Center for Advanced Multimodal Mobility Solutions and Education

Project ID: 2019 Project 03

Analyzing Cycling Behavior During Different Time Periods Using Crowdsourced Bicycle Data

Final Report

by

Wei Fan (ORCID ID: https://orcid.org/0000-0001-9815-710X)
Zijing Lin (ORCID ID: https://orcid.org/0000-0001-6529-5725)

Wei Fan, Ph.D., P.E.
Director, USDOT CAMMSE University Transportation Center
Professor, Department of Civil and Environmental Engineering
The University of North Carolina at Charlotte
EPIC Building, Room 3261, 9201 University City Blvd, Charlotte, NC 28223
Phone: 1-704-687-1222; Email: wfan7@uncc.edu

for

Center for Advanced Multimodal Mobility Solutions and Education
(CAMMSE @ UNC Charlotte)
The University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223

September 2020

ACKNOWLEDGEMENTS

This project was funded by the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE @ UNC Charlotte), one of the Tier I University Transportation Centers that were selected in this nationwide competition, by the Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT), under the FAST Act. The authors are also very grateful for all of the time and effort spent by DOT and industry professionals to provide project information that was critical for the successful completion of this study.

DISCLAIMER

Table of Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

Cycling has been more and more prevalent among citizens especially in bike friendly cities where planners and policy makers have been promoting non-motorized travel mode. Compared to driving, cycling is healthier and is able to reduce energy consumption. Therefore, it is essential to analyze cycling behavior to better understand the needs of bicyclists. Since cycling behavior during different time periods can be distinctive, it is critical to discover the differences of cycling behavior in each time period.

To analyze cycling behavior, data including bicycle volume on each road segment, road characteristics, time of day, day of week are quite indispensable. The methods for data collection are diverse. In the past, traditional manual count data and travel survey data were the most commonly used ones for data collection. However, crowdsourcing is becoming more popular, since crowdsourced data can be cost effective and time saving compared to the other two data collection methods. In addition, traditional manual count data and travel survey data cannot provide spatial and temporal information which is critical for relevant research studies. Although these two data collection methods have been used for most of the previous research efforts, crowdsourced data has been selected by researchers to conduct relevant studies recently because of its ability to address the data gap for decision making and policy development.

This research concentrates on the analysis of cycling behavior utilizing crowdsourced bicycle data collected from Strava in the City of Charlotte. The cycling behavior of Strava users in the City of Charlotte during different time periods is compared. From the link-based cycling behavior prospective, several discrete choice models have been developed to model the preference of roadway segments along their cycling routes during different time periods. Factors including road characteristics, bike facilities, day of week etc. have been carefully examined to gain a better understanding of the variables that have significant impacts on link-based cycling behavior. From the route-based cycling behavior prospective, a method is provided to guide researchers to analyze cycling route choice including a choice set generation method and a Path Size Logit model for future route choice analysis.

Chapter 1. Introduction

1.1 Problem Statement

To obtain better and healthier lives, citizens are trying to spend more time on outdoor activities. And for traveling, more people prefer to select cycling for both commuting and recreational trips especially those with short distances. Cycling, as a healthier and greener non-motorized travel mode, has been encouraged by city planners and policymakers to help reduce energy consumption, decrease traffic emissions, and improve public health. However, there are several concerns for people to choose cycling over other travel modes in terms of safety issues, and environmental issues, etc. Compared with other road users, cyclists can be more vulnerable. Therefore, it is essential to analyze the impacts of different factors on cycling behavior, so that recommendations can be made to provide a better cycling environment.

One of the most useful ways to improve cycling condition is to construct bicycle facilities which can provide a safer and more comfortable cycling environment for the potential cyclists. The convenience brought by leveraging the well-constructed bicycle facilities may increase the bicycle level of service.

According to the Charlotte Department of Transportation (CDOT) Bicycle Program developed in 2017 (City of Charlotte Department of Transportation, 2017), the City of Charlotte has been making great efforts to become a bicycle friendly city for the past fifteen years. To promote cycling, a comprehensive bike plan has been implemented and improvements have been made to the policies. Since the first mile of bike lanes was constructed in 2001, the bike network in Charlotte has been expanding. There are more than 90 miles of bike lanes, 40 miles of greenways and off-street paths, and 55 miles of signed bike routes in the City of Charlotte (City of Charlotte Department of Transportation, 2017). Although the city bike network has been growing rapidly, there are still 62% of the residents in Charlotte who do not think biking is easy for them according to a survey conducted by CDOT (City of Charlotte Department of Transportation, 2017). However, more than half of the citizens would like to select cycling as their travel modes more than they currently do. Therefore, it is not difficult to infer that the cycling condition in the City of Charlotte still needs to be improved. And it can be expected that, more people will select cycling after the improvement of cycling environment.

To understand how to improve cycling condition and promote cycling among potential cyclists, factors need to be analyzed to examine the significant impacts on cycling behavior for both link-based route choice and path-based route choice. Therefore, data including bicycle volume on each road segment, road characteristics, sociodemographic information, temporal characteristics, etc. are essential for analyzing link-based route choice, while data including cycling trajectories and cycling route related information (distance, road characteristics, bike facilities, etc.) are necessary for analyzing path-based route choice.

The data collection methods for these research studies usually include three most commonly used ones which are traditional manual count data collected from the manual count machines, the travel surveys, and the crowdsourced data from the third party. Previously, most of the research efforts were conducted utilizing the first two data collection methods. But these two

methods can be expensive and time-consuming. The crowdsourced data, on the other hand, are timesaving and cost effective. In addition, this kind of data can provide spatial and temporal information that data collected using traditional methods cannot provide. Therefore, crowdsourced data have been widely used by researchers and many public agencies. Recently, crowdsourced data collected from smartphone applications (Strava etc.) have been prevalent among researchers since it has increased the availability of data collection and provided a feasible way to bridge the data gap for relevant research studies and decision making.

Crowdsourcing is an advanced data collection method. It has the advantages for researchers and practitioners to collect data from a large range of people in a time-saving and cost-efficient way. Usually, crowdsourcing involves crowd itself through an internet-based platform during an outsourcing procedure. It obtains useful information from the interested group and is utilized by scholars and planners to solve the relevant problems which can benefit the interested group back. The thought of crowdsourcing was first brought by Howe in his article "The rise of crowdsourcing" back in 2006 (Howe, 2006). With the development of crowdsourcing, researchers have utilized crowdsourced data to conduct research studies for different aspects including model development, travel behavior analysis, traffic demand estimation, bicycle facility evaluation, and road safety analysis.

With the development of GPS enabled smartphones, it is more convenient to collect crowdsourced data from smartphone applications. The first smartphone application for cycling data collection is CycleTracks developed by San Francisco County Transportation Authority in 2009 (San Francisco County Transportation Authority, 2013). Later, based on the first smartphone application, several different applications including Strava, Cycle Atlanta, and ORcycle etc. have been developed to conduct studies for various locations and research aspects. Obviously, the data provided by these applications can be distinctive. Some offer original cycling trajectories that need to be matched to the map, while others provide the aggregated data that have been preprocessed by specialists.

Based on the crowdsourced data collected from the smartphone applications, multiple models can be developed which include ordered probit models, ordered logit models, partial proportional odds models, path size logit models, expanded path size logit models, recursive models, and C-logit models. These models have been adopted for bicycle travel related research studies in terms of route choice behavior analysis, bicycle volume estimation and forecasting, bicyclist injury risk and safety analysis, air pollution exposure assessment, cycling comfort and level of service evaluation, etc. To develop these models, information besides crowdsourced data is still needed which contains road characteristics, sociodemographic factors, geometry features, air pollution measures, cyclist involved crash data, and temporal attributes, etc.

This research is intended to systematically analyze the cycling activities during different time periods for both link-based cyclist route choice and path-based cyclist route choice. Crowdsourced data utilized in this research are collected from Strava smartphone application which contain the Strava user counts on each road segment in the whole Charlotte network and the OD matrix for Strava users. To complete the link-based cyclist route choice behavior analysis, factors including road geometry, demographic characteristics, bicycle facilities, road features, and temporal data, etc. are carefully examined to develop several discrete choice models for different time periods. Data processing and combination procedures can be conducted

with ArcGIS and SAS. To provide a method to guide researchers to analyze path-based cyclist route choice, a choice set generation method has been selected and a Path Size Logit model for future route choice analysis has been presented.

1.2 Objectives

The objective of this report is to analyze the cycling behavior during different time periods in the City of Charlotte using crowdsourced bicycle data collected from Strava smartphone application, and compare different cycling behavior to provide specific recommendations accordingly on what can be done to help increase bicycle volume and build a better environment for bicycle riding. Route choice models will be developed for each time period, and the differences of each model will be identified which might not be explicitly accounted for in previous research. The proposed work in this report is to fulfill the following objectives:

1. To review and synthesize past experiences in cycling behavior analysis;

2. To compile the data needed for this project from all the available sources including Strava smartphone application data, roadway characteristics data, and other potential useful data for the follow-up work;

3. To analyze the crowdsourced bicycle data and conduct descriptive analysis;

4. To develop link-based cyclist route choice models using multiple discrete choice models;

5. To identify and compare the differences of cycling behavior between various time period;

6. To provide a choice set generation method for cyclist route choice behavior analysis;

7. To present the structure of a Path Size Logit model showing the method to analyze cyclist route choice for potential future studies.

1.3 Expected Contributions

To better understand the factors affecting bicyclist route choice behavior during different time periods, models need to be developed for both link-based route choice analysis and path-based route choice analysis. Along that line, the expected contributions of this research can be summarized as follows:

1. Present a systematic method for existing research efforts based on crowdsourced bicycle data;

2. Develop several discrete choice models to analyze link-based cyclist route choice behavior and identify the best model structure for this case study, examine and compare the impact factors for different time periods;

3. Provide a practical method to generate choice set for preparation of path-based route choice modeling.

4. Present the Path Size Logit model to give a method of analyzing path-based route choice for potential future studies.

1.4 Report Overview

The remainder of this report is organized as follows: Chapter 2 presents a comprehensive review of the state-of-the-art and state-of-the-practice on the link-based and path-based route choice behavior analysis using both traditional data collection methods and crowdsourced bicycle data. Chapter 3 discusses the bicycle count data and the OD matrix data collected from Strava application and other relevant supporting data. Chapter 4 conducts a descriptive analysis based on the data collected in Chapter 3. Chapter 5 develops several discrete choice models for analyzing link-based cyclist route choice behavior in City of Charlotte during different time periods. Impact factors have been compared across different time periods. Chapter 6 presents the structure of a Path Size Logit model showing the method to analyze cyclist route choice for potential future studies. Finally, Chapter 7 concludes this report with a summary and a discussion of the directions for future research.

Chapter 2. Literature Review

2.1 Introduction

Cycling behavior has been studied for decades to provide guidance to policy makers and planners for active transportation development and management, with the support of traditional data collection methods such as travel surveys and manual counts. Nowadays, the information and communication technologies have been developed, which leads us to a new era of big data. Numerous sources of novel data, including crowdsourced data collected from the smartphone have emerged and been utilized for the transportation research especially in travel behavior analysis area. The use of the innovative crowdsourcing data collection method, compared with the previous traditional data collection method, shows a lot of unique features and advantages. Thus, many researchers in relative research fields have been attracted to apply the crowdsourced data to their travel behavior research which has brought a certain amount of progress to date. And this is only the beginning of the benefit from crowdsourcing, this kind of data collection method still have great potential to be exploited for the further advanced transportation research studies.

This chapter provides a summary of the review of previous research efforts regarding the crowdsourcing data collection method and the potential use of the crowdsourced data on relative transportation research studies, especially route choice behavior analysis. The comprehensive review will greatly help in gaining a clearer understanding of the methods of modeling cyclist route choice behavior based on crowdsourced bicycle data for future research studies.

The remainder of this chapter is structured as follows. Section 2.2 gives a brief introduction to the existing data collection methods including open data, big data and stated preference, revealed preference travel surveys and other traditional transportation survey methods. Section 2.3 summarizes and introduces the smartphone crowdsourcing applications and the potential use of crowdsourced bicycle data for relevant research studies. Section 2.4 reviews the link-based cyclist route choice behavior analysis methods based on both traditional data collection methods and crowdsourced bicycle data. Section 2.5 provides a detailed description of the choice set generation methods prepared for path-based route choice behavior analysis. Section 2.6 summarizes the previous research in terms of the path-based route choice analysis methods. Finally, section 2.7 concludes the whole chapter.

2.2 Data Collection Methods

2.2.1 Crowdsourcing

Crowdsourcing is an innovative method which introduce new developments for the process of data collection. With the evolving of crowdsourcing, the definition of crowdsourcing has changed over the years. It was first brought up by Howe in 2006 in his article named “The Rise of Crowdsourcing”. According to his statement, crowdsourcing is defined as follow:

“Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.” (Howe, 2006)

Based on this new concept, lots of researchers who are interested in this kind of data collection method provide their own interpretations of crowdsourcing (Estellés-Arolas and González-Ladrón-de-Guevara, 2012). Usually, the definitions of crowdsourcing contain three main features that represent this data collection method including the crowd that provides critical information, the outsourcing procedure that spread out the data, and the internet-based platform that enables the accomplishment of crowdsourcing (Saxton, 2013). Some definitions of crowdsourcing are listed below:

(1) Crowdsourcing is an online production model that help solve the problem in recent years (Brabham, 2008)
(2) Crowdsourcing is an integration of the users or consumers that creates value in internal processes (Kleemann et al., 2008).
(3) Crowdsourcing is a new online production model that collaborates the networked people to solve the problem and complete a task (Vukovic, 2009).
(4) Crowdsourcing is an outsourcing procedure of a task or job that invites a larger group of innovators to provide a solution (Liu and Porter, 2010).
(5) Crowdsourcing is a procedure that motivates individuals to participate into the tasks voluntarily and allow both researchers and the crowd to find the solutions for the tasks (Schenk and Kishore, 2011).
(6) Crowdsourcing uses a passionate crowd or loosely bound public to solve the problems (Wexler, 2011).

2.2.2 Open Data

Open data is a kind of public or private dataset that anyone can get access to freely through the internet with no restriction or cost. Usually, the open data are released by local government or institutions for relevant research studies.

There are certain requirements that open data should meet. One of the requirements is to ensure the free usage of data and the ability to reuse and redistribute data. The definition of "open" indicates that this kind of data is not restricted to a specific field or by any individual. Thus, according to Attard et al. (2015), the open data that are already published should be platform independent, information reusable, machine readable, and public available without any restrictions. To conclude, the open data means the type of data that are available through internet without any extra charges or limitations (Reiche and Höfig, 2013). Under this circumstance, open data are seen as the critical motivator of open government (Kučera et al., 2013).

2.2.3 Big Data

Big data are prevalent for multiple aspects of research studies recently which may require advanced data processing, essential data cleaning procedure, data integration with other supporting datasets to provide critical information for decision making.

With the development of technology, the importance of big data has been revealed. Generally speaking, big data refer to datasets that are too large to perceive, difficult to acquire, complex

to manage, and hard to processed by the traditional application software within a reasonable amount of time. Different definitions for big data are provided from various points of view by researchers, technological and scientific companies, data analysts, and technical specialists.

Apache Hadoop defines big data as the large datasets that cannot be extracted, managed, and processed by normal computers within an acceptable scope in 2010 (Chen et al., 2014). Similarly, an IDC report defines big data as a new generation of technologies that are designed to obtain the information from large volumes of data with wide diversities through an efficient data extraction and analysis procedure (Gantz and Reinsel, 2011). Thus, four main features of the big data can be identified which are large in volume, great in variety, fast in variation, and high in value.

2.2.4 Traditional Survey Methods

There are many traditional survey methods that have been utilized for data collection. Stated preference survey (i.e., SP survey) and revealed preference survey (i.e., RP survey) are the two kinds of survey methods that are commonly used by researchers. The SP survey designs the investigation based on assumed values, since the content of the questions is intentionally made up and has not taken place. This feature gives SP survey the advantage of flexibility. On the contrary, RP survey is designed to acquire results of choices from the respondents under certain selection conditions. Unlike SP survey, the content of investigation in RP survey has already taken place. In other words, the results of RP survey are the reflection of the actual choice behavior (Guan, 2004).

Other survey methods (both paper-based and web-based) that have been used massively include traditional household survey (Kagerbauer et al., 2015), workplace survey, longitudinal and panel survey, transit on-board ridership surveys, commercial vehicle (truck) surveys and external station survey.

2.3 Smartphone Crowdsourcing Applications and Their Potential Use

As stated in Section 2.2.1, there are numerous definitions of crowdsourcing. This section will concentrate on the smartphone crowdsourcing applications that are related to cycling and introduce and summarize the potential use of this kind of data.

The first smartphone application designed for cycling data collection is CycleTracks which was developed by the San Francisco County Transportation Authority (SFCTA) in 2009 (SFCTA, 2013). The GPS-enabled smartphones were utilized to collect the cycling trajectory. In addition, demographic information and trip purposes were also collected from the users.

Based on the first smartphone application, AggieTrack was developed by Texas A&M University for collecting the travel information from the users within the university area (Hudson et al., 2012). Data including travel mode, trip purposes, classification (student, faculty or staff), etc. were collected after the generation of each trip.

In addition, Cycle Atlanta was also developed based on the CycleTracks smartphone application (Misra et al., 2014). Different from the previous applications, Cycle Atlanta provided

several additional features with a different user interface. Data other than the cycling trip related information were collected such as issues encountered by cyclists during their cycling trips, bicycle parking, and the locations of certain infrastructures. Similarly, demographic information is collected.

Based on Cycle Atlanta, RenoTracks was developed during 2013 (RenoTracks 2013). Different from Cycle Atlanta, RenoTracks added the "CO_2 Saved" counter calculating the carbon dioxide emission reduction when selecting bicycle as the travel mode instead of automobile.

Strava is another smartphone application that has been widely used by numerous cyclists. Speed, distance, and trip time are displayed on the personal record dashboard. Graphical representations of the route profile and plan overview are also provided. The unique function for Strava enables the users to compete with other cyclists who bike on the same segment by tracking performance of the Strava users. This functionality helps Strava become more social and attracts lots of cyclists. Other popular smartphone applications that are used recently include Mon RésoVélo (Jackson et al., 2014), MapMyRide, MyTracks, and ORcycle (Broach et al., 2012).

These smartphone crowdsourcing applications offer massive data for researchers to conduct various research studies in terms of link-based and path-based cyclist route choice behavior analysis which will be reviewed in detail in the following sections. In addition, this kind of crowdsourced data can be utilized for other research areas as well including cycling safety, cycling activities associated with air pollution exposure, bicycle level of service, and health impact assessment, etc.

Raihan et al. (2017) investigated the impact of roadway characteristics and bicycle facilities on bicycle safety. In order to examine the association between bicycle crash frequencies and the impact factors (roadway characteristics and bicycle facilities), Crash Modification Factors (CMFs) were developed utilizing a robust cross-sectional analysis. This research focused on the urban facilities where 98 percent of the bicycle crashes occurred. The CMFs developed in this research provided the quantitative results of the impact of roadway characteristics on bicycle safety which was not studied by lots of researchers. In addition, bicycle exposure was considered based on the cycling data obtained from Strava application.

Sun and Mobasheri (2017) conducted a study on the air pollution exposure for both commuting and non-commuting trips. Spatial patterns of non-commuting cycling trips were identified. Cycling behavior was analyzed based on the number of non-commuting trips for different environmental characteristics. Data utilized in this research study were collected from Strava Metro. According to the Strava nodes data, compared with commuting trips, non-commuting trips tend to be occurred in the outskirts of the city. Cyclists biking for non-commuting trips have a higher likelihood to be exposed to low levels of air pollution while comparing to the commuting trips. In addition, the method adopted in this research study was examined to have good fitness for estimating the number of non-commuting trips.

Strava data were utilized by engineers and planners in Foresite Group (2015) to investigate the representativeness of the crowdsourced data. The correlation between the Bicycle Level-of-Service (BLOS) grades were evaluated with traditional methods and the crowdsourced

data. Results revealed that Strava data may not have the ability to represent all the cyclists, and the majority of the cycling trips are recreational activities.

To identify the location of cycling activities especially recreational trips, Griffin and Jiao (2015) analyzed the data collected from Travis County, Texas. Bicycle volumes were estimated based on the residential and employment density, the land use categories, bicycle infrastructures and terrain. Locations that were selected for recreational cycling trips were identified. The method developed in this research study provided guidance for health impact analysis studies.

2.4 Link-based Cyclist Route Choice Behavior Analysis

Many researchers have conducted their studies by using crowdsourced data. GPS enabled smartphones to provide researchers new opportunities to collect data from a broader group of people and use them to conduct the cyclists' route choice analysis. The existing use of crowdsourced data for link-based cyclist route choice behavior analysis is presented as follows.

Moore (2015) conducted a study to analyze the impact of various factors on cycling route choice based on the crowdsourced bicycle data collected from Strava application. An ordinal logistic regression model was developed to examine the effect of impact factors on the cyclists' route choice. GIS was applied to conduct a qualitative analysis in order to investigate the specific areas and facilities to discover their differences from other facilities. Results revealed that the selection of a road segment is highly associated with the road characteristics and the land use.

Griffin and Jiao (2016) collected data from both CycleTracks smartphone application and the Strava fitness application to conduct a data comparison between crowdsourced bicycle data and the manual count bicycle data. Five specific locations were selected in the downtown Austin, Texas. All the data were compiled and compared in GIS for these five locations.

To explore the relationship between manual count data collected in Victoria, British Columbia, Canada and crowdsourced bicycle data from Strava application, a generalized linear model was developed by Jestico et al. (2016). The bicycle volumes were categorized into several levels, and a regression model was developed for the prediction of bicycle volume level. The maps that illustrate the distribution of bicycle volumes were created. Results revealed that the bicycle trips recorded by Strava are similar to the commuting trips in the urban areas of the mid-size North American cities.

A data comparison was conducted by Watkins et al. (2016) to find out the differences between Cycle Atlanta and Strava data in terms of the sociodemographic information, total cycling trips on each road segment, and the cycling trips during each time of day. In addition, the manual count data were compared to the crowdsourced bicycle data from Cycle Atlanta in both AM and PM peak hours. The percentage of the manual count data collected by Cycle Atlanta was calculated based on data selected from 78 intersections. The data comparison results indicated that noticeable differences exist in the populations of the crowdsourced data. Thus, the bicycle data collected from smartphone applications should be carefully utilized before conducting relevant research studies.

Hochmair et al. (2017) utilized the crowdsourced bicycle data collected from Strava application in the Miami-Dade County area to analyze the impact of demographic information,

network characteristics (especially bicycle facilities), and place specific features on bicycle ridership. A series of linear regression models were developed to predict the bicycle kilometers traveled for both commuting and non-commuting trips, and trips occurred on both weekdays and weekends. Eigenvector spatial filtering was adopted to avoid bias and model spatial autocorrelation. Results showed that Strava data performs well for the analysis of the impact of explanatory variables on bicycle volumes for commuting and non-commuting trips and during different days of week. In addition, Strava data revealed the broad coverage of spatial and temporal information and can be utilized as a critical supplement to bicycle volume estimation in large areas.

Route choice analysis was conducted by LaMondia and Watkins (2017) based on the crowdsourced bicycle data collected from Strava, Cycle Dixie and Cycle Atlanta. The impact factors were identified by modeling the bicycle facility preferences. In addition, cyclists' route segment choice and route choice were analyzed. Results revealed that sociodemographic information, road characteristics, and land use have a significant impact on the route segment choice.

Proulx and Pozdnukhov (2017) developed a novel method with geographically weighted data fusion for bicycle volume estimation utilizing crowdsourced data from Strava smartphone application and Bay Area Bikeshare data. It can be found that the method of Geographically Weighted Data Fusion can improve predictive accuracy for link-level bicycle volume estimation.

Zimmermann et al. (2017) analyzed the link-based cyclist route choice based on the GPS data in the network with more than 40,000 road segments in the City of Eugene. A recursive logit (RL) model following the research conducted by Fosgerau et al. (2013) was developed which did not require the choice set generation procedure. The results showed the advantages of this method in terms of the link flow prediction and accessibility measures. Compared to the path-based route choice models, this method is better in computational time and may avoid paradoxical results which is consistent with Nassir et al. (2014).

To conclude, a summary of the link-based route choice analysis studies is provided below in TABLE 2.1.

Table 2.1 Summary of Link-based Route Choice Analysis

Year	Author	Data	Methods	Results
2015	Moore	Data from Strava application	Ordinal logistic regression model	Roadway characteristics and surrounding land-use have a significant impact on whether or not a particular street segment would be used.
2016	Griffin and Jiao	Data from CycleTracks, Strava application, and traffic counts	Ordinary least squares regression	Crowdsourced data are appropriate for bicycle volume evaluation.
2016	Jestico et al.	Data from Strava and manual	Generalized linear model	In mid-size North American cities within urban areas, the routes

		counting data		recorded in crowdsourced fitness application tend to be similar with those of the commuter cyclists.
2016	Watkins et al.	Data from Cycle Atlanta, Strava, and actual cyclist trips	Data comparison	The smartphone application data should be carefully used considering the likely bias.
2017	Hochmair et al.	Data from Strava application	Linear regression models	Strava data can be used to examine the impact of explanatory variables on estimated bicycle volume.
2017	LaMondia and Watkins	Data collected using the Strava, Cycle Dixie and Cycle Atlanta	Route suitability score and preference models	Demographics, roadway characteristics and surrounding land-use have a significant impact on route choice.
2017	Proulx and Pozdnukhov	Crowdsourced data from Strava and usage data from Bay Area Bikeshare	Geographically Weighted Data Fusion	The method of Geographically Weighted Data Fusion can improve predictive accuracy for link-level bicycle volume estimation.
2017	Zimmermann et al.	GPS observations in the city of Eugene	Link-based bike route choice model (recursive logit model)	Cyclists are sensitive to distance, traffic volume, slope, crossings and the presence of bike facilities.

2.5 Choice Set Generation Methods

In a path-based route choice modeling procedure, there are usually two steps. First, possible alternative routes within the roadway network are needed to be generated to comprise the choice set. After that, the probability of a certain route being chosen from the generated choice set is calculated based on the route choice model. Thus, this section will introduce various methods to accomplish the first step of the route choice modeling which is choice set generation.

In a bicycle network, there are numerous alternative routes for bicyclists to choose either for their commute trips or their recreational trips. Since the purpose of this project is to analyze bicyclists' route choice behavior, the preparation work (choice set generation) is essential. This task of the project is to ideally identify all the biking routes that any traveler might consider. In particular, algorithmic rules for generating the observed biking routes to avoid biases in the model estimation procedure is critical in this task.

There are many previous methods for the design of a path generation algorithm. One of the well-known methods is called the K-shortest Path algorithm which generates the first "k" shortest paths for a given origin-destination pair in a roadway network. There are two popular heuristics which are link penalty and link elimination methods (De La Barra et al., 1993).

According to these two heuristics, the link penalty method gradually increases the impedance of all links on the shortest path, while the link elimination method removes the links on the shortest paths in sequence to generate new routes.

The labeling approach is also a choice set generation method. It allows the availability for multiple link attributes including travel time, distance, cost, etc. that produce alternative routes (Ben-Akiva et al., 1984). In this method, the routes may be labeled based on the criteria such as "minimize time", "minimize distance", "minimize cost", "maximize the use of expressways", etc.

In addition, simulation methods produce alternative feasible paths by drawing impedances from different probability distributions. The distribution type (for example, Gaussian, Gumbel, Poisson), distribution parameters, number of draws and the seed of the pseudo-random number generator are design variables. (Bekhor et al., 2006)

Many researchers have applied different methods of choice set generation to get ready for the route choice analysis. The existing choice set generation methods that include but are not limited to the ones introduced above are presented as follows.

Bekhor et al (2006) utilized the simulation methods to produce alternative paths which form the choice set. A Gaussian distribution with a mean and standard deviation calculated from travel times was used. (The choice of the Gaussian distribution was primarily for computational convenience, rather than for any theoretical reason.) Up to 48 draws were simulated for each observation, as this was estimated to take roughly the same computational time as the link elimination and link penalty algorithms.

The choice set generation approach used in the research conducted by Hess et al. (2015) was developed by Rieser-Schssler et al. (2013) specifically for route generation in high-resolution networks and successfully applied to different bike and route choice problems. The ability to apply this non-behavioral approach easily across different context and countries is a clear advantage, with only an application-specific cost function being needed for each study. The method employs a link elimination approach which means that links of the current least cost path are eliminated before the next least cost path is searched. This is repeated until the required number of routes is found.

Broach et al. (2009) developed a sophisticated choice set generation algorithm based on multiple permutations of labeled path attributes, which seems to out-perform comparable implementations of other route choice set generation algorithms.

Bierlaire et al. (2010) sampled the path alternatives using a biased random walk algorithm, with arc weights at each node set by the ratio of the length of the shortest path to the destination using any arc and using the target arc. The sampling bias was subsequently corrected in the choice model.

Menghini et al. (2010) employed a breadth-first search link elimination approach. It searches for the shortest path between origin and destination and removes the links in turn. These shortest paths became in turn the starting points for the next iteration of link elimination. The algorithm kept track of the networks generated and retained only unique and connected networks

and in turn shortest paths for the choice set. The depth, i.e. number of links removed, was increased until the desired number of distinct routes in the choice set had been generated or the original shortest path was exhausted.

Frejinger et al. (2009) presented a new paradigm for choice set generation in the context of route choice model estimation. The choice sets were assumed to contain all paths connecting each origin-destination pair. Although this is behaviorally questionable, this assumption was made in order to avoid bias in the econometric model. These sets were in general impossible to generate explicitly. Therefore, an importance sampling approach was proposed to generate subsets of paths suitable for model estimation. Using only a subset of alternatives requires the path utilities to be corrected according to the sampling protocol to obtain unbiased parameter estimates. A sampling correction was derived for the proposed algorithm.

2.6 Path-based Cyclist Route Choice Behavior Analysis

Based on the methods that generate appropriate choice set, the path-based cyclist route choice behavior analysis can be conducted. Several previous research studies concentrating on the path-based route choice behavior are summarized as follows.

Stinson and Bhat (2003) examined the explanatory variables that have a significant impact on the commuting cycling trips. Two categories of factors were considered which include route level and link level attributes. Data used in this research study were collected based on a stated preference survey completed through the internet. Empirical models were developed, and results were concluded that the most critical factor for commuting trips is travel time. Other factors that affect the route choice significantly included bicycle facilities (e.g., bike lanes and separate paths), traffic condition, and pavement quality. Policy implications were provided for bicycle facility planning based on these results.

Dill and Gliebe (2008) studied the impact of different types of bicycle facilities on bicycle activities. GPS data were utilized with a sample of 164 cyclists in Portland, OR from March to November 2007. The cyclists selected in this research study usually bike more than one day per week. Four major sets of research questions were addressed with the GPS data. Results revealed that most of the cycling trips generated by the participants are for utilitarian purposes. Approximately half of the cycling trips occurred during AM and PM peak hours. Bicycle facilities were preferred by cyclists biking for utilitarian purposes. The main factors that affect the route choice were cycling distance and the traffic condition.

Sener et al. (2009) examined a comprehensive set of attributes that influence bicycle route choice. The data used in the analysis was drawn from a web based stated preference survey of Texas bicyclists. The results of the study emphasized the importance of a comprehensive evaluation of both route-related attributes and bicyclists' demographics in bicycle route choice decisions. The empirical results indicated that travel time (for commuters) and motorized traffic volume were the most important attributes in bicycle route choice. Other route attributes with a high impact included number of stop signs, red light, and cross-streets, speed limits, on-street parking characteristics, and whether there existed a continuous bicycle facility on the route.

Winters et al. (2010) investigated differences in total distance, road type used, and built environment features for shortest-path routes versus actual routes for utilitarian bicycle trips and

car trips in Metro Vancouver, Canada. Regardless of mode, people did not detour far off the shortest route: detour ratios (actual distance/shortest distance) were similar. Compared with shortest-path routes, cyclists spent significantly less of their travel distance along arterial roads and significantly more along local roads, off-street paths, and routes with bike facilities. As expected, car trips were more likely to be along highways and less likely to be along local roads than predicted by the shortest route.

Charlton et al. (2011) introduced the CycleTracks smartphone application and its use for recording cycling trips by cyclists. The cycling data in terms of cyclist-related and trip-related information were collected via this smartphone application. The potential data bias was discussed, and route choice model was developed based on this dataset.

Following the research of the first bicycle route choice model built with GPS data in Zurich, Hood et al. (2011) analyzed the cyclist route choice and developed a route choice model based on the crowdsourced bicycle data collected from CycleTracks. A "doubly stochastic" choice set generation method was adopted in this research based on the study conducted by Bovy and Fiorenzo Catalano (2007). Instead of using the multinomial logit model that has the independence of irrelevant alternatives property, a path size logit model was developed with the path size factor of Ben-Akiva and Bierlaire (1999). The model estimation results indicated that bike lanes have a positive impact on cycling, while steep slope and turning have a negative impact.

To analyze the bicyclists' route choice, especially the preference for bicycle facilities, Broach et al. (2012) developed a route choice model base on the GPS data of 1449 cycling trips occurred in Portland, Oregon. Three choice set generation methods (K-shortest paths, route labeling, and simulated shortest paths) were compared, and a modified method of route labeling was developed and utilized for this research study. A path size logit model was built for cyclist route choice analysis, and route choice differences between commuters and non-commuters were identified. The model results showed that factors including trip distance, intersection control, turning, traffic volume, and slope have significant impact on cyclists' route choice. In addition, trip distance was more important to commuters than non-commuters.

Chen and Chen (2013) examined recreational cyclists' preferences for bicycle routes in Taiwan using the stated preference method. The multinomial logit model was employed to estimate the relative influences of facility attributes on bicycle route choice behavior, while the latent class model was adopted in order to better understand the differences in preferences. Using data collected from 232 recreational cyclists in Taiwan, the results indicated that bicycle facility attributes, such as basic facilities and maintenance equipment, tourist information centers, and attractions had significant effects on recreational cyclists' preferences. Cyclists with high levels of recreation specialization appeared to be more likely to choose challenge and endurance routes than those with low recreational specialization.

Casello and Usyukov (2014) estimated the utility/generalized cost function of the path alternatives for each cyclist based on the GPS data that record the cycling activities. Four non-chosen alternatives were generated for the choice set. Two multinomial logit models were developed for the route choice analysis. The explanatory variables including the length of trips, automobile speed, slope, and the bike lanes were examined to see whether they have significant

impacts on cyclists' route choice. The predictive powers of the multinomial logit models were tested based on the 181 trips which were not used for the model parameter estimation. The results showed that vehicle speeds and the presence of bike lanes are factors that affect cyclists' route choice significantly.

Yeboah et al. (2015) used the GPS tracks and travel diary data from 79 cyclists around Newcastle upon Tyne in North East England as well as the OpenStreetMap (OSM) as the transportation network to conduct route choice analysis. Factors based on the previous relevant literature were examined to test the impact on commuting cycling trips. The results showed that OSM combined with GPS data of cycling trajectory performs well for bicyclists' route choice research. The transportation network restrictions including one-way road, turning restrictions and access for the selected routes and the shortest paths were significant. In other words, it is critical to consider route directness for both restricted and unrestricted transportation networks.

Bergman and Oksanen (2016) collected the crowdsourced bicycle data from Sports Tracker and utilized this data and OpenStreetMap to provide automatic route choices. The Sports Tracking data was pre-processed, and path choice set was generated. An advanced Hidden Markov Model (HMM)-based algorithm and a simple geometric point-to-curve method were utilized and compared to conduct the map-matching procedure. Results showed that HMM-based algorithm provides better matching performance.

Grond (2016) conducted a research study on the influence of physical and environmental factors of the network on cyclists' route choice. This study can provide a better understanding of the impact of physical infrastructure which will guide the city planners for bicycle facility investment. Data utilized in this research study were collected from the cycling application with cycling trips recorded from August 23 and September 23, 2015 in the City of Toronto. GPS tracks were matched to the GIS network dataset containing road characteristics. A path size multinomial logit model was developed for route choice analysis. Factors including bicycle facilities, road characteristics, demographic information of the cyclists were carefully examined in the model.

Khatri et al. (2016) utilized GPS data collected from Grid Bikeshare recording 9,101 trips created by 1,866 bikeshare users in Phoenix, Arizona to analyze the cyclists' route choice behavior, especially the impact of bicycle facility. Only direct utilitarian trips were considered in this research study, and circuitous trips or recreational trips were removed from the dataset. The results showing route choice behavior of register users and casual users were compared. It was found that registered users prefer roads with low traffic volume and bicycle facilities, and their trip distance tend to be shorter. A path size logit model was developed subsequently to model cyclists' route choice. Results revealed that cyclists are sensitive to the trip distance and prefer using bicycle facilities. Casual users disliked left turns compared to right turns. Explanatory variables including one-way road, AADT, and trip distance were found to have a negative impact on route choice, while the number of signalized intersections was likely to affect cyclists' route choice positively.

To conclude, a summary of the path-based route choice analysis studies is provided below in TABLE 2.2.

Table 2.2 Summary of Path-based Route Choice Analysis

Year	Author	Data	Methods	Results
2008	Dill and Gliebe	GPS data of Portland, OR	USGS Digital Elevation Model	The majority of the bicycle travels were for utilitarian purposes. About half of the trips occurred during morning and evening peak travel times. Distance and traffic volume have a negative impact on route choice.
2009	Sener et al.	A web based stated preference survey of Texas bicyclists	Panel mixed multinomial logit	Travel time (for commuters) and motorized traffic volume are the most important attributes in bicycle route choice.
2010	Winters et al.	A survey conducted in 2006 in Metro Vancouver	Logistic model	Road infrastructure and bicycle-specific aspects of the built environment influence people's travel patterns: that car drivers detour from shortest routes to fast roads and cyclists deviate from shortest routes to routes with better bicycle facilities.
2011	Charlton et al.	Data from CycleTracks application	Bicycle route choice model	Cyclists are sensitive to slope, presence of bike lanes or bike route designations. The route choice behavior is also influenced by trip purpose and gender.
2011	Hood et al.	Data from CycleTracks application	Path Size Multinomial Logit model	Bike lanes are preferred compared to other types of bicycle facilities, while steep slopes are disfavored. Length and turns have negative impact on route choice. Surprisingly, traffic volume, speed, number of lanes, crime rates and nightfall have no impact on route choice.
2012	Broach et al.	The GPS data collected in Portland, Oregon	Path-Size Logit (PSL) model	Cyclists are sensitive to distance, turn frequency, slope, intersection control and volumes. For commuters, they are more sensitive to distance than non-commuters.
2013	Chen and Chen	Data collected from 232 recreational cyclists in Taiwan	Multinomial logit model and latent class model	Bicycle facility attributes, such as basic facilities and maintenance equipment, tourist information centers, and attractions had significant effects on recreational

				cyclists' preferences.
2014	Casello and Usyukov	724 cycling trip GPS data	Multimodal logit models	Cyclists consider both vehicle speeds and the presence or absence of a bike lane during route choice process.
2015	Yeboah et al.	OpenStreetMap, GPS tracks (7 days) and travel diary data	Four-step method for generating routes	Network restrictions for both observed and shortest paths are significant.
2016	Bergman and Oksanen	OpenStreetMap and mobile application data from Sports Tracker	advanced HMM-based algorithm	HMM-based algorithm has better matching results in terms of the number of the correctly matched road segments.
2016	Grond	GPS dataset from the City of Toronto's cycling app	path-size multinomial logit model	Steep hills, high traffic volumes, left turns without signalized intersections and right turns at signalized intersections have negative impact on route choice.
2016	Khatri et al.	GPS data from Grid Bikeshare in Phoenix, Arizona	Path Size Logit Model	The proportion of one way segments, AADT and length of trip have a negative influence on route choice and number of signalized intersections has a positive influence on selecting routes.

2.7 Summary

This chapter provides a comprehensive review of the previous research on both linked-based and path-based cyclist route choice behavior analysis especially those based on crowdsourced bicycle data. It is intended to give a better understanding of crowdsourcing, and existing research efforts utilizing crowdsourced data which will provide a useful reference for future studies.

Chapter 3. Collecting Crowdsourced Data and Other Supporting Data

3.1 Introduction

Collecting data including crowdsourced bicycle data from Strava and other relevant supporting data is the first step of this research study. Chapter 3 provides an introduction of the collected Strava data as wells as the critical supporting data that will be utilized for the modeling in the model development sections.

The following sections in Chapter 3 are organized as follows. Section 3.2 introduces Strava in detail to give an overview of this smartphone application. Section 3.3 presents the Strava data that are collected for the research study. Section 3.4 shows the data view to Strava data for different aspects including street view, intersection view, OD view, and the heatmap view. Section 3.5 presents the other essential supporting data that will be used for the link-based route choice behavior. Finally, Section 3.6 concludes this chapter with a summary.

3.2 Introduction to Strava

Smartphone applications like Strava tend to generate route data that are saved in databanks together with the demographic details of the user derived from the application. These route data contain sensitive information, such as the user's place of residence or workplace, which can also be connected to profile information such as name, age, gender, and other freely given information. When passing on data to third parties, vendors are obliged to anonymize this information in accordance with the data protection laws and general conditions of business. In consequence, the buyer acquires data that have already been aggregated and is not allowed for any tracing back to the people that created the data. Anonymized demographic information such as gender and age are permitted to remain in the dataset. The data from global vendors of smartphone applications offer the largest range and number of possible users. Considerable differences can emerge within the user structure. The data are obtained second by second, saved at the end of the journey and transmitted to a server. The data can then be viewed by users on their smartphones and shared with others. This social factor feeds the user's motivation, in sporty applications such as Strava, to share the route just traveled with others or to keep a training journal.

The routing data used in this project are collected from Strava smartphone application developed by a technology company recording the cyclist travel trajectory with the GPS located in their smartphones. A screenshot of the application interface can be seen in Figure 3.1, which also shows some of the information that the app displays to the user after a route has been recorded. The application is available for use by any person who has a GPS device and access to the internet, with the majority of users comprised of cyclists and runners. As the cyclist uses the app, information such as duration, speed, elevation change, and distance are collected, along with the GPS route information. This allows the user to be able to look and see not only where they went but they can also analyze how well they performed and compared with other users. The accuracy of the GPS data from both apps depends on the connection to the GPS satellites, with more satellites available the better the accuracy. Having an unobstructed signal to the satellites is

also important to have high-quality accuracy, with dense tree foliage and tall buildings obscuring and scattering the GPS signal.

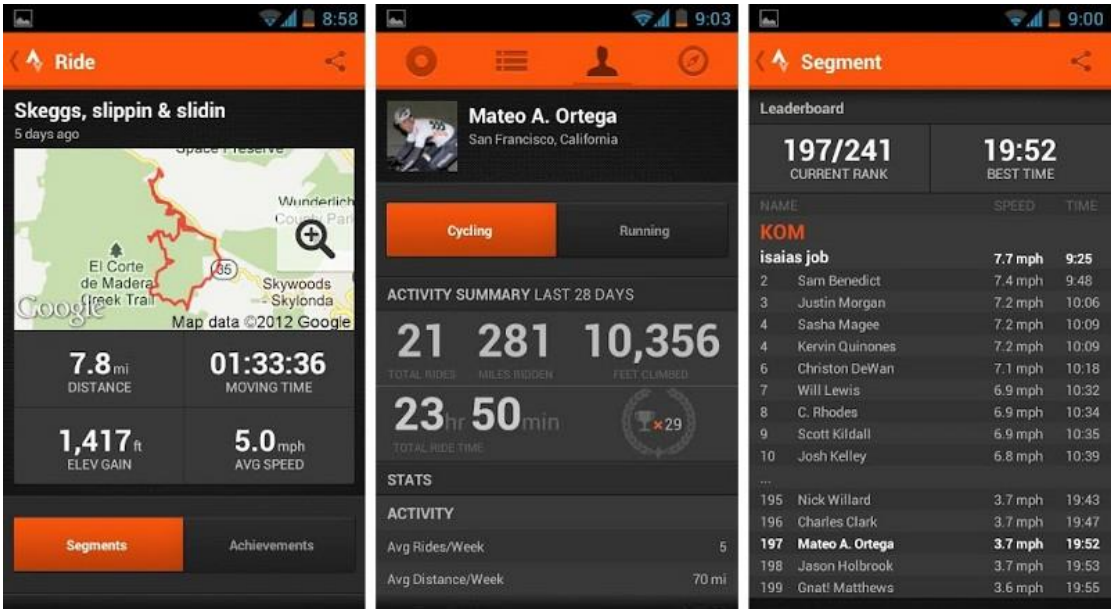


Figure 3.1: Strava App Screen Shots

3.3 Strava Data

The GPS data collected from the Strava users usually include the biking information on the network at both the link-level and the intersection-level. The link-level data set contains the Strava user counts on each roadway segment and the intersection-level data set includes the number of cyclists for each intersection as well as their waiting times. To record the cycling route of the Strava users, the OD matrix data set is provided.

The data offered by Strava Metro usually contain three main components including core data, roll-ups, and reports. The core data provide cycling information in each minute in the city network at both the link-level and intersection-level. In addition, it provides the OD pairs for the cycling trips. The roll-ups data are the aggregated data developed from the core data to obtain cycling information for different times and trip purposes. And the reports of the data show a summary of the cyclists' demographic information. The detailed data deliveries of Strava Metro can be found below.

3.3.1 Core Data

1. Link-level data set: Database file that presents the cycling information (especially bicycle counts) on each roadway segment during the time period of the delivery.

2. Intersection-level data set: Database file shows the cyclist counts and waiting time at each intersection during the time period of the delivery.

3. OD data: Origin/Destination file provides the cycling trip information including the OD pairs during the time period of the delivery.

3.3.2 Roll-ups

The roll-up data are the categorized core datasets that are processed by Strava Metro. For the link-level and intersection-level core dataset, several roll-ups are provided to summarize the views that present total counts, hour groupings, monthly use, weekday/weekend, and seasonality. In addition, other views of the roll-ups can be generated by researchers based on the specific research needs.

The seasonality and hour groupings categorized for this research studies in the City of Charlotte are shown as follows.

On season: From March to October
Off-season: From November to February
Early AM hours: 12:00 am - 5:59 am (labeled as_0)
AM peak hours: 6:00 am - 8:59 am (labeled as_1)
Mid-day hours: 9:00 am - 2:59 pm (labeled as_2)
Peak afternoon hours: 3:00 pm - 5:59 pm (labeled as_3)
Evening hours: 6:00 pm - 7:59 pm (labeled as_4)
Late evening hours: 8:00 pm - 11:59 pm (labeled as_5)

3.3.3 Reports

1. Demographics: A report that summarizes the cyclist demographic information in terms of different age and gender.

2. Summary: The total Strava user counts and the cycling activities recorded during the time period of the delivery.

3.4 Data View

Metro Data view is another way for researchers to visualize the cycling information on the total biking activities, total cyclists, and the commuters aggregated to the street level, intersection and the origin-destination polygonal geometry. The default data view shows the total cycling activities in the whole network. Researchers can find the useful information (e.g., total activities) by selecting the intersection button in the map interface. When selecting the intersection view, the median waiting time of crossing an intersection can be shown. In addition, a heatmap is provided to have an overview of the number of cycling activities in the whole network. The four data views that Strava Metro provided for the City of Charlotte can be seen as follows.

3.4.1 Street

The Street Data illustrating the cyclist counts in the City of Charlotte can be found in Figure 3.2, with dark blue color showing the lowest number of rides, and dark red representing the highest activity counts. The researchers can find the levels of activity counts by different colors corresponding to the number of rides shown in the legend. By hovering on the legend, researchers can find the percent distribution of number of streets. In addition, the counts of cyclists on a specific road segment can be seen by hovering on that street.

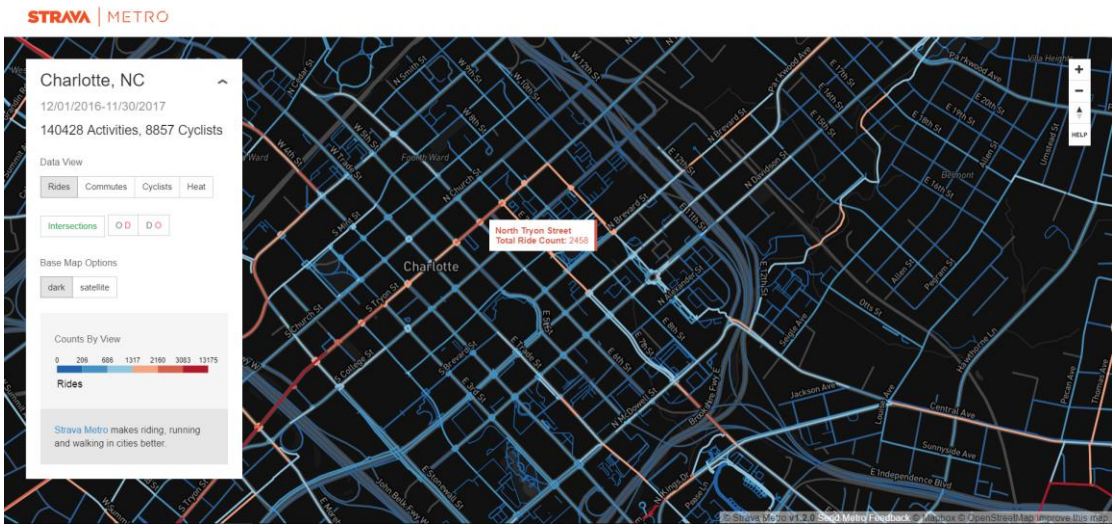


Figure 3.2: Charlotte Metro Data View 2017 Sample: Total activity counts from December 01, 2016 to November 30, 2017

3.4.2 Intersections

The intersection view can be presented by selecting the intersection button on the map interface. The default map view will not provide the intersection-level data when the intersection selection is off.

There are multiple map interfaces that can be selected to show different data information. To view the counts of activities, rides button should be selected. Also, clicking on the cyclists button, the number of cyclists at the intersection can be shown. Cycling for different trip purposes can also be displayed by clicking on the commutes button to show the number of commuting trips at each intersection.

The visualization view of the intersection map interface can provide an overview of the rides, commutes, and bicyclist counts at each intersection, with larger nodes representing the higher counts, and brighter nodes depicting the longer intersection crossing time. When hovering on the specific intersection that a researcher might be interested in, the map will show the exact cycling activity data. The detailed intersection data view in the City of Charlotte can be found in Figure 3.3.

Figure 3.3: Charlotte Intersection Metro Data View 2017 Sample: Total activity counts from December 01, 2016 to November 30, 2017

3.4.3 Origin and Destination

The origin and destination data view shows a cycling trip generated by a bicyclist with an origin/destination polygon layer based on a contiguous 350-meter hexagonal bin. Similar to the intersection button, the default data view will not show the OD pair information when the OD button is not selected.

Like the intersection view, different cycling information data can be obtained by selecting the toggle buttons shown on the map interface. By selecting the “Rides” button, the total number of cycling activities started within the polygon can be shown. To view the bicyclist counts, researchers can click on the cyclist button to obtain data regarding the number of bicyclists departed within the polygon.

Similarly, the visualization of the OD map indicates an overview of the rides, commutes, and bicyclists within the polygon with darker polygons representing the fewer counts and lighter polygons depicting the higher counts. When hovering on the polygon of a specific area, researchers can see the exact data of the trip origin. To view the destination polygons associated with the selected origin polygon, researchers can click on the origin polygon. To distinguish between origins and destinations, the destination polygons will be shown in pink. The darker color represents polygons with more rides and the lighter color depicts less rides. The origin destination data view in the City of Charlotte can be seen in Figure 3.4.

Figure 3.4: Charlotte Origin Destination Metro Data View 2017 Sample

3.4.4 Heat Map

The heat map view shows a visualization of the GPS points which are aggregated to the road segments. To view the heatmap of the streets or intersection, the “Heat button” should be selected. Streets with higher cycling activity counts will be shown in brighter lines, while streets with a fewer number of cycling activities will be shown inn darker lines. The heat map view of the City of Charlotte can be seen in Figure 3.5.

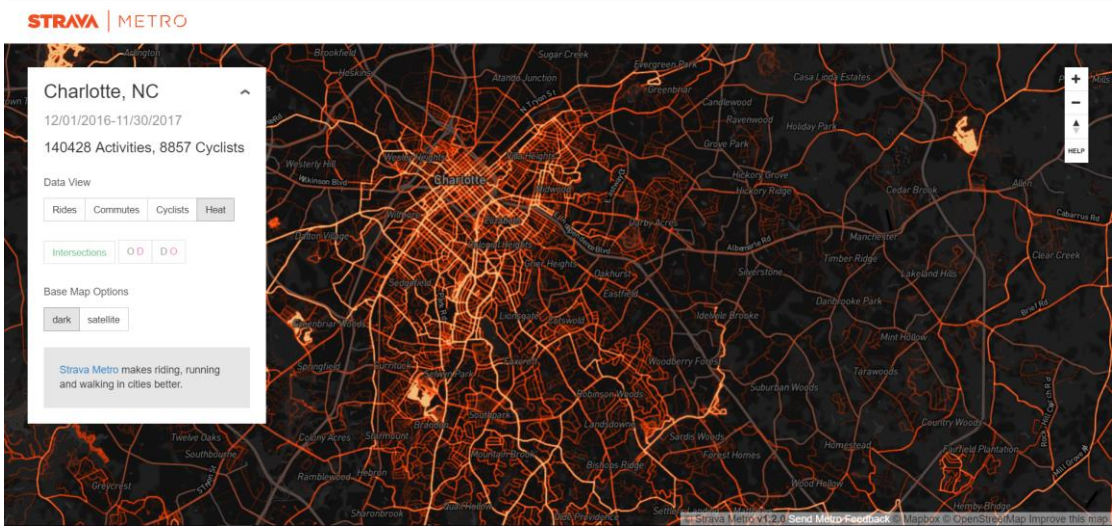


Figure 3.5: Charlotte Heat Map View

3.5 Other supporting data

3.5.1 Bicycle facilities

The bicycle facilities might have a potential impact on the cycling behavior. Therefore, the information on the existing bicycle facilities in the City of Charlotte are collected for the

cycling behavior analysis. A bicycle facility map showing the bike lanes, off street paths, signed bike routes, suggested bike routes, greenways, and the low comfort suggested bike routes can be seen in Figure 3.7. Please note that this map can be found on the following website: http://charlotte.maps.arcgis.com/apps/PanelsLegend/index.html?appid=00e8015ea3e54607a880fe31cc7e2fbf.

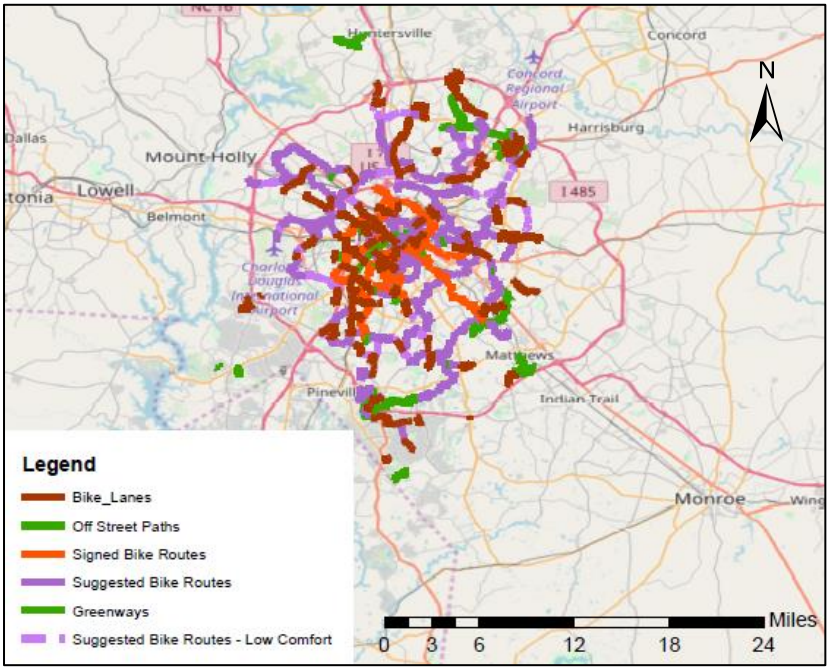


Figure 3.6: Bike Facilities in the City of Charlotte

3.5.2 Population

The population data collected from the US census data set can be seen in Figure 3.8. Please note that this data are found on the following website: http://www.arcgis.com/home/webmap/viewer.html?url=https://services1.arcgis.com/yfahUFAYAdeS5rmM/ArcGIS/rest/services/Enriched%20Enriched%20Charlotte%20Blocks/FeatureServer&source=sd.

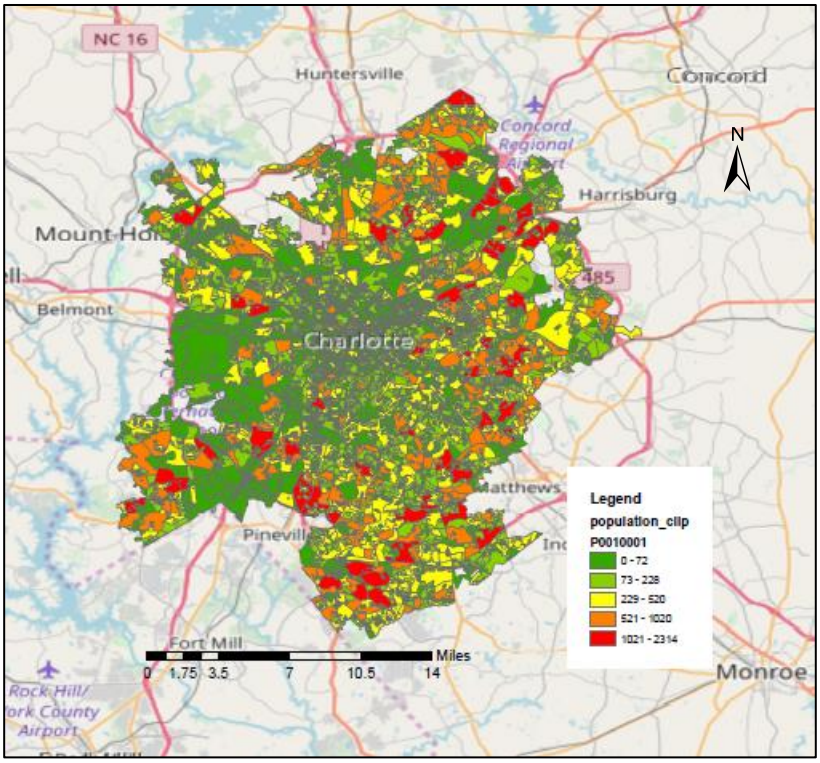


Figure 3.7: Total Population in the City of Charlotte

3.5.3 Slope

The slope cell data shown in Figure 3.9 are collected from the ArcGIS online dataset. Please note that researchers can find this data by adding data from ArcGIS online with "Lidar2017_Slope".

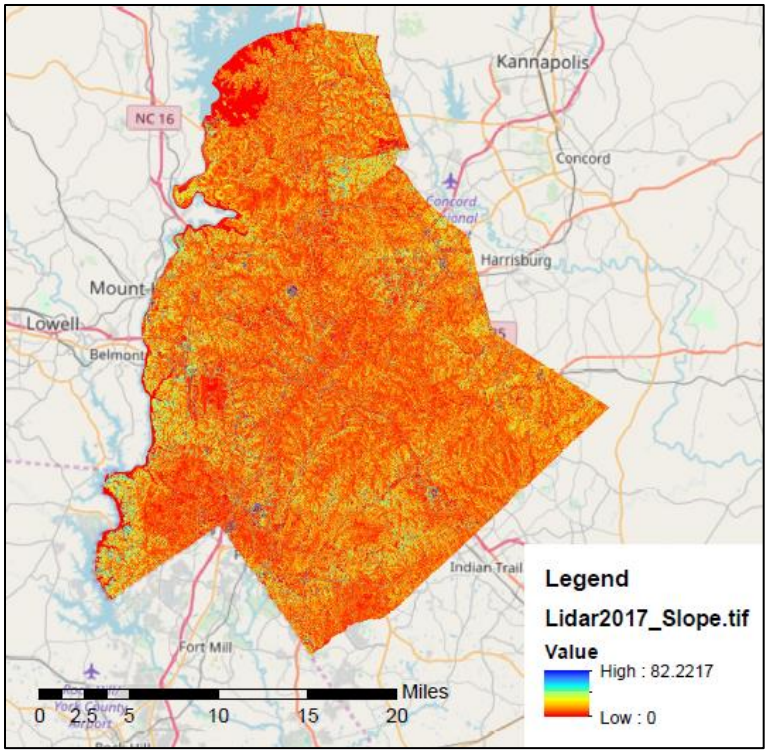


Figure 3.8: Slope in City of Charlotte

3.6 Summary

This chapter shows the data collected for this research such as Strava data that contains cycling information, demographic data, bicycle facility data, and slope data in the City of Charlotte. These data will be utilized in the cycling behavior modeling in the model development chapter.

Chapter 4. Data Descriptive Analyses

4.1 Introduction

This chapter analyzes the crowdsourced bicycle data collected from Strava. Descriptive analyses are conducted based on the Strava data in terms of demographics, trip purposes, bicycle volume for different months, time of day, and day of week, and origin and destination of cycling trips.

The sections in Chapter 4 are organized as follows. Section 4.2 presents the demographic information on Strava users. Section 4.3 describes the cycling trips for different trip purposes. Section 4.4 shows the bicycle count data by different month of year, weekday and weekend, and time of day. Section 4.5 provides the origin/destination information for Strava users' cycling trips. Finally, Section 4.6 concludes the chapter with a summary.

4.2 Demographics

According to the data collected from Strava, there were 8,857 cyclists using Strava applications to record their cycling trips during December 2016 to November 2017 in the City of Charlotte. 140,428 trips were generated by these Strava users.

From the cyclist demographic information report provided by Strava, most of the cyclists are male accounting for 80.49% of the total Strava users in the City of Charlotte. Only 14.91% of the cyclists are female. In addition, 407 cyclists prefer to not present their gender in the application. The number of cyclist counts in the City of Charlotte from December 2016 to November 2017 is presented in Figure 4.1.

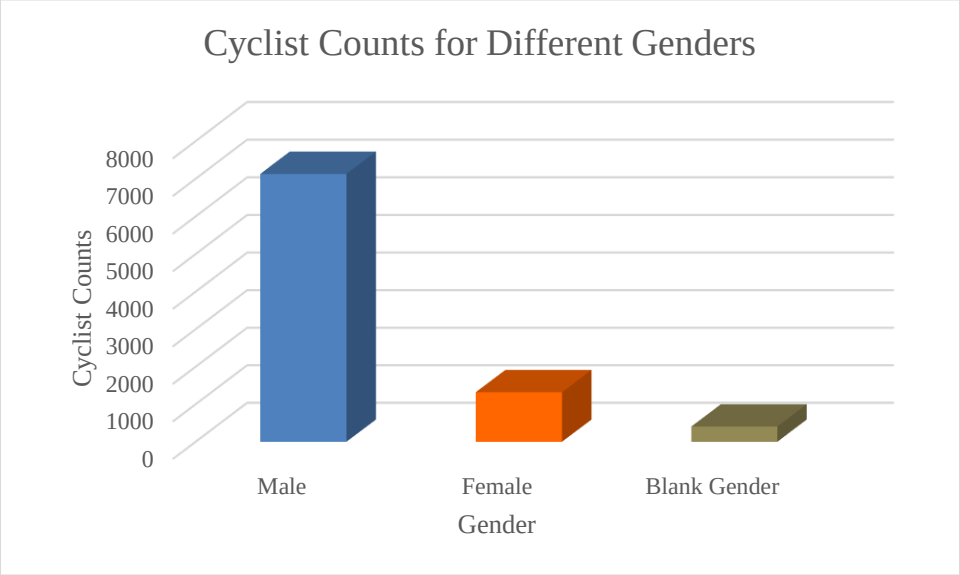


Figure 4.1: Strava User Counts for Different Genders

The number of Strava users from different age groups can be found in the demographic information report. According to the data, the ages of the Strava users range from under 25 to

over 95 which cover both young and old cyclists. The portion of cyclists from different age groups is presented in Figure 4.2. From the figure, it can be seen that the majority of the cyclists are between 25 and 54. Cyclists over 65 are very few. However, there are 1578 cyclists who do not provide their ages to Strava.

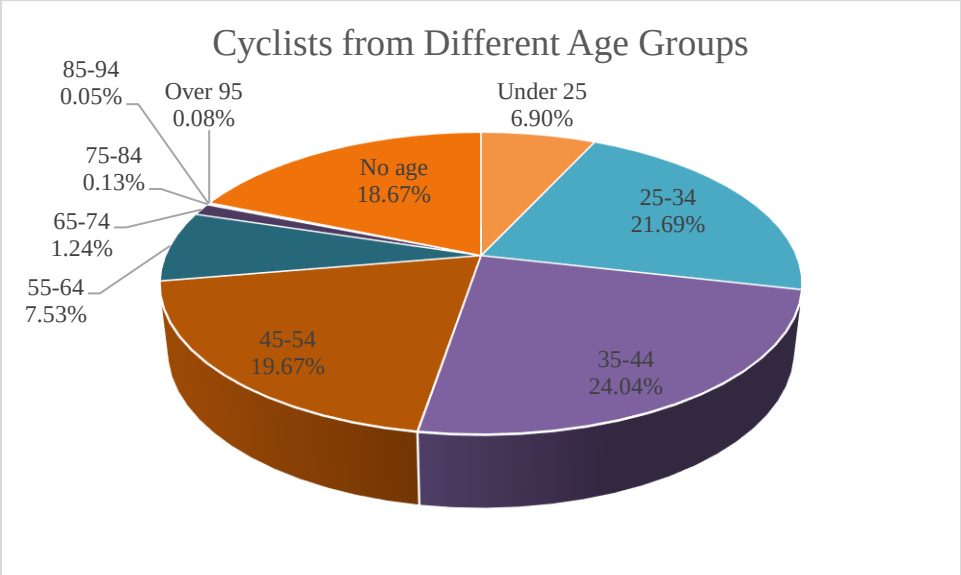


Figure 4.2: Portion of Cyclists from Different Age Groups

4.3 Trip Purpose

The trip purposes of the Strava users are categorized into two parts which are commute trips and non-commute trips. The majority of the cycling trips generated by Strava users are non-commute trips. Figure 4.3 shows the comparison of the number of commute trips and non-commute trips in the City of Charlotte during December 2016 to November 2017. From the figure, it can be seen that the number of commute trips is only 25,737, while the number of non-commute trips is 114,691.

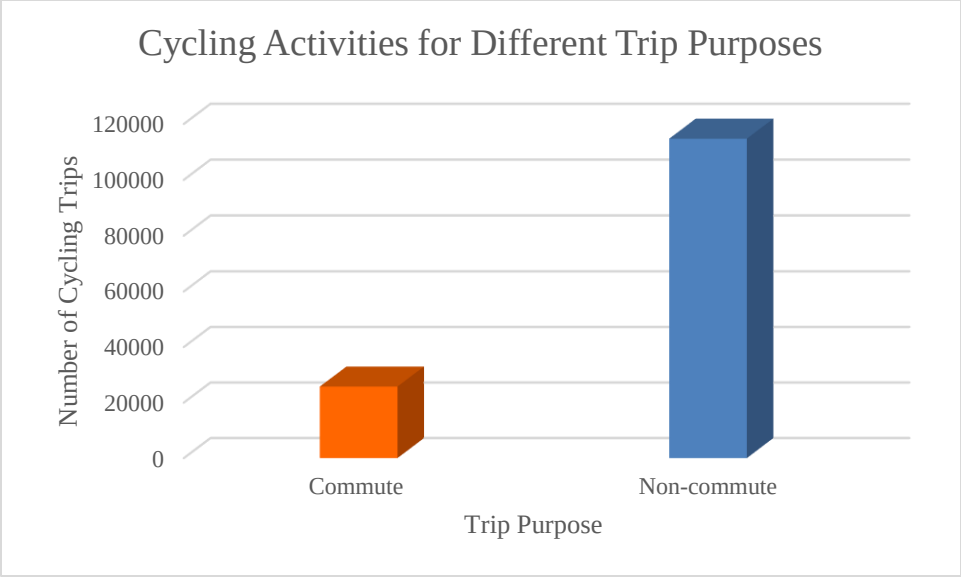


Figure 4.3: Cycling Activities for Different Trip Purposes

To view the distribution of the commute cyclist counts on each road segment in the City of Charlotte, a map is presented in Figure 4.4. From this figure, it can be found that the road segments associated with high bicycle volume are located in the center city where the business district of Charlotte is located. The reason that high volume of commute trips occurred in center city is probably related to the following facts: 1) It is difficult to drive in the center city since there are a lot of one-way roads; 2) The parking issue can be severe in the center city.

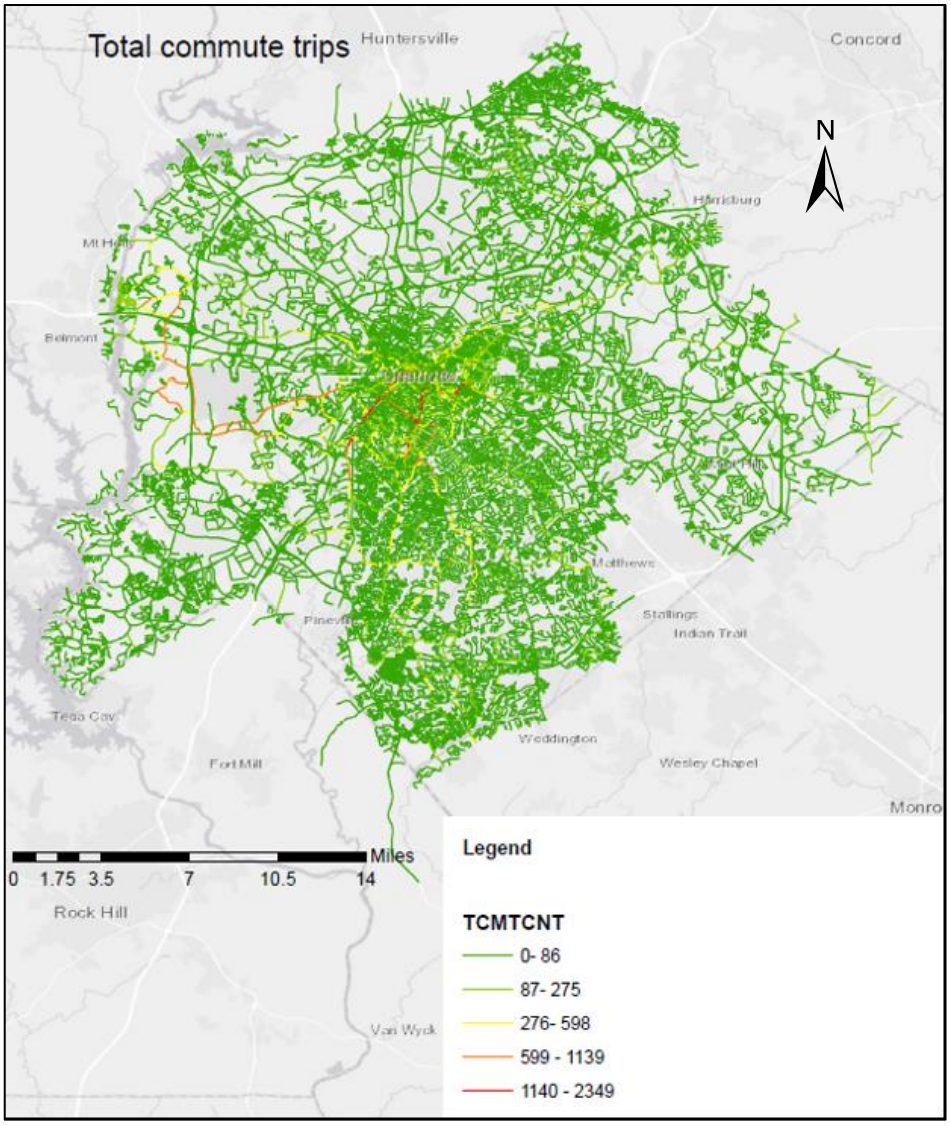


Figure 4.4: Total Commute Trips

4.4 Cyclist Counts

4.4.1 Total Cyclist Counts

To have an overview of the cyclist distribution in the City of Charlotte, a map that presents the number of cyclist counts on each roadway segment from December 2016 to November 2017 is shown in Figure 4.5. It can be seen that most of the road segments have low cyclist counts.

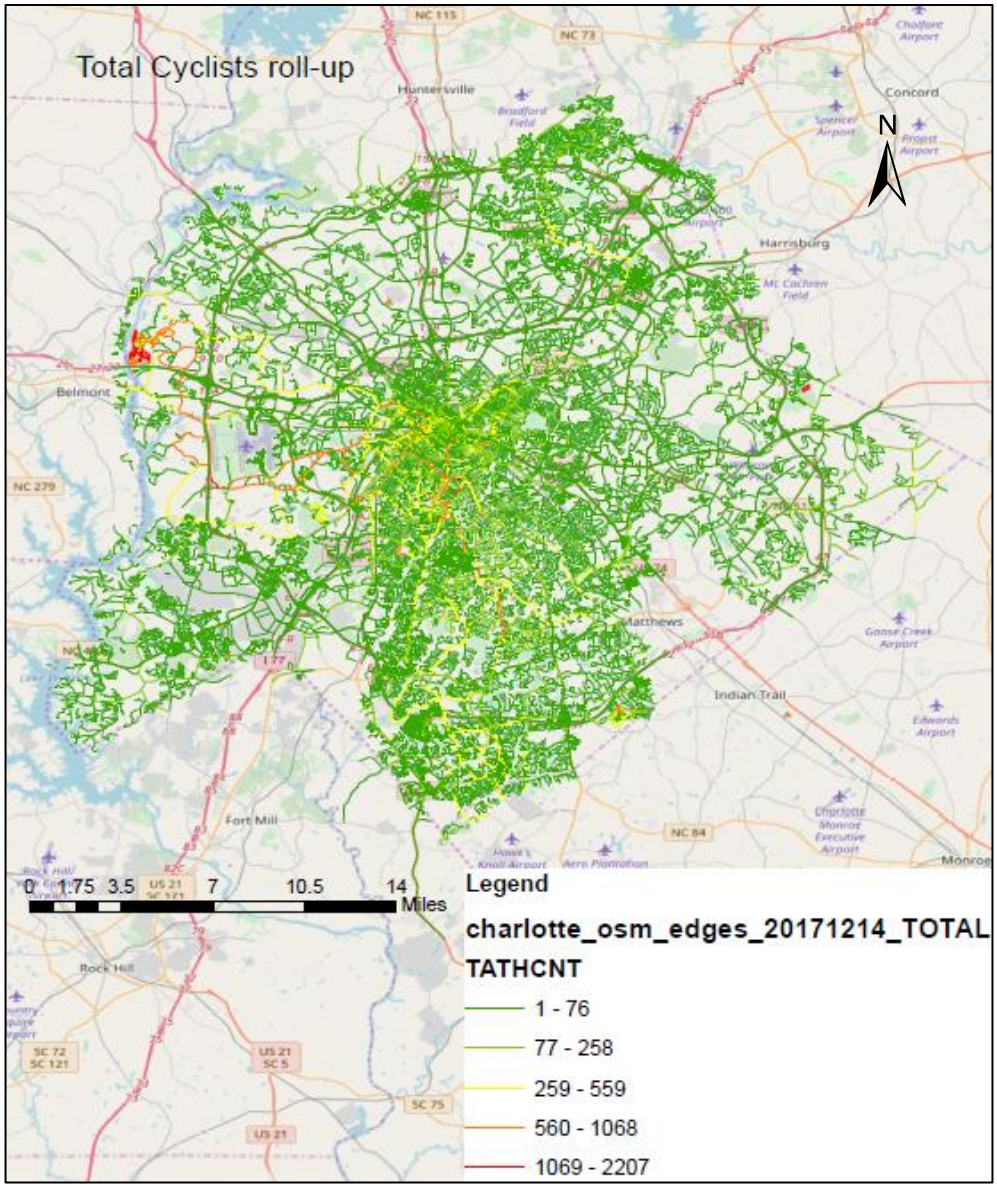
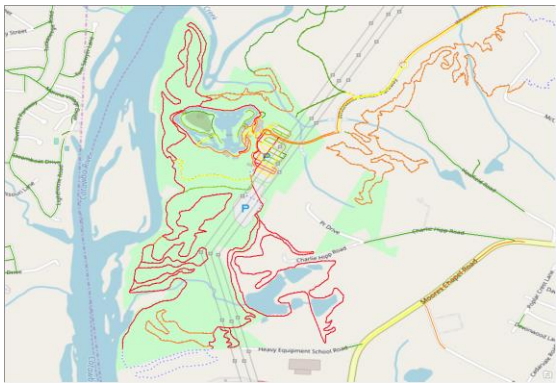


Figure 4.5: Total Cyclist Counts

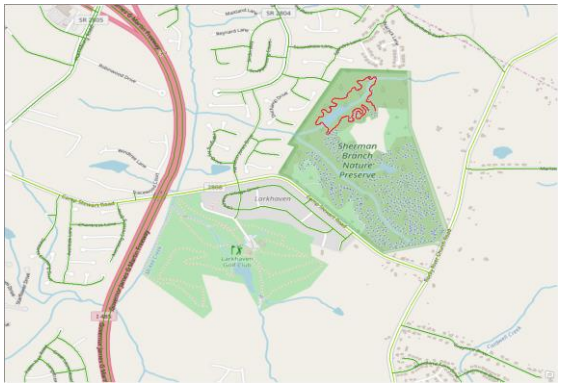
From the total cyclist counts presented in the above figure, four locations with high bicycle volumes are identified and shown in Figure 4.6 which are greenway, school, airport, and park. These locations are popular among Strava users in the City of Charlotte.



4.6.a Greenway

4.6.b School

4.6.c Airport

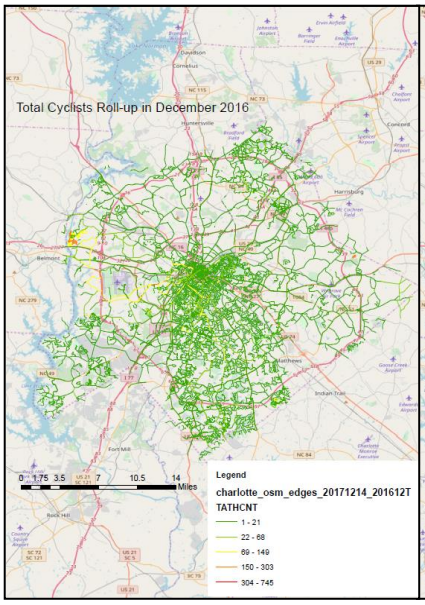


4.6.d Park

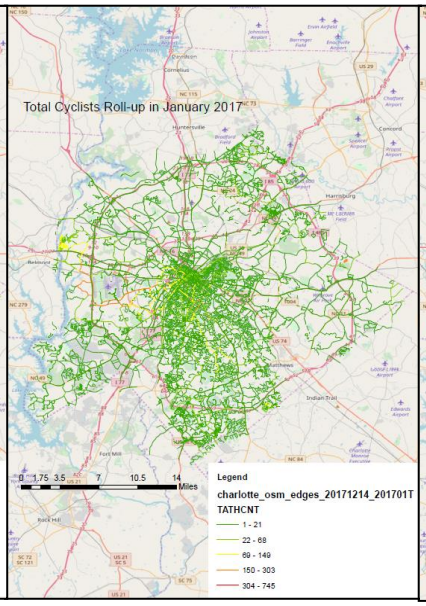
Figure 4.6: Four Popular Cycling Locations

4.4.2 Month of Year

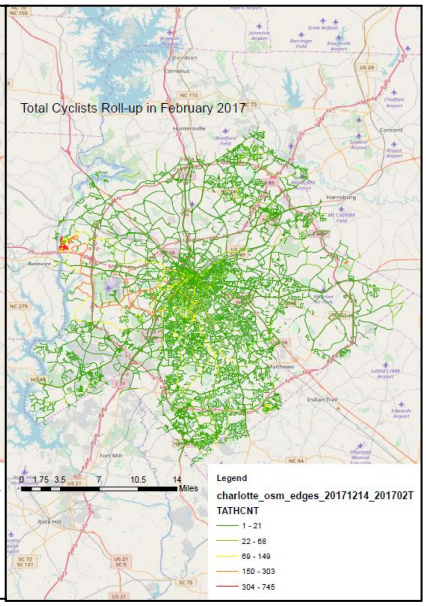
To discover the variation trend of the cyclist distribution in the City of Charlotte from December 2016 to November 2017, maps are created to illustrate the bicyclist counts on each road segment in the whole network in Figure 4.7. Since cycling activities have a strong relationship with the weather condition, the cyclists' behavior for each month of year may vary with the change of temperature and the specific weather condition throughout the year.



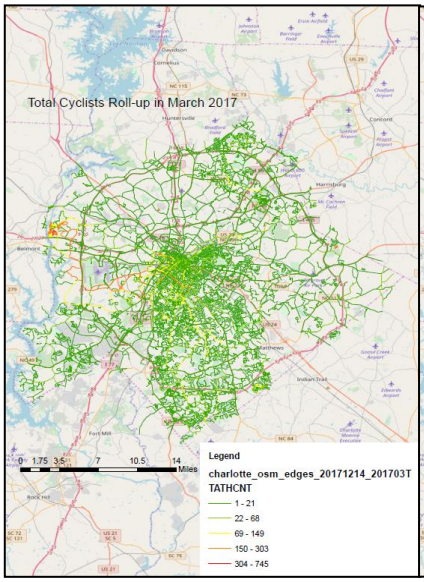
4.7.a December 2016



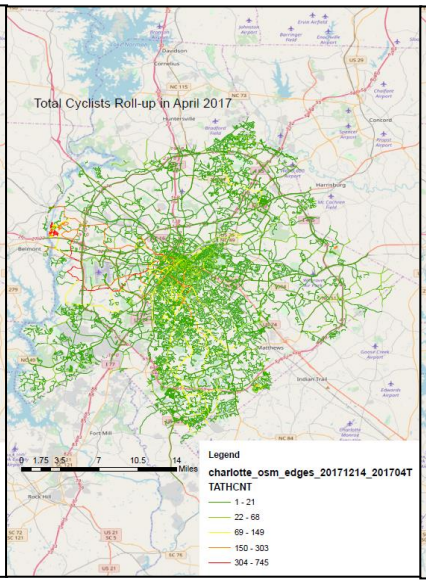
4.7.b January 2017



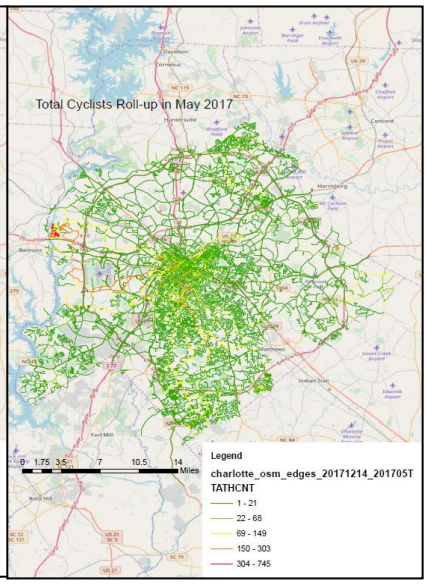
4.7.c February 2017



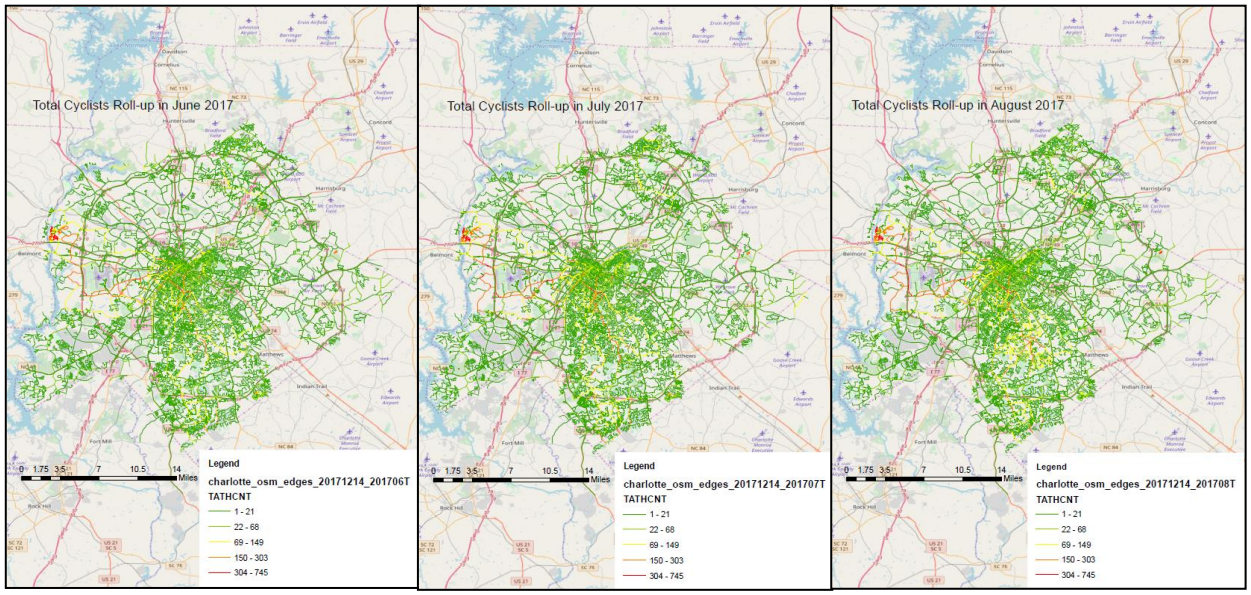
4.7.e March 2017



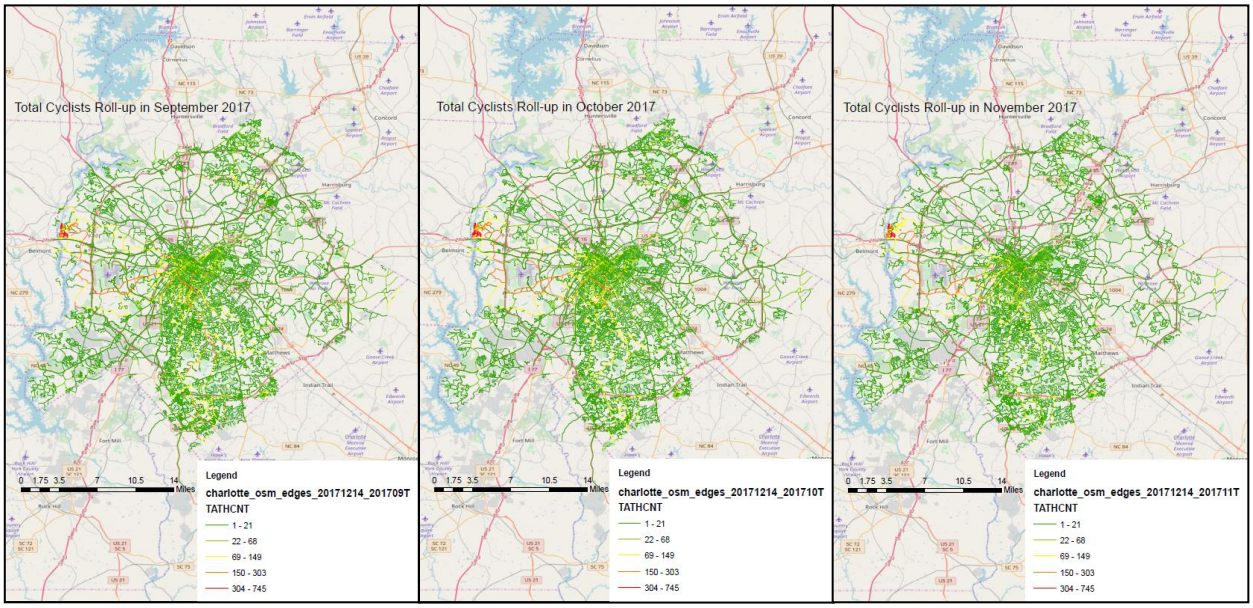
4.7.f April 2017



4.7.g May 2017



4.7.h June 2017 4.7.i July 2017 4.7.j August 2017



4.7.k September 2017 4.7.l October 2017 4.7.m November 2017

Figure 4.7: Total Bicycle Volume in Each Month

The total bicycle volumes in the whole network for each month in the investigation year are presented in Figure 4.8. Comparing the twelve maps shown in Figure 4.7 and the bar chart in Figure 4.8, the characteristics of cycling behavior in twelve months are concluded as follows:

1. The common feature of the cycling behavior over the twelve months is the consistency of the four popular cycling locations which are greenway, school, airport, and park.

2. The actual on-season months for cycling in the City of Charlotte are from April to October. The total bicycle volumes in the whole network are increasing from

December 2016 and reach a peak in July 2017. With the variation of the temperature and weather condition, the total bicycle volumes begin to decrease from July 2017 to November 2017.

3. The variances of the bicycle volumes for different locations in each month are not the same.

4. Greenways are popular among Strava users and the bicycle volume on greenway starts to increase from February and decrease in December. For the uptown area and the roads near airport area, the bicycle volume increases from April and decreases in October. For the bicycle volume in the park, it remains high volume from August to November.

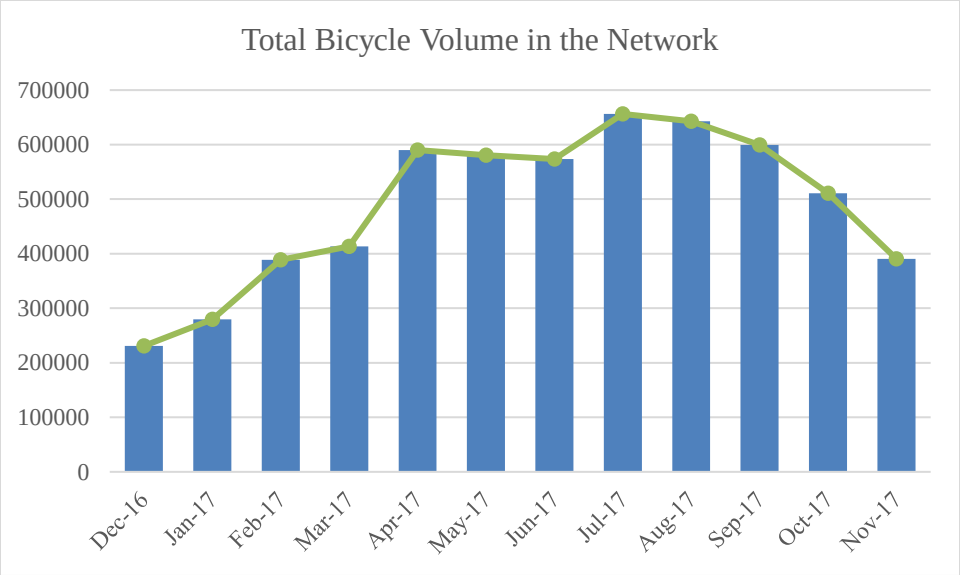


Figure 4.8: Total Bicycle Volume in the Network

4.4.3 Weekdays and Weekends

The cycling activities occurred on weekdays and weekends are different. To see the volume difference between weekdays and weekends on each road segment, a map is generated in Figure 4.9 where red lines represent the higher bicycle volume on weekends and green lines depict the higher volume on weekdays. According to Figure 4.9, the uptown area in the City of Charlotte appears to have more green lines which indicates more weekday cycling trips in this location.

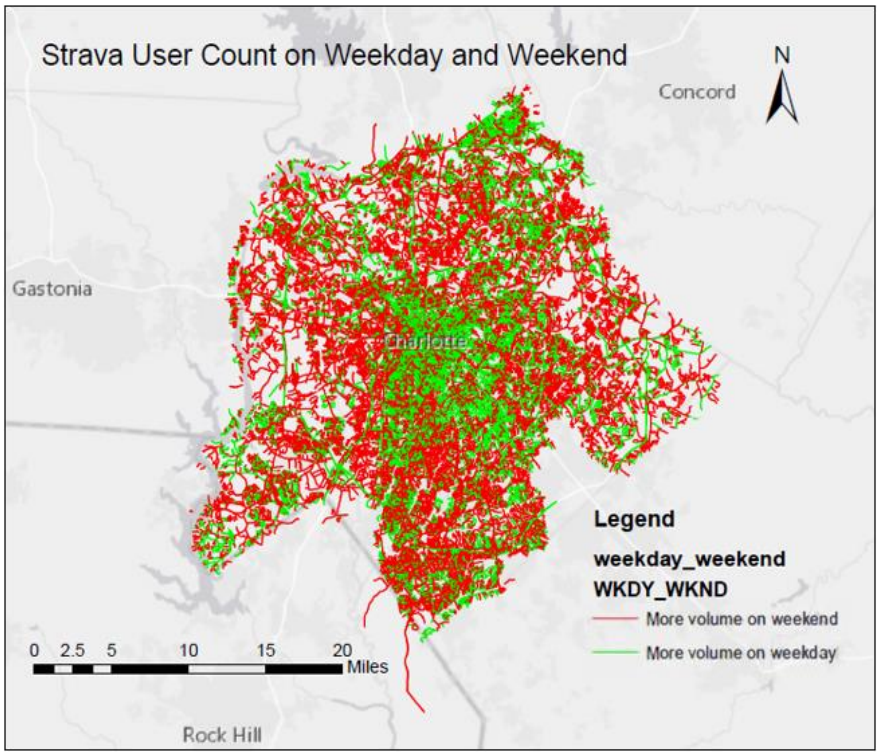


Figure 4.9: Total Bicycle Volume on Weekdays and Weekends

4.4.4 Time of Day

The bicycle volume for each road segment varies with different time of day. The variation of bicycle volume is presented in Figure 4.10. From the figure, one can see that most of the cycling activities occurred from 5 am in the morning to 7 pm in the evening. Two cycling peaks are identified in this figure which are around 8 am and 6 pm. The bicycle volume at 5 am is higher than the volume at 6 am and 7 am. It can be assumed that cyclists choose to bike early in the morning before working hour. There is a decrease in the middle of the day. Two assumptions can be made. First, the temperature around noon is high. Second, workers are busy during the day.

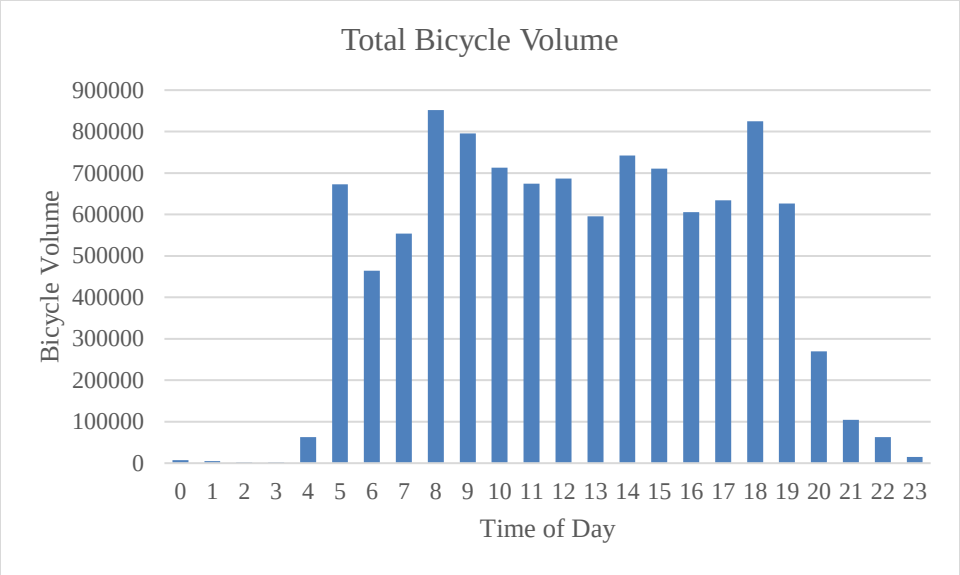


Figure 4.10: Total Bicycle Volume for Different Time of Day

4.5 Origin/Destination

According to the origin and destination data provided by Strava Metro, the total number of unique OD pairs is 23,617. Among these OD pairs, the most popular one is from Polygon ID 2857 back to the same polygon where the parking lot of the US National Whitewater Center is located. The location of this polygon is presented in Figure 4.11. There are multiple greenways around this area, cyclists can drive to park at this location and bike on the greenways nearby.

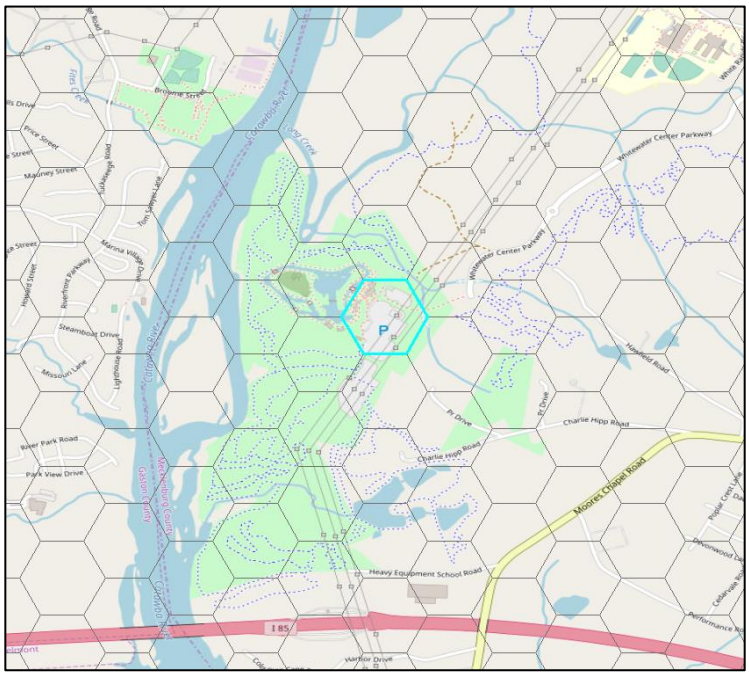


Figure 4.11: The Location of the Most Popular OD Pair

During the investigation year, a total of 2,384 unique bicyclists select to start their trips from the location highlighted in Figure 4.11 and end their trips at the same location. 11,602 cycling trips are generated by these bicyclists which are all non-commute trips. The number of the bicyclists and cycling trips for this OD pair during different time periods can be found in Table 4.1.

Table 4.1 Number of Bicyclists and Trips during Different Time Periods

Time Period	00:00 – 05:59	06:00 – 8:59	09:00 – 14:59	15:00 – 17:59	18:00 – 19:59	20:00 – 23:59
Number of Bicyclists	1	395	1867	1097	482	28
Number of Cycling Trips	1	763	5750	3862	1167	59

According to the table above, the numbers of bicyclists and cycling trips vary with different time periods. Most of the bicyclists select to bike from 9 am to 6 pm. The variation is shown in Figure 4.12 and the portion of cycling trips occurred in each time period is presented in Figure 4.13. According to Figure 4.12, both the numbers of bicyclists and cycling trips increase from 00:00 to 15:00 and then begin to decrease. From Figure 4.13, nearly half of the cycling trips occurred from 9 am to 3 pm. Only 7.09% trips occurred before 9 am and after 8 pm.

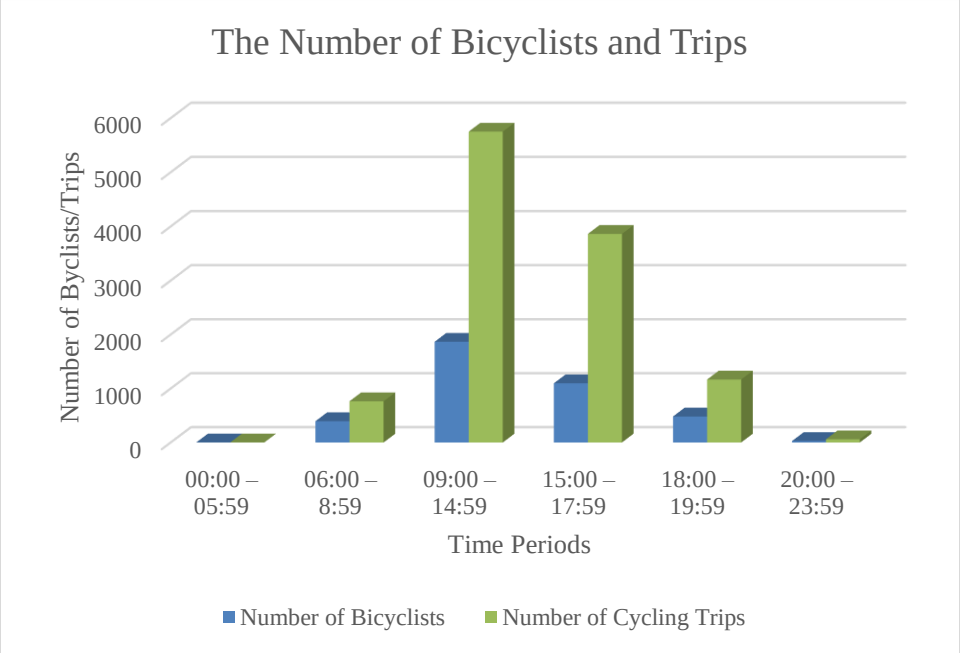


Figure 4.12: The Variation of the Bicyclist/Trip Number for Different Time Periods

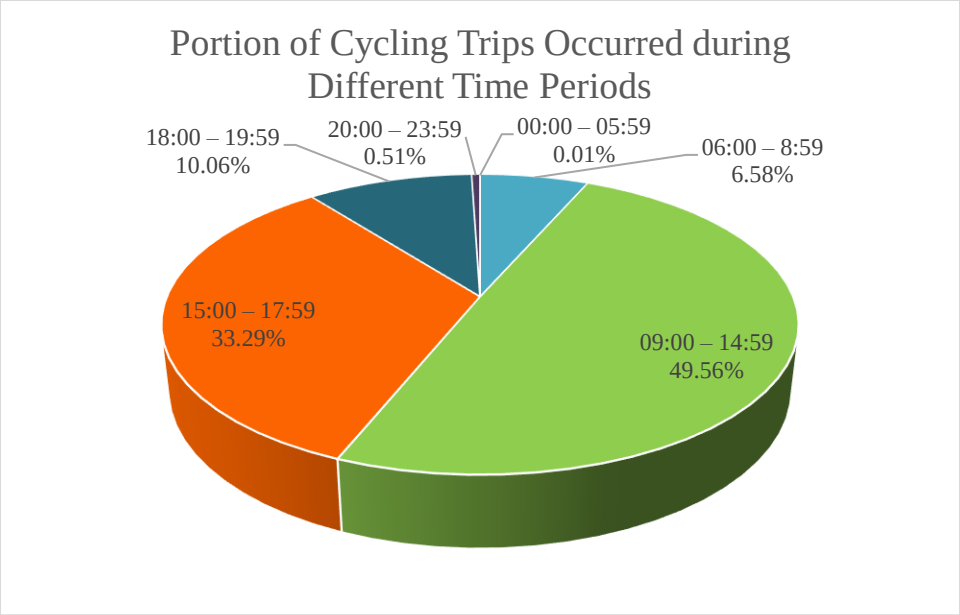


Figure 4.13: The Portion of Cycling Trips Occurred in Each Time Period

Comparing the commute trips and non-commute trips, most of the commute trips with the same OD pair are generated by a unique bicyclist, while several non-commute trips occurred at popular locations are generated by multiple bicyclists. That is to say, the commuters have distinctive commute trips and the non-commuters have similar recreational trips.

The detailed analyses based on the total cyclist counts, total commute counts, and the activity counts on weekdays and weekends in each origin and destination polygon are presented in the following sections.

4.5.1 Total Cyclist Counts

To have an overview of the origins selected by the Strava users in the City of Charlotte from December 2016 to November 2017, a map that illustrates the number of cyclists in each origin polygon is presented in Figure 4.14. It can be seen that the majority of the preferred origins are located in the center city, the Renaissance Park near airport, around the US National Whitewater Center (the western part of the city), Colonel Francis J. Beatty Regional Park (the southern part of the city), and the Sherman Branch Nature Preserve (the eastern part of the city) surrounded by multiple greenways. Most of the trips started in the center, northern and southern parts of the city.

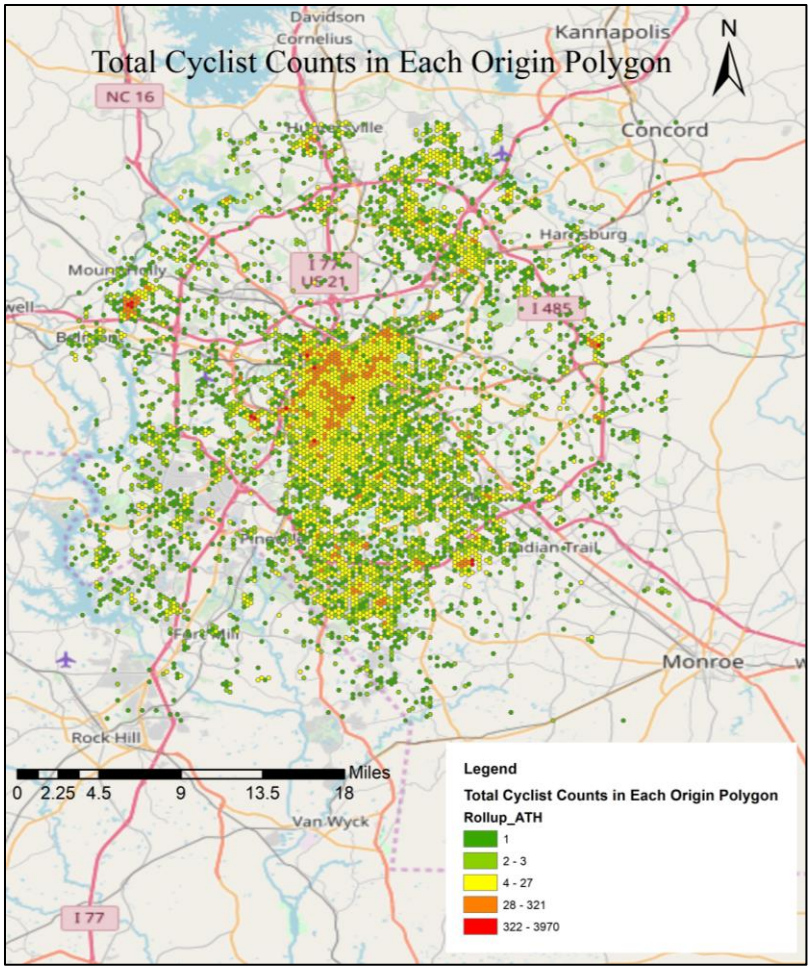


Figure 4.14: Total Cyclist Counts in Each Origin Polygon

Similarly, the total cyclist counts in each destination polygon in the City of Charlotte are shown in Figure 4.15. Comparing the cyclist counts in the origin and destination polygons, the locations of destination polygons associated with high number of cyclist counts remain the same as the locations of the preferred origin polygons.

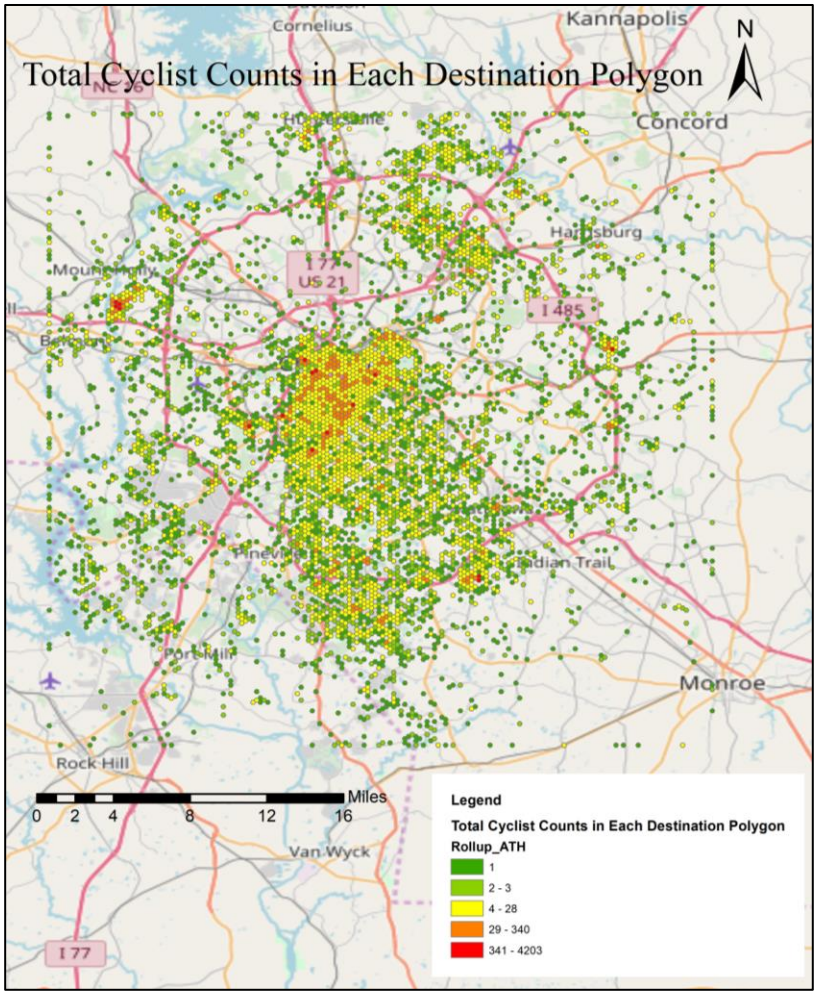


Figure 4.15: Total Cyclist Counts in Each Destination Polygon

4.5.2 Total Commute Counts

To see the difference of popular origin and destination locations between the total trips and the commute trips, the total commute counts within each origin and destination polygon are aggregated and presented in the following figures.

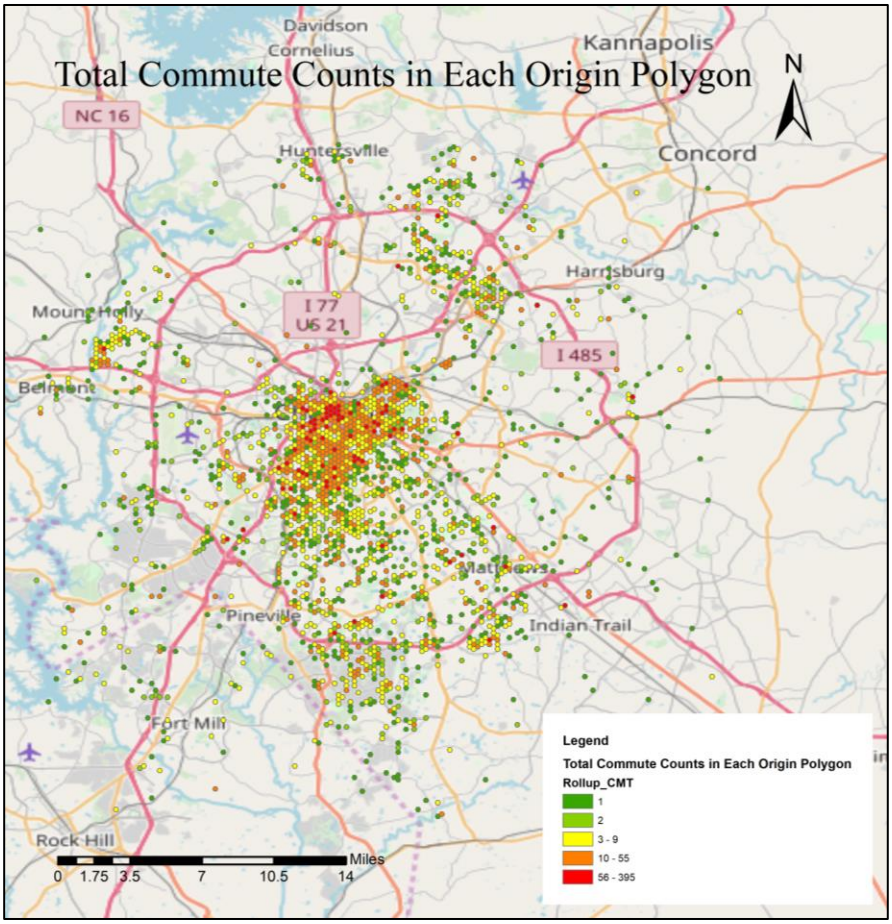


Figure 4.16: Total Commute Counts in Each Origin Polygon

In Figure 4.16, most of the commute trips start from the center city. Compared to the total cyclist counts in each origin polygon containing both commute and non-commute trips, the locations near parks are no longer associated with high number of commute origins. Most of the origin polygons concentrate in the uptown area, and some others spread out in the northern and southern parts of the city.

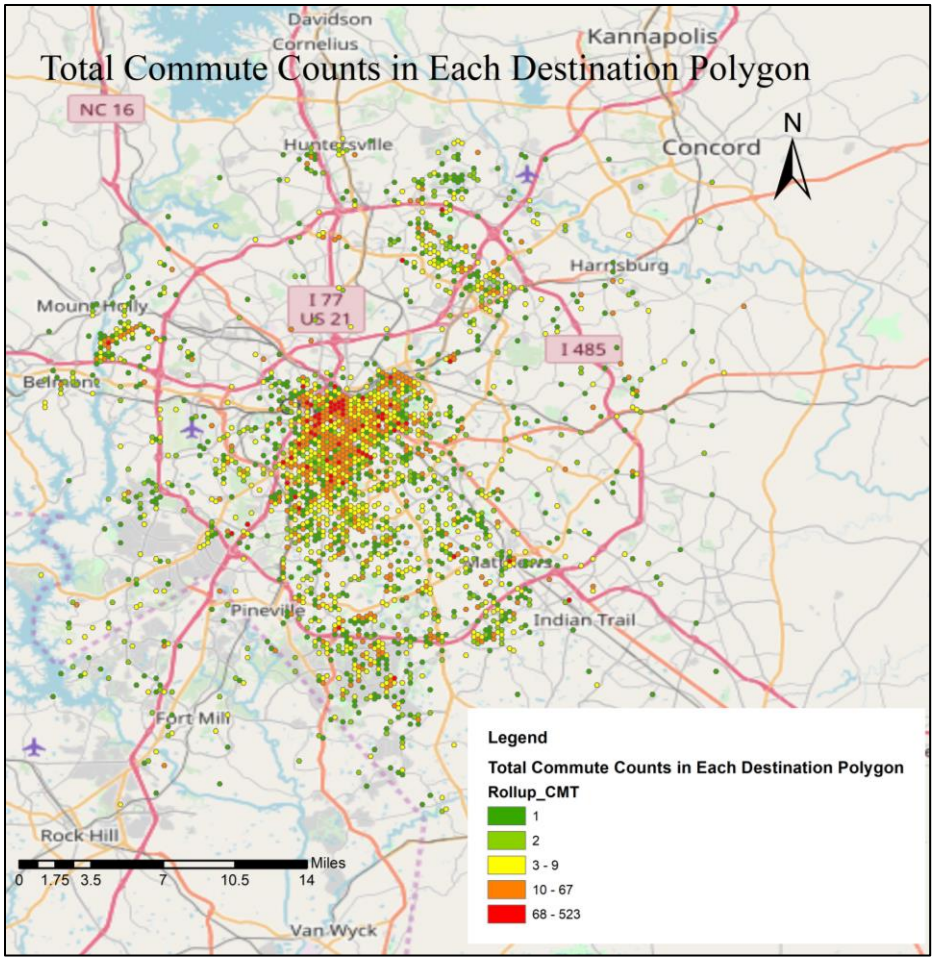


Figure 4.17: Total Commute Counts in Each Destination Polygon

Similar result can be found in Figure 4.17. Most of the destination polygons associated with high commute counts are located in the center city. Other selected destinations are spread out in the southern and northern parts of the city.

4.5.3 Total Activity Counts on Weekdays and Weekends

The cycling activities occurred on weekdays and weekends can be different. In order to discover the differences of the preferred origins and destinations for cycling trips on weekdays and weekends, two figures demonstrating the comparison of total activity counts on weekdays or weekends for each origin and destination polygon are presented in Figure 4.18 and Figure 4.19.

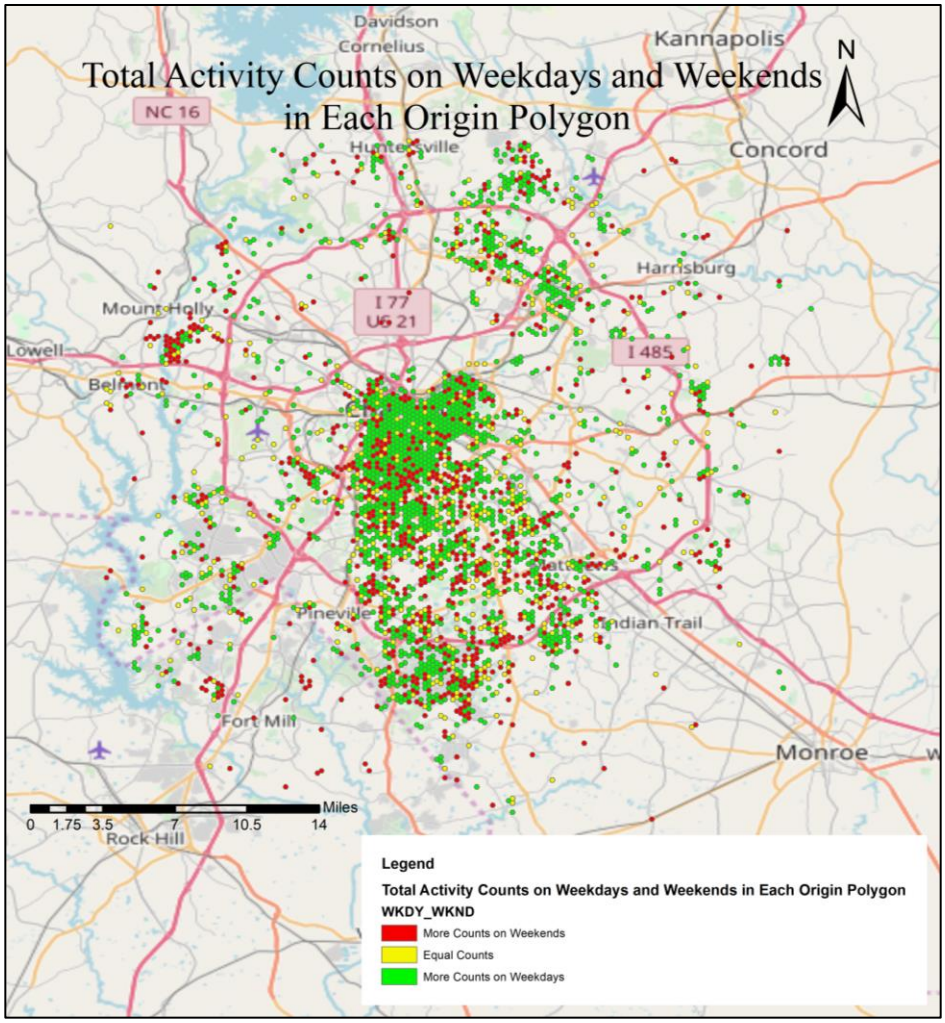


Figure 4.18: Total Activity Counts on Weekdays and Weekends in Each Origin Polygon

In Figure 4.18 and Figure 4.19, red polygons indicate more activity counts on weekends, green polygons represent more activity counts on weekdays, and yellow polygons demonstrate equal counts. According to Figure 4.18, more cycling trips start at the center and northern part of the city on weekdays, and more cycling activities start at the locations near parks on weekends. This result is consistent with the analysis based on the commute and non-commute trips.

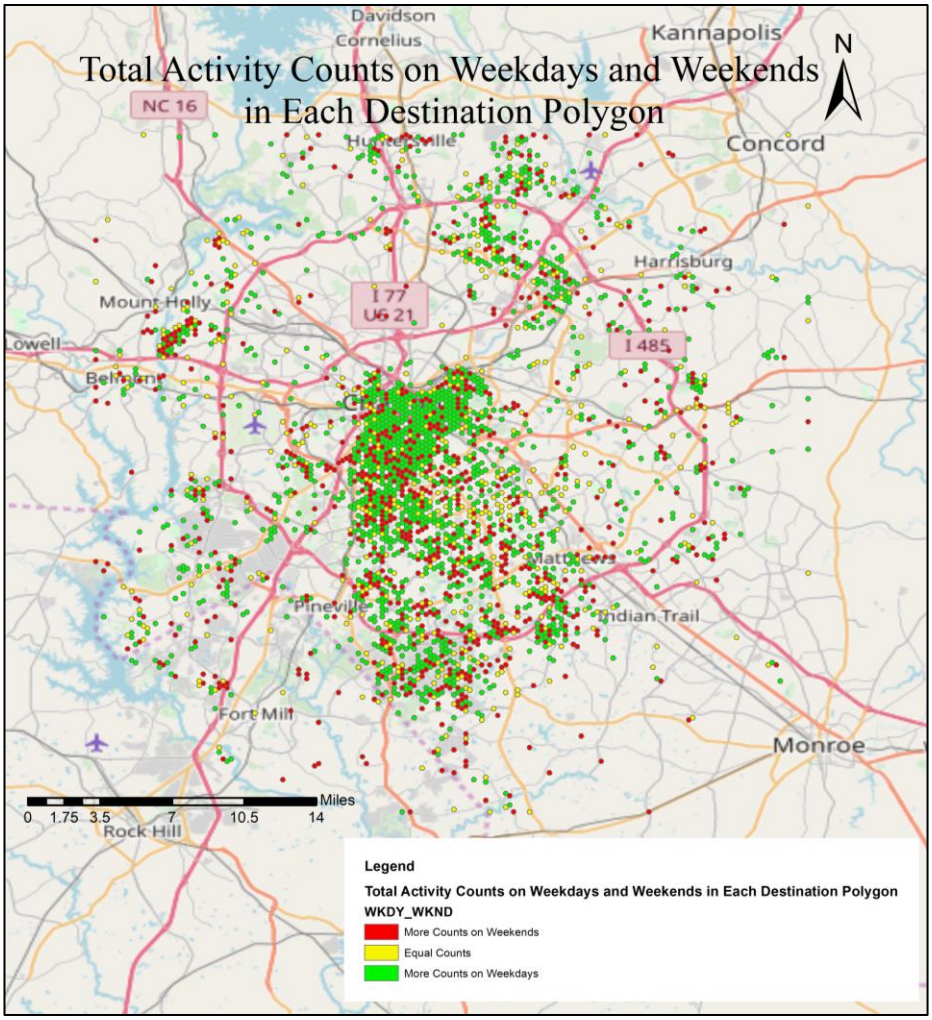


Figure 4.19: Total Activity Counts on Weekdays and Weekends in Each Destination Polygon

Similar result can be found in Figure 4.19. The destination polygons associated with higher activity counts on weekdays are located in the center city. Compared to Figure 4.18, more destination polygons with higher activity counts on weekends occurred in the center city.

4.6 Summary

This chapter provides the descriptive analyses based on the crowdsourced bicycle data collected from Strava Metro. Demographic information on the Stava users and the trip purposes are analyzed based on the delivery report. Link-based bicycle counts in different month of year, on both weekdays and weekends, and for commute and non-commute trips are presented in several heatmaps. In addition, cycling information regarding different OD pairs is also provided.

Chapter 5. Modeling Link-based Cyclist Route Choice Behavior

5.1 Introduction

This chapter is based on the 2018 USDOT Project 03 "Evaluating the Potential Use of Crowdsourced Bicycle Data in North Carolina". The data processing method is adopted from the procedure provided in Chapter 6 of the 2018 USDOT Project report. Discrete choice models are developed to analyze the link-based cyclist route choice behavior and model comparison is conducted to identify the best fit for this behavior analysis. To examine the different impacts of explanatory variables on link-based route choice during selected time periods, discrete choice models are developed separately.

The following sections are organized as follows. Section 5.2 through Section 5.5 provide the models developed for link-based cyclist route choice behavior including ordered logit (ORL) model, partial proportional odds (PPO) model, multinomial logit (MNL) model, and mixed logit (MXL) model respectively. Section 5.6 compares the models developed in the previous sections and identifies the best model structure for this research study. Section 5.7 develops two models for different selected time periods. Model result comparison is provided in this section. Finally, Section 5.8 concludes this chapter with a summary.

5.2 Ordered Logit Model

5.2.1 ORL Model Structure

The ordered logit model is one of the traditional discrete choice models that is utilized for ordinal dependent variable analysis. In this research study, the number of bicycle counts for each road segment is divided into five categories which are low (0-39), low-average (40-79), average (80-119), high-average (120-159), and high (160-200). In the ORL model, the level of bicycle counts on a road segment is denoted as y_i which is associated with the latent variable y_i^*. The model specification is presented as follows:

$$y_i^* = \beta X_i + \varepsilon_i \qquad \text{Eq. (1)}$$

where y_i^* demonstrates the latent bicycle volume, X_i denotes a vector of the explanatory variables contributing to the bicycle volume, β represents the coefficients that will be estimated, and ε_i stands for the error term which is Gumbel distributed.

In this research study, the continuous latent variable y_i^* is divided by the cut-points θ_j (j = 1, 2, …, *J*) into *J* intervals (J = 5 for this scenario) and the bicycle volume is shown as follows:

$$y_i = \begin{cases} 1, -\infty \le y_i^* \le \theta_1 \\ 2, \theta_1 < y_i^* \le \theta_2 \\ 3, \theta_2 < y_i^* \le \theta_3 \\ 4, \theta_3 < y_i^* \le \theta_4 \\ 5, \theta_4 < y_i^* \le +\infty \end{cases} \qquad \text{Eq. (2)}$$

Thus, the probability of the level of bicycle counts on each road segment can be presented as follows:

$$P_i(j) = \begin{cases} \mathrm{F}(\theta_1 - \beta_j X_i), j = 1 \\ \mathrm{F}(\theta_j - \beta_j X_i) - \mathrm{F}(\theta_{j-1} - \beta_j X_i), j = 2, \dots, j-1 \\ 1 - \mathrm{F}(\theta_{J-1} - \beta_j X_i), j = J \end{cases} \quad \text{Eq. (3)}$$

where F(.) represents the cumulative standard logistic distribution function.

5.2.2 ORL Model Results

To analyze the level of bicycle counts on each road segment and examine the factors affecting the link-based route choice behavior of the bicyclists in the City of Charlotte, an ordered logit model is developed. Explanatory variables are carefully selected for this ORL model which include temporal variables, road characteristics, sociodemographic information, geometry, and bicycle facilities. The detailed variable description is presented in Table 5.1.

Table 5.1 Explanatory Variable

Variable	**Description**
Temporal Variables	
Hour_0	If cycling time is during 00:00-05:59, then Hour_0 = 1.
Hour_1	If cycling time is during 06:00-08:59, then Hour_1 = 1.
Hour_2	If cycling time is during 09:00-14:59, then Hour_2 = 1.
Hour_3	If cycling time is during 15:00-17:59, then Hour_3 = 1.
Hour_4	If cycling time is during 18:00-19:59, then Hour_4 = 1.
Hour_5	If cycling time is during 20:00-23:59, then Hour_5 = 1.
Weekday	If bike on a weekday, then weekday = 1.
Road Characteristics	
Speed Limit	The posted speed limit on a roadway segment.
RouteClass1	Interstate
RouteClass2	US route
RouteClass3	NC route
RouteClass4	Secondary route
MPLength	The length of the segment in miles.
ThruLaneCo	The number of through lanes.
Oneway	If the road segment is one way, then oneway = 1
Sociodemographic Characteristics	

Variable	Description
TOTPOP_CY	Total population in each census block.
MEDAGE_CY	The median age in each census block.
MEDHINC_CY	Median household income in each census block.
Total_Hous	Total households in each census block.
TotalFamil	Total families in each census block.
FamilyPove	Family poverty rate in each census block.
Geometry	
Slope	The slope of a road segment at intersection.
Bicycle Facilities	
B_offstree	Off street paths
B_bikelane	Bike lanes
B_signedbi	Signed bike lanes
B_suggeste	Suggested bike routes
B_suggest0	Suggested bike routes with low comfort
B_greenway	Greenway

All the factors presented in Table 5.1 are included in the ordered logit model to determine the probability of each segment being selected by the Strava users. The maximum likelihood estimation method is utilized to estimate the model parameters and the thresholds in the ordered logit model. This process is conducted in SAS 9.4. To keep the variables that have a significant impact on the level of bicycle counts on each road segment, the backward selection demand is used in the model estimation procedure. A summary of the backward selection results is presented in Table 5.2. The model estimation results, and the fit statistics are shown in Table 5.3 and Table 5.4 respectively.

Table 5.2 Summary of Backward Elimination

Summary of Backward Elimination				
Step	**Effect Removed**	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
1	**B_Hour_0**	1	0.0000	0.9993
2	**B_offstree**	1	0.0000	0.9951
3	**B_Hour_4**	1	0.0027	0.9586
4	**SpeedLimit**	1	0.1222	0.7266
5	**FamilyPove**	1	0.4548	0.5001

Summary of Backward Elimination				
Step	**Effect Removed**	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
6	**TOTPOP_CY**	1	0.4030	0.5255
7	**B_bikelane**	1	0.6974	0.4037

Table 5.3 Ordered Logit Model Estimation Results

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**5**	1	1.6165	1.0468	2.3847	0.1225
Intercept	**4**	1	3.6935	1.0534	12.2937	0.0005
Intercept	**3**	1	4.0366	1.0565	14.5970	0.0001
Intercept	**2**	1	5.4232	1.0882	24.8353	<.0001
B_weekday		1	-4.2510	0.3204	176.0312	<.0001
B_Hour_1		1	1.0789	0.4326	6.2192	0.0126
B_Hour_2		1	1.1850	0.4193	7.9859	0.0047
B_Hour_3		1	2.9484	0.4137	50.7871	<.0001
MPLength		1	1.0827	0.4673	5.3673	0.0205
ThruLaneCo		1	0.6786	0.0853	63.2215	<.0001
MEDAGE_CY		1	0.0244	0.0115	4.4958	0.0340
MEDHINC_CY		1	0.000032	2.773E-6	129.7401	<.0001
Total_Hous		1	0.00119	0.000345	11.8828	0.0006
TotalFamil		1	-0.00133	0.000470	8.0179	0.0046
Slope		1	-0.0506	0.00959	27.8175	<.0001
B_signedbi		1	-1.1172	0.1814	37.9421	<.0001
B_suggeste		1	0.7100	0.3414	4.3260	0.0375
B_suggest0		1	-1.8420	0.3542	27.0457	<.0001
B_greenway		1	2.6567	1.0285	6.6720	0.0098
RouteClass1		1	-0.6356	0.2719	5.4624	0.0194
RouteClass2		1	0.8828	0.2390	13.6409	0.0002
RouteClass3		1	-0.3567	0.1395	6.5407	0.0105
Oneway		1	0.9971	0.1553	41.2258	<.0001

Table 5.4 Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	7480.648	5802.726
SC	7522.162	6114.085
-2 Log L	7472.648	5742.726

According to the backward elimination summary in Table 5.2, variables including time period from 00:00 to 05:59 and from 18:00 to 19:59, speed limit, off street paths, bike lanes, speed limit, total population, and family poverty rate do not have a significant impact on the level of bicycle counts on each road segment. Based on the model estimation results presented in Table 5.3, variables including weekday, total family, slope, signed bike lanes, suggested bike routes with low comfort, interstate route, and NC route all have a negative impact on the level of bicycle counts, while other variables which are time period from 6:00 to 17:59, segment length, number of through lanes, median age, median household income, total household, suggested bike routes, greenway, US route, and one-way road all have a positive impact on the level of bicycle counts. The detailed interpretation of the impact of each factor on the level of bicycle counts will be provided in Section 5.6. AIC and -2LogL presented in Table 5.4 are indicators that measure the fitness of the model which will be used for model comparison in Section 5.6.

5.3 Partial Proportional Odds Model

5.3.1 PPO Model Structure

The partial proportional odds model is developed based on the ordered logit model. In ordered logit model, the proportional odds (PO) assumption is subjected. It can be interpreted that the estimated parameters are restricted to be same across all the alternatives. However, this assumption is unrealistic. To relax the assumption, the PPO model is developed.

The explanatory variables associated with each road segment are categorized into two groups. One contains parameters satisfying the PO assumption, which is presented as vector X_i, the other includes parameters that violate the PO assumption which is shown as vector Z_i. The variables that violate the PO assumption are able to affect the response variables differently, while others remaining fixed parameters have the same effect across different levels. Thus, the PPO model with logit function is presented as follows:

$$P(Y_i \geq j) = \frac{\exp[\theta_j - (X_i'\beta_j + Z_i'\gamma_j)]}{1+\exp[\theta_j - (X_i'\beta_j + Z_i'\gamma_j)]} \quad \text{Eq. (4)}$$

where j denotes the level of bicycle counts on each road segment and Y_i represents the bicycle counts for road segment i, β and γ represents the coefficients that will be estimated, and θ_j demonstrates the threshold for jth cumulative logit.

To examine whether the explanatory variables violate the PO assumption or not, the Wald Chi-square tests are utilized during the model development. This procedure helps divide the explanatory variables into two groups which belong to either vector X_i or vector Z_i.

5.3.2 PPO Model Results

This PPO model is built based on the ORL model developed in Section 5.2. A series of Wald Chi-square are conducted to test the explanatory variables that violate the PO assumption. These variables are presented in Table 5.5.

Table 5.5 Linear Hypotheses Testing Results

Label	Wald Chi-Square	Pr > ChiSq
Hour_1_po	38.4832	<.0001
ThruLaneCo_po	10.1651	0.0172
MEDHINC_CY_po	33.7202	<.0001
Total_hous_po	25.5679	<.0001
TotalFamil_po	37.5464	<.0001
B_suggeste_po	12.4505	0.0060
RouteClass2_po	27.5757	<.0001
oneway_po	17.0930	0.0007

Thus, variables including time period from 6 am to 9 am, the number of through lanes, median household income, total households, total families, suggested bike routes, US routes, and one-way road violate the PO assumption and have different effects across different levels.

The PPO model estimation results and the fit statistics are presented in Table 5.6 and Table 5.7.

Table 5.6 Partial Proportional Odds Model Estimation Results

Analysis of Maximum Likelihood Estimates					
Parameter	**Level**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	5	2.9121	2.1919	1.7651	0.1840
Intercept	4	8.2183	1.2527	43.0387	<.0001
Intercept	3	9.9807	5.1830	3.7081	0.0541
Intercept	2	10.7126	1.5216	49.5631	<.0001
Weekday		-7.0154	1.2122	33.4937	<.0001
Hour_1	5	-0.2021	0.1664	1.4750	0.2246

Analysis of Maximum Likelihood Estimates					
Parameter	**Level**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Hour_1	4	3.1647	0.5676	31.0867	<.0001
Hour_1	3	0.3418	1.7064	0.0401	0.8412
Hour_1	2	-0.0473	2.3269	0.0004	0.9838
Hour_3		1.7205	0.1034	276.6263	<.0001
ThruLaneCo	5	0.5160	0.0711	52.7303	<.0001
ThruLaneCo	4	-0.2532	0.2544	0.9905	0.3196
ThruLaneCo	3	-0.1234	0.4373	0.0796	0.7778
ThruLaneCo	2	-0.5763	1.2314	0.2190	0.6398
MEDHINC_CY	5	0.000031	2.66E-6	138.3050	<.0001
MEDHINC_CY	4	0.000034	8.519E-6	15.7006	<.0001
MEDHINC_CY	3	0.000154	0.000022	50.5355	<.0001
MEDHINC_CY	2	0.000109	0.000035	9.8945	0.0017
Total_Hous	5	0.00105	0.000334	9.8621	0.0017
Total_Hous	4	0.00859	0.00280	9.4126	0.0022
Total_Hous	3	0.0277	0.00534	26.9469	<.0001
Total_Hous	2	0.0373	0.0206	3.2672	0.0707
TotalFamil	5	-0.00120	0.000458	6.8452	0.0089
TotalFamil	4	-0.0122	0.00377	10.4502	0.0012
TotalFamil	3	-0.0389	0.00617	39.7655	<.0001
TotalFamil	2	-0.0530	0.0253	4.3881	0.0362
Slope		-0.0575	0.00891	41.5729	<.0001
B_signedbi		-1.0671	0.1841	33.6052	<.0001
B_suggeste	5	2.8458	0.9343	9.2777	0.0023
B_suggeste	4	2.8330	1.1958	5.6128	0.0178
B_suggeste	3	-3.4416	1.9230	3.2029	0.0735
B_suggeste	2	-0.4743	2.3583	0.0404	0.8406
B_suggest0		-4.0556	0.9381	18.6881	<.0001
B_greenway		3.5327	1.4672	5.7973	0.0161
RouteClass2	5	1.4188	0.2791	25.8462	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**Level**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
RouteClass2	4	-3.7311	1.2443	8.9915	0.0027
RouteClass2	3	1.8386	2.2886	0.6454	0.4218
RouteClass2	2	0.3602	2.9464	0.0149	0.9027
Oneway	5	0.8081	0.1259	41.1903	<.0001
Oneway	4	3.4399	0.8487	16.4278	<.0001
Oneway	3	5.2436	1.3215	15.7451	<.0001
Oneway	2	1.1925	3.5373	0.1136	0.7360

Table 5.7 Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	7480.648	5521.322
SC	7522.162	5957.225
-2 Log L	7472.648	5437.322

Based on the PPO model estimation results presented in Table 5.6, variables that satisfy the PO assumption including weekday, time period from 15:00 to 17:59, slope, signed bike lanes, suggested bike routes with low comfort, and greenways remain the same interpretation as the previous developed ORL model. Other variables seem to have different effects across the outcomes. The detailed model interpretation and model comparison will be presented in Section 5.6.

The model fit statistics provided in Table 5.7 indicate that the -2 LogL for the PPO model is less than that of the ORL model and is less than the constant-only model. It means the PPO model has a better fitness for the level of bicycle counts. To better examine the goodness of fit for this PPO model, the likelihood ratio index ρ^2 is utilized and presented in the following equation:

$$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(c)} \quad \text{Eq. (5)}$$

where $LL(\hat{\beta})$ is the log-likelihood value at convergence and $LL(c)$ represents the log-likelihood value for constant-only model. Based on the results presented in Table 5.7, the likelihood ratio index ρ^2 is 0.27. According to Train (2009)'s research study, a better model is associated with a higher value of ρ^2, and it is good enough to have ρ^2 from 0.2 to 0.4 in real world case studies. Therefore, it can be concluded that the PPO model is good enough to analyze the link-based route choice behavior for the Strava users in the City of Charlotte.

5.4 Multinomial Logit Model

5.4.1 MNL Model Structure

The multinomial logit model developed in this section is used to analyze the link-based bicyclist route choice behavior. The MNL model is usually based on the random utility theory (Train, 2009). It assumes that the alternative which yields the maximum utility is always selected. The utility function of the MNL model comprises an observed utility and an unobserved error term, which are shown in Equation (1).

$$U_{in} = V_{in} + \varepsilon_{in} \qquad \text{Eq. (6)}$$

where U_{in} is the utility function of the level of bicycle counts i for the road segment n, V_{in} is the observed utility of level i for the segment n, ε_{in} is the unobserved error term of level i for the segment n. V_{in} is usually taken as a linear utility function as shown in Equation (7).

$$V_{in} = \beta_0 + \sum_{k=1}^{N} \beta_k X_{ink} \qquad \text{Eq. (7)}$$

where X_{ink} is the kth attribute variable of level of bicycle counts i for road segment n, N is the total number of the attributes, β_0 is the constant term, and β_k is the coefficient of the kth attribute variable.

It is assumed that ε conforms to a Gumbel distribution, and attributes are independent of each other. Then the probability of the level of bicycle counts for each road segment for this research study can be derived as follows:

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \qquad \text{Eq. (8)}$$

5.4.2 MNL Model Results

The MNL model estimation result is presented in Table 5.8, in which the parameter estimates are shown for each level of the bicycle counts. One category is selected as the base case for this MNL model which is the low level of the bicycle counts. Variables that do not have significant impacts on the bicycle counts at 0.05 level are removed from the model utilizing the backward selection method.

Table 5.8 Multinomial Logit Model Estimation Results

Parameter Estimates					
Parameter	**Level**	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**
Constant2	2	2.3112	0.4306	5.37	<.0001
Constant3	3	-2.4150	1.0291	-2.35	0.0189
Constant4	4	5.9278	0.5980	9.91	<.0001

Parameter Estimates					
Parameter	**Level**	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**
Constant5	5	6.8923	0.7711	8.94	<.0001
Weekday	5	-4.1488	0.3084	-13.45	<.0001
Hour_2	2	-1.7464	0.4134	-4.22	<.0001
Hour_2	3	-1.4990	0.5230	-2.87	0.0042
Hour_2	5	1.2087	0.5109	2.37	0.0180
Hour_3	5	1.8764	0.4731	3.97	<.0001
Hour_4	4	-3.8902	0.4753	-8.19	<.0001
MPLength	5	1.5708	0.4601	3.41	0.0006
ThruLaneCo	5	0.5906	0.0775	7.62	<.0001
TOTPOP_CY	3	0.000278	0.000121	2.31	0.0211
MEDHINC_CY	3	0.0000402	7.6282E-6	5.27	<.0001
MEDHINC_CY	5	0.0000360	2.7568E-6	13.06	<.0001
Total_hous	4	0.005706	0.001417	4.03	<.0001
Total_hous	5	0.006635	0.001381	4.81	<.0001
TotalFamil	4	-0.007300	0.001749	-4.17	<.0001
TotalFamil	5	-0.008146	0.001691	-4.82	<.0001
Slope	4	0.0477	0.009090	5.25	<.0001
B_suggest0	4	1.1859	0.1312	9.04	<.0001
RouteClass2	4	-2.2344	0.3437	-6.50	<.0001
RouteClass4	5	0.4541	0.1313	3.46	0.0005
oneway	5	1.0411	0.1432	7.27	<.0001

According to the MNL model estimation results presented in Table 5.8. Variables that have significant impacts on bicycle counts contain weekday, time period from 9:00 to 14:59, time period from 15:00 to 17:59, time period from 18:00 to 19:59, the length of segment, the number of through lanes, total population, median household income, total households, total families, slope, suggested bike routes with low comfort, US route, secondary route, and one-way road. The explanatory variables kept in the MNL model are similar to the ORL and PPO model but not exactly the same. The detailed model result interpretation and comparison will be presented in Section 5.6.

The MNL model fit summary is shown in Table 5.9. From the table, it can be seen that the log-likelihood value at convergence is -2774. Therefore, -2 LogL is calculated which equals to 5548. This value will be used for the model comparison in Section 5.6.

Table 5.9 Model Fit Summary

Number of Observations	237673
Number of Cases	1188365
Log Likelihood	-2774
Log Likelihood (LogL(c))	-3736
AIC	5596
Schwarz Criterion	5845

5.5 Mixed Logit Model

The mixed logit model is different from the multinomial logit model because it allows explanatory variables to affect the mean of the random parameter distribution (Bhat 1998, Revelt and Train 1998, Bhat 2000, McFadden and Train 2000, Hensher and Greene 2003) and it can address the unobserved heterogeneity. Similar to MNL model, the linear utility function of the mixed logit model is shown in the following equation:

$$U_{in} = \beta_{in} X_{in} + \varepsilon_{in} \qquad \text{Eq. (9)}$$

where U_{in} denotes the utility function of the level of bicycle counts i on each road segment n, β_{in} means a vector of coefficient estimates which are allowed to vary, X_{in} represents a vector of explanatory variables which affect the level of bicycle counts, and ε_{in} is the error term.

According to the research conducted by Train (2009), the mixed logit model structure is shown in the following equation:

$$P_{in} = \int \frac{\exp(\beta_i X_{in})}{\sum_{i=1}^{I} \exp(\beta_i X_{in})} f(\beta \mid \varphi) d\beta \qquad \text{Eq. (10)}$$

where $f(\beta|\varphi)$ is the probability density function of β, and φ represents the parameter vector that shows the mean and variance of the density function. The distribution of β can be flexible or fixed, and can be any (e.g., normal, lognormal, uniform or triangular) distribution (Train 2009). In this research, the normal distribution is selected. If all the parameters are fixed, the mixed logit model will collapse into a simple multinomial logit model.

The MXL model is developed based on the MNL model. Subsequently, all variables in multinomial logit models are assumed to be randomly distributed at first and normal distribution is employed for all the variables in the MXL model. Then, a backward selection process is applied to determine the normally distributed parameters in the MXL model. Parameters will be

fixed if the standard deviation is not significantly different from zero at a level of significance of 0.5. 200 Halton draws are utilized during the simulation-based model estimation process. It is verified by some scholars that 200 Halton draws are sufficient and accurate for mixed logit model development (e.g., Koppelman et al. 2003). However, the number of observations (237673) is extremely large for the estimation of MXL model which is not time efficient. Therefore, the peak hour data are selected to analyze the link-based route choice behavior and the MXL models will be developed in Section 5.7.

5.6 Model Comparison

This section compares the results of ORL, PPO, and MNL models developed in the previous sections. Indicators utilized for the model comparison include -2Log-likelihood, the Akaike's information criterion (AIC), the Bayesian information criterion (BIC), and likelihood ratio index ρ^2.

5.6.1 Indicators for Model Comparison

The most commonly used indicators for model comparison are -2Log-likelihood, AIC, BIC, and ρ^2. To compare the models with same structure (e.g., ORL and PPO), all the indicators can be used. However, to compare models with different structures, it is not appropriate to utilize the likelihood values.

The values of AIC and BIC are calculated with the following equations:

$$AIC = 2p - 2LL \quad \text{Eq. (11)}$$

$$BIC = p\ln(Q) - 2LL \quad \text{Eq. (12)}$$

where p represents the number of parameters in the model, Q is the number of observations and LL denotes the log-likelihood value of the model.

Therefore, the four indicators for each model developed in the previous sections are presented in Table 5.10.

Table 5.10 Indicators for Model Comparison

Model	No. of Obs (Q)	No. of Vars. (p)	-2LogL	AIC	BIC	ρ^2
ORL	237673	23	5743	5789	6028	0.2315
PPO	237673	42	5437	5521	5957	0.2724
MNL	237673	24	5548	5596	5845	0.2575

Comparing the traditional ORL model to PPO model, the PPO has a smaller value of -2LogL than that of the ORL model, which indicates that PPO model outperforms the ORL model for fitting the bicycle count data in the City of Charlotte. To compare the three models with different structures, AIC and BIC values are utilized. Based on the values of AIC, PPO has

the smallest value among the three models, which reveals the best fitness of PPO model for this data. However, the BIC value of PPO is not the smallest. According to the BIC values, the MNL model performs better than PPO model, and PPO model is better than ORL model. The implication derived from the value of ρ^2 demonstrates that the PPO model with the largest value performs better than the other two models. The reason that the BIC value of PPO model is larger than MNL's can be interpreted that PPO model has more estimated parameters than MNL model. The trade-off between better fitness of the model and the number of variables should be carefully considered and examined. In this research study, with the consideration of the four indicators, conclusion can be provided that PPO model fits better for this link-based route choice behavior analysis.

5.6.2 Model Result Comparison

Based on the model estimation results in Table 5.3, Table 5.6, and Table 5.8, variables that have significant impacts on the link-based cyclist route choice behavior are identified and interpreted for all three models including ORL model, PPO model, and MNL model. The detailed analysis is provided as follows:

1. Temporal variables:
   The cycling behavior varies with different time in terms of weekday/weekend and time of day. According to the model estimation results of three models, weekdays have a negative impact on the bicycle counts for each road segment especially for the category of high-level bicycle counts. It can be interpreted that Strava users in the City of Charlotte prefer to bike on weekends. And on weekdays, the probability of the high-level bicycle count occurrence will decrease. The conclusion of this result is probably related to the high proportion of the non-commute trips in the Strava dataset. Different time of day will have different impacts on the bicycle counts since the link-based route choice varies with the change of time. The time period from 06:00 to 17:59 has a positive overall impact on the bicycle counts, while time period from 18:00 to 19:59 has a negative impact on the bicycle counts. To be specific, time period from 06:00 to 08:59 has a positive impact on average-high level. Time period from 09:00 to 14:59 has a negative impact on the low-average and average level, while it has a positive impact on the high level of bicycle counts. Time period from 15:00 to 17:59 has a positive impact on the high level of bicycle counts. And time period from 18:00 to 19:59 has a negative impact on average-high level. To conclude, cyclists prefer to bike during daytime, and time period from 06:00 to 17:59 is associated with high likelihood of above average bicycle counts. Researchers can assume that: First, the light condition is better during the daytime. Second, cyclists choose to bike during daytime considering the safety issue.

2. Road characteristics:
   Road characteristics are highly related to the cycling conditions which make the road characteristic factors significantly affect the link-based route choice. The explanatory variables that have a significant impact on the level of bicycle counts include the length of the road segment, number of through lanes, Interstate, US route, NC route, secondary

route, and one-way road. From the model estimation results, the length of the road segment has a positive impact on the bicycle counts. In other words, cyclists prefer to bike on long-distance road segments. This is probably because cyclists are willing to bike on roads with bicycle facilities (e.g., greenways) which tend to be long-distance road segments. The number of through lanes have a positive impact on the high-level bicycle counts for each road segment. It can be interpreted that cyclists tend to select road segments with greater number of through lanes as a part of their cycling routes. Interstate and NC route have a negative impact on the bicycle counts. In addition, US route will positively affect the high-level bicycle counts, however, negatively influence the average-high level. Secondary routes are associated with high-level bicycle counts. Therefore, it can be concluded that more bicycle counts are likely to occur on US routes and secondary routes. One-way road segments have a positive impact on the bicycle counts especially for high-level category. This result is probably related to the cycling preference in the uptown area where numerous one-way roads exist.

3. Sociodemographic characteristics:
   Several sociodemographic characteristics have different impacts on the level of bicycle counts on each road segment in the City of Charlotte. According to the model estimation results, explanatory variables that have significant impacts on bicycle counts contain total population, median age, median household income, total household, and total families. Based on the MNL model estimation results, the total population in the certain areas (census blocks) has a positive impact on the average level of the bicycle counts which indicates that a large population will be associated with average level of the bicycle counts. Locations with higher median age have a positive impact on bicycle counts. It can be interpreted that cyclists prefer to bike in the area with higher median age. The median household income factor may affect the bicycle counts differently across different levels. To be specific, the median household income has a positive impact on the average and above average levels, while it has a negative impact on the low-average level. An assumption can be made that the uptown area has higher median income and the bicycle counts in the uptown location are higher since bicyclists prefer to bike in the center city area. Interestingly, the total households and total families affect the level of bicycle counts differently. The total households affect the higher levels of bicycle counts positively, while the total families affect the higher levels of bicycle counts negatively. It can be assumed that cyclists prefer to select locations with more rental apartments and less family house neighborhood.

4. Geometry:
   The slope is one of the impact factors that affect the bicycle counts significantly. In the three discrete choice models, this variable is examined to discover the correlation between the probability of selecting the road segment as a part of the cycling route and

the slope. The model estimation results reveal that slope has a negative impact on the level of bicycle counts on each road segment. It is not hard to understand that bicyclists prefer to bike on flat segments instead of steep segments.

5. Bicycle facilities:
   Bicycle facilities are the critical consideration for cycling activities. Bicyclists may have different preferences for different bicycle facilities which are able to provide higher cycling safety. Based on the model estimation results, bike facilities including signed bike lanes, suggested bike routes (both regular and low comfort), and greenways will have a significant impact on the bicycle counts. Signed bike lanes will affect the level of bicycle counts negatively, while greenways will increase the likelihood of higher level of bicycle counts. The suggested bike routes with low comfort will have a negative impact on bicycle count levels expect for average-high level. And suggested bike routes have a positive impact on bicycle counts especially for the high-level category. It can be interpreted that greenways and suggested bike routes may have a better road condition compared to the other types of the bicycle facilities.

5.7 Modeling Link-based Route Choice for Different Time Periods

Based on the methodology described in Section 5.6, two mixed logit models are developed to analyze the link-based route choice for different time periods (am peak hours and pm peak hours). The model estimation procedure is conducted in SAS 9.4. The MXL logit models developed in this section are based on the MNL models built for different time periods. The MXL model developed for AM peak hours collapses into a MNL model. Therefore, the indicators for different time periods are presented in Table 5.11.

Table 5.11 Indicators for Different Time Periods

Time Periods	Model	No. of Obs (Q)	No. of Vars. (p)	-2LogL	AIC	BIC	ρ^2
AM Peak Hours	**MNL**	43444	24	798.71	846.71	1055.01	0.1632
PM Peak Hours	**MXL**	48447	13	1789.96	1815.96	1930.21	0.1690

In Section 5.7.1 and Section 5.7.2, the MNL model and the MXL for AM peak hours and PM peak hours respectively are presented. The analysis of the model estimation results demonstrates the impacts of different explanatory variables on the link-based route choice behavior for both peak hours.

5.7.1 AM Peak Hours

To analyze the link-based cyclist route choice behavior for AM peak hours, a MXL model is developed with low level of bicycle counts selected as the base. However, standard deviations of all the levels in the MXL model are not significantly different from zero at the

0.05 level. Therefore, this MXL model collapses into a MNL model. And the MNL model estimation results are shown in Table 5.12.

Table 5.12 MNL Model Estimation Results for AM Peak Hours

Parameter Estimates					
Parameter	**Level**	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**
Constant	2	4.0770	0.9704	4.20	<.0001
Constant	3	-11.5363	2.6558	-4.34	<.0001
Constant	4	-1.3841	3.0688	-0.45	0.6520
Constant	5	1.3761	1.8488	0.74	0.4567
Weekday	5	-1.8047	0.4723	-3.82	0.0001
MPLength	2	-12.1937	4.4586	-2.73	0.0062
SpeedLimit	4	-0.1408	0.0620	-2.27	0.0232
ThruLaneCo	3	2.1545	0.8328	2.59	0.0097
ThruLaneCo	4	2.3905	0.6926	3.45	0.0006
ThruLaneCo	5	2.0612	0.6235	3.31	0.0009
MEDHINC_CY	3	0.0000820	0.0000186	4.40	<.0001
MEDHINC_CY	4	0.0000422	0.0000156	2.70	0.0069
MEDHINC_CY	5	0.0000667	0.0000142	4.70	<.0001
Total_hous	2	-0.002737	0.001125	-2.43	0.0150
Total_hous	5	0.003217	0.001249	2.58	0.0100
TotalFamil	5	-0.005797	0.001417	-4.09	<.0001
FamilyPove	3	6.1878	2.8829	2.15	0.0318
FamilyPove	5	6.1456	1.7811	3.45	0.0006
B_bikelane	2	1.9884	0.8153	2.44	0.0147
B_bikelane	3	3.3529	0.8581	3.91	<.0001
B_greenway	2	3.4877	1.0441	3.34	0.0008
oneway	3	3.4908	1.0794	3.23	0.0012
oneway	4	2.3318	0.9354	2.49	0.0127
oneway	5	2.4732	0.7732	3.20	0.0014

1. Temporal variables:

Similar to the MNL model developed for the whole dataset, weekday has a negative impact on the high-level bicycle counts on each road segment. Same results can be concluded that the cyclists in the City of Charlotte prefer to bike on weekends. Weekdays will probably decrease the likelihood of the occurrence of high-level bicycle counts.

2. Road characteristics:
   The explanatory variables that have significant impacts on the level of bicycle counts are different from the variables in the MNL developed with the whole dataset. According to the model estimation results presented in Table 5.12, the road characteristic variables that have a significant impact on the level of bicycle counts contain the length of road segment, speed limit, number of through lanes, and one-way road. The length of the road segment has a negative impact on the low-average level of bicycle counts which indicates that low-average level of bicycle counts is likely to be associated with shorter road segments. The posted speed limit on a road segment will affect the bicycle count level (high-average) negatively. It is not hard to imagine cyclists prefer to bike on roads with a lower speed limit. Greater number of through lanes increases the likelihood of high level of bicycle counts (average and above). It can be interpreted that cyclists tend to select roads with more through lanes. In addition, the one-way road remains to have a positive impact on the high level of bicycle counts (average and above) which demonstrates that cyclists prefer to bike on one-way roads.

3. Sociodemographic characteristics:
   Changes are also found in the sociodemographic variables that have significant impacts on the level of bicycle counts for AM peak hours. Based on the results represented in Table 5.12, median household income, total households, total families, and family poverty rate will affect the bicycle counts significantly. The median household income has a positive impact on the average and above average levels which indicates that cyclists prefer to bike in the areas with higher household income. This result is consistent with the interpretation of the variable from models based on the whole dataset. Total households have a negative impact on low-average level of bicycle counts, while this variable affects the high level positively. This result reveals that the area with more households increases the likelihood of high-level bicycle counts and decreases the probability of low-average level. The impact of the total families remains the same as the MNL model developed using the whole dataset. The family poverty rate in this MNL is identified to have a positive impact on both average and high level of bicycle counts which means that cyclists prefer to bike at the locations with high poverty rates.

4. Bicycle facilities:
   The bicycle facilities that have significant impacts on bicycle counts are different from the previous MNL model. Only bike lanes and greenways will affect the level of bicycle

counts significantly. They both have positive impact on the low-average or average level. It can be interpreted that bike lanes and greenways increase the likelihood of low-average or average level of bicycle counts. It can be assumed that a lot of cycling trips occurred during AM peak hours are in the center city where few cyclists bike on these two types of bicycle facilities.

5.7.2 PM Peak Hours

To explore the difference of impact factors between the cycling activities occurred during AM peak hours and PM peak hours, the MXL model is developed and the model estimation results are presented in Table 5.13.

Table 5.13 MXL Model Estimation Results for PM Peak Hours

Parameter Estimates					
Parameter	**Level**	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**
Constant	2	1.0470	0.5182	2.02	0.0433
Constant	3	-1.5170	0.8442	-1.80	0.0723
Constant	4	0.1042	1.0485	0.10	0.9209
Constant	5	8.8208	0.7556	11.67	<.0001
SpeedLimit	4	-0.0518	0.0159	-3.25	0.0012
TOTPOP_CY	5	-0.000402	0.000135	-2.97	0.0030
MEDAGE_CY	4	0.0765	0.0253	3.02	0.0025
MEDHINC_CY	4	-0.000104	0.0000117	-8.86	<.0001
Total_hous_M	3	0.001810	0.002016	0.90	0.3691
Total_hous_S	3	-0.002175	0.000682	-3.19	0.0014
Total_hous	4	0.004110	0.001235	3.33	0.0009
Total_hous	5	0.005762	0.000981	5.87	<.0001
Slope	4	-0.0799	0.0216	-3.70	0.0002

Compared to the MNL developed for the cycling behavior during AM peak hours, the explanatory variables that remain to have significant impacts on the bicycle counts during PM peak hours include speed limit, median household income, and total households. In addition, different from the impact factors for cycling behavior during AM peak hours, total population, median age, and slope are found to affect the level of bicycle counts significantly during PM peak hours.

Speed limit still has a negative impact on the level of bicycle counts which is consistent with the results of cycling behavior during AM peak hours. Different link-based route choice behavior is found in terms of the impact of total population. During PM peak hours, cyclists

prefer to bike on roads located in the area with low population which is opposite to the results concluded from the models based on the whole dataset. The median age variable has a positive impact on the high-average level which remains the same as what was mentioned before. However, median household income has a negative impact on the average-high level of bicycle counts which indicates that cyclists prefer to bike in the area with low household income. Total households still have a positive impact on average and above average levels, and slope still remains a negative impact on high-average level.

5.8 Summary

This chapter developed several discrete choice models including ordered logit model, partial proportional model, multinomial logit model, and mixed logit model to analyze the link-based cyclist route choice behavior. Model comparison is conducted to select the best model structure for this research study. The link-based route choice behavior of different time periods including AM peak hours and PM peak hours is analyzed based on the mixed logit model. Impact factors that are associated with different levels of bicycle counts in the City of Charlotte are identified.

Chapter 6. Methods for Analyzing Path-based Cyclist Route Choice

6.1 Introduction

This Chapter provides a method to analyze the path-based cyclist route choice. The labeling method is selected for the choice set generation procedure which is the preparation of cyclist route choice analysis. The structure of Path Size Logit model is presented as a guidance for modeling path-based route choice behavior. The rest of this Chapter is organized as follows. Section 6.2 explains the choice set generation method. Section 6.3 introduces the Path Size Logit model. Finally, Section 6.4 concludes the chapter with a summary.

6.2 Choice Set Generation

There are several choice set generation methods that have been utilized as the preparation for the cyclist route choice analysis. One of the most prevalent methods is the labeled route method. It can be conducted in the ArcGIS 10.4 by the Network Analyst extension. The alternative routes from the selected origin to the destination are generated based on the maximum or minimum values of certain attributes. For the unique OD pair, the alternatives are comprised of the created nonchosen alternatives and the route that is actually chosen by the specific cyclist.

In this study, the alternative routes for the pair of origin and destination are generated following the criteria listed below:

1. Minimize the distance of the cycling route from origin to the destination;
2. Maximize the usage of bicycle facilities along the cycling route from origin to the destination;
3. Minimize the number of intersections for the cycling route from origin to the destination;
4. Minimize the proportion of one-way road segments along the cycling route from origin to the destination;

An example of the generation method for shortest cycling path from origin to destination is presented in the following figure.

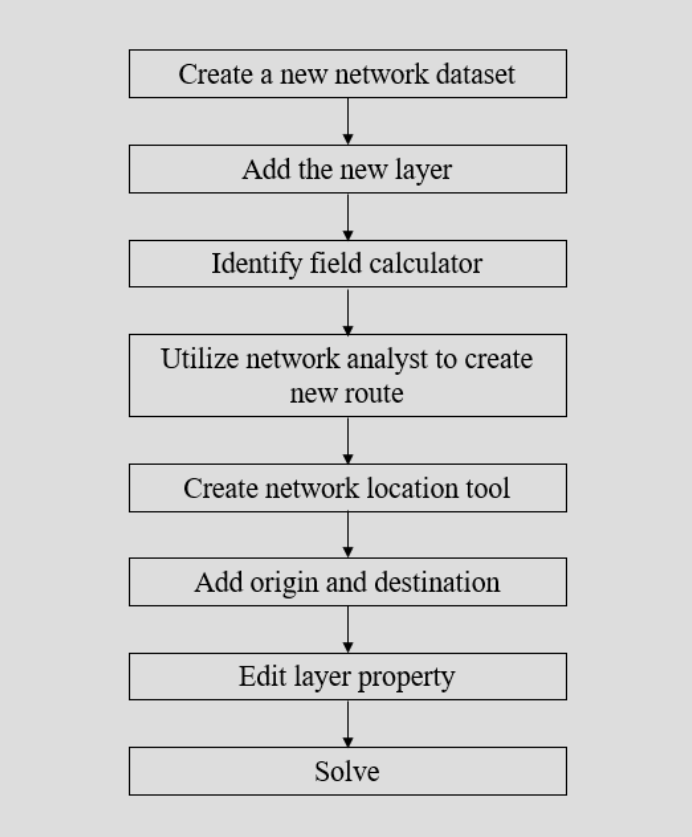


Figure 6.1: An Example of Choice Set Generation Procedure

6.3 Path Size Logit Model

To estimate models based on the generated route alternatives, some types of the discrete choice models can be utilized. The basic model for analyzing the path-based route choice behavior of cyclists is the multinomial logit model mentioned in the previous section. A classical conditional maximum likelihood estimation method can be used for developing the MNL model.

In this path-based route choice research study, the probability of a cyclist choosing the alternative route *i* from the available alternative routes in choice set C_n is presented as follows:

$$P(i|C_n) = \frac{\exp(V_{in})}{\sum_{j \in C_n} \exp(V_{jn})} \quad \text{Eq. (13)}$$

where *i* denotes the chosen alternative route, *j* represents the alternative routes in choice set C_n, V_{in}/V_{jn} are the deterministic utility of alternative route *i*/*j* for individual *n*.

However, the limitation of MNL model is revealed in terms of the independence of irrelevant alternatives (IIA) property. In this situation, the provided alternative routes in the choice set are required to be mutually exclusive. In other words, overlapping routes are not allowed to exist while using MNL model to estimate the route choice. Neglecting this IIA property will end with overestimating the overlapping routes.

To release this restriction, an appropriate correction should be introduced for the utilities of alternative routes to account for the correlation. A path size (PS) factor that corrects the utilities can reflect the correlation among all routes. The PS factor is presented in the following equation:

$$PS_{in} = \sum_{a \in T_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_n} (\frac{L_i}{L_j})^{\gamma} \delta_{aj}} \quad \text{Eq. (14)}$$

where L_a denotes the length of link a, L_i represents the length of route i, T_i demonstrates the set of links in route i, δ_{aj} equals one if link a is used in route j, otherwise, δ_{aj} equals zero, γ indicates the long-path correction factor. For most cycling cases, this factor is assumed to be zero.

The PS attribute is included in the deterministic utility of the route alternatives and then the Path Size Logit (PSL) model is developed. Thus, the probability of alternative route i selected by a cyclist from the choice set C_n is presented in the following equation:

$$P(i|C_n) = \frac{\exp(V_{in} + \ln(PS_{in}))}{\sum_{j \in C_n} \exp(V_{jn} + \ln(PS_{in}))} \quad \text{Eq. (15)}$$

From the PSL model, it can be seen that the utility of each alternative route is changed. The new form of the utility function of each route is presented as follow:

$$U_i = \beta X_i + \beta_{PS} \times \ln(PS) \quad \text{Eq. (13)}$$

where X_i is the vector of attribute variables of route i, and β is the coefficients that need to be estimated.

Because the limitation of the Strava data, the exact cycling trajectory information cannot be obtained for this path-based route choice analysis. Only aggregated data and the moving direction within designed polygons can be extracted. Therefore, for further study on the path-based route choice, cycling trajectories are essential for model development and research analysis.

6.4 Summary

This chapter provides the methods for analyzing cyclists' path-based route choice behavior. The labeled route method is selected to demonstrate the choice set generation procedure for this research study. An example of choice set generation conducted in ArcGIS 10.4 is presented. Based on the generated choice set, the PS factor is introduced in the MNL model for the correction, and a PSL model structure is presented to give a guidance for the route choice behavior analysis.

Chapter 7. Summary and Conclusions

7.1 Introduction

Cycling has gained more attention from the citizens and planners recently, since it can provide benefits not only for the society but also for the environment. By promoting cycling especially for short-distance trips, Charlotte has been making every effort to become a bike-friendly city. As an ideal travel mode, cycling is able to improve public health, reduce energy consumption, and alleviate air pollution, etc.

To increase the mode share of cycling, research studies are needed to conduct in order to explore the impacts on both link-based and path-based route choice behavior. One of the most critical issues that need to be considered for the route choice analysis is the data collection method. Traditional data collection methods including travel surveys and data from manual count machines can be time-consuming and expensive. The novel crowdsourced data address the issues brought by traditional data collection methods and provide the temporal and spatial information of cycling to bridge the gap.

Based on the crowdsourced bicycle data collected from the Strava application, this research study is conducted to analyze the link-based cyclist route choice behavior in the City of Charlotte and to present a method for the future investigation on path-based route choice behavior.

The primary objective of this research study is to model the link-based route choice behavior for the cyclists in the City of Charlotte. Different discrete choice models are developed including ORL model, PPO model, MNL model, and MXL model. A model comparison is conducted to identify the best model structure for this research study. MXL model and MNL model are utilized to compare the cycling behavior for different time periods. The impact of explanatory variables in terms of temporal variables, road characteristics, sociodemographic information, geometry, and bike facilities are analyzed. In addition, a method is provided for the future studies on path-based route choice behavior. The labeled route method is selected for the choice set generation procedure. An example of the choice set generation conducted using ArcGIS 10.4 is provided for the preparation of the route choice analysis. Based on the previous research study, a PSL model is proposed to analyze the path-based route choice behavior. Finally, the limitation of the crowdsourced bicycle data collected from Strava is discussed.

The rest of this chapter is organized as follows. Section 7.2 provides a brief review of the methods used to analyze the link-based route choice behavior for the cyclists in the City of Charlotte. The model estimation results are concluded in this section, and the model comparison result indicating the best model structure for this research study is summarized. Different link-based route choice behavior for cycling during both AM and PM hours are identified. Section 7.3 discusses the limitation of this research study and provides the future research directions in terms of the path-based route choice behavior analysis using PSL model.

7.2 Summary and Conclusions

As mentioned before, a comprehensive literature review regarding the concept of crowdsourcing, the introduction of crowdsourced bicycle data, and the use of crowdsourced data for different aspects of research studies including both link-based and path-based route choice behavior analysis, etc. is conducted to understand the usage of crowdsourced bicycle data and the modeling methods for route choice in previous research studies.

Based on the crowdsourced data collected from Strava, the descriptive analyses are conducted in terms of the demographic information on Strava users, cycling activities for different trip purposes, the cyclist counts on each road segment in the City of Charlotte for each month of year, weekdays/weekends, and time of day, the origin and destination of cycling trips for the most popular one, and the total cyclist counts in each origin/destination polygon for different trip purposes, and on weekdays and weekends, etc.

Several discrete choice models are developed to analyze the link-based cyclist route choice behavior in the City of Charlotte. Models including ORL model, PPO model, MNL model, and MXL model are compared to identify the best fit for this Strava dataset. According to the model estimation results, variables including weekday, total family, slope, signed bike lanes, suggested bike routes with low comfort, interstate route, and NC route are found to have a negative impact on the level of bicycle counts, while other variables which are time period from 6:00 to 17:59, segment length, number of through lanes, median age, median household income, total household, suggested bike routes, greenway, US route, and one-way road are identified to have a positive impact on the level of bicycle counts in the ORL model. In the PPO model, variables including time period from 6 am to 9 am, the number of through lanes, median household income, total households, total families, suggested bike routes, US routes, and one-way road violate the PO assumption and have different effects across different levels. Variables that satisfy the PO assumption including weekday, time period from 15:00 to 17:59, slope, signed bike lanes, suggested bike routes with low comfort, and greenways have the same interpretation as the ORL model. The explanatory variables that have a significant impact on bicycle counts in MNL model contain weekday, time period from 9:00 to 14:59, time period from 15:00 to 17:59, time period from 18:00 to 19:59, the length of segment, the number of through lanes, total population, median household income, total households, total families, slope, suggested bike routes with low comfort, US route, secondary route, and one-way road which are similar to the ORL and PPO model. By calculating the indicators (-2LogL, AIC, BIC, and ρ^2) for model comparison, PPO model is determined to be best model structure for this link-based route choice behavior analysis. To explore the different route choice behavior for both AM and PM peak hours, two MXL models are developed. Impact factors that are associated with different levels of bicycle counts in the City of Charlotte are identified.

7.3 Directions for Future Research

This section summarizes the limitation of this research study and provides the directions for the future research study. The limitations of this research study are listed as follows:

1. Strava data limitations:

(1) The crowdsourced bicycle data collected from Strava only contain a large portion of the cyclists in the City of Charlotte. The models developed based on this dataset only reveal the cycling behavior of the Strava users in the City of Charlotte. The factors affecting link-based route choice may vary with different sources of data.

(2) Most of the cyclists using Strava application are male cyclists accounting for 80.49% of the total Strava users in the City of Charlotte which may lead to an unavoidable bias to the route choice analysis.

(3) The majority of the cycling trips generated by Strava users are non-commute trips which may be different from the cycling behavior for commute trips.

(4) The Strava data are aggregated before providing for research studies. No actual cycling trajectory information can be obtained for path-based route choice analysis.

2. Link-based route choice models:

(1) Some variables may have a potential impact on the link-based route choice behavior, such as traffic volumes. However, the traffic volume data are not available for this case study.

(2) Some supporting data (e.g., roadway characteristics data) are not available for certain roadway segments, and thus the records with blank information are removed from the dataset.

Based on the summarized limitations of this research study and the literature review on relevant topics, the directions for future studies are provided as follows:

1. Other models besides ORL, PPO, MNL, and MXL models should be developed and tested to see the fitness for the link-based route choice. And a more comprehensive model comparison can be conducted based on the new models.

2. The differences between commute trips and non-commute trips should be identified by modeling route choice models separately.

3. The cycling activities occurred in various locations can be different. Comparison can be conducted for route choice behavior for different locations (e.g., urban or rural areas).

4. Crash frequency or severity can be considered to examine their impacts on the cyclists' route choice behavior. In addition, cyclist injury risk factors can be computed for the safety analysis.

5. Other choice set generation methods can be utilized to compare with the labeled route method for the path-based route choice analysis.

6. Cycling trajectory data should be collected to complete the path-based route choice analysis. Other models (e.g., expanded path size logit model) should be used and compared with the PSL model.

References

1. Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). "A systematic review of open government data initiatives." *Government Information Quarterly*, *32*(4), pp. 399-418.
2. Bekhor, S., Ben-Akiva, M. E., and Ramming, M. S. (2006). "Evaluation of choice set generation algorithms for route choice models." *Annals of Operations Research*, *144*(1), pp. 235-247.
3. Ben-Akiva, M., Bergman, M., Daly, A., and Ramaswamy, J. (1984). "Modelling Inter Urban Route Choice Behaviour." In J. Volmuller and R. Hamerslag (eds.), *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, VNU Press, Utrecht, pp. 299-330.
4. Ben-Akiva, M. and Bierlaire, M. (1999). "Discrete choice methods and their applications to short term travel decisions." R. Hall, ed., *Handbook of Transportation Science*, Kluwer Academic Publishers, Norwell, MA, chapter 2, pp. 5-34.
5. Bergman, C., and Oksanen, J. (2016). "Conflation of OpenStreetMap and Mobile Sports Tracking Data for Automatic Bicycle Routing." *Transactions in GIS*, 20(6), pp. 848-868.
6. Bhat, C.R. (1998). "Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling." *Transportation Research Part A: Policy and Practice*, 32(7), pp. 495-507.
7. Bhat, C.R. (2000). "Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling." *Transportation science*, 34(2), pp. 228-238.
8. Bierlaire, M., Chen, J., and Newman, J. (2010). *Modeling route choice behavior from smartphone GPS data* (No. REP_WORK).
9. Bovy, P. and Fiorenzo-Catalano, S. (2007). "Stochastic route choice set generation: behavioral and probabilistic foundations." *Transportmetrica,* 3, pp. 173-189.
10. Brabham, D. C. (2008). "Crowdsourcing as a model for problem solving: An introduction and cases." *Convergence,* 14(1), pp. 75-90.
11. Broach, J., Gliebe, J., and Dill, J. (2009). "Development of a multi-class bicyclist route choice model using revealed preference data."
12. Broach, J., Dill, J., and Gliebe, J. (2012). "Where do cyclists ride? A route choice model developed with revealed preference GPS data." *Transportation Research Part A: Policy and Practice*, *46*(10), pp. 1730-1740.
13. Casello, J., and Usyukov, V. (2014). "Modeling cyclists' route choice based on GPS data." *Transportation Research Record: Journal of the Transportation Research Board,* 2430, pp. 155-161.
14. Charlton, B., Hood, J., Sall, E., and Schwartz, M. A. (2011). "Bicycle route choice data collection using GPS-enabled smartphones." In *Transportation Research Board 90th Annual Meeting.* No. 11-2652.

15. Chen, C., and Chen, P. (2013). “Estimating recreational cyclists’ preferences for bicycle routes-Evidence from Taiwan.” *Transport Policy, 26*, pp. 23-30.
16. Chen, M., Mao, S., and Liu, Y. (2014). “Big data: A survey.” *Mobile Networks and Applications* 19(2), pp. 171-209.
17. City of Charlotte Department of Transportation. (2017, May 22). Charlotte BIKES Bicycle Plan. Available at http://charlottenc.gov/Transportation/Programs/Documents/Charlotte%20BIKES%20Final.pdf
18. De la Barra, T., Perez, B., and Anez, J. (1993). “Multidimensional Path Search and Assignment.” In *Proceedings of the 21st PTRC Summer Meeting*, pp. 307-319.
19. Dill, J., and Gliebe, J. (2008). “Understanding and measuring bicycling behavior: A focus on travel time and route choice.”
20. Estellés-Arolas, E., and González-Ladrón-De-Guevara, F. (2012). “Towards an integrated crowdsourcing definition,” *Journal of Information science*, *38*(2), pp. 189-200.
21. Foresite Group. (2015). Available at http://www.fg-inc.net/crowdsourced-data-for-bike-pedestrian-facility-planning/.
22. Fosgerau, M., Frejinger, E., and Karlstrom, A. (2013). “A link based network route choice model with unrestricted choice set.” *Transportation Research Part B: Methodological*, 56, pp. 70-80.
23. Frejinger, E., Bierlaire, M., and Ben-Akiva, M. (2009). “Sampling of alternatives for route choice modeling.” *Transportation Research Part B: Methodological*, *43*(10), pp. 984-994.
24. Gantz, J., and Reinsel, D. (2011). “Extracting value from chaos.” *IDC iview*, 1142(2011), pp. 1-12.
25. Griffin, G. P. and Jiao J. (2016). “Crowdsourcing Bicycle Volumes: Exploring the Role of Volunteered Geographic Information and Established Monitoring Methods.” *Journal of the Urban and Regional Information Systems Association*, 27(1).
26. Griffin, G. P., and Jiao, J. (2015). “Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus.” *Journal of Transport & Health*, 2(2), pp. 238-247.
27. Grond, K. (2016). *Route Choice Modeling of Cyclists in Toronto*. Diss. University of Toronto (Canada).
28. Guan, H. *Discrete choice Modeling*. 2004.
29. Hensher, D.A., and Greene, W.H. (2003). “The mixed logit model: the state of practice.” *Transportation*, 30(2), pp. 133-176.
30. Hess, S., Quddus, M., Rieser-Schüssler, N., and Daly, A. (2015). “Developing advanced route choice models for heavy goods vehicles using GPS data.” *Transportation Research Part E: Logistics and Transportation Review*, 77, pp. 29-44.
31. Hochmair, H. H., Bardin, E., and Ahmouda, A. (2017). “Estimating Bicycle Trip Volume for Miami-Dade County from Strava Tracking Data.” No. 17-06577.

32. Hood, J., Sall, E., and Charlton, B. (2011). “A GPS-based bicycle route choice model for San Francisco, California.” *Transportation letters*, 3(1), pp. 63-75.
33. Howe, J. (2006). “The rise of crowdsourcing.” *Wired magazine, 14*(6), pp. 1-4.
34. Hudson, J. G., Duthie, J. C., Rathod, Y. K., Larsen, K. A., and Meyer, J. L. (2012). *Using smartphones to collect bicycle travel data in Texas* (No. UTCM 11-35-69). Texas Transportation Institute. University Transportation Center for Mobility.
35. Jackson, S., Miranda-Moreno, L. F., Rothfels, C., and Roy, Y. (2014). “Adaptation and implementation of a system for collecting and analyzing cyclist route data using smartphones.” *Transportation Research Board 93rd Annual Meeting*. No. 14-4637.
36. Jestico B., Nelson T. and Winters M. (2016). “Mapping ridership using crowdsourced cycling data.” *Journal of Transport Geography,* 52, pp. 90-97.
37. Kagerbauer, M., Hilgert, T., Schroeder, O., and Vortisch, P. (2015). “Household travel survey of intermodal trips - Approach, challenges and comparison.” *Transportation research procedia, 11*, pp. 330-339.
38. Khatri, R., Cherry, C. R., Nambisan, S. S., and Han, L. D. (2016). “Modeling route choice of utilitarian bikeshare users with GPS data.” *Transportation Research Record: Journal of the Transportation Research Board,* 2587, pp. 141-149.
39. Kleemann, F., Voß, G. G., and Rieder, K. (2008). “Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing.” *Science, technology & innovation studies*, 4(1), pp. 5-26.
40. Koppelman, F., Bhat, C., Sethi, V., and Williams, B. (2003). *A Self-instructing Course in Mode Choice Modeling*. US Department of Transportation, Federal Highway Administration.
41. Kučera, J., Chlapek, D., and Nečaský, M. (2013). “Open government data catalogs: Current approaches and quality perspective.” *International Conference on Electronic Government and the Information Systems Perspective*. Springer, Berlin, Heidelberg, pp. 152-166.
42. LaMondia, J., and Watkins, K. (2017). Using Crowdsourcing to Prioritize Bicycle Route Network Improvements.
43. Liu, E., and Porter, T. (2010). “Culture and KM in China.” *Vine,* 40(3/4), pp. 326-333.
44. McFadden, D., and Train, K. (2000). “Mixed MNL models for discrete response.” *Journal of applied Econometrics*, 15(5), pp. 447-470.
45. Menghini, G., Carrasco, N., Schüssler, N., and Axhausen, K. W. (2010). “Route choice of cyclists in Zurich.” *Transportation research part A: policy and practice*, 44(9), pp. 754-765.
46. Misra, A., Gooze, A., Watkins, K., Asad, M., and Le Dantec, C. (2014). “Crowdsourcing and its application to transportation data collection and management.” *Transportation Research Record: Journal of the Transportation Research Board*, 2414, pp. 1-8.
47. Moore, Michael. (2015). “Modeling Factors Influencing Commuter Cycling Routes: A Study of GPS Cycling Records in Auburn, Alabama.”
48. Nassir, N., Ziebarth, J., Sall, E., and Zorn, L. (2014). “Choice set generation algorithm suitable for measuring route choice accessibility.” *Transportation Research Record: Journal*

of the Transportation Research Board, 2430, pp. 170-181.
49. Proulx, F. R. and Pozdnukhov, A. (2017). "Bicycle Traffic Volume Estimation Using Geographically Weighted Data Fusion." Available at http://faculty.ce.berkeley.edu/pozdnukhov/papers/Direct_Demand_Fusion_Cycling.pdf.
50. Raihan, M. A., and Priyanka Alluri PHD, P. E. (2017). "Impact of Roadway Characteristics on Bicycle Safety." *Institute of Transportation Engineers. ITE Journal,* 87(9), pp. 33.
51. Reiche, K. J., and Höfig, E. (2013). "Implementation of metadata quality metrics and application on public government data." Accepted for *IEEE - Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual*. pp. 236-241.
52. RenoTracks. *RenoTracks*. (2013). Available at http://renotracks.nevadabike.org/.
53. Revelt, D., and Train, K. (1998). "Mixed logit with repeated choices: households' choices of appliance efficiency level." *Review of economics and statistics*, 80(4), pp. 647-657.
54. Rieser-Schüssler, N., Balmer, M., and Axhausen, K. W. (2013). "Route choice sets for very high-resolution data." *Transportmetrica A: Transport Science*, *9*(9), pp. 825-845.
55. San Francisco County Transportation Authority. The CycleTracks Smartphone Application. (2013). Available at http://www.sfcta.org/modeling-and-travel-forecasting/cycletracks-iphone-andandroid/cycletracks-smartphone-application.
56. Saxton, G. D., Oh, O., and Kishore, R. (2013). "Rules of crowdsourcing: Models, issues, and systems of control." *Information Systems Management, 30*(1), pp. 2-20.
57. Schenk, E., and Guittard C. (2011). "Towards a characterization of crowdsourcing practices." *Journal of Innovation Economics & Management,* 1, pp. 93-107.
58. Sener, I. N., Eluru, N., and Bhat, C. R. (2009). "An analysis of bicycle route choice preferences in Texas, US." *Transportation*, *36*(5), pp. 511-539.
59. Stinson, M., and Bhat, C. (2003). "Commuter bicyclist route choice: Analysis using a stated preference survey." *Transportation Research Record: Journal of the Transportation Research Board,* 1828, pp. 107-115.
60. Sun, Y., and Mobasheri, A. (2017). "Utilizing Crowdsourced data for studies of cycling and air pollution exposure: A case study using Strava Data." *International journal of environmental research and public health,* 14(3), pp. 274.
61. Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.
62. Vukovic, M. (2009). "Crowdsourcing for enterprises." Accepted for *IEEE - Services-I, 2009 World Conference*, pp. 686-692.
63. Watkins, K., Ammanamanchi, R., LaMondia, J., and Le Dantec, C. A. (2016). "Comparison of Smartphone-based Cyclist GPS Data Sources." In *Transportation Research Board 95th Annual Meeting.* No. 16-5309.
64. Wexler, M. N. (2011). "Reconfiguring the sociology of the crowd: exploring crowdsourcing." *International Journal of Sociology and Social Policy,* 31(1/2), pp. 6-20.
65. Winters, M., Teschke, K., Grant, M., Setton, E., and Brauer, M. (2010). "How far out of the way will we travel? Built environment influences on route selection for bicycle and car

travel." *Transportation Research Record: Journal of the Transportation Research Board*, (2190), pp. 1-10.
66. Yeboah, G., and Alvanides, S. (2015). "Route Choice Analysis of Urban Cycling Behaviors Using OpenStreetMap: Evidence from a British Urban Environment." In *OpenStreetMap in GIScience*, Springer International Publishing, pp. 189-210.
67. Zimmermann, M, Mai, T., and Frejinger, E. (2017). "Bike route choice modeling using GPS data without choice sets of paths." *Transportation research part C: emerging technologies*, 75, pp. 183-196.