August 2019

Deciphering Public Transit Ridership in Baton Rouge: Spatial Disaggregation Approaches

Xuan Kuai

xuankuai91@gmail.com

DECIPHERING PUBLIC TRANSIT RIDERSHIP IN BATON ROUGE: SPATIAL
DISAGGREGATION APPROACHES

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements of the degree of
Doctor of Philosophy

in

The Department of Geography and Anthropology

by
Xuan Kuai
B.E., Wuhan University, China, 2013
M.S., Louisiana State University, 2015
December 2019

This dissertation is dedicated to my wife Xueying. I am deeply grateful for having you in my life.

ACKNOWLEDGEMENTS

First and foremost, I want to express my gratitude to my advisor Dr. Fahui Wang, for your insightful guidance and enduring support throughout my graduate studies. I am forever grateful for what you have taught me to be a better geographer and a better person. The knowledge you taught, the ideas you proposed, and the hours you devoted paved my way to the completion of this dissertation. I appreciate all the support and encouragement you gave me to pursue the Ph.D. degree 275 miles away from the campus. It has been an honor and a great pleasure to work under your advisory. This is the last milestone of this beautiful journey.

My appreciation is also extended to my committee members, Dr. Lei Wang and Dr. Michael Leitner at Department of Geography and Anthropology, and Dr. Chester Wilmot at Department of Civil and Environmental Engineering, for all your insightful suggestions and constructive critiques.

I gratefully acknowledge Mr. Pong Wu at the Capital Region Planning Commission, Mr. Joshua Albritton at the Louisiana Department of Transportation and Development, Mr. Micah Anthony at the Capital Area Transit System, and Ms. Melinda Morang at the ESRI for her help with data and software.

I am also thankful for the help and friendship from my previous manager Mr. Warren Kron at the City of Baton Rouge Parish of East Baton Rouge, my current manager Mr. Brooks Andrews at the Houston-Galveston Area Council, Dr. Qunshan Zhao at University of Glasgow, and Dr. Yujie Hu at University of South Florida.

The final and deepest appreciation goes to my parents, for your everlasting support, encouragement, and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Background: Surveys across the U.S. reveal that commuters driving personal vehicles spend a significant amount of time in traffic, while public transit, as an efficient commuting mode, has been largely underutilized.

Purpose: What causes a low public transit ridership? How could public transit ridership be explained by demographic, socio-economic and spatial characteristics of neighborhood? This study answers these questions by deciphering the relationships between public transit ridership and various factors in a medium-size city in southern U.S. – Baton Rouge, Louisiana.

Methods: Non-spatial and spatial data in a larger areal unit (e.g., block group) are used to infer demographic, socio-economic and spatial variables in a smaller areal unit (e.g., census block) to gain a sharper spatial resolution in the analysis of public transit ridership in geographic information systems (GIS). First, the ecological inference method is used to disaggregate demographic and socio-economic data from the block group level to the census block level. Secondly, Monte Carlo simulation and transit schedule data are used to improve the estimation of travel time by private vehicle and public transit, respectively, based on which commuting time ratio of these two is calibrated at the census block level. Regression analyses including ordinary least square (OLS) regression, geographically-weighted regression (GWR) and semi-parametric GWR (SGWR) are used to explain the variability of public transit ridership using demographic, socio-economic, and spatial variables at the census block level.

Results: A stepwise regression process selects six variables from 25 original variables representing different aspects of demographic, socio-economic, and spatial characteristics at the census block level. The final model includes both global and localized effects on public transit ridership. Recent immigrants and carless population are positively related to public transit

ridership. White population concentration is negatively related to public transit ridership. These relationships are found to be consistent across the study area. The relationships between public transit ridership and income, commuting time ratio, and accessibility to employment via public transit vary across the study area, and some of these variables even show opposite effects in specific pockets in contrast to their area-wide average effects.

CHAPTER 1. INTRODUCTION

Daily commute from residence to workplace is an important part of people's everyday life, and it accounts for a significant portion of intra-urban transportation. On average, the amount of time that American workers spend on commuting is longer than the paid hours they take for vacations and federal holidays. According to the 2011-2015 5-Year American Community Survey (ACS) data, the average one-way travel time to work was 25.4 minutes – that is equivalent to 27 8-hour workdays annually. In Louisiana, workers spent 24.9 minutes per day on commute, only slightly less than the national average. The 2015 Urban Mobility Scoreboard by Texas A&M Transportation Institute reports that 19.0% of commuting time (i.e., 42.0 hours a year) was wasted on road congestion (Schrank et al. 2015).

Burgeoning private vehicle ownership and use largely explain the escalating road congestions (Stopher 2004). High percentage of daily commute by private vehicle not only aggravates road congestions, but also damages environment and human health. For example, less walking or biking increases the rate of obesity and hypertension, and more driving rises the risks of death and injury in car accidents (Hoehner et al. 2012; Litman 2005). Emissions from vehicles, on the other hand, contribute to greenhouse effect and air pollution, the latter of which is detrimental to human health (Simonson et al. 1968; Tallis 2014).

Among other approaches, promoting the use of public transit, whose market share in daily commute has been squeezed to minimum for years, is an important pathway to the reduction of private vehicle commute. The 2009 National Household Travel Survey (NHTS) by the Federal Highway Administration reports that the ridership of public transit for home-to-work commute purpose was held below 4.0% in a twenty-year span from 1990 to 2009, while the percentage of trips by private vehicle was steadily maintained over 90% during the same period

(Santos et al. 2011). In Louisiana, the 2011-2015 ACS shows that a dismal 1.4% of workers aged 16 and over commuted by public transit, whereas 92.2% drove private vehicles to work.

Public transit has been offered as an option to help to reshape the quality and form of urban growth (Bernick and Cervero 1997; Calthorpe 1993). Extensive debate over the past two decades has established that public transit needs to be implemented alongside with supportive policies to encourage transit-oriented development (TOD). TOD is broadly conceptualized as the development of high density, pedestrian-friendly, and mixed-use neighborhoods (Atkinson-Palombo and Kuby 2011), with multiple objectives such as, for transit operators, maximizing revenue for transit agency through lucrative ground leases, maximizing public transit ridership, or revitalizing station areas (Babalik-Sutcliffe 2002; Belzer et al. 2004; Willson and Menotti 2007), and for general public, reducing traffic congestion. Promoting public transit and TOD could form a virtuous circle that benefits not only public transit agency, but also urban development, and ultimately, urban community.

What factors explain transit ridership? Demographics, private vehicle ownership, land use, parking availability and fare, and transit quality, frequency, fares, etc. all play a role. However, the relative importance of these factors, as well as the interaction between them, is not well understood (Taylor and Fink 2003). This study attempts to explore some possible answers to this question. This study chooses East Baton Rouge Parish, Louisiana to analyze the relationship between public transit ridership and various demographic, socio-economic, and spatial characteristics at the census block level. The study area is a medium-size metropolitan area in the U.S. with considerable traffic congestion during rush hours, although the currently underused public transit system could have relieved the traffic. This study attempts to understand how public transit ridership could be explained by neighborhood characteristics, and the findings

2

can have important implications for policy-making and planning. Improving public transit ridership helps alleviate traffic congestion, improve community health and environment, and achieve better living.

This dissertation has eight chapters. Following this introduction, Chapter 2 provides a critical review of the researches on public transit ridership. It mainly discusses the theories and methods in four categories:

1. The choice of explanatory factors on public transit ridership, including both non-spatial (demographic and socio-economic) and spatial factors;

2. The methods of disaggregating non-spatial data from larger areal unit to smaller areal unit, or the solutions to the "ecological inference problem";

3. The methods of simulating the commuting trips at smaller areal unit based on data in larger areal unit;

4. The measures of travel time, including driving private vehicle and riding public transit.

Chapter 3 describes the study area and data used. It offers an overview of the geography, transportation infrastructure, including road network and public transit system, and demographic and socio-economic landscape of the study area. It also documents the data sources and the corresponding years.

Chapter 4 first describes the demographic, socio-economic, and spatial variables used in this study. It then proposes a method of data disaggregation to interpolate non-spatial data that are only available in larger areal units (e.g., block group) to a smaller areal unit (i.e., census block).

Chapter 5 explains the concept and method to simulate individual commuting trips based on the census tract-to-census tract trip flow data. This is enabled by the Monte Carlo simulation method that accounts for commuting trip flow and land use patterns in the study area.

Chapter 6 discusses how to use transit schedule data to calculate transit-based commuting time via transit network.

Chapter 7 uses multivariate regression analyses to explain public transit ridership using the aforementioned demographic, socio-economic and spatial factors.

Chapter 8 summarizes the major findings of the study, highlights the methodological contributions, and discusses the possible implication in public policy and urban planning. It also outlines the directions for future extensions to this study.

CHAPTER 2. LITERATURE REVIEW

This chapter reviews the literatures pertaining four major components of this study: factors influencing public transit ridership, non-spatial (i.e., demographic and socio-economic) data disaggregation, trip simulation, and measures of travel time.

2.1. External factors influencing public transit ridership

Studies on factors influencing transit ridership can be generally grouped into two main categories: descriptive analysis and causal analysis (Taylor and Fink 2003). *Descriptive analysis* uses qualitative survey and interview data to identify factors that may affect transit ridership, and focuses on traveler's attitudes and perceptions. Such information is highly subjective and dependent on respondents' perceptions and assumptions. It is hard to quantify (TRB 1998), and the data collection processes are often not outlined in detail (Bianco, Dueker, and Strathman 1998). Such methodological and interpretative defects could easily yield biases and unreliable results. Descriptive analysis is usually employed by transit operators for purposes such as service adjustments, marketing, planning, and fare policy. *Causal analysis*, on the other hand, employs a wider variety of data to conduct empirical studies with more robust results. It is also more feasible to generalize findings from causal analysis to other public transit systems and study areas (Spillar 1989; Hartgen and Horner 1997; Taylor et al. 2009). This study falls under the broad category of causal analysis, on which the remaining section focuses.

Transportation researchers often model transit demand or evaluate existing transit systems. Most traditional causal analysis studies focus on the metropolitan area level (Kain and Liu 1999; Taylor et al. 2009), and compare different transit systems by related geographical, environmental, and demographic and socio-economic characteristics in different metropolitan

areas. Only a few exceptions examine intra-urban variations by analyzing factors such as population density and land uses. However, these studies usually choose transit stops or stations as the primary object and explain ridership by the geographic and socio-economic characteristics of their surrounding service areas (Cervero and Seskin 1995; Kuby, Barranda, and Upchurch 2004; Gutierrez, Cardozo, and Garcia-Palomares 2011). These studies certainly provide valuable insights on evaluating the effectiveness of a public transit system, but lack depth on analyzing ridership on commuter's side (e.g., the propensity of using public transit service by different population groups).

Factors believed to affect public transit ridership are generally categorized into internal factors and external factors (Taylor and Fink 2013). *Internal factors* are those controlled by transit agencies and operators, including service quality, fares and service frequency (Chen, Varley, and Chen 2011). Studies on internal factors help operators to manage public transit supply and improve ridership. *External factors* refer to those that are exogenous to transit system itself, such as distributions of population and employment, and demographic and socio-economic factors (Gomez-Ibanez 1996; Taylor et al. 2009). Studies on external factors are useful for policy makers and transportation planners to identify transit needs and encourage transit use. This study only examines external factors. As public transit's market has been continuously squeezed by private vehicle, its importance to commuters with limited access to private vehicles (the poor, the disabled, children and elderlies) and travelers to large employment centers with limited and expensive parking has become more significant. To capture these two groups of public transit users, demographic, socio-economic and spatial factors are the main types of external factor discussed in this study. The following part discusses the specific external factors used in previous literatures.

Due to its wide availability, the Census data is a popular choice for researchers to measure external factors. A host of demographic and socio-economic factors are identified to explain public transit ridership at aggregate level. In a study on western American cities (Spillar 1989), areas with higher population density and lower income level tend to have higher public transit ridership. Additional factors such as percentages of college students and recent immigrants are found to be positively related to public transit ridership among most U.S. urbanized areas (Taylor et al. 2009). Private vehicle ownership is also a crucial factor that reduces public transit ridership (Sharaby and Shiftan 2012), and improvements in public transit may not suppress the impact of increasing use of private vehicle (Kitamura 1989).

Besides demographic and socio-economic variables derived from Census data (i.e., characteristics of the "origin"), variables related to employment (i.e., the "destination" of a commuting trip) are also relevant. A study of public transit commute to CBD in U.S. cities finds a positive relationship between employment growth and public transit ridership (Hendrickson 1986). This observation is echoed by a study on Chicago's public transit system from 1976 to 1995 (Chung 1997) and another study of 54 most populous Metropolitan Statistical Areas in U.S. from 2000 to 2005 (Armbruster 2010). However, others argue that the positive effect of employment growth on public transit ridership is offset by the impact of rising income (on a per capita basis) and suburbanization (Liu 1993; Gomez-Ibanez 1996). Such inconsistency may be related to the multicollinearity problem among the explanatory variables in the multivariate regression models in those studies.

In addition to the aforementioned non-spatial variables, commuter's travel behavior could also be influenced by various spatial factors. In fact, urban and transportation planners have more direct control over land use and deployment of public transit system than demographic and

socio-economic factors of neighborhoods. Residential and employment densities have long been thought to have positive relationships with transit use (Pushkarev and Zupan 1977; Hendrickson 1986; Spillar 1989; Chung 1997). Decentralized residential and employment locations are difficult to serve with traditional fixed-route public transit and have low patronage, while densely populated neighborhoods and compactly developed business districts tend to attract more public transit use (Crane 2000; Ewing and Cervero 2010). As discussed previously, most existing studies on resident and employment densities either examine at large scales, such as a set of metropolitan areas, or focus on the service area of each transit stop. Little emphasis is placed on transit accessibility of commuter's neighborhood (e.g., census block). In addition to the straightforward measures such as the aforementioned residential and employment densities, the difference in travel efficiency by competing commute modes is pivotal for a commuter's mode choice, and has been largely missed in most studies. Furthermore, walking distance in transit trip (e.g., between home or workplace and transit stops, or between transit stops to transfer) is an important factor affecting whether to choose public transit or not. Most studies consider a walking distance threshold to construct catchment area for public transit (Neilson and Fowler 1972; Alshalalfah and Shalaby 2007; Crowley, Shalaby, and Zarei 2009; Guerra, Cervero, and Tischler 2012), instead of a gradual influence of increased distance on reduced transit ridership propensity.

One notable study in the Netherlands considers both non-spatial and spatial factors to explain travel mode choice in resident worker communities (Limtanakool, Dijst, and Schwanen 2006). However, that is on medium- and long-distance inter-urban trips. A public transit system in the U.S. mainly serves one or a few adjacent metropolitan areas. On the methodological front, one common problem in the multivariate model used in the existing literature is multicollinearity

among those external factors. For example, places of higher residential density also tend to share certain demographic and socio-economic characteristics, such as lower income and lower car ownership. Untangling the mutual effects of non-spatial and spatial variables on one another and on public transit ridership is a challenge (Gomez-Ibanez 1996; Crane 2000; Taylor and Fink 2003) that has motivated this study.

2.2. Data disaggregation

Some external factors, especially non-spatial factors, can only be accessed in larger areal units. Such data need to be disaggregated to smaller areal unit in order to gain better spatial resolution. An influential study on data disaggregation is by King (1997). It uses a regression-randomization process to estimate the parameter values for subunits using a "statistical approach" within the value ranges predefined by a "method of bounds". His model assumes that the correlation between two variables is constant in both aggregate area level and disaggregate area level, and also assumes a variance function that fits an important feature of aggregate data and is usually available in Census data. This data disaggregation method avoids commonly-found estimation biases in many of his precedents (e.g., Flanigan and Zingale 1985; Dykstra 1986), and has been widely adopted by social scientists. More technical details will be discussed in Chapter 3, which builds upon King's method with some refinements specially designed for this study for better results.

2.3. Trip simulation

Similar to non-spatial data, spatial data of commuting trips between larger areal units also need to be disaggregated to improve accuracy of estimation. Trip simulation is largely dominated

by the popular four-step model of travel demand model: trip generation, trip distribution, mode choice and trip assignment (Wang 2015). Hu, Wang, and Wilmot (2017) uses Monte Carlo simulation method to simulate individual intra-urban commuting trips that are consistent with aggregate trip flow data derived from the Census data. Their case study, which is also in Baton Rouge, Louisiana, shows that this method is very promising. Their study utilizes the Census and land cover data to improve simulation of trip origins and destinations, and uses origin-destination (OD) flow data between census tracts from Census Transportation Planning Package (CTPP) to guide the simulation of individual commuters. This model is further improved by adjusting the parameters of the algorithm for the maximal fitness between simulated pattern and observed traffic count data provided by a local government agency. This study uses the more accurate land use data to further improve the simulation of commuting trips.

2.4. Measures of travel time

As previously mentioned, most studies on public transit ridership fall short in accurately measuring travel time difference between different travel modes. A study in Seattle suggests that travel time is the primary factor in influencing a resident's choice of travel mode (Frank et al. 2008). Longer driving time by private vehicle tend to sway workers to public transits, and similarly, longer transit-riding time is associated with reduced public transit use. Most studies rely on survey data such as the aforementioned CTPP or NHTS to gather information on travel time difference between different transportation modes. Once again, these survey data often lack a reasonable geographic resolution to infer meaningful travel behavior. Survey-based travel data also suffer from biases from multiple sources (Spurr, Chapleau, and Piché 2014). Recent advancement in applying network theory in geographic information systems (GIS) enables one to

estimate driving time with reasonable accuracy. The same cannot be assumed for transit travel time estimation since a fixed route transit system follows its posted schedule instead of the shortest or fastest route (Kuai and Zhao 2017). A recent study in Seattle (Tallis 2014) uses a schedule-aware tool for transit-based network analysis to compute in-vehicle time as well as waiting, walking, and transfer time. This study implements the state-of-art transit travel time simulation tool to measure travel time via public transit network.

CHAPTER 3. DATA AND VARIABLE DEFINITIONS

3.1. Data for resident workers and employment

This study selects East Baton Rouge Parish, Louisiana, as the study area. A "parish" is a county-equivalent administrative unit in the state of Louisiana. According to the data provided by Capital Region Planning Commission (CRPC) (2013) – the metropolitan planning organization of the Baton Rouge metropolitan area, East Baton Rouge Parish holds a total population of 443,598, 47.6% of which (211,184) are workers aged above 16 not working at home. This study is primarily based on the census block level. The study area has 9,270 census blocks with non-zero resident workers. Among the four incorporated cities, the city of Baton Rouge is the parish seat, and is the largest both in terms of area and number of resident workers. It is also the second-largest city in the state of Louisiana. The cities of Baker and Zachary also have minor concentrations of resident workers. The city of Central is mostly rural with population sparsely scattered. Figure 1 is based on resident worker data at the census block level in the study area (hereafter, simply referred to as "Baton Rouge").
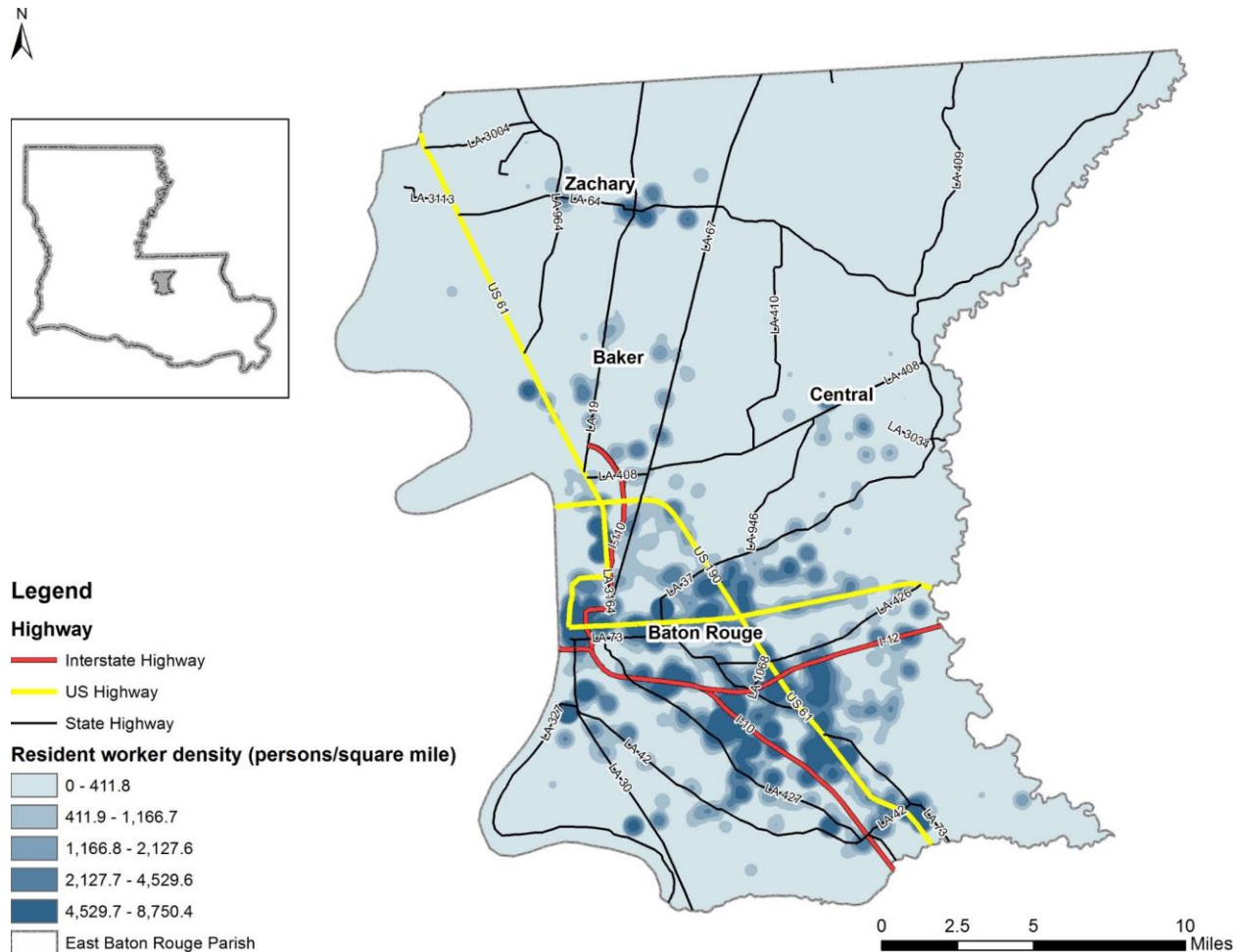
Figure 1. Resident worker distribution in Baton Rouge 2013

Also based on the CRPC data at the census block level, Figure 2 shows the density

distribution of employment (the white-shaded area represents urban area). Downtown Baton

Rouge, in the west of the study area, has the highest concentration of employment, with the State

and the City-Parish governments being the two largest employers. Employment density generally

declines with increasing distance from downtown. There are two other more noticeable

employment centers: Louisiana State University (LSU), located at the south of downtown, and

the regional medical center (i.e., Our Lady of the Lake Regional Medical Center (OLOL) and

Baton Rouge General Hospital), located southeast of downtown. LSU and OLOL are the largest

and second largest employers of the entire study area, respectively. This makes the study area

unique from other metropolitan areas in U.S. that it is multi-centered. Similar to the pattern of

resident workers, the three satellite cities (Baker, Central and Zachary) have far less

employment.
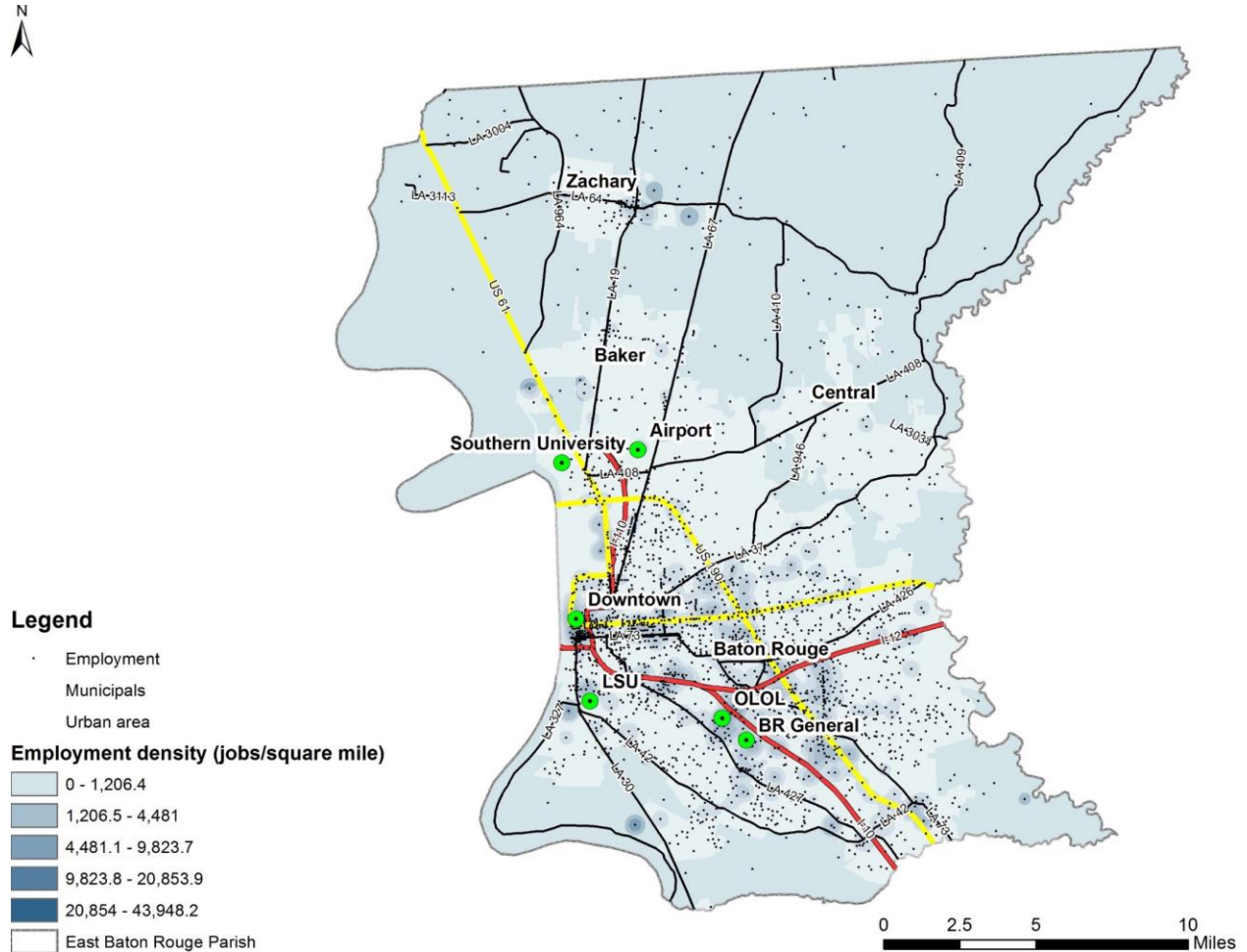
Figure 2. Employment distribution in Baton Rouge 2013

3.2. Data for transportation and land uses

The public transportation system in Baton Rouge includes road network and public transit

system. The road network data is based on the 2012 ESRI StreetMap data that includes all levels

of roads and streets with essential information such as speed limits and directions. Given the year

(2013) for the CRPC data for resident workers and employment, the 2012 ESRI StreetMap data

is the closest match in date for road network data available to this study. This study uses the General Transit Feed Specification (GTFS) data to access information on the public transit system. GTFS is an open-source data format developed by Google Developers to describe fixed-route transit services (Antrim and Barbeau 2013). The 2015 GTFS data, provided by Baton Rouge's public transit authority – the Capital Area Transit System (CATS), include detailed information of transit routes, stops, and schedules. The transit system mainly covers the cities of Baton Rouge and Baker, and has 30 regular-servicing, fixed bus routes with a total length of approximately 280 miles and 1676 bus stops. In 2014, it served more than 90,000 passengers. On weekdays, it operates from as early as 5:00 a.m. to as late as 11:00 p.m.

The 2012 land use data is acquired from the local government (City of Baton Rouge and Parish of East Baton Rouge), shown in Figure 3.
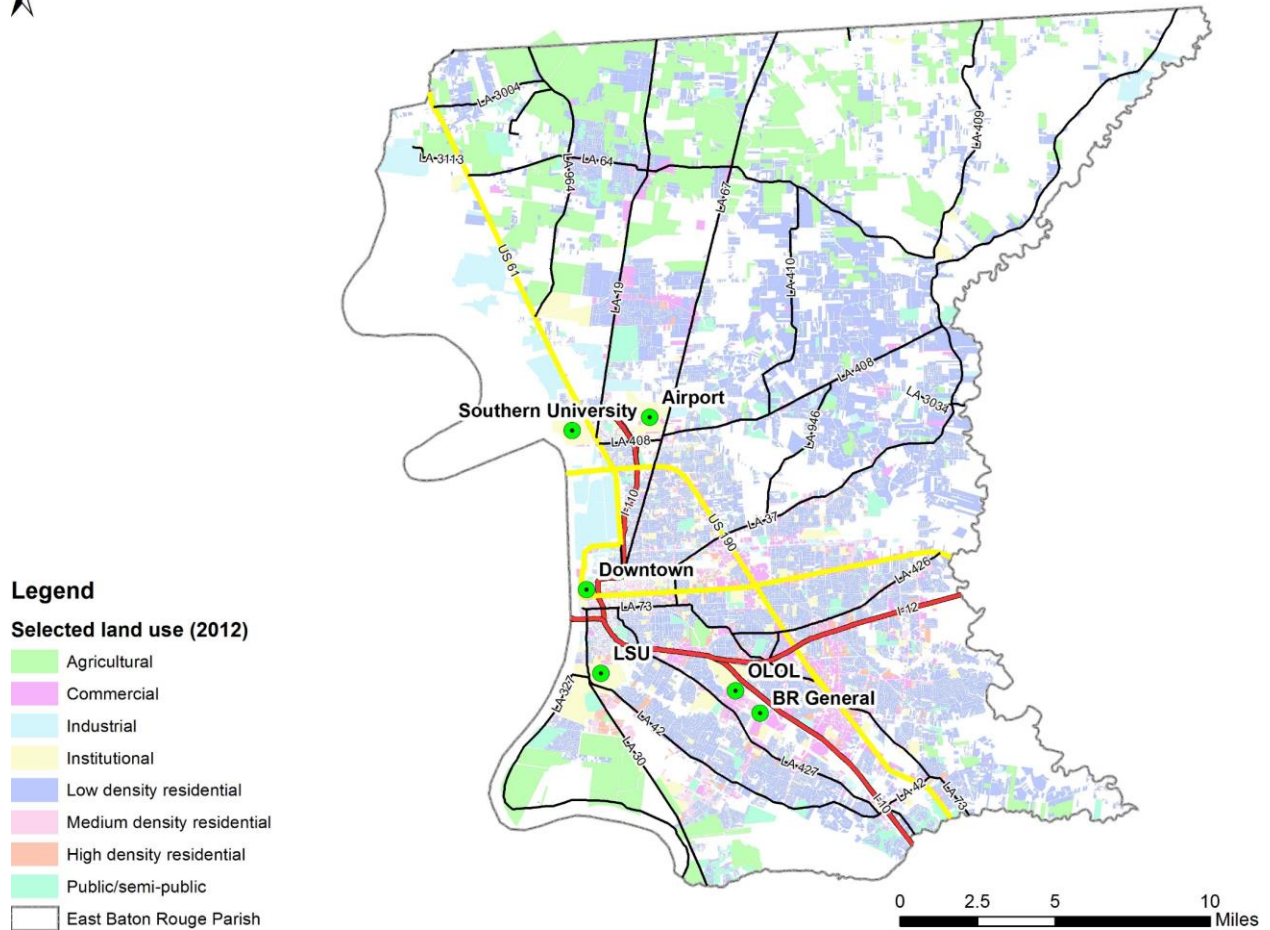
Figure 3. Selected land uses in Baton Rouge 2012

The majority of land uses is residential, among which low-density residential land use

accounts for the highest percentage and is distributed throughout the study area. Medium- and

high-density residential land uses mainly are located in and around the downtown area.

Commercial land use, where much employment is located, mainly appears along major

transportation corridors. Industrial land use is found along the east bank of Mississippi River.

Notable occupants of institutional land use include four major employers in the study area: LSU

in the southwest, the aforementioned regional medical center, Southern University in the north

by Mississippi River, and Baton Rouge Metropolitan Airport to the east of Southern University.

Undeveloped and unpopulated lands are shown as blank on the map.

The data sets used in this study are documented in Table 1.

Table 1. Summary of data sets

Data set	Source	Spatial resolution	Year(s)
Demographics and housing	CRPC, ACS	Census block	2013
Socio-economic status	ACS	Block group	2011 – 2015
Employment count	CRPC	Census block	2013
Trip flow	CTPP	Census tract	2006 – 2010
Road network	ESRI StreetMap	-	2012
Public transit system	CATS	-	2015
Land use	Local government	Parcel	2012
Traffic count	LaDOTD	-	2014

3.3. Variable selection and definition

This dissertation attempts to decipher the relationship between public transit ridership and demographic, socio-economic, and spatial factors that represent different characteristics of neighborhood. The dependent variable is public transit ridership at the census block level. As discussed in Chapter 2, public transit ridership is associated with both internal and external factors. Because this study is primary interested in how a neighborhood's patronage of public transit is affected by its own characteristics, factors like gasoline price and parking cost are thus not considered. Based on the existing literatures and considering the information in the data available, 16 non-spatial and 9 spatial variables at the census block level are selected as explanatory variables. These variables do not necessarily have direct causal relationships with public transit ridership. Also, different variables reflecting similar aspect of neighborhood characteristic are included for testing purpose. Nevertheless, these variables characterize neighborhood's demographic and socio-economic structure and spatial location that indicate its residents' propensity for commuting by public transit. Findings on how public transit ridership is

related to these variables can help transportation planners and policy makers to predict transit needs and make adjustment and plans accordingly.

The non-spatial variables include a set of demographic, socio-economic, and housing condition variables. For example, in terms of demographic variables, this study includes median age of neighborhood which may influence the acceptable walking distance from home to bus stop (Alshalalfah and Shalaby 2007). Another common set of demographic variables include race and ethnicity, gender, and immigrant status (Blumenberg and Shiki 2007). To quantify immigrant status, this study uses "households speaking limited English" as a proxy for recent immigrants. Neighborhood's financial status is a major category of socio-economic variables, and is represented by poverty level and income level, as well as unemployment rate. Education attainment, measured by population with high school diploma, is another factor that could affect a commuter's perspective towards how to commute. Availability of private vehicles is one of the most intuitive and significant variables that affects a commuter's behavior. Percentage of renter-occupied housing unit that quantifies the concentration of home renters is a joint result of demographic and socio-economic conditions (Kuby, Barranda, and Upchurch 2004). Other housing condition variables are newly added in this study, including housing unit with multiple occupants per room and housing unit without complete kitchen and plumbing facilities.

For spatial factors, this study first considers the relative convenience of commuting by public transit versus private vehicle for census block, here measured by the average ratio of travel time by public transit to travel time by private vehicle for every trip originated from census block:

$$v = \frac{\overline{T^t}}{T^d} \tag{3.1}$$

where t_T and t_D stand for time for public transit and private vehicle, respectively, and their specific formulations will be explained in Chapter 6. The second variable denotes the distance from census block to its nearest bus stop to measure census block's proximity to public transit system. In addition, total number of bus stop(s) within a 1-mile radius of census block is also included as a measure of availability of public transit system. In both cases and in the following sections, the location of census block is represented by its centroid. The next variable is accessibility to jobs via public transit network. It is measured as number of jobs within a 1-hour transit catchment area:

$$v = \sum_{i \in (t_i \leq 60)} E_i \tag{3.2}$$

where t_i is the travel time (in minutes) from census block to employment location i via the public transit network, and E_i stands for the number of employments in i. As discussed in Chapter 2, population density is a popular factor considered in previous literature. Resident density, worker density and housing unit density of census block are 3 similar but slightly different measures of population density. Furthermore, a binary land use variable, coded as:

$$v = \begin{cases} 0 & \text{Low-density residential} \\ 1 & \text{Medium- or high-density residential} \end{cases} \tag{3.3}$$

is used, as this term is often adopted in planning and policy making process. These four variables together help to capture a more comprehensive picture of transit needs in terms of residential land use. Lastly, the presence of sidewalk may impact a neighborhood's walkability to bus stop(s), and is measured by the total length of sidewalks within a 1-mile radius of census block. The non-spatial and spatial variables discussed above are mapped in Figure A1-A7 in Appendix.

19

3.4. Dependent variable and effective sample

The common walking distance (network distance) to a bus station is 0.5 miles (Kuby, Barranda, and Upchurch 2004), thus 1 mile is considered as the maximum range for pedestrians to account for any possible edge effect. This study includes only 4,045 worker-populated census blocks that has its centroid within the 1-mile range from its nearest bus station. That is to say, the 4,045 census blocks assemble the *effective sample* for this study.

Public transit ridership for the effective sample census blocks in 2015 is shown in Figure 4. Overall, public transit ridership in the study area is low, and the network has limited coverage in the cities of Baton Rouge and Baker in the middle part of the study area. Even as the network extends to the southern part of the parish, the ridership there is near zero. The system is seriously underutilized.
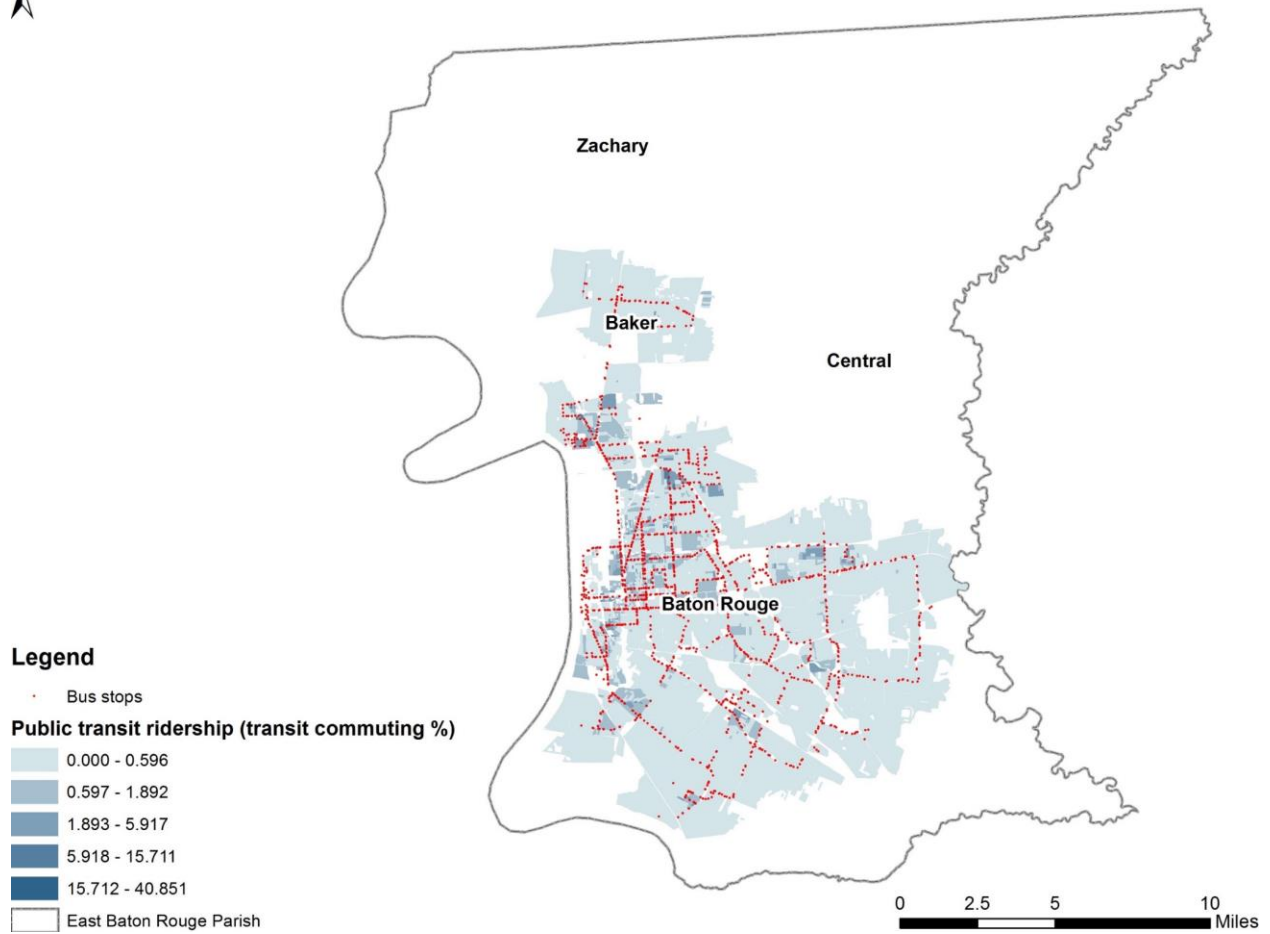
Figure 4. The CATS ridership 2015

CHAPTER 4. DISAGGREGATION OF SOCIO-DEMOGRAPHIC VARIABLES

As outlined previously in Table 1, most socio-economic variables from Census in this study such as poverty rate and median household income are at the block group level, and a small number of variables such as basic demographics and housing conditions from the local agencies are at the census block level. It is often desirable for studies to be at a smaller geographic areal unit for better resolution. It is even more so for spatial variables such as the calculation of travel time in order to achieve reasonable precision of location. Transformation of data from a larger areal unit (e.g., block group) to a smaller areal unit (e.g., census block) is termed "data disaggregation". The task begins with the disaggregation of demographic and socio-economic variables, a process also referred to as "ecological inference" (King 1997).

4.1. Solution to the ecological inference problem

To explain the solution to ecological reference problem, this dissertation borrows King's (1997) example – estimating the voting patterns of different racial groups in geographic unit called "voting precinct", as depicted in Table 2.

Table 2. Non-White and White voting turnouts of a precinct

Race	Turned out	Did not turn out	
Non-White	β_i^b	$1 - \beta_i^b$	X_i
White	β_i^w	$1 - \beta_i^w$	$1 - X_i$
	T_i	$1 - T_i$	

Table 2 depicts the problem to estimate voting turnouts of non-White and White that occur in one of p precincts in a voting district, denoted by i ($i \in [1, p]$). The notations are explained as follows:

T_i: proportion of voters turned out (known);

X_i: proportion of non-White voters (known);

β_i^b: proportion of non-White voters turned out (unknown);

β_i^w: proportion of White voters turned out (unknown).

Researchers can only observe the "marginals" – the final column representing the total number of non-White/White voters, and the final row representing the total number of voters turned out/did not turn out. β_i^b and β_i^w are the variables of interest. The goal is to infer the cell entries β_i^b and β_i^w from the aggregate marginals. The basic model that describes the relationship between these variables is:

$$T_i = \beta_i^b X_i + \beta_i^w (1 - X_i) \tag{4.1}$$

This accounting identity is a statement of fact that holds for each one of the p precincts in the data set, forming a system with p equations (one equation for each precinct) and $2p$ unknowns (one set of β_i^b and β_i^w for each precinct). It is assumed that β_i^b and β_i^w are modeled as if they are generated by a truncated normal (TN) distribution conditional on X_i:

$$P(\beta_i^b, \beta_i^w) = TN(\beta_i^b, \beta_i^w | \mathfrak{B}, \Sigma) \tag{4.2}$$

where the mean vector $\mathfrak{B}$ and the variance matrix Σ of (β_i^b, β_i^w) are:

$$\mathfrak{B} = \begin{pmatrix} \mathfrak{B}^b \\ \mathfrak{B}^w \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_b^2 & \sigma_{bw} \\ \sigma_{bw} & \sigma_w^2 \end{pmatrix} \tag{4.3}$$

The means, $\mathfrak{B}^b$ and $\mathfrak{B}^w$, are based on the district-level averages of non-White and White turnout rates of voters: $\mathfrak{B}^b = E(\beta_i^b)$ and $\mathfrak{B}^w = E(\beta_i^w)$. This assumption can be verified to a sufficient degree in the aggregate data set. Limited to the scope of this study, this dissertation does not discuss the observable implications that provides diagnostic test to verify these assumptions. Besides, King also adds other two assumptions to this identity that are either proven to be unnecessary or "does not have major consequences for most aspects of this model" (King 1997).

The solution to this model is a combination of what King refers to as "statistical approach" and "method of bounds". Define β_i^b and β_i^w as:

$$\begin{aligned} \beta_i^b &= \mathcal{B}^b + \varepsilon_i^b \\ \beta_i^w &= \mathcal{B}^w + \varepsilon_i^w \end{aligned} \tag{4.4}$$

where $\mathcal{B}^b$ and $\mathcal{B}^w$ are their means, ε_i^b and ε_i^w are error terms that $E(\varepsilon_i^b) = 0$ and $E(\varepsilon_i^w) = 0$.

Now, Equation 4.1 can be written as:

$$T_i = \mathcal{B}^b X_i + \mathcal{B}^w (1 - X_i) + \varepsilon_i \tag{4.5}$$

where $\varepsilon_i = \varepsilon_i^b X_i + \varepsilon_i^w (1 - X_i)$.

To determine the bounds of β_i^b and β_i^w, first write Equation 4.5 as:

$$\varepsilon_i = T_i - \mathcal{B}^b X_i + \mathcal{B}^w (1 - X_i) \tag{4.6}$$

As $E(\varepsilon_i | X_i) = 0$, the bounds of β_i^b and β_i^w can be derived by assuming homogeneous non-White or White precincts. With X_i at its extremes (0 and 1), β_i^b and β_i^w fall within the permissible bounds:

$$\begin{aligned} \beta_i^b &\in \left[\max\left(0, \frac{T_i - (1 - X_i)}{X_i} \right), \min\left(\frac{T_i}{X_i}, 1 \right) \right] \\ \beta_i^w &\in \left[\max\left(0, \frac{T_i - X_i}{1 - X_i} \right), \min\left(\frac{T_i}{1 - X_i}, 1 \right) \right] \end{aligned} \tag{4.7}$$

With the parameter bounds determined, the key is to estimate the variance matrix Σ. Then through a randomized variation on ε_i^b and ε_i^w, β_i^b and β_i^w are estimated as:

$$\begin{aligned} \beta_i^b &= \mathcal{B}^b + \varepsilon_i^b \\ \beta_i^w &= \mathcal{B}^w + \varepsilon_i^w \end{aligned} \tag{4.8}$$

where $\varepsilon_i^b \sim N(0 | \sigma_i^b)$ and $\varepsilon_i^w \sim N(0 | \sigma_i^w)$.

To sum up, King's solution to ecological inference problem uses Bayesian constructs to derive posterior distribution of unknown parameters based on known means of $\hat{\beta}$ and estimates

of variances and covariances (Anselin and Cho 2002). Thus, heterogeneity of parameters is conceptualized as random variations around a certain common mean value.

4.2. Data disaggregation of socio-economic variables

This study introduces an altered version of King's solution to ecological inference problem. The task is to disaggregate the socio-economic factors that are only available at block group level to census block level.

Here, the estimation of variable "number of renter-occupied housing unit" is used as example. This method only disaggregates one variable at a time (e.g., number of renter-occupied housing unit), and does not consider a second dimension (e.g., race, as in King's example) to estimate the variable of interest. Table 3 illustrates the problem, showing numbers of renter-occupied housing unit, owner-occupied housing unit, and total occupied housing unit across n census blocks within one block group.

Table 3. Housing unit counts by tenure at census block level

Census block	Renter-occupied	Owner-occupied	Occupied unit total
1	x_1^r	$x_1 - x_1^r$	x_1
2	x_2^r	$x_2 - x_2^r$	x_2
…	…	…	…
i	x_i^r	$x_i - x_i^r$	x_i
…	…	…	…
n	x_n^r	$x_n - x_n^r$	x_n
Categorical total	x^r	$x - x^r$	x

The notations for the variables in Table 3 are:

x: total number of occupied housing unit in the block group (known);

x_i: number of occupied housing unit in census block i (known);

x^r: total number of renter-occupied housing unit in the block group (known);

x_i^r: number of renter-occupied housing unit in census block i (unknown); and

$x_i - x_i^r$: number of owner-occupied housing unit in census block i (unknown).

According to Equation 4.5, the count of renter-occupied housing unit is estimated as:

$$x_i^r = (x_i^r)_0 + \varepsilon_i^r \tag{4.9}$$

where $(x_i^r)_0$ is the preliminary estimated value of x_i^r, and ε_i^r is an error term. This is the basic accounting identity to estimate the total number of renter-occupied housing unit x_i in census block i.

This data disaggregation method uses an external variable that is available at both block group and census block levels to predict the preliminary value of an unknown variable at census block level. Here, the external variable used is number of occupied housing unit X. First, at block group level, assume that number of renter-occupied housing unit X^r could be expressed using X: $X^r = kX$. k could be estimated using a bivariate linear regression between X^r and X at block group level. Next, assume that the relationship between X^r and X at block group level also applies to x_i^r and x_i at census block level, so that the preliminary value is:

$$(x_i^r)_0 = kx_i \tag{4.10}$$

With the preliminary value established, a disaggregating coefficient β_i^r is introduced as the ratio of the preliminary number of renter-occupied housing unit in the census block to the total number of renter-occupied housing unit in its corresponding block group:

$$\beta_i^r = \frac{(x_i^r)_0}{x^r} \tag{4.11}$$

The next step is to simulate the error term ε_i^r. ε_i^r can be simulated by randomized generation, as it follows a truncated normal distribution:

$$P(\varepsilon_i^r) = TN(\varepsilon^r|0, \sigma_i^r) \tag{4.12}$$

where σ_i^r is the standard deviation of the probability density distribution function of ε_i^r.

As the original ACS data provides margin of error (ϵ^r) for renter-occupied housing unit count at block group level, σ_i^r is calculated as the block group-level standard deviation σ^r multiplied by the disaggregating coefficient β_i^r:

$$\sigma_i^r = \beta_i^r \sigma^r = \beta_i^r \cdot \frac{\epsilon^r}{1.645} \tag{4.13}$$

The denominator 1.645 comes from the fact that U.S. Census Bureau uses a 90% confidence interval to calculate the margin of error ϵ_i that equals to approximately 1.645 times of its standard error (Berkley 2017).

Next, the permissible bounds of x_i^r could be written as:

$$x_i^r \in \left\{ \max[0, (x_i^r)_0 - \beta_i^r \epsilon^r], \min[(x_i^r)_0 + \beta_i^r \epsilon^r, 1] \right\} \tag{4.14}$$

Using all known variables, Equation 4.9 estimates the number of renter-occupied housing unit in census block as:

$$x_i^r = k x_i + \varepsilon_i^r \tag{4.15}$$

where

$$P(\varepsilon_i^r) = TN\left(\varepsilon_i^r | 0, \beta_i^r \cdot \frac{\epsilon^r}{1.645} \right) \tag{4.16}$$

With the number of renter-occupied housing unit estimated, the variable representing the percentage of renter-occupied housing unit is written as:

$$v = \frac{x_i^r}{x_i} \times 100\% \tag{4.17}$$

In summary, this data disaggregation method combines "statistical approach" and "method of bounds" to solve the ecological inference problem. Statistical approach contains two components: using regression method to establish preliminary value, and then using randomized variation to readjust preliminary value within a value range. It also takes advantage of the margin

of error information contained by the ACS data to integrate the method of bounds. This data disaggregation method has two assumptions:

1. *Linearity*: The variable of interest can be represented by an external variable, available in both larger and smaller areal units. This method uses this assumption to establish preliminary values for the variable to be estimated.

2. *Normality*. The variable in smaller areal unit follows a normal distribution within a variance range defined by the margin of error. This method uses this assumption to simulate values across small areas randomly.

4.3. Disaggregation results

Unknown population and housing unit counts at census block level are estimated by the procedures illustrated in the previous section, and are then used to define ratio variables such as female headed household percentage, percentage of renter-occupied housing unit, carless population percentage, poverty rate, etc. Note that in Equation 4.17, the denominator x_i is a known value for all census blocks, as the ACS data contains total population and housing unit counts at census block level (see Table 1). Non-ratio or non-percentage variables such as median age, median household income, and per capita income are directly estimated by Equations 4.9 – 4.16.

Here, we use a variable "White population percentage" available at both block group and census block levels to validate this data disaggregation method. Figure 5 plots the correlation between simulated and true values. The Pearson correlation coefficient between the two values is 0.837, and thus highly correlated (with 4,045 census blocks).
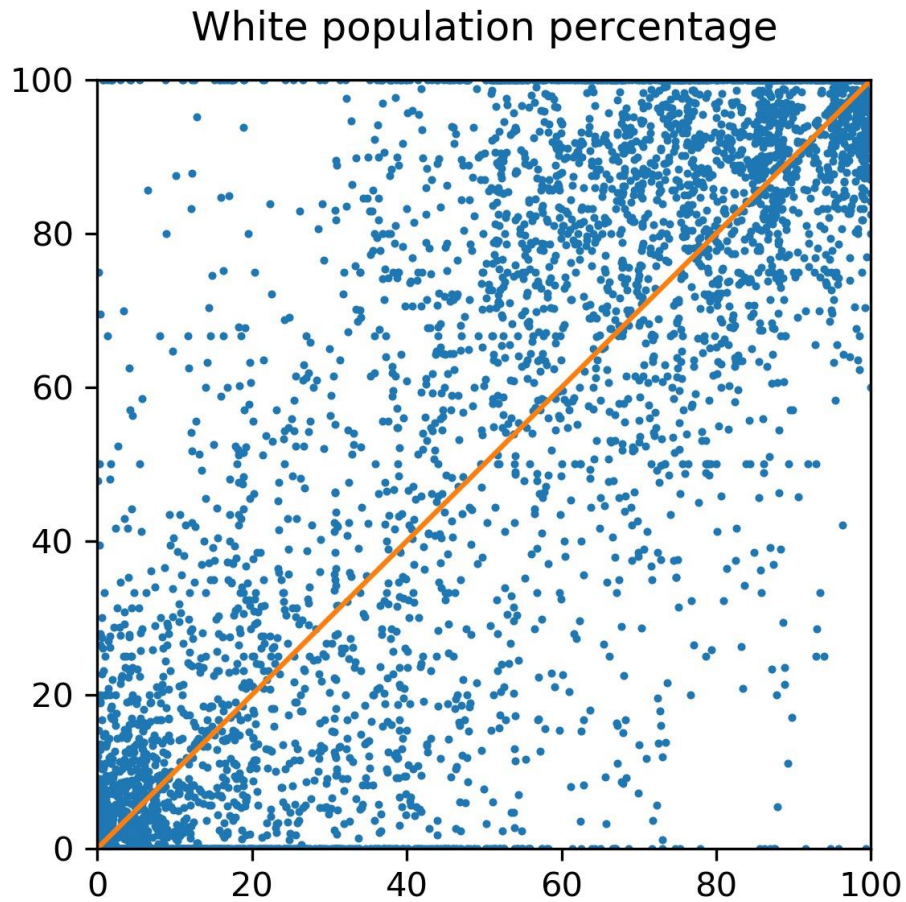
Figure 5. White population percentage: simulated versus true

Table 4 summarizes the basic statistics of the dependent and explanatory variables used in this study. Among the 16 demographic and socio-economic variables, 14 are disaggregated by this data disaggregation method from block group level to census block level, and two are directly extracted from the CRPC data. For convenience to read, the 9 spatial variables are also included here, and their definitions and calibrations will be discussed in the next two chapters (i.e., Chapters 5 and 6).

Table 4. Basic statistics of variables at census block level (number of observations 4,045)

Variable	Mean	Minimum	Maximum
Dependent variable (directly from local agencies)			
Workers commuting by public transit (%)	0.408	0.000	40.851
Explanatory variables (directly from local agencies)			
White population (%)	25.000	0.000	26.369
Vacant housing unit (%)	5.000	0.000	8.453
Explanatory variables (disaggregated from block group)			
Median age (years)	35.264	0.410	117.126
Female headed householder (%)	18.800	0.000	100.000
Population with high school diploma (%)	6.865	1.735	229.451
Household speaking limited English (%)	0.534	0.035	9.893
Poverty rate (%)	1.545	0.004	85.883
Median household income (dollars)	40,171.798	308.349	303,944.519
Per capita income (dollars)	26,086.832	568.408	91,713.952
Household receiving food stamp (%)	1.168	0.000	32.447
Unemployment rate (%)	2.502	0.010	70.272
Renter-occupied housing unit (%)	25.000	0.000	100.000
Housing unit with >1 persons/bedroom (%)	4.775	0.000	23.387
Carless housing unit (%)	0.615	0.000	43.570
Housing unit with incomplete plumbing device (%)	0.161	0.000	7.981
Housing unit with incomplete kitchen (%)	0.201	0.000	11.113
Spatial variables			
Bus-riding-driving time ratio	6.832	4.028	11.984
Distance to nearest bus stop (meters)	287.705	11.497	1,606.619
Number of jobs within 1-hour transit service area	3,043.906	1.000	25,104.000
Number of nearby bus stop	46	1.000	186.000
Population density (per acre)	6.989	0.013	938.331
Worker density (per acre)	2.099	0.011	118.396
Housing unit density (per acre)	2.708	0.000	423.121
Nearby sidewalk length (meters)	37,258.783	173.515	96,249.783
Low-density residential land use	-	-	-

In summary, the data disaggregation method proposed in this study applies King's solution to the ecological inference problem with external controls. First, it establishes a

preliminary value for the variable of interest. It uses an external variable that is available in both larger and smaller areal units to predict the variable of interest that is only available in larger areal unit based on a regression. Then, it establishes a value range for the variable of interest using the margin of error information from the ACS data and assumes that the variable of interest follows a truncated normal distribution within that range. Within value range, the estimate of the variable of interest is randomly generated. The data disaggregation results are then validated with real data and could be considered acceptable. Thus, this data disaggregation method provides acceptable results at the smaller areal unit and could be used in the following analyses.

CHAPTER 5. MONTE CARLO SIMULATION OF COMMUTING TRIPS

Apart from demographic and socio-economic factors, travel time is another critical factor that could affect people's travel behavior. Estimating travel time begins with retrieving commute trips between residence and workplace units. Similar to demographic and socio-economic data, trip count data is only available between larger areal units such as census tract in the study area. If one uses census tract centroid to represent commute origins and destinations, the estimation of travel time may suffer from serious zonal effect, especially in suburban and rural areas where census tracts have large areal size. This study uses Monte Carlo simulation method to simulate individual residence-to-workplace commute trips, and then aggregate them back to estimate travel time between census blocks. By doing so, much zonal effect could be mitigated and more accurate travel time could be calculated.

5.1. Principles of Monte Carlo simulation

Monte Carlo simulation is a numerical analysis technique that generates random numbers to explore the distribution of individuals in a featured system when the distribution pattern is not directly available. Random numbers are generated by following a pre-defined probability distribution function (PDF) that describes the probabilities of occurrence of different outcomes in a system. Some commonly used PDFs include uniform, normal, lognormal and discrete distributions (Hu, Wang, and Wilmot 2017), as illustrated in Figure 6.

Take discrete distribution as an example. Instead of a continuous curve, a group of numbers within a certain range share a same occurring probability, and numbers in different groups may have different chances of occurrence. For example, in the context of this study, the

occurrence possibility of commuting trip may be 40% for a commuting distance range between 0 to 3 miles, 30% for 3 to 5 miles, 20% for 5 to 10 miles and 10% for beyond 10 miles.
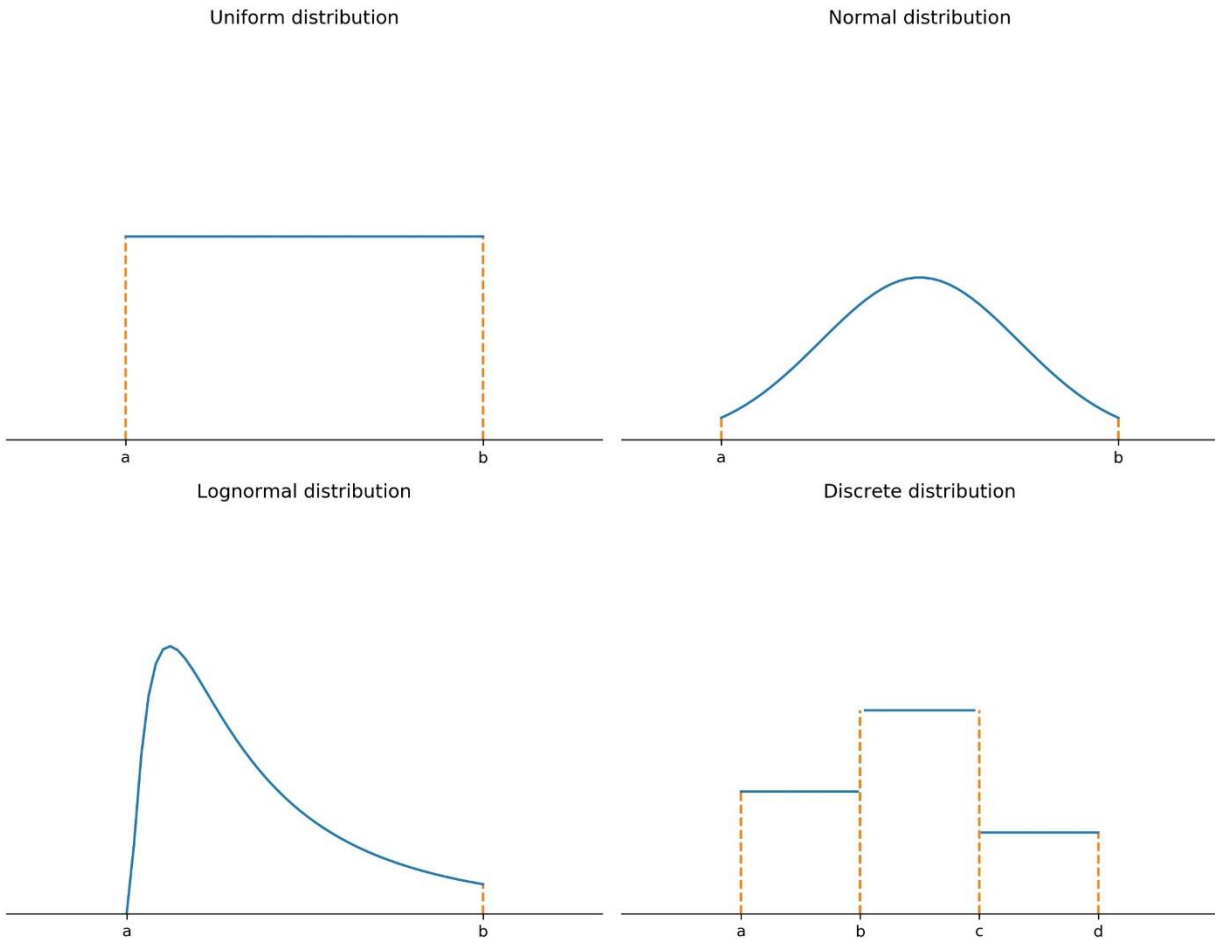
Figure 6. Uniform, normal, lognormal and discrete distributions

In Monte Carlo simulation, a set of random results generated according to a predefined PDF is called an iteration. For instance, an iteration includes a set of residence locations, a set of workplace locations and the commuting trips between them. The simulation process is iterated for a large number of times until the set conforms to certain prior knowledge (e.g., observed traffic count data).

5.2. Monte Carlo simulation of residences (O) and workplaces (D)

Corresponding to trip generation in the four-step travel demand simulation model, the first step is to simulate individual locations of trip ends from zonal resident (origin) and employment (destination) data. An earlier study (Hu and Wang 2015) used resident worker and employment count data at census tract level, and used National Land Cover Database data to help to improve the accuracy of trip end simulation. This study makes further improvement by utilizing higher-quality resident worker and employment counts at census block level and land use data at sub-census block level from the local government agencies to further confine the feasible location sets for simulated trip ends. Specifically, resident worker points are only generated in residential (low-, medium- and high-density) land use areas and employment points are only generated in commercial, industrial and institutional land use areas.

Monte Carlo simulation of trip ends begins by generating random geographic coordinates across the study area, following a PDF of two-dimensional discrete distribution:

$$P(X_{i_j}, Y_{i_j}) = \begin{cases} p_i & \text{Point inside residential/employment land use} \\ 0 & \text{Point outside residential/employment land use} \end{cases} \tag{5.1}$$

where $P(X_{i_j}, Y_{i_j})$ stands for the occurring probability of trip end point i in census block j during an iteration of Monte Carlo simulation, X and Y are geographic coordinates, and p_j represents the occurring probability, which is constant within the corresponding land use area in census block j.

Figure 7 illustrates the residential and employment land uses in the study area and simulated trip ends. Specifically, the numbers of simulated origins and destinations in each census block are proportional to the resident worker and employment counts in each census block reported by CRPC.
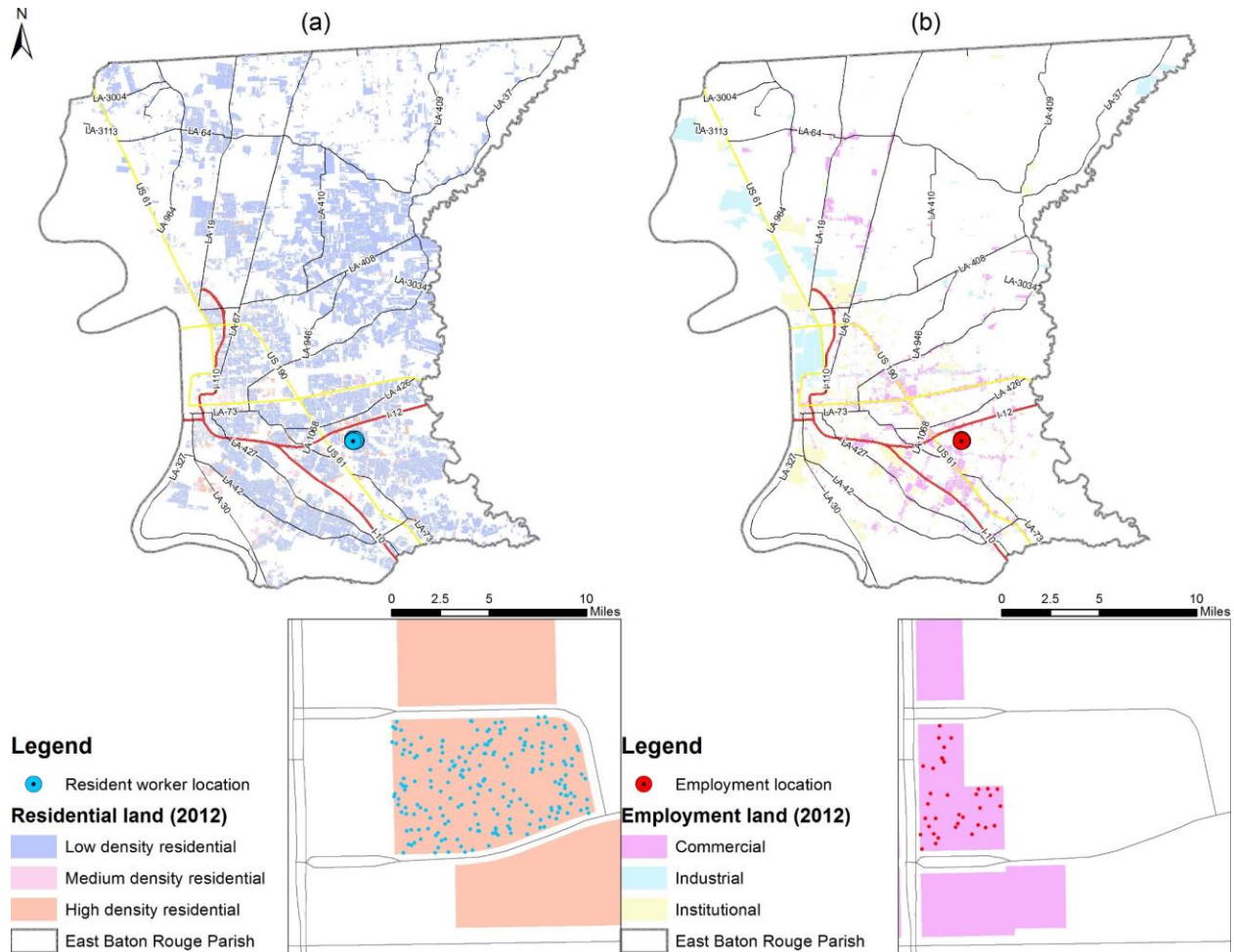
Figure 7. Land uses and simulated trip ends in Baton Rouge: (a) residential, (b) employment

5.3. Monte Carlo simulation of commuting trips

After individual resident worker and employment locations are generated, the next step is to pair resident worker and employment locations to simulate commuting trips. Monte Carlo simulation here generates individual commuting trips that are consistent with observed inter-zonal traffic flow data. This is implemented in four steps:

1. From a residential zone containing m resident worker locations (origins), randomly choose one denoted as O_i, where $i \in [1, m]$.

2. Similarly, from an employment zone containing n employment locations (destinations), randomly choose one denoted as D_j, where $j \in [1, n]$.

3. Pair O_i and D_j as a trip, and cumulate the trip count C between i and j.

4. Iterate the previous three steps until the cumulated trip count C reaches a given inter-zonal trip count C_0.

The process above is repeated for every valid pair of residential and employment zones. A valid pair is one with non-zero commuting trips as reported in CTPP data, and the simulated trip counts are assigned proportionally to the reported commuting volumes on corresponding OD pairs. This study uses Traffic Simulation Modules for Education (TSME) program developed by Hu and Wang (2015) to implement both simulation processes of trips ends and trips. Monte Carlo simulations of trips facilitate the disaggregation of OD commuting trip volumes at the aggregate (census tract) level to individual trips.

5.4. Validating simulation results

The purpose of simulating individual commuting trips in this study is to improve accuracy when estimating travel time. With travel time on individual OD pairs in place, travel time between any pairs of zones (e.g., between residential and employment census blocks in this study) can be easily calculated. This section uses observed traffic count data (provided by LaDOTD) to validate the simulation results. By aggregating the simulated trips passing through a specific road segment, we are able to obtain the simulated traffic count at a particular location and compare to actual recorded traffic data.

Figure 8 shows traffic monitor locations in the study area. Most are located along major arterials such as the interstate and state highways. There are 211 traffic monitor stations with

reported traffic counts in 2014. The traffic counts are recorded on a 48-hour cycle, and then converted to annual average daily traffic (AADT) count. AADT counts are inclusive of all lanes of travel, in both directions.
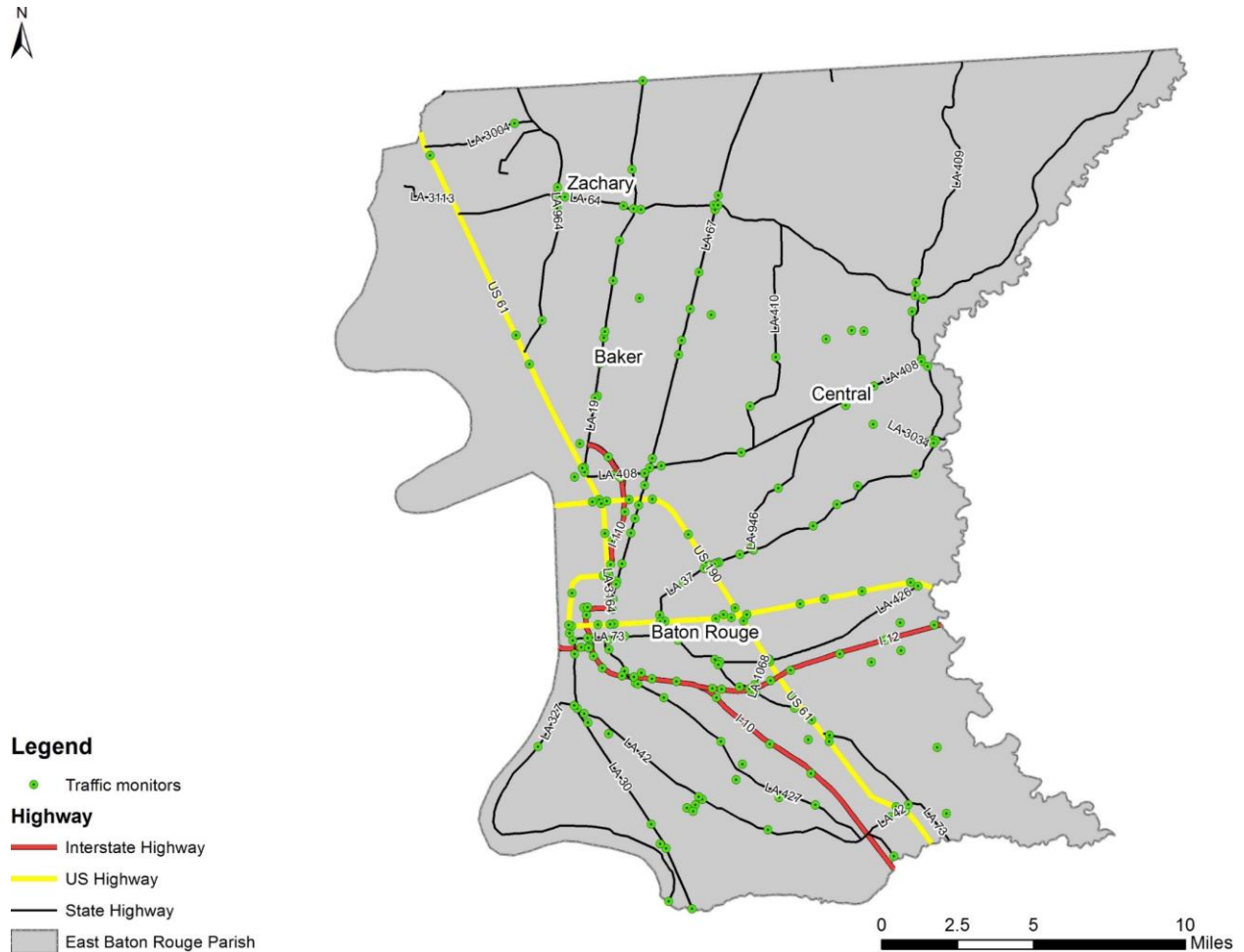
Figure 8. Traffic monitor locations

As shown in Figure 9, the simulated traffic counts and the AADT counts are largely consistent. The simulation tends to underestimate traffic in the lower range of traffic volumes and overestimate traffic in the upper range of traffic volumes. The Pearson correlation coefficient is as high as 0.919, significantly better than the study by Hu and Wang (2015) with a reported Pearson correlation coefficient of 0.660.
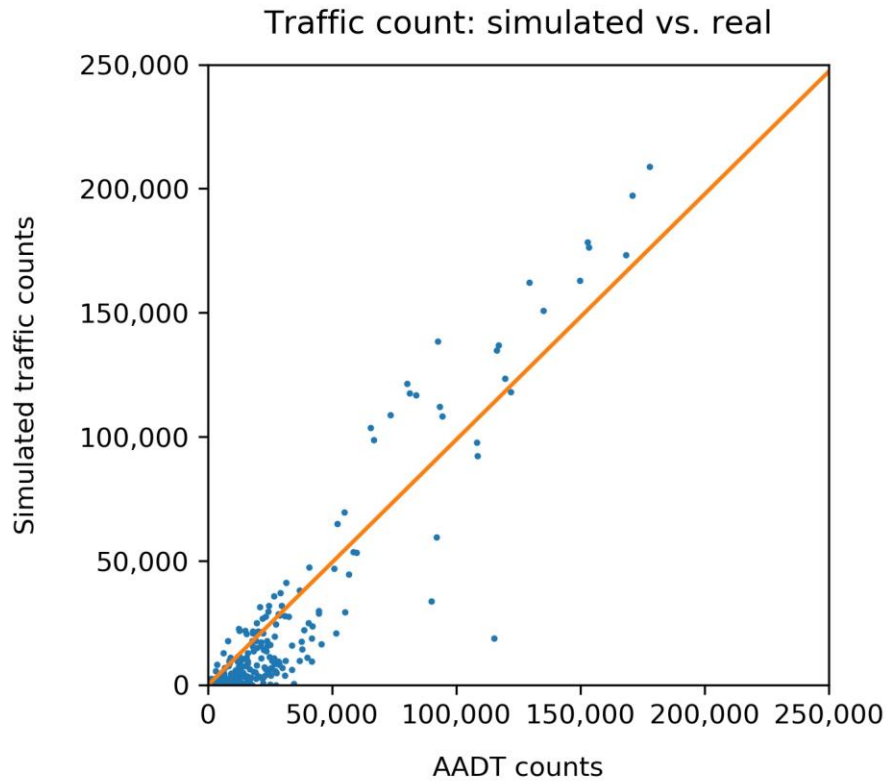
Figure 9. Simulated traffic counts vs. observed AADT in Baton Rouge

In summary, this study disaggregates inter-zonal commuting trip volume data that is originally in census tract level to individual commuting trips using the four-step model. This model is based on Monte Carlo simulation to generate trip ends and form trip routes. Land use data is used to constrain trip end locations so that the simulation results are more reliable. By validating the simulation results to the actual traffic count data, the trip simulation method used in this study proves to be able to produce usable simulation to commute pattern, paving way to access an important factor: travel time.

CHAPTER 6. TRAVEL TIME ESTIMATION

Based on the commuting trips simulated from Chapter 5, this chapter discusses how to estimate

travel time of trips by private vehicle and by public transit since the difference between them is a

key element in commuter's mode choice. The Network Analyst module in ArcGIS is used to

estimate driving time by private vehicle through road network, and this process is fairly routine

(see Wang (2015)) and thus not discussed here. CATS provides a bus transit system covering a

limited area in the study area. This system has fixed routes and stops, and operates under a fixed

schedule. A reliable estimation of travel time by transit calls for the use of a schedule-aware

method – for this study, the GTFS model. This chapter focuses on estimating the travel time via

transit network.

6.1. The GTFS data model for transit

Since its first launch in 2005, GTFS has been the most popular data format to describe

fixed-route transit systems (Antrim and Barbeau 2013). Implemented in Google Transit, its main

functionality is transit trip routing, providing information on transit routes, transfers and travel

time and distance for trips via a public transit network. GTFS data model represents a fixed-route

transit system in a series of tables in the form of comma-delimited text files. These tables use

data with pre-defined field names to describe multiple components of transit system, such as

agency basics, transit stops and routes, schedules, etc. Among those tables, six are necessary to

create a functioning GTFS data set (https://developers.google.com/transit/):

1. File "agency.txt" contains basic information about transit *agency* such as its unique
   ID, full name, URL, and time zone.

2. File "stops.txt" contains information about each transit *stop* such as its ID, name, and geographic coordinates in latitude and longitude.

3. File "routes.txt" contains information about each transit *route* such as its ID, affiliated transit agency ID, short and full names, and type (e.g., bus, subway).

4. File "trips.txt" contains information about each transit *trip* that belongs to every transit routes of every transit agency within the transit system such as its route ID, service calendar ID, and trip ID.

5. File "stop_times.txt" contains information about the *stop times* a vehicle arrives at and departs from each individual transit stop for each individual trip.

6. File "calendar.txt" contains *operation calendar types* of weekly schedules (e.g., business day only, weekend only, etc.) to be referenced by the transit trips table.

An additional route shape table "shapes.txt" describes how transit routes are drawn in GIS for visualization purposes. As shown in Figure 10, all tables are interconnected to each other by the common fields (keys), similar to a relational database, so that essential trip information such as routes, departure/arrival and transfer locations and times, trip lines, and especially travel time can be derived by routing analysis applications.
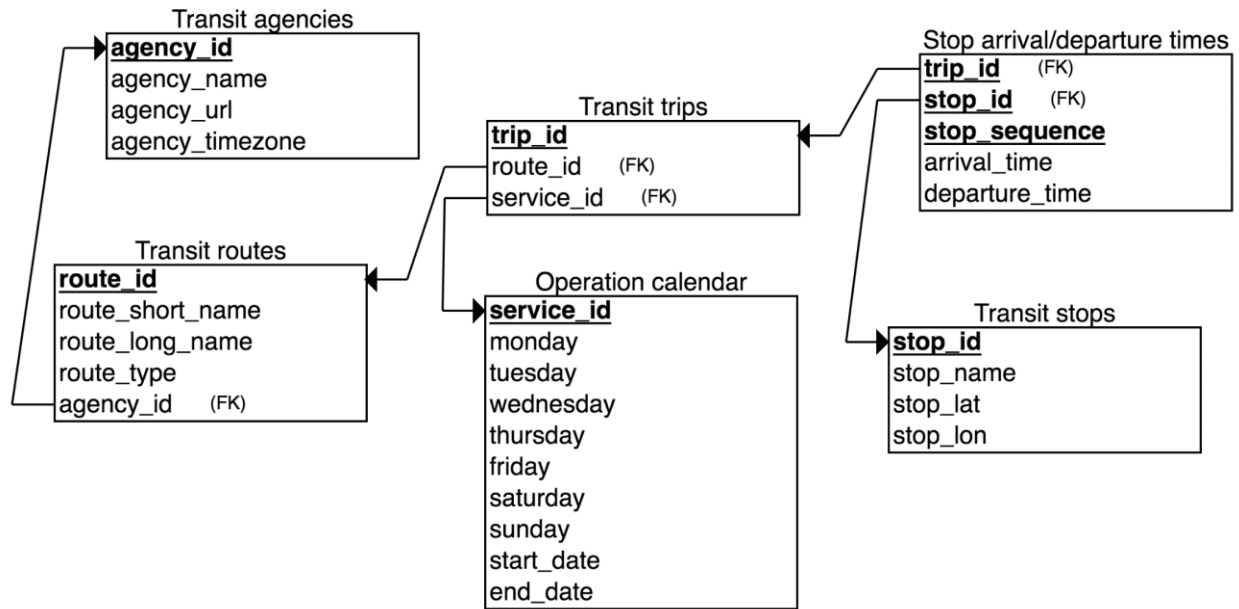
Figure 10. Interconnected GTFS tables

6.2. Transit travel time estimation by "Yay, transit!"

This study uses an application "Yay, transit!" from Esri to implement GTFS into the Network Analyst module of ArcGIS. "Yay, transit!" translates GTFS text files into operational transit data in ArcGIS by building spatial components of transit network and interpolate temporally conscious transit operations (Tallis 2014). This section details the step-by-step implementation.

This process begins with building a transit network. It first generates transit lines, stop points and a database of schedules from the aforementioned GTFS text files. Figure 11 shows the bus stops and routes of CATS. Secondly, it creates small connector features between transit lines/stops and road network. Specifically, it snaps each bus stop to the closest point of the closest street, and builds a connector line between them to apply a time delay for boarding and un-boarding bus (see Figure 12). Last, it creates a transit network dataset in ArcGIS that includes road network, transit lines, connector lines, and bus stops and their snapped points.
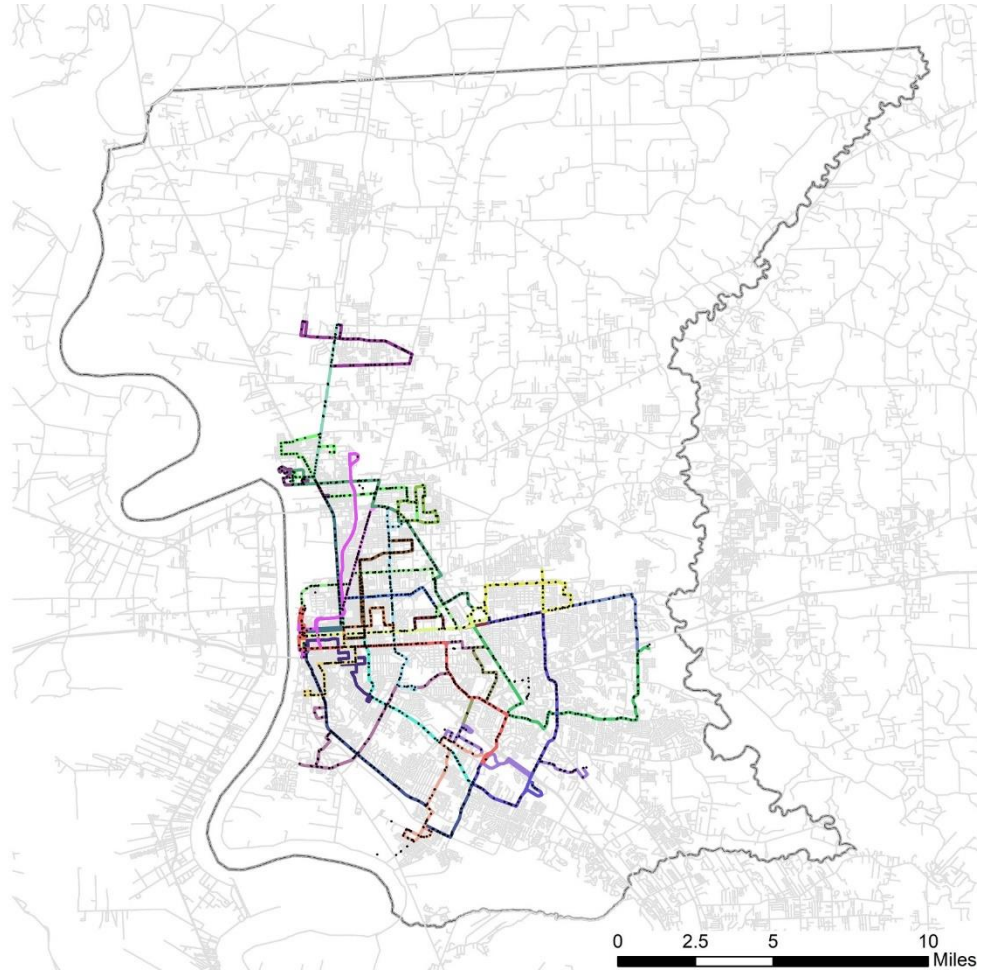
Figure 11. CATS system map

Copy stop and snap to street   Create connector line   Create vertex on street feature
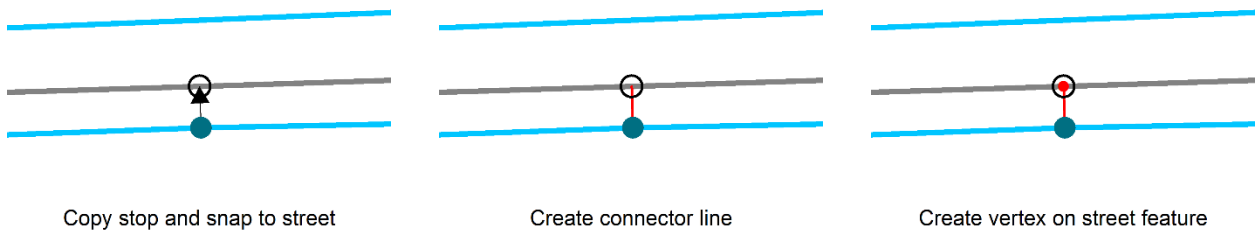
Legend

● New vertices   ● Bus stops   ○ Snapped points   ━━ Transit lines   ━━ Connectors   ━━ Streets

Figure 12. Build connections between bus stops and streets

Once the transit network dataset is built, travel time can be calculated. "Yay, transit!" decomposes the travel time of a trip T into four parts (Farber, Morang, and Widener 2014):

$$T = T^{wk} + T^{wt} + T^{b} + T^{r} \qquad (6.1)$$

where T^{wk} represents the walking time, including home-to-stop, stop-to-work, and transferring between two bus stops, T^{wt} denotes the waiting time, T^b is a constant boarding/un-boarding time for each bus trip, and T_i^r stands for the bus-riding time (shown in Figure 13). Both T^{wt} and T^r are queried from the GTFS schedules.
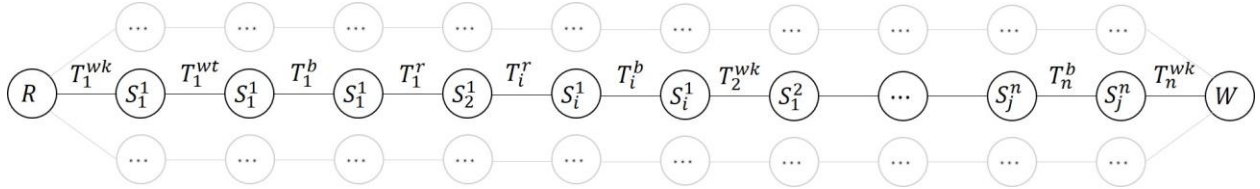
Figure 13. Querying a complete transit commuting trip route from residence to workplace

Figure 14 uses an example to further illustrate the transit routing problem:

1. Suppose a passenger leaves home at 7:30 a.m., the solver finds one possible walking route to a nearby stop 1.1, and calculates the walking time that $T_1^{wk} = 2$. This is the impedance value of the first walking edge, and the passenger arrives at the starting stop at 7:32 a.m.

2. The solver queries the GTFS database to get the arrival time of the next bus, which is 7:38 a.m., and calculates the impedance value of the waiting edge $T_1^{wt} = 6$, which is the difference between the bus arrival time and the passenger's initial arrival time at the stop.

3. The impedance value of the boarding edge T_1^b is set as a constant: 15 seconds, or 0.25 minutes.

4. By querying the GTFS database again for arrival time for the next stop 1.2, which is 7:39 a.m., the solver calculates the impedance value of the first riding edge (in light blue) $T_1^r = 1$.

5. Now the passenger needs to transfer to another route (in dark blue) by un-boarding the bus at stop 1.2 and walk to stop 2.1. Another boarding edge (un-boarding, 0.25

minutes) and walking edge (from stop 1.2 to stop 2.1) $T_2^{wk} = 1$ is added to the graph, and the passenger arrives at stop 2.1 at 7:40 a.m.

6. The process repeats Steps 2-5 to add more waiting (9 minutes), boarding and riding edges (2 minutes) until the passenger finally reaches the last stop (stop 2.3) at 7:51 a.m.

7. One last walking edge (4 minutes) is added for walking from the final stop to workplace, and the passenger arrives at the workplace at 7:55 a.m.
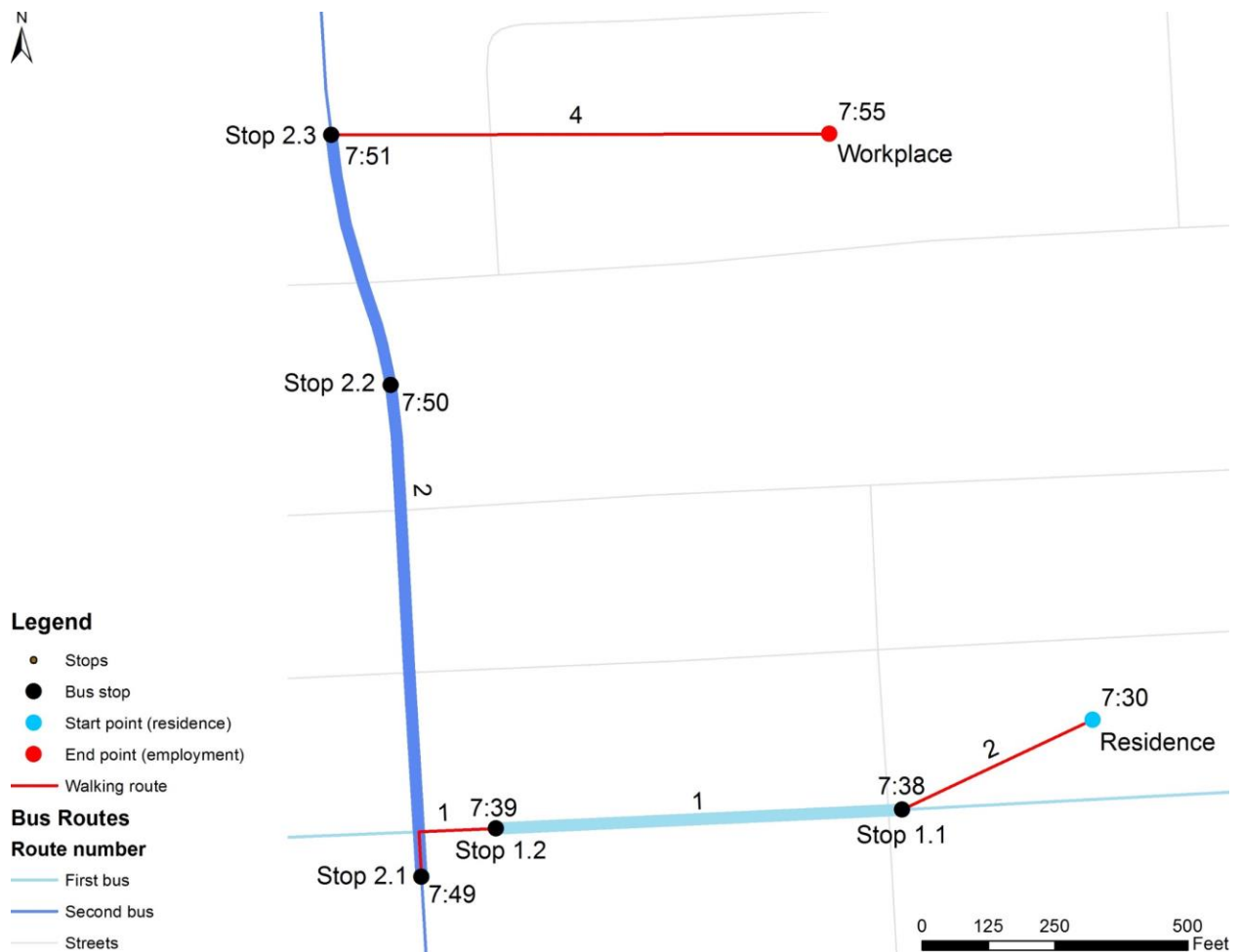

Figure 14. An example of a complete transit commuting trip from residence to workplace

6.3. Transit-to-driving travel time ratio

The difference in travel time between transit and private vehicle is measured as ratio. The transit-to-driving travel time ratio for a census block as a whole is calculated as the travel time ratio for every commuting trip originated from this census block:

$$v = \frac{\overline{T^t}}{T^d} \tag{6.2}$$

where T^t and T^d stand for the transit commuting time and driving time, respectively. Figure 15 maps out the variation of transit-to-driving travel time ratio across the study area. Commuting with private vehicle is much faster and more convenient across the study area, as riding public transit usually involves walking, waiting, and boarding/un-boarding, and bus typically would typically travel slower. In the southern half of Baton Rouge (south of US Highway 190), the map shows a trend that travel time ratios of census blocks along Interstate Highways 10 and 12 are generally higher than distant ones. The public transit system in the study area mostly bypasses freeways, so that the commuting time advantage for census blocks near freeway would be more significant, making public transit a less appealing choice for commuters in those neighborhoods. In comparison, in the northern half of the study area (north of US Highway 190), the travel time ratio is not as high, as there are no nearby freeways, and there are more bus routes and stops (see Figure 11).

In summary, the travel time of public transit is still measured based on the traditional Dijkstra's algorithm (Dijkstra 1959). But the traverse time of each trip segment is not simply calculated as distance divided by speed, like one would do for a private vehicle trip. Instead, the traverse time is dynamically determined by querying GTFS transit schedule (for waiting and riding), or by calculating walking time (for walking). This leads to more accurate estimate of

transit travel time comparing to survey or other indirect approaches, and provides a more reliable

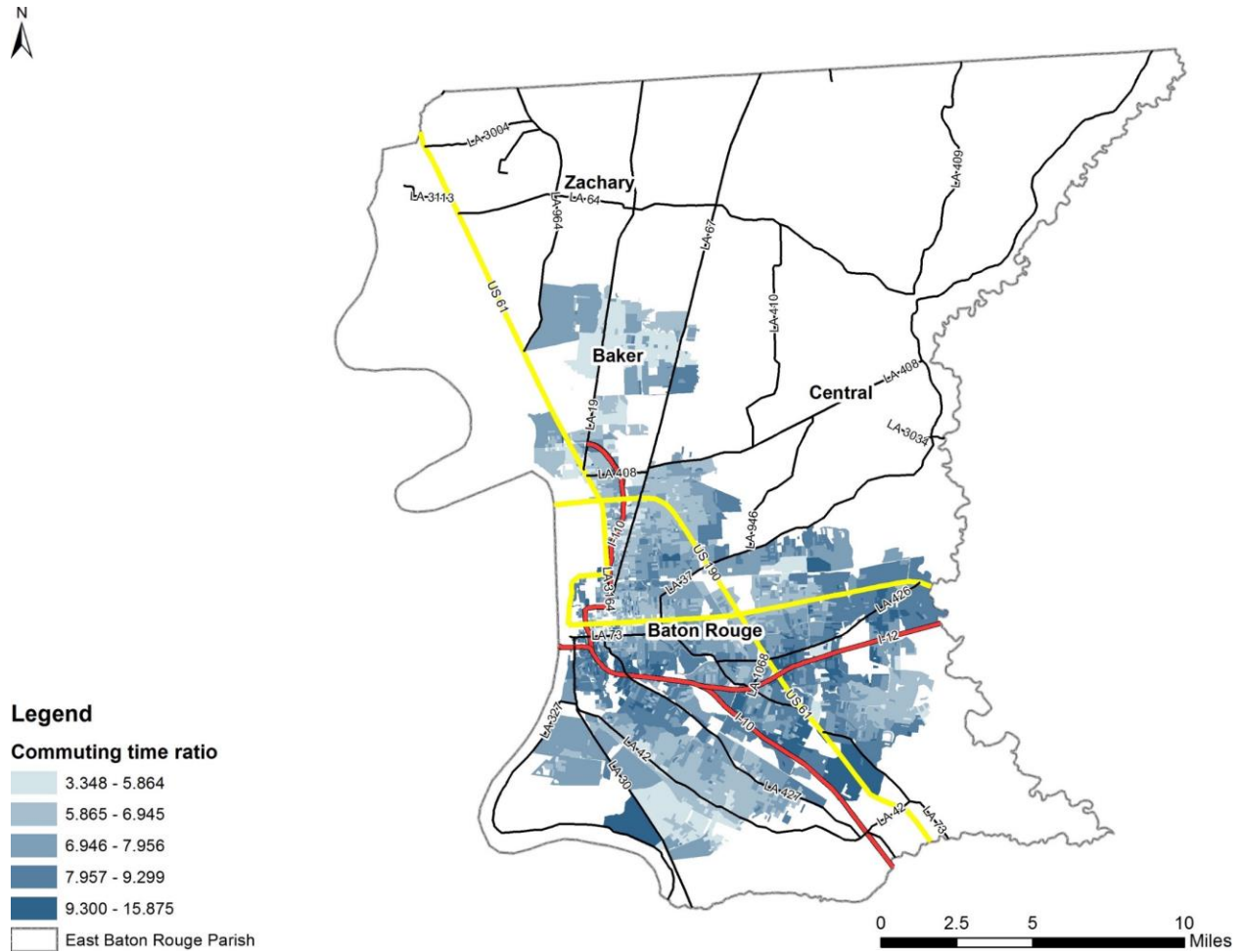factor to evaluate commuter's behaviors.

Figure 15. Transit-to-driving travel time ratio in Baton Rouge

CHAPTER 7. GLOBAL AND LOCALIZED NEIGHBORHOOD EFFECTS ON PUBLIC TRANSIT RIDERSHIP

Like many issues, public transit ridership in neighborhood may be explained by "who they are" (i.e., non-spatial factors) and "where they are" (i.e., spatial factors). As discussed in Chapters 2 and 3, this study focuses on the factors at census block level. These factors are often referred to as "external factors" as they are external to public transit system. Many other factors such as fare and quality of service of transit system, gasoline price, accessibility to parking facilities, and weather are not considered due to the scope of this study. More importantly, due to a lack of access to data of individual commuters (e.g., age, sex, gender, education attainment, income, occupation, etc.), this study is limited to census block level. While this study makes all efforts to disaggregate data from larger areal units (census tract or block group) down to a small areal unit (census block), it still remains at aggregate areal level and thus is not completely immune from possible criticism of "ecological fallacy" (Robinson 1950). Nevertheless, the disaggregation approaches, implemented for measures of both non-spatial and spatial factors, have prepared data at a sharper spatial resolution and of a larger sample size, and thus helped mitigate some of the problems (e.g., limit loss of information or variability of variables in data aggregation). When individual-level data is available, future work will employ a multilevel modeling approach in detecting both individual commuter behavior and neighborhood effects (e.g., Antipova, Wang, and Wilmot (2011)).

7.1. Selecting explanatory variables by stepwise regression

As outlined in Table 4, this research selects 25 variables to explain the variation of public transit ridership across the study area at census block level. These 25 variables represent several different aspects of neighborhood characteristics including demographic and socio-demographic

(or non-spatial), and spatial factors. All variables are log-transformed for the following analyses. A forward stepwise regression is used to refine the variable list to minimize the possibility of unnecessary inclusion of variables. The selection results are reported in Table 5. Variables that meet both of the following criteria are entered as explanatory variables in the following regression analyses:

1. Probability of F-to-enter ≤ 0.05. This is the default criteria. If the addition of a variable does not inflate the probability of the model's F-test to above 0.05. The variable would be accepted.

2. Contribution to adjusted $R^2 \geq 0.010$. It is possible that some variables that meet the previous criterion do not contribute enough to the model. This study adds another layer of control to retain the variables that increases to the model's adjusted R^2 by more than 0.010, in order to keep the model as concise as possible.

The stepwise regression retains 6 variables (denoted by * in Table 5) out of 25 as the explanatory variables in the following regression models. Among the excluded variables, non-spatial ones like age, householder gender, education attainment, and spatial ones such as nearby sidewalk length are not statistically strong indicators of public transit ridership. Note that neither population density nor worker density is a statistically significant indicator of census block's public transit ridership, as well as housing unit density and residential land use, which is contradictory against most existing studies (Pushkarev and Zupan 1977; Hendrickson 1986; Spillar 1989; Chung 1997). Another scenario of exclusion of variable is because there is another variable retained to represent a group of highly correlated variables. For example, median household income is retained as the representative of per capita income, poverty rate and percentage of renter-occupied housing units.

Table 5. Stepwise regression results

Code	Variable	p-value	R^2 change
	Median age (years)	0.582	-
White	*White population (%)	0.000	0.013
	Female headed householder (%)	0.253	-
	Population with high school diploma (%)	0.671	-
English	*Household speaking limited English (%)	0.000	0.012
	Poverty rate (%)	0.012	0.000
Income	*Median household income (dollars)	0.000	0.012
	Per capita income (dollars)	0.368	-
	Household receiving food stamp (%)	0.599	-
	Unemployment rate (%)	0.131	-
	Vacant housing unit (%)	0.225	-
	Renter-occupied housing unit (%)	0.025	0.000
	Housing unit with >1 persons/bedroom (%)	0.523	-
Carless	*Carless housing unit (%)	0.000	0.188
Plumb	*Housing unit with incomplete plumbing device (%)	0.000	0.002
Kitchen	*Housing unit with incomplete kitchen (%)	0.000	0.008
Time	*Bus-riding-driving time ratio	0.000	0.105
Distance	*Distance to nearest bus stop (meters)	0.000	0.001
Jobs	*Number of jobs within 1-hour transit service area	0.000	0.034
	Number of nearby bus stop	0.217	-
	Population density (per acre)	0.142	-
	Worker density (per acre)	0.756	-
	Housing unit density (per acre)	0.203	-
	Nearby sidewalk length (meters)	0.606	-
	Low-density residential land use	0.228	-

*: retained variables

7.2. Assessing neighborhood global effects

After variable selection, 6 selected non-spatial and spatial variables are included in the regression models. The initial analysis is to access the global effect of these variables on public transit ridership by building a global model with OLS regression to access their global effects on public transit ridership across census blocks. The global model is:

$$y_i = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_9 v_9 + \varepsilon_i \qquad (7.1)$$

where y is the dependent variable (public transit ridership), β_0 is the model intercept, and β_1

through β_9 are the parameters for explanatory variables v_1 through v_9, respectively. The

coefficient estimators B, in matrix notation, are solved as:

$$B = (V^T V)^{-1} V^T Y \qquad (7.2)$$

where V is a matrix for the explanatory variables, and Y is a vector for the dependent variable.

The results of OLS regression are reported in Table 6. The variance inflation factor (VIF)

for all variables are well below the threshold of 2.500, showing that multicollinearity between

the variables is relatively low.

Table 6. OLS regression results

Variable	Coefficient estimate	Standard error	t-statistic	P-value	VIF
Intercept	6.372	0.446	14.300	0.000	-
White	-0.083	0.012	-6.764	0.000	1.362
English	0.296	0.031	9.459	0.000	1.091
Income	-0.236	0.033	-7.08	0.000	1.427
Carless	0.271	0.021	12.943	0.000	1.482
Time	-3.798	0.153	-24.887	0.000	1.150
Jobs	0.235	0.016	14.671	0.000	1.058
Adjusted R^2: 0.364; AICc: 13,246.548					

The global results suggest that higher public transit ridership tend to be associated with

less White population, lower median household income, and certainly fewer carless population in

census block. On the other hand, more linguistically isolated census blocks, which may indicate

more recent immigrants, as well as census blocks with worse housing conditions (i.e., housing

units with incomplete plumbing or kitchen facilities), would probably have more people

commuting by public transit. For spatial factors, greater commuting time difference between

riding public transit and driving private vehicle is related to lower public transit ridership.

Distance to the nearest bus stop is surprisingly positively related to public transit ridership. The last factor, number of jobs within a 1-hour transit service area, shows a positive relationship, meaning that for a census block, if more jobs available with public transit, more commuters tend to choose public transit. As shown in Figure 16, the spatial distribution of residuals from the global model shows some level of clustering. This is further confirmed by its global Moran's I of 0.151 with a z-score of 52.338 and a p-value of 0.000, which indicates statistically significant spatial autocorrelation.
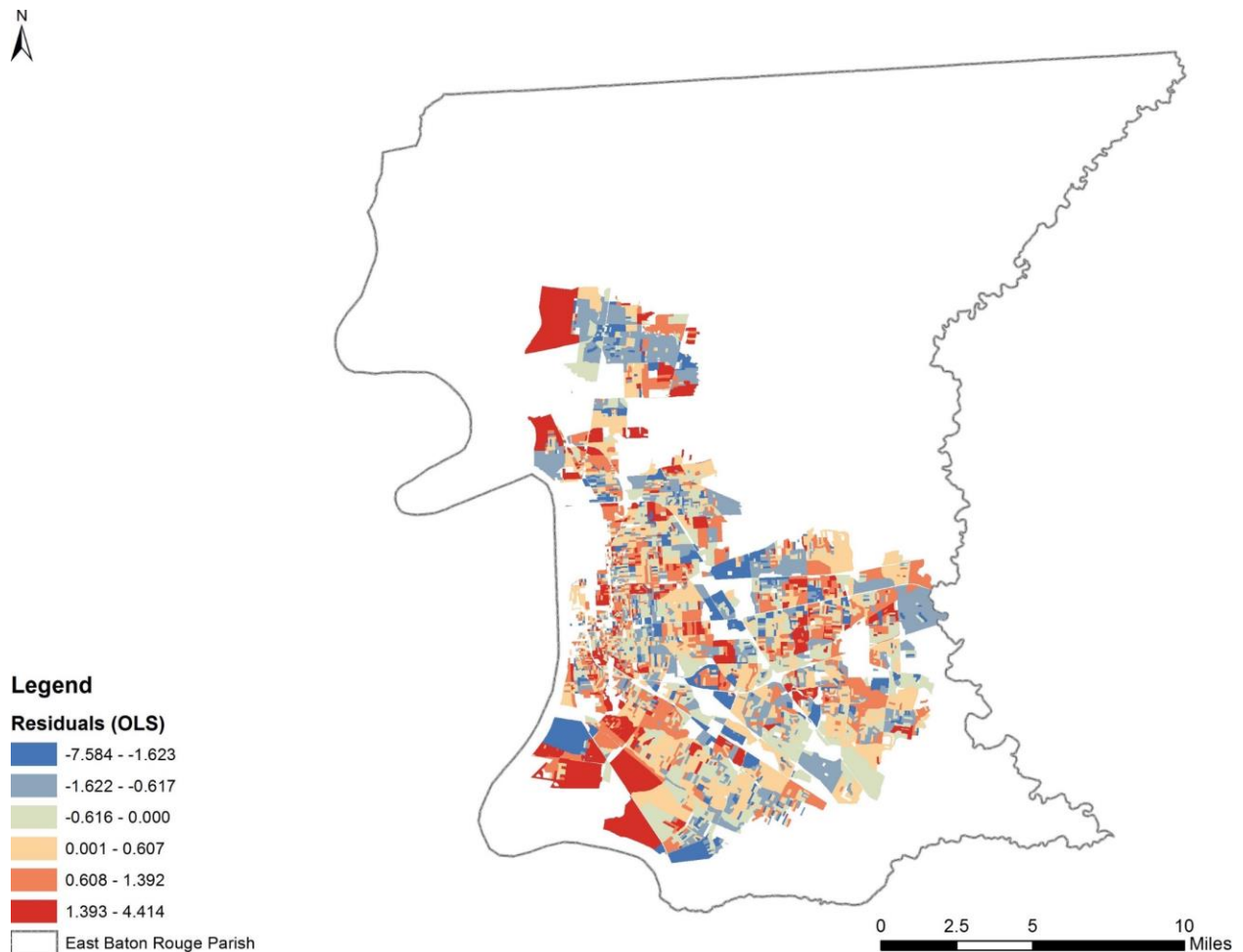

Figure 16. Residuals from the OLS model

The presence of spatial autocorrelation indicates that the errors derived from the OLS regression are systemically related to each other, and similar values are next to each other, as

suggested by the positive Moran's I. This violates a major assumption of OLS regression and raises questions whether the estimators of coefficients for the corresponding independent variables are unbiased and reliable.

7.3. Assessing neighborhood localized effects

There are several approaches developed in the literature that can control or mitigate the effect of spatial autocorrelation. For example, spatial error model and spatial lag model can be estimated by a maximum likelihood method (Wang 2015). However, those are broadly defined as global spatial regression models that detect only the global effects of those explanatory variables. In other words, they assume that the effect of each explanatory variable is uniform across the entire study area. Recent literature suggests that it is common, especially in socio-economic applications in geographic or spatial data, that the relationship between explanatory and dependent variables varies across study area. A localized model by geographically weighted regression (GWR) is developed to analyze the spatially varying relationships and permits spatial non-stationarity in the regression coefficients (Fotheringham, Brunsdon, and Charlton 2002). GWR stems from OLS regression with varying coefficients. The GWR model here is written as:

$$y_i = \beta_{0_i} + \beta_{1_i} v_1 + \beta_{2_i} v_2 + \cdots + \beta_{9_i} v_9 + \varepsilon_i \tag{7.3}$$

where β_{0_i} through β_{9_i} are weighted parameters, and i indexes census block. In matrix notations, these parameters (denoted as vector B) are estimated by implementing a spatial weight matrix W_i to Equation 7.2:

$$B = (V^T W_i V)^{-1} V^T W_i Y \tag{7.4}$$

The same 9 independent variables are used in the localized model. The model yields an adjusted R^2 of 0.558 and an AICc of 12,190.100. Comparing to the global model whose adjusted

R^2 is 0.364 and AICc is 13,246.548, the localized model captures spatial effects and has a better model fit. Furthermore, given the residuals mapped in Figure 17(a), as well as its global Moran's I of 0.002 with a z-score of 0.827 and p-value of 0.408, the spatial distribution of the residuals from the localized model is confidently random. Therefore, the use of GWR in place of OLS regression is well warranted, and overall the localized effects of neighborhood variables are significant.

Legend

Residuals (GWR)

■	-5.517 - -1.197
■	-1.196 - -0.471
■	-0.470 - 0.000
■	0.001 - 0.374
■	0.375 - 0.953
■	0.954 - 3.715
☐	East Baton Rouge Parish

Legend

Local R-squared

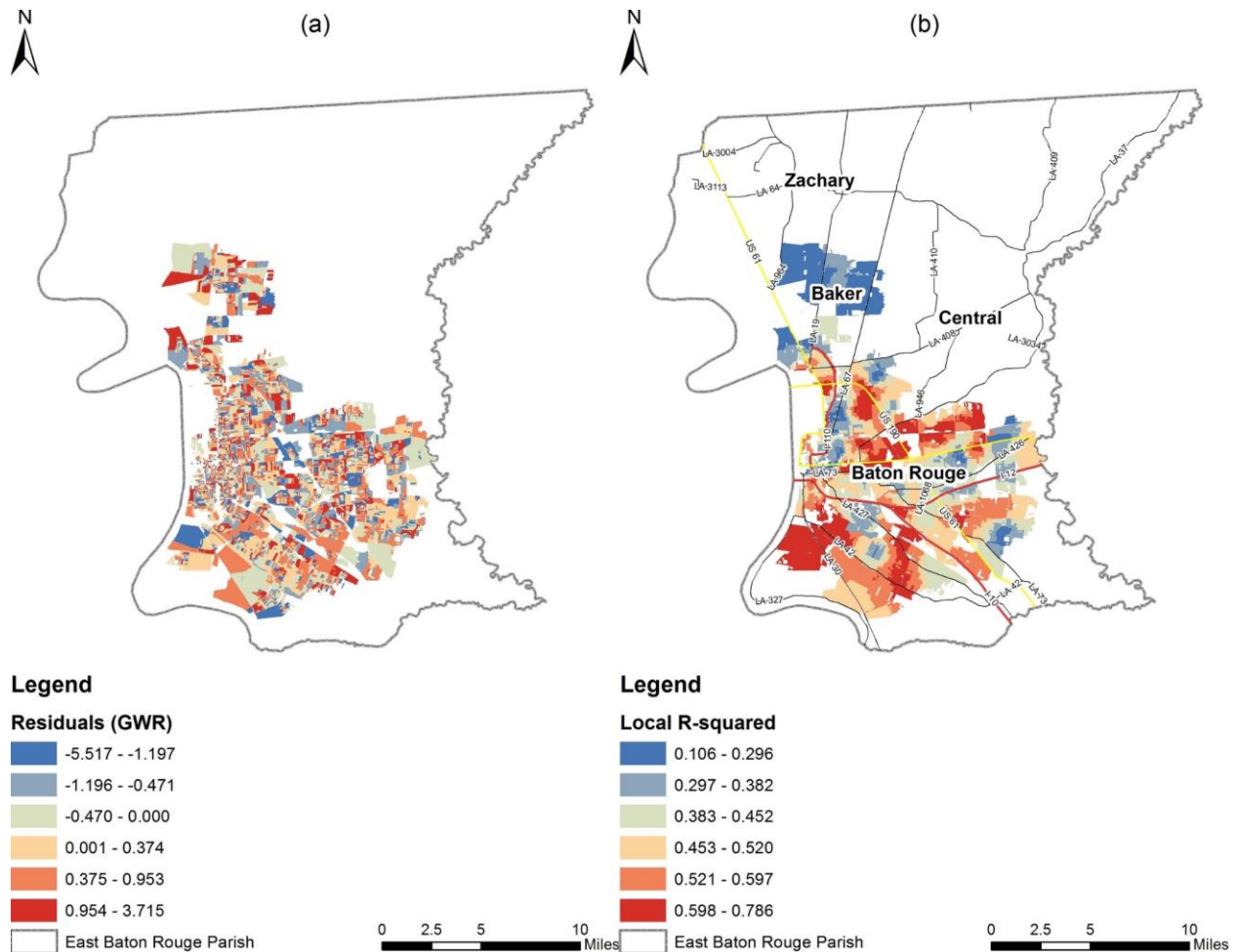|	
■	0.106 - 0.296
■	0.297 - 0.382
■	0.383 - 0.452
■	0.453 - 0.520
■	0.521 - 0.597
■	0.598 - 0.786
☐	East Baton Rouge Parish

Figure 17. Residuals and Local R^2 of the GWR model

The local R^2 of the localized model is mapped in Figure 17(b). For most areas with public transit access, the demographic, socio-economic, and spatial factors can explain fairly enough variances in public transit ridership by the localized model, while in some other areas –

notably Baker, the Southern University area, the east of ExxonMobil refinery, and the region between Interstate Highway 12 and US Highway 190, the model fit is not as sufficient.

Table 7 reports the summary statistics of coefficient estimates from the localized model. Although the means and medians report the same trends of relationships between public transit ridership and the explanatory variables, some census blocks reports opposite relationships. For example, White population percentage has a maximum coefficient estimate of 0.681 albeit a negative global estimate.

Table 7. GWR model's coefficient estimates

Variable	Mean	Std. dev.	Minimum	Median	Maximum
Intercept	9.894	7.544	-16.266	9.784	40.391
White	-0.032	0.251	-1.180	-0.018	0.681
English	0.186	0.327	-1.029	0.187	1.440
Income	-0.163	0.421	-1.837	-0.128	1.159
Carless	0.141	0.234	-0.563	0.135	0.857
Time	-5.623	2.579	-14.096	-5.100	-0.117
Jobs	0.129	0.566	-3.853	0.073	4.100
Adjusted R^2: 0.558; AICc: 12,190.100; optimal bandwidth: 131.293					

The coefficient estimates for all six variables and the intercept are mapped in Figures 18 and 19. Figure 18 reports the spatial distribution patterns for the coefficient estimates of the four non-spatial variables: (a) White population percentage, (b) percentage of households speaking limited English, (c) median household income, and (d) percentage of carless population. Figure 19 shows the spatial distribution patterns for the coefficient estimates of the two spatial variables: (a) bus-riding-to-driving time ratio, and (b) number of jobs within a 1-hour public transit service area.
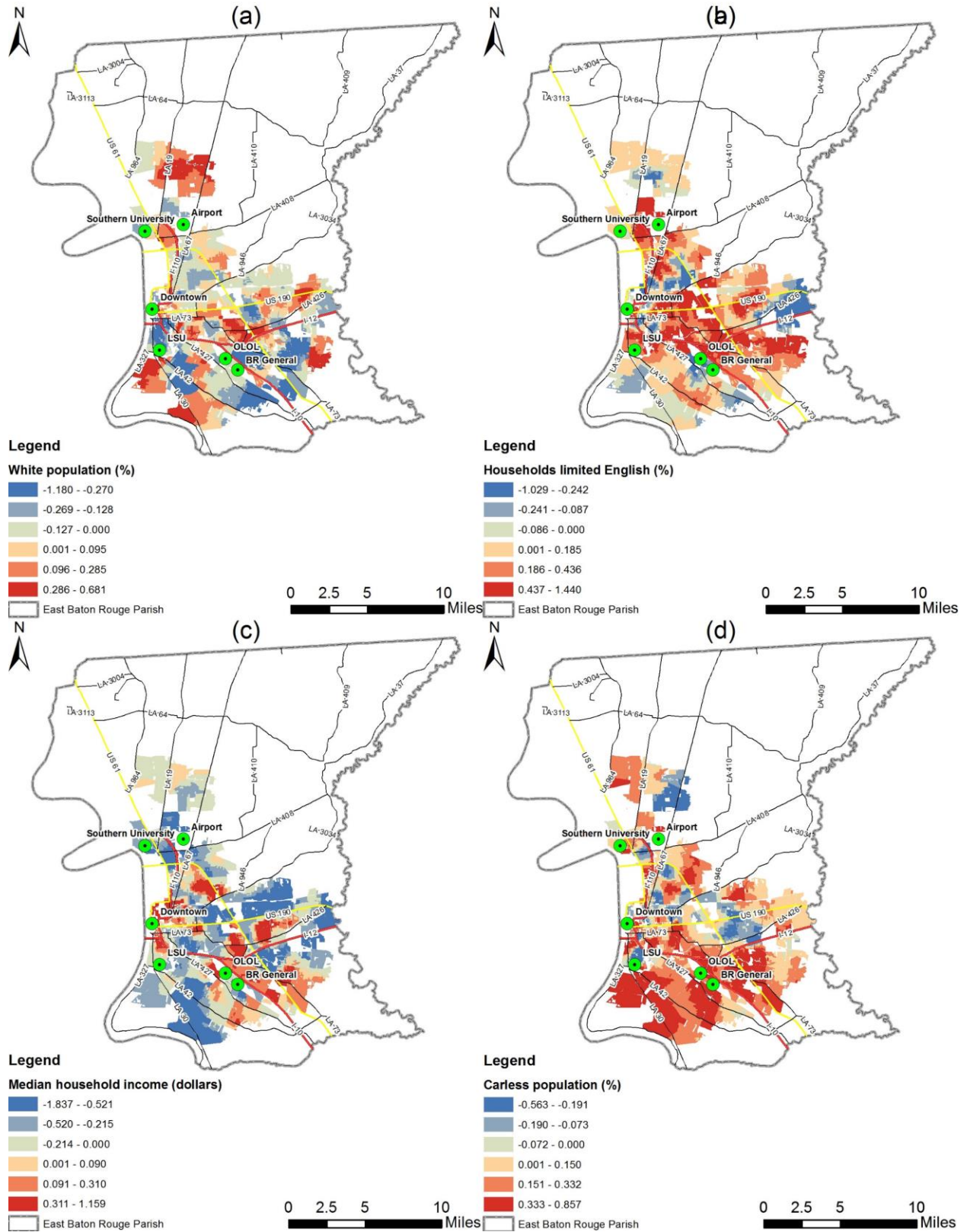
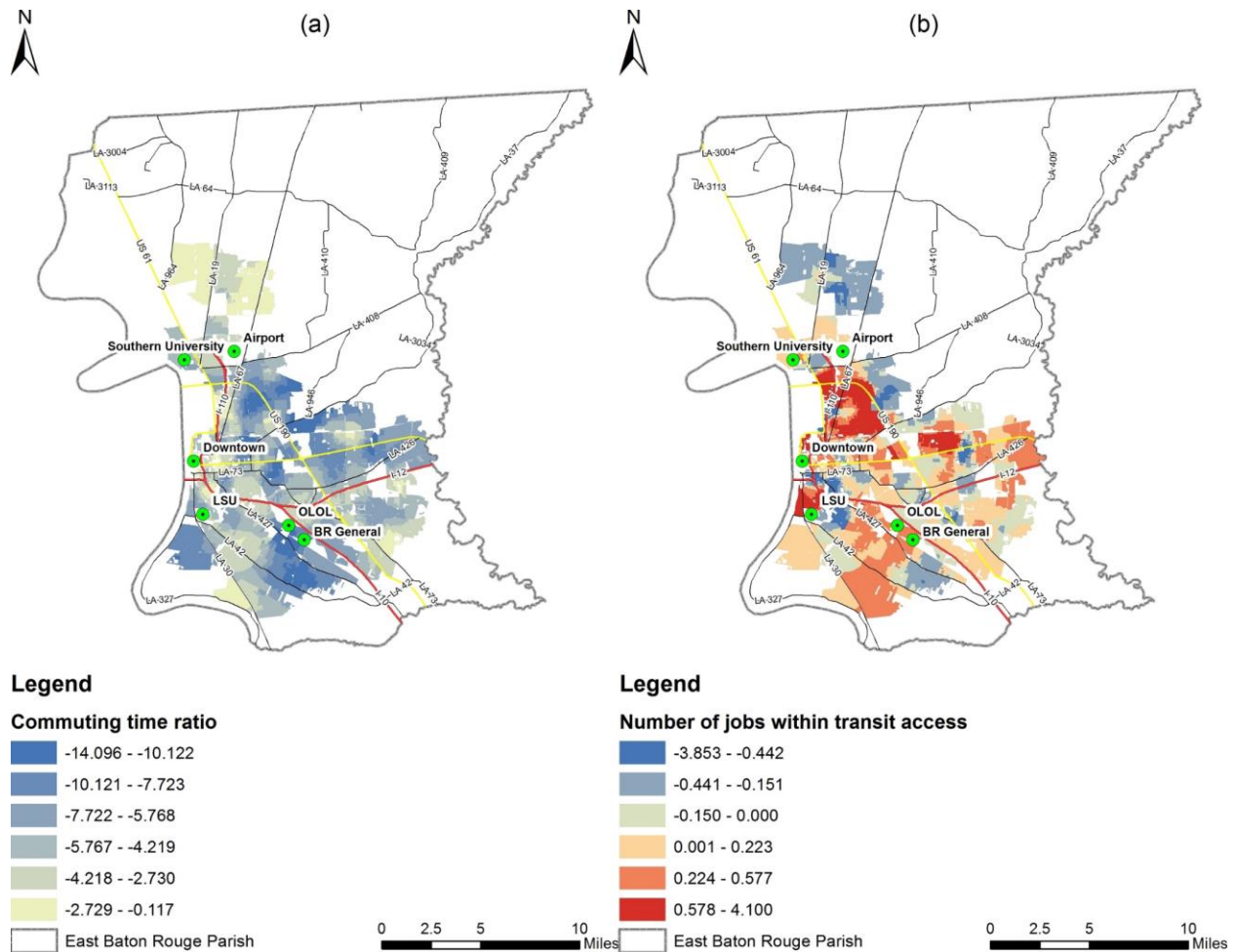Figure 18. Coefficient estimates for non-spatial variables

Figure 19. Coefficient estimates for spatial variables

7.4. Separating global and localized neighborhood effects

The localized model reports localized parameter estimates for the explanatory factors, and statistically improves the model performance when comparing to the global model. However, both models fail to consider the possibility of a combination of global and localized effect in the same model. Some factors could be used to explain public transit ridership in a constant manner across the study area, while others may have different implications on public transit ridership for different neighborhoods. This is important for this study, because if a variable actually with global effect is incorrectly modeled with localized effect, or vice versa,

future effort in improving public transit ridership based on such model could be fruitless, or even cause further decrease for some neighborhoods. Semi-parametric GWR (SGWR) is the most recent development of GWR that captures such difference by including both geographically fixed and varying parameters in the same model. Thus, factors with global effect would be modeled with constant coefficient across space, and factors with localized effect would be modeled with varying coefficient based on location. SGWR is implemented in this study using the GWR4.09 software.

To determine whether a variable is global or localized, a geographical variability test is performed (Mashhoodi, Stead, and van Timmeren 2019). This test is based on conducting a series of SGWR analyses with exactly one variable being global in each model, and then comparing their model performances to the original GWR model. For example, to test whether v_k is a global variable, the following test model:

$$y_i = \beta_{0_i} + \beta_k v_{k_i}^G + \sum_5 \beta_{l_i} v_{l_i}^L + \varepsilon_i \tag{7.6}$$

is compared to the ordinary GWR model described in Equation 7.3. In Equation 7.6, β_k is the global (constant) coefficient of the corresponding global variable v_k^G that is tested for geographical variability, and β_{l_i} is the localized (varying) coefficient of the corresponding remining localized variable v_l^L. If the test model's AICc is lower than that of the ordinary GWR model (i.e., a positive "DIFF of Criterion"), v_k should be considered a global variable in SGWR model. Otherwise, v_k should be considered a localized variable. Model performance indices other than AICc may also be used, including adjusted R^2, cross-validation (CV), and global Moran's I. Table 8 reports the results of geographical variability tests for the model variables:

Table 8. Geographical variability test results

Variable	DIFF of Criterion	Type of variable
Intercept	-3,018.216	Local
White	8.990	Global
English	22.854	Global
Income	-1,224.451	Local
Carless	5.031	Global
Time	-2,327.762	Local
Jobs	-1,443.892	Local

By employing SGWR, the model described in Equation 7.3 is further developed to a combination of three global and three localized variables, including the intercept being localized (Nakaya et al. 2005):

$$y_i = \beta_{0_i} + \sum_3 \beta_k v_{k_i}^G + \sum_3 \beta_{l_i} v_{l_i}^L + \varepsilon_i \qquad (7.5)$$

According to the geographical variability test, the variables representing White population percentage, percentage of linguistically isolated households, and carless population percentage are identified as global variables. Median household income, along with the two spatial variables: bus-riding-to-driving time ratio, and number of jobs within a 1-hour public transit service area are remained as localized variables. With global and localized variables differentiated, the results of the SGWR analysis are reported in Table 9. The SGWR analysis reports standardized coefficients so that the question of which variables are more related to public transit ridership could be answered by simply comparing their absolute values.

Table 9. SGWR model's coefficient estimates (standardized)

Global variable	Coefficient estimate	Std. error	t-statistic	Probability
White	-0.097	0.033	-2.923	0.011
English	0.118	0.020	5.841	0.000
Carless	0.168	0.024	6.866	0.000

(table cont'd)

Localized variable	Mean	Std. dev.	Minimum	Median	Maximum
Intercept	-1.936	0.945	-5.425	-1.967	2.607
Income	-0.092	0.371	-1.656	-0.083	1.410
Time	-0.831	0.463	-2.827	-0.744	0.101
Jobs	0.173	1.048	-5.708	0.117	6.656
Adjusted R^2: 0.574; AICc: 12,055.300; optimal bandwidth: 74.055					

The parameter estimates of the global factors: White population percentage, percentage of linguistically isolated households, and carless population percentage are all statistically significant at the 0.05 level. The AICc of the SGWR model further decreased to 12,055.300. In addition, Figure 20(a) maps the standard residuals from the SGWR model showing even more randomized pattern (global Moran's I: -0.002, z-score: 0.622, p-value: 0.534). Figure 20(b) maps the local R^2 of the SGWR model. For most areas with public transit access, the demographic, socio-economic, and spatial factors can explain fairly enough variances in public transit ridership with the SGWR model, despite the fact that the model fit not as sufficient in some other areas – notably Baker, the Southern University area, east of ExxonMobil refinery, and the region between Interstate Highway 12 and US Highway 190. Judging from the statistical diagnostics, the SGWR model reports better model performances comparing to the previous models (see Table 10).

Table 10. Diagnostics of the OLS, GWR, and SGWR models

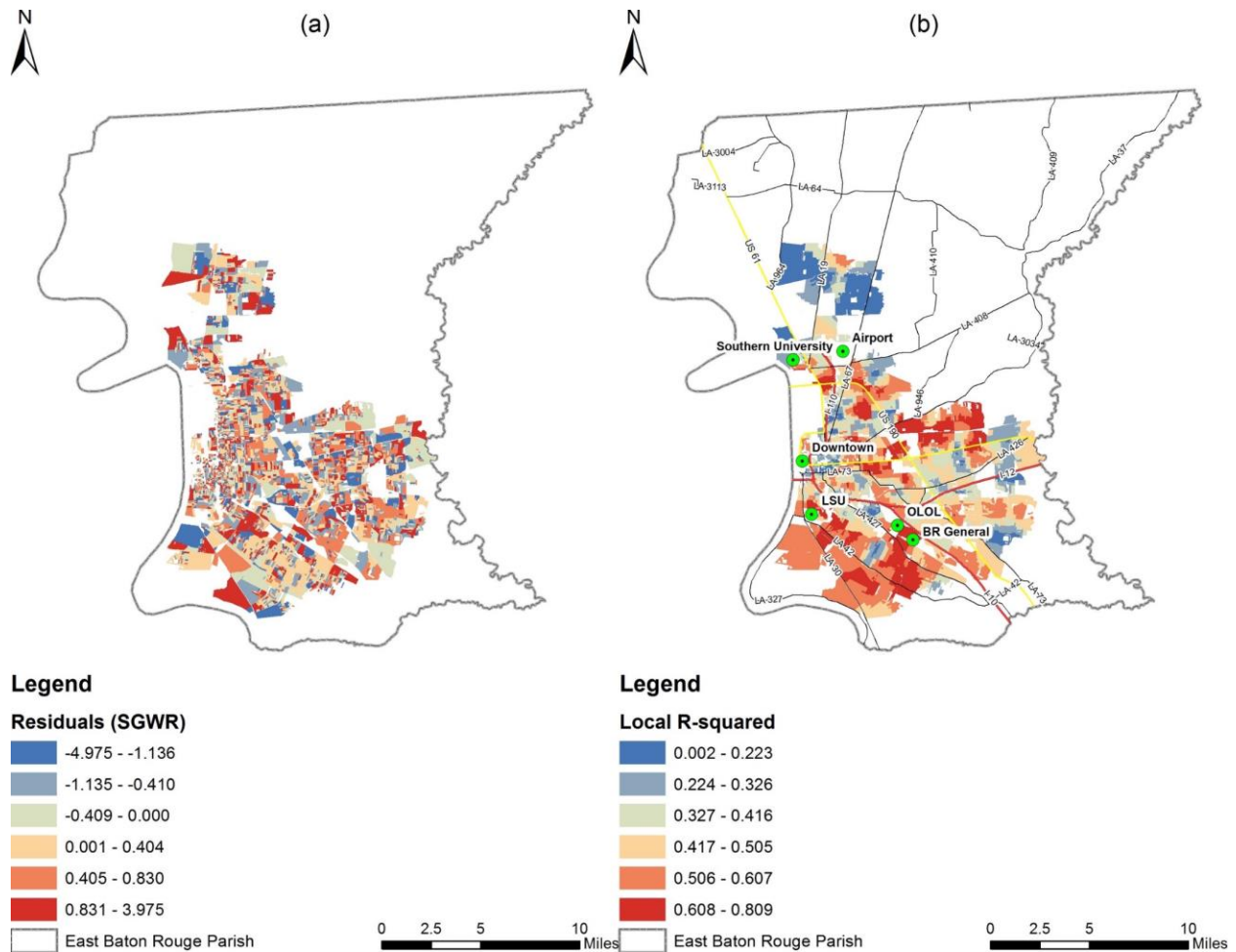|	OLS	GWR	SGWR
Adjusted R^2	0.364	0.558	0.574
AICc	13246.548	12,190.100	12,055.300
Residuals Moran's I	0.151	0.002	-0.002
p-value	0.000	0.408	0.534

Figure 20. Residuals and local R^2 of the SGWR model

Generally speaking, the parameter estimates of the global factors appear to have similar trend to what global model reports. White population percentage is negatively related to public transit ridership, meaning that commuters in neighborhoods with higher concentration of White population are more willing to drive to work. More linguistically isolated neighborhoods, or neighborhoods with more recent immigrants, are related to higher public transit ridership. In terms of car ownership, certainly more carless population would indicate more commuters taking public transit. The global results point to positive relationship between public transit ridership and disadvantages in demographic and socio-economic status. While it is a social welfare to provide public transit services to minorities and people with low mobility, it might also help to

increase public transit ridership in the advantageous neighborhoods by promoting the benefits of commuting with public transit, such as reducing air pollution, or forcing reasonable amount of walking that is beneficial to health, etc. Such policies could be implemented to the advantageous neighborhoods to dig the potential of public transit commuters. Among the three global factors, the results show that car ownership has the most impact on public transit ridership, with a parameter estimate of 0.168. When compared to the three localized factors: income, commuting time ratio, and access to jobs via public transit, the parameter estimate of carless percentage is greater than their parameter estimates in only 41.9%, 3.1%, and 21.2% of census blocks in the study area, respectively. For the other two global factors with less contribution, White population has greater impact than the localized factors in 25.5%, 1.7%, and 12.3% of census blocks, respectively, and recent immigrant level has greater impact than the localized factors in 30.4%, 2.0%, and 14.9% of census blocks. The comparisons speak that car ownership is the most contributive global factor, and the global factors do not have greater impacts comparing to the localized factors.

Figure 21 – 23 maps the spatial distribution patterns for the parameter estimates of the localized variables. The spatial distribution patterns generally show similar trends to those of the four corresponding variables in the ordinary GWR model, as discussed in the previous section. The nature of the relationships between public transit ridership and these variables vary across space. Figure 21 reports the spatial distribution patterns for the parameter estimate of median household income. The mean and median are -0.092 and -0.083, respectively. This is consistent with the global model, meaning that generally, richer neighborhoods tend to drive to work. However, the map shows that for a lot of neighborhoods, higher income level would instead be related to higher public transit ridership in 38.5% of census blocks. The distribution of the

61

converse trend represented by the warmer colors is dispersed, and the notable areas include the east and south ExxonMobil refinery, the north of downtown, the north of LSU, the Sherwood Forest region, and the Old Jefferson area.
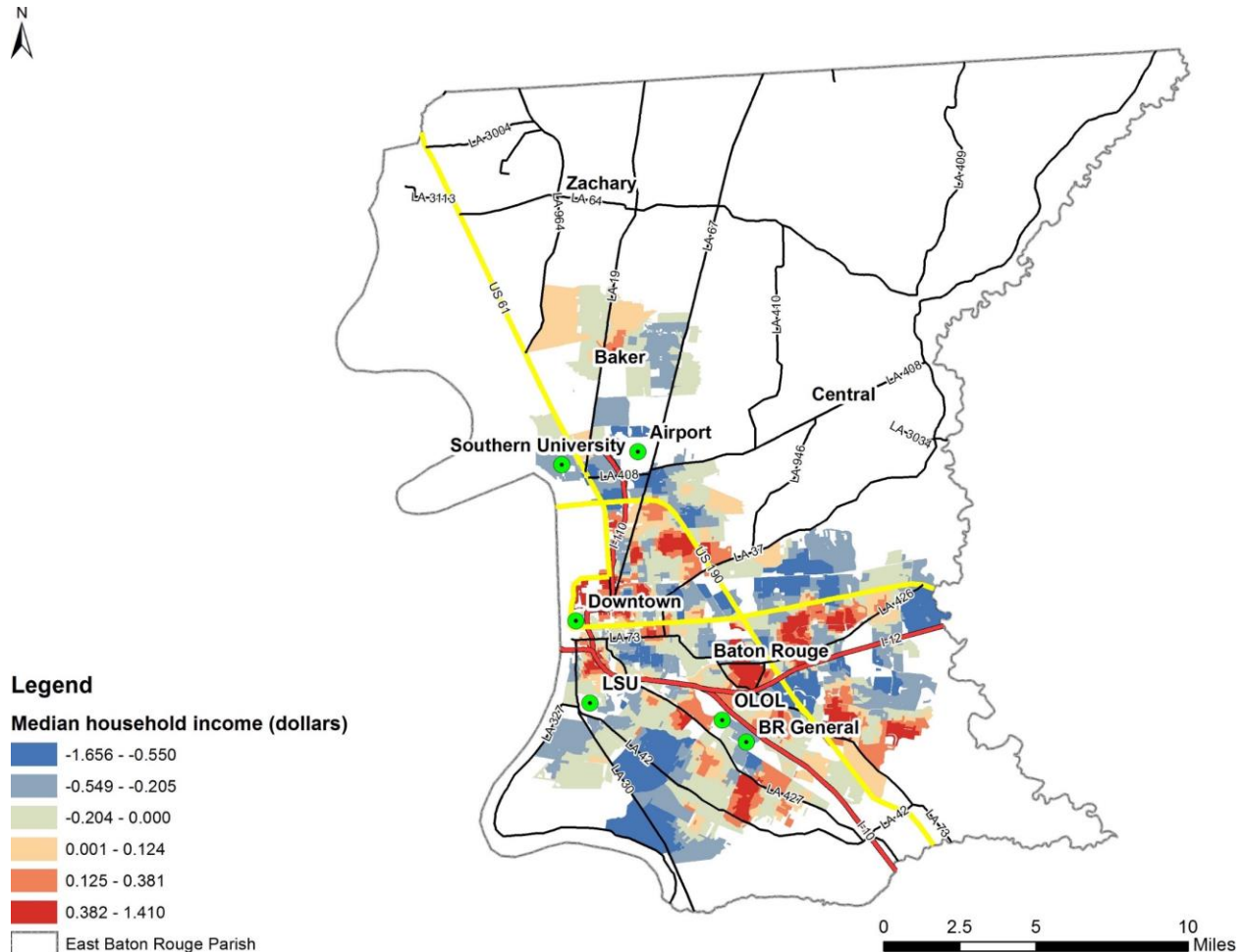

Figure 21. Coefficient estimates for median household income

As previously discussed, commuting time difference between riding public transit and driving private vehicle is perhaps one of the most important factors in commuting mode choice. Figure 22 reports the spatial distribution patterns for the parameter estimate of commuting time ratio. The mean and median are -0.831 and -0.744, respectively, which coherent with the global estimate. An easy conclusion could be reached that the greater commuting time difference between public transit and private vehicle, the more likely that commuters would be discouraged

62

to use public transit system. The distribution of local coefficient estimates shows perfect

consistency with the global trend, albeit some variations in degree, that for all census blocks,

extended commuting time via public transit system makes it a less preferable option for
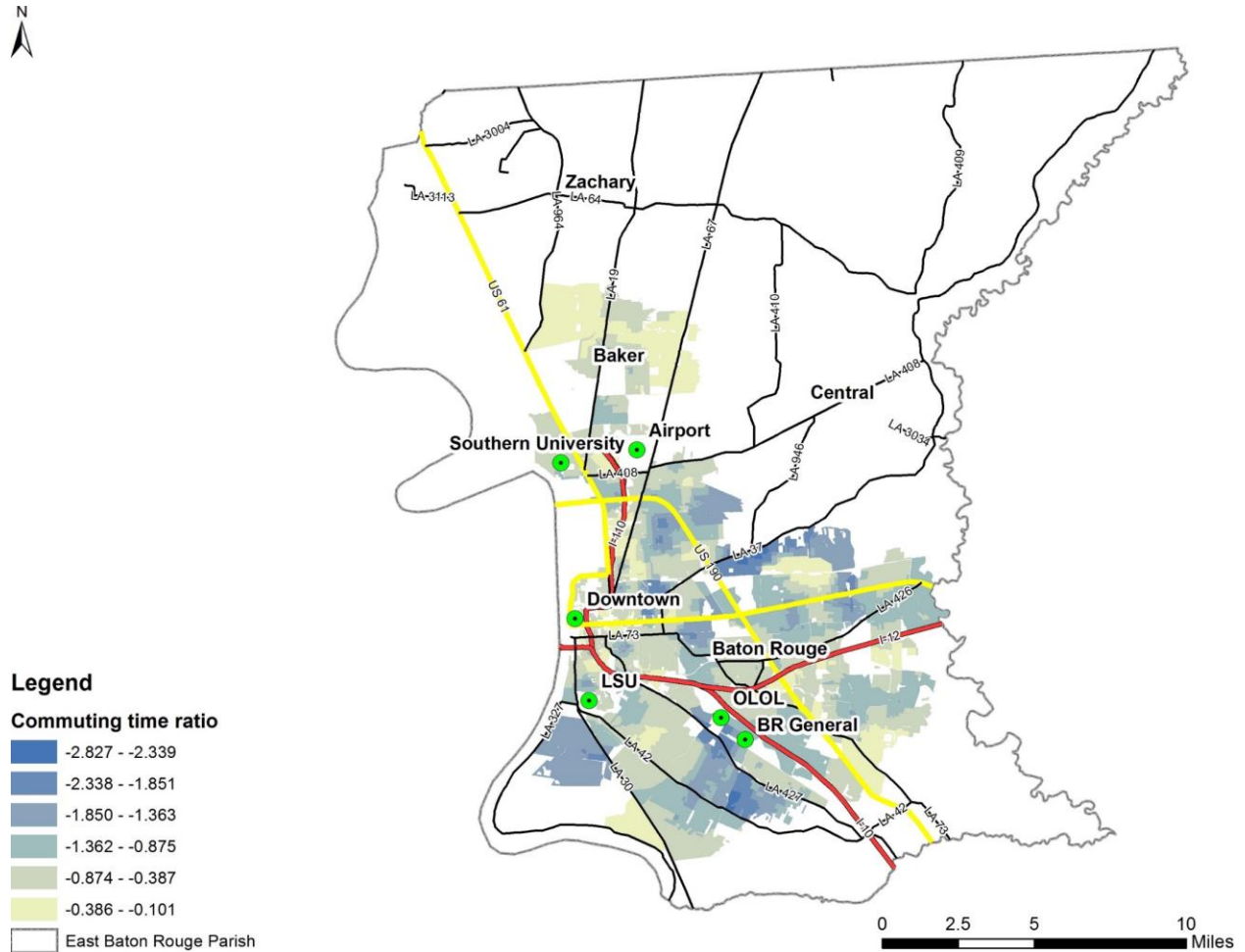
commute.

Figure 22. Coefficient estimates for commuting time ratio

Number of jobs within a 1-hour transit service area of census block measures

accessibility to employment via public transit. Figure 23 maps the spatial distribution patterns for

this factor. The mean and median are 0.173 and 0.119, respectively, which is similar to the

global model. From the map, most of the study area also have positive coefficient estimates,

meaning that for most neighborhood (58.6%), the more jobs are connected by public transit

system, the more commuters would choose public transit. Most area in Baker is again the most

notable exception, along with the area east of LSU, and the neighborhoods at the east and south
sides of Baton Rouge. In these regions, more jobs accessible via public transit system might
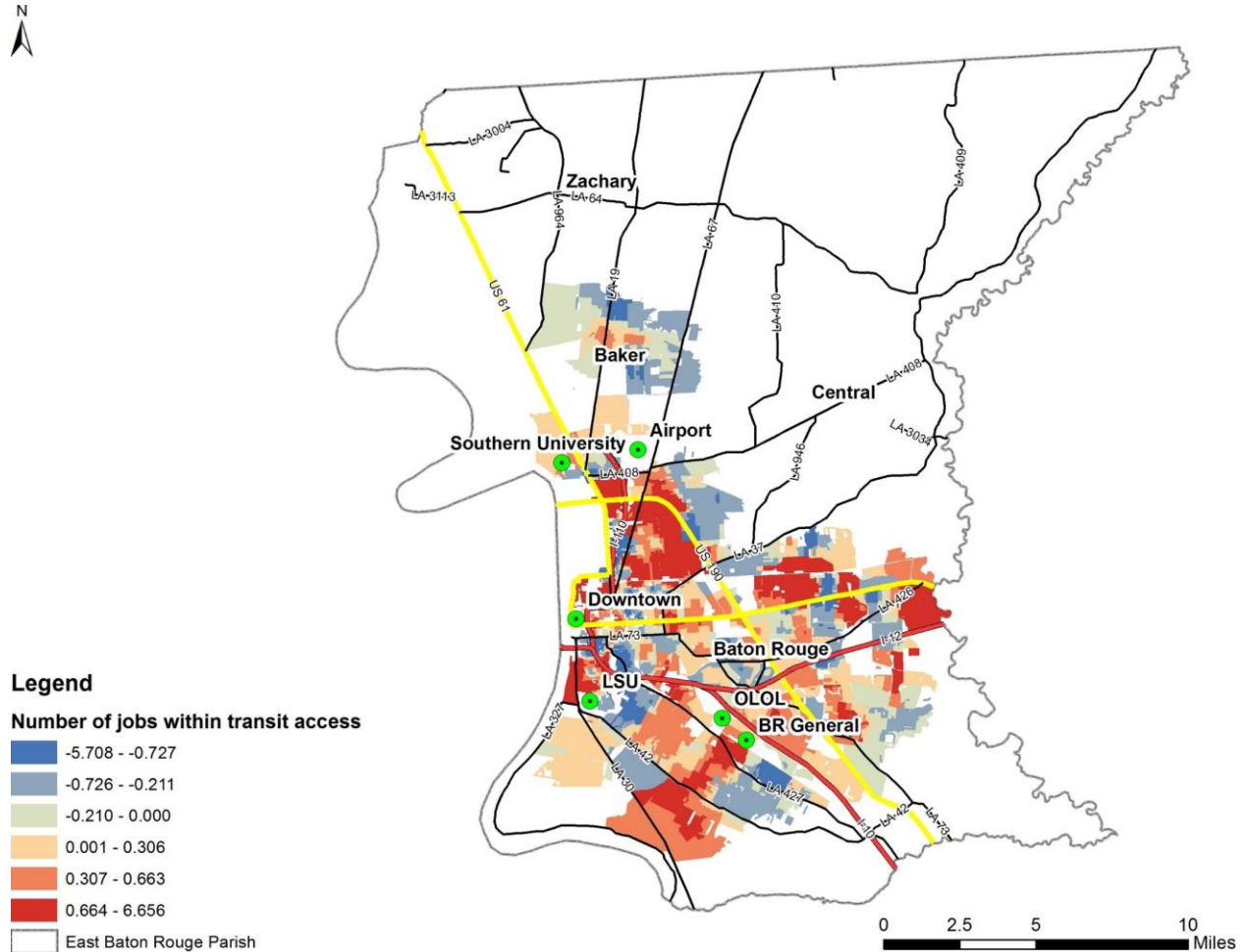instead indicate lower public transit ridership.

Figure 23. Coefficient estimates for number of jobs within 1-hour transit service area

From the standardized parameter estimates reported in Table 9, it could be concluded that
commuting time ratio has the greatest impact on public transit ridership (mean: -0,831, median: -
0.744) – about four times greater than the second-place factor number of jobs within a 1-hour
transit service area. Figure 24 displays where the three localized factors have the greatest impact
in the study area. Commuting time ratio is the most dominant localized factor in more than 60%
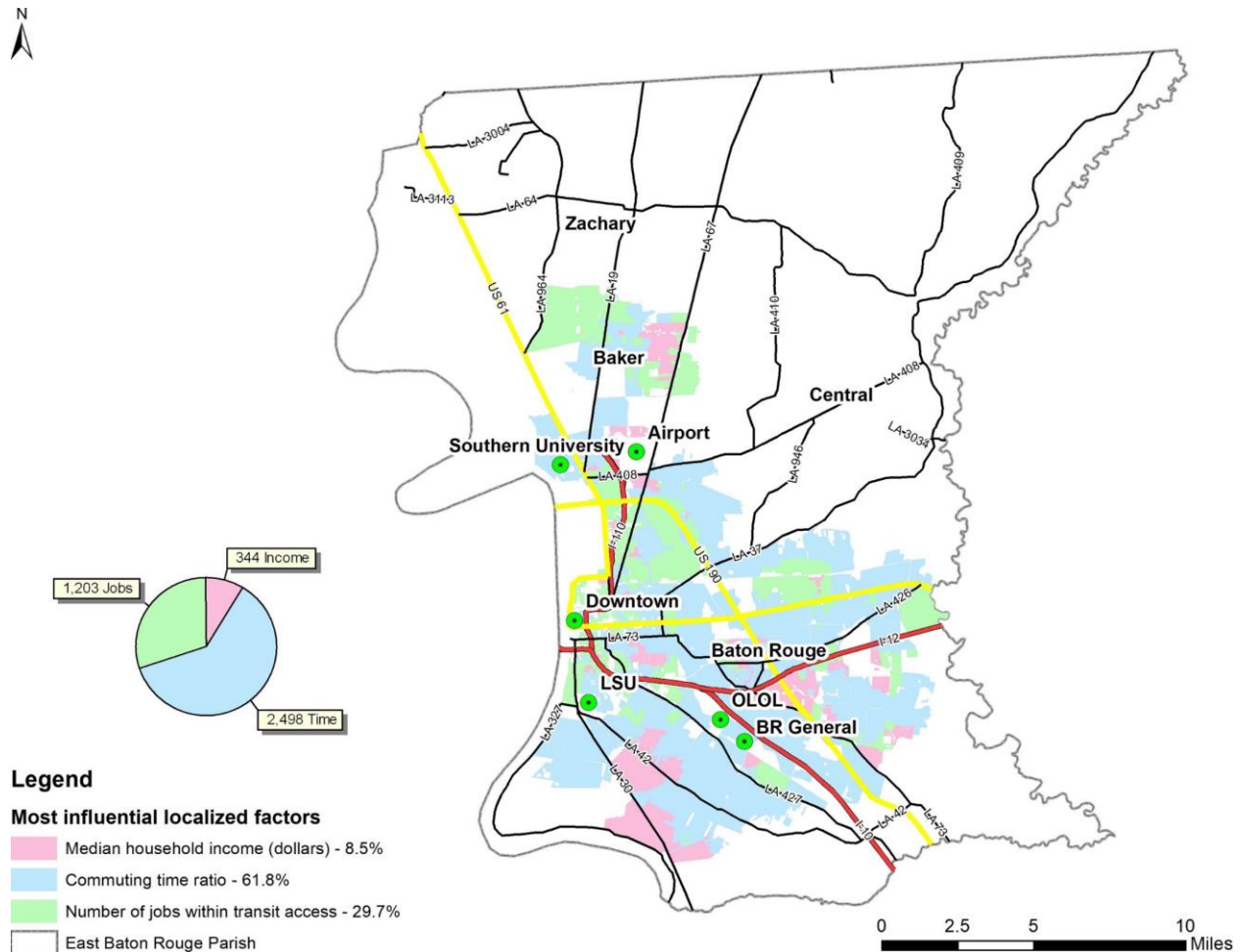of the study area.

Figure 24. Most influential localized factors

This chapter uses regression models to analyze the relationships between public transit ridership and a wide spectrum of demographic, socio-economic and spatial factors that represent different aspects of census block. Only a certain number of factors are considered to have statistically significant correlation to public transit ridership in census block. Some factors are proven to have generally the same correlations with public transit ridership across the study area (i.e., global/homogenous variables), while others' relationships are more localized (i.e., localized/heterogenous variables). Comparing the global terms to the global model, and the localized terms to the localized model, the SGWR model shows consistent findings. Although

the findings are similar, the SGWR model should still be considered an advancement from the previous two models for the following reasons:

1. The model clearly differentiates factors that have global effect on public transit ridership, and ones that have localized effect. The model could potentially help the policy makers and transit operators to target where and how public transit ridership could be improved, and implement strategies for different kinds of commuters in different places.

2. SGWR improves model performance in terms of increasing the model fit, and minimizing the information loss.

CHAPTER 8. CONCLUSIONS AND DISCUSSION

This chapter summarizes major findings from this study, highlights significant contributions and discusses possible extensions for future work. This dissertation attempts to examine how public transit ridership can be explained by neighborhood characteristics such as demographic, socio-economic, and spatial variables. The case study is based on East Baton Rouge Parish, Louisiana in 2013.

8.1. Major findings and contributions

This section provides a brief recap of *four major tasks* accomplished, corresponding to Chapters 4-7, respectively:

1. A disaggregation method is used to interpolate demographic and socio-economic variables from a larger areal unit (block group) to a smaller areal unit (census block), and therefore integrate variables at two different scales into one (i.e., census block).

2. The integration of high-quality resident worker and employment data and land use data with the CTPP data enables more accurate simulations of resident worker and employment locations, and hence individual commuting trips. Improvement in simulations is validated by the recorded traffic count data. Better simulation of commuting trip leads to more accurate measurement of travel time.

3. In addition to calibrating travel time by private vehicle in GIS, this study implements GTFS model into GIS to compute travel time via public transit system based on transit schedule.

4. Variables prepared from the three tasks above are fed into a series of regression models to identify neighborhood effects of demographic, socio-economic and spatial

variables on public transit ridership. The regression analysis advances from a stepwise regression to extract effective global explanatory variables, to a GWR to detect localized effects, and finally to a SGWR to separate global and localized effects.

Some *major findings* are highlighted below:

First, a number of variables have constantly been identified by the existing literature to exert strong influence on public transit ridership, but fail to be validated by this study. These include: age, population or housing density, presence of sidewalks, and residential land use. A likely reason for such discrepancies in findings is that this study examines the relationship in neighborhoods (here census blocks), whereas the traditional intra-urban studies focus on how public transit ridership is affected by these factors in surrounding areas of transit stops. Take population density for example, a traditional intra-urban study may explain public transit ridership at a transit stop by population density within its 1-mile service area. Obviously a higher population density means more population (within the same 1-mile area), and thus more transit riders there. However, as in this study, a densely populated area does not necessarily generate a higher percentage of public transit commuters. Another possible cause is multicollinearity, as discussed in section 7.1. One variable's statistical significance in a regression model may be spurious if this variable is highly correlated with other truly influential variables in the model. For this reason, this study has made a conscious effort to extract a large number of variables. Since some of them are related, the stepwise regression is used to retain the significant variables while filtering out the lesser ones. For example, median household income is retained over per capita income, poverty rate, percentage of renter-occupied housing units, and housing conditions.

For the retained variables, the final model is able to sort out which effects are global and which are localized. For example, commuters in socio-economically disadvantageous neighborhoods tend to be more likely to ride public transit to work (i.e., non-White concentrated, linguistically isolated, inferior housing conditions, etc.), and such effects are constant across the study area. This is consistent with most of existing studies. More importantly, this study shows that overall a lower income level tends be positively related to a higher public transit ridership, but the relationship is far from uniform and can be even reversed in certain regions of the study area. For instance, neighborhoods in extreme poverty may experience a very low labor participation rate, which may be a result of poor accessibility to public transit (say, away from transit routes or lack of transit fare affordability). In this case, low transit ridership is both an outcome and a contributor of poverty. The effects of spatial factors on public transit ridership are all localized. Greater disadvantage of public transit in commuting time pushes commuters away to driving private vehicle. The effect of commuting time ratio is significant, and is more significant in certain areas than others. Neighborhoods that are better connected to jobs by public transit might have more commuters willing to commute by public transit, but the entirely opposite can be found elsewhere in the study area. More in-depth field work is needed to uncover the underlying dynamics in specific parts of the study area. Commuting time ratio has the strongest effect on public transit ridership by a wide margin, and dwarfs the effect by demographic and socio-economic factors. This has significant implication in service planning and policy making: improving public transit efficiency and reliability is vital for promoting public transit ridership.

On the *methodological* front, this study makes several notable *contributions*.

69

1. It considers both spatial and non-spatial factors at census block level to explain public transit ridership. While most non-spatial factors have been considered by existing studies, this study disaggregates these data to facilitate analysis at a smaller areal unit (census block) is the first of kind in applying ecological inference method in transportation-related studies.

2. It defines multiple spatial factors, among which the calibration of transit-to-driving time ratio is considered a major strength of this study. The pursuit for more accurate measurement is enabled by access to better data and more sophisticated simulation techniques powered by geospatial computation in a GIS environment.

3. This study taps into the advanced SGWR model to finalize the regression analysis. While the SGWR method was developed over a decade ago, its implementation and availability for public usage was fairly recent. As discussed under "major findings", it enables this study to sort out what are global effects and what are localized effects, and where the localized effects are. Such findings may help policy makers and transit operators to target where and how public transit ridership might be needed and improved, and implement different strategies for different kinds of commuters in different places. In short, when it comes to public policy, it rejects the notion that "one size fits all" and promotes "place-adaptable policy".

8.2. Limitations and possible future improvements

First of all, this study does not focus on individual behaviors due to the nature of aggregated data in areal units.

This study area, Baton Rouge, is a medium-size U.S. city where only a very small portion of commuters use public transit system. Small values of transit ridership make its variability sensitive to data and measurement errors. This may limit the replicability of the findings in other cities.

This study uses the aforementioned data disaggregation method to simulate variable values in a smaller areal unit from data in a larger areal unit. The data disaggregation method makes certain assumptions that may require additional justification (e.g., one unknown variable is linearly correlated to one known variable). Future studies can use other data sources for improving the estimation. For trip simulation, this study uses land use data to improve the simulation of trip ends. Future work can further advance the simulation by utilizing parcel data, building footprint data, and others.

For driving time by private vehicle, this study simply assumes that drivers follow speed limits. Future study can use other methods (e.g., Google Maps API) to derive more realistic travel time that accounts for traffic conditions. On measuring riding time on public transit, this study relies on a fixed transit schedule. However, delays are also common for public transits in the study area. This study also omits LSU's own shuttle service, Tiger Trails, available for anyone on and around the campus. However, it does not operate on a fixed schedule, and therefore one cannot use the GTFS data adopted by this study to estimate travel time.

REFERENCES

Alshalalfah, BW, and Amer Shalaby. 2007. "Case Study: Relationship of Walk Access Distance to Transit with Service, Travel, and Personal Characteristics." Journal of Urban Planning and Development 133 (2): 114–18.

Anselin, Luc, and Wendy K Tam Cho. 2002. "Spatial Effects and Ecological Inference." Political Analysis 10 (3): 276–97.

Antipova, Anzhelika, Fahui Wang, and Chester Wilmot. 2011. Urban Land Uses, Socio-Demographic Attributes and Commuting: A Multilevel Modeling Approach. Vol. 31.

Antrim, Aaron, and Sean Barbeau. 2013. "The Many Uses of GTFS Data–opening the Door to Transit and Multimodal Applications." Tampa, FL: Location-Aware Information Systems Laboratory at the University of South Florida.

Armbruster, Brendan. 2010. "Factors Affecting Transit Ridership at the Metropolitan Level 2002-2007." Master's thesis, Georgetown University.

Atkinson-Palombo, Carol, and Michael J Kuby. 2011. "The Geography of Advance Transit-Oriented Development in Metropolitan Phoenix, Arizona, 2000–2007." Journal of Transport Geography 19 (2): 189–99.

Babalik-Sutcliffe, Ela. 2002. "Urban Rail Systems: Analysis of the Factors behind Success." Transport Reviews 22 (4): 415–47.

Belzer, Dena, Gerald Autler, Judith Espinosa, Sharon Feigon, and Gloria Ohland. 2004. "The Transit-Oriented Development Drama and Its Actors." The New Transit Town: Best Practices in Transit-Oriented Development, 42–55.

Berkley, Jennifer. 2017. "Using American Community Survey Estimates and Margin of Error." Washington, DC: U.S. Census Bureau.

Bernick, Michael, and Robert Cervero. 1997. Transit Villages in the 21st Century.

Bianco, Martha, Kenneth Dueker, and James Strathman. 1998. Parking Strategies to Attract Auto Users to Transit. Transportation Research Board.

Blumenberg, Evelyn, and Kimiko Shiki. 2007. "Transportation Assimilation: Immigrants, Race and Ethnicity, and Mode Choice."

Calthorpe, Peter. 1993. The next American Metropolis: Ecology, Community, and the American Dream. Princeton architectural press.

Cervero, Robert, and Samuel Seskin. 1995. "An Evaluation of the Relationships between Transit and Urban Form." TCRP Research Results Digest, no. 7.

Chen, Cynthia, Don Varley, and Jason Chen. 2011. "What Affects Transit Ridership? A Dynamic Analysis Involving Multiple Factors, Lags and Asymmetric Behaviour." Urban Studies 48 (9): 1893–1908.

Chung, Kyusuk. 1997. "Estimating the Effects of Employment, Development Level, and Parking Availability on CTA Rapid Transit Ridership: From 1976 to 1995 in Chicago." Paper presented at Metropolitan Conference on Public Transportation Research, Chicago, IL, June 6.

Crane, Randall. 2000. "The Influence of Urban Form on Travel: An Interpretive Review." CPL Bibliography 15 (1): 3–23.

Crowley, David, Amer Shalaby, and Hossein Zarei. 2009. "Access Walking Distance, Transit Use, and Transit-Oriented Development in North York City Center, Toronto, Canada." Transportation Research Record: Journal of the Transportation Research Board, no. 2110: 96–105.

Dijkstra, Edsger. 1959. "A Note on Two Problems in Connexion with Graphs." Numerische Mathematik 1 (1): 269–71.

Dykstra, Robert. 1986. "Ecological Regression Estimates: Alchemist's Gold?" Social Science History 10 (1): 85–86.

Ewing, Reid, and Robert Cervero. 2010. "Travel and the Built Environment: A Meta-Analysis." Journal of the American Planning Association 76 (3): 265–94.

Farber, Steven, Melinda Morang, and Michael Widener. 2014. "Temporal Variability in Transit-Based Accessibility to Supermarkets." Applied Geography 53: 149–59.

Flanigan, William, and Nancy Zingale. 1985. "Alchemist's Gold: Inferring Individual Relationships from Aggregate Data." Social Science History 9 (1): 71–91.

Fotheringham, Stewart, Chris Brunsdon, and Martin Charlton. 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester: Wiley.

Frank, Lawrence, Mark Bradley, Sarah Kavage, James Chapman, and Keith Lawton. 2008. "Urban Form, Travel Time, and Cost Relationships with Tour Complexity and Mode Choice." Transportation 35 (1): 37–54.

Gomez-Ibanez, Jose. 1996. "Big-City Transit Rider Snip, Deficits, and Politics: Avoiding Reality in Boston." Journal of the American Planning Association 62 (1): 30–50.

Guerra, Erick, Robert Cervero, and Daniel Tischler. 2012. "Half-Mile Circle: Does It Best Represent Transit Station Catchments?" Transportation Research Record: Journal of the Transportation Research Board, no. 2276: 101–9.

Gutierrez, Javier, Osvaldo Daniel Cardozo, and Juan Carlos García-Palomares. 2011. "Transit Ridership Forecasting at Station Level: An Approach Based on Distance-Decay Weighted Regression." Journal of Transport Geography 19 (6): 1081–92.

Hartgen, David T, and Mark Horner. 1997. Comparative Performance of Major US Bus Transit Systems, 1988-1995. Charlotte, NC: Center for Interdisciplinary Transportation Studies, University of North Carolina at Charlotte.

Hendrickson, Chris. 1986. "A Note on Trends in Transit Commuting in the United States Relating to Employment in the Central Business District." Transportation Research Part A: General 20 (1): 33–37.

Hoehner, Christine, Carolyn Barlow, Peg Allen, and Mario Schootman. 2012. "Commuting Distance, Cardiorespiratory Fitness, and Metabolic Risk." American Journal of Preventive Medicine 42 (6): 571–78.

Hu, Yujie, and Fahui Wang. 2015. Monte Carlo method and its application in urban traffic simulation. In: Fahui Wang (ed.) *Quantitative Methods and Socio-Economic Applications in GIS*. Boca Raton, FL: CRC Press, pp. 259–277

Hu, Yujie, and Fahui Wang. 2016. "Temporal Trends of Intraurban Commuting in Baton Rouge, 1990–2010." Annals of the American Association of Geographers 106 (2):470–79.

Kain, John, and Zvi Liu. 1999. "Secrets of Success: Assessing the Large Increases in Transit Ridership Achieved by Houston and San Diego Transit Providers." Transportation Research Part A: Policy and Practice 33 (7): 601–24.

King, Gary. 1997. A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton, NJ: Princeton University Press.

Kitamura, Ryuichi. 1989. "A Causal Analysis of Car Ownership and Transit Use." Transportation 16 (2): 155–73.

Kuai, Xuan, and Qunshan Zhao. 2017. "Examining Healthy Food Accessibility and Disparity in Baton Rouge, Louisiana." Annals of GIS 23 (2): 1–14.

Kuby, Michael, Anthony Barranda, and Christopher Upchurch. 2004. "Factors Influencing Light-Rail Station Boardings in the United States." Transportation Research Part A: Policy and Practice 38 (3): 223–47.

Limtanakool, Narisra, Martin Dijst, and Tim Schwanen. 2006. "The Influence of Socioeconomic Characteristics, Land Use and Travel Time Considerations on Mode Choice for Medium- and Longer-Distance Trips." Journal of Transport Geography 14 (5): 327–41.

Litman, Todd. 2005. "Terrorism, Transit and Public Safety Evaluating the Risks." Journal of Public Transportation 8 (4): 33–45.

Liu, Zhi. 1993. "Determinants of Public Transit Ridership: Analysis of Post-World War II Trends and Evaluation of Alternative Networks." PhD diss., Havard University.

Mashhoodi, Bardia, Dominic Stead, and Arjan van Timmeren. 2019. "Spatial Homogeneity and Heterogeneity of Energy Poverty: A Neglected Dimension." Annals of GIS 25 (1): 19–31.

Nakaya, Tomoki, Stewart Fotheringham, Martin Charlton, and Chris Brunsdon. 2009. "Semiparametric Geographically Weighted Generalised Linear Modelling in GWR 4.0."

Neilson, Gordon, and William Fowler. 1972. "Relation between Transit Ridership and Walking Distances in a Low-Density Florida Retirement Area." Highway Research Record, no. 403.

Pushkarev, Boris, and Jeffrey Zupan. 1977. Public Transportation and Land Use Policy. Bloomington, IN: Indiana Univ Press.

Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." American Sociological Review 15 (3): 351–57.

Santos, Adella, Nancy McGuckin, Hikari Yukiko Nakamoto, Danielle Gray, and Susan Liss. 2011. "Summary of Travel Trends: 2009 National Household Travel Survey." Washington, DC: U.S. Department of Transportation.

Schrank, David, Bill Eisele, Tim Lomax, and Jim Bak. 2015. "2015 Urban Mobility Scorecard." College Station, TX: Texas A&M Transportation Institute.

Sharaby, Nir, and Yoram Shiftan. 2012. "The Impact of Fare Integration on Travel Behavior and Transit Ridership." Transport Policy 21: 63–70.

Simonson, Ernst, Charles Baker, Neal Burns, Charles Keiper, Otto Schmitt, and Stirling Stackhouse. 1968. "Cardiovascular Stress (Electrocardiographic Changes) Produced by Driving an Automobile." American Heart Journal 75 (1): 125–35.

Spillar, Robert. 1989. "The Effects of Population Density, Income, and Transit Coverage Per Capita Transit Ridership in Western American Cities." PhD diss., Seattle, WA: University of Washington.

Spurr, Tim, Robert Chapleau, and Daniel Piché. 2014. "Use of Subway Smart Card Transactions for the Discovery and Partial Correction of Travel Survey Bias." Transportation Research Record: Journal of the Transportation Research Board, no. 2405: 57–67.

Stopher, Peter. 2004. "Reducing Road Congestion: A Reality Check." Transport Policy 11 (2): 117–31.

Tallis, Federico. 2014. "Evaluating Transit and Driving Disaggregated Commutes through GTFS in ArcGIS." Master's thesis, University of Southern California.

Taylor, Brian, and Camille Fink. 2003. "The Factors Influencing Transit Ridership: A Review and Analysis of the Ridership Literature." Berkeley, CA: University of California Transportation Center.

Taylor, Brian, and Camille Fink. 2013. "Explaining Transit Ridership: What Has the Evidence Shown?" Transportation Letters 5 (1): 15–26.

Taylor, Brian, Douglas Miller, Hiroyuki Iseki, and Camille Fink. 2009. "Nature And/Or Nurture? Analyzing the Determinants of Transit Ridership across US Urbanized Areas." Transportation Research Part A 43: 60–77.

TRB (Transportation Research Board). 1998. "Strategies to Attract Auto Users to Public Transportation." Washington, DC: National Research Council.

Wang, Fahui. 2015. Quantitative Methods and Socio-Economic Applications in GIS. Boca Raton, FL: CRC Press.

APPENDIX. SUPPLEMENTARY MAPS
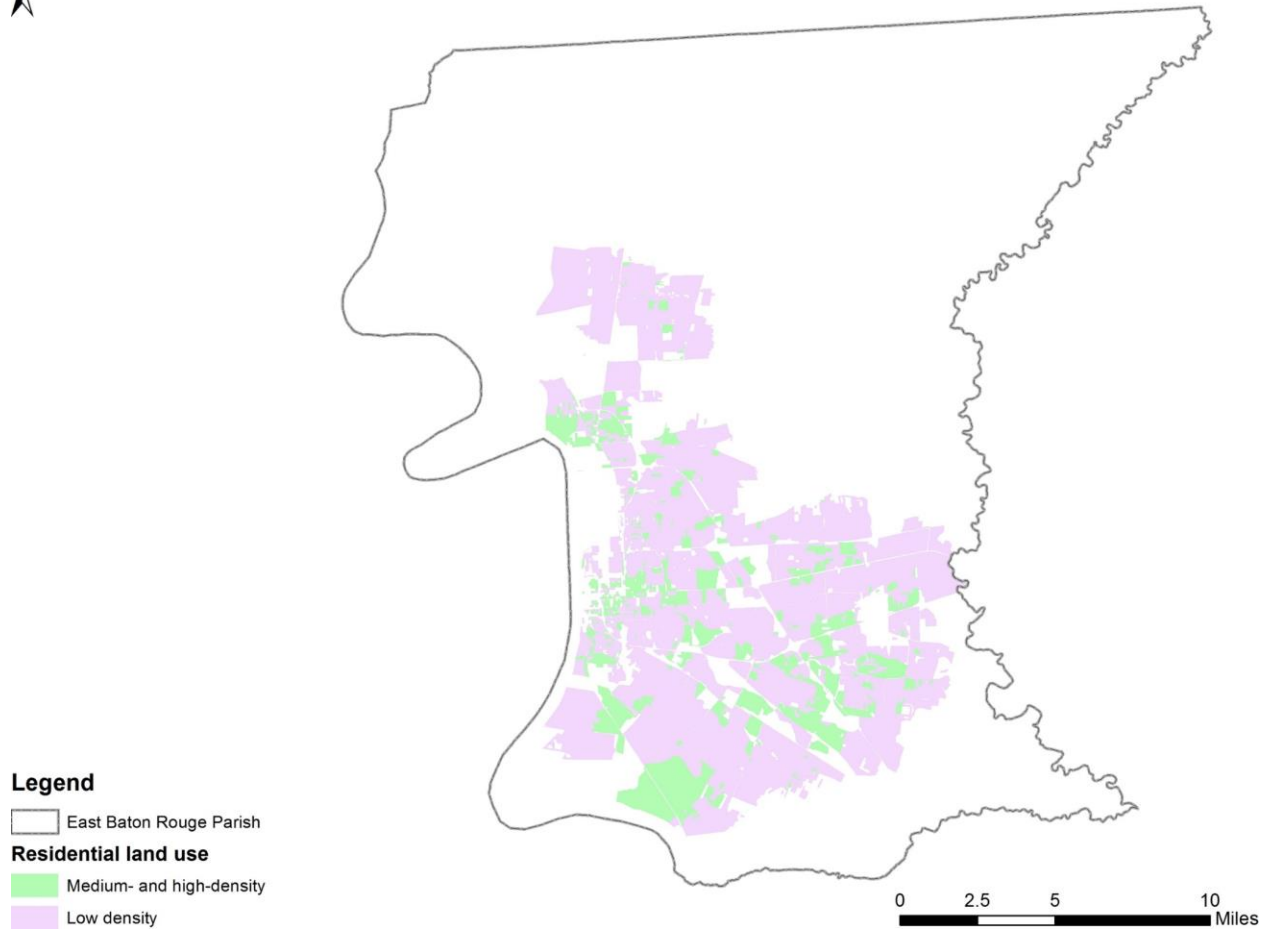
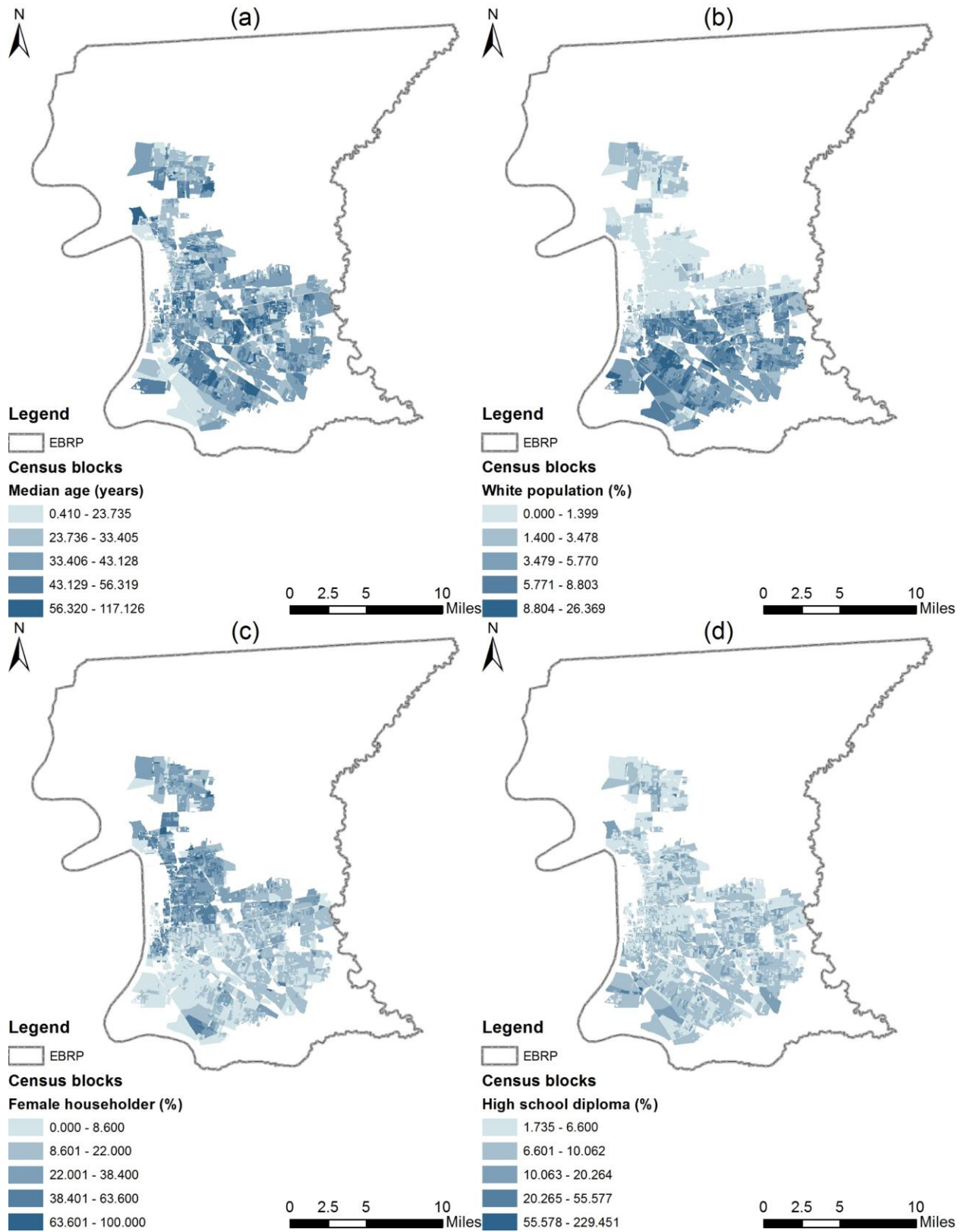Figure A.1. Residential land use: low density, and medium and high density

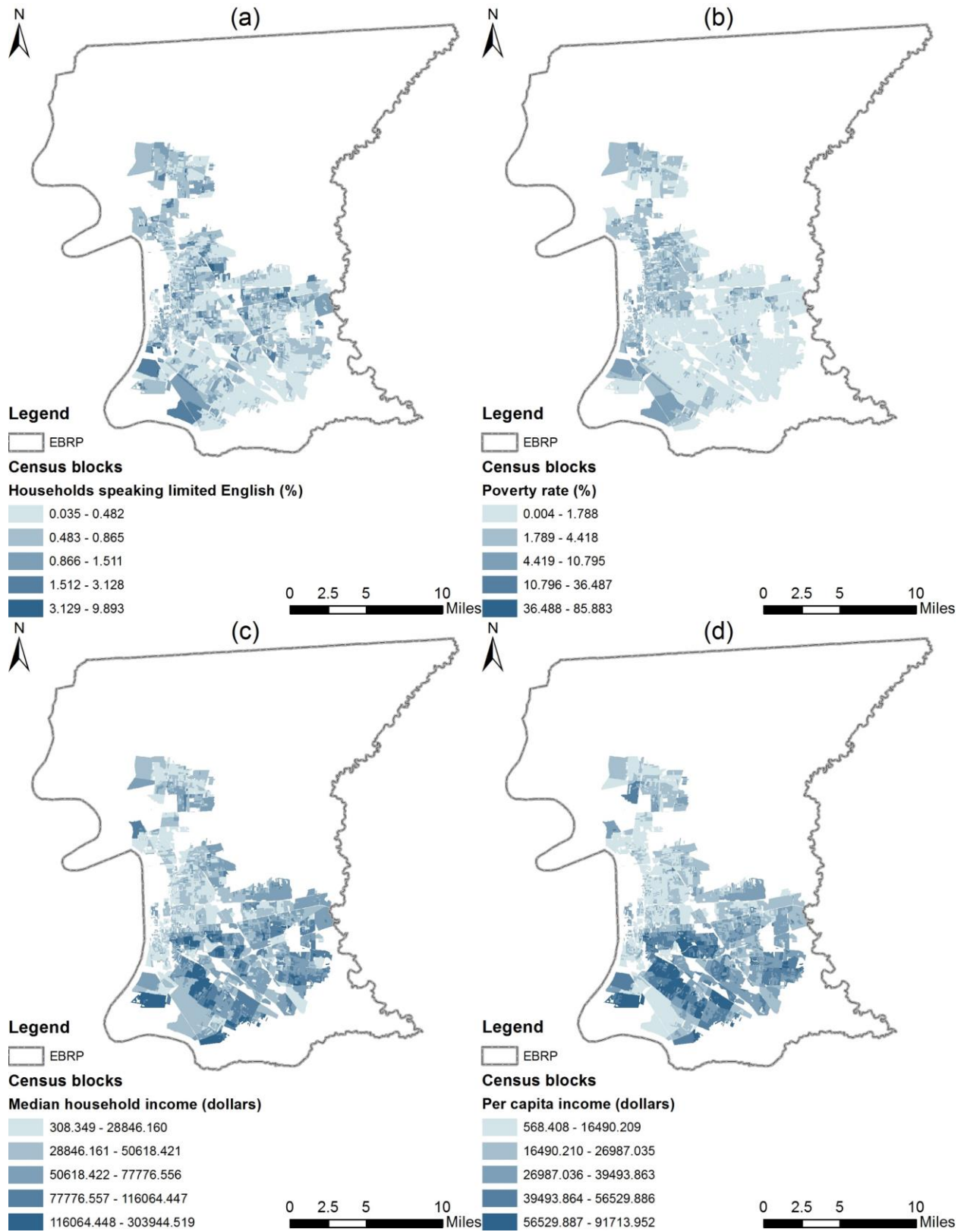Figure A.2. Demographic variables at the census block level

Figure A.3. Socio-economic variables at the census block level
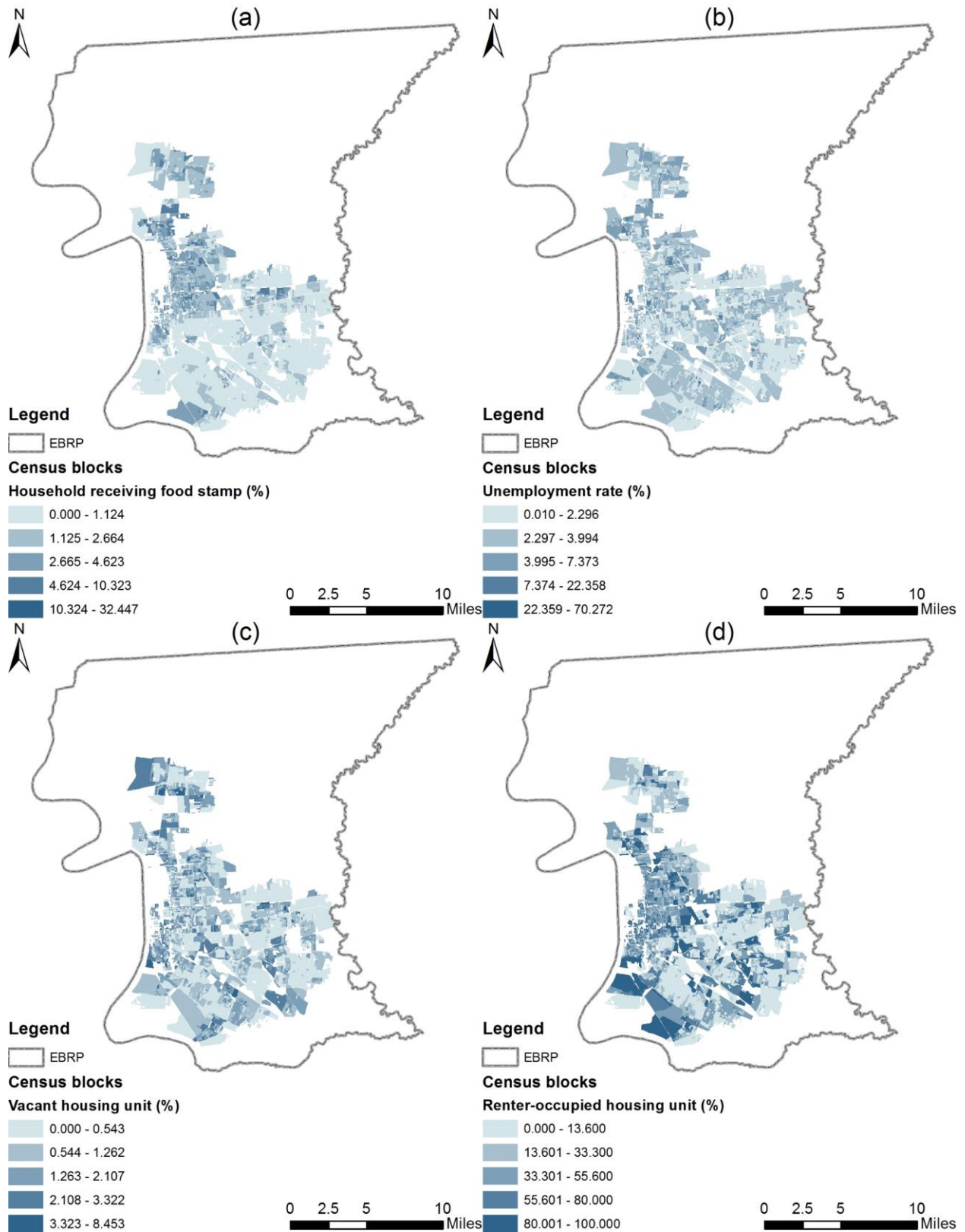
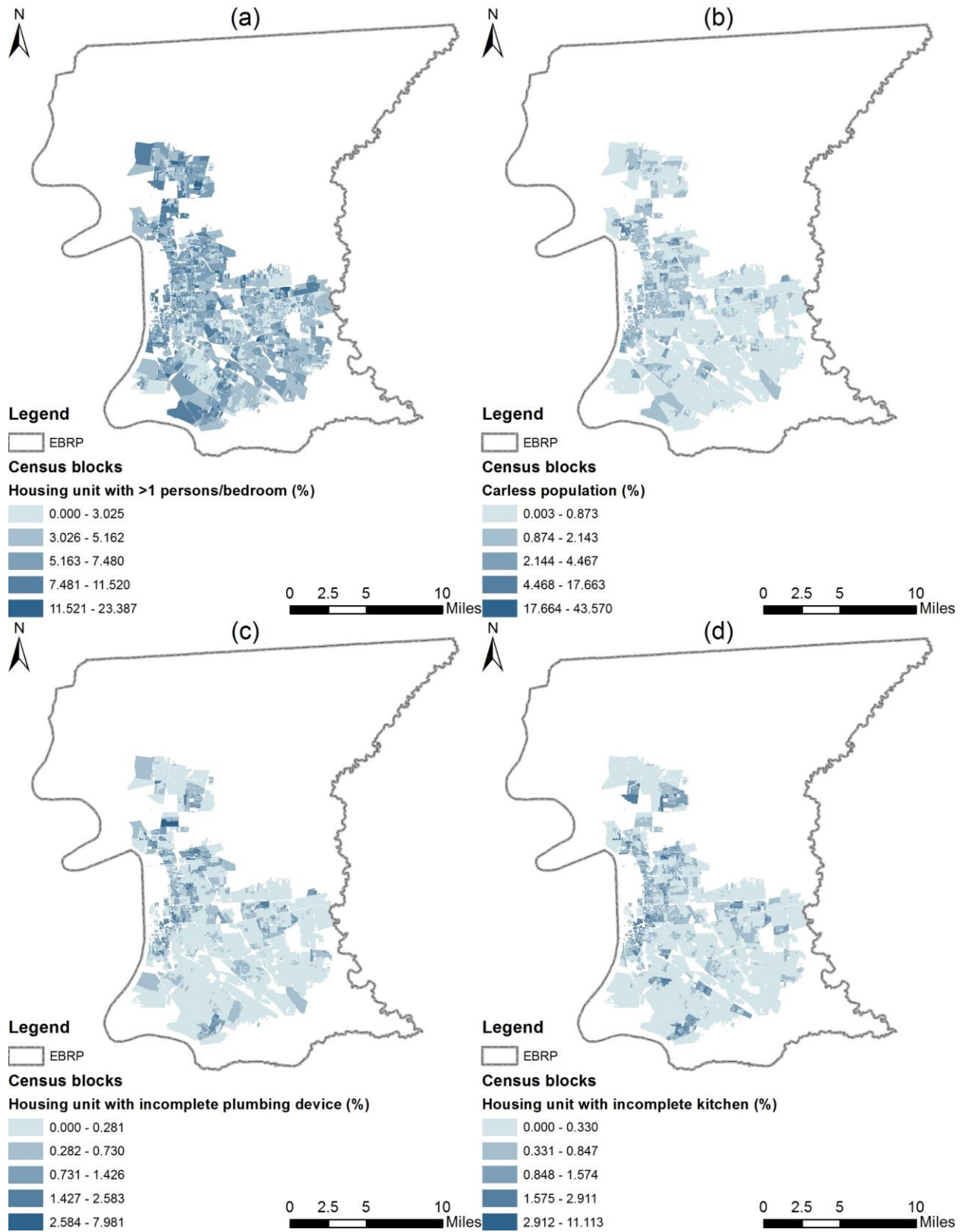Figure A.4. Socio-economic variables at the census block level

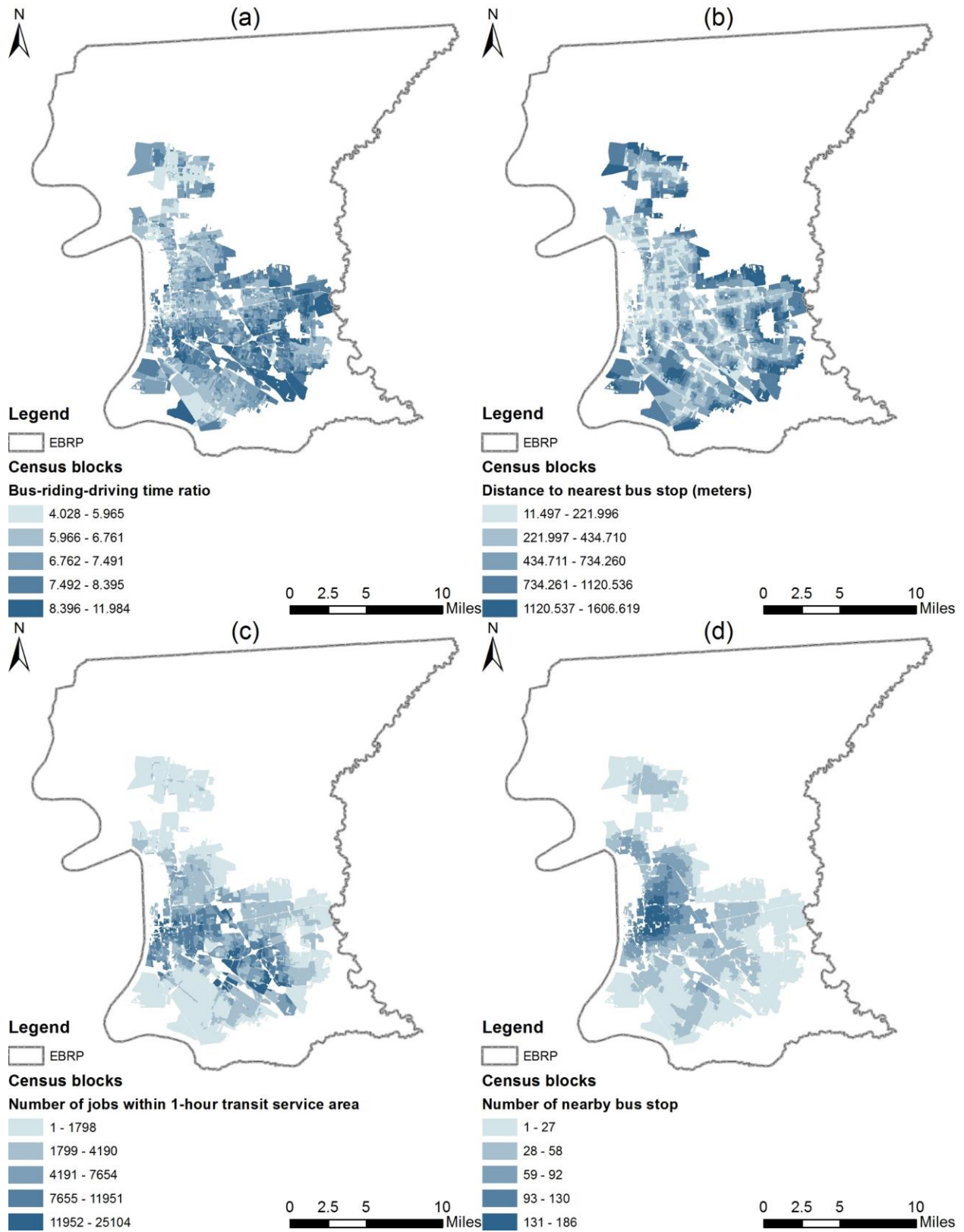Figure A.5. Socio-economic variables at the census block level

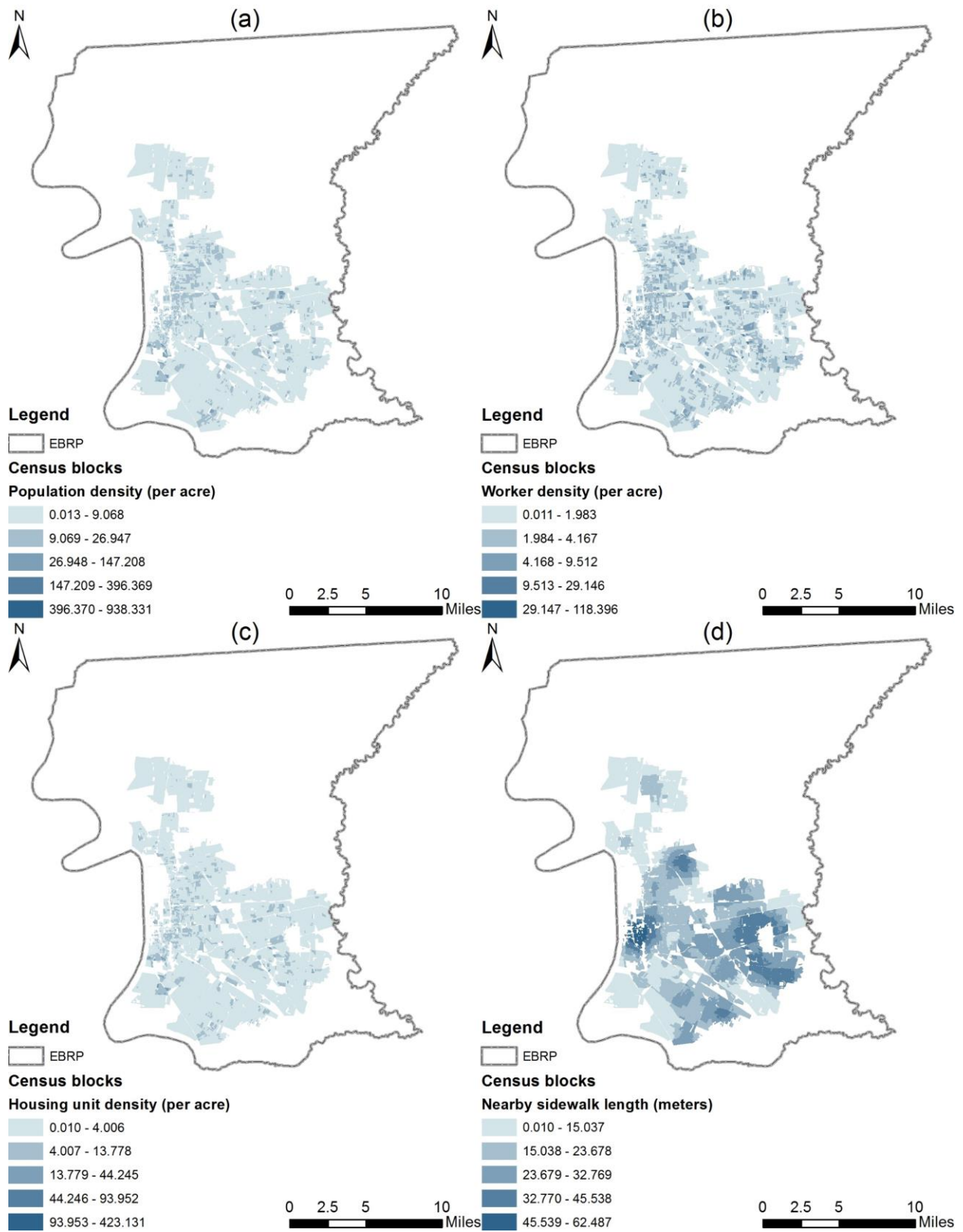Figure A.6. Spatial variables at the census block level

Figure A.7. Spatial variables at the census block level

VITA

Xuan Kuai was born in Wuhan, China in 1991. His research interests include geographic information science, urban geography, and transportation geography. He earned his Bachelor of Engineering degree in geographic information science from Wuhan University, China in 2013, and his Master of Science degree in geography from Louisiana State University (LSU) in 2015. Since 2016, he has been a Doctor of Philosophy student in geography at LSU while working at the City of Baton Rouge and East Baton Rouge Parish (2016 – 2018) and at the Houston-Galveston Area Council (2018 – current).