT_PCT	259.31	26	HISP_PCT (96.30), NH_ASN_PCT (14.81), NH_BLK_PCT (14.81)

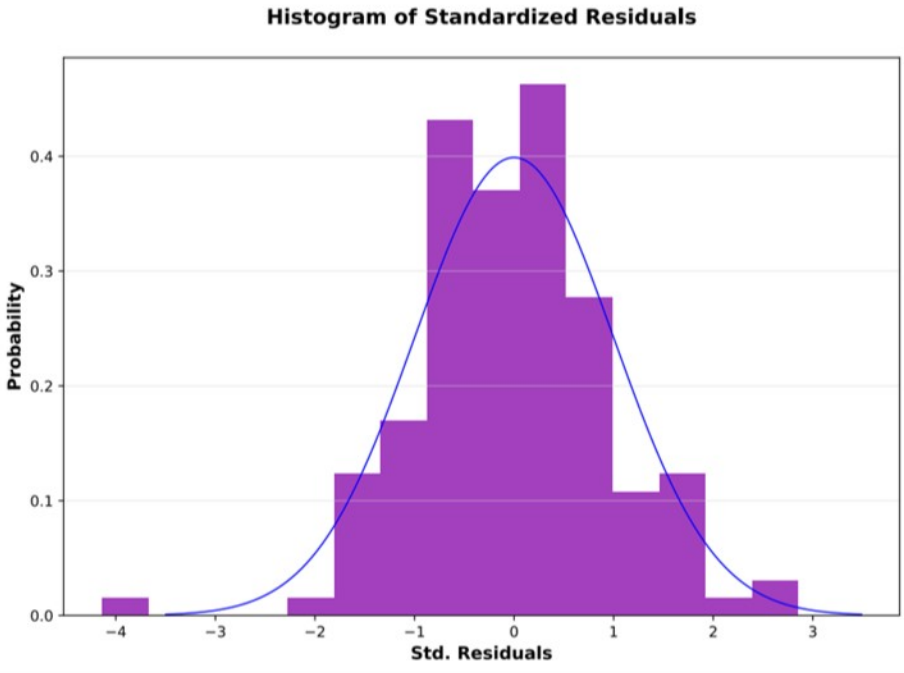

Figure 4. Select diagnostics for initial exploratory regression model

distributed. However, the histogram of standardized residuals showed a convincing normal distribution, which suggested the low Jacque-Bera p-values may not be the result of a non-normal distribution of residuals. Before this could be confirmed, the selection of variables required more iterations. No models passed the default threshold for spatial autocorrelation, which suggested that one or more explanatory variables was missing from the model or that one or more variables have geographic relationships that are not explained by a linear regression model. The low VIF values were the result of multicollinearity between the various racial demographic variables.

Several iterations of exploratory regression were run to test combinations of the independent variables. The variables for percent non-Hispanic Asian and percent non-Hispanic black were removed because the data distribution was widely dispersed and did not show strong linear trends. With these variables removed, both percent non-Hispanic white and percent Hispanic retained high VIF scores and the multicollinearity causing them were remedied by removing the variable for percent Hispanic. The assumption was that because non-Hispanic whites and Hispanics make up the two largest demographics in Worcester and the geographic distribution of these populations is divided, both variables were explaining the same linear trend: The positive linear trend between non-Hispanic whites and voter turnout and the negative linear trend between Hispanics and voter turnout are essentially the inverse of one another. Removing percent Hispanic from the model and leaving percent non-Hispanic white eliminated the multicollinearity and showed consistently low VIF scores.

The overall performance of the model improved significantly when including only income, distance to polls, vehicles per person, and percent non-Hispanic white. The significance of distance to polling place remained low and split between positive and negative linear. Again, despite normally distributed standardized residuals, the model returned significant Jacques-Bera p-values. This model also continued to fail the spatial autocorrelation threshold. Together, these indicated the need to investigate one or more missing explanatory variables. Additionally, the dispersed data points for distance to polling place and overall weak linear trend indicated that despite being the variable of interest, distance to polling place may not be a compelling explanatory variable for voter turnout in Worcester. The normally-distributed standardized residuals (when distance was examined in isolation) suggested that an exponential or logarithmic

Percentage of Search Criteria Passed

Search Criterion	Cutoff	Trials	# Passed	% Passed
Min Adjusted R-Squared	> 0.50	15	12	80.00
Max Coefficient p-value	< 0.05	15	8	53.33
Max VIF Value	< 7.50	15	15	100.00
Min Jarque-Bera p-value	> 0.10	15	0	0.00
Min Spatial Autocorrelation p-value	> 0.10	13	0	0.00

Summary of Variable Significance

Variable	% Significant	% Negative	% Positive
VEH_PP	100.00	0.00	100.00
NH_WHT_PCT	100.00	0.00	100.00
INCOME	75.00	0.00	100.00
DISTANCE	37.50	50.00	50.00

Summary of Multicollinearity

Variable	VIF	Violations	Covariates
INCOME	1.95	0	--------
DISTANCE	1.33	0	--------
VEH_PP	1.67	0	--------
NH_WHT_PCT	2.07	0	--------

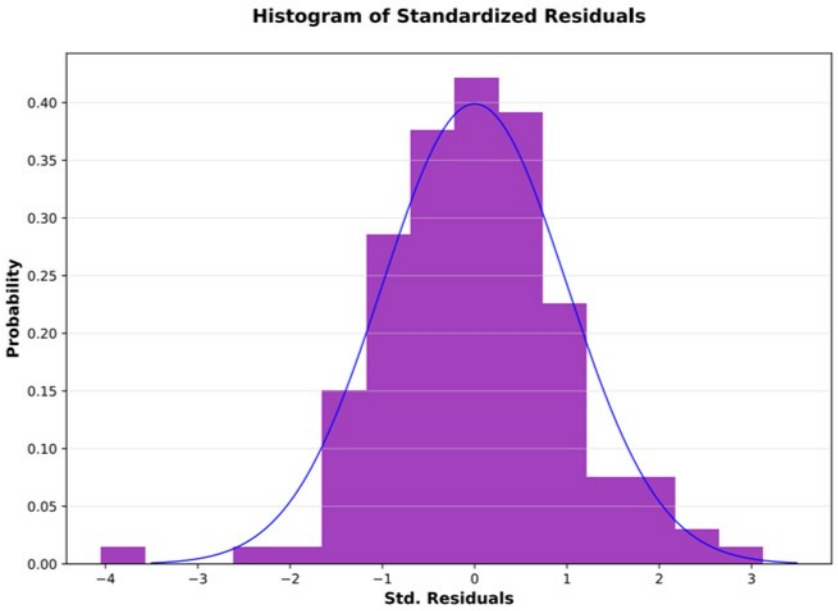

Figure 5. Select diagnostics from refined exploratory regression model

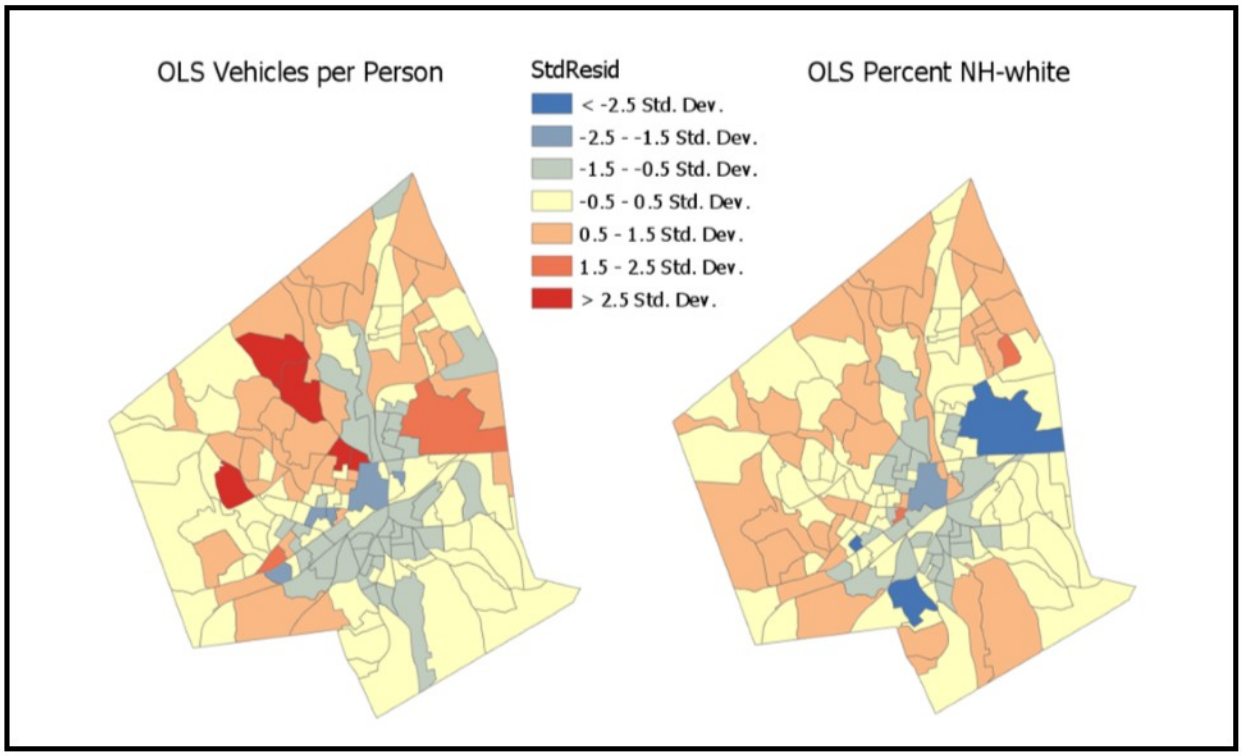

Figure 6. Standardized residuals for high-performing OLS variables

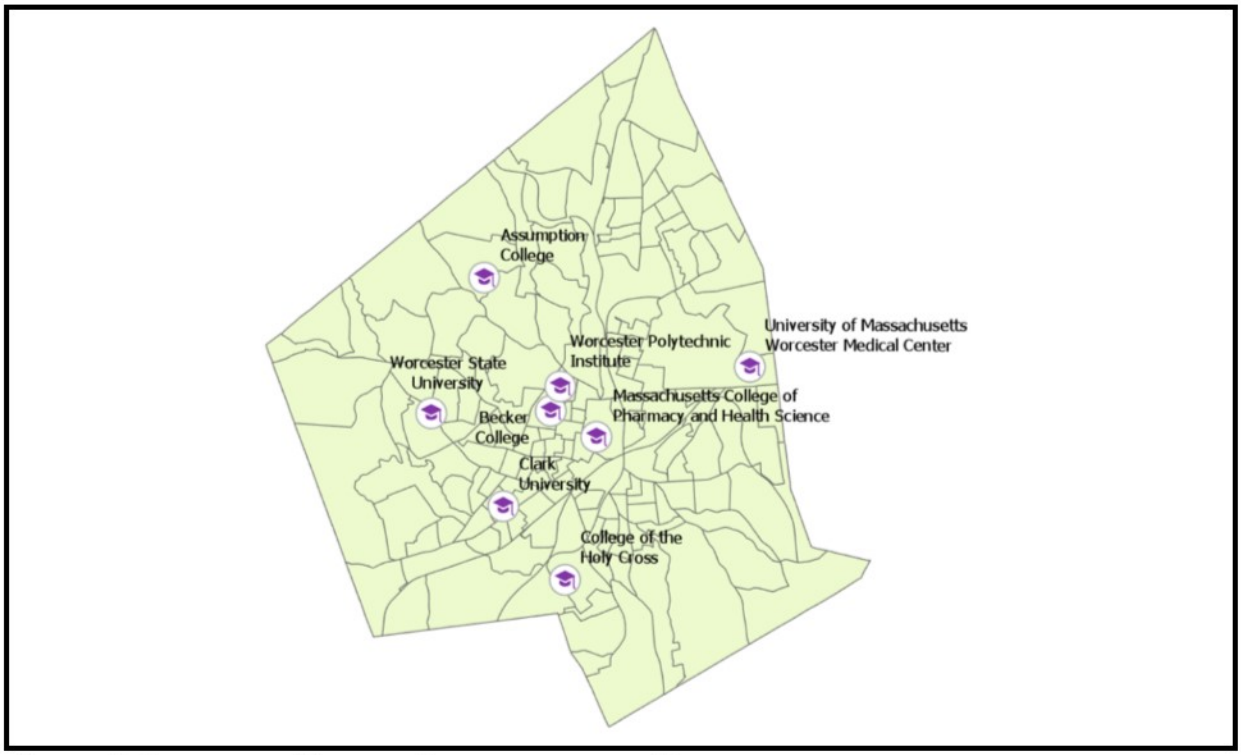

Figure 7. Locations of large college campuses in Worcester

transformation would not improve model performance of this variable.

To identify a missing explanatory variable, outlying residuals in Ordinary Least Squares for high-performing variables were identified. OLS was run on the two highest-performing variables in the existing model: vehicles per person and percent non-Hispanic white. The census block groups associated with outlying standardized residuals were overlaid on a map of Worcester in Google Earth Pro. Upon examination of the facilities within each block group, a characteristic common to the outlying residuals appeared to be the presence of a large college campus. Demographics for college-enrolled undergraduates and graduates were added to the model, but this variable returned low significance and was divided between positive and negative relationships with turnout. It is possible that this census demographic may not be a reliable count of college-enrolled or college-age residents because of the way in which college students are counted on the census (Cohn, 2010).

In an effort to capture the variable which makes these census tracts unique, demographics for age were included in the model. On the assumption that census tracts containing a large college campus have a higher proportion of young people, percent of population ages 18-30 was added to the model. The addition of this variable did not drastically change the performance of the model. Instead of identifying college-enrolled or young people, percent of population 56 and older was added to the model. The model including vehicles per person, percent non-Hispanic white, and percent age 56 and older as explanatory variables is covered in more detail in the Results section. These variables were also modeled using Geographically-weighted regression to assess the influence of geography on the relationships between the independent and dependent variables. Local R-squared values were mapped to locate areas of the city where the model fit best. Correlation coefficients were mapped to locate areas of the city with the strongest relationship between the explanatory variables and voter turnout.

RESULTS

Superficial analysis of the variables through scatterplots and choropleth maps sheds some light on aspects of Worcester's demography. Figures 8 through 13 show distribution of variables by census tract using geometric breaks, which balances frequency of observation and class width. Voter turnout in Worcester exhibits distinct geographic disparity between the northwest side of the city and the city center area (see Figure 8). Exactly how the independent variables examined in this study relate to the dependent variable is unclear when visually analyzing the choropleth maps. Nonetheless, some geographic patterns appear: Voter distance to polling place varies throughout the city, with the somewhat larger block groups farther from the city center having longer distances (Figure 9); Median household income is highest on the northwest side of the city and lowest in the downtown and eastern sections (Figure 10); Vehicle availability is generally much higher farther from downtown and lowest in the downtown area (Figure 11); Hispanic and non-Hispanic white populations are geographically the inverse of one another, with the highest proportions of white living on the north side of Worcester and the highest proportions of Hispanics living in the city center (Figures 12 and 13). Regression analysis provides indicators of how well these variables actually predict voter turnout in Worcester.

Linear Regression Model

The combination of vehicles per person, percent non-Hispanic white, and percent age 56 and older formed the best-performing linear regression model using Ordinary Least Squares to predict voter turnout in Worcester. Distance to polling place was not a significant explanatory variable in Worcester despite significance in regression models created in previous studies. Percent non-Hispanic Asian and black were removed due to collinearity with percent Hispanic and non-Hispanic white and widely dispersed data points without a strong trend. Percent Hispanic was also removed because of collinearity with percent non-Hispanic white because, as the two largest racial/ethnic populations in the city, their respective trends are essentially the inverse of one another. It is important to note that percent Hispanic showed a strong negative relationship with voter turnout, while percent non-Hispanic white showed a strong positive relationship. Percent college-enrolled and percent ages 18 to 30 did not improve model performance. Despite an overall positive relationship between median income and voter turnout, this variable did not consistently show high significance.

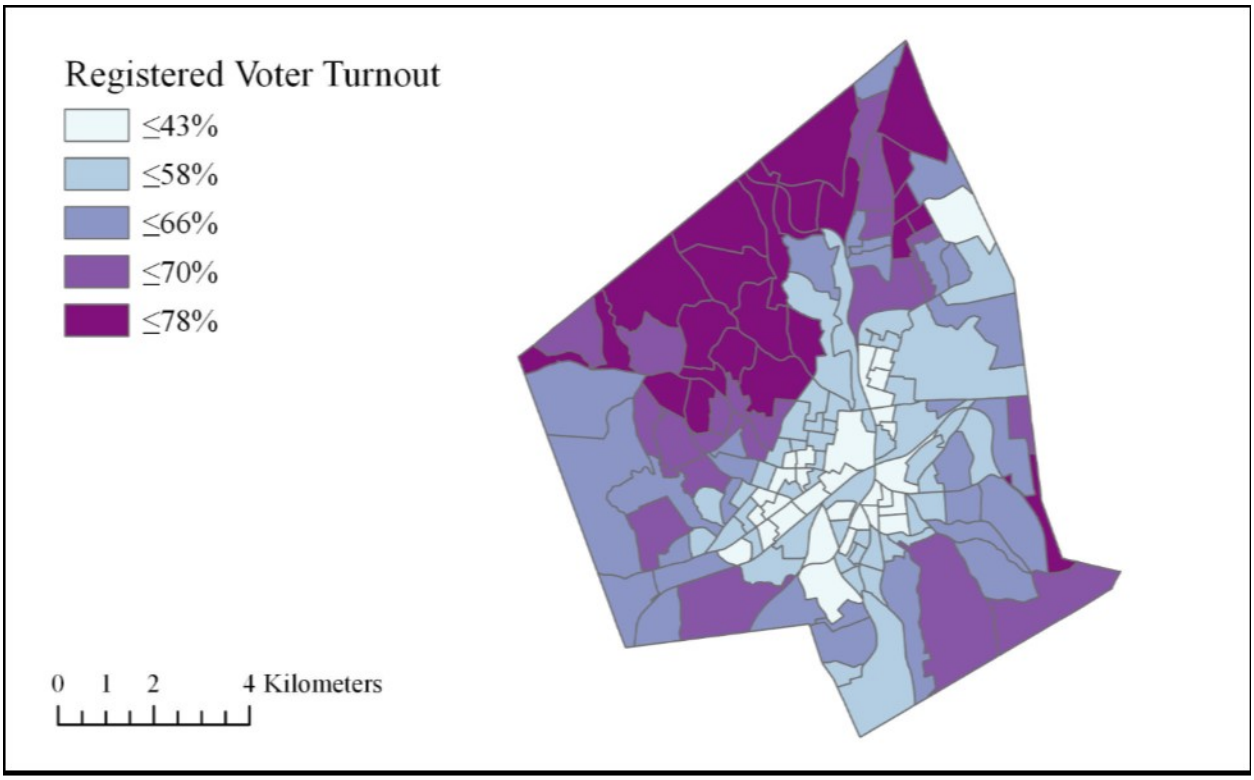

Figure 8. Percent voter turnout (dependent variable)

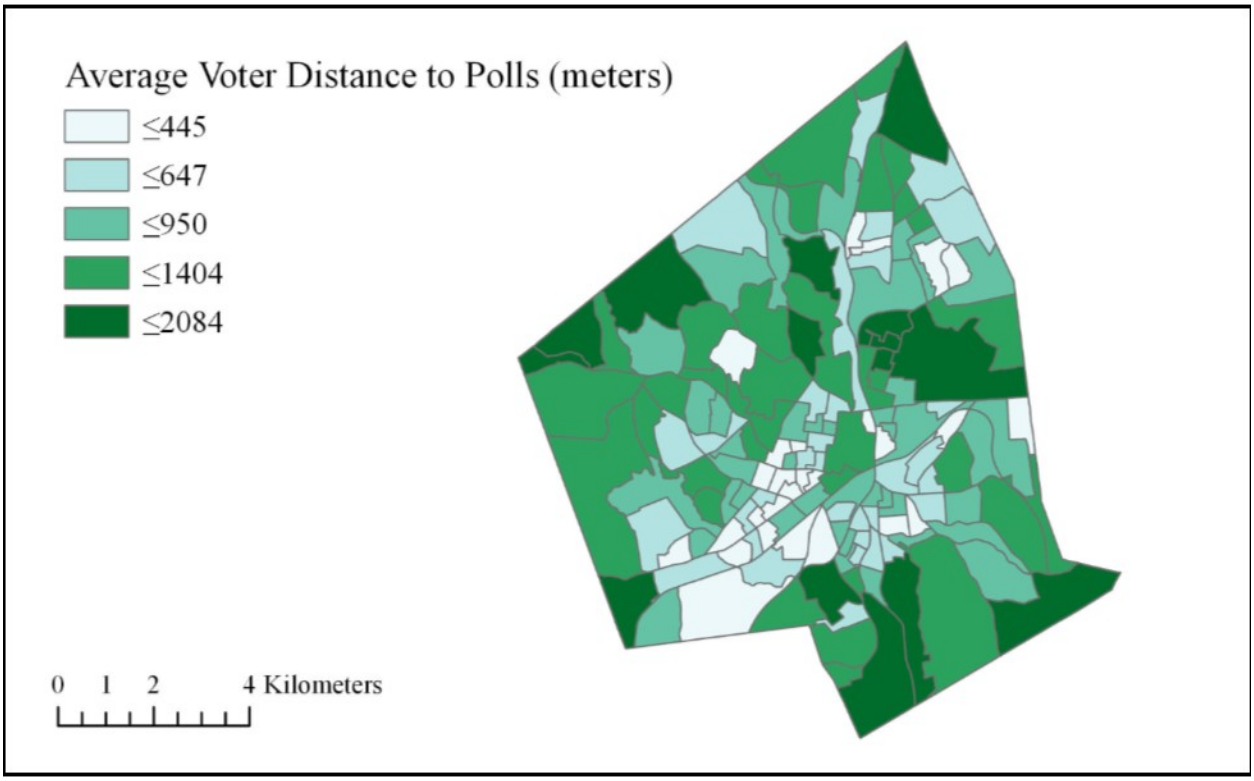

Figure 9. Average distance from voters' home address to polling place

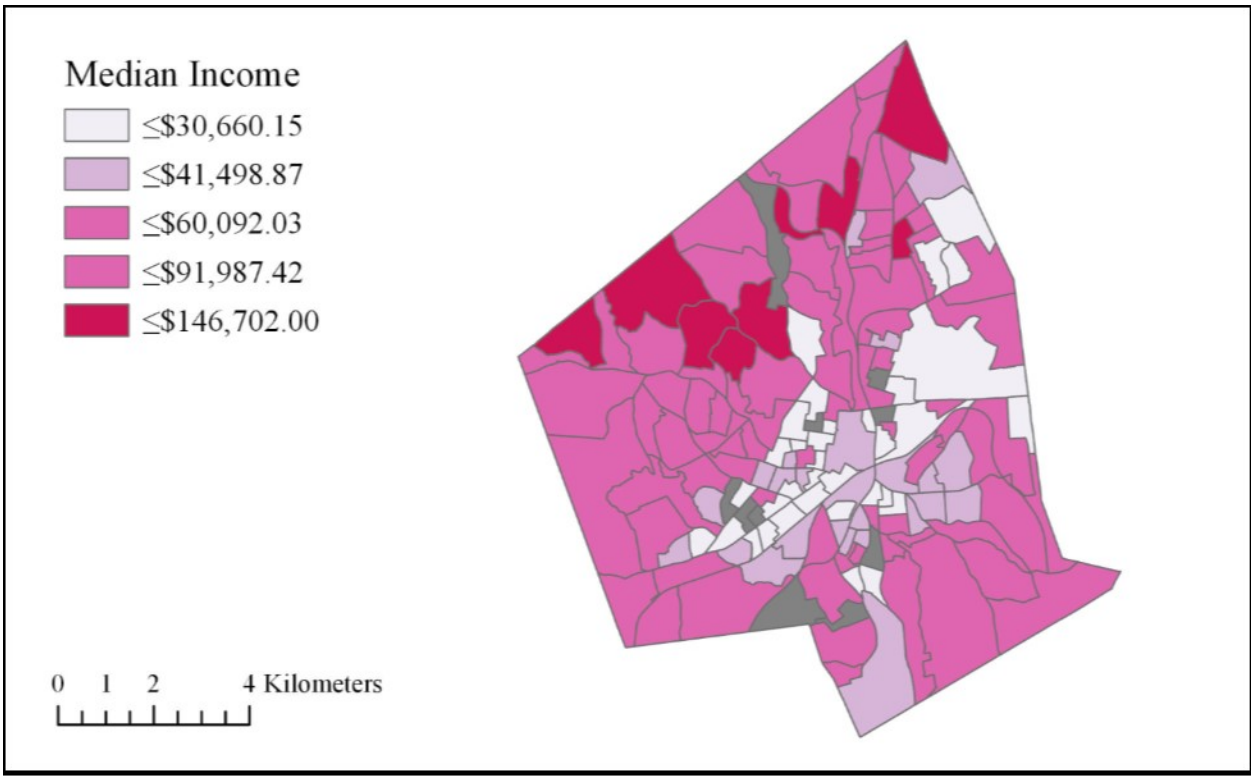

Figure 10. Median household income

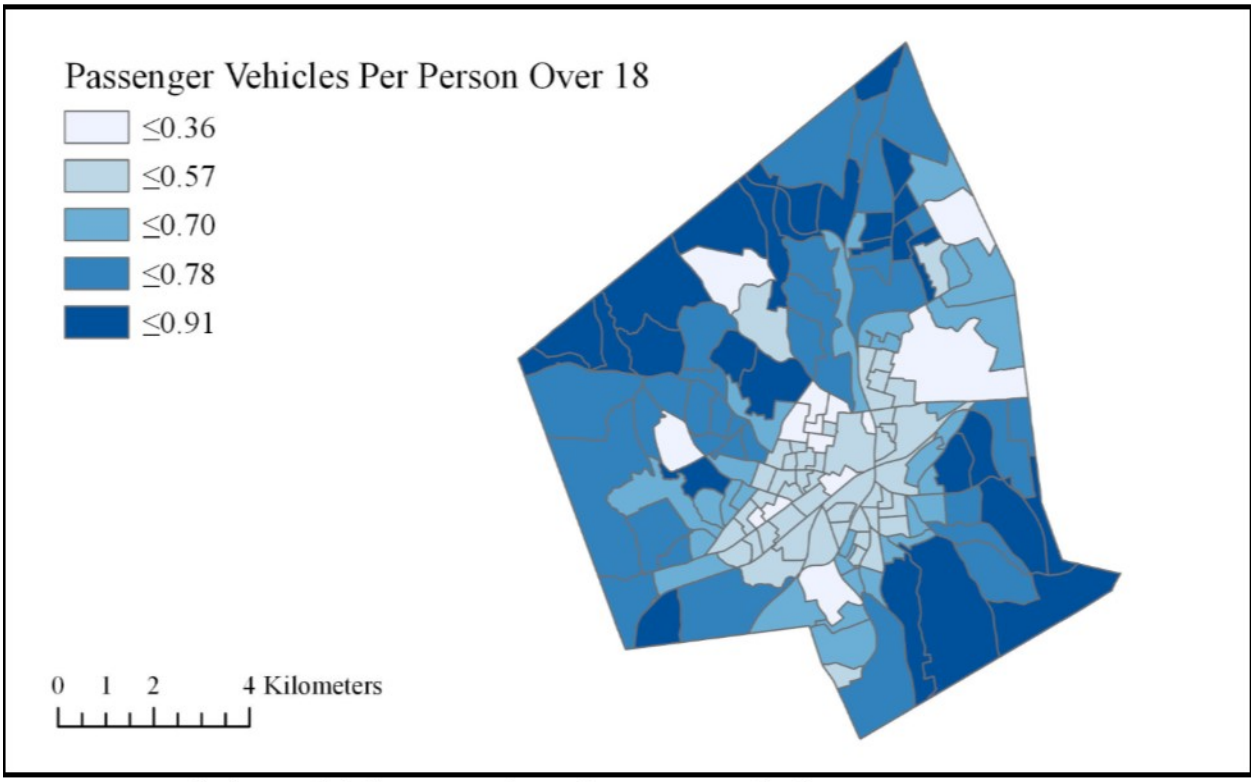

Figure 11. Vehicle availability as number of passenger vehicles per person over 18

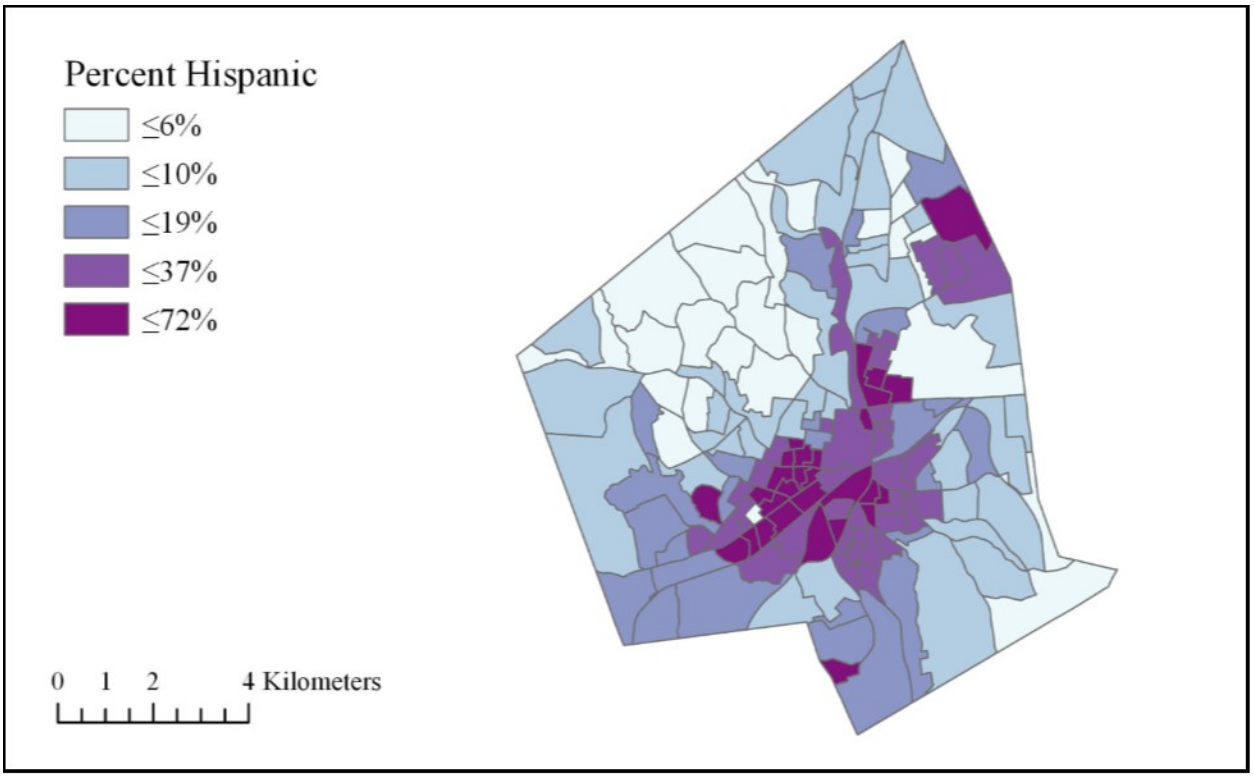

Figure 12. Percent of population identifying as Hispanic

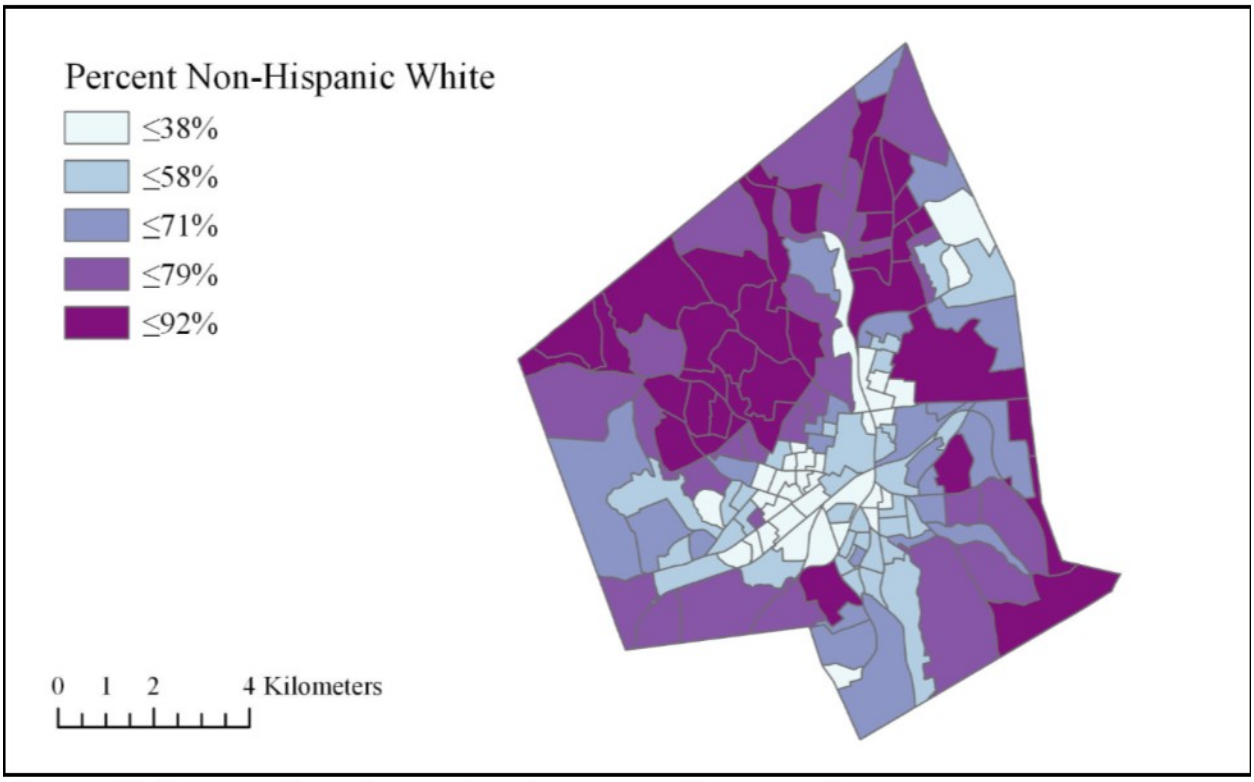

Figure 13. Percent of population identifying as non-Hispanic white

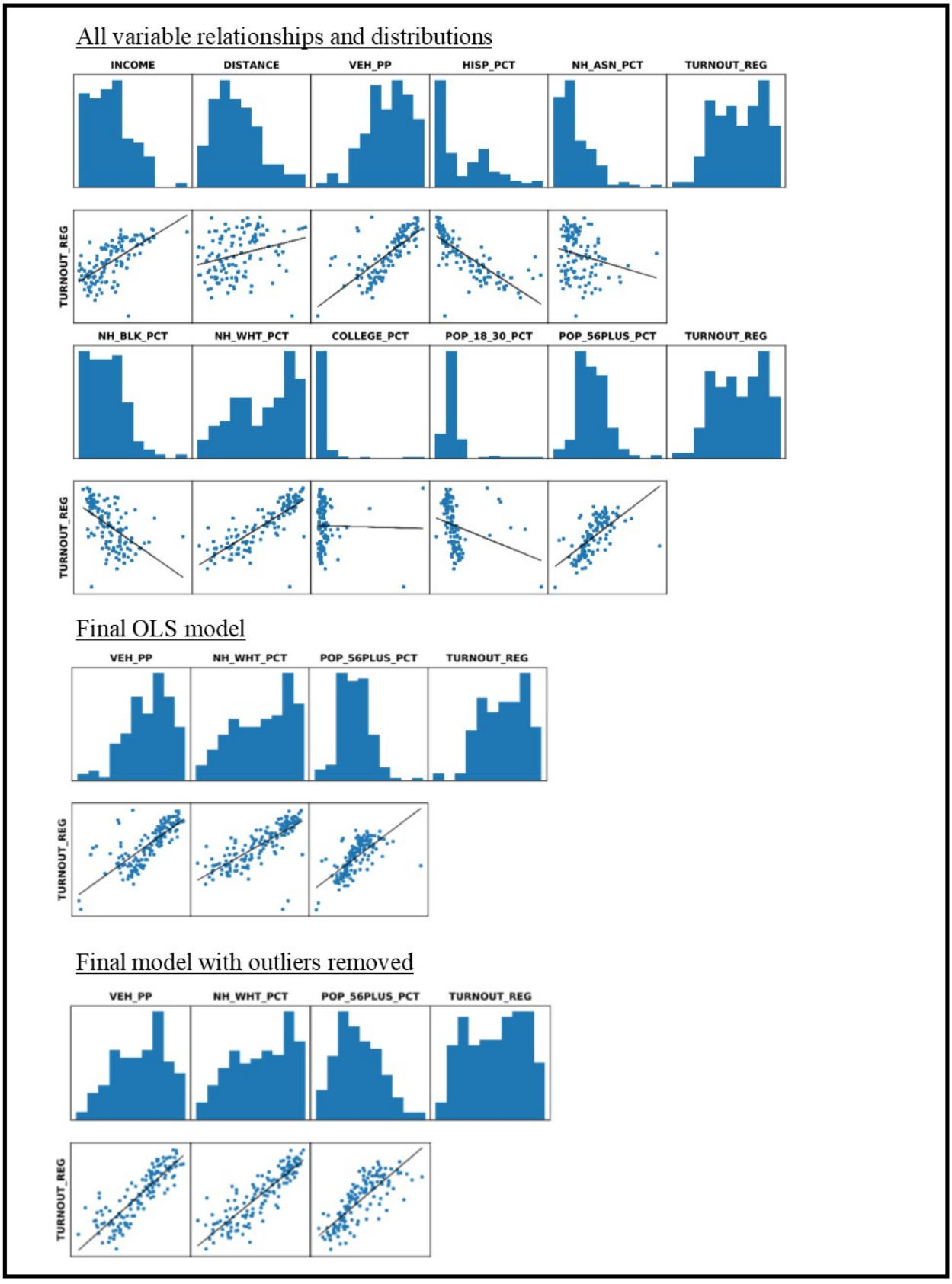

Figure 14. OLS model variable scatterplots and histograms

Despite the OLS model combining vehicles per person, percent non-Hispanic white, and percent age 56 and older as explanatory variables performing well, it still had a significant Jacques-Bera p-value. However, as was the case in previous combinations of explanatory variables, residuals in the model were normally distributed. The significant Jaque-Bera p-value was a result of outliers in the data. A second OLS model was created after removing five outlying data points from the vehicles per person variable, two from percent non-Hispanic white, and one from percent age 56 and older. This model achieved a Jacque-Bera p-value of 0.85, indicating that the data outliers were the cause of the significant p-value in the unaltered model. Although the Koenker (BP) statistic was significant in the unaltered model, the robust probabilities showed that all three variables remained significant. Both variations of the model exhibited spatial clustering of residuals when tested for Spatial Autocorrelation. The significance of the Koehker statistic and clustering of residuals strongly indicate that a local geographic model could better explain the relationship between the explanatory variables and voter turnout.

Multiple R-Squared	Adjusted R-Squared	AICc
0.787646	0.783252	-408.539450

Joint F-Statistic	Joint Wald Statistic	Koenker (BP) Statistic	Jarque-Bera Statistic
179.273721 ($p < 0.01$)	592.019605 ($p < 0.01$)	77.650987 ($p < 0.01$)	21.915083 ($p < 0.01$)

Global Moran's I z-score	Global Moran's I p-value	Result
3.126193	0.001771	Less than 1% likelihood clustered residuals are result of random chance

Variable	Coefficient	Std. Error	t-Stat	Prob	VIF
Vehicles per person	0.296456	0.034866	8.502623	$p < 0.01$	1.635112
Percent non-Hispanic white	0.253119	0.028798	8.789567	$p < 0.01$	1.596927
Percent age 56 and older	0.256901	0.059563	4.313081	$p < 0.01$	1.677926

Variable	Robust Std. Error	Robust t-Stat	Robust Probability
Vehicles per person	0.068706	4.314866	0.000033
Percent non-Hispanic white	0.058821	4.303180	0.000034
Percent age 56 and older	0.103485	2.482494	0.014179

Table 1. Ordinary Least Squares model statistics and scores

Geographically-Weighted Regression Model

GWR results showed that the significant variables in OLS tests were also significant when geographic weights were applied to the regression model. Various combinations of other independent variables in GWR models ruled out some variables that could have potentially been significant with geographic weights included. Distance to polling place showed insignificant coefficients, supporting the OLS results and implying that a significant geographic pattern is unlikely. Median income in GWR models also produced very low coefficients. Spatial autocorrelation in Global Moran's I showed no significant clustering of residuals.

R-squared	Adjusted R-squared	AICc	Moran's I z-score	Moran's I p-value
0.893973	0.869847	571.307556	0.409592	0.682105

Table 2. Geographically-weighted regression model statistics and scores

Local variable coefficients for vehicles per person, percent non-Hispanic white, and percent age 56 and older in the GWR model showed variations throughout the city (Figure 14). All three variables had maximum local coefficients close to 0.5, but the lowest coefficients were around 0.1. Local R-squared values ranged from 0.5-0.9 and show that the model performance varies by geography. The model performs best in the southern part of the city and northeast corner. It performs worst in the far west and northwest areas, as well as southeast of the city center (Figure 15). These maps were also categorized using geometric breaks.

The geographically-weighted regression model produced modestly higher R-squared and adjusted R-squared values than the OLS model, but also a higher AICc value. Most importantly, the Moran's I p-value was not significant, indicating that with geographic weights incorporated into the regression model, residuals were randomly distributed and the model was performing well despite potentially lacking additional explanatory variables.

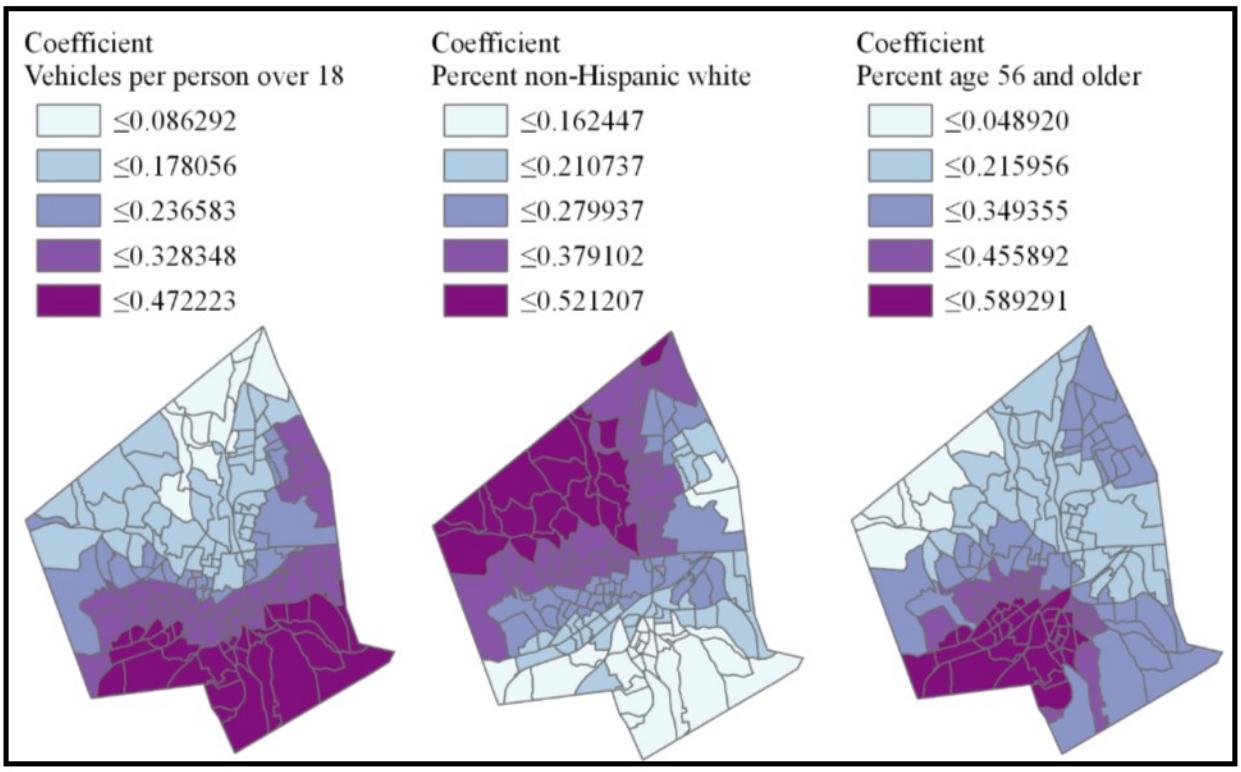

Figure 15. GWR local variable coefficients

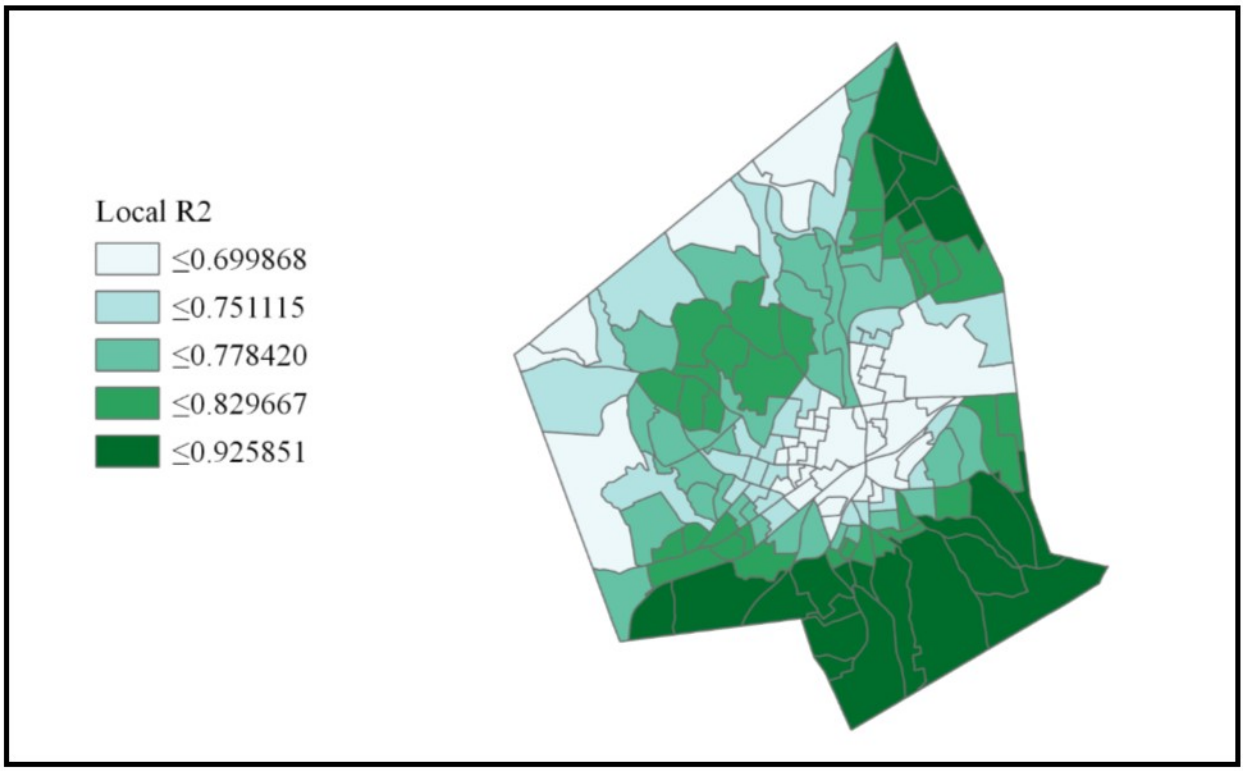

Figure 16. GWR local R-squared

DISCUSSION

Whether distance to polling place, as a measure of personal cost of traveling to the polls in time, money, and energy expended, exhibits a logarithmic relationship with turnout was shown to be the wrong question. Although previous studies corroborated that a significant logarithmic model predicts voter turnout based on voter distance to polling place, the geographic areas used in these studies may not have been representative of all geographies. The logarithmic model is explained by the rapidly increasing personal cost for people within walking distance to polls and a gradual increase in personal cost once the distance increases to those great enough that most people will drive a car. The explanation as to why a given model best fits a set of variables may also vary by cultural region. Indeed, this seems to be the case in Worcester. Average voter distance to polling place in Worcester lacked a significant linear relationship with voter turnout in OLS models. The distribution of voter distance data points showed that a logarithmic model would also be a poor fit. As it turns out, voter turnout in Worcester cannot be reliably predicted by how far someone has to travel in order to vote regardless of transformation.

The lack of significance which distance has in relation to voter turnout in Worcester is, in itself, a significant finding. First, the ability to rule out a variable of interest is valuable. Second, the question of *why* it is not significant is just as important. This question points to other variables of interest and may explain why other explanatory variables in this study had significant correlations to voter turnout. Of the independent variables considered in this study, access to vehicular transportation, race, and age were the most significant. It appears that having a car or cars in your family in Worcester is one of the biggest deciding factors as to whether or not you will go vote. A previous study evaluating vehicle availability along with travel distance found that having access to a vehicle increased the likelihood of voting, but at a certain distance vehicle ownership no longer made a difference (Haspel and Knotts, 2005).

Worcester's hilly, twisting and winding road network may reveal the explanation as to why vehicle availability was significant. Large portions of the city are not conducive to walking. It makes sense that cars are the preferred mode of transportation when it is frequently impossible to travel a straight-line distance to a destination. The regional attitude toward public transportation may play a role as well. Outside of New York City, the northeast is not known for exemplary public transportation. The car is king in Massachusetts, even if it means an hours-long

commute in heavy traffic on congested roads. It is not a stretch of the imagination to think that this attitude carries over to errands, shopping, or the occasional trip to the polls.

Areas with a high percent of non-Hispanic whites also showed strong correlation with higher voter turnout, and areas with a high percentage of Hispanics consistently exhibited a strong negative correlation with voter turnout. These are the primary racial/ethnic components in Worcester, as Hispanics numbered approximately 37,000 and whites numbered 107,000 in the 2010 census. These populations are also geographically segregated, which makes the relationship with turnout essentially the inverse of one another – so much so that they were statistically collinear. This confirms with statistical significance the findings of the earlier report by Conroy, Hansen, and Holbrook (2017) that Hispanics are voting much less than whites in Worcester.

Further Research

While this study has reiterated a racial divide in voter turnout, it has also highlighted the importance of how transportation relates to voter participation. Although distance itself may not be indicative of voting likelihood, access to polling place is likely significant because the availability of vehicle(s) was a significant predictor of voter turnout. This suggests that other variables which reflect polling place access, such as public transportation routes, parking availability, and polling facility capacity may also be significant indicators for voter turnout, and could be a fruitful area of further study. A broader analysis incorporating multiple transportation types could reveal finer details about the importance of travel cost. Vehicle access could be examined at a finer scale, such as by census block.

The effects of a disproportionate electorate could be evaluated based on city council representation and target specific communities with low participation. The implications of early voting, absentee voting, and same-day registration could be assessed. As in previous studies considering travel distance to polls and voter turnout, this study did not incorporate every possible explanatory variable. It is highly likely that additional variables could improve model performance, especially those relating to polling place access. It is also likely that some of these variables may be difficult to capture as data or are too abstract to easily quantify.

The methods and tools used in this study were based on analysis of relevant peer-reviewed literature and current best practices in GIS. The structure of the GIS used in this study

is such that data from alternative elections or geographies can be substituted for the election and location used here. The statistical models can be replicated using equivalent GIS software. Unexplored explanatory variables can easily be easily added and tested for significance. With access to GIS resources such as an address geocoding database and geostatistical analysis, the structure and workflow used in this study can be adapted to future research on variables which influence voter turnout.

Conclusions

Democracy is most representative when more people participate. Is Worcester proportionally represented by its major demographic components? The evidence suggests it is not. But voter turnout will not increase simply by understanding it. Those who are not informed or engaged must be educated and motivated. It is up to elections commissions, activists, community leaders, and politicians themselves to energize the under-engaged in the electorate. In a city with such a large Hispanic population, the lack of voter participation in this community means their voices are not being heard. In the democratic process, your vote is your voice, and disproportionate voting among racial or cultural groups means that the group voting in higher proportion will have more of a say. Disproportionate representation means that public policy in the state and the nation, but most directly in Worcester city government, will continue to favor older white people with cars who go out on election day and cast their vote.

REFERENCES

Blais, A. (2006). What Affects Voter Turnout? *Annual Review of Political Science, 9*, 111-125. doi: 10.1146/annurev.polisci.9.070204.105121

Brady, H., and McNulty, J. (2011). Turning Out to Vote: The Costs of Finding and Getting to the Polling Place. *The American Political Science Review, 105*(1), 115-134.

City of Worcester. (2017). *GIS Data - City of Worcester, MA*. Retrieved from worcesterma.gov: http://gisdata.worcesterma.gov/

Cohn, D. (2010, March 15). *College Students Count in the Census, but Where?* Retrieved from Pew Research Center: http://www.pewsocialtrends.org/2010/03/15/college-students-count-in-the-census-but-where/

Conroy, T., Hansen, W., and Holbrook, J. (2017). *A Study of "Eligible" Voters in Worcester, Massachusetts.* Worcester State University. Worcester, MA: Worcester State University.

DeSilver, D. (2017, 15 May). *U.S. trails most developed countries in voter turnout*. Retrieved from Pew Research: http://www.pewresearch.org/fact-tank/2017/05/15/u-s-voter-turnout-trails-most-developed-countries/

Encyclopedia Britannica. (2004, April 15). *Gerrymandering*. Retrieved from Encyclopedia Britannica: https://www.britannica.com/topic/gerrymandering

ESRI. (2018, April). *ArcGIS Pro Help*. Retrieved from arcgis.com: http://pro.arcgis.com/en/pro-app/help/main/welcome-to-the-arcgis-pro-app-help.htm

Gerber, A., Green, D., and Larimer, C. (2008). Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. *American Political Science Review, 102*(1), 33-48. doi:10.1017/S000305540808009X

Geys, B. (2006). Explaining voter turnout: A review of aggregate-level research. *Electoral Studies, 25*(4), 637-663. Retrieved from https://doi.org/10.1016/j.electstud.2005.09.002

Gimpel, J., and Schuknecht, J. (2003). Political participation and the accessibility of the ballot box. *Political Geography, 22*(5), 471-488. doi:10.1016/S0962-6298(03)00029-5

Gimpel, J., Dyke, J., and Shaw, D. (2006). Distance, Turnout, and the Convenience of Voting. *Social Science Quarterly, 86*(3), 531.

Haspel, M., and Knotts, H. (2005). Location, Location, Location: Precinct Placement and the Costs of Voting. *The Journal of Politics, 67*(2), 560-573. doi:10.1111/j.1468-2508.2005.00329.x

MAPC Boston. (2018). *Open Data*. Retrieved from Metropolitan Area Planning Council: https://www.mapc.org/learn/data/

MassGIS. (2017). *MassGIS Data Layers*. Retrieved from mass.gov: https://www.mass.gov/service-details/massgis-data-layers

McNulty, J., Dowling, C., and Ariotti, M. (2009). Driving Saints to Sin: How Increasing the Difficulty of Voting Dissuades Even the Most Motivated Voters. *Political Analysis, 17*(4), 435-455.

Sui, D., and Hugill, P. (2002). A GIS-based spatial analysis on neighborhood effects and voter turn-out: a case study in College Station, Texas. *Political Geography, 21*(2), 159-173.

U.S. Census Bureau. (1994, November). *Geographic Areas Reference Manual.* Retrieved from United States Census Bureau: https://www2.census.gov/geo/pdfs/reference/GARM/GARMcont.pdf

U.S. Census Bureau. (2016, March 10). *Census in the Constitution*. Retrieved from United States Census Bureau: https://www.census.gov/programs-surveys/decennial-census/about/census-constitution.html

U.S. Census Bureau. (2018). Retrieved from American FactFinder: https://factfinder.census.gov/

U.S. Census Bureau. (2018). *About Congressional Districts*. Retrieved from United State Census Bureau: https://www.census.gov/geo/maps-data/data/aboutcd.html

Worcester Historical Museum. (2018). *Transportation*. Retrieved from Worcester Historical Museum: http://www.worcesterhistory.org/worcesters-history/worcester-in-the-19th-century/transportation/