Modeling Transit Dependency Index And The Analysis on The Intersecting Transit-Dependent Groups — A Spatial Microsimulation Approach

CP 6950 Capstone Project

Jian Pang

Abstract

This research is primarily focused on building a methodology framework to model a Composite Transit Dependency Index (CTDI) that incorporates various Transit-Dependent groups. The application of Spatial Microsimulation in this research helps better identify intersecting demographic groups that contribute to the overall Transit Dependency of an area. By performing Multivariate Linear Regression, the TDIs are also found to be able to predict the number of outbound trips of a census tract to some level of extent. And the results of the regression can be used into forming the Composite Transit Dependency Index.

Existing Research Review

Transit Dependency

It is widely accepted that in transit planning, riders are categorized as either "choice riders" and "captive riders"[1]. The choice riders refer to those who have alternatives besides transit for their traveling purposes, and they choose transit out of preference. The captive riders refer to those that do not have a private vehicle available or cannot drive (for any reason) and who must use transit to make a desired trip"[2]. Classical research findings show that the two groups significantly distinct from each other in the willingness to pay for the transit service, while the choice rider is more sensitive, and the latter is less sensitive[3].

Recent study has shown that the rigid segmentation of "captive riders" and "choice riders" cannot precisely describe the passenger's choice behavior, as it is influenced not only by the alternatives, but also their socioeconomic situations that blur the boundary between the two groups. Not all captive riders are in a situation where they have to rely solely on transit, but they also have other alternatives that fit their needs as well. And price sensitivity has been proved not to be the major distinction between the two groups. People tend to share similarities when making choices on transit across the two groups, which starts to blur the boundary between the two. A more detailed market segmentation method has been used in some research to reveal the choice behavior of transit passengers. A new way of categorization was created between transit riders and non-transit riders. Both choice riders and captive riders are in the transit riders category, and similarities are found between the two subgroups: they prize reliability, travel time,

type of service, and comfort. Among choice riders, there is also an overlapping part with the non-transit riders because they value more about safety and comfort about the service[4].

A more detailed segmentation in ridership is crucial to understanding the demands for transit. But the definition on transit riders should be inspected also from the dimension of transit supply. Research has also discovered that the “captive ridership” is substantially correlated with urban development density. Higher urban density means higher transit ridership that can support a large network service to provide accessibility by transit to a larger range of activities. Both higher residential densities and larger markets support being able to provide quality transit service because they tend to reduce the access time and enable more frequent service, which reduces the wait times. However, the growth in overall socioeconomic status and the shift in urban land-development trends indicate both significant declines in transit dependency and continued sprawling development[5]. Therefore, transit development faces a growing challenge to provide the needs for lower-density development and appeal to both choice and captive riders to become active transit users. Another consequence of the low-density urban development is that a large socioeconomically disadvantaged population are placed in the area where there is a gap between transit supply and transit demand. These population lack adequate public transit service are deemed Transit-Dependent population[6].

American Public Transportation Association (APTA) defines Transit-Dependent Population as people in the transit-dependent market have no personal transportation, no access to such transportation, or are unable to drive. Included are those with low incomes, the disabled, elderly, children, families whose travel needs cannot be met with only one car, and those who opt not to own personal transportation[7].

APTA lays out a large framework in researching and modeling transit dependency, but existing research has adopted more flexible ways to measure transit-dependent population according to their respective focus. Chandra et al.(2005) have defined the subgroups of transit riders based on the spatial and temporal characteristics of their travel behaviors. In their effort to identify the transit-dependent group, they propose not only American Community Survey data to extract socioeconomic characteristics, but also the National Travel Survey Data to extract the trip purpose and trip destination, which makes it a more holistic approach. Junfeng J, in his research on identifying transit deserts in Texas, uses the formula that was created by U.S. Department of

Transportation to calculate the transit-dependent population. This approach focuses particularly on the group with no vehicle ownership and the underaged population[5].

A. *Household drivers = (population age 16 and over) – (persons living in group quarters)*
B. *Transit-dependent household population = (household drivers) – (vehicles available) * national level carpooling ratio*
C. *Transit-dependent population = (transit-dependent household population) + (population ages 12–15) + (non-institutionalized population living in group quarters)*

Fang and Thomas (2013) map out the demographic groups that demand the public transit in Miami-Dade County using a framework closest to the ATPA's. They are able to identify and locate four demographic groups that demand transit service using four criteria: Housing Affordability, Employment, Household Income and Transit Coverage [8].

While it is possible to identify multiple transit-dependent demographic groups using various data source, the identification scheme treats each group as an independent group, and overlooks the sharing characteristics and even the overlapping in population across different groups. It is also difficult to position "transit-dependent population" within a certain range of the "transit rider to non-transit rider" spectrum as is discussed above, since they vary extensively on socioeconomic characteristics as well as personality. Under the definition given by APTA, it is still unclear how to exactly measure the level of transit dependency of an area because it does not provide a composite index.

Multidimensional Transit Dependency

While APTA's definition on Transit Dependency already covers extensive groups of population, it is still limited to using one criteria at a time to do the population searching. Very often, researchers and policy makers want to know how different demographic characteristics co-influence people's behavior. And although that can be achieved using Linear Regression by using multiple demographic characteristics as independent variables, the coefficients in the result are only telling what is the aggregated influence of multiple independent variables on the dependent variable. The "aggregated influence" is different from the "co-influence". A simple example would be the Mathematical concept of "The Set", where Set A ∪ Set B is equivalent to

the aggregated influence, and Set A ∩ B is the co-influence. The co-influence of multiple demographic characteristics can only be explored by individual observations, which means the researcher needs to do the sampling or a full survey, whereas exploring the aggregate influence only requires the statistics for the whole group.

In social research, the co-influence of socioeconomic factors is essential in analyzing the socioeconomically "vulnerable population". The concept of "vulnerable population" is put forward by the World Health Organization in 2002, which describes a demographic group being unable to anticipate, cope with, resist and recover from the impacts of disasters. Multiple research on public health has discovered that the vulnerable population across different demographic groups often overlap each other. A report on overlapping vulnerability of the immigrant workers that was released by American Society of Safety Engineers (ASSE) and National Institute for Occupational Safety and Health (NIOSH) discover that Hispanic immigrants, especially young immigrants are subject to higher risk in occupation due to multiple social vulnerability factors: language, age, education, etc [9]. Transit-dependent population also fit in the vulnerable group concept as they are defined by both endogenous factors such as sex and race, which are usually personal characteristics, but also exogenous factors such as poverty, employment status, which are usually socioeconomic characteristics. Verbich and El-Geneidy (2016) have discovered that fare vendors in neighborhoods with low median household income and/or with a high proportion of unemployed residents are predicted to sell more weekly fares than vendors in neighborhoods with high household income and low rates of unemployment [10]. This research proves that in those neighborhoods, transit-dependent population are subject to both income and employment disadvantages.

The concept of "intersectionality" is a framework to analyze of socially marginalized group under an interlocking social stratification system. This framework is first applied in feminism research to investigate the oppression of women of minority groups, and has now been widely adopted by social researchers. Intersectionality has been explained in multidimensional poverty, which interprets poverty not only as in materialistic poverty, but also health-wise and education-wise [11]. Similarly, as is pointed out in vulnerable population research and the APTA definition, Transit-Dependent population are highly likely to face multiple socioeconomic and physical disadvantages at the same time, The goal of this research is to find if there is also intersecting transit dependency, i.e., whether an area that with a population that belong to two or

more socially disadvantaged groups at the same time will have higher transit dependency. To analyze this correlation, a composite transit-dependent index need to be constructed, which is the second goal of this research.

Research Methodology

Research Assumptions

As is explained in the last chapter, the goal of the research is to test if there is an intersecting transit dependency, and how overlapping Transit-Dependent groups affect the overall transit dependency of an area. Two main assumptions are made in order to perform the experiment.

1) A person that belongs to the Transit-Dependent population will have more difficulties in making a trip than a Non-Transit-Dependent person, and therefore, is inclined to make fewer trips on average.
2) A person that belongs to multiple disadvantaged Transit-Dependent groups will have even greater difficulties in making a trip than a person that belongs to a single disadvantaged group, and therefore, is inclined to make fewer trips on average.

Research Area, Research Resolution

I select the 20 counties within the Atlanta Regional Commission (ARC) jurisdiction as the site of the research. Within the 20 counties, there are 10 considered as core ARC counties — Cherokee, Clayton, Cobb, DeKalb, Douglas, Fayette, Fulton, Gwinnett, Henry, and Rockdale, and meanwhile, ARC also serves as the Metropolitan Planning Organization for the rest 10 counties — Barrow, Bartow, Carroll, Coweta, Forsyth, Hall, Newton, Paulding, Spalding, and Walton. The transit service for the whole region is principally provided by Metropolitan Atlanta Rapid Transit Authority (MARTA). The 20 counties are also all within the larger Atlanta Metropolitan Statistical Area, which indicates a certain level of homogeneity from the perspective of economic structure and urban commuting flows.

The resolution of the research is the smallest unit of the analysis. In this research, the resolution is census tract, mainly for the consideration that at the census tract level, there are

more choices in ACS variables than block groups, and at the same time, maintaining a moderately high resolution for a metropolitan region of 20 counties.

Choice of Transit Dependency Variables

To define the transit-dependent population more accurately, I include the variables from APTA definition, as well as adding the race, and employment status to identify more subgroups. Among the variables, race, age and disability are endogenous factors, while poverty, Household vehicle ownership, employment status are exogenous factors (Table 1).

For the race constraint, I only consider if a person belongs to a non-white ethnicity, which already fits the needs of this research, without more specific sub-categorizing. For the age constraint, population between 5 to 17 years old are considered as potentially transit-dependent because most of them can't afford a vehicle yet, while people over age 65 are also considered potentially transit dependent due to the general consideration on their physical activeness. For the poverty constraint, I use the household income to national poverty line ratio, and is divided into three categories: below 0.5, 0.5 to 1 and above 1. For household vehicle ownership, only three categories are provided: no vehicle, 1 vehicle and 2 vehicles and above because the focus is primarily on the auto-less population.

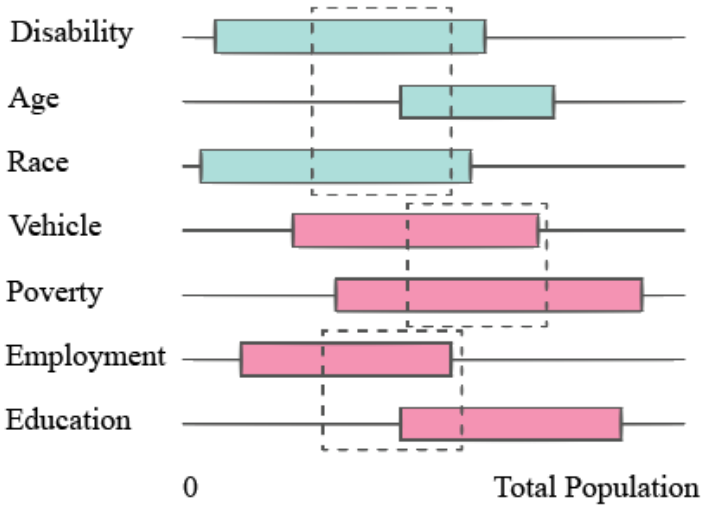
Table 1. Variables Selected for The Research

Endogenous Variable	Race	White Only
		Minority
	Age	Under 5
		5 - 17
		18 - 34
		35 - 64
		Above 65
	Disability	Disabled
		Not Disabled
Exogenous Variables	Poverty (by Household Income to Poverty Line Ratio)	Below 0.5
		0.5 - 1
		Above 1
	Household Vehicle Ownership	No Vehicle
		1 Vehicle
		2 Vehicles

		And Above
	Employment Status	Employed
		Unemployed
	Education Attainment	Below High School
		High School and Above
Intersecting Variables	Vehicle Ownership by Age	
	Vehicle Ownership by Disability	
	Vehicle Ownership by Poverty	
	Poverty by Employment Status	
	Poverty by Race	
	Poverty by Disability by Age	

Cross-joining the variables will generate the intersecting categories (Figure 1). This research is not going to test all the combinations, but only concerned with the cross groups that are likely to be observed in reality. As these variables are divided into personal characteristics and socioeconomic characteristics, I will cross-join the variables on either side. Household vehicle ownership is a crucial factor in transit dependency index, so I will cross-join with the "Age" and "Disability" variables to inspect population that can't properly drive due to their age constraint. Another factor that limits vehicle ownership is poverty, so I am also going to cross-join "Household Vehicle Ownership" with the "Poverty variable". Within socioeconomic variables, variables that are associated in reality can be cross-joined. For example, cross-joining "Poverty" with "Employment Status" will help identify if the population at the bottom of the socioeconomic status; and cross-joining "Poverty" with "Race" will help identify if there are areas that have high poverty rates and meanwhile, with a majority of minority ethnicities. Last but not least, cross-joining "Poverty" with "Disability and Age" will help identify most commonly, the aging communities within impoverished neighborhoods.

Figure 1. Intersecting Demographic Groups (Concept Diagram)



In total, using the 6 initial variables and the 6 cross-joined variables, with each of them representing a transit-dependent group, there are 12 demographic groups.

Demographic Data Source and Spatial Microsimulation

To analyze the intersecting demographic groups, population data that allows cross-table lookup needs to be obtained. American Community Survey census data is widely used in researching the demographic characteristics, which also includes some common cross-table categories, e.g. “age by sex”, “sex by race”, etc. However, it does not allow manually create cross-table categories. The American Community Survey Public Use Microdata Samples (PUMs), however, is the sample data from the survey, and is preserved and can be downloaded in an id-case format as a wide table. The PUMs has both options of choosing 1-year sample or 5-year sample, and the former is a 1 percent sample from each Public Use Microdata Area (PUMA), while the latter is a 5 percent sample by combining multiple 1-year data with appropriate adjustments to the weights and inflation adjustment factors.

With the PUMs data, I apply the spatial microsimulation approach to generate a synthetic population dataset. Spatial microsimulation is “A method for generating spatial microdata — individuals allocated to zones—by combining individual and geographically aggregated datasets.

In this interpretation, 'spatial microsimulation' is roughly synonymous with 'population synthesis' " [12]. Spatial microsimulation takes into the sample data as seeds, and generates a weight for each person in the sample base on the person's characteristics and the corresponding population at the census tract level. Therefore, besides the PUMs data that is used as sample data, spatial microsimulation also requires the marginal data for each Transit-Dependent group. The marginal data is obtained from the ACS census data. The initial 6 Transit-Dependent groups are used as constraints when performing the microsimulation.

Transit Dependency Index Construction

The construction of the Transit Dependency Index is inspired by the Environmental Justice Index (EJ Index) that incorporates the vulnerable population and the impact of environmental hazards using an intuitive formula:

EJ Index *= (The Environmental Indicator) * (Demographic Index for Block Group – Demographic Index for U.S.) * (Population count for Block Group)*

This core idea of this formula is to first, identify a certain demographic group for an area that falls below the national average statistics as the vulnerable population, and second, multiply by the environmental impact assess the combined effect of population size and the scale of the hazard normalized by distance. The demographic index is the half of the sum of percent low-income and percent minority, the two demographic factors that were explicitly named in Executive Order 12898 on Environmental Justice. The demographic index can also be replaced by other demographic indicators such as percent low-income and percent minority.

In measuring transit dependency, I follow the construction of the EJ Index. The variables mentioned in the above section will directly be used to define a Transit-Dependent group. When measuring the intersecting demographic groups, instead of applying the arithmetic mean of the demographic ratio, I can always use the synthetic data to generate the index for the group needed for analysis. Due to the scope of the research that is limited in the Atlanta region, I choose to use the demographic indicator for the State of Georgia, rather than the U.S., for providing a more accurate context when making comparisons. An accessibility index will replace the environmental indicator to be the multiplier in front of the demographic index. This accessibility index is the product of both accessibility to transit stops and the employment access within the ARC 20 counties, which requires the Longitudinal Origin-Destination Employment Statistics

dataset. This will be explained in detail in the Calculating Transit Dependency Index Chapter. With 13 different transit-dependent groups, each census tract, as well as Georgia, will have a Transit Dependency Index (TDI) for each of the groups.

TDI (for One Transit-Dependent group) *= (Accessibility Index) **
*(Demographic Indicator for Census Tract -Demographic Indicator for Georgia) **
(Population Count for Census Tract)

The calculation result of the Transit Dependency Index indicates the level of transit dependency of the census tract. If the index value is negative, it is because the percent of the Transit-Dependent population are lower than the state average, and therefore is less relying on transit. If the index value is positive, it is because the percent of the Transit-Dependent population in the tract is greater than the state average, and therefore, has an excessive Transit-Dependent demand.

Testing for The Intersecting Group Assumption

Based on the research assumptions, in which people are inclined to make fewer trips on average as they fit in more Transit-Dependent groups. Therefore, I use linear regression method to test if the two assumption hold. For the dependent variable, I use the employment Origin-Destination data obtained from the LODES dataset. For the independent variables, I use the sum of the TDI for each of the Transit-Dependent groups.

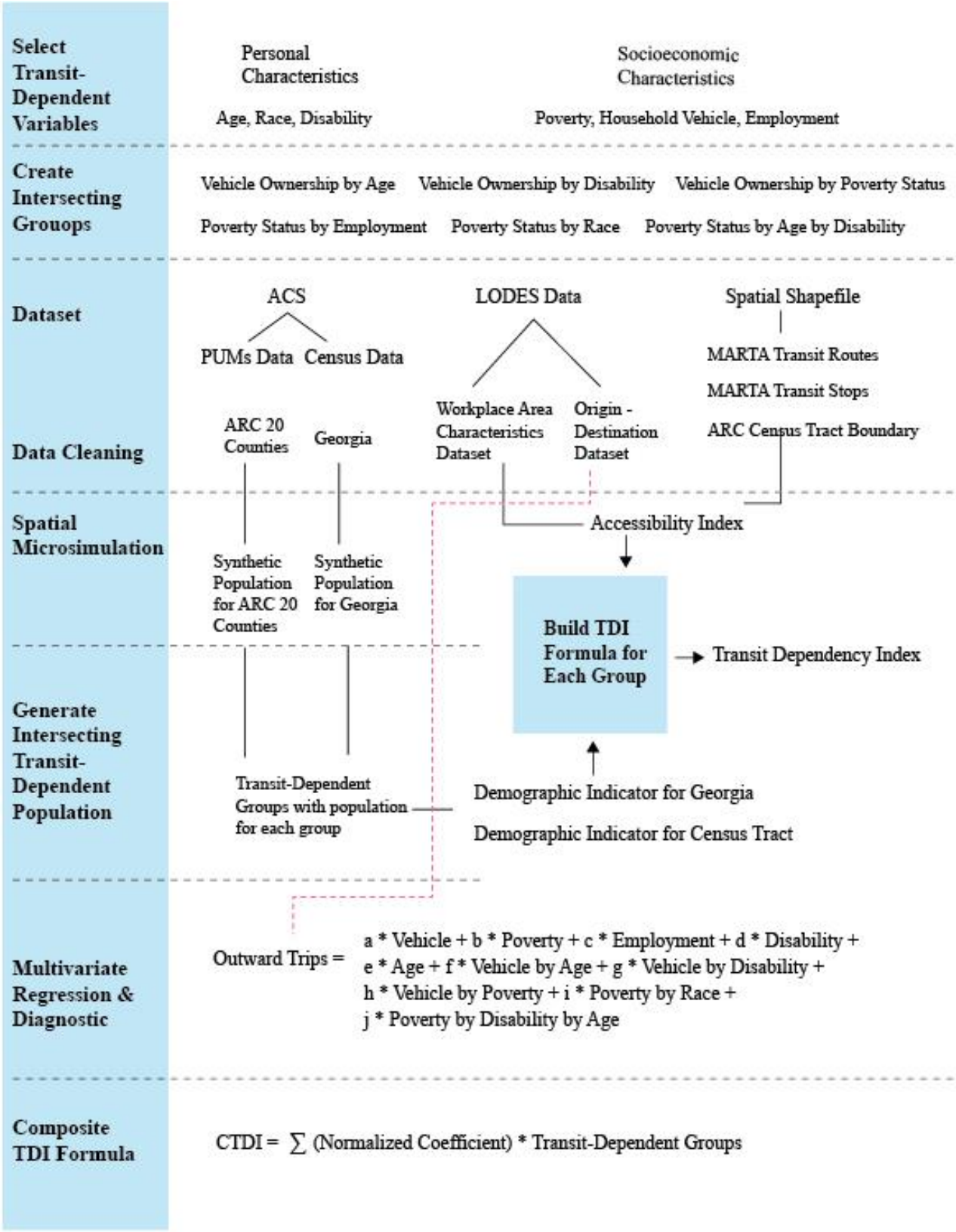
*Number of Outbound Trips = a*TDI (Group 1) +...+ z*TDI (Group n)*

If the regression results show good statistical significance for the Transit-Dependent groups, it means they have an impact on the total trips made from the census tract outwards. The scale of the impact depends on the scale parameter that is associated with each of the TDI variables.

Composite Transit Dependency Index Construction

After the regression is examined, I will build a composite transit dependent index based on the scale parameters generated from the regression by normalizing each of them into fractional numbers that will sum up to 1. Next, a Composite Transit Dependency Index (CTDI) can be generated. Finally, a complete structure of the research methodology is shown in the flowchart below.

Figure 2. Methodology Framework



Spatial Microsimulation

Data Cleaning, Data Restructuring

The ACS PUMs data is downloaded using the "lodown" package in R, and the ACS census data is downloaded using the "tidyverse" package in R. Next, both datasets will be filtered for Georgia as well as for all the 20 counties in ARC. They will also be filtered using the 6 Transit-Dependent groups variables—Disability, Age, Race, Household Vehicle Ownership, Poverty Status and Employment Status.

Spatial microsimulation is performed in R using the "ipfp" package [13]. The algorithm behind the package has some strict requirement on input data quality. Before performing spatial microsimulation, there are several flaws with both the PUMs data and the census data. Some of the flaws are do not affect data quality, but the others do and could potentially cause the errors to occur in microsimulation.

The first major problem is the "NA"s in the PUMs data. The algorithm will not take NA value unless it's been re-coded to other values. The second major problem is the conditional variable in the census data that does not cover the whole population. For example, Education Attainment in the census dataset only records for population 18 and above. This can be problematic because firstly, the total marginal population across different variables has to be consistent (e.g., the total population across different age group should be the same as the total population for the disabled and non-disabled people), or the algorithm will produce abnormal weights. Second, the categories for a certain variable has to be consistent in both the sample data and the marginal data as the simulation algorithm will adjust weights according to the categories. The solutions of the input data flaws are listed in Table 2 and Table 3.

Table 2. Flaws with PUMs Data And Solutions

PUMs Data	Solution
NA values	If all the variables for this person is "NA", then remove the record. If "NA" only appears in a certain variable, use Stat Match package to perform imputation.
Some in continuous form, not categorical form	Recategorize those variables into the categories that are discussed in the Methodology Chapter.
Household Vehicle Ownership is not in the person PUMs file, but household PUMs file	Join the household PUMs file to the person PUMs file based on the unique identifier for each sampled household and obtain the vehicle ownership for person.

Table 3. Flaws with Census Data And Solutions

Census Data	Solution
Tracts with 0 population or 0 households	Inspect these tracts on the map, and see if these are non-residential tracts. If so, delete these tracts so microsimulation will not be performed on them.
Some variables are conditional variables that don't cover the whole population	Re-code the conditional variable for the remaining population, but also set up a category that matches the criteria correspondingly in the sample data.

Next step in data reconstruction is to expand the sample data, namely converting every category under the same variable into a new variable. The reason for this process is because the marginal data uses every category as an individual variable, and the microsimulation algorithm will take each of them as a constraint (Figure 3). For example, there are 5 categories under the "Age" variable—"Age under 5", "Age 5 to 17", "Age 18 to 34", "Age 35 to 64" and "Age 65 and above". Each category is going to be expanded into a binary variable so that a person's age category will be determined by six variables (Figure 4).

	GEOID	ptotal	htotal	age14	age15_17	age18_24	age25_64
1	13013180103	4697	1486	1072	214	214	1525
2	13013180104	2032	662	550	87	143	756
3	13013180105	2907	853	687	179	195	1008
4	13013180106	2773	912	592	99	280	965
5	13013180107	3904	1264	729	155	425	1466
6	13013180108	2825	902	613	95	160	1106
7	13013180203	3743	1158	776	211	498	1305
8	13013180204	3020	837	805	176	248	770
9	13013180205	4397	1473	1165	53	393	1347
10	13013180206	2864	1036	427	170	302	913
11	13013180301	2268	750	423	134	228	765
12	13013180302	5434	2021	1201	260	439	1643
13	13013180303	4871	1395	1166	248	496	1510
14	13013180401	5724	1925	1436	307	262	1994

Figure 3. Marginal Data

	age14	age15_17	age18_24	age25_64	age65_74	age75
1	0	0	0	0	1	0
2	0	0	0	0	1	0
3	0	0	0	1	0	0
4	0	0	0	1	0	0
5	1	0	0	0	0	0
6	1	0	0	0	0	0
7	0	0	0	1	0	0
8	0	0	0	1	0	0
9	1	0	0	0	0	0
10	1	0	0	0	0	0

Figure 4. Sample Data

Performing Microsimulation

The algorithm behind the "ipfp" package is Iterative Proportional Fitting. Its algorithm is expressed in the formula below.

$$w(i,z,t+1) = w(i,z,t) * \frac{cons\_t(z,c,ind(i,c))}{\sum_{j=1}^{n_{ind}} w(j,z,t) * I(ind(j,c) = ind(i,c))}$$

In the formula, "ind" is a two-dimensional array (a matrix) in which each row represents an individual and each column a variable. The value of the cell ind(i,j) is therefore the category of the individual "i" for the variable "j". cons_t(i,j,k) is the number of individuals corresponding to the marginal for the zone 'i', in the variable "j" for the category "k".

The IPF algorithm will proceed zone per zone. For each zone, each individual will have a weight of "representatively" of the zone. The weight matrix will then have the dimension "number of individual x number of zone". "w(i,j,t)" corresponds to the weight of the individual "i" in the zone "j" (during the step "t"). For the zone "z", we will adapt the weight matrix to each constraint "c". This matrix is initialized with a full matrix of 1 and then, for each step "t" [12].

The microsimulation generates a weight for each person under every constraint (Figure 5). The next step is to check if the weights will sum up to match the marginal population. After aggregating the weights to every census tracts, I check the correlation between the aggregated weights and the marginal population in R, and the Pearson'r value is 0.9998, indicating that it is a successful spatial microsimulation. Nevertheless, the weights generated are fractional, so I perform an integerise process to convert the weights to integer so the weight can represent the number of people in reality. The integerised weights is a vector, and each element is the original

sample id repeated number of times based on their weights. Transform this vector into a dataframe, and join it back to the orignal sample data, I manage to generate a full synthetic population dataset.

Figure 5. Weights Generated by IPF

1	0.00299503999806	0.002397870350	0.004538968683	0.002184153299	0.006331841802	0.0018972062755	0.014550472763
2	0.00299503999806	0.002397870350	0.004538968683	0.002184153299	0.006331841802	0.0018972062755	0.014550472763
3	0.00002427867946	0.000036163059	0.000019723191	0.000061906398	0.000044778686	0.0000217248694	0.000270062003
4	0.00008721652874	0.000039547606	0.000045689775	0.000062729863	0.000144934456	0.0000182132048	0.000298781902
5	0.00014856106014	0.000050657674	0.000068263130	0.000091928212	0.000170339294	0.0000276338476	0.000437802866
6	0.00015448710665	0.011755424340	0.003324171608	0.007525324258	0.002341003970	0.0028477125193	0.008462310564
7	0.00011721664295	0.000054792526	0.000107789394	0.000109873736	0.000219496316	0.0000553565392	0.000380233857
8	0.00001129553831	0.000008902900	0.000017135841	0.000009392407	0.000042755871	0.0000072642648	0.000056705261
9	0.00007985565996	0.000050929682	0.000095996702	0.000066226262	0.000137293115	0.0000295538424	0.000321520335
10	0.00007985565996	0.000050929682	0.000095996702	0.000066226262	0.000137293115	0.0000295538424	0.000321520335
11	0.00000042479509	0.000006413880	0.000004058951	0.000022087135	0.000015399113	0.0000106446928	0.000039540333
12	0.00011721664295	0.000054792526	0.000107789394	0.000109873736	0.000219496316	0.0000553565392	0.000380233857
13	0.00013066486929	0.000071974432	0.000105967800	0.000136358663	0.000170735087	0.0000649315760	0.000343103418
14	0.00011721664295	0.000054792526	0.000107789394	0.000109873736	0.000219496316	0.0000553565392	0.000380233857
15	0.00013066486929	0.000071974432	0.000105967800	0.000136358663	0.000170735087	0.0000649315760	0.000343103418
16	0.00022256918253	0.000092194136	0.000158321936	0.000199828401	0.000200662389	0.0000985169436	0.000502746849
17	0.00001594310328	0.000009811925	0.000033387762	0.000018303363	0.000096213312	0.0000145792370	0.000182813405

Construct Composite Transit Dependency Index

Demographic Index

Both the Non-Intersecting Transit-Dependent population and the intersecting Transit-Dependent population are created by aggregating the synthetic population at the census tract level using cross-table filters in R. Next, the percentage of each of the Transit-Dependent groups in census tract total population is calculated, which will be the demographic index for every census tract in the ARC. In the Transit Dependency Index, the statistics of Georgia is used as a comparison to the ARC counties, and therefore, I calculate the percentage for each Transit-Dependent group using the population for the entire state, which will be the demographic index for Georgia as an average.

Demographic Index for ARC census tracts *= Population for one Transit-Dependent Group in the census tract / Total population of the census tract.* —①

Demographic Index for Georgia *= Population for one Transit-Dependent Group in the entire state/ Total population of the state.* —②

Table 4. Different Transit-Dependent Group Statistics Comparison

	Age 5-17 & Above 65	No Vehicle	Disable	Poverty	Unemployed	Minority
Georgia	53.24%	4.62%	12.89%	19.81%	50.88%	46.16%
ARC	52.95%	4.29%	10.35%	16.09%	47.98%	52.10%
	No Vehicle & Age 5-17 & Above 65	No Vehicle & Disabled	No Vehicle & In Poverty	In Poverty & Unemployed	In Poverty & Minority	In Poverty & Disabled & Age 5-17 & Age above 65
Georgia	3.15%	0.43%	2.59%	13.72%	10.08%	1.55%
ARC	2.94%	0.33%	2.19%	10.76%	9.71%	1.10%

The following maps (Figure 5 – Figure 16) show the difference between the demographic index for every census tract in the ARC, and the State of Georgia average demographic index. The value shown on the maps' lengend is the percentage of the Transit Dependent population in the census tract subtracting the state average percentage.

As is shown on the maps, the disabled population is located in the outer region of the ARC. Counties like Bartow, Carroll and Spalding have a proportion of disabled population higher than the Georgia average (Figure 6). For minority population, areas in south Fulton County, Clayton County and part of Dekalb County have a percentage higher than the Georgia average (Figure 7). For the group of age between 5 to 17 and above 65, most area are close to the Georgia average (Figure 9). For unemployment population, the outer counties are higher than the Georgia average (Figure 10). For the no-vehicle-ownership intersecting groups (Figure 11 -Figure 13), it is obvious to see that the south Fulton County and part of Dekalb county all have higher percentage of people that are higher than the Georgia average.

Figure 6.

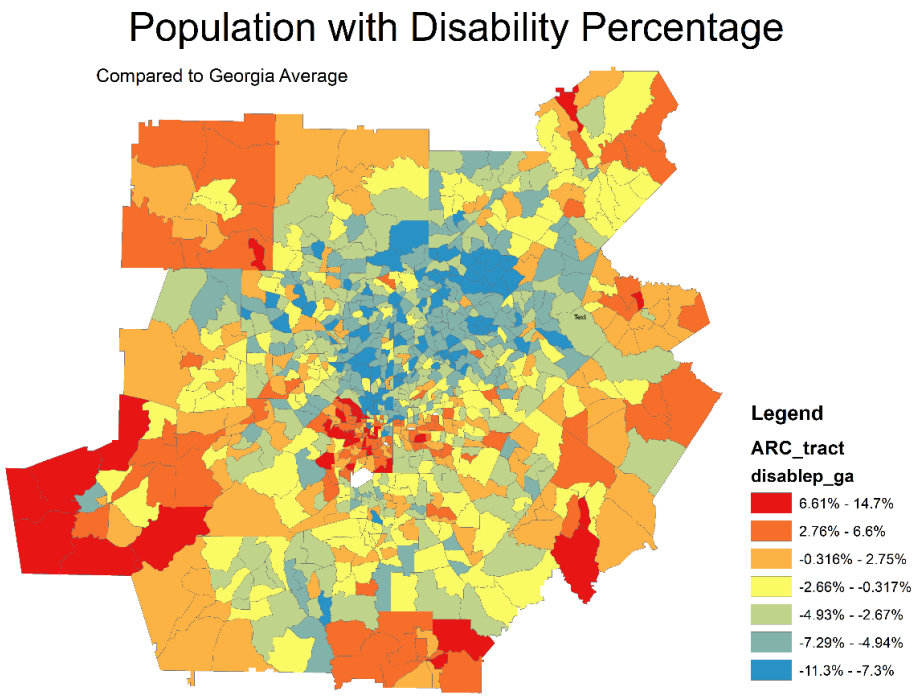


Figure 7.

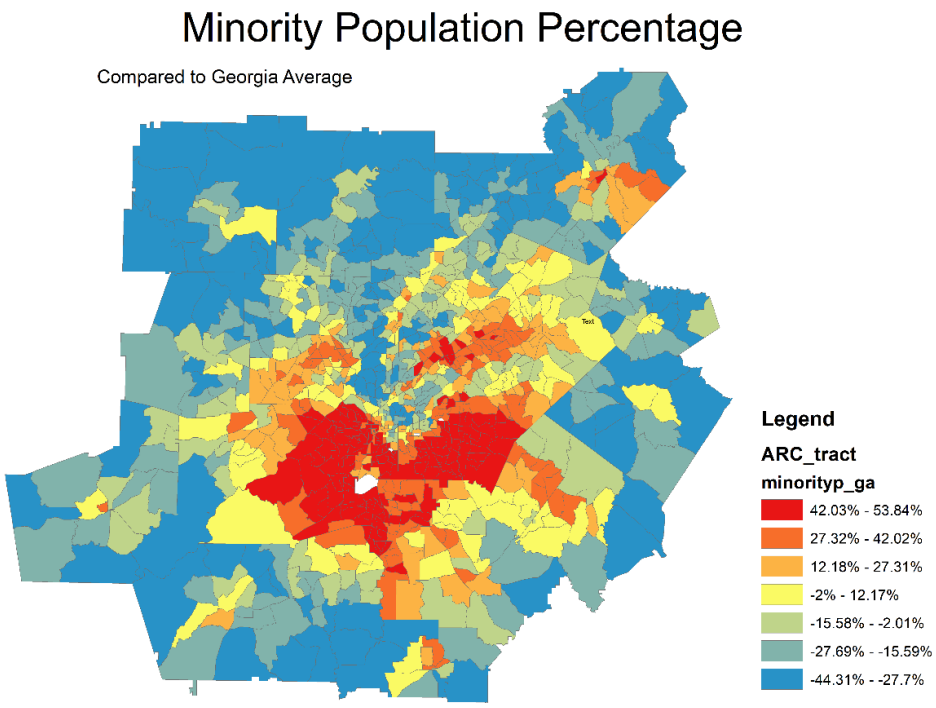


Figure 8.

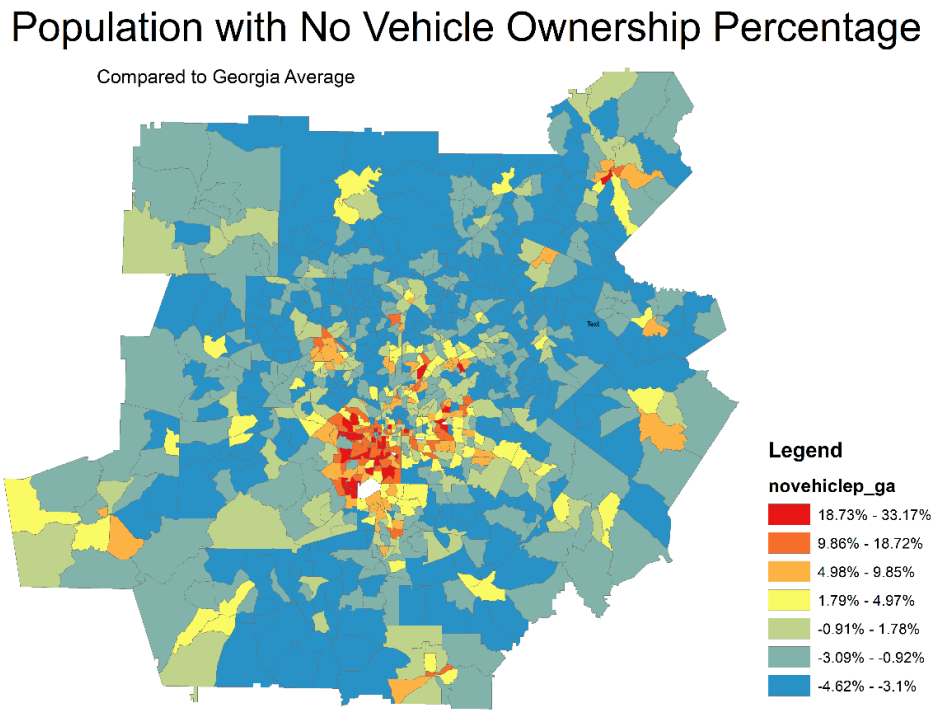


Figure 9.

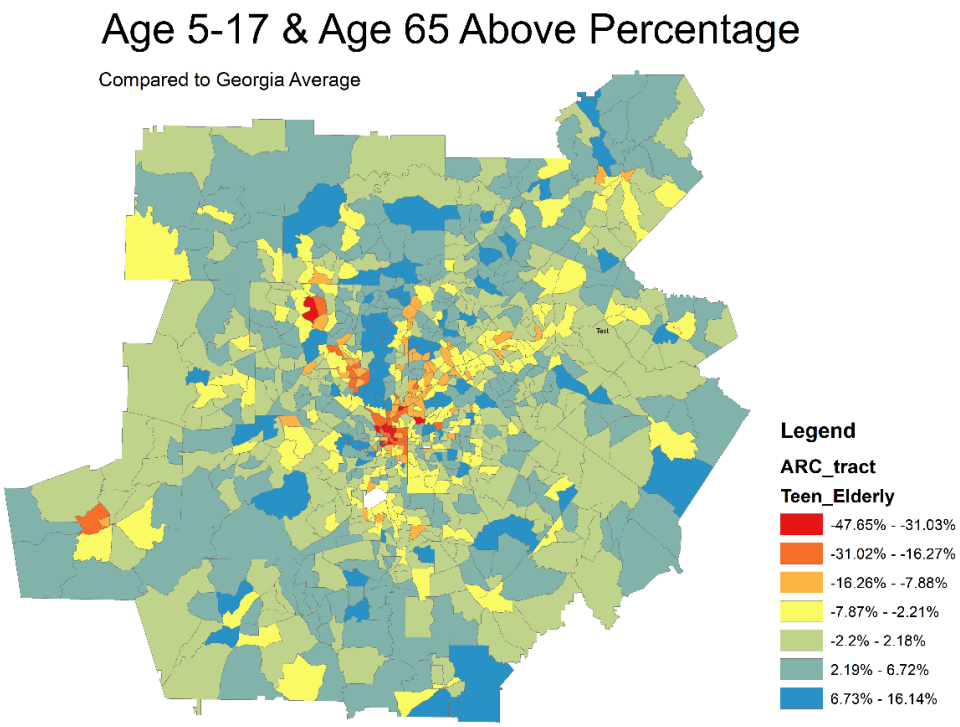


Figure 10.

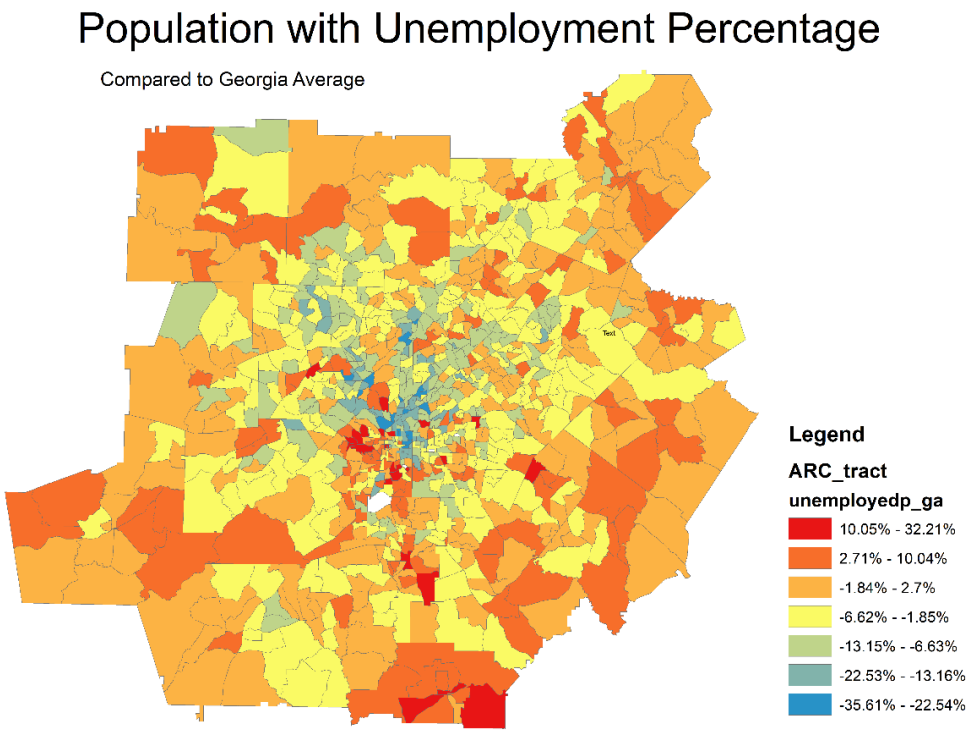


Figure 11.

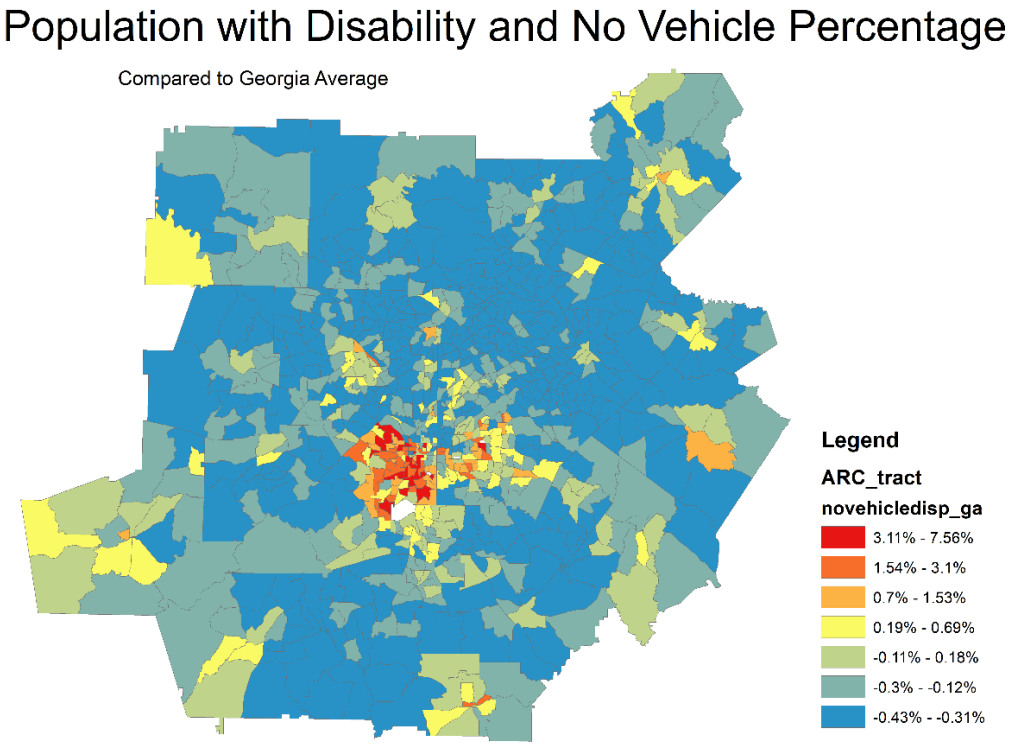


Figure 12.

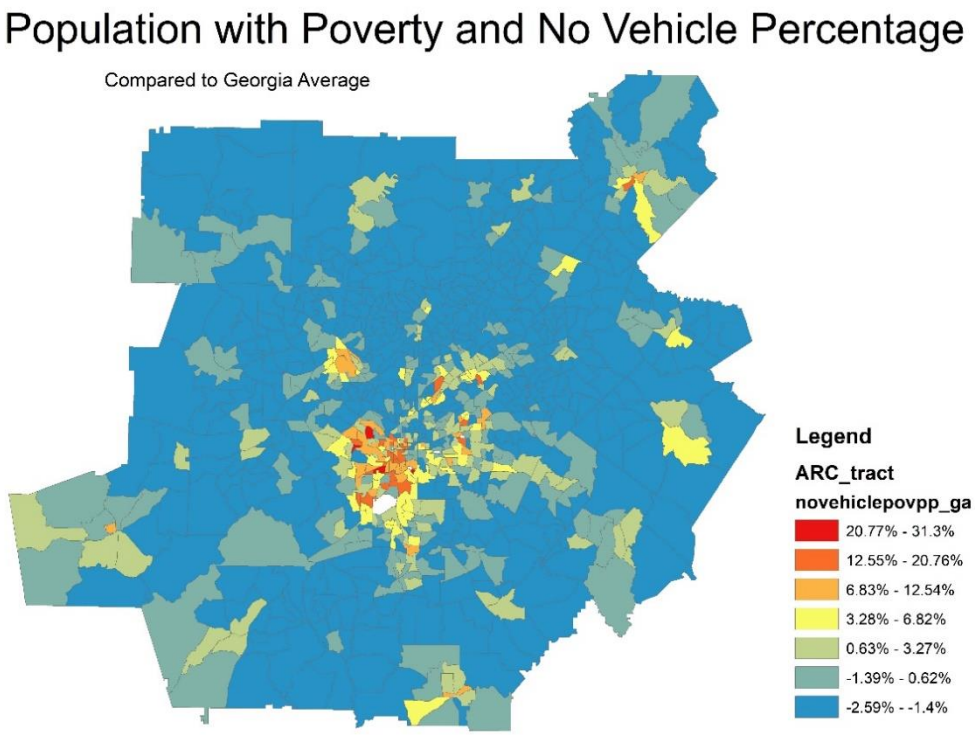
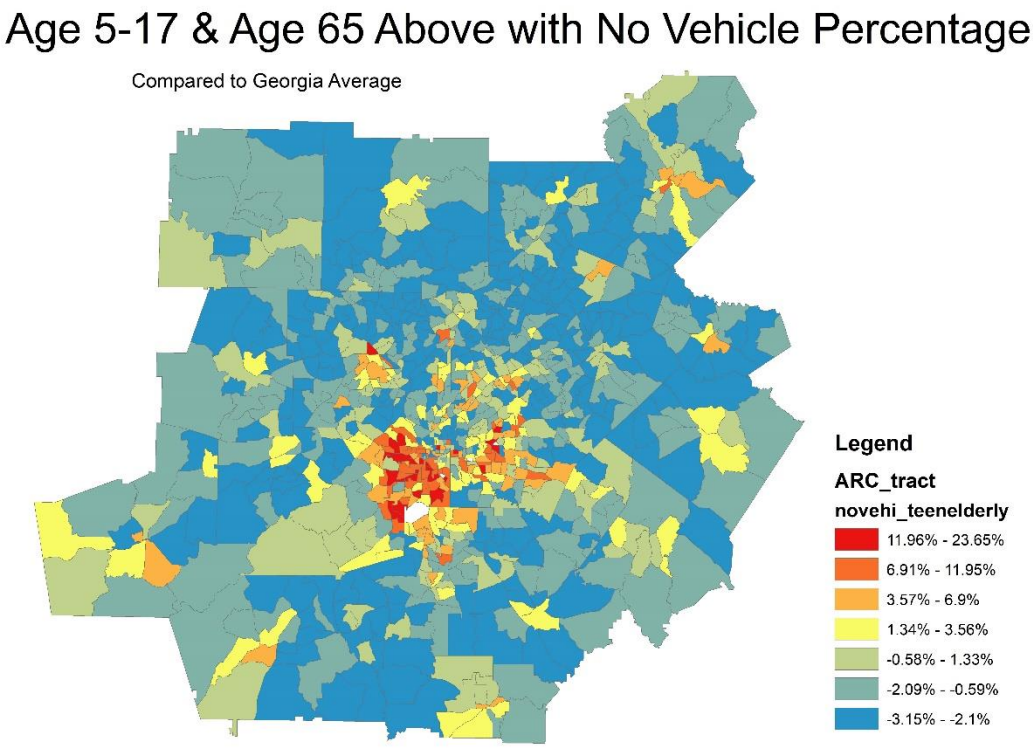


Figure 13.



Accessibility Index

The Accessibility Index is a combinatory index of both employment accessibility and the transit service accessibility (Figure 17, Figure 18). Here, the employment accessibility is the major part of the index, and the transit access is mainly an adjustment factor. There are two reasons to incorporate these two factors. First, the work trip takes up most of the total trips for a census tract. This situation applies better to the areas in the far suburb, where people spend much time commuting to and from work. The second reason is that usually in an urban context, places with larger employment are also more likely to become attractive destinations for non-work trips. For one census tract, employment accessibility is derived from the LODES Workplace Area Characteristics dataset, using the number of employees for an outside census tract and divided by the Euclidean distance between the centroids of both tracts. Repeat the same process for the rest of the census tracts in ARC counties, and calculate the mean for all the outside census tracts.

Employment Accessibility from One Census Tract to All The Other Census Tracts *=*

Sum of all tracts((Number of Employees/Euclidean Distance of

Two Centroids of Both Tracts)) / n — ③

Transit accessibility is measured by the service area of the transit station, and the ratio of the service area to the census tract. For MARTA rail stations, the service area are buffers of 1 mile, while for MARTA bus stops, the service area are buffers of 0.5 mile.

Transit Accessibility *= Buffer Area of The Transit Stop / Area of the Census Tract* —④

For census tracts that do not have transit coverage, the accessibility index is purely the employment access. For census tracts that have transit coverage, the accessibility is (1+Transit Coverage) * Employment Accessibility.

Accessibility Index *= Employment Accessibility from One Census Tract to All The Other*

*Census Tracts * (1+Transit Accessibility)* —⑤

Figure 17. Employment Accessibility

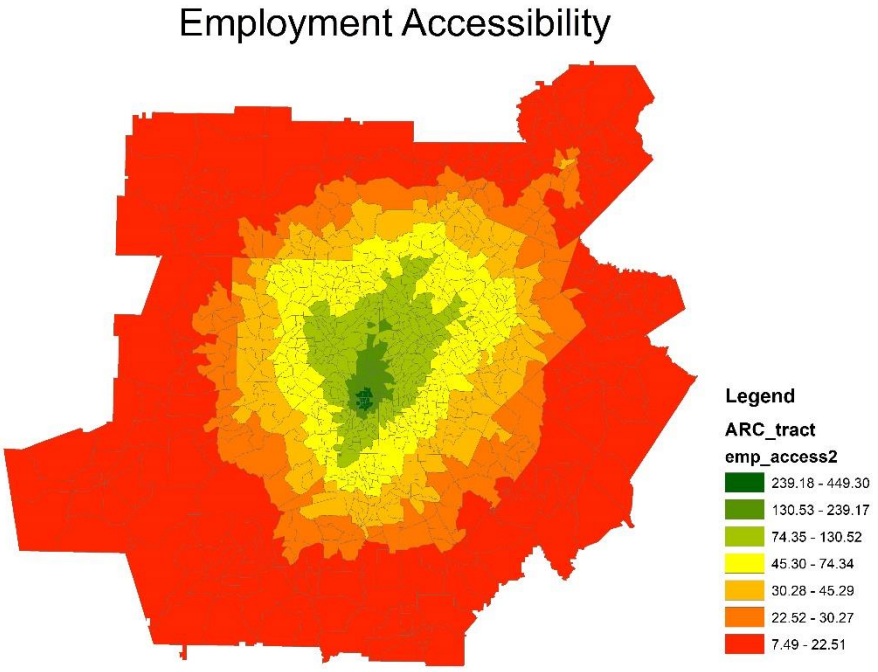
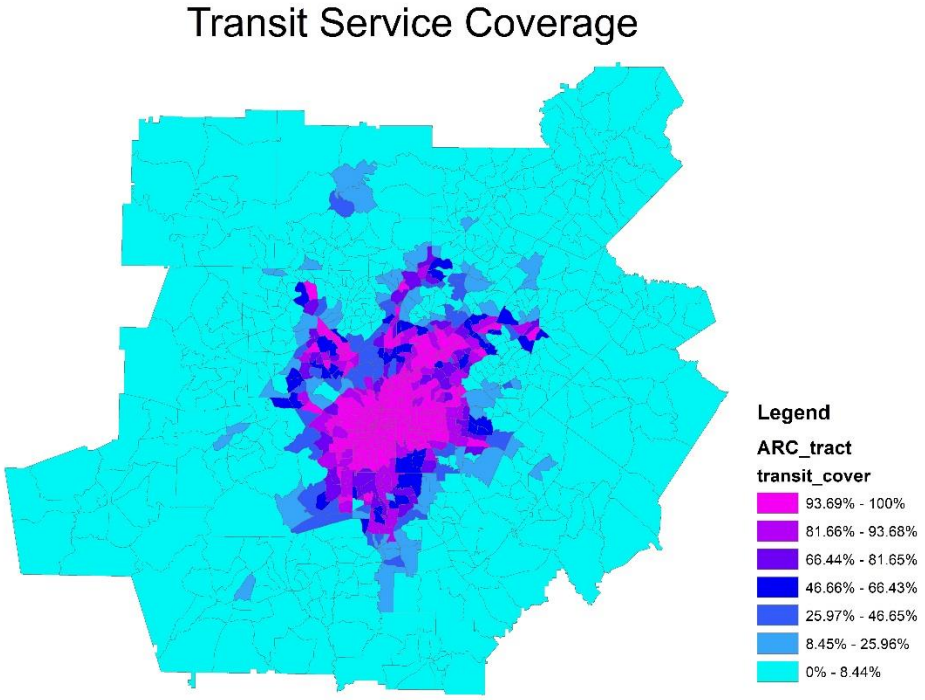


Figure 18. Transit Coverage



Transit Dependency Index

With formula ①②③④⑤, a TDI score can be calculated using the formula as below.

TDI (for One Transit-Dependent group) *=*

*(Accessibility Index) * (Demographic Indicator for Census Tract - Demographic Indicator for Georgia) * (Population Count for Census Tract)*

Below are the maps that show the spatial distribution of each Transit-Dependent group.

Analysis: Do Transit Dependency and Intersecting Transit Dependency Affect Trips?

Two Models

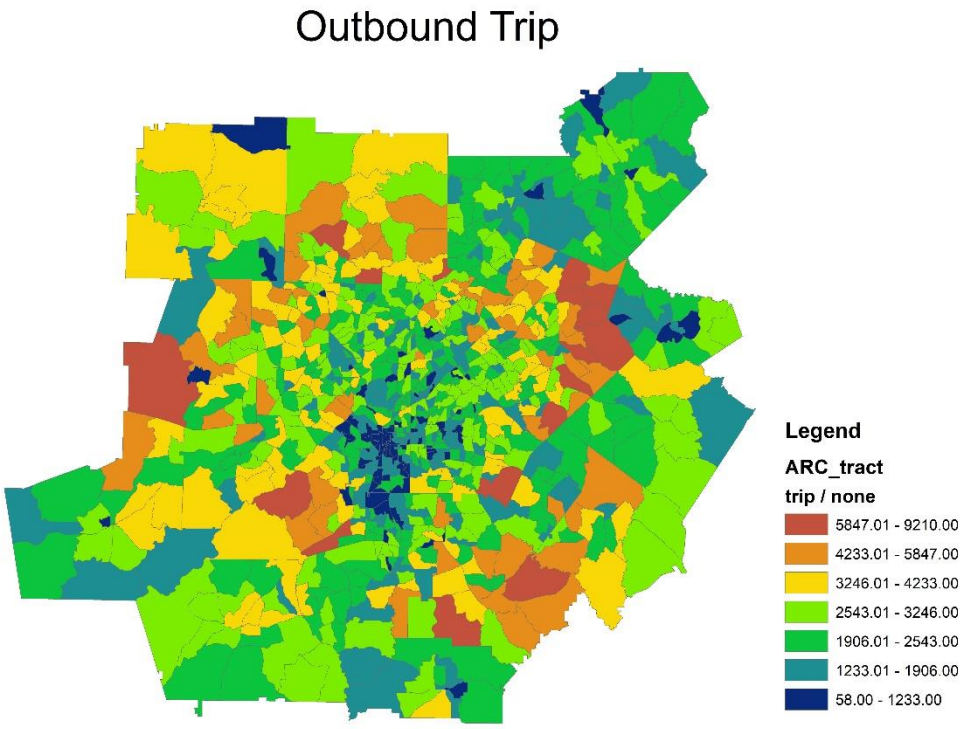
With the generation of TDI for each Transit-Dependent group, I build multivariate regression models to see if there is a correlation between Transit Dependency and the number of outbound trips that are made from a census tract (Figure 19.). The first multivariate model only contains the Non-intersecting Transit-Dependent groups. The regression model is expressed as below.

*Outbound Trips = a * TDI(Vehicle) + b * TDI(Poverty) + c * TDI(Employment)+ d * TDI(Disability) + e * TDI(Age)*

The second model adds in the Intersecting Transit-Dependent Groups, and the regression model is expressed as below.

*Outbound Trips = a * TDI(Vehicle) + b * TDI(Poverty) + c * TDI(Employment) + + d * TDI(Disability) +e * Age+ f * Vehicle by Age + g * Vehicle by Disability + h * TDI(Vehicle by Poverty) + i * TDI(Poverty by Race) + j * TDI(Poverty by Disability by Age)*

Figure 19. Outbound Trip



Regression Results

The first model's regression result is shown below. At 99% confidence level, the "Disabled" group and the "Poverty" group are statistically significant. At 95% level, the "Age 5 to 17 & 65 Above" group are statistically significant. The disable group has the highest impact among all group. One unit increase in the "Disabled" group's TDI will reduce the outbound trip by 0.188.

Table 6. Regression Results for The First Model

Term	Estimate	Std.error	Statistic	P.value
(Intercept)	2444.489	49.942	48.947	0.000
TDI_age517_65	0.011	0.006	1.834	0.067
TDI_novehicle	-0.010	0.006	-1.602	0.110

TDI_disable	-0.188	0.024	-7.701	0.000
TDI_poverty	-0.004	0.001	-3.666	0.000
TDI_unemployed	-0.001	0.003	-0.311	0.756
TDI_minority	0.001	0.001	1.314	0.189
Residuals: Min 1Q Median 3Q Max -2036.9 -737.2 -170.6 491.0 6347.6				
Multiple R-squared: 0.121, Adjusted R-squared: 0.1154				

In the second model result, the “Disabled”, “Minority” and the “Disabled with No Vehicle” groups are statistically significant at 99% confidence level. The “Age 5 to 17 & 65 Above” group is also statistically significant at 95% confidence level. Other variables are not statistically significant.

Table 7. Regression Results for The Second Model

Term	**Estimate**	**Std.error**	**Statistic**	**P.value**
(Intercept)	2443.297	50.361	48.515	0.000
TDI_age517_65	0.013	0.006	2.086	0.037
TDI_novehicle	-0.025	0.027	-0.926	0.355
TDI_disable	-0.233	0.030	-7.670	0.000
TDI_poverty	0.003	0.003	1.125	0.261
TDI_unemployed	0.002	0.004	0.511	0.610
TDI_minority	0.003	0.001	3.967	0.000

TDI_novehage51765	-0.062	0.094	-0.656	0.512
TDI_novehdis	1.955	0.529	3.696	0.000
TDI_novehpoverty	-0.006	0.034	-0.178	0.859
Residuals: Min 1Q Median 3Q Max -2401.1 -691.2 -161.6 486.4 6418.2				
Multiple R-squared: 0.1543, Adjusted R-squared: 0.1433				

Diagnosis And Discussion

The second model contains intersecting Transit-Dependent groups, which are variables that correlate with the initial six single variables. This is because, within the single variables, the intersecting records are already included. By examining the second model's result, it is also clear to see that the insignificant variables are associated with the "Vehicle Ownership" and "Poverty Status" variables. In order to test for collinearity problems, I use the Variance-Inflation factors (VIF) method (Table 7).

Table 7. VIF Diagnostic Result

TDI_age517_65	**TDI_novehicle**	**TDI_disable**	**TDI_poverty**
2.135	28.116	2.356	19.550
TDI_unemployed	**TDI_minority**	**TDI_novehage51765**	**TDI_novehdis**
3.295	1.834	11.290	5.861
TDI_novehpoverty	**TDI_povertyemp**	**TDI_povertyminor**	**TDI_povpdisage51765**
9.244	30.368	35.903	1.966

The test result shows that the “Poverty and Unemployed”, “Poverty and Minority”, “Poverty” and the “No Vehicle” variables have collinearity issues because of their high VIF values. Improvements of this model is made by eliminating these groups one at a time. The VIF will be checked after each round of elimination. After eliminating the three variables, the final model do not have the collinearity issue. The final regression model is expressed as below (Table 3).

*outbound Trips = b * Poverty + c * Employment + d * Disability + e * Age+*
*f * Vehicle by Age + g * Vehicle by Disability + h * Vehicle by Poverty +*
*+ j * Poverty by Disability by Age*

Table 8. Regression Results for the Final Model

Term	**Estimate**	**Std.error**	**Statistic**	**P.value**
(Intercept)	2439.555	49.936	48.853	0.000
TDI_age517_65	0.011	0.006	1.893	0.059
TDI_disable	-0.220	0.029	-7.519	0.000
TDI_poverty	-0.004	0.001	-3.522	0.000
TDI_unemployed	0.001	0.003	0.330	0.741
TDI_minority	0.001	0.001	2.509	0.012
TDI_novehage51765	-0.157	0.057	-2.755	0.006
TDI_novehdis	2.080	0.469	4.430	0.000
TDI_novehpoverty	-0.060	0.022	-2.734	0.006
TDI_povpdisage51765	-2.953	4.035	-0.732	0.464
Residuals: Min 1Q Median 3Q Max				

-2690.2 -702.5 -155.5 469.3 6344.0
Multiple R-squared: 0.1419, Adjusted R-squared: 0.1336

Table 9. VIF for the Final Model

TDI_teenelderly	**TDI_disable**	**TDI_poverty**	**TDI_unemployed**
1.887	2.161	2.207696	2.350053
TDI_minority	**TDI_novehteenelderly**	**TDI_novehdis**	**TDI_novehpoverty**
1.386	4.101	4.565473	3.723734
TDI_povpdisteenelderly			
1.761			

The result of the final model shows an improvement in the variables' significance. Except for the "Unemployed" and "Age 5-17 and Above 65, Disabled with Poverty", other variables are all statistically significant at 95% confidence level.

Among the statistically significant variables, one unit increase in TDI for the "No Vehicle and Disabled" group will result in 2.08 increase in trips. This means that as this Transit-Dependent group expands, it still has a positive effect on the number of trips. This is finding is opposed to the assumptions of this research that are made in the first place. After checking the spatial distribution of this group, it is found out that the areas with high TDI are mostly in South Downtown Atlanta, and also some tracts in the boundary counties of ARC. Downtown Atlanta is the place where least outbound trip is happening so it is likely that there are some other factors or demographic groups that are left out of the model that are needed to explain this paradoxical finding. Similarly, the "Age 5-17 and Above 65" group also has a positive impact on the number of trips as TDI increases. However, comparing to the "Age 5-17 and Above 65 with No Vehicle" group, which has a negative coefficient that are larger than the former's coefficient when

converted into absolute value, the positive effect of the former group are likely to cancel out, meaning that overall, the increase in this “Age Transit-Dependent” group will have a negative impact on the total trips outwards.

For the insignificant variables, the “Unemployment” group is spatially scattered inside the whole ARC region, however, with the higher percentage concentrating on the boundary counties of the ARC region. In contrast, those counties also have high outbound trips, which can be the reason why the model won’t fit between observation and modeling result. The “Age 5-17 and Above 65, Disabled with Poverty” is a triple intersecting variable, and the average population percentage in ARC region is 1.1%. Therefore, the overlapping among these three groups are comparatively small, and may not show a significant impact on trips.

Overall, the regression model is able to explain the correlation between TDI and outbound trips to some level of extent. The intersecting Transit-Dependent groups are also found contributing to the regression model, which means that Transit Dependency in the ARC region is a cross-demographic issue.

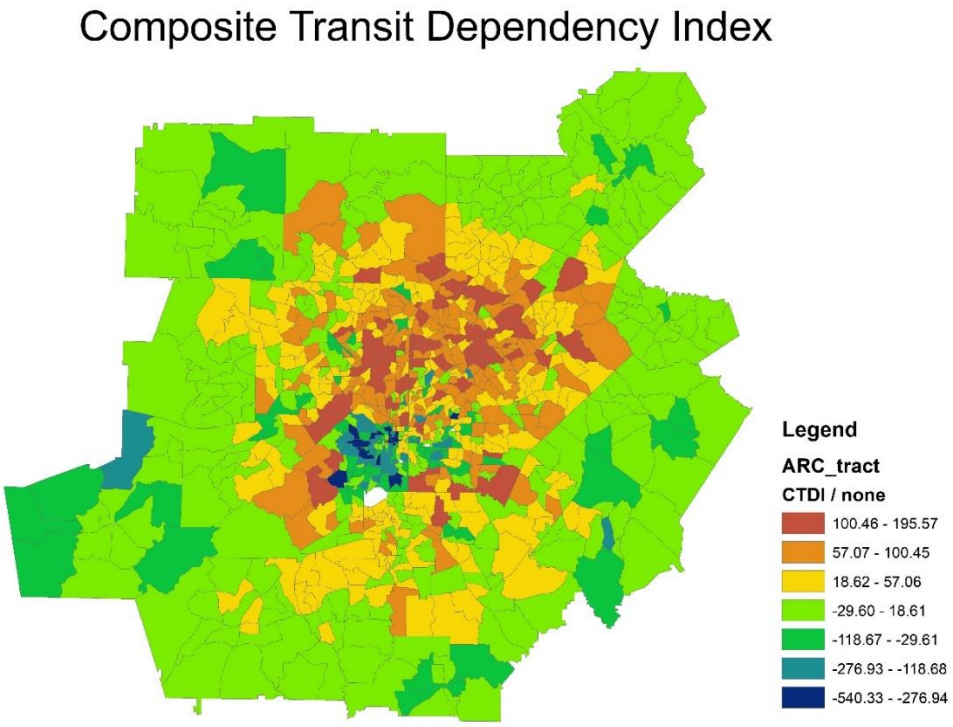
The Composite Transit Dependency Index

Based on the regression model result in the last chapter, it is not yet convincing that a Composite Transit Dependency Index (CTDI) can be modeled, given that there are still flaws with the regression model, and statistically insignificant variables should not be incorporated into the model.

Nevertheless, as a demonstration of the goal of this research, that is, the workflow of how to construct a CTDI model, there is a necessity to complete the final process to respond to the methodology framework that has been laid out before.

As is explained in the Methodology Chapter, the CTDI model takes into account the remaining variables that have been filtered through the above regression diagnostic process. The coefficient for each Transit-Dependent group will be normalized so that the sum of all coefficients’ absolute value is 1. The CTDI value is calculated by summing up the TDI for all groups. Below is the map that presents the spatial pattern of Composite Transit Dependency Index.

Figure 20. Composite Transit Dependency Index



Conclusion

The research demonstrates the comprehensive framework of modeling a Composite Transit Dependency Index for a geography region from data source gathering to model execution. It proves that there is a way to incorporate different demographic groups into a composite index. This is beneficial to the large-scale analysis as it produces intuitive results. There are still flaws and limitations to the regression model that will require vast statistical examinations to make improvements.

In the final CTDI map, it is obvious to see that the census tracts with high overall transit dependency are located in the near suburbs of the City of Atlanta, and almost cover up the middle-ring of the whole ARC region. In recent years, these suburb counties such as Gwinnett and north Fulton have gone through a rapid urban development with the rise in immigration, among whom are a large number of socially marginalized groups. This mapping finding is consistent with the reality, which reveals the potential application of the TDI modeling in the future use of transit planning.

Bibliography

1. *Krizek, K., & El-Geneidy, A. (2007, 09). Segmenting Preferences and Habits of Transit Users and Non-Users. Journal of Public Transportation, 10(3), 71-94.*

2. *Litman T. (2004, 10). Rail Transit In America A: Comprehensive Evaluation of Benefits Report Summary, Victoria Transport Policy Institute.*

3. *Maitra, B., Dandapat, S., & Chintakayala, P. (2015, 06). Differences between the Perceptions of Captive and Choice Riders toward Bus Service Attributes and the Need for Segmentation of Bus Services in Urban India. Journal of Urban Planning and Development, 141(2), 04014018.*

4. *Krizek, K., & El-Geneidy, A. (2007, 09). Segmenting Preferences and Habits of Transit Users and Non-Users. Journal of Public Transportation, 10(3), 71-94.*

5. *Steve E. Polzin, Xuehao Chu, and Joel R. Rey (2007). Density and Captivity in Public Transit Success: Observations from the 1995 Nationwide Personal Transportation Study. Transportation Research Record, 1735, Paper No. 00-0891*

6. *Jiao J. and Dillivan M. (2013), "Transit deserts: the gap between demand and supply", Journal of Public Transportation, Vol. 16 No. 2, pp. 23-39*

7. *Lovely, M., & Brand, D. (1982). Atlanta transit pricing study: moderating impact of fare increases on poor. Transportation Research Record: Bus Operations and Performance, 857, 39–44*

8. *Zhao F, Gustafson T. (2013). Transportation Needs of Disadvantaged Populations: Where, When, and How? Federal Transit Administration Report No. 0030.*

9. *NIOSH, ASSE [2015]. Overlapping vulnerabilities: the occupational safety and health of young workers in small construction firms. By Flynn MA, Cunningham TR, Guerin*

RJ, Keller B, Chapman LJ, Hudson D, Salgado C. Cincinnati, OH: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, DHHS (NIOSH) Publication No. 2015-178.

10. *Verbich, D., & El-Geneidy, A. (2017, 02). Public transit fare structure and social vulnerability in Montreal, Canada. Transportation Research Part A: Policy and Practice, 96, 43-53.*

11. *United Nations Development Program. (2016). Human Development Report.*

12. *Lovelace R, Dumont M. (2018). Spatial Microsimulation with R. CRC Press. Chapter 1.3.*

13. *Alexander W Blocker. (2016, 02). Package 'ipfp'. Retrieved from https://www.rdocumentation.org/packages/ipfp/versions/1.0.1/topics/ipfp*