Center for Advanced Multimodal Mobility Solutions and Education

Project ID: 2018 Project 03

EVALUATING THE POTENTIAL USE OF CROWDSOURCED BICYCLE DATA IN NORTH CAROLINA

Final Report

by

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

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

for

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

September 2019

ACKNOWLEDGEMENTS

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

DISCLAIMER

Table of Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

Cycling, 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 cyclist's route choice on each roadway segment, so as to quantify the impact of certain attributes on bicycle volume and further provide better cycling environment for cyclists to encourage non-motorized travel.

To map ridership, data including network characteristics, sociodemographic factors, time of day, and day of week, are quite indispensable. There have been multiple 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 to collect the data of interest. However, such methods are expensive and time-consuming. Crowdsourced data, on the contrary, are cost effective and time-saving, and therefore they have been widely collected and used by many public agencies and private sectors. 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 focuses on evaluating the potential use of crowdsourced bike data and compare them with the traditional bike counting data that are collected in the City of Charlotte. Using the bike data both from the Strava smartphone cycling application and from the bicycle count stations, the bicycle volume models are developed. Based on the results, a predictive model is concluded, 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. An ordered probit model is developed to analyze the Strava users' cycling route segment choice. Finally, recommendations are made in order to help improve the cycling environment and increase the bicycle volume in the future.

Chapter 1. Introduction

1.1 Problem Statement

With the increase in the 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 the city, in return, will benefit from it to have healthier and more energetic population.

According to the Charlotte Department of Transportation (CDOT) Bicycle Program (City of Charlotte Department of Transportation, 2017), Charlotte is making every effort to offer an inclusive and comfortable cycling environment for all the potential bicyclists including all ages and abilities to provide them 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, policy-makers, 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 bicycle network in the City of Charlotte has grown to contain more than 90 miles of bicycle lanes, 55 miles of signed routes and 40 miles of greenways and off-street paths (City of Charlotte Department of Transportation, 2017). According to a cycling survey conducted by CDOT (City of Charlotte Department of Transportation, 2017), 51% of the residents in Charlotte would like to travel by bicycle 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 condition in Charlotte. 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 bicycle volume on road segments, data including network characteristics, sociodemographic information, location-specific elements, temporal factors, 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 the 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 time-saving, and therefore they have been widely collected and used by many public agencies and private sectors. Among all the crowdsourced data, data collected from smartphone applications including Strava 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 obtain data from a broader range of people in a shorter and more cost-efficient way. This method was first introduced by Howe in his "The Rise of Crowdsourcing" article (Howe, 2006). 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 crowd to address common issues that affect them all. Recently, crowdsourcing techniques have developed rapidly. Some studies regarding its use in transportation have shown its immense potential in augmenting or replacing 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, path size logit models, expanded path size logit models, C-logit models, and recursive models. These models can be utilized to analyze the bicycle travels in terms of bicycle volume estimation, bicyclist route choice 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 models and a preferred bicyclist route segment choice model. Crowdsourced bicycle data from Strava smartphone application are collected and combined with other relevant data (including NC road characteristics data, demographic data, slope data, manual count data from count stations in the City of 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, an ordered probit model is developed to analyze the Strava users' cycling route choice in the City of Charlotte. Finally, the conclusion is made to summarize the whole study, and directions for future research are also provided.

1.2 Objectives

The objective of this report is to evaluate the potential use of crowdsourced bicycle data in the City of Charlotte and to utilize the promising crowdsourced bicycle data to develop bicycle volume prediction models as well as to generate a ridership map using the bicycle volume prediction model. The proposed work in this report is to fulfill the following objectives:

1. To compile bicycle data from all the available sources including Strava, bicycle manual count data, NC road characteristics data, demographic data, slope data, temporal data, and bicycle facility data as preparation of the follow-up work;

2. To combine all the collected data using ArcGIS and SAS for model estimation;

3. To develop linear regression 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 an ordered logit model to analyze the impact of different variables on Strava bicycle count in the City of Charlotte.

1.3 Expected Contributions

To provide a better cycling environment and encourage more potential bicyclists to bicycle 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. Along that line, the expected contributions of this research can be 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 an ordered probit model to analyze the factors contributing to Strava bicycle count 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.

1.4 Report Overview

The remainder of this report is organized as follows: Chapter 2 presents a comprehensive review of the state-of-the-art and state-of-the-practice on the potential use of crowdsourced bicycle data. Chapter 3 discusses the bicycle count data collected from Strava application and other relevant supporting data. Chapter 4 conducts a descriptive analysis based on the data

collected in Chapter 3. Chapter 5 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 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 6 develops an ordered probit model to analyze the Strava users' cycling route segment choice. Finally, Chapter 7 concludes this report with a summary and a discussion of the directions for future research.

Chapter 2. Literature Review

2.1 Introduction

Data collection is one of the basic yet most important parts in modeling. In some cases, the most time-consuming process of the whole research is the data collection. Therefore, selecting an effective and advanced method for data collection is essential.

This chapter provides a review of previous research efforts regarding the crowdsourcing data collection method and the potential use of the crowdsourced data. The comprehensive review will greatly help gain a clearer understanding of the potential use of crowdsourced data in various aspects of future research studies.

The remainder of this chapter is structured as follows. Section 2.2 gives a brief introduction to several data collection methods including crowdsourcing, open data (including the concept of big data), stated preference survey, revealed preference survey, and traditional survey methods. Section 2.3 presents a list of smartphone crowdsourcing applications and the information collected from each crowdsourcing application. Section 2.4 summarizes the potential use of crowdsourced data, which includes the use of crowdsourced data for route choice analysis, volume estimation, traffic safety analysis, pollution exposure research, and health influence, etc. Section 2.5 provides the currently prevalent route choice analysis methods. Finally, section 2.6 concludes the whole chapter.

2.2 Data Collection Methods

2.2.1 Crowdsourcing

Crowdsourcing is an advanced and innovative method, which brings new improvements and developments in data collection. The definition of crowdsourcing has evolved over the years. It was first introduced by Howe in his "The Rise of Crowdsourcing" article. He defined crowdsourcing as follows:

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

(2) "The application of Open Source principles to fields outside of software." (Howe, 2008)

Subsequently, various interpretations of crowdsourcing in distinctive fields are presented. According to Estellés-Arolas and González-Ladrón-de-Guevara (2012), different researchers have different definitions for crowdsourcing. Most of the definitions of crowdsourcing involve three main features, which include the crowd itself, the outsourcing procedure, and an internet-based platform (Saxton, 2013). For example, Schenk (2011) presented that crowdsourcing implied individuals participate voluntarily to achieve the task, which would tend to motivate both the experts and the individuals to find solutions for the tasks. Brabham (2008) stated that "crowdsourcing is an online, distributed problem-solving and production model that has emerged in recent years". Other relevant definitions for crowdsourcing are summarized in TABLE 2.1.

Table 2.1 Summary of Crowdsourcing Definitions

Author	Year	Crowdsourcing Definitions
Kleeman et al.	2008	"A form of the integration of users or consumers in internal processes of value creation. The essence of crowdsourcing is the intentional mobilization for commercial exploitation of creative ideas and other forms of work performed by consumers."
Vukovic	2009	"A new online distributed problem-solving and production model in which networked people collaborate to complete a task."
Liu and Porter	2010	"The outsourcing of a task or a job, such as a new approach to packaging that extends the life of a product, to a large group of potential innovators and inviting a solution. It is essentially open in nature and invites collaboration within a community."
Wexler	2011	"Focal entity's use of an enthusiastic crowd or loosely bound public to provide solutions to problems."

2.2.2 Open Data

Open data is a type of public or private dataset that can be freely accessed through the internet without any restrictions or extra cost. Typically, open data are provided by local governments or institutions.

The "Open" definition states eleven requirements that this kind of data should meet. The representative requirements basically demonstrate how to enable the free use, reuse, and redistribution of data. In addition, open data are referred to as data without any restriction to the specific use in one field or for an individual. Therefore, published open data 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). Open data indicate only the data available for the general public freely without any cost or limitations (Reiche and Höfig, 2013), and open data are becoming a critical enabler of open government (Kučera et al., 2013).

2.2.3 Big Data

Big data refer to large datasets that might need advanced data processing, data cleaning or integration with data from other datasets in order to provide useful information and generate support for decision making.

Currently, although it has been revealed that the importance of big data cannot be neglected, there are still different interpretations and definitions towards it. Generally, big data refer to the datasets that are hard to perceive, acquire, manage, and processe by regular software within a tolerable amount of time. Based on different research interests, various definitions for big data have been created by academic researchers, data analysts, technical enterprises, and scientific practitioners.

In 2010, Apache Hadoop defined big data as “datasets which could not be captured, managed, and processed by general computers within an acceptable scope.” (Chen et al., 2014). In 2011, an IDC report (Gantz and Reinsel, 2011) defined big data the following way: “technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis.” Based on this definition, four main characteristics can be concluded: great volume, large variety, rapid variation, and huge value.

2.2.4 Stated Preference Survey and Revealed Preference Survey

2.2.4.1 Stated Preference Survey

Stated preference survey (SP survey) is conducted to obtain the decision-making result of a respondent under certain conditions. The most important feature of the SP survey is that the content of the investigation is something that is intentionally made up and has not really taken place yet.

In the SP survey, the scenario is purposely designed in which the characteristics of choice are all assumed values. SP surveys have the advantage of being able to arbitrarily design the scenarios according to future conditions. This is particularly advantageous for the analysis of options that have not been established in the past in the area under analysis. In addition, since many people under the same conditions can be surveyed, it is possible to study differences in decision results due to differences in individual attributes.

2.2.4.2 Revealed Preference Survey

Revealed preference survey (RP survey) refers to the survey of completed selective behavior. The purpose of the RP survey is to obtain the results of the respondent’s choice under a certain selection condition. The most important feature of the RP survey is that the content of the investigation is what has already taken place.

In the RP survey, the result of choice is determined by the actual choice behavior and the choice condition. Regardless of the respondent being aware of the influencing factor or how they value these factors, the choice behavior is under the influence of all these factors, which means the phenomenon itself is hidden in the choice of results.

In the RP survey, the following issues often exist, which may result in inaccuracy (Guan, 2004).

(1) Fuzzy choice information: The respondents may have a fuzzy memory when they trace back to previous pieces of information.

(2) Alternative choice information ambiguity: Insufficient information about the conditions of alternative choices and varied volume of information are the major causes of ambiguity.

2.2.5 Traditional Survey Methods

Like the travel survey methods mentioned in Section 2.2.4 (stated preference survey and revealed preference survey), there are other traditional survey methods that have been used massively. One of the most common is the traditional household survey (Kagerbauer et al. 2015). Usually, participants are asked to answer questions that include information regarding their travel patterns. Most of the surveys are paper-based which can be time-consuming for both filling out the questionnaires and collecting useful information. Therefore, web-based travel surveys were introduced later with smart filter management features. However, there are still a lot of disadvantages associated with such travel surveys. For instance, young participants may outnumber older ones, since they can get access to the internet more easily. The receiving rate of the questionnaire can be low. 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.

2.3 Smartphone Crowdsourcing Applications

As stated in Section 2.2.1, there are numerous definitions of crowdsourcing. This section will concentrate on the smartphone crowdsourcing applications that are related to cycling, and several of them are listed below.

2.3.1 CycleTracks

The CycleTracks application was the first smartphone application designed for cycling data collection. It was developed by the San Francisco County Transportation Authority (SFCTA) in 2009 (SFCTA, 2013) and was designed to assess the cyclist demand on various bicycle facilities in San Francisco. CycleTracks uses the Global Positioning System (GPS) sensor in users' smartphones to record the cyclists' trip trajectories. It can also collect some user demographic information optionally. The demographic information is utilized to study the difference between each user based on individual attributes. The application is available to download freely for both iPhone and Android users. The user can enter his/her trip purpose (such as commute, school, work-related, exercise, social, shopping) into the comments field optionally during each trip. This trip purpose information can be utilized to fulfill the data gap of a cycling route or trip. Afterwards, users can view their trips with collected cycling information and trip trajectories on the application programming interface (API). In addition, users can fill in their demographic information (e.g., age, gender, email address, home, work, or school ZIP code, and cycling frequency) willingly when editing their personal profile.

2.3.2 AggieTrack

Since the code of CycleTracks is open source, some other applications have been developed based on it. Borrowed from CycleTracks, AggieTrack was developed by Texas A&M University to collect the travel patterns of the users in the university area (Hudson et al. 2012). The users are given the option to input their transportation mode and travel purpose after their trips. In addition, AggieTrack asks the users to choose their classification such as student, faculty or staff. Other information (e.g. about whether they live on campus or whether they own a private car) is also collected.

2.3.3 Cycle Atlanta

Cycle Atlanta was developed by a Georgia Tech research team together with the City of Atlanta and the Atlanta Regional Commission based on the code for the CycleTracks smartphone application (Misra et al. 2014). In addition to all the functions performed by CycleTracks, Cycle Atlanta added some additional features and developed a new user interface. It allows users to provide such information as pavement issues, traffic signal issues, bicycle parking and water fountain locations. They can also select the issues they want to report through a categorical list from the application. Like CycleTracks, Cycle Atlanta can also collect demographic information about its users such as ethnicity, household income and age.

2.3.4 RenoTracks

RenoTracks is built based on the Cycle Atlanta application which has all the functions added since CycleTracks. This smartphone application was created during the "Hack 4 Reno" coding convention in 2013, dedicated to develop an innovative data collection method for cycling information, especially from Reno cyclists (RenoTracks 2013). In addition to several significant new features added in Cycle Atlanta, RenoTracks further developed a customized user interface and added an additional "CO_2 Saved" counter, which calculates the carbon dioxide saved by using a bicycle instead of traveling by automobile.

2.3.5 Mon RésoVélo

Mon RésoVélo is developed by a joint cooperation between McGill University's Civil Engineering department and the City of Montréal based on CycleTracks and Cycle Atlanta to collect cycling information especially for the City of Montréal, Quebec. Unlike Cycle Atlanta and RenoTracks, Mon RésoVélo does not provide the information of "deterrent and amenity reporting"; however, several application functions are improved compared to the previous ones (Jackson et al. 2014). According to the application developers, Mon RésoVélo has optimized the underlying model for GPS data collection to break entire cycling trips to several segments to manage stopping and pausing more easily, reduce GPS connection loss, and solve the issues of forgetting to stop GPS collection when finishing a cycling trip (Jackson et al. 2014). Finally, Mon RésoVélo added a greenhouse gas emissions calculator considering local conditions (Jackson et al. 2014). Different from RenoTracks which only involves trip duration and average speed, a calorie counter was also included in Mon RésoVélo to correct for a cyclist's weight.

2.3.6 MapMyRide

MapMyRide is one of the applications developed by MapMyFitness that collects cycling information of the users worldwide. The series of applications in MapMyFitness are built for different modes of activities (e.g., bike, run, walk, hike, swim), and are customized for collecting specific activity information respectively. In addition, various external sensors (such as heart rate monitors and bicycle cadence detectors) can be linked to the applications to collect relevant fitness statistics. One of the features of this recreational application is the ability to view other users' activity routes to obtain fitness suggestions or ideas for different types of trips in terms of fun, challenging, and comfortable routes. Other than the application, there is a web application version of MapMyRide that can show the summary statistics and

the cycling trajectory of a user's trip. Users can utilize the web application to plan their trips ahead in order to select the preferred route before cycling.

2.3.7 Strava

Strava Metro has developed both a smartphone application and a web application for iPhone and Android users to record cycling information especially for recreational trips. Similar to MayMyRide, users may track their cycling trajectories using GPS-enabled smartphones and view the routes and corresponding information afterward via the smartphone or web application. Summary statistics including average speed, trip distance, and travel time are presented, as well as the overview of the cycling route. The unique feature of Strava application is that it provides the performance and comparison of multiple users on the same road segment, which enables cyclists to compete with the maximum speed, shortest segment time, and other top statistics. This functionality makes Strava popular among cyclists from worldwide and outperforms other recreational applications (such as MapMyRide and MyTracks) in terms of more social ability.

2.3.8 MyTracks

MyTracks is a separate application designed by Google in 2011 to collect cycling information. The advantage of this smartphone application is the massive mapping software capabilities. However, the shortage of MyTracks is the limitation to Android platform only. Similar to MapMyFitness and some other crowdsourcing applications since CycleTracks, MyTracks is able to track GPS points of various modes including cycling, running, hiking, and driving, etc. Like Strava, MyTracks also provides summary statistics such as average and maximum speed, trip distance, and elevation climbed at the end of a specific trip. With the convenience of Google's suite of cloud-based software, summary statistics and trip information can be exported to and presented in Google Earth or investigated in other available software.

2.3.9 ORcycle

ORcycle was developed by researchers at Portland State University and Oregon DOT to collect cycling data from the users. It was launched in November 2014 for both Android and iOS platforms. The users' trip trajectories can be collected utilizing this application. Based on the Python scripts developed by Broach et al. (2012) for cycling activity research study using GPS data, the original trip trajectories can be matched to the street network in Portland metropolitan area. Other information regarding the algorithms of Python scripts can be found in the research studies conducted by Schuessler and Axhausen (2009a, 2009b).

Since CycleTracks, Cycle Atlanta and Strava are the most commonly used applications, TABLE 2.2 provides a summary and gives a comparison of these three applications.

Table 2.2 Summary of Smartphone Crowdsourcing Applications

Data Category	**Information Collected**	**CycleTracks**	**Cycle Atlanta**	**Strava**
Year App was developed		2009	2012	2009
GPS trajectory	Trip record	√	√	√

Demographics	Age	√	√	√
	E-mail address	√	√	
	Gender	√	√	√
	Household income		√	
	Rider type		√	
	Rider history		√	
	Cycling frequency	√	√	
	Winter cyclist status			
	Home ZIP	√	√	
	Work ZIP	√	√	
	School ZIP	√	√	
Trip related	Travel purpose	√	√	√
Infrastructure reporting	Water fountain		√	
	Secret passage		√	
	Public restroom		√	
	Bike shops		√	
	Bike parking presence		√	
	Other cycling asset		√	
Improvement request	Pavement issue		√	
	Traffic signal complaint		√	
	Enforcement issue		√	
	Bike parking request		√	
	Bicycle lane design issue		√	
	Other improvement request		√	
Useful information for users	Bicycle route map			√
	Trip routes	√	√	√
	Average speed	√	√	√
	Trip distance	√	√	√
	Trip related information of other users on the same segment			√
	User count on each segment			√
	User count on both direction			√
	User count during each time			√

	period			

2.4 Potential Use of Crowdsourced Data

Many researchers have conducted their studies by using crowdsourced data. The existing use of crowdsourced data for different research purposes is presented as follows.

2.4.1 Crowdsourced data for route choice analysis

Dill and Gliebe (2008) conducted a research study to investigate the impact of various infrastructure types (such as bicycle lanes or paths) on cycling. Data were collected using global positioning system (GPS) technology from March to November 2007 in the Portland, OR region. A sample of 164 adult cyclists' cycling trajectories were recorded to address the four major sets of research problems, including the cycling time, location, and purpose; the difference between cycling route and the shortest network distance; the cyclist route choice behavior and the impact of rider characteristics; and the difference between cycling and driving travel time. It can be concluded that most of the cycling trips occurred during daylight hours and half of them occurred during peak hours. Cyclists prefer using bike facilities and streets with less vehicle traffic to the shortest paths.

Charlton et al. (2011) introduced the smartphone application (i.e., CycleTracks) to show a method for recording bicyclists' trips using CycleTracks. Data were collected via this application which included bicyclist demographic information and trip characteristics. Potential data bias was discussed before the model development. The risk of utilizing this biased data was taken, and the route choice model was developed based on the bicyclist route choice data. Results showed that bicyclists are sensitive to slope and the presence of bike facilities. In addition, gender and trip purpose are two primary contributing factors to cycling. Different gender and trip purpose will have different impacts on cycling preference. According to the results, compared to male cyclists, female dislike up-slope more. And for commuting trips, cyclists dislike up-slope more than other trip purposes.

Hood et al. (2011) collected data from the CycleTracks to develop a bicyclist route choice model based on the first one that was built from a large GPS dataset in Zurich. This research applied an innovative choice set generation method which is a "doubly stochastic" method that was developed by Bovy and Fiorenzo Catalano (2007) instead of using the previous stochastic path generation and labelling method. The Path Size measure was adopted from Ben-Akiva and Bierlaire (1999) and utilized to overcome the Independence of Irrelevant Alternatives (IIA) property of the Multinomial Logit (MNL) models. Results indicated that bike lanes have a positive impact on cycling in San Francisco, while slope or turning has a negative impact.

Broach et al. (2012) used 1449 trips from the GPS data that were collected in Portland, Oregon to develop a model to estimate the route choice for the better analysis of cyclists' preference for facility types. After comparing the three choice set generation methods including the K-shortest paths, simulated shortest paths, and route labelling, this study switched to develop a modified method of route labeling that was more suitable in this case. Similar to Hood (2011)'s research, Broach et al. (2012) developed a Path-Size Logit (PSL) model to analyze the impact of route and facility characteristics (e.g., specific bike facility

types and bridge facility) on cyclists' route choice and identified the behavioral differences between commuters and non-commuters. Cyclists' sensitivity to distance, turn frequency, slope, intersection control, and volumes were revealed in the final results. It was found that commuters are more sensitive to distance than non-commuters.

Casello and Usyukoy (2014) collected the GPS data of 724 cycling trips to estimate the bicyclists' evaluation of the path alternatives. Path attributes including length, grade, auto speed, and bike lanes were compiled for model estimation. Two Multinomial Logit models were developed to analyze the bicyclists' route choice. 181 trips that were not selected in the model calibration process were utilized to test the predictive powers of the models. Under the best situation, the model can predict 65% of the actual paths that were taken in all trips accurately.

Moore (2015) collected data from the Strava smartphone application to investigate the contributing factors to bicyclists' route choice in the City of Auburn, Alabama. Based on the crowdsourced bicycle data, an ordinal logistic regression model was built to analyze the impact of various attributes on bicyclists' route choice decisions. To present the spatial information on cycling routes and investigate the influence of cycling locations and facilities, a qualitative analysis was conducted using GIS. Based on the analysis, it can be concluded that roadway characteristics and specific land use have a significant impact on the preference for a particular road segment.

Yeboah et al. (2015) conducted a route choice analysis based on the transportation network information that was gathered from OpenStreetMap (OSM). Seven days of 79 bicyclists' GPS tracks and travel diary data in North East England were collected for the route choice analysis. Specific factors (including distance, time, and road types etc.) were examined to discover the influence on bicyclists' route choice. Following the research conducted by Papinski and Scott (2011), a four-step method was adopted for the cycling route generation. The research findings suggested that the OSM transportation network information can be used for cycling research, especially when being compiled with GPS trajectory data. In addition, the observed routes are usually not the shortest paths and the route directness is one of the significant factors affecting route choice in the bicycle network.

Bergman and Oksanen (2016) utilized OpenStreetMap (OSM) and mobile application data from Sports Tracker to provide automatic route choices for cyclists. After pre-processing the Sports Tracking data and generating the route choice set extracted from OSM, the map-matching was conducted with an advanced Hidden Markov Model (HMM)-based algorithm. A comparison was made between this method and a simple geometric point-to-curve method. Results revealed that the HMM-based algorithm had better matching results in terms of the number of the correctly matched road segments.

Grond (2016) analyzed the impact of the road physical and environmental network on bicyclists' route choice. 5713 trips of 554 bicyclists from August 23 to September 23, 2015 were recorded in a high-quality GPS dataset from the City of Toronto's cycling application and utilized in this study. Two choice sets were generated separately using a simple labelling approach and a modification of the labelling technique which was proposed by Broach et al. (2010). A path-size multinomial logit model was developed to examine the impact of cycling

facilities and roadway characteristics. In addition, bicyclists' demographic characteristics and attitudes were also taken into consideration to investigate the different route choice behavior.

Khatri et al. (2016) collected the GPS data from Grid Bikeshare in Phoenix, Arizona which contain 9101 trips involving 1866 bicyclists to analyze the bicyclists' utilitarian route choice behavior with a Path Size Logit Model. The bikeshare users were categorized into two groups which are registered users and casual users. It is revealed that registered users prefer shorter trips with low-volume road segments and the presence of bicycle facilities. From the modelling results, it was found that one-way road segments, Annual average daily traffic (AADT), and length of trip all have a negative impact on bicyclists' route choice, while the number of signalized intersections affects the route choice positively. This result might be different from what are expected, since signalized intersections may increase the total travel time and therefore could have a negative impact on cycling safety. However, on the other hand, compared to unsignalized intersections, signalized intersections will protect cyclists for crossing large roadways. In addition, a reasonable signal timing plan may not add too much delays to the total travel time.

LaMondia and Watkins (2017) used the crowdsourced bicycle data collected from three smartphone applications including Strava, Cycle Dixie and Cycle Atlanta to examine the contributing factors to bicyclists' route choice. Factors including bicycle facilities were investigated to examine the prioritization preferences of bicyclists. Bicyclists were categorized into different rider types, and several logistic regression models were developed based on these rider types. The modelling results showed that demographic characteristics, roadway features and the nearby land use are the primary factors affecting bicyclists' route choice behavior.

Zimmermann et al. (2017) collected GPS data from a network that contains over 40,000 links in the City of Eugene, Oregon to analyze the bicyclist route choice. A link-based bicycle route choice model called recursive logit (RL) model was developed following Fosgerau et al. (2013)'s work. This approach showed its advantage in that there is no requirement for choice set generation compared to the path-based route choice model which requires one to generate choice sets. An accessibility measure was derived from the bike route choice model which was examined previously by Nassir et al. (2014). Results revealed that bicyclists are sensitive to trip distance, slope, traffic volume, number of crossings and cycling facilities.

To conclude, a summary of the route choice analysis studies utilizing crowdsourced data is provided below in TABLE 2.3.

Table 2.3 Summary of Route Choice Analysis Utilizing Crowdsourced Data

Year	Author	Data	Methods	Results
2008	Dill and Gliebe	GPS data of Portland, OR	USGS Digital Elevation Model	The majority of the bicycle travels were for utilitarian purposes. About half of the trips occurred during morning and evening peak travel times. Distance and traffic volume have a negative impact on route choice.

2011	Charlton et al.	Data from CycleTracks application	Bicycle route choice model	Cyclists are sensitive to slope, presence of bike lanes or bike route designations. The route choice behavior is also influenced by trip purpose and gender.
2011	Hood et al.	Data from CycleTracks application	Path Size Multinomial Logit model	Bike lanes are preferred compared to other types of bicycle facilities, while steep slopes are disfavored. Length and turns have negative impact on route choice. Surprisingly, traffic volume, speed, number of lanes, crime rates and nightfall have no impact on route choice.
2012	Broach et al.	The GPS data collected in Portland, Oregon	Path-Size Logit (PSL) model	Cyclists are sensitive to distance, turn frequency, slope, intersection control and volumes. For commuters, they are more sensitive to distance than non-commuters.
2014	Casello and Usyukov	724 cycling trip GPS data	Multimodal logit models	Cyclists consider both vehicle speeds and the presence or absence of a bike lane during route choice process.
2015	Moore	Data from Strava application	Ordinal logistic regression model	Roadway characteristics and surrounding land-use have a significant impact on whether or not a particular street segment would be used.
2015	Yeboah et al.	OpenStreetMap, GPS tracks (7 days) and travel diary data	Four-step method for generating routes	Network restrictions for both observed and shortest paths are significant.
2016	Bergman and Oksanen	OpenStreetMap and mobile application data from Sports Tracker	advanced HMM-based algorithm	HMM-based algorithm has better matching results in terms of the number of the correctly matched road segments.
2016	Grond	GPS dataset from the City of Toronto's cycling app	path-size multinomial logit model	Steep hills, high traffic volumes, left turns without signalized intersections and right turns at signalized intersections have negative impact on route choice.
2016	Khatri et al.	GPS data from Grid Bikeshare in Phoenix, Arizona	Path Size Logit Model	The proportion of one way segments, AADT and length of trip have a negative influence on

				route choice and number of signalized intersections has a positive influence on selecting routes.
2017	LaMondia and Watkins	Data collected using the Strava, Cycle Dixie and Cycle Atlanta	Route suitability score and preference models	Demographics, roadway characteristics and surrounding land-use have a significant impact on route choice.
2017	Zimmermann et al.	GPS observations in the city of Eugene	Link-based bike route choice model (recursive logit model)	Cyclists are sensitive to distance, traffic volume, slope, crossings and the presence of bike facilities.

2.4.2 Crowdsourced data for bicycle volume estimation

Griffin and Jiao (2016) selected five specific monitoring locations with recorded bicycle counts in downtown Austin, Texas. The data collected by CycleTracks smartphone applications, and crowdsourced data derived from the Strava fitness application, together with the traffic counts, were compiled and compared in GIS at these five locations.

Jestico et al. (2016) utilized the Geographic Information systems (GIS) and developed a generalized linear model to identify the relationship between crowdsourced data from Strava fitness application and manual counting data in Victoria, British Columbia, Canada. A regression model was developed to predict categories of bicycle volumes and to create maps. The result showed that 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'.

Hochmair et al. (2017) used Strava activity tracking data collected in the Miami-Dade County area to identify which sociodemographic factors, network measures (in particular on-road bicycle facilities), and place specific characteristics influence bicycle ridership. For this purpose, a series of linear regression models were developed to predict bicycle kilometers traveled for different trip puposes (e.g., commuring and non-commuting), as well as bicycle kilometers traveled on weekdays and weekends. Spatial autocorrelation was modeled using eigenvector spatial filtering, and parameter estimation bias can be avoided. Based on the model results, it is confirmed that Strava data with high spatial and temporal resolution and coverage showed its advantages in examining the impact of contributing factors on estimated bicycle volume for different trip purposes and day of week. Therefore, Strava data can be considered as a sufficient supplement for bicycle volume estimation especially in large study areas.

Proulx and Pozdnukhov (2017) developed a novel geographically weighted data fusion-based method to estimate bicycle traffic volumes in a network by fusing crowdsourced data from Strava and usage data from Bay Area Bikeshare.

2.4.3 Crowdsourced data for other research purposes

Engineers and planners in Foresite Group (2015) utilized Strava's data to analyze the representativeness of the crowdsourced data, and the correlation between the Bicycle Level-of-Service (BLOS) grades evaluated by using traditional methods and the crowdsourced data. Results showed that Strava's data may not represent the general cycling population and that most of its users tend to be recreational cyclists.

Griffin et al. (2015) examined the bicycle ridership especially for fitness purposes based on data collected from Travis County, Texas. Bicycle volume was estimated considering the influence of different types of land use, residential and employment density, bike facilities, and terrain which determines the cycling places chosen by bicyclists. Although there is a limitation of using the smartphone application data (i.e., the partial collection of bicycle counts), it is promised to use this kind of data for non-motorized transportation planning and health impact assessment.

Watkins et al. (2016) compared Cycle Atlanta with Strava in terms of user demographic data, the time-of-day of cyclist trips and the number of trips by road segments. In addition to comparing these two GPS data, they also compared the data from Cycle Atlanta with actual cyclist trips in both AM peak and PM peak hours. 78 intersections were selected to calculate the actual percentage of manual counts recorded in Cycle Atlanta. In some cases, there were noticeable differences in the populations of cyclists recording trips in both applications. Such GPS data should be carefully used before conducting further analysis and modeling.

Raihan et al. (2017) conducted a study to investigate the impact of roadway characteristics on bicycle safety. This study was intended to assist Florida Department of Transportation (FDOT) to quantify the influence of bicycle facilities and roadway characteristics on the frequencies of bicyclist-involved crash. This research mainly focused on urban facilities where 98% of bicycle crashes occurred. Florida-specific Crash Modification Factors (CMFs) for bicycle crash were developed using a robust cross-sectional analysis. These CMFs enable researchers to investigate the impact of roadway characteristics on bicycle safety analysis. In addition, this is the first time to consider bicycle exposure using Strava smartphone application data.

Sun and Mobasheri (2017) conducted a research study on cycling activities and air pollution exposure based on the crowdsourced bicycle data collected from Strava application. Cycling behavior was analyzed by estimating the non-commuting bicycle volume considering the influence of environmental characteristics. Strava node data in Glasgow, United Kingdom was utilized for a case study to bridge the data gap regarding the lack of trip purposes. Results were found that non-commuting trips are more likely to occur in outskirts of the city compared to commuting trips. In addition, bicyclists bike for non-commuting trips are more likely to be exposed to low levels of air pollution compared to commuting trips.

2.5 Summary

This chapter provides a comprehensive review of the previous research on crowdsourcing data and its application for numerous research such as route choice analysis and bicycle volume

estimation. It is intended to give a better understanding of crowdsourcing, and existing research efforts utilizing crowdsourced data which will provide a useful reference for future studies.

Chapter 3. Collecting Crowdsourced Data and Other Supporting Data

3.1 Introduction

As mentioned before, the first step of this research is to collect crowdsourced bicycle data from Strava application and other relevant supporting data. Chapter 3 gives an overview of the collected Strava bicycle data, and other essential supporting data for the later model development.

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 provides the data view to Strava data in the City of Charlotte. Section 3.5 shows the other relevant supporting data collected for this research. Finally, Section 3.6 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 routing data used in this project are collected from Strava smartphone application developed by a technology company recording the cyclist travel trajectory with the GPS located in their smartphones. A screenshot of the application interface is presented in Figure 3.1, which shows some of the information about distance, time, and speed, etc. that Strava offers to the user after his or her trip. The application is available for use by any person who has a GPS enabled smartphone that can have access to the internet. Most of the users are cyclists or runners. When a cyclist or runner uses the Strava application, his/her trip information such as distance, duration, elevation change, and average speed is collected. In addition, the GPS route information will also be recorded 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. The accuracy of the GPS data depends on the connection to the GPS satellites, which will certainly get better when the number of the available GPS satellites increases.

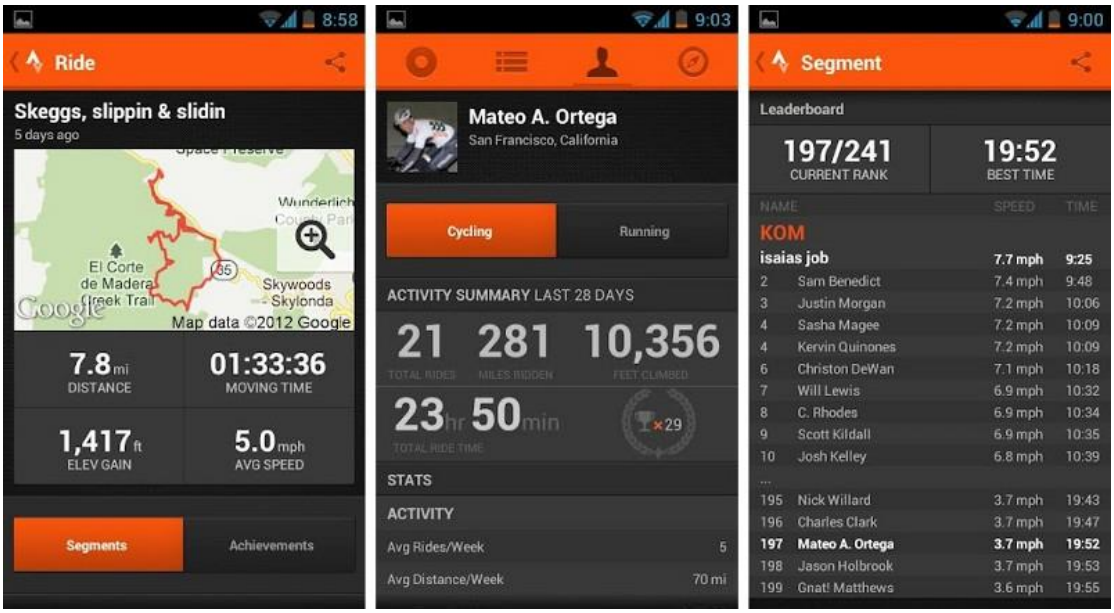


Figure 3.1: Strava App Screen Shots

3.3 Strava Metro Delivery

The GPS data from Strava smartphone application usually includes a data set for the nodes of the network and a data set for the links. The data set for the links includes bicycle volumes on each link and the cyclists' average speeds, and the nodes' dataset contains both the information in the link dataset and also the waiting times. In addition, origin-destination-matrices can be calculated based on the data. The data offered by Strava usually contain the following three components.

3.3.1 Core Data

1. Street-level database file with minute-to-minute cycling information on each road segment for the time period of the delivery.

2. Intersection database file with minute-to-minute cycling information on nodes (intersections) for the time period of the delivery.

3. Origin/Destination file with OD pairs for all trips during the time period of the delivery.

3.3.2 Roll-ups

Street/Intersections roll-ups: each core dataset will have a roll-up data file that summarizes views that show monthly use, weekend/weekday, seasonality, hour groupings, and total counts. Also, based on the original data, users can generate their own roll-ups based on their specific demand.

For the cycling data in Charlotte, the seasonality and hour groupings are categorized 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: report of Strava users included in each data delivery in terms of age and gender.

2. Tourism: breakdown of the home location of all Strava users at county level over the time frame if requested.

3.4 Data View

Metro DataView is an interactive visualization that provides information about total activities, commutes and number of cyclists aggregated to the street segment, intersection, and origin-destination polygonal geometry. In addition, GPS point heat can be seen at the street level and intersection behavior times will be shown when the intersection view is selected. The default data view presents the total activities in the whole network. One can switch between views to see the activities occurred on streets or intersections using the “intersection button” found in the upper left corner of the map interface. The four default data views are introduced as follows.

3.4.1 Street

Streets in the Metro DataView are symbolized using different colors, with dark blue representing lowest counts and dark red indicating highest counts. The street legend showing the number of rides that corresponds to the specific color is presented on the bottom left of the interface. To see the percent distribution of number of streets in each data class, one can hover on the street legend. Also, to see the exact counts of cyclists on a specific street segment, one can hover on that street. The Street Data View is shown in Figure 3.2.

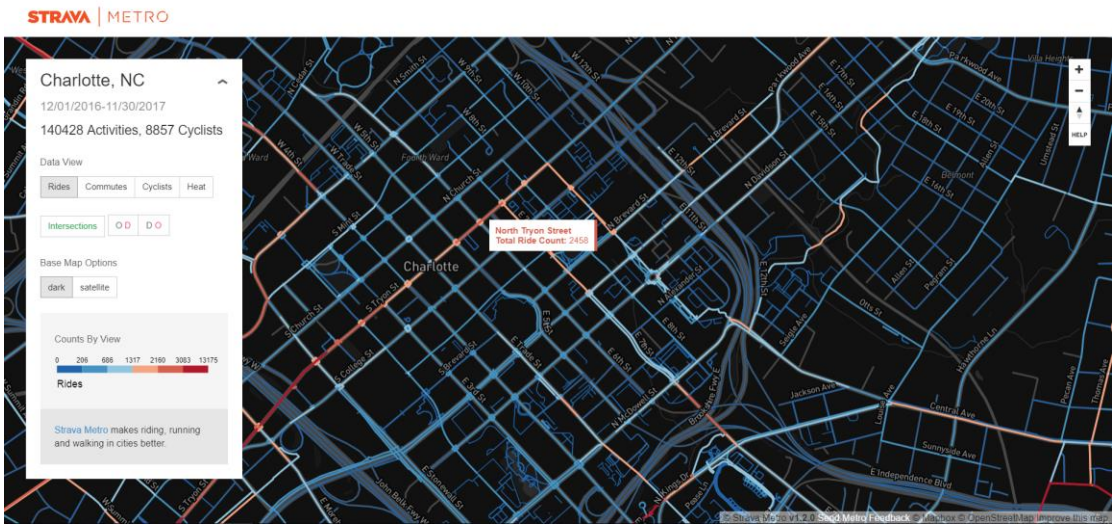


Figure 3.2: Charlotte Metro Data View 2017 Sample: Total activity counts from December 01, 2016 to November 30, 2017

3.4.2 Intersections

To view the intersection nodes in the whole network, users can click on the “intersection button” on the left of the map interface to switch the nodes on or off.

The default set will not show the intersection information with the “intersection button” being off. This will avoid information overload when such data are not of interest to users.

Different data information can be shown when different buttons on the map interface are selected. Clicking on the “rides button” will show the counts of activities at the intersection. Clicking on the “commutes button” will display the counts of the commutes at the intersection. Clicking on the “cyclists button” will show the number of cyclists at the intersection.

When looking at the visualization view, larger nodes depict higher rides/commutes/cyclist counts, and brighter nodes represent longer intersection crossing times. For example, a small dark node symbolizes an intersection with few rides crossing it and a short median intersection crossing time. To view the exact data associated with the specific intersection, users can hover on that node. The Intersections Data View can be seen in Figure 3.3.

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

3.4.3 Origin & Destination

To view the origin and destination of a trip generated by a cyclist, users can click on the "Origin Destination" button to switch on the origin/destination polygons of which the layer is based on a contiguous 350-meter hexagonal bin.

Same as the "intersection" button, the default will not show the OD information with the button being selected to be off, which avoids information overload.

Similar to the intersection view, there are also various data information that can be provided when different toggle buttons are selected. Clicking on the "Rides" button, users can obtain the counts of activities started within the polygon. Clicking on the "commutes button", users can get the counts of the commutes started within the polygon. Clicking on the "cyclists button", users can acquire the number of cyclists started within the polygon.

Also, when looking at the visualization view, darker polygons depict fewer rides/commutes/cyclists and lighter polygons represent more rides/commutes/cyclists. To see the exact data associated with the trip origin, users can hover on the polygon that contains the origin. To see all the destination polygons associated with the origin, users can click on the origin polygon. The destination polygons will be shown in pink, with darker ones indicating areas with more rides, and lighter ones representing less rides. The Origin Destination Data View is shown in Figure 3.4.

Figure 3.4: Charlotte Origin Destination Metro Data View 2017 Sample

3.4.4 Heat Map

The “Heat Map” view will provide the users a visualized view of GPS points aggregated to streets and intersections. The brighter lines indicate streets with higher number of activities, while darker lines show streets with lower number of activities. The Heat Map View is shown in Figure 3.5.

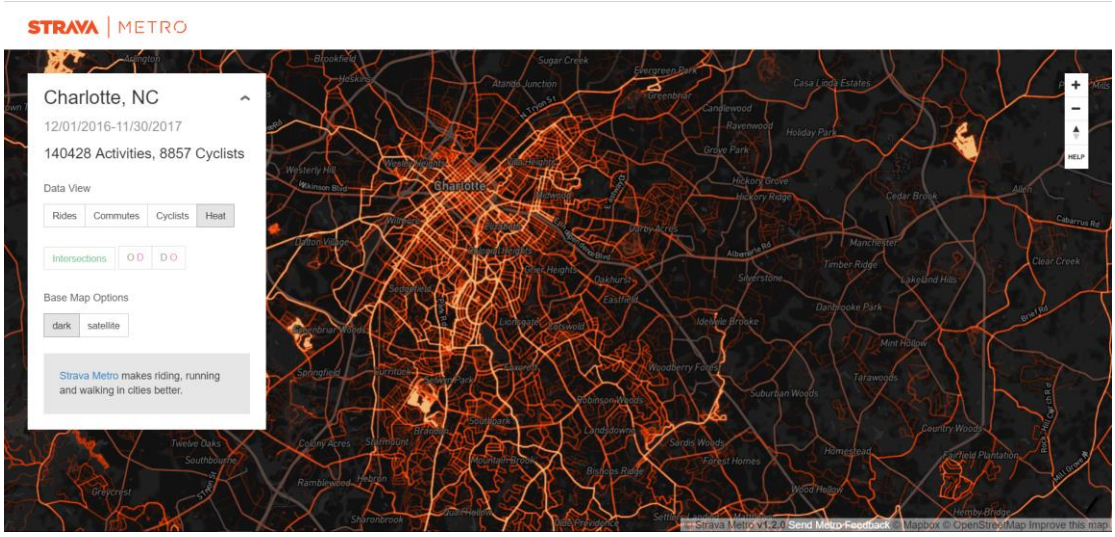


Figure 3.5: Charlotte Heat Map View

3.5 Other supporting data

3.5.1 Manual Count Data

Bicycle volume data from a manual count station is collected for this research. The locations of the count stations are shown in the following figure. From the figure, one can see that

most of the count stations are located in the center city. Note that the data can be found on this website: https://itre.ncsu.edu/focus/bike-ped/nc-nmvdp/.

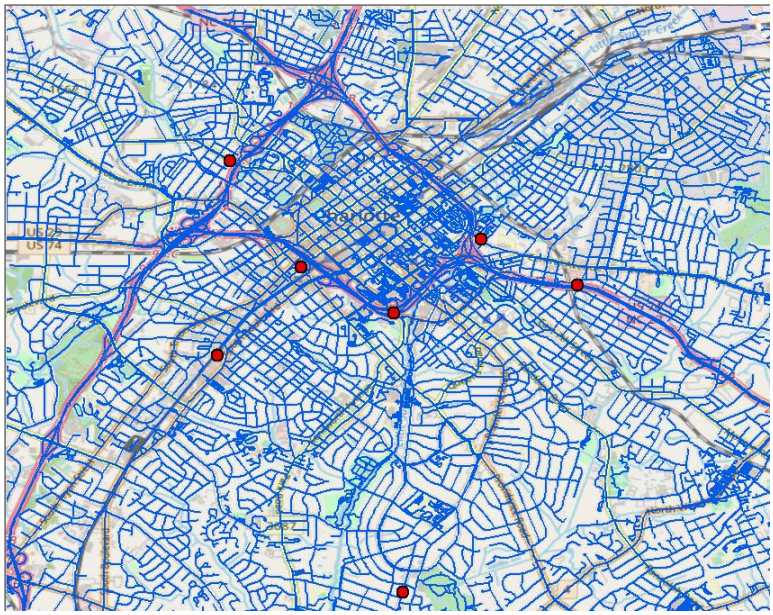


Figure 3.6: Manual Count Station Locations

3.5.2 Bicycle facilities

The bicycle facilities in the City of Charlotte are shown in the following figure. According to the data, the bicycle network in the City of Charlotte contains more than 90 miles of bicycle lanes, 55 miles of signed routes, and 40 miles of greenways and off-street paths. Note that the bicycle facility map can be found on this website: http://charlotte.maps.arcgis.com/apps/PanelsLegend/index.html?appid=00e8015ea3e54607a880fe31cc7e2fbf.

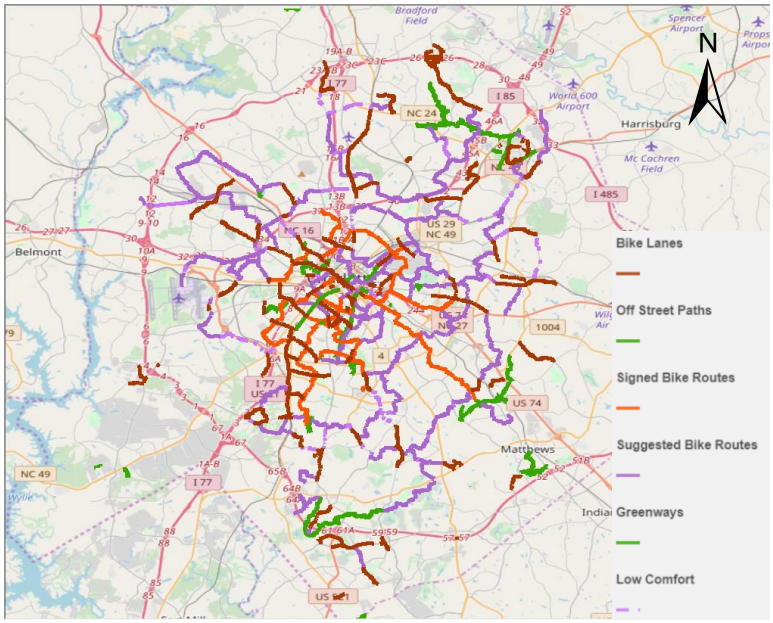


Figure 3.7: Bike Facilities in the City of Charlotte

3.5.3 Population

The population data are collected from the US census dataset which is presented in the following figure. Note that this data can be found on this website: http://www.arcgis.com/home/webmap/viewer.html?url=https://services1.arcgis.com/yfahUFAYAdeS5rmM/ArcGIS/rest/services/Enriched%20Enriched%20Charlotte%20Blocks/FeatureServer&source=sd.

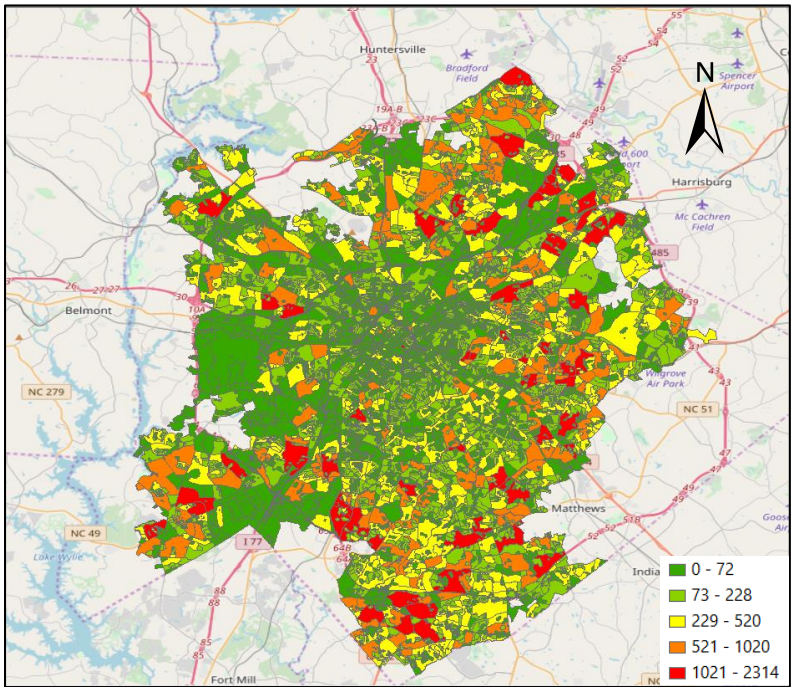


Figure 3.8: Total Population in the City of Charlotte

3.5.4 Slope

The slope cell data are collected from the ArcGIS online dataset which is shown in the following figure. Note that this data can be found via ArcGIS by selecting "Lidar2017_Slope" in "add data from ArcGIS online".

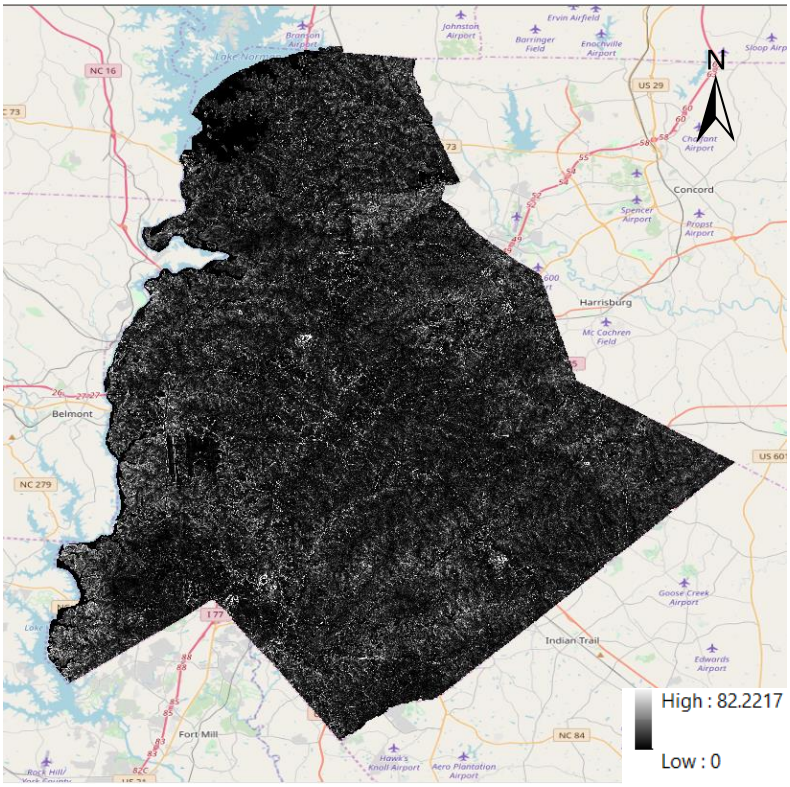


Figure 3.9: Slope in City of Charlotte

3.6 Summary

The objective of this chapter is to present the data collected in this research including Strava bicycle data, bicycle manual count data from count stations in the City of Charlotte, demographic data, road characteristic data, slope data, and bicycle facility data that are utilized for the model development in the following chapters.

Chapter 4. Data Descriptive Analyses

4.1 Introduction

This chapter provides the descriptive analyses of the collected data. Data comparison is conducted between bicycle manual count data from the count stations in the City of Charlotte and the Strava bicycle data collected from the smartphone application.

The following sections are organized as follows. Section 4.2 describes the Strava bicycle count data in terms of the heatmap based on different month of year, trip purpose, weekday and weekend. Section 4.3 gives a data comparison between bicycle manual count data from the count stations in the City of Charlotte and the Strava bicycle data from the smartphone application. Finally, Section 4.4 concludes this chapter with a summary.

4.2 Strava Data Analysis

4.2.1 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 are 140,428 miles. The proportion of Strava users' gender is presented in Figure 4.1.

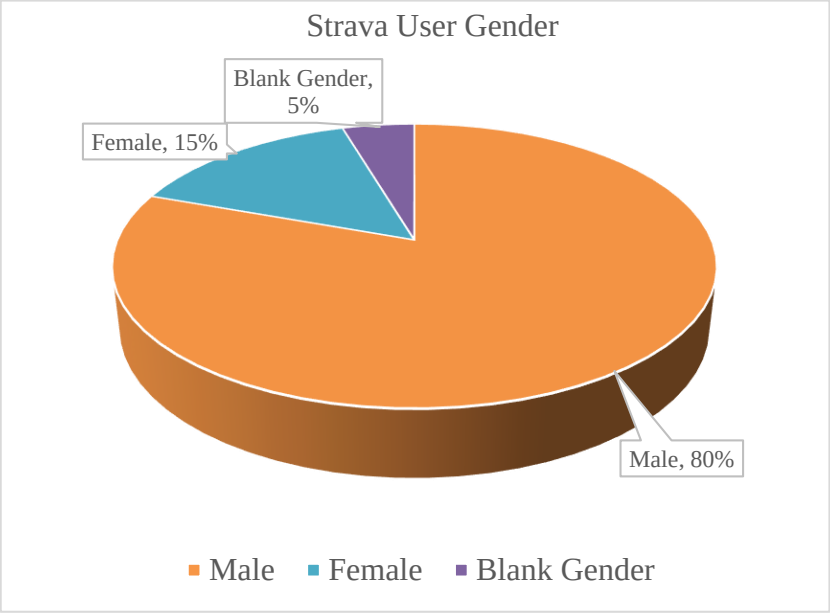


Figure 4.1: Strava User Gender

From the age data, one can see that cyclists of all ages are 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. The distribution of the cyclists from different age groups for both male and female cyclists is presented in Figure 4.2. From the figure, one can see that most of the cyclists are between 25 and 54.

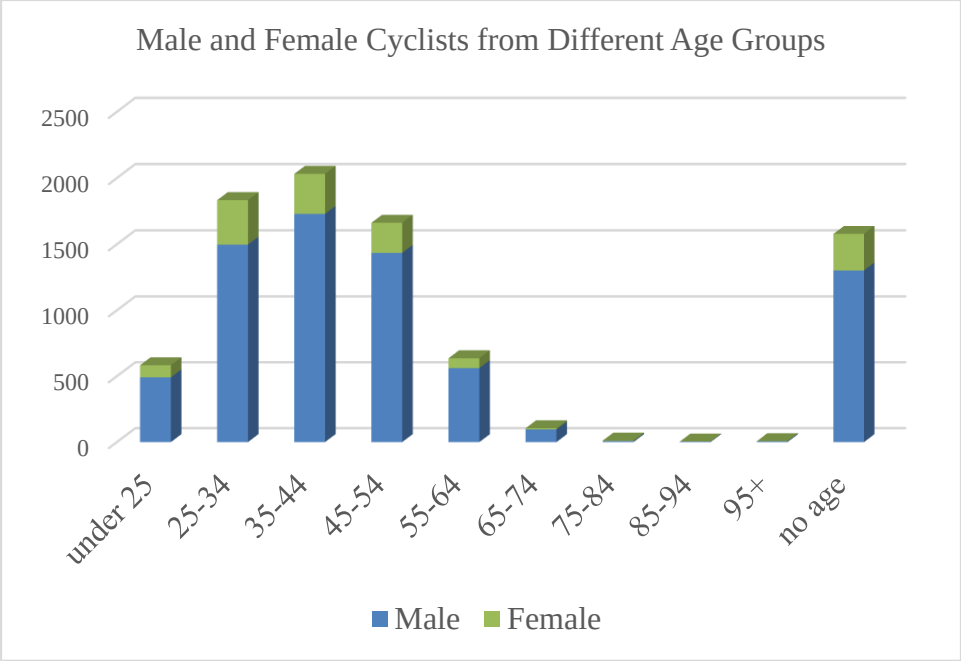


Figure 4.2: Male and Female Cyclists from Different Age Groups

4.2.2 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.

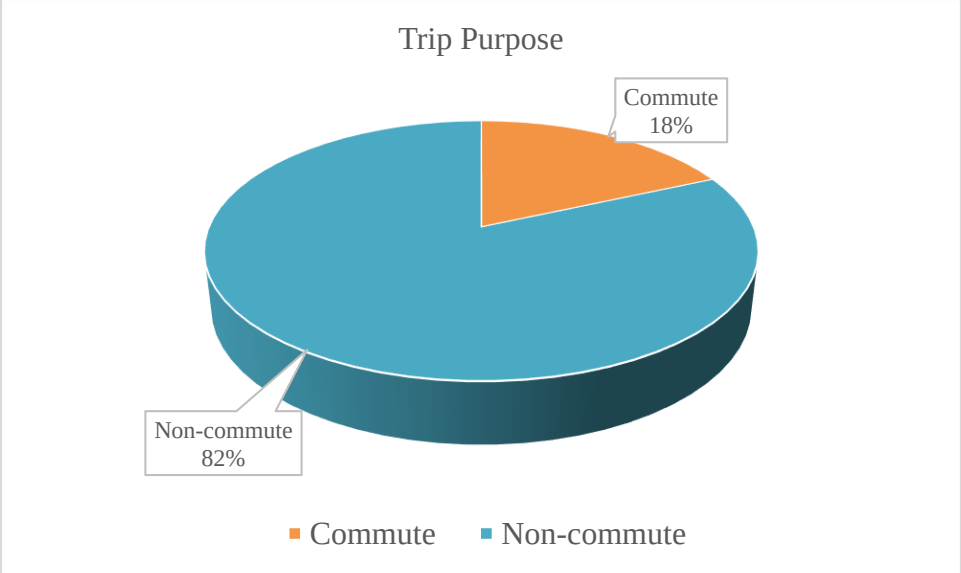


Figure 4.3: Cyclist Counts for Different Trip Purposes

4.2.3 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 4.4.

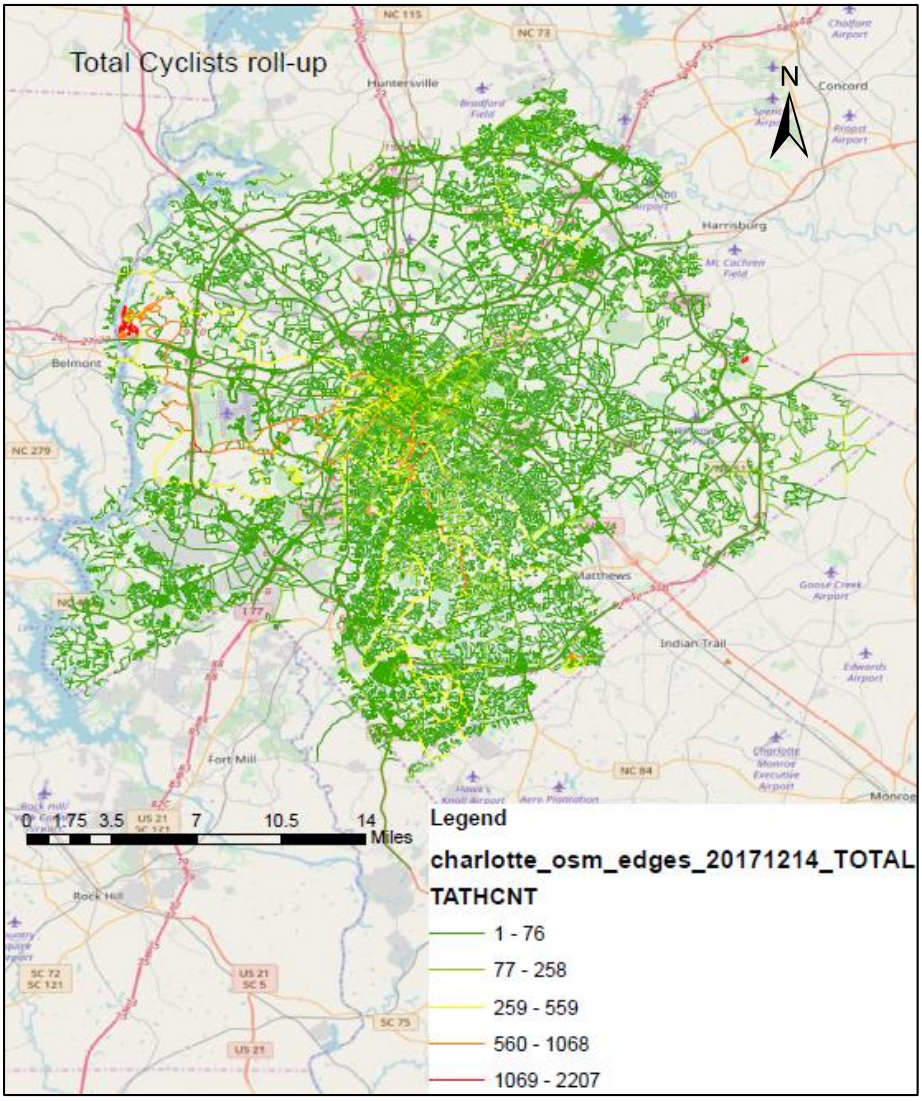
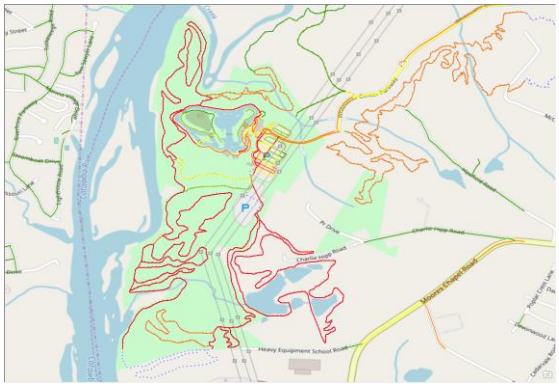


Figure 4.4: Total Cyclists Roll-ups

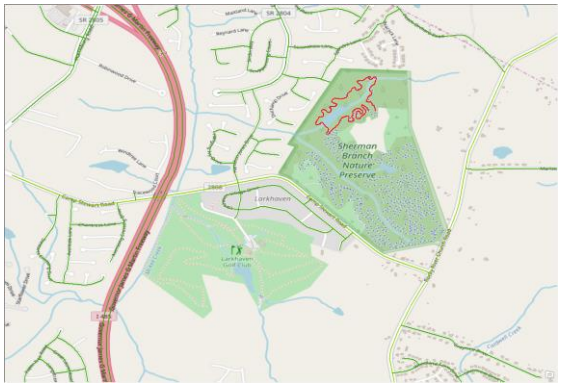
Based on the bicycle volume in Figure 4.4, four locations where the volume of Strava users are high are identified which involve greenway, school, airport, and park. These are the popular cycling locations among Strava users.



4.5.a Greenway

4.5.b School

4.5.c Airport



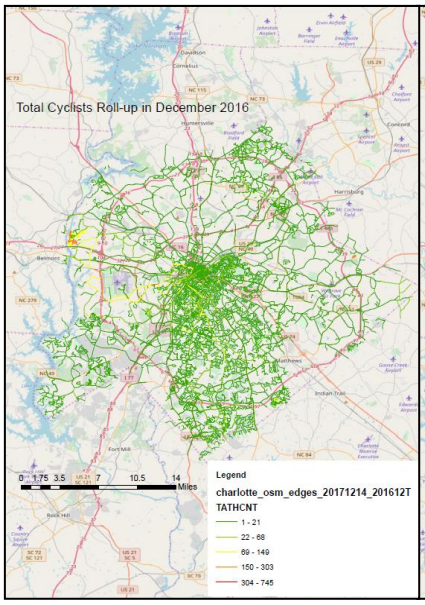
4.5.d Park

Figure 4.5: Four Popular Cycling Locations

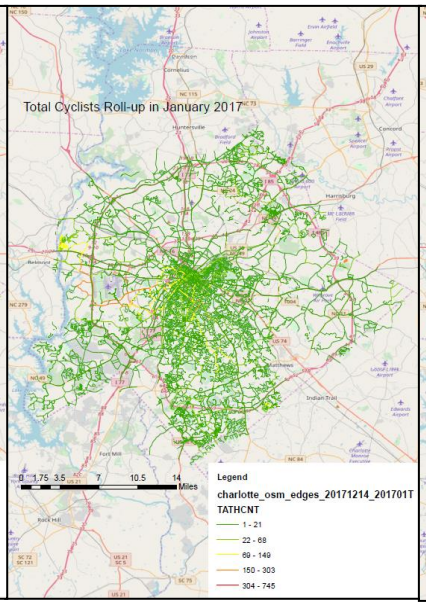
The Strava bicycle counts under different situations are presented in detail as follows.

4.2.3.1 Month of Year

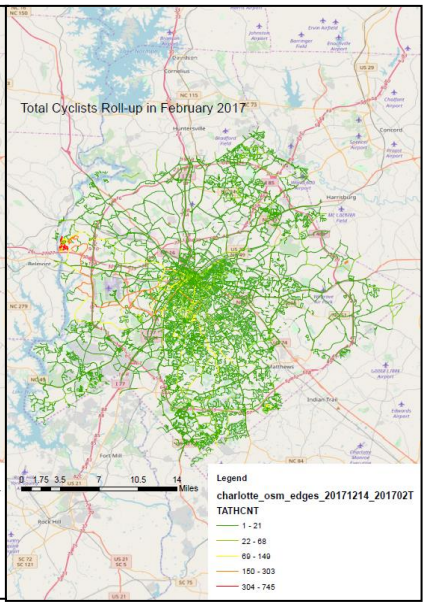
Cycling is a kind of activity which is highly related to the weather condition. Therefore, bicycle counts in different month of year vary with the temperature. Total bicycle volume on each road segment in twelve months of a year is presented in the following figures.



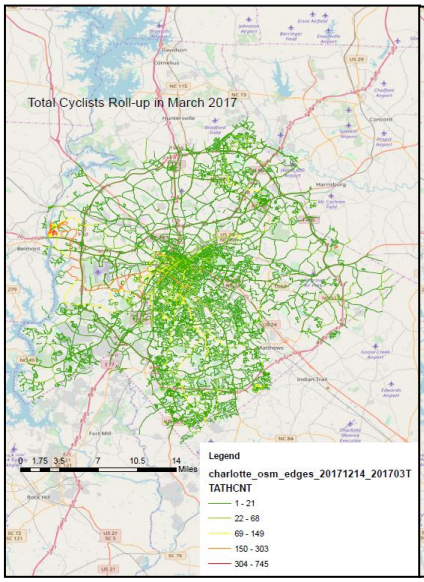
4.6.a December 2016



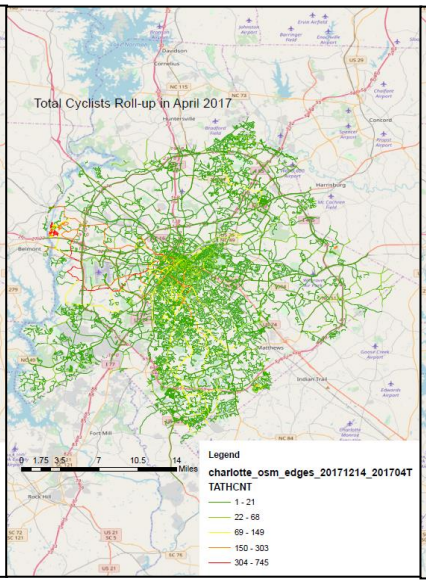
4.6.b January 2017



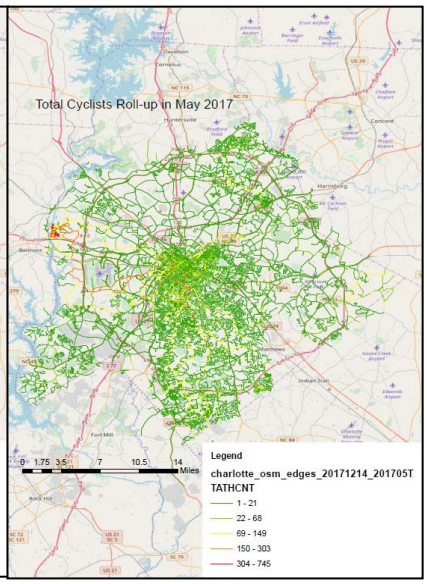
4.6.c February 2017



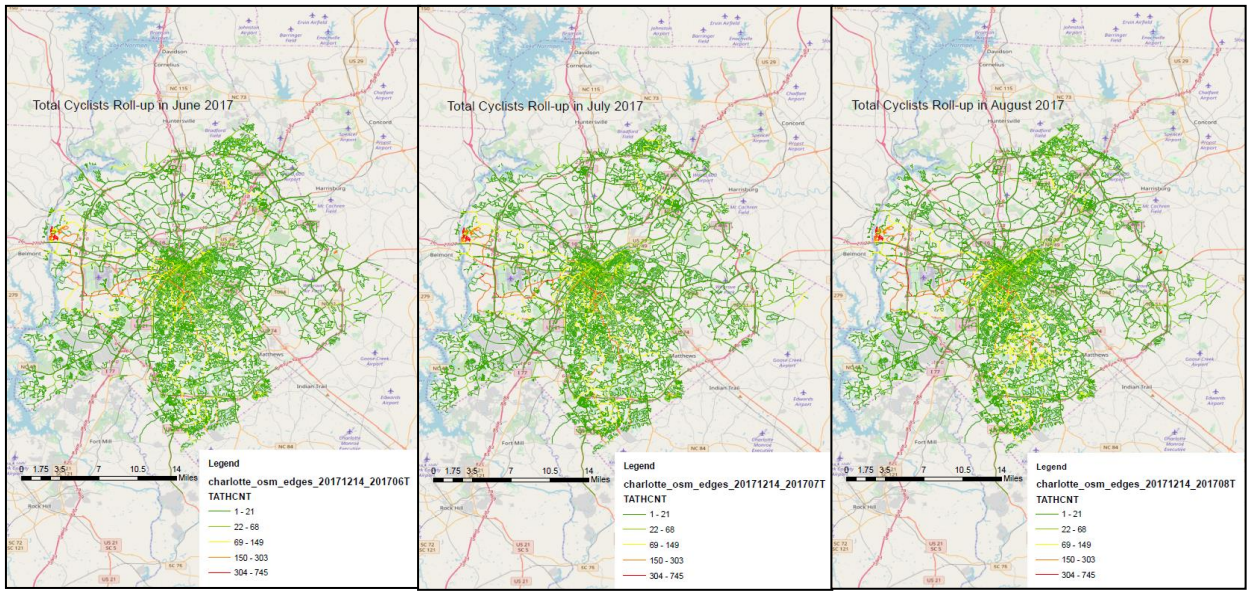
4.6.e March 2017



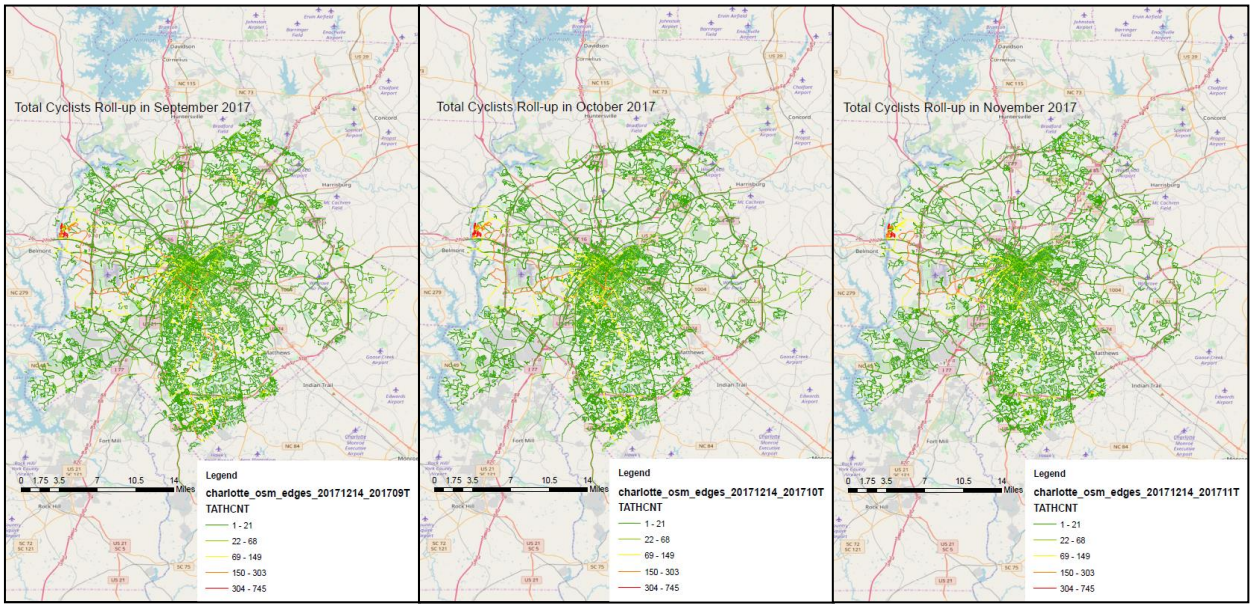
4.6.f April 2017



4.6.g May 2017



4.6.h June 2017 4.6.i July 2017 4.6.j August 2017



4.6.k September 2017 4.6.l October 2017 4.6.m November 2017

Figure 4.6: 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 concluded 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 greenway 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.

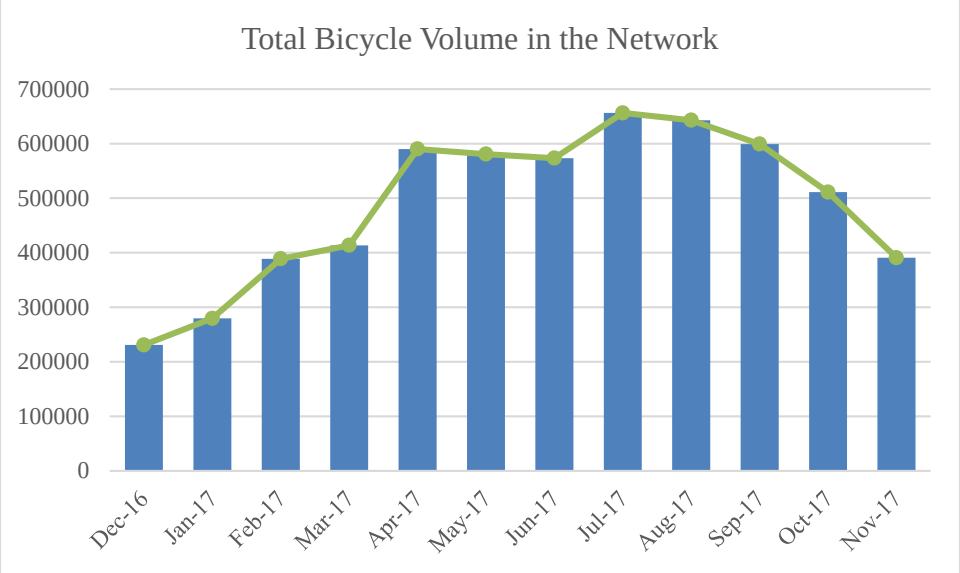


Figure 4.7: Total Bicycle Volume in the Network

4.2.3.2 Weekdays & Weekends

There is a noticeable difference between bicycle volume on weekdays and weekends in which red lines indicate higher volume on weekends and green lines represent higher volume on weekdays. From the figure, one can also see that more weekday cycling trips occurred in the uptown area, and for most areas that are not located in the uptown area, more weekend cycling trips occurred. In addition, around the boundary of the City of Charlotte, there are more cycling trips that occurred on weekdays than those on weekends. In summary, 45.64% of the cycling activities occurred during weekdays. Among these trips, 21.16% are commuting trips.

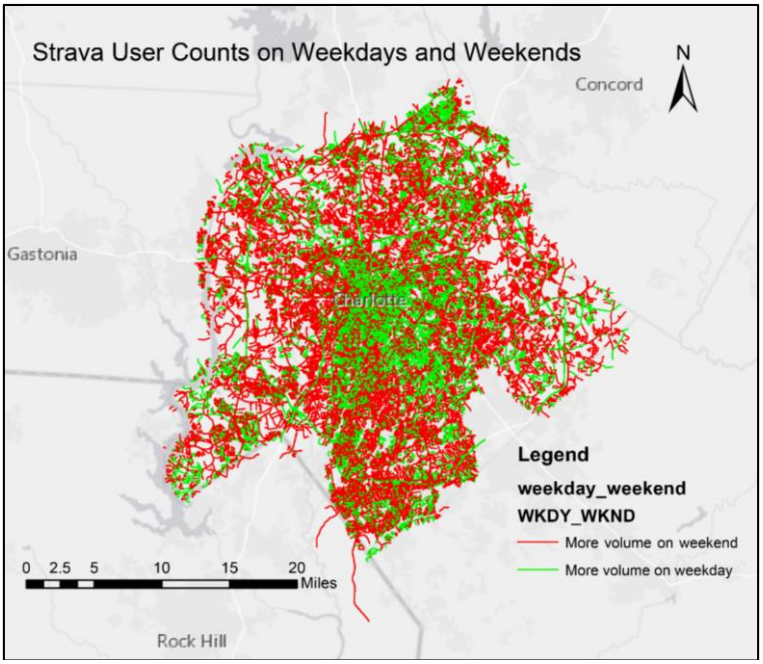


Figure 4.8: Total Bicycle Volume on Weekdays and Weekends

4.2.3.3 Time of Day

Total bicycle volume varies with different time of day. The volume on each road segment during a specific time of day is presented in the following figure. According to the figure, cyclists prefer to bike from 05:00 to 19:00.

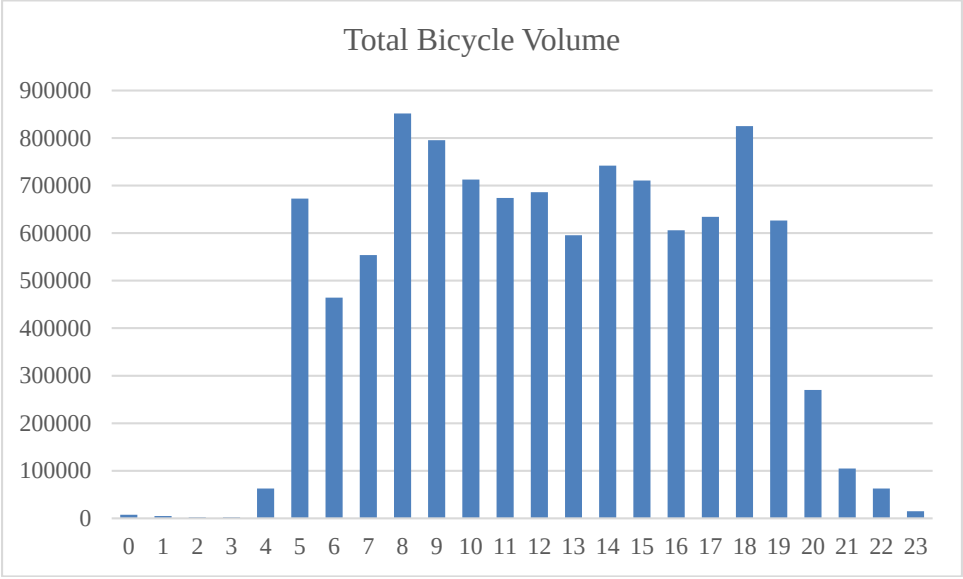


Figure 4.9: Total Bicycle Volume for Different Time of Day

4.2.3.4 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.

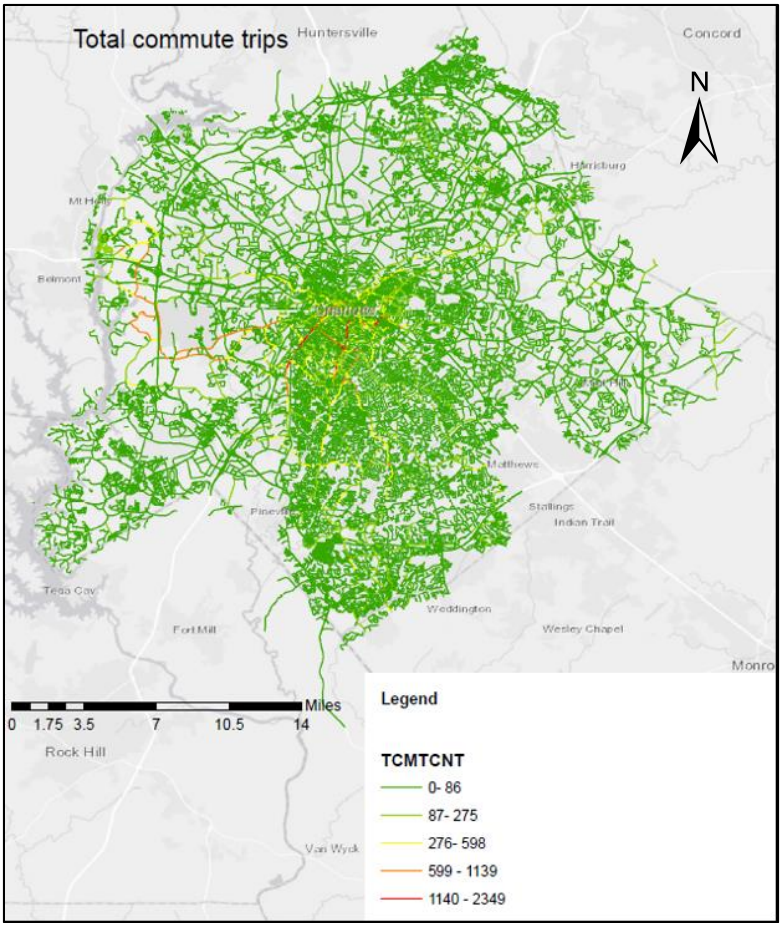


Figure 4.10: Total Commute Trips

4.3 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.

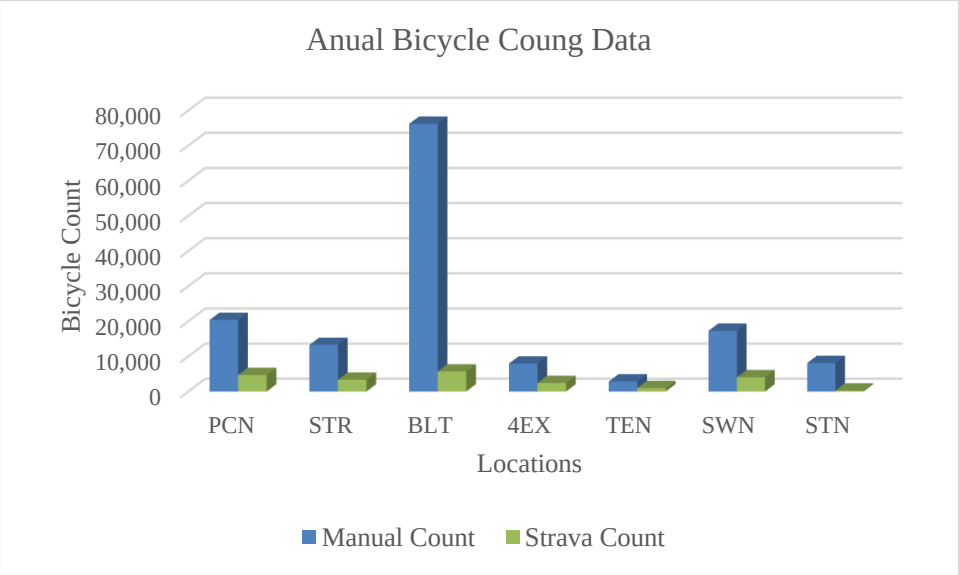


Figure 4.11: Comparison of Manual and Strava Count

4.4 Summary

This chapter provides the descriptive analyses based on the data collected in Chapter 3. Strava data analysis is conducted by creating several heatmaps for bicycle volume in different month of year, weekdays and weekends, and for different trip purposes. A data comparison between bicycle manual count data from the count station in the City of Charlotte and Strava bicycle data from the smartphone application is also provided.

Chapter 5. Developing Bicycle Volume Models

5.1 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 5.2 introduces the methods of data processing with ArcGIS and SAS. Section 5.3 presents the bicycle volume models and the model estimation results. Section 5.4 provides the bicycle volume prediction for most of the road 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 5.5 concludes this chapter with a summary.

5.2 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 time of day. In order 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 the following figure.

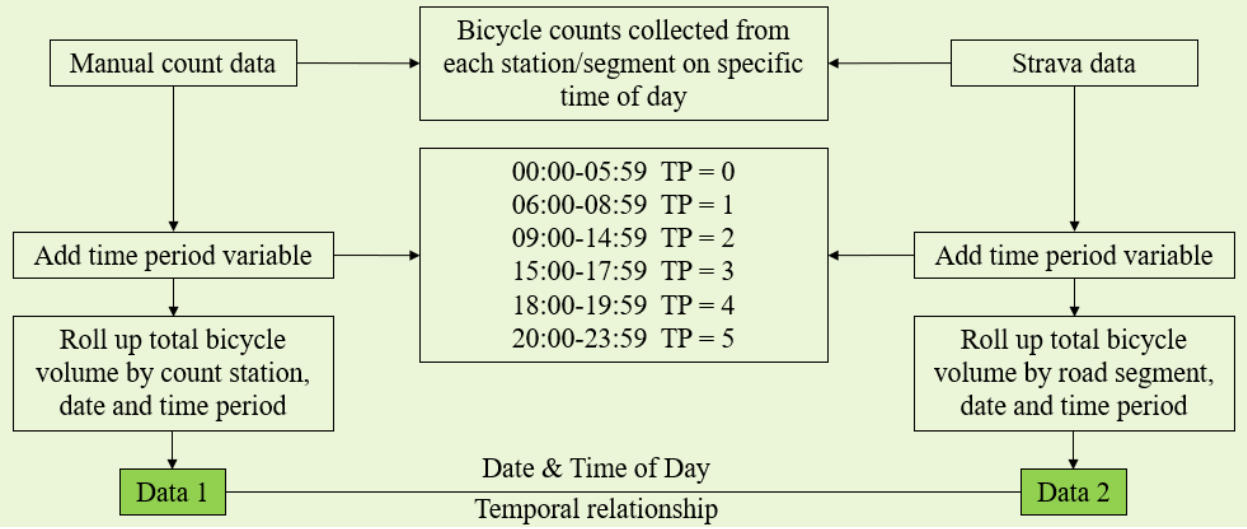


Figure 5.1: First Step of the Data Processing Procedure in SAS

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 which shares the same road segment ID with Data 2 in step 1 is 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 the following figure.

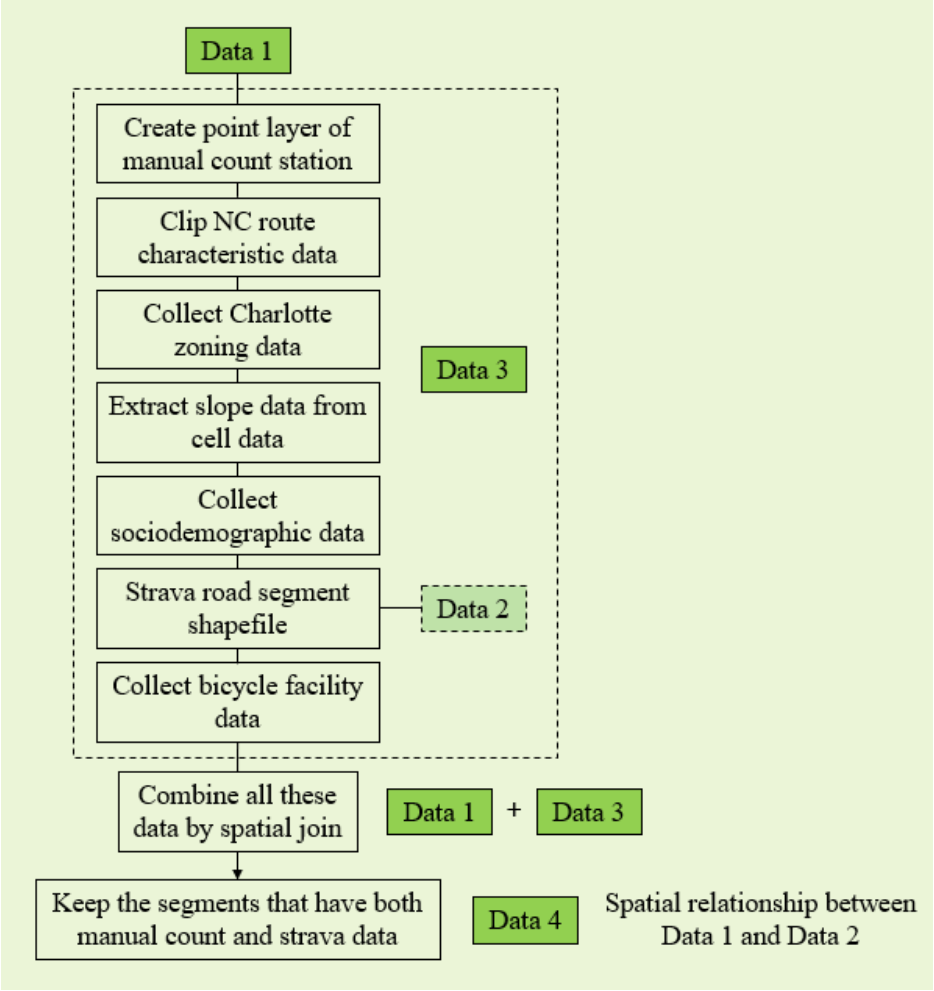


Figure 5.2: Second Step of the Data Processing Procedure in ArcGIS

Step 3:

Now that Data 4 contain the spatial relationship between Data 1 and Data 2 and information of 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 the following figure.

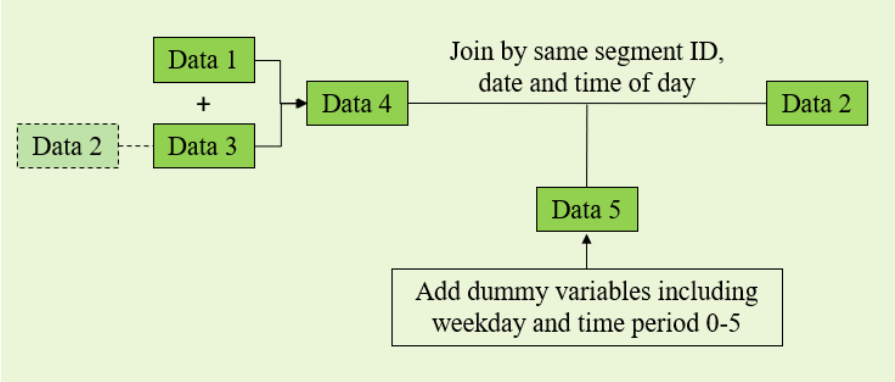


Figure 5.3: Third Step of the Data Processing Procedure in SAS

5.3 Bicycle Volume Regression Models

5.3.1 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 5.1 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 an increase of a Strava user on a road segment can lead to an increase of 7.63 total bicyclist count on the specific 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 are affected by not only the Strava data. 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.

5.3.2 Multiple Linear Regression Model

To examine the impact of various 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 5.2.

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 5.2 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.

Variable Type	Variable Label	Description
	Hour_5	If cycling time is during 20:00-23:59, then Hour_5 = 1.
	Weekday	If bike on a weekday, then weekday = 1.
Bicycle facilities	Off_Street_Paths	Off street paths
	Bike_Lanes	Bike lanes
	Signed_bike_lanes	Signed bike lanes
	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 5.3.

Table 5.3 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 5.3, variables including weekday, time period except 00:00-06:00 am, Strava user count, bike lanes, off-street paths, 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.

5.4 Bicycle Volume Prediction

Based on the model estimation results from the multiple linear regression model, a bicycle volume prediction on all the road segment in 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]

Therefore, the average annual daily bicycle (AADB) prediction can be calculated using the following equation, and an AADB prediction of most of the road segments in the City of Charlotte is presented in Figure 5.4.

AADB = Bicycle volume/365

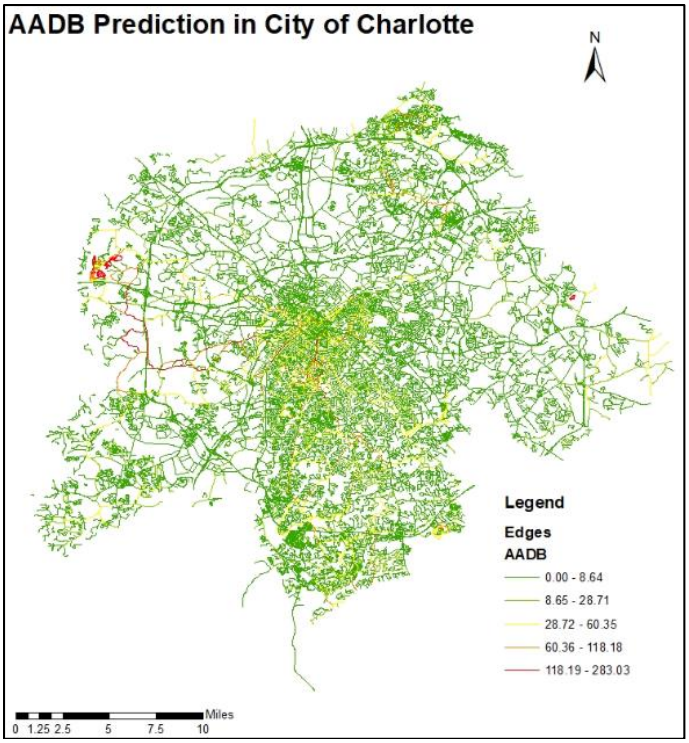


Figure 5.4: AADB Prediction in City of Charlotte

5.5 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 6. Modeling Strava Users' Cycling Route Segment Choice

6.1 Introduction

This Chapter develops an ordered probit model to analyze the Strava users' cycling route segment choice. The rest of this chapter is organized as follows. Section 6.2 provides a data processing method to combine all the needed data for the development of a preferred ordered probit model. Section 6.3 presents the preferred ordered probit model. The model estimation results are provided, and result analysis is also conducted in this section. Finally, Section 6.4 concludes this chapter with a summary.

6.2 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 characteristic 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 characteristic 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 6.1.

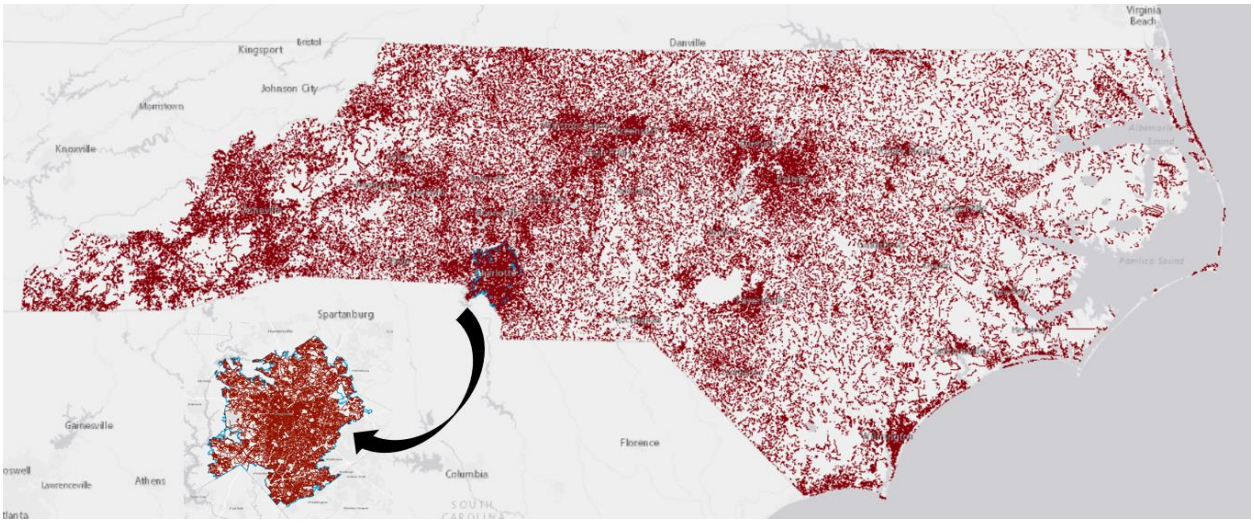


Figure 6.1: Clip in ArcGIS

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 6.2.

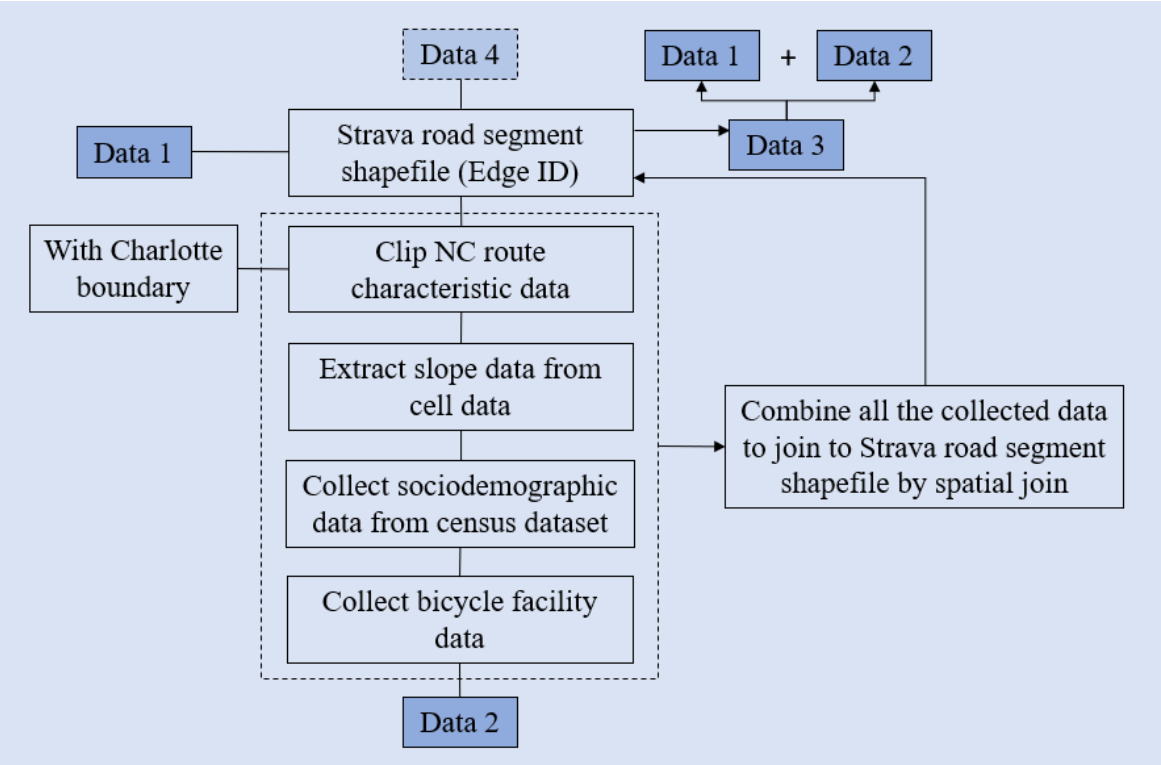


Figure 6.2: Data Processing in ArcGIS

Step 2:

This step is accomplished in SAS. First, the Strava bicycle count data (Data 4) collected from each road segment on 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 need to first convert the day of year variable to date using DATEJUL in SAS, and leading zeros are added to make sure 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 6.3.

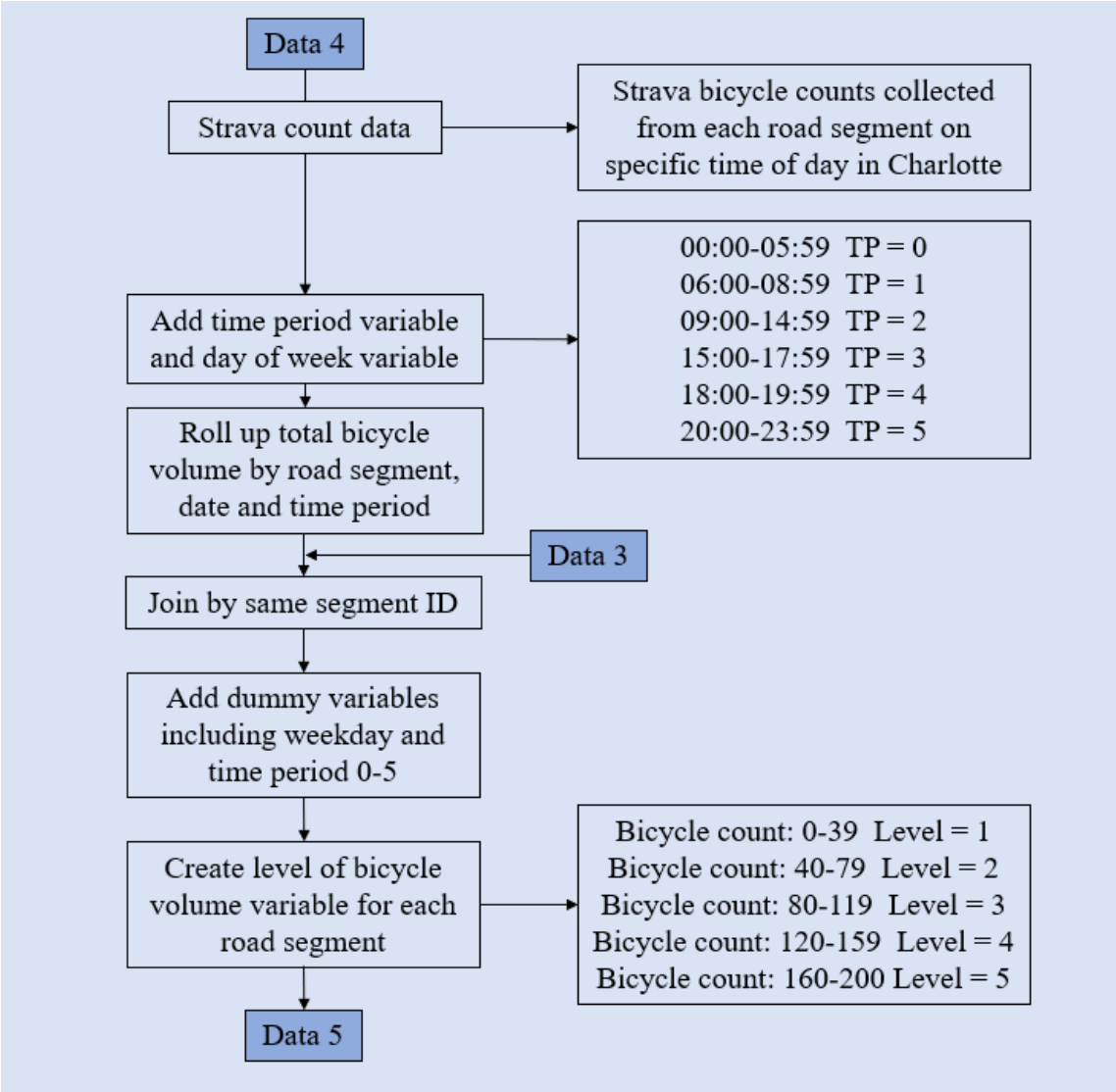


Figure 6.3: Data Processing in SAS

6.3 Ordered Probit Model Development

6.3.1 Ordered Probit Model

The Ordered Probit Model (ORP) model has been utilized for analysis of the dependent variables that are ordinal in nature (for example, bicycle volume). The model specification is shown as follows:

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

where y_i^* denotes the latent bicycle volume i, X_i represents the explanatory variables that contribute to the bicycle volume, β is a vector of the parameters that need to be estimated, and ε_i stands for the error term which follows a normal distribution.

In this study, the latent variable y_i^* has been divided by threshold θ_j(j = 1, 2, …, J) into J intervals, and the bicycle volume can be presented as follows:

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

Thus, the probability of bicycle volume j can be presented as follows:

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

where Φ is the cumulative standard normal distribution function.

In this study, the bicycle volume denotes the number of bicyclists (strava users) on each road segment which is collected by using the strava application. The counts of the bicyclists on each road segment in the City of Charlotte are divided into five groups (from the lowest to the highest bicycle volume).

6.3.2 Model Results

Based on the data processed in Section 6.2, an ordered probit model is developed to analyze the Strava users' route segment choice. Variables related to time period, weekday, roadway characteristics, sociodemographic features, slope, and bicycle facilities are examined and included in this model. The variable descriptions are shown in Table 6.1.

Table 6.1 Variable Description

Variable Type	Variable Label	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.

Variable Type	Variable Label	Description
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.
	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

The multinomial variables (level of bicycle volume in this research) are inherently ordered. Although the outcome is discrete, the multinomial logit models might lack the ability to account for the ordinal structure of the dependent variable. Therefore, the ordered probit model is developed here to determine the likelihood of each road segment being selected as a part of a cycling route. Model parameters and threshold in the ordered probit model are estimated using the maximum likelihood estimation method which is performed in SAS. A backward selection is utilized to keep all the variables that have a significant impact on the Strava users' route segment choice. The removed insignificant variables are summarized in Table 6.2. And the model estimation results are present in Table 6.3 and Table 6.4.

Table 6.2 Summary of Backward Elimination

Step	Effect Removed	Wald Chi-Square	Pr > ChiSq

1	Hour_0	0.0000	1.0000
2	B_offstreet	0.0001	0.9934
3	Hour_4	0.0036	0.9519
4	SpeedLimit	0.1184	0.7307
5	B_bikelane	0.1396	0.7087
6	Poverty	0.6591	0.4169
7	TOTPOP_CY	1.4344	0.2310
8	RouteClass1	2.1943	0.1385
9	RouteClass3	2.8771	0.0898
10	MEDAGE_CY	2.5959	0.1071

Table 6.3 Ordered Probit Model Estimation Results

Analysis of Maximum Likelihood Estimates				
Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept_1	1.7024	0.2631	41.8727	<.0001
Intercept_2	2.3836	0.2652	80.8116	<.0001
Intercept_3	2.4763	0.2659	86.7190	<.0001
Intercept_4	2.8432	0.2725	108.9019	<.0001
weekday	-1.2939	0.0814	252.7491	<.0001
Hour_1	0.3303	0.1181	7.8177	0.0052
Hour_2	0.2885	0.1143	6.3687	0.0116
Hour_3	0.9084	0.1128	64.8323	<.0001
MPLength	0.3466	0.1656	4.3817	0.0363
ThruLane	0.1738	0.0248	49.1024	<.0001
MEDHINC_CY	0.000011	9.108E-7	141.8903	<.0001
Total_HH	0.000528	0.000123	18.2768	<.0001
TotalFamily	-0.00060	0.000169	12.8010	0.0003
Slope	-0.0208	0.00342	37.0266	<.0001
B_signed	-0.4031	0.0701	33.0406	<.0001
B_suggested	0.2403	0.1109	4.6918	0.0303
B_suggest0	-0.6727	0.1181	32.4429	<.0001
B_greenway	0.7137	0.2839	6.3190	0.0119

RouteClass2	0.2887	0.0743	15.1020	0.0001
Oneway	0.2441	0.0446	29.9310	<.0001
Number of observations: 237673.				

Table 6.4 Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	7480.648	5885.329
SC	7522.162	6169.689
-2 Log L	7472.648	5825.329

The model fit statistics are used to determine whether the developed model is better than a constant only model. According to Table 6.4, -2LogL for this model is less than that of the constant only model, which means this ordered probit model has a better fitness for modeling the Strava users' route segment choice. Likelihood ratio index ρ^2 is calculated in the following equation:

$$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(c)} \quad \text{Eq. (4)}$$

For the ordered probit model that considers both intercept and covariates, the likelihood ratio index ρ^2 equals to 0.22. According to Train (2009)'s research, a higher ρ^2 indicates a better model. However, ρ^2 from 0.2 to 0.4 is good enough for real world case studies (Train, 2009). Therefore, the ordered probit model developed for the likelihood of a road segment being chosen as a part of the cycling route is robust.

Based on the model estimation results in Table 6.3, variables including weekday, time period from 06:00 to 18:00, segment length, number of through lanes, median household income, total household, total family, slope, signed bicycle road, suggested bicycle route both normal and low comfort, greenway, route class 2, and one way, have a significant impact on the route segment choice. To be specific, the analysis is conducted in detail as follows:

1. Temporal variables:
   Cycling choice will vary with the change of time in terms of day of week and time of day. For different time of day, the traffic condition on each road segment will be different which will have a significant impact on bicycle usage. In addition, the light condition and temperature variance throughout the day also affect the cycling choice. Based on the results from the model estimation, weekdays have a negative impact on the route segment choice. More Strava users choose to bike on weekends compared to weekdays. The time period from 6:00 to 18:00 has a positive impact on cyclists' route choice, which means compared to the nighttime, Strava users prefer to bike from early in the morning till 18:00. Two assumptions can be made: First, some portion of these Strava users have flexible working hours, and they will bike during the daytime. Second, some of the trips are occurred during weekends, and these Strava users preferred biking with better light condition and safer biking environment.

2. Road characteristics:

One of the major factors that affect cyclist route choice is road characteristics. The road characteristics that have a significant impact on Strava users' route segment choice include the length of a road segment, the number of through lanes, US route, and road direction of one way. From the model estimation results presented in Table 4, the length of road segment has a positive impact on cyclists' route choice. It can be interpreted that cyclists prefer to bike on longer road segments. Therefore, longer road segment will be more likely to be selected as a part of the cyclists' routes. Greater number of through lanes has a positive impact on cyclists' route choice, which means Strava users tend to select a road segment with greater number of through lanes as a part of their cycling routes. Different route classes will have an impact on cyclists' route choice. According to the model estimation results, US route will positively affect Strava users' route segment choice, suggesting that US routes are more popular among Strava users compared with other route classes including interstate route, NC route, and secondary route. In addition, one-way road segments are preferred. That is probably due to the preference for cycling in uptown areas where there are lots of one-way routes there. Surprisingly, speed limit does not have a significant impact on the route choice of Strava users. An assumption can be made that few cyclists will bike on highways and that the speed limit on the local streets does not change much. That might be the reason that speed limit does not have a significant impact on cyclist route choice.

3. Sociodemographic characteristics:
   Different sociodemographic characteristic variables have different impacts on cyclist route choice. Based on the model estimation results in Table 4, median household income, total household number, and total families in a census block can significantly affect the Strava users' route segment choice. Results reveal that areas with high median household income has a positive impact on the Strava users' cycling choice. This result is consistent with the cyclist route choice research conducted by LaMondia and Watkins (2017). It can be interpreted that some of the areas with high median income may be located in the center city where a higher bicycle volume exists. In addition, higher safety is always associated with areas with high median household income. Interestingly, total households and total families have different impacts on cyclists' route choice. According to the model estimation results, total households have a positive impact on Strava users' route choice, while total families have a negative impact. It can be assumed that the center city has a lower number of families but a higher bicycle volume. Cyclists tend to choose areas with higher rental apartments, and less family house neighborhoods.

4. Geometry characteristics:
   This variable is examined to identify the correlation between the likelihood of choosing a road segment as a part of the cyclists' route and the slope of a road segment. Results show that the slope has a negative impact on Strava users' route segment choice. It is not hard to imagine that cyclists prefer to select a flat segment instead of a steep segment as one part of their cycling routes.

5. Bike facilities:
   Bike facilities are another critical consideration for the cyclists' route choice. Better bike facilities will provider higher cycling safety which is perhaps the most important issue

while cycling. Bike facilities examined in this ordered probit model include off-street paths, bike lanes, signed bike lanes, and suggested bike routes. Suggested bike routes with low comfort and greenway of which four types of facilities have a significant impact on cyclists' route choice. Greenway and suggested bike routes will increase the likelihood of the road segment being selected as a part of the cycling route, while signed bike lanes and suggested bicycle route with low comfort will decrease the probability of being chosen. That is because greenway and suggested bike routes usually have a better road condition, and the other two types of bicycle facilities may not provide a good enough road condition.

To conclude, variables which are associated with providing higher safety and comfort level will have a positive effect on the likelihood of a road segment being selected as a part of the cycling route, and vice versa.

6.4 Summary

This chapter develops an ordered probit model to analyze the Strava users' cycling route choice. Different variables are evaluated to quantify their impacts on cycling route choice in the City of Charlotte. Results reveal that weekday, time of day, road characteristics, demographic data, slope, and bicycle facilities will have significant impacts on cycling route choice.

Chapter 7. Summary and Conclusions

7.1 Introduction

As an efficient way of alleviating traffic congestion and improving air quality, cycling has been encouraged for short distance trips to provide a healthier and greener travel. To motivate cycling in the City of Charlotte, research need to be conducted on studying the contributing factors to the bicycle volume and route choice. One of the most critical issues for the conduct of such research studies is that the traditional data collection methods have some limitations and their data collecting process can be time-consuming and expansive.

Recently, crowdsourcing has become prevalent in transportation planning and management. It offers shared platforms and systems to invite a large number of interested people to address common issues that affect them all. Recently, crowdsourcing techniques have been developed rapidly. Some studies regarding its use in transportation have shown its immense potential in augmenting or replacing the normal data collection methods. Since crowdsourcing has many advantages in data collection, it is leveraged in this research study.

Based on the crowdsourced bicycle data collected from the Strava smartphone application, this research is conducted to estimate the bicycle volume on most of the road segments in the City of Charlotte and analyze the Strava users' route segment choice in the City of Charlotte.

The primary objective of this research is to systematically develop bicycle volume models and a bicyclist route segment choice model. 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 count stations in the City of Charlotte, temporal data, and bicycle facility data. 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. 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. In addition, an ordered probit model is developed to analyze the Strava users' cycling route segment choice in the City of Charlotte.

The rest of this chapter is organized as follows. Section 7.2 provides a brief review of the methods used to develop the bicycle volume prediction models and the Strava users' route choice model. Model estimation results are concluded in this section and policy-related recommendations are provided here. Section 7.3 details the directions that should be considered in the future research in order to improve the cycling environment in the City of Charlotte.

7.2 Summary and Conclusions

As mentioned above, bicycle volume models and a bicyclist route segment choice model are developed in this research. Crowdsourced bicycle data from Strava smartphone application

are collected to combine with a series of relevant data that contain NC road characteristics data, demographic data, slope data, manual count data from count stations in the City of Charlotte, temporal data, and bicycle facility data. 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. Bicycle volume on most of the road segments in the City of Charlotte is predicted based on the model estimation results. A bicycle ridership map is created to have a graphical view of the bicycle counts. In addition, an ordered probit model is developed to analyze the Strava users' cycling route segment choice in the City of Charlotte.

Based on the results of this research, variables including weekday, time period except 00:00-06:00 am, Strava user counts, bike lanes, off-street paths, and suggested bike routes have a significant impact 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. In terms of the variables that have significant impacts on Strava user count on a road segment, time period from 06:00 to 18:00, road segment length, number of through lanes, median household income, total households in a census block, cycling on a suggested bike route, greenway, US route, and one-way road have a positive impact on Strava user counts on a road segment, while variables including cycling on weekdays, total families in a census block, slope, signed bike route, and suggested bike routes with low comfort have a negative impact on the Strava user counts on a road segment.

According to the model results 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 the traffic in Charlotte uptown area is bad, especially during peak hours, and the bicycle volume is higher there compared to other locations, constructing more off street paths will 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.

If the above policy-related recommendations are followed, better bike environment can be provided for the citizens in Charlotte to improve their life quality 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 the 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 central 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. Therefore, biases might exist when predicting the bicycle volume.

   (3) These 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) Some variables are assumed to have impacts on the bicycle volume, such as traffic volumes. However, no traffic volume data can be collected for the case study conducted in this research.

2. Strava users' route segment choice:

   (1) Some roadway segments that lack supporting data like roadway characteristics data are removed from the dataset.

   (2) Variables including traffic volumes may have some impacts on Strava users' route segment choice. However, such data are not available in this study.

Based on the limitations of this research 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 tested to see if there is a better fitness for modeling the bicycle volume and the route segment choice.

3. Crash frequency or severity can be examined to see if they have a negative impact on bicycle volume or route segment choice. In addition, research on how to solve these

safety problems will need to be conducted in order to provide a better bicycle environment for cyclists.

References

1. Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). "A systematic review of open government data initiatives." *Government Information Quarterly*, *32*(4), pp. 399-418.
2. Ben-Akiva, M. and Bierlaire, M. (1999). "Discrete choice methods and their applications to short term travel decisions." R. Hall, ed., *Handbook of Transportation Science*, Kluwer Academic Publishers, Norwell, MA, chapter 2, pp. 5-34.
3. Bergman, C., and Oksanen, J. (2016). "Conflation of OpenStreetMap and Mobile Sports Tracking Data for Automatic Bicycle Routing." *Transactions in GIS*, 20(6), pp. 848-868.
4. Bovy, P. and Fiorenzo-Catalano, S. (2007). "Stochastic route choice set generation: behavioral and probabilistic foundations." *Transportmetrica,* 3, pp. 173-189.
5. Brabham, D. C. (2008). "Crowdsourcing as a model for problem solving: An introduction and cases." *Convergence,* 14(1), pp. 75-90.
6. Broach, J., Dill, J., and Gliebe, J. (2012). "Where do cyclists ride? A route choice model developed with revealed preference GPS data." *Transportation Research Part A: Policy and Practice*, *46*(10), pp. 1730-1740.
7. Casello, J., and Usyukov, V. (2014). "Modeling cyclists' route choice based on GPS data." *Transportation Research Record: Journal of the Transportation Research Board*, 2430, pp. 155-161.
8. Charlton, B., Hood, J., Sall, E., and Schwartz, M. A. (2011). "Bicycle route choice data collection using GPS-enabled smartphones." In *Transportation Research Board 90th Annual Meeting.* No. 11-2652.
9. Chen, M., Mao, S., and Liu, Y. (2014). "Big data: A survey." *Mobile Networks and Applications* 19(2), pp. 171-209.
10. City of Charlotte Department of Transportation. (2017, May 22). Charlotte BIKES Bicycle Plan. Available at http://charlottenc.gov/Transportation/Programs/Documents/Charlotte%20BIKES%20Final.pdf
11. Dill, J., and Gliebe, J. (2008). "Understanding and measuring bicycling behavior: A focus on travel time and route choice."
12. Estellés-Arolas, E., and González-Ladrón-De-Guevara, F. (2012). "Towards an integrated crowdsourcing definition," *Journal of Information science*, *38*(2), pp. 189-200.
13. Foresite Group. (2015). Available at http://www.fg-inc.net/crowdsourced-data-for-bike-pedestrian-facility-planning/.
14. Fosgerau, M., Frejinger, E., and Karlstrom, A. (2013). "A link based network route choice model with unrestricted choice set." *Transportation Research Part B: Methodological*, 56, pp. 70-80.

15. Gantz, J., and Reinsel, D. (2011). "Extracting value from chaos." *IDC iview* 1142(2011), pp. 1-12.
16. Griffin, G. P. and Jiao J. (2016). "Crowdsourcing Bicycle Volumes: Exploring the Role of Volunteered Geographic Information and Established Monitoring Methods." *Journal of the Urban and Regional Information Systems Association*, 27(1).
17. Griffin, G. P., and Jiao, J. (2015). "Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus." *Journal of Transport & Health*, 2(2), pp. 238-247.
18. Grond, K. (2016). *Route Choice Modeling of Cyclists in Toronto*. Diss. University of Toronto (Canada).
19. Guan, H. *Discrete choice Modeling*. 2004.
20. Hochmair, H. H., Bardin, E., and Ahmouda, A. (2017). "Estimating Bicycle Trip Volume for Miami-Dade County from Strava Tracking Data." No. 17-06577.
21. Hood, J., Sall, E., and Charlton, B. (2011). "A GPS-based bicycle route choice model for San Francisco, California." *Transportation letters*, 3(1), pp. 63-75.
22. Howe, J. (2006). "The rise of crowdsourcing." *Wired magazine, 14*(6), pp. 1-4.
23. Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.
24. 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.
25. Jackson, S., Miranda-Moreno, L. F., Rothfels, C., and Roy, Y. (2014). "Adaptation and implementation of a system for collecting and analyzing cyclist route data using smartphones." *Transportation Research Board 93rd Annual Meeting*. No. 14-4637.
26. Jestico B., Nelson T. and Winters M. (2016). "Mapping ridership using crowdsourced cycling data." *Journal of Transport Geography*, 52, pp. 90-97.
27. Kagerbauer, M., Hilgert, T., Schroeder, O., and Vortisch, P. (2015). "Household travel survey of intermodal trips - Approach, challenges and comparison." *Transportation research procedia, 11*, pp. 330-339.
28. Khatri, R., Cherry, C. R., Nambisan, S. S., and Han, L. D. (2016). "Modeling route choice of utilitarian bikeshare users with GPS data." *Transportation Research Record: Journal of the Transportation Research Board*, 2587, pp. 141-149.
29. Kleemann, F., Voß, G. G., and Rieder, K. (2008). "Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing." *Science, technology & innovation studies*, 4(1), pp. 5-26.
30. Kučera, J., Chlapek, D., and Nečaský, M. (2013). "Open government data catalogs: Current approaches and quality perspective." *International Conference on Electronic Government and the Information Systems Perspective*. Springer, Berlin, Heidelberg, pp. 152-166.
31. LaMondia, J., and Watkins, K. (2017). Using Crowdsourcing to Prioritize Bicycle Route

Network Improvements.
32. Liu, E., and Porter, T. (2010). "Culture and KM in China." *Vine,* 40(3/4), pp. 326-333.
33. Misra, A., Gooze, A., Watkins, K., Asad, M., and Le Dantec, C. (2014). "Crowdsourcing and its application to transportation data collection and management." *Transportation Research Record: Journal of the Transportation Research Board,* 2414, pp. 1-8.
34. Moore, Michael. (2015). "Modeling Factors Influencing Commuter Cycling Routes: A Study of GPS Cycling Records in Auburn, Alabama."
35. Nassir, N., Ziebarth, J., Sall, E., and Zorn, L. (2014). "Choice set generation algorithm suitable for measuring route choice accessibility." *Transportation Research Record: Journal of the Transportation Research Board,* 2430, pp. 170-181.
36. Proulx, F. R. and Pozdnukhov, A. (2017). "Bicycle Traffic Volume Estimation Using Geographically Weighted Data Fusion." Available at http://faculty.ce.berkeley.edu/pozdnukhov/papers/Direct_Demand_Fusion_Cycling.pdf.
37. Raihan, M. A., and Priyanka Alluri PHD, P. E. (2017). "Impact of Roadway Characteristics on Bicycle Safety." *Institute of Transportation Engineers. ITE Journal,* 87(9), pp. 33.
38. Reiche, K. J., and Höfig, E. (2013). "Implementation of metadata quality metrics and application on public government data." Accepted for *IEEE - Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual*. pp. 236-241.
39. RenoTracks. *RenoTracks*. (2013). Available at http://renotracks.nevadabike.org/.
40. 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.
41. Saxton, G. D., Oh, O., and Kishore, R. (2013). "Rules of crowdsourcing: Models, issues, and systems of control." *Information Systems Management*, *30*(1), pp. 2-20.
42. Schenk, E., and Guittard C. (2011). "Towards a characterization of crowdsourcing practices." *Journal of Innovation Economics & Management,* 1, pp. 93-107.
43. Schuessler, N., and Axhausen, K. W. (2009a). "Map-matching of GPS traces on high-resolution navigation networks using the Multiple Hypothesis Technique (MHT)." *Arbeitsberichte Verkehrsund Raumplanung,* 568, pp. 1-22.
44. Schuessler, N., and Axhausen, K. W. (2009b). "Processing raw data from global positioning systems without additional information." *Transportation Research Record: Journal of the Transportation Research Board,* 2105, pp. 28-36.
45. Stinson, M., and Bhat, C. (2003). "Commuter bicyclist route choice: Analysis using a stated preference survey." *Transportation Research Record: Journal of the Transportation Research Board,* 1828, pp. 107-115.
46. Sun, Y., and Mobasheri, A. (2017). "Utilizing Crowdsourced data for studies of cycling and air pollution exposure: A case study using Strava Data." *International journal of environmental research and public health,* 14(3), pp. 274.
47. Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

48. Vukovic, M. (2009). “Crowdsourcing for enterprises.” Accepted for *IEEE - Services-I, 2009 World Conference*, pp. 686-692.
49. Watkins, K., Ammanamanchi, R., LaMondia, J., and Le Dantec, C. A. (2016). “Comparison of Smartphone-based Cyclist GPS Data Sources.” In *Transportation Research Board 95th Annual Meeting.* No. 16-5309.
50. Wexler, M. N. (2011). “Reconfiguring the sociology of the crowd: exploring crowdsourcing.” *International Journal of Sociology and Social Policy*, 31(1/2), pp. 6-20.
51. Yeboah, G., and Alvanides, S. (2015). “Route Choice Analysis of Urban Cycling Behaviors Using OpenStreetMap: Evidence from a British Urban Environment.” In *OpenStreetMap in GIScience*, Springer International Publishing, pp. 189-210.
52. Zimmermann, M, Mai, T., and Frejinger, E. (2017). “Bike route choice modeling using GPS data without choice sets of paths.” *Transportation research part C: emerging technologies*, 75, pp. 183-196.