9-14-2015

Spatial and Temporal Measures of Mismatch between Transit Supply and Employment for Low Income and Auto Dependent Populations

Sina Kahrobaei
University of Connecticut - Storrs, sina.kahrobaei@uconn.edu

Spatial and Temporal Measures of Mismatch between Transit Supply and Employment for Low Income and Auto Dependent Populations

Sina Kahrobaei

MCRP, University of Texas at Arlington, 2012

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

At the

University of Connecticut

2015

APPROVAL PAGE

Master of Science Thesis

Spatial and Temporal Measures of Mismatch between Transit Supply and Employment for Low Income and Auto Dependent Populations

Presented by

Sina Kahrobaei, MCRP

Major Advisor_____
Nicholas E. Lownes

Associate Advisor_____
Karthik C. Konduri

Associate Advisor_____
Amy C. Burnicki

University of Connecticut

2015

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor, Dr. Nicholas Lownes, for providing me with all the necessary guidance to complete this research. I am extremely thankful and indebted to him for his help, patience, and motivation. I also would like to express my sincere thanks to Dr. Karthik Konduri for sharing expertise and valuable guidance on this research, especially to work with PopGen model. My sincere thanks also go to Dr. Amy Burnicki for her help and knowledge share.

I would like to thank my friend and colleague Kelly Bertolaccini for her help working with GTFS data. I also thank my friends Abed Ghanbari, Hamed Ahangari, Hojjat Seyyedi, Farzin Maniei, Doray Hill Jr., Mahdis Ahmaripour, Wendy Medina, and Amir Erfanian for their continued support and help.

Finally, my sincere gratitude to my family for their support. I am especially grateful to my mother, Marzieh Shahoseini, for supporting me spiritually throughout writing this thesis and my life in general.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

PROBLEM CONTEXT

In the 19th century, the morphology of the U.S. cities was mono-centric, small, and dense, typically around a port or a train station. Due to high intra-urban transport costs for people and goods, residence were located near or even within the central business district area. From the second half of the 20th century, however, the emergence of new transport modes coupled with decrease in intra-urban costs enabled people to move to the suburbs (*1*). The inner-city minorities, however, were exceptions in this regard since they could not move to the suburb in spite of suburbanization (*2*).

Population suburbanization was followed by job suburbanization in that labor force continuously attracted jobs to the periphery of the U.S. cities. While jobs, especially low-paid jobs, were moving to the suburb, low-skilled minority workers trapped in inner cities, began to be affected by distant job locations. That is the heart of what is known as spatial mismatch.

Spatial mismatch is a mismatch between where low-income households reside and where the jobs are located, so in essence, spatial mismatch is an imbalance between *demand* of low income population to work and *supply* of jobs. While earlier literature used to measure the spatial mismatch in distance (spatial component) the newer studies measure it in time (temporal component) as well, so spatial mismatch can be seen through both *spatial* and *temporal* lenses.

Transit service could be utilized in narrowing spatial mismatch gap by improving job access for low income households. This improvement in job access might be seen through each of those spatial or temporal lenses. Temporal aspects of transit service tend to look at the

alignment of transit service with the demand and the needs that change over the course of a day. Spatial aspect of transit service is mostly reflected in terms of access to transit stops or routes.

The focus of this research is on the Low Income and High Car Ownership (LIHCO) or Forced Car Ownership (FCO) households, who are involuntary choice low-income families that lack alternative transport options and at the same time have high car ownership (*3*).

SIGNIFICANCE OF PROBLEM

While the spatial mismatch is generally about low income access to jobs, LIHCO population is at the center of this analysis. As location theories suggest, there is a trade-off between transportation costs and land cost. This trade-off creates a continuum with different levels of access to activities and land costs leaving the cheapest lands with limited accessibility at one end and the most expensive lands with high accessibility to goods and services at another. It is believed that LIHCO families live in less expensive areas but farther from the activities so are forced to have more cars to get round. That is the very difference between general Low Income (LI) families and LIHCO families and the reason that puts the later group at the heart of this research. Providing quality transit service to LIHCO families releases the necessity to have more cars and narrows the spatial mismatch gap.

In this regard, attention should be paid to specific needs of low income households, which have different travel patterns from the others. Low income households typically take shorter trips and are less likely to travel during peak periods. They also travel less by any measure than non-low-income persons. Finally, low-income workers are less likely to commute during the a.m. peak (*4*). Two questions that this research intends to answer are:

- Does the local bus service in New Haven County, Connecticut meet the demands for commute during different time intervals (whether question)?
- Where does the local bus service in New Haven County, Connecticut meets the demand for commute during different time intervals (where question)?

CONTRIBUTION OF RESEARCH

Spatial mismatch is essentially a mismatch between employment demands of low income individuals and job supplies. If transit is selected as the tool to narrow this gap, transit service should properly match the needs of low income population. This coordination between job demand and transit supply should incorporate both spatial and temporal components. An example for a spatial type mismatch is a bus rapid transit line that connects high-income neighborhoods to an industrial park with ample low-income jobs. A temporal mismatch in this example would be connecting the low-income areas to the industrial park by BRT but in A.M. peak hours knowing that most low-income riders commute during the mid-day hours.

Several empirical studies investigate the transit accessibility or quality of service (*5, 6, 7, 8, 9, 10, 3, 11, 12, 13, 14*). Among them, some account for both spatial and temporal aspects of both supply and demand sides of transit service (*10, 14, 13, 12, 11, 7, 6, 5*) and, further down, a few explicitly investigate the transit needs and supplies of low income individuals or welfare-recipient population (*10, 13, 12*).

It is noteworthy that the spatial component of spatial mismatch requires applying spatial statistics techniques in examining the match between the areas of population (demand) and the

levels of transit supply. The reason being that both measures of transit demand such as population locations and measures of transit supply such as frequency of transit service are essentially geographic phenomena that tend to correlate in space. For example, it is more likely to have many bus stops with high-frequent service close together in an area than a combination of low-frequent to high-frequent bus services. Applying sole conventional statistics methods might lead to dismissing such interactions in space hence distorting or falsifying the inferences or results (*15*).

Some scholars have stressed the importance of spatial statistics in calculating the levels of transit accessibility. Kawabata and Shen (*16*) utilize spatial regression models and find higher degrees of association between shorter commute times and greater job accessibility for public transit than driving alone. Similarly, Wang and Chen (*17*) use spatial regression models and show locations with higher share of zero-vehicle housing units have better job accessibility by transit. Including income as an equity variable, Griffin and Sener (*18*) apply spatial statistics techniques for equity analysis of transit service in large auto-oriented cities in the U.S. They find that the regions with extensive rail and bus service are most likely to provide low income and all workers served by transit more equitably.

This study is intended to research the levels that transit supply meets the employment demands of worker population in general and Low Income and High Car Ownership (LIHCO) population *(3)* in specific. It identifies LIHCO population as a sub-population of low income households that are forced to own more cars as a consequence of their residential location choices bounded by their limited income. It is assumed that providing quality transit service in

accordance to their spatial and temporal needs will release the pressure for having more cars and grant them at least better employment accessibility.

This research utilizes global and local spatial statistics measures, namely spatial autocorrelation, in conjunction with conventional statistics to investigate whether and where the local bus service in New Haven County, Connecticut meets the demands for commute during different time intervals over a course of a day.

This research adds to the literature in that it combines different spatial global and local measures to identify the temporal and spatial match between transit supply and employment demands of population. It might be useful in that it enables the transit operator to analyze whether its resources are spatially or temporally distributed in accordance with the commute needs of workers in general and LIHCO workers in specific.

CHAPTER 2: REVIEW OF LITERATURE

SPATIAL MISMATCH AND LIHCO HOUSEHOLDS

Spatial Mismatch

Spatial mismatch is defined as a mismatch between where low-income households reside and where jobs are located. The term was originally coined after the work of John Kain (*2*), an economist at Harvard University, on Chicago and Detroit metropolitan areas to investigate the relationship between the housing segregation and non-white employment. He pointed out the huge difference in unemployment rates between black inner-city communities and America as a whole. Kain (*2*) also concluded that jobs moving away coupled with inability of inner-city poor residents to move closer to jobs (primarily because of housing segregation) to be the main reason for racial division in the nation.

On the same year, “National Advisory Commission on Civil Disorders” (also known as Kerner Commission), similarly found geographic disparity in job growth as one of the primary causes for civil disorders *(1)*. While matching black workers to jobs was proposed as the general solution, creating better transportation to connect ghettos to jobs was one of the three proposals.

The literature of the Spatial Mismatch Hypothesis (SMH) bifurcates to those that investigate the validity of the hypothesis and those that propose solutions to the presumed valid spatial mismatch hypothesis. The former has generated mixed and contradicting results both supporting and refuting the existence of spatial mismatch hypothesis (*19, 2, 20*) though the majority of the empirical work provides either strong or moderate support for the hypothesis (*1, 20*). The later mainly suggests three different policy solutions of helping black households to

move to suburbs, incentivizing jobs to move to the central city, or informational and other access solutions to connect people and jobs (*20*).

One aspect that adds more complexity to the normative research about existence of SMH is the timeframe in which it was originally proposed. The second half of the 20th century saw a rapid suburbanization of first people and then jobs (*1*), which gave rise to the intuition of spatial mismatch. However, technological and socio-economical forces have led morphological and demographical changes such as emergence of poly-centric cities and enormous metropolitan areas over time such that seems to challenge SMH's validity.

Reviewing recent literature, however, suggests that SMH remains valid even with the changes in technology and socio-economic trends. In other words, although the spatial pattern of mismatch has been changed, the spatial mismatch still continues to exist (*21*). For example, although the focus of Kain was on African-Americans, several studies show that spatial mismatch is an issue for other ethnical groups (Hispanics, Asians) but only to a lesser extent (*21*). More recent studies show that spatial segregation has shifted from being based on race to being based on income (*22*), or what has been called socioeconomic status or economic class.

The other deviation is that the clear dichotomy between the central city and the suburbs no longer prevails in the nowadays poly-centric and decentralized metropolitan areas (*21*). The low-income job seekers residing in the suburbs, however, face spatial mismatch as well as those in the inner city (*21*).

Low Income and High Car Ownership

The focus of this research is on the Low Income and High Car Ownership (LIHCO) or Forced Car Ownership (FCO) households. These terms were first defined in the relation to the UK rural areas (*23*) and in the Australian literature in context of "transport poverty" (*24*). The concept indicates the low-income families that lack alternative transport options and at the same time have high car ownership thus dedicate a large portion of their income to transportation (*3*). Curie et al. identify households with a weekly income below $AUS 500/week living in outer urban Melbourne and running more than 2 cars as FCO households.

LIHCO is defined against Low Income and No Car Ownership (LINCO) groups, who are simply low income and have no cars. As early location theories suggest, there is a trade-off between transportation cost and land cost. LINCO households locate in inner-city areas, which are more expensive, yet offer more services within walking distances. LIHCO households, however, live in outer less expensive areas but farther from the activities so are forced to have more cars to get round (*3*). In the current poly-centric metropolitan areas, however, dividing between inner-cities and suburbs is challenging, so any presumption about the location of LIHCO households might be misleading. Rather, LIHCO households should be identified by their lower incomes and higher number of vehicles owned.

Public Transportation to Tackle Spatial Mismatch

There are three main strategies proposed to alleviate the spatial mismatch problem: bringing jobs to low-income people, bringing low income people to jobs, and better connecting low income population to jobs. While the first two solutions suppress the causes of the spatial mismatch, the

third tries to reduce the disadvantages associated with disconnection from jobs and has received more attention (*20, 1*).

Quality transit service provision is the main tactic for the third strategy. Several studies have investigated the merits of public transit to resolve spatial mismatch. The outcomes of the empirical study of Sanchez (*13*) partially support policies for improving access to public transit to improve urban employment. He used 1990 census data to examine differences in rates of labor-force participation in Atlanta and Portland who lived within a quarter-mile walking distance of a transit stop versus those who did not and found those residing near transit stops had higher rate of employment.

Gao and Johnston (*25*) use separate car and transit scenarios. They measure the marginal utility of obtaining a car or having improved transit service. They utilize Sacramento's travel demand model and apply a marginal utility indicator called "traveler benefit" in this regard. The results showed that zero-car households gained benefits in job access through each scenario. In the car scenario, however, their gain was accompanied by a loss in traveler benefits for the households already owning cars because of slight increase in congestion.

Transit service not only provides low income households with affordable access to jobs but also releases the necessity of owning more cars among them. Kim and Kim (*26*) developed an ordered logit model to predict the effect of access to public transit on automobile ownership while using the inverse square root of transit distance as a measure for transit accessibility. Using samples from "Nationwide Personal Transportation Survey" (NPTS) 1995, they found that transit access has a large negative effect on the number of automobiles owned.

QUALITY TRANSIT SERVICE OUTLINES

Transit Performance Measures

Quality transit service can positively impact spatial mismatch, though characterizing quality transit service is typically a non-trivial task. Measuring the quality of a transit service requires having quantitative performance measures, by which different aspects of service can be evaluated. Transit performance measures may be different with regards to who is performing the evaluation. In the eyes of operators, measures such as funds, revenues and ridership may be of primary importance. Transit users, however, may find frequency or in-vehicle time to be more representative of the quality of a transit system.

There are a number of key characteristics for effective performance-measurement systems. Performance measures should reflect different aspects of relevant issues. In the context of this research, this translates to having both spatial and temporal measures. Also a reasonable number of measures should be selected. Brown (*27*) describes this as choosing between “the vital few measures and the trivial many”.

In addition to the general characteristics of the measures, compatibility between measures is important. Measures vary in their structure and often include “individual measures”, “ratios”, and “indexes” mixed with raw performance data. An individual measure can be measured directly such as “ridership” and “frequency”. Although they are easy to calculate, a large number of them might be required to portray a complete picture of transit performance. Ratios are calculated by dividing one individual measure by another such as “cost per revenue mile” or

"passenger per seat". While they are not much more difficult to calculate than individual measures, they facilitate comparisons between routes, areas, and agencies.

Indexes are generally used to simplify the reporting of potentially complex measures and to produce a single output measure. These outputs are mostly normalized to fit a scale for ease of presentation.

Transit Accessibility

Transportation Review Board's (TRB) TCRP report 165: Transit Capacity and Quality of Service Manual, Third Edition, provides insights into the quality of service concepts and methods (*28*). The document defines the quality of service as "the overall measured or perceived performance of transit service from the passenger's point of view".

The document categorizes the quality of service into two availability and comfort and convenience subsets. Availability is an enabling factor that determines whether or not a transit service is even an option for a particular trip. Spatial availability at trip origin and destination, information availability, temporal availability, and capacity availability are all transit availability factors. Assuming the transit service is available, a potential passenger in the next step weighs the comfort and convenience of transit service compared to other modes. Unlike availability, comfort and convenience aspects of service quality is not a pass or fail question.

The manual suggests a number of more important and easily quantifiable measures for fixed-route transit quality in practice at the expense of excluding those aspects that are not easy to forecast such as safety and security under the term quality of service framework. The proposed

availability measures are frequency, service span, and access; while passenger load, reliability, and travel time take part the proposed comfort and convenience measures.

Frequency determines how often service is provided. Service span determines the potential market that transit serves. The longer the span, the more potential passengers can use the transit service. There are a number of ways to gain access to transit including walking, bicycling, auto drop-off, and auto park-and-ride; among which walking is the dominant access mode to local bus services. The outcomes of studies in some North American cities in the past few decades show that on average75 to 80% of passengers walk 0.25 mile (400 meters) or less to local bus stops *(28)*. Setting an access distance to transit stops forms service coverage areas, which depending on the desired level of detail could be air distance-based or walk distance-based.

As previously stated, comfort and convenience aspects of quality of service framework include passenger load, reliability, and travel time. However, forecasting the reliability and load factor needs to have automated or manual field data collection in place, which might come at a price for smaller transit operators with limited budget.

Transit travel time compared to automobile travel time is an important factor that largely influences the decision of potential passengers to whether use transit *(28)*. While the manual mentions travel time, average speed, and travel time rate as useful metrics for travel time factor, it suggests *transit-auto travel time ratio* as a metric that reflects the passenger's point of view.

Stressing "travel time" as the dominant factor influencing the travelers' view on transit quality of service, Fu and Xin (*5*) draw on the TCQSM second edition and try to quantify the

quality of a transit system into a single measure called Transit Service Indicator (TSI). The authors state that Optimal Strategy Method (OSM) is used to calculate the transit travel time between each pair of random end points in origin and destination areas (zones).

Another heavy data-driven transit accessibility tool is Transit Accessibility Measure (TAM) by Bhat et al. (*6*) for Texas Department of Transportation (TxDOT) to measure transit accessibility for fixed-route transit systems. Considering both the demand and the supply sides of transit, the software package is comprised of two indices, namely Transit Accessibility Index (TAI) and Transit Dependence Index (TDI). While the former reflects the level of transit service supply, the latter reflects the potential level of transit needs.

The authors take a "utility-based" approach and build a multinomial model of transit path choices. For each individual rider with an origin and a destination, the choice set is considered to be all paths that start within 2 miles around the origin of the trip and finish within 2 miles around the destination of that trip assuming that riders are willing to walk at most 2 miles to access transit.

One critique to the built model is that the utility-based transit accessibility models need detailed on-board surveys, which in many cases are cumbersome especially for smaller transit providers with limited budgets. Additionally, and more importantly, although the paper proposes temporal accessibility measures, among the others, to be included in the multinomial choice model, they turn to be statistically insignificant and were consequently removed from the final model.

To include the temporal aspects in transit accessibility analysis, Polzin et al. (*7*) propose Time-of-Day-based Transit Accessibility Analysis tool. The days of service are first divided to a series of desired time periods. The supply side of transit is represented by "frequencies" of service in each period and on each route. The demand side of the temporal component is incorporated by calculating the tolerable wait time in a period on the route and the fraction of daily travels that fall within a period on that route.

The spatial component takes account for the fraction of the population that falls in the transit buffer in each zone and for each route. However, instead of the raw population counts, the model uses an "equivalent population" variable that sums the raw population of the zone with the normalized employment counts of that to account for both employment and population as the transit ride generators. A general critique to this method is that transit accessibly scores are calculated for single transit stations not for origin-destination pairs.

The method developed by Mamun et al. (*8*), "Transit Opportunity Index" (TOI), accounts for spatial and temporal components and for pair of origin and destination zones. Dividing between "transit access" and "transit connectivity", the paper defines accessibility as the ability of travelers to reach transit facilities. It is calculated by multiplying the relative buffer area around transit lines to the zone's area by relative frequency of services to the population of that zone (temporal component).

The connectivity parameter is measured by a binary "connectivity parameter" and "decay factor". Connectivity parameter takes value of 1 if two zones are directly connected and 0 otherwise. The decay function captures the discomfort associated with longer travel times.

Finally, the TOI is calculated by multiplying the accessibility parameter by the connectivity parameter for each pair of the trip ends and for each route.

Similar to Mamun et al. *(8)*, the transit accessibility measure by Hart and Lownes (*9*) accounts for both spatial and temporal components and pair of origins and destinations. Focusing on New Haven, CT as the case study, this paper shows that number of low income jobs accessible from residential locations and late night transit frequency at the residential locations are negatively associated with the number of LIHCO households as the percentage of a block group's low income population.

The paper refines the analysis area for destinations to those within a quarter mile of a bus stop. Low income jobs are also categorized. A low income job spot is accessible from a location of residence if the destination work zone could be reached between the hours of 7-8 AM in less than 60 minutes. To account for temporal aspects of accessibility, the cumulative transit frequency for the off-peak hours (9PM-12AM) is found for each block group.

The "multiple regression model" with two predictors suggests that higher number of LIHCO households correlates to the areas with lower late night service frequency and lesser accessibility to low income jobs.

CHAPTER 3: METHODOLOGY

AN OVERVIEW OF THE SELECTED METHODOLOGY

Methodology Summary

Figure 1 outlines the general flow of data analysis in this research, where rectangles show datasets and circles represent commercial software packages or open source models/programming languages. PopGen model uses PUMS dataset and Census Summary Files and outputs household income, number of vehicles, and time of departure to work. In the next stage, Census Tiger/Line is joined to geocode these outputs and Census Summary File P1 table is joined to add block group population.

On the supply side, Python programming language is used to generate hourly available seat per capita ratios and average travel time ratios. These two outputs are then combined with demand-side output from the last stage for conventional statistical and spatial statistical analyses. Pearson chi-square test and bivariate Pearson correlation is conducted in SPSS environment while bivariate Moran's I and bivariate LISA are constructed in GeoDa. Full description of each of the above stages is discussed in chapter 3.

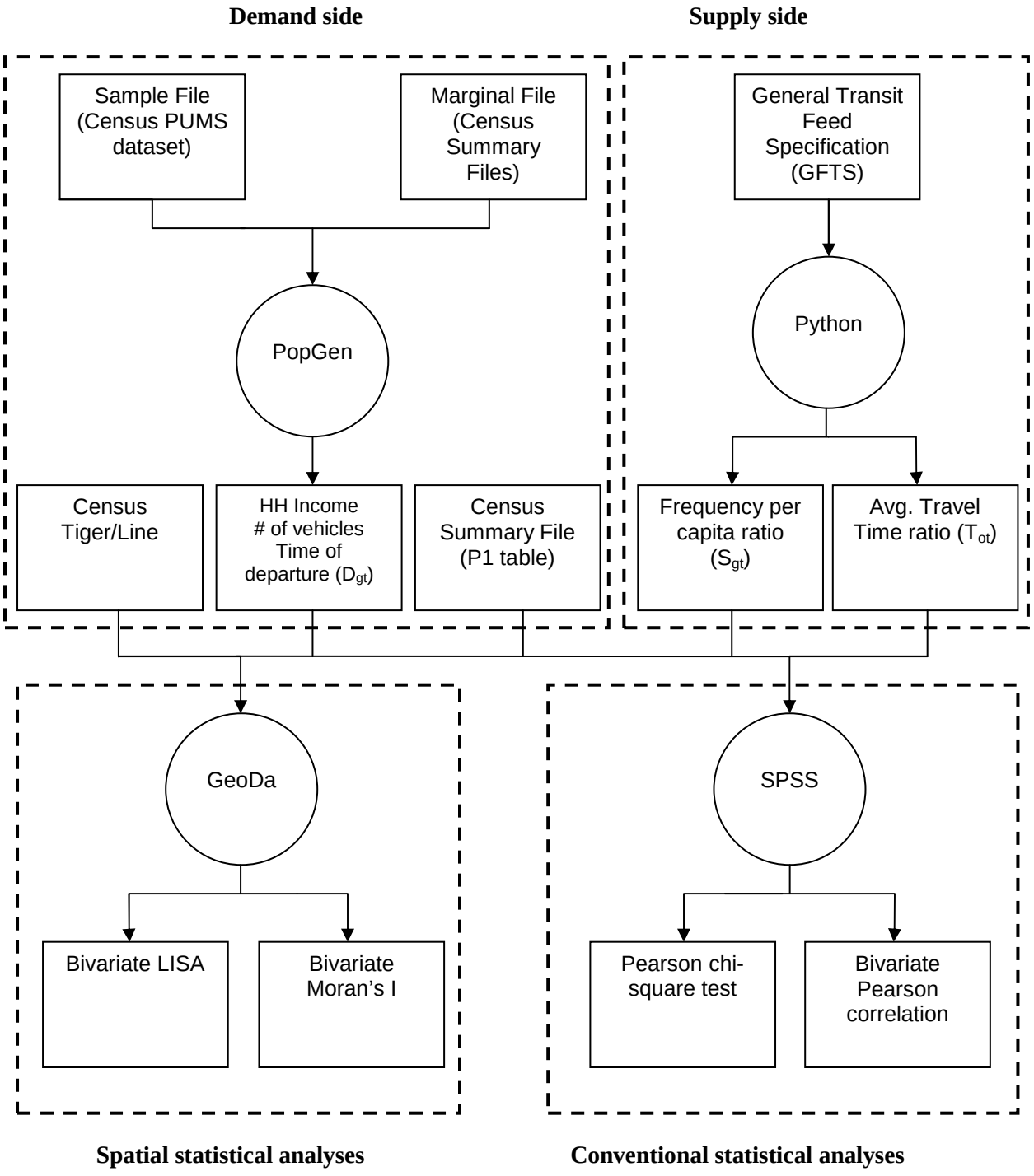


Figure 1 Outline of the general flow of different analyses

Selected Transit Quality of Service Measures

Transit access, transit frequency, and transit travel time are selected to assess how well the transit system performs. Since the common state of practice is to consider 0.25 mile (400 meters) Euclidean distances as service coverage areas, in this research, transit access is characterized by buffer zones with a 0.25-mile radius around bus stops. More specifically, transit access is calculated as a ratio of each buffer zone that fall within each analysis zone (Census block group). All transit access, transit frequency, and transit travel times are calculated from General Transit Feed Specification (GTFS), which is a common format for public transportation schedules and associated geographic information.

Portraying the Demand Side of Transit

If spatial mismatch is defined as an imbalance between transit supply and employment demands of population, in addition to transit performance measures, portraying the workers commute demand is the other indispensable aspect to properly lay out the problem. A handful of studies investigate the demand side of transit service (*10, 14, 13, 12, 30, 7, 6, 5*), among which majority use employment data as a proxy for demand (*10, 14, 13, 12, 30, 5*).

In this sense, and in its simplest form, one might think of transit accessibility in the lines of the Newton's law of *universal gravitation*, which relates the interaction of two bodies in the universe to their masses and the reciprocal of the distance between them. Transit performance measures are intrinsically pair-wise measures that gauge one aspect of transit service between a binary trip ends of trip origin and trip destination. The potential of transit trip production at origin end could be conceived by the levels that transit is spatially and temporally provided in

that zone. In other words, transit access and transit frequency could be taught as transit trip production impetuses.

Alternatively, the potential of transit trip attraction at destination end could be conceived by the levels that transportation is needed at destination zone. Considering employment as the transit trip attraction stimulus is even more relevant in spatial mismatch discourse, where job accessibility of working population is under study. Finally, the transit travel time between transit trip origin and transit trip destination can be deemed as impedance or friction factor in the Newton's gravity model framework.

Although employment at a transit trip destination might represent the extent at which workers need the transit service from their residence to that destination so incorporate the demand side of transit service, it lacks the temporal component that portrays the fluctuation of different levels of demand during course of a day. This last point deserves even more attention while investigating the spatial mismatch and working with low-income sub population, which have different travel patterns *(26, 28, 4)*. Thus there is a need for a demand-side measure that is capable of showing the differences in employment demand during different time intervals. Such measure draws more realistic picture of employment demand during the course of a day than a single aggregate value for the whole day.

Reviewing the literature suggests that temporal distribution of travel demand plays an important role in determining the importance of service provided in each time period of day, however, as Polzin et al. (*7*) state, temporal dimension has lagged behind the spatial dimension

with respect to its explicit recognition and incorporation in transit accessibility measurement studies.

The method that Polzin et al. (*7*) use has two temporal measures of “tolerable wait time” and “frequencies”. The mechanism for incorporating the temporal demand is dividing the days of service into a series of desired time periods and then weighing the tolerable wait time and frequencies within each period on each route by the amount that riders use the transit in a particular time and on a particular route. This allows service in periods of high demand to be weighted more heavily than that in periods of low demand into calculating the overall accessibility index.

In this research, proportion of working population departing to work at different time intervals is used as a good demand side measure to capture differences in employment demand during different time intervals. Using a simultaneous mode and departure-time choice model, Hendrickson and Plank (*31*) found that departure time decisions are even more elastic than the mode choices so could be used toward transit system management strategies. McKenzie and Rapino (*32*) state information about when workers leave for work plays an integral role in the regional transportation planning process, including an understanding of public transportation infrastructure.

Equation 1 illustrates the calculation of transit accessibility A_{ij} in its simplest form (*a*) and then the method this research adopts (*b*). The main difference lies in how to incorporate the demand side of the transit accessibility. While the existing literature use employment data at the

transit trip destinations E_j, this research uses the proportion of workers who commute during different time intervals W_i.

In equation 1, A_{ij} represents overall transit accessibility, B_i *is* transit buffer, F_{ij} represents transit frequency, E_j shows employment at transit destinations, W_i shows proportion of workers departing to work at each time interval, T_{ij} denotes transit travel time, and finally *i* and *j* represent origin and destination of transit trips respectively.

$$A_{ij} \approx \frac{B_i \times F_{ij} \times E_j}{T_{ij}} \quad (1.a)$$

$$A_{ij} \approx \frac{B_i \times F_{ij} \times W_i}{T_{ij}} \quad (1.b)$$

DATASETS AND MODELS

Datasets

This research uses several datasets from different sources and at different levels of detail. Table 1 summarizes the datasets, tables, and files used in this research. Census Tiger/Line shapefiles contain geography entity codes (GEOIDs) that are used to link the Census Bureau's demographic data to their corresponding block groups.

All transit access, transit frequency, and transit travel times are calculated from General Transit Feed Specification (GTFS), which is a common format for public transportation schedules and associated geographic information. A GTFS feed is composed of a series of text files collected in a ZIP file. Each text file models a particular aspect of transit data including calendar schedules, routes, trips, stops, and stop times.

Census Public Use Microdata Sample (PUMS) and American Community Survey (ACS) detailed tables are used toward generating the household income, household number of vehicles, and time of departure to work up to the person level. This is done by utilizing PopGen, which is an open-source synthetic population generator. PUMS files contain individual records of the characteristics for a five percent sample of people and housing units. Microdata are generally individual records which contain information collected about each person and housing unit. Information which could identify a household or an individual is excluded in order to protect the confidentiality of respondents.

Tables from summary files (detailed tables) provide the most detailed data on all topics and geographic areas from the decennial population and housing census, the economic censuses and the American Community Survey. Decennial census detailed tables are identified and labeled using established guidelines. Table identification begins with a letter that refers to the type of data in the table, and then a number is assigned sequentially as the tables are produced. In this research P1 table is used ("P" are population tables). American Community Survey (ACS) detailed tables begin with the letters B for base tables, and C for collapsed tables. The collapsed tables cover the same topics as the base table, but with fewer details. This research uses are series of different base tables as one of the inputs of PopGen model.

Dataset	Version/year	File/table used	Level of detail
Census Tiger/Line	2010 Census		Block group
Census summary files (detailed tables)	2010 census	P1	Block group
General Transit Feed Specification (GTFS)	2014	agency, calendar, calendar_dates, routes, trips, stops, stop_times, shapes	Miscellaneous
Census Public Use Microdata Sample (PUMS)	ACS 2013, 5-year estimates	person file household file	Person Household
Census summary files (detailed tables)	ACS 2013, 5-year estimates	B19001, B25044, B08302, B09019, B01003	Block group

Table 1 Datasets, files, and tables used

PopGen: Synthetic Population Generator

This research uses a synthetic population generator, called PopGen, to generate the required person- and household-level variables. Population generators generally intend to synthetically generate rich and detailed data that is not available from conventional data sources. This detailed data might originally not have been measured by the conventional surveys or suppressed for the privacy reasons hence not available to public and researchers. Most of the time, however, the detailed data is available for a small sample of the whole region, which enables researchers to synthetically generate the detailed data for the whole region by expanding the sample to match some known aggregate variables of the region in an iterative process.

Similarly this research uses a population generator for two main reasons. One, in order to identify the LIHCO households, the number of vehicles by poverty status is required. However, that is not available at levels of detail lower than Census tracts from relevant sources such as Census Transportation Planning Package (CTPP). This research runs most of the analyses at the

block group level, which is finer than census tracts, so it needs PopGen to identify LIHCO at that level of detail. Second and more important than the first one, there is no dataset that has cross-tabulation of time of departure to work and LIHCO status. By using PopGen, LIHCO households are identified in New Haven County and time of departure to work is specified for each individual in each household. This data is used in table 5 to investigate whether LIHCO individuals have different temporal employment demands from non-LIHCO individuals in the first place.

PopGen is originally developed by Ye et a. (*33*) to address some weaknesses of conventional synthetic population generators by a heuristic approach, called "Iterative Proportional Updating" (IPU) algorithm. Both conventional and IPU population generators obtain the marginal distribution of desired variables for the whole region from the summary files and expand the sampled cross-tab of variables to match the marginal distribution. The conventional generators, however, just control for household-level attributes to match while IPU controls for both household-level and person-level attributes. In other words, the IPF procedure watches the household composition while assigning the weights and might re-assign weights based on household composition. To measure how well the model predicts the cell frequencies, Ye et al. (*33*) use the chi-square test statistic.

$$\chi^2 = \sum_j \left[\frac{(n_j - c_j)^2}{c_j} \right] \qquad (2)$$

In the above formula, n_j denotes the frequency of the synthetic person of the j^{th} person type and c_j is the j^{th} IPF-estimated person-type constraint. It is noteworthy that the model

generates chi-square values for each geography unit. For example, if the geographic resolution is set to block groups, each block group has one chi-square test statistic. Also, it is important to note that the "person-type" or "household-type" definition refers to the unique persons or households with unique sets of values for different person-level or household-level variables. The number of person-type of household-type constraints is the combination of the categories of different person-level or household-level variables. For example, if a person file has 2 categories of gender and race and gender could be either male or female and race could be white, black, or Asian, there are 6 person-type constraints (Figure 2).

Using the chi-square test statistic, a "Pearson Chi-square test for goodness of fit" can be performed on the data. While the null hypothesis states that the difference between the estimated frequencies of different person types and the constraints in each geography unit is by chance, the alternative hypothesis states that the difference is statistically meaningful. Considering a level of significance (α), the null hypothesis might or might not be rejected. This research will include a geography unit in further analysis even if the null hypothesis for that geography is rejected. The number of geographies that do not pass the test, however, will be identified and marked as a caveat of the model.

Person Attributes	
Gender	Male
	Female
Race	White
	Black
	Asian
	Number of constraints = 2*3=6

Figure 2 Example of number of person-type constraints calculation

Table 2 shows different person-level and household-level variables/tables that this research uses. As stated before, two datasets of sample file and marginal file are needed to run the PopGen model. It is noteworthy that groupquarters do not have any actual variables, but a dummy variable that specifies the count of groupquarters in the marginal files. They, however, need to be extracted from regular household records for the purpose of modeling.

The first row of Table 2 shows the sample file categories. Since the marginal files have comparatively limited number of categories, the sample file categories are aggregated based on them. To identify the groupquarters among other records in “household” PUMS file, the “TYPE” variable is used. The second row of the table shows the marginal files at different levels. Census table “B09019” is used to calculate the groupquarter counts for the marginal files. Also, the original number of categories for the time of departure to work in marginal files was 14. However, to deal with “N/A” data records (non-workers or who worked at home), a dummy category is added to both sample and marginal files. Table “B01003” (total population) is used for marginal file calculations in this regard.

	Household-level variables		Person-level variables		Groupquarter-level variables	
	Variable/table name	# of categories	Variable/table name	# of categories	Variable/table name	# of categories
Sample File (census PUMS dataset)	HINCP* (household income in past 12 months)	Not categorized (continuous)	JWDP** (time of departure for work)	150	Groupquarter	1
	VEH* (vehicles available)	7 (from zero car to plus 6)				
Marginal File (census summary files/detailed tables)	Table B19001 (Household income in the past 12 months)	16 (from less than $ 10,000 to $200,000 and more)	Table B08302 (time leaving home to go to work)	15	Groupquarter	Equal to count of groupquarters in each block group
	Table B25044 (tenure by vehicles available)	6 (from zero to plus 5)				

* From "household" PUMS file
** From "person" PUMS file

Table 2 Household and person level variables used in PopGen model

SOLUTION PROCEDURES

Transit Frequency

Frequency is one of the "availability" measures of the transit service quality. While it is intrinsically a transit stop's (and a transit line's) characteristic, this characteristic needs to be applied to the population zone (Census block group). In this research, the frequency measure is presented as a ratio of hourly available seats per capita.

Equations 3 to 7 show the calculations for seats per capita ratio (S_{gt}) for each block group g at time interval t, where i denotes the bus stop, l specifies the bus route, g is the Census block

group (population zone), and U is the bus capacity, which is deemed to be 45 seats assuming typical city buses.

$$F_{it} = \sum_{l} F_{itl} \tag{3}$$

$$C_{igt} = \frac{F_{it}U}{P_g}, \ \forall i,g: B_{ig} \neq 0 \tag{4}$$

$$R_{ig} = \frac{B_{ig}}{B_i} \tag{5}$$

$$S_{igt} = R_{ig}C_{igt} \tag{6}$$

$$S_{gt} = \sum_{i} S_{igt} \tag{7}$$

Equation 3 sums all frequencies throughout all bus routes for each bus stop at time interval t. Equation 4 calculates the hourly available seats per capita for people who access bus stop i from block group g and at time interval t (C_{igt}). It inputs the frequency of bus stop i from the last stage and multiplies that by bus capacity (45 seats) and then divides that by population of block group g provided that service area of bus stop i overlaps with block group g. That is to normalize the hourly available seats by the population of the block group.

As suggested by the literature, an average access area with 400-meter radius is considered as the access area around each bus stop, however, all this access area might not fall under a single block group. It is hard to judge which bus stop provides better service frequency if only a part of the access area around the bus stop serves the block group. Equations 5 and 6

resolve this problem by calculating the proportion of the bus stop service area within each block group to the whole service area and adjusting C_{igt} values accordingly.

Equation 5 calculates the ratio of the access area of bus stop i that falls within block group g (B_{ig}) to the whole access area of bus stop i (B_i). Then in equation 6, this ratio is multiplied by C_{igt} to output hourly seats per capita values of bus stop i for block group g at time interval t (S_{igt}). One assumption in equations 4 to 6 is that the population is uniformly distributed throughout each block group. The final step is to sum all hourly seats per capita values of the bus stop i for the block group g at time interval t over all bus stops with intersecting access areas with block group g. The outcome is called hourly available seats per capita ratio (S_{gt}) for block group g at time interval t.

Transit Travel Time

Transit travel time is one of the "comfort and convenience" measures of the transit service quality. This research considers both trip ends in calculating the travel time, and then sums travel time values over different destinations of a single origin to output the transit travel time for that origin. If transfer is necessary, a fixed 20-minute transfer time is assumed and added to travel time. Like frequencies, one particular challenge here is to expand the travel time between a pair of origin and destination bus stops to a pair of origin and destination block groups. The transit travel time is calculated as a ratio, and the final travel times are presented in minutes in this research. The lower transit travel time, the better the transit service.

Equations 8 to 10 show different calculations for transit travel time, where T shows the transit travel time, l denotes the bus route, i and j are origin and destination bus stops, o and d

represent origin and destination block groups, *t* represent the time interval, and finally *n()* notation specifies the cardinality of different sets.

$$T_{ijt} = \frac{\sum_{l} T_{ijtl}}{n(l)} \tag{8}$$

$$T_{odt} = \frac{\sum_{i}\sum_{j} T_{ijt}}{n(i)\times n(j)} \ , \ \forall i,j,o,d: B_{io} \neq 0 \ AND \ B_{jd} \neq 0 \tag{9}$$

$$T_{ot} = \frac{\sum_{d} T_{odt}}{n(d)} \tag{10}$$

Equation 8 calculates the average transit travel time between origin stop *i* and destination stop *j* at time interval *t* over different routes connecting those origin-destination pairs.

Equation 9 in the next step calculates the travel time between origin block group *o* and destination block group *d* at time interval *t*, using the origin-destination bus stops *i* and *j*. For each origin block group o, origin bus stops *i* are all bus stops with intersecting access areas with block group *o* ($B_{io} \neq 0$). Similar to frequency calculations, the access area of a bus stop is considered to be 400 meters. Likewise, for each destination block group *d*, destination bus stops *j* are all bus stops with intersecting access areas with block group *d* ($B_{jd} \neq 0$). Equation 9 sums all travel times between these candidate *i-j* pairs for each *o-d* block group at time interval *t* and then divides the result by the count of these pairs. The outcome is the average travel time between origin block group *o* and destination block group *d* at time interval *t* (T_{odt}).

The last step is to sum all pair-wise transit travel times between origin *o* and destination *d* over the destinations to get an average travel time ratio for each block group *o* at time interval *t* in the study area.

Time of Departure to Work

This research uses proportion of working population departing to work at different time intervals as a demand side measure to capture differences in employment demand during different time intervals. Equation 11 shows the calculations. The first step is to calculate the working population in block group *g* who depart to work at time interval *t* (W_{gt}). This variable is previously calculated by summing the number of working individuals who depart to work at time interval *t*, produced by PopGen model, over the block group *g*.

The next step is to normalize this by total working population of block group *g* to obtain the proportion of the working population in block group *g* that depart to work at time interval *t* (D_{gt}).

$$D_{gt} = \frac{W_{gt}}{\sum_{t} W_{gt}} \tag{11}$$

Conventional Statistical Analyses

This study is intended to research the levels that transit supply meets the employment demands of worker population in general and LIHCO population in specific. One important question in this regard is to first investigate whether LIHCO individuals have different temporal employment demands from non-LIHCO population. To find the primary answer of this question Pearson Chi-

square test for goodness of fit is constructed on the data. Two categorical variables are LIHCO status (with two categories of LIHCO and non-LIHCO) and time of departure to work (with four categories of AM Peak, Mid-day, Evening, and Early AM).

The output of the popGen model is the only needed source of data for running this analysis. PopGen generates the cross-tabulation between income, number of cars, and time of departure to work for each worker individual in New Haven County and that data is used in running Pearson Chi-square test at worker individual level.

The intersection of LIHCO status and time of departure to work variables creates a 2x4 or 4x2 contingency table. Calculation for the test statistic is shown in equation 12, where i and j denote the rows and columns and O_{ij} shows the observed counts in row i and column j of the table.

$$\chi^2 = \sum_{i,j} \left[\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \right] \tag{12}$$

In equation 12, E_{ij} shows the expected counts for row *i* and column *j* of the table and is calculated using equation 13, where *R* is the row total and *C* is the column total of the observed counts and *n* is the observed count of each cell.

$$E_{ij} = \frac{R_i C_j}{n_{ij}} \tag{13}$$

While the null hypothesis (H_0) states that the difference between expected number of LIHCO or non-LIHCO departing to work at each time interval and the actual numbers is by

chance, the alternative hypothesis (H_1) states that the difference is statistically meaningful. Considering a level of significance (α), the null hypothesis might or might not be rejected. If the null hypothesis is rejected, it is concluded that LIHCO individuals have significantly different temporal employment demands from non-LIHCO individuals.

The outcome of this analysis feeds into the spatial mismatch discourse as it pertains to transit service provisions. For example, if it is concluded that LIHCO individuals commute to work during the Mid-day hours significantly higher than non-LIHCO individuals, providing LIHCO block groups with high frequent mid-day service deserves more consideration.

Another important consideration is to investigate the association between the levels of transit supply (frequency and in-vehicle travel time) and proportion of LIHCO population. That is to speculate how well the population zones with higher LIHCO population are served by transit supply in each time interval. This can be done by constructing a *bivariate Pearson correlation*. If two-tailed significance test is used, the null hypothesis (H_0) states that the population correlation coefficient is 0 (no association) and the alternative hypothesis (H_0) states that it is not 0. A special attention might be given to the time intervals with significantly more LIHCO employment demands in hopes of more LIHCO temporal demand and supply match.

Additionally, correlations between pairs of transit supply measures (frequency and in-vehicle travel time) and transit demand measure (proportion of the working population) could also be investigated by using the same test. It is indicative of general association between transit supply and employment demand.

Spatial Statistics Analyses

Spatial component of spatial mismatch requires applying spatial statistics techniques in examining the match between the areas of population (demand) and the levels of transit supply. All frequency, travel time, working population, income, and car-ownership data are essentially spatial data and are subject to spatial dependency (*34*). Applying sole conventional statistics methods might lead to dismissing such interactions in space hence distorting or falsifying the inferences or results (*15*).

Spatial dependence exists when a value observed in one location depends on the values observed at neighboring locations. There are two main reasons for that in the literature. First, data collection of observations associated with spatial units may reflect measurement error. This happens when the boundaries for which information is collected do not accurately reflect the nature of the underlying process generating the sample data. In this research, selecting the block group as the unit of analysis was due to the factors other than the changes in the values of the LIHCO density, frequency, or travel time observations, so this type of spatial dependency might exist.

Second, underlying socio-economic process might lead to clustered (or alternatively dispersed) distribution of variable values through spatial interactions, diffusion, or spill-over effect. For example, because of numerous external socio-economic factors, LIHCO households might live in nearby block groups in the New Haven County.

In statistics, the official term for spatial dependence is *spatial autocorrelation*. Autocorrelation is the correlation of a variable with itself, and spatial autocorrelation is defined

as correlation of a variable with itself through the space. The problem with autocorrelation is that it violates the independence assumption among samples in Ordinary Least Square (OLS) estimation of linear regression models as well as in bivariate Pearson correlation test:

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j$$

Following each of those two reasons for spatial autocorrelation, two different solutions are proposed. *Spatial Error* models assume autocorrelation in the error term, meaning the residuals are auto-correlated. The model structure takes the below form. In this case, the error term ε has a spatially weighted component $\lambda W\varepsilon$ and a random error term u.

$$Y = X\beta + \varepsilon$$

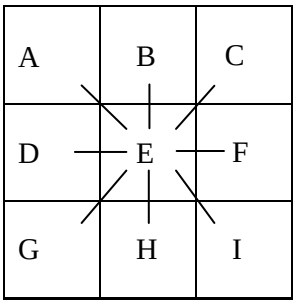
$$\varepsilon = \lambda W\varepsilon + u$$

Spatial Lag models assume that the response variable is not only dependant of the predictors but on itself as well. An additional term in the model is used to represent this:

$$Y = X\beta + \rho WY + \varepsilon$$

In the above formula, W is the "spatial weights matrix" and ρ is an additional coefficient to be estimated, which measures the autocorrelation. Spatial weights matrix indexes the relative locations of all zones *i* and j and is the main mechanism that specify how neighboring zones correlate in the space. There are two main weighting systems in this regard. Contiguity-based weights assign a weight of 1 to zone j if it is adjacent to zone i and 0 otherwise. Distance-based weights measure the actual distance between polygon centroids. Further, polygon contiguity might be rook, bishop, or queen resembling the chess moves. The *order number* then specifies

what level of the nearest neighbors to be included. Figure 3 shows a first order queen contiguity and its corresponding spatial weights matrix.



	A	B	C	D	E	F	G	H	I
A	–								
B	0	–							
C	0	0	–						
D	0	0	0	–					
E	1	1	1	1	–				
F	0	0	0	0	1	–			
G	0	0	0	0	1	0	–		
H	0	0	0	0	1	0	0	–	
I	0	0	0	0	1	0	0	0	–

Figure 3 First order queen contiguity and the spatial weights

Important questions are how to find whether spatial correlation exists. There are a few methods to check the spatial autocorrelation, among which “Global Moran’s I” is the most common one and used in this research (Equation 13).

$$I = \frac{n\sum_i\sum_j w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_i (x_i-\bar{x})^2 \sum_i\sum_j w_{ij}} \tag{14}$$

From equation 14, x is the variable of interest, w_{ij} is an element of the spatial weights matrix, which takes *1* if zone *i* and *j* are adjacent and *0* otherwise, and *n* is the number of zones. The Moran's I can be visualized as the slope in a scatterplot of the spatially lagged variable versus the original variable. The spatially lagged value of measure *y* is a sum of the spatial weights multiplied by values for observations at neighboring locations, expressed as *Wy*, where *W* is the spatial weights matrix (*34*). If no spatial correlation exists, the expected value of "Moran's I" is as the below.

$$E(I) = \frac{-1}{n-1} \tag{15}$$

Statistical significance test for "Moran's I" could be based on a normal distribution assumption. However, while independence of variable values is under serious question (iid assumption), that procedure does not seem to be appropriate. Generally, spatial data hardly follow a normal distribution, so a more adequate way for statistical significance test of the Moran's I is through Monte Carlo process (permutations). The final result is a z-score, which is similar to that of a normal distribution assumption. The test statistic is calculated using the below formula. The null hypothesis (H_0) is that there is no spatial correlation $I=E(I)$ while the alternative hypothesis (H_1) is that spatial correlation exits $I<>E(I)$.

$$Z = \frac{I - E(I)}{Se(I)} \tag{16}$$

If $I>E(I)$, positive autocorrelation or "clustered pattern" exists, if $I<E(I)$ negative autocorrelation or "dispersed pattern" exists. If the null hypothesis cannot be rejected, it means

that the spatial autocorrelation identified by Moran's I is not statistically significant and only occurs by chance. In that case, any event has an equal probability of occurring at any location, or, in other words, position of any event is independent of the position of any other.

The Moran scatterplot portrays the values of observation versus the spatially lagged values of observation and can be divided to four different quadrants each representing a spatial association that exists: high values surrounded by similarly high values (High-High); low values surrounded by dissimilarly high values (Low-High); low values surrounded by similarly low values (Low-Low); and high values surrounded by dissimilarly low values (High-Low).

Moran's I is a global measure that checks the spatial autocorrelation in the entire study area. However, autocorrelation might exist in some parts of the region but not in others, or even positive in some areas and negative in others. For that reason and to be able to investigate the autocorrelation in more details, there is a need for a local version, capable of calculating autocorrelation for each areal unit in the data. Local Indicators of Spatial Association (LISA) is the local version of the *Moran's I,* which measures the spatial correlation locally for each areal unit based on neighboring polygons of that unit (*34*).

$$I = \frac{n(x_i - \bar{x})\sum_j w_{ij}(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2 \sum_i \sum_j w_{ij}} \tag{17}$$

In equation 17, x is the variable of interest, w_{ij} is an element of the spatial weights matrix, which takes *1* if zone i and j are adjacent and *0* otherwise, and n is the number of zones. Building on the Moran scatterplot, the Moran significance map incorporates information about the

significance of local spatial patterns and the Moran cluster map shows the type of spatial patter (HH, LH, LL, and HL).

What is discussed so far about spatial autocorrelation refers to the spatial pattern of a particular phenomenon, which is depicted as the correlation of that phenomenon's observations with the weighted average of observations of the same phenomenon at neighboring locations. In a lot of cases, however, the spatial relationship between two different variables is in question. The Bivariate Moran's I and the Bivariate LISA investigate the global and local spatial correlation of two variables respectively.

The bivariate spatial autocorrelation is the correlation between one variable at a location and the weighted average of different spatially lagged variable at all neighboring locations as defined by the spatial weights matrix (*34*). The bivariate Moran's I can be visualized as the slope in a scatterplot of the spatially lagged values of one variable on another variable.

CHAPTER 4: APPLICATION ENVIRONMENT

AN OVERVIEW OF THE APPLICATION AREA

Socio-economic Characteristics

The application area of this research is New Haven County, which is located at the south central part of the state of Connecticut. The population of the New Haven County was 824,008 in 2000 and that increased to 862,477 in 2010 showing a 4.67% increase in 10 years. The population of the whole state of Connecticut showed a little higher rate of increase (4.95%) during the same period reaching to 3,574,097 in 2010 from 3,405,565 in 2000 (U.S. Census Bureau)

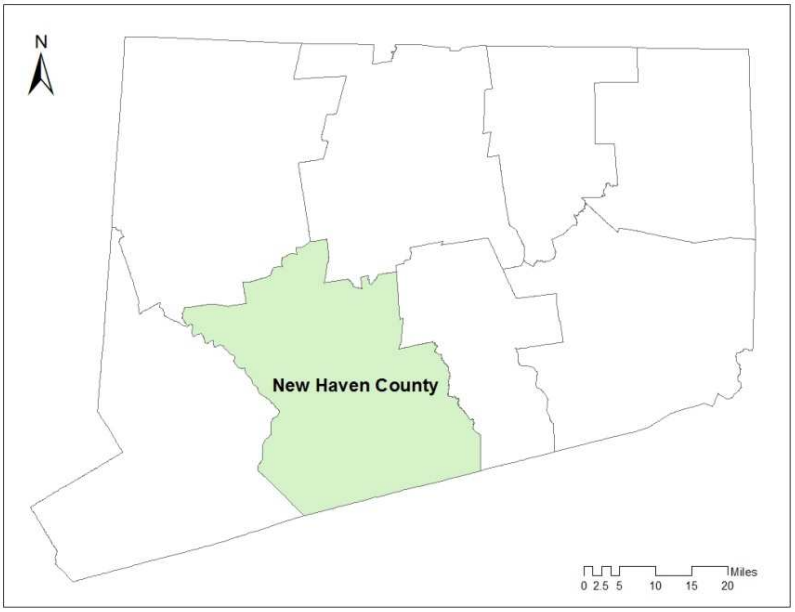


Figure 4 New Haven County in the state of Connecticut

The median household income in New Haven County in 2013 is 61,996 while Connecticut and the U.S. have median household income of 69,461 and 53,046 respectively for the same year based on the ACS 5-year estimates. Other socio-economic data including time of

departure to work, number of vehicles available, and household income are shown in table 3. In terms of income, it seems that the New Haven County stands somewhere between Connecticut and the whole nation. State of Connecticut has more middle or high income and less low income families than New Haven County and the whole America. Similar order applies to car ownership as well, though the differences among three geographies are smaller.

In terms of time of departure to work, majority of workers leave to work during the AM peak hours throughout all three geographies. New Haven County shows a larger percentage leaving during the day and late at night.

		New Haven County	State of Connecticut	U.S.
Time of departure to work				
	Early AM (12:00-6:00AM)	8.7%	9.0%	12.7%
	AM (6:00-9:00AM)	66.7%	67.6%	63.1%
	Mid-day & evening (9:00AM-12:00AM)	24.7%	23.4%	24.3%
Number of vehicles				
	No vehicle	4.1%	3.5%	4.4%
	One vehicle	21.8%	20.0%	21.4%
	Two vehicles	41.6%	43.0%	42.2%
	Three or more vehicles	32.5%	33.6%	31.9%
Household Income (2013)				
	Low income (<$25,000)	20.9%	18.0%	23.4%
	Lower-middle income (>$25,000 and <$50,000)	20.4%	18.8%	23.9%
	Middle income (>$50,000 and <$100,000)	29.6%	29.6%	30.1%
	High income (>$100,000)	29.1%	33.7%	22.6%

Table 3 Selected socio-economic characteristics among New Haven County, Connecticut, and the U.S. (ACS 2013 5-year estimates)

Public Transportation Profile: CTTrasnit

The CTTrasnit is the main operator in New Haven County offering service (bus transit) to more than 530,000 people in 456 square miles. Although the buses are owned by the Connecticut Department of Transportation (ConnDOT), other companies operate the system under contract. It operates over 22 local routes in New Haven area in the cities such as Meriden, Waterbury, Wallingford, Milford and others.

Towns and Cities

The state of Connecticut is divided into 169 towns and these towns are grouped into eight counties including the New Haven County. Villages are named localities within towns but have no separate corporate existence from the towns they are in. The New Haven County includes 27 towns. The city of New Haven, which is the home to Yale University, is the largest city in the New Haven County and the second largest city in the state according to 2010 census. Figure 5 shows different towns in New Haven County.

Figure 5 Different towns in New Haven County

CHAPTER 5: RESULTS

OUTCOMES OF THE ANALYSES

Temporal is one of the aspects of the analyses in this research, so there is a need to exactly define time interval during a day for different calculations. The initial thought was to categorize the time of day by 4 different categories of AM peak (6:00-9:00AM), Mid-day (9:01AM-3:00PM), PM peak (3:01-6:00PM), and off-peak (6:01PM-5:59AM). However, the special categorization of the census summary files does not allow such grouping, so the below categorization is selected:

- 6:00–8:59AM……AM Peak
- 9:00AM–3:59PM……..Mid-day
- 4:00–11:59PM……..Evening
- 12:00–5:59AM……..Early AM

In terms of the application area, there is a constraint that limits the sample of all block groups in New Haven County for further analysis. Frequency and travel time calculations only consider the block groups that are partly or totally located in the 400-meter buffer area around the bus stops. Only 403 block groups out of 608 block groups in New Haven County have the above stated condition and hence are selected for analysis. In terms of LIHCO categorization, households that make less than $25,000 a year and have more than a car are considered to be low income and high car ownership.

PopGen Model Outputs

This research uses PopGen 1.1 model to generate one person-level (car ownership) and two household-level (household income and working population departing to work) variables in the New Haven County. Table 4 shows the statistics on outputs of the model. The model generates the same amount of households as actual. However, the synthesized population count is 1.06 percent less than the actual. Among all 628 block groups in New Haven County, 474 have significant variable estimation at 0.05 level of significance (cf. Equation 2), however, all 628 block groups will be considered in analyses.

	All results (block groups)		
	Actual	Estimated	% change
Persons	862,611	853,464	-1.06
Households*	356,667	356,667	0
Number of block groups	628	628	
Avg. number of iterations before convergence		47	

* Actual households and groupquarters are combined to simplify the comparisons with estimations.

Table 4 Statistics on outputs of PopGen model

Results of Conventional Statistical Analyses

Using statistical Pearson Chi-square test and bivariate Pearson correlation analysis, this section tries to first testify whether LIHCO individuals have different temporal employment demands from non-LIHCO individuals and then find the direction and the magnitude of association among different levels of transit supply and LIHCO density as well as that among transit supply and employment demands of working population.

Table 5 shows the results of Pearson chi-square test on working population who depart to work at different time intervals and LIHCO status. The advantage of using PopGen allows running this test at person level and on 394,771 different synthesized worker individuals in New Haven County, so the only source of data for this Pearson Chi-square test is the output of PopGen model, which generates a cross-tabulation between income levels, number of cars owned, and time of departure to work at individual level. It is noteworthy that from 853,464 estimated persons in New Haven County, 394,771 persons estimated to be workers who work outside of their homes and, with having a value in "time of departure to work" field, are considered as valid records for Chi-square test analysis. In general, with a *p-value* of *0.00*, the test rejects the null hypothesis at *0.01*-level of significance. This means that the difference between expected and actual number of LIHCO or non-LIHCO individuals who depart to work at different time intervals is not by chance and indeed is statistically meaningful. In other words, LIHCO individuals do have different temporal employment demands from non-LIHCO individuals.

The direction of change between actual and expected cell values in table 5 is not the same for all time intervals. Generally, LIHCO populations were expected to depart to work more during 6:00-9:00AM than actual counts. This finding complies with other literature that finds low-income workers less likely to commute during the AM peak (*4*). The table also shows that LIHCO populations relatively depart to work more during mid-day and evening hours than what was expected. The proportion of the LIHCO population that departs to work during early AM hours is not that different from what was expected.

		Non LIHCO	LIHCO	Total
Working population who leave to work in AM peak ($W_{AM\ Peak}$)	Expected	257,641.5	5,351.5	262,993.0
	Actual count	258,516.0	4,477.0	262,993.0
	% change	0.34	-16.34	
Working population who leave to work in Mid-day ($W_{Mid\text{-}day}$)	Expected	71,743.8	1,490.2	73,234.0
	Actual count	70,939.0	2,295.0	73,234.0
	% change	-1.12	54.01	
Working population who leave to work in the Evening ($W_{Evening}$)	Expected	23,764.4	493.6	24,258.0
	Actual count	23,684.0	574.0	24,258.0
	% change	-0.34	16.29	
Working population who leave to work in early AM ($W_{Early\ AM}$)	Expected	33,588.3	697.7	34,286.0
	Actual count	33,599.0	687.0	34,286.0
	% change	0.03	-1.53	
Total (C_j)	Expected	386,738.0	8,033.0	394,771.0
	Actual count	386,738.0	8,033.0	394,771.0
Valid cases (N)		394,771.0		
Pearson Chi-square test statistic		603.07		
Degrees of freedom		3.0		
p-value		0.00		

Table 5 Contingency table between working population who leave to work at different time intervals and LIHCO status

Table 6 shows the Pearson correlation coefficients among different LIHCO population ratios, frequency per capita ratios, average travel time ratios, and working population ratios throughout different times of day. In addition to LIHCO ratios, both Low Income (LI) population ratios and High Car Ownership (HCO) population ratios are also presented in the table. The level of detail for the data in the table is block group. The analysis is run on 403 block groups in New Haven County, Connecticut.

The significant correlations are marked in the table. From table 6, the normalized count of the LIHCO population in the block groups does not generally show significant association with either frequency per capita or in-vehicle travel time ratios.

LI population ratios and HCO population ratios, however, show a direct and an inverse relationship with frequency per capita ratios for different times of day respectively. In other words, block groups with higher population of low income population are served by more frequent transit service and block groups with lower population of high-car ownership families are less served by high frequent service. This distribution of bus service seems to properly match the socioeconomic status of population since, contrary to people with more than one car, low income households presumably have fewer cars and are more dependent on transit hence need more frequent service.

LI population ratios and HCO population ratios also show similar association with in-vehicle travel time ratios. In other words, block groups with higher population of low income population are served by longer transit service and block groups with higher population of high-car ownership families are served by shorter transit service. Knowing that low income people typically take shorter trips (*4*), this outcome might be a sign that low income block groups are not properly served in terms of transit travel times. HCO populations on the other hand seem to have better transit service in terms of the time they have to spend in the buses to get around.

Table 6 also helps to statistically identify whether transit supply and employment demands of working population correlate. In other words, it helps to investigate the levels that transit supply might meet the employment demands of worker population. To identify such associations, all pair-wise correlation coefficients of working population ratios and frequency per capita ratios or average travel time ratios are considered for each time interval.

Among all these correlations, the association between employment demand and bus frequency during AM hours (6:00 – 8:59am) and during Evening hours (4:00 – 11:59pm) is counterintuitive. For both stated time intervals, the association is negative suggesting that there is less frequent bus service available for block groups with higher needs for transportation. This might be a sign for mismatch between transportation demands of AM and Evening workers (i.e. those who go to work from 6:00am to midnight) and bus service supply (more specifically bus frequency).

	S_{AM}	$S_{Mid\text{-}day}$	$S_{Evening}$	$S_{Early\ AM}$	T_{AM}	$T_{Mid\text{-}day}$	$T_{Evening}$	$T_{Early\ AM}$	D_{AM}	$D_{Mid\text{-}day}$	$D_{Evening}$	$D_{Early\ AM}$
LIHCO	.092	.103*	.088	.105*	.114*	.096	.012	-.034				
LI	.206**	.216**	.205**	.233**	.193**	.181**	.106*	.067				
HCO	-.214**	-.209**	-.212**	-.255**	-.134**	-.148**	-.133**	-.173**				
S_{AM}									-.109*			
$S_{Mid\text{-}day}$										.117*		
$S_{Evening}$											-.133**	
$S_{Early\ AM}$												.120*
T_{AM}									-.120*			
$T_{Mid\text{-}day}$										.063		
$T_{Evening}$											.066	
$T_{Early\ AM}$												-.010

* Significant at 0.05 level
** Significant at 0.01 level
LIHCO_ Low Income and High Car Ownership normalized by block group population
LI_ Low Income normalized by block group population
HCO_ High Car Ownership normalized by block group population

Table 6 Bivariate correlation results between different variables

Results of Spatial Statistical Analyses

In this section first the spatial locations and the values of different variables including frequency, travel time, working population, income, and car-ownership data are presented and later the spatial interactions of them will be discussed. Figure 6 shows the location, the proportion of low-income (LI) population, and the proportion of high car ownership (HCO) population of 403 block groups under study in New Haven County. The low income density values are presented by different shading while high car ownership density values are shown by block group heights.

As figure 6 shows, low income block groups are mostly located in the central area (City of New Haven). However, high car ownership block groups are mostly scattered throughout the county in the cities of Woodbridge, North Haven, and Wallingford. One noticeable observation from figure 6 is that there is little overlap between low income block groups and high car ownership block groups, meaning that most of the low income population own not many cars and vice versa.

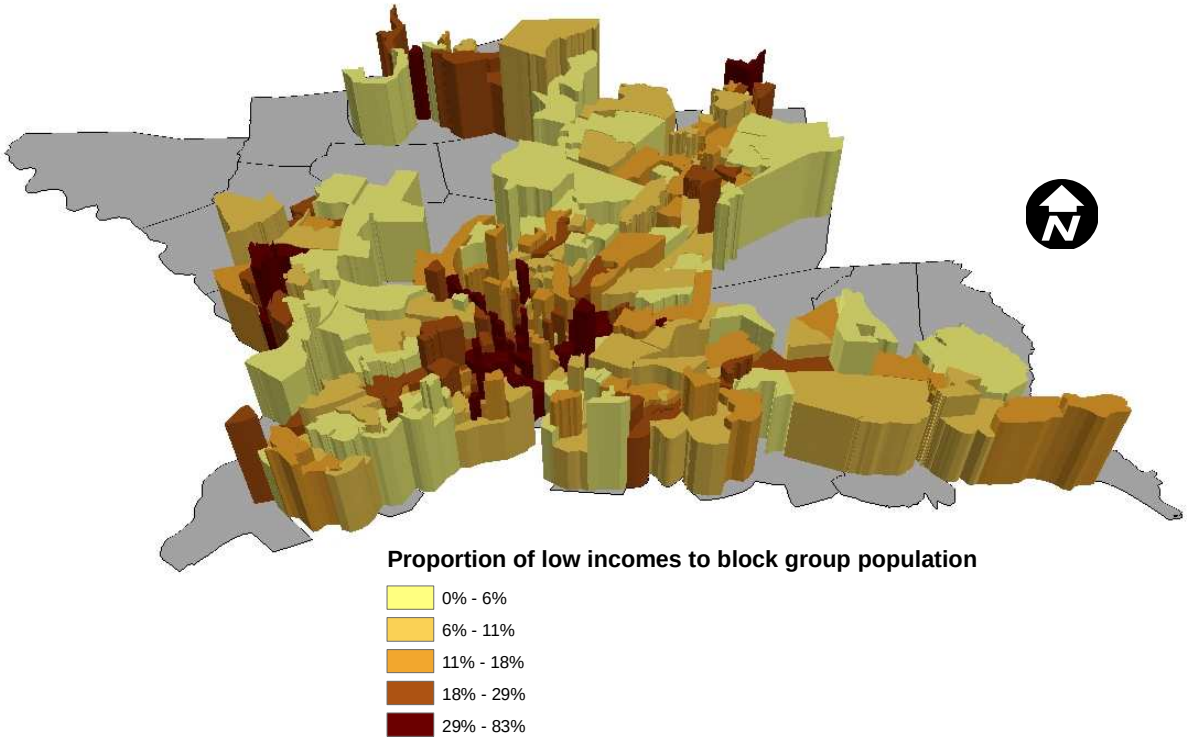


Figure 6 Low income population and high car ownership population in New Haven County

Figures 7 and figure 8 show hourly available seats per capita ratios and average travel time ratios for different time intervals. From figure 7, more frequent bus service areas are located in the central parts and north-south corridor than east-west corridor. Cities such as Orange, West Haven, New Haven, North Haven, and Wallingford are more frequently served with bus service throughout the day.

Among all time intervals, evening hours are best served and early AM hours are least served with bus service. There are up to 8 seats per capita for some block groups of New Haven County during one evening hour. On the other hand and in the best case, there is as few as one seat for three persons for each block group during one hour of early AM service.

Apparently, this comparison fails to portray the whole picture of frequency of bus service in New Haven County because population of each block group is not a good temporal indicator for different levels of demand for bus service. In other words, higher frequent bus service during the evening hours is probably due to presumed more demand for transit at that timeframe. Proportion of workers commuting at different times of day from each block group might be capable to shed some light on this assumption.

Figure 8 shows average bus travel times from each block group in New Haven County. There is little variability in the ranges of travel times among different times of service although mid-day hours show a little more dispersion in terms of the travel time values with 45 minutes as the average travel time from some block groups.

Travel time ranges of values and geographic dispersion patterns are very alike between AM and Evening hours, which both mostly show higher travel times in Cheshire, Wallingford, Hamden, North Haven, New Haven, and Guilford. The pattern of transit travel times during mid-day and early AM hours are different from AM and evening hours and are more dispersed.

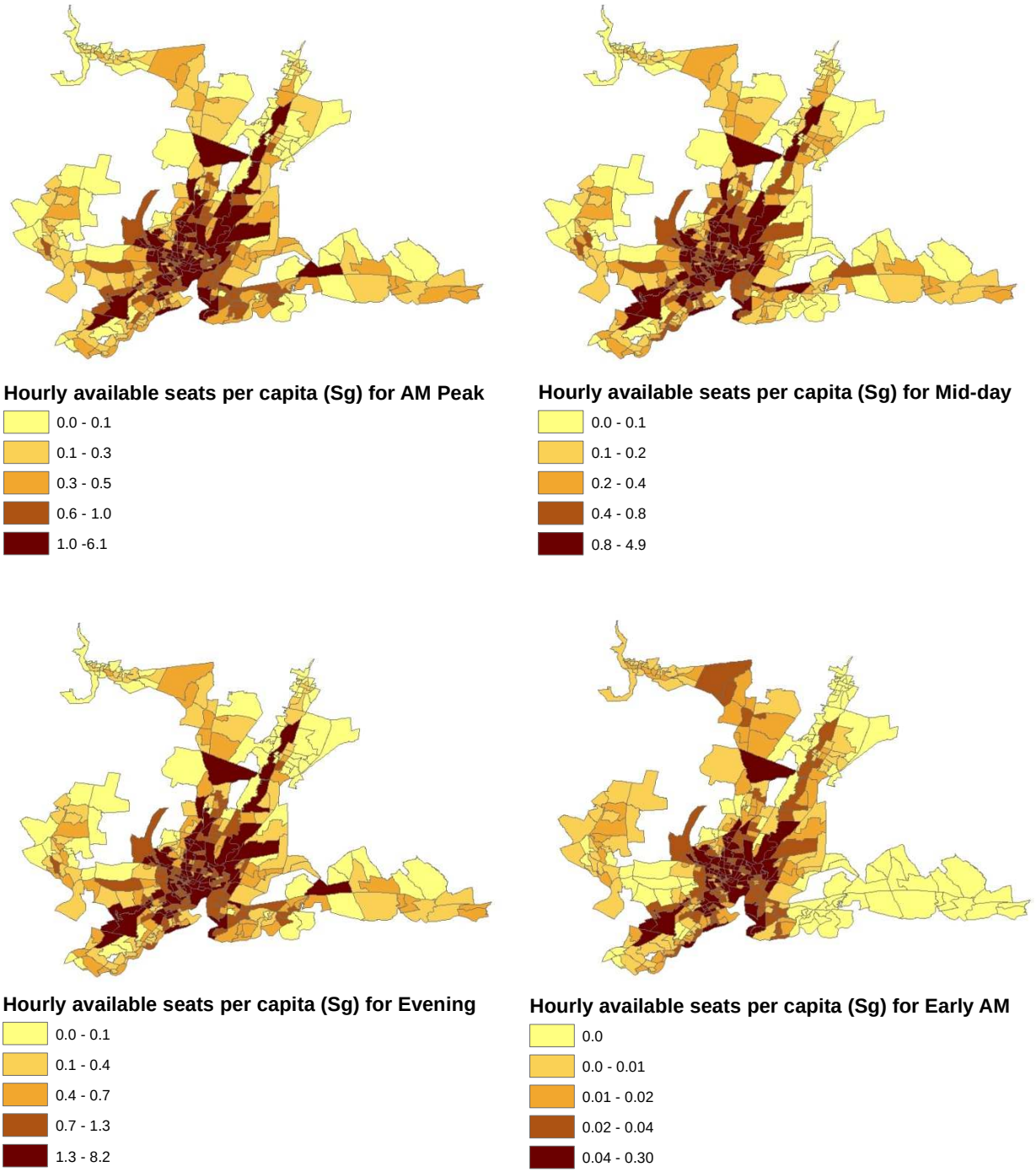


Figure 7 Hourly available seats per capita during different time intervals

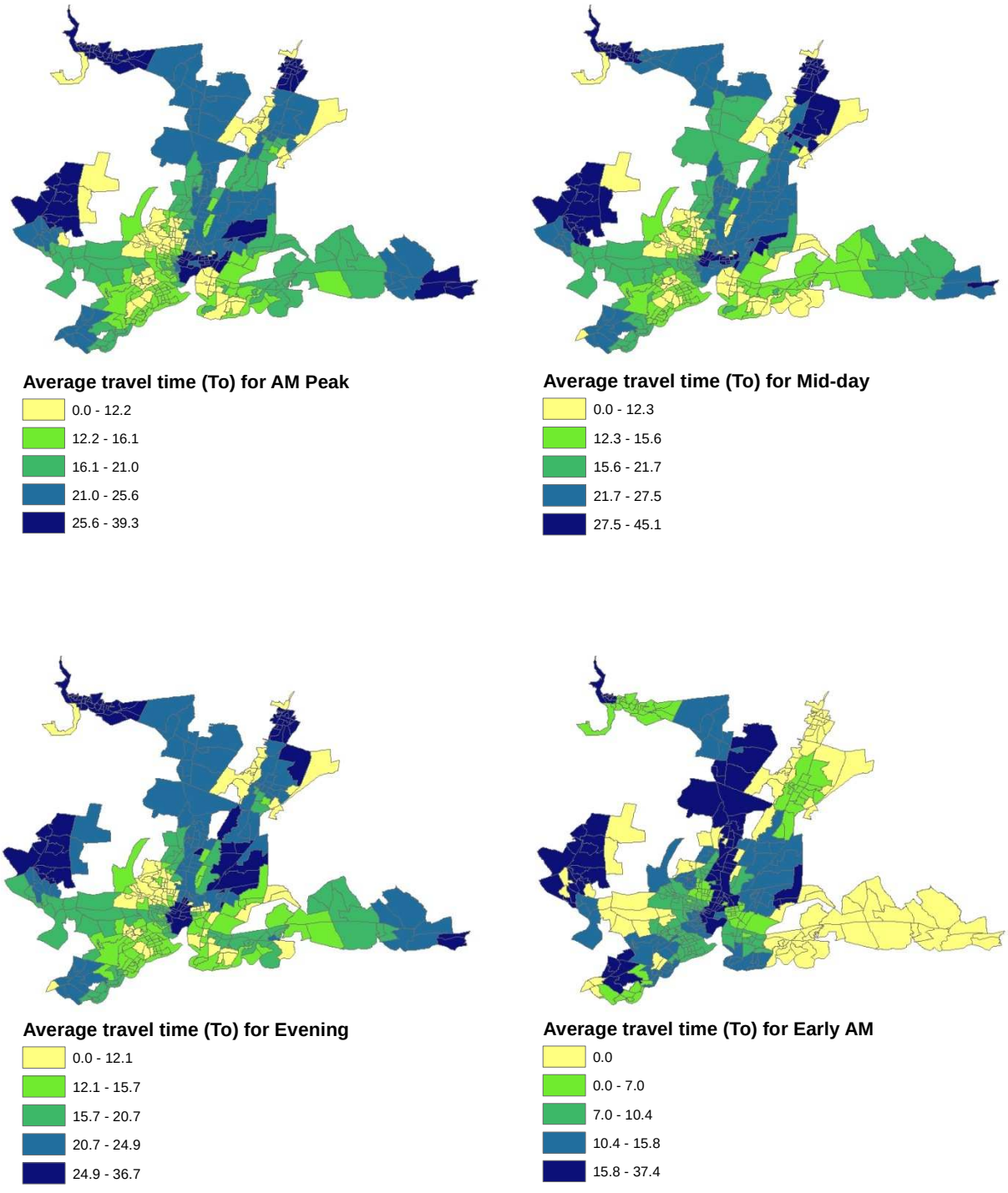


Figure 8 Average travel times from block groups during different time intervals

Figure 9 shows proportion of working population who depart to work during different time intervals. Higher ranges of workers commute during the AM hours than any other time interval during the day followed by mid-day hours. Up to 90% of workers in some block groups such as Hamden, Cheshire, Guilford, Woodbridge, and Orange commute during the AM hours.

One interesting observation is that the ranges of workers who commute during the evening hours is the lowest, however, it was shown that frequency of bus service is the highest for evening hours. This could be suggestive of having excessive transit service than needed for the evening hours. On the other hand, bivariate Pearson correlation analysis showed less frequent bus service is available for block groups with higher needs for transportation during the same period of time. One explanation to the above paradox might be that although the evening time of day is generally served with better transit frequency, the provided service does not match the demand locally. In other words, there is a spatial mismatch between where the buses are available and where workers need the service. Local spatial statistics measures will help to further investigate this assumption.

Another interesting piece of information is the high levels of proportion of working population who depart to work during AM hours. It was also shown than block groups with higher needs for transportation during AM hours are less served by frequent buses (bivariate Pearson correlation results). Altogether, this could be a potential sign for spatial mismatch between transportation needs of AM workers and bus frequency of service although, knowing that LIHCO significantly commute less than non-LIHCO in AM hours, this is not generally a

concern for LIHCO families. Further global and local spatial statistics analysis will shed some light on this assumption.

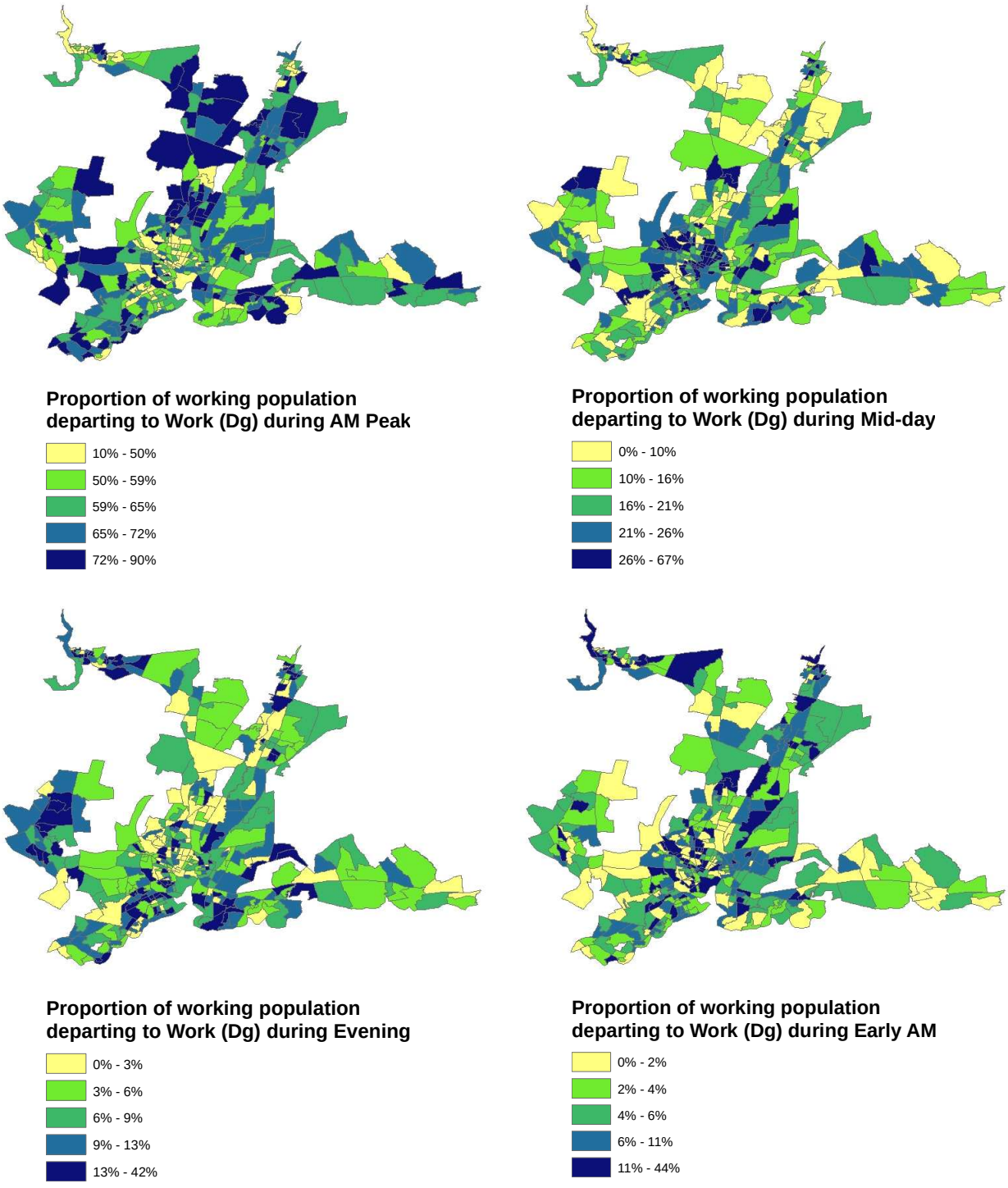


Figure 9 Proportion of working population who depart to work during different time intervals

Table 7 shows the results of bivariate Moran's I between proportion of LIHCO population in block groups and lagged bus frequency ratios as well as between proportion of LIHCO population in block groups and lagged average travel times during different times of day. The weight matrix is created using the 8-nearest neighbors and the z-values are created using 999 permutations.

The results show significant positive autocorrelation at 0.01-level between proportion of LIHCO population and lagged bus travel times during AM and Mid-day hours. This is suggestive of a general cluster pattern. In other words, block groups with high proportion of LIHCO population are surrounded by high travel time ratios and block groups with low proportion of LIHCO population are surrounded by shorter travel times. This can be suggestive of a spatial mismatch between LIHCO residence location and bus travel time for AM and Mid-day hours.

From Pearson Chi-square analysis, LIHCO found to significantly commute less than non-LIHCO in AM hours. Therefore, among those two, spatial mismatch between LIHCO residence locations and travel times during AM hours is less of a concern than that during the Mid-day hours.

The only significant autocorrelation between LIHCO population and lagged average bus frequency is during the early AM hours. However, the positive autocorrelation, which is suggestive of a cluster pattern for bus frequency and transportation demand is intuitive.

		LIHCO				
		Bivariate Moran's I	E(I)	Spatial Pattern	z-value*	Significant**
AM						
	S_g	0.0318	-0.0025	Cluster	1.89	No
	T_o	0.0916	-0.0025	Cluster	5.31	Yes
Mid-day						
	S_g	0.0465	-0.0025	Cluster	2.56	No
	T_o	0.0901	-0.0025	Cluster	5.40	Yes
Evening						
	S_g	0.0352	-0.0025	Cluster	2.08	No
	T_o	0.0213	-0.0025	Cluster	1.25	No
Early AM						
	S_g	0.0633	-0.0025	Cluster	3.66	Yes
	T_o	-0.0391	-0.0025	Dispersion	-2.27	No

* Through randomization with 999 permutations

** At 0.01 level of significance

Table 7 Bivariate Moran's I between proportion of LIHCO population and lagged bus frequency ratios as well as proportion of LIHCO population and lagged average travel

To investigate the spatial dependency of LIHCO population and bus travel times at local level *Moran Cluster Maps* are created. Figure 10 shows Moran cluster maps between proportions of LIHCO population in block groups and lagged average travel times during AM and Mid-day hours. The figure shows that high proportion of LIHCO population block groups that are surrounded by higher bus travel times (High-High) are mostly located in the northern and eastern areas (Waterbury, Seymour, Meriden) for both time intervals. This occurrence and other Low-Low locations have led to a general cluster pattern for LIHCO and bus in-vehicle travel time.

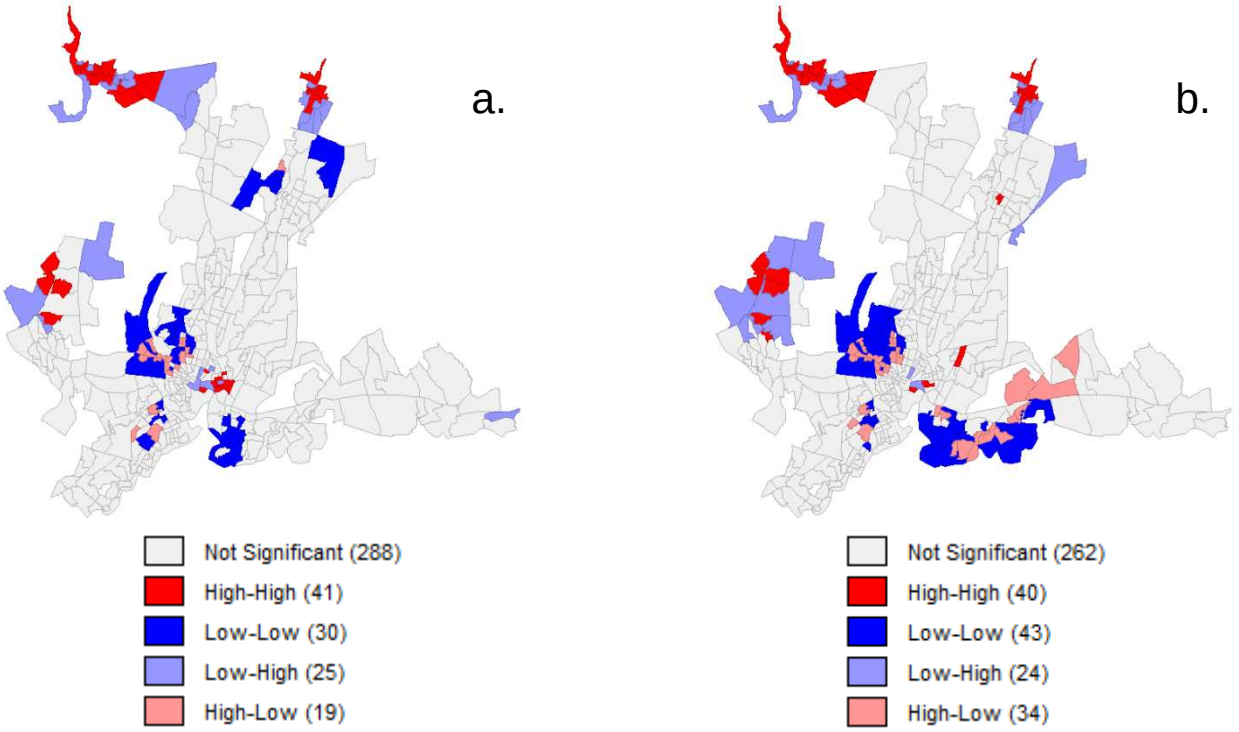


Figure 10 Bivariate LISA between a. LIHCO vs. lagged T_{AM} b. LIHCO vs. lagged $T_{Mid\text{-}day}$

Table 8 shows the results of bivariate Moran's I between proportion of commuting workers and lagged bus frequency ratios as well as between proportion of commuting workers and lagged average travel times during different times of day. The weight matrix is created using the 8-nearest neighbors and the z-values are created using 999 permutations.

The results show significant negative autocorrelation at 0.01-level between proportion of commuting workers and bus frequency ratios during AM and Evening hours. This can be suggestive of a spatial mismatch between employment demands of workers and bus frequency supply during AM and Evening hours. In addition to this recent result, the outcome of both descriptive maps and bivariate Pearson correlation were suggestive of such mismatch. Therefore,

this research finds strong indication of spatial mismatch between proportion of commuting workers and bus frequency ratios during AM and Evening hours.

There is a difference in such mismatch between AM time of day and Evening time of day. As stated before, there is well-frequent bus service available for workers during the Evening hours, but the problem lies in the spatial distribution of bus service for Evening hours since high proportion of commuting workers are surrounded by lower bus frequency ratios (High-Low) and vice versa (Low-High) for Evening hours.

For AM hours, there are both temporal shortage and spatial mismatch of bus service. In other words, neither frequent bus service is provided for high volume of departing commuters nor the provided service is distributed in accordance with the spatial needs of workers for AM hours. Again, since LIHCO households found not to commute significantly during the AM hours, the stated mismatch between transportation demands of workers and bus frequency during AM hours in not a concern for LIHCO families.

Table 8 also shows significant positive autocorrelation at 0.01-level between proportion of commuting workers and bus travel times during Evening hours, which is counter intuitive.

		Bivariate Moran's I	E(I)	Spatial Pattern	z-value*	Significant**
D_{AM}	S_{AM}	-0.1336	-0.0025	Dispersion	-7.53	Yes
	T_{AM}	-0.0613	-0.0025	Dispersion	-3.55	Yes
$D_{Mid\text{-}day}$	$S_{Mid\text{-}day}$	0.1663	-0.0025	Cluster	9.86	Yes
	$T_{Mid\text{-}day}$	0.0168	-0.0025	Cluster	0.95	No
$D_{Evening}$	$S_{Evening}$	-0.0861	-0.0025	Dispersion	-4.90	Yes
	$T_{Evening}$	0.0740	-0.0025	Cluster	4.23	Yes
$D_{Early\ AM}$	$S_{Early\ AM}$	0.0801	-0.0025	Cluster	4.5	Yes
	$T_{Early\ AM}$	-0.0418	-0.0025	Dispersion	-2.45	No

* Through randomization with 999 permutations
** At 0.01 level of significance

Table 8 Bivariate Moran's I between proportion of commuting workers and lagged bus frequency ratios as well as between proportion of commuting workers and lagged average travel times

Figure 11 shows Moran cluster maps between proportions of commuting workers and lagged bus frequency for AM and Evening hours. The figure shows that, during Am hours, high proportion of commuting workers that are surrounded by lower bus frequency ratios (High-Low) are mostly located in northern areas of New Haven County such as Cheshire, Wallingford, Meriden, and Waterbury and during Evening hours are mostly located in northern and western parts of the County such as Seymour, Ansonia, Derby, and Waterbury.

According to figure 10, towns of Meriden and Waterbury have also had high population of LIHCO surrounded by long bus travel trips during AM hours, so it seems that those towns are not properly served by bus service during the AM hours.

Since the temporal bus frequency of service for AM hours is plenty, redistribution of bus service from Low-High areas to High-Low areas within AM hours could be considered. As figure 11.a shows, some parts of the City of New Haven show *Low-High* pattern so could be good candidates for such redistribution.

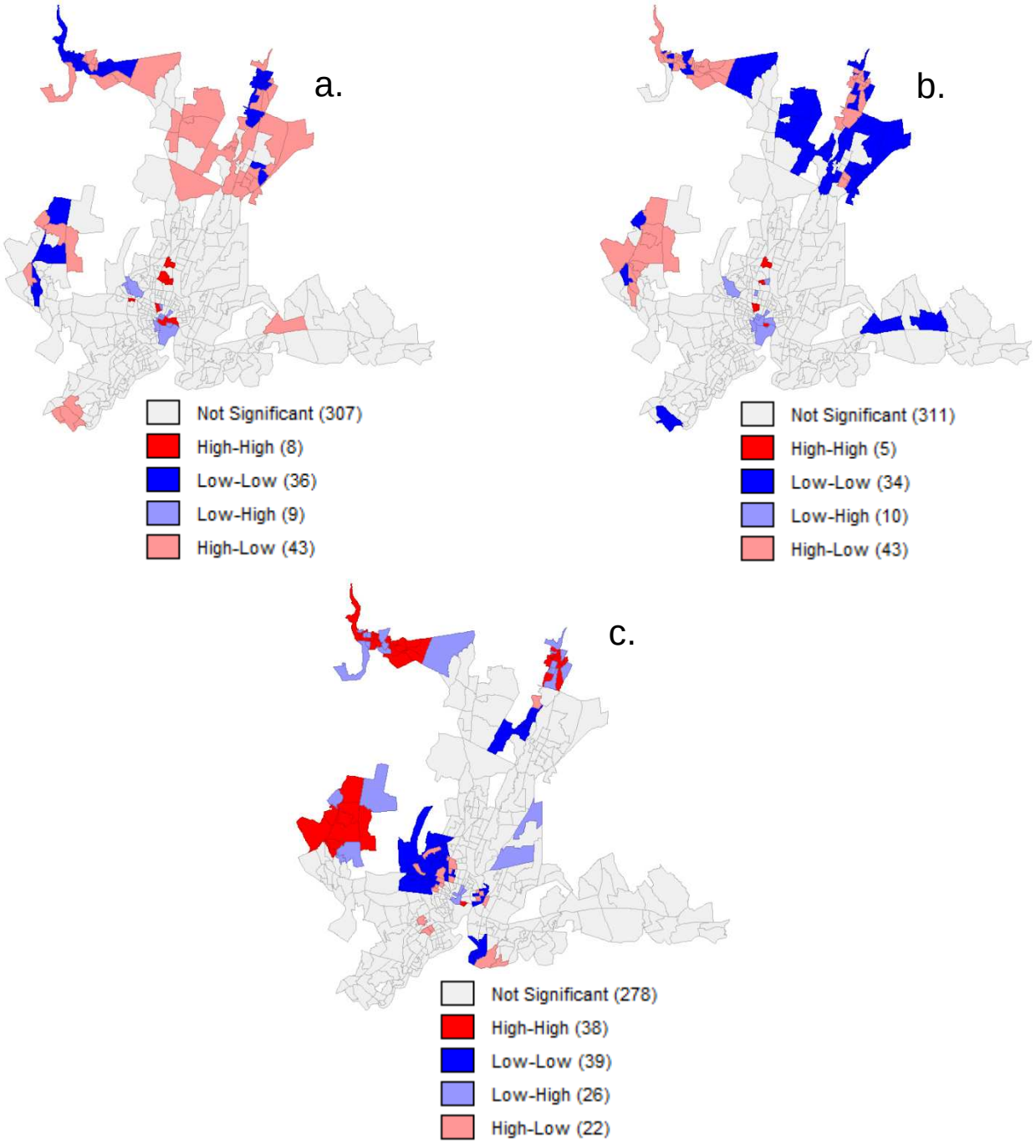


Figure 11 Bivariate LISA between a. D_{AM} vs. lagged S_{AM} b. $D_{Evening}$ vs. lagged $S_{Evening}$ c. $D_{Evening}$ vs. lagged $T_{Evening}$

CHAPTER 6: CONCLUSIONS

INTERPRETING THE RESULTS

Results

This study was intended to research the levels that transit supply meets the employment demands of worker population in general and Low Income and High Car Ownership (LIHCO) population in specific. It identified LIHCO households as a sub-population of low income households that are forced to own more cars to get around (as a consequence of their residential location choice) in the face of their limited income so finds them at more immediate needs for transit service. The primary objective for this study was to investigate whether and where the local bus service in New Haven County, Connecticut meets the demands for commute during different time intervals over a course of a day.

Examining the spatial mismatch, this research utilized global and local spatial statistics techniques in conjunction with conventional statistics. The results of Pearson Chi-square test for goodness of fit suggest that LIHCO individuals do have different temporal employment demands from non-LIHCO individuals. Generally, LIHCO populations were expected to depart to work more during 6:00-9:00AM than actual counts. This finding complies with other literature that finds low-income workers less likely to commute during the AM peak (*4*). The table also shows that LIHCO populations relatively depart to work more during mid-day and evening hours than what was expected. The proportion of the LIHCO population that departs to work during early AM hours is not that different from what was expected.

The primary results of bivariate Pearson correlation did not show significant association between proportion of LIHCO population and either *bus frequency ratios* or *average bus travel times*. However, it showed negative correlations between proportion of workers departing to work either in AM hours or Evening hours and bus frequency ratios at the same time intervals. This suggests that workers who go to work in AM hours or Evening hours are inappropriately served with less frequent bus service, so it might be an indication for mismatch.

Creating descriptive maps from different bus frequency, bus travel time, and time of departure to work variables revealed interesting observations. One of them is that the ranges of workers who commute during the evening hours is the lowest, however, another descriptive map showed that frequency of bus service is the highest for evening hours. This could be suggestive of having excessive transit service than needed for the evening hours. On the other hand, bivariate Pearson correlation analysis showed less frequent bus service is available for block groups with higher needs for transportation during the same period of time. One explanation to the above paradox might be that although the evening time of day is generally served with better transit frequency, the provided service does not match the demand locally. In other words, there is a spatial mismatch between where the buses are available and where workers need the service in evening hours.

The results of bivariate Moran's I also showed significant negative autocorrelation between proportion of commuting workers and bus frequency ratios during evening hours. Therefore, this research finds strong indication of spatial mismatch between proportion of commuting workers and bus frequency ratios during evening hours.

It was shown that, during Evening hours, high proportion of commuting workers that are surrounded by lower bus frequency ratios (High-Low) are mostly located in northern and western parts of the County such as Seymour, Ansonia, Derby, and Waterbury.

Descriptive maps also showed high levels of proportion of working populations who depart to work during AM hours. The results of bivariate Pearson correlation revealed that block groups with higher needs for transportation during AM hours are less served by frequent buses. Altogether, this could be a potential sign for spatial mismatch between transportation needs of AM workers and bus frequency of service. The results of bivariate Moran's I showed significant negative autocorrelation between proportion of commuting workers and bus frequency ratios during AM hours. Therefore, this research finds strong indication of spatial mismatch between proportion of commuting workers and bus frequency ratios during AM hours. However, knowing that LIHCO significantly commute less than non-LIHCO in AM hours, this is not generally a concern for LIHCO families.

It was shown that during Am hours, high proportion of commuting workers that are surrounded by lower bus frequency ratios (High-Low) are mostly located in northern areas of New Haven County such as Cheshire, Wallingford, Meriden, and Waterbury.

The results of bivariate Moran's I show significant positive autocorrelation between proportion of LIHCO population and lagged bus travel times during AM and Mid-day hours. In other words, block groups with high proportion of LIHCO population are surrounded by high travel time ratios and block groups with low proportion of LIHCO population are surrounded by

shorter travel times. This can be suggestive of a spatial mismatch between LIHCO residence location and bus travel time for AM and Mid-day hours.

The Moran cluster maps showed that high proportion of LIHCO population block groups that are surrounded by higher bus travel times (High-High) are mostly located in the northern and eastern areas (Waterbury, Seymour, Meriden) for both time intervals.

Practical Implications

This research finds indications of mismatch between bus frequency supply and employment demands of people in New Haven County during Evening and AM service hours. Therefore, this study suggests bus frequency of service especially during Evening and AM hours is worth investigating. Some northern parts of New Haven County such as Waterbury show high demand for work during AM and Evening hours and poor bus frequency so deserve more consideration.

The study also finds indications of mismatch between proportion of LIHCO population and bus travel times during AM and Mid-day hours. Some areas such as Waterbury show higher proportion of LIHCO population and higher travel times so consequently deserve more attention.

Possible Takeaways for Readers

Spatial statistics techniques showed improvements to the conventional statistical techniques while working with geographical phenomena. Capturing the latent spatial dependency among variables and exposing such correlation to the analyses helped the methodology. Among spatial statistics techniques, local measures such as Local Indicators of Spatial Association (LISA) helped to answer where questions in addition to general whether questions.

Synthetic population generators such as PopGen showed remarkable capabilities to produce high-detailed data with good estimates. Using these models is especially helpful when a cross-tabulation of different variables at finer levels is needed. Census Public Use Microdata Samples (PUMS) are valuable sources for detailed socioeconomic data and can be used as sample seed in synthetic population generators.

CHAPTER 7: RECOMMENDATIONS FOR FUTURE STUDY

FUTURE RECOMMENDATIONS

Limitations of the Used Method/Model

Synthetic population generators such as PopGen enable cross-tabulation of different variables up to person level, which is very beneficial. However, not all areal units have significant variable estimation. In this research, out of 608 block groups in New Haven County, only 474 had significant variable estimation by PopGen.

The selected methodology for calculating the accessibility to a transit station was to consider a 400-meter buffer area around each bus stop as the distance people walk to reach bus service. In reality, however, people might walk longer distances or use other modes than walking such as bicycling or driving (park and ride stations) to reach transit stops. One implication of selecting such methodology was to exclude 199 block groups out of 608 total block groups in New Haven County since they did not have access to bus stops based on 400-meter buffer area imposition.

Because of ease in data acquisition, this research selected bus frequency and bus travel times as the supply-side variables into transit accessibility calculations. These two performance measures might be calculated directly from the GTFS feeds so are publicly available. However, in order to draw a complete picture of a transit service, more performance measures are needed.

Ways to Expand

In addition to travel time and transit frequency of service, there are other performance measures to characterize a transit service such as passenger load and reliability. Future research might consider more performance measures to improve the argument.

This analysis summed all different variables including transit travel time, transit frequency, and proportion of workers at different time interval over the origin block groups. Since all other socioeconomic characteristics of assumed transit users are also stored in block groups, this method provides a simple procedure to further analyze service performance measures and riders' characteristics in one place and at the block group level. In the process of averaging transit performance measures for origin block groups, some useful data might be lost, so a general improvement to this method is to skip such aggregation and perform route-based analysis.

New Datasets

Performing route-based analysis need enriched data that track riders from origin to their destination and enable a cross-tabulation of rider socioeconomic characteristics and transit line attributes. Travel surveys provide such detailed data, however they are expensive to obtain. Another economical alternative way is to use OD tables of Longitudinal Employer Household Dynamics (LEHD) data, which include a cross-tabulation of socioeconomic characteristics of both origin and destination up to Census block level. Although these data are not at person level but at block level and there is no trip data associated with them, with some assumptions about transit trips, this method might improve the results from this research.

REFERENCES

1. Gobillon, L., Selod, H., & Zenou, Y. (2007). The mechanisms of spatial mismatch. *Urban Studies*, 44(12): 2401-2427.
2. Kain, J. (1968). Housing segregation, negro employment, and metropolitan decentralization. *The Quarterly Journal of Economics,* 83(2): 175-197.
3. Currie G., et al. (2010). Investigating links between transport disadvantage, social exclusion and well-being in Melbourne – Updated results. *Research in Transportation Economics*, 29: 287-295.
4. Giuliano, G. (2005). Low income, public transit, and mobility. *Transportation Research Record: Journal of the Transportation Research Board,* 1927: 63-70.
5. Fu, L. & Xin, Y. (2007). A new performance index for evaluating transit quality of service. *Journal of Public Transportation,* 10(3): 47-69.
6. Bhat, C. et al. (2006). Metropolitan area transit accessibility analysis tool. TxDOT Project 0-5178: Measuring access to public transportation service
7. Polzin, S. E., Pendyala, R. M., & Navari, S. (2002). Development of time-of-day-based transit accessibility analysis tool. *Transportation Research Record,* 1799: 35-41.
8. Mamun, S. A. et al. (2013). A method to define public transit opportunity space. *Journal of Transport Geography,* 28: 144-154.
9. Hart, N. & Lownes, N. E. (2013). Urban core transit access to low income jobs. *Transportation Research Record: Journal of the Transportation Research Board,* 2357: 58-65.
10. Grengs, J. (2010). Job Accessibility and the modal mismatch in Detroit. *Journal of Transport Geography,* 18: 42-54.

11. Alam, B. M. (2009). Transit accessibility to jobs and employment prospects of welfare recipients without cars: A study of Broward County, Florida, using geographic information systems and econometric model. *Transportation Research Record: Journal of the Transportation Research Board,* 2110: 78-86.
12. Lee, J. (2013). Perceived neighborhood environment and transit use in low-income populations. *Transportation Research Record: Journal of the Transportation Research Board,* 2397: 125-134.
13. Sanchez, T. W. (1999). The connection between public transit and employment: The cases of Portland and Atlanta. *Journal of the American Planning Association,* 65(3): 284-296.
14. Yi, C. (2006). Impact of public transit on employment status: Disaggregate analysis of Houston, Texas. *Transportation Research Record: Journal of the Transportation Research Board,* 1986: 137-144.
15. Diao, M. (2014). Selectivity, spatial autocorrelation and the valuation of transit accessibility. *Urban Studies,* 52(1): 159-177.
16. Kawabata, M. & Shen, Q. (2007). Commuting inequality between cars and public transit: The case of the San Francisco Bay Area, 1990-2000. *Urban Studies,* 44(9): 1759-1780.
17. Wang, C. & Chen, N. (2015). A GIS-based spatial statistical approach to modeling job accessibility by transportation mode: Case study of Columbus, Ohio. *Journal of Transport Geography,* 45: 1-11.
18. Griffin, G. P. & Sener, I. N. (2014). Equity analysis of transit service in large auto-oriented cities in the United States. 94th Annual Meeting of the Transportation Research Board.

19. Holzer, H. J. (1991). The spatial mismatch hypothesis: What has the evidence shown? *Urban Studies,* 28(1): 105-122.

20. Ihlanfeldt, K. R. & Sjoquist, D. L. (1998). The spatial mismatch hypothesis: A review of recent studies and their implications for welfare reform. *Housing Policy Debate,* 9(4): 849-892.

21. Hu, L. & Giuliano, G. (2011). Beyond the inner city: New form of spatial mismatch. *Transportation Research Record: Journal of the Transportation Research Board,* 2242: 98-104.

22. Fischer, M. J. (2003). The relative importance of income and race in determining residential outcomes in U.S. urban areas, 1970-2000. *Urban Affairs Review,* 38(5): 669-696.

23. Banister, D. (1994). Internalising the social costs of transport. OECD/ECMT Seminar, Paris.

24. Gleeson, B. & Randolph, B. (2002). Social advantage and planning in the Sydney context. *Urban Policy and Research,* 20(1): 101-107.

25. Gao, S. & Johnston, R. A. (2009). Public versus private mobility for low-income households. *Transportation Research Record: Journal of the Transportation Research Board,* 2125: 9-15.

26. Kim, H. S. & Kim, E. (2004). Effects of public transit on automobile ownership and use in household of USA. RURDS, 16(3): 245-262.

27. Brown, M. (1996). Keeping score: Using the right metrics to drive world class performance, Quality Resources, New York, NY.

28. Kittelson & Associates, Inc. et al. (2013). Transit capacity and quality of service manual, 3rd edition.

29. Grengs, J. (2004). Measuring change in small-scale transit accessibility with geographic information systems: Buffalo and Rochester, New York. *Transportation Research Record: Journal of the Transportation Research Board,* 1887: 10-17.

30. Alam, B. M., Thompson, G. L., & Brown, J. R. (2010). Estimating transit accessibility with an alternative method. *Transportation Research Record: Journal of the Transportation Research Board,* 2144: 62-71.

31. Hendrickson, C. & Plank, E. (1984). The flexibility of departure times for work trips, *Transportation Research Part A,* 18A(1): 25-36.

32. McKenzie, B. & Rapino, M. (2011). Commuting in the United States: 2009, American Community Survey Reports, ACS-15. U. S. Census Bureau, Washington, DC.

33. Ye, X. et al. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic population. 88th Annual Meeting of the Transportation Research Board.

34. Anselin, L. (1995). Local indicators of spatial association – LISA. Geographical Anlysis, 27: 93-115.