Assessing the impacts of bicycle infrastructure on the willingness to commute via bicycle

by

Tyler W. Tripp

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF REGIONAL AND COMMUNITY PLANNING

Department of Landscape Architecture and Regional and Community Planning
College of Architecture, Planning & Design

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2023

Approved by:

Major Professor
Gregory Newmark

Copyright

Abstract

Past research has been conducted into factors that guide the travel mode choice of individuals through surveys, usership data, simulations, etc. Studies which focus on mode choice factors have often used surveys. Due to these survey-based methods, there are many sources of potential bias that may inflate the importance of factors and/or the willingness of respondents to change existing behaviors. The following study aims to reduce potential bias by using publicly available Public Use Microdata Samples (PUMS) data as well as a variety of spatial/temporal data. PUMS data is used to determine the effects of demographic factors on the probability of biking, while spatial/temporal data is used to determine the effects of population density, employment density, median property value, median housing cost, median income, crime rate, and bike accessibility. The bike accessibility measure is based on 20-minute isochrones with impedance correlating to the proportion of the population willing to use a particular type of infrastructure. The data used within the study represents 17 unique Public Use Microdata Areas (PUMAs) within Chicago, Illinois over the course of eight years (2012-2019).

This study produces a logistic regression model that estimates the probability of choosing to bicycle to work given a variety of demographic and environmental factors. Based on the model's output, it was shown that one-point bikeability index increase correlates with a 0.33-percent increase in the likelihood of an individual choosing to bike. Men were to be 2.61 times more likely to bike to work than women. Contrary to prior research, black/African American and Hispanic persons were found to be less likely to bike to work than the rest of the population, possibly due to local variation in the effects of demographics or unequal distribution of infrastructure. Age had a negative correlation with likelihood of biking to work and education had a positive correlation.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank several individuals for their assistance in guiding me through this process. Firstly, I would acknowledge Professor Gregory Newmark, who has been very helpful in guiding my progress as my Faculty Advisor. You have been a tremendous resource for me given the extensive experience you have guiding other students through transportation-related projects. You have also been very helpful in making sure that I had resources and opportunities that allow me to focus fully on my research. I would also like to thank the other two members of my advisory committee, Professor Shakil Kashem and Professor Susmita Rishi, for their patience and assistance with aspects of my project. Both of you have been invaluable resources. I would also like to thank Professor Husain Aziz, who's own research is referenced in my work, for his willingness to meet with me and provide assistance in determining appropriate statistical methods for my research. Finally, I would like to thank my parents, Melissa and Dale Tripp, for being supportive of me throughout this process and always being a second and third set of eyes for reviewing my document as it has come together. Thank you all; I would not have been able to produce a product of this quality without help from all of you!

Introduction

It is widely held that alternative transportation, and in particular, active transportation, has many benefits including reduction of greenhouse gas emissions, reduction of vehicle congestion, increased physical activity, reduced infrastructure costs, reduced transportation costs, etc. Past research has sought to quantify the effects of shifting mode choice towards active transportation, some even quantifying benefits in an economic perspective (Xia et al., 2013). Due to these known benefits, research has been conducted into the primary factors that guide the travel mode choice of individuals through surveys, usership data, simulations, etc. Largely, studies have focused on demographics, attitudes, and environmental characteristics, but lack information on how infrastructure contributes to ridership, particularly when it comes to cycling. My research aims to include demographic and environmental characteristics as well as a direct measurement of the accessibility of jobs via bicycle infrastructure. By adding and focusing on a bicycle accessibility index (bikeability index), I aim to provide a tool to guide investment decisions and predict the overall change in ridership based on changes to the built environment. In doing this, I provide a repeatable procedure that can be used at the local, regional, or national level to guide local decision making and produce a more robust model in the future.

The other significant aspect of my research is the reduction of potential survey bias present in mode-choice studies. Most past research focusing on bicycle-mode-choice factors have utilized surveys asking respondents rate the importance of mode-choice factors. Due to these survey methods, there are many sources of potential bias in these studies that may generate results that may skew the importance of various factors. My study aims to reduce potential bias by using publicly available Public Use Microdata Samples (PUMS) as well as a variety of spatial/temporal data to represent bicycle use trends given a variety of factors, including

bikeability. The surveys used in this study (including PUMS and the American Community Survey (ACS)) rely on questions of fact rather than opinion-based questions and are therefore less susceptible to bias. To guide my analysis, I have reviewed existing literature related to bicycle-mode-choice with particular focus on research that relates to the built environment. Additionally, I reviewed literature to assist in creating an appropriate measure of bikeability.

PUMS data is used to determine the effects of demographic factors on the probability of biking, while spatial/temporal data is used to account for environmental characteristics. Spatial/temporal data includes population density, employment density, median property value, median housing cost, median income, crime rate, and a bicycle accessibility index based on 20-minute isochrones weighted by infrastructure type with impedance representing infrastructure desirability. The data used within the study represents 17 unique Public Use Microdata Areas (PUMAs) within Chicago, Illinois over the course of eight years (2012-2019). The final output of this study is a logistic regression model estimating the probability of biking work given several demographic and environmental factors, with particular focus on the bikeability index.

Based on the model's output, it was shown that a one-point change in the bicycle accessibility index contributes to a 0.33-percent increase in the likelihood of an individual choosing to bike to work. A test of this shows that strategic investment can lead to tens or hundreds of new cyclists with investments in short paths. Additionally, it was determined that men were found to be 2.61 times more likely to bike to work than women. Contrary to prior research, black/African American and Hispanic persons were found to be less likely to bike to work than the rest of the population, possibly due to local variation in the effects of demographics or unequal distribution of bikeable infrastructure. Age was found to have a negative correlation with likelihood of biking to work and education had a positive correlation.

Review of Existing Literature

A substantial amount of research has been done regarding mode-choice decisions of people who use alternative transportation. One study summarizes that for studies that focus on mode-choice factors "findings are generally obtained from cross-sectional studies, intervention studies, stated preference studies, revealed preference studies and summarized meta-analysis" (Hardinghaus et al., 2021). In my review, I found that the primary method that has been used to determine mode-choice factors has been analysis of surveys that ask the respondent what factors are important to them (Guinn & Stangl, 2014; Handy & Xing, 2011; Hardinghaus et al., 2021; Heinen et al., 2011; Piatkowski & Marshall, 2015; Ye et al., 2020). Furthermore, most studies either omitted infrastructure considerations, based this factor on self-reported opinions of respondents, or included infrastructure minimally as a control for other factors (Guinn & Stangl, 2014; Handy & Xing, 2011; Heinen et al., 2011; Li et al., 2013; Ye et al., 2020). Since infrastructure is a direct factor that governments can control, this is the primary variable of interest for my research. Past research has found a positive correlation between both bike lanes per square mile and per capita spending on bicycle infrastructure and bicycle commuter share (Dill & Carr, 2003). This same study recommended additional research to explain the relationship between cycling and infrastructure.

There have been many studies that attempt to analyze the mode-choice of commuters. A commonality between these studies has been the inclusion of socio-demographic characteristics (Aziz, Nagle, et al., 2018; Aziz, Park, et al., 2018; Dill & Carr, 2003, 2003; Guinn & Stangl, 2014; Handy & Xing, 2011; Heinen et al., 2011; Li et al., 2013; Piatkowski & Marshall, 2015; Pritchard et al., 2019; Song et al., 2017; H. Wang et al., 2016; Ye et al., 2020). Many of these studies include these characteristics as controls when analyzing other factors as is the case in my

study. Commonly included factors include gender/sex, race/ethnicity, age, home ownership, car ownership, income, education, etc. All sources that included gender as an explanatory variable found men more likely to bike than women (Aziz, Nagle, et al., 2018; Handy & Xing, 2011; Li et al., 2013; Piatkowski & Marshall, 2015; Song et al., 2017; Ye et al., 2020). The effects of age varied in several studies. Ye et al. (2020) found a positive correlation between age and biking, whereas Aziz, Nagle, et al. (2018) found a positive correlation for individuals under 25 and a negative correlation for those over 65. Finally, Li et al. (2013) found the highest probability of cycling at ages 20-40. Education had a positive correlation in two studies (Piatkowski & Marshall, 2015; Song et al., 2017). Income was found to have a negative correlation with biking in two of the reviewed studies (Piatkowski & Marshall, 2015; Ye et al., 2020). Car availability was negatively correlated with biking (Piatkowski & Marshall, 2015).

Several studied covered perceptions and attitudes of cyclists were covered in several of the studies as explanatory variables. This is potentially beneficial it can explain the variation between the mode choice of people with similar or identical demographics (Aziz, Nagle, et al., 2018; Handy & Xing, 2011; Piatkowski & Marshall, 2015). Though there are potential issues with the subjectivity in opinion/perception (discussed later), there is also the consideration that choosing to bike is often a subjective choice. There are exceptions to this, as low-income individuals may not have vehicles, but when considering how to get more people to bike, the consideration of subjective decision making is important. Heinen et al. (2011) focus their attention on opinions and attitudes associated with various aspects of cycling to work including factors such as perceived status effect, environmental benefits, relaxation, lifestyle suitability, etc. In total, they identified 14 factors which respondents rated on a Likert scale to analyze individual attitudes toward cycling. Unsurprisingly, they found that people with more positive

perceptions/attitudes towards the various aspects of cycling to work were significantly more likely to bike to work (Heinen et al., 2011). The primary effects on ridership were (broadly) based on direct trip-based benefits, awareness (of less direct benefits such as environmental benefits and physical health), and safety. They also found positive correlations between habit, the subjective norm, and perceived behavioral control and the likelihood of cycling to work. Finally, they determined that most of the explanatory power of the model rested with perception and attitude factors rather than socio-demographic characteristics. Guinn & Stangl (2014) surveyed individuals and asked them to rank factors that determine whether to use active transportation. The study found that buffering between cyclists and vehicles and the presence of bike lanes the fourth and fifth most important factors out of a total of 14 options, coming after exercise, environment, and "visually appealing".

Individual perceptions/attitudes are often influenced by the social environment. Though there has not been substantial research regarding social environment influences, two studies did include some consideration for these effects. Handy & Xing (2011) included a "supervisor disapproval" variable in their analysis, which was only significant at $\alpha = 0.10$. This begins to demonstrate some of the effect that the social environment can have. Aziz, Park, et al. (2018) takes a unique approach to consideration of the social environment. Rather than relying on surveys, they used a simulated population which considers potential interpersonal interactions at home and at work which influence mode-choice decisions. The agent-based model starts with probabilities of cycling using socio-demographic, built environment, and perception factors similar to other studies, but then runs a simulation of social interactions. They then ran scenarios including widening of sidewalks and lengthening of bike lanes, showing an increase in active transportation from these interventions. This model is likely more robust than other methods of

prediction due to the consideration of the social environment, though adding other potential social interactions such as social media could improve it further (Aziz, Park, et al., 2018). Though my model will not include social environment factors, these factors can explain much of the variation in mode choice between people of similar socio-demographic backgrounds.

As stated earlier, many studies omit built environment factors or include them minimally. Studies that specifically focused on the effects of bicycle infrastructure include Aziz, Park, et al. (2018), Aziz, Nagle, et al. (2018), Pritchard et al. (2019), Song et al. (2017), and H. Wang et al. (2016). Studies that included infrastructure as a supplemental factor include Handy & Xing (2011) and Piatkowski & Marshall (2015). Guinn & Stangl (2014), though not measuring built environment factors, shows that they are important mode-choice factors. Piatkowski & Marshall (2015) found a statistically significant correlation between the likelihood of biking and trip distance, origin link-to-node ratio, security and comfort, and relative convenience. Other built environment characteristics and attitude/perception variables did not appear statistically significant. In this instance, built environment characteristics were overly divided into individual components which left little ability to interpret the overall effects of built environment improvements. The only considerations in Handy & Xing (2011) were distance from work, parking cost, and whether biking to work was perceived as dangerous. Therefore, very little can be inferred from this study in terms of the effects of the built environment.

Aziz, Park, et al. (2018), Aziz, Nagle, et al. (2018), Pritchard et al. (2019), Song et al. (2017), and H. Wang et al. (2016) focused specifically on the effects of bicycle infrastructure. Aziz, Nagle, et al. (2018) found that a one-percent increase in the total bike lane proportion was correlated with a 1.13% increase in the probability of someone choosing to bike by using a random parameter (or mixed) logistic. Aziz, Park, et al. (2018) again found a correlation

between active transportation infrastructure built and the number of people biking by simulating an increase in width of sidewalks and length of bike lanes by 7%, 10%, and 12. Since this study is simulation based, the validity of simulation and the reasonableness of the assumptions made will affect the outcomes. Song et al. (2017) found that people are more likely to continue using active transportation after having used it before within their community. Unfortunately, this study was unable to establish a statistically significant relationship between passive exposure (i.e. proximity to new infrastructure) and modal shift towards active transportation. Another study, Wang et al. (2016), tested bicycle network level of traffic stress (LTS) against the number of trips taken by bike. Though this study is primarily intended to test LTS as a measurement/prediction tool, they found a negative correlation between LTS and the number of bicycle trips taken. Again, since LTS is a measure of built environment characteristics, it can be inferred that improvements in the built environment are positively correlated with bicycle mode choice based on these results. Pritchard et al. (2019), was unable to show a statistically significant increase in bicycle mode share following the addition of a 400-meter contraflow bike lane. As the authors state, this is likely due to the small sample size of the study (n = 39), and/or the minimal scale of the intervention. This shows that future studies should strive to have large sample sizes and include more and/or larger interventions.

As previously mentioned, many of the analyzed studies are survey-based. Surveys are a rational method for obtaining data about important factors for people when considering using active transportation because the data is received directly from those people. There are, however, several challenges when utilizing this type of data: the primary challenge being sources of bias. First, there is potential bias caused by the surveying procedure itself. Two studies targeted bicyclists and pedestrians for their surveys (Guinn & Stangl, 2014; Piatkowski &

Marshall, 2015). In Guinn & Stangl (2014), surveyors identified four locations with high levels of bicycle and pedestrian traffic and asked respondents to complete a written or online survey. In Piatkowski & Marshall (2015), they surveyed individuals who participated in Bike to Work Day (BTWD) and had signed up to receive more information via email. In both instances, the surveys capture individuals that have used active transportation and would be more likely to do so in the future. Because of these methods, they do not obtain feedback regarding what factors may contribute to non-cyclists and non-walkers choosing to bike/walk in the future. Targeting cyclists and pedestrians keeps surveying costs low, but "… this method only provides a singular point of view… [Automobile drivers] may respond differently to the motivating factors" (Guinn & Stangl, 2014). Even in studies where active transportation users are not specifically targeted, more respondents can be expected from active transportation users if they know that the survey pertains to active transportation. One study found that "although [they] designed the survey to be relevant to all individuals, not just bicyclists, it is possible that individuals who do not bicycle were less inclined to complete the survey" (Handy & Xing, 2011).

Another potential source of bias comes in survey responses. These surveys rely on respondents to accurately represent their priorities when considering mode choice. The key form of bias in these types of surveys is social desirability or conformity bias, or the desire to answer such that the respondent is seen in a positive light. In Piatkowski & Marshall (2015), respondents are asked to rate barriers to cycling. The primary barrier according to this survey was "too few bike lanes" while one of the least influential was getting too sweaty. Though it is possible that presence of bike lanes *is* the most influential reason, it could also be reasoned that other factors (such as getting too sweaty) are more important but access to additional bike lanes was the response that would be seen in the best light by the surveyor. Though the results of this

study may still be accurate, this presents the importance of confirming ridership behavior using other methods. Other studies that rely on perception surveys are similarly susceptible to conformity bias, though reviewed studies did generally avoid said bias through specific phrasing of questions (Guinn & Stangl, 2014; Handy & Xing, 2011; Heinen et al., 2011; Ye et al., 2020). A similar cause of bias is identifiable in Handy & Xing (2011) where they ask for the respondents to provide their "*perceptions* of the availability of safe routes to the work destination." This type of questioning is present in several of the studies. Because of the questions about the perceptions of the built environment, they may miss out on the reality of what the built environment contains based on respondents' levels of exposure.

Time is an important consideration in changing mode choice. Attitudes and perceptions of cycling are likely to change as factors in the physical and social environments change. Most reviewed studies that were cross-sectional, focusing on one moment in time. In particular, survey-based studies tend to be cross-sectional (Aziz, Nagle, et al., 2018; Guinn & Stangl, 2014; Handy & Xing, 2011; Heinen et al., 2011; Ye et al., 2020). This makes sense, as 1) surveys are both expensive and time consuming to conduct, 2) for survey-based studies to be longitudinal the output would be unavailable until the study period is complete, and 3) it is unlikely in most cases that the same respondents will be available to repeat the survey at a later date, not that that is always a requirement. Only one of the survey-based studies includes data intentionally drawn from two years, however they stated that it was not a panel study, and they made no distinction between the years (Piatkowski & Marshall, 2015). In addition to survey-based studies, other studies, even those based on publicly accessible data, are also often cross-sectional. Though there are cost benefits to the cross-sectional design for studies, drawbacks are significant. Song et al. (2017) explains that a shortcoming of cross-sectional studies is that they "[limit] their

capacity to support causal inference about the effectiveness of the interventions." Particularly in the case measuring the effects of built environment characteristics, it is worthwhile to include a measured period of time, as other factors may cause difference in use trends in different communities. Handy & Xing (2011) agree with this point stating that "if an individual lives in a community with a strong bicycle culture and with good bicycle infrastructure, his preferences for bicycling increase over time." Song et al. (2017) surveyed the same respondents before and after the addition of new bicycle infrastructure to determine if there was a causal relationship and found a modal shift towards cycling after installation. Similarly, Pritchard et al. (2019) surveyed the same group of respondents twice (before and after the addition of a new bike lane) to report on their travel behavior over four weeks and supplemental GPS data from participants cellphones. Unfortunately, this study was not able to show a statistically significant modal shift, likely due to a small sample size.

This review gave me several insights about how my own research should be conducted. First, I intend to avoid surveys that utilize opinion-based responses. In some ways my study is limited by not including individual attitudes and perceptions in the model. There are two main reasons why I have selected to omit these factors, however: 1) using Public Use Microdata Sample (PUMS) data, there is no feasible way to attach attitudes and perceptions to individual records; and 2) the goal of my research is to *verify* that increased infrastructure leads to a higher probability of an individual cycling without relying on their subjective indications their priorities. The second major consideration for my own research is to make sure to include a direct measure of infrastructure since most existing studies do not. Finally, my study will use a contiguous dataset containing multiple years to better establish a causal relationship. I also must make sure to have a sufficient sample size and many interventions of various scales.

Creating a Decentralized Transportation Accessibility Index

In creating the final predictive model, an issue arose in determining the best way to quantify the quality of the bicycle network. Additionally, there is no well-accepted universal definition of bikeability (Kellstedt et al., 2021). In Kellstedt et al. (2021) existing bicycle index methods were categorized. They identified 17 studies, in which five used self-report surveys, five used geospatial data, and seven used some form of direct observation. Due to the subjective nature of self-report surveys (discussed earlier), the challenges in conducting a survey of this scale, and the desire to validate use trends on the built environment, this method will not be utilized. Similarly, time constraints will not allow for a full analysis of Chicago's streets for the eight years of my study, and Google Street View does not have data that is updated annually for every street. This leaves the final possible method, use of geospatial data. Fortunately, the City of Chicago provided bicycle route data and public street data can complete the network.

Since use of geospatial data is most appropriate for this study, it is important to determine how to create a decentralized accessibility index. With that said, one method for creating the index is the suitability analysis method. In this method, a variety of factors which contribute to the suitability of the environment for a particular use are given weights and combined into a standardized index. Replication of the U.S. Census' National Walkability Index was considered, due to the widespread use and the accessibility of the methods to recreate this type of index. This method combines measures of intersection density, proximity to transit stops, and diversity of land uses (Thomas et al., 2014). Though this method provides a general sense of the transportation quality, there are several reasons that the suitability analysis method is insufficient for this study. For one, proxy measures of accessibility such as employment mix are similar to factors that already appear in my regression model, and more direct measures of job accessibility

are desired. Bike Score® is another suitability-analysis-based method. This index combines three major factors in their analysis; a bike lane score, a hill score, and a destinations and connectivity score (Kellstedt et al., 2021). Though this method is not tied to origin-destination pairs, it does capture a general concept of connectivity using a distance decay function and considering land uses that a cyclist may need to commute to (Walk Score, 2021). Due to the ease of production of suitability indices, this method was used in many of the reviewed studies that used a walkability/bikeability index in their analysis (Al Shammas & Escobar, 2019; Arellana, Saltarín, Larrañaga, Alvarez, et al., 2020; Arellana, Saltarín, Larrañaga, González, et al., 2020; Frank et al., 2010; Hardinghaus et al., 2021; Kellstedt et al., 2021; Thomas et al., 2014).

The other major method for creating a decentralized transportation accessibility index was using origin-destination transportation modeling procedures. Often, this type of analysis is done from one or a few locations (often transit stops) to determine accessibility to those particular locations. This type of analysis usually determines the total number of residents or the number of jobs/points of interest within walking/biking distance of a transit. For example, Cyril et al. (2019) determined the accessibility from transit stops in Trivandrum, India to identified points of interest (POIs) and weighted them by the distance from the stops to determine accessibility values for each stop. This method is promising, though some method of decentralizing the indices would need to be added.

Other studies determine accessibility from many origin points, sometimes aggregating this data. One of the more complex and accurate methods for this process is the calculation of level of access between origin-destination pairs (Gholamialam & Matisziw, 2019; Mavoa et al., 2012; Mohammad et al., 2020; Reggiani et al., 2022; Ryu et al., 2021; C.-H. Wang & Chen,

2015). This is highly effective since it provides measures for every point-to-point combination, rather than determining how many points are within a certain travel time/distance. Several studies have multi-criteria models which consider different priorities for prospective cyclists (Gholamialam & Matisziw, 2019; Reggiani et al., 2022; Ryu et al., 2021). These studies produce several potential paths between origin and destination based on comfort level of the cyclist or other prioritizing factors. All of these paths are then used to produce the accessibility index. This method, though very robust, is perhaps one of the most complex methods for generating a bicycle accessibility index.

Alternatively, a method similar to that used in Cyril et al. (2019) can be used for many points rather than just a few. Mohammad et al. (2020) provides an example of aggregating land uses within a 300-meter buffer (reasonable walking distance). By combining and adapting these two methodologies, a service area representing bikeable infrastructure can be generated. My research uses this combined method to select all destinations within a reasonable biking distance from a given origin, repeated for every origin in my data. This process is easier to execute than origin-destination pairing, and a finer spatial resolution (more points) can be used as a result. Additionally, this method will produce a more direct measure of bikeability in the context of job accessibility than would result from creating a suitability-type index. The process used to create the index, including methods for weighting infrastructure types for service area generation, is further described in the "Methods" section of this paper.

Methodology

This study uses a variety of demographic and spatial variables from various sources to create a binary logistic regression predicting the probability of an individual choosing to bike to work. Chicago is the primary study area. I use individual demographic data and household data from Public Use Microdata Samples (PUMS) and spatial data from the American Community Survey (ACS) Cook County Assessors' Assessments, City of Chicago Crime Data, U.S. Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) Work Area Characteristics (WAC) Data, and a bicycle accessibility index created using City of Chicago Bike Routes Data and TIGER/Line Roads Data. A PUMA_Yr variable was created and attached to each of the records within the PUMS dataset in order to add spatial/temporal data from other datasets. Each of the spatial datasets was then aggregated to the proper PUMA and year and given its own PUMA_Yr identifier.

To measure bicycle accessibility, an index was created using roads and bicycle routes. Each segment was given a speed limit based on the desirability of the infrastructure (i.e. using speed as a proxy for infrastructure desirability). 20-minute isochrones (time-based service areas) were created for each block group and average jobs per person within the isochrones were aggregated to the PUMA and year of their origin.

Each of the spatial/temporal variables were merged with the PUMS data based on their PUMA_Yr. All variables were then used to create a binary logistic regression with a dependent variable determining the likelihood of an individual choosing to bike to work. The variables in the model were transformed and variables that were not shown to be statistically significant were removed, yielding a final model. Model assumptions were then tested to determine its effectiveness and accuracy.

Site Selection

Though this research has been designed so that it may be repeatable in other communities, Chicago, Illinois was selected as the primary site for this research for several reasons. First is the generally high accessibility of data. Bicycle routes with the years of installation are required and the City of Chicago was able to provide this data. Additional spatial data was available through the Chicago Data Portal. Additionally, since Chicago has a high population density, publicly accessible PUMS data was available at a finer spatial resolution. Additionally, Chicago had good data continuity in terms of regular addition of new bicycle infrastructure.

Many projects have been implemented due to high levels of advocacy from organizations such as the Active Transportation Alliance and a favorable political climate for active transportation in this area. The Mayor of Chicago between 2011 and 2019 was Rahm Emanuel, an advocate of bicycle transportation and a cyclist himself. Emanuel's administration implemented the Divvy bikeshare program in 2013. Additionally, during his administration, the City of Chicago "built or renovated 40 L stations, finished the new-and-improved Riverwalk, revamped the 18-mile Lakefront Trail, and approved 200 miles of bike lanes" (Smith, 2019). The Active Transportation Alliance, too, has been influential during the period of the study, lobbying politicians including Emanuel, and getting him to agree to install 100 miles of bike lanes in his first term (Active Transportation Alliance, 2021).

Certain aspects of Chicago may lead to different results than might be seen around the rest of the country. For one, Chicago is well known for its inclement weather conditions and winters that are harsh enough to be an obstacle/deterrent to many cyclists. Additionally, local variation in demographics compared to the rest of the U.S. may lead to results that are different

from our typical understanding given the demographics of the U.S. as a whole. Some of the effects of these variations can be observed later in my study. Chicago has a poverty rate of 17.3% compared to an 11.6% national average, putting it at nearly 1.5 times the national average (U.S. Census Bureau, 2020a, 2020b). Chicago's climate includes average highs in the low 80's in the summer and around freezing in winter. Chicago is also relatively flat, making it ideal for cycling.

Enumeration Units (Spatial Resolution)

Data in this study is processed at the Public Use Microdata Area (PUMA) level. PUMAs are constructed from Census tracts and contain approximately 100,000 residents (at the time of their creation). There are two reasons that this unit is being used: 1) individual records from Public Use Microdata Samples (PUMS) are only available at the PUMA level; and 2) smaller spatial resolutions would likely not have enough (or any) cyclist records to produce meaningful interpretations of the data. Chicago contains 17 PUMAs (seen in Figure 3.1). The mean area of these 17 PUMAs is 13 square miles.

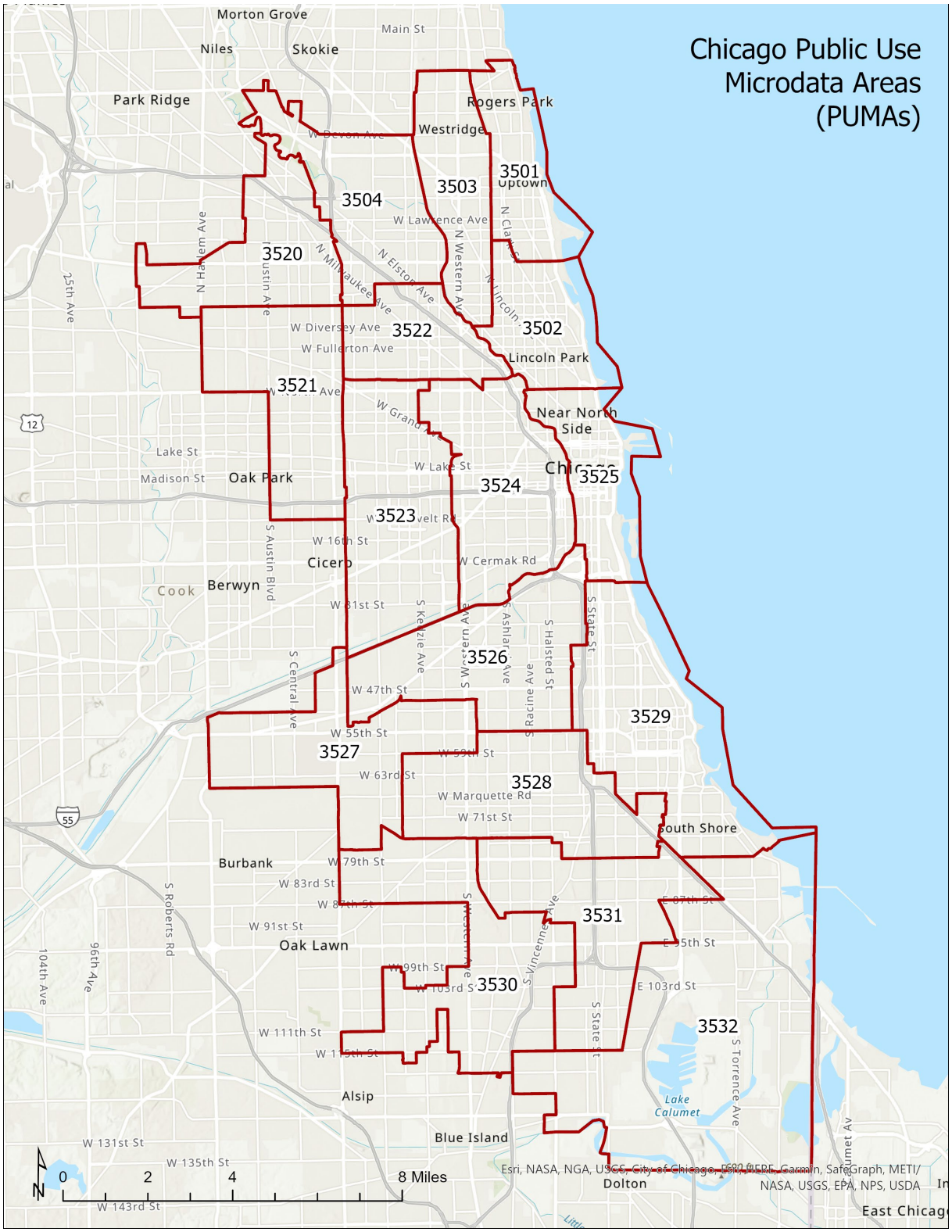


Figure 3.1. Chicago PUMAs Reference.

This map shows the locations of the 17 Public Use Microdata Areas (PUMAs) within the study area of Chicago, Illinois, as well as the underlying road infrastructure.

Data

Public Use Microdata Samples (PUMS) are the primary data used for demographics. This data is aggregated to create the results for the U.S. Annual Community Survey (ACS). PUMS data is at the individual record level which is critical for the analysis done in this study. The PUMS data provides both the primary transportation mode that the individual uses to travel to work, and the socio-demographic data used in this study. PUMS data was subset to not include years outside of 2012-2019, records outside of the study area, records indicating unemployment (using the *EMP* variable), and records indicating a 0-minute commute time (working from home). Figure 3.2 shows the percent of the population choosing to bike to work over time based on PUMS records. Figure 3.3 shows the average percent of the population choosing to bike to work for each PUMA over all years in the study with existing infrastructure overlayed. Both figures use person's weight, *PWGTP*, in the calculation to ensure an accurate account. Figures 3.4-3.11 (at the end of this section) show the spatial trends of key spatial variables used in this analysis.

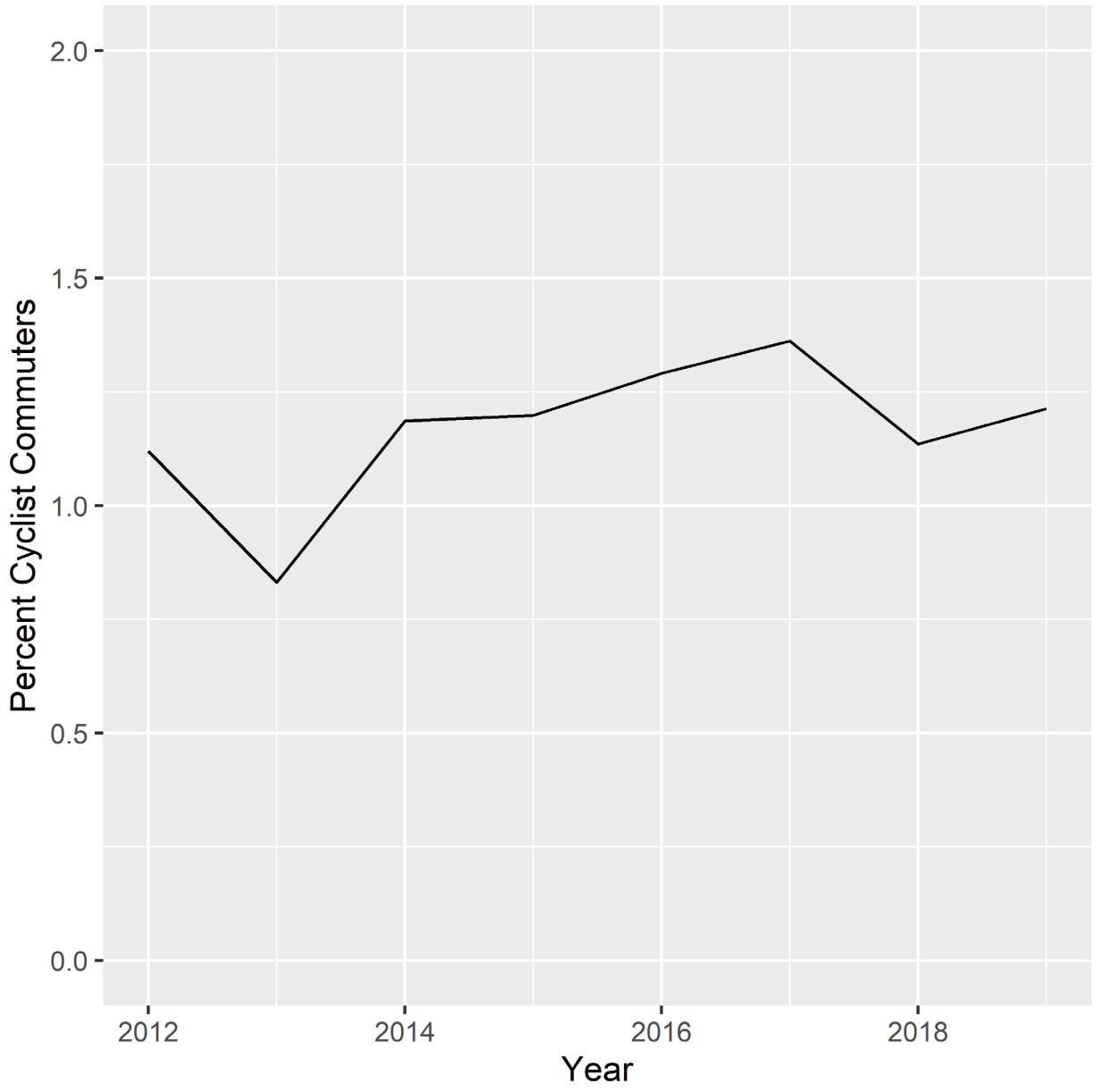


Figure 3.2. Percent of Population Biking Over Time

This graph shows the change in bicycle mode share within the study area over time based on survey records from PUMS data. As can be seen, between 2012 and 2019, the percent of cyclists has varied from a low of around 0.8% and a high of around 1.3% with the lowest use occurring in 2013, possibly due to record weather conditions that year. Otherwise, ridership has not varied greatly over time. Total number of cyclists increases slightly over time as the population increases, but a graph of total cyclists did not reveal any substantially different trend.

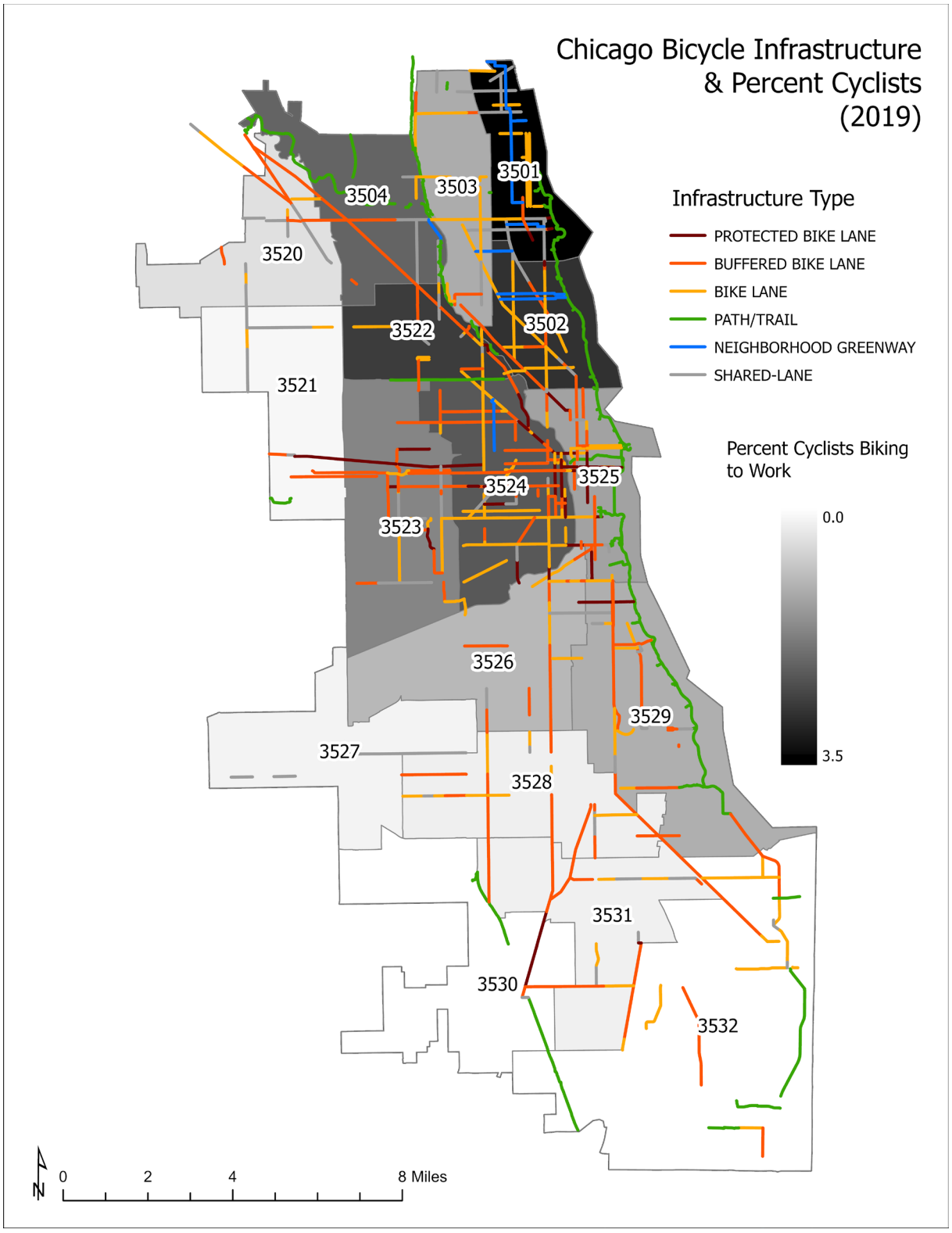


Figure 3.3. Bicycle Infrastructure with Percent Cyclists

This map shows all of the bicycle facilities present in 2019 and the percent of individuals biking to work in 2019. We can observe a higher percent of cyclists where it appears that there is more bicycle infrastructure.

The variables utilized from the PUMS data are described in further detail below in Table 3.1. Some variables from the PUMS data required alteration to improve their usefulness and/or interpretability in the regression equation. These alterations are described in further detail in Table 3.2. Categories for Age were created based on observation of natural break points in the effects of the variable when categorizing age into groups of every ten years.

Unaltered Demographic Variables from PUMS Data:

Variable Name	Definition	Values (if not continuous)
HINCP	Household income	Continuous
MV	Years at current residence	1: 12 months or less 2: 13-23 months 3: 2-4 years 4: 5-9 years 5: 10-19 years 6: 20-29 years 7: 30+ years
VEH	Number of vehicles for household	0-5: 0-5 vehicles 6: 6+ vehicles

Table 3.1. Description of PUMS Unaltered Demographic Variables

This table provides each of the variables used in this study from the PUMS data that were used in their original form for use in the regression equation, as well as their definitions and their values.

Altered Data from PUMS:

New Name	PUMS Name	Original Categories Description	New Categories
Bike	JWTR (pre-2019) JWTRNS (2019+)	Categorical variable describing means of transportation to work (12 categories)	1: Bike to work 0: All other modes
TEN	TEN	1: Owned w/ mortgage/loan 2: Owned free and clear 3: Rented 4: Occupied (no rent)	1: Renting 0: Owning or no payment
Education	SCHL	1: No schooling 2: Preschool/Nursery school	No Degree Undergraduate Degree

		3: Kindergarten 4-14: 1st-11th grade 15: 12th (no diploma) 16: Highschool diploma 17: GED/Alternative diploma 18: <1 year of college 19: >1 year of college (no degree) 20: Associate's degree 21: Bachelor's degree 22: Master's degree 23: Degree beyond bachelor's 24: Doctorate degree	Graduate Degree
Age	AGEP	Continuous	< 20 20-60 60-70 70+
DIS	DIS	1: Disability present 2: No disability	1: Disability present 0: No disability
Hispanic	HISP	Hispanic origin (24 categories) 1: Not Spanish/Hispanic/Latino	1: Hispanic/Latino 0: Not Hispanic/Latino
RACE	RAC1P	1: White alone 2: Black/African American alone 3: American Indian alone 4: Alaska Native alone 5: American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other races 6: Asian alone 7: Native Hawaiian and Other Pacific Islander alone 8: Some Other Race alone 9: Two or More Races	White Black Native American Asian Other
SEX	SEX	1: Male 2: Female	Male Female
ADULTS	NP - NRC	*NP:* No. of persons in household *NRC:* No. of related children	Continuous (no. of adults in household)

HousingCost	coalesce(SMOCP, RNTP)	*SMOCP:* Selected monthly owner costs *RNTP:* Monthly rent	Continuous (monthly housing cost of owning or renting)
PopDens	Σ(PWGT)/SqMi	*PWGT:* Person's weight SqMi (Not from PUMS): Area of PUMA in square miles	Continuous (residential density of PUMA)

Table 3.2. Description of PUMS Altered Demographic Variables

This table provides each of the variables used in this study from the PUMS data that were altered for use in the regression equation, their new variable names if a new name was used, and their old and altered values.

In addition to the PUMS data, data from the American Community Survey (ACS), the City of Chicago, the Cook County Assessor's Office, and U.S. Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) data were used in this study. Each of these supplemental data sources are described below:

American Community Survey (ACS): The American Community Survey is the aggregated result of the PUMS data. Using this dataset spatially aggregated demographic and household characteristics can be added to the model to indicate spatial characteristics rather than individual characteristics. The default variables were altered as described in the table below:

New Name	ACS Code(s)/Name	Description
MedHouseholdIncome	B19013_001: Median Household Income	Median household income
MedHousingCost	B25105_001: Selected monthly owner costs (SMOC) B25064_001: Monthly rent	Median housing cost
NonWhiteProp	B02001_002: Count of white persons B02001_001: Count of all persons	Proportion of population that it nonwhite

Table 3.3. Description of PUMS Altered Demographic Variables

This table provides each of the variables used in this study from the ACS data that were used to develop the regression equation.

Bike Routes Data (City of Chicago): Upon request, the City of Chicago was able to provide a GIS shapefile which includes bicycle routes, infrastructure type, and installation date. In creating networks for each year, sections designated as upgrades rather than new installations were not included in added infrastructure for a given year due to the inability to reasonably determine the type and extent of the upgrade. Since most of these upgraded sections are protected or buffered bike lanes, they likely are being upgraded from existing non-buffered/protected bike lanes.

All Roads (TIGER/Line Shapefiles): This data includes all roads for a particular state (Illinois and Indiana in this instance) for a given year. This data will be part of the network dataset for cyclists.

LEHD LODES (U.S. Census Bureau): Residence Area Characteristics (RAC) and Workplace Area Characteristics (WAC) datasets are used in this study. These are block-level datasets. The WAC dataset is used to determine the employment density for each PUMA, and RAC is used in generating the Bikeability Index.

Crimes 2001 to Present (City of Chicago): This dataset includes all reported crimes within the Chicago urban boundary from 2001 to 2020 as well as their location and other descriptive attributes. Homicide is in its own category. Other crimes were reclassified as described below:

Violent Crime: Battery, Assault, Robbery, Criminal Sexual Assault, Kidnapping, Intimidation, Human Trafficking, Ritualism, Domestic Violence

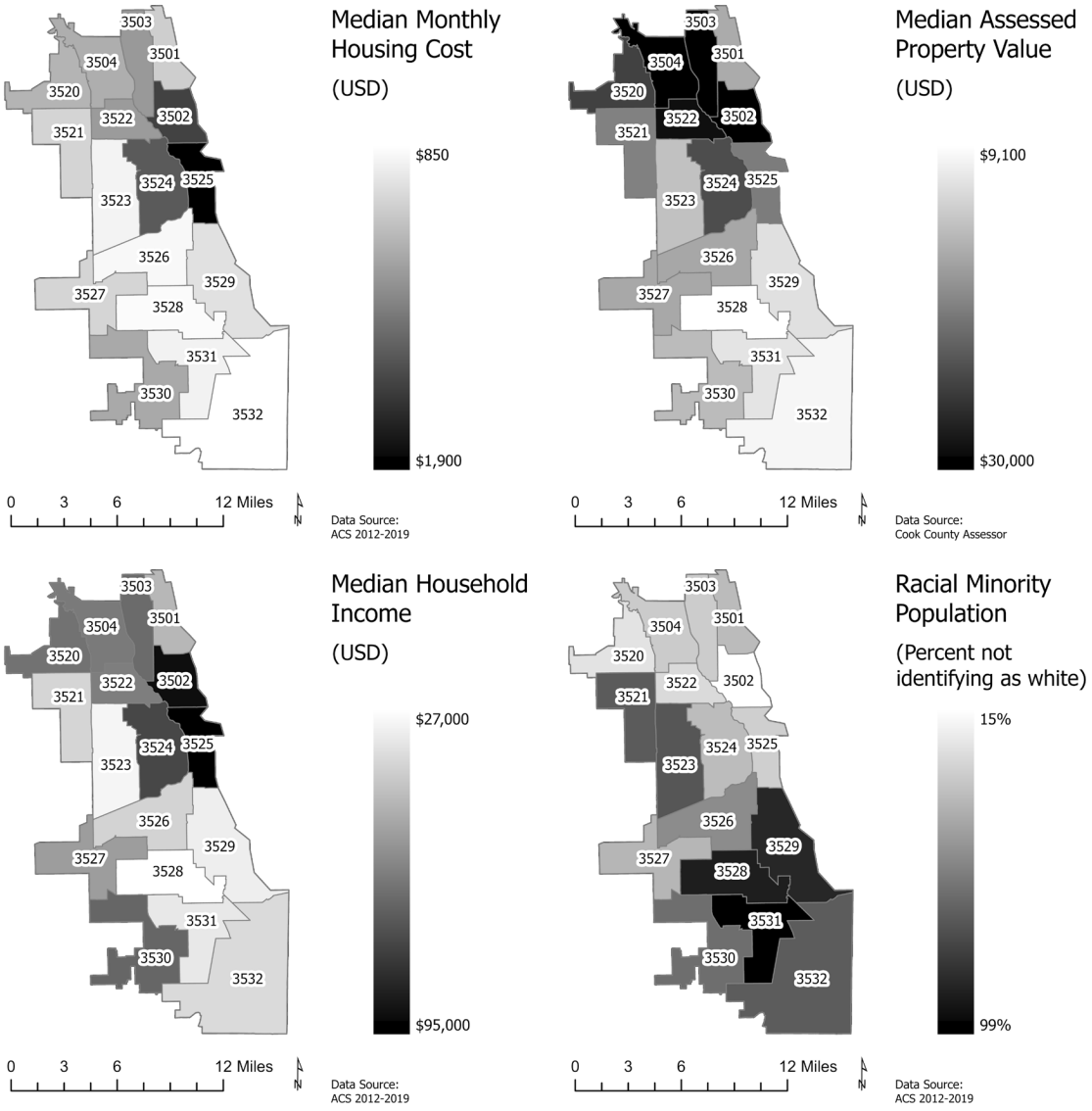
Property Crime: Theft, Burglary, Criminal Damage, Motor Vehicle Theft, Arson

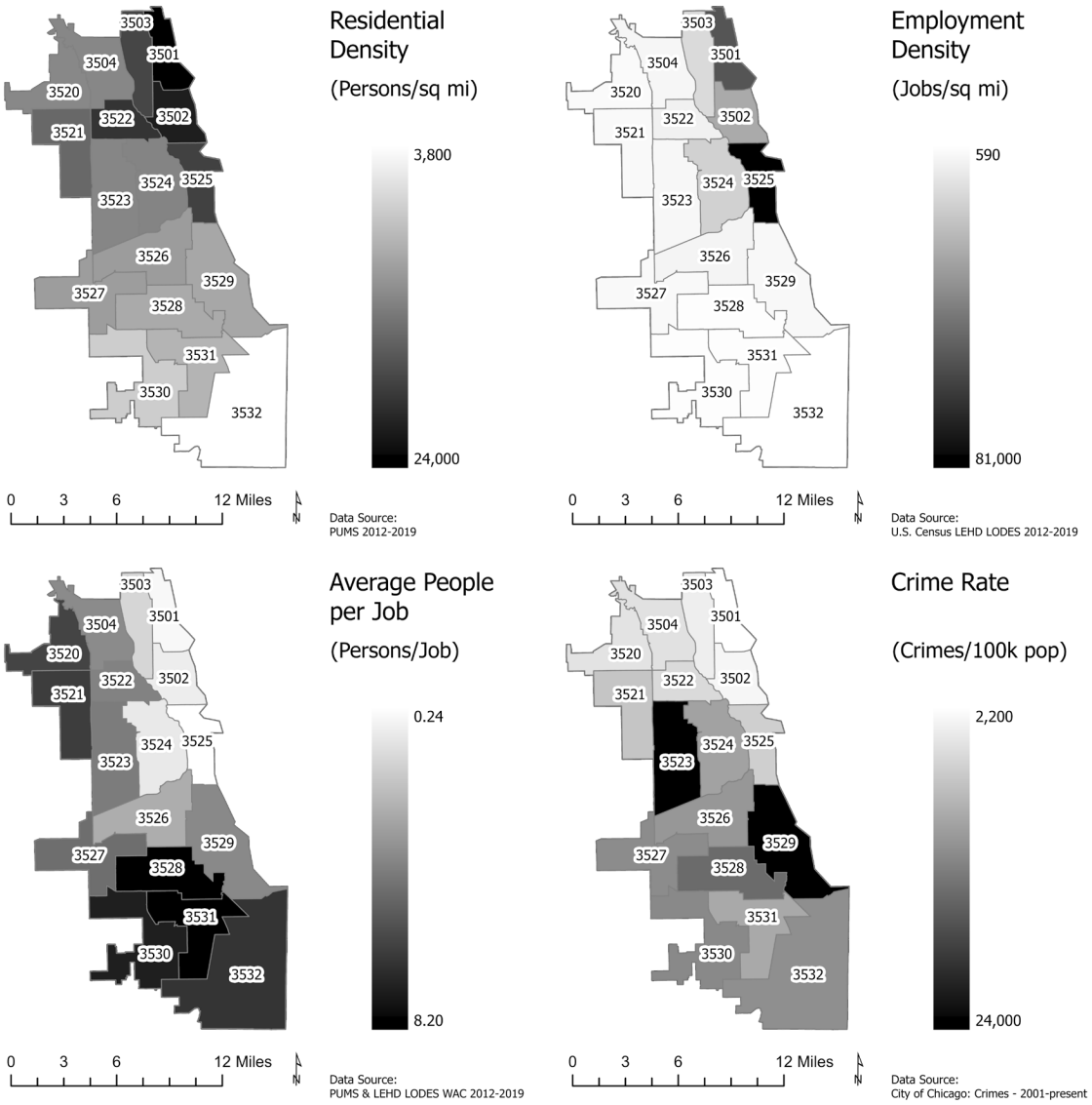
Nonviolent Crime: Deceptive Practice, Narcotics, Weapons Violation, Criminal Trespass, Sex Offense, Obscenity, Public Peace Violation, Prostitution, Gambling, Liquor

Law Violation, Stalking, Public Indecency, Concealed Carry License Violation, Other Narcotic Violation, Interference with a Public Officer

Other: Other Offense, Non-Criminal, Offense Involving Children

Cook County Assessor's Assessments: This dataset includes assessments of all properties within Cook County.





Figures 3.4 – 3.11. Maps of Spatial Variables in Model

This map shows all of spatial variables used to develop the regression. These trends show that there is likely substantial collinearity between spatial variables, which is not uncommon. Since these variables are being utilized as controls this high collinearity can be ignored given the willingness to sacrifice interpretability of these variables.

Procedure

The primary goal of this project is to generate a binary logistic regression model determining the likelihood of an individual choosing to bike to work using the available data (outlined above). To perform this analysis, spatial/temporal data must be merged to the existing individual records within the PUMS.

The first step of this process was to obtain all of the necessary data, both demographic and spatial. PUMS data was chosen for the source of demographic data. Spatial data was selected by considering what data might be influential in determining mode choice and determining what data was readily available. PUMS data from Illinois was obtained for each year from 2012 to 2019. All years of PUMS data were merged into a single dataset and subset to include only the needed data. A new field, PUMA_Yr, was created to be used as an identifier for the Public Use Microdata Area (PUMA) and year of the record to merge spatial/temporal data.

R and ArcGIS batch geoprocessing/model builder were used to convert geospatial data into tabular data associated with a particular PUMA_Yr. This process included aggregating data to the PUMA level based on existing spatial identifiers as outlined below:

	Aggregation Method	**Aggregation Steps**
Residential Density (PUMS)	Sum/Area(mi^2)	Records to PUMA
LEHD LODES WAC	Sum/Area(mi^2)	Tract to PUMA
Crimes	Sum/100,000 people	Point to Community Area Community Area to PUMA
Assessments	Median	Records to PUMA
ACS	Starts at PUMA level	PUMA
Bike Routes	Accessibility Index	Block Group to PUMA

Table 3.4. Spatial Aggregation Steps and Methods

This table provides the steps used to aggregate spatial data to the PUMA level for use with the PUMS dataset.

Bicycle Accessibility Index Creation

In order to effectively measure the bikeability of each PUMA, some approximation needed to be produced. Though measures such as miles of infrastructure divided by area are quick approximations of bikeability, these measures fail to consider the overall density of the area, the presence of jobs, and other nuances that affect the effectiveness of the infrastructure. I determined that the most accurate measure of bikeability would be based on the number of jobs within a reasonable bike commute. To determine the "reasonable" travel time I used to PUMS data to find the median commute time for cyclists in Chicago (20 minutes).

Network datasets were made for use in ArcGIS Network Analyst. Highways were excluded from these datasets. The road and bicycle facility data were classified with the following infrastructure types:

Bicycle Facilities (City of Chicago)	Roads (TIGER/Line)
Off-street trails and paths	Alleys
Protected bike lanes	City streets
Buffered bike lanes	Secondary roads and service drives
Bike lanes (unprotected)	Primary roads and ramps
Neighborhood greenways	
Shared lanes	

Table 3.5. Facility Types in Network Datasets

This table provides the infrastructure classifications of the merged network datasets.

Speed is used as a proxy for the percentage of the overall population willing to use each type of infrastructure. This method operates under the assumption that network impedance can be roughly equated with the proportion of cyclists willing to use the infrastructure. This does not affect the way that the index operates in my model, but this does mean that this can only be a

rough estimate of the average number of jobs accessible. To determine the proxy speeds, the commonly accepted *Four Types of Cyclists* classifications were used to determine the percent of the population likely to use each type of infrastructure. These "four types of cyclists" include "strong and fearless", "enthused and confident", "interested but concerned", and "no way, no how", and each of these categories have different levels of comfort on different types of infrastructure (Geller, 2006). One of the most recent follow-up studies found that the nationally 7% of cyclists fall in the "strong and fearless" category, 5% fall in the "enthused and confident" category, and 51% fall in the "interested but concerned" category (Dill & McNeil, 2016). The remaining 37% fall in the "no way, no how" category. The same study asked members of each category (except "no way, no how") to rate their comfort on six different types of infrastructure on a scale of 1-4 where 4 equals very comfortable and 1 equals very uncomfortable. I reclassified the infrastructure types to match those in Dill & McNeil (2016). To determine the average share of each category of cyclist using each type of infrastructure I included the whole proportion of people that reported that they were "very comfortable" using the infrastructure and half of the proportion of people that reported that they were "somewhat comfortable".

These proportions were then multiplied by the proportion of each type of cyclist (excluding "no way, no how"). These new proportions for each type of cyclist once the "no way, no how" category is excluded are 11.1% strong and fearless, 7.9% enthused and confident, and 81.0% interested but concerned. The sums were then totaled for all types of cyclists to provide an initial weight equivalent to the total proportion of cyclists willing to use a particular type of infrastructure. A final weight was created by dividing the initial weights by the maximum initial weight to make the new maximum equal to 1 and multiplied by 20 (assumption is that max speed will be approximately 20mph). These weights and speeds are shown in Table 3.7.

Dill & McNeil 2016	Chicago Bike Facilities & TIGER Roads	Strong & Fearless	Enthused & Confident	Interested but Concerned
Path or trail	PATH/TRAIL	0.745	0.765	0.730
Bike boulevard	GREENWAYS	0.805	0.835	0.660
Quiet residential street	SHARED LANE	0.795	0.740	0.630
	ALLEY	0.795	0.740	0.630
Major street, protected bike lane	PROTECTED BIKE LANE	0.785	0.795	0.450
avg	BUFFERED BIKE LANE	0.805	0.898	0.305
Major street, striped bike lane	BIKE LANE	0.825	1.000	0.160
Major street, no bike lane	SECONDARY ROAD	1.000	0.200	0.080
	PRIMARY ROAD/RAMP	1.000	0.200	0.080
	CITY STREET	1.000	0.200	0.080

Table 3.6. Infrastructure Reclassification for "Four Types of Cyclists" and Use Proportions

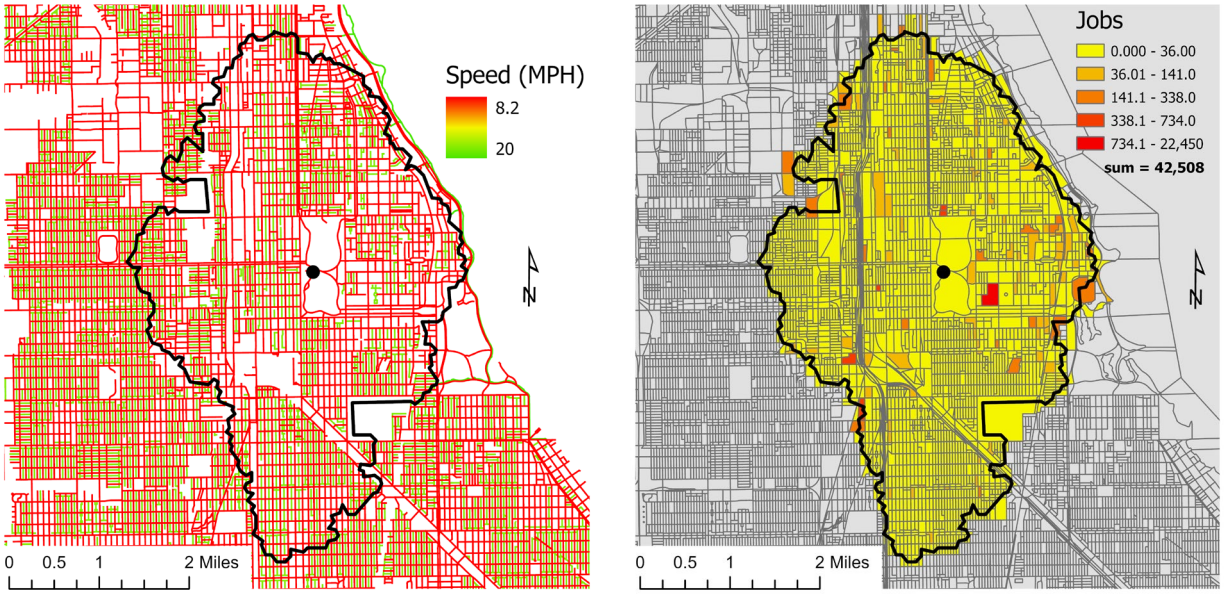
Infrastructure reclassification and proportion of each type of cyclist using each type of infrastructure. Weights for buffered bike lanes are determined by averaging weights from striped and protected bike lanes since it does not cleanly match with a category from the Dill & McNeil, 2016 study.

Dill & McNeil 2016	Chicago Bike Facilities & TIGER Roads	Initial Weight	Final Weight (Initial/0.734)	Speed (Final x 20)
Path or trail	PATH/TRAIL	0.734	1.000	20.000
Bike boulevard	GREENWAYS	0.690	0.939	18.790
Quiet residential street	SHARED LANE	0.657	0.895	17.893
	ALLEY	0.657	0.895	17.893
Major street, protected bike lane	PROTECTED BIKE LANE	0.515	0.701	14.013
avg	BUFFERED BIKE LANE	0.408	0.555	11.099
Major street, striped bike lane	BIKE LANE	0.301	0.409	8.185
Major street, no bike lane	SECONDARY ROAD	0.192	0.261	5.222
	PRIMARY ROAD/RAMP	0.192	0.261	5.222
	CITY STREET	0.192	0.261	5.222

Table 3.7. Weighting of Infrastructure Types

This table shows the initial and final weights of each type of infrastructure based on the "four types of cyclists" classification as well as proxy speeds determined by multiplying the final weight by a maximum speed of 20mph.

To create the "facilities" used in Network Analyst, WAC and RAC data were aggregated and merged with TIGER/Line block group files with total workers and jobs within the block group. Centroids were created from the resulting file. 20-minute isochrones were generated for each year using these facilities and the network datasets. Figure 3.12 provides an example of an isochrone generated using this method, and Figure 3.13 shows how these isochrones are used to select blocks and aggregate the total number of jobs accessible.



Figures 3.12 & 3.13. Isochrone Generation and Job Aggregation Example

Figure 3.12 (left) shows an example 20-minute isochrone generated in ArcGIS. Figure 3.13 (right) shows an example of how a generated isochron is used to select all of the blocks accessible to the starting point. This is then used to determine the sums of all jobs within the service area/isochrone. This process is repeated for all block groups within the study area and aggregated to the PUMA level.

To create meaningful indices for each PUMA, I multiplied the number of jobs within the isochrone by the number of workers within the starting block group, divided that by the total number of workers in the PUMA, and summed all of those values for each block group within

the PUMA. Overall, this process entailed applying the following equation to each PUMA for each year in the study:

$$\sum W_{BG} * \frac{R_{BG}}{R_{PUMA}} = Index\ Value$$

R_{BG} = tot. no. of people living in the block group for which a service area was created

R_{PUMA} = tot. no. of people living in the PUMA containing the start block

W_{BG} = tot. no. of jobs within a given block group's service area

The index was divided by 1,000 for easier interpretation of the coefficient in the final logistic model. Figure 3.14 provides the distribution of the indices once they have been merged with the PUMS records. Figure 3.15 shows the final indices over time for all PUMAs. Any decrease can be explained by the loss of jobs in an area. Increases may be caused by either an increase in jobs within an area or greater connectivity. Figure 3.16 shows the index for each PUMA averaged for all years in the study.

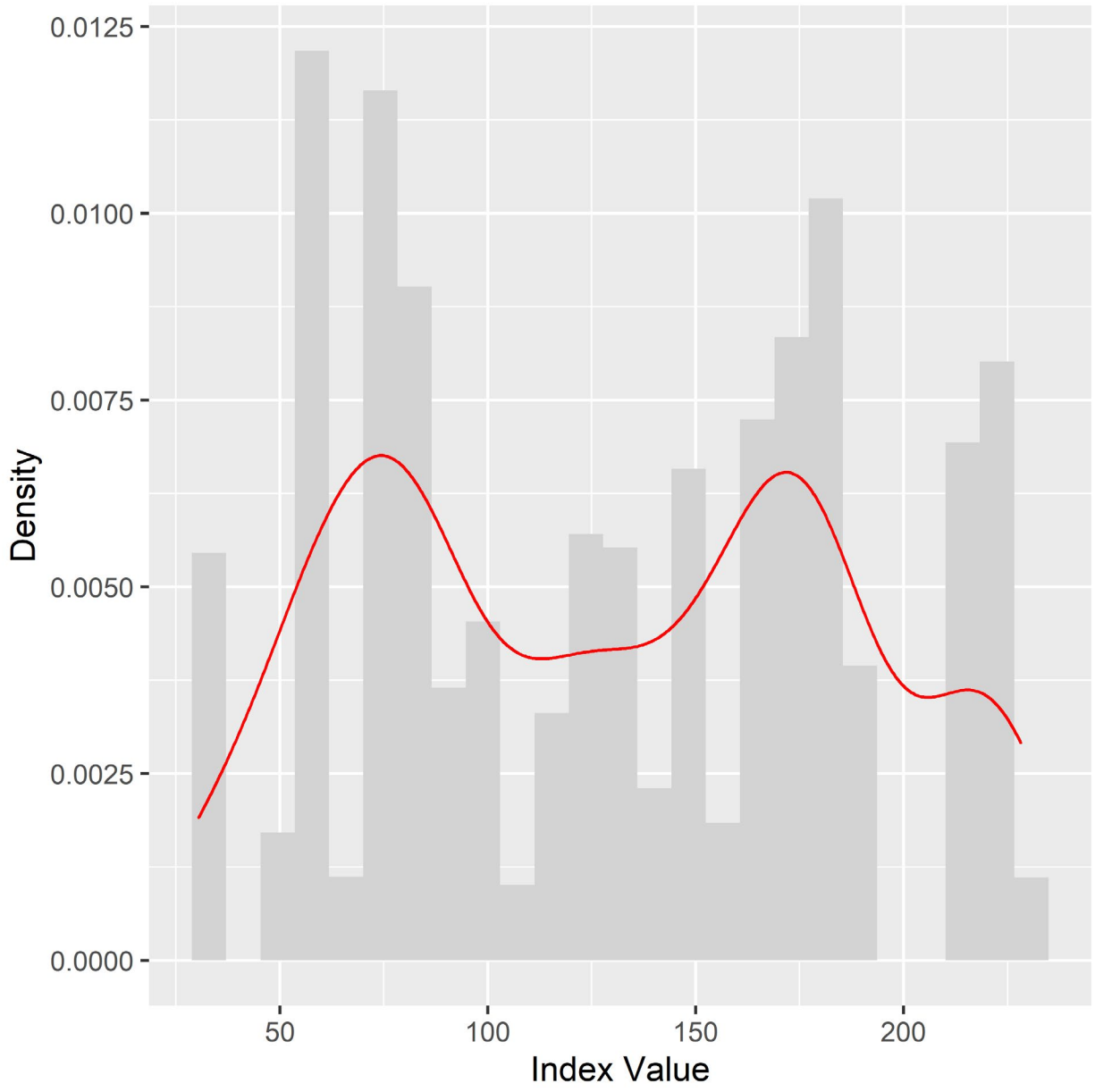


Figure 3.14. Bicycle Accessibility Index Histogram

This histogram and density plot provides the distribution of the bicycle accessibility indices for all records within the PUMS dataset. Though the distribution is not normal and appears to have two primary modes, there does not appear to be an issue with any major outliers. In order maintain an easily interpretable result, the bicycle index variable will not be transformed.

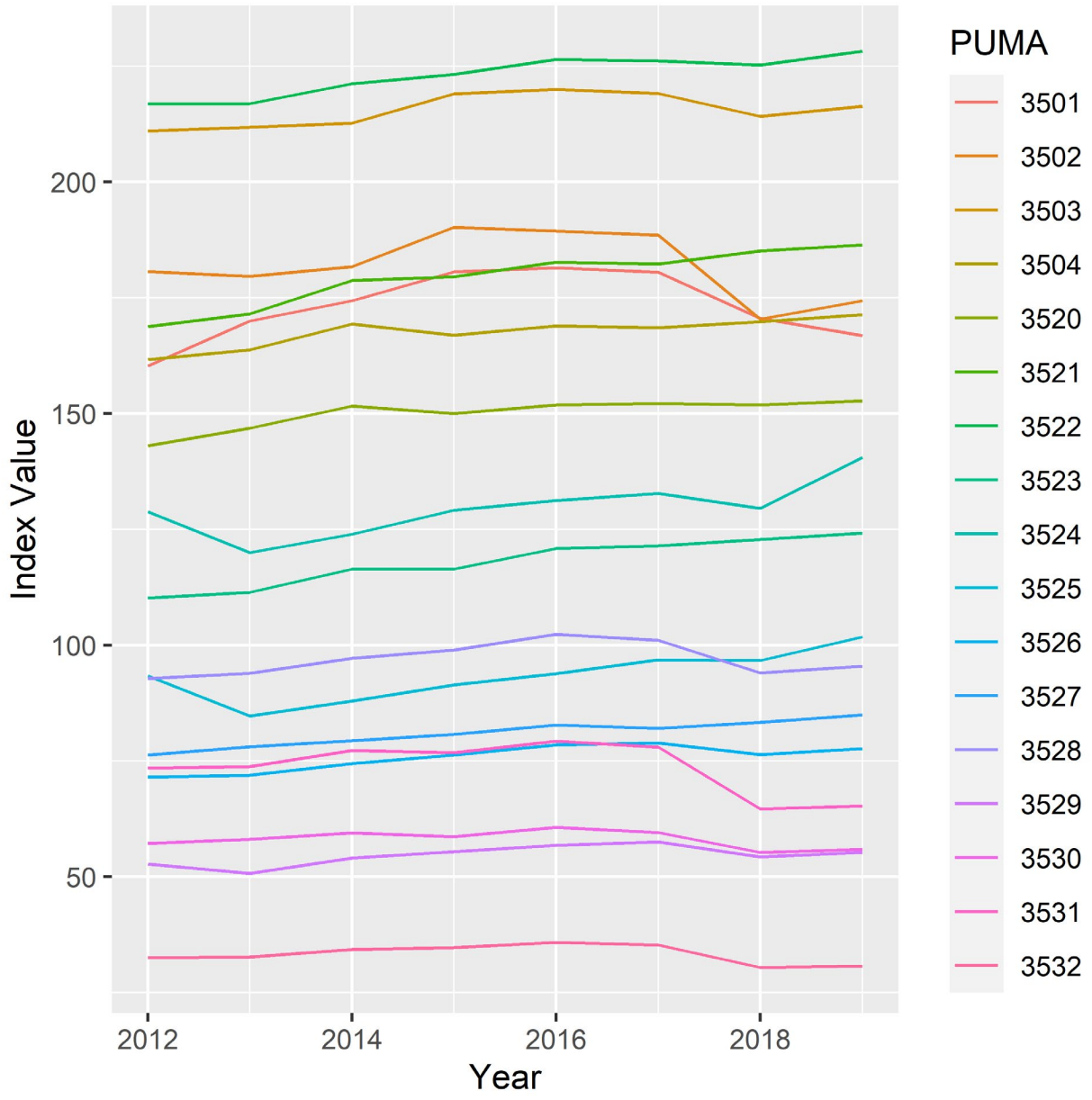


Figure 3.15. Bicycle Accessibility Index Over Time

This graph shows the change bikeability index for each PUMA over time. Any decrease can be explained by the loss of jobs in an area. Increases may be caused by either an increase in jobs within an area or greater connectivity.

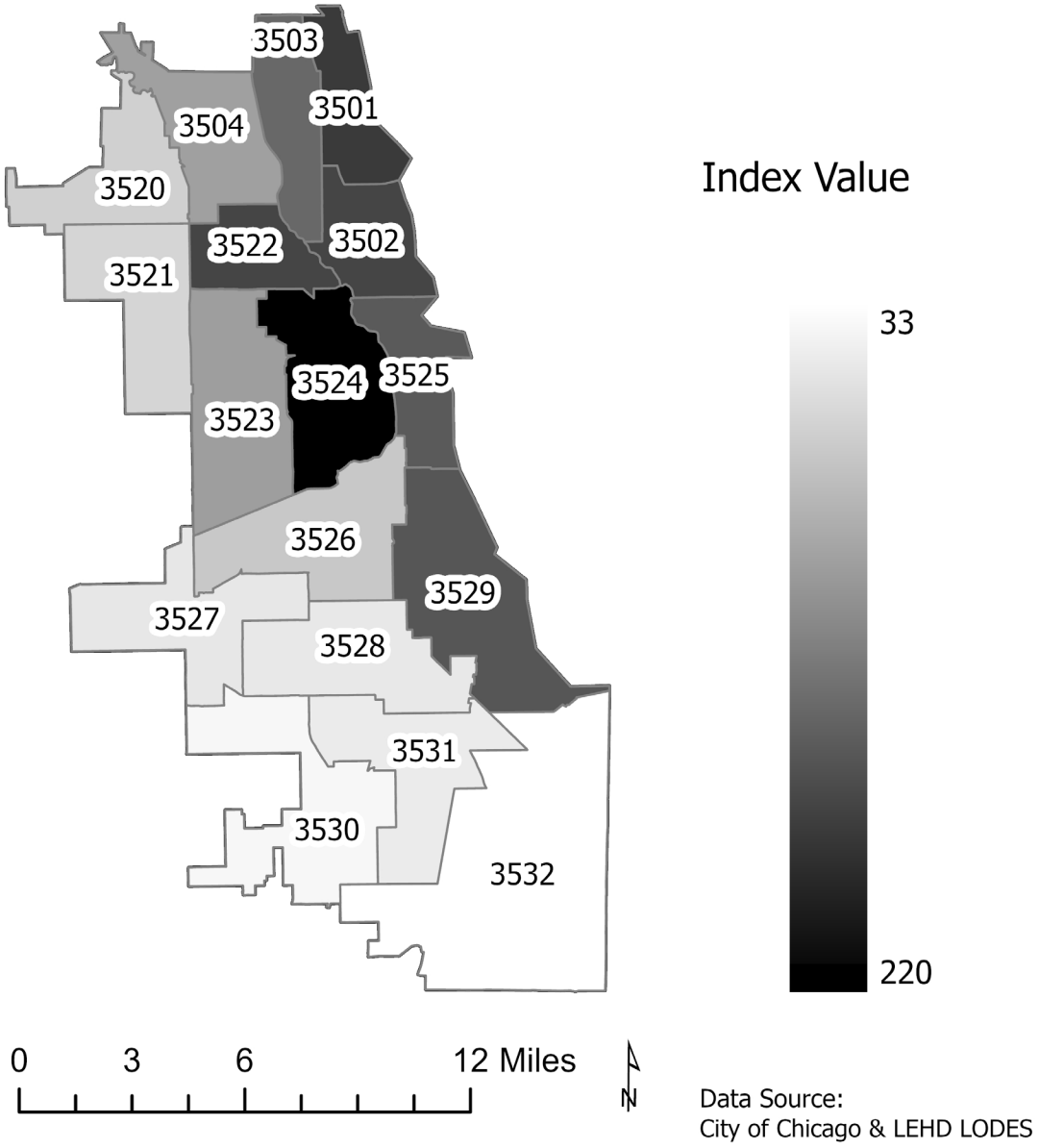


Figure 3.16. Bicycle Accessibility Index Map

This map shows the average bikeability index for each PUMA averaged for all years between 2012 and 2019. When compared to Figure 3.3, we can see a similar distribution with the percentage of individuals cycling.

Logistic Regression Procedure

I merged all spatial/temporal data (including bikeability) to the PUMS data by PUMA_Yr using R. I then used R to create a binary logistic regression where biking to work = 1, and other modes = 0. Independent variables include all of the demographic and spatial/temporal data. Prior to producing the model, I transformed each independent variable using natural log (ln), ln(1+x), reciprocal (1/x), square root ($x^{1/2}$), square (x^2), or exponential (e^x). For each variable I selected the transformation that has both or some balance of the lowest skewness and strongest correlation with the *Bike* variable possible. Reduction of skewness reduces the potential impact of outliers. Correlation with the *Bike* variable is not entirely necessary but more correlation suggests a greater effect on the binary selection process. The transformations chosen should also have no infinite values and should avoid negative values. Rather than a transformation, the age variable was put into bins of < 20 years, 20-60 years, 60-70 years, and 70+ years based on change points identified when binning into groups of 10 years. In addition to transformations, several interaction terms were also tested. Next, I selected variables in a pseudo-backward selection procedure for demographic/household variables and a pseudo-forward/stepwise selection process for spatial variables to maximize the predictive quality of the model (based on AIC) and to ensure that collinearity with the bikeability index variable is kept at a minimum (VIF $\leq$ 2.5). The bikeability index variable was added first and each variable was added in order of correlation with the index variable from highest to lowest. Should a variable be statistically insignificant, or increase the AIC, it was removed. If the VIF of the bikeability index variable was increased over 2.5, the variable with the highest p-value was removed. After all variables were attempted, variables the were removed were readded based on the previous steps until no additional variables could be added. This procedure is outlined in detail in Table 3.8.

Variable	AIC	p-value	Index VIF	Index |r|	p-value (removed)	Removal Reason
DIS	6174.9	0.107				
Age	6174.9					
Education	6174.9					
Sex	6174.9					
RACE	6174.9					
Hispanic	6174.9	0.000				
TEN	6174.9	0.869				
MV	6174.9					
HousingCost	6174.9	0.208				
log(HINCP/ADULTS)	6174.9	0.069				
log(HINCP)	6174.9	0.081				
HousingCost/HINCP	6174.9	0.852				
VEH/ADULTS	6174.9	0.000				
[TEN]	6173.0				0.869	high p-value
[HousingCost/HINCP]	6171.1				0.851	high p-value
[MV]	6170.8				0.017 - 0.939	high p-value
[HousingCost]	6170.1				0.255	high p-value
[DIS]	6170.2				0.131	high p-value
[log(HINCP/ADULTS)]	6173.7				0.020	high colinearty, higher p-value
[log(HINCP)]	6172.7				0.310	high p-value
Index1000	6166.9	0.005	1.08	1.000		
PopDens	6167.9	0.296	2.00	0.804		
[PopDens]	6166.9		1.08		0.296	high p-value
log(ViolentCrimeRate)	6122.0	0.000	2.26	0.754		
log(NonviolentCrimeRate)	6113.7	0.001	2.22	0.723		
[log(ViolentCrimeRate)]	6111.9		2.10		0.690	high p-value
log(PropertyCrimeRate)	6112.6	0.246	2.19	0.714		
[log(PropertyCrimeRate)]	6111.9		2.10		0.246	high p-value
1000/MedHousingCost	6111.0	0.086	2.09	0.650		
NonWhiteProp	6104.4	0.003	2.29	0.612		
log(WorkDens)	6075.1	0.000	2.36	0.491		
[1000/MedHousingCost]	6074.2		2.33		0.284	high p-value
PopDens/WorkDens	6065.2	0.001	2.46	0.381		
[log(WorkDens)]	6063.2		2.34		0.808	high p-value
MedHouseholdIncome/1000	6065.1	0.670	2.49	0.378		
[MedHouseholdIncome/1000]	6063.2		2.34		0.670	high p-value
MedHousingCost/MedHouseholdIncome*100	6061.4	0.049	2.53	0.290		
[NonWhiteProp]	6061.8		2.13		0.116	Index VIF > 2.5; high p-value
PopDens	6063.6	0.656	2.74	0.804		
[PopDens]	6061.8		2.13		0.656	Index VIF > 2.5; high p-value
log(ViolentCrimeRate)	6062.9	0.328	2.26	0.754		
[log(ViolentCrimeRate)]	6061.8		2.13		0.328	high p-value
log(PropertyCrimeRate)	6058.6	0.024	2.51	0.714		
[log(NonviolentCrimeRate)]	6059.3		2.57		0.100	Index VIF > 2.5, highest p-value
log(NonviolentCrimeRate)	6058.6		2.51	0.723		Removal increased Index VIF
[log(PropertyCrimeRate)]	6061.8		2.13		0.024	Index VIF > 2.5, second highest p-value
1000/MedHousingCost	6059.4	0.037	2.14	0.650		
NonWhiteProp	6058.0	0.066	2.53	0.612		
[NonWhiteProp]	6059.4		2.14		0.066	Index VIF > 2.5, highest p-value
log(WorkDens)	6054.6	0.009	2.43	0.491		
MedHouseholdIncome/1000	6043.5	0.000	2.93	0.378		
[log(NonviolentCrimeRate)]	6043.6		1.64		0.152	Index VIF > 2.5; high p-value

No further variables could be added at this point that were statistically significant and maintained an index VIF under 2.5

Table 3.8. Procedure for Regression Variable Selection

This table outlines the order in which variables were added and removed and why they were removed if applicable.

Logistic Regression Quality Assessment

Though the model has been produced at this point it is important to test the logistic regression model assumptions to test the reasonableness of the model (Wolfe, 2018). First, we must determine if the model has an appropriate outcome structure (ensure the outcome is binary). This is met by my model's dependent variable where 1 = biking to work and 0 = all other transportation options. Next, we must verify observation independence (absence of matched data and autocorrelation). For this I have used the Durbin-Watson D Test and the Breusch-Godfrey Test for autocorrelation (results in Table 4.1). Third, I must confirm the absence of multicollinearity in any variable I wish to directly interpret. For this assumption I have produced the variance inflation factors (VIFs) for each of the independent variables in the logistic regression model. For a logistic regression, linearity of independent variables and log odds is expected. A visual analysis of the scatter plots of the independent variables compared to the log odds was utilized to test this assumption. Finally, there is the assumption of a large sample size. For this, I must determine if the data available meets the rule of thumb: 10 cases with the least frequent outcome for each independent variable. In my final logistic regression, there are a total of 11 independent variables. 662 records of cyclists were in my dataset after subsetting, exceeding the 110 required. Finally, I must determine if my variable of interest (my bike index) has a statistically significant relationship with the odds of someone choosing to bike to work using the p-value for this coefficient in the model output.

Results

The final output of my research is a binary logistic regression model which includes demographic variables from the PUMS dataset and spatial/temporal data from the various other sources mentioned previously. In this model *Bike* is the dependent variable and includes two possible values: 1 = choosing to bike as the primary mode of transportation to work and 0 = some other mode of transportation. Transformations were made to several variables. Binning was used for the *Education* and *Age* variables as using them as continuous variables led to inconsistent effects. Education was binned into categories of No Degree, Undergraduate Degree, and Graduate Degree. Age was binned into categories of < 20, 20-60, 60-70, 70+ based on apparent changes in coefficients when tested in ranges of ten years.

Similar to previous studies, sex was shown to significantly influence the likelihood biking to work, men being 2.61 times more likely to bike than women. Interestingly, and perhaps counterintuitively, Hispanic individuals were only 36.9% as likely to bike to work as non-Hispanic individuals. Similarly, the racial group that was least likely to choose to bike to work were black and African American individuals. These trends could be a result of inequitable access to infrastructure and/or regional variation in the effects of race on willingness/necessity to bike as past studies have found that "compared to non-Hispanic Caucasians, minorities are more likely to [use active transportation]" (Aziz, Park, et al., 2018). Educational attainment was positively correlated with the probability of biking to work, possibly due a greater understanding of the benefits of cycling. Age had a negative correlation with likelihood of choosing to bike to work. Those under 20 years of age were substantially more likely to bike, likely due to a combination of a lack of vehicular access and good physical ability. Presence of more vehicles per person in a household unsurprisingly led to a lower likelihood of choosing to bike to work.

	Variable	Coef	Std. Error	Pr(>|z|)		Odds Ratio	VIF	Linearity with Log Odds
	(Intercept)	5.959	1.730	0.001	***	387.041		
Bicycle Accessibility Index	Index1000	0.003	0.001	0.001	***	1.003	1.636	Linear
Other Spatial/Temporal Data	PopDens/WorkDens *(People per Job)*	-0.324	0.052	0.000	***	0.723	6.940	Linear
	log(WorkDens)	-0.432	0.093	0.000	***	0.649	7.171	Linear
	1000/MedHousingCost	-4.466	0.825	0.000	***	0.011	14.881	Linear
	MedHousingCost/MedHouseholdIncome*100	0.776	0.210	0.000	***	2.173	4.963	Nonlinear
	MedHouseholdIncome/1000	-0.033	0.008	0.000	***	0.967	21.212	Linear
Personal Demographics	Age						1.151	
	20-59	-1.048	0.193	0.000	***	0.351		
	60-69	-1.372	0.257	0.000	***	0.254		
	70+	-1.858	0.541	0.001	***	0.156		
	Education						1.651	
	Undergraduate Degree	0.361	0.117	0.002	**	1.435		
	Graduate Degree	0.599	0.124	0.000	***	1.820		
	Sex						1.010	
	male	0.959	0.089	0.000	***	2.609		
	RACE						1.307	
	native american	1.056	0.471	0.025	*	2.874		
	asian	-1.039	0.172	0.000	***	0.354		
	black	-1.990	0.208	0.000	***	0.137		
	Hispanic	-0.998	0.125	0.000	***	0.369	1.495	
Household Characteristics	VEH/ADULTS	-1.728	0.126	0.000	***	0.178	1.073	Linear

Psuedo R-Squared Values

AIC:	6043.6
McFadden:	0.133
McFaddenAdj:	0.128
CoxSnell:	0.020
Nagelkerke:	0.142
Tjur:	0.029

Autocorrelation Tests

	p-value
Durbin-Watson Test	0.205
Breusch-Godfrey Test	0.427

Model Goodness of Fit Test

	p-value
Hosmer-Lemeshow Test	0.108

Table 4.1. Logistic Regression Output

This table provides the estimated coefficients for each independent variable used in initial model (prior to variable elimination) and the final model, their significance values, and assumption tests for the final model.
Note: *** means significance at the $\alpha = 0.001$ level, ** means $\alpha = 0.01$, and * means $\alpha = 0.0$

Due to the model assumption being (at least mostly) met, we can assume that the model itself is fairly accurate and robust. The primary coefficient of this concern is for the *Index1000* variable. This model does not differentiate whether an increase in this value is achieved by adding more jobs close to a person's residence, the improvement/expansion of bicycle infrastructure near their residence, or some combination thereof, as ultimately a similar effect is achieved. To interpret the coefficient of this variable we must apply an exponential transformation to the coefficient (e^(coef)). By doing this we can see that $e^{0.0032960} \approx 1.0033$. Therefore, a one-point increase in the accessibility index (an estimated 1,000 additional jobs within biking distance) correlates with a 0.33-percent increase in the likelihood of someone choosing to bike to work.

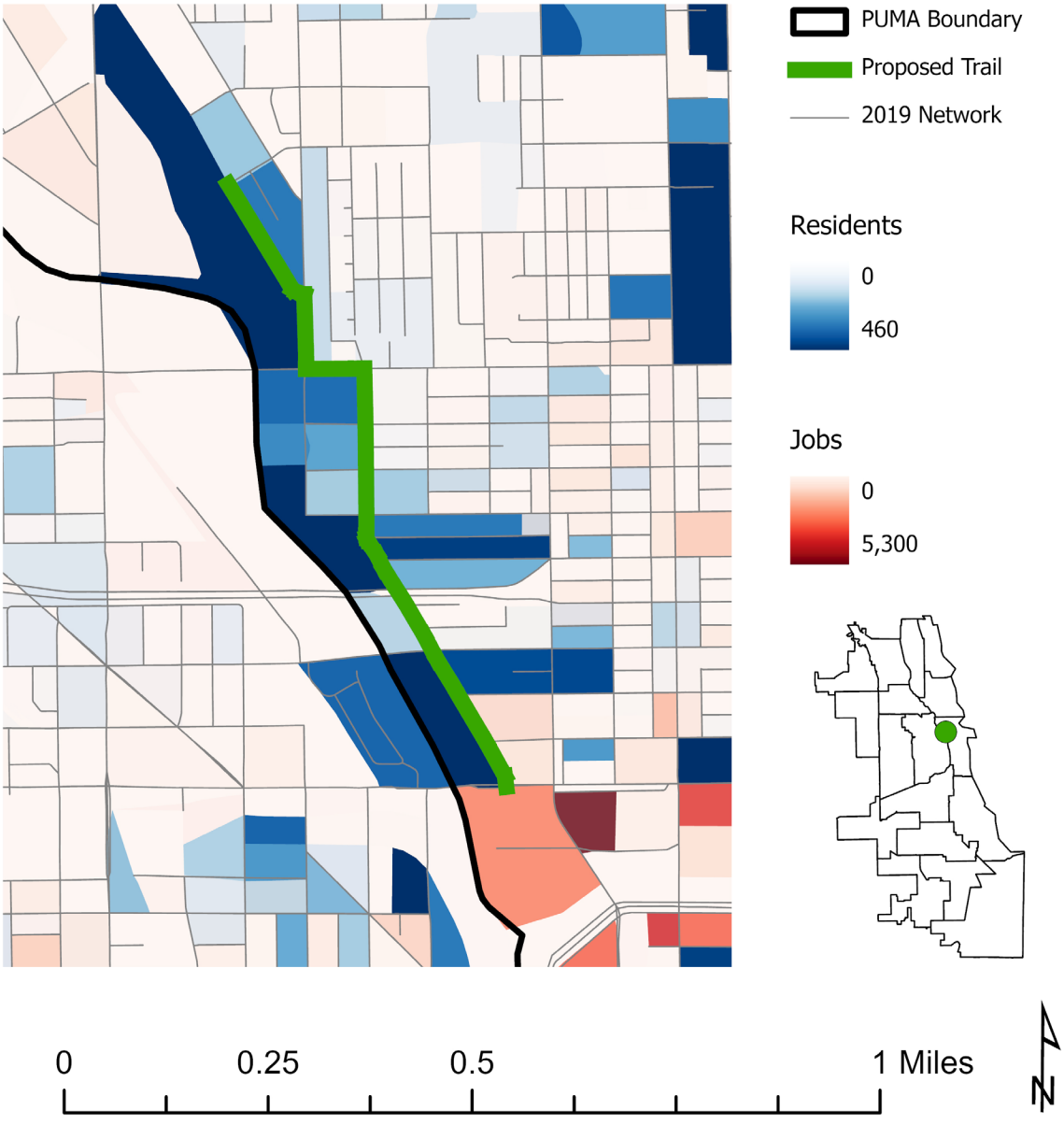
Discussion and Recommendations

A major goal of this research is to provide a repeatable procedure which can be used to guide investment decisions. A 0.33-percent increase in probability of biking to work given access to an estimated additional thousand jobs may not sound like much, but using this information a municipality could, for example, test the change in the amount of ridership due to an infrastructure project prior to any investment being made, much like modern travel demand modelling can predict rerouting of vehicles. Additionally, the procedure used in this study can be used locally to determine local variations in the likelihood of cycling. The following procedure could be used to maximize the accuracy of a local prediction in ridership change:

1. Run the procedure outlined in this study using local data to produce a logistic regression for use as a prediction equation.
2. Alter the network dataset to include the proposed project(s) to produce a new accessibility index and add the revised index to the dataset.
3. Run a prediction of the odds of every individual within the dataset, or, since we are assuming that the other variables are remaining constant, simply take the predictions of the existing dataset for the closest year of the intervention times e^(coef(ΔIndex1000)). The sum of this field (multiplied by person's weight) will represent the predicted total cyclists within the study area after the intervention has been implemented. Additionally, it is recommended to factor in the proportion of records omitted for the regression procedure due to incomplete records. The difference of these totals will represent the estimated number of cyclists gained as a result of the intervention.

This process can be used to directly compare the estimated impacts of different infrastructure projects to help prioritize investments, and can be used by advocacy groups, consultants, etc. to

promote the impacts of investment. An example of this testing procedure was performed for a theoretical 0.89-mile trail connecting a high-residential area to a high-job area, as seen in Figure 5.1. This theoretical intervention resulted in an estimated 67 individuals biking to work that would not have otherwise done so, not including potential new recreational cyclists.



Figures 5.1. Example of Intervention Tested with Logistic Regression Test Procedure

This theoretical intervention connects an area of high residential density to an area of high job density via an off-street trail. Using the procedure outlined in this study to test the effects of this intervention, it was determined that an estimated 67 new commuter cyclists would result from this addition.

This procedure can also be altered to fit data that is locally available. Locally produced demographic and locational data can. For many smaller communities that do not contain multiple PUMAs, demographic data at a finer spatial resolution should be used for effective analysis. Ultimately, any spatial dataset that would likely have some influence on willingness to bike can be added to the model using the same procedure and can lead to a more effective predictive model. So long as a community has bike route data which includes year of installation, other spatial data can be added to the analysis dependent on what is influential and locally available.

This research can be further applied at a regional or national level to bolster the dataset used to produce the model and effectively obtain a national-level understanding of the degree to which infrastructure influences bicycle mode choice. There were some results that I have attributed to local variations, specifically decreased probability of Hispanic and black/African American minorities choosing to bike to work. A national study would address some of this variation. A national-level study could be done one of two major ways: 1) the study could use the same variables for every city and a combined national dataset could be used to run a single regression, or 2) the procedure can be repeated for each community within the study depending on the variety of spatial data that is available and comparisons between communities can be used to determine overall trends. Both of these have benefits and drawbacks, though the second option is likely more implementable, as the first option requires that all cities have the same spatial data available. In either case, a meaningful national average of the impact of additional bicycle infrastructure on the probability of biking to work could be able to be determined.

Additional ways to improve this research in the future would include considering what types of jobs cyclists are likely to commute to. LEHD LODES data includes breakdowns of the

counts of jobs of different types and this data could be used to improve the model. Additionally, this study's accessibility index assumes that travel impedance can be roughly equated to the proportion of people willing to travel on different types of infrastructure. Further testing might determine the level of accuracy of this method. A more direct measure of the trend that was not used in this study would be to generate different isochrones based on different infrastructure quality levels and weighting each of those isochrones by the proportion of cyclists willing to travel those routes. This is a possible method for future research but would require the creation of seven sets of isochrones for each year in the study (one set for each infrastructure category).

Conclusion

Past studies regarding bicycle mode choice factors have often relied on survey respondents reporting the value of differing factors. Though this is a good way to obtain this information from the people making the decisions, there are a variety of potential sources of bias using this method. Additionally, many studies either exclude built-environment characteristics or only include the built environment as a tangential factor that does not receive significant focus in their research. My study was created to 1) be a tool to guide investment decisions, and 2) verify actual change in behavior based on changes in overall bikeability rather than rely on self-reported claims of behavioral change from surveys. Through this study I was able to show a 0.33-percent increase in the probability of an individual choosing to bike given a one-point increase in the bikeability index. Additionally, some demographic characteristics such as race/ethnicity had results contrary to previous understanding, likely due to local variation in behavior. Furthermore, my research has confirmed the results of other studies in regard to sex/gender being one of the most influential demographic factors determining the likelihood of cycling, with men being substantially more likely to bike to work than women. This research opens up many possibilities for use of this process as an investment-decision-making tool, and future research that can determine national mode-choice trends.

References

Active Transportation Alliance. (2021). Our Accomplishments. *Active Transportation Alliance*. https://activetrans.org/about-us/our-organization/our-accomplishments

Al Shammas, T., & Escobar, F. (2019). Comfort and Time-Based Walkability Index Design: A GIS-Based Proposal. *International Journal of Environmental Research and Public Health*, *16*(16), 2850. https://doi.org/10.3390/ijerph16162850

Arellana, J., Saltarín, M., Larrañaga, A. M., Alvarez, V., & Henao, C. A. (2020). Urban walkability considering pedestrians' perceptions of the built environment: A 10-year review and a case study in a medium-sized city in Latin America. *Transport Reviews*, *40*(2), 183–203. https://doi.org/10.1080/01441647.2019.1703842

Arellana, J., Saltarín, M., Larrañaga, A. M., González, V. I., & Henao, C. A. (2020). Developing an urban bikeability index for different types of cyclists as a tool to prioritise bicycle infrastructure investments. *Transportation Research Part A: Policy and Practice*, *139*, 310–334. https://doi.org/10.1016/j.tra.2020.07.010

Aziz, H. M. A., Nagle, N. N., Morton, A. M., Hilliard, M. R., White, D. A., & Stewart, R. N. (2018). Exploring the impact of walk–bike infrastructure, safety perception, and built-environment on active transportation mode choice: A random parameter model using New York City commuter data. *Transportation*, *45*(5), 1207–1229. http://dx.doi.org/10.1007/s11116-017-9760-8

Aziz, H. M. A., Park, B. H., Morton, A., Stewart, R. N., Hilliard, M., & Maness, M. (2018). A high resolution agent-based model to support walk-bicycle infrastructure investment decisions: A case study with New York City. *Transportation Research Part C: Emerging Technologies*, *86*, 280–299. https://doi.org/10.1016/j.trc.2017.11.008

Cyril, A., Mulangi, R. H., & George, V. (2019). DEVELOPMENT OF A GIS-BASED COMPOSITE PUBLIC TRANSPORT ACCESSIBILITY INDEX. *Journal of Urban and Environmental Engineering*, 235–245. https://doi.org/10.4090/juee.2019.v13n2.235245

Dill, J., & Carr, T. (2003). Bicycle Commuting and Facilities in Major U.S. Cities: If You Build Them, Commuters Will Use Them. *Transportation Research Record*, *1828*, 116–123.

Dill, J., & McNeil, N. (2016). *Revisiting the Four Types of Cyclists: Findings from a National Survey*.

Frank, L. D., Sallis, J. F., Saelens, B. E., Leary, L., Cain, K., Conway, T. L., & Hess, P. M. (2010). The development of a walkability index: Application to the Neighborhood Quality of Life Study. *British Journal of Sports Medicine*, *44*(13), 924. https://doi.org/10.1136/bjsm.2009.058701

Geller, R. (2006). *Four Types of Cyclists*.

Gholamialam, A., & Matisziw, T. C. (2019). Modeling Bikeability of Urban Systems. *Geographical Analysis*, *51*(1), 73–89. https://doi.org/10.1111/gean.12159

Guinn, J. M., & Stangl, P. (2014). Pedestrian and bicyclist motivation: An assessment of influences on pedestrians' and bicyclists' mode choice in Mt. Pleasant, Vancouver. *Urban, Planning and Transport Research*, *2*(1), 105–125. https://doi.org/10.1080/21650020.2014.906907

Handy, S., & Xing, Y. (2011). Factors Correlated with Bicycle Commuting: A Study in Six Small US Cities. *International Journal of Sustainable Transportation - INT J SUSTAIN TRANSP*, *5*. https://doi.org/10.1080/15568310903514789

Hardinghaus, M., Nieland, S., Lehne, M., & Weschke, J. (2021). More than Bike Lanes—A Multifactorial Index of Urban Bikeability. *Sustainability*, *13*(21), 11584. https://doi.org/10.3390/su132111584

Heinen, E., Maat, K., & Wee, B. van. (2011). The role of attitudes toward characteristics of bicycle commuting on the choice to cycle to work over various distances. *Transportation Research Part D: Transport and Environment*, *16*(2), 102–109. https://doi.org/10.1016/j.trd.2010.08.010

Kellstedt, D. K., Spengler, J. O., Foster, M., Lee, C., & Maddock, J. E. (2021). A Scoping Review of Bikeability Assessment Methods. *Journal of Community Health*, *46*(1), 211–224. https://doi.org/10.1007/s10900-020-00846-4

Li, Z., Wang, W., Yang, C., & Jiang, G. (2013). Exploring the causal relationship between bicycle choice and trip chain pattern. *Transport Policy*, *29*, 170–177. https://doi.org/10.1016/j.tranpol.2013.06.001

Mavoa, S., Witten, K., McCreanor, T., & O'Sullivan, D. (2012). GIS based destination accessibility via public transit and walking in Auckland, New Zealand. *Journal of Transport Geography*, *20*(1), 15–22. https://doi.org/10.1016/j.jtrangeo.2011.10.001

Mohammad, A., Farshidreza, H., Hamid, M., & Javad, M. T. M. (2020). Investigating the relationship between housing policy and accessibility, based on developing a multi-perspectives accessibility index: A case study in Tehran, Iran. *Journal of Housing and the Built Environment*, *35*(4), 1237–1259. https://doi.org/10.1007/s10901-020-09738-4

Piatkowski, D. P., & Marshall, W. E. (2015). Not all prospective bicyclists are created equal: The role of attitudes, socio-demographics, and the built environment in bicycle

commuting. *Travel Behaviour and Society*, *2*(3), 166–173. https://doi.org/10.1016/j.tbs.2015.02.001

Pritchard, R., Bucher, D., & Frøyen, Y. (2019). Does new bicycle infrastructure result in new or rerouted bicyclists? A longitudinal GPS study in Oslo. *Journal of Transport Geography*, *77*, 113–125. https://doi.org/10.1016/j.jtrangeo.2019.05.005

Reggiani, G., van Oijen, T., Hamedmoghadam, H., Daamen, W., Vu, H. L., & Hoogendoorn, S. (2022). Understanding bikeability: A methodology to assess urban networks. *Transportation*, *49*(3), 897–925. https://doi.org/10.1007/s11116-021-10198-0

Ryu, S., Chen, A., Su, J., & Choi, K. (2021). A multi-class, multi-criteria bicycle traffic assignment model. *International Journal of Sustainable Transportation*, *15*(7), 524–540. https://doi.org/10.1080/15568318.2020.1770906

Smith, R. (2019, May 14). *How did Mayor Rahm Emanuel change transportation?* Curbed Chicago. https://chicago.curbed.com/2019/5/14/18618611/mayor-transportation-legacy-chicago-rahm-emanuel

Song, Y., Preston, J., & Ogilvie, D. (2017). New walking and cycling infrastructure and modal shift in the UK: A quasi-experimental panel study. *Transportation Research Part A: Policy and Practice*, *95*, 320–333. https://doi.org/10.1016/j.tra.2016.11.017

Thomas, J., Zeller, L., & Reyes, A. R. (2014). National Walkability Index: Methodology and User Guide. *PLoS ONE*, *9*(1), e85295. https://doi.org/10.1371/journal.pone.0085295

U.S. Census Bureau. (2020a). *U.S. Census Bureau QuickFacts: Chicago city, Illinois*. https://www.census.gov/quickfacts/fact/table/chicagocityillinois/IPE120221#IPE120221

U.S. Census Bureau. (2020b). *U.S. Census Bureau QuickFacts: United States*. https://www.census.gov/quickfacts/fact/table/US/PST045221

Walk Score. (2021). *Bike Score Methodology*. https://www.walkscore.com/bike-score-methodology.shtml

Wang, C.-H., & Chen, N. (2015). A GIS-based spatial statistical approach to modeling job accessibility by transportation mode: Case study of Columbus, Ohio. *Journal of Transport Geography*, *45*, 1–11. https://doi.org/10.1016/j.jtrangeo.2015.03.015

Wang, H., Palm, M., Chen, C., Vogt, R., & Wang, Y. (2016). Does bicycle network level of traffic stress (LTS) explain bicycle travel behavior? Mixed results from an Oregon case study. *Journal of Transport Geography*, *57*, 8–18. https://doi.org/10.1016/j.jtrangeo.2016.08.016

Wolfe, A. (2018). *Logistic and Linear Regression Assumptions: Violation Recognition and Control*. 21.

Xia, T., Zhang, Y., Crabb, S., & Shah, P. (2013, July 16). *Cobenefits of Replacing Car Trips with Alternative Transportation: A Review of Evidence and Methodological Issues* [Review Article]. Journal of Environmental and Public Health; Hindawi. https://doi.org/10.1155/2013/797312

Ye, M., Zeng, S., Yang, G., & Chen, Y. (2020). Identification of contributing factors on travel mode choice among different resident types with bike-sharing as an alternative. *IET Intelligent Transport Systems*, *14*(7), 639–646. https://doi.org/10.1049/iet-its.2019.0581

Appendix A - R Codes/Scripts

The following codes were created to complete various analysis and data management tasks throughout the completion of this project.

Appendix A Script A.1 – Combination and Subsetting of PUMS Data

The following code was used to combine multiple years of both household and individual PUMS data into a single dataset as well as subset the data to include data from only those PUMAs used in the study.

Appendix A Script A.2 – Population Data Management

The following code was used to calculate population density for each PUMA within the study area using person's weight from PUMS data and spatial measurements obtained using ArcGIS Pro.

```
setwd("R:\\Data")

#Reads final altered data from R-Script "PUMS_Merging"
PUMS <- read.csv("Altered Data\\PUMS.csv", header=T, sep=",", stringsAsFactors = F)

#Reads final altered data from ArcGIS Project "PUMA_Spatial_Characteristics", Layer "Chicago_PUMAs_UTM16N"
PUMS_Spatial <- read.csv("Altered Data\\Chicago_PUMA_Areas_UTM16N.csv", header=T, sep=",",
stringsAsFactors = F)

#_____________________________________________________________________________

#Changes PUMA variable name in PUMS_Spatial for merging
PUMS_Spatial$PUMA <- PUMS_Spatial$PUMACE10

#Merges PUMS with the spatial data
PUMS <- merge(PUMS, PUMS_Spatial, by ="PUMA")

#_____________________________________________________________________________

#Calculates population density per representative person using PWGTP (Person's Weight)
PUMS$PopDensP <- (PUMS$PWGTP/PUMS$AreaSqMi)

library(dplyr)

#Calculates sum of density per person to determine total population density within the PUMAs
#Function dependent on package "dplyr"
ResDens <- (PUMS %>%
              group_by(PUMA_Yr) %>%
              summarize(PopDens = sum(PopDensP)) %>%
              ungroup)

#_____________________________________________________________________________

#Removes unneeded data from R workspace
rm(PUMS, PUMS_Spatial)

#_____________________________________________________________________________

write.csv(ResDens,"Altered Data\\SpatialTemporal\\ResDens.csv", row.names = FALSE)
```

Appendix A Script A.3 – Job Data Management

The following code was used to calculate job density using counts from US Census LEHD LODES Work Area Characteristics data and spatial measurements obtained using ArcGIS Pro.

```
setwd("R:\\Data")

#Reads block and tract to PUMA data from the Census
Blocks <- read.csv("Original Data\\Census LEHD LODES\\Chicago_Blocks.csv")
Tracts.PUMAs <- read.csv("Original Data\\Tract_to_PUMA_2010.csv")

#Reads final altered data from ArcGIS Project "PUMA_Spatial_Characteristics", Layer "Chicago_PUMAs_UTM16N"
PUMS_Spatial <- read.csv("Altered Data\\Chicago_PUMA_Areas_UTM16N.csv", header=T, sep=",",
stringsAsFactors = F)

#Reads original work area characteristic (WAC) data
WAC_2012 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2012.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2013 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2013.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2014 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2014.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2015 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2015.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2016 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2016.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2017 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2017.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2018 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2018.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2019 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2019.csv", header=T,
sep=",", stringsAsFactors = F)

#_____________________________________________________________________________

#Adds year variable to each of the WAC datasets
WAC_2012$Year <- "2012"
WAC_2013$Year <- "2013"
WAC_2014$Year <- "2014"
WAC_2015$Year <- "2015"
WAC_2016$Year <- "2016"
WAC_2017$Year <- "2017"
WAC_2018$Year <- "2018"
WAC_2019$Year <- "2019"

library(plyr)
#Combines all of the WAC dataframes into a single WAC dataframe
#Function dependent on package "plyr"
WAC <- rbind.fill(WAC_2012, WAC_2013, WAC_2014, WAC_2015, WAC_2016, WAC_2017, WAC_2018,
WAC_2019)
```

```
#Removes unneeded data from R workspace
rm(WAC_2012, WAC_2013, WAC_2014, WAC_2015, WAC_2016, WAC_2017, WAC_2018, WAC_2019)

#_____________________________________________________________________________

#Changes GEOID variable name and type for merging
WAC$GEOID <- as.character(WAC$w_geocode)

#library(tidyverse)
#WAC <- WAC %>% select("GEOID", "C000", "Year")

#_____________________________________________________________________________

#Changes GEOID variable name and type for merging
Blocks$GEOID <- as.character(Blocks$GEOID10)
#Changes Tract GEOID variable name for merging
Blocks$TRACTCE <- Blocks$TRACTCE10

#Merges WAC and block files for tract reference
WAC <- merge(WAC, Blocks, by = "GEOID")
#Merges WAC and tract files for PUMA reference
WAC <- merge(WAC, Tracts.PUMAs, by = "TRACTCE")

#Removes unneeded data from R workspace
rm(Blocks, Tracts.PUMAs)

#Subsets WAC data to include only the PUMAs used in the study
WAC <- subset(WAC, PUMA5CE=="3501" |
            PUMA5CE=="3502" | PUMA5CE=="3503" |
            PUMA5CE=="3504" | PUMA5CE=="3520" |
            PUMA5CE=="3521" | PUMA5CE=="3522" |
            PUMA5CE=="3523" | PUMA5CE=="3524" |
            PUMA5CE=="3525" | PUMA5CE=="3526" |
            PUMA5CE=="3527" | PUMA5CE=="3528" |
            PUMA5CE=="3529" | PUMA5CE=="3530" |
            PUMA5CE=="3531" | PUMA5CE=="3532")

#Changes Tract GEOID variable name for merging
WAC$PUMACE10 <- as.character(WAC$PUMA5CE)

#Merges WAC data with the spatial data
WAC <- merge(WAC, PUMS_Spatial, by = "PUMACE10")

#Removes unneeded data from R workspace
rm(PUMS_Spatial)

#_____________________________________________________________________________

#Adds a new field "PUMA_Yr" that will be used to join to PUMS dataset
WAC$PUMA_Yr <- paste0(WAC$PUMACE10,"_",WAC$Year)

#Calculates work density of a single block/PUMA
```

```
WAC$PartWorkDens <- (WAC$C000/WAC$AreaSqMi)

#Removes package "plyr" as it interferes with the function below
detach("package:plyr", unload=TRUE)

library(dplyr)

#Calculates sum of density per block to determine total population density within the PUMAs
#Function dependent on package "dplyr"
WorkDens <- (WAC %>%
        group_by(PUMA_Yr) %>%
        summarize(WorkDens = sum(PartWorkDens)) %>%
        ungroup)

#_____________________________________________________________________________

write.csv(WorkDens,"Altered Data\\SpatialTemporal\\WorkDens.csv", row.names = FALSE)
```

Appendix A Script A.4 – Assessor Data Management

The following code was used to calculate the mean and median property values for each PUMA within the study area using the Cook County Assessor's Assessments data.

```
setwd("R:\\Data")

#_____________________________________________________________________________

#Reads original Assessor data
Assessments <- read.csv("Original Data\\Assessments\\Assessor_-_Historic_Assessed_Values.csv", header=T,
sep=",", stringsAsFactors = F)

library(tidyverse)

#Reduces data to only the needed fields
#Function dependent on package "tidyverse"
Assessments <- Assessments %>% select("pin", "tax_year", "certified_tot")

#Subsets assessment data to only the used years to reduce processing requirements
Assessments <- subset(Assessments, tax_year=="2012" |
                      tax_year=="2013" | tax_year=="2014" |
                      tax_year=="2015" | tax_year=="2016" |
                      tax_year=="2017" | tax_year=="2018" |
                      tax_year=="2019")

#_____________________________________________________________________________


library(data.table)
library(bit64) #Specifically used for "pin" field in Cook County parcel data

#Reads only the needed fields from the parcel data due to the large file size
#Function dependent on packages "data.table" and "bit64"
Parcels <- fread("Original Data\\Assessments\\Assessor_-_Parcel_Universe.csv", sep=",", select = c("pin",
"census_puma_geoid"))

#Subsets parcel data to only the used PUMAs to reduce processing requirements
Parcels <- subset(Parcels, census_puma_geoid=="1703501" |
                    census_puma_geoid=="1703502" | census_puma_geoid=="1703503" |
                    census_puma_geoid=="1703504" | census_puma_geoid=="1703520" |
                    census_puma_geoid=="1703521" | census_puma_geoid=="1703522" |
                    census_puma_geoid=="1703523" | census_puma_geoid=="1703524" |
                    census_puma_geoid=="1703525" | census_puma_geoid=="1703526" |
                    census_puma_geoid=="1703527" | census_puma_geoid=="1703528" |
                    census_puma_geoid=="1703529" | census_puma_geoid=="1703530" |
                    census_puma_geoid=="1703531" | census_puma_geoid=="1703532")

#_____________________________________________________________________________
```

```
#Reformats pins for joining
Assessments$pin <- as.character(Assessments$pin)
Parcels$pin <- as.character(Parcels$pin)

#_____________________________________________________________________________

#Merges assessment and parcel data
Assessments <- merge(Assessments, Parcels, by="pin")

#Removes unneeded data from R workspace
rm(Parcels)

#_____________________________________________________________________________

#Removes first three digits to leave only the PUMA number
Assessments$PUMA <- substring(Assessments$census_puma_geoid, 4)

Assessments$PUMA_Yr <- paste0(Assessments$PUMA,"_",Assessments$tax_year)

#_____________________________________________________________________________

library(dplyr)

#Calculates mean property values by PUMA_Yr
#Function dependent on package "dplyr"
AssessmentsMean <- (Assessments %>%
        group_by(PUMA_Yr) %>%
        summarize(MeanPropertyValue = mean(certified_tot)) %>%
        ungroup)

#Calculates median property values by PUMA_Yr
#Function dependent on package "dplyr"
AssessmentsMedian <- (Assessments %>%
        group_by(PUMA_Yr) %>%
        summarize(MedianPropertyValue = median(certified_tot)) %>%
        ungroup)

#Creates data frame which includes both mean and median property values by PUMA_Yr
AssessmentsAggregated <- merge(AssessmentsMean, AssessmentsMedian, by = "PUMA_Yr")

#Removes unneeded data from R workspace
rm(Assessments, AssessmentsMean, AssessmentsMedian)

#_____________________________________________________________________________

write.csv(AssessmentsAggregated,"Altered Data\\SpatialTemporal\\PropertyValues.csv", row.names = FALSE)
```

Appendix A Script A.5 – Crime Data Management

The following code was used to reclassify and calculate the density of various categories of criminal activity for each PUMA within the study area.

```
setwd("R:\\Data")

#Reads original crime data
Crimes <- read.csv("Original Data\\Crimes_Chicago_2001-present.csv", header=T, sep=",", stringsAsFactors = F)

#Reads community area data derived from spatial data in GIS
CommunityAreas <- read.csv("Altered Data\\GIS CSVs\\CommunityAreas.csv", header=T, sep=",", stringsAsFactors
= F)

#_____________________________________________________________________________

#identify unique values for crime types
#unique(Crimes[c("Primary.Type")])

#Adds count variable to crime data for aggregation
Crimes$Count <- seq.int(c(1))

#Changes name of community area field for merging
CommunityAreas$Community.Areas <- CommunityAreas$area_num_1

#Merges crime data with community area data which includes PUMAs
Crimes <- merge(Crimes,CommunityAreas,by="Community.Areas")

#Removes unneeded data from R workspace
rm(CommunityAreas)

#_____________________________________________________________________________

library(tidyverse)

#Function dependent on package "tidyverse"
Crimes$Year <- as.POSIXct(Crimes$Date, format = "%m/%d/%Y %H:%M:%S")
Crimes$Year <- format(Crimes$Year, format="%Y")

#Subsets crime data to include only the years used in the study
Crimes <- subset(Crimes, Year=="2012" |
                 Year=="2013" | Year=="2014" |
                 Year=="2015" | Year=="2016" |
                 Year=="2017" | Year=="2018" |
                 Year=="2019")

#Subsets crime data to include only the PUMAs used in the study
Crimes <- subset(Crimes, PUMA=="3501" |
           PUMA=="3502" | PUMA=="3503" |
           PUMA=="3504" | PUMA=="3520" |
```

```
        PUMA=="3521" | PUMA=="3522" |
        PUMA=="3523" | PUMA=="3524" |
        PUMA=="3525" | PUMA=="3526" |
        PUMA=="3527" | PUMA=="3528" |
        PUMA=="3529" | PUMA=="3530" |
        PUMA=="3531" | PUMA=="3532")

#Adds a new field "ComArea_Yr" that will be used to merge the various community data
Crimes$ComArea_Yr <- paste0(Crimes$Community.Areas,"_",Crimes$Year)

#_____________________________________________________________________________

#Reclassifies crimes into simplified crime types
Crimes$Type <- factor(Crimes$Primary.Type, levels = c("HOMICIDE",

                                                      "BATTERY",
                                                      "ASSAULT",
                                                      "ROBBERY",
                                                      "CRIM SEXUAL ASSAULT",
                                                      "CRIMINAL SEXUAL ASSAULT",
                                                      "KIDNAPPING",
                                                      "INTIMIDATION",
                                                      "HUMAN TRAFFICKING",
                                                      "RITUALISM",
                                                      "DOMESTIC VIOLENCE",

                                                      "THEFT",
                                                      "BURGLARY",
                                                      "CRIMINAL DAMAGE",
                                                      "MOTOR VEHICLE THEFT",
                                                      "ARSON",

                                                      "DECEPTIVE PRACTICE",
                                                      "NARCOTICS",
                                                      "WEAPONS VIOLATION",
                                                      "CRIMINAL TRESPASS",
                                                      "SEX OFFENSE",
                                                      "INTERFERENCE WITH PUBLIC OFFICER",
                                                      "PUBLIC PEACE VIOLATION",
                                                      "PROSTITUTION",
                                                      "GAMBLING",
                                                      "LIQUOR LAW VIOLATION",
                                                      "STALKING",
                                                      "CONCEALED CARRY LICENSE VIOLATION",
                                                      "OBSCENITY",
                                                      "PUBLIC INDECENCY",
                                                      "OTHER NARCOTIC VIOLATION",

                                                      "OTHER OFFENSE",
                                                      "OFFENSE INVOLVING CHILDREN",
                                                      "NON - CRIMINAL",
                                                      "NON-CRIMINAL",
                                                      "NON-CRIMINAL (SUBJECT SPECIFIED)"),
```

```
            labels = c("Homicide",

"ViolentCrime","ViolentCrime","ViolentCrime","ViolentCrime","ViolentCrime","ViolentCrime","ViolentCrime","Viol
entCrime","ViolentCrime","ViolentCrime",
                       "PropertyCrime","PropertyCrime","PropertyCrime","PropertyCrime","PropertyCrime",

"NonviolentCrime","NonviolentCrime","NonviolentCrime","NonviolentCrime","NonviolentCrime","NonviolentCrim
e","NonviolentCrime","NonviolentCrime","NonviolentCrime","NonviolentCrime","NonviolentCr
ime","NonviolentCrime","NonviolentCrime","NonviolentCrime",
                       "Other","Other","Other","Other","Other"))

#_____________________________________________________________________________

#Creates binary variable for each of the crime types
Crimes$Homicide <- factor(Crimes$Type, levels =
c("Homicide","ViolentCrime","PropertyCrime","NonviolentCrime","Other"), labels = c(1,0,0,0,0))
Crimes$Homicide <- as.numeric(as.character(Crimes$Homicide))

Crimes$ViolentCrime <- factor(Crimes$Type, levels =
c("Homicide","ViolentCrime","PropertyCrime","NonviolentCrime","Other"), labels = c(0,1,0,0,0))
Crimes$ViolentCrime <- as.numeric(as.character(Crimes$ViolentCrime))

Crimes$PropertyCrime <- factor(Crimes$Type, levels =
c("Homicide","ViolentCrime","PropertyCrime","NonviolentCrime","Other"), labels = c(0,0,1,0,0))
Crimes$PropertyCrime <- as.numeric(as.character(Crimes$PropertyCrime))

Crimes$NonviolentCrime <- factor(Crimes$Type, levels =
c("Homicide","ViolentCrime","PropertyCrime","NonviolentCrime","Other"), labels = c(0,0,0,1,0))
Crimes$NonviolentCrime <- as.numeric(as.character(Crimes$NonviolentCrime))

Crimes$OtherCrime <- factor(Crimes$Type, levels =
c("Homicide","ViolentCrime","PropertyCrime","NonviolentCrime","Other"), labels = c(0,0,0,0,1))
Crimes$OtherCrime <- as.numeric(as.character(Crimes$OtherCrime))

#_____________________________________________________________________________

#Aggregates each variable by community area
CrimesAgComArea.Homicide <- (Crimes %>%
                group_by(ComArea_Yr) %>%
                summarize(Homicide = sum(Homicide)) %>%
                ungroup)

CrimesAgComArea.ViolentCrime <- (Crimes %>%
                group_by(ComArea_Yr) %>%
                summarize(ViolentCrime = sum(ViolentCrime)) %>%
                ungroup)

CrimesAgComArea.PropertyCrime <- (Crimes %>%
                group_by(ComArea_Yr) %>%
                summarize(PropertyCrime = sum(PropertyCrime)) %>%
                ungroup)

CrimesAgComArea.NonviolentCrime <- (Crimes %>%
```

```
            group_by(ComArea_Yr) %>%
            summarize(NonviolentCrime = sum(NonviolentCrime)) %>%
            ungroup)

CrimesAgComArea.OtherCrime <- (Crimes %>%
            group_by(ComArea_Yr) %>%
            summarize(OtherCrime = sum(OtherCrime)) %>%
            ungroup)

CrimesAgComArea.Count <- (Crimes %>%
            group_by(ComArea_Yr) %>%
            summarize(Crimes = sum(Count)) %>%
            ungroup)

CrimesAgComArea.AreaSqMi <- (Crimes %>%
            group_by(ComArea_Yr) %>%
            summarize(Area.ComArea = mean(SqMi)) %>%
            ungroup)

CrimesAgComArea.PUMAs <- (Crimes %>%
            group_by(ComArea_Yr) %>%
            summarize(PUMA = mean(PUMA)) %>%
            ungroup)

Crimes$Year <- as.numeric(Crimes$Year)
CrimesAgComArea.Year <- (Crimes %>%
            group_by(ComArea_Yr) %>%
            summarize(Year = mean(Year)) %>%
            ungroup)

#Merges all aggregated files relating to Community Areas
CrimesAgComArea <- merge(CrimesAgComArea.Count,CrimesAgComArea.AreaSqMi,by="ComArea_Yr")

CrimesAgComArea <- merge(CrimesAgComArea,CrimesAgComArea.PUMAs,by="ComArea_Yr")

CrimesAgComArea <- merge(CrimesAgComArea,CrimesAgComArea.Year,by="ComArea_Yr")

CrimesAgComArea <- merge(CrimesAgComArea,CrimesAgComArea.Homicide,by="ComArea_Yr")

CrimesAgComArea <- merge(CrimesAgComArea,CrimesAgComArea.ViolentCrime,by="ComArea_Yr")

CrimesAgComArea <- merge(CrimesAgComArea,CrimesAgComArea.PropertyCrime,by="ComArea_Yr")

CrimesAgComArea <- merge(CrimesAgComArea,CrimesAgComArea.NonviolentCrime,by="ComArea_Yr")

CrimesAgComArea <- merge(CrimesAgComArea,CrimesAgComArea.OtherCrime,by="ComArea_Yr")

#Removes unneeded data from R workspace
rm(CrimesAgComArea.Count, CrimesAgComArea.AreaSqMi, CrimesAgComArea.PUMAs, CrimesAgComArea.Year,
CrimesAgComArea.Homicide, CrimesAgComArea.NonviolentCrime,CrimesAgComArea.PropertyCrime,
CrimesAgComArea.ViolentCrime, CrimesAgComArea.OtherCrime, Crimes)

#_____________________________________________________________________________
```

```
#Adds a new field "PUMA_Yr" that will be used to join to PUMS dataset
CrimesAgComArea$PUMA_Yr <- paste0(CrimesAgComArea$PUMA,"_",CrimesAgComArea$Year)

#Aggregates each variable by PUMA_Yr
CrimesAgPUMA.Count <- (CrimesAgComArea %>%
            group_by(PUMA_Yr) %>%
            summarize(Crimes = sum(Crimes)) %>%
            ungroup)

CrimesAgPUMA.AreaSqMi <- (CrimesAgComArea %>%
            group_by(PUMA_Yr) %>%
            summarize(Area2 = sum(Area.ComArea)) %>%
            ungroup)

CrimesAgPUMA.Year <- (CrimesAgComArea %>%
            group_by(PUMA_Yr) %>%
            summarize(Year = mean(Year)) %>%
            ungroup)

CrimesAgPUMA.Homicide <- (CrimesAgComArea %>%
            group_by(PUMA_Yr) %>%
            summarize(Homicide = sum(Homicide)) %>%
            ungroup)

CrimesAgPUMA.ViolentCrime <- (CrimesAgComArea %>%
            group_by(PUMA_Yr) %>%
            summarize(ViolentCrime = sum(ViolentCrime)) %>%
            ungroup)

CrimesAgPUMA.PropertyCrime <- (CrimesAgComArea %>%
            group_by(PUMA_Yr) %>%
            summarize(PropertyCrime = sum(PropertyCrime)) %>%
            ungroup)

CrimesAgPUMA.NonviolentCrime <- (CrimesAgComArea %>%
            group_by(PUMA_Yr) %>%
            summarize(NonviolentCrime = sum(NonviolentCrime)) %>%
            ungroup)

CrimesAgPUMA.OtherCrime <- (CrimesAgComArea %>%
            group_by(PUMA_Yr) %>%
            summarize(OtherCrime = sum(OtherCrime)) %>%
            ungroup)

#Merges all aggregated files relating to PUMA by PUMA_Yr
CrimesAggregated <- merge(CrimesAgPUMA.Count,CrimesAgPUMA.AreaSqMi,by="PUMA_Yr")

CrimesAggregated <- merge(CrimesAggregated,CrimesAgPUMA.Year,by="PUMA_Yr")

CrimesAggregated <- merge(CrimesAggregated,CrimesAgPUMA.Homicide,by="PUMA_Yr")

CrimesAggregated <- merge(CrimesAggregated,CrimesAgPUMA.ViolentCrime,by="PUMA_Yr")
```

```
CrimesAggregated <- merge(CrimesAggregated,CrimesAgPUMA.PropertyCrime,by="PUMA_Yr")

CrimesAggregated <- merge(CrimesAggregated,CrimesAgPUMA.NonviolentCrime,by="PUMA_Yr")

CrimesAggregated <- merge(CrimesAggregated,CrimesAgPUMA.OtherCrime,by="PUMA_Yr")

#Removes unneeded data from R workspace
rm(CrimesAgPUMA.NonviolentCrime, CrimesAgPUMA.PropertyCrime, CrimesAgPUMA.ViolentCrime,
CrimesAgPUMA.Homicide, CrimesAgPUMA.Year, CrimesAgPUMA.AreaSqMi, CrimesAgPUMA.Count,
CrimesAgPUMA.OtherCrime, CrimesAgComArea)

#_____________________________________________________________________________

#Determines the density of various types of crime in each PUMA_Yr
CrimesAggregated$CrimeDens <- (CrimesAggregated$Crimes/CrimesAggregated$Area2)
CrimesAggregated$HomicideDens <- (CrimesAggregated$Homicide/CrimesAggregated$Area2)
CrimesAggregated$ViolentCrimeDens <- (CrimesAggregated$ViolentCrime/CrimesAggregated$Area2)
CrimesAggregated$PropertyCrimeDens <- (CrimesAggregated$PropertyCrime/CrimesAggregated$Area2)
CrimesAggregated$NonviolentCrimeDens <- (CrimesAggregated$NonviolentCrime/CrimesAggregated$Area2)
CrimesAggregated$OtherCrimeDens <- (CrimesAggregated$OtherCrime/CrimesAggregated$Area2)

#_____________________________________________________________________________

#Reduces data to only the needed fields
#Function dependent on package "tidyverse"
CrimesAggregated <- CrimesAggregated %>% select("PUMA_Yr", "CrimeDens", "HomicideDens",
"ViolentCrimeDens", "PropertyCrimeDens", "NonviolentCrimeDens", "OtherCrimeDens")

#_____________________________________________________________________________

write.csv(CrimesAggregated,"Altered Data\\SpatialTemporal\\Crimes.csv", row.names = FALSE)
```

Appendix A Script A.6 – Bicycle Accessibility Index Generation

The following code was used to create a bicycle accessibility index using a combination of Census TIGER roads files, bicycle facility data provided by the City of Chicago, and Census LEHD LODES Work Area Characteristics (WAC) and Residential Area Characteristics (RAC) data. The original isochrones were generated for each year use ArcGIS Network Analyst.

Pre-GIS LEHD Origin-Destination Employment Statistics (LODES) Management

The following code was used merge all LEHD LODES Work Area Characteristics (WAC) and Residential Area Characteristics (RAC) data into a single dataset so that it can be easily merged to a TIGER/Line block shapefile.

```
setwd("R:\\Data")

#Reads csv containing all blocks within Illinois
Blocks <- read.csv("Altered Data\\GIS CSVs\\IL_Blocks.csv")

#Reads CSVs containing Work Area Characteristic data for every year in the study
WAC_2012 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2012.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2013 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2013.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2014 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2014.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2015 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2015.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2016 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2016.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2017 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2017.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2018 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2018.csv", header=T,
sep=",", stringsAsFactors = F)
WAC_2019 <- read.csv("Original Data\\Census LEHD LODES\\WAC\\il_wac_S000_JT00_2019.csv", header=T,
sep=",", stringsAsFactors = F)

#Reads CSVs containing Residential Area Characteristic data for every year in the study
RAC_2012 <- read.csv("Original Data\\Census LEHD LODES\\RAC\\il_rac_S000_JT00_2012.csv", header=T, sep=",",
stringsAsFactors = F)
RAC_2013 <- read.csv("Original Data\\Census LEHD LODES\\RAC\\il_rac_S000_JT00_2013.csv", header=T, sep=",",
stringsAsFactors = F)
RAC_2014 <- read.csv("Original Data\\Census LEHD LODES\\RAC\\il_rac_S000_JT00_2014.csv", header=T, sep=",",
stringsAsFactors = F)
RAC_2015 <- read.csv("Original Data\\Census LEHD LODES\\RAC\\il_rac_S000_JT00_2015.csv", header=T, sep=",",
stringsAsFactors = F)
```

```
RAC_2016 <- read.csv("Original Data\\Census LEHD LODES\\RAC\\il_rac_S000_JT00_2016.csv", header=T, sep=",",
stringsAsFactors = F)
RAC_2017 <- read.csv("Original Data\\Census LEHD LODES\\RAC\\il_rac_S000_JT00_2017.csv", header=T, sep=",",
stringsAsFactors = F)
RAC_2018 <- read.csv("Original Data\\Census LEHD LODES\\RAC\\il_rac_S000_JT00_2018.csv", header=T, sep=",",
stringsAsFactors = F)
RAC_2019 <- read.csv("Original Data\\Census LEHD LODES\\RAC\\il_rac_S000_JT00_2019.csv", header=T, sep=",",
stringsAsFactors = F)

#_____________________________________________________________________________

#Reduces data to include only the field representing total jobs within the block
WAC_2012$WAC_2012 <- WAC_2012$C000
WAC_2013$WAC_2013 <- WAC_2013$C000
WAC_2014$WAC_2014 <- WAC_2014$C000
WAC_2015$WAC_2015 <- WAC_2015$C000
WAC_2016$WAC_2016 <- WAC_2016$C000
WAC_2017$WAC_2017 <- WAC_2017$C000
WAC_2018$WAC_2018 <- WAC_2018$C000
WAC_2019$WAC_2019 <- WAC_2019$C000

#Reduces data to include only the field representing total residents within the block
RAC_2012$RAC_2012 <- RAC_2012$C000
RAC_2013$RAC_2013 <- RAC_2013$C000
RAC_2014$RAC_2014 <- RAC_2014$C000
RAC_2015$RAC_2015 <- RAC_2015$C000
RAC_2016$RAC_2016 <- RAC_2016$C000
RAC_2017$RAC_2017 <- RAC_2017$C000
RAC_2018$RAC_2018 <- RAC_2018$C000
RAC_2019$RAC_2019 <- RAC_2019$C000

#_____________________________________________________________________________

#Merges all work data into a single dataset
WAC <- merge(WAC_2012, WAC_2013, by = "w_geocode", all = TRUE)
WAC <- merge(WAC, WAC_2014, by = "w_geocode", all = TRUE)
WAC <- merge(WAC, WAC_2015, by = "w_geocode", all = TRUE)
WAC <- merge(WAC, WAC_2016, by = "w_geocode", all = TRUE)
WAC <- merge(WAC, WAC_2017, by = "w_geocode", all = TRUE)
WAC <- merge(WAC, WAC_2018, by = "w_geocode", all = TRUE)
WAC <- merge(WAC, WAC_2019, by = "w_geocode", all = TRUE)

#Merges all residential data into a single dataset
RAC <- merge(RAC_2012, RAC_2013, by = "h_geocode", all = TRUE)
RAC <- merge(RAC, RAC_2014, by = "h_geocode", all = TRUE)
RAC <- merge(RAC, RAC_2015, by = "h_geocode", all = TRUE)
RAC <- merge(RAC, RAC_2016, by = "h_geocode", all = TRUE)
RAC <- merge(RAC, RAC_2017, by = "h_geocode", all = TRUE)
RAC <- merge(RAC, RAC_2018, by = "h_geocode", all = TRUE)
RAC <- merge(RAC, RAC_2019, by = "h_geocode", all = TRUE)

#Removes all unneeded files
```

```
rm(WAC_2012, WAC_2013, WAC_2014, WAC_2015, WAC_2016, WAC_2017, WAC_2018, WAC_2019, RAC_2012,
RAC_2013, RAC_2014, RAC_2015, RAC_2016, RAC_2017, RAC_2018, RAC_2019)

#_____________________________________________________________________________

#Ensures all GEOIDs are in the same format for effective merging
WAC$GEOID <- as.character(WAC$w_geocode)
RAC$GEOID <- as.character(RAC$h_geocode)
Blocks$GEOID <- as.character(Blocks$GEOID10)

#Merges WAC and RAC data to the block data
LODES_WACandRAC <- merge(Blocks, RAC, by = "GEOID", all = TRUE)
LODES_WACandRAC <- merge(LODES_WACandRAC, WAC, by = "GEOID", all = TRUE)

#Removes all unneeded files
rm(Blocks, WAC, RAC)

#_____________________________________________________________________________

library(tidyverse)

#Reduces variables to only include those needed for analysis
LODES_WACandRAC <- LODES_WACandRAC %>% select("GEOID", "WAC_2012", "WAC_2013", "WAC_2014",
"WAC_2015", "WAC_2016", "WAC_2017", "WAC_2018", "WAC_2019", "RAC_2012", "RAC_2013", "RAC_2014",
"RAC_2015", "RAC_2016", "RAC_2017", "RAC_2018", "RAC_2019")

#Turns any null WAC and RAC values to zeros for analysis
LODES_WACandRAC[is.na(LODES_WACandRAC)] <- 0

#_____________________________________________________________________________

write.csv(LODES_WACandRAC,"Altered Data\\SpatialTemporal\\Intermediate
Data\\LODES_WACandRAC_Illinios.csv", row.names = FALSE)
```

Determining Jobs Accessible Per Facility (Step 1)

The following code uses a Census block shapefile generate in ArcGIS which contains the Work Area Characteristics and Residential Area Characteristics data as well as isochrones generated using ArcGIS Network Analyst to determine the total number of jobs accessible within a 20-minute bike for each facility in a given year. This process requires repeating for each year in the study and the "find" (Control + F) function can be used to replace all instances of a year with a different year.

```
#Packages needed for the following spatial analysis
library(sf)
library(rgdal)
library(rgeos)

setwd("R:\\Data\\Altered Data\\GIS for R")

Blocks <- st_read("LODES_within_StudyArea.shp")
#Service areas created using ArcGIS Network Analyst
ServiceAreas <- st_zm(st_read("SA_2019_20min.shp"))

#_____________________________________________________________________________

#Creates variables to be used as placeholders for data in the iterative process
FacilityID <- c(NULL)
Sum_WAC <- c(NULL)
WAC_by_Facility <- data.frame(FacilityID, Sum_WAC)

#Values must be added so the datasets have at least one record to replace
FacilityID <- c(1)
Sum_WAC <- c(1)
Output <- data.frame(FacilityID, Sum_WAC)

#_____________________________________________________________________________

#Identifies and creates dataset for unique Facility values for iteration
Facility_IDs <- unique(ServiceAreas$FacilityID)

#Iterates for each facility in the service areas
for(Facility in Facility_IDs) {

  #Subsets to single facility
  SA_by_Facility <- subset(ServiceAreas, FacilityID == Facility)

  #Determines whether blocks are within the servicea area
  InSA <- lengths(st_intersects(Blocks, SA_by_Facility)) > 0

  #Makes dataframe with only blocks within selected service area
```

```
  Blocks_InSA <- Blocks$RAC_2019[InSA]

  #Changes existing Output dataframe to represent aggregate data about WAC counts
  Output$FacilityID <- Facility

  #Determines the sum of jobs accessible from facility
  Output$Sum_WAC <- sum(Blocks_InSA)

  #Appends Output to WAC_by_Facility for complete dataset
  WAC_by_Facility <- rbind(WAC_by_Facility, Output)
}

#_____________________________________________________________________________________

write.csv(WAC_by_Facility,"WAC_per_Facility_SA_2019.csv", row.names = FALSE)

rm(WAC_by_Facility, WAC_by_Facility_fix)
```

Creating Aggregated Indices (Step 2)

The following code uses data created in the previous code and aggregates this data to provide averages for each PUMA in a given year. This process requires repeating for each year in the study and the "find" (Control + F) function can be used to replace all instances of a year with a different year.

```
#Library allows us to load and edit shapefiles
library(sf)

setwd("R:\\Data\\Altered Data\\GIS for R")

#Reads output of previous code
WAC_by_Facility <- read.csv("WAC_per_Facility_SA_2019.csv")

#Reads shapefile containing each of the starting facilities that still contains WAC and RAC data
LODES <- read.csv("Chicago_LODES_BlockGroup.csv")

#_____________________________________________________________________________

#Renames Facility IDs for merging datasets
LODES$FacilityID <- LODES$ï..OID_

#Merges facilities with job totals with their original block groups containing RAC data and PUMA identifiers
WAC_by_BlockGroup <- merge(WAC_by_Facility, LODES, by="FacilityID")

library(tidyverse)
#Aggregates residents to the PUMA level so that jobs accessible may be averaged
RAC_Aggregated <- (WAC_by_BlockGroup %>%
                     group_by(PUMA) %>%
                     summarize(PUMA_RAC = sum(SUM_RAC_2019)) %>%
                     ungroup)

#Puts WAC per Block Group and RAC per PUMA in one dataset
WAC_by_BlockGroup <- merge(WAC_by_BlockGroup, RAC_Aggregated, by="PUMA")
Bike_INDEX <- WAC_by_BlockGroup %>% select("PUMA", "BlckGrp", "Sum_WAC", "SUM_RAC_2019",
"PUMA_RAC")

#_____________________________________________________________________________

#These two functions together determine the average number of jobs accessible per PUMA
Bike_INDEX$PreAggregate_INDEX <-
Bike_INDEX$Sum_WAC*(Bike_INDEX$SUM_RAC_2019/Bike_INDEX$PUMA_RAC)
Bike_INDEX_2019 <- (Bike_INDEX %>%
                      group_by(PUMA) %>%
                      summarize(IndexValue = sum(PreAggregate_INDEX)) %>%
                      ungroup)

#Creates Year Variable for PUMA_Yr Identifier
```

```
Bike_INDEX_2019$Year <- "2019"
#Creates PUMA_Yr Identifier
Bike_INDEX_2019$PUMA_Yr <- paste0(Bike_INDEX_2019$PUMA,"_",Bike_INDEX_2019$Year)

#_____________________________________________________________________________

write.csv(Bike_INDEX_2019,"Bike_INDEX_2019.csv", row.names = FALSE)
```

Merging Indices from All Years to One Dataset (Step 3)

```
setwd("R:\\Data\\Altered Data\\GIS for R")

Bike_INDEX_2012 <- read.csv("Bike_INDEX_2012.csv")
Bike_INDEX_2013 <- read.csv("Bike_INDEX_2013.csv")
Bike_INDEX_2014 <- read.csv("Bike_INDEX_2014.csv")
Bike_INDEX_2015 <- read.csv("Bike_INDEX_2015.csv")
Bike_INDEX_2016 <- read.csv("Bike_INDEX_2016.csv")
Bike_INDEX_2017 <- read.csv("Bike_INDEX_2017.csv")
Bike_INDEX_2018 <- read.csv("Bike_INDEX_2018.csv")
Bike_INDEX_2019 <- read.csv("Bike_INDEX_2019.csv")

#_____________________________________________________________________________

#Binds all years of the Bike Index into a single dataset
Bicycle_Accessibility_INDEX <- rbind(Bike_INDEX_2012, Bike_INDEX_2013, Bike_INDEX_2014, Bike_INDEX_2015,
Bike_INDEX_2016, Bike_INDEX_2017, Bike_INDEX_2018, Bike_INDEX_2019)

setwd("R:\\Data\\Altered Data\\SpatialTemporal")
write.csv(Bicycle_Accessibility_INDEX,"Bike_INDEX.csv", row.names = FALSE)
```

Appendix A Script A.7 – Data Merging

The following code was used to merge all spatial/temporal data previously described with the PUMS dataset (see A.1 for code to create single dataset for PUMS data).

```
setwd("R:\\Data\\Altered Data")

PUMS <- read.csv("PUMS.csv", header=T, sep=",", stringsAsFactors = F)

#_____________________________________________________________________________

#Makes a dummy variable for tenure where 1 = renting
PUMS$TEN <- ifelse(PUMS$TEN == "3", 1, 0)

#Recodes dis to be 1 = presence of disability and 0 = no disability
PUMS$DIS <- ifelse(PUMS$DIS == "1", 1, 0)

#Recodes race to more intuitive groups
PUMS$RACE <- factor(PUMS$RAC1P, levels = c(1:8),
                    labels = c("white", "black", "native american", "native american", "native american", "asian",
"asian", "other"))

#Recodes employment to more intuitive labels
PUMS$EMP <- factor(PUMS$FES, levels = c(1:8),
                   labels = c("both employed", "married, husband employed", "married, wife employed", "married,
neither employed", "single male employed", "single male unemployed", "single female employed", "single female
unemployed"))

#Makes new field determining number of adults in household
PUMS$ADULTS <- (PUMS$NP - PUMS$NRC)

#Makes binary variable with biking = 1 and all other modes = 0
PUMS$Bike <- factor(PUMS$JWTR, levels = c(1:12),
            labels = c(0,0,0,0,0,0,0,0,1,0,0,0))

#Creates binary variable where Hispanic = 1 and Nonhispanic = 0
PUMS$Hispanic <- factor(PUMS$HISP, levels = c(1:24),
            labels = c(0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1))

#Converts sex from numeric to categorical (nominal) data
PUMS$Sex <- factor(PUMS$SEX, levels = c(1:2),
            labels = c("Male", "Female"))

PUMS$Education <- factor(PUMS$SCHL, levels = c(1:24),
                         labels = c("No Degree", #1
                                    "No Degree", #2
                                    "No Degree", #3
                                    "No Degree", #4
                                    "No Degree", #5
                                    "No Degree", #6
```

```
                "No Degree", #7
                "No Degree", #8
                "No Degree", #9
                "No Degree", #10
                "No Degree", #11
                "No Degree", #12
                "No Degree", #13
                "No Degree", #14
                "No Degree", #15
                "No Degree", #16
                "No Degree", #17
                "No Degree", #18
                "No Degree", #19
                "Undergraduate Degree", #20
                "Undergraduate Degree", #21
                "Graduate Degree", #22
                "Graduate Degree", #23
                "Graduate Degree", #24
                NULL
        ))

library(dplyr)

#perform binning 20-year breaks for age
PUMS <- PUMS %>% mutate(Age = cut(AGEP, breaks=c(0, 20, 60, 70, 110)))

PUMS <- PUMS %>% mutate(HousingCost = coalesce(SMOCP, RNTP))

#_____________________________________________________________________________

#Reads each of the sources of spatial data
BikeIndex <- read.csv("SpatialTemporal\\Bike_INDEX.csv", header=T, sep=",", stringsAsFactors = F)
WAC_Tot <- read.csv("SpatialTemporal\\Intermediate Data\\Total_SA_Jobs_by_Yr.csv")
CrimeData <- read.csv("SpatialTemporal\\Crimes.csv", header=T, sep=",", stringsAsFactors = F)
PropertyValues <- read.csv("SpatialTemporal\\PropertyValues.csv", header=T, sep=",", stringsAsFactors = F)
ResDens <- read.csv("SpatialTemporal\\ResDens.csv", header=T, sep=",", stringsAsFactors = F)
WorkDens <- read.csv("SpatialTemporal\\WorkDens.csv", header=T, sep=",", stringsAsFactors = F)
ACS <- read.csv("SpatialTemporal\\ACS_Data.csv", header=T, sep=",", stringsAsFactors = F)

#Merges all sources of spatial data to PUMS dataset by PUMA_Yr
PUMS <- merge(PUMS, BikeIndex, by ="PUMA_Yr")
PUMS <- merge(PUMS, CrimeData, by ="PUMA_Yr")
PUMS <- merge(PUMS, PropertyValues, by ="PUMA_Yr")
PUMS <- merge(PUMS, ResDens, by ="PUMA_Yr")
PUMS <- merge(PUMS, WorkDens, by ="PUMA_Yr")
PUMS <- merge(PUMS, ACS, by ="PUMA_Yr")

#_____________________________________________________________________________

library(tidyverse)

#Reduces variables to only include those needed for analysis
```

```
Final_Dataset <- PUMS %>% select("PUMA_Yr", "PWGTP", "HINCP", "TEN", "MV", "SCHL", "RACE", "VEH", "AGEP",
"DIS", "PUMA.x", "Year.x", "ADULTS", "EMP", "WorkDens", "PopDens", "MeanPropertyValue",
"MedianPropertyValue", "Bike", "JWMNP", "IndexValue", "CrimeRate", "ViolentCrimeRate", "PropertyCrimeRate",
"NonviolentCrimeRate", "OtherCrimeRate", "JWTR","Hispanic", "Sex", "Age", "Education", "PUMA_Yr",
"MedianRent", "MedPropertyValue", "MedHousingCost", "MedHouseholdIncome", "PropBlack", "NonWhiteProp",
"HousingCost", "JWMNP")

#Removes records containing null values
Final_Dataset <- na.omit(Final_Dataset)


#_____________________________________________________________________________________

#Adjusts index value to be thousands of jobs accessible rather than single jobs for easier interpretation
Final_Dataset$Index1000 <- Final_Dataset$IndexValue/1000

#_____________________________________________________________________________________

#Removes unemployed individuals from dataset since they cannot bike to work
Final_Dataset_Employed <- subset(Final_Dataset, EMP != "married, neither employed" & EMP != "single male
unemployed" & EMP != "single female unemployed")

write.csv(Final_Dataset_Employed,"FinalDataset_Employed.csv", row.names = FALSE)
```

Appendix A Script A.8 – Variable Transformation Testing

The following code was used to create possible transformations for all continuous variables in the dataset and output spreadsheets which include the skewness, minimum, maximum, and correlation coefficient of each of the transformations. This data was used to test the appropriateness of each of the transformations.

```
setwd("R:\\Data\\Altered Data")

MyData <- read.csv("FinalDataset_Employed.csv", header=T, sep=",", stringsAsFactors = F)

#_____________________________________________________________________________

#Standardizes household income variable over the number of adults in a household

MyData$INC_S <- MyData$HINCP/MyData$ADULTS

#Standardizes vehicles variable over the number of adults in a household
MyData$VEH_S <- MyData$VEH/MyData$ADULTS

MyData$MedHousingCostProp <- MyData$MedHousingCost/MyData$MedHouseholdIncome

MyData$People_per_Job <- MyData$PopDens/MyData$WorkDens

#Defines variables for interative process
Variables <- c("WorkDens", "PopDens",
          "MedianPropertyValue", "Index1000", "CrimeRate", "ViolentCrimeRate",
          "PropertyCrimeRate", "NonviolentCrimeRate",
          "OtherCrimeRate", "INC_S", "VEH_S", "MedHouseholdIncome", "MedianRent", "MedPropertyValue",
"PropBlack", "HousingCost", "MedHousingCost", "MedHousingCostProp", "People_per_Job")


#Interative process produces fields for and calculates possible transformations for each variable
for(Variable in Variables) {
  NewColName <- as.name(paste0(Variable,"_sq"))
  MyData[[NewColName]] <- MyData[[Variable]]^2

  NewColName <- as.name(paste0(Variable,"_sqrt"))
  MyData[[NewColName]] <- sqrt(MyData[[Variable]])

  NewColName <- as.name(paste0(Variable,"_exp"))
  MyData[[NewColName]] <- exp(MyData[[Variable]])

  NewColName <- as.name(paste0(Variable,"_ln"))
  MyData[[NewColName]] <- log(MyData[[Variable]])

  NewColName <- as.name(paste0(Variable,"_log1p"))
  MyData[[NewColName]] <- log1p(MyData[[Variable]])
```

```
  NewColName <- as.name(paste0(Variable,"_recip"))
  MyData[[NewColName]] <- 1/(MyData[[Variable]])
}

#Omits null values from the dataset
MyData <- na.omit(MyData)

#Writes a CSV including all possible transformations
write.csv(MyData,"FinalDataset_Employed_Transformed.csv", row.names = FALSE)
#MyData <- read.csv("FinalDataset_Employed_Transformed.csv",header=T,sep=",",stringsAsFactors = F)

#_____________________________________________________________________________

#Library used to calculate skewness
library(moments)

#Pulls only numeric variables from the database so that skewness and correlation can be calculated
NumericData <- MyData[,sapply(MyData, is.numeric)]

#Creates variables to be used as placeholders for data in the iterative process
Name <- c(NULL)
Skewness <- c(NULL)
Min <- c(NULL)
Max <- c(NULL)
#Creates empty dataframe for appending values to
Measures <- data.frame(Name, Skewness, Min, Max)

#Values must be added so the datasets have at least one record to replace
Name <- c(1)
Skewness <- c(1)
Min <- c(1)
Max <- c(1)
#Creates one-record dataframe for temporarily holding values
Output <- data.frame(Name, Skewness, Min, Max)

#Iteratively determines skewness and the minimum and maximum value of each transformation in the dataset
for(Column in names(NumericData)) {
  Output$Name <- Column
  Output$Skewness <- skewness(NumericData[[Column]])
  Output$Min <- min(NumericData[[Column]])
  Output$Max <- max(NumericData[[Column]])

  #Appends new values to "Measures" dataframe
  Measures <- rbind(Measures, Output)
}

#Outputs "Measures" dataframe as a csv for analysis
write.csv(Measures,"Measures.csv", row.names = FALSE)

#_____________________________________________________________________________

#Determines and outputs the correlation coefficients for each of the transformed independent variables
Correlations <- cor(MyData[,sapply(MyData, is.numeric)])
```

```
#Turns correlation variable into a dataframe
Correlations <- as.data.frame(Correlations)

#Outputs "Correlations" dataframe as a csv for analysis
write.csv(Correlations,"Correlations.csv", row.names = TRUE)
```

Appendix A Script A.9 – Creating and Testing a Binary Logistic Regression

```
setwd("R:\\Data\\Altered Data")

MyData <- read.csv("FinalDataset_Employed_Transformed.csv",header=T,sep=",",stringsAsFactors = F)

library(tidyverse)

#_____________________________________________________________________________

WorkDens <- 0

IndexCorrelations <- data.frame(WorkDens)

IndexCorrelations$WorkDens <- cor(MyData$Index1000, log(MyData$WorkDens))
IndexCorrelations$PopDens <- cor(MyData$Index1000, MyData$PopDens)
#IndexCorrelations$MedPropertyValue <- cor(MyData$Index1000, sqrt(MyData$MedPropertyValue))
#IndexCorrelations$MedianRent <- cor(MyData$Index1000, 1/(MyData$MedianRent))
IndexCorrelations$MedHousingCost <- cor(MyData$Index1000, 1/(MyData$MedPropertyValue))
#IndexCorrelations$MedRentPropOfIncome <- cor(MyData$Index1000,
(MyData$MedianRent/MyData$MedHouseholdIncome))
IndexCorrelations$MedHCostPropOfIncome <- cor(MyData$Index1000,
log(MyData$MedHousingCost/MyData$MedHouseholdIncome))
IndexCorrelations$ViolentCrime <- cor(MyData$Index1000, log(MyData$ViolentCrimeRate))
IndexCorrelations$NonviolentCrime <- cor(MyData$Index1000, log(MyData$NonviolentCrimeRate))
IndexCorrelations$PropertyCrime <- cor(MyData$Index1000, log(MyData$PropertyCrimeRate))
IndexCorrelations$NonWhiteProp <- cor(MyData$Index1000, MyData$NonWhiteProp)
IndexCorrelations$MedHouseholdIncome <- cor(MyData$Index1000, MyData$MedHouseholdIncome)
IndexCorrelations$People_per_Job <- cor(MyData$Index1000, MyData$People_per_Job)

#_____________________________________________________________________________

MyData$Education <- factor(MyData$Education, levels = c("No Degree", "Undergraduate Degree", "Graduate
Degree"))
MyData$RACE <- factor(MyData$RACE, levels = c("white", "native american", "asian", "black", "other"), labels =
c("other", "native american", "asian", "black", "other"))

#_____________________________________________________________________________

#Runs binary logistic regression given all desired demographic and spatial variables
LogitRegression <- glm(Bike ~

                         #Index
                           Index1000 +
                         #Density
                           I(PopDens/WorkDens) +
                           I(log(WorkDens)) +
                           #PopDens +
                         #Spatial Household Characteristics
                           I(1000/MedHousingCost) +
                           I(MedHousingCost/MedHouseholdIncome*100) +
                           #Renting/Ownership Separated
```

```
                #I(sqrt(MedPropertyValue)) +
                #I(1/(MedianRent)) +
                #I(MedianRent/MedHouseholdIncome) +
              #Crime
                #I(log(CrimeRate)) +
                #I(log(ViolentCrimeRate)) +
                #I(log(PropertyCrimeRate)) +
                #I(log(NonviolentCrimeRate)) +
                #I(log(NonviolentCrimeRate + PropertyCrimeRate)) +
              #Spatial Demographics
                #NonWhiteProp +
                I(MedHouseholdIncome/1000) +

              #Personal Demographics
                #DIS +                          #not significant in demographic + index model
                as.factor(Age) +
                as.factor(Education) +
                as.factor(Sex) +
                as.factor(RACE) +
                Hispanic +
              #Household Characteristics
                #TEN +                          #Safe to remove (not significant in any iteration)
                #as.factor(MV) +                    #Safe to remove (not significant in any iteration)
                #HousingCost +                      #Overall housing cost (may represent value)
                #I(HousingCost/ADULTS) +                #Housing cost per person
                #I(log(HINCP/ADULTS)) +
                #I(log(HINCP)) +
                #I(HousingCost/HINCP) +
                I(VEH/ADULTS) #+
                 ,

               family="binomial", data = MyData)

summary(LogitRegression)
#write.csv(LogitRegression$coefficients,"Logit.csv", row.names = TRUE)

#Libary required for Variance Inflation Factors (VIFs)
library(car)
#Outputs VIFs for each variable
vif(LogitRegression)

#Libary required for Pseudo R-Squared values
library(DescTools)
#Creates output of all possible pseudo r-squared values
PseudoR2(LogitRegression, which = "all")

#Libary required for Durbin-Watson and Breusch-Godfrey Tests
library(lmtest)
#Durbin-Watson Test
lmtest::dwtest(LogitRegression)
#Breusch-Godfrey Test
lmtest::bgtest(LogitRegression)
```

```
library(glmtoolbox)

hltest(LogitRegression, verbose = TRUE)

#_____________________________________________________________________________

#Adds field with calculated probability of choosing to bike
MyData$BikeChance <- predict(LogitRegression, type = "response")
#Adds field with Log Odds
MyData$LogOdds <- log(MyData$BikeChance/(1-MyData$BikeChance))

#_____________________________________________________________________________

library(ggplot2)

#Remaining code produces graphs to determine variable linearity with Log Odds

Index1000_LogOdds <- ggplot(MyData, aes(LogOdds, Index1000)) +
  geom_point(size = 1.5, alpha = 0.03) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
ggsave(plot = Index1000_LogOdds, width = 5, height = 5, dpi = 600, filename = "Index1000_LogOdds.png")

PeoplePerJob_LogOdds <- ggplot(MyData, aes(LogOdds, I(log(PopDens/WorkDens)))) +
  geom_point(size = 1.5, alpha = 0.03) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
ggsave(plot = PeoplePerJob_LogOdds, width = 5, height = 5, dpi = 600, filename = "PeoplePerJob_LogOdds.png")

WorkDens_LogOdds <- ggplot(MyData, aes(LogOdds, I(log(WorkDens)))) +
  geom_point(size = 1.5, alpha = 0.03) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
ggsave(plot = WorkDens_LogOdds, width = 5, height = 5, dpi = 600, filename = "WorkDens_LogOdds.png")

MedHousingCost_LogOdds <- ggplot(MyData, aes(LogOdds, I(1000/MedHousingCost))) +
  geom_point(size = 1.5, alpha = 0.03) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw() #+
  #ylim(0, 0.005)
ggsave(plot = MedHousingCost_LogOdds, width = 5, height = 5, dpi = 600, filename =
"MedHousingCost_LogOdds.png")

HousingBurden_LogOdds <- ggplot(MyData, aes(LogOdds, I(MedHousingCost/MedHouseholdIncome*100))) +
  geom_point(size = 1.5, alpha = 0.03) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
ggsave(plot = HousingBurden_LogOdds, width = 5, height = 5, dpi = 600, filename =
"HousingBurden_LogOdds.png")

MedHouseholdIncome_LogOdds <- ggplot(MyData, aes(LogOdds, MedHouseholdIncome/1000)) +
  geom_point(size = 1.5, alpha = 0.03) +
  geom_smooth(method = "lm", color = "red") +
```

```
  theme_bw()
ggsave(plot = MedHouseholdIncome_LogOdds, width = 5, height = 5, dpi = 600, filename =
"MedHouseholdIncome_LogOdds.png")

VEH_LogOdds <- ggplot(MyData, aes(LogOdds, I(VEH/ADULTS))) +
  geom_point(size = 1.5, alpha = 0.03) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
ggsave(plot = VEH_LogOdds, width = 5, height = 5, dpi = 600, filename = "VEH_LogOdds.png")

#_____________________________________________________________________________________

Cyclists <- subset(MyData, Bike == 1)
```