Travel Behavior and People's Attitude from Social Media: An Exploratory Case Study in New York City

A thesis submitted by

WENCONG XU

In partial fulfillment of the requirements of the degree of Master of Science

in

Environmental Policy and Planning

Tufts University

August 2018

Advisor: Justin Hollander, Reader: Sumeeta Srinivasan

Abstract

This thesis utilized Twitter data to determine if people's sentiment had connections to the travel mode choice they made. Whether the Transit Oriented Development (TOD), with catchment radius of 2 kilometer and 0.5 mile, boosts people's sentiment toward non-vehicle travel modes was also assessed. Moreover, a model was built to predict vehicle usage rate in transit sheds using a Artificial Neural Network (ANN). The result tells the association between people's sentiment toward travel modes and people's travel behavior was very weak ($R^2 < 0.03$). People's overall sentiment is statistically significant to people's travel mode usage. The models effectively predicted the vehicle usage rate in the transit sheds and they gave us accuracy scores larger than 0.87. This thesis utilized social media to provide another angle to understand people's travel behavior. Also, with the models planners and developers can effectively predict one TOD areas' vehicle usage rate and develop better policies.

Acknowledgements

I would first like to thank my Thesis Advisor and Reader, Justin Hollander and Sumeeta Srinivasan for their patience, support, and willingness to make suggestions on my research framework and questions throughout the writing process. I also want to pay tribute to Statistics and Research Technology Specialist, Kyle M. Monahan who has given me a lot of advice on how to use Python, R and other programming languages that have been so important to this thesis. Lastly, I want to express many thanks to my parents, wife and friends, all of whom encouraged me to finish this thesis through their financial and emotional support.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Transit Oriented Development (TOD), rooted in the New Urbanism and Smart Growth, thrived at the end of the 20th century and was supposed to be a method to constrain low-density urbanization and suburbanization by focusing on providing transit service along with high density and mixed-use development to encourage transit ridership and none-vehicle travel modes (Nasri and Zhang 2014a). However, as we can see today in most of metropolises around the world, especially in the United States, suburbanization and congestion are becoming more and more serious even though there are about 200 established methods of TODs with nearly 4000 sites offering the potential for various forms of TOD practice (Reconnecting America, 2009). Although some studies gave evidence that people living in more urbanized areas seem prefer to use public transportation as it is easier for them to access those transportation facilities (Cervero and Kockelman 1997), the average non-vehicle household share was at low of 0.5% in 93 US metropolitan areas in 2010 (Bureau n.d.).

According to Google Scholar, over 15,900 reports and papers related to the keywords “transit” and “travel mode choice” were published during 2017, and they tried to find solutions to enhance non-vehicle travel use in TOD catchment areas. A large number of those studies focused on the impact of the built-

environment to the choice of travel mode, only a few closely examined people's will to use non-vehicle travel modes. The conventional way to collect data on people's attitudes toward travel mode is by sending surveys or e-mails, which are small in sample size and updated at low frequencies (Shan Jiang et al. 2016). With the rise of ubiquitous sensing technologies, digital resources, and open-source data like Twitter and Facebook, we can harness digital records of traces that people leave in cyber-physical spaces.(Daqing Zhang, Philipose, and Yang 2011; D. Zhang, Guo, and Yu 2011). Social media like Twitter and Facebook are used worldwide, and they allow people to publish real-time information about anything they want to share. Published tweets are useful for us to analyze urban issues, like individual users and transportation agencies publishing real-time traffic information about traffic jams and traffic incidents through social media platforms (Lv et al. 2017). We can also try to analyze people's behaviors and sentiments by using social media's GPS and their contents.

New York City, which has one of the largest metro systems in the world while also struggling with traffic congestion, is a good research target as it has resourceful open data and a large number of transits ("NYC's Public Transit Somehow Ranked Best in the Nation | New York Post" n.d.). Nevertheless, New York City is losing its position as the world's best city for public transport (Santora 2017). Based on a New York Times report, the average vehicle speeds in Manhattan below 60th Street declined 12% from 2010 to 2015, and the

subway ridership dropped to 5.712 million from 5.817 million in September 2016. So, the Big Apple is a very useful case study for me to analyze the effect of TOD.

Through a case study in NYC Metropolitan Statistical Area (MSA), this thesis evaluated TOD's effects on travel mode choice using people's attitudes and behaviors. This work aimed to uncover the inner connections between the usage of each travel mode and people's sentiments in TOD catchment areas. The relationship between built environment factors combined with people's sentiment and vehicle usage rate in transit sheds was also measured. The questions we want to answer for this thesis are:

- What is the current usage rate for each kind of travel mode in the NYC MSA's TOD catchment areas?
  - (1) How to define TOD catchment areas by their service radius?
  - (2) What are the travel modes we need to include?
  - (3) How to measure each of the travel mode's usage in TOD catchment areas?
- What's the relationship between people's sentiment and travel mode choice in transit zones?
  - (1) What is the relationship between sentiment measured by social media and mode of travel?

(2) How does social media measurement compare to the Regional Household Travel Survey (RHTS)?

(3) What is the relationship between people's overall sentiment and travel mode choice in transit zones?

- What are the other factors which can impact people's travel behavior and comparing factors like built-environment, what role are people's sentiments playing to impact their travel choices?

Depending on the questions raised above, the hypothesis we developed are:

(1) People's sentiment scores can replace travel surveys to predict people's travel behavior in TOD catchment areas.

(2) TOD really changes people's minds about travel mode choice.

(3) By including more indicators and using Machine Learning, we can correctly predict people's travel behavior in TOD catchment areas.

To answer those questions ahead and test the hypothesis, this thesis focused on four kinds of travel mode: subway, bicycle, walking and automobiles. We studied 422 transit catchment areas located in New York City (NYC). Among them, 186 of the transit zones enclose all kinds of travel mode usage data. Additionally, 422 transit zones contain at least two of those travel modes data. Other zones lack of data and spaces without a transit node was set as control

experiment examples.

In terms of Twitter data, 2,042,301 tweets from April 15 to July 30 in 2016 were collected by the Urban Attitudes Lab at Tufts University. After sorting these original tweets (Appendix) by those containing each travel mode keywords, we get the distribution of tweets below:

Table 1: Tweets' count with key words of each of travel mode and sentiment

Travel Mode	**Tweets**
Vehicle	6893
Subway	6518
Cycling	3773
Walking	3772

Other than that, travel mode usage data including annual subway entry data from Metropolitan Transportation Authority (MTA), bicycle usage data from Citi Bike, pedestrian data and vehicle usage data from the U.S. Census were used as a baseline measure of real usage for each travel mode. Census tract level trip statistics from 2010 to 2011 Regional Household Travel Survey data for vehicle, bicycle, subway and walking were also collected, which represented the travel survey data. (Table 2)

Table 2: Data Sets Description

Travel Mode	**Source**	**Description**
Bicycle	Citi Bike	6,237,130 Citi Bike usage records for 619 bicycle stations from April to July in 2016 were collected
Walking	2016 American Community Survey	Times people take walk as their travel mode in 2016 for 1,672 block groups in

		NYC.
Transit	NYC MTA 2016 ridership statistics	The average day-ridership for 422 transit stations in 2016 are collected.
Vehicle	2016 American Community Survey	Times people use vehicle reported in 2016 for 1,672 block groups in NYC.
Vehicle, Subway, bicycle and walking trips	2010 to 2011 Regional Household Travel Survey	188,199 records for unlinked personal trips or trip segments, where either the "from" or the "to" based on each kind of travel modes.

In addition, other variables which are considered to impact people's travel behavior derived from EPA's Smart Location Database and NYC Department of Transportation (DOT) were collected. The Smart Location Database contains over 90 different indicators associated with built environment, land use diversity, road density, accessibility, etc. in block group level (US EPA 2014). This article chose 16 indicators, which had coefficient correlation value with vehicle usage rate larger than 0.3, as parameters to analyze their relationship with vehicle usage rate in TOD catchment areas (Table 3). Also, the parking signs' data from DOT was included.

Table 3: Other Included Factors

Data Name	**Original Data Name**	**Description**	**Source**
Auto0	Pct_AO0	Percent of zero-car households in CBG	American Community Survey (ACS)
Auto1	Pct_AO1	Percent of one-car households in CBG	ACS
Auto2	Pct_AO2p	Percent of two-car households in	ACS

		CBG	
High_Wage	R_HiWageWk	# of workers earning $3333/month or more (home location), 2010	Census LEHD, 2010
SJ5	E5_Svc10	Service jobs within a 5-tier employment classification scheme (LEHD: CNS12 + CNS14 + CNS15 + CNS16 + CNS19)	Census LEHD, 2010
GRD	D1a	Gross residential density (HU/acre) on unprotected land	Derived from other SLD variables
GED	D1c	Gross employment density (jobs/acre) on unprotected land	Derived from other SLD variables
GSED	D1c5_Svc10	Gross service (5-tier) employment density (jobs/acre) on unprotected land	Derived from other SLD variables
GAD	D1d	Gross activity density (employment + HUs) on unprotected land	Derived from other SLD variables
RD	D2r_JobPop	Regional Diversity. Standard calculation based on population and total employment: Deviation of CBG ratio of jobs/pop from regional average ratio of jobs/pop	Derived from other SLD variables
AutoND	D3aao	Network density in terms of facility miles of auto-oriented links per square mile	NAVSTREETS
Auto45min	D5ar	Jobs within 45 minutes auto travel time, time decay (network travel time) weighted	NAVSTREETS
Transit45min	D5br	Jobs within 45-minute transit commute, distance decay (walk network travel time, GTFS schedules) weighted	NAVSTREETS
RCIauto	D5cri	Regional Centrality Index – Auto: CBG D5cr score relative to max CBSA D5cr score	Derived from other SLD variables

RCItransit	D5dri	Regional Centrality Index – Transit: CBG D5dr score relative to max CBSA D5dr score	Derived from other SLD variables
ParD	Parking Signs and Locator	Parking sign density (400,429 Parking Signs data for the five borough in NYC)	DOT

In addition to the datasets, the software we used in this thesis are listed below:

Table 4: Analysis Software

Software	**Use**
ArcMap	Calculate variables I'm going to analyze and geographic them on maps according to TOD areas.
GeoDa	Provide spatial data analysis, such as spatial autocorrelation statistics for aggregate data and basic spatial regression for data in TOD areas.
R Studio	Calculate sentiment score for each kind of travel mode and clean raw datasets.
Python 3.4	Build Artificial Neural Network (ANN) and provide the environment to train the model.

This thesis illustrates how tweets relevant to the travel mode choice and how the built-environment in TOD areas in NYC impact vehicle usage rate. The first chapter gives an overview of the entire thesis and shows the questions we study, as well as the datasets used. The motivation behind this research is also discussed. In the second chapter, I discuss other relevant work which informs the research scope of this work. This chapter includes five aspects which are helpful to guide my research and improve its overall utility. Figuring out how

scholars define Non-Vehicle Travel mode in TOD catchment areas can provide us with a clue as to which travel modes we need to focus on. Understanding different travel modes' access radius can help us define the research radius for TOD areas. The literature about travel behavior of people in TOD catchment areas, and other factors such as built-environment and people's sentiment, which may affect or useful for predicting people's behaviors are also carefully discussed. The literature review chapter also introduces the importance of big data for our research. These articles give us a hint about what factors we need to investigate and how other scholars included or addressed these factors. The third chapter, Methodology, represents the workflow to be followed. This chapter also includes the method of data collection, determination of sentiment scores, research areas and machine learning. The current usage of the travel modes, bicycle, walking, vehicle and subway, are measured and discussed in chapter four. Chapter five explores the connections between people's sentiment score and travel behavior in TOD catchment areas based on different TOD radius. The sixth chapter, using ANN, builds a model with built-environment variables and people's overall sentiment score to predict the vehicle usage rate in transit sheds. The last chapter will provide a conclusion about my analysis and the limitations of this research.

Chapter 2: Literature Review

Considering the research questions I discussed earlier, this literature review will provide a brief overview of scholarly articles on the topic, which provide a foundation for this thesis. The literature review includes articles focused on travel modes, catchment radius, people's behavior and built environment. Since we are going to analyze people's sentiment and behavior, the connections between those two elements and the necessity of big-data techniques are also investigated.

1. Non-Vehicle Travel modes in TOD catchment areas

Travel modes like walking and cycling are seen as important substitutes for vehicles. Based on research in west Australia, 92% of the people living in the service area of Subiaco Station agree that walking is an important travel mode in their daily life and cycling also gets nearly 40% of the residents' support (Griffiths and Curtis 2017). Nevertheless, both walking and cycling can bring advantages to our daily life. Scholars like Lilah Besser and her colleagues confirmed in their article that higher frequency of transit use can enhance people's walking distance so that they meet their suggested daily physical activity.(Besser and Dannenberg 2005) On the other hand, since bicycles enormously extend TOD's catchment area and make the transit nodes more accessible (O'Sullivan and Morrall 1996), they can make the city more

accessible and more equal. A research in Wuhan points out that the transit can enhance the property value in 400 meters from the station. The closer to the station, the more expensive the property value will be (Xu, Zhang, and Aditjandra 2016). Transit systems rely heavily on three groups for their core ridership: low income households, people of color and renters(Stephanie Pollack, Barry Bluestone, and Chase Billingham n.d.), who are usually pushed out of the catchment areas by gentrification. Bicycle usage can effectively benefit those groups.

2. Transit's Service Radius

Different travel modes have different levels of accessibility toward their destinations. For example one can only walk 0.5 miles in 5 minutes, but he or she can reach a lot further by cycling. Thus, TOD catchment areas' radius may be highly affected by which tools people choose to use. By setting different radii for subway service areas can effectively help us know people's travel behavior. Scholars from Toronto University studied 69 subway stations in Toronto and used the Transportation Tomorrow Survey (TTS) to measure the catchment areas for transit nodes. They found out that for each kind of travel mode, the transit station might have different catchment radius. For instance, the bus catchment area is about 160 square kilometers, which is nearly 80 times bigger than the pedestrian catchment area, 2.03 square Kilometers and

the car catchment area can reach to over 1000 square kilometers (Xi, Saxe, and Miller 2016). In the US, scholars and specialists usually use 0.5 mi radius (805m) to define a transit node's catchment radius because it is a relatively more walkable and pedestrian friendly distance (Mamun et al. 2013). A quarter mile also widely used to define a service radius and some scholars announce that it is more fitful for small transportation stations like BRT station and bus stations.(O'Sullivan and Morrall 1996) On the other hand, an article from South Korea points out that in Seoul, bicycles can extend the transit node's catchment radius to 1.96 km and 2.13 km for origin (home)-to-station and station-to-work trips (Lee, Choi, and Leem 2016). Taylor and Mahmasasani also said in their article that public transport passengers had a bicycle access distance of 2.4 km, and added that if a bike path or parking lot becomes available, the access distance could be extended up to 4.8 km(B.~Taylor and Mahmassani 1997).

3. People's travel behavior in transit zones

What kind of travel mode do people like to use in subway sheds? A few studies, which focused on people's travel behavior in TOD catchment areas, had been completed by scholars using different methods. For instance, Nasri and Zhang announced that the TOD indeed reduced the Vehicle Miles Traveled (VMT) among the residents living in the TOD areas in Washington D.C. and

Baltimore. Based on their research, the TOD mode can reduce residents' VMT by around 38% in Washington, D.C. and 21% in Baltimore compare to those areas with similar land use but without a Rapid Transit Station.(Nasri and Zhang 2014b) Also, Shirke and his colleagues point out that in developing countries' cities with a high density population like Mumbai, TOD is a very effective way to reduce vehicle use and increase job density in Transit Influence Area (TIA) by building a travel choice model. They also found even though the density increase in the future in TIA, the congestion on the road would not increase (Shirke et al. 2017). However, some scholars hold other ideas. Kamruzzaman and his coworkers investigated over 6000 individuals who lived in TOD catchment areas in Brisbane and found out that "They did not naturally adjust their preferences according to their surrounding land use patterns and continue their predisposed travel behavior," and 38% of the residents didn't want to live in this area and were not eager to take the rapid transit as scholars expected (Kamruzzaman et al. 2016). What's more, Ettema and Nieuwenhuis designed a survey among 355 recently relocated households in Dutch TOD locations, and they concluded that there was a weak relationship between people's location choice and people's travel attitudes, which means people who chose to live in TOD catchment areas probably didn't want to use the rapid transit as their main transportation tool (Ettema and Nieuwenhuis 2017).

4. Built environment and people's behavior in TOD catchment areas

The relationship between travel behavior and built environment is one of the hottest research points in academia: "In travel research, such influences have often been named with words beginning with 7D", which encompass diversity, density, design, destination accessibility, distance to transit, demand management and demographics (Ewing and Cervero 2010). Cervero and Ewing think the density of population and employment, the diversity of land use, the design of intersections, the networks of the streets, the destination accessibility of jobs and other attractions and the distance to transit stations are the key factors which impact people's travel mode choice (Ewing, Deanna, and Li 1996; Ewing and Cervero 2010). Some other scholars, like Xuedong Lu and Daniel Hess illustrated that demographics and parking supply played important roles in people's travel mode choice(Lu and Pas 1999; Hess 2001).

Other experts tried to find the potential relationship between built environment and people's behavior in TOD catchment areas. Kim and his colleagues used OLS regression model to analyze six sites, and the result showed that "adopting TOD strategies to increase the level of land-use mix, develop comfortable sidewalks, and improve street connectivity and transit access could be more effective to promote walking than to increase

development density in overly intensified urban areas" (Kim, Sohn, and Choo 2017). Similarly, scholars like Park, Choi and Lee also show their point of view in their articles that creating wider and more continuous sidewalks with fewer driveway and making the blocks smaller and street narrower are essential to encourage people walking to the transit stations (Park, Choi, and Lee 2017). Other than that, Olaru and Curtis stated transit nodes' accessibility had a positive relationship with residents' reducing car based travel and encourage people to choose other travel modals (Olaru and Curtis 2015). Langlois and some researchers' perspective about this issue was that the more walkable an environment is, the more likely people living in it will use active travel modes. They compared people who were living in TOD areas and people who didn't and found out the percentage of residents lived in TOD areas' physical activity is 15% higher than people who didn't. (Langlois et al. 2016) On the other hand, some scholars think built environment around the transit node is not the main factor impacting the usage of the rapid transit. Mamdoohi and Janjany analyzed the factors affecting the frequency of travel by metro in a half-mile radius from stations and found out nearly half of the users taking the transit in Tehran were from the bus transfer and the prioritized factors were habits and waiting time. (Mamdoohi and Janjany 2016). Nevertheless, Langlois argued in one of his other articles that new people who moved to the TOD's catchment areas were less likely to use the rapid transit to commute and shopping but for

amenities and leisure trips. (Langlois et al. 2015)

5. Attitudes and Behavior

Since we are going to connect people's sentiment and their travel behavior, it is significant to know the general connections between them. The Planned Behavior theory tells us "Intentions to perform behaviors of different kinds can be predicted with high accuracy from attitudes toward the behavior"(Ajzen 1991). In a TOD catchment area, people with strong preferences for driving a car will lead to a high car use rate and decrease the level of physical activity (Langlois et al. 2016). In a case study in China, the authors asserted that people's subjective-norm and attitude toward reducing car usage is significant to reducing the car-transport (Liu et al. 2017). Social media, as an integral part of daily routine in the world, offers an open platform where people can share their attitudes and feelings about various of topics, which means to some extent, it can represents people's attitude (Öztürk and Ayvaz 2018). It can be used to analyze a wide range of aspects of our society like public health, political events and economic prediction and so on (Rout et al. 2018; Wang, Paul, and Dredze 2015; X. Zhang et al. 2018). Twitter, as a widely used social media platform, is an easy resource for us to access and manage.

6. Big-data and People's Behavior

Big data nowadays is widely used on many aspects of urban and social

issues, especially in transportation analysis. Satellite data is one of the common tools to assess the transportation issues because it can effectively locate people's real-time location and track people's travel route, mode and behavior in a specific region (S. Jiang, Ferreira, and Gonzalez 2017; Semanjski et al. 2017; Zhou et al. 2017). However, the GPS data that gets saved in the smartphone is hard for us to access in most of the cases, and the data population scholars can use in academic setting is also very limited. Tweets, defined as a specific category of big data, provide us a relatively easier way to analyze social issues by using big-data (Justin B. Hollander author 2016; Pavlicek and Novak 2015). The coordinate information saved in each tweet can help us understand people's movement and behavior.

Chapter 3: Methodology

1. Research Area

The research areas I focused on are 422 TOD catchment areas in the NYC which are located around the 422 subway stations launched by the MTA. (Figure 1) Based on the catchment radius, 0.5 mi and 2 km, the research areas can also be divides into two categories. Other than that, another kind of research area will be the area outside all the TOD's coverage (Figure 2).

(1) Catchment areas with 0.5 mi radius which include data of tweets, census tract level travel survey data, block group level pedestrian and vehicle usage as well as transit station.

(2) TOD catchment areas with 2 kilometer radius which contains data of tweets, Citi Bike facilities, travel survey data by census tract, pedestrian and vehicle statistical block groups and transit stations.

(3) The third one is places including the data of tweets, travel survey data by census tract, pedestrian and vehicle statistical block groups but other datasets.

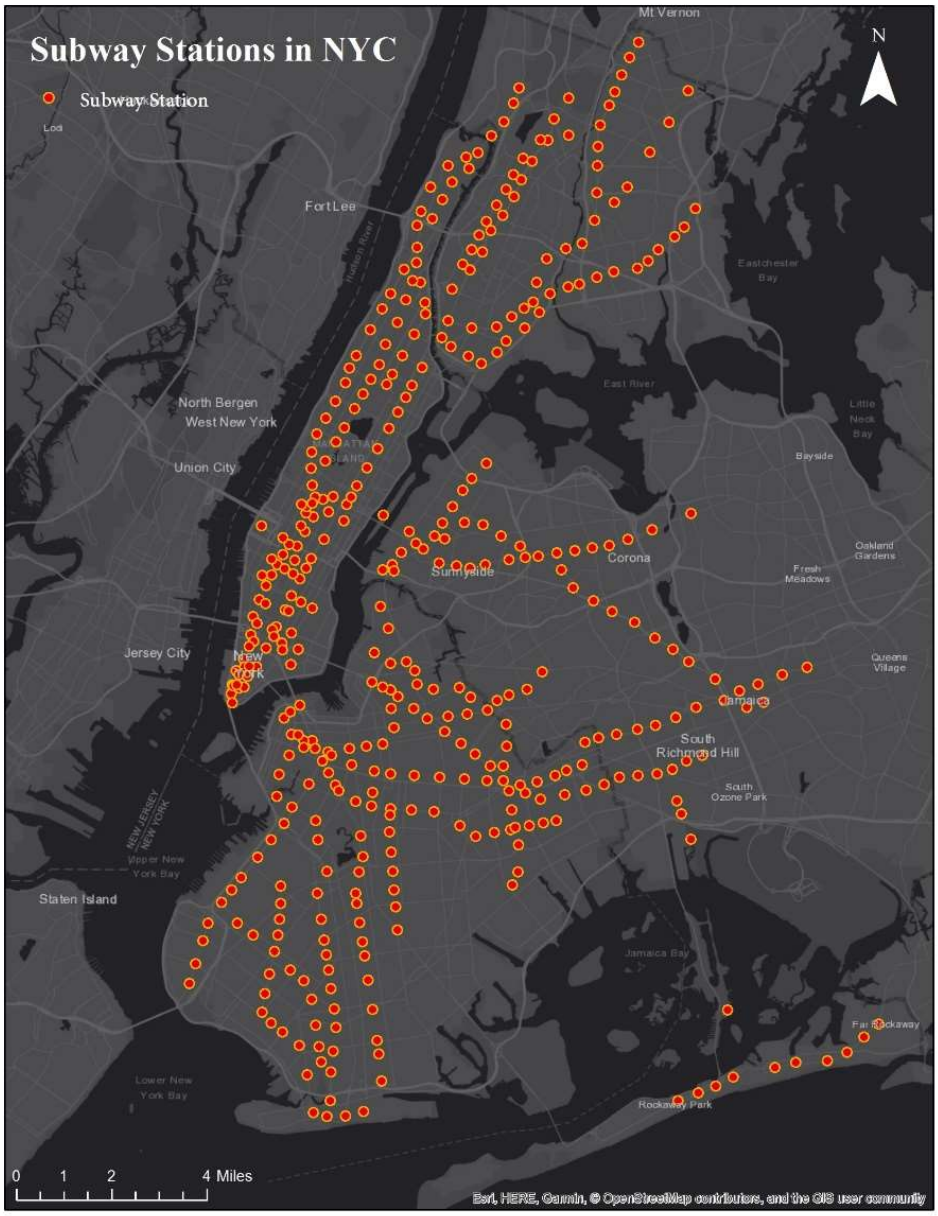

Figure 1: Subway Stations in NYC

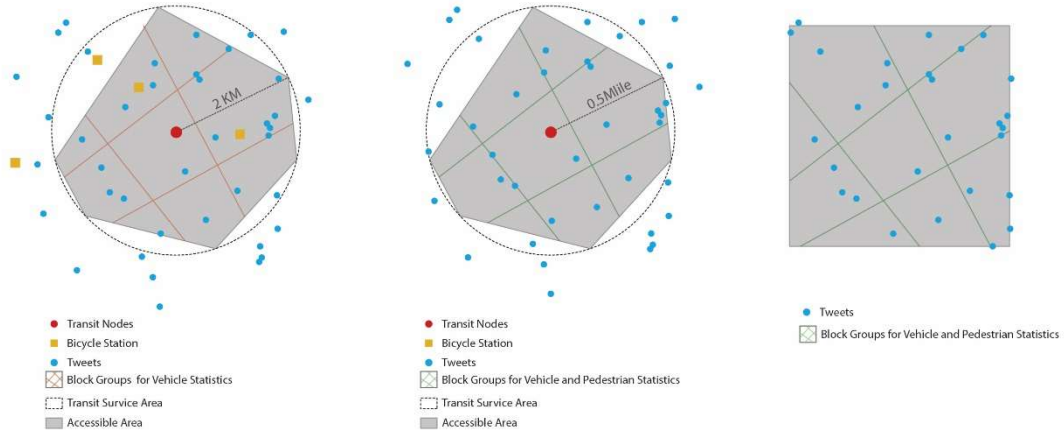

Figure 2: Research Areas

2. General Work Method

The general method for this thesis is followed by the steps listed below (Figure 3):

Firstly, we compared the travel mode choice and people's sentiment for each of the travel mode in transit sheds to test if people's sentiment score could impact their travel behavior.

Secondly, we merged all of people's sentiment score into the overall sentiment, then, tested the relationship between people's overall sentiment and travel mode usage.

The third, combining people's sentiment and built-environment data, we built a model to predict the vehicle usage rate in TOD catchment areas by using ANN.

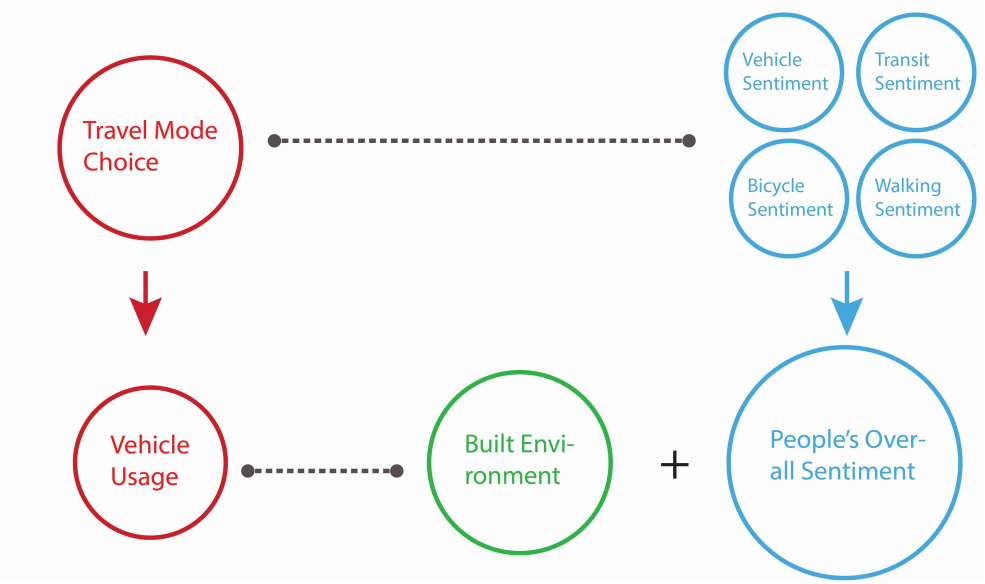

Figure 3: General Method

3. Specific Work Method

To find the connections between people's sentiment, travel behavior and built-environment in TOD catchment areas, all Twitter data, built-environment data and travel mode usage data covered by the transit service area need be assigned to each TOD catchment areas because the TOD catchment areas are relatively bigger than the blocks, block groups and census tracts. Therefore, the block group level travel mode usage data which comes from the American Community Survey (ACS), times people travel to work, could be assigned to each catchment area. RHTS data in census tract level with transit usage, vehicle usage and non-motorized modes usage were assigned to each catchment area as well as the built environment data which contained 16 parameters.

However, the statistical points for people's sentiment is not perfectly

distributed in our research area. It will lead to biases if we simply join those points in to the TOD catchment areas. The travel mode survey data from RHTS and real usage data in ACS can also lead to biases because those data are collected based on different geography sizes and different population sizes. The boundaries of the TOD catchment areas and the boundaries of those statistical geographical areas don't match. Therefore, the block groups and census tracts were converted to points first. Then all points were rasterized by using Inverse distance weighted (IDW) interpolation which is a ArcGIS tool used to measure values surrounding each statistical points to predict value for un-sampled locations.

Lastly, all the raster data was joined into the 0.5 mi radius and 2 km radius transit zones by using Zonal Statistics to Table tool. Transit ridership data was also joined into the TOD's catchment areas based on their mean value of ridership. Since the built environment data was saved as block level, which was far smaller than the TOD catchment area, those blocks were converted to points to join in the sheds directly (Figure 4).

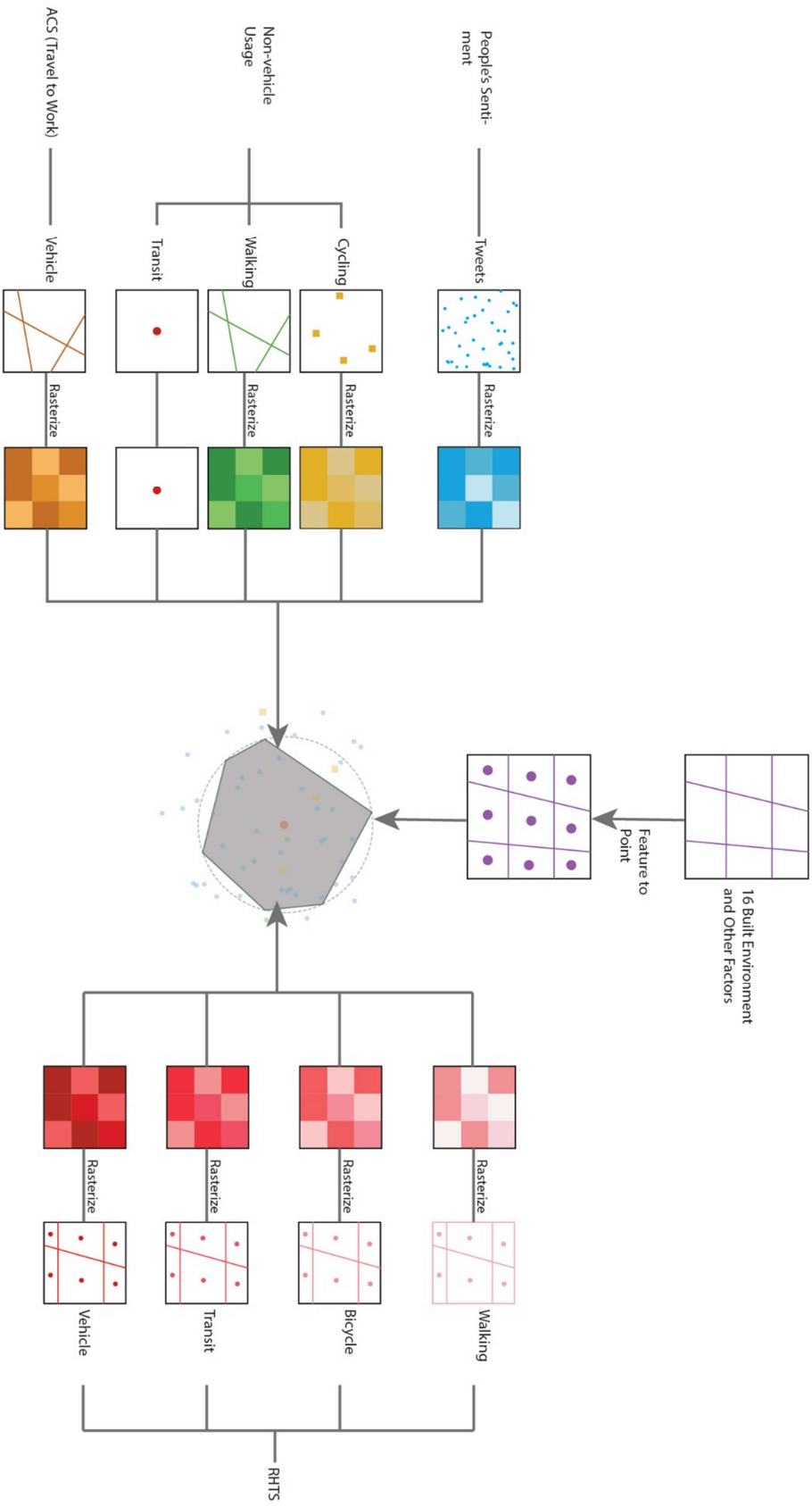

Figure 4: Specific Work Method

- People's sentiment and travel mode choice

Connections between people's sentiment score for different kinds of travel mode and the usage of each kind of travel mode drawn from ACS data, MTA and Citi Bike were tested by using ordinary least squares regression. RHTS data which represents the travel survey, was also measured with travel modes' usage to evaluate if they are associated with each other. (Figure 5) Finally, with the results of the measurement above, we can draw a conclusion on whether people's sentiment can actually take the place of the travel survey.

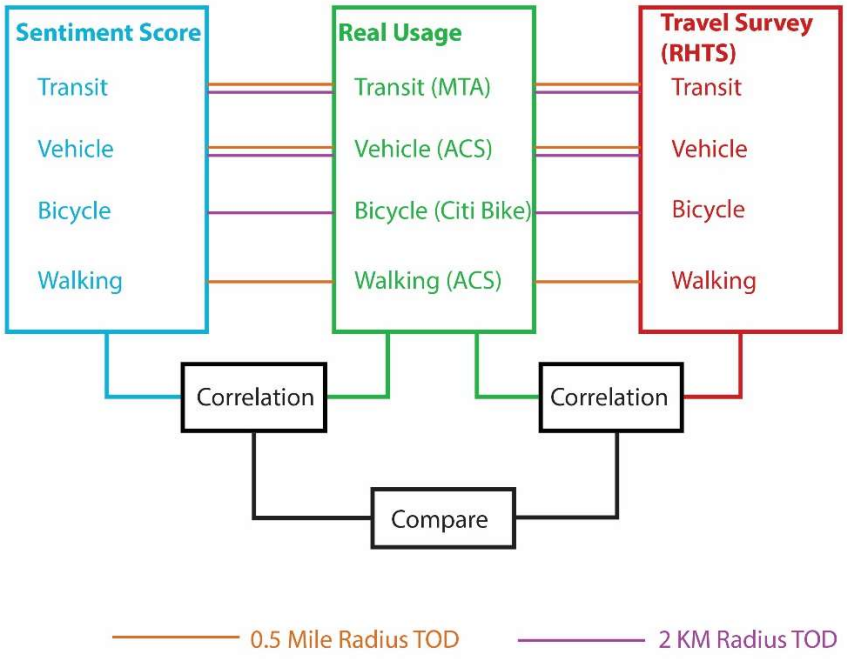

Figure 5: Compare Method

What's more, the overall sentiment score in each catchment area was calculated to compare to real travel mode usage. Whether people's general sentiment had a relationship with their travel behavior was tested. Other than

that, people's sentiment toward each kind of travel mode inside and outside the TOD catchment areas were also measured to investigate if the transit nodes really change people's mind about traveling.

- Built Environment and Vehicle Usage Rate

Based on the literature review in Chapter 2, the 7D concept is one of the most important forms of guidance while we are choosing our variables. Thus, 16 out of 90 indicators were chosen from Smart Location Database and the parking sign density data was chosen from NYC DOT (Table 5).

Table 5: Indicator Choose

D	**Indicators**
Density	D1a, D1c, D1c5_Svc10, D1d
Diversity	D2r_JobPop
Design	D3aao
Destination Accessibility	D5ar, D5br, D5cri, D5dri
Distance to Transit	--
Demand Management	Parking Signs and Locator
Demographics	Pct_AO0, Pct_AO1, Pct_AO2p, R_HiWageWk, E5_Svc10

To analyze the relationship between the built environment and the vehicle usage rate, 16 former selected factors, except for the overall sentiment score, were filtered so that the error of variable covariant would be exempted. We filtered our variables by using Exploratory Regression tool in ArcMap, which tries all possible combinations of explanatory variables to see which models pass all of the necessary OLS diagnostics. The maximum Variance Inflation Factor (VIF) value, which represents redundancy among explanatory variables

is set to be 7.5. The result is shown in Figure 6. We can see Ave_E5_SVC, AVE_DIC, AVE_D1C5_S, AVE_D1D, AVE_D5AR, AVE_D5BR, AVE_D5CRI, AVE_D5DRI are the variables that have high VIF values for all 0.5 mi radius sheds (Figure 6). For 2km radius sheds, only All_TW_2KM, AVE_PCT__1, AVE_D1A, AVE_D3AAO, passed the VIF test and other variables were all failed (Figure 7).

Summary of Multicollinearity*

Variable	VIF	Violations	Covariates
AVE_ALL_TW	1.30	0	--------
AVE_PCT_AO	5.13	0	--------
AVE_PCT__1	3.13	0	--------
AVE_PCT__2	3.67	0	--------
AVE_R_HIWA	4.44	0	--------
AVE_E5_SVC	11.87	1957	AVE_D1C (7.59), AVE_D1D (7.59), AVE_D5DRI (4.94), AVE_D5BR (4.94), AVE_D5CRI (4.31), AVE_D5AR (4.31)
AVE_D1A	2.68	0	--------
AVE_D1C	77.65	2671	AVE_D1C5_S (7.59), AVE_E5_SVC (7.59), AVE_D1D (7.59), AVE_D5BR (6.75), AVE_D5DRI (6.75), AVE_D5CRI (5.90), AVE_D5AR (5.90)
AVE_D1C5_S	11.87	1957	AVE_D1C (7.59), AVE_D1D (7.59), AVE_D5DRI (4.94), AVE_D5BR (4.94), AVE_D5CRI (4.31), AVE_D5AR (4.31)
AVE_D1D	91.48	2671	AVE_D1C5_S (7.59), AVE_E5_SVC (7.59), AVE_D1C (7.59), AVE_D5BR (6.75), AVE_D5DRI (6.75), AVE_D5CRI (5.90), AVE_D5AR (5.90)
AVE_D2R_JO	3.92	0	--------
AVE_D3AAO	1.96	0	--------
AVE_D5AR	19.21	1874	AVE_D1C (5.90), AVE_D1D (5.90), AVE_D5BR (5.72), AVE_D5DRI (5.72), AVE_E5_SVC (4.31), AVE_D1C5_S (4.31)
AVE_D5BR	30.46	2125	AVE_D1C (6.75), AVE_D1D (6.75), AVE_D5CRI (5.72), AVE_D5AR (5.72), AVE_E5_SVC (4.94), AVE_D1C5_S (4.94)
AVE_D5CRI	19.19	1874	AVE_D1C (5.90), AVE_D1D (5.90), AVE_D5BR (5.72), AVE_D5DRI (5.72), AVE_E5_SVC (4.31), AVE_D1C5_S (4.31)
AVE_D5DRI	30.41	2125	AVE_D1C (6.75), AVE_D1D (6.75), AVE_D5CRI (5.72), AVE_D5AR (5.72), AVE_E5_SVC (4.94), AVE_D1C5_S (4.94)
P_D	3.03	0	--------

* At least one model failed to solve due to perfect multicollinearity.
Please review the warning messages for further information.

Figure 6: Summary of Multicollinearity(0.5 mile)

Summary of Multicollinearity*

Variable	VIF	Violations	Covariates
ALL_TW_2KM	1.79	0	--------
AVE_PCT_AO	13.37	3245	AVE_D2R_JO (13.95), AVE_D1D (11.06), AVE_D1C (10.61), AVE_R_HIWA (8.69), AVE_D5BR (8.28), AVE_D5DRI (8.28), AVE_D5CRI (8.14), AVE_D5AR (8.14), AVE_D1C5_S (8.02), AVE_E5_SVC (8.02), AVE_PCT__2 (3.48)
AVE_PCT__1	6.55	0	--------
AVE_PCT__2	7.74	570	AVE_PCT_AO (3.48), AVE_R_HIWA (2.12), AVE_D1D (2.03), AVE_D2R_JO (1.98), AVE_D1C (1.76), AVE_D1C5_S (1.74), AVE_E5_SVC (1.74), AVE_D5BR (1.54), AVE_D5DRI (1.54), AVE_D5AR (1.51), AVE_D5CRI (1.37)
AVE_R_HIWA	9.65	2534	AVE_D2R_JO (10.46), AVE_D1D (9.57), AVE_D1C (8.83), AVE_PCT_AO (8.69), AVE_D5DRI (7.81), AVE_D5BR (7.65), AVE_D1C5_S (7.11), AVE_E5_SVC (7.11), AVE_D5CRI (6.67), AVE_D5AR (6.62), AVE_PCT__2 (2.12)
AVE_E5_SVC	28.11	2114	AVE_D2R_JO (9.45), AVE_PCT_AO (8.02), AVE_D1C (7.59), AVE_D1D (7.59), AVE_R_HIWA (7.11), AVE_D5DRI (5.52), AVE_D5BR (5.52), AVE_D5CRI (5.38), AVE_D5AR (5.38), AVE_PCT__2 (1.74)
AVE_D1A	5.63	0	--------
AVE_D1C	56.47	2805	AVE_D2R_JO (12.71), AVE_PCT_AO (10.61), AVE_R_HIWA (8.83), AVE_D1C5_S (7.59), AVE_E5_SVC (7.59), AVE_D1D (7.59), AVE_D5BR (7.31), AVE_D5DRI (7.31), AVE_D5CRI (7.14), AVE_D5AR (7.14), AVE_PCT__2 (1.76)
AVE_D1C5_S	28.11	2114	AVE_D2R_JO (9.45), AVE_PCT_AO (8.02), AVE_D1C (7.59), AVE_D1D (7.59), AVE_R_HIWA (7.11), AVE_D5DRI (5.52), AVE_D5BR (5.52), AVE_D5CRI (5.38), AVE_D5AR (5.38), AVE_PCT__2 (1.74)
AVE_D1D	95.39	2888	AVE_D2R_JO (12.99), AVE_PCT_AO (11.06), AVE_R_HIWA (9.57), AVE_D1C5_S (7.59), AVE_E5_SVC (7.59), AVE_D1C (7.59), AVE_D5BR (7.54), AVE_D5DRI (7.54), AVE_D5CRI (7.33), AVE_D5AR (7.33), AVE_PCT__2 (2.03)
AVE_D2R_JO	11.82	3774	AVE_PCT_AO (13.95), AVE_D1D (12.99), AVE_D1C (12.71), AVE_R_HIWA (10.46), AVE_D5BR (9.91), AVE_D5DRI (9.91), AVE_D5CRI (9.89), AVE_D5AR (9.89), AVE_D1C5_S (9.45), AVE_E5_SVC (9.45), AVE_PCT__2 (1.98)
AVE_D3AAO	5.25	0	--------
AVE_D5AR	28.52	2123	AVE_D2R_JO (9.89), AVE_PCT_AO (8.14), AVE_D1D (7.33), AVE_D1C (7.14), AVE_R_HIWA (6.62), AVE_D5BR (5.72), AVE_D5DRI (5.72), AVE_D1C5_S (5.38), AVE_E5_SVC (5.38), AVE_PCT__2 (1.51)
AVE_D5BR	66.44	2176	AVE_D2R_JO (9.91), AVE_PCT_AO (8.28), AVE_R_HIWA (7.65), AVE_D1D (7.54), AVE_D1C (7.31), AVE_D5CRI (5.72), AVE_D5AR (5.72), AVE_D1C5_S (5.52), AVE_E5_SVC (5.52), AVE_PCT__2 (1.54)
AVE_D5CRI	28.60	2123	AVE_D2R_JO (9.89), AVE_PCT_AO (8.14), AVE_D1D (7.33), AVE_D1C (7.14), AVE_R_HIWA (6.67), AVE_D5BR (5.72), AVE_D5DRI (5.72), AVE_D1C5_S (5.38), AVE_E5_SVC (5.38), AVE_PCT__2 (1.37)
AVE_D5DRI	66.58	2176	AVE_D2R_JO (9.91), AVE_PCT_AO (8.28), AVE_R_HIWA (7.81), AVE_D1D (7.54), AVE_D1C (7.31), AVE_D5CRI (5.72), AVE_D5AR (5.72), AVE_D1C5_S (5.52), AVE_E5_SVC (5.52), AVE_PCT__2 (1.54)
PA_DEN	4.07	0	--------

* At least one model failed to solve due to perfect multicollinearity.
Please review the warning messages for further information.

Figure 7: Summary of Multicollinearity (2km)

The variables with a VIF value over 7.5 were located. The decision of choosing which variables to drop was made according to their relationship with vehicle usage rate. The one with relatively lower correlation coefficient value with vehicle usage rate was eliminated (Figure 8, Figure9). The ones with stronger relationship to vehicle use were kept. Then, these kept variables were put into the Exploratory Regression again, and this process was maintained until no pair of variables were tested having VIF value higher than 7.5.

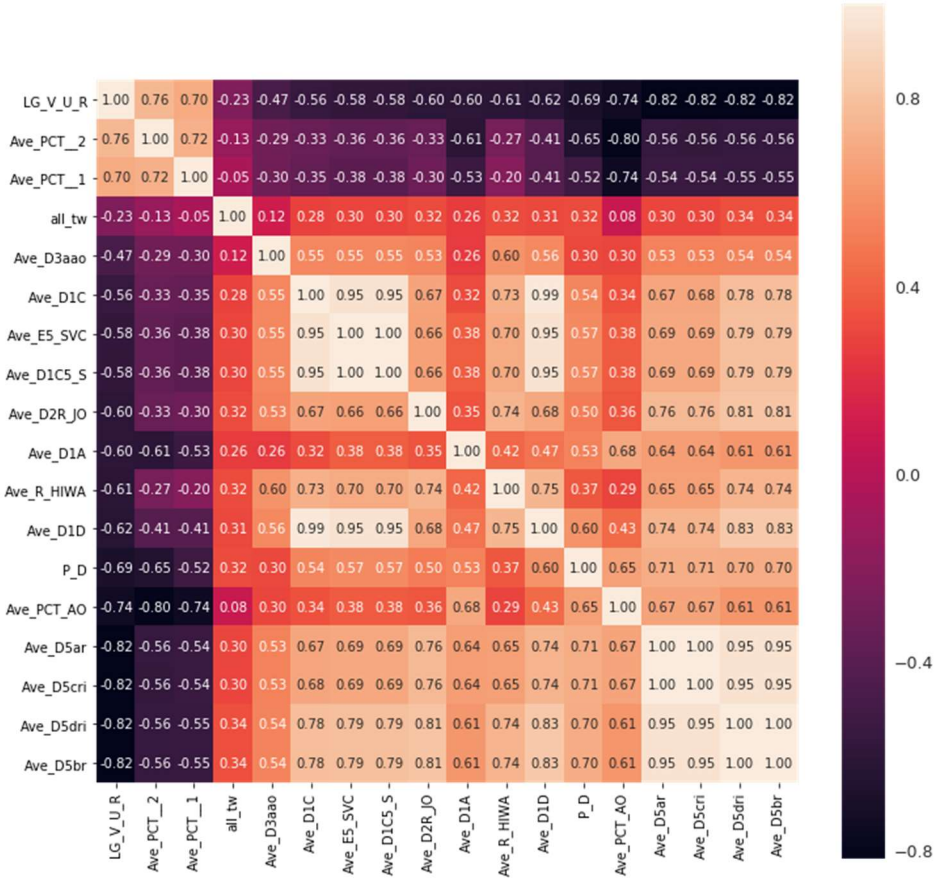

Figure 8: R Value Matrix (0.5 mile)

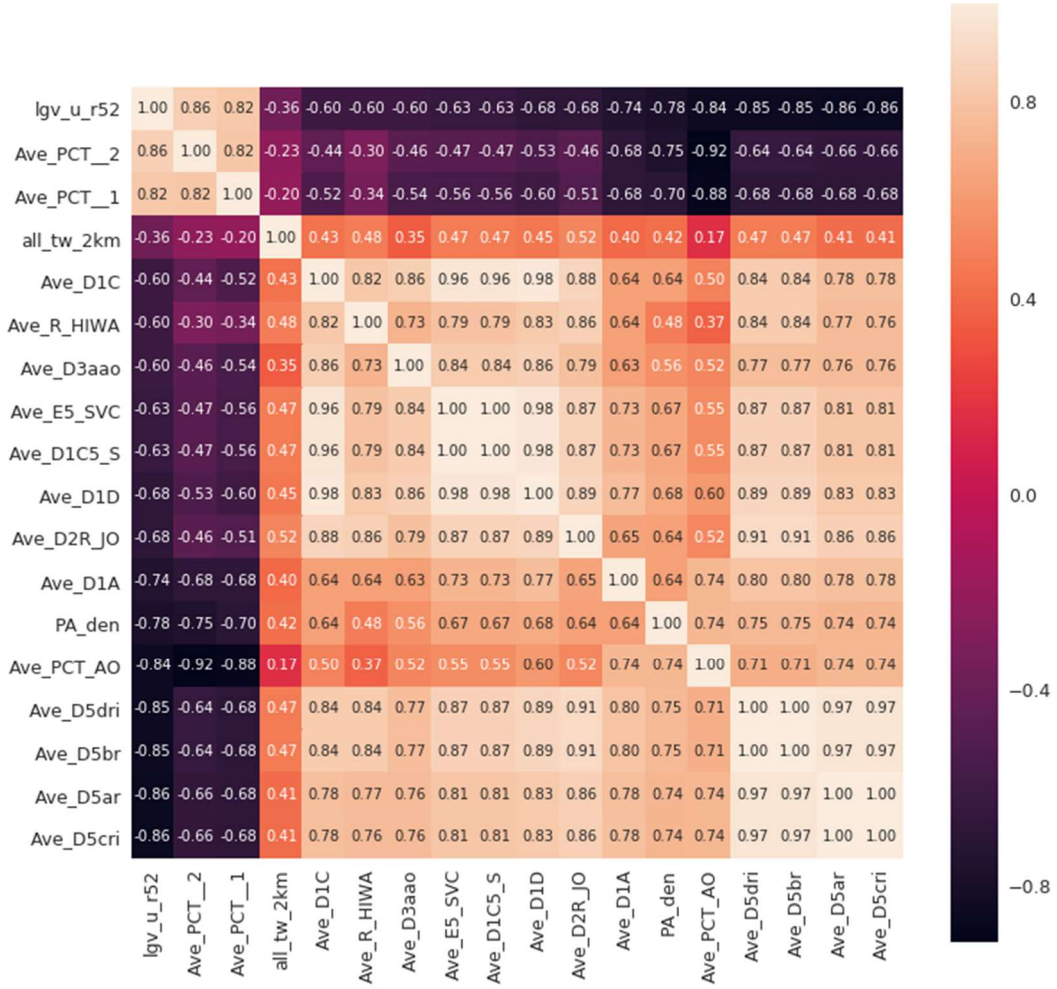

Figure 9: R Value Matrix (2km)

The variables we used to predict vehicle usage rate are listed in Table 6. The indicator AVE_D3AAO was also eliminated since was missing over 35% of observations:

Table 6: Selected Variables

Catchment Areas	**Chosen Variables**
0.5 mile	AVE_ALL_TW, AVE_PCT_AO, AVE_PCT__1, AVE_PCT__2, AVE_R_HIWA, AVE_E5_SVC, AVE_D1A, AVE_D1C5_S, AVE_D2R_JO, AVE_D5BR, AVE_D5DRI, P_D
2 km	AVE_ALL_TW, AVE_PCT__1, AVE_PCT__2, AVE_E5_SVC, AVE_D1A, AVE_D1C5_S,

	AVE_D5BR, AVE_D5DRI, AVE_D5AR, AVE_D5CRI, PA_den

- "Affin" Dictionary

By using Afinn sentiment keyword analysis, which is a list of 1476 English words rated for valence with an integer between minus five (negative) and plus five (positive), we can easily quantify people's sentiment. For instance, "evil" will be scored -3, "rejoice" will be scored 4. For this thesis, we scored people's emotion words in tweets and words related to each travel mode will be selected to filter the raw tweets. Table 7 illustrates the key words we chose related to each kind of travel mode.

Table 7: Key Words Related to Travel Modes

Travel Mode	**Keywords**
Bicycle	Bike, Bicycle, Bicycling, Cycling
Vehicle	SUV, Auto, Motor, Van, Car, Cars, Automobile, Drive, Driving, Vehicle, Truck
Subway	Subway, MTA, Rapid transit, Metro
Walk	Walk, Walked, Walking, On Foot, Pedestrian

Using statistical analysis software, R, the Affin dictionary was loaded and located 3,000 to 7,000 tweets for each of the travel modes. These tweets are the ones contain both key words for travel modes, like bicycle, vehicle, walk, etc. and sentiment words, like happy, glad, love, etc. Figure 10 represents a small set sample of the filtered dataset (Figure 10). Column B is the Twitter ID, column C is the original ID in the big dataset, column D is the user's ID, column

E is Tweets, and columns F and G display the longitude and latitude of the tweets, column H is the time the tweets were posted and the I column is the overall sentiment score we calculated by summarizing all keyword's individual score.

A	B	C	D	E	F	G	H	I
	ID1	X	ID2	Tweets	Longtitude	Latitude	Time	score
1	8.555E+17	112	167425739	991.2 GT3. 500HP and available with a manual transmission. Next car territory.â€¦ https	-73.94828042	40.64677197	2017-04-21 19:02:53	0
2	8.555E+17	138	28674721	Sweet New Promo Picture for Extinct. Or a portrait of me driving on the 405. You pick. I	-74.0007613	40.7207559	2017-04-21 19:02:57	2
3	8.555E+17	361	470009021	#romanholiday/Pleasantville mashup coming soon to a drive-in near you... @ New Yorl	-74.0064	40.7142	2017-04-21 19:12:43	0
4	8.555E+17	391	14772499	SPICY goat pepper soup meat pies. Real Nigerian food from a street truck! Only in NYC	-73.9611	40.68025	2017-04-21 19:14:01	0
5	8.5551E+17	1259	761640096	Little something to brighten chrisb515 day.... @ 260 auto repair inc https://t.co/JRAJkM	-73.9838028	41.2036362	2017-04-21 19:49:34	0
6	8.5552E+17	1893	2940230107	RS5! Photo by rs.ix | Follow audiwhips --- #audi #audiloverr #audilover #carporn #car #	-74.0647	40.7114	2017-04-21 20:15:58	0
7	8.5552E+17	1963	42640432	Overturned vehicle left lane blocked in #Union on I-78 EB approaching X52 stop and g	-74.24601	40.7052	2017-04-21 20:19:05	-3
8	8.5552E+17	2199	42640432	Overturned vehicle left lane blocked in #Union on I-78 Express Ln EB approaching X52	-74.24601	40.7052	2017-04-21 20:31:00	-3
9	8.5552E+17	2233	19637118	When you walk into the parking garage and your car is parked at the front of the line.â€	-73.99159	40.72965	2017-04-21 20:32:46	0
10	8.5552E+17	2323	42640432	Disabled truck in #AreaOfTheLowerEastRiverCrossingsNyc on The Brooklyn Brg EB at Th	-74.00429	40.71291	2017-04-21 20:36:35	0

Figure 10: Sample Filtered Tweets

Although using sparse Twitter data is not a robust platform for finding the routes people move, we can still easily discover the connections between people's attitudes and the urban space. For instance, after analyzing these tweets, we can know where people have positive attitudes or negative attitudes toward bicycles in the city and which TOD catchment area contends people's highest sentiment toward subway.

- Interpolation Analysis

Because all of the point data, tweets, RHTS data, ACS data and bicycle usage data, can lead to error if we assign them into catchment areas directly. So, we chose to use Inverse Distance Weighting (IDW) Interpolation tool in GIS. It can help us generate raster for point data. The advantage of this tool is it can evaluate the value of non- value pixel around a valued pixel based on spatial autocorrelation. For this thesis, we used IDW to estimate the average

sentiment score and travel mode usage in missing data spaces. The detailed settings are shown in Figure 11. To naturalize error caused by over impact of near pixel, the minimum number of weighted neighborhood data was set to be 2 and the maximum weighted neighborhood data was set to be 6.

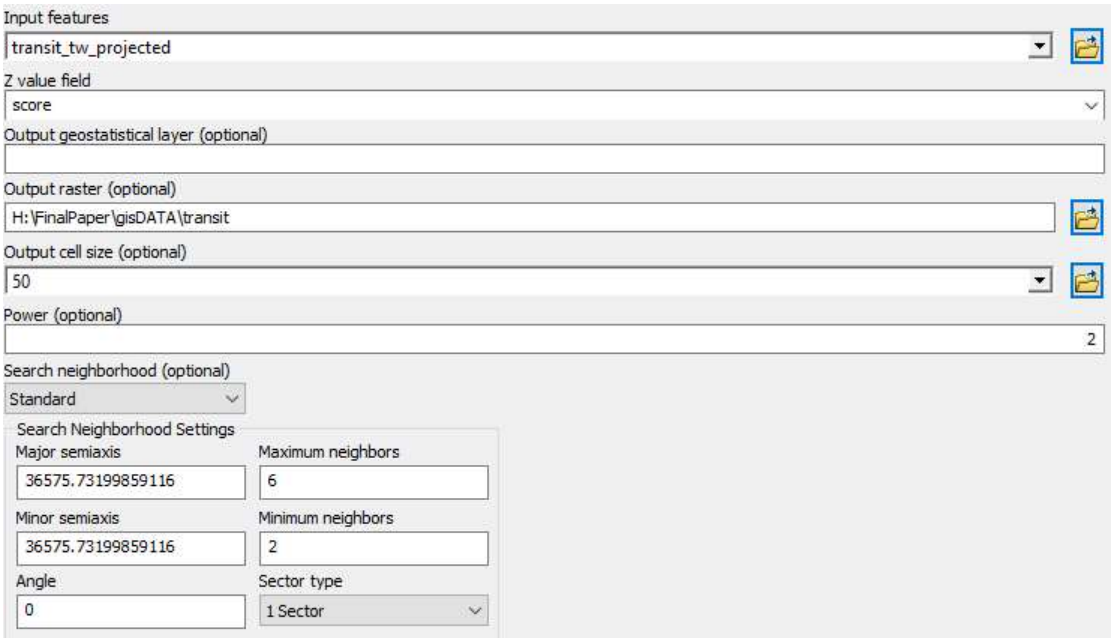

Figure 11: IDW Setting

Then, all rasterized data was assigned into TOD catchment areas with 0.5 mi radius and 2 km radius by using Feature Statistics to Table tool in ArcGIS. The mean value of each data set was calculated.

- Network Analysis

The traditional way to define TOD catchment areas are circles centered by transit stations. It is rough because people have to move along the roads. So, the catchment shed will be generated by the network around the transit stations depend on the route length of 0.5 mi radius and 2 km radius, which

represents the accessibility of walking and biking. 422 transit stations were located as original locations to build service areas around. To make sure every service area has same catchment radius, instead of breaking by the catchment area borders, they were set to overlap each other.

- Local Indicator of Spatial Association (LISA)

Based on Anselin's paper, the LISA means "for each observation, it gives an indication of the extent of significant spatial clustering of similar values around that observation"(Anselin 1995). In this thesis, I implemented the LISA for people's travel behavior and sentiment score to see if there were spatial relationships of them for different travel modes. Local Moran's I value can in some level indicate the neighboring features' similarity index. Positive value for I represents that a feature has neighboring features with similarity high or low attribute values.("Anselin Local Moran's I" n.d.) Negative value for I indicates that a feature has neighboring features with dissimilar values. By analyzing the patterns of those clustering neighborhoods we can know the spatial distribution of travel preference and the hot spots of people's sentiment.

- Artificial Neural Network (ANN)

When we try to analyze the connections between variables, the common way in the academic world is generating an Ordinary least Squares (OLS) regression to minimize the squared distances between the observed and the

predicted dependent variables. However, for a non-linear model, OLS is no longer the best way to predict the outcome. Some scholars like Hichem Omrani, Deepak Agrawal and his co-workers points out that Artificial Neural Network is a better model than Multinomial logistic regression (MNL) and Support Vector Machine (SVM) to forecast the dependent variable especially in the realm of travel mode and market share.(Agrawal and Schorling 1996; Omrani 2015) The advantage of the ANN model is that it can generate invisible layers by calculating weights for each of the variables. According to Heaton, "neural networks without a hidden layer will not model any non-linear function, but with a hidden layer they can approximate/estimate the relationship between the influencing factors and the outcome"(Heaton 2008). Plus, weights not only exist between input variables and hidden layer, but also exist between hidden layers and the outcomes. This process creates virtual factors to improve the whole model. For this work, the selected variables will be joined into the network as inputs to predict the outcome of vehicle usage rate (Figure 12).

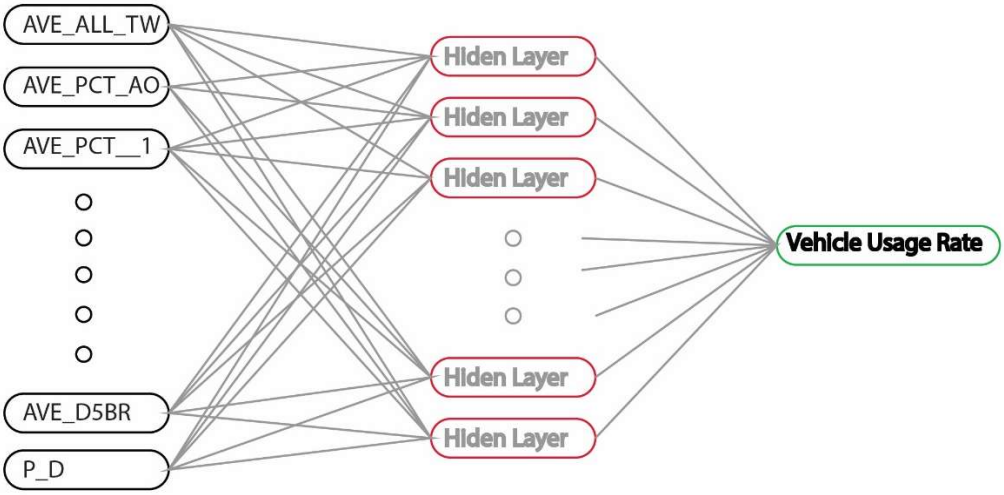

Figure 12: Neural Network

Multi-layer Perceptron (MLP) is a supervised learning algorithm in ANN which can create a model with a function:

$$f(x): R^m \; to \; R^o$$

by training on a dataset. The letter “m” is the number of dimensions for input and the letter “o” is the number of dimensions for output. Given a set of indicators X1, X2, X3…, and a target y, the model can learn a non-linear function for classification or regression.

The advantages of MLP is it can not only predict non-linear models but also perform real-time predictions. In my thesis, we made of the first advantage.

Chapter 4: Travel Modes in TOD Areas

1. Subway Ridership

There are 472 subway stations located in NYC, which is the largest number of public transit subway stations of any system in the world. MTA holds the annual records for all passengers (other than NYC Transit employees) who enter the subway system for 422 stations from 2011 to 2016. The entry records of each of the transit stations were based on people who entered from outside the subway system, but it didn't cover the people who transfered from another subway lane. In my paper, I used the average workday ridership in 2016 as a research object. After analyzing the ridership data, we learned that the maximum ridership belongs to the Times Square Station, which had 202,363 average entries for each work day in 2016. On the contrary, the minimum ridership, 289 entries for each day, was held by Beach 105 St. Station. The mean ridership for all 422 stations of workday in 2016 was 13,402 (Table 8).

Table 8: Subway Ridership in 2016 Workday Description

	Value	**Note**
Maximum	202363	Time Square Station
Minimum	289	Beach 105 St. Station
Mean	13402	-
Standard Deviation	19052	-

As we can see in Figure 13, most of the high ridership subway stations are concentrated in the middle of Manhattan. Some are gathered in high subway

route density areas like Atlantic Terminal Station area in Brooklyn and lower Manhattan areas. Some others are distributed along line 7 and F line in Queens. In general, Manhattan island's subway stations have relatively higher workday ridership than the stations located in Queens and Brooklyn. In the east of NYC, terminal stations have relatively higher average workday ridership.

To ensure that the variable we use to do the regression later is normally distributed, I transformed the ridership into log(Ridership).

$$Ridership\ Variable = \log Ridership$$

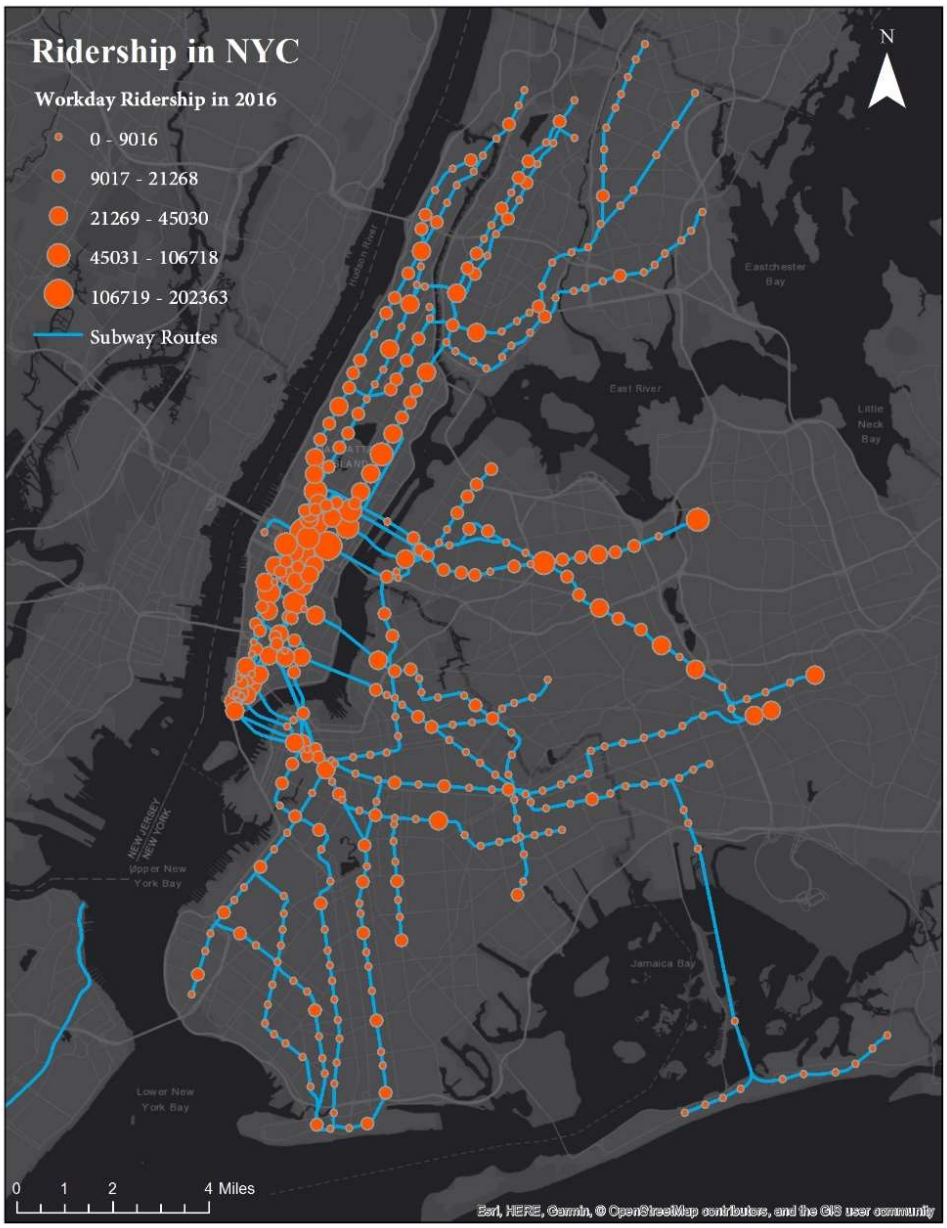

Figure 13:Subway Ridership in NYC

2. Bicycle Usage

According to Citi Bike, there are 619 share bicycle stations in NYC. The Citi Bike records every ride trip of the share bicycle for each station they have. The data had been processed to remove trips that were taken by staff as they

serviced and inspected the system and any trips that were below 60 seconds in length. We can see in the data description in Table 9, the max share bicycle usage for four months, 59989 trips, happened at Pershing Square and the minimum usage station, Sedgwick Ave Station only had 8 trips.

Table 9: Bicycle Usage Description

	Value	**Note**
Maximum	59989	Pershing Sq. North
Minimum	8	Sedgwick Ave
Mean	10076	-
Standard Deviation	8470	-

In terms of the geographical analysis, we found out that the middle and lower Manhattan areas had relatively high bicycle usage. Bronx, Brooklyn and Queens also had a few high bicycle usage areas that were close to Manhattan Island's. The hottest points were located to the east of Time Square and around Borough of Manhattan Community College (Figure 14).

After rasterizing the bicycle usage data, it was assigned to 186 TOD catchment areas with 2 km radius which at least contain one Citi Bike station in their catchment areas. The result shows that the max bicycle use TOD area is 14 St./8 Ave, which is located in central Manhattan and has 20870 trips from April 2016 to July 2016. On the other hand, the lowest bicycle usage TOD area is Mosholu Pkwy, which is located in the Bronx with only 160 bike trips for the whole four months. In general, people in central Manhattan and lower Manhattan seem to prefer riding a bicycle but people in other boroughs do not.

Figure 15 illustrates how those TOD catchment areas are geographically gathered. It shows 72 high bicycle usage TOD areas are gathered together in central and lower Manhattan and other 92 low bicycle usage catchment areas are concentrated in Queens, Brooklyn and upper Manhattan. Weighting by threshold of 2 km, the local Morans'I value for the bicycle usage of these areas is 0.95 which means they have high spatial relationship.

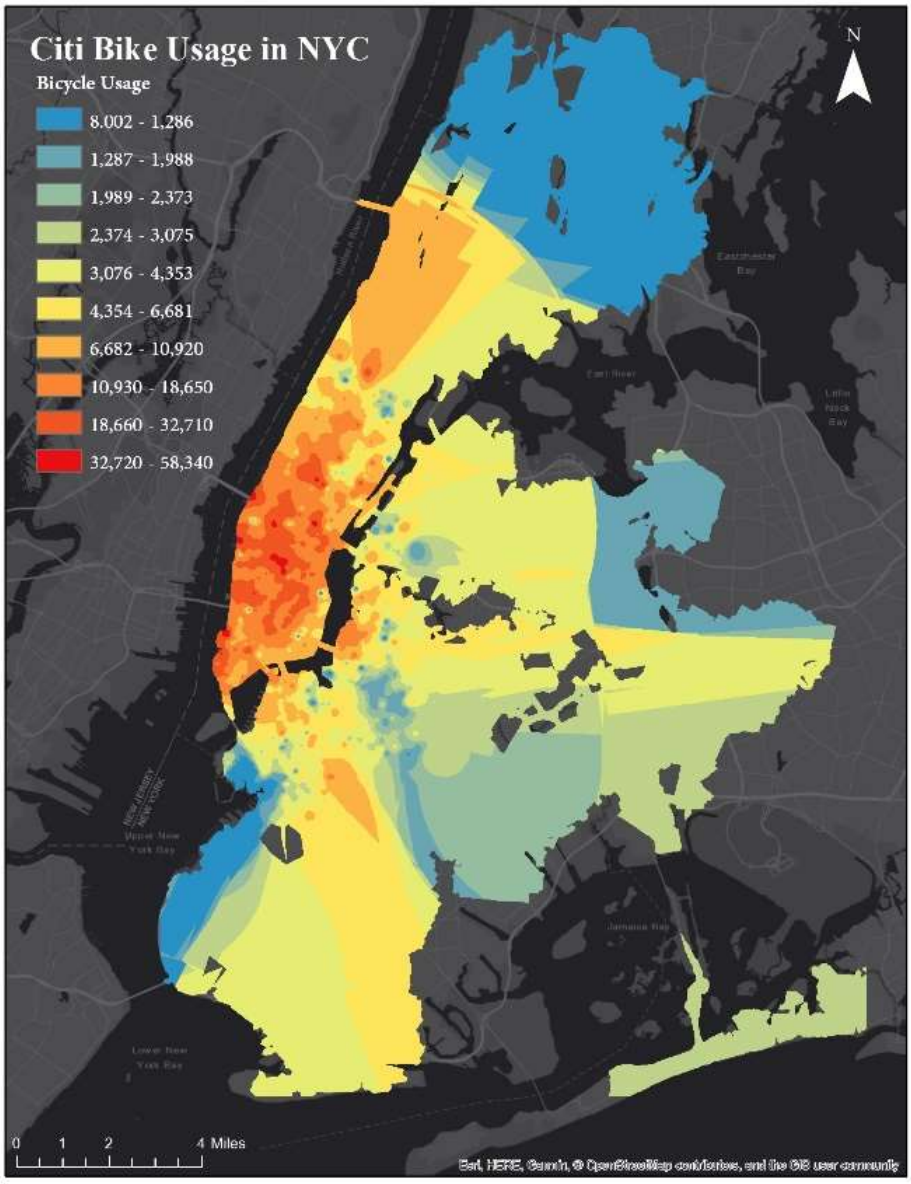

Figure 14: Geospatial Relationship

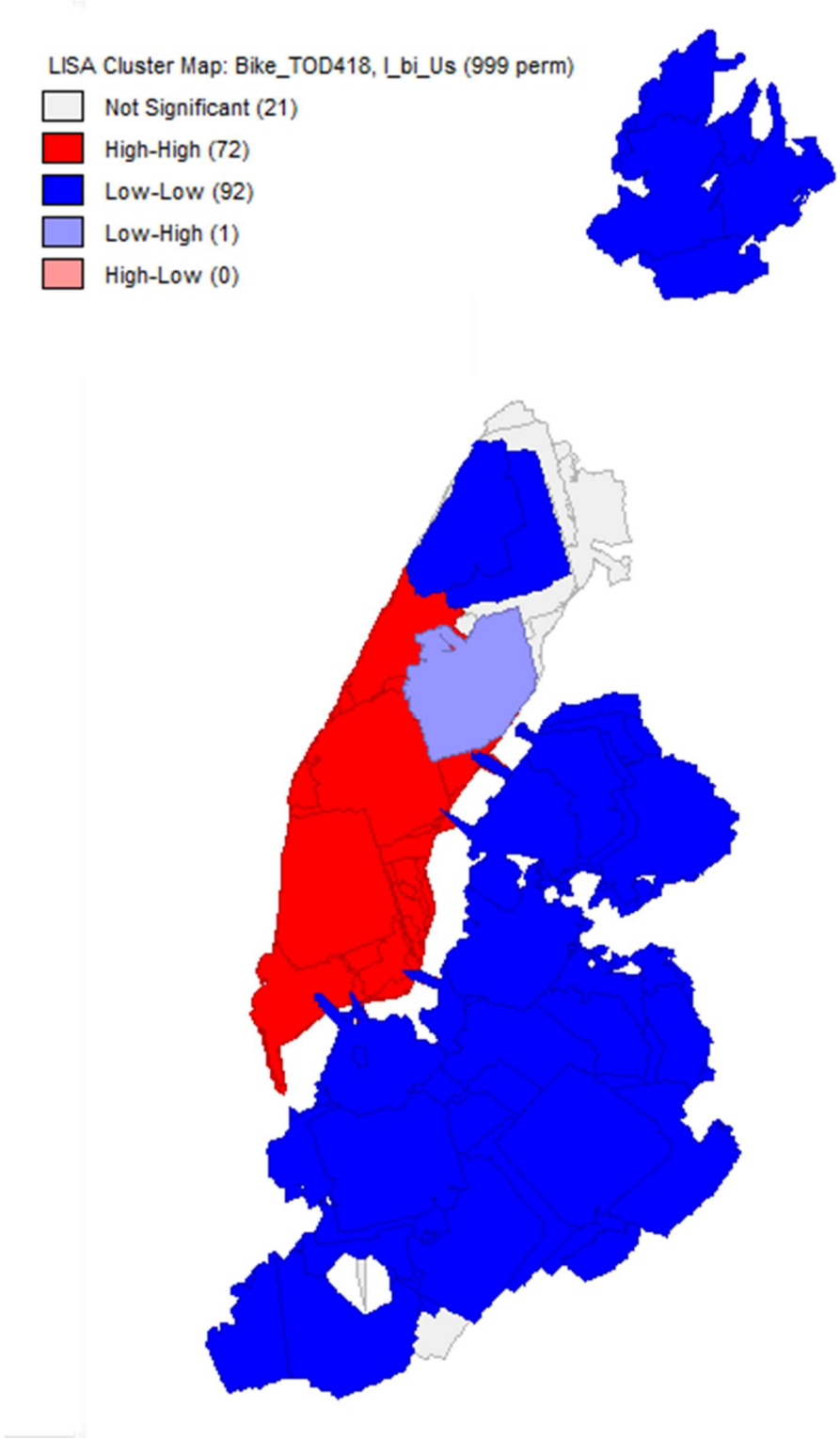

Figure 15: Citi Bike Usage in NYC

3. Vehicle Usage

The vehicle usage data came from the U.S. census data set. In this article, 11238 block groups' data were collected. Most of the block groups are in NYC and others are located in New Jersey. Since different block groups have different spatial areas, vehicle usage in each block group should be normalized by their spatial value to waive the error. What's more, as the American Community Survey data was collected by self-report, the vehicle usage should also be normalized by the total travel mode population to eliminate bias.

$$V_{ia} = \frac{Rsterize\frac{VU_i}{S_i}}{Rasterize\frac{T_i}{S_i}}$$

V_{ia}: Vehicle Usage Rate for a TOD catchment area

VU_i: Vehicle Usage for a Block Group

S_i: Block Group Area

T_i: Total Travel Mode Usage for a Block Group

By analyzing the vehicle usage rate data for each TOD catchment areas with a different radius, we found that people prefer to travel by vehicle in 2km zones rather than 0.5 mi zones. In 0.5 mi radius catchment areas, the max vehicle usage rate in 2016, 71.09%, appeared at Beach 90 St. Station and the minimum was 4.03% at South Ferry/Whitehall St. Station. On the other hand, in 2 km radius transit zones, the max vehicle usage rate is 58.59% which

appears at Broad Channel Station, and the minimum is 6.71% at New York University Station. The mean value for the vehicle travel mode rate for 0.5 mi zones and 2 km zones are 17.83 and 18.84%. (Table 10)

Table 10: Vehicle Usage Rate Description

	Maximum	**Minimum**	**Mean**	**Maximum Station**	**Minimum Station**
0.5 Mile Radius	71.09%	4.03%	17.83%	Beach 90 St.	South Ferry/Whitehall St.
2 km Radius	58.59%	6.71%	18.84%	Broad Channel	New York University

Figure 16 shows the vehicle usage density in NYC. We discovered that most share of the Manhattan area had relatively low vehicle usage density especially in Middle Manhattan and Lower Manhattan areas. Also, people west of the Brooklyn and Queens areas had relatively low vehicle usage rates. Relatively high vehicle usage density appears around the Central park area, Upper West Side, Upper East Side, most of areas in Upper Bronx and east of Queens and Brooklyn.

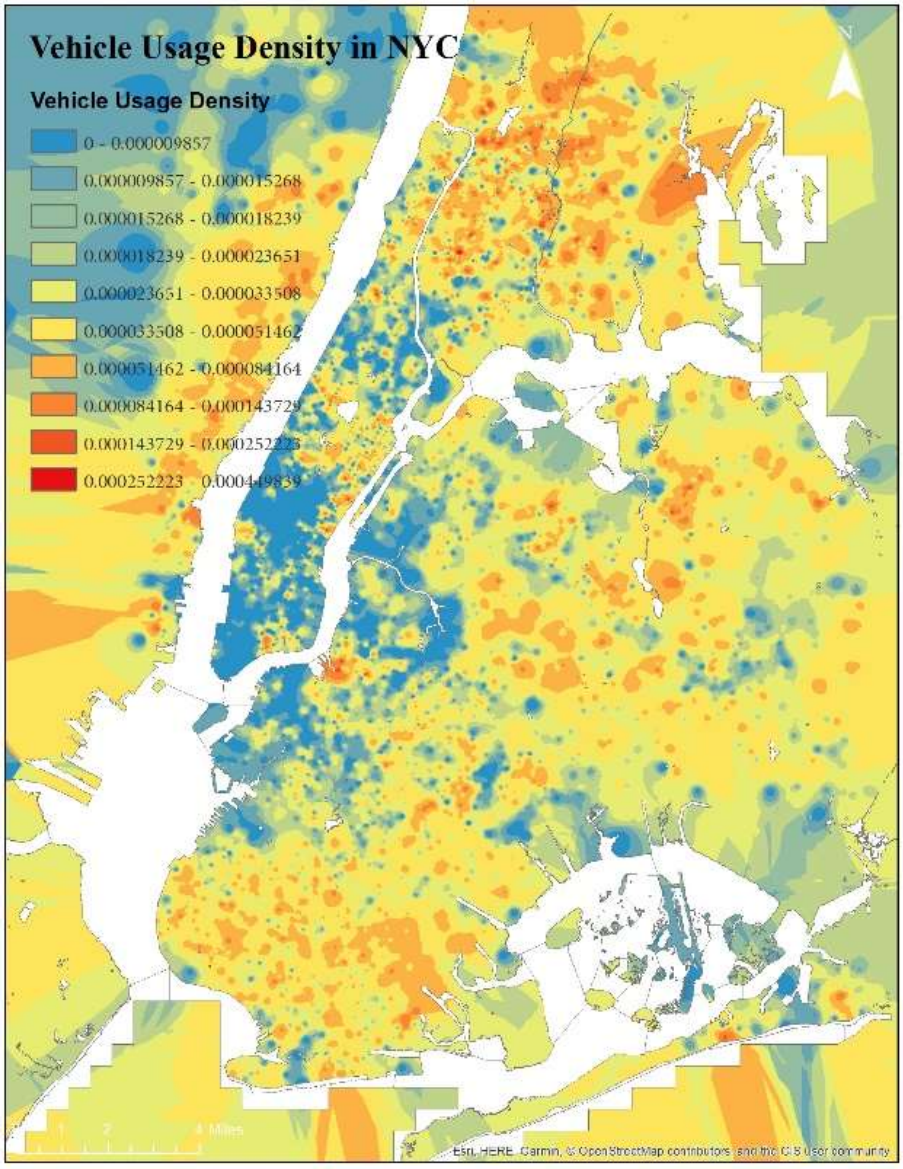

Figure 16: Vehicle Usage Density in NYC

We also found that TOD areas with low vehicle usage, both for 2km radius and 0.5 mi radius, were highly concentrated in middle and lower Manhattan areas. Both of these two kinds of areas' local Morans'I value are larger than 0.8 which mean they are highly spatial correlated. For 0.5 mi radius zones, 35 high-high areas were found and they are gathered near the edge of NYC's boundary far away from the low vehicle usage places. On the other hand, if we expand

our TOD radius to 2km, the high-high areas are gathered more obviously and increased to 111 units from 35 units. Most of the TOD catchment areas in the east Queens, Brooklyn and upper Bronx were high drive rate places, whereas the low auto-use areas grew from 90 units to 183 units as well (Figure17, Figure18).

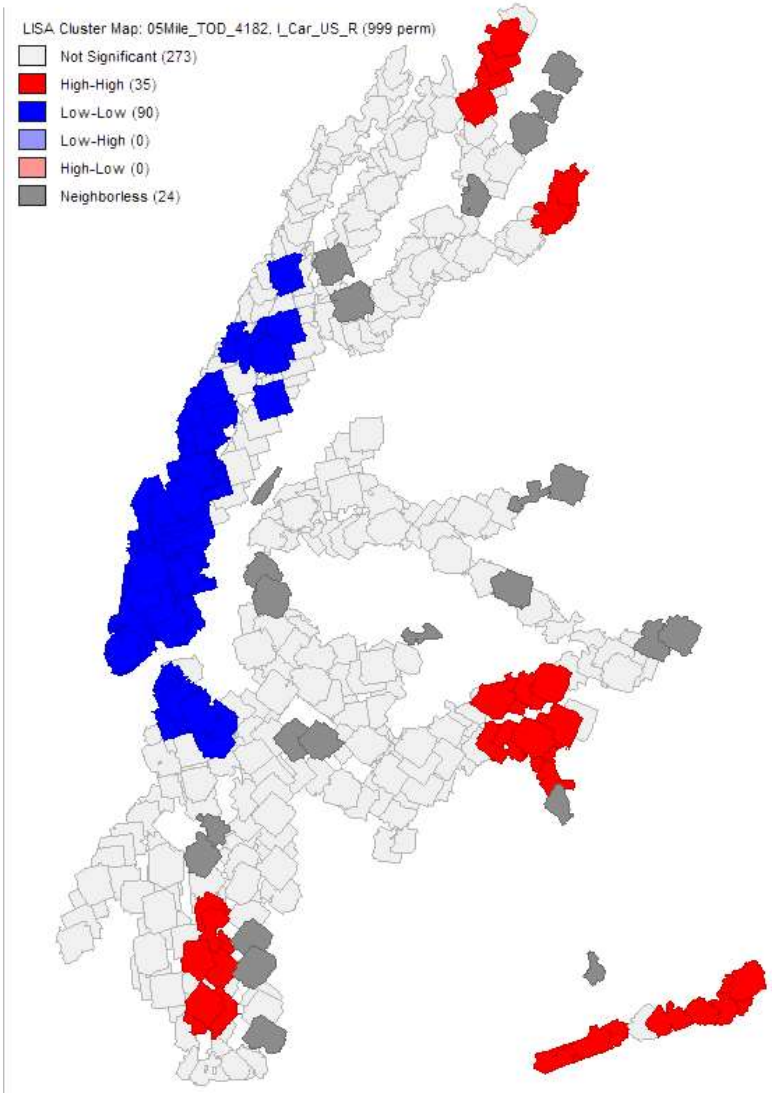

Figure 17: Spatial Relationship for Vehicle Use Rate (0.5 Mile)

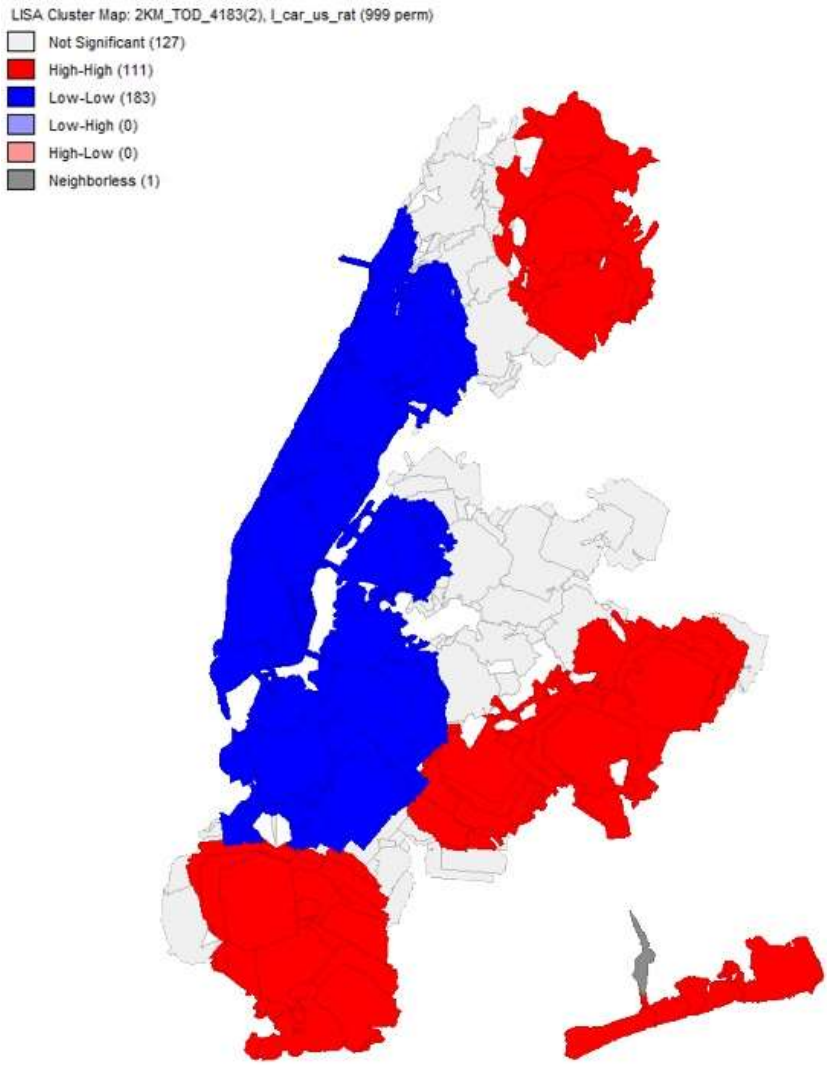

Figure 18: Spatial Relationship for Vehicle Use Rate (2 km)

4. Walking

Travel by walking data for block groups' level was also drawn from the U.S. census data set. Similar to the vehicle data, it was normalized by the geographical area and total usage for all travel modes. Because the walking

distance for people in 5 minutes is about 0.5 mi, so the walking rate was only assigned to TOD catchment areas with 0.5 mi radius. Based on the data analysis, the maximum walking rate appears at Broad Channel Station with 46.77% of people who commute by walking. The minimum walking rate is in Zerega Ave Station's catchment area with only 2.73% of people who chose walking as their commute travel mode. (Table 11)

Table 11: Walking Data Description

	Value	**Note**
Maximum	46.77%	Broad Channel Station
Minimum	2.73%	Zerega Ave Station
Mean	13.06%	-
Standard Deviation	9.38%	-

Figure 19 explains the walking density in NYC. We can see it is still the Manhattan area that have relatively high walking travel density. Transit service areas also have high walking rate density are the Navy Yard and the Kensington areas in Brooklyn. The low walking density areas are located at central Queens. The spatial relationship map (Figure 20) verifies our results. It shows the high walking ratio TOD catchment areas are gathered in middle and lower Manhattan and two other places in Brooklyn. The low walking ratio places are gathered in east Brooklyn and Queens.

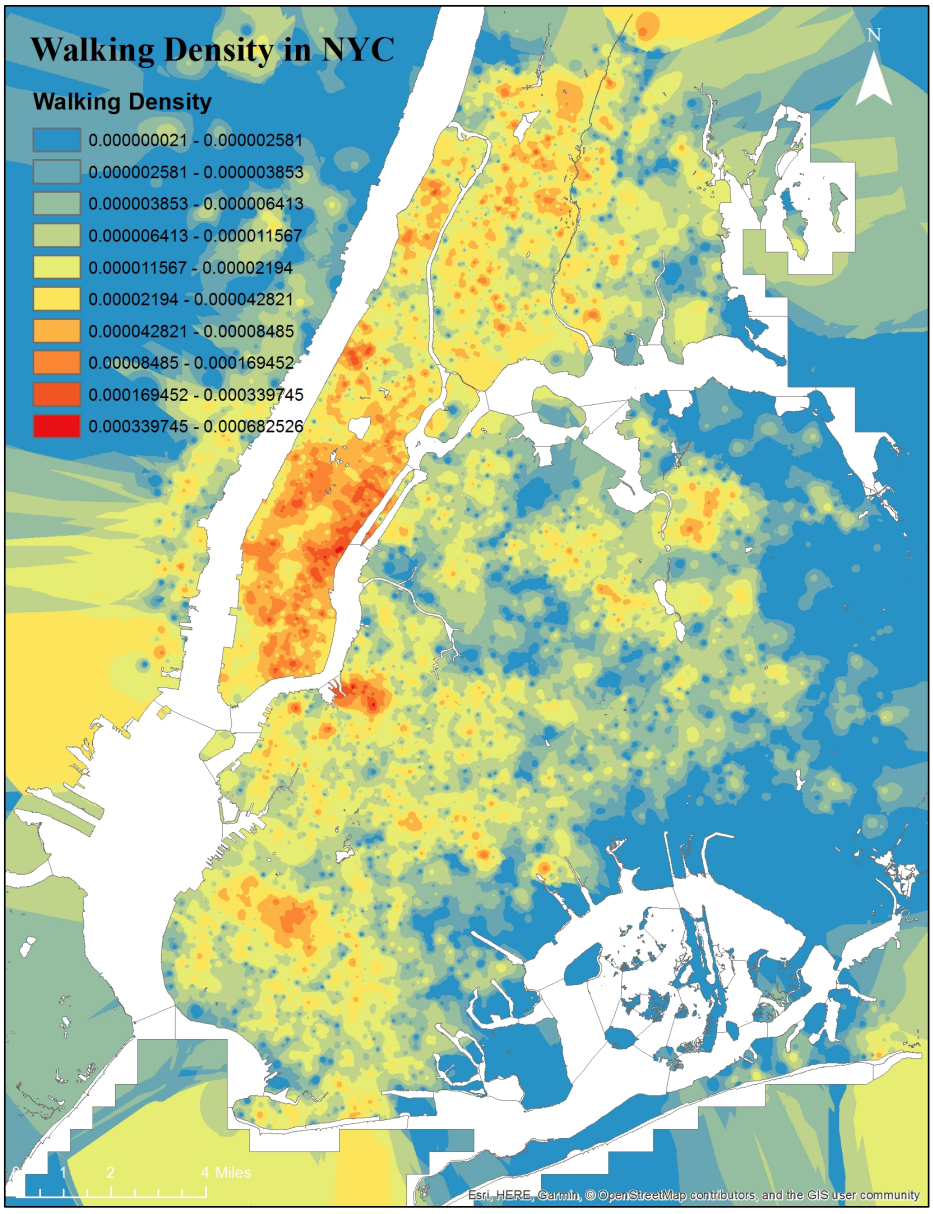

Figure 19: Walking Density in NYC

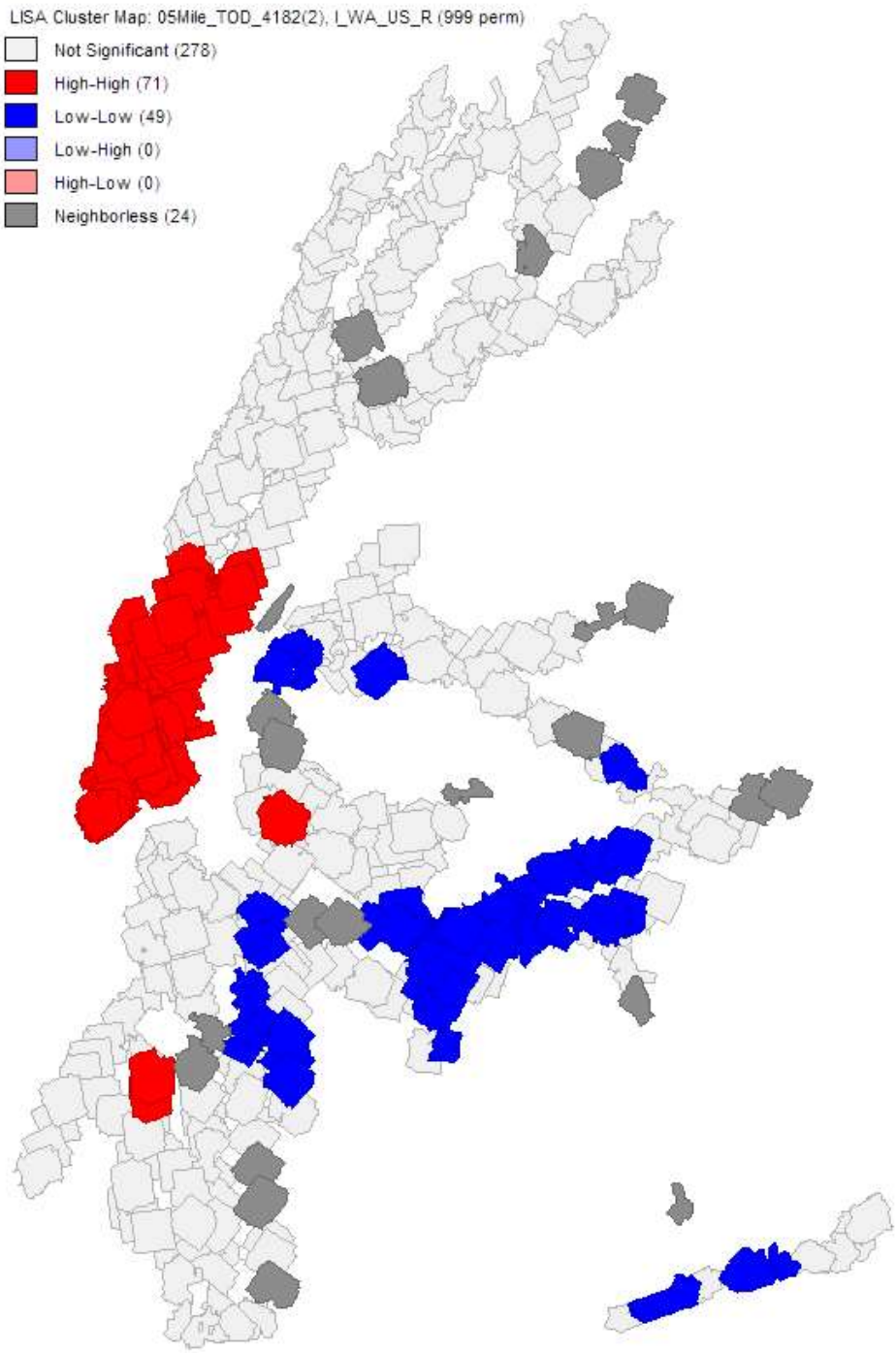

Figure 20: Spatial Relationship of Walking Ratio

5. Differences of Travel Survey and Reality

The travel survey data, drawn from 2010 to 2011 Regional Household Travel Survey, contains personal level all travel modes' trip records. However,

the data population is not as perfect as we expected. Over 50% of the census tract in RHTS has no bicycle and subway records. Thus, we dropped those TOD areas with limited data. Also, to make the travel survey data and the real usage data more comparable, we converted the vehicle and walking usage into percentages to make them comparable to other travel usage data. To see if different catchment radii have different outcomes, 0.5 mi radius and 2 km radius transit sheds were analyzed separately. (Table 12, Table 13)

Table 12: Travel Survey-Real Usage Comparison in 0.5 Mile Catchment Areas

Real Usage	**Travel Survey**	**Available TOD areas**	**Coefficient**	**P value**	**R2**
Log Subway Ridership	Log Reported Subway Usage	369	0.66	0.00	0.32
Log Vehicle Usage Rate	Log Vehicle Usage Rate	420	0.61	0.00	0.09
Log Walking Rate	Log Walking Rate	416	0.65	0.00	0.09

For 0.5 mi subway service areas, 369 sheds include travel survey data of the subway usage. The coefficient of the travel survey for subway and the real subway usage data is 0.66. The R squared value is a low value of 0.32, which illustrates the relationship of the predict variable and predicted variable is weak. Then, according to the analysis of 420 transit sheds, the result gave us a negative outcome, which the correlation coefficient for the vehicle use is only 0.1 and the coefficient is 0.52. In terms of the walking travel mode, there is a big difference between the reality and the travel survey. The R squared value

is only 0.09 for all 416 statistical TOD catchment areas which contain data. However, even though the R squared value for all these three modes in 0.5 transit sheds are low, they are still statistically significant as the P values are close to 0. The coefficient for these three regressions are all close to 0.6, which means the travel survey under-estimated the real usage of each kind of travel modes for 40%.

Table 13: Travel Survey-Real Usage Comparison in 2 km Catchment Areas

Real Usage	**Travel Survey**	**Available TOD areas**	**Coefficient**	**P value**	**R^2**
Subway Ridership	Reported Subway Usage	421	0.8	0.00	0.43
Vehicle Usage Rate	Vehicle Usage Rate	422	0.81	0.00	0.35
Bicycle Usage	Bicycle Usage	165	1.62	0.00	0.29

For 2 km radius transit sheds, the outcome shows that there are 165 TOD areas that have travel survey data for bicycle. The value of correlation coefficient with the real bicycle usage data is 0.29. After expanding the catchment radius for subway stations from 0.5 mi to 2 km, we witness an increase of the correlation coefficient value for vehicle usage from 0.09 to 0.35. The highest R squared value belongs to the subway regression, which reaches 0.43. We can see in 2 km radius transit sheds that the relationship between the travel survey and the real usage become stronger. Both coefficient values for subway data and vehicle data are close to 1, which means those two travel

modes' travel surveys are relatively accurate.

From the analysis above, we know that the travel survey indeed associates with the real usage of the travel modes. The association for the dependent variables and independent variables in 2 km catchment radius areas is stronger than 0.5 mi catchment radius areas'. Whereas, since the correlation coefficient value are all lower than 0.5, the travel survey may still not be a good tool to predict people's travel behavior in transit sheds.

Chapter 5: People's Sentiment Analysis and Travel Mode Choice

1. People's Sentiment Analysis and Travel Mode Choice

- People's sentiment toward subway

As we can see in table 14, people's sentiment scores toward subway travel in both 0.5 mi radius catchment areas and 2 km radius catchment areas are all positive. The sentiment score toward the subway for all people in 2 km sheds is higher than 0.5 mi ones. Although the average sentiment score in the subway sheds are very close, 0.5 mi radius' sheds have more extreme sentiment score values than 2km sheds since the standard deviation of them is far higher than the 2km's.

Table 14: Subway Attitude Description

Category	**Score (0.5 mile)**	**Score (2km)**
Minimum	-2.648	-0.722
Maximum	1.965	1.020
Sum	55.362	58.802
Mean	0.131	0.139
Standard Deviation	0.456	0.271

After assigning the mean sentiment score of transit into the subway sheds with 0.5 mi, we used GeoDa to explore the spatial connections between them. In 0.5 mi radius sheds, we found that there were 67 highly connected catchment areas, 42 of which were High-High clustered, which means they all had relatively high sentiment scores and spatial connections. We also found

that the higher sentiment areas are gathered in the middle and lower Manhattan, Boerum Hill and Bushwick in Brooklyn and 167 St to 174 St in Bronx areas. On the other hand, 25 transit sheds are considered low sentiment score cluster areas. Most of them are gathered in Cypress Hills, Woodhaven and Clinton Hill in Brooklyn and Harlem area and West Farm area in Bronx. (Figure 21) For 2 km transit sheds, 115 areas are positively spatially related and they were mostly gathered in most the Manhattan area, west of Queens and Brooklyn and Rockaway Park areas. The low sentiment core areas are located in the middle of Brooklyn and the Bronx. (Figure 22)

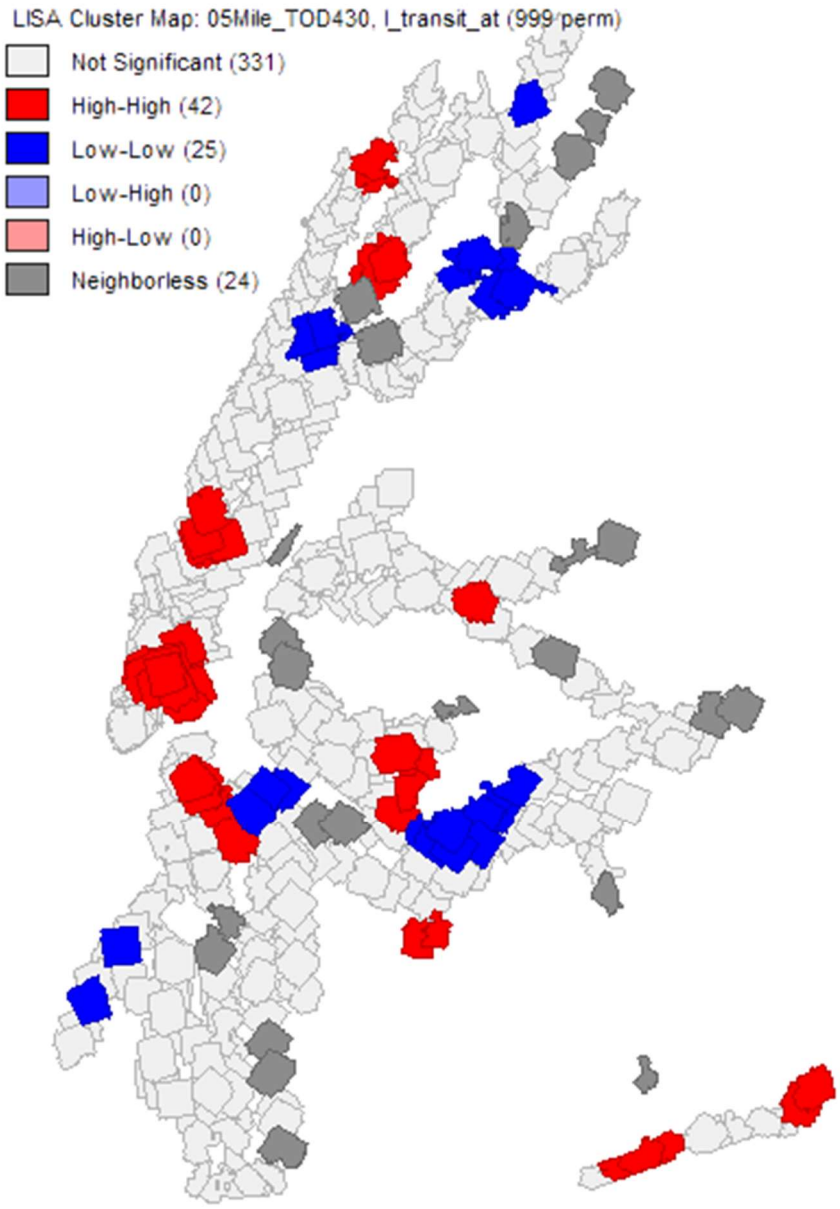

Figure 21: Spatial Relationship of People's Sentiment of Subway (0.5 mile)

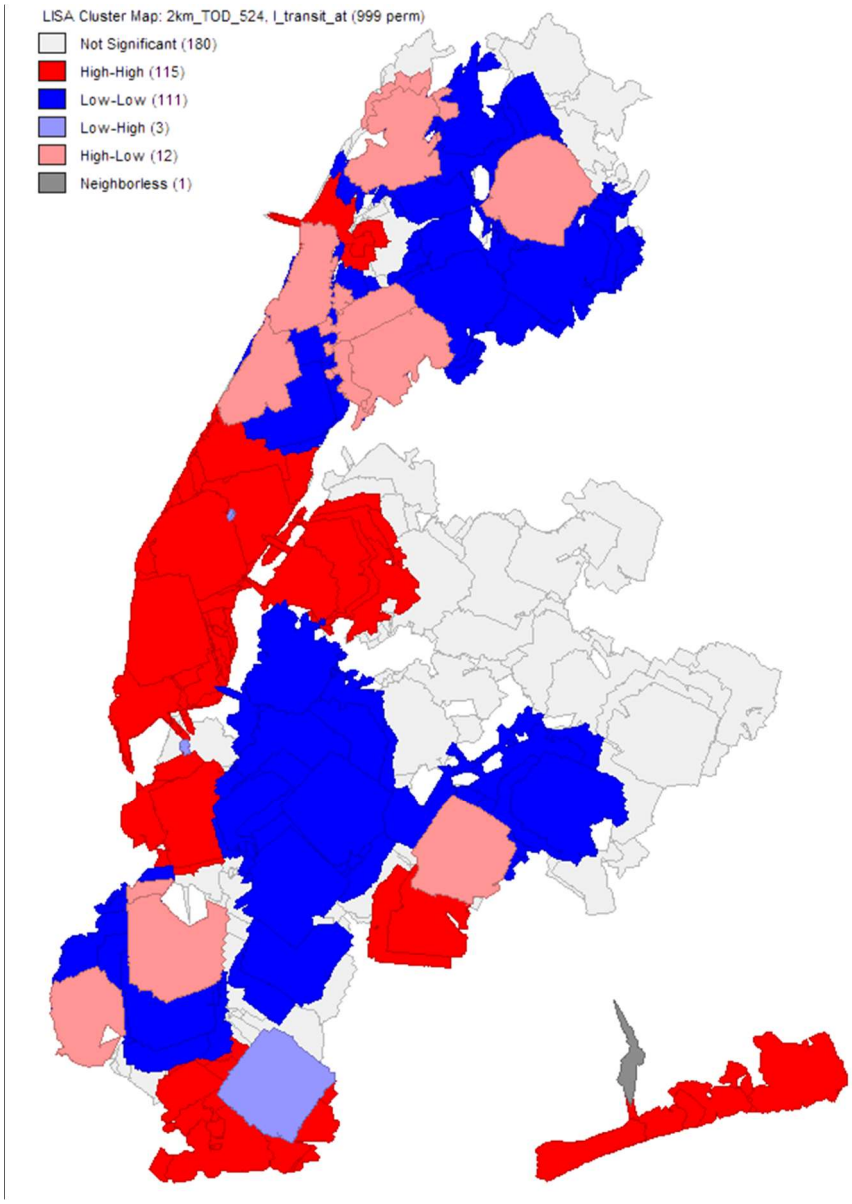

Figure 22: Spatial Relationship of People's Sentiment of Subway (2km)

- People’s sentiment toward walking

Table 15 illustrates the sentiment score in 0.5 radius catchment areas of all 422 subway stations. Compared to the subway, people have a relatively more positive sentiment toward walking and the discrimination from one person’s attitude to another is very big as the standard deviation is 0.9.

Table 15: Walking Attitude Description

Category	**Score (0.5 mile)**
Minimum	-5.946
Maximum	4.394
Sum	263.053
Mean	0.623
Standard Deviation	0.914

Because 0.5 mi was always treated as the walking accessible distance for subway stations, we chose 0.5 mi transit sheds as our walking sentiment analysis target. By using Univariate Local Moran's I spatial analysis tool in GeoDa, we found out that unlike the subway sentiment score's space distribution, the gathered high sentiment score areas were located randomly in Queens, Bronx and Brooklyn and most of the TOD catchment areas were not significantly correlated. The low sentiment score areas were concentrated in north Bronx and east of Queens. (Figure 23)

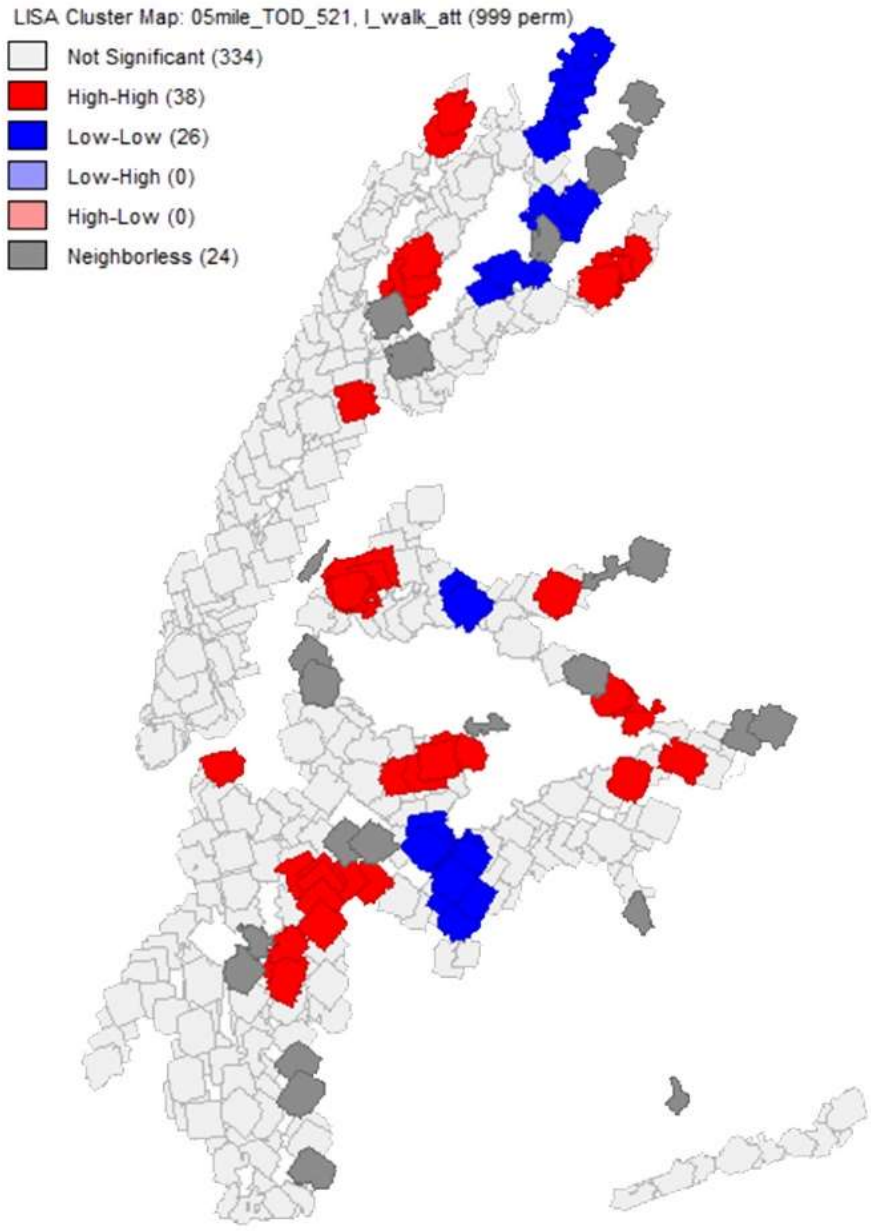

Figure 23: Spatial Relationship of People's Sentiment of walking (0.5 mile)

- People’s Sentiment Toward Vehicle

Something similar occurred with the transit attitude analysis, where we

separately analyzed people's sentiment score in both 0.5-mi radius transit service areas and 2 km radius transit services areas. The results showed that in 0.5 mile sheds, people have relatively positive attitudes toward vehicles rather than people in 2 km sheds, but people's sentiment scores are closer to each other in 2km sheds than 0.5 mi sheds.(Table 16)

Table 16: Vehicle Attitude Description

Category	**Score (0.5 mile)**	**Score (2km)**
Minimum	-3.497	-1.183
Maximum	2.571	1.742
Sum	62.991	44.293
Mean	0.149	0.105
Standard Deviation	0.800	0.462

Similar to the walking clustering distribution, the high vehicle sentiment score cluster catchment areas for 0.5 radius sheds are randomly located in NYC. (Figure 24) Low-Low clustering areas are mostly distributed in the Bronx and a few places in Brooklyn and Queens except Manhattan. On the contrary, 2 km radius TOD catchment areas with high vehicle sentiment score are gathered in Manhattan, upper Bronx and middle Brooklyn. (Figure 25)

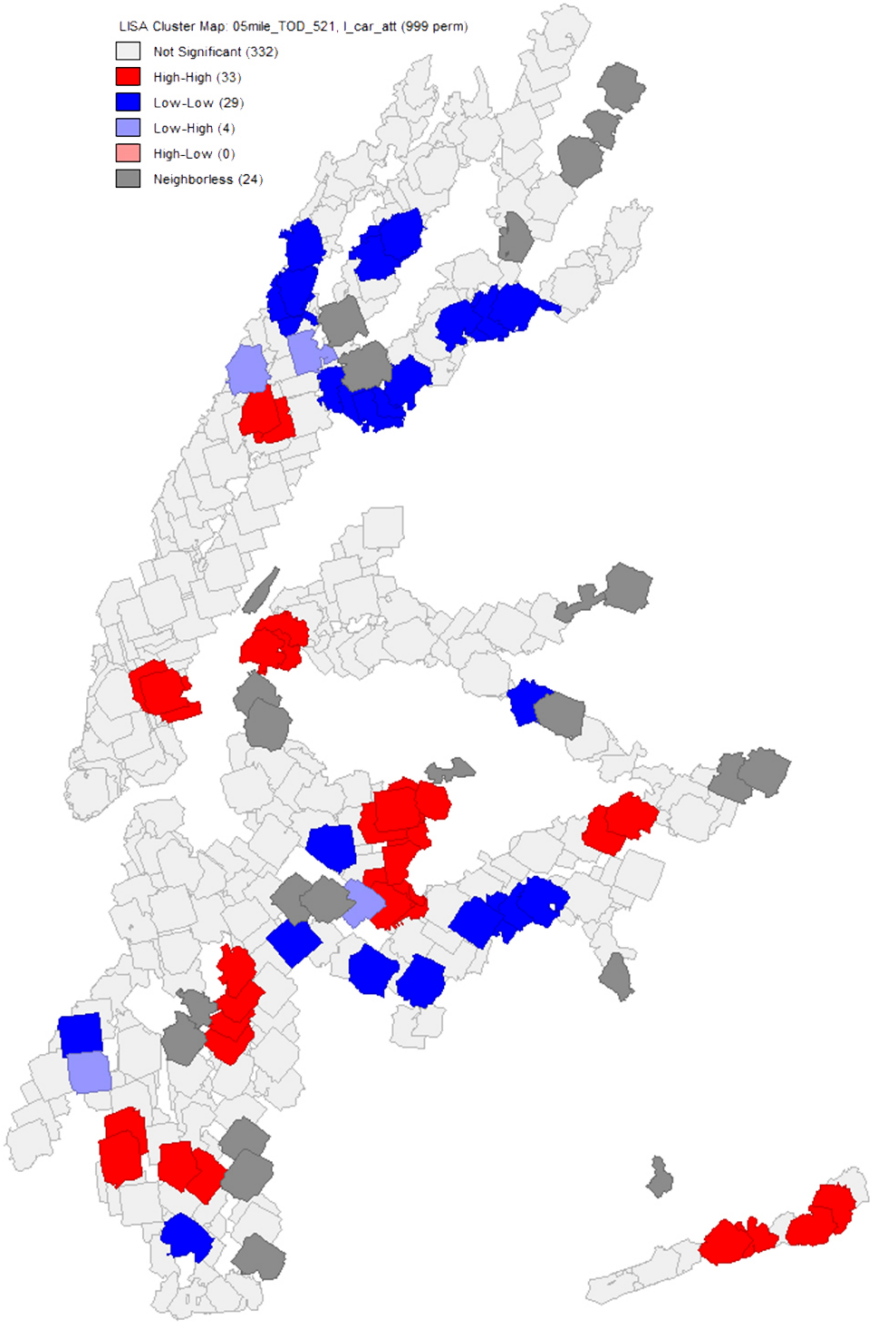

Figure 24: Spatial Relationship of People's Sentiment of Vehicle (0.5 mile)

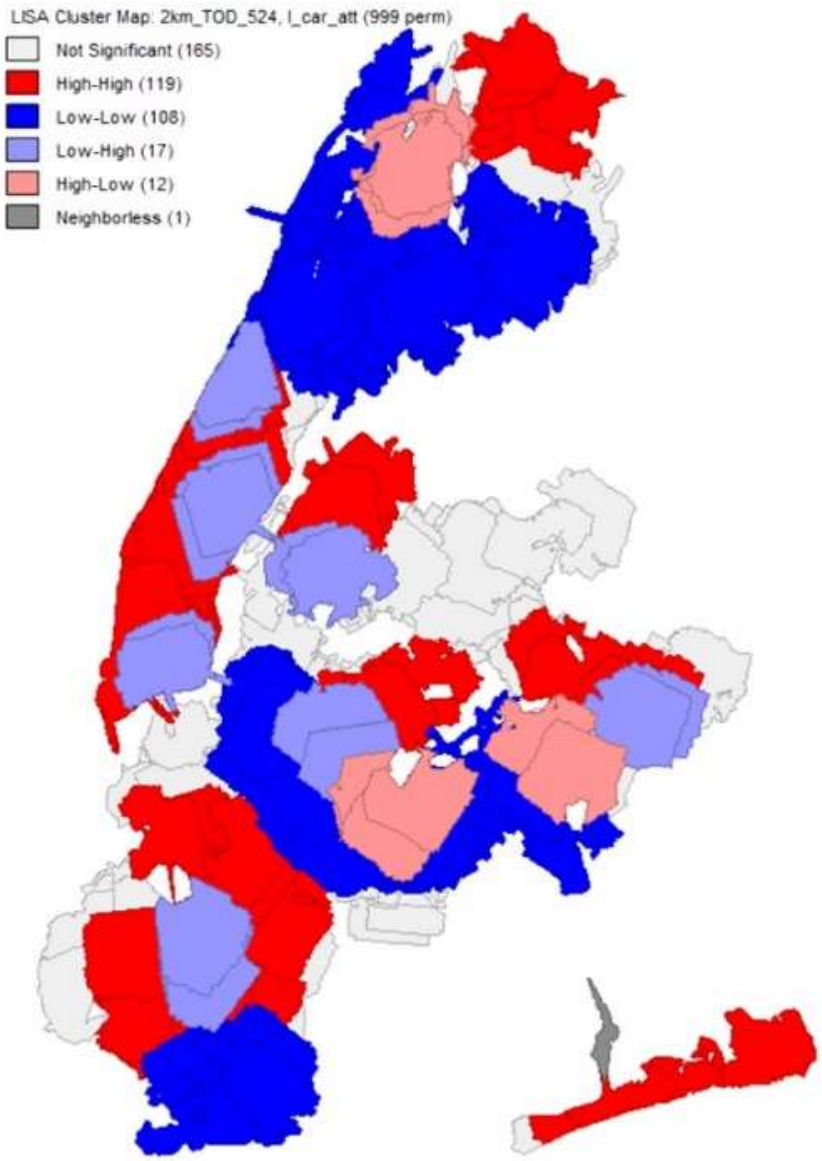

Figure 25: Spatial Relationship of Sentiment Score of Vehicle (2 km)

- People's Sentiment Toward Biking

People's sentiment score description about biking is listed in Table 17. The average sentiment score towards biking is higher than other travel modes with the exception of walking. The standard deviation, which is relatively low at 0.46, means people have concentrated attitude toward bicycles in 2 km radius

subway sheds.

Table 17: Bicycle Attitude Description

Category	**Score (2km)**
Minimum	-1.086
Maximum	3.380
Sum	145.168
Mean	0.344
Standard Deviation	0.468

Since biking can significantly improve people's accessibility toward subway stations from 0.5 mi to 2 km, we used 2 km radius catchment areas to analyze people's sentiments toward cycling. As we can see in Figure 26, the High-High clustering areas are concentrated in Manhattan and Queens, the Low-Low spaces are mostly in Brooklyn. Unlike sentiments about transit, there are some lower sentiment score islands in lower Manhattan, which is very wired to me since there are many accessible bicycle stations there.

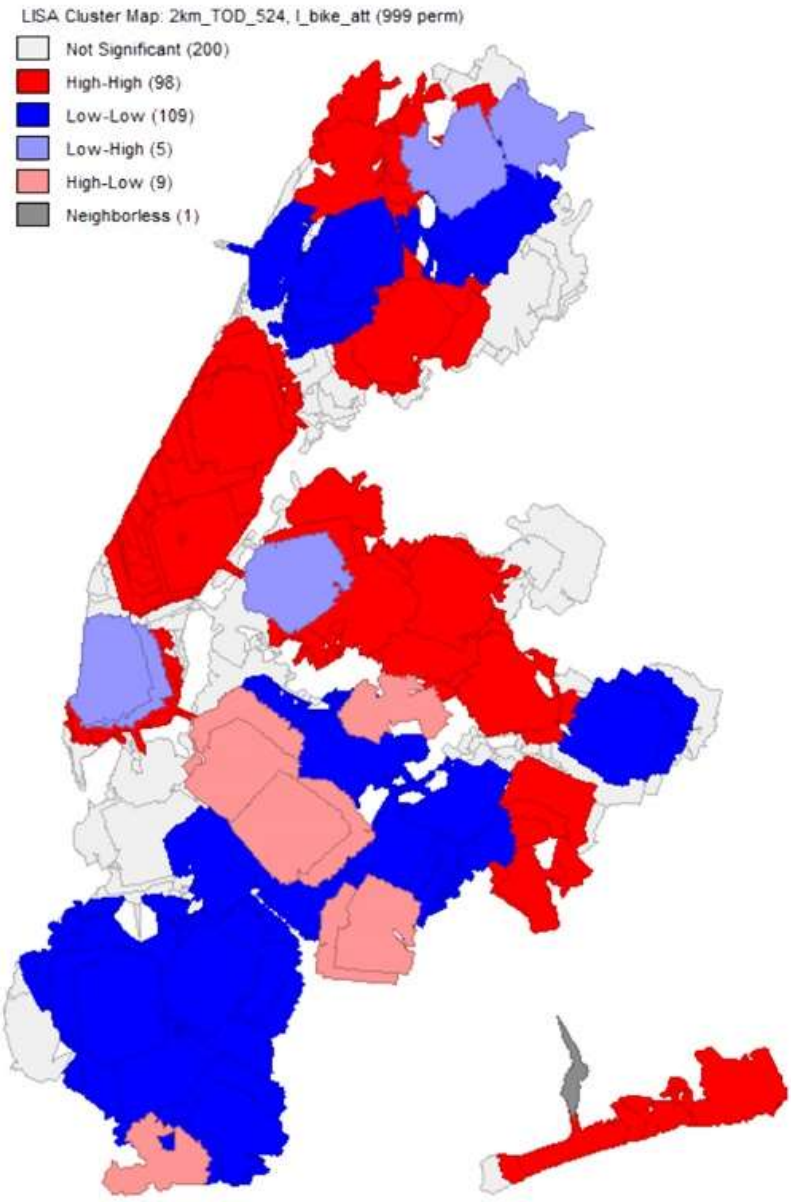

Figure 26: Spatial Relationship for People's Sentiment of Bicycle

2. The relationship between people's sentiment and the usage of each travel mode

As we know from the Literature Review chapter, there should be

connections between people's attitude and behavior toward a specific issue. So, we used the sentiment score of each travel mode to compare to the real usage in both 0.5 mi radius subway sheds and 2 km radius subway sheds. For 0.5-mi radius sheds, all 422 TOD catchment areas were located. Because all sentiment score data are normally distributed, they were not converted to log value. After applying the regression to each pair of variables, we found out that except for the subway, other travel modes had very low connections between the real usage and people's sentiment score. All of the R squared values for these three pair of data sets are lower than 0.1 which indicates they are not good predictors for people's travel behavior in 0.5 mi radius TOD catchment areas.(Table 18)

Table 18: People's Sentiment Score and Behavior (0.5 mile)

Sentiment Score	**Real Usage**	**Available TOD areas**	**Coefficient**	**P value**	**R^2**
Mean Score for Subway	Log Subway Usage	422	0.26	0.02	0.01
Mean Score for Vehicle	Log Vehicle Usage Rate	422	-0.008	0.82	0.00
Mean Score for Walking	Log Walking Rate	422	-0.002	0.34	0.002

For 2km subway sheds, there aren't many differences within the 0.5 mi sheds. As we can see in Table 19, except for the subway, people's sentiments are not statistically significant with the real usage of vehicle and bicycle. The P values are all higher than 0.025 and the correlation coefficient value are all

pretty low.

Table 19: People's Sentiment Score and Behavior (2 km)

Sentiment Score	**Real Usage**	**Available TOD areas**	**Coefficient**	**P value**	**R^2**
Mean Score for Subway	Log Subway Usage	422	0.69	0.00	0.03
Mean Score for Vehicle	Log Vehicle Usage Rate	422	-0.05	0.29	0.003
Mean Score for Bicycle	Log Bicycle Usage	165	-0.21	0.37	0.005

These results illustrate that, compared to the travel survey, people's sentiment score likely is not a better predictor of people's travel behavior in TOD catchment areas for both 0.5 mi and 2 km radius in NYC, compared to random chance alone. Also, they are not significantly associated with the real usage of each kind of travel modes in subway sheds.

Table 20: Overall Sentiment and Travel Behavior

Sentiment Score	**Real Usage**	**Available TOD areas**	**Coefficient**	**P value**	**R^2**
Overall Mean Score (0.5 mile)	Log Subway Usage (0.5 mile)	422	0.52	0.00	0.04
Overall Mean Score (0.5 mile)	Log Vehicle Usage Rate (0.5 mile)	422	-0.32	0.00	0.05
Overall Mean Score (0.5 mile)	Log Walking Rate (0.5 mile)	422	0.54	0.00	0.14
Overall Mean Score (2 km)	Log Subway Usage (2 km)	422	1.17	0.00	0.09
Overall Mean Score (2 km)	Log Vehicle Usage Rate (2 km)	422	-0.75	0.00	0.13
Overall Mean	Log Bicycle	165	2.58	0.00	0.38

Score (2 km)	Usage (2 km)				

Positive results were shown while analyzing the relationship between people's overall sentiment and travel mode usage. The outcome shows people's overall sentiments are statistically significant with all kinds of travel modes we choose. Although the R squared value are still relatively low, high sentiment score is still associated with high non-vehicle usage. Especially in 2km catchment areas, people's positive sentiment is associated with lower vehicle usage, and as one sentiment score increases, 0.75 unit of vehicle usage rate decreases. Also, higher sentiment scores are associated with higher subway usage in 2 km subway sheds than 0.5-mi subway sheds. The strongest association appears between bicycle usage and people's sentiment score. High sentiment scores are strongly associated with high bicycle usage in 2km radius transit service areas. In general, happy people are more likely to choose non-vehicle travel modes than un-happy people. Biking is the most sensitive travel mode to people's change in attitude and the second most sensitive travel mode is subway.

3. People's Sentiment about Travel Modes Inside and Outside Transit Catchment Areas

Even though people's positive attitude to travel modes isn't strongly associated with higher travel mode usage, can we say the TOD mode fails? Table 21 explains how people feel inside and outside of TOD catchment areas.

Most of the tweets related to the subway are concentrated within a 0.5 mi radius of subway stations, while only 6 percent of the tweets are located out of the catchment areas. After expanding the catchment radius to 2km, the number of tweets for the subway only increase by 165, which means people in transit catchment areas with a 0.5-mi radius are the ones who care more about the subway. Also, people in the catchment areas have more positive sentiment toward the subway, or at least tweet more often about it. The sentiment score for subway within 0.5 mi radius and 2km radius catchment areas are the same. When stepping out of the 2km circle, the sentiment score dramatically decreases. On the contrary, people have a relatively more positive attitude toward vehicles within subway sheds. The results also show that the closer to the subway station, the more positive people feel towards vehicles. In terms of cycling, people seem to have a slightly more positive feeling within a 2km radius transit shed rather than outside of it. The average sentiment score about biking is 0.46 in the service areas and it is 0.45 out of subway service areas. Lastly, the table shows that people have more positive feelings toward walking outside the 0.5 radius subway sheds (Table 21). The sentiment score decreases for 0.24 point after stepping into the catchment areas.

Table 21: Sentiment Comparison Inside and Outside TOD Catchment Areas

	Within 0.5 mile radius shed	**Outside 0.5 mile radius shed**	**Within 2 km radius shed**	**Outside 2 km radius shed**

Mean Sentiment Score for Subway	0.24	0.12	0.24	-0.03
Tweets Count for Subway	5442	372	5607	207
Mean Sentiment Score for Bicycle	-	-	0.46	0.45
Tweets Count for Bicycle	-	-	2297	808
Mean Sentiment Score for Vehicle	0.16	-0.48	-0.01	-0.4
Tweets Count for Vehicle	2992	1591	3919	664
Mean Sentiment Score for Walking	0.8	1.04	-	-
Tweets Count for Walking	2140	573	-	-
Over All Sentiment Score	1.95	1.66	1.90	1.68
Over all Tweets Count	407,292	363,763	462,167	308,888

Looking at the big picture here, the overall sentiment score in TOD catchment areas is higher than the scores outside them, both for a 0.5 mi radius and 2 km radius. This outcome illustrates that people living in subway sheds probably are happier than people living outside them, and the closer to the shed center, the happier people get.

Chapter 6: Built Environment and Vehicle Usage Rate

1. Model Building

As illustrated in the methodology chapter, 12 indicators and 11 indicators of built-environment for 0.5 mi radius sheds and 2 km radius sheds were located. They were used to predict the vehicle usage rate in TOD catchment areas. The data set were divided into two groups with a population ratio of 2:8. The larger part will be training data set and the other one will be test data set. The training data set will be used to train a model iteratively until it reaches the maximum accuracy of predicting the vehicle use. Then the model will be tested by the test data to ensure it is workable.

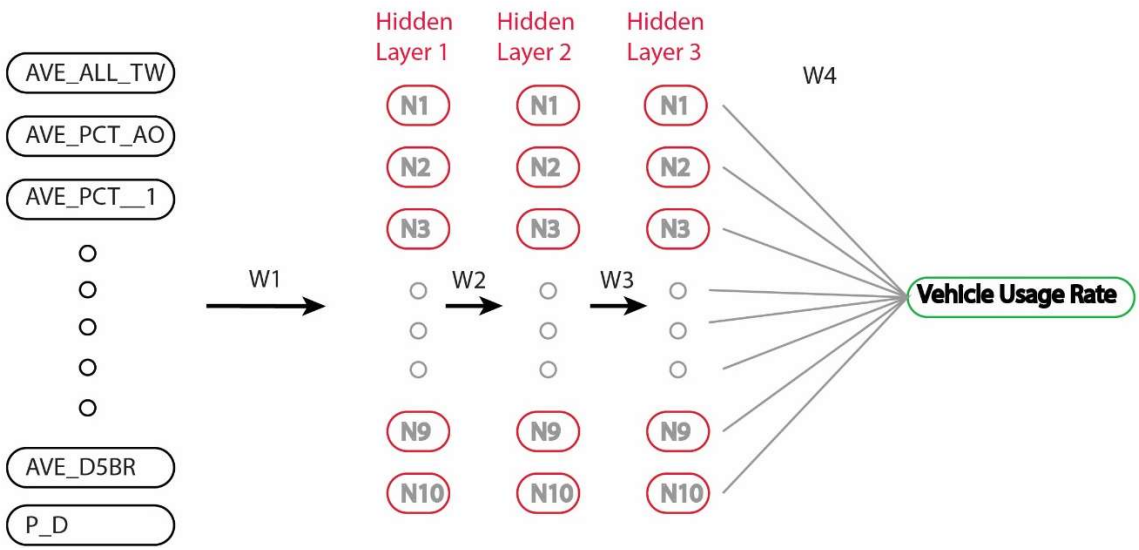

Figure 27: MLP Algorithm with Three Hidden Layers and 10 Nodes

Since the hidden layers are useful for unpacking behavior where they might not be represented well by a single perceptron, I gradually increased the

number of hidden layers and the neural nodes to achieve the best prediction. By using MLP, I found three hidden layers and ten nodes for each layer could impressively increase the accuracy of the learning model (Figure 27). Weights were calculated for each of the inputs and nodes, which means in the learning process weight for each of the variables was given four times. Inputs multiplied the weights and then passed down to the first hidden layer and multiplied? new weights to pass to the next hidden layer until reaching the outcome. This whole process was iterated for 2,000 times to arrive at final estimate.

2. Model Evaluation

The model for 0.5 mi radius subway sheds was trained successfully and it gave us positive results. By using this model, we can predict vehicle usage rate in TOD catchment areas in NYC well. The correlation coefficient value between vehicle usage rate predicted by the model and the vehicle usage rate in reality is 0.904, which is far higher than the R squared value calculated by traditional multivariable regression. (Figure 28)

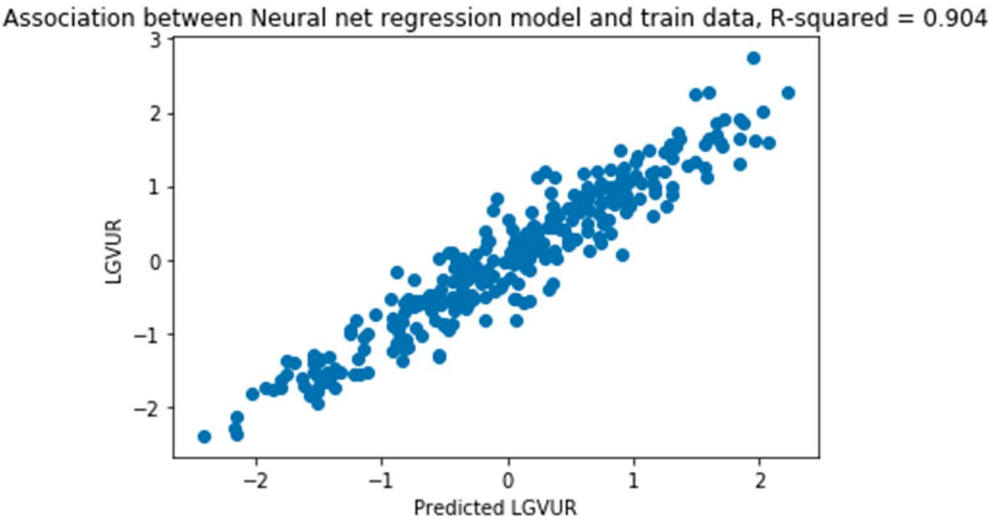

Figure 28: Association Between Neural Net Regression Model and Train Data (0.5 mile)

To test if the model can be used to predict other subway sheds with a 0.5-mi radius in NYC, test data was joined. The result shows although not as good as the original trained samples, the model still can be effectively used to measure other TOD catchment areas' vehicle usage rates in NYC. The R squared value reaches 0.87 when measuring the association between predicted vehicle usage rate and real vehicle usage rate in test data. (Figure 29)

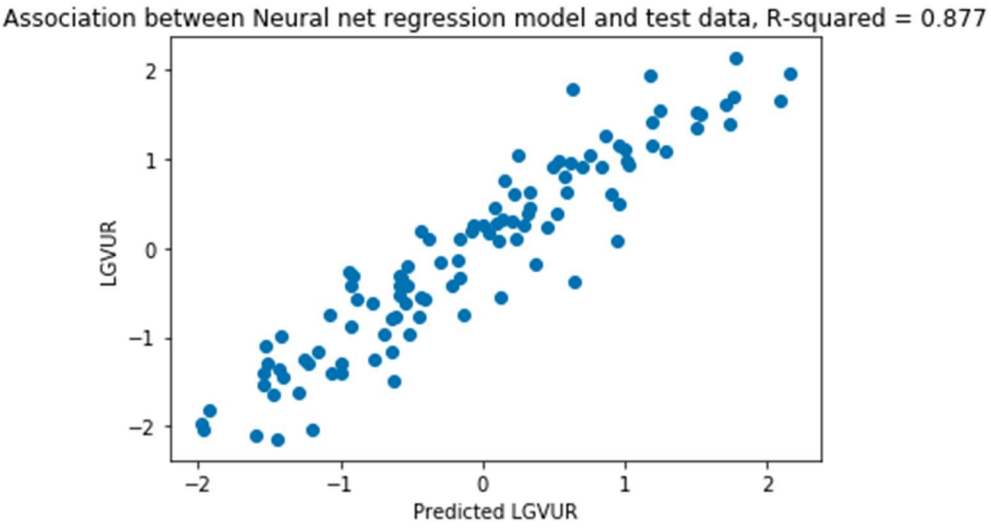

Figure 29: Association Between Neural Net Regression Model and Test Data (0.5 mile)

On the other hand, for 2 km radius subway sheds, the model we generated is even more accurate than the 0.5 mi's for predicting the vehicle usage rate. As we can see in Figure 30, the R squared value reaches 0.94 when testing the relationship between original vehicle usage rate and predicted vehicle usage rate with training data.

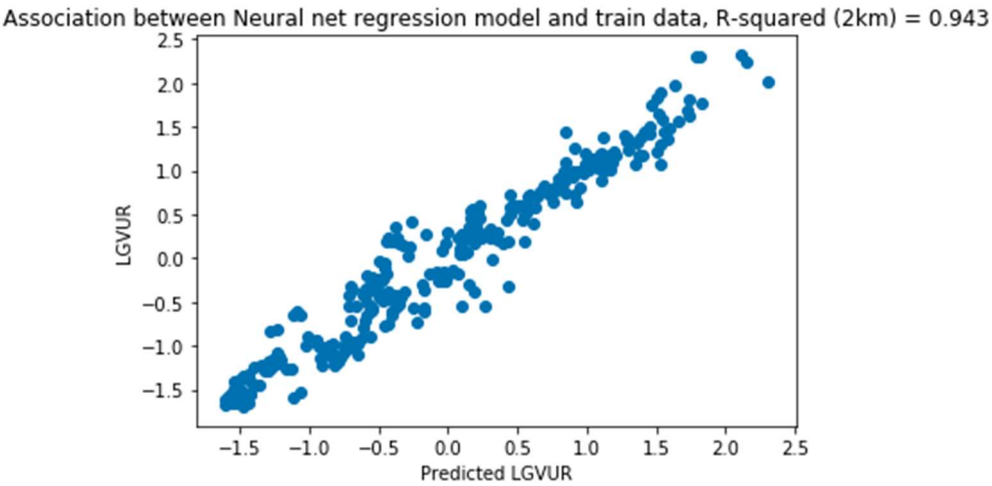

Figure 30: Association Between Neural Net Regression Model and Train Data (2km)

Furthermore, in the test part, the model also performed better than the

0.5-mi radius' model. The R squared value is 0.928 (Figure 31), which means the model is very accurate when predicting vehicle usage rate in 2km radius' transit catchment areas in NYC.

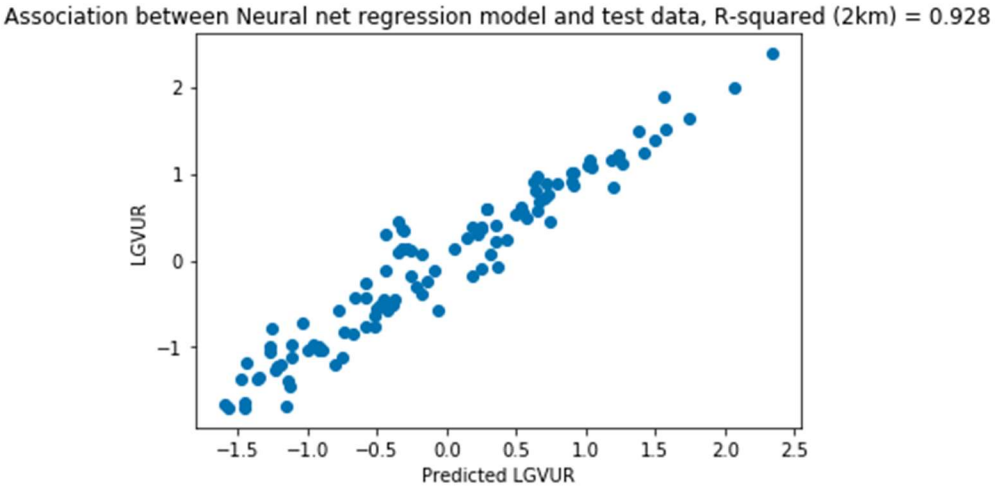

Figure 31: Association Between Neural Net Regression Model and Test Data (2km)

- The Role of Sentiment

Unfortunately, people's sentiment score seems to play a very small part in the model of 0.5 mi radius shed. As we can see in Figure 32, the relationship between the overall sentiment score and the predicted vehicle usage rate in training data is very weak. The plots are distributed randomly and the R value is low to -0.247. (Figure 32)

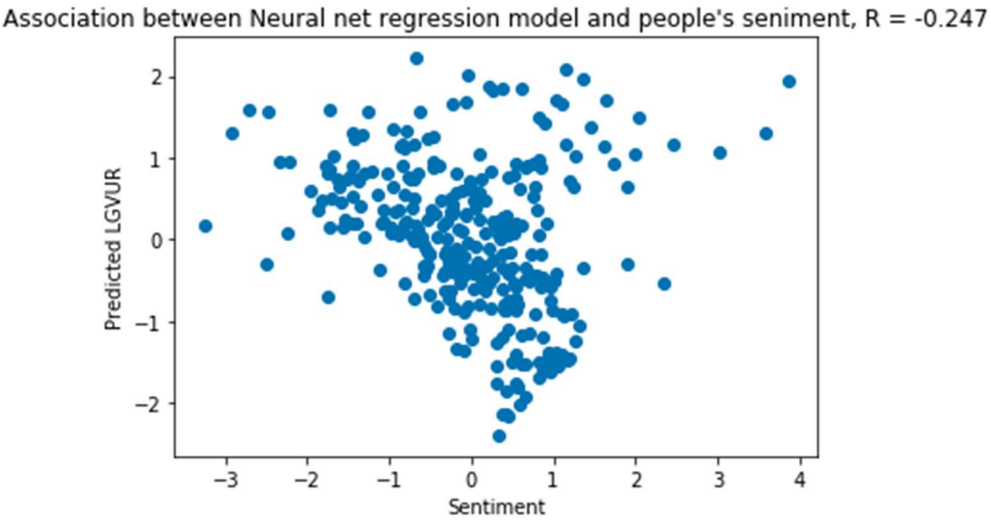

Figure 32: Association Between Neural Net Regression Model and People's Sentiment for Train Data (0.5 mile)

The test part for 0.5 mi radius data is similar to the train part. The association between people's sentiment and predicted vehicle usage rate is very weak too. (Figure 33)

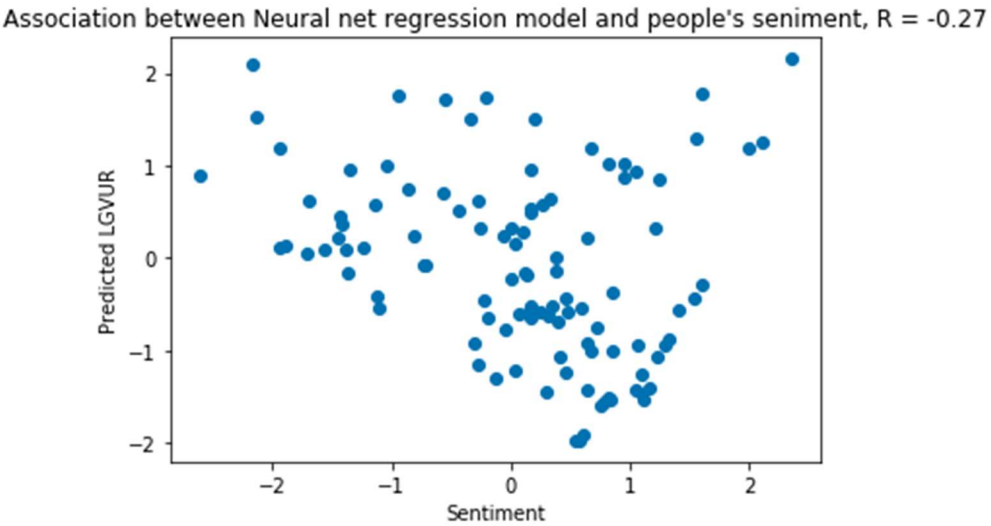

Figure 33: Association Between Neural Net Regression Model and People's Sentiment for Test Data (0.5 mile)

Surprisingly, we find that people's sentiment score becomes more

important when we expand the catchment radius to 2 km. Figure 34 illustrates that as people's sentiment grow higher, predicted vehicle usage rate in 2km radius subway will decrease and the correlation coefficient value reaches 0.32, higher than the 0.5-mi radius model (Figure 34). Furthermore, when we use the model to deal with other sheds, the sentiment factor shows its power. In the test part, people's sentiment score is highly associated with the predicted vehicle usage rate. The R value for that part is high to -0.53. (Figure 35)

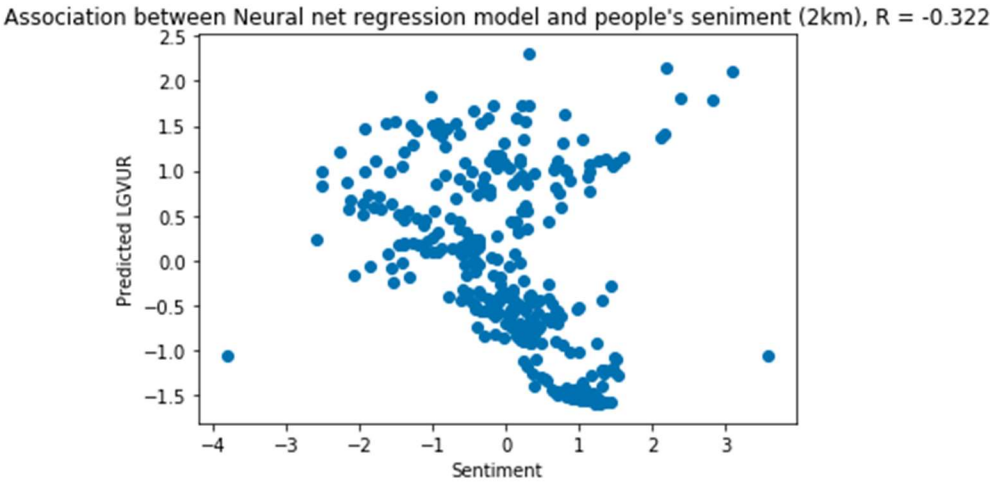

Figure 34: Association Between Neural Net Regression Model and People's Sentiment for Train Data(2km)

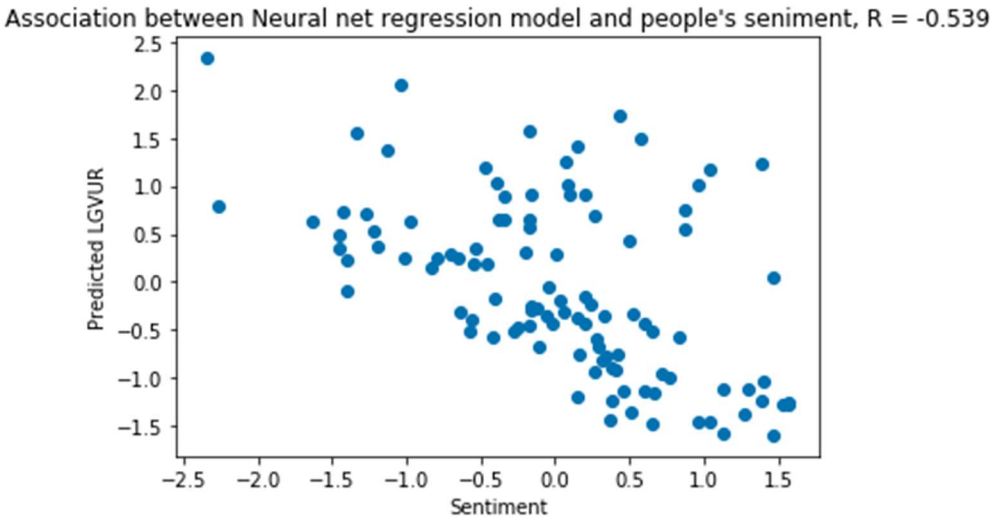

Figure 35: Association Between Neural Net Regression Model and People's Sentiment for Test Data(2km)

To sum up, these two model with 12 and 11 factors can effectively predict the vehicle usage rate in TOD catchment areas. Thus, we concluded that, for 0.5 mi radius transit sheds, people's overall sentiment score, percent of zero-car households in CBG, percent of one-car households in CBG, percent of two-car households in CBG, workers earning $3333/month or more, Service jobs within a 5-tier employment, gross residential density, gross service (5-tier) employment density (jobs/acre), deviation of CBG ratio of jobs/pop from regional average ratio of jobs/pop, jobs within 45-minute transit commute, Regional Centrality Index and parking sign density were the key factors impact vehicle usage rate. For 2km radius transit sheds, people's overall sentiment score, percent of one-car households in CBG, percent of two-car households in CBG, Service jobs within a 5-tier employment, gross residential density, Gross

service (5-tier) employment density (jobs/acre), Jobs within 45 minutes auto travel time, time decay, jobs within 45-minute transit commute, Regional Centrality Index (transit), Regional Centrality Index (auto) and parking sign density were the key factors impact vehicle usage rate.

Chapter 7 : Conclusion and Limitation

Before moving to more detailed conclusion, let's review the questions we raised at the beginning of this thesis:

Firstly, we wanted to know the current usage rate for each kind of travel mode in the NYC MSA's TOD catchment areas. Learning from the articles planners and scholars published, we decided the catchment radius for our TOD catchment areas should be set to 0.5 mi and 2 km. We also knew the travel modes we should focus on were subway, biking, walking and using vehicle from the literature review chapter. Then, as we discussed in the methodology chapter, all the data from Citi Bike, MTA and ACS were assigned into the transit sheds. Thus we knew what was the usage rate for each kind of travel mode we wanted to focus on.

Secondly, we wanted to figure out the relationship between people's sentiment and travel mode choice in transit zones. We joined the RHTS data into our research areas. The relationship between people's travel mode choice and three other variables, people's sentiment toward each kind of travel mode, people's overall sentiment and RHTS, were tested by using OLS regression. The ones with P value closer to 0 and R squared value closer to 1 should be seen as variables have stronger association with people's travel mode usage. Our outcome showed that in general, the RHTS is still the parameter which has lower P value and higher R squared value with people's travel mode usage.

People's overall sentiment score is statistically significant with the travel mode usage but the R squared values are relatively lower than the RHTS. On the other hand, people's sentiment toward each kind of travel mode has relatively higher P value and Lower R squared value comparing to the other two.

Finally, we asked our selves what are the other factors, except for people's sentiment, can impact people's travel behavior? Also, what role are people's sentiments playing among all factors which impact people's travel mode choice. To answer these questions, we built two models by using ANN for 0.5 mi radius and 2 km radius catchment areas. The input variables were 11 and 12 built environment factors which may impact the vehicle usage rate and people's overall sentiment. The accuracy of test part in these two models are 0.877 and 0.928. People's overall sentiment score plays an important role in the 2 km model as the correlation coefficient value it has with the predicted vehicle usage rate is -0.5.

Based on the exploration of our questions, we can evaluate our initial hypothesis:

(1) Depending on my study for the TOD catchment areas in NYC, the tweets are not a suitable replacement for travel surveys.

(2) TOD indeed boosts people's sentiment score toward each kind of travel mode in catchment areas.

(3) The models we built have an accuracy higher than 0.9, and people's

sentiment plays an important part in the 2 km model.

1. Conclusion

The more detailed conclusions are:

In reality, central Manhattan and lower Manhattan's TOD sheds have relatively lower vehicle usage rate and higher non-vehicle usage, which represents higher subway use, higher bicycle use and a higher walking rate.

The travel surveys were indeed associated with the real usage of the travel modes. They are more accurate in 2 km catchment radius areas than in 0.5 mi catchment radius areas. Whereas, since the correlation coefficient value are all lower than 0.5, the travel survey may not a good tool to predict people's travel behavior in transit sheds.

People's mean sentiment scores toward each kind of travel mode are all positive and biking gets the highest mean sentiment score among all. Manhattan area's TOD sheds don't get higher sentiment scores for non-vehicle travel modes than vehicle. On the other hand, for 2 km radius TOD catchment areas, high sentiment score sheds for subway and bicycle, as well as most of the higher score sheds for vehicle are gathered in central Manhattan.

Based on this analysis, compared to the travel survey, people's sentiment score is not a better predictor of people's travel behavior in TOD catchment areas for both 0.5 mi and 2 km radius, as their association with the real travel

mode usage is very weak.

People's overall sentiments are statistically significant with each kind of travel mode we chose. Although the R squared values are all relatively low, high sentiment score is still associated with high non-vehicle usage. In general, happy people are probably more likely to choose non-vehicle travel modes than unhappy people. Biking is the most sensitive travel mode to people's sentiment change and the second most sensitive travel mode is subway.

The subway usage indeed associate with people's sentiment toward subway in 2 km radius catchment areas but other travel modes are not. What's more, within 0.5 mi sheds, people even hold relatively more positive attitude toward vehicle than people outside the sheds. However, people who located in the sheds seem happier than the people outside them as the overall sentiment score in the sheds is higher.

By using machine learning, we successfully generate two models to predict the vehicle usage rate for both 0.5 mi radius subway sheds and 2 km radius subway sheds. The R squared value for 0.5 mi radius model is around 0.9 and the R squared value for 2km radius model is even higher. The results show with machine learning, we can predict the vehicle usage rate in TOD catchment areas more effectively and precisely than traditional multiple-regression.

People's sentiment doesn't play an important role in the 0.5 mi radius model but it performs better in the 2km radius model. This means in 2 km

radius' subways sheds, people's sentiment is more likely to impact the vehicle usage rate.

2. Limitation

- Data limitation

The tweets we get have limitations for both time and population. Although we collected over two million tweets, few were left after we filtered them by using sentiment words and travel mode words. It can hardly be called big-data since only a couple of thousands tweets are located for each kind of travel mode. People's attitudes toward different kinds of travel mode may be sensitive to the time of a day. Thus, such limited data can't perfectly capture the whole day's sentiment change. Plus, the tweets we got are from April to July in 2016, which didn't merge the data collecting time of American Community Survey and RHTS. We used the ACS data to represent the real usage rate for vehicle and walking, but we all know that's not the reality. ACS has its own limitation too since it is based on self-reporting, which can lead to bias. The data population for RHTS is also very limited. What's more, four kinds of travel mode were analyzed, but if we want to understand more about non-vehicle use in research areas, more travel modes are needed.

Another limitation in this thesis is we probably underestimate the error caused by missing data. The smaller our catchment radius is, the more

influence our outcome will get from the data shortage. For instance, since the travel survey data and the ACS data are saved in census tract and block group level, one statistical unit data missing will significantly lower the average travel mode usage in 0.5 mi radius transit sheds no matter if we rasterize them or not. So, the reason why in 0.5-mi radius TOD catchment areas the correlation coefficient for the travel survey and the reality is lower than 2km is perahps due to a lack of data. Also, the model we created by ANN can also be impact by that problem so that the accuracy of 0.5 mi model is lower than the 2km one.

- Method Limitation

Based on the 7D concept, we chose 16 indicators from 90 in Smart Location Database and another one from NYC DOT. However, we are not sure if there are other variables that also strongly associate with vehicle usage rate in subway sheds. Other than that, we used VIF value to determine which two variables are highly correlated and we dropped the one with lower R squared value with y. The minimum requirement for VIF value was set up by ourselves, in my case 7.5, which was a relatively loose requirement. If we shrink it, more variables will be dropped, and it is hard to say which value is appropriate for my research. Other than that, the model we built is basically grounded on NYC. Although we tested the model by separating a few data from the total, those

data are still coming from NYC. NYC is a relatively unique transportation complex compared to other cities in the US, or in the world as it is one of the most populated, richest and busiest cities on earth. We probably ignored some biases when building our model. If this model can work effectively in other cities is a question that still needs to be answered.

3. Recommendations and Implications

As mentioned above, this thesis has a few limitations. However, I believe with more research and exploration, we can improve our outcome and reduce the biases we drew. Firstly, we can change our way of collecting tweets using python so that only tweets within a certain distance from subway stations can be located. This method can help us boost the data mining efficiency, enlarge the data population and collect the data more precisely. What's more, I suggest future study use artificial intelligence, which includes facial recognition technology, to collect data with people's travel mode choice, like walking, cycling, driving, etc. since CCTV is everywhere in our lives. Lowering the VIF value to enhance the minimum requirement to eliminate covariant variables is another suggestion I would make. In the meantime, more built-environment factors need to be explored to see whether there are other variables relevant to the vehicle usage rate. At last, I suggest scholars in the future conduct similar investigations in other metropolitan areas around the world to evaluate the

accuracy of my model. Other cities' TOD areas' data can be add in to improve that model if necessary.

For all these times, planners and scholars focused more on the built-environment when researching people's travel behavior. This thesis, gives our specialists, like planners in EPA, TOD experts, a hint that there are other factors that can also impact people's travel mode choice, namely, people's sentiment score. Although we can hardly predict people's travel behavior only by using their sentiment score, we still find people's attitude toward non-vehicle travel mode in TOD catchment areas is higher than people's attitude toward vehicles, which can lead us to future studies related to other research realms, like public health, environmental science, etc.

Another practical significance of my thesis is that we built two models which can guide planers and policy makers to make decisions. Although before development it's hard to answer how many cars a household may have or what people's sentiment score is in the future, we can still evaluate the residential density, job density, RCI and parking restrictions by creating proper zoning laws and policies. Thus, planners and policy makers can to some level restrict TOD catchment areas' vehicle usage rates in the future. What's more, we all know that machine learning can not only be used in predicting vehicle usage rate in transit sheds, it can also be used to predict other variables, like people's sentiment or vehicle occupation rate. Therefore, it is possible for future

planners to get all variables before the development so that will make my thesis more meaningful.

Bibliography

Agrawal, Deepak, and Christopher Schorling. 1996. "Market Share Forecasting: An Empirical Comparison of Artificial Neural Networks and Multinomial Logit Model." *Journal of Retailing* 72 (4): 383–407. https://doi.org/10.1016/S0022-4359(96)90020-2.

Ajzen, Icek. 1991. "The Theory of Planned Behavior." *Organizational Behavior and Human Decision Processes*, Theories of Cognitive Self-Regulation, 50 (2): 179–211. https://doi.org/10.1016/0749-5978(91)90020-T.

"Anselin Local Moran's I." n.d. GISC. Accessed May 4, 2018. /tutorials/gis-techniques/spatial-statistics/anselin-local-morans-i/.

Anselin, Luc. 1995. "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27 (2): 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

B.~Taylor, Dean, and Hani Mahmassani. 1997. "Analysis of Stated Preferences for Intermodal Bicycle-Transit Interfaces." *Transportation Research Record Journal of the Transportation Research Board* 1556 (January): 86–95. https://doi.org/10.3141/1556-11.

Besser, Lilah M., and Andrew L. Dannenberg. 2005. "Walking to Public Transit: Steps to Help Meet Physical Activity Recommendations." *American Journal of Preventive Medicine* 29 (4): 273–80. https://doi.org/10.1016/j.amepre.2005.06.010.

Bureau, US Census. n.d. "State and Metropolitan Area Data Book: 2010." Accessed February 28, 2018. https://www.census.gov/library/publications/2010/compendia/databooks/smadb10.html.

Cervero, Robert, and Kara Kockelman. 1997. "Travel Demand and the 3Ds: Density, Diversity, and Design." *Transportation Research Part D: Transport and Environment* 2 (3): 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6.

Ettema, Dick, and Roy Nieuwenhuis. 2017. "Residential Self-Selection and Travel Behaviour: What Are the Effects of Attitudes, Reasons for Location Choice and the Built Environment?" *Journal of Transport Geography* 59 (Supplement C): 146–55. https://doi.org/10.1016/j.jtrangeo.2017.01.009.

Ewing, Reid, and Robert Cervero. 2010. "Travel and the Built Environment." *Journal of the American Planning Association* 76 (3): 265–94. https://doi.org/10.1080/01944361003766766.

Ewing, Reid, Marybeth Deanna, and Shi-Chiang Li. 1996. "Land Use Impacts on Trip Generation Rates." *Transportation Research Record: Journal of the Transportation Research Board* 1518 (January): 1–6.

https://doi.org/10.3141/1518-01.

Griffiths, Brittany, and Carey Curtis. 2017. “Effectiveness of Transit Oriented Development in Reducing Car Use: Case Study of Subiaco, Western Australia.” *Urban Policy and Research* 35 (4): 391–408. https://doi.org/10.1080/08111146.2017.1311855.

Heaton, Jeff. 2008. *Introduction to Neural Networks for Java, 2Nd Edition*. 2nd ed. Heaton Research, Inc.

Hess, Daniel. 2001. “Effect of Free Parking on Commuter Mode Choice: Evidence from Travel Diary Data.” *Transportation Research Record: Journal of the Transportation Research Board* 1753 (January): 35–42. https://doi.org/10.3141/1753-05.

Jiang, S., J. Ferreira, and M. C. Gonzalez. 2017. “Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore.” *IEEE Transactions on Big Data* 3 (2): 208–19. https://doi.org/10.1109/TBDATA.2016.2631141.

Jiang Shan, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. 2016. “The TimeGeo Modeling Framework for Urban Mobility without Travel Surveys.” *Proceedings of the National Academy of Sciences* 113 (37): E5370–78. https://doi.org/10.1073/pnas.1524261113.

Justin B. Hollander. 2016. *Urban Social Listening: Potential and Pitfalls for Using Microblogging Data in Studying Cities*. Palgrave Pivot. New York : Secaucus: Palgrave Macmillan, Springer distributor. http://link.springer.com/10.1057/978-1-137-59491-4.

Kamruzzaman, Md, Douglas Baker, Simon Washington, and Gavin Turrell. 2016. “Determinants of Residential Dissonance: Implications for Transit-Oriented Development in Brisbane.” *International Journal of Sustainable Transportation* 10 (10): 960–74. https://doi.org/10.1080/15568318.2016.1191094.

Kim, Taehyun, Dong-Wook Sohn, and Sangho Choo. 2017. “An Analysis of the Relationship between Pedestrian Traffic Volumes and Built Environment around Metro Stations in Seoul.” *KSCE Journal of Civil Engineering* 21 (4): 1443–52. https://doi.org/10.1007/s12205-016-0915-5.

Langlois, Myriam, Dea van Lierop, Rania A. Wasfi, and Ahmed M. El-Geneidy. 2015. “Chasing Sustainability.” *Transportation Research Record: Journal of the Transportation Research Board* 2531 (January): 83–92. https://doi.org/10.3141/2531-10.

Langlois, Myriam, Rania A. Wasfi, Nancy A. Ross, and Ahmed M. El-Geneidy. 2016. “Can Transit-Oriented Developments Help Achieve the Recommended Weekly Level of Physical Activity?” *Journal of*

Transport & Health, Special Issue: Public Transport and Health, 3 (2): 181–90. https://doi.org/10.1016/j.jth.2016.02.006.

Lee, Jaeyeong, Keechoo Choi, and Yountaik Leem. 2016. "Bicycle-Based Transit-Oriented Development as an Alternative to Overcome the Criticisms of the Conventional Transit-Oriented Development." *International Journal of Sustainable Transportation* 10 (10): 975–84. https://doi.org/10.1080/15568318.2014.923547.

Liu, Yuwei, Hong Sheng, Norbert Mundorf, Colleen Redding, and Yinjiao Ye. 2017. "Integrating Norm Activation Model and Theory of Planned Behavior to Understand Sustainable Transport Behavior: Evidence from China." *International Journal of Environmental Research and Public Health* 14 (12). https://doi.org/10.3390/ijerph14121593.

Lu, Xuedong, and Eric I. Pas. 1999. "Socio-Demographics, Activity Participation and Travel Behavior." *Transportation Research Part A: Policy and Practice* 33 (1): 1–18. https://doi.org/10.1016/S0965-8564(98)00020-2.

Lv, Y., Y. Chen, X. Zhang, Y. Duan, and N. L. Li. 2017. "Social Media Based Transportation Research: The State of the Work and the Networking." *IEEE/CAA Journal of Automatica Sinica* 4 (1): 19–26. https://doi.org/10.1109/JAS.2017.7510316.

Mamdoohi, Amir Reza, and Amir Janjany. 2016. "Modeling Metro Users' Travel Behavior in Tehran: Frequency of Use." *TeMA Journal of Land Use, Mobility and Environment*, no. 10.6092/1970-9870/3933 (October): 47–58.

Mamun, Sha A., Nicholas E. Lownes, Jeffrey P. Osleeb, and Kelly Bertolaccini. 2013. "A Method to Define Public Transit Opportunity Space." *Journal of Transport Geography* 28 (April): 144–54. https://doi.org/10.1016/j.jtrangeo.2012.12.007.

Nasri, Arefeh, and Lei Zhang. 2014a. "The Analysis of Transit-Oriented Development (TOD) in Washington, D.C. and Baltimore Metropolitan Areas." *Transport Policy* 32 (March): 172–79. https://doi.org/10.1016/j.tranpol.2013.12.009.

———. 2014b. "The Analysis of Transit-Oriented Development (TOD) in Washington, D.C. and Baltimore Metropolitan Areas." *Transport Policy* 32 (Supplement C): 172–79. https://doi.org/10.1016/j.tranpol.2013.12.009.

"NYC's Public Transit Somehow Ranked Best in the Nation | New York Post." n.d. Accessed February 27, 2018. https://nypost.com/2017/10/29/nycs-public-transit-somehow-ranked-best-in-the-nation/.

Olaru, Doina, and Carey Curtis. 2015. "Designing TOD Precincts: Accessibility and Travel Patterns." *European Journal of Transport and Infrastructure*

Research 15 (January): 6–26.

Omrani, Hichem. 2015. "Predicting Travel Mode of Individuals by Machine Learning." *Transportation Research Procedia*, 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands, 10 (January): 840–49. https://doi.org/10.1016/j.trpro.2015.09.037.

O'Sullivan, Sean, and John Morrall. 1996. "Walking Distances to and from Light-Rail Transit Stations." *Transportation Research Record: Journal of the Transportation Research Board* 1538 (January): 19–26. https://doi.org/10.3141/1538-03.

Öztürk, Nazan, and Serkan Ayvaz. 2018. "Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis." *Telematics and Informatics* 35 (1): 136–47. https://doi.org/10.1016/j.tele.2017.10.006.

Park, Sungjin, Keechoo Choi, and Jae Seung Lee. 2017. "Operationalization of Path Walkability for Sustainable Transportation." *International Journal of Sustainable Transportation* 11 (7): 471–85. https://doi.org/10.1080/15568318.2016.1226996.

Pavlicek, Antonin, and Richard Novak. 2015. *"Big Data" from the Perspective of Data Sources*. Edited by R. Nemec and F. Zapletal. Ostrava: Vsb-Tech Univ Ostrava.

Rout, Jitendra Kumar, Kim-Kwang Raymond Choo, Amiya Kumar Dash, Sambit Bakshi, Sanjay Kumar Jena, and Karen L. Williams. 2018. "A Model for Sentiment and Emotion Analysis of Unstructured Social Media Text." *Electronic Commerce Research* 18 (1): 181–99. https://doi.org/10.1007/s10660-017-9257-8.

Santora, Marc. 2017. "Subway Ridership Falls as M.T.A. Scrambles to Improve Service." *The New York Times*, November 15, 2017, sec. N.Y. / Region. https://www.nytimes.com/2017/11/15/nyregion/subway-ridership-falls-as-mta-scrambles-to-improve-service.html.

Semanjski, Ivana, Sidharta Gautama, Rein Ahas, and Frank Witlox. 2017. "Spatial Context Mining Approach for Transport Mode Recognition from Mobile Sensed Big Data." *Computers, Environment and Urban Systems* 66 (November): 38–52. https://doi.org/10.1016/j.compenvurbsys.2017.07.004.

Stephanie Pollack, Barry Bluestone, and Chase Billingham. n.d. "Maintaining Diversity In America's Transit-Rich Neighborhoods:" Accessed December 13, 2016. https://www.yumpu.com/en/document/view/35157884/maintaining-diversity-in-americas-transit-rich-neighborhoods.

US EPA, OA. 2014. "Smart Location Mapping." Data and Tools. US EPA. February 27, 2014. https://www.epa.gov/smartgrowth/smart-location-

mapping.

Wang, Shiliang, Michael J. Paul, and Mark Dredze. 2015. “Social Media as a Sensor of Air Quality and Public Response in China.” *Journal of Medical Internet Research* 17 (3): e22. https://doi.org/10.2196/jmir.3875.

Xi, Yang (Luna), Shoshanna Saxe, and Eric Miller. 2016. “Accessing the Subway in Toronto, Canada.” *Transportation Research Record: Journal of the Transportation Research Board* 2543 (January): 52–61. https://doi.org/10.3141/2543-06.

Xu, Tao, Ming Zhang, and Paulus T. Aditjandra. 2016. “The Impact of Urban Rail Transit on Commercial Property Value: New Evidence from Wuhan, China.” *Transportation Research Part A: Policy and Practice* 91 (September): 223–35. https://doi.org/10.1016/j.tra.2016.06.026.

Zhang, D., B. Guo, and Z. Yu. 2011. “The Emergence of Social and Community Intelligence.” *Computer* 44 (7): 21–28. https://doi.org/10.1109/MC.2011.65.

Zhang, Daqing, Matthai Philipose, and Qiang Yang. 2011. “Introduction to the Special Issue on Intelligent Systems for Activity Recognition.” *ACM Trans. Intell. Syst. Technol.* 2 (1): 1:1–1:4. https://doi.org/10.1145/1889681.1889682.

Zhang, Xi, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang, and Philip S. Yu. 2018. “Improving Stock Market Prediction via Heterogeneous Information Fusion.” *Knowledge-Based Systems* 143 (March): 236–47. https://doi.org/10.1016/j.knosys.2017.12.025.

Zhou, Chaoran, Hongfei Jia, Jingxin Gao, Lili Yang, Yixiong Feng, and Guangdong Tian. 2017. “Travel Mode Detection Method Based on Big Smartphone Global Positioning System Tracking Data.” *Advances in Mechanical Engineering* 9 (6): 1687814017708134. https://doi.org/10.1177/1687814017708134.

Reconnecting America. (2009). Case Studies for Transit Oriented Development, Available on line at: http://www.reconnectingamerica.org/assets/Uploads/tools.pdf

Appendix

R code for scoring people's sentiment:

For bicycle:

```
df <- read.csv("Twitter_417_724.csv", stringsAsFactors = FALSE)
bike <- NULL
bike <-grepl("bike|bicycle|bicycling|cycling",df$Tweets)
df <- mutate(df,bike)
df_new <- subset(df,bike == TRUE)
df_new <- as.tibble(df_new)

tidy_Tweets <- df_new %>%
   unnest_tokens(word,Tweets)
afinn <- get_sentiments("afinn")
tidy_Tweets_afinn <-tidy_Tweets %>%
   inner_join(afinn)
df_sentiment_only<- tidy_Tweets_afinn %>%
   group_by(ID1) %>%
   summarise(score=sum(score))
ungroup()
total <- merge(df_new,df_sentiment_only,by="ID1",all = T)
total[is.na(total)]<-0
```

For vehicle:

```
df_car <- read.csv("Twitter_417_724.csv", stringsAsFactors = FALSE)
auto <- NULL
auto <- grepl("\\bSUV\\b|\\bauto\\b|\\bmotor\\b|\\bvan\\b|\\bcar\\b|\\bcars\\b|automobile|\\bdrive\\b|driving|vehicle|truck",df_car$Tweets)
df_car <- mutate(df_car,auto)
df_car <- subset(df_car, auto == TRUE)
df_car <- as.tibble(df_car)
df_car[8] = NULL
tidy_Tweets1 <- df_car %>%
   unnest_tokens(word,Tweets)
afinn <- get_sentiments("afinn")
tidy_Tweets_afinn1 <-tidy_Tweets1 %>%
   inner_join(afinn)
df_sentiment_only2<- tidy_Tweets_afinn1 %>%
   group_by(ID1) %>%
```

```
    summarise(score=sum(score))
  ungroup()
  total_car <- merge(df_car,df_sentiment_only2,by="ID1",all = T)
  total_car[is.na(total_car)]<-0

  For Subway:
  df <- read.csv("Twitter_417_724.csv", stringsAsFactors = FALSE)
  Subway <- NULL
  Subway <- grepl("subway|MTA|rapid transit",df$Tweets)
  df<-mutate(df,Subway)
  df_new<-subset(df,Subway == TRUE)
  df_new <- as.tibble(df_new)
  tidy_Tweets <- df_new %>%
    unnest_tokens(word,Tweets)
  afinn <- get_sentiments("afinn")
  tidy_Tweets_afinn <-tidy_Tweets %>%
    inner_join(afinn)
  df_sentiment_only<- tidy_Tweets_afinn %>%
    group_by(ID1) %>%
    summarise(score=sum(score))
  ungroup()
  total <- merge(df_new,df_sentiment_only,by="ID1",all = T)
  total[is.na(total)]<-0

  For Walking:
  df_p <- read.csv("Twitter_417_724.csv", stringsAsFactors = FALSE)
  Pedestrian <- NULL
  Pedestrian <- grepl("\\bwalk\\b|\\bwalked\\b|\\bon
foot\\b|pedestrian",df_p$Tweets)
  df_p <- mutate(df_p,Pedestrian)
  df_pn <- subset(df_p,Pedestrian == TRUE)
  df_pn <- as.tibble(df_pn)
  tidy_Tweets2 <- df_pn %>%
    unnest_tokens(word,Tweets)
  afinn <- get_sentiments("afinn")
  tidy_Tweets_afinn2 <-tidy_Tweets2 %>%
    inner_join(afinn)
  df_sentiment_only2<- tidy_Tweets_afinn2 %>%
    group_by(ID1) %>%
    summarise(score=sum(score))
```

Overall Sentiment:

```
df <- read.csv("Twitter_417_724.csv", stringsAsFactors = FALSE)
df_t <- as.tibble(df)
tidy_Tweets <- df_t %>%
   unnest_tokens(word,Tweets)
afinn <- get_sentiments("afinn")
tidy_Tweets_afinn <-tidy_Tweets %>%
   inner_join(afinn)
df_sentiment_only<- tidy_Tweets_afinn %>%
   group_by(ID1) %>%
   summarise(score=sum(score))
ungroup()
df_2 <- df %>%
   inner_join(df_sentiment_only)
```

Python Data for Building the Model:

```
x05 = TOD05mile.values[:,1:13]
x05[:5]

y05 = TOD05mile.values[:,0:1]
y05[:5]

x2k = TOD2km.values[:,1:12]
y2k = TOD2km.values[:,0:1]
from sklearn.model_selection import train_test_split

X= x05
y = y05

X2 = x2k
y2 = y2k

X.shape
y.shape

X_train, X_test, y_train, y_test = train_test_split(X,y)
X2_train, X2_test, y2_train, y2_test = train_test_split(X2,y2)

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler(copy=True, with_mean=True, with_std=True)
```

```
scaler.fit(X_train)
scaler.fit(X2_train)

#Take that fit and transform the training AND test data (we have to do both)
#0.5mile
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)

scaler.fit(y_train)
y_train = scaler.transform(y_train)
y_test = scaler.transform(y_test)

#2km
X2_train = scaler.transform(X2_train)
X2_test = scaler.transform(X2_test)

scaler.fit(y2_train)
y2_train = scaler.transform(y2_train)
y2_test = scaler.transform(y2_test)



#Always check
plt.hist(y_train)

plt.hist(X_train)

from sklearn.neural_network import MLPRegressor
mlp = MLPRegressor(hidden_layer_sizes=(10,10,10),shuffle=True,
tol=0.0001, validation_fraction=0.1)

mlp.fit(X_train,y_train)

mlp =
MLPRegressor(max_iter=2000,hidden_layer_sizes=(10,10,10),shuffle=True,
tol=0.0001, validation_fraction=0.1)
mlp.fit(X_train,y_train)

#2km
mlp =
MLPRegressor(max_iter=2000,hidden_layer_sizes=(10,10,10),shuffle=True,
```

```
tol=0.0001, validation_fraction=0.1)
    mlp.fit(X2_train,y2_train)

    #0.5mile
    predictions_2 = mlp.predict(X_train)
    print(predictions_2)
    predictions = mlp.predict(X_test)
    print(predictions)

    #2km
    predictions_3 = mlp.predict(X2_train)
    print(predictions_3)
    predictions_4 = mlp.predict(X2_test)
    print(predictions_4)

    # 0.5 mile
    mlp.score(X_test,y_test)
    mlp.score(X_train,y_train)

    # 2km
    mlp.score(X2_test,y2_test)
    mlp.score(X2_train,y2_train)

    # 0.5 mile
    scores = round(mlp.score(X_train, y_train),3)
    plot = plt.scatter(predictions_2, y_train)
    plt.title(str("Association between Neural net regression model and train
data, R-squared = " + str(scores)))
    plt.xlabel('Predicted LGVUR', fontsize=10)
    plt.ylabel('LGVUR', fontsize=10)
    plot

    scores2 = round(mlp.score(X_test, y_test),3)
    plot2 = plt.scatter(predictions, y_test)
    plt.title(str("Association between Neural net regression model and test
data, R-squared = " + str(scores2)))
    plt.xlabel('Predicted LGVUR', fontsize=10)
    plt.ylabel('LGVUR', fontsize=10)
    plot2

    Rsq_4 = np.corrcoef(X_train[:,0], predictions_2,rowvar=0)
```

```
plot3 = plt.scatter(X_train[:,0], predictions_2)
plt.title(str("Association between Neural net regression model and people's seniment, R = -0.247" ))
plt.xlabel('Sentiment', fontsize=10)
plt.ylabel('Predicted LGVUR', fontsize=10)
plot3

Rsq_5 = np.corrcoef(X_test[:,0], predictions,rowvar=0)
plot4 = plt.scatter(X_test[:,0], predictions)
plt.title(str("Association between Neural net regression model and people's seniment, R = -0.27" ))
plt.xlabel('Sentiment', fontsize=10)
plt.ylabel('Predicted LGVUR', fontsize=10)
plot4

# 2km
scores3 = round(mlp.score(X2_train, y2_train),3)
plot5 = plt.scatter(predictions_3, y2_train)
plt.title(str("Association between Neural net regression model and train data, R-squared (2km) = " + str(scores3)))
plt.xlabel('Predicted LGVUR', fontsize=10)
plt.ylabel('LGVUR', fontsize=10)
plot5

scores4 = round(mlp.score(X2_test, y2_test),3)
plot6 = plt.scatter(predictions_4, y2_test)
plt.title(str("Association between Neural net regression model and test data, R-squared (2km) = " + str(scores4)))
plt.xlabel('Predicted LGVUR', fontsize=10)
plt.ylabel('LGVUR', fontsize=10)
plot6

Rsq_6 = np.corrcoef(X2_train[:,0], predictions_3,rowvar=0)
plot7 = plt.scatter(X2_train[:,0], predictions_3)
plt.title(str("Association between Neural net regression model and people's seniment (2km), R = -0.322" ))
plt.xlabel('Sentiment', fontsize=10)
plt.ylabel('Predicted LGVUR', fontsize=10)
plot7
```

```
Rsq_7 = np.corrcoef(X2_test[:,0], predictions_4,rowvar=0)
plot8 = plt.scatter(X2_test[:,0], predictions_4)
plt.title(str("Association between Neural net regression model and people's seniment, R = -0.539" ))
plt.xlabel('Sentiment', fontsize=10)
plt.ylabel('Predicted LGVUR', fontsize=10)
plot8
```