EVALUATING THE POTENTIAL USE OF CROWDSOURCED BICYCLE DATA FOR CYCLING ACTIVITIES AND SAFETY ANALYSIS

by

Zijing Lin

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Infrastructure and Environmental Systems

Charlotte

2020

Approved by:

_____

Dr. Wei Fan

_____

Dr. Martin Kane

_____

Dr. David Weggel

_____

Dr. Jay Wu

_____

Dr. Jing Yang

ABSTRACT

ZIJING LIN. Evaluating the potential use of crowdsourced bicycle data for cycling activities and safety analysis. (Under the direction of DR. WEI FAN)

Cycling, as a healthier and greener travel mode, has been encouraged for short-distance trips by city planners and policymakers. Since cycling provides an efficient way to improve public health, alleviate traffic congestion, and reduce energy consumption, it is essential to analyze the contributing factors to the cycling activities on each roadway segment and bicyclist injury risk, so as to quantify the impact of certain attributes on bicycle volume as well as biking safety and further provide better cycling environment for cyclists to encourage non-motorized travels.

To map ridership, data including network characteristics, sociodemographic factors, and temporal characteristics, are quite indispensable. There have been multiple bicycle volume data collection methods and the most commonly used ones include traditional manual counts, travel surveys, and crowdsourced data from the third party. Most of the previous research efforts used the first two methods mentioned above to collect the data of interest. However, such methods are expensive and time-consuming. Crowdsourced data, on the contrary, are cost effective and timesaving, and therefore they have been widely collected and used by many public agencies and private sectors in recent years. Among all the crowdsourced data, data collected from smartphone applications including Strava, CycleTracks, ORcycle, etc. have become more and more prevalent. Crowdsourcing has increased the availability of data collection and provided an efficient way to bridge the data gap for decision making and performance measures.

This research concentrates on evaluating the potential use of crowdsourced bike data and comparing them with the traditional bike counting data that are collected in the city of Charlotte, NC. Using the bike data from both the Strava smartphone cycling application and the bicycle count stations, the bicycle volume models are developed. Based on the results, a bicycle volume predictive model is presented, and a map illustrating the bicycle volume on most of the road segments in the City of Charlotte is generated. In addition, to gain a better understanding of the attributes that have an impact on cycling, other supporting data are also collected and combined with the Strava bicycle count data. Multiple discrete choice models are developed to analyze the Strava users' cycling activities. Furthermore, bicyclist injury risk analysis is also conducted to explore the impact factors affecting biking safety by developing a series of safety performance functions. Several indicators for model comparison are utilized to select the best fitting model for bicyclist injury risk modeling. Finally, recommendations are made in order to help improve the cycling environment and safety and increase the bicycle volume in the future.

ACKNOWLEDGEMENTS

I would like to express my greatest gratitude to my advisor, Dr. Wei Fan, who has been continuously supporting and helping me during my Ph.D. study and other relevant research studies. Dr. Wei Fan has guided me to obtain technical skills, provided me the ability to develop and conduct research studies individually, and encouraged me to explore various topics that I am interested in. I am grateful for his kindness, patience, and motivation.

Also, it is important to have the guidance and support from my Ph.D. committee members including Dr. Martin Kane, Dr. David Weggel, Dr. Jay Wu, and Dr. Jing Yang. I would like to thank the committee members for their kind comments and useful suggestions made to this dissertation, which have helped me tremendously to improve the quality of this research.

Furthermore, I would like to thank my peers in the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE) lab for the great help and support in my research study at UNC Charlotte. The academic collaboration within the lab really helped improve our performance and provided inspiration to each other.

Finally, I would like to thank my family: my mother Suhong Zhang and my father Zhishui Lin and friends: Ziyi Sun, Yanwei Ma, Xinzi Ji, Irene, and Xiaofan Feng, who have been my support and helped me significantly in my life abroad.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AADB	annual average daily bicycle
AADT	annual average daily traffic
AIC	Akaike information criterion
BIC	Bayesian information criterion
BLOS	bicycle level of service
CDOT	Charlotte Department of Transportation
CLMPO	Central Lane Metropolitan Planning Organization
DAF	daily adjustment factor
DR	deceleration rate
EB	empirical Bayes
GLMM	generalized linear mixed model
GPS	global positioning system
LTS	level of traffic stress
MAF	monthly adjustment factor
MNL	multinomial logit
MXL	mixed logit
NBRM	negative binomial regression model
ORL	ordered logit
PO	proportional odds
PPO	partial proportional odds
PRM	Poisson regression model
RP	revealed preference

SFCTA	San Francisco County Transportation Authority
SP	stated preference
SPF	safety performance function
USDOT	United States Department of Transportation
VGI	volunteered geographic information
ZINB	Zero-inflated Negative Binomial
ZIP	Zero-inflated Poisson

CHAPTER 1: INTRODUCTION

1.1. Problem Statement and Motivation

With the increase in traffic demand, cities all over the world begin to encourage use of non-motorized travel modes, such as cycling, especially for short distance trips. It has been well known that cycling is an efficient way to provide healthier and greener travel which can help alleviate traffic congestion, reduce emissions, decrease energy consumption, and improve public health. In a safe and comfortable traveling environment, cycling will become a normal and common choice for travelers to get around, and in return, the city will benefit from it to have healthier and more energetic population.

According to the Charlotte Department of Transportation (CDOT) Bicycle Program (CDOT, 2017), Charlotte is making every effort to offer an inclusive and comfortable cycling environment for all potential bicyclists. The program provides people of all ages and abilities the convenience to use their bicycles for traveling, fitness, and fun. Therefore, studies on identifying what attributes might have an impact on cycling are highly desirable and even become essential for city planners, policymakers, and researchers.

Charlotte has been taking significant steps to become a bicycle-friendly city during the past fifteen years. A comprehensive bicycle plan has been adopted, and changes to the policies have been made that lead to changes on the ground for bicyclists. The first mile of bicycle lanes was constructed in 2001. With the changes in bicycle plans and policies, the bike network in Charlotte has increased to contain more than 90 miles of bicycle lanes, 40 miles of greenways and off-street paths, and 55 miles of signed routes (CDOT, 2017). According to a cycling survey conducted by CDOT (2017), 51% of the

residents in Charlotte would be willing to travel by bike more than they currently do. However, a majority of 62% of the respondents in this survey do not think it is easy to bicycle in Charlotte. This survey results clearly indicate that there is still a lot to do in order to improve cycling conditions in Charlotte. In addition, cyclists in the United States, compared to other developed countries, have a higher probability to suffer from fatal injuries (Pucher and Dijkstra, 2003). Based on the North Carolina Crash Data, 11,266 crashes were cyclist-involved crashes from 2007 to 2018, of which 250 cyclists were fatally injured. It is clearly indicated that exploring the impact factors on cyclist injury and conducting injury risk analysis are meaningful and essential. It is expected that, when the cycling environment is properly improved, more travelers will choose cycling as their travel mode.

To evaluate the factors that affect cycling activities on road segments and analyze bicyclist injury risk, data including network characteristics, sociodemographic information, location-specific elements, temporal factors, crash records and bicycle counts are essential. There have been multiple data collection methods and the most commonly used ones include traditional manual counts, travel surveys, and crowdsourced data from a third party. Most of the previous research efforts used the first two methods to collect the data of interest. However, such methods are expensive and time-consuming. Crowdsourced data, on the contrary, are cost effective and timesaving, and therefore have been widely collected and used by many public agencies and private sectors in recent years. Among all the crowdsourced data, data collected from smartphone applications including Strava, CycleTracks, and ORcycle, etc. have become more and more prevalent.

Crowdsourcing has increased the availability of data collection and provided an efficient way to bridge the data gap for decision making and performance measures.

As an advanced data collection method, crowdsourcing enables practitioners and scholars to collect data from a broader range of people in a shorter and more cost-efficient way. This method was first introduced by Howe (2006) in his "The Rise of Crowdsourcing" article. Crowdsourced data can greatly help planners develop models, analyze the travel behavior, estimate the traffic demand, evaluate bike facilities, and explain road traffic safety such as collisions. Different research efforts have been made with different definitions for crowdsourcing. According to Brabham (2008), crowdsourcing is "a strategic model to attract an interested, motivated crowd of individuals capable of providing solutions superior in quality and quantity to those that even traditional forms of business can."

Crowdsourcing is especially helpful and beneficial to transportation planning and management. It offers shared platforms and systems to invite a large amount of interested crowds to address common problems that influence them all. Recently, crowdsourcing techniques have developed rapidly. Some studies regarding its use in transportation have shown its tremendous potential in enhancing or taking the place of the traditional data collection methods. Since crowdsourcing has many advantages in data collection, it is leveraged in this research study.

With the availability of crowdsourced data, many models have been developed, such as linear regression models, ordered logit models, ordered probit models, (expanded) path size logit models, recursive models, C-logit models, and safety performance functions. These models can be applied to analyze the bicycle travels in terms of bicycle

volume estimation, bicyclist cycling behavior analysis, bicycle safety assessment and other topics including air pollution exposure studies, bicycle level of service evaluation and bicycling comfort analyses. To conduct these research studies, crowdsourced data are not sufficient. Other data including road characteristics, demographic features, geographic factors, temporal information, air pollution measures, and bicyclist involved crash records, etc. are needed to be compiled and integrated. In the data fusion process, software such as ArcGIS, SAS, SPSS, or R can be used to accomplish the task of data processing.

This research is intended to systematically develop bicycle volume prediction models, model cycling activities, and conduct bicyclist injury risk analysis with safety performance functions. Crowdsourced bicycle data from the Strava smartphone application are collected and combined with other relevant data (including NC road characteristics data, demographic data, slope data, manual count data from continuous count stations in Charlotte, temporal data, and bicycle facility data). Data comparison is conducted to demonstrate the difference between manual count data and Strava's bicycle count data. Data processing and combination procedures are completed using ArcGIS and SAS. Based on the combined data, two linear regression models are developed. The relationship between manual count data and Strava data as well as other relevant data is built. Bicycle volume on most of the road segments in the City of Charlotte is predicted using the developed model. A bicycle ridership map is created to display a graphical representation of the bicycle counts. In addition, a series of discrete choice models are developed to analyze the Strava users' cycling activities in the City of Charlotte. The bicyclist injury risk analysis is conducted based on the validated bicycle volume. Finally,

the conclusion is made to summarize the whole study, and directions for future research are also provided.

1.2. Study Objectives

The objective of this research is to evaluate and utilize the potential use of crowdsourced bicycle data in Charlotte to develop bicycle volume prediction models, model cycling activities, and conduct injury risk analysis with safety performance functions, as well as to map bicycle ridership and analyze biking safety influence. The detailed objectives are listed as follows:

1. To compile bicycle data from all the available sources including Strava data, bicycle manual count data, NC road characteristics data, demographic data, slope data, temporal data, bicycle facility data, and bicyclist involved crash records as preparation of the follow-up work;
2. To combine all the collected data using ArcGIS and SAS for model estimation;
3. To develop bicycle volume prediction models based on the combined data;
4. To calculate the predicted bicycle volume based on the developed models, and generate a bicycle ridership map for most of the road segments in the City of Charlotte;
5. To develop discrete choice models to explore the impact of different variables on Strava bicycle count in the city of Charlotte;
6. To develop safety performance functions based on bicycle volume for bicyclist safety analysis.

1.3. Expected Contributions

To provide a better cycling environment and encourage more potential bicyclists to bike in the City of Charlotte, models need to be developed to analyze the factors that affect bicycle volume on each roadway segment. Prediction of the bicycle volume on most of the roadway segments in the City of Charlotte should be conducted and used to provide guidance for the bicycle facility construction and improvement in the future. The impacts of biking safety need to be analyzed. Along that line, the expected contributions of this research are summarized as follows:

1. Present a systematic method for developing models to analyze the relationship between bicycle manual count data and Strava's bicycle count data that can be applied to other regions;
2. Generate a bicycle ridership map in the City of Charlotte to give an overview of the predicted bicycle volume that can be used as a reference for future bicycle facility construction/ improvement;
3. Develop discrete choice models to analyze the factors contributing to Strava bicycle counts in the City of Charlotte. Based on the model estimation results, the factors that have a positive impact on bicycle volume can be identified and used as the basis for bicycle policy recommendation;
4. Provide a method to develop safety performance functions for bicyclist injury risk analysis and mapping bicycle-vehicle crashes.

1.4. Research Overview

The research structure is organized as follows and Figure 1.1 will illustrate the whole research contents in summary.

In Chapter 1, the background of this research study is introduced, and the motivation of modeling cycling activities and conducting safety analysis are discussed. In addition, the objectives and expected contributions are described and presented in this chapter.

Chapter 2 provides a comprehensive review of the state-of-the-art and state-of-the-practice on the potential use of crowdsourced bicycle data. The data collection methods utilized for relevant research studies including crowdsourcing and other traditional data collection methods are summarized. Representative smartphone applications for crowdsourcing are presented and their use for different aspects of research is discussed. Methods for bicycle volume estimation and prediction, and bicyclist injury risk analysis are summarized.

Chapter 3 gives an overview of the collected data and conducts a descriptive analysis based on the data collected from Strava smartphone application in terms of users' demographics, different trip purposes, and Strava counts for different times of day, weekdays and weekends, months of year, and trip purposes. A simple data comparison between bicycle counts collecting from manual count stations and Strava application is provided. In addition, other supporting data are introduced in this chapter as well.

Chapter 4 presents a method for data processing and develops two linear regression models to analyze the relationship between bicycle manual count data and Strava data as well as other relevant attributes. The bicycle volume on most road segments in the City of Charlotte is predicted using the developed model. A bicycle ridership map is also created to display a graphical representation of the bicycle counts.

Chapter 5 develops a series of discrete choice models for conducting the analysis of impacts on cycling activities. In addition, the model comparison is conducted based on several indexes, and the best-fit model is also identified.

Chapter 6 provides a method to develop safety performance functions for bicyclist injury risk analysis. The method is based on the bicycle volume from previous chapter and other factors including bicycle facilities, annual average daily traffic (AADT), road characteristics, and the presence of bus stops. The indicators for model comparison are utilized to identify the best-fit model for bicyclist injury risk analysis.

Chapter 7 concludes this research with a summary of the methods for modeling cycling activities and conducting bicyclist injury risk analysis, and a discussion of the directions for future research.

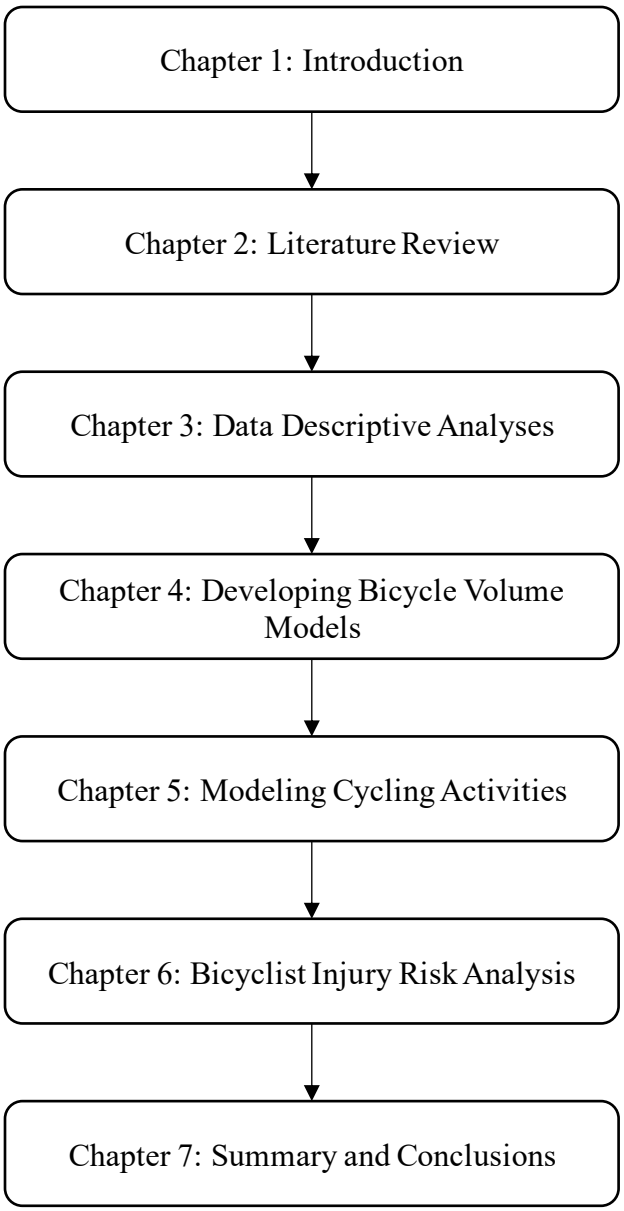


Figure 1.1: Research Structure

CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

This chapter presents a comprehensive literature review of current state-of-the-art and state-of-the-practice of relevant non-motorized transportation research studies, especially bicycle volume estimation and prediction, its impacts on bicycle activity, and bicyclist injury risk analysis. This literature review also summarizes the data utilized for the research studies, methods applied for bicycle volume estimation, prediction, and injury risk analysis and results concluded from previous and ongoing research.

The remainder of this chapter is structured as follows. Section 2.2 introduces different types of data collection methods such as crowdsourcing, open data, big data, and other traditional data collection ones including travel survey and count data. Section 2.3 summarizes the most prevalent smartphone crowdsourcing applications (e.g., CycleTracks, Cycle Atlanta, Mon RésoVélo, Strava, and ORcycle) and their use in different research aspects. Section 2.4 details the bicycle volume estimation and prediction methods based on both traditional data collection methods and crowdsourcing. Section 2.5 presents the approach to bicyclist injury risk analysis based on different types of data. Finally, section 2.6 concludes this chapter with a summary.

2.2. Data Collection

This section summarizes both the advanced and traditional data collection methods utilized for relevant research studies. An introduction to each type of data and the advantages and disadvantages of novel data and traditional data are provided.

2.2.1. Crowdsourcing

Crowdsourcing is an innovative sourcing model which brings new developments to data collection and the data-driven research studies. Crowdsourcing techniques have evolved rapidly since they emerged approximately ten years ago. Crowdsourcing was first introduced by Howe (2006) in "The Rise of Crowdsourcing", which was published in Wired Magazine in 2006 and was defined as follows:

"Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively) but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers." (Howe, 2006)

"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, 2008)

Crowdsourcing is a mixture of two components which are crowd and outsourcing. Based on the definition of crowdsourcing provided by Howe (2006), numerous scholars have been interested in the new concept of data collection method. Different definitions have emerged based on their understanding of crowdsourcing. According to Brabham (2008), crowdsourcing is "a strategic model to attract an interested, motivated crowd of individuals capable of providing solutions superior in quality and quantity to those that even traditional forms of business can". Later in Brabham's (2013) book, crowdsourcing was defined as "an online, distributed problem-solving and production model that

leverages the collective intelligence of online communities to serve specific organizational goals". Vukovic (2009) defined crowdsourcing as "a new online distributed problem-solving and production model in which networked people collaborate to complete a task". Instead of interpreting crowdsourcing as a model that solves the problems of the crowd through an online platform, Chanal and CaronFasan (2008) defined crowdsourcing as "the opening of the innovation process of a firm to integrate numerous and disseminated outside competencies through web facilities". Kleeman et al. (2008) found that the spirit of crowdsourcing is the intentional mobilization. The authors defined crowdsourcing as "a form of the integration of users or consumers in internal processes of value creation". To explain it simply, La Vecchia and Cisternino (2010) describe crowdsourcing as "a tool for addressing problems in organizations and business".

With the development of crowdsourcing, researchers have analyzed various definitions of crowdsourcing from different articles/papers to find out the features and common elements. Estellés-Arolas and González-Ladrón-De-Guevara (2012) reviewed and summarized the studies on crowdsourcing in terms of the information about the crowd and crowdsourcer, the tasks need to be conducted by the crowd, the benefit for the crowd and crowdsourcer, and the process of crowdsourcing. A definition of crowdsourcing integrated using the critical elements extracted from the previous literature was created which defined crowdsourcing as "a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task" (Arolas and González-Ladrón-De-Guevara, 2012). Other analysis as well as the summary of crowdsourcing can be

found in (Świeszczak and Świeszczak, 2016, Estellés-Arolas, Navarro-Giner, and González-Ladrón-de-Guevara, 2015, Hosseini et al., 2014).

To summarize, most of the crowdsourcing definitions contain three main features which are the crowd itself, the outsourcing procedure, and an internet-based platform (Saxton, 2013). It can be interpreted that crowdsourcing implied that individuals participate voluntarily to achieve the task which would tend to motivate both the experts and the individuals to find solutions to the tasks (Schenk, 2011). Table 2.1 presents a summary of the definitions of crowdsourcing in chronological order.

Table 2.1: Summary of Crowdsourcing Definitions

Author	Year	Definition
Howe	2006	"The act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call."
Brabham	2008	"A strategic model to attract an interested, motivated crowd of individuals capable of providing solutions superior in quality and quantity to those that even traditional forms of business can."
Chanal and CaronFasan	2008	"The opening of the innovation process of a firm to integrate numerous and disseminated outside competencies through web facilities."
Howe	2008	"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."
Kleeman et al.	2008	"A form of the integration of users or consumers in internal processes of value creation."
Vukovic	2009	"A new online distributed problem-solving and production model in which networked people collaborate to complete a task."
La Vecchia and Cisternino	2010	"A tool for addressing problems in organizations and business."
Brabham	2013	"An online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals."

With the development of crowdsourcing, it has brought improvement and benefit in data collection. This type of innovative data collection method shows its potential to augment the traditional data collection methods. Recently, Misra et al. (2014) studied how crowdsourcing was applied in transportation research area. In addition, as the

number of GPS-enabled smartphones increases, crowdsourcing with smartphones (Chatzimilioudis et al., 2012) sees more possibilities in transportation related research studies. A comprehensive summary of the existing smartphone applications utilized for different aspects of transportation research will be provided in the following section.

2.2.2. Open Data

Open data is another critical type of data that researchers might use for their studies. It can be easy to interpret that open data is open for anyone to use freely, and to reuse or redistribute it flexibly (Kitchin, 2014). In other words, an open data format should be "platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information" (Attard et al., 2015). Therefore, open data should be available for anyone with no additional restrictions or limitations.

Most of the open data are provided by institutions or local government. The government-related data is also called open government data which is a specific type of open data (Kučera et al., 2013). This type of data is provided by the government and is released openly to the public which usually contains public transportation information, crash records, population, infrastructure, and land use, etc.

2.2.3. Big Data

Big data is a general type of data that refers to large volumes of data from various sources that need to be cleaned and pre-processed before being utilized for research studies (McAfee et al., 2012). The primary attributes of big data are the '3Vs' which are volume (representing the size of the data), velocity (indicating the speed of the data collection or generation), and variety (referring to a synthetic range of sources) (Laney,

2001). Besides these three Vs, other researchers (Kitchin, 2014) have added other attributes to define big data including veracity demonstrating the quality of the data.

However, most of the big data utilizing in the transportation research area are under the "volume" feature, since the data sources of a large number of data are from a single application, internet platform or data provider. In the transportation research area, the expansion and development of the smart card system for transit in several major cities (Pelletier et al., 2011), the increasing popularity of smartphone applications, the availability of GPS devices, and the broad range of online information (Romanillos et al., 2016) have made great contribution to the development of big data.

2.2.4. Traditional Data Collection Methods

Traditional traffic data collection methods are the basic approach for data extraction which may not be replaced by some of the advanced data collection methods, since traditional data can provide accurate and useful information for relevant transportation research studies. Basically, there are two categories of traditional data collection methods which are data collected from traffic counting equipment and different kinds of travel surveys.

The commonly used traffic counting equipment include piezo-electric sensors, inductive loops, microwave, radar, video image detection, and manual observation, etc. (Skszek, 2001). Using the equipment to collect data may cost a lot for installation and may be time consuming during the whole time of the collection process.

The travel survey method can be divided into two categories which are web-based and paper-based travel surveys. The most well-known traditional travel survey is the household survey (Kagerbauer et al. 2015). Information relevant to their travel patterns

are collected through questionnaires. The process of filling out paper-based surveys and selecting useful and suitable answers usually takes a lot of time. The web-based travel surveys, on the other hand, are utilized later with smart filter management features. However, bias and other issues associated with this type of travel survey cannot be addressed and neglected. One of the problems with data collected from travel surveys may come from the respondents. Since young participants are able to get access to the internet more easily compared to old respondents, the proportion of young respondents might be higher than older ones. In addition, not all the questionnaires can be collected back, as the receiving rate can be lower than travel surveys conducted in person. Other traditional transportation survey methods such as workplace surveys, longitudinal and panel surveys, transit on-board ridership surveys, commercial vehicle (truck) surveys and external station surveys, usually have similar disadvantages.

From another perspective, travel surveys can be divided as two categories, which are stated preference surveys (i.e., SP surveys) and revealed preference surveys (i.e., RP surveys) (Guan, 2004). The SP survey is to receive decision-making results of the respondents in terms of certain different conditions. And the RP survey refers to the survey of completed selective behavior. The differences between these two kinds of travel survey are: (1) the questions of SP survey usually contain the investigation content which has not really occurred yet or is intentionally designed for the specific research topics, while the RP survey is a questionnaire regarding investigation questions which has already taken place; (2) the scenario in SP survey can be designed flexibly with assumed values of choices and attributes that are needed for the research studies, while the results of choices and choice conditions in the RP survey are based on actual travel

choice behavior. According to the features of these two types of surveys, the advantages of them are revealed. SP surveys are able to arbitrarily design the questionnaires and the corresponding scenarios for the future conditions which will benefit transportation planning and design especially for upcoming constructions and establishments. RP surveys are able to show the results or phenomenon hidden in each individual's choice which reflects the contribution of the impact factors and how individuals value these factors.

To summarize the traditional data collection methods, Figure 2.1 shows a clear structure of the traditional data collection methods mentioned in this section.

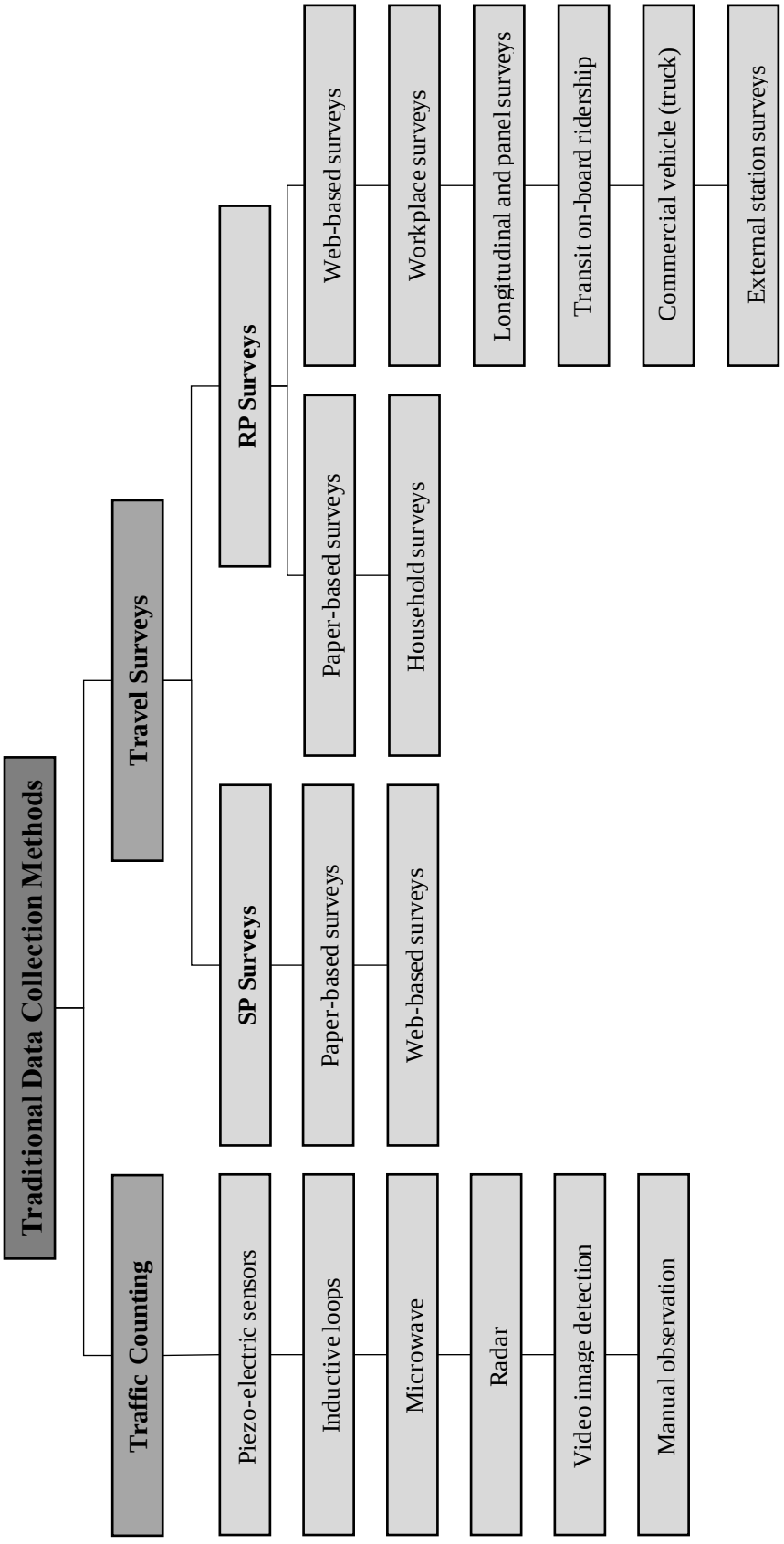


Figure 2.1: Traditional Data Collection Methods

2.3. Smartphone Crowdsourcing Applications

As mentioned in Section 2.2.1, there are numerous forms of crowdsourcing. Based on the literature review on bicycle-related research utilizing crowdsourced data, smartphone crowdsourcing applications are most prevalent for this innovative data collection method. There are multiple smartphone crowdsourcing applications that have been utilized for relevant research studies. This section provides a comprehensive summary of the cycling applications and their use to conduct different aspects of research studies.

2.3.1. CycleTracks

The CycleTracks application is the first smartphone crowdsourcing application developed for collecting crowdsourced bicycle data for bicycle-related research studies (Blanc et al., 2016). It was designed by the San Francisco County Transportation Authority (SFCTA) in 2009 to utilize the built-in GPS in smartphones to collect cycling information and users space trajectories. In addition, some of the users' demographic information can be collected (users optional answers to the demographic questions) to analyze the distinctive individual attributes for cycling behavior. The reported demographic information can be gender, age, home zip code, commute locations, and the frequency of cycling activities, etc. The comments field is provided for users to report their cycling trip purposes (e.g., commute, non-commute, recreational, exercise, shopping, school, work, social, etc.) during each trip (SFCTA, 2013). This information is also optionally filled in by CycleTracks users (Charlton et al., 2011).

Most of the studies used CycleTracks to analyze cyclists cycling behavior. Charlton et al. (2011) collected the cycling data from CycleTracks from November 12,

2009 to April 18, 2010 to analyze these cyclists' route choice in San Francisco. A total of 7,096 cycling trips generated by 1,083 cyclists were collected and selected as the chosen routes in the modeling procedure. A doubly-stochastic choice set generation method was utilized for modeling cyclists' route choice. The impacts of the length of the route, turns per mile, proportion of the route on wrong-way links, proportion on bike paths, bike lanes, and bike routes, infrequent cyclists, and average up-slope were considered in the path size multinomial logit model (Hood et al., 2011). Results revealed that the length of the route, turns per mile, proportion of route on wrong-way links, and average up-slope affected the cyclist's route choice negatively, while other explanatory variables had positive impacts on route choice.

Chen and Shen (2016) collected data from CycleTracks to analyze the impacts of road characteristics and land use on cyclists' route choice. Labeling route approach and the K-shortest mean approach were utilized to create the route choice set to conduct the cyclist route choice analysis. A path size logit model was developed, and results were concluded that cyclists selected their cycling routes based on the consideration of utility, cycling safety, and suitability. Subsequently, Chen et al. (2018) explored the influences of built environment on cyclist route choice based on the same dataset. Another comprehensive discrete choice model (i.e., path-size-based mixed logit model) was developed in this research study.

As the first application developed for cycling studies, researchers have compared this dataset with other data sources including traditional count data and data collected from other smartphone applications. Griffin and Jiao (2019) compared CycleTracks with traditional count data from five selected locations in Austin, Texas, and data provided by

Strava fitness application. The relationship between CycleTracks and count data, and the relationship between Strava and count data were examined utilizing ordinary least squares regression. In addition, spatial autocorrelation was also evaluated with OpenGeoDa software.

Based on the first smartphone crowdsourcing applications, other applications designed for cycling including AggieTracks, Cycle Atlanta, Mon RésoVélo, RenoTracks, CyclePhilly, Toronto Cycling App, ORcycle, CycleSac, and C-Vill Bike mAPP, etc. were developed subsequently (Blanc et al., 2016). Some of the applications that were utilized for bicycle-related research studies are introduced in detail as follows.

2.3.2. AggieTracks

As mentioned in the previous section, AggieTracks was developed based on the open source code of CycleTracks by Texas A&M University to collect the cycling information on the users in the university area (Hudson et al. 2012). Travel purposes were collected after the cycling trips by filling the corresponding questions optionally. Classifications (e.g., student, faculty, or staff) were asked to be identified by the AggieTracks users. Additional information such as users' living locations (on campus or not) and the car ownership was also collected. Since this application was developed to track cycling patterns within the university area, few research studies choose to utilize this data source.

2.3.3. Cycle Atlanta

Similar to AggieTracks, Cycle Atlanta was developed based on the open source code of CycleTracks application by the Georgia Tech research team (Figliozzi and Blanc, 2015). In addition to the cycling route data, Cycle Atlanta can collect other information,

such as demographic information including gender, email, age, ethnicity, income, zip codes of home, school, or workplace, etc., and selection information including issues (e.g., pavement issue, traffic signal issue, bicycle lane design issue, enforcement request, bicycle parking request, and custom entry) and amenities (e.g., water fountains, bike parking facilities, bike shops, and public restrooms).

Like other data collected from smartphone applications, cycling information data extracted from Cycle Atlanta were compared with other types of cycling data including manual count data as well as data from other applications. Watkins et al. (2016) conducted a study to compare data collected from Cycle Atlanta and Strava in terms of demographic data, cycling trip information, time of day, and different road segments to examine the ability of GPS data from smartphone applications for mapping cyclist ridership. In addition, manual count data were compared with data from Cycle Atlanta to investigate the proportion of Cycle Atlanta users to the total cyclists.

Later on, Cycle Atlanta data were utilized for route choice modeling, street segment choice, evaluating bicycle level of service (BLOS), and measuring level of traffic stress (LTS). In a USDOT final report named “Using Crowdsourcing to Prioritize Bicycle Route Network Improvements”, LaMondia and Watkins (2017) conducted a research study on calculating BLOS, measuring LTS, modeling bicyclist route choice, and route segment choice using data collected from three smartphone applications which were Strava, CycleDixie and Cycle Atlanta. An ordinal logistic regression model was developed to analyze the route segment choice of Cycle Atlanta users. Explanatory variables including roadway characteristics, access groups, and socio-demographic accessibility were included in the model. To analyze the willingness of a bicyclist to

choose a detour over the shortest route, a binary logistic choice model was developed based on the alternative choice (i.e., shortest route) created by the A-star algorithm.

In addition, different perspectives of bicyclist route choice modeling research studies were conducted. Misra and Watkins (2018) investigated the differences of bicyclist route choice between different genders and age groups. Multiple path size logit models were developed for different segmentations in terms of age and gender. Also, a pooled path size logit model based on the entire Cycle Atlanta cycling data were developed for comparison. Results revealed that traffic characteristics including annual average daily traffic (AADT) and speed might affect the cyclist route choice differently for different genders and ages of cyclists.

2.3.4. Cycle Lane

Cycle Lane is another smartphone application that is built on the code developed for CycleTracks (Roll, 2014). In order to collect the bicycle trip information, Central Lane Metropolitan Planning Organization (CLMPO) developed this Cycle Lane application in 2011. Demographic information (including age and gender) on the cyclists using this application can be collected through questions that are asked to fill out in the application. Additional information such as the frequency of riding is also collected before cycling.

Zimmermann et al. (2017) modeled the bicyclist route choice using the Cycle Lane data to analyze the trade-offs that cyclists made while selecting cycling routes. Zimmermann et al. developed a recursive logit model for the bicyclist route choice modeling since this type of link-based route choice model does not require one to generate route choice sets compared to the path-based models such as path size logit

model. According to the results concluded, this recursive logit model may save a lot of computational time.

2.3.5. Mon RésoVélo

Mon RésoVélo is also a smartphone application for collecting bicyclist route information in the City of Montreal based on CycleTracks as well as Cycle Atlanta. Cycling trip information including travel distance, travel time, cycling route are collected. In addition, socio-demographic information and other attributes of the cyclists using this application are obtained through an anonymous questionnaire. Different from the two applications that Mon RésoVélo was built on, this application adds an emission tool and a calorie calculator, which is a new development for the smartphone crowdsourcing application for cycling (Jackson et al., 2014).

Based on the GPS cycling trip data from Mon RésoVélo, deceleration rates at intersections and on road segments were extracted by Strauss et al. (2017) to explore the relationship between the deceleration rate (DR) and the number of injuries. The site ranking based on the deceleration rate and the expected injury number were compared utilizing Spearman's rank correlation coefficient.

In addition, with the benefit of this innovative smartphone application, many other research studies were conducted based on the data extracted from Mon RésoVélo. Strauss and Miranda-Moreno (2017) utilized the GPS cycling data from Mon RésoVélo to identify the performance measures in terms of delay, speed, and travel time on road segments and at intersections in the whole city network on the island of Montreal. To examine the impacts of geometric design and built environment on cycling speed on each road segment, a linear regression model was adopted. The model results demonstrated

that cycling speeds were higher along arterials than on local streets, and cyclists biked faster on road segments with bicycle infrastructure. Furthermore, impact factors including geometry characteristics, built environment features, travel purposes, peak hours were found to affect the cycling speed significantly.

2.3.6. RenoTracks

RenoTracks is a smartphone application that builds based on Cycle Atlanta. Therefore, all features were included in RenoTracks since CycleTracks. Similar functions can be provided in this application including recording cycling information, collecting travel distance, calculating travel speed, reporting issues, and collecting demographic data from cyclists, etc. (RenoTracks 2013). In addition to these features adopted from the previous smartphone crowdsourcing applications, RenoTracks created a customized user interface as well as a "CO_2 Saved" calculator to compute the CO_2 that could be saved compared to traveling by automobile.

2.3.7. ORcycle

Portland State University and Oregon Department of Transportation developed a ORcycle based on the code for CycleTracks to collect cycling information from the application users. This application was released for both Android and iOS platforms in November 2014. Using this application, cycling data included bicycle trip trajectories, user information, infrastructure issues, and crashes.

Therefore, with the help of ORcycle, useful data can be collected to design and upgrade better bicycle facilities and analyze the impacts on cyclists' comfort levels. Blanc and Figliozzi (2016) leveraged ORcycle application to collect data for cyclists' comfort level modeling. Factors including bicycle facilities, sources of stress associated

with the cycling routes, travel purposes, distance, cycling frequency, and temporal characteristics were considered in the model. Ordinal logistic regression models were developed to examine the influence on cyclists' comfort levels. Based on the model results, bicycle boulevards, separated cycling paths, sources of stress associated with the cycling routes, trip purposes, and cycling distance were found to affect cyclists' comfort levels significantly.

ORcycle data can also be utilized for safety analysis. Blanc and Figliozzi (2017) investigated the impact factors on the urgency of a perceived potential safety issue. Based on the statistical models, application users are usually reliable for the reported urgency of safety issue and the infrastructure problems. The factors affected safety urgency and type include user gender and income levels, traffic volumes, speed, and waiting times at signalized intersections.

2.3.8. MapMyRide

MapMyRide is a sub-application of MapMyFitness, which is created to get the most from the users' bike ride and track their cycling trips especially for recreational travel purposes. This application allows users from worldwide to plan their cycling route, track their GPS trajectories, share links with others, and provide user information. Cyclists using MapMyRide can view others' cycling route through this application to follow the popular cycling routes for comfortable and challenging activities. In addition to the smartphone application, MapMyRide also provides a web version which can present and summarize the statistics and the ridership of the users' cycling trips (Figliozzi and Blanc, 2015).

As a smartphone application that can collect cycling data from the entire United States, MapMyRide provides data for physical activity patterns investigation. Hirsch et al. (2014) utilized data collected from MapMyFitness to analyze the users' physical activity patterns. It was concluded that this set of applications is a critical and useful platform to explore travel patterns within large geographic and temporal scales.

2.3.9. Strava

Strava is the one of the best cycling applications for 2019 especially for tracking recreational cycling trips in large cities (Best cycling apps, 2019). Similar to MapMyRide, Strava allows users to track their cycling routes through the GPS-enabled smartphone and view and share the trip trajectories afterwards via website or application. Summary statistics including travel speed, trip distance, activity time, and other cycling route information are provided and displayed. The unique features that Strava has are the ability to track cycling performance of multiple cyclists on the same segment which enables Strava users to compete with each other for the least segment time, highest speed, etc. This particular functionality attracts numerous cyclists worldwide to use this smartphone application for recording their cycling trips which provides Strava a large dataset in extensive geographic and temporal scales.

With the large dataset, Strava has become one of the most prevalent smartphone applications to collect cycling information from a variety of users. Multiple bicycle-related research studies in different aspects were conducted based on Strava data.

Sun and Mobasheri (2017) utilized Strava data to analyze the spatial patterns of cycling activities for different travel purposes and air pollution exposure in a large spatial scale. The improved Multidirectional Optimum Ecotope-Based algorithm was utilized to

identify the clusters associated with high non-commuting rate. Ordinary least squares, support vector machine, random forest, and multilayer perceptron neutral network methods were used to analyze the Strava users' non-commuting cycling activities. Results were found that more non-commuting cycling trips occurred in the outskirts of the city. In addition, bicyclists biking for commuting were found to have a higher probability to suffer from more severe air pollution.

Other research studies conducted based on Strava data include non-motorized transport planning (Selala and Musakwa, 2016), cycling patterns and trends (Musakwa and Selala, 2016), cycling behavior (Sun et al, 2017), bicycle trip volume (Hochmair et al., 2019), etc.

To summarize the literature reviewed in this Section 2.3 regarding the summaries of smartphone crowdsourcing applications developed for cycling information collection and research studies based on the data extracted from the smartphone applications, two tables are presented as follows.

Table 2.2: Summary of Smartphone Crowdsourcing Applications

Year	Applications	Developer	Information Collected	Emphasis
2008	MapMyRide	MapMyFitness	Demographic information Travel purpose Cycling trajectories	Physical activity patterns analysis
2009	CycleTracks	SFCTA	Demographic information Travel purpose Cycling trajectories	Route choice modeling
2009	Strava	Strava Metro	Demographic information Travel purpose Cycling counts	Non-motorized transport planning Air pollution exposure Cycling patterns and behavior Bicycle volume Active travel and health
2011	AggieTracks	Texas A&M University	Trip purpose On campus living Car ownership	Travel patterns analysis
2011	Cycle Lane	CLMPO	Demographic information Travel purpose Cycling trajectories	Route choice modeling Bicyclist behavior analysis

2012	Cycle Atlanta	Georgia Tech	Demographic information Travel purpose Issue reporting Amenity reporting	Route choice modeling Bicycle volume LTS
2013	Mon RésoVélo	The City of Montreal	Demographic information Travel purpose Cycling trajectories Calorie Emissions	Level of service measures Safety analysis
2013	RenoTracks	2013 Hack4Reno Team	Demographic information Travel purpose Cycling trajectories "CO_2 Saved"	Cycling data analysis
014	ORcycle	Portland State University and ODOT	Demographic information Travel purpose Cycling trajectories Infrastructure issues Crashes	Cyclists' comfort level Route choice modeling Crash and injury risk modeling

Table 2.3: Summary of Research Topics Based on Crowdsourced Bicycle Data

Year	Author	Data Source	Study Area	Data Size	Methods	Research Area
2011	Hood et al.	CycleTracks	San Francisco	7,096 cycling trips generated by 1,083 cyclists	Path size multinomial logit model	Route choice modeling
2014	Hirsch et al.	MapMyFitness	Winston-Salem, NC	43,872 unique workouts by 3,094 unique users	Statistical analyses	Physical activity patterns analysis
2016	Blanc and Figliozzi	ORcycle	Portland, OR	729 trips from 170 users	Ordinal logistic regression models	Cyclists' comfort levels
2016	Chen and Shen	CycleTracks	Seattle	543 observations	Path size logit model	Route choice modeling
2017	LaMondia and Watkins	Cycle Atlanta Strava CycleDixie	Auburn, AL Atlanta, GA	5,201 trips generated by 458 cyclists	Ordinal logistic regression model, Binary logistic choice model	Route segment and path choice modeling
2017	Strauss et al.	Mon RésoVélo	Montreal	More than 10,000 trips generated by about 1,000 cyclists	Spearman's rank correlation coefficient	Safety measure
2017	Strauss and Miranda-Moreno	Mon RésoVélo	Montreal	More than 10,000 trips generated by about 1,000 cyclists	Linear regression model	Performance measures

2017	Sun and Mobasheri	Strava Scottish Air Quality Database	Glasgow, United Kingdom	287,833 cycling activities contributed by 13,684 users	Ordinary least squares, multilayer perceptron neutral network, random forest, support vector machine	Cycling activities and air pollution exposure
2017	Zimmermann et al.	Cycle Lane	Eugene	648 bike trips from 103 users	Recursive logit model	Route choice modeling
2018	Chen et al.	CycleTracks	Seattle	3,310 routes created by 197 cyclists	Path-size-based mixed logit model	Route choice modeling
2018	Misra and Watkins	Cycle Atlanta	Atlanta, GA	About 20,000 trips by 1,495 users	Path size logit models	Route choice modeling
2019	Griffin and Jiao	Traditional Count Data, CycleTracks, Strava	Austin, Texas	183,070 counts and 111 CycleTracks records, 4,372 counts and 209 Strava records	Ordinary least squares regression and spatial autocorrelation	Bicycle volume

2.4. Bicycle Volume

This section reviews the research studies regarding different methods of bicycle volume estimation and prediction based on different types of data (obtained from both traditional data collection methods and crowdsourcing). The potential impact factors that might affect bicycle volume or cycling activities significantly are summarized through the review of the state-of-the-art and the state-of-the-practice.

2.4.1. Research Based on Traditional Data Collection Methods

Although crowdsourcing is an innovative data collection method, the importance of traditional data collection methods cannot be neglected. Manual count data and automated counting data are the basic traditional data collected for annual average daily traffic (AADT) estimation. Many research studies are conducted based on this kind of data.

To synthesize the approach to estimating AADT with non-motorized traffic monitoring, Lu et al. (2017) utilized three types of automated counters including pneumatic tube, radio beam, and passive infrared to collect long-term counts, and collected manual counts for short duration. A strong correlation was found between these two types of data. NB models were developed for each site to estimate bicycle and pedestrian volume. In addition, to estimate annual average daily traffic, day of year scaling factors were applied for both non-motorized traffic counts. The volume of bicycles and pedestrians were found to be positively affected by street functional class, certain facilities for bicyclists and pedestrians, and proximity to campus.

Chen et al. (2017) investigated the impacts of built environment explanatory variables on bicycle volume. A dataset of five-year bicycle volume in Seattle,

Washington was utilized to develop a generalized linear mixed model (GLMM) assumed to follow a Poisson distribution in order to model the variation of bicycle counts over time. Model results indicated that exploratory variables including temporal characteristics such as weekends and peak hours, bicycle facilities, non-winter seasons, employment density were likely to affect bicycle volume positively. Lower bicycle volume was associated with steep areas, while areas with more mixed land use, water bodies, and workplaces were found to be high bicycle volume locations.

Miranda-Moreno et al. (2013) classified the bicycle volume data collected from 38 sites in five North American cities into four categories including recreational, mixed recreational, mixed utilitarian, and utilitarian. The variation of bicycle volume in terms of different times of day, days of week, months, and seasons was analyzed using standardized hourly, daily, and monthly indexes, as well as traffic distribution indexes.

Esawey (2014) conducted a research study on estimating AADB with daily adjustment factors (DAFs) and monthly adjustment factors (MAFs). Bicycle count data collected from 12 permanent counting stations in Vancouver were utilized for adjustment factor calculation. Subsequently, the calculated factors were used to estimate annual average daily bicycle counts at other counting stations.

The standard K factor is another type of adjustment factor using for bicycle volume estimation and calculation. Esawey and Mosa (2015) developed the standard K factors (i.e., Kp/d and Kp/AADB) and provided an example of AADB calculation using the developed standard K factors. Furthermore, the estimation accuracy based on the K factors was examined.

To address the issue of missing bicycle count data at counting stations, Esawey et al. (2015) developed an innovative model called autoencoder neural network to fill in data gaps and estimate missing daily bicycle volume using available data from nearby and at the same location. The model parameters that might influence the estimation accuracy were assessed and the sensitivity analysis was conducted.

Considering the impacts of weather and seasonal factors, Schmiedeskamp and Zhao (2016) investigated the relationship between these factors and bicycle volume based on the automated bicycle counts collected from Seattle, Washington. A NB model was developed, and counterfactual simulation was used to estimate quantities of interest. Model results demonstrated that variables including season, holidays, day of week, temperature, and precipitation might affect the bicycle volume significantly.

Similarly, Lewin (2011) also analyzed the impact of temporal and weather factors on bicycle volume. A standard linear regression model was developed based on the detector data from two permanent bicycle count stations on multi-use paths in Boulder, CO. The variables included in the model were carefully selected considering the temporal patterns of bicycle volume and weather correlation results. The bicycle volume was then estimated using the developed linear regression model.

To conclude, a summary of the studies on bicycle volume estimation and analysis using traditional manual count data or automated count data is provided below in Table 2.4.

Table 2.4: Summary of Bicycle Volume Studies Using Bicycle Count Data

Year	Author	Bicycle Data	Study Area	Methods	Variables
2011	Lewin	Detector data	Boulder, CO	Linear regression model	Temperature, weather condition (e.g., rain and snow), weekend
2013	Miranda-Moreno et al.	Long-term automated counting data	Montreal, Ottawa, Vancouver, Portland, and San Francisco	Standardized hourly, daily, and monthly indexes, and traffic distribution indexes	Bicycle volume in terms of time of day, day of week, month of year, and seasonality
2014	Esawey	Bicycle volume from inductive loop counters	Vancouver, British Columbia	DAFs and MAFs	Bicycle volume in terms of time of day, day of week, month of year, and seasonality
2015	Esawey and Mosa	Bicycle volume from inductive loop counters	Vancouver, British Columbia	$K_{p/d}$ and $K_{p/AADB}$	Bicycle volume in terms of time of day, day of week, month of year, and seasonality
2015	Esawey et al.	Bicycle volume from inductive loop counters	Vancouver, British Columbia	Autoencoder neural network models	Daily bicycle volume
2016	Schmiedeskamp and Zhao	Automated bicycle counts	Seattle, Washington	Negative binomial model	Hours of daylight, university in session, holiday, temperature, precipitation, day of week, season,
2017	Chen et al.	SDOT[1] bicycle count data	Seattle, Washington	GLMM	Winter, peak hours, weekends, land use, bicycle facilities, road characteristics, steep areas, demographics

2017	Lu et al.	Automated count data and validation counts	Blacksburg, VA	Negative binomial regression models	Daily temperature variation, daily max temperature, precipitation, windspeed, weekend, proximity to university

[1] Seattle Department of Transportation

2.4.2. Research Based on Crowdsourcing

Many researchers have conducted their studies using crowdsourced data. GPS enabled smartphones provide researchers new opportunities to collect data from a broader group of people and use them to conduct the research on bicycle volume estimation and prediction. The existing use of crowdsourced data for this research area is presented as follows.

Moore (2015) conducted a study to examine the impact of various factors on bicycle counts using 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 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 the manual count and crowdsourced 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 to predict 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.

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) used 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 that they can be utilized as a critical supplement to bicycle volume estimation in large areas.

Cycling activity 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, land use, and road characteristics, have significant impacts 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.

To conclude, a summary of the studies on bicycle volume estimation and prediction as well as cycling activity analysis is provided below in Table 2.5.

Table 2.5: Summary of Bicycle Volume Research Based on Crowdsourced Data

Year	Author	Bicycle Data	Methods	Results
2015	Moore	Data from Strava application	Ordinal logistic regression model	Land use, demographic information, and road characteristics have significant impacts on cycling route choice.
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 counting data	Generalized linear model	In mid-size North American cities within urban areas, the routes 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	Variables that have significant impacts on cycling route choice include land use, demographics, and road characteristics.
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.

2.5. Bicyclist Injury Risk Analysis

Bicyclist injury risk analysis is another critical research concentration that needs to be studied to better understand the variables contributing to high injury risk and consequently help provide greener and safer cycling environment and promote biking in large bicycle-friendly cities.

Many research studies have been conducted to explore the bicyclist injury risk using different functions and models from various perspectives. Strauss et al. (2013) are the researchers who were interested in bicyclist activity and injury risk, and a series of studies were conducted with multiple modeling approaches and different types of data.

Strauss et al. (2013) used a Bayesian modeling approach to analyzing cycling activity and bicyclist injury risk at signalized intersections simultaneously. Impact factors contributing to both bicyclist injury risk and bicycle volume were identified. This two-equation modeling method reveals the potential existence of endogeneity and unobserved heterogeneities and can also be applied to find the high-risk locations. The data utilized for this research study included bicycle volume data and motor-vehicle counts collected at 647 signalized intersections by Montreal Department of Transportation, geometric design, built environment, bicycle facilities, and bicyclist injury data. Temporal and weather adjustment factors were applied for manual bicycle counts normalization to calculate AADB. Results revealed that higher bicycle volume would lead to more bicyclist injuries yet lower bicyclist injury risk. In addition, total crosswalk length and bus stops were found to increase the likelihood of bicyclist injuries, while raised medians might have the opposite influence.

Later on, a research study was conducted by Strauss et al. (2014) to analyze multimodal injury risk including motor-vehicle, pedestrian, and bicyclist injury risk and activities at signalized intersections as well as non-signalized intersections. Like the previous research, a Bayesian modeling approach was utilized for safety and volume analysis simultaneously based on the same dataset along with the injury and volume data collected from 435 more non-signalized intersections. Afterwards, the Bayesian multivariate Poisson models were calibrated and the explanatory variables contributing to injury frequency were determined. A comparison of injury risk for different modes for both intersection types was conducted. Results found that motor-vehicle traffic is the primary cause of all multimodal injuries at signalized intersections and non-signalized intersections. Additionally, bicyclists and pedestrians have a much higher injury risk on average compared to motorists at signalized intersections. Factors including built environment and some geometric design were found to have a significant impact on injury risk for all three types of road users.

Furthermore, with the development of crowdsourcing, smartphone GPS data collected from numerous applications were utilized for estimating bicycle volume as well as bicyclist injury risk analysis. Strauss et al. (2015) introduced an approach to estimating bicycle volume and map ridership and bicyclist injury risk in the whole city network in Montreal for both roadway segments and intersections based on data collected from Mon RésoVélo smartphone application as well as the manual count data. An extrapolation function approach was applied to combine the manual count bicycle data with crowdsourced bicycle data for bicycle volume estimation. Then, safety performance functions (SPFs) were developed based on the estimated AADB to validate the predicted

AADB by comparing the parameter coefficients with the previous SPFs using manual count data. After calibration, the annual average daily bicycle function can be adopted to predict bicycle volume at intersections and on all the road segments within the whole city network. Then, statistical models were utilized to compute empirical Bayes (EB) for bicyclist injury risk analysis. Injury risk maps can be generated to illustrate the distribution of bicyclist injury. According to the results, more injuries and higher injury risk occurred at signalized intersections compared to non-signalized intersections. On average, more injuries occurred on segments with cycle tracks, yet the injury risk per bicyclist was lower because of the presence of cycle tracks.

In addition to Mon RésoVélo smartphone application, data from Strava can also be utilized for the bicyclist injury risk analysis. According to a research study conducted by Wang et al. (2016), bicycle safety performance functions including negative binomial model, Poisson regression model, and zero-inflated negative binomial model, were developed based on crowdsourced bicycle data. After model estimation, the best model for SPF was identified utilizing the likelihood ratio test and Vuong non-nested hypothesis test. The comparison results revealed that negative binomial model outperforms Poisson regression model, and normal negative binomial model performs better than the zero-inflated negative binomial model.

Similarly, Saad et al. (2019) estimated safety performance functions for bicyclist injury risk analysis at intersections based on the crowdsourced bicycle data collected from Strava application. Strava data were adjusted before being utilized as the input of safety performance functions. Models based on the original Strava data, the Strava data with manual bicycle count data adjustments, and Strava data with adjusted population

were developed and compared. Negative binomial models were developed to predict bicycle crash at intersection. The model estimation results demonstrated that the adjusted Strava data with adjusted population and manual counts perform best in bicyclist injury analysis. In addition, impact factors including signal control system, bicycle lanes, and intersection size etc. would affect bicyclist injury at intersections.

Chen (2017) utilized a data-driven method to build the bicycle safety performance functions in both micro and macro scales using Strava smartphone application data, automatic bicycle count data and reported crash data. Negative Binomial model, Poisson model, Zero-inflated Negative Binomial model, and Zero-inflated Poisson model were developed to predict intersection crash frequency. A likelihood ratio test was utilized to identify the explanatory variables that affect crash frequency significantly. Similarly, the safety performance functions were developed for corridor crash frequency as well. Crash severity distributions were adopted in the bicycle crash frequency prediction models.

Another approach to identifying injury risk factors other than developing SPFs using smartphone applications is to collect volunteered geographic information (VGI) from cyclists through websites or applications. von Stülpnagel and Krukar (2018) assessed this type of crowdsourced data as well as the authoritative data as the indicators for biking risk analysis. Volunteered bicyclists were asked to conduct the laboratory-based virtual reality experiments to estimate their risk perception. Bicyclists were divided into two groups for separate cases. The first group was tested as experienced and frequent bicyclists who are not familiar with the selected test locations. The second group, on the contrary, was tested as bicyclists who are both experienced and familiar with the test locations. After that, the indicators of biking risk were obtained from the volunteered

geographic information. Therefore, based on the indicators from VGI and collected authoritative data, biking risk perception was estimated using linear mixed-effect models. The model results revealed that the semantic severity described for cycling hazard and the public response to the hazard might affect the risk perception significantly. Based on the authoritative data, a Space Syntax analysis was conducted which demonstrated the bicyclist sensitivity to street size and complexity.

Jestico (2016) utilized crowdsourced bicycle data to conduct research on bicycle ridership and cycling safety analysis. The bicyclist safety and injury risk were analyzed based on the bicycle volume in the certain area estimated using crowdsourced bicycle data collected from Strava. Manual count data at intersections during peak hours were also collected and used to compare with the crowdsourced data with a generalized linear model. Results indicated that time of year, slope, traffic speeds, and on street parking might affect the bicycle volume significantly. Based on the estimated bicycle volume, bicyclist injury risk was analyzed using Poisson generalized linear model based on the incident reports obtained from www.BikeMaps.org to examine the impacts of various factors. Results revealed that motor-vehicle and bicycle volumes and lack of deceleration factors were found to affect accident frequency significantly.

Al-Fuqaha et al. (2017) developed and utilized a smartphone application called BikeableRoute to analyze the risk factors based on crowdsourcing. This application enables bicyclists to report hazards during their cycling trips as well as to track their cycling information. The data collected from this application included risk report generated by bicyclists, user evaluation on the bike ability of cycling routes, and cycling information such as distance, cycling time, and speed. Using the data from

BikeableRoute, risk factors were categorized into three groups which are infrastructure-related, facility-related, and traffic-related factors. An ordered probit model was developed to analyze the perception of narrow bicycle lanes in terms of different ages and skill levels. Results revealed that bicyclists from different age groups have different perception of hazard. Table 2.6 provides a summary of research on bicyclist injury risk analysis.

Table 2.6: Summary of Research on Bicyclist Injury Risk Analysis

Year	Author	Research Objectives	Bicycle Data	Study Area	Methods	Variables
2013	Strauss et al.	Bicyclist activity and injury risk analysis	Manual bicycle counts and bicyclist injury data	Montreal, Quebec, Canada	Two-equation Bayesian modelling approach	(1) Bicyclist injury model: Bicycle volume, vehicle right turn and left turn flows, bus stops, crosswalk length, raised median; (2) Bicycle volume model: Employment, schools, metro stations, land use, bicycle facility length, three approaches.
2014	Strauss et al.	Multimodal injury risk analysis	Manual bicycle counts and bicyclist injury data	Montreal, Quebec, Canada	Bayesian multivariate Poisson models	(1) Bicyclist injury risk at signalized intersection: Bicycle volume, vehicle right turn and left turn volume, bus stops, crosswalk length, raised median; (2) Bicyclist injury risk at non-signalized intersection: Bicycle volume, vehicle volume, number of lanes.
2015	Strauss et al.	Bicyclist activity and injury risk analysis	Manual bicycle counts, smartphone GPS data, and bicyclist injury data	Montreal, Quebec, Canada	Extrapolation function and negative binomial SPF model	(1) AADB: Bicycle facilities (cycle path, cycle track, bicycle lane, etc.), distance to downtown; (2) Injury models for signalized intersections: Bicycle volume, bus stops, three approaches; (3) Injury models for non-signalized intersections: Bicycle volume, arterial or collector, three approaches; (4) Injury models for segments: Bicycle volume, arterial or collector, downtown boroughs.

2016	Jestico	Ridership trends and safety	Manual bicycle counts, Strava data, and incident reports from BikeMaps.org	The Capital Regional District (CRD), British Columbia (BC), Canada	Generalized linear model with a Poisson distribution	(1) Bicycle volume model: Strava counts, slope, population density, pavement widths, on-street parking, speed limit, bike facilities (e.g., paved trails and paited bike lanes), and month. (2) Incident model: Bicyclist and vehicle volume, speed reduction factors.
2016	Wang et al.	Bicycle safety analysis	Strava data and bicycle crash data	Seattle, Washington; Portland, Oregon	Poisson, NB, ZINB, etc.	AADT & AADB
2017	Al-Fuqaha et al.	Non-motorized behavior analysis and risk factor identification	Crowdsourced data from BikeableRoute	Kalamazoo, Michigan	Web survey and order probit model	Bicyclist skill level, age, gender, bicycle facility (e.g., narrow bicycle lane).
2017	Chen	Crash frequency prediction	Strava data, automatic bicycle count data, and reported crash data	Portland, Oregon and Eugene-Springfield, Oregon	NB, Poisson, ZINB, and ZIP models	(1) Crash frequency for intersections: Strava counts, AADT, network density, directions, bike lane, total lanes, signal, leg number. (2) Crash frequency for corridors: Signal/mile, median, bus route number, on-street parking, two-way left turn lane.
2018	von Stülpnagel and Krukar	Risk perception	Crowdsourced and authoritative data	Munich and Freiburg, Germany	Linear mixed-effect models and Space Syntax analysis	Semantic severity, number of votes, street size, traffic volume, complexity, accident category, familiarity.

2019	Saad et al.	Bicycle safety analysis	Strava data and bicycle crash data	Orange County, Florida	Negative binomial models	TEV[1], bicycle exposure, intersection size, number of legs, signal control system, bike lane, speed limit, median width, sidewalk width,.

[1] Total entering volume

2.6. Summary

A comprehensive review and the current and previous research studies regarding different kinds of data collection methods including crowdsourcing, open data, big data, and other traditional data collection methods were presented at the first section of this chapter. Then, the most prevalent smartphone crowdsourcing applications and their use on relevant research studies were summarized. Furthermore, the methods that were applied by researchers to estimate and predict bicycle volume were provided. Finally, bicyclist injury risk analyses conducted based on different types of data were discussed. This is to provide a solid reference and assistance in bicycle volume estimation and prediction, and injury risk analysis in future chapters.

CHAPTER 3: DATA DESCRIPTIVE ANALYSES

3.1. Introduction

As mentioned before, the first step of this research is to collect crowdsourced bicycle data from Strava application and other relevant supporting data. This chapter gives an overview of the collected Strava bicycle data, and other essential supporting data for the later model development. The descriptive analyses of the collected Strava data are also provided. Data comparison is conducted between bicycle manual count data from the continuous count stations in Charlotte and the Strava bicycle data collected from the smartphone application.

The following sections are organized as follows. Section 3.2 gives a brief introduction to Strava. Section 3.3 presents the Strava metro delivery. Section 3.4 shows the other relevant supporting data collected for this research. Section 3.5 describes the Strava bicycle count data in terms of the heatmap based on different months of year, trip purposes, weekdays and weekends. Section 3.6 gives a data comparison between bicycle manual count data from the continuous count stations in Charlotte and the Strava bicycle data from the smartphone application. Finally, Section 3.7 concludes this chapter with a summary.

3.2. Introduction to Strava

The field of possible GPS data has certainly been changing over time. The most commonly used solutions today are the data from smartphone application with completely different user structures and data types (such as Strava), data from bicycle hire systems, or data collected from local initiatives. Most of the smartphone applications including Strava tend to record route data directly collected from the users that utilized

this application, together with the demographic information about the users derived from the application. Such data contain various aspects of sensitive information, such as the user's place of residence or workplace. Such information can also be related to profile information such as name, gender, age, and other freely provided information. When passing data on to third parties, it is obliged to anonymize the users' sensitive information according to the data protection laws and general conditions of business. Therefore, the buyers will only receive data that have already been aggregated by the vendors and cannot trace back to the people that generated the data. Anonymized demographic information (such as gender and age) is aggregated and permitted to remain in the dataset. Such data generated from global vendors of smartphone applications will provide information about the largest range and number of possible users. Considerable differences can exist within the user structure. The route data are collected from each user on a second to second basis, saved at the end of the trip and transmitted to a server. The saved data can then be viewed by users on their smartphones and shared their trip information with others. This allows the application (such as Strava) users to share their recent routes with others or keep a training journal.

The cycling data utilized in this research 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 showing the information about cycling distance, time, and speed, etc. is presented in Figure 3.1. Most of the users are cyclists or runners. When a cyclist or runner uses the Strava application, his/her trip information including distance, elevation change, trip duration, and average speed is recorded. In addition, the cycling route will also be saved in the application. This allows

users to be able to look and see their cycling trajectory, and how well they performed each time, and even compared with other users on the same segment/route.

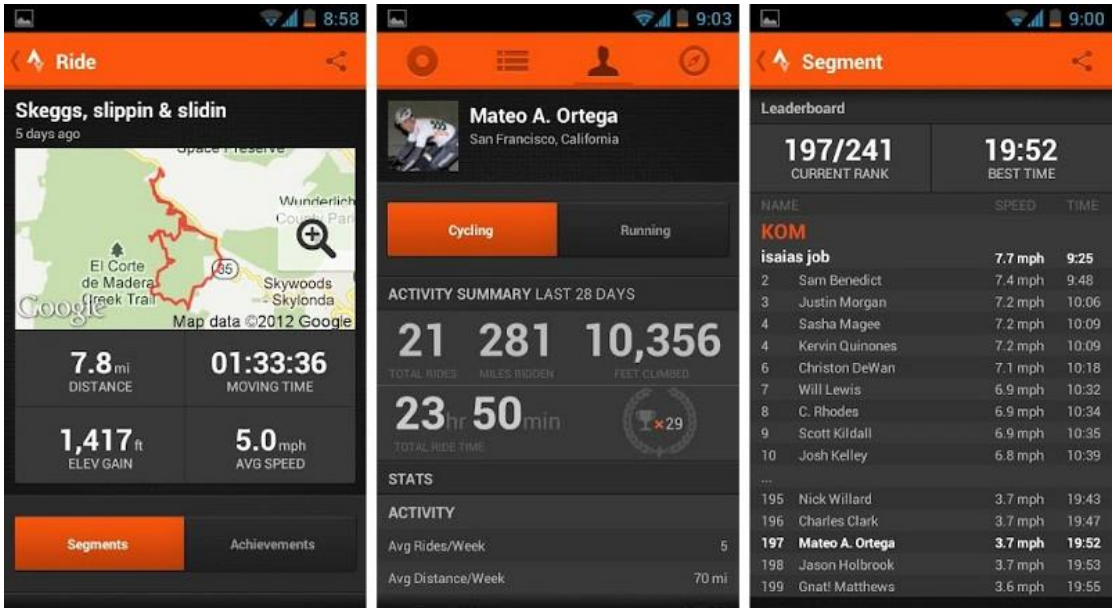


Figure 3.1: Strava App Screen Shots

3.3. Strava Data

The GPS data collected from the Strava users usually include the biking information for both the link-level and the intersection-level of the network. The link-level data set contains the Strava user counts on roadway segments and the intersection-level data set includes cyclist counts at intersections 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 of each minute in the city network at both 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. 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 that shows the cyclist counts and waiting time at each intersection during the time period of the delivery.

3. OD data: Origin/Destination file that 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 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 that were recorded during the time period of the delivery.

3.4. Other Supporting Data

Other supporting data collected for the following bicycle volume estimation and prediction, cycling activity modeling, and injury risk analysis include bicycle counts from manual count stations, road characteristics (e.g., route class, length of segment, number of through lanes, and road direction, speed limit), demographic characteristics (e.g., total population, median age in census blocks, household income, total families, and poverty rate), slope, bicycle facilities (e.g., off-street paths, bike lanes, signed bike lanes, suggested bike routes, suggested bike routes with low comfort, and greenway), zoning data, bus stops, sidewalk, AADT, and bicyclist involved crash data. The following figures show the bicycle facilities, total population, slope, and bicycle-vehicle crashes distribution in the City of Charlotte.

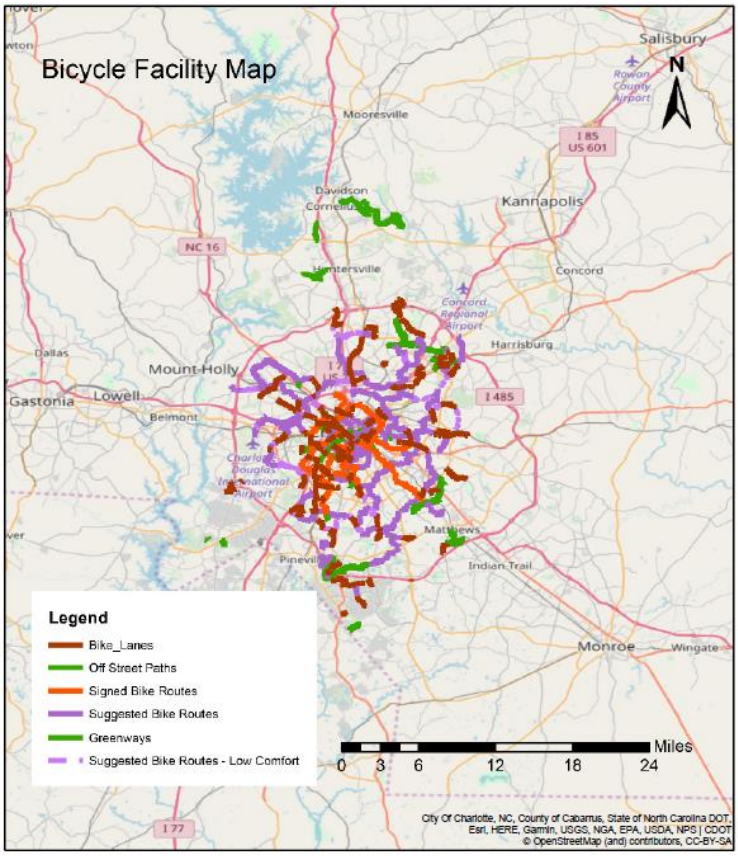


Figure 3.2: Bike Facilities in the City of Charlotte

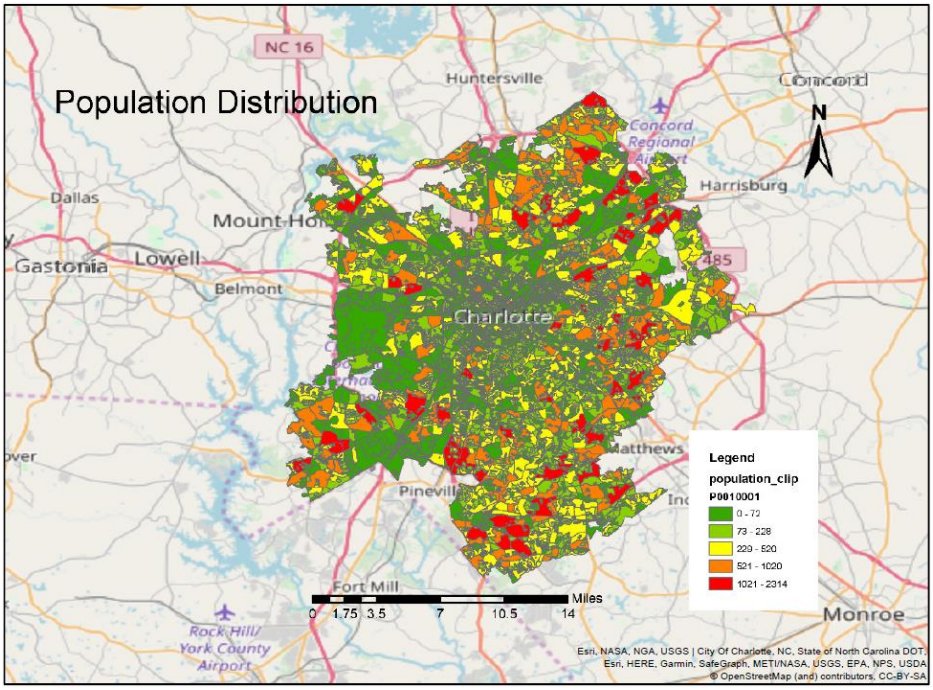


Figure 3.3: Total Population in the City of Charlotte

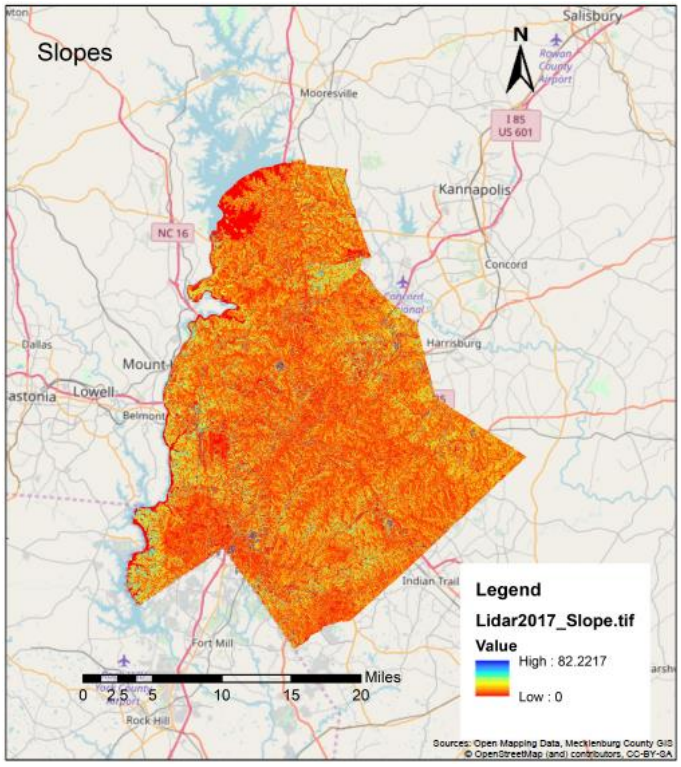


Figure 3.4: Slope in the City of Charlotte

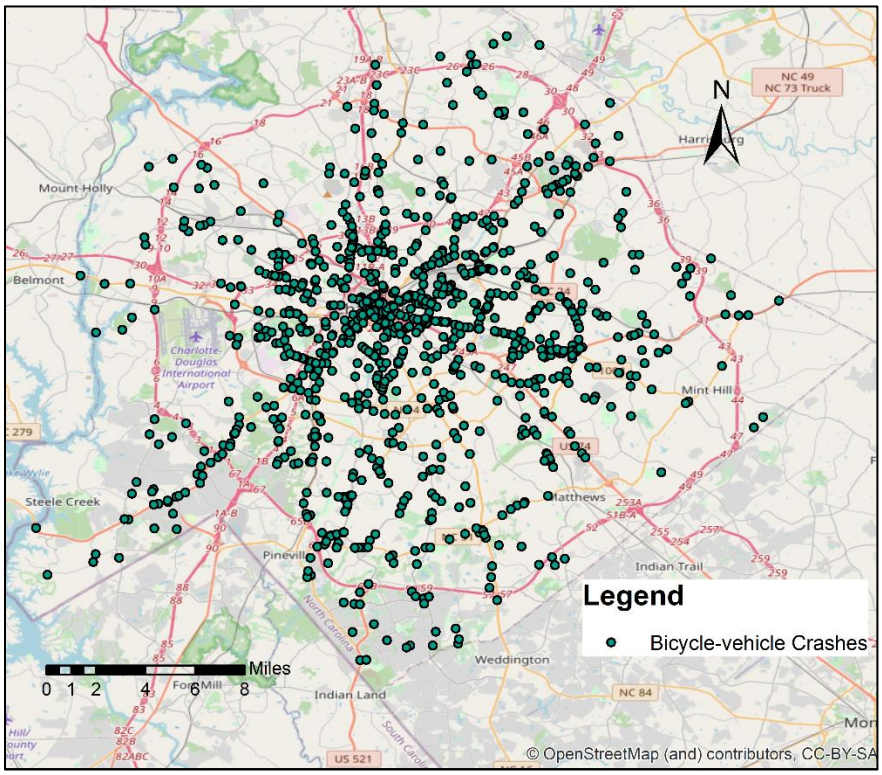


Figure 3.5: Bicycle-vehicle Crashes Occurred in the City of Charlotte

The crash data utilized in this research are the bicyclist-involved crash data collected in the City of Charlotte from 2007 to 2017. The data are obtained from North Carolina Department of Transportation. There are 1183 observations contained in the dataset with most of the bicycle-vehicle crashes (1149) occurred in the urban areas. To have a clear view of the crash number within each census block, Figure 3.6 is generated as follows.

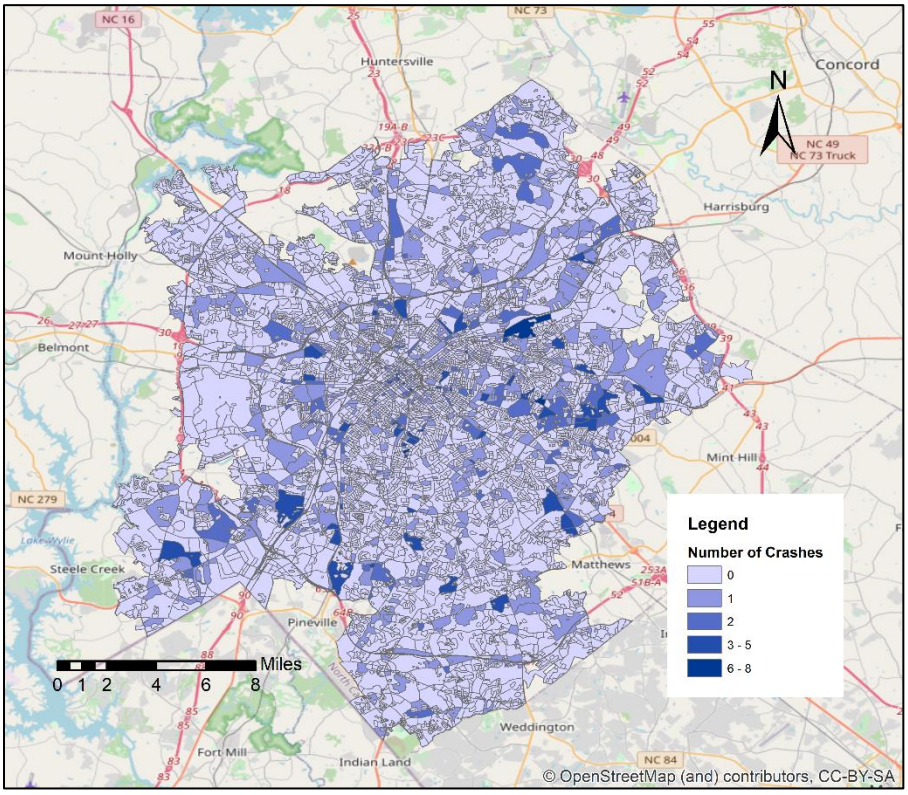


Figure 3.6: Number of Bicycle-vehicle Crashes within Census Blocks

3.5. Strava Data Analysis

3.5.1. Demographics

The total cyclists using Strava application are 8,857 with a majority of 7,129 male cyclists. Their total cycling trips from December 2016 to November 2017 were 140,428 miles. The proportion of Strava users' gender is presented in Figure 3.7.

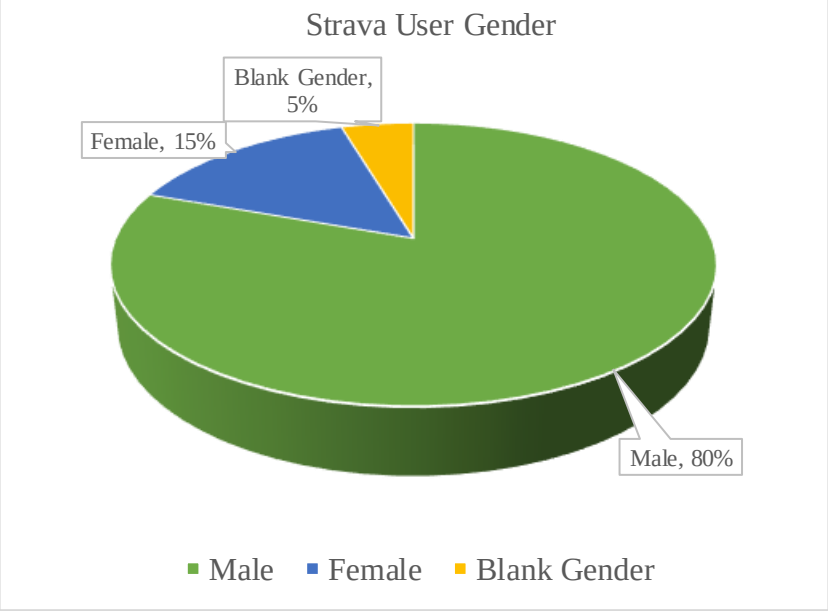


Figure 3.7: Strava User Gender

From the age data, one can see that cyclists of all ages were using Strava application to record their trips. This data indicate that a large number of cyclists, both young and old, are familiar with Strava application based on the fact that age groups of the Strava users range from under 25 to over 95 as presented in the figure below. Cyclists from different age groups for both male and female cyclists is presented in Figure 3.8. From the figure, one can see that most of the cyclists are between 25 and 54.

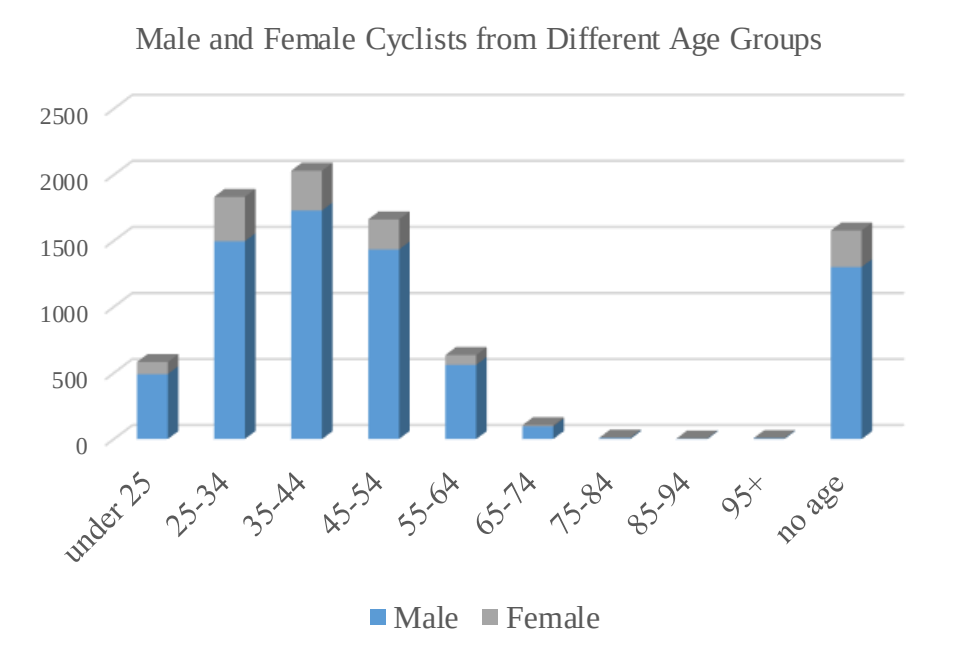


Figure 3.8: Male and Female Cyclists from Different Age Groups

3.5.2. Trip purpose

According to the data, among a large number of cyclists recording their cycling trips with Strava application, most of the trips are recreational trips. The proportion of commute trips and non-commute trips is shown in the following pie chart where commute trips account for only 18.33% of the total cycling trips and non-commute trips account for 81.67% of the total cycling trips.

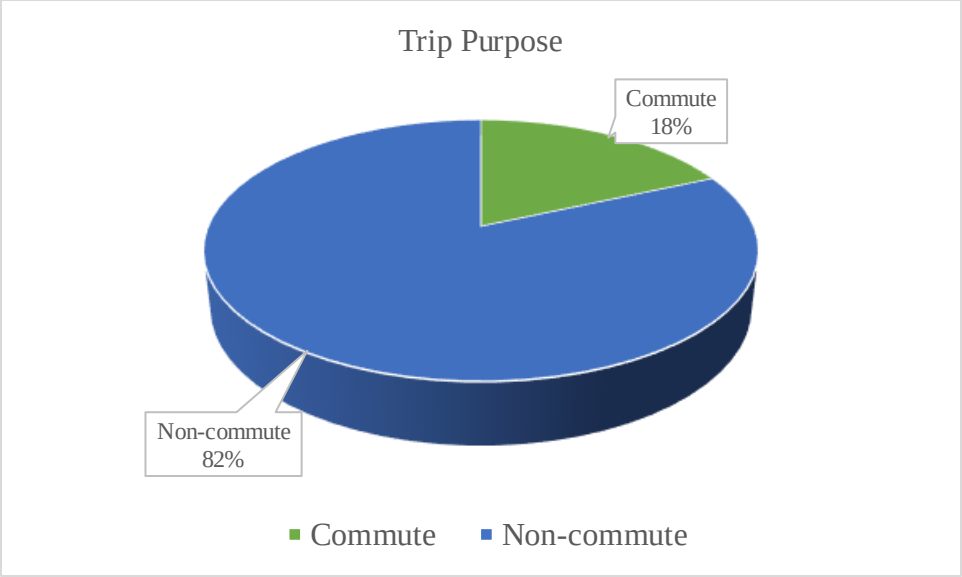


Figure 3.9: Cyclist Counts for Different Trip Purposes

3.5.3. Strava Count

The Strava bicycle counts vary from month to month, from day to day, and from hour to hour. Therefore, comparisons are conducted to identify the difference between each different aspect. Before comparison, a map that illustrates the total cyclists on each road segment for the whole year is presented, as shown in Figure 3.10.

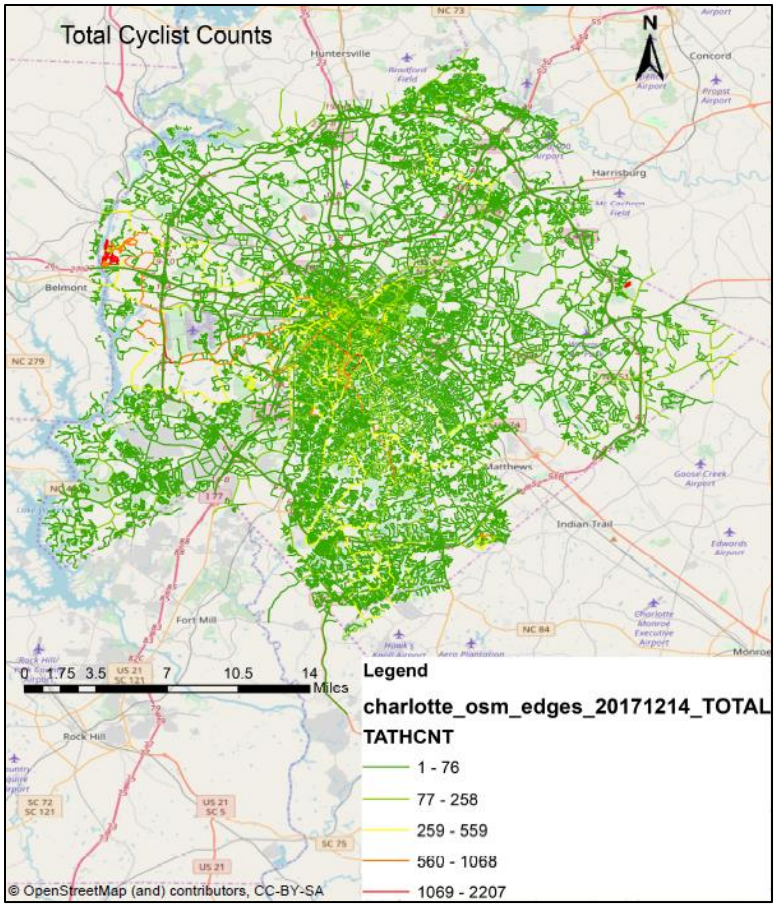
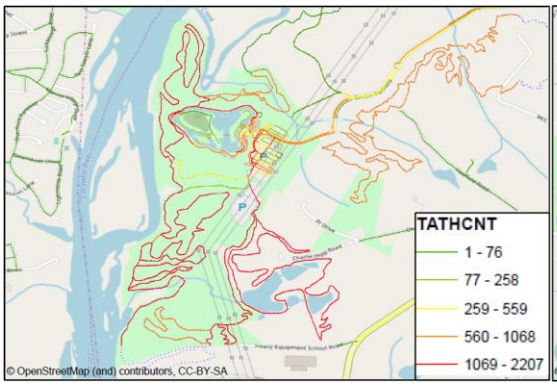
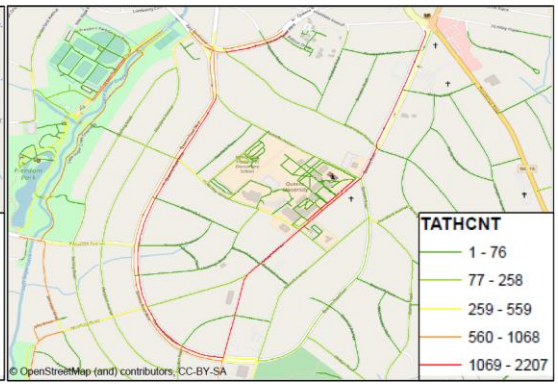


Figure 3.10: Total Cyclists Roll-ups

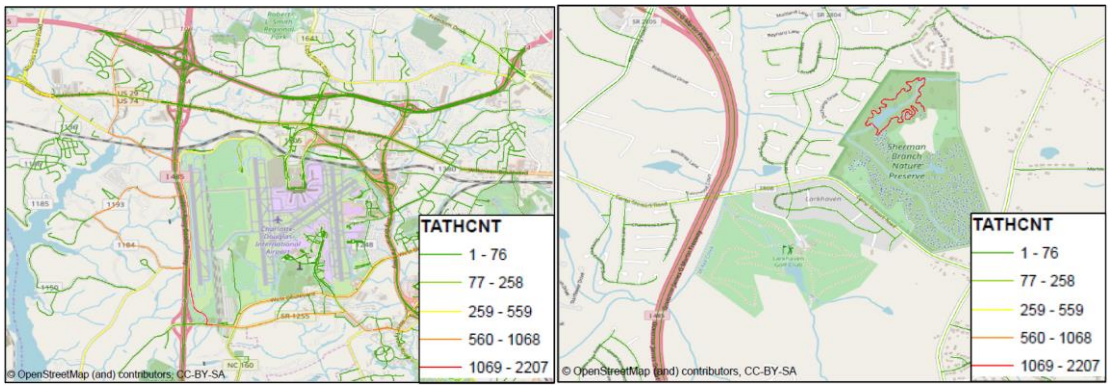
Based on the bicycle volume in Figure 3.10, four locations where the volume of Strava users are high are identified which involve greenway, school, airport, and park. These are the popular cycling locations among Strava users.



3.11.a Greenway



3.11.b School



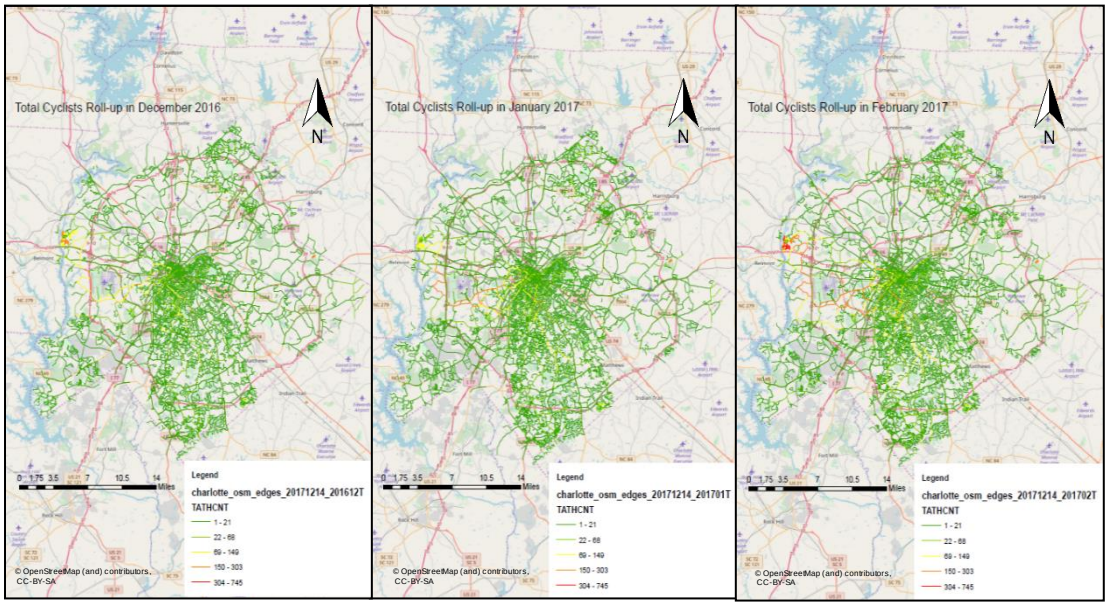
3.11.c Airport 3.11.d Park

Figure 3.11: Four Popular Cycling Locations

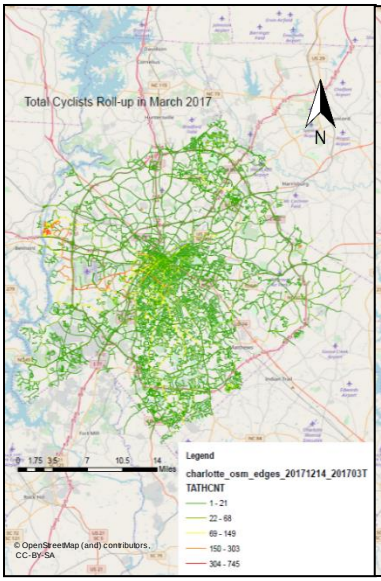
The Strava bicycle counts under different situations are presented in detail as follows.

3.5.3.1 Month of Year

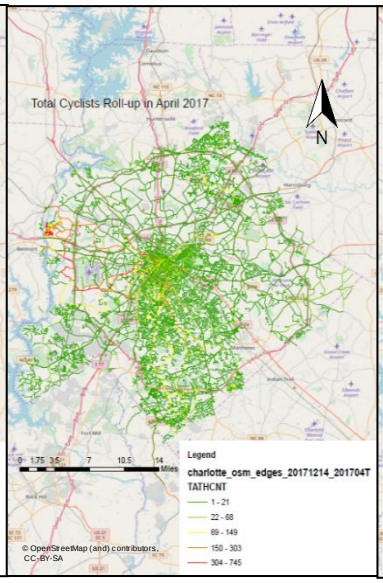
Cycling is a kind of activity which is highly related to the weather condition. Therefore, bicycle counts in different months of year vary with the temperature. Total bicycle volume on each road segment in twelve months of a year is presented in the following figures.



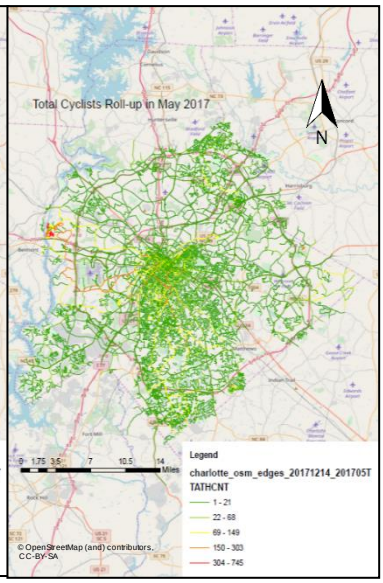
3.12.a December 2016 3.12.b January 2017 3.12.c February 2017



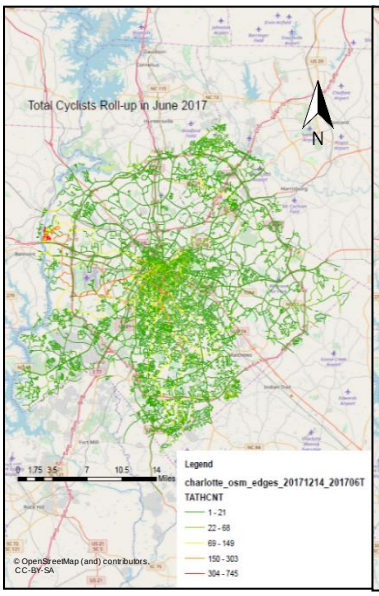
3.12.e March 2017



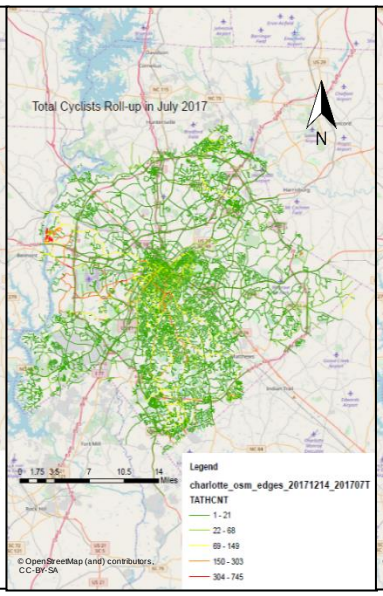
3.12.f April 2017



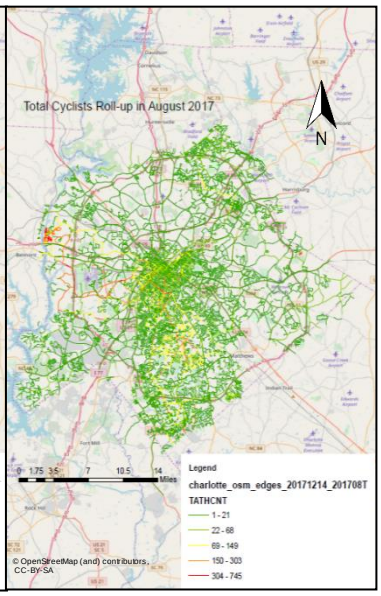
3.12.g May 2017



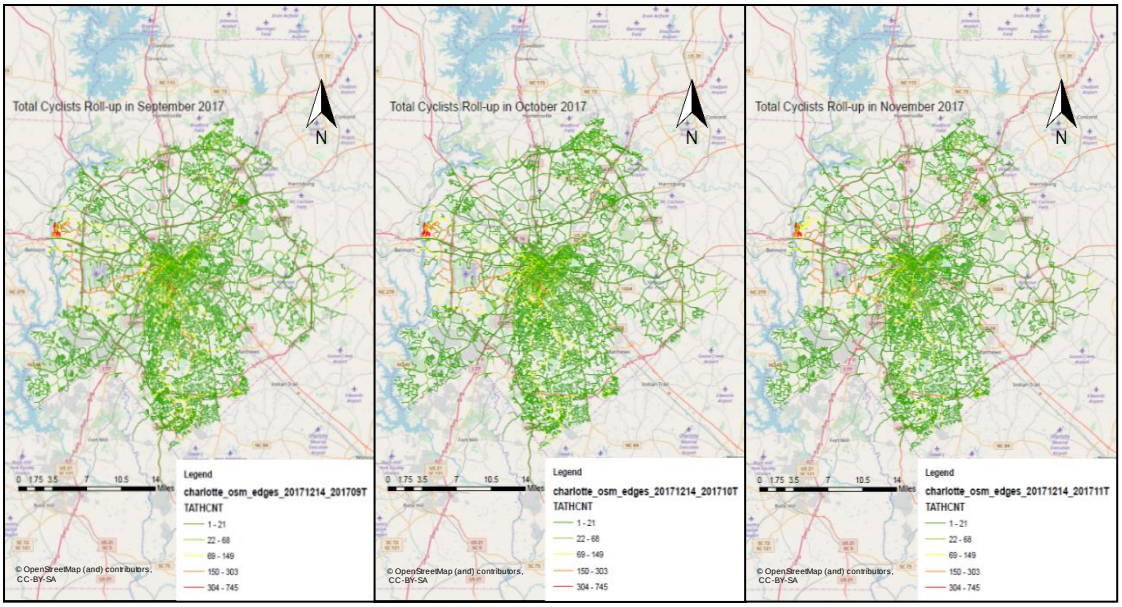
3.12.h June 2017



3.12.i July 2017



3.12.j August 2017



3.12.k September 2017 3.12.l October 2017 3.12.m November 2017

Figure 3.12: Total Bicycle Volume in Each Month

Based on the twelve maps generated to show the total bicycle volume on each road segment, several results can be presented as follows:

1. The four popular cycling locations remain the same over the twelve months.
2. The total bicycle volume on each road segment begins to increase in February and decreases in December.
3. Different locations have different variance in total bicycle volume.
4. The total bicycle volume on greenways begins to increase in February and decrease in December. However, the total bicycle volume in the uptown area and around airport begins to increase in April and decrease in October. And the park area has high bicycle volume from August to November.

A total bicycle volume change for the whole year can be seen in the following figure.

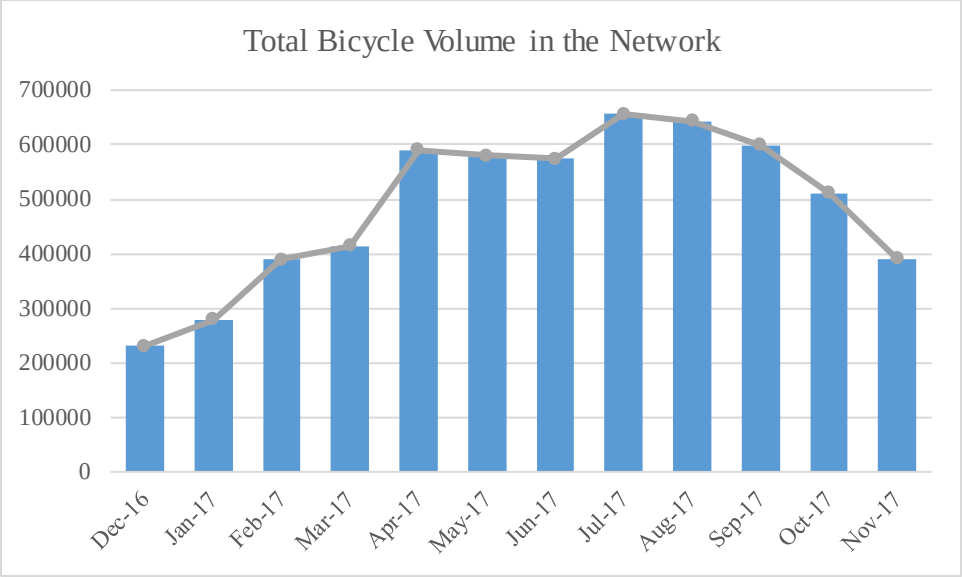


Figure 3.13: Total Bicycle Volume in the Network

3.5.3.2 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 3.14 where red lines represent the higher bicycle volume on weekends and green lines depict the higher volume on weekdays. According to Figure 3.14, the uptown area in the City of Charlotte appears to have more green lines which indicates more weekday cycling trips in this location.

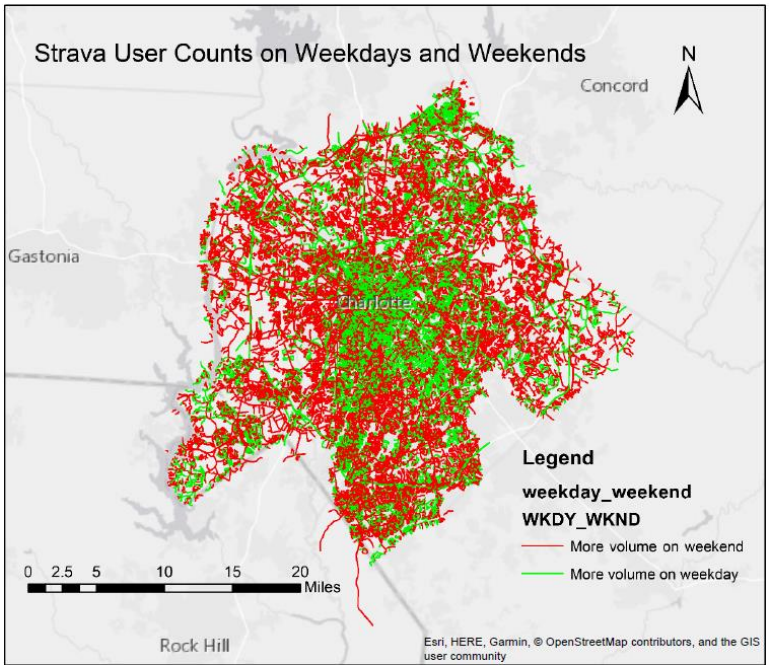


Figure 3.14: Total Bicycle Volume on Weekdays and Weekends

3.5.3.3 Time of Day

The bicycle volume for each road segment varies with different times of day. The variation of bicycle volume is presented in Figure 3.15. 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 possible reasons can be identified. First, the temperature around noon is high. Second, workers are busy during the day.

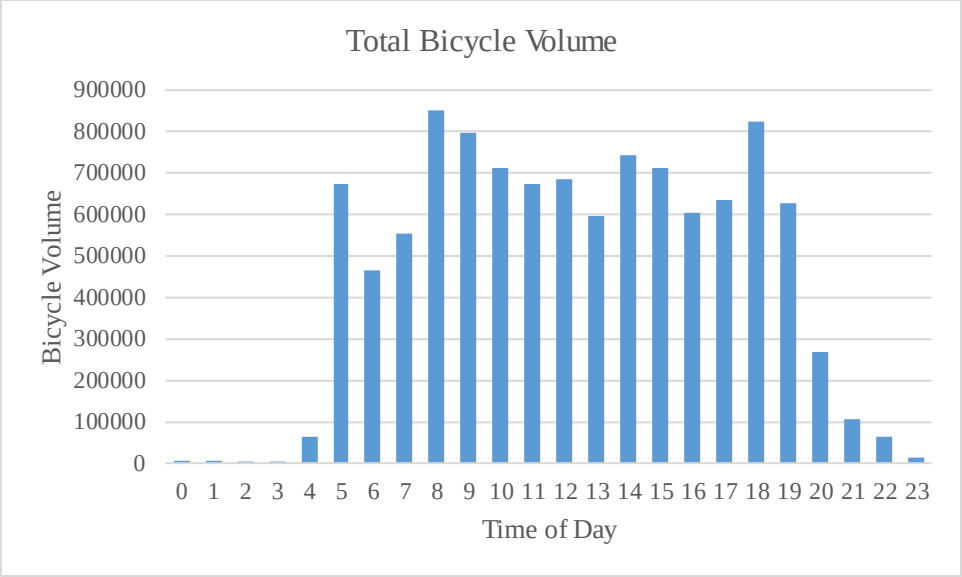


Figure 3.15: Total Bicycle Volume for Different Times of Day

3.5.3.4 Trip Purpose

The trip purpose has an impact on the total bicycle volume on each road segment. The commute trips are presented in the following figure.

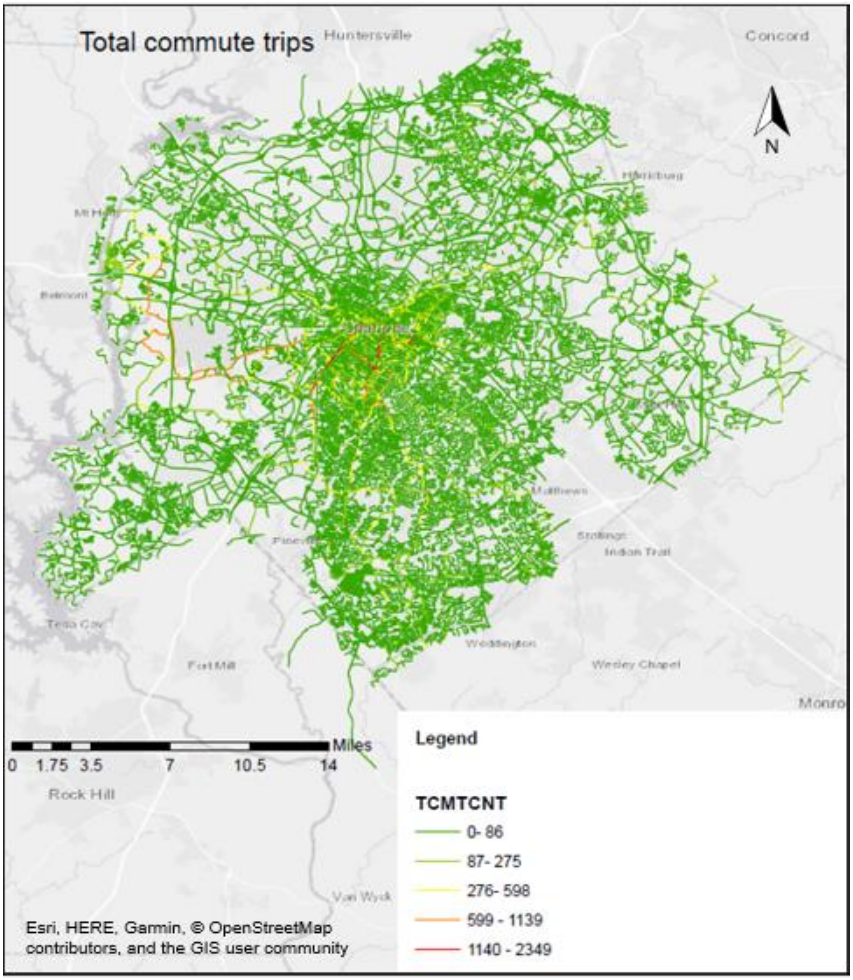


Figure 3.16: Total Commute Trips

3.6. Data Comparison

Difference remains between manual count data and Strava data. Since crowdsourced data usually involves a large number of people, the coverage of the road segment that is being used can be broad. On the contrary, installing manual count stations are costly and the coverage has to be limited. In other words, only the bicycle count at some locations can be collected. In addition, Strava data contain the bicycle trip time and the trip purpose (commuting or recreation), while manual count data cannot collect such information. In this research, the bicycle manual count from different count stations and

Strava user count at the same locations are compared in the following figure. In the figure, one can see that the manual count is greater than the Strava count.

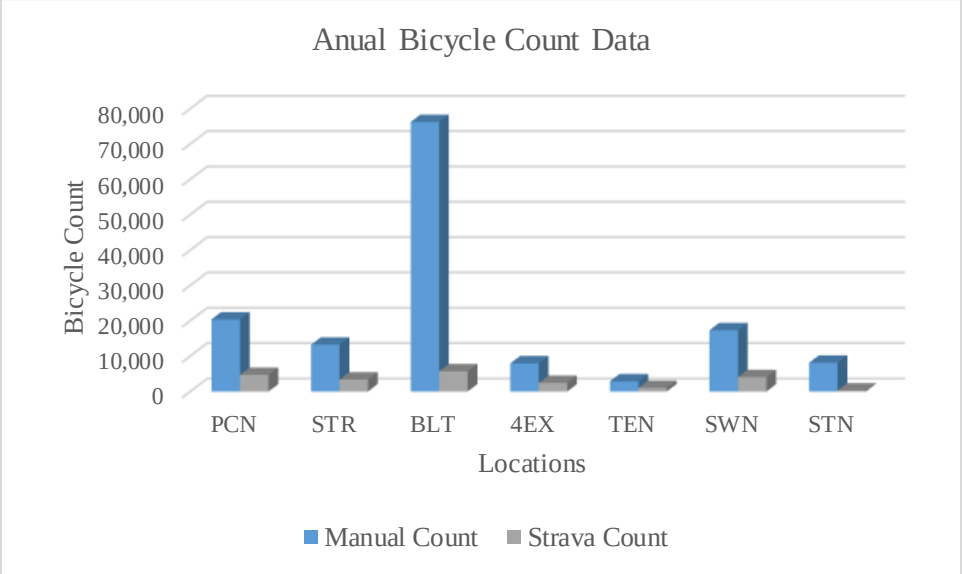


Figure 3.17: Comparison of Manual and Strava Counts

3.7. Summary

This chapter provides an overview of the data collected for this research. The descriptive analyses based on the data collected are conducted by creating several heatmaps for bicycle volume in different months of year, weekdays and weekends, and for different trip purposes. A data comparison between bicycle manual count data from the continuous count station in Charlotte and Strava bicycle data from the smartphone application is also provided.

CHAPTER 4: DEVELOPING BICYCLE VOLUME MODELS

4.1. Introduction

This chapter provides a method to combine all the collected data for the development of the bicycle volume models utilizing ArcGIS and SAS. After the data processing procedure, two bicycle volume models are developed to quantify the relationship between bicycle manual count data and Strava bicycle data as well as other relevant variables. Model results are analyzed and bicycle volume on most of the road segments in the City of Charlotte is calculated based on the model estimation results. In addition, a map illustrating the bicycle ridership in the City of Charlotte is created.

The following sections are organized as follows. Section 4.2 introduces the methods of data processing with ArcGIS and SAS. Section 4.3 presents the bicycle volume models and the model estimation results. Section 4.4 provides the bicycle volume prediction for most of the roadway segments in the City of Charlotte and creates a map to give an overall view of the bicycle ridership in the City of Charlotte. Finally, Section 4.5 concludes this chapter with a summary.

4.2. Data Processing

The data processing in this Chapter is conducted utilizing ArcGIS and SAS. Three steps are followed to obtain the final combined data which can be seen in detail as follows:

Step 1:

This step is done in SAS. First, the bicycle manual count data are collected from the count stations and the Strava bicycle volume data from the smartphone application. Both the data contain the bicycle counts on a specific roadway segment during different

times of day. To analyze the bicycle volume during different time periods, a time period variable is added to the data, where TP = 0 represents time from 00:00 to 05:59, TP = 1 represents time from 06:00 to 08:59, TP = 2 represents time from 09:00 to 14:59, TP = 3 represents time from 15:00 to 17:59, TP = 4 represents time from 18:00 to 19:59, and TP = 5 represents time from 20:00 to 23:59. Then, the total bicycle volume is summed up by each count station/road segment for the manual count data and Strava data separately to get Data 1 and Data 2. From this step, one can see that Data 1 and Data 2 have a temporal relationship in terms of date and time of day. The detailed data processing procedure for this step is presented in Figure 4.1.

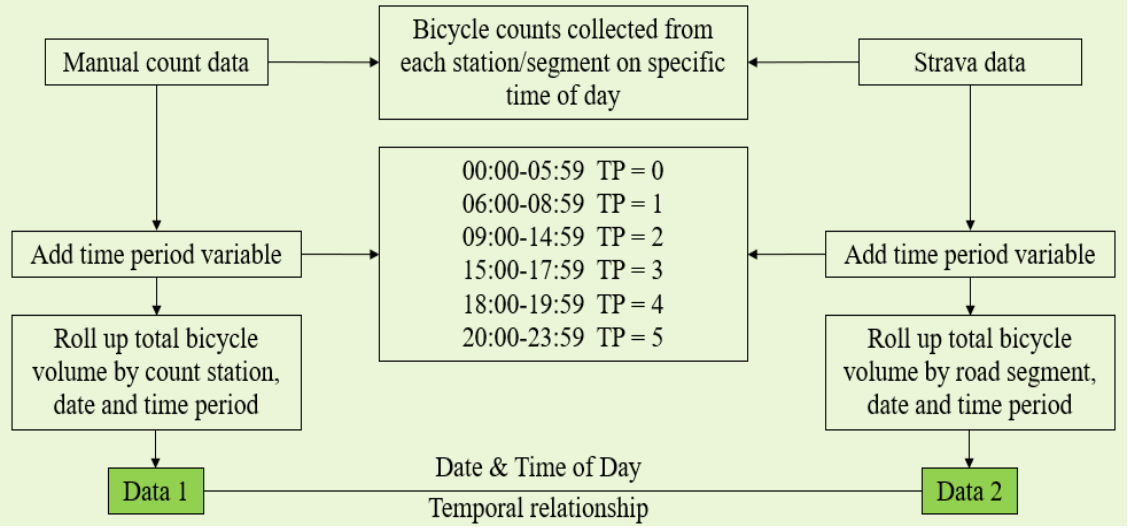


Figure 4.1: First Step of the Data Processing Procedure in SAS

Step 2:

This step is accomplished in ArcGIS. First, a point layer containing manual count station information that was compiled in step 1 is created (which is called Data 1) here. Second, other relevant supporting data including NC route characteristic data, Charlotte zoning data, slope cell data, sociodemographic data, bicycle facility data, and Strava road segment shapefile that shares the same road segment ID with Data 2 in step 1 are added

to ArcGIS. Before combining all the supporting data with the manual count station point layer, a data preprocessing is conducted. The NC route characteristic data are filtered by Charlotte boundary and the slope information is extracted from the cell data. Third, all the processed supporting data are combined together with Data 3 by spatial join in ArcGIS. Finally, Data1 and Data 3 are combined and the segments that have both manual count and Strava data are kept to create Data 4 that show the spatial relationship between Data 1 and Data 2. The detailed data processing procedure for this step is shown in Figure 4.2.

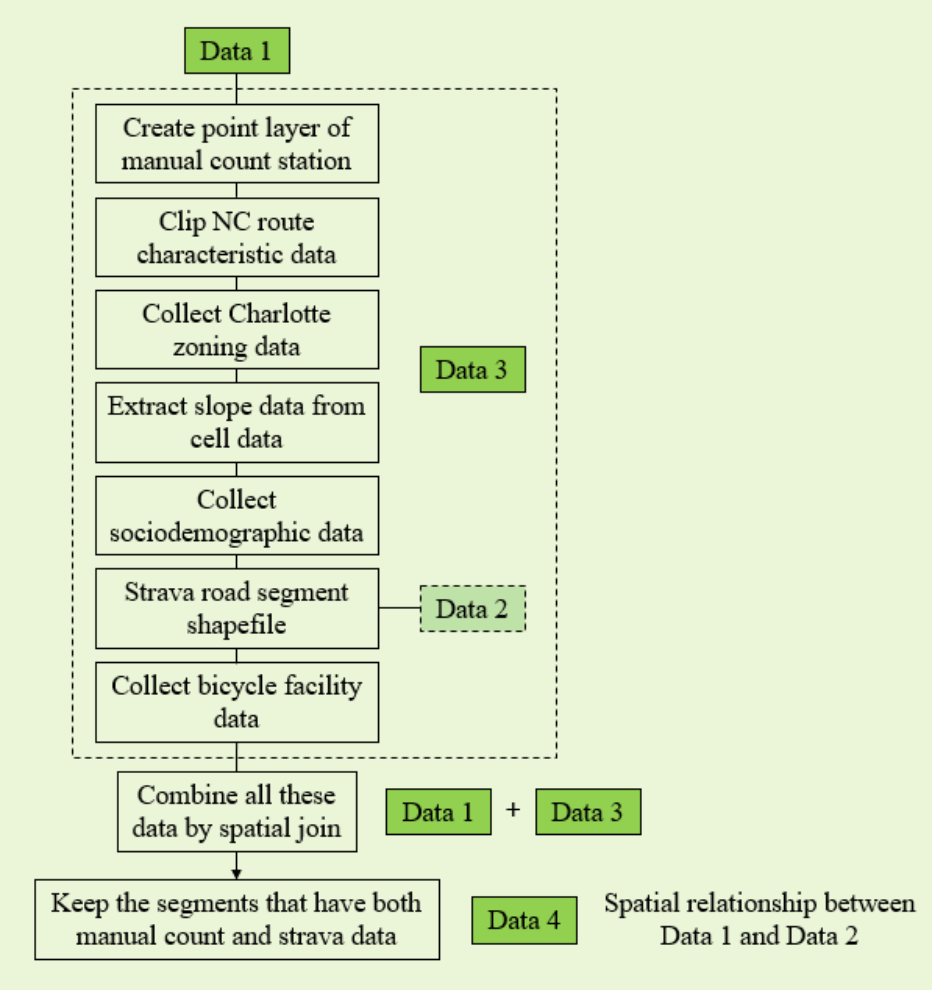


Figure 4.2: Second Step of the Data Processing Procedure in ArcGIS

Step 3:

Now that Data 4 contain the spatial relationship between Data 1 and Data 2 and information on Data 1, and one will still need to add the temporal relationship to it to obtain the final dataset. Thus, Data 4 and Data 2 are imported in SAS to create Data 5 by joining them with the same road segment ID, date, and time of day. Finally, dummy variables including weekdays and six time periods are added to Data 5. The detailed data processing procedure is shown in Figure 4.3.

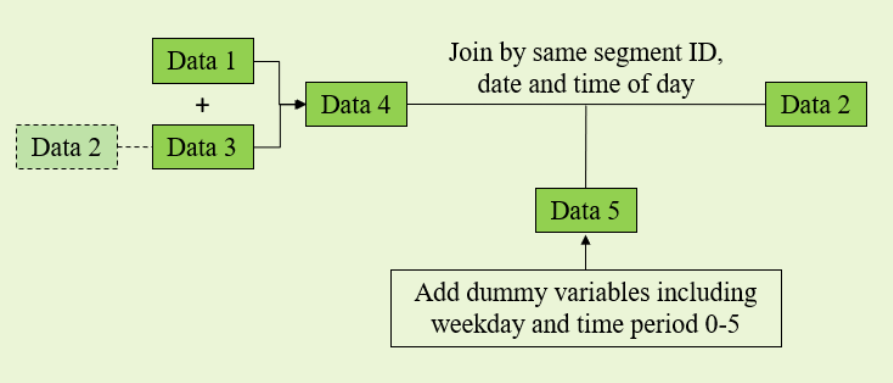


Figure 4.3: Third Step of the Data Processing Procedure in SAS

4.3. Bicycle Volume Regression Models

4.3.1. Simple Linear Regression Model

To assess the relationship between Strava data and bicycle manual count data, a simple linear regression model is developed, with manual count data being the dependent variable and Strava count data as the independent variable. The model estimation is conducted by using SAS, and the results are presented in the following table.

Table 4.1: Simple Linear Regression Model Estimation Results

Variable	Label	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	8.78724	0.82601	10.64	<.0001
BikeCount	Strava	7.62895	0.17815	42.82	<.0001
R-Square		0.3562	Adj R-Square	0.3560	

Results reveal that total bicyclist counts on the specific road segment are about 7.63 times as high as the number of Strava users on the same road segment. However, according to the values of R square (0.3562) and adjusted R square (0.3560), the predictive accuracy of this model is low. That is probably because the manual count data could be determined by many other factors that are not accounted for in this model. Therefore, to estimate the impacts of other variables on bicycle manual count data on each road segment, a multiple linear regression model is conducted below.

4.3.2. Multiple Linear Regression Model

To investigate the influence of contributing factors on manual count including Strava user count, a multiple linear regression model is formulated as shown below, and the variables considered in this model are presented in Table 4.2.

Manual Count = f (N, G, S, Z, T, B, C)

where:

N = Network characteristics data which include speed limit, segment length and through lane.

G = Slope.

S = Sociodemographic data which include total population, median household income and median age.

Z = Zoning data including residential, business and mixed use.

T = Temporal data including different time periods and weekday.

B = Bicycle facility data including off-street paths, bike lanes, signed bike lanes, suggested bike routes, suggested bike routes with low comfort, and greenway.

C = Strava bicycle count.

Table 4.2: Variable Description

Variable Type	Variable Label	Description
Network Characteristics	Speed Limit	The posted speed limit on a roadway segment.
	Segment length	The length of the segment in miles.
	Through lane	The number of through lanes.
Geometry	Slope	The slope of a road segment at intersection.
Sociodemographic characteristics	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.
Zoning	Residential	Charlotte zoning with residential land use.
	Business	Charlotte zoning with business land use.
	Mixed use	Charlotte zoning with mixed use land use.
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

Variable Type	Variable Label	Description
		= 1.
	Weekday	If bike on a weekday, then weekday = 1.
	Off_Street_Paths	Off street paths
	Bike_Lanes	Bike lanes
	Signed_Bike_Lanes	Signed bike lanes
Bicycle facilities	Suggested_Bike_Routes	Suggested bike routes
	Suggested_Bike_Routes_Lowcomfort	Suggested bike routes with low comfort
	Greenway	Greenway
Strava data	BikeCount	Strava user count on a road segment.

The parameter estimation of this multiple linear regression model is conducted in SAS, and the model estimation results are present in Table 4.3.

Table 4.3: Multiple Linear Regression Model Estimation Results

Variable	Parameter Estimate	Standard Error	t Value
Intercept	4.53707	2.15121	2.11
Hour_1	5.05005	1.93230	2.61
Hour_2	25.29360	1.92312	13.15
Hour_3	24.72827	1.89316	13.06
Hour_4	16.67931	1.97334	8.45
Hour_5	10.91207	2.17965	5.01
weekday	-9.16515	1.11290	-8.24
BikeCount	6.32098	0.15380	41.10
Bike_Lanes	-22.10636	1.20746	-18.31
Off_Street_Paths	22.60260	1.23021	18.37
Suggested_Bike_Routes	-13.94757	2.41011	-5.79
R-Square	0.6084	Adj R-Square	0.6073

Based on the model estimation results in Table 4.3, variables including weekday, time period except 00:00-06:00 am, Strava user count, off-street paths, bike lanes, and suggested bike routes have a significant impact on the manual count. Specific analysis is conducted in detail as follows:

Time period except 00:00-06:00 am has a positive impact on the total bicycle volume on a road segment, which means cycling activity starts early in the morning and ends late at night. Cyclists prefer to bike on weekends compared to weekdays. This is probably because cyclists may need to work on weekdays which gives them less time for cycling. Another possible reason is that most of the cycling trips may be recreational trips. Therefore, weekday has a negative impact on the manual bicycle count. According to the results, different bicycle facilities have different impacts on the total bicycle volume on road segments. Interestingly, bike lanes and suggested bike routes have negative impacts on the manual count, while off-street path has a positive impact on it. It can be interpreted that compared with other bicycle facilities, off-street paths are the most popular ones among cyclists in the City of Charlotte. The values of R square (0.6084) and adjusted R square (0.6073) of this multiple linear regression model are higher than the simple linear regression model, which indicates that this model has a higher prediction accuracy than the previous one.

4.4. Bicycle Volume Prediction

Based on the model estimation results from the multiple linear regression model, a bicycle volume prediction on most of the road segments in the city of Charlotte with availability of Strava data and bike facility data is computed using the following equation:

Bicycle volume = 4.53707 + 5.05005 * [Hour_1] + 25.29360 * [Hour_2] + 24.72827 * [Hour_3] + 16.67931 * [Hour_4] + 10.91207 * [Hour_5] - 9.16515 * [Weekday] + 6.32098 * [BikeCount] - 22.10636 * [Bike_Lanes] + 22.60260 * [Off_Street_Paths] - 13.94757 * [Suggested_Bike_Routes]

To obtain the annual average daily bicycle (AADB) prediction, the predicted bicycle volume on each road segment calculated using the equation above are rolled up for the whole year, which provides the aggregated whole year bicycle volume (V_T) on each road segment in the city of Charlotte. Therefore, the AADB prediction can be calculated using the following equation:

$$AADB = V_T / 365$$

An AADB prediction of most of the road segments in the city of Charlotte is presented in Figure 4.4.

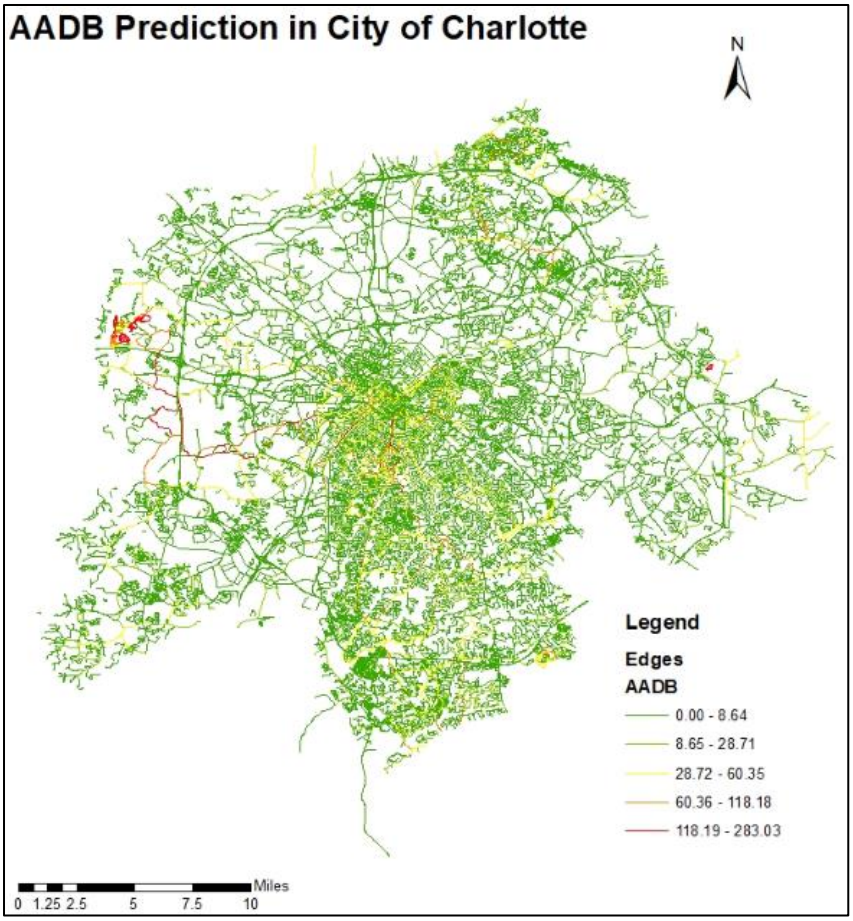


Figure 4.4: AADB Prediction in the City of Charlotte

4.5. Summary

This chapter provides a method to combine all the collected data for the development of the bicycle volume models utilizing the ArcGIS and SAS. After the data processing, two bicycle volume models are developed to quantify the relationship between bicycle manual count data and Strava bicycle data as well as other relevant variables. Model results are analyzed and predicted bicycle volume on most of the road segments in the city of Charlotte is calculated using the developed estimation model. In addition, a map illustrating the bicycle ridership in the city of Charlotte is also created.

CHAPTER 5: MODELING CYCLING ACTIVITIES

5.1. Introduction

In this chapter, discrete choice models are developed to model cycling activities and model comparison is conducted to identify the best-fit model for this cycling activity analysis. To examine the different impacts of explanatory variables on cycling activities during selected time periods, discrete choice models are developed separately.

The following sections are organized as follows: Section 5.2 provides a data processing method to combine all the needed data for the later development of discrete choice models. Section 5.3 through Section 5.6 provide the models developed for cycling activities including ordered logit (ORL) model, partial proportional odds (PPO) model, multinomial logit (MNL) model, and mixed logit (MXL) model respectively. Section 5.7 compares the models developed in the previous sections and identifies the best model structure for this research study. Section 5.8 develops two models for different selected time periods and a model comparison is provided in this section. Finally, Section 5.9 concludes this chapter with a summary.

5.2. Data Processing

The data processing procedure is conducted utilizing ArcGIS and SAS. Two steps are needed to obtain the final combined data which can be seen in detail as follows:

Step 1:

This step is done in ArcGIS. First Strava road segment shapefile is added in ArcGIS (named Data 1 later). This data contain basic information on the Strava road based on the Open Street Map with a column that records the road segment ID (i.e., Edge

ID). This ID is used to relate it to the Strava user count data (Data 4) to match the bicycle volume data to the Strava roadway network.

In addition, other relevant supporting data including NC route characteristics data, slope cell data, sociodemographic data and bicycle facility data are also added to ArcGIS. Before combining all the data together, data preprocessing is conducted. For the NC route characteristics data, only the data in the City of Charlotte are selected to accelerate the data processing speed later. Therefore, Charlotte boundary data are added to clip the NC route data as shown in Figure 5.1.

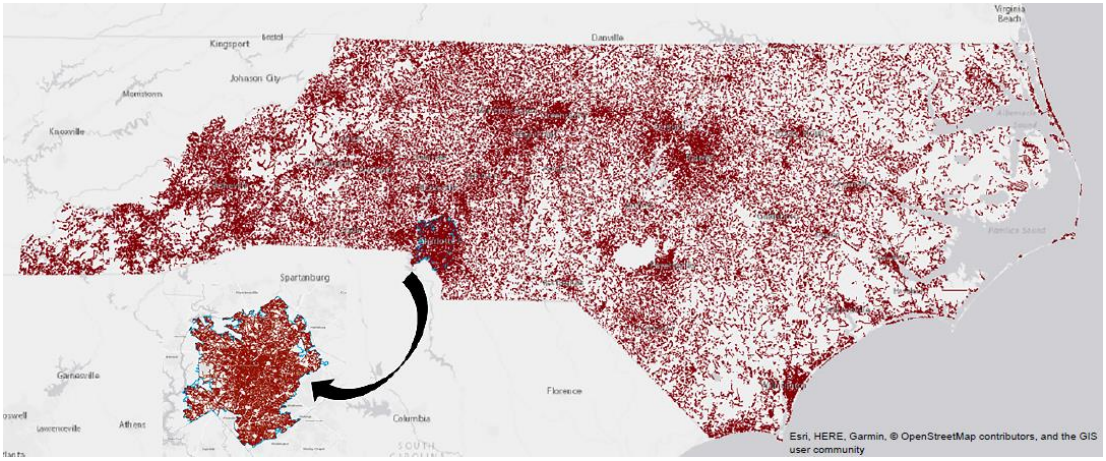


Figure 5.1: Clip in ArcGIS

To obtain the data from the slope cell data, the “Extract” tool in ArcGIS is utilized to export the slope data. After all the data preprocessing, the Data 2 are acquired as a combination of four supporting data. Then, Data 2 are combined with Data 1 in order to join to Strava road segment shapefile. Finally, Data 3 are obtained as a combination of Data 1 and Data 2. The detailed data processing procedure for this step is shown in Figure 5.2.

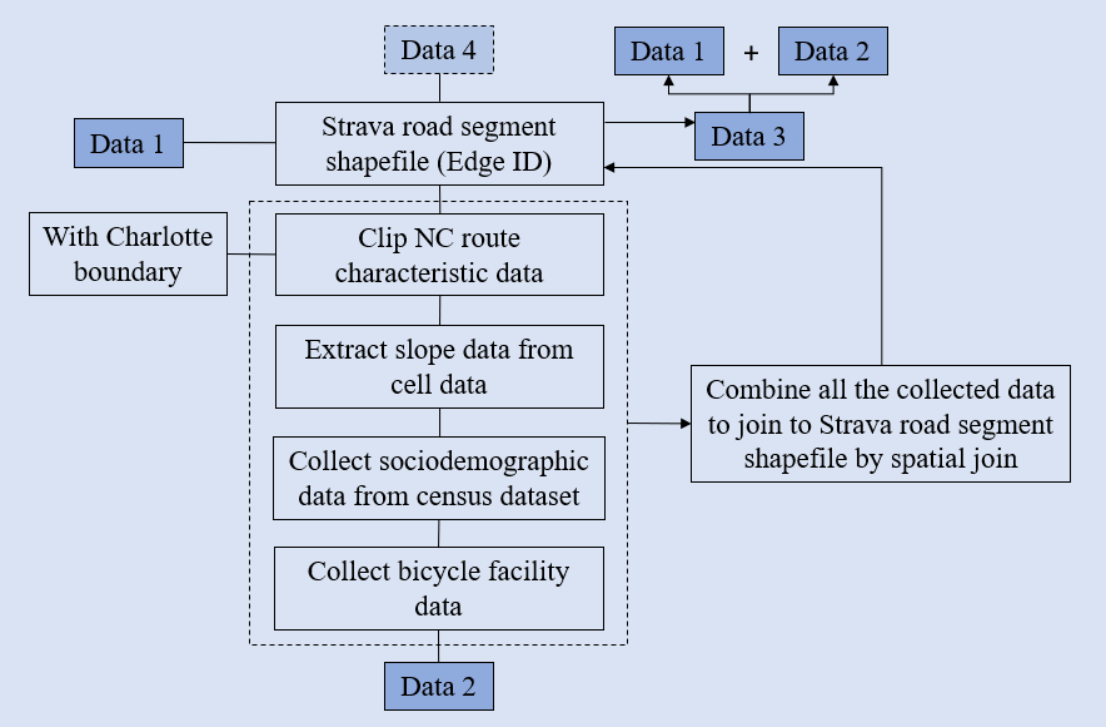


Figure 5.2: Data Processing in ArcGIS

Step 2:

This step is accomplished in SAS. First, the Strava bicycle count data (Data 4) collected from each road segment during a specific time of day in the City of Charlotte are imported in SAS. A column with six time periods from 0 to 5 is created where TP = 0 represents time from 00:00 to 05:59, TP = 1 represents time from 06:00 to 08:59, TP = 2 represents time from 09:00 to 14:59, TP = 3 represents time from 15:00 to 17:59, TP = 4 represents time from 18:00 to 19:59, and TP = 5 represents time from 20:00 to 23:59. In order to add the day of week variable, one needs to first convert the day of year variable to date using DATEJUL in SAS, and leading zeros are added to make sure that the day is consistent with 3 digits. Based on the date, the WEEKDAY function is used to obtain the day of week from the SAS data value. And then, a roll-up bicycle volume table is created by road segment, date, and time period.

After preprocessing the Strava count data, Data 3 from the previous step are joined to Data 4 by the same segment ID. Dummy variables including weekday and time period 0 – 5 are added to the data. The variable that indicates the level of the bicycle volume on each road segment is created. Five categories are set up with bicycle counts from low (0-39), low-average (40-79), average (80-119), high-average (120-159), to high (160-200). And finally, Data 5 are obtained for the future model development. The detailed data processing procedure for this step is shown in Figure 5.3.

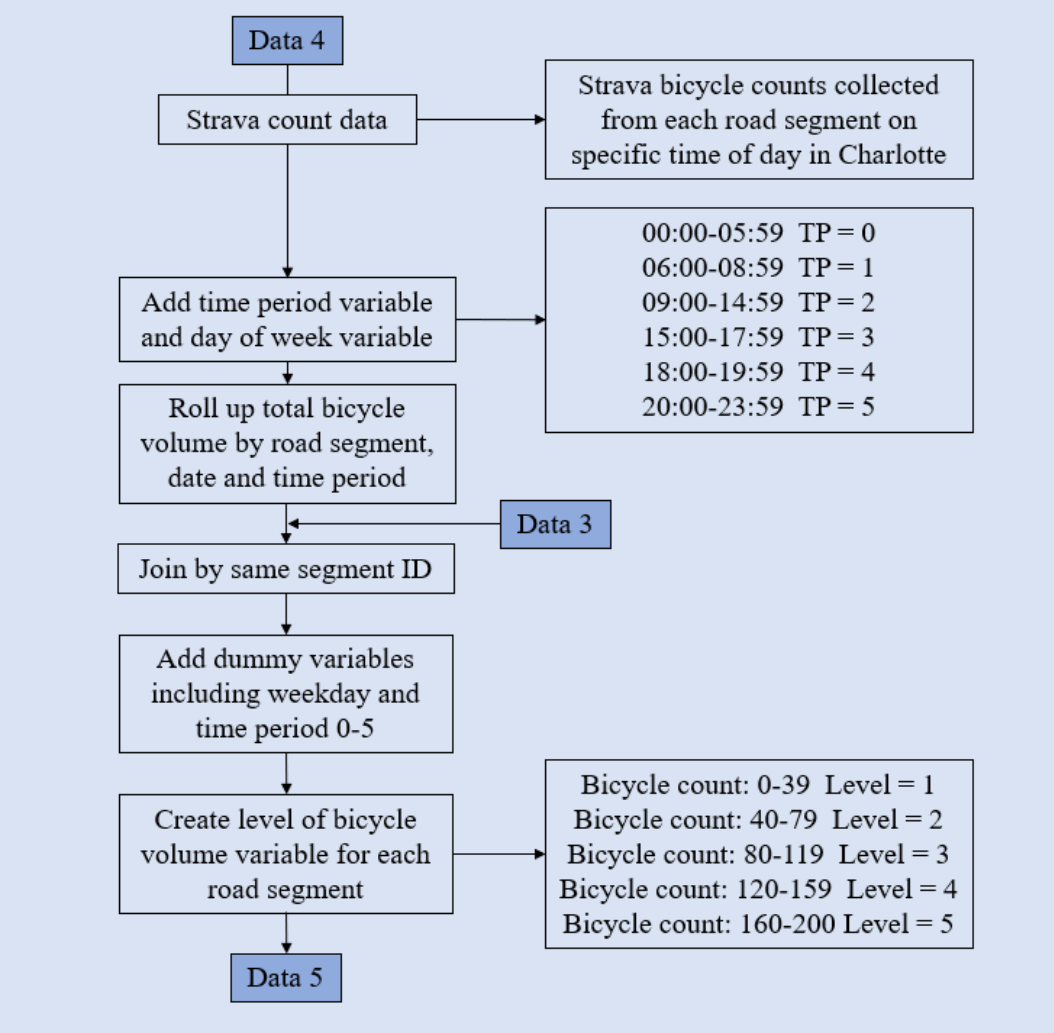


Figure 5.3: Data Processing in SAS

5.3. Ordered Logit Model

5.3.1. ORL Model Structure

The ordered logit model is one of the traditional discrete choice models using for ordinal dependent variable analysis. In this research study, the number of bicycle counts on 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 ordered logit model, the level of bicycle counts on a road segment is denoted as y_i, which is associated with the variable y_i^*. The model specification is presented as follows:

$$y_i^* = \beta X_i + \varepsilon_i$$

where y_i^* demonstrates the 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, the continuous 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:

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

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

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

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

5.3.2. ORL Model Results

To analyze the level of bicycle counts on each road segment and examine the impact factors on cycling activities 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.
ThruLane	The number of through lanes.
Oneway	If the road segment is one way, then oneway = 1
Sociodemographic Characteristics	
TOTPOP_CY	Total population in each census block.
MEDAGE_CY	The median age in each census block.

Variable	Description
MEDHINC_CY	Median household income in each census block.
Total_HH	Total households in each census block.
TotalFamily	Total families in each census block.
Poverty	Family poverty rate in each census block.
Geometry	
Slope	The slope of a road segment at intersection.
Bicycle Facilities	
B_offstreet	Off street paths
B_bikelane	Bike lanes
B_signed	Signed bike lanes
B_suggested	Suggested bike routes
B_suggest0	Suggested bike routes with low comfort
B_greenway	Greenway

All the factors presented in Table 5.1 are considered in the ordered logit model to determine the probability of segments being selected by the Strava users. The maximum likelihood method is utilized to conduct the model estimation and to determine the thresholds in the ordered logit model. This process is conducted in SAS 9.4. To keep the variables that affect the level of bicycle counts on each road segment significantly, the backward selection demand is used for the model estimation. The backward selection results are 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	Wald Chi-Square	Pr > ChiSq
1	Hour_0	0.0000	0.9993
2	B_offstreet	0.0000	0.9951

Summary of Backward Elimination			
Step	Effect Removed	Wald Chi-Square	Pr > ChiSq
3	Hour_4	0.0027	0.9586
4	SpeedLimit	0.1222	0.7266
5	Poverty	0.4548	0.5001
6	TOTPOP_CY	0.4030	0.5255
7	B_bikelane	0.6974	0.4037

Table 5.3: Ordered Logit Model Estimation Results

Analysis of Maximum Likelihood Estimates					
Parameter	Level	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	5	1.6165	1.0468	2.3847	0.1225
Intercept	4	3.6935	1.0534	12.2937	0.0005
Intercept	3	4.0366	1.0565	14.5970	0.0001
Intercept	2	5.4232	1.0882	24.8353	<.0001
Weekday		-4.2510	0.3204	176.0312	<.0001
Hour_1		1.0789	0.4326	6.2192	0.0126
Hour_2		1.1850	0.4193	7.9859	0.0047
Hour_3		2.9484	0.4137	50.7871	<.0001
MPLength		1.0827	0.4673	5.3673	0.0205
ThruLane		0.6786	0.0853	63.2215	<.0001
MEDAGE_CY		0.0244	0.0115	4.4958	0.0340
MEDHINC_CY		0.000032	2.773E-6	129.7401	<.0001
Total_HH		0.00119	0.000345	11.8828	0.0006
TotalFamily		-0.00133	0.000470	8.0179	0.0046
Slope		-0.0506	0.00959	27.8175	<.0001
B_signed		-1.1172	0.1814	37.9421	<.0001
B_suggested		0.7100	0.3414	4.3260	0.0375
B_suggest0		-1.8420	0.3542	27.0457	<.0001
B_greenway		2.6567	1.0285	6.6720	0.0098

Analysis of Maximum Likelihood Estimates					
Parameter	Level	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
RouteClass1		-0.6356	0.2719	5.4624	0.0194
RouteClass2		0.8828	0.2390	13.6409	0.0002
RouteClass3		-0.3567	0.1395	6.5407	0.0105
Oneway		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 significant impacts 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 negative impacts 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 positive impacts on the level of bicycle counts. The detailed interpretation of the influence 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.

5.4. Partial Proportional Odds Model

5.4.1. PPO Model Structure

The partial proportional odds (PPO) model is developed based on the ordered logit (ORL) model. In the ORL 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 mentioned above, 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 violating the assumption which is shown as vector Z_i. The variables violating 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:

$$P(Y_i \geq j) = \frac{\exp\left[\theta_j - \left(X_i'\beta_j + Z_i'\gamma_j\right)\right]}{1 + \exp\left[\theta_j - \left(X_i'\beta_j + Z_i'\gamma_j\right)\right]}$$

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 PO assumption is violated or not, the Wald Chi-square tests are utilized during the model development. This procedure helps divide the explanatory variables into two groups which are categorized in either vector X_i or vector Z_i.

5.4.2. PPO Model Results

This PPO model is built based on the ORL model developed in Section 5.2. A series of Wald Chi-square tests are conducted to determine if the explanatory variables 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
ThruLane_po	10.1651	0.0172
MEDHINC_CY_po	33.7202	<.0001
Total_HH_po	25.5679	<.0001
TotalFamily_po	37.5464	<.0001
B_suggested_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 affect different levels variously.

The PPO model estimation results and the fit statistics are shown 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

Analysis of Maximum Likelihood Estimates					
Parameter	Level	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
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
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
ThruLane	5	0.5160	0.0711	52.7303	<.0001
ThruLane	4	-0.2532	0.2544	0.9905	0.3196
ThruLane	3	-0.1234	0.4373	0.0796	0.7778
ThruLane	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_HH	5	0.00105	0.000334	9.8621	0.0017
Total_HH	4	0.00859	0.00280	9.4126	0.0022
Total_HH	3	0.0277	0.00534	26.9469	<.0001
Total_HH	2	0.0373	0.0206	3.2672	0.0707
TotalFamily	5	-0.00120	0.000458	6.8452	0.0089
TotalFamily	4	-0.0122	0.00377	10.4502	0.0012
TotalFamily	3	-0.0389	0.00617	39.7655	<.0001
TotalFamily	2	-0.0530	0.0253	4.3881	0.0362
Slope		-0.0575	0.00891	41.5729	<.0001
B_signed		-1.0671	0.1841	33.6052	<.0001
B_suggested	5	2.8458	0.9343	9.2777	0.0023
B_suggested	4	2.8330	1.1958	5.6128	0.0178
B_suggested	3	-3.4416	1.9230	3.2029	0.0735
B_suggested	2	-0.4743	2.3583	0.0404	0.8406

Analysis of Maximum Likelihood Estimates					
Parameter	Level	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
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
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 shown in Table 5.6, variables that satisfy the PO assumption include 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 are allowed 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:

$$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(c)}$$

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 model the cycling activities for the Strava users in the City of Charlotte.

5.5. Multinomial Logit Model

5.5.1. MNL Model Structure

The multinomial logit model developed in this section is used to analyze cycling activities. In this model, it assumes that the alternative which yields the maximum utility is always selected, which is called random utility theory (Train, 2009). The utility function comprises an observed utility and an unobserved error term, which are shown in the following equation:

$$U_{in} = V_{in} + \varepsilon_{in}$$

where U_{in} is the utility function of the level of bicycle counts i for the road segment n, V_{in} denotes the observed utility of level i for the segment n, ε_{in} represents the unobserved error term of level i for the segment n. V_{in} is usually taken as a linear utility function as shown in the following equation:

$$V_{in} = \beta_0 + \sum_{k=1}^{N} \beta_k X_{ink}$$

where X_{ink} represents the kth explanatory variable of level of bicycle counts i for road segment n, N denotes the number of the explanatory variables, β_0 indicates the constant term, and β_k expresses the estimated coefficient of the kth explanatory variables.

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:

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}}$$

5.5.2. MNL Model Results

The MNL model estimation result is shown 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
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

Parameter Estimates					
Parameter	Level	Estimate	Standard Error	t Value	Approx Pr > |t|
Hour_4	4	-3.8902	0.4753	-8.19	<.0001
MPLength	5	1.5708	0.4601	3.41	0.0006
ThruLane	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_HH	4	0.005706	0.001417	4.03	<.0001
Total_HH	5	0.006635	0.001381	4.81	<.0001
TotalFamily	4	-0.007300	0.001749	-4.17	<.0001
TotalFamily	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 being kept in the MNL model are similar to those in the ORL and PPO models but are 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, 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

5.6. Mixed Logit Model

The MXL model is different from the MNL model because it allows explanatory variables to influence the mean of the random parameter distribution (Bhat, 1998; Revelt and Train, 1998; McFadden and Train, 2000; Bhat, 2000; Hensher and Greene, 2003) and it can address the unobserved heterogeneity. Similar to MNL model, the linear utility function of the MXL model is shown as follows:

$$U_{in} = \beta_{in} X_{in} + \varepsilon_{in}$$

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 MXL model structure is shown in the following equation:

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

where $f(\beta|\varphi)$ represents the probability density function of β, φ denotes the parameter vector, which shows the mean and variance of the density function. The coefficent β can be flexible or fixed, and can be any (e.g., normal, uniform, lognormal 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 built 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 different from zero at 0.05 level of significance. 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 (237,673) is extremely large for the estimation of MXL model which is not time efficient. Therefore, the peak hour data are selected to analyze cycling activities and the MXL models will be developed in Section 5.8.

5.7. 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, likelihood ratio index ρ^2, the Akaike's information criterion (AIC), and the Bayesian information criterion (BIC).

5.7.1. Indicators for Model Comparison

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

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

$$AIC = 2p - 2LL$$

$$BIC = pln(Q) - 2LL$$

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

Therefore, the four indicators for each model developed in the previous sections are shown 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 the PPO model, the PPO has a smaller value of -2LogL than that of the ORL model, which indicates that the PPO model outperforms the ORL model for fitting the bicycle count data in Charlotte. To compare

the three models with different structures, AIC and BIC values are utilized. Based on the values of AIC, the partial proportional odds model shares the smallest, which reveals the best fitness of the PPO model. However, the BIC value of PPO is not the smallest. According to the BIC values, the MNL model performs better than the PPO model, and the PPO model is better than the 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 the PPO model is larger than the MNL's can be interpreted that the PPO model has more estimated parameters than the MNL model. The trade-off between better fitness of the model and the variable number should be carefully considered and examined. In this research study, with the consideration of the four indicators, conclusion can be provided that the PPO model fits best for this cycling activity analysis.

5.7.2. 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 cycling activities 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 from 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 might be related to the high proportion of the non-commute trips in the Strava dataset. Different times of day will have different impacts on the bicycle counts since cycling activities vary 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 affects the high level of bicycle counts positively. Time period from 15:00 to 17:59 affects the high level of bicycle counts positively. 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 therefore 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 characteristics factors significantly affect the cycling activities. The explanatory variables that influence the level of bicycle counts significantly 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 bicyclists are willing to bike on roadway segments 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 a 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) affects the average level of the bicycle counts positively, which indicates that high 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 affects the average and above average levels positively, 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 affects the level of bicycle counts on each road segment negatively. 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 all have significant influences on the bicycle counts. Signed bike lanes affect the level of bicycle counts negatively, while greenways increase the likelihood of higher level of bicycle counts. The suggested bike routes with low comfort 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.

5.8. Modeling Cycling Activities for Different Time Periods

Applying the methodology mentioned in Section 5.6, two MXL models are developed to analyze the cycling activities 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. The indicators for different time periods are shown 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 5.8.1 and Section 5.8.2, 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 cycling activities for both peak hours.

5.8.1. AM Peak Hours

To analyze the cycling activities 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 different from zero at the significance level of 0.05. In other words, the coefficients in this model are fixed. Therefore, this MXL model collapses into a MNL model, and the MNL model estimation results are presented 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
ThruLane	3	2.1545	0.8328	2.59	0.0097
ThruLane	4	2.3905	0.6926	3.45	0.0006
ThruLane	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_HH	2	-0.002737	0.001125	-2.43	0.0150
Total_HH	5	0.003217	0.001249	2.58	0.0100
TotalFamily	5	-0.005797	0.001417	-4.09	<.0001
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. The 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 affects the level of bicycle counts significantly are different from the variables in the MNL developed with the whole dataset. According to the model results presented in Table 5.12, the road characteristic variables that have significant impacts on the level of bicycle counts contain the length of road segment, number of through lanes, speed limit, and one-way road. The length of the road segment affects the low-average level of bicycle counts negatively, 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 affects the bicycle count level (high-average) negatively. It is not hard to imagine cyclists prefer to bike on roads with lower speed limits. A greater number of through lanes increases the likelihood of high-level 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 influence the high level of bicycle counts (average and above) positively, which demonstrates that cyclists prefer to bike on one-way roads.

3. Sociodemographic characteristics:

Changes are also found in the sociodemographic variables having 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, and total families all affect the bicycle counts significantly. The median household income affects the average and above average levels of bicycle counts positively, 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 influences low-average level of bicycle counts negatively, 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 decrease the probability of low-average level. The impact of the total families remains the same as the MNL model developed with the whole dataset.

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 affect the level of bicycle counts significantly. They both have a positive impact on the low-average or average level of bicycle counts. 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 lots of cycling trips that occurred during AM peak hours are in the center city where few cyclists bike on these two types of bicycle facilities.

5.8.2. 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 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_HH_M	3	0.001810	0.002016	0.90	0.3691
Total_HH_S	3	-0.002175	0.000682	-3.19	0.0014
Total_HH	4	0.004110	0.001235	3.33	0.0009
Total_HH	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 affects the level of bicycle counts negatively, which is consistent with the results of cycling behavior during AM peak hours. Different cycling behavior is found in terms of the impact of total population. For example, during PM peak hours, cyclist prefer to bike on roads located in areas with low population, which is opposite to the results concluded from the models based on the whole dataset. The median age variable affects the high-average level of bicycle counts positively, which remains the same effect as mentioned before. However, median household income influences the average-high level of bicycle counts negatively, 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 the slope still has a negative impact on high-average level of bicycle counts.

5.9. Summary

This chapter develops several discrete choice models including ORL model, PPO model, MNL model, and MXL model to analyze the cycling activities. Model comparison is conducted to choose the best model structure for this study, and PPO model outperforms the other discrete choice models. The cycling behavior in 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: BICYCLIST INJURY RISK ANALYSIS

6.1. Introduction

This chapter develops a series of safety performance functions to analyze bicyclist injury risk. The rest of this chapter is organized as follows. Section 6.2 provides the data preparation procedure for the later bicyclist injury risk analysis. Section 6.3 through Section 6.6 present the methodology for analyzing the impacts of cycling safety including Negative Binomial (NB) model, Poisson model, Zero-inflated Negative Binomial (ZINB) model, and Zero-inflated Poisson (ZIP) model. Section 6.7 compares the model estimation results utilizing the goodness of fit and summarizes the model results with different impacts of various explanatory variables. Finally, Section 6.8 concludes this chapter with a summary.

6.2. Data Preparation

The data preparation procedure is similar to the data processing procedure presented Chapter 4 and Chapter 5. This process is conducted mainly using ArcGIS. The primary function used in ArcGIS is spatial join that helps researchers join multiple layers by the same location with different spatial and relative information. Based on the literature review as well as the data availability, the following information including bicycle volume, bicycle-vehicle crashes, road characteristics, sidewalk information, bicycle facilities, bus stops, and AADT is collected for the model development of safety performance functions. The detailed data description and sources are shown in Table 6.1

Table 6.1: Data Description and Sources

Data	Description	Sources
Strava	Bicycle volume data (December 2016 to November	Strava Metro

Data	Description	Sources
	2017) including bicycle counts on each road segment in Charlotte and the Charlotte road network shapefile	
Bike Crashes	Bicycle-vehicle crashes occurred in the city of Charlotte from 2007 to 2017	NCDOT
Road Characteristics	North Carolina road characteristics	NCDOT
Sidewalks	The sidewalk information in the city of Charlotte	Charlotte Open Data Portal
PBIN	Bicycle facilities in North Carolina	NCDOT
AADT	Annual average daily traffic information in North Carolina	NCDOT

To obtain the final combined dataset including the information mentioned above, all the data are imported in ArcGIS, and "spatial join" is utilized to identify the spatial relationships between each dataset. To be specific, the Strava road segment shapefile created based on the OpenStreetMap is used as the base of all the spatial join/table join. First, layers including road characteristics, AADT, sidewalks, bus stops, and bicycle facilities are joined spatially to the base layer (Strava road segment shapefile). Second, Strava data including the bicycle volume on each segment and all the spatial joined layers are compiled together with the same road segment ID to obtain the combined road shapefile. Finally, each bicycle-vehicle crash is assigned to its closest road segment, and the bicycle crash counts on each road segment are rolled up to generate the final complete data for the development of safety performance functions. The data preparation procedure can be seen in Figure 6.1.

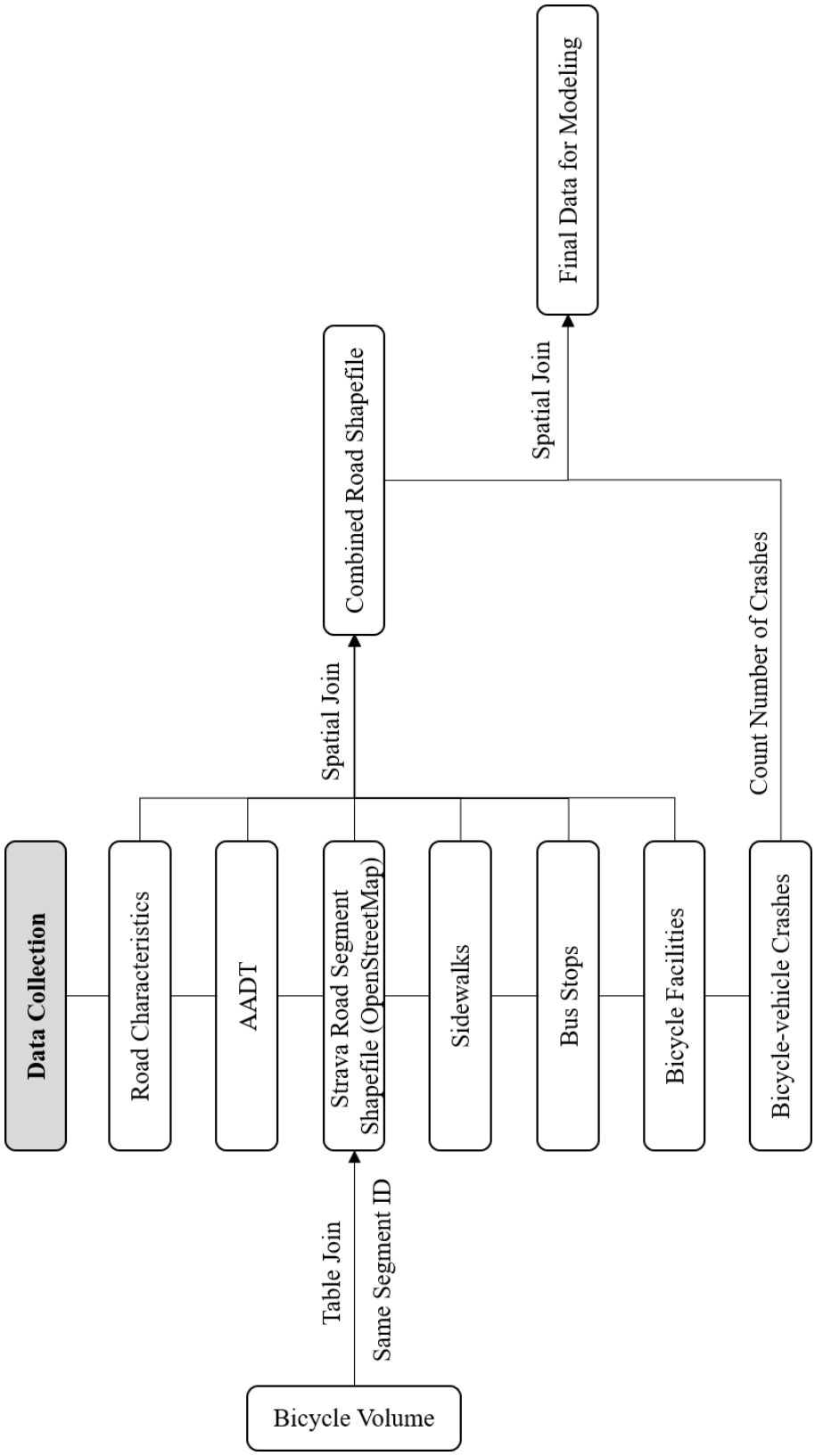


Figure 6.1: Data Preparation Procedure

Before using the combined data, the data is imported in SAS to remove the observations with missing values and convert variables into dummy variables. The detailed explanatory variables considered in the following safety performance functions and their descriptions are presented in Table 6.2.

Table 6.2: Explanatory Variables

Variable	Description
Volume Variables	
AADB	Annual average daily bicycle counts on each road segment
AADT	Annual average daily traffic collected from AADT count stations
Road Characteristics	
Oneway	If the road segment is one way, then oneway = 1, dummy variable
MPLength	The length of the segment in miles.
Functional Classification1	Interstate, dummy variable
Functional Classification2	Principal Arterial, dummy variable
Functional Classification3	Minor Arterial, dummy variable
Functional Classification4	Major Collector, dummy variable
Functional Classification5	Minor Collector, dummy variable
Median	The presence of a median, dummy variable
MedianWidth	The width of the median
SpeedLimit	The posted speed limit on a roadway segment
Sidewalk	The presence of a sidewalk, dummy variable
SidewalkWidth	The width of the sidewalk
Bus_Stop	The presence of a bus stop
Bicycle Facilities	
Bike_Lane	The presence of a bike lane, dummy variable

Variable	Description
Paved_Shoulder	The presence of a paved shoulder, dummy variable

6.3. Poisson Model

Poisson regression model is known as one of the most prevalent models for estimating count data. Many researchers have applied this method to numerous studies regarding transportation count data. In this case, bicycle-vehicle crash counts are studied. Thus, Poisson regression model is applied as a safety performance function to analyze bicyclist injury risk. This Poisson regression model has an assumption, which is the mean equals to its variance, which can be expressed in the following equation:

$$VAR[y_i] = E[y_i]$$

where *VAR* denotes the variance; y_i indicates that segment *i* has *y* times of crashes happened in the studied time period; *E* represents the expected mean. The number of *y* crashes follows a Poisson distribution with a condition mean and the characteristics of an individual are related to the number of crashes. The expected value of crashes *y* and the association with the considered explanatory variables are shown in the following equation:

$$\mu_i = EXP(\beta X_i)$$

where *EXP* means the exponential; β denotes the estimated coefficient corresponding to the independent variable X_i; μ_i is the expected value of the dependent variable representing the total number of bicycle-vehicle crashes happened at a specific segment.

The probability of a segment i experiencing bicycle-vehicle crashes during the certain research period is shown as the following equation:

$$P(y_i) = \frac{EXP(-\mu_i)\mu_i{}^{y_i}}{y_i!}$$

where $P(y_i)$ represents the probability of y_i crashes occurred on a segment *i*; μ_i denotes the Poisson parameter for the specific segment, which equals to $E[y_i]$.

6.4. Negative Binomial Model

Although Poisson regression is a prevalent method for modeling transportation count data, it has the assumption mentioned in the above section that the mean equals to the variance. This assumption may bring bias to the model estimation results. In addition, bicycle crash count data are usually over-dispersed based on the previous research studies, which shows a higher variance than the sample mean. Hence, NB model is developed to address the over-dispersed issue. The following equation shows the relationship between the dependent and independent variables:

$$\mu_i = EXP(\beta X_i + \varepsilon_i)$$

where ε denotes the random error term that represents the unobserved attributes neglected in the NB model. It is assumed that the error term has no correlation with *X*. $EXP(\varepsilon_i)$ means a disturbance term that follows Gamma distribution, where mean equals to 1 and variance equals to α. With this distinctive term, the variance is not restricted to be the same as the value of the mean. This can be expressed in the following equation:

$$VAR[y_i] = E[y_i]\left[1 + \alpha E[y_i]\right] = E[y_i] + \alpha E[y_i]^2$$

As is seen in the above equation, it can be interpreted that if the overdispersion parameter α equals to 0, the variance will be the same as the value of the mean. The probability function of the NB model is shown by the following equation:

$$P(y_i|X_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\mu_i}{\frac{1}{\alpha}+\mu_i}\right)^{y_i}$$

where Γ represents the gamma distribution function.

6.5. Zero-inflated Poisson Model

One of the critical phenomena that cannot be neglected is that the number of observations with zero crash during a certain study period can be an issue to the model estimation. It can be found that zero crash may occurred on numerous roadway segments. This problem is common since many road segments have no crash record.

In order to solve the zero-state issue, Zero-inflated Negative Binomial model and Zero-inflated Poisson model are developed based on the zero model from the method of modeling with zero. These two models separate the model estimation process into two splitting means for zero counts and non-zero counts respectively.

It is assumed in the Zero-inflated Poisson model that the crashes $Y = (y_1, y_2, \dots, y_n)$ occurred on road segments are independent and the probability functions for zero count and non-zero counts are shown in the following equations:

$$y_i = 0\ with\ probability\ p_i + (1 - p_i)\exp(-u_i)$$

$$y_i = y\ with\ probability\ \frac{(1 - p_i)\exp(-u_i){u_i}^y}{y!}$$

where p_i is the probability of experiencing zero observation, y_i is the number of crashes occurred on a specific road segment during research period, where $u_i = \exp(\beta X_i)$. The variance is shown in the following equation:

$$VAR[y_i|X_i, Z_i] = u_i(1 - p_i)(1 + u_i p_i)$$

6.6. Zero-inflated Negative Binomial Model

Similar to ZIP model, ZINB model also splits the underlying data generating process into two regimes. It is an extension of Negative Binomial model, which solves the zero-state problem.

The ZINB model is presented in the following equations:

$$y_i = 0\ with\ probability\ p_i + (1 - p_i)(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + u_i})^{\frac{1}{\alpha}}$$

$$y_i = y\ with\ probability\ (1 - p_i)[\frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + u_i}\right)^{\frac{1}{\alpha}}\left(\frac{u_i}{\frac{1}{\alpha} + u_i}\right)^{y_i}]$$

where the disturbance term following Gamma distribution has the mean of 1 and the variance of α. The variance of Zero-inflated Negative Binomial model is shown as follows:

$$VAR[y_i|X_i, Z_i] = u_i(1 - p_i)(1 + u_i(p_i + \alpha))$$

6.7. Model Result Analysis

To analyze the bicyclist injury risk on road segments and explore the impact factors on the bicycle-vehicle crash counts in the city of Charlotte, several safety performance functions including NB model, Poisson model, ZINB model, and ZIP model are developed. Explanatory variables (presented in Table 6.2) are carefully selected for the model estimation based on the literature review as well as the data availability.

All the explanatory variables presented in Table 6.2 are first included in the safety performance functions to analyze the probability of certain crash counts. The maximum likelihood method is applied to conduct the model parameter estimation. SAS 9.4 is utilized to conduct the model estimation procedure. To keep the variables that affects the

crash counts on the roadway segments significantly, the backward selection demand is utilized. The final model results for the four safety performance functions with significant variables only are shown in the following tables.

Table 6.3: Poisson Model Estimation Results

Analysis of Maximum Likelihood Parameter Estimates						
Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	-3.4211	0.0502	-3.5195	-3.3227	4643.27	<.0001
AADB	0.0002	0.0000	0.0002	0.0002	65.91	<.0001
Interstate	-0.4781	0.2473	-0.9627	0.0066	3.74	0.0532
Principal_Arterial	0.6010	0.1034	0.3984	0.8036	33.80	<.0001
Minor_Arterial	0.5042	0.1046	0.2992	0.7092	23.24	<.0001
Major_Collector	0.4612	0.1159	0.2340	0.6884	15.83	<.0001
Minor_Collector	0.5449	0.3055	-0.0538	1.1437	3.18	0.0745
Bus_Stop	1.2603	0.0787	1.1061	1.4146	256.41	<.0001
Bike_Lane	0.6181	0.1103	0.4020	0.8342	31.43	<.0001

Table 6.4: Negative Binomial Model Estimation Results

Analysis of Maximum Likelihood Parameter Estimates						
Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	-3.4578	0.0551	-3.5658	-3.3499	3944.33	<.0001
AADB	0.0002	0.0000	0.0002	0.0003	48.85	<.0001
Interstate	-0.4791	0.2580	-0.9847	0.0265	3.45	0.0633
Principal_Arterial	0.6338	0.1192	0.4001	0.8675	28.25	<.0001
Minor_Arterial	0.5029	0.1209	0.2659	0.7399	17.30	<.0001
Major_Collector	0.5144	0.1352	0.2494	0.7794	14.48	0.0001
Minor_Collector	0.6838	0.3469	0.0039	1.3638	3.89	0.0487
Bus_Stop	1.3159	0.0937	1.1322	1.4996	197.08	<.0001
Bike_Lane	0.6653	0.1355	0.3998	0.9309	24.11	<.0001

Table 6.5: Zero-inflated Poisson Model Estimation Results

Analysis of Maximum Likelihood Parameter Estimates						
Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	-2.0783	0.1007	-2.2756	-1.8810	426.12	<.0001
Interstate	-0.5033	0.2543	-1.0017	-0.0049	3.92	0.0478
Principal_Arterial	0.5817	0.1143	0.3577	0.8057	25.90	<.0001
Minor_Arterial	0.4544	0.1154	0.2283	0.6806	15.51	<.0001
Major_Collector	0.4507	0.1288	0.1984	0.7031	12.26	0.0005
Minor_Collector	0.6943	0.3548	-0.0011	1.3896	3.83	0.0504
Bus_Stop	1.2172	0.0904	1.0400	1.3945	181.22	<.0001
Bike_Lane	0.6704	0.1256	0.4242	0.9166	28.49	<.0001
Analysis of Maximum Likelihood Zero Inflation Parameter Estimates						
Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1.0395	0.1190	0.8064	1.2726	76.37	<.0001
AADB	-0.0004	0.0001	-0.0005	-0.0002	19.64	<.0001

Table 6.6: Zero-inflated Negative Binomial Model Estimation Results

Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	-2.957305	0.063632	-46.48	<.0001
Interstate	-0.472699	0.259744	-1.82	0.0688
Principal_Arterial	0.537115	0.118766	4.52	<.0001
Minor_Arterial	0.318137	0.118765	2.68	0.0074
Major_Collector	0.402069	0.130228	3.09	0.0020
Bus_Stop	1.111739	0.092795	11.98	<.0001
Bike_Lane	0.659989	0.128543	5.13	<.0001
Inf_Intercept	0.704238	0.169307	4.16	<.0001
Inf_AADB	-0.113304	0.029687	-3.82	0.0001

To compare the four safety performance functions, the indicators for model comparison mentioned before in Section 5.7 is adopted. Therefore, the indicators for each model including NB model, Poisson model, ZINB model, and ZIP model are presented in the following table.

Table 6.7: Indicators for Model Comparison

Model	No. of Obs (Q)	No. of Vars. (p)	-2LogL	AIC	BIC
Poisson Model	15664	9	6312	6329	6398
NB Model	15664	10	6156	6176	6252
ZIP Model	15664	10	6188	6208	6285
ZINB Model	15664	10	6090	6110	6186

As mentioned before in Chapter 5, smaller values of the indicators represent better fitness. Comparing the four models with the values of -2LogL, AIC, and BIC, ZINB model outperforms the other three safety performance functions. This model comparison result is not hard to infer, since the estimation procedure of ZINB model is a splitting data modeling process that consider the zero-state issue. The crash data utilized in this research study contain a lot of road segments with zero crashes, which may lead to biases when developing traditional Poisson model or Negative Binomial model. Therefore, it is confirmed that ZINB is the best fit for this bicyclist injury risk analysis.

Summarizing the model estimation results presented in Table 6.3, Table 6.4, Table 6.5, and Table 6.6, variables that have significant impacts on bicyclist injury risk including annual average daily bicycle counts, interstate roads, minor arterials, principal arterials, minor collectors, major collectors, the presence of bus stops, and the presence of

a bicycle lane are identified and will be interpreted in detail. The explanation of the impacts of significant variables on bicyclist injury risk is provided as follows:

1. Volume variables:

As expected, the annual average daily bicycle counts affect the crashes occurred on a road segment significantly. The number of bicycle counts on a road segment has a positive impact on the bicyclist injury risk. In other words, if the road segment has more bicycle counts, the probability of higher injury risk on this road segment is greater. In the ZINB and ZIP models, the annual average daily bicycle counts are included in the zero-inflation parameter estimation. In this process, the effect of the AADB is different from that of the Poisson model and Negative Binominal model. It can be interpreted that the higher bicycle counts on a road segment, the smaller probability of obtaining zero bicycle-vehicle crashes.

2. Road characteristics:

Interstate roads, minor arterials, principal arterials, minor collectors, and major collectors all have significant impacts on the bicyclist injury risk. It can be seen that the function classification of a road segment is the major impact on the bicycle-vehicle crash counts. The interstate roads have a negative impact on bicyclist injury risk, while minor arterials, principal arterials, minor collectors, and major collectors affect the cycling safety positively. This result indicates that the likelihood of higher crash counts on minor arterials, principal arterials, minor collectors, and major collectors is higher, while the probability of crashes occurred on interstate roads is lower.

In addition, the presence of bus stops on a road segment has a positive impact on bicyclist injury risk, which indicates that the presence of bus stops may increase the

probability of more bicycle-vehicle crashes. It can be imagined that if a bus stop is located on a road segment, the conflict of bicyclists and bus may increase the probability of a bicycle-vehicle crash.

3. Bicycle facilities:

The presence of a bike lane on a roadway segment affects the bicyclist injury risk significantly. Interestingly, it is likely to increase the probability of crashes, which might be different from the expectation. This result may be related to the bicycle facility condition, and the higher likelihood of more cycling activities on bike lanes.

6.8. Summary

This chapter develops several safety performance functions including NB model, Poisson model, ZINB model, and ZIP model to analyze the bicyclist injury risk. Model comparison is conducted to select the best model structure for this research study. Impact factors that are associated with the number of bicyclist-involved crashes occurred on roadway segments in city of Charlotte are identified and interpreted.

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 to explore the impacts on bicycle volume on a road segment in the whole city network and the bicyclist injury risk. One of the most critical issues that need to be considered for the bicycle volume estimation and prediction, cycling activity modeling, and bicyclist injury risk 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 can address the issues brought by traditional data collection methods and provide the temporal and spatial information on cycling to bridge the data gap.

Based on the crowdsourced bicycle data collected from the Strava application, this research study is conducted to estimate the bicycle volume on most of the road segments in the City of Charlotte, to analyze cycling activities, and to develop safety performance functions to analyze cycling safety.

The rest of this chapter is organized as follows. Section 7.2 provides a brief review of the methods used to conduct the bicycle volume prediction, cycling activity modeling, and safety analysis based on the novel crowdsourced bicycle data. The model

results are concluded in this section, and the model comparison results indicating the best model structure for this research study are summarized. Different cycling activities during AM and PM hours are concluded in this section and policy-related recommendations are provided here. Section 7.3 discusses the limitation of this study and provides the future research directions in order to improve the research and also further enhance cycling environment and safety.

7.2. Summary and Conclusions

The primary objectives of this research are to estimate and predict the bicycle volume on each roadway segment, model the cycling activities based on crowdsourced bicycle data, and conduct cycling safety analysis. 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.

Crowdsourced bicycle data from Strava smartphone application are combined with a series of other relevant data including NC road characteristics data, demographic data, slope data, manual count data from continuous count stations in Charlotte, temporal data, and bicycle facility data, etc. Data comparison is conducted to demonstrate the differences between manual count data and Strava bicycle count data. Data process and combination procedures are completed using ArcGIS and SAS. Based on the combined data, two linear regression models are developed. The relationship between manual count data and Strava data as well as other relevant data is analyzed. To be specific, variables including weekday, time period except 00:00-06:00 am, Strava user counts, off-street

paths, bike lanes, and suggested bike routes have significant impacts on the total bicycle volume on a road segment, where cycling during time period except 00:00-06:00 am, Strava user counts and cycling on off-street paths have positive impacts on the total bicycle volume, while cycling on weekdays and bicycle facilities including bike lanes and suggested bike routes have negative impacts on total bicycle volume. Bicycle volume on most of the road segments in the City of Charlotte is predicted using the developed model. A bicycle ridership map is created to have a graphical view of the bicycle counts for the whole road network.

Several discrete choice models are developed to analyze cycling activities in the City of Charlotte. Models including ORL model, PPO model, MNL model, and MXL model are compared to select the best-fit model for this cycling activity analysis. 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 affect the level of bicycle counts negatively, 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 affect the level of bicycle counts positively 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 affect different levels variously. 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 ORL model. The

explanatory variables that have significant impacts 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 the best model structure for this cycling activity analysis. To explore the different cycling activities for both AM and PM peak hours, a MNL model and a MXL model are developed. Impact factors that are associated with different levels of bicycle counts in the City of Charlotte are identified.

In addition, several safety performance functions are developed to analyze bicyclist injury risk on road segments in the city of Charlotte. Models including NB model, Poisson model, ZINB model, and ZIP model are compared to identify the best fit for this cycling safety analysis. ZINB is identified to outperform the other three models. Variables including AADB, minor arterials, principal arterials, minor collectors, major collectors, and the presence of bus stops and a bike lane on a road segment all have positive impacts on bicyclist injury risk, while the interstate roads affects the number of bicycle-vehicle crashes on a road segment negatively.

According to the bicycle volume estimation model results and the bicyclist injury risk analysis obtained and conclusions made in this research, some policy-related recommendations can be provided as follows:

1. Based on the modeling results that bicyclists prefer off street paths, planners can design more off-street paths to offer better bike environment for bicyclists in the City of Charlotte.
2. To promote biking to work, the locations of the off-street paths need to be constructed in the uptown area. Since there are a lot of traffic in Charlotte uptown area, especially during peak hours, and the bicycle volume is higher there compared to other locations, constructing more off street paths should attract more bicyclists to choose to bike other than private cars and public transit (for short-distance trips).
3. According to the modeling results, the predicted bicycle volume on road segments related to parks and greenways in the City of Charlotte has a higher number. To encourage recreational bicycle trips, the bicycle facilities in the park or greenway area should be improved.
4. It is important to identify the right of way on a roadway segment with bus stops. It is recommended to constructed separated bike facilities especially for bicyclists to avoid crashes.

If the above policy-related recommendations are followed, better bike environment and cycling safety can be provided for the citizens in Charlotte to improve their quality of life and to mitigate traffic congestion to some extent.

7.3. Directions for Future Research

In this section, some of the limitations of this research are pointed out and the directions for future research are also provided. The limitations of this research can be summarized as follows:

1. Bicycle volume:

(1) The bicycle manual count data have a limitation in the model development. The availability of more count data from the bicycle count stations may improve the model results.

(2) The manual count data are collected from the count stations which are located in the center city. Most of the bicycle trips might be related to commuting trips based on the trip locations. Since a large portion of the manual count data that are used to predict bicycle volume might be commuting trips, and the bicycle volume in the uptown area might be higher than other locations in the City of Charlotte, biases might exist when predicting the bicycle volume.

(3) The two bicycle volume regression models are developed in the urban cycling environment. Situations may vary in the rural area and in other metropolitan centers, and as such, the model developed for predicting the bicycle volume may not be representative in those cases.

(4) 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.

2. Cycling activities:

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 limitations of this study and the literature review on relevant studies, some improvements can be made in future studies.

1. Since more manual count stations are under construction now, with more bicycle manual count data, the bicycle volume regression models can be improved.
2. Other models can be developed and tested to see if there is a better fitness for relevant research studies.
3. The cycling activities occurred in various locations can be different. Comparison can be conducted for cycling activities in different locations (e.g., urban or rural areas).
4. Bicyclist injury risk at intersections can be examined since it is more likely to experience crashes at intersections. It can be analyzed and compared with the cycling safety on roadway segments.

REFERENCES

Al-Fuqaha, A., Oh, J. S., Kwigizile, V., Mohammadi, S., and Alhomadat, F. 2017. Integrated Crowdsourcing Platform to Investigate Non-Motorized Behavior and Risk Factors on Walking, Running, and Cycling Routes (No. TRCLC 15-06). Western Michigan University. Transportation Research Center for Livable Communities.

Attard, J., Orlandi, F., Scerri, S., and Auer, S. 2015. A systematic review of open government data initiatives. Government Information Quarterly, 32(4): 399-418.

Best cycling apps – 16 of the best iPhone and Android apps to download. 2019. Available at https://www.bikeradar.com/advice/buyers-guides/best-cycling-apps/

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): 495-507.

Bhat, C.R. 2000. Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. Transportation science, 34(2): 228-238.

Blanc, B., and Figliozzi, M. 2016. Modeling the impacts of facility type, trip characteristics, and trip stressors on cyclists' comfort levels utilizing crowdsourced data. Transportation Research Record, 2587(1): 100-108.

Blanc, B., and Figliozzi, M. 2017. Safety Perceptions, Roadway Characteristics, and Cyclists' Demographics: A Study of Crowdsourced Smartphone Bicycle Safety Data (No. 17-03262).

Blanc, B., Figliozzi, M., and Clifton, K. 2016. How representative of bicycling populations are smartphone application surveys of travel behavior? Transportation research record, 2587(1): 78-89.

Brabham, D. C. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. Convergence, 14(1): 75-90.

Brabham, D. C. 2013. Crowdsourcing. Mit Press.

Chanal, V., and Caron-Fasan, M. L. 2008. How to invent a new business model based on crowdsourcing: the Crowdspirit® case.

Charlton, B., Sall, E., Schwartz, M., and Hood, J. 2011. Bicycle route choice data collection using GPS-enabled smartphones. In Transportation Research Board 90th Annual Meeting, 23-27 January 2011.

Chatzimilioudis, G., Konstantinidis, A., Laoudias, C., and Zeinalipour-Yazti, D. 2012. Crowdsourcing with smartphones. IEEE Internet Computing, 16(5): 36-44.

Chen, C. 2017. Crowdsourcing Data-driven Development of Bicycle Safety Performance Functions (SPFs): Microscopic and Macroscopic Scales.

Chen, P., and Shen, Q. 2016. A gps-based analysis of built environment influences on bicyclist route preferences (No. 16-1948).

Chen, P., Shen, Q., and Childress, S. 2018. A GPS data-based analysis of built environment influences on bicyclist route preferences. International journal of sustainable transportation, 12(3): 218-231.

Chen, P., Zhou, J., and Sun, F. 2017. Built environment determinants of bicycle volume: A longitudinal analysis. Journal of Transport and Land Use, 10(1).

City of Charlotte Department of Transportation. 2017. Charlotte BIKES Bicycle Plan. Available at https://charlottenc.gov/Transportation/Programs/Documents/Charlotte%20BIKES%20Final.pdf

El Esawey, M., Mosa, A. I., and Nasr, K. 2015. Estimation of daily bicycle traffic volumes using sparse data. Computers, Environment and Urban Systems, 54: 195-203.

Esawey, M. E. 2014. Estimation of annual average daily bicycle traffic with adjustment factors. Transportation Research Record, 2443(1): 106-114.

Esawey, M. E., and Mosa, A. I. 2015. Determination and application of standard K factors for bicycle traffic. Transportation research record, 2527(1): 58-68.

Estellés-Arolas, E., and González-Ladrón-De-Guevara, F. 2012. Towards an integrated crowdsourcing definition. Journal of Information science, 38(2): 189-200.

Estellés-Arolas, E., Navarro-Giner, R., and González-Ladrón-de-Guevara, F. 2015. Crowdsourcing fundamentals: definition and typology. In Advances in crowdsourcing. Springer, Cham.

Figliozzi, M. A., and Blanc, B. P. 2015. Evaluating the use of crowdsourcing as a data collection method for bicycle performance measures and identification of facility improvement needs.

Figliozzi, M. A., and Blanc, B. P. 2015. Evaluating the use of crowdsourcing as a data collection method for bicycle performance measures and identification of facility improvement needs.

Griffin, G. P., and Jiao, J. 2019. Crowdsourcing Bicycle Volumes: Exploring the role of volunteered geographic information and established monitoring methods.

Guan, H. Discrete choice Modeling. 2004.

Hensher, D.A., and Greene, W.H. 2003. The mixed logit model: the state of practice. Transportation, 30(2): 133-176.

Hirsch, J. A., James, P., Robinson, J. R., Eastman, K. M., Conley, K. D., Evenson, K. R., and Laden, F. 2014. Using MapMyFitness to place physical activity into neighborhood context. Frontiers in public health, 2: 19.

Hochmair, H. H., Bardin, E., and Ahmouda, A. 2019. Estimating bicycle trip volume for Miami-Dade county from Strava tracking data. Journal of Transport Geography, 75: 58-69.

Hood, J., Sall, E., and Charlton, B. 2011. A GPS-based bicycle route choice model for San Francisco, California. Transportation letters, 3(1): 63-75.

Hosseini, M., Phalp, K., Taylor, J., and Ali, R. 2014. The four pillars of crowdsourcing: A reference model. In 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS), 1-12. IEEE.

Howe, J. 2006. The rise of crowdsourcing. Wired magazine, 14(6): 1-4.

Howe, J. 2008. Crowdsourcing: How the power of the crowd is driving the future of business. Random House.

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.

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 (No. 14-4637).

Jestico, B., Nelson, T., and Winters, M. 2016. Mapping ridership using crowdsourced cycling data. Journal of transport geography, 52: 90-97.

Jestico, B. 2016. Crowdsourced data as a tool for cycling research on ridership trends and safety in the Capital Regional District (Doctoral dissertation).

Kagerbauer, M., Hilgert, T., Schroeder, O., and Vortisch, P. 2015. Household travel survey of intermodal trips–Approach, challenges and comparison. Transportation research procedia, 11: 330-339.

Kitchin, R. 2014. Big data should complement small data, not replace them. LSE Impact blog, 27.

Kitchin, R. 2014. The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Kleemann, F., Voß, G. G., and Rieder, K. 2008. Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing. Science, technology and innovation studies, 4(1): 5-26.

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.

Kučera, J., Chlapek, D., and Nečaský, M. 2013. Open government data catalogs: Current approaches and quality perspective. In International conference on electronic government and the information systems perspective, 152-166. Springer, Berlin, Heidelberg.

La Vecchia, G., and Cisternino, A. 2010. Collaborative workforce, business process crowdsourcing as an alternative of BPO. In International Conference on Web Engineering, 425-430. Springer, Berlin, Heidelberg.

LaMondia, J., and Watkins, K. 2017. Using Crowdsourcing to Prioritize Bicycle Route Network Improvements.

Laney, D. 2001. 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70): 1.

Lewin, A. 2011. Temporal and weather impacts on bicycle volumes (No. 11-2536).

Lu, T., Buehler, R., Mondschein, A., and Hankey, S. 2017. Designing a bicycle and pedestrian traffic monitoring program to estimate annual average daily traffic in a small rural college town. Transportation research part D: transport and environment, 53: 193-204.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., and Barton, D. 2012. Big data: the management revolution. Harvard business review, 90(10): 60-68.

McFadden, D., and Train, K. 2000. Mixed MNL models for discrete response. Journal of applied Econometrics, 15(5): 447-470.

Miranda-Moreno, L. F., Nosal, T., Schneider, R. J., and Proulx, F. 2013. Classification of bicycle traffic patterns in five North American Cities. Transportation research record, 2339(1): 68-79.

Misra, A., and Watkins, K. 2018. Modeling Cyclist Route Choice using Revealed Preference Data: An Age and Gender Perspective. Transportation Research Record, 2672(3): 145-154.

Misra, A., Gooze, A., Watkins, K., Asad, M., and Le Dantec, C. A. 2014. Crowdsourcing and its application to transportation data collection and management. Transportation Research Record, 2414(1): 1-8.

Moore, M. 2015. Modeling Factors Influencing Commuter Cycling Routes: A Study of GPS Cycling Records in Auburn, Alabama.

Musakwa, W., and Selala, K. M. 2016. Mapping cycling patterns and trends using Strava Metro data in the city of Johannesburg, South Africa. Data in brief, 9: 898-905.

Pelletier, M. P., Trépanier, M., and Morency, C. 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies, 19(4): 557-568.

Proulxa, F. R., and Pozdnukhovb, A. 2017. Bicycle Traffic Volume Estimation Using Geographically Weighted Data Fusion. Manuscript submitted for publication.

Pucher, J., and Dijkstra, L. 2003. Promoting safe walking and cycling to improve public health: lessons from the Netherlands and Germany. American journal of public health, 93(9): 1509-1516.

RenoTracks. RenoTracks. 2013. Available at http://renotracks.nevadabike.org/.

Revelt, D., and Train, K. 1998. "Mixed logit with repeated choices: households' choices of appliance efficiency level." Review of economics and statistics, 80(4): 647-657.

Roll, J. CycleLane Smart Phone Application Data Summary. Central Lane Metropolitan Planning Organization, Eugene, Ore., 2014. http://www.lcog.org/documentcenter/view/3577.

Romanillos, G., Zaltz Austwick, M., Ettema, D., and De Kruijf, J. 2016. Big data and cycling. Transport Reviews, 36(1): 114-133.

Saad, M., Abdel-Aty, M., Lee, J., and Cai, Q. 2019. Bicycle Safety Analysis at Intersections from Crowdsourced Data. Transportation Research Record, 0361198119836764.

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.

Saxton, G. D., Oh, O., and Kishore, R. 2013. Rules of crowdsourcing: Models, issues, and systems of control. Information Systems Management, 30(1): 2-20.

Schenk, E., and Guittard, C. 2011. Towards a characterization of crowdsourcing practices. Journal of Innovation Economics Management, (1): 93-107.

Schmiedeskamp, P., and Zhao, W. 2016. Estimating daily bicycle counts in Seattle, Washington, from seasonal and weather factors. Transportation research record, 2593(1): 94-102.

Selala, M. K., and Musakwa, W. 2016. The potential of strava data to contribute in non-motorised transport (Nmt) planning in Johannesburg.

Skszek, Sherry L. State-of-the-art report on non-traditional traffic counting methods. No. FHWA-AZ-01-503. Arizona. Dept. of Transportation, 2001.

Strauss, J., and Miranda-Moreno, L. F. 2017. Speed, travel time and delay for intersections and road segments in the Montreal network using cyclist Smartphone GPS data. Transportation research part D: transport and environment, 57: 155-171.

Strauss, J., Miranda-Moreno, L. F., and Morency, P. 2013. Cyclist activity and injury risk analysis at signalized intersections: A Bayesian modelling approach. Accident Analysis & Prevention, 59: 9-17.

Strauss, J., Miranda-Moreno, L. F., and Morency, P. 2014. Multimodal injury risk analysis of road users at signalized and non-signalized intersections. Accident Analysis and Prevention, 71: 201-209.

Strauss, J., Miranda-Moreno, L. F., and Morency, P. 2015. Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. Accident Analysis & Prevention, 83: 132-142.

Strauss, J., Zangenehpour, S., Miranda-Moreno, L. F., and Saunier, N. 2017. Cyclist deceleration rate as surrogate safety measure in Montreal using smartphone GPS data. Accident Analysis & Prevention, 99: 287-296.

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): 274.

Sun, Y., Du, Y., Wang, Y., and Zhuang, L. 2017. Examining associations of environmental characteristics with recreational cycling behaviour by street-level Strava data. International journal of environmental research and public health, 14(6): 644.

Świeszczak, M., and Świeszczak, K. 2016. Crowdsourcing–what it is, works and why it involves so many people? World Scientific News, 48: 32-40.

Thomas, L., and Levitt, D. 2017. North Carolina Bicycle Crash Facts 2008-2012. Available at http://www.pedbikeinfo.org/pbcat_nc/_biketypefacts.cfm.

Train, K. E. 2009. Discrete choice methods with simulation. Cambridge university press.

von Stülpnagel, R., and Krukar, J. 2018. Risk perception during urban cycling: An assessment of crowdsourced and authoritative data. Accident Analysis & Prevention, 121: 109-117.

Vukovic, M. 2009. Crowdsourcing for enterprises. In 2009 congress on services-I, IEEE, 686-692.

Wang, H., Chen, C., Wang, Y., Pu, Z., and Lowry, M. B. 2016. Bicycle Safety Analysis: Crowdsourcing Bicycle Travel Data to Estimate Risk Exposure and Create Safety Performance Functions.

Watkins, K., Ammanamanchi, R., LaMondia, J., and Le Dantec, C. A. 2016. Comparison of smartphone-based cyclist GPS data sources (No. 16-5309).

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: 183-196.