RESEARCH & DEVELOPMENT

Bicycle Volume: Counting Machine Validation & Correction, Estimating & Forecasting, and Analysis of Injury Risk

Wei (David) Fan, Ph.D., P.E., Zijing Lin, Ph. D., Shaojie Liu, Ph.D. Candidate
Center for Advanced Multimodal Mobility Solutions and Education
Department of Civil and Environmental Engineering
University of North Carolina at Charlotte

Sarah Searcy, Program Manager, Blythe Carter, Research Associate
Institute for Transportation Research and Education
North Carolina State University

NCDOT Project 2020-43
FHWA/NC/2020-43
August 2021

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 cycling and bicyclist injury risk, to quantify the impact of certain attributes on bicycle volume as well as biking safety, and further provide better cycling environment for cyclists to encourage non-motorized travels.

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

This project focuses on evaluating the potential use of crowdsourced bike data and comparing them with the traditional bike counting data that are collected in the City of Charlotte, North Carolina. Bicycle volume models were developed using bike data from both the Strava smartphone cycling application and permanent continuous bicycle count stations. Based on the results, a bicycle volume predictive model is presented, as well as a map illustrating the bicycle volume on most of the road segments in the City of Charlotte.

In addition, this research project investigated the correction factor calculation methodology used in the North Carolina Non-Motorized Volume Data Program (NC NMVDP) to adjust bicycle and pedestrian count data collected by permanent continuous counters across the state. The analysis examines the impacts of rounding on corrected count data, appropriate temporal aggregations for applying linear correction factors, the minimum number of non-zero observations required to properly calibrate an Eco-Counter system, differences between the magnitude of correction factors as calculated using historic programmatic processes and those calculated with linear regression methods, and methods for accounting for accuracy, tolerance, and uncertainty in count data recorded by an Eco-Counter system.

This research also developed bicycle volume prediction models using both the simple linear regression model and multiple linear regression model. In the simple linear regression model, the bicycle volume from counter stations are approximately 4.46 times those of the bicycle volume from the Strava data, with an intercept value of 5.72. In the multiple linear regression model, factors that have a significant impact on bicycle volumes are identified: five time periods, weekday, Strava counts, bike lane, and off-street path.

Furthermore, bicyclist injury risk analysis is also conducted to explore the impact factors affecting biking safety by developing a series of safety performance functions, including Poisson Model, Negative Binomial (NB) model, Zero-inflated Poisson (ZIP) model, and Zero-inflated Negative Binomial (ZINB) model. Several indicators for model comparison were used to select the best fitting model for bicyclist injury risk modeling. The ZINB model has the optimal performance compared to the other three models. Based on the results from the ZINB model, interstate, principal arterial, minor arterial, major collector, bus stop, bike lane, and annual average daily bicycle volume have significant impacts on bicycle injury risk.

Finally, recommendations are made to help improve the cycling environment and cycling safety and to increase bicycle volume in the future. The research findings suggest more bike infrastructure should be built in uptown areas, especially in the vicinity of greenways and parks to reduce cyclist crash risks, separated bike lanes away from bus stops should be constructed.

NCDOT Research Project No. RP 2020-43

BICYCLE VOLUME: COUNTING MACHINE VALIDATION & CORRECTION, ESTIMATING & FORECASTING, AND ANALYSIS OF INJURY RISK

Final Report

Submitted to

North Carolina Department of Transportation

by:

Wei Fan, Ph. D., P.E.
Zijing Lin, Ph.D.
Shaojie Liu, Ph.D. Candidate

Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE)
Department of Civil and Environmental Engineering
The University of North Carolina at Charlotte
9201 University City Boulevard
Charlotte, NC 28223-0001

Sarah Searcy, Program Manager
Blythe Carter, Research Associate

Institute for Transportation Research and Education
909 Capability Drive, Suite 3600
Research Building IV
Raleigh, North Carolina 27606

July 2021

Technical Report Documentation Page

1. Report No. FHWA/NC/2020-43	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Bicycle Volume: Counting Machine Validation & Correction, Estimating & Forecasting, and Analysis of Injury Risk		5. Report Date July 2021
		6. Performing Organization Code
7. Author(s) Wei Fan, Zijing Lin, Shaojie Liu, Sarah Searcy and Blythe Carter		8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Civil and Environmental Engineering, The University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223 Institute for Transportation Research and Education, 909 Capability Drive, Suite 3600, Research Building IV, Raleigh, North Carolina 27606		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 2020-43
12. Sponsoring Agency Name and Address North Carlina Department of Transportation Research & Development 1549 Mail Service Center Raleigh, NC 27699-1549		13. Type of Report and Period Covered Final Report: August 2019 – September 2021
		14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

Cycling is an environmentally friendly mode of transportation that can potentially mitigate traffic congestion and improve air pollution. City planners and decision-makers have begun to incorporate the needs of cyclists when building infrastructures. However, factors like annual average daily bicycle volume (AADB) and cycling injury risk should be taken into consideration during the planning. As technology develops, cost-effective crowdsourced bicycle data and traditional bicycle count data from permanent continuous count stations are useful tools in investigating the bicycle volume and determining injury risk to cyclists. This research project used bicycle count data from permanent continuous counters and data from a smartphone application (i.e., Strava). Upon installation, a permanent continuous bicycle or pedestrian counter undergoes validation to determine a correction factor. This factor accounts for systematic equipment error including false positive bicycle counts that are attributed to motor vehicles and natural occlusions of a pedestrian sensor when a pair of pedestrians walk side-by-side, among other possible errors. The research team analyzed the validation and correction factor calculation methodology used in the North Carolina Non-Motorized Volume Data Program (NC NMVDP) by examining the impacts of rounding on corrected count data, appropriate temporal aggregations for applying linear correction factors, the minimum number of non-zero observations required to properly calibrate an Eco-Counter system, differences between the magnitude of correction factors as calculated using historic programmatic processes and those calculated with linear regression methods, and methods for accounting for accuracy, tolerance, and uncertainty in count data recorded by an Eco-Counter system. The research team also developed a bicycle volume prediction model based on both the corrected bicycle count data recorded by the permanent continuous count stations in the NC NMVDP and crowdsourced bicycle data from the smartphone applications. The research team also generated a ridership map according to the prediction results. In addition, the cyclist crash data were integrated with the crowdsourced bicycle data using ArcGIS software. Locations with high crash risks were identified and injury risk models were developed. Important factors that affect crash risks are selected and discussed.

17. Key Words Counting machine validation, bicycle volume prediction, injury risk analysis, Strava, crowdsourced data	18. Distribution Statement		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 101	22. Price

DISCLAIMER

The contents of this report reflect the views of the authors and not necessarily the views of the University of North Carolina at Charlotte (UNC Charlotte), Institute of Transportation Research and Education (ITRE), or the North Carolina Department of Transportation (NCDOT). The authors are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of either UNC Charlotte, NCDOT, ITRE, or the Federal Highway Administration (FHWA) at the time of publication. This report does not constitute a standard, specification, or regulation.

ACKNOWLEDGMENTS

The authors acknowledge the North Carolina Department of Transportation (NCDOT) for providing financial support for this project. Special thanks are extended to John Vine-Hodge, Curtis Bradley, Joseph Furstenberg, Paul Black, Alex Riemondy, Kent L Taylor, Brian Murphy, and Paul Benton of NCDOT for providing excellent support, guidance, and valuable input for the successful completion of this project. Without the help of all the above individuals, this project could not have been completed in such a successful manner.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Problem Statement and Motivation

Numerous sources of novel data, including crowdsourced data collected from smartphones, have emerged and been used for transportation research. The innovative crowdsourcing data collection method offers unique features and advantages when compared with the traditional data collection method. Thus, many researchers in relevant research fields have applied the crowdsourced data method to their research. This is only the beginning of the benefit from crowdsourcing; and this kind of data collection method still has great potential to further advance transportation research studies.

In order to estimate bicycle volume on each roadway segment and to encourage cycling, studies need to examine the contributing factors to bicycle volume and the correlation between bicycle data from permanent continuous count stations and crowdsourced bicycle data. One of the critical issues for the conduct of such research studies is that the traditional data collection methods have some limitations, and their data collection process can be time-consuming and expensive (Boss et al., 2018; Musakwa & Selala, 2016).

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 (Misra et al., 2014). Howe (2006) first introduced this method in "The Rise of Crowdsourcing" and discussed how planners can use crowdsourced data to develop models, analyze travel behavior, estimate traffic demand, evaluate bike facilities, and explain road traffic safety.

Different research efforts have been made with different definitions for crowdsourcing (Estellés-Arolas & González-Ladrón-de-Guevara, 2012). 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" (see page 79). Usually, the definitions of crowdsourcing contain three main features: The users of the crowdsourcing application who provide critical information, the outsourcing procedure that spreads out the data, and the internet-based platform that enables the accomplishment of crowdsourcing (Saxton, 2013).

The concept of crowdsourcing has also been used within various research areas (Whitla, 2009; Boulos et al., 2011; Brabham, 2012; Overeem et al., 2013; Marti et al., 2012), and numerous smartphone applications have been developed for transportation planning and management, including CycleTracks, Strava, MapMyRide, etc. (Figliozzi and Blanc, 2015). These applications use GPS technology to provide users' information regarding their trip trajectories which will benefit researchers' and planners' studies.

Therefore, crowdsourcing is especially helpful and beneficial to transportation planning and management. It offers shared platforms and systems to invite a large amount of interested application users to address common issues that affect them all. Recently, crowdsourcing techniques have developed rapidly. Some studies regarding crowdsourced data use in

transportation have shown its immense potential in augmenting traditional data collection methods since it can provide temporal and spatial information to help researchers analyze the travel behavior of application users and increase the usage of the software.

1.2 Study Objectives

The objective of this research is to validate the bicycle volume data collected from counting machines and determine the correction factor, and to evaluate and apply the potential use of crowdsourced bicycle data in the City of Charlotte. This data will be used to develop bicycle volume estimation and prediction models, conduct injury risk analysis with safety performance functions, map bicycle ridership, and analyze biking safety influence. The proposed work in this research is to fulfill the following objectives:

1. To analyze and evaluate the methodology used in the North Carolina Non-Motorized Volume Data Program (NC NMVDP) to validate the bicycle count data recorded by permanent continuous counters and to determine correction factors;
2. To compile bicycle data from all the available sources: Strava data, bicycle count data collected from permanent continuous count stations, NC road characteristics data, demographic data, slope data, temporal data, annual average daily traffic (AADT), bicycle facility data, and bicyclist involved crash records as preparation for the follow-up work;
3. To combine all the collected data using ArcGIS and SAS for model estimation;
4. To develop bicycle volume estimation and prediction models based on the combined data;
5. 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;
6. To develop safety performance functions based on bicycle volume for bicyclist safety analysis.

1.3 Expected Contributions

In order to provide a better cycling environment and to encourage more potential bicyclists to bike in the City of Charlotte, this research pairs the verified and validated bicycle counts recorded by permanent continuous count stations in the NC NMVDP with Strava data to develop models to analyze the factors that are associated with the actual 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 bicycle facility construction and improvement in the future. The impacts of biking safety also need to be analyzed. Given these goals, the expected contributions of this research are as follows:

1. Refine the method used to validate and correct bicycle count data collected by permanent continuous counters in the NC NMVDP;
2. Present a systematic method for developing models to analyze the relationship between bicycle count data from count stations and Strava's bicycle count data;
3. Generate a bicycle ridership map of 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;

4. Provide a method to develop safety performance functions for analyzing bicyclist injury risk and mapping bicycle-vehicle crashes.

1.4 Research Overview

This report is organized as follows:

Chapter 1 introduces the background of this research study and discusses the motivation of modeling cycling activities and conducting safety analysis. In addition, the objectives and expected contributions are described and presented.

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. This chapter also summarizes the data collection methods used for relevant research studies including crowdsourcing and other traditional data collection methods. Representative smartphone applications for crowdsourcing are presented and their use for different aspects of research is discussed. Methods for bicycle volume estimation and prediction, counting machine validation and correction, and bicyclist injury risk analysis are summarized.

Chapter 3 gives an overview of the collected data and conducts a descriptive analysis of the data collected from the Strava smartphone application in terms of users' demographics, different trip purposes, and total Strava counts. Chapter 3 also provides a simple data comparison between bicycle counts collecting from permanent continuous count stations and the Strava application. In addition, other supporting data are introduced in this chapter.

Chapter 4 introduces the bicycle volume estimation and validation procedures. Specifically, this chapter discusses the impact of rounding on corrected count data, appropriate temporal aggregations for applying linear correction factors, the minimum number of non-zero observations required to properly calibrate an Eco-Counter system, differences between the magnitude of correction factors, and methodologies for accounting for accuracy, tolerance, and uncertainty in count data recorded by an Eco-Counter system.

Chapter 5 presents a method for data processing and develops two linear regression models to analyze the relationship between bicycle count data from permanent continuous count stations and Strava, as well as other relevant attributes. The bicycle volume on most road segments in the City of Charlotte is predicted using the developed model. A bicycle ridership map is also created to display a graphical representation of the bicycle counts.

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

Chapter 7 concludes this research with a summary of the methods for estimating and predicting bicycle volume on each road segment in the city network and conducting bicyclist injury risk analysis.

2 LITERATURE REVIEW

2.1 Introduction

This chapter presents a comprehensive literature review on the current state-of-the-art and state-of-the-practice of relevant non-motorized transportation research studies, especially counting machine validation and correction, bicycle volume estimation and prediction, impacts on bicycle activity, and bicyclist injury risk analysis. This literature review will summarize the data used for the research studies, methods applied for counting machine validation and correction, bicycle volume estimation and prediction, and injury risk analysis, as well as results concluded from previous and ongoing research.

The remainder of this chapter is structured as follows: Section 2.2 introduces different types of data collection, such as crowdsourcing, open data, and big data, as well as other traditional data collection methods including travel surveys and count data. Section 2.3 summarizes the most prevalent smartphone crowdsourcing applications (e.g., CycleTracks, Cycle Atlanta, Mon RésoVélo, Strava, and ORcycle) and their use on different aspects of research. Section 2.4 details the bicycle volume estimation and prediction methods based on both traditional data collection methods and crowdsourcing. Section 2.5 reviews the methods for counting machine validation and correction. Section 2.6 presents the approach to bicyclist injury risk analysis based on different types of data. Finally, section 2.7 concludes this chapter with a summary.

2.2 Data Collection

This section summarizes advanced data collection methods and the traditional data used for relevant research studies. This section introduces each type of data and the advantages and disadvantages of novel data and traditional data.

2.2.1 Crowdsourcing

Crowdsourcing is an innovative sourcing model which brings new developments to data collection and data-driven research studies. Crowdsourcing techniques have evolved rapidly since they emerged approximately ten years ago. The concept of crowdsourcing was first introduced by Howe (2006) in his article “The Rise of Crowdsourcing” published in Wired Magazine in which he defines crowdsourcing as the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.” (Howe, 2008) He expands his definition as follows:

> “Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively) but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers.” (Howe, 2008, page 1)

Crowdsourcing is a mixture of two components: crowd and outsourcing. Based on the definition of crowdsourcing provided by Howe (2006), numerous scholars have been interested in the new concept of this data collection method. Different definitions have emerged based on their understanding of crowdsourcing. According to Brabham (2008), crowdsourcing is "a strategic model to attract an interested, motivated crowd of individuals capable of providing solutions superior in quality and quantity to those that even traditional forms of business can." Later in Brabham's (2013) book, crowdsourcing was defined as "an online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals." Vukovic (2009) defined crowdsourcing as "a new online distributed problem-solving and production model in which networked people collaborate to complete a task." Instead of interpreting crowdsourcing as a model that solves the problems of the crowd through an online platform, Chanal and CaronFasan (2008) defined crowdsourcing as "the opening of the innovation process of a firm to integrate numerous and disseminated outside competencies through web facilities." Kleeman et al. (2008) found the essence of crowdsourcing to be intentional mobilization and defined crowdsourcing as "a form of the integration of users or consumers in internal processes of value creation." To explain it simply, La Vecchia and Cisternino (2010) described crowdsourcing as "a tool for addressing problems in organizations and business."

With the development of crowdsourcing, researchers have analyzed various existing definitions of crowdsourcing to discern the basic features and common elements. Estellés-Arolas and González-Ladrón-De-Guevara (2012) reviewed and summarized the research studies on crowdsourcing in terms of the information about the crowd and crowdsourcer, the tasks that need to be conducted by the crowd, the benefit for the crowd and crowdsourcer, and the process of crowdsourcing. An integrated definition of crowdsourcing based on the critical elements extracted from the previous literature was created which defined crowdsourcing as "a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task" (Arolas and González-Ladrón-De-Guevara, 2012). Other analyses and summaries of crowdsourcing can be found in Świeszczak and Świeszczak, 2016; Estellés-Arolas, Navarro-Giner, & González-Ladrón-de-Guevara, 2015; and Hosseini et al., 2014.

To summarize, most of the crowdsourcing definitions contain three main features which are the crowd itself, the outsourcing procedure, and an internet-based platform (Saxton, 2013). It means that crowdsourcing implies that individuals participate voluntarily to achieve the task which would tend to motivate both the experts and the individuals to find solutions to the tasks (Schenk, 2011). Table 2-1 presents a summary of the existing crowdsourcing definitions in chronological order.

Table 2-1 Summary of Existing Crowdsourcing Definitions

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

The development of crowdsourcing has brought improvements and benefits in data collection. This type of innovative data collection shows its potential to augment traditional data collection methods. Recently, Misra et al. (2014) studied the use of crowdsourcing in transportation. In addition, as the number of GPS-enabled smartphones increases, crowdsourcing with smartphones (Chatzimilioudis et al., 2012) will see more possibilities in transportation-related research studies. A comprehensive summary of the existing smartphone applications used for different aspects of transportation research areas is provided in the following section.

2.2.2 Open Data

Open data is another type of data that researchers can use for their studies. Open data is open for anyone to use freely, and to reuse or redistribute flexibly (Kitchin, 2014). In other words, an open data format should be "platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information" (Attard et al., 2015). Therefore, open data is available for anyone without any additional costs or limitations.

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

2.2.3 Big Data

Big data is a general type of data that refers to large volumes of data from various sources which need to be cleaned and pre-processed before being used for research studies (McAfee et al.,

2012). The main attributes of big data are the '3Vs' which are volume - representing the size of the data, velocity - indicating the speed of the data generation or collection, and variety - referring to a synthetic range of sources (Laney, 2001). Besides these three Vs, other researchers (Kitchin, 2014) have added other attributes to define big data including veracity, demonstrating the quality of the data.

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

2.2.4 Traditional Data Collection Methods

Traditional traffic data collection methods can provide accurate and useful information for relevant transportation research studies. The two most common traditional traffic data collection methods are traffic counting equipment and travel surveys.

Commonly used traffic counting equipment includes piezo-electric sensors, inductive loops, microwave, radar, video image detection, and manual observation, etc. (Skszek, 2001). Using the equipment to collect data may be more expensive and time-consuming compared to other methods.

Travel survey methods can be divided into two groups: web-based and paper-based. The most well-known type of travel survey is the household survey (Kagerbauer et al. 2015). Information relevant to household travel patterns is collected through questionnaires. The process of filling out paper-based surveys and selecting useful and suitable answers can be time-consuming. Web-based travel surveys, on the other hand, are used as an alternative that provides smart filter management features. However, bias and other issues associated with this type of travel survey cannot be ignored. One of the problems with data collected from travel surveys may come from the respondents. Since young participants can access the internet more easily compared to older respondents, the proportion of young respondents might be higher than older ones for web-based surveys. In addition, not all the questionnaires may be returned, as the receiving rate can be lower than travel surveys conducted in person. Other traditional transportation survey methods such as workplace surveys, longitudinal and panel surveys, transit onboard ridership surveys, commercial vehicle (truck) surveys, and external station surveys, usually have similar disadvantages.

Travel surveys can be categorized as stated preference surveys (i.e., SP surveys) and revealed preference surveys (i.e., RP surveys) (Guan, 2004). The SP survey receives the decision-making results of the respondents in terms of certain different conditions. RP surveys refer to the survey of completed selective behavior. The differences between these two kinds of travel surveys are: (1) the questions on an SP survey usually contain investigation content that has not occurred yet or is intentionally designed for specific research topics, while RP surveys contain investigation questions about behaviors that have already taken place; and (2) the scenario in an SP survey can be designed flexibly with assumed values of choices and attributes that are needed for the

research studies, while the results of choices and choice conditions in the RP survey are based on actual travel choice behavior. When considering these features of the two types of surveys, their advantages are revealed. With SP surveys, researchers can arbitrarily design the questionnaires and the corresponding scenarios for future conditions which will benefit transportation planning and design, especially for upcoming constructions and establishments. With RP surveys, researchers can show the results or phenomenon hidden in each individual's choice which reflects the contribution of the impact factors and how individuals value these factors.

Figure 2-1 below shows a clear structure of the traditional data collection methods mentioned in this section.

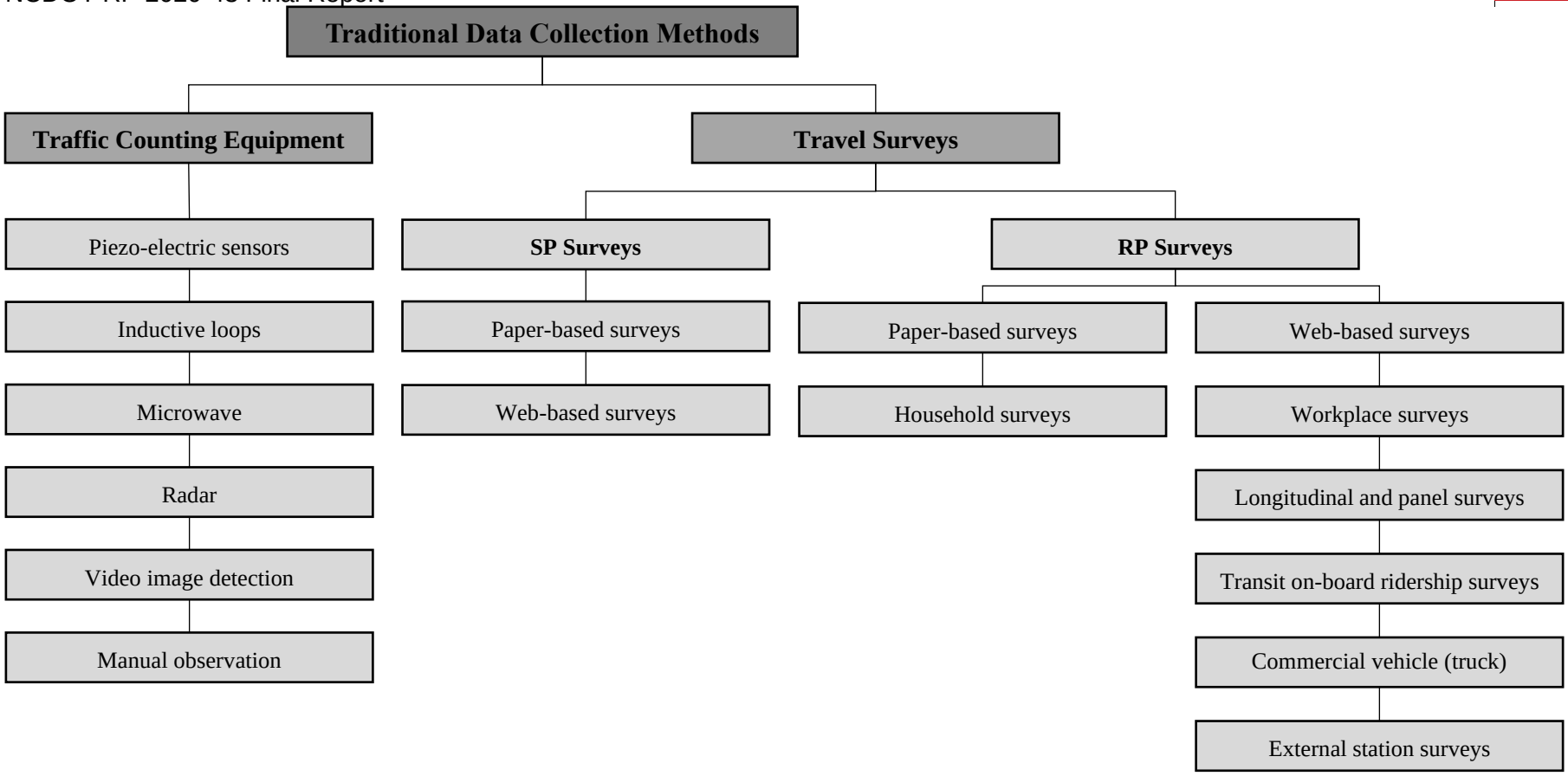

Figure 2-1 Traditional Data Collection Methods

2.3 Smartphone Crowdsourcing Applications

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

2.3.1 CycleTracks

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

Most of the studies used CycleTracks to analyze cyclists' route choice behavior. Charlton et al. (2011) collected the cycling data from CycleTracks from November 12, 2009, to April 18, 2010, to analyze these cyclists' route choices in San Francisco. A total of 7,096 cycling trips generated by 1,083 cyclists were collected and selected as the chosen routes in the modeling procedure. A doubly-stochastic choice set generation method was used for modeling cyclists' route choice. The impacts of the length of the route, turns per mile, the proportion of the route on wrong-way links, proportion on bike paths, bike lanes, and bike routes, infrequent cyclists, and average up-slope were considered in the path size multinomial logit model (Hood et al., 2011). Results revealed that the length of the route, turns per mile, the proportion of route on wrong-way links, and average up-slope affected the cyclist's route choice negatively, while other explanatory variables had positive impacts on route choice.

Chen and Shen (2016) collected data from CycleTracks to analyze the effects of land use and roadway characteristics on cyclists' route choices. The labeling route approach and the K-shortest mean approach were used to generate the route choice set for cyclist route choice analysis. A path size logit model was developed, and results were concluded that cyclists selected their cycling routes based on the consideration of utility, cycling safety, and suitability. Subsequently, Chen et al. (2018) examined the influences of the built environment on cyclist route choice using the same dataset. Another comprehensive discrete choice model (i.e., path-size-based mixed logit model) was developed for this research study.

As the first application developed for cycling studies, researchers have compared this dataset with other data sources including traditional count data and data collected from other smartphone

applications. Griffin and Jiao (2019) compared CycleTracks with traditional count data from five selected locations in Austin, Texas, and data provided by Strava fitness application. The relationship between CycleTracks and count data, as well as the relationship between Strava and count data were examined utilizing ordinary least squares regression. Additional spatial autocorrelation was also evaluated using OpenGeoDa software.

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

2.3.2 AggieTracks

AggieTracks was developed by Texas A&M University based on the open-source code of CycleTracks to collect cycling information on users in the university area (Hudson et al. 2012). Data on travel purposes were collected from cyclists who completed optional questions after their cycling trips. Cyclists using AggieTracks also entered classification status e.g., student, faculty, or staff. Additional information such as users' living locations (on or off campujs) and car ownership was also collected. Since this application was developed to track cycling patterns within the university area, few research studies choose to utilize this data source.

2.3.3 Cycle Atlanta

Like AggieTracks, Cycle Atlanta was developed based on the open-source code of the CycleTracks application by a research team from Georgia Tech collaborating with the City of Atlanta and the Atlanta Regional Commission (Figliozzi and Blanc, 2015). In addition to the cycling route data, Cycle Atlanta can collect demographic information including age, email, gender, ethnicity, home income, home zip code, work zip code, or school zip code, etc., Cycle Atlanta also notes selections including issues (e.g., pavement issue, traffic signal issue, bicycle lane design issue, enforcement request, bicycle parking request, and custom entry) and amenities (e.g., water fountain, public restroom, bicycle shop, bicycle parking, and custom entry).

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

Cycle Atlanta data were used for route choice modeling, street segment choice, bicycle level of service (BLOS), and level of traffic stress (LTS). In a USDOT final report named "Using Crowdsourcing to Prioritize Bicycle Route Network Improvements", LaMondia and Watkins (2017) conducted a comprehensive research study on calculating BLOS, measuring LTS, modeling bicyclist route choice, and route segment choice using data collected from three

smartphone applications: Strava, CycleDixie and Cycle Atlanta. An ordinal logistic regression model was developed to analyze the route segment choice of Cycle Atlanta users. Explanatory variables including roadway characteristics, access groups, and socio-demographic accessibility were included in the model. To analyze the willingness of a cyclist to choose a detour over the shortest route, a binary logistic choice model was developed based on the alternative (i.e., shortest route path) generated by the A-star algorithm.

Misra and Watkins (2018) investigated the differences in bicyclist route choice between different genders and age groups. Multiple path size logit models were developed for different segments in terms of age and gender. A pooled path size logit model based on the entire data set collected from Cycle Atlanta was developed for comparison. Results revealed that traffic characteristics including speed and annual average daily traffic (AADT) might affect the cyclist's route choice differently for male and female cyclists as well as young and old cyclists.

2.3.4 Cycle Lane

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

Zimmermann et al. (2017) modeled the bicyclist route choice based on the data collected from Cycle Lane to analyze the trade-offs cyclists make while selecting their cycling routes. The researchers applied a recursive logit model for the bicyclist route choice modeling since this type of link-based route choice model does not require generating route choice sets, compared to the path-based models such as the path size logit model. According to the results, this recursive logit model may save computational time.

2.3.5 Mon RésoVélo

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

Based on the GPS cycling trip data from Mon RésoVélo, deceleration rates at intersections and on-road segments were extracted by Strauss et al. (2017) to investigate the relationship between the deceleration rate and the number of injuries. The ranking of sites based on the deceleration rate and the expected injury number were compared using Spearman's rank correlation coefficient.
With the benefit of this innovative smartphone application, many other research studies were conducted based on the data extracted from Mon RésoVélo. Strauss and Miranda-Moreno (2017)

used the GPS cycling trip data from the Mon RésoVélo application to identify performance measures in terms of speed, delay, and travel time at both intersections and on-road segments in the whole city network on the island of Montreal. To examine the impacts of geometric design and built environment on cycling speed on each road segment, a linear regression model was developed. The model results demonstrated that cycling speeds were higher along arterials than on local streets, and cyclists biked faster on road segments with bicycle infrastructure. Furthermore, impact factors including geometry characteristics, built environment features, travel purposes, and peak hours were found to affect cycling speed significantly.

2.3.6 RenoTracks

RenoTracks is a smartphone application that builds on the Cycle Atlanta application. This application offers similar functions to CycleTracks, including recording cycling information, collecting travel distance, calculating travel speed, reporting issues, and collecting demographic data from cyclists. (RenoTracks 2013). In addition to the features adopted from previous smartphone crowdsourcing applications, RenoTracks added a customized user interface and a "CO_2 Saved" counter to calculate the carbon dioxide that would be saved compared to traveling by automobile.

2.3.7 ORcycle

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

With the help of ORcycle, useful data can be collected to design and upgrade bicycle facilities and analyze the impacts on cyclists' comfort levels. Blanc and Figliozzi (2016) leveraged the ORcycle application to collect data for cyclists' comfort level modeling. Factors including bicycle facilities, sources of stress associated with the cycling routes, travel purposes, distance, cycling frequency, and temporal characteristics were considered in the model. Ordinal logistic regression models were developed to estimate the influence on cyclists' comfort levels. Based on the model results, bicycle boulevards, separated cycling paths, sources of stress associated with the cycling routes, trip purposes, and cycling distance were found to affect cyclists' comfort levels significantly.

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

2.3.8 MapMyRide

MapMyRide is one of the smartphone applications developed by MapMyFitness to get the most from the users' bike rides and track their cycling trips, especially for recreational travel purposes. This application allows users worldwide to plan their cycling route, track their GPS trajectories,

share links with others, and provide users with information. Cyclists using MapMyRide can view others' cycling routes to follow popular cycling routes for comfortable and challenging activities. In addition to the smartphone application, MapMyRide also provides a web version that can present and summarize the statistics and ridership of the users' cycling trips (Figliozzi and Blanc, 2015).

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

2.3.9 Strava

Similar to MapMyRide, Strava allows users to track their cycling routes through the GPS-enabled smartphone and view and share the trip trajectories afterward via a website or application. Summary statistics including travel speed, trip distance, activity time, and another cycling route information are provided and displayed. Strava's unique features are the ability to track the cycling performance of multiple cyclists on the same segment which enables Strava users to compete with each other for the least segment time, highest speed, etc. This particular functionality attracts numerous cyclists worldwide to use this smartphone application for recording their cycling trips which provides Strava a large dataset in extensive geographic and temporal scales.

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

Sun and Mobasheri (2017) used Strava data to analyze the spatial patterns of cycling activities for different travel purposes and air pollution exposure on a large scale. The improved Multidirectional Optimum Ecotope-Based algorithm was used to identify the clusters associated with a high non-commuting rate. Ordinary least squares, multilayer perceptron neutral network, random forest, and support vector machine methods were used to analyze the Strava users' non-commuting cycling activities. Results showed that more non-commuting cycling trips occurred on the outskirts of the city. In addition, cyclists biking for commuting were found to be more likely to be exposed to higher levels of air pollution.

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

Table 2-2 and Table 2-3 summarize the literature reviewed in this section related to the crowdsourcing applications developed for collecting cycling information and research studies based on the data extracted from smartphone applications.

Table 2-2 Summary of Smartphone Crowdsourcing Applications

Year	Applications	Developer	Information Collected	Emphasis
2008	MapMyRide	MapMyFitness	Demographic information Travel purpose Cycling trajectories	Physical activity patterns analysis
2009	CycleTracks	SFCTA	Demographic information Travel purpose Cycling trajectories	Route choice modeling
2009	Strava	Strava Metro	Demographic information Travel purpose Cycling counts	Non-motorized transport planning Air pollution exposure Cycling patterns and behavior Bicycle volume Active travel and health
2011	AggieTracks	Texas A&M University	Trip purpose On campus living Car ownership	Travel patterns analysis
2011	Cycle Lane	CLMPO	Demographic information Travel purpose Cycling trajectories	Route choice modeling Bicyclist behavior analysis
2012	Cycle Atlanta	Georgia Tech	Demographic information Travel purpose Issue reporting Amenity reporting	Route choice modeling Bicycle volume LTS
2013	Mon RésoVélo	The City of Montreal	Demographic information Travel purpose Cycling trajectories Calorie Emissions	Level of service measures Safety analysis
2013	RenoTracks	2013 Hack4Reno Team	Demographic information Travel purpose Cycling trajectories "CO_2 Saved"	Cycling data analysis
2014	ORcycle	Portland State University and ODOT	Demographic information Travel purpose Cycling trajectories Infrastructure issues Crashes	Cyclists' comfort level Route choice modeling Crash and injury risk modeling

Table 2-3 Summary of Research Topics Based on Crowdsourced Bicycle Data

Year	Author	Data Source	Study Area	Data Size	Methods	Research Area
2011	Hood et al.	CycleTracks	San Francisco	7,096 cycling trips generated by 1,083 cyclists	Path size multinomial logit model	Route choice modeling
2014	Hirsch et al.	MapMyFitness	Winston-Salem, NC	43,872 unique workouts by 3,094 unique users	Statistical analyses	Physical activity patterns analysis
2016	Blanc and Figliozzi	ORcycle	Portland, OR	729 trips from 170 users	Ordinal logistic regression models	Cyclists' comfort levels
2016	Chen and Shen	CycleTracks	Seattle	543 observations	Path size logit model	Route choice modeling
2017	LaMondia and Watkins	Cycle Atlanta Strava CycleDixie	Auburn, AL Atlanta, GA	5,201 trips generated by 458 cyclists	Ordinal logistic regression model, Binary logistic choice model	Route segment and path choice modeling
2017	Strauss et al.	Mon RésoVélo	Montreal	Over 10,000 trips recorded by almost 1,000 cyclists	Spearman's rank correlation coefficient	Safety measure
2017	Strauss and Miranda-Moreno	Mon RésoVélo	Montreal	Over 10,000 trips recorded by almost 1,000 cyclists	Linear regression model	Performance measures
2017	Sun and Mobasheri	Strava Scottish Air Quality Database	Glasgow, United Kingdom	287,833 cycling activities contributed by 13,684 users	Ordinary least squares, multilayer perceptron neutral network, random forest, support vector machine	Cycling activities and air pollution exposure
2017	Zimmermann et al.	Cycle Lane	Eugene	648 bike trips from 103 users	Recursive logit model	Route choice modeling
2018	Chen et al.	CycleTracks	Seattle	3,310 routes created by 197 cyclists	Path-size-based mixed logit model	Route choice modeling
2018	Misra and Watkins	Cycle Atlanta	Atlanta, GA	About 20,000 trips by 1,495 users	Path size logit models	Route choice modeling
2019	Griffin and Jiao	Traditional Count Data, CycleTracks, Strava	Austin, Texas	183,070 counts and 111 CycleTracks records, 4,372 counts and 209 Strava records	Ordinary least squares regression and spatial autocorrelation	Bicycle volume

2.4 Bicycle Volume

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

2.4.1 Research Based on Traditional Data Collection Methods

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

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

Chen et al. (2017) investigated the impacts of built environment explanatory variables on bicycle volume with a dataset of five-year bicycle volume in Seattle, Washington. A generalized linear mixed model (GLMM) assumed to follow a Poisson distribution was developed to model the variation of bicycle volume over time. Model results indicated that exploratory variables, including non-winter seasons, temporal characteristics such as weekends and peak hours, bicycle facilities, and employment density, were likely to affect bicycle volume positively. Lower bicycle volume was associated with steep areas, while areas with more mixed land use, water bodies, and workplaces were found to be high bicycle volume locations.

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

Esawey (2014) conducted a research study on estimating the annual average daily bicycle (AADB) with both daily adjustment factors (DFs) and monthly adjustment factors (MFs). Bicycle volume data collected from 12 permanent counting stations in the City of Vancouver were used for adjustment factor calculation. Subsequently, the calculated factors were used to estimate AADB at other counting stations. Later, Esawey and Mosa (2015) developed two variations of the standard K factors, which is another type of adjustment factor used for bicycle volume estimation and calculation, and provided an example of AADB calculation using the

developed standard K factors. (i.e., Kp/d and Kp/AADB). They also provided an example of AADB calculation using the developed standard K factors. The estimation accuracy based on the two variations of the K factors was also examined.

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

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

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

To conclude, a summary of the studies on bicycle volume estimation and analysis using traditional manual count data or automated count data is provided below in Table 2-4.

Table 2-4 Summary of Bicycle Volume Studies Using Bicycle Count Data

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

[1] Seattle Department of Transportation

2.4.2 Research Based on Crowdsourcing

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

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

Griffin and Jiao (2019) collected data from both the CycleTracks smartphone application and the Strava fitness application to conduct a data comparison between crowdsourced bicycle data and manual count bicycle data. Five specific locations were selected in downtown Austin, Texas. All the data were compiled and compared in GIS for these five locations.

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

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

Hochmair et al. (2019) used the crowdsourced bicycle data collected from the Strava application in the Miami-Dade County area to analyze the impact of demographic information, network characteristics (especially bicycle facilities), and place-specific features on bicycle ridership. A series of linear regression models were developed to predict the bicycle kilometers traveled for both commuting and non-commuting trips, and trips that occurred on both weekdays and weekends. Eigenvector spatial filtering was adopted to avoid bias and model spatial autocorrelation. Results showed that Strava data performs well for the analysis of the impact of explanatory variables on bicycle volumes for commuting and non-commuting trips and during different days of the week. In addition, Strava data revealed the

broad coverage of spatial and temporal information and that they can be used as a critical supplement to bicycle volume estimation in large areas.

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

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

To conclude, a summary of the studies on bicycle volume estimation and prediction as well as cycling activity analysis is provided below in Table 2-5.

Table 2-5 Summary of Bicycle Volume Research Based on Crowdsourced Data

Year	Author	Bicycle Data	Methods	Results
2015	Moore	Data from Strava application	Ordinal logistic regression model	Roadway characteristics and surrounding land-use have a significant impact on whether or not a particular street segment would be used.
2016	Jestico et al.	Data from Strava and manual counting data	Generalized linear model	In mid-size North American cities within urban areas, the routes recorded in crowdsourced fitness application tend to be similar with those of the commuter cyclists.
2016	Watkins et al.	Data from Cycle Atlanta, Strava, and actual cyclist trips	Data comparison	The smartphone application data should be carefully used considering the likely bias.
2017	LaMondia and Watkins	Data collected using the Strava, Cycle Dixie and Cycle Atlanta	Route suitability score and preference models	Demographics, roadway characteristics and surrounding land-use have a significant impact on route choice.
2017	Proulx and Pozdnukhov	Crowdsourced data from Strava and usage data from Bay Area Bikeshare	Geographically Weighted Data Fusion	The method of Geographically Weighted Data Fusion can improve predictive accuracy for link-level bicycle volume estimation.
2019	Griffin and Jiao	Data from CycleTracks, Strava application, and traffic counts	Ordinary least squares regression	Crowdsourced data are appropriate for bicycle volume evaluation.
2019	Hochmair et al.	Data from Strava application	Linear regression models	Strava data can be used to examine the impact of explanatory variables on estimated bicycle volume.

2.5 Counting Machine Validation and Correction Methods

2.5.1 Introduction

The North Carolina Non-Motorized Volume Data Program (NC NMVDP) was established in 2013 to test a regional bicycle and pedestrian count data collection protocol and to determine how to replicate the methodology across the state of North Carolina (ITRE, 2016). The Institute for Transportation Research and Education (ITRE) worked in partnership with NCDOT, local agencies (municipalities and regional planning agencies), and the technology vendor Eco-Counter to install and maintain a permanent sensor network on sidewalks, bike lanes, shared lanes, and shared-use paths. The purpose of these permanent counters was to establish statistically valid expansion factor groups, measure existing trends, and model future increases in non-motorized volumes at the site, corridor, and regional levels.

Eco-Counter MULTI systems were installed on representative facilities to continuously monitor bicyclist and pedestrian activity over time. The counting systems were composed of inductive loops to detect bicyclists and passive infrared sensors to detect pedestrians. ITRE's validation protocol is meant to account for the systematic error (i.e., over or undercounting) associated with the use of a particular counting technology by calculating correction factors to adjust raw counts closer to the ground truth. Video cameras were installed to record activity at a counting location when a counting system is first installed, when a sensitivity setting is changed, or when any part associated with the counting system is changed (loops, sensors, logger, smart-connect, or y-connect). Historically, analysts review two days of video for each site to produce a manual count of pedestrian and bicycle users who pass through the sensors' detection zones. Counts by mode are aggregated into 15-minute bins and related to the counts recorded by the Eco-Counter for the same period reviewed by the analysts to produce a correction factor. The correction factor is applied as a linear multiplicative factor to hourly count totals recorded by a sensor. Corrected hourly counts are then rounded to the nearest whole number and reported as the final adjusted hourly count for a particular sensor. Development of the historic NC NMVDP correction factor calculation methodology is outlined in the NC NMVDP Phase 1 Final Report (ITRE, 2016).

This research project investigates several research questions related to the historic correction factor calculations produced in the NC NMVDP. These include:

- Impacts of rounding on corrected count data;
- Appropriate temporal aggregations for applying linear correction factors;
- Minimum number of non-zero observations required to properly calibrate an Eco-Counter system;
- Differences between the magnitude of correction factors as calculated using historic programmatic processes and those calculated with linear regression methods; and
- Proposed methodologies for accounting for accuracy, tolerance, and uncertainty in count data recorded by an Eco-Counter system.

2.5.2 Estimating Errors and Uncertainty

Accuracy, tolerance, and uncertainty are essential concepts in determining the error of any measurement. While accuracy and tolerance standards are common in traffic monitoring, uncertainty measures are less available. The Federal Highway Administration (FHWA) states that “calibration/validation procedures (even if conducted on a limited scale) should be used to ensure that (non-motorized) count data is within the bounds of acceptable accuracy" (Jessberger, 2017). Calibration of motorized traffic monitoring devices is recommended to be performed “at a minimum of every year as detailed in the FWHA TMG (Traffic Monitoring Guide)” (FHWA, 2014), and field counts should be within 5-10% of the true count (FHWA, 2014).

While FHWA provides guidance on best practices determining and ensuring the accuracy of motorized traffic monitoring devices, specific calibration protocols are typically set by state DOTs and are variable. State practices vary from no calibration, following manufacturer suggestions for calibration, to multiple calibration studies performed on equipment each year (Jessberger, 2017). For example, the Pennsylvania Department of Transportation (PennDOT) conducts a quality assurance test at least once every three years on their Automatic Traffic Recorders (ATRs) to determine if the percent error of their tests is within +/- 2%. Jessberger (2017) notes that “traffic monitoring program data quality requirements hinder some interagency departments from sharing traffic monitoring sites, installation, equipment, and data integration challenges due to technology configuration, cultural and institutional coordination” (Jessberger, 2017). The traffic counting industry is currently working to establish methods for achieving standard accuracies and associated confidence intervals for federally mandated traffic monitoring data (Jessberger, 2017).

2.5.3 Defining Accuracy, Tolerance, and Uncertainty

Accuracy is defined as the closeness between the measured value and the true value. For the NC NMVDP, this relates to the bicyclists and pedestrians detected by an Eco-Counter device and the bicyclists and pedestrians observed by an analyst in video recordings during a validation study. Tolerance is defined as the maximum range of error that is acceptable in a device. A standard tolerance for bicycle and pedestrian counting technologies has not been established. The NC NMVDP defines the maximum range of error for counts recorded by Eco-Counter devices as +/- 40% of the matched manual counts, which is much higher than the typical motorized traffic monitoring tolerance. Uncertainty is a non-negative parameter that characterizes the dispersion of values attributed to a measurement. Uncertainty parameters are typically derived from a standardized figure unique to the type and model of the measuring device. However, standardized uncertainty measures for bicycle and pedestrian counting technologies do not currently exist.

Di Leo et al. (2007) suggest that the baseline data source for uncertainty in traffic monitoring is a manual observation of traffic flow. They suggest the following methodology for calculating the uncertainty in traffic monitoring:

“Following a type A approach, at first deterministic errors must be estimated (i.e., through a comparison with a reference instrument), to provide a correction “c” with its uncertainty u_c; then the standard deviation σ_m of a set of measurements, carried out keeping constant the measurand

and all the controllable influence parameters, and corrected by "c," allows the uncertainty u to be calculated as:

$$u^2 = \sigma_m^2 + u_c^2 \qquad \text{Eq. 2-1}$$

where u_c is the difference between the corrected sensor data and the observed data, σ_m is standard deviation of measurements, and u is the uncertainty value."

While uncertainty measures were not incorporated into previous correction factor studies for the NC NMVDP, count data from historic validation studies can be used to develop uncertainty measures for all counting periods within a given validation study. NCDOT does not generate statistical measures of uncertainty for counting devices on a regular basis as it would require a large commitment of resources.

2.6 Bicyclist Injury Risk Analysis

Bicyclist injury risk analysis is another critical research topic that needs to be studied. This would enable researchers and relevant agencies to better understand the impact factors contributing to high injury risk and consequently help provide a greener and safer cycling environment and promote biking in large bicycle-friendly cities.

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

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

A research study was conducted by Strauss et al. (2014) to analyze multimodal injury risk including motor-vehicle, pedestrian, and bicyclist injury risk and activities for both signalized and non-signalized intersections. Like the previous research, a Bayesian modeling approach was used for safety and volume analysis simultaneously based on the same dataset, along with the injury and volume data collected from 435 more non-signalized intersections. Afterward, the Bayesian multivariate Poisson models were calibrated and the explanatory variables contributing

to injury frequency were determined. A comparison of injury risk for different modes for both intersection types was conducted. Results showed that motor-vehicle traffic is the primary cause of all multimodal injuries for both signalized and non-signalized intersections. In addition, bicyclists and pedestrians have a much higher injury risk on average compared to motorists at signalized intersections. Factors including some geometric design and built environment characteristics were found to have a significant impact on injury risk for all three kinds of road users.

Furthermore, with the development of crowdsourcing, smartphone GPS data collected from numerous applications were used for bicycle volume as well as bicyclist injury risk analysis. Strauss et al. (2015) proposed a method to estimate bicycle volume and map ridership and bicyclist injury risk in the whole city network in Montreal for both roadway segments and intersections based on data collected from Mon RésoVélo smartphone application as well as the manual count data. An extrapolation function approach was applied to combine the manual count bicycle data with crowdsourced bicycle data for bicycle volume estimation. Then, safety performance functions (SPFs) were developed based on the estimated AADB to validate the predicted AADB by comparing the parameter coefficients with the previous SPFs using manual count data. After calibration, the AADB function can be applied to predict bicycle volume on all the road segments and intersections within the city network. Statistical models were used to compute empirical Bayes (EB) for bicyclist injury risk analysis. Injury risk maps can be generated to illustrate the distribution of bicyclist injuries. According to the results, more injuries and higher injury risk occurred at signalized intersections compared to non-signalized intersections. On average, more injuries occurred on segments with cycle tracks, yet the injury risk per bicyclist was lower because of the presence of cycle tracks.

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

Similarly, Saad et al. (2019) estimated safety performance functions for bicyclist injury risk analysis at intersections based on the crowdsourced bicycle data collected from the Strava application. Strava data were adjusted before being used as the input in safety performance functions. Models based on the original Strava data, the Strava data with field observation data adjustments, and Strava data with adjusted population were developed and compared. Negative binomial models were developed for bicycle crash prediction at intersections. The model estimation results demonstrated that the adjusted Strava data with both population and field observation perform best in bicyclist injury analysis. In addition, impact factors including signal control systems, bicycle lanes, and intersection size, etc. would affect bicyclist injury at intersections.

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

Another approach to injury risk factors other than developing SPFs using smartphone applications is to collect volunteered geographic information (VGI) from cyclists through websites or applications. von Stülpnagel and Krukar (2018) assessed this kind of crowdsourced data as well as the authoritative data as indicators for biking risk analysis. Bicyclists were recruited to voluntarily participate in laboratory-based virtual reality experiments to estimate their risk perception. Participants were divided into two groups for separate cases. The first group was made up of experienced and frequent bicyclists who are not familiar with the selected test locations. The second group was composed of bicyclists who are both experienced and familiar with the test locations. After the experiments were conducted, indicators of biking risk were extracted from the VGI. Based on the indicators from both VGI and collected authoritative data, biking risk perception was estimated using linear mixed-effect models. The model results revealed that the semantic severity described for cycling hazard and the public response to the hazard might affect the risk perception significantly. Based on the authoritative data, a Space Syntax analysis was conducted which demonstrated bicyclist sensitivity to street size and complexity.

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

Al-Fuqaha et al. (2017) developed a smartphone application called BikeableRoute to analyze risk factors using crowdsourced data. BikeableRoute enables bicyclists to report hazards during their cycling trips as well as to track their cycling information. The data collected from this application included risk reports generated by bicyclists, user evaluation on the biking ability of cycling routes, and cycling information such as speed, cycling time, and distance. Based on the data from BikeableRoute, risk factors were categorized into three groups: infrastructure-related, facility-related, and traffic-related factors. An ordered probit model was developed to analyze the perception of narrow bicycle lanes in terms of different ages and skill levels. Results revealed that bicyclists' perceptions of hazards vary across different age groups.

Table 2-6 summarizes the research studies conducted based on manual count bicycle data and crowdsourced bicycle data from smartphone applications for bicyclist injury risk analysis.

Table 2-6 Summary of Research on Bicyclist Injury Risk Analysis

Year	Author	Research Objectives	Bicycle Data	Study Area	Methods	Variables
2013	Strauss et al.	Bicyclist activity and injury risk analysis	Manual bicycle counts and bicyclist injury data	Montreal, Quebec, Canada	Two-equation Bayesian modelling approach	(1) Bicyclist injury model: Bicycle volume, vehicle right turn and left turn flows, bus stops, crosswalk length, raised median; (2) Bicycle volume model: Employment, schools, metro stations, land use, bicycle facility length, three approaches.
2014	Strauss et al.	Multimodal injury risk analysis	Manual bicycle counts and bicyclist injury data	Montreal, Quebec, Canada	Bayesian multivariate Poisson models	(1) Bicyclist injury risk at signalized intersection: Bicycle volume, vehicle right turn and left turn volume, bus stops, crosswalk length, raised median; (2) Bicyclist injury risk at non-signalized intersection: Bicycle volume, vehicle volume, number of lanes.
2015	Strauss et al.	Bicyclist activity and injury risk analysis	Manual bicycle counts, smartphone GPS data, and bicyclist injury data	Montreal, Quebec, Canada	Extrapolation function and negative binomial SPF model	(1) AADB: Bicycle facilities (cycle path, cycle track, bicycle lane, etc.), distance to downtown; (2) Injury models for signalized intersections: Bicycle volume, bus stops, three approaches; (3) Injury models for non-signalized intersections: Bicycle volume, arterial or collector, three approaches; (4) Injury models for segments: Bicycle volume, arterial or collector, downtown boroughs.
2016	Jestico	Ridership trends and safety	Manual bicycle counts, Strava data, and incident reports from BikeMaps.org	The Capital Regional District (CRD), British Columbia (BC), Canada	Generalized linear model with a Poisson distribution	(1) Bicycle volume model: Strava counts, slope, population density, pavement widths, on-street parking, speed limit, bike facilities (e.g., painted bike lanes and paved multiuse trails), and month. (2) Incident model: Bicyclist and vehicle volume, speed reduction factors.
2016	Wang et al.	Bicycle safety analysis	Strava data and bicycle crash data	Seattle, Washington; Portland, Oregon	NBRM, PRM, zero-inflated negative binomial regression model, etc.	AADT & AADB

Year	Author	Research Objectives	Bicycle Data	Study Area	Methods	Variables
2017	Al-Fuqaha et al.	Non-motorized behavior analysis and risk factor identification	Crowdsourced data from BikeableRoute	Kalamazoo, Michigan	Web survey and order probit model	Bicyclist skill level, age, gender, bicycle facility (e.g., narrow bicycle lane).
2017	Chen	Crash frequency prediction	Strava data, automatic bicycle count data, and reported crash data	Portland, Oregon and Eugene-Springfield, Oregon	Poisson, NB, ZIP and ZINB models	(1) Crash frequency for intersections: Strava counts, AADT, network density, directions, bike lane, total lanes, signal, leg number. (2) Crash frequency for corridors: Signal/mile, median, two-way left turn lane, bus route number, on-street parking.
2018	von Stülpnagel and Krukar	Risk perception	Crowdsourced and authoritative data	Munich and Freiburg, Germany	Linear mixed-effect models and Space Syntax analysis	Semantic severity, number of votes, street size, traffic volume, complexity, accident category, familiarity.
2019	Saad et al.	Bicycle safety analysis	Strava data and bicycle crash data	Orange County, Florida	Negative binomial models	TEV[1], bicycle exposure, intersection size, signal control system, number of legs, bike lane, sidewalk width, median width, speed limit.

[1] Total entering volume

2.7 Summary

A comprehensive review and synthesis of the current and historical research studies related to different kinds of data collection methods including crowdsourcing, open data, big data, and other traditional data collection methods were presented in the first section of this chapter. Then, the most prevalent smartphone crowdsourcing applications and their use by relevant research studies were summarized. The methods that were applied by researchers to estimate and predict bicycle volume were provided and the methods for counting machine validation and correction were summarized. Then, existing studies on counting machine validation and correction methods were presented. Finally, bicyclist injury risk analyses conducted based on different types of data were discussed. This literature review is intended to provide solid reference and guidance in counting machine validation and correction, bicycle volume estimation and prediction, and injury risk analysis in future studies.

3 DATA DESCRIPTION

3.1 Introduction

The first step of this research was to collect crowdsourced bicycle data from the Strava application and other relevant supporting data. This chapter gives an overview of the collected Strava bicycle data and other essential supporting data for later model development. Data comparison was conducted between bicycle count data from permanent continuous count stations in the City of Charlotte and the Strava bicycle data collected from the smartphone application.

in this chapter, section 3.2 introduces the Strava data in terms of delivery contents, Strava users' demographic information, and cycling trip distribution. Section 3.3 summarizes the supporting data collected for model development. Section 3.4 compares the actual bicycle count data and Strava data. Finally, Section 3.5 concludes this chapter with a summary.

3.2 Strava Data

This study utilizes the crowdsourced data collected from Strava Metro to conduct research on bicycle volume estimation and prediction. The City of Charlotte in North Carolina was selected as the study area. This data contains the Strava users' cycling information on each road segment. A total of 140,428 cycling trips were recorded from 8,857 cyclists from December 2016 to November 2017, accounting for 1.03% of the total population in Charlotte in 2017.

The core data are the primary parts of the delivery provided by Strava Metro, which includes three main components: bicycle volume for link level, intersection level, and origin-destination matrix. In this project, the link-level data were used for further model development.

To provide options for researchers to leverage this innovative data, Strava also offers roll-up datasets. These datasets are the summary of bicycle volumes on each road segment/intersection during the time period of delivery. With these aggregated data, researchers can discover bicycle volumes for monthly use, weekday/weekend, on-season/off-season, hour groupings, and total bicycle counts. In addition, researchers can customize the roll-up datasets flexibly in terms of their needs.

In this project, the hour groupings are categorized as follows:

- Early AM hours: 00:00 to 05:59 (labeled as_0)
- AM peak hours: 06:00 to 08:59 (labeled as_1)
- Mid-day hours: 09:00 to 14:59 (labeled as_2)
- Peak afternoon hours: 15:00 to 17:59 (labeled as_3)
- Evening hours: 18:00 to 19:59 (labeled as_4)
- Late evening hours: 20:00 to 23:59 (labeled as_5)

The report file provided by Strava can be used to obtain the Strava users' demographic information. Based on the data in the report, the gender and age summary of Strava users recorded during the research period (from December 2016 to November 2017) can be seen in Figure 3-1, Figure 3-2, and Figure 3-3.

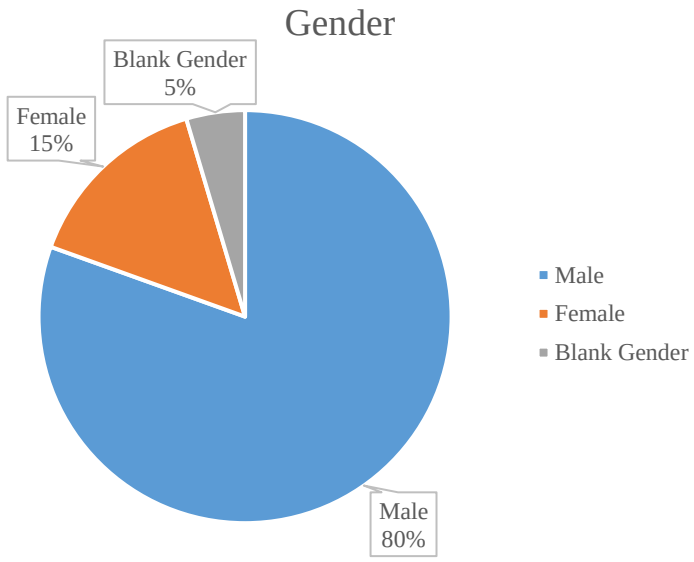

Figure 3-1 Gender Proportion of Strava Users'

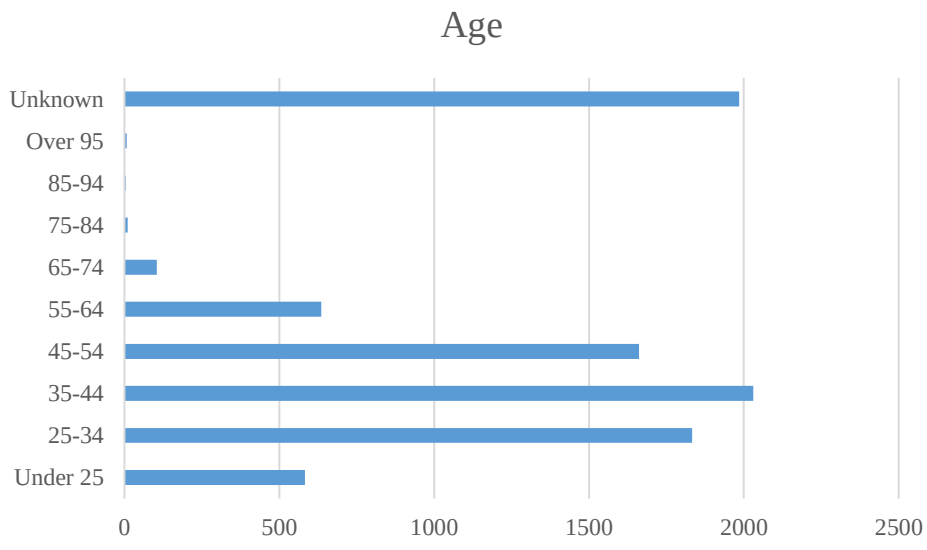

Figure 3-2 Number of Strava Users from Different Age Groups

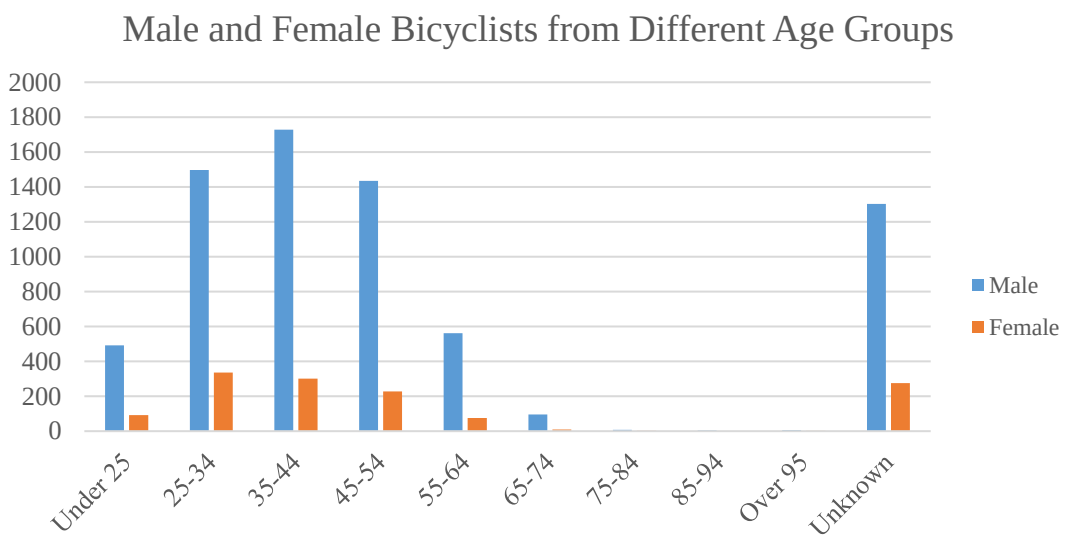

Figure 3-3 Number of Male and Female Strava Users from Different Age Groups

In addition to the demographic information, a map showing the Strava users' bicycle ridership was generated using ArcGIS with the availability of link-level Strava bicycle volume data. Figure 3-4 shows the total bicycle counts on each road segment in the City of Charlotte during the research period (from December 2016 to November 2017).

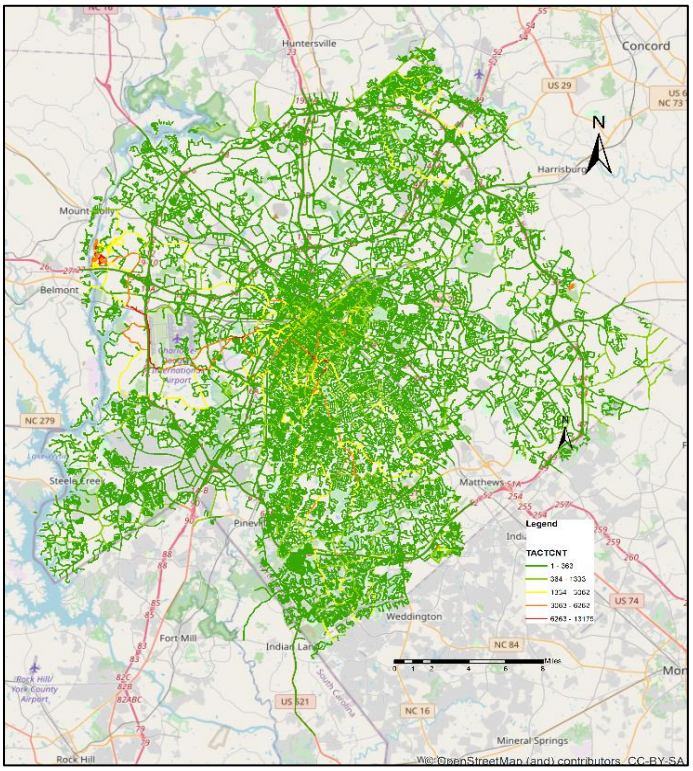

Figure 3-4 Total Bicycle Volume Distribution Map

Based on the trip purpose information provided by Strava data, 82% of cycling trips are non-commute trips, and 18% of cycling trips are commute trips. Two distribution maps illustrating the locations of commute and non-commute trips can be seen in the following figures. It can be seen in Figure 3-5 and Figure 3-6 that a higher number of commute trips occur in the uptown area, while a higher number of non-commute trips occur near parks and greenways.

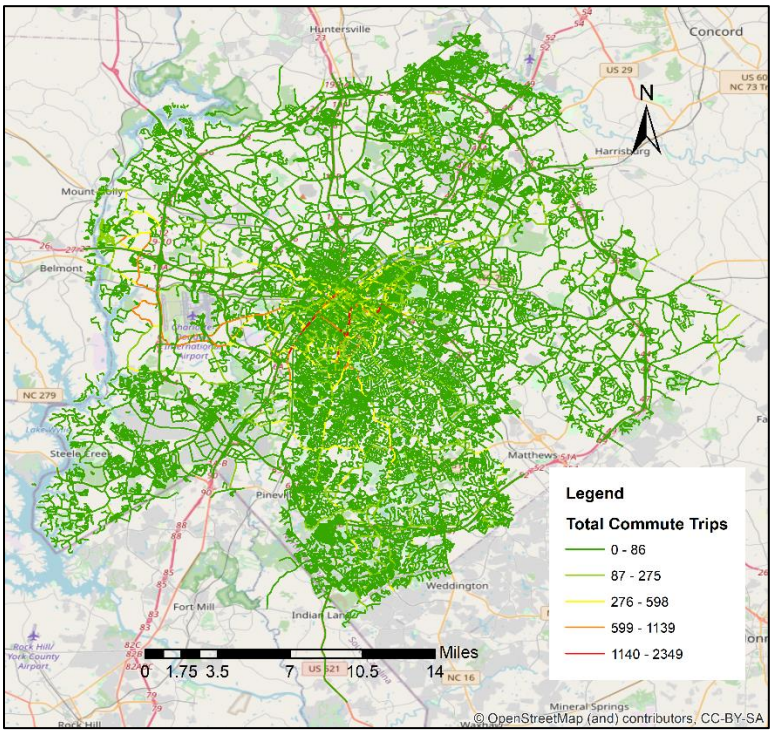

Figure 3-5 Total Commute Trips

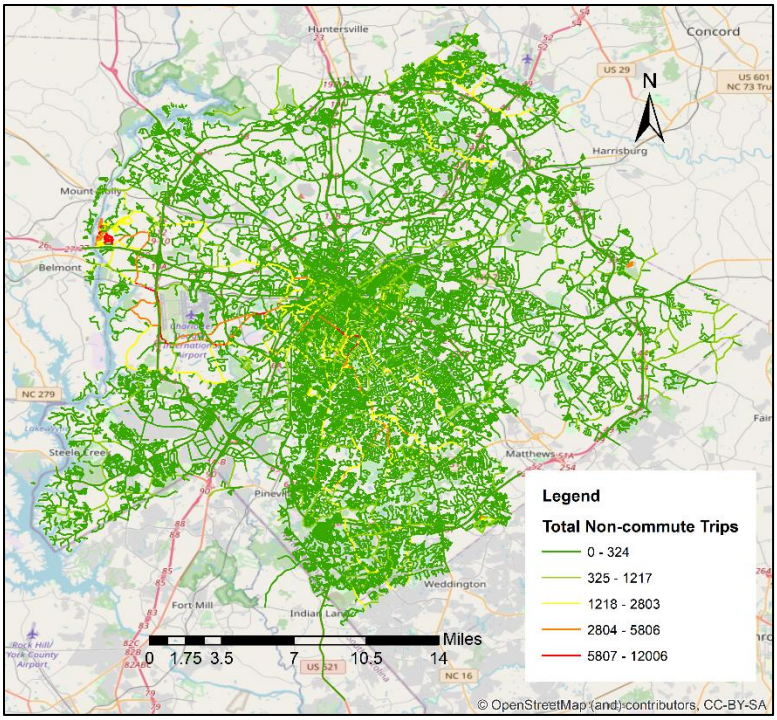

Figure 3-6 Total Non-commute Trips

3.3 Other Supporting Data

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

Figure 3-7 shows the locations of the six permanent continuous count stations in the North Carolina Non-Motorized Volume Data Program (NC NMVDP) that are in Charlotte and were evaluated for the analysis.

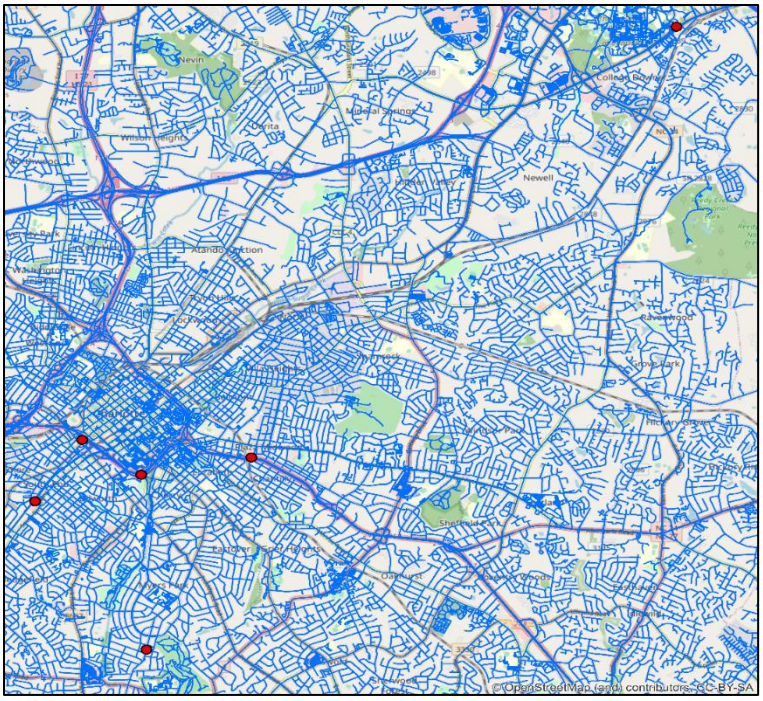

Figure 3-7 The Locations of the Continuous Count Stations

The crash data used in this project are bicyclist-involved crash data collected in the City of Charlotte from 2007 to 2017. The data were obtained from the North Carolina Department of Transportation. There are 1183 observations contained in the dataset with most of the bicycle-vehicle crashes (1149) occurring in urban areas. The distribution of bicycle-vehicle crashes in the City of Charlotte can be seen in Figure 3-8.

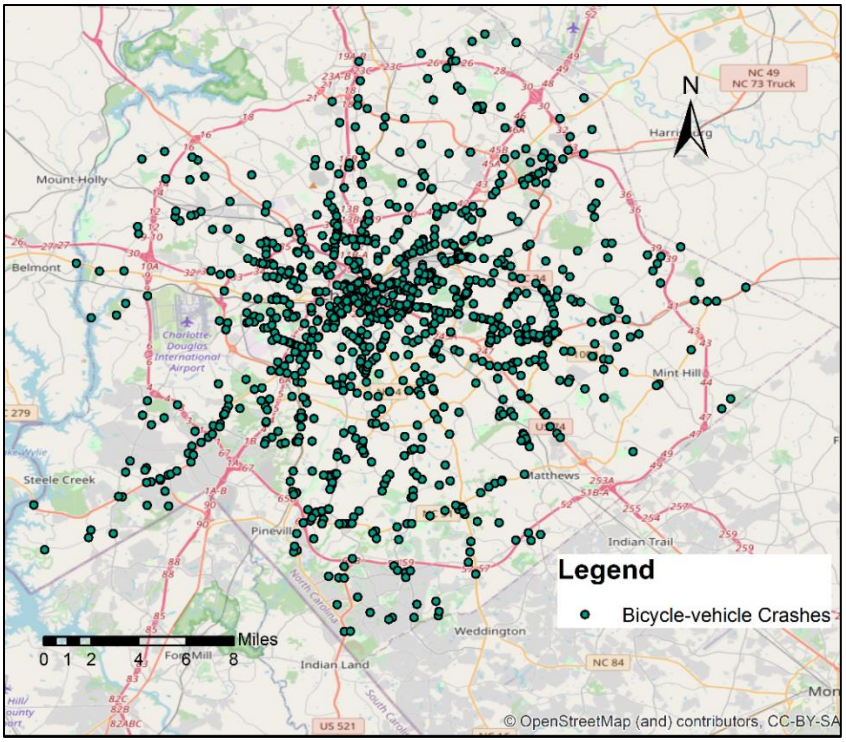

Figure 3-8 The Distribution of Bicycle-vehicle Crashes in the City of Charlotte

Figure 3-9 was generated to give a view of the crash numbers within each census block.

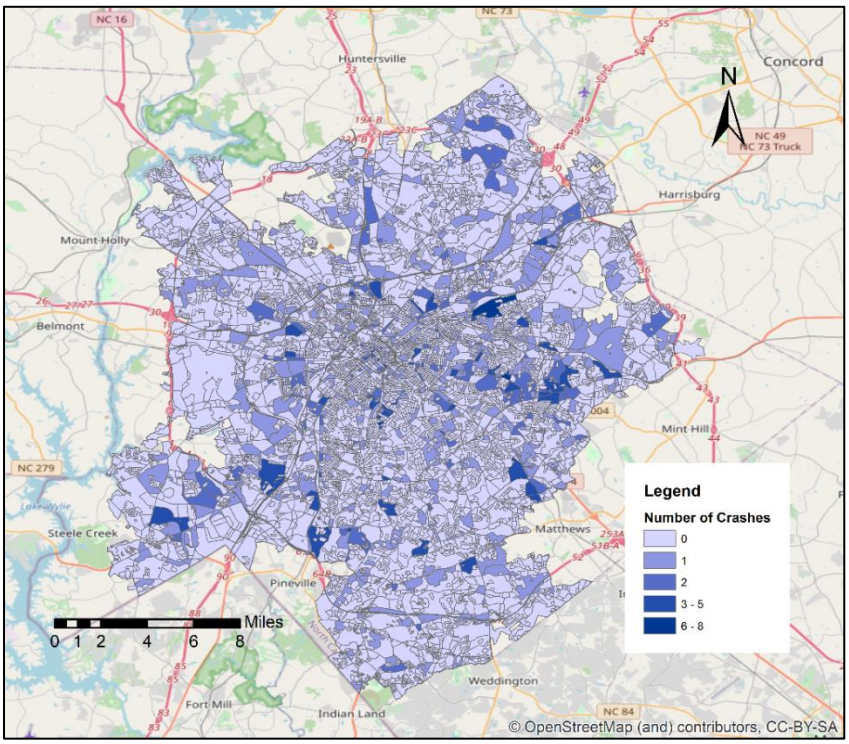

Figure 3-9 Number of Bicycle-vehicle Crashes within Census Blocks

3.4 Data Comparison

The difference remains between bicycle count data from the permanent continuous count stations and Strava data. Since crowdsourced data usually involves a large number of people, the coverage of the road segment that is being studied can be large. On the other hand, installing permanent continuous count stations can be costly and the geographic coverage limited since bicycle counts are collected at only a sample of locations. In addition, Strava data contains bicycle trip time and the trip purpose (commuting or recreation), while bicycle count data from permanent continuous count stations cannot collect such information. The bicycle counts from the different count stations and Strava user counts at the same locations are compared in Figure 3-10 which shows that the bicycle counts from count stations are greater than the Strava counts.

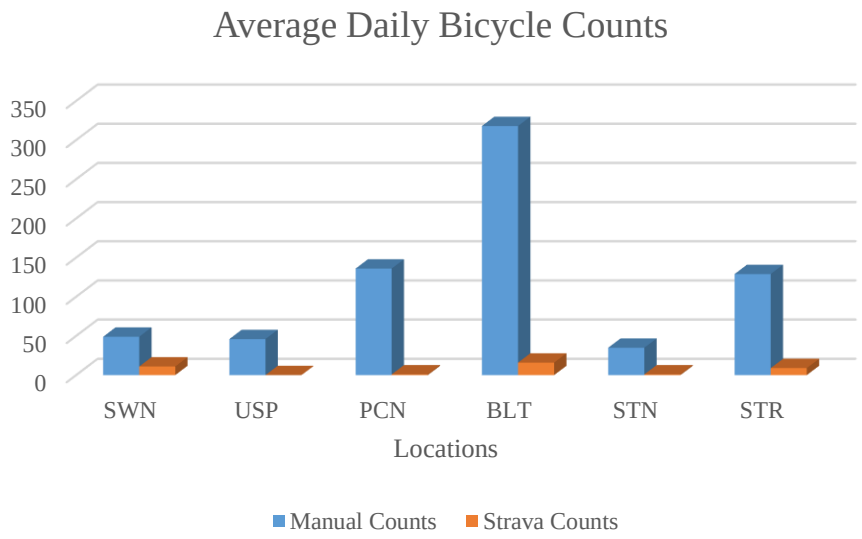

Figure 3-10 Comparison of Actual Bicycle Counts and Strava Counts

3.5 Summary

This chapter provides an overview of the data collected for this research. The descriptive analyses based on the data collected were conducted by creating several distribution maps for bicycle volume for different trip purposes. Data comparison between bicycle count data from permanent continuous count stations in the City of Charlotte and Strava bicycle data from the smartphone application is also provided.

4 BICYCLE VOLUME VALIDATION ANALYSIS

4.1 Introduction

This chapter introduces the validation and correction factor calculation methodology used in the North Carolina Non-Motorized Volume Data Program (NC NMVDP) to assess and adjust bicycle and pedestrian count data collected by permanent continuous counters across the state. Specifically, section 4.2 discusses the variation due to rounding, while section 4.3 introduces the correction factor model validation methodologies. Section 4.4 discusses the results. Finally, section 4.5 summarizes this chapter and provides relevant recommendations based on the research findings.

4.2 Variation Due to Rounding

The research team wanted to know whether significant differences in AADT estimations result based on how corrected counts are rounded. The distribution of volumes over the hours of the day may impact the magnitude that counts are adjusted using a correction factor. For example, suppose a validation study shows that the correction factor as derived from the historical method is 1.15. This signifies a 15% undercount error in the counts recorded by a sensor when compared with the manual counts. However, the spread of volume over the day impacts the daily volume estimate. When the total daily volume is evenly distributed across the hours of the day, rounding of the corrected hourly data results in a lower volume estimate than that if the total daily volume is concentrated in a single hour or a few hours of the day. In this example, an AADT estimate may be 14% different between one methodology and another (see Table 4-1).

Table 4-1 Correction Factor Application Variability Example

Time	Example 1 Sensor Data	Example 1 Corrected Data	Example 2 Sensor Data	Example 2 Manual Data
7:00	3	3 (3.45)	0	0
8:00	3	3 (3.45)	0	0
9:00	3	3 (3.45)	0	0
10:00	3	3 (3.45)	0	0
11:00	3	3 (3.45)	0	0
12:00	3	3 (3.45)	0	0
13:00	3	3 (3.45)	0	0
14:00	3	3 (3.45)	0	0
15:00	3	3 (3.45)	0	0
16:00	3	3 (3.45)	0	0
17:00	3	3 (3.45)	42	48(48.30)
18:00	3	3 (3.45)	0	0
19:00	3	3 (3.45)	0	0
20:00	3	3 (3.45)	0	0
Total	**42**	**42 (48.30)**	**42**	**48 (48.30)**

4.3 Correction Factor Model Validation

Three linear regression models were tested during Phase 1 (Pilot) of the NC NMVDP using counts from thirteen validation studies.

- Regression Model 1: Corrected Hourly Count = b1(AHC)
- Regression Model 2: Corrected Hourly Count = b1(AHC) + b2(AHC)^2-
- Regression Model 3: Corrected Hourly Count = b1(AHC) + b2(AHC)^2 + b3(AHC)^3

Where AHC stands for Automated Hourly Count. Based on the testing results, Model 1 was chosen because only minor improvements in model fit were achieved in Models 2 and 3. However, linear regression models with a constant term were not tested.

For the current study, linear regression models were tested using count data from the NC NMVDP validation studies that were performed between 2014 and 2019. Two linear regression models were calculated using raw counts that were aggregated into 15-minute and one-hour intervals. The linear regression models were then used to correct raw counts that were aggregated into 15-minute, one-hour, and daily intervals. AADT values were estimated from the corrected data. The linear regression models were also applied to AADT estimates calculated from raw counts (ITRE, 2016). Pearson's Correlation Coefficient values were calculated for each linear regression model. Weighted Average Percent Deviation (WAPD) was also calculated for all studies to calculate a reliable estimate of the accuracy of the sensors. WAPD is less sensitive to deviations in error due to low volume observations.

4.3.1 AADT Calculations

Impacts on count data due to rounding were measured by calculating the difference between AADT estimates where the correction factor is applied to the AADT value and AADT estimates where the correction factor is applied to other levels of temporal aggregation (daily, hourly, and 15-minute binned data) prior to calculating the AADT value. Methodologies for calculating AADT and correction factors are outlined in the next sections.

AASHTO AADT Methodology, no correction factors applied to counts

$$AADT = \frac{1}{12}\sum_{m=1}^{12}\left[\frac{1}{7}\sum_{j=1}^{7}\left(\frac{1}{n_{jm}}\sum_{i=1}^{n_{jm}} VOL_{ijm}\right)\right] \qquad \text{Eq. 4-1}$$

where:

n = number of days in a year (365 or 366)

VOL_{ijm} = daily volume for the ith occurrence of the jth day of the week within the mth month

i = occurrences of day j in month m for which traffic data are available

j = day of week (1 to 7)

m = month of year (1 to 12)

n_{jm} = number of occurrences of day j in month m for which traffic data are available

Linear Regression, applied to AADT

$$AADT = \sum_{p=1}^{s}\left[\beta_1 * \left(\frac{1}{12}\sum_{m=1}^{12}\left[\frac{1}{7}\sum_{j=1}^{7}\left(\frac{1}{n_{jm}}\sum_{i=1}^{n_{jm}} VOL_{ijm}\right)\right]\right) + \beta_0\right]_p \quad \text{Eq. 4-2}$$

AADT is rounded to the nearest whole number.

where:

s = number of unique sensors that comprise each screenline modal count
β_1 = linear regression coefficient, unique for each sensor
β_0 = intercept term, unique for each sensor
d = number of time intervals in one day
- d = 1 for daily volume
- d = 24 for hourly volume
- d = 96 for fifteen-minute volume

n = number of days in a year (365 or 366)
VOL_{ijm} = daily volume for the *i*th occurrence of the *j*th day of the week within the *m*th month
i = occurrences of day j in month m for which traffic data are available
j = day of week (1 to 7)
m = month of year (1 to 12)
n_{jm} = number of occurrences of day j in month m for which traffic data are available

Linear Regression, applied to volume data

$$AADT = \left(\frac{1}{12}\sum_{m=1}^{12}\left[\frac{1}{7}\sum_{j=1}^{7}\left(\frac{1}{n_{jm}}\sum_{i=1}^{n_{jm}}\left(\sum_{p=1}^{s}\left(\sum_{k=1}^{d} Round(\beta_1 * VOL_k + \beta_0)\right)_p\right)_{ijm}\right)\right]\right) \quad \text{Eq. 4-3}$$

where:

s = number of unique sensors that comprise each screenline modal count
β_1 = linear regression coefficient, unique for each sensor
β_0 = intercept term, unique for each sensor
VOL_k = time interval volume for the *k*th occurrence for one day
d = number of time intervals in one day
- d = 1 for daily volume
- d = 24 for hourly volume
- d = 96 for fifteen-minute volume

i = occurrences of day j in month m for which traffic data are available
j = day of week (1 to 7)
m = month of year (1 to 12)

n_{jm} = number of occurrences of day *j* in month *m* for which traffic data are available

4.3.2 Correction Factor Methodology

$$\frac{\sum_{i=1}^{2}(\sum_{j=1}^{m} VOL_{manual})_i}{\sum_{i=1}^{2}(\sum_{j=1}^{m} VOL_{eco})_i} \qquad \text{Eq. 4-4}$$

where:

VOL_{eco}= Total 15-minute volume as recorded by Eco-Counter device

VOL_{manual} = Total 15-minute volume as recorded by staff reviewing video observation of equipment installation site

j = 15-minute observation period

m = all 15-minute time intervals during daylight hours over the two-day observation period, for the "Two-Day Observation" correction factor method; or

m = first 30 non-zero 15-minute time intervals or all 15-minute time intervals during daylight hours over two-day observation period, whichever is less, for "First 30" correction factor method

Correction Factor, applied to AADT

$$AADT = \sum_{p=1}^{s}\left[CF_P * \left(\frac{1}{12}\sum_{m=1}^{12}\left[\frac{1}{7}\sum_{j=1}^{7}\left(\frac{1}{n_{jm}}\sum_{i=1}^{n_{jm}} VOL_{ijm}\right)\right]\right)\right]_p \qquad \text{Eq. 4-5}$$

where:

CF_P = Correction factor for the pth sensor

s = number of unique sensors that comprise each screenline modal count

n = number of days in a year (365 or 366)

VOL_{ijm} = daily volume for the ith occurrence of the jth day of the week within the mth month

i = occurrences of day *j* in month *m* for which traffic data are available

j = day of week (1 to 7)

m = month of year (1 to 12)

n_{jm} = number of occurrences of day j in month *m* for which traffic data are available

Correction Factor, applied to volume data.

$$AADT = \left(\frac{1}{12}\sum_{m=1}^{12}\left[\frac{1}{7}\sum_{j=1}^{7}\left(\frac{1}{n_{jm}}\sum_{i=1}^{n_{jm}}\left(\sum_{p=1}^{s}\left(\sum_{k=1}^{d} CF_p * VOL_k\right)_p\right)_{ijm}\right)\right]\right) \qquad \text{Eq. 4-6}$$

where:

s = number of unique sensors that comprise each screenline modal count

CF = correction factor, unique for each sensor
VOL_k = time interval volume for the *k*th occurrence for one day
d = number of time intervals in one day
d = 1 for daily volume
d = 24 for hourly volume
d = 96 for fifteen-minute volume

Weighted Mean Percentage Error (WMPE)

$$\sum_{t=1}^{n}\left(\frac{VOL_{eco} - VOL_{manual}}{\sum_{t=1}^{n} VOL_{manual}}\right) \qquad \text{Eq. 4-7}$$

where:
VOL_{eco} = Total 15-minute volume as recorded by Eco-Counter device
VOL_{manual} = Total 15-minute volume as recorded by staff reviewing video observation of equipment

Notes on Data Processing:

- All data were collected using Eco-Counter MULTI systems. Eco-Counter MULTI systems include passive infrared sensors to detect pedestrian traffic and inductive loops to detect bicycle traffic.
- Data was flagged using the NM COAST process developed by the ITRE NC NMVDP team
- Only days of data where all sensors were not flagged as invalid by the QA/QC process were included in the modal screenline and considered in the AADT calculation.
- Years of data when more than sixty days of data were removed from consideration for the AADT calculation due to error flags are not included in this analysis.
- An observation period is considered a "non-zero" observation if the sum of all observations as recorded by the Eco-Counter device and all observations made by a video review analyst is greater than zero.

4.4 Results and Findings

4.4.1 Number of Observations Required to Calculate a Consistent Correction Factor

Comparisons of the differences between the correction factors calculated based on all two-day 15-minute interval observations and the correction factors calculated based on the first 30 15-minute interval observations are provided in the figures below. This analysis was performed to determine if accurate correction factors could be calculated using fewer observations than current practices, which could save analysts time in future validation studies. Figures 4-1 and 4-2 show that most validation studies are only within five percentage point differences, yet a significant number of studies have larger differences between the two methods (28 bicycle sensor studies and 45 pedestrian studies). This shows that less than two days of validation data may be

appropriate for most sites, but collecting more than 30 non-zero observations may be necessary for certain counting locations to ensure sufficient variability is captured in the validation study. Tables 4-2 and 4-3 summarize the validation studies with differences greater than 10% between the two correction factor methods.

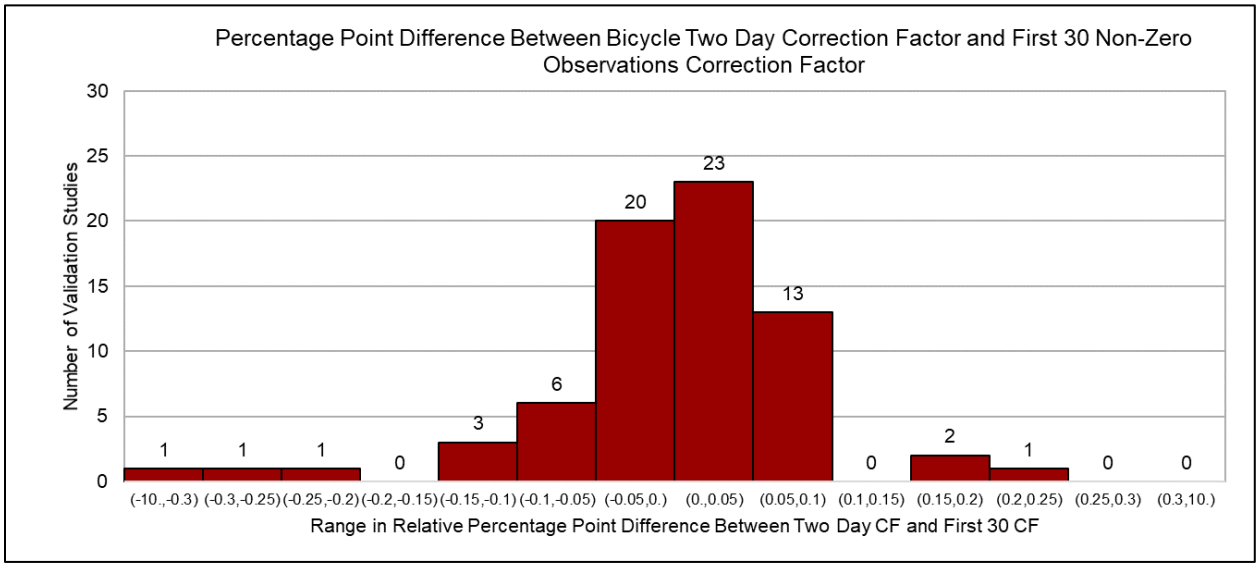

Figure 4-1 Percentage Point Difference Between Bicycle Two Day Correction Factor and First 30 Non-Zero Observations Correction Factor

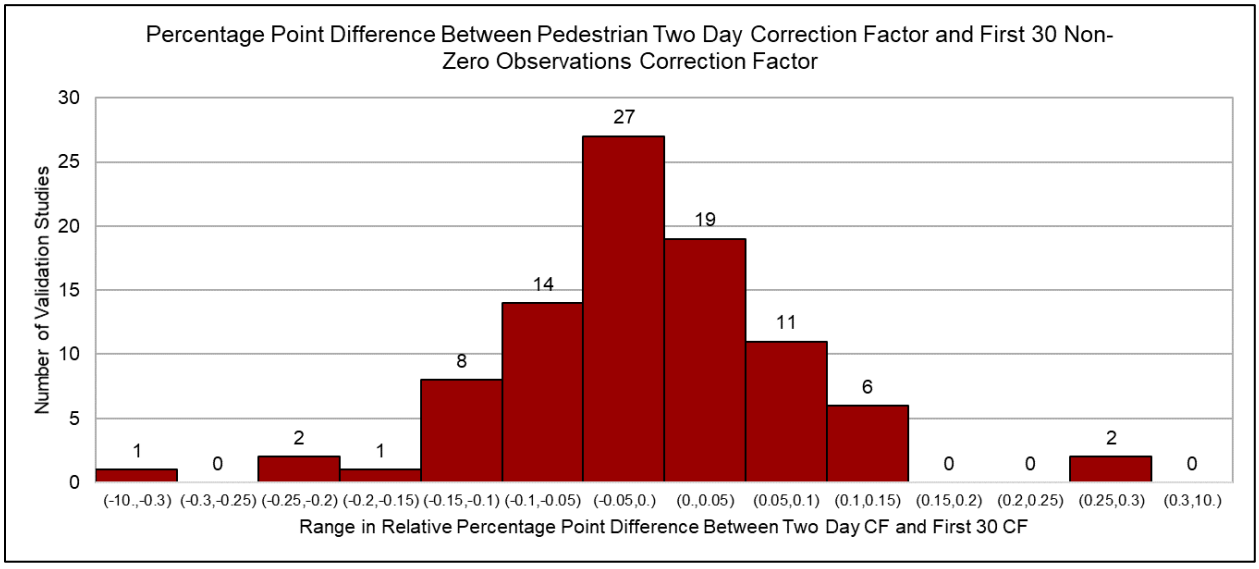

Figure 4-2 Percentage Point Difference Between Pedestrian Two Day Correction Factor and First 30 Non-Zero Observations Correction Factor

Table 4-2 Outlier Studies with > 10 Percentage Point Differences in Correction Factors Calculated from First 30 Observations Versus All Two Day Observations

Site Name	Mode	Two Day Correction Factor	First 30 Non-Zero Observations Correction Factor	Difference	Two Day Non-Zero Time Intervals Observed	Two Day Total Users Observed
GSO_ELM_B_E_RD	Bike	1.03	1.24	0.21	40	34
CRY_NHC_B_S_RD	Bike	1.45	1.62	0.17	42	84
CLT_USP_B_P_SU	Bike	0.93	1.09	0.16	57	113
BRV_BGW_B_P_SU	Bike	0.99	0.88	-0.11	64	132
GSO_SPR_B_N_RD	Bike	0.89	0.78	-0.11	51	68
CHL_MLK_B_E_RD	Bike	0.75	0.63	-0.13	38	21
DRH_MAI_B_S_RD	Bike	2.51	2.27	-0.24	84	226
CRB_OLD_B_W_RD	Bike	0.80	0.51	-0.29	90	164
CRY_NHC_B_N_RD	Bike	3.43	2.57	-0.86	49	96

Table 4-3 Outlier Studies with >10 Percentage Point Differences in Correction Factors Calculated from First 30 Observations Versus All Two Day Observations

Sensor Name	Mode	Two Day Correction Factor	First 30 Non-Zero Observations Correction Factor	Difference	Two Day Non-Zero Time Intervals Observed	Two Day Total Users Observed
BRV_BGW_P_P_SU	Pedestrian	0.95	0.80	-0.15	79	261
CHL_MLK_P_E_SW	Pedestrian	1.27	1.15	-0.12	89	320
CHL_MLK_P_W_SW	Pedestrian	1.3	1.43	0.13	67	143
CLT_BLT_P_P_SU	Pedestrian	1.32	0.97	-0.35	109	2782
CLT_ELZ_P_E_SW	Pedestrian	1.12	0.87	-0.25	103	557
CLT_ELZ_P_W_SW	Pedestrian	0.99	0.76	-0.23	99	429
CLT_ELZ_P_W_SW	Pedestrian	1.15	1.25	0.1	75	727
CLT_TEN_P_E_SW	Pedestrian	1.41	1.28	-0.13	85	270
CRB_LCB_P_P_SU	Pedestrian	0.79	0.93	0.14	111	616
CRY_BCT_P_P_SU	Pedestrian	1.18	1.32	0.14	97	237
DCK_TRL_P_P_SU	Pedestrian	1.59	1.45	-0.14	103	2386
DRH_ATD_P_P_SU	Pedestrian	1.19	1.31	0.12	88	402
DRH_ATD_P_P_SU	Pedestrian	0.84	1.13	0.29	99	543
DRH_ATT_P_P_SU	Pedestrian	1.21	1.10	-0.11	106	827
DRH_MAI_P_N_SW	Pedestrian	1.2	1.06	-0.14	103	501
DRH_MAI_P_S_SW	Pedestrian	0.85	0.95	0.1	50	77
GSO_ELM_P_E_SW	Pedestrian	1.35	1.21	-0.14	110	2450
GSO_ELM_P_W_SW	Pedestrian	1.18	1.03	-0.15	118	3947
SAN_EFG_P_P_SU	Pedestrian	0.72	0.98	0.26	86	323
W-S_STR_P_P_SU	Pedestrian	1.08	0.92	-0.16	76	279

4.4.2 Correlation Between Interval Observations and Eco-Visio Counts

Pearson's Correlation Coefficients (*r*) were calculated between the users observed by data analysts and the users recorded by Eco-Counters for all 15-minute intervals and hourly intervals. The statistical results for pedestrians are presented in Figure 4-3, Figure 4-4, Table 4-4, and Table 4-5, while the results for cyclists are presented in Figure 4-5, Figure 4-6, Table 4-6, and Table 4-7. This analysis informs whether the observations recorded by the Eco-Counter are well-correlated with the users observed by data analysts and how under- and over-counting errors can be smoothed over an hourly interval.

Thirteen 15-minute time interval validation studies and eight 1-hour time interval validation studies had a correlation coefficient less than 0.7. Of these studies, twelve 15-minute time interval validation studies and seven 1-hour time interval validation studies had correction

factors within the tolerance set by the NC NMVDP. This shows that the effects of counting device errors can be smoothed over the two-day validation observation period.

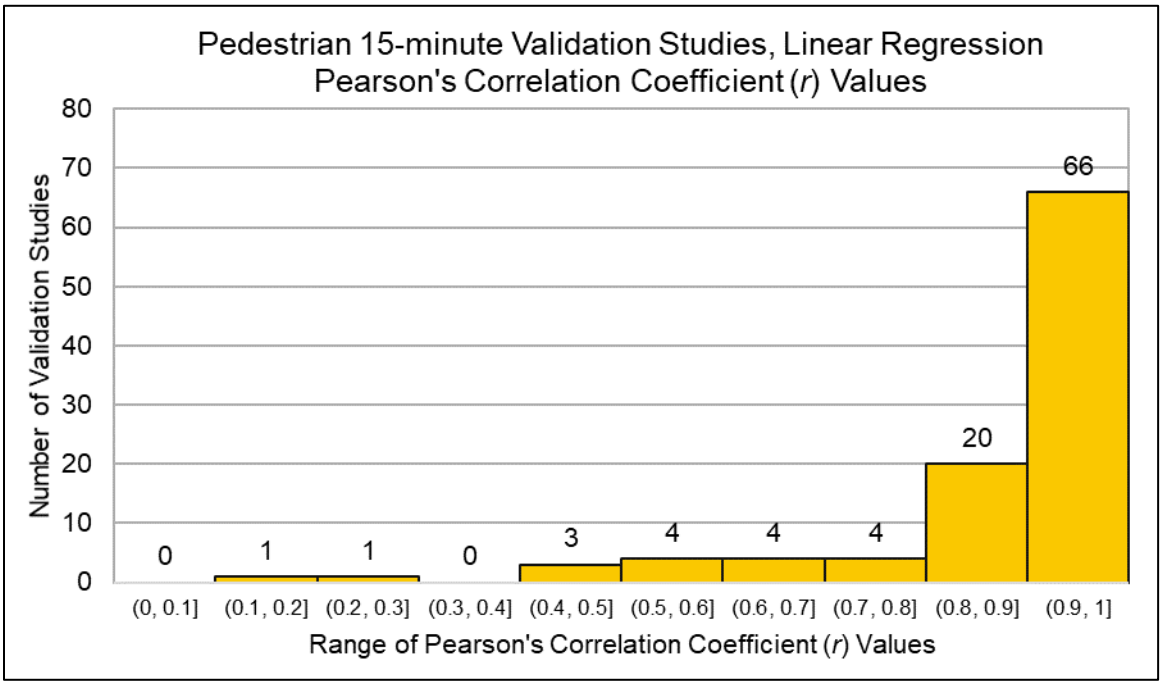

Figure 4-3 Pedestrian 15-minute Validation Studies, Linear Regression (Pearson's Correlation Coefficient Values)

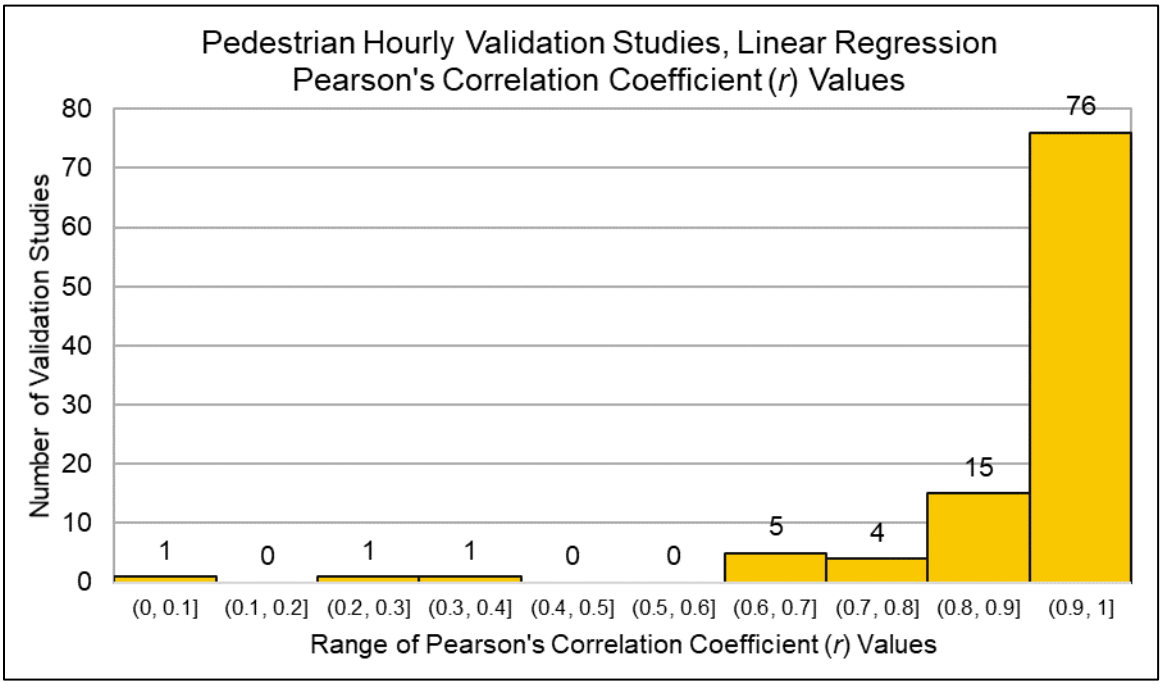

Figure 4-4 Pedestrian Hourly Validation Studies, Linear Regression (Pearson's Correlation Coefficient Values)

Table 4-4 Outlier Studies in Pearson's Correlation Coefficient (r < 0.7) for Pedestrian 15-minute Validation Studies

Sensor Name	Time Intervals Observed	Total Users Observed	Correction Factor	a(lin)	b(lin)	*r*
GSO_WAL_P_S_SW	127	1224	2.10	1.43	3.05	0.69
BRV_BGW_P_P_SU	95	261	0.95	0.68	0.79	0.67
SAN_EFG_P_P_SU	112	323	0.72	0.41	1.23	0.64
DRH_MAI_P_S_SW	108	77	0.85	0.40	0.38	0.60
CLT_ELZ_P_W_SW	104	429	0.99	0.61	1.59	0.60
DRH_ATD_P_P_SU	104	543	0.84	0.40	2.74	0.54
CHL_MLK_P_E_SW	99	320	1.27	0.60	1.71	0.51
CLT_4EX_P_S_SW	92	78	1.03	0.39	0.53	0.50
CLT_TEN_P_W_SW	108	23	1.05	0.45	0.12	0.50
CLT_ELZ_P_E_SW	104	557	1.12	0.50	2.96	0.41
CLT_STN_P_S_SW	92	95	1.17	0.47	0.62	0.40
CRY_BCT_P_P_SU	114	237	1.18	0.33	1.50	0.25
CLT_NTR_P_S_SW	103	219	0.97	0.13	1.83	0.14

Table 4-5 Outlier Studies in Pearson's Correlation Coefficient (r < 0.7) for Pedestrian Hourly Validation Studies

Sensor Name	Time Intervals Observed	Total Users Observed	Correction Factor	a(lin)	b(lin)	*r*
CHL_MLK_P_E_SW	25	320	1.27	0.92	3.52	0.70
CLT_TEN_P_W_SW	27	23	1.05	0.72	0.27	0.70
GSO_WAL_P_S_SW	32	1224	2.10	1.69	7.33	0.70
SAN_EFG_P_P_SU	28	323	0.72	0.36	5.76	0.67
CLT_STN_P_S_SW	23	95	1.17	0.75	1.50	0.61
DRH_ATD_P_P_SU	26	543	0.84	0.25	14.76	0.35
CLT_4EX_P_S_SW	24	78	1.03	0.26	2.41	0.26
CLT_NTR_P_S_SW	26	219	0.97	0.07	7.82	0.07

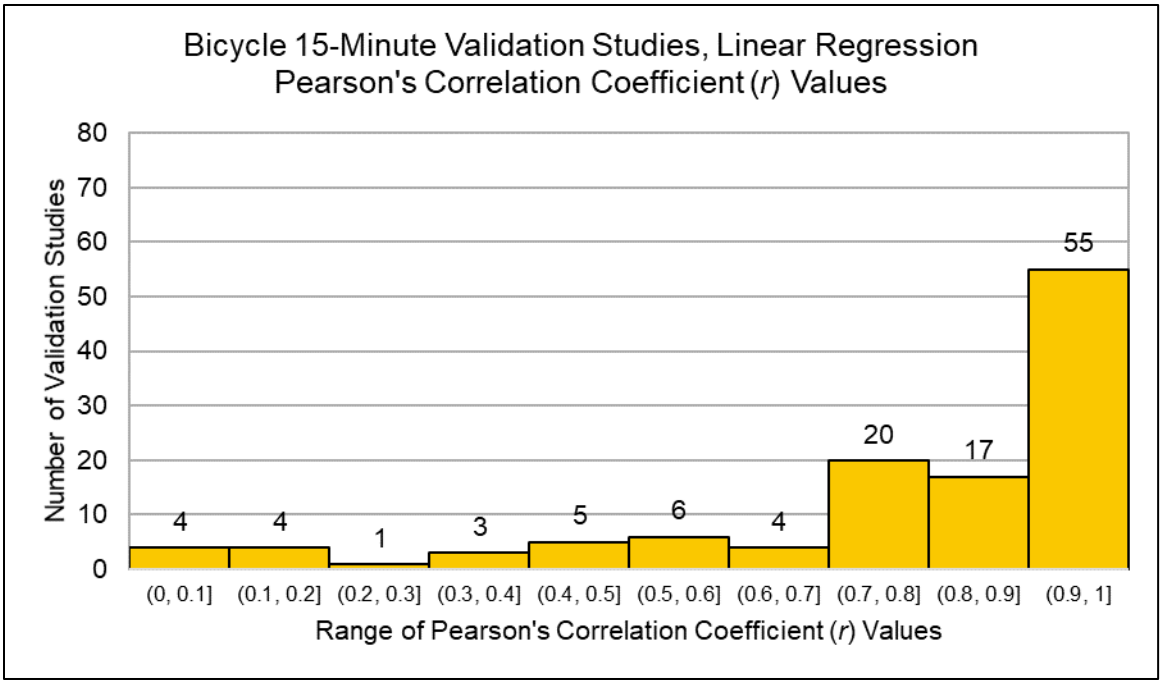

Figure 4-5 Bicycle 15-minute Validation Studies, Linear Regression (Pearson's Correlation Coefficient Values)

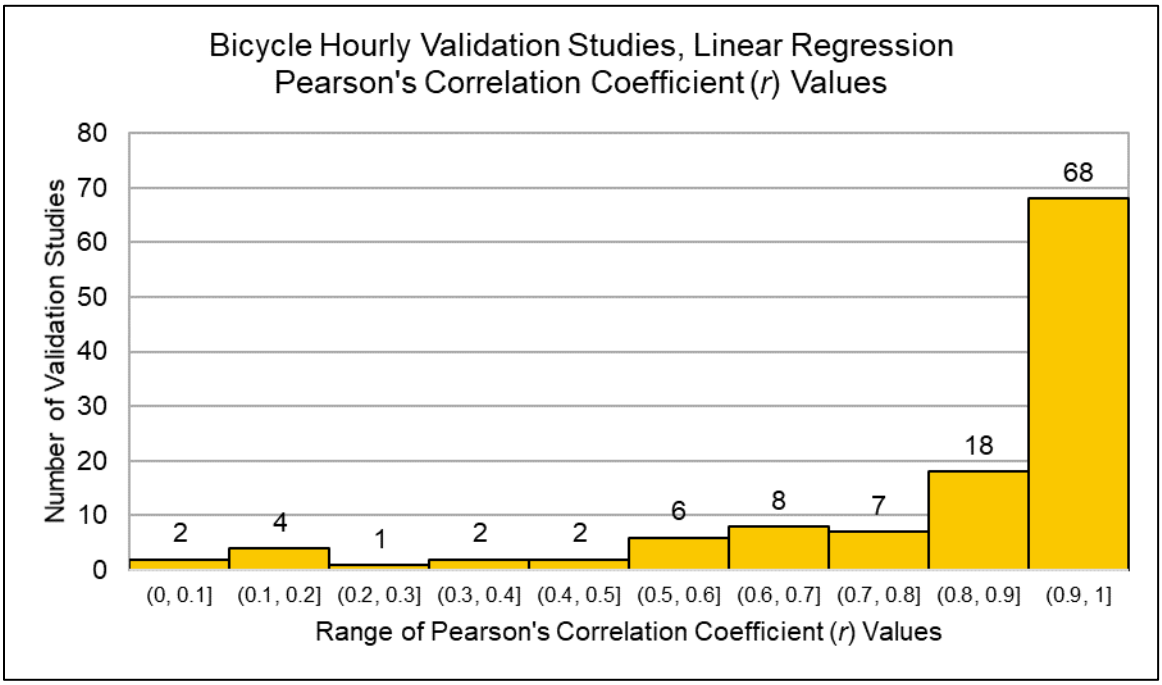

Figure 4-6 Bicycle Hourly Validation Studies, Linear Regression (Pearson's Correlation Coefficient Values)

Table 4-6 Outlier Studies in Pearson's Correlation Coefficient (r < 0.7) for Bicycle 15-minute Validation Studies

Sensor Name	Time Intervals Observed	Total Users Observed	Correction Factor	a(lin)	b(lin)	*r*
APX_SAL_B_W_RD	112	32	1.23	0.04	1.07	0.69
CRY_NHC_B_S_RD	120	84	1.45	0.19	1.06	0.68
DVD_MAI_B_W_RD	112	41	1.28	0.11	0.89	0.63
BRV_BGW_B_P_SU	95	120	1.03	0.54	0.59	0.61
W-S_4TH_B_S_RD	127	57	1.06	0.17	0.66	0.59
CLT_TEN_B_E_RD	112	17	0.41	0.04	0.32	0.55
CLT_TEN_B_E_SW	112	17	0.41	0.04	0.32	0.55
CLT_TEN_B_E_RD	114	33	3.30	0.20	1.06	0.55
CRY_NHC_B_N_RD	120	96	3.43	0.50	1.27	0.53
GSO_WAL_B_N_RD	91	8	0.67	0.04	0.40	0.52
APX_SAL_B_E_RD	112	10	1.25	0.05	0.58	0.47
CLT_STN_B_N_SW	104	47	0.66	0.21	0.35	0.41
DRH_COR_B_S_RD	112	27	3.86	0.19	0.88	0.41
DVD_MAI_B_E_RD	112	36	0.97	0.14	0.54	0.41
GSO_ELM_B_E_RD	100	34	1.03	0.20	0.44	0.40
W-S_4TH_B_N_RD	127	70	0.36	0.29	0.17	0.39
CLT_STN_B_S_RD	92	3	0.75	0.02	0.24	0.39
CLT_4EX_B_S_RD	80	36	3.00	0.33	0.82	0.31
DVD_GRF_B_S_RD	97	5	1.67	0.04	0.29	0.23
CLT_4EX_B_N_RD	92	24	12.00	0.23	1.27	0.20
DVD_MAI_B_E_RD	104	11	5.50	0.10	0.40	0.18
CHL_MLK_B_E_RD	99	21	0.75	0.18	0.13	0.14
W-S_POL_B_N_RD	148	72	0.13	0.48	0.03	0.13
CHL_MLK_B_W_RD	99	12	0.38	0.11	0.03	0.06
CRY_BCT_B_P_SU	114	138	1.05	1.16	0.04	0.04
DRH_COR_B_N_RD	112	25	1.92	0.21	0.07	0.04
CLT_STN_B_N_RD	104	15	0.60	0.14	0.01	0.02

Table 4-7 Outlier Studies in Pearson's Correlation Coefficient (r < 0.7) for Bicycle Hourly Validation Studies

Sensor Name	Time Intervals Observed	Total Users Observed	Correction Factor	a(lin)	b(lin)	*r*
DRH_MAI_B_N_RD	28	89	1.47	3.26	0.85	0.68
CRY_NHC_B_N_RD	30	49	3.43	1.68	1.63	0.68
W-S_4TH_B_S_RD	32	57	1.06	0.36	0.84	0.63
DVD_GRF_B_S_RD	28	13	1.27	0.16	0.86	0.62
APX_SAL_B_E_RD	28	12	1.25	0.10	0.90	0.61
CLT_TEN_B_E_RD	28	30	0.41	0.11	0.34	0.61
CLT_TEN_B_E_SW	28	30	0.41	0.11	0.34	0.61
GSO_WAL_B_N_RD	23	14	0.67	0.09	0.50	0.61
DVD_MAI_B_W_RD	26	20	0.95	0.32	0.51	0.60
GSO_ELM_B_E_RD	26	40	1.03	0.57	0.58	0.60
RAL_RBT_B_P_SU	28	58	0.94	0.93	0.56	0.57
CLT_STN_B_N_SW	26	55	0.66	0.69	0.41	0.54
W-S_4TH_B_N_RD	32	83	0.36	0.77	0.23	0.53
CLT_TEN_B_E_RD	30	21	3.30	0.71	1.16	0.52
DVD_MAI_B_E_RD	28	35	0.97	0.49	0.60	0.46
DVD_MAI_B_E_RD	26	12	5.50	0.33	1.17	0.45
DVD_GRF_B_S_RD	26	7	1.67	0.13	0.54	0.36
CLT_STN_B_S_RD	23	4	0.75	0.10	0.20	0.34
CHL_MLK_B_W_RD	25	34	0.38	0.30	0.14	0.25
W-S_POL_B_N_RD	28	97	0.15	2.10	0.03	0.16
DRH_COR_B_S_RD	28	18	3.86	0.85	0.46	0.15
CLT_4EX_B_N_RD	24	12	12.00	0.91	1.09	0.12
CLT_STN_B_N_RD	26	29	0.60	0.49	0.09	0.11
CLT_4EX_B_S_RD	20	18	3.00	1.59	0.34	0.08
CHL_MLK_B_E_RD	25	38	0.75	0.78	0.06	0.06
DRH_COR_B_N_RD	28	24	1.92	0.97	-0.17	-0.12

4.4.3 Rounding Error Due to Correction Factor Application

Correction factors calculated using the two-day validation studies were applied to volume totals at 15-minute time intervals, hourly time intervals, and daily time intervals and then compared to the yearly corrected AADT value. This analysis was performed to determine the effect of rounding a corrected count on the overall summary statistics of a counting site. Only years of data where the following criteria were met were included in the analysis:

1. All sensors were well-functioning for at least 300 days in the year.
2. All sensors had the same correction factor during the duration of the year; no maintenance requiring an additional validation study was required of any sensor.

The results indicate that the effects of rounding increased as the time interval that the correction factor was applied to decreased; rounding had no impact on the AADT values when correction factors were applied at the daily levels and had a greater effect on the AADT calculations when applied to 15-minute intervals. Bicycle sensors were also more heavily impacted by changes in AADT values due to rounding than pedestrian sensors since bicycle volume patterns are more likely to be lower more consistently throughout the day. The impacts of rounding are also more pronounced for lower volume sites than higher volume sites. Low volume bicycle site AADT calculations with correction factors applied to the 15-minute time interval were 2% to 12% different than the corrected AADT value. Medium volume pedestrian sites AADT calculations with the correction factors applied to the 15-minute time interval were between 2% to 7%. This analysis shows that rounding errors due to correction factor application can have a significant impact on AADT estimates due to the lower volume nature of bicycle and pedestrian counting sites.

The following tables show the AADT values of the uncorrected and corrected datasets as well as the percent difference between the uncorrected (baseline) AADT value and the AADT values produced by the corrected datasets when correction factors are applied at the daily, hourly, and 15-minute time intervals (Day, Hour, 15Min). The charts summarize the difference in percentage points between yearly AADT figures developed by datasets with correction factors applied at the daily level and datasets with correction factors applied at hourly and 15-minute levels (Equations 4-8 and 4-9).

$$Diff_{Hourly} = \frac{(AADT_{Daily\ Corrected\ Data} - AADT_{Uncorrected\ Baseline})}{AADT_{Uncorrected\ Baseline}} - \frac{(AADT_{Hourly\ Corrected\ Data} - AADT_{Uncorrected\ Baseline})}{AADT_{Uncorrected\ Baseline}}$$

Eq. 4-8

$$Diff_{Hourly} = \frac{(AADT_{Daily\ Corrected\ Data} - AADT_{Uncorrected\ Baseline})}{AADT_{Uncorrected\ Baseline}} - \frac{(AADT_{15-Minute\ Corrected\ Data} - AADT_{Uncorrected\ Baseline})}{AADT_{Uncorrected\ Baseline}}$$

Eq. 4-9

These equations measure the impact that applying correction factors at different aggregation levels of volume data have on final AADT calculations. Results for individual years at individual stations are also summarized in the following figures and tables. Processes with higher

incidences of years when the magnitude of difference in AADT values between methodologies is greater than +/- 5% indicate processes that are more sensitive to impacts of applying correction factors to different aggregation levels.

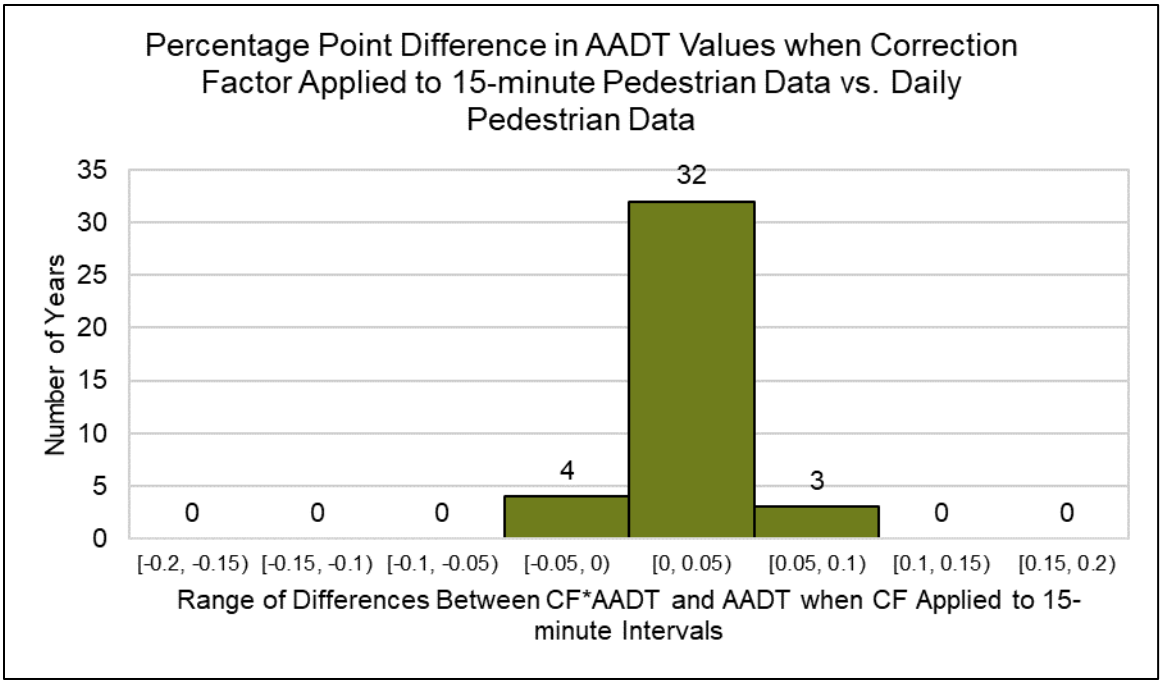

Figure 4-7 Difference in AADT Values when Correction Factor Applied to 15-minute Pedestrian Data

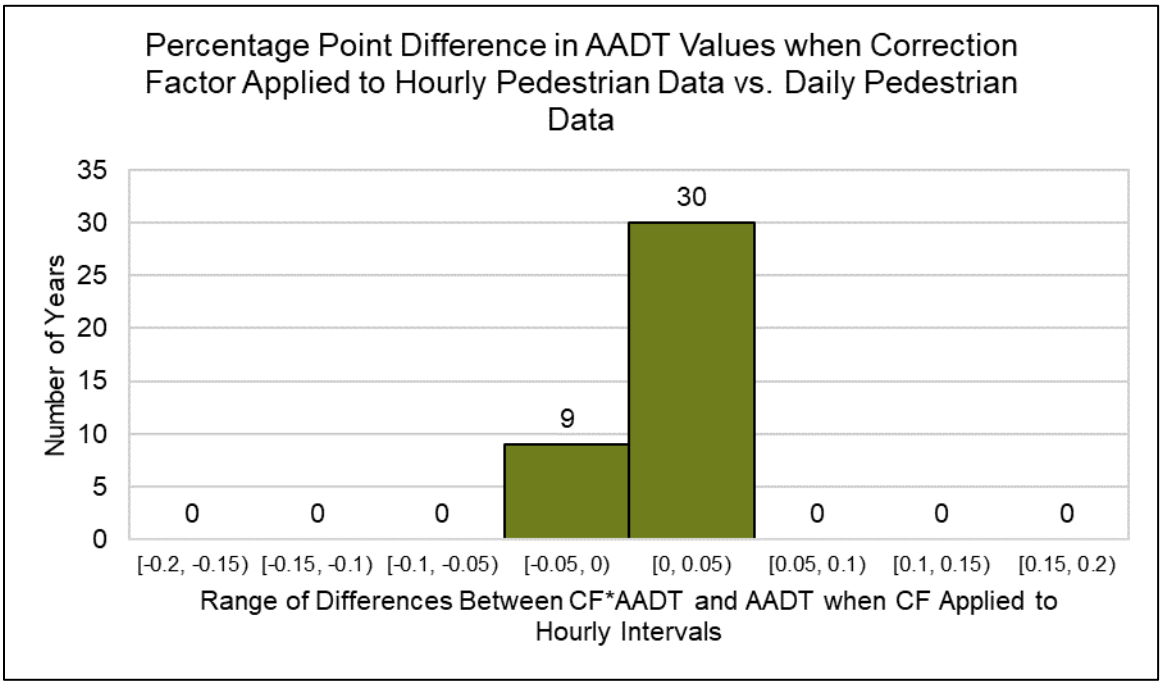

Figure 4-8 Difference in AADT Values when Correction Factor Applied to Hourly Pedestrian Data

Table 4-8 Difference between AADT Values when Correction Factor is Applied to Daily Totals, Hourly Totals, and 15-Minute Totals; Medium Volume Pedestrian Sites (AADT < 500)

Station	*n*	Year	Uncorrected AADT	Day AADT	Percent Difference	Hour AADT	Percent Difference	15-Min AADT	Percent Difference
CLT_4EX	324	2017	132	144	9%	141	7%	135	2%
CRB_OLD	361	2016	63	67	6%	65	3%	63	0%
CLT_SWN	312	2017	251	280	12%	279	11%	266	6%
CRB_OLD	365	2015	67	71	6%	70	4%	68	1%
GSO_LDG	321	2015	173	200	16%	199	15%	193	12%
GSO_LDG	333	2017	174	202	16%	201	16%	195	12%
GSO_LDG	326	2016	191	221	16%	220	15%	214	12%
CLT_USP	364	2019	201	241	20%	241	20%	235	17%
DRH_ATT	353	2016	277	344	24%	343	24%	337	22%
CLT_USP	364	2018	226	271	20%	271	20%	266	18%
DRH_ATT	361	2017	287	355	24%	354	23%	349	22%
WLK_YDK	365	2017	193	220	14%	220	14%	217	12%
W-S_END	359	2015	256	336	31%	335	31%	333	30%
W-S_END	323	2018	268	347	29%	345	29%	344	28%
CLT_ELZ	323	2017	473	502	6%	503	6%	497	5%
W-S_END	349	2017	266	353	33%	352	32%	351	32%
W-S_END	334	2016	291	387	33%	386	33%	385	32%
CRB_LCB	366	2016	466	368	-21%	368	-21%	372	-20%
CRB_LCB	365	2015	454	359	-21%	360	-21%	363	-20%
CRB_LCB	321	2017	440	347	-21%	348	-21%	351	-20%

Table 4-9 Difference between AADT Values when Correction Factor is Applied to Daily Totals, Hourly Totals, and 15-Minute Totals; High Volume Pedestrian Sites (AADT > 500)

Station	*n*	Year	Uncorrected AADT	Day AADT	Percent Difference	Hour AADT	Percent Difference	15-Min AADT	Percent Difference
CHL_MLK	363	2015	642	691	8%	690	7%	675	5%
CHL_MLK	320	2016	640	689	8%	688	8%	674	5%
GSO_SPR	365	2019	778	806	4%	804	3%	792	2%
GSO_SPR	365	2018	787	815	4%	813	3%	801	2%
GSO_SPR	357	2015	808	837	4%	835	3%	823	2%
GSO_SPR	366	2016	831	862	4%	860	3%	848	2%
GSO_SPR	365	2017	856	887	4%	885	3%	873	2%
GSO_WAL	359	2015	850	968	14%	968	14%	963	13%
DVD_MAI	360	2019	1736	2081	20%	2078	20%	2073	19%
GSO_ELM	363	2015	2603	3251	25%	3251	25%	3242	25%
DVD_MAI	353	2018	1636	1962	20%	1959	20%	1957	20%
W-S_4TH	364	2015	3436	4284	25%	4285	25%	4276	24%
GSO_WAL	360	2017	991	1124	13%	1124	13%	1122	13%
GSO_WAL	352	2016	1035	1174	13%	1175	14%	1173	13%
W-S_4TH	362	2017	3268	4079	25%	4079	25%	4078	25%
GSO_ELM	366	2016	2552	3202	25%	3203	26%	3202	25%
W-S_4TH	365	2016	3422	4269	25%	4270	25%	4269	25%
W-S_4TH	358	2018	3027	3765	24%	3766	24%	3765	24%
CLT_NTR	305	2017	674	646	-4%	649	-4%	665	-1%

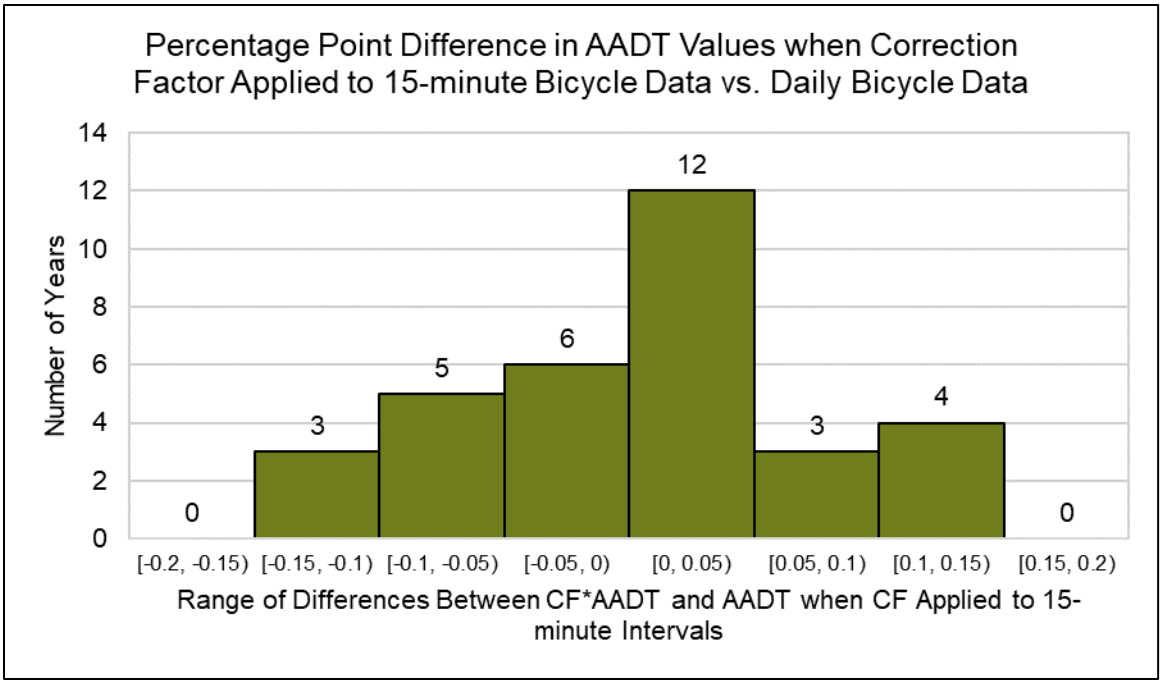

Figure 4-9 Difference in AADT Values When Correction Factor Applied to 15-minute Bicycle Data

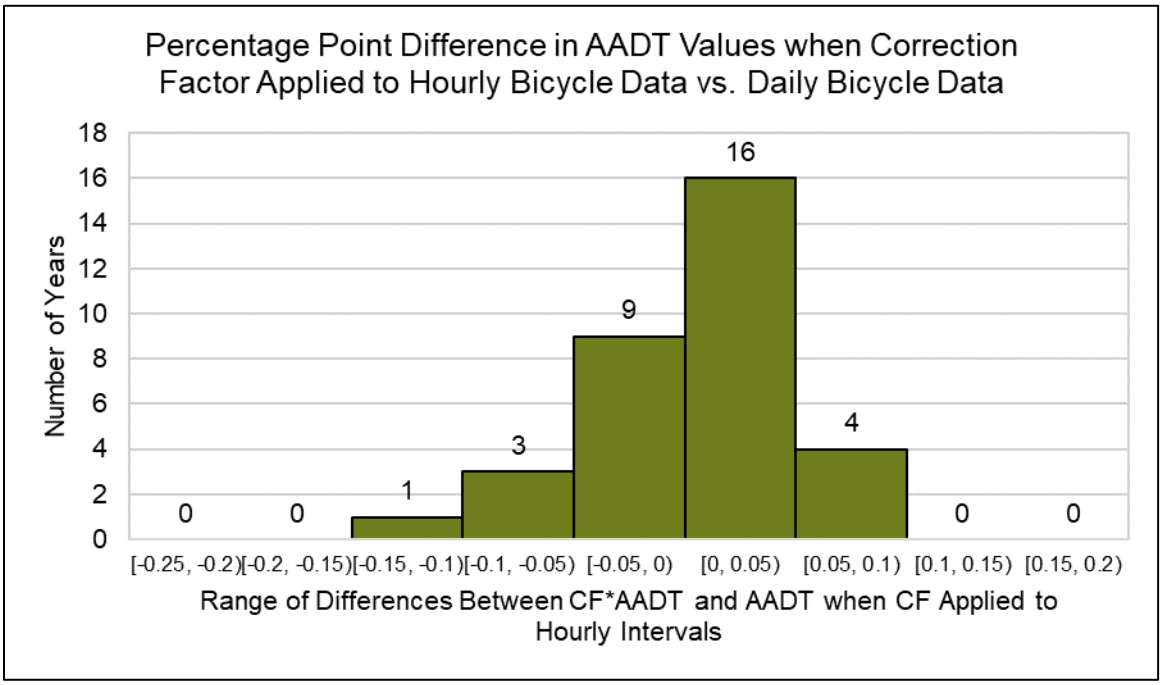

Figure 4-10 Difference in AADT Values when Correction Factor Applied to Hourly Bicycle Data

Table 4-10 Difference between AADT Values when Correction Factor is Applied to Daily Totals, Hourly Totals, and 15-Minute Totals; Low Volume Bicycle Sites (AADT < 50)

Station	*n*	Year	Uncorrected AADT	Day AADT	Percent Difference	Hour AADT	Percent Difference	15-Min AADT	Percent Difference
CLT_SWN	338	2017	43	50	16%	48	12%	46	7%
CLT_USP	358	2018	41	38	-7%	40	-2%	41	0%
CLT_USP	362	2019	47	44	-6%	46	-2%	47	0%
GSO_LDG	304	2018	49	48	-2%	49	0%	49	0%
GSO_WAL	365	2015	35	43	23%	41	17%	38	9%
GSO_WAL	364	2016	33	40	21%	38	15%	36	9%
WLK_YDK	360	2017	32	33	3%	32	0%	32	0%
W-S_4TH	365	2018	50	58	16%	55	10%	52	4%
W-S_END	364	2015	35	29	-17%	31	-11%	33	-6%
W-S_END	366	2016	30	25	-17%	27	-10%	28	-7%
W-S_END	365	2017	28	23	-18%	26	-7%	27	-4%
W-S_END	349	2018	26	22	-15%	24	-8%	25	-4%

Table 4-11 Difference between AADT Values when Correction Factor is Applied to Daily Totals, Hourly Totals, and 15-Minute Totals; Medium Volume Bicycle Sites (50 < AADT < 250)

Station	*n*	Year	Uncorrected AADT	Day AADT	Percent Difference	Hour AADT	Percent Difference	15-Min AADT	Percent Difference
CHL_MLK	361	2015	118	98	-17%	102	-14%	108	-8%
CHL_MLK	364	2016	123	101	-18%	105	-15%	112	-9%
CRB_OLD	365	2015	172	163	-5%	163	-5%	168	-2%
CRB_OLD	366	2016	104	100	-4%	100	-4%	102	-2%
CRB_OLD	365	2017	73	71	-3%	71	-3%	73	0%
CRB_OLD	364	2018	61	59	-3%	60	-2%	61	0%
DRH_ATT	351	2016	243	250	3%	249	2%	245	1%
GSO_ELM	348	2015	74	80	8%	78	5%	75	1%
GSO_ELM	348	2016	71	77	8%	75	6%	72	1%
GSO_LDG	349	2015	62	62	0%	62	0%	62	0%
GSO_LDG	363	2016	56	56	0%	56	0%	56	0%
GSO_LDG	361	2017	56	55	-2%	56	0%	56	0%
GSO_SPR	329	2015	110	113	3%	114	4%	111	1%
GSO_SPR	364	2017	102	105	3%	106	4%	103	1%
GSO_SPR	365	2018	82	85	4%	85	4%	83	1%
GSO_SPR	365	2019	74	77	4%	77	4%	75	1%
W-S_4TH	365	2018	50	58	16%	55	10%	52	4%
W-S_4TH	363	2019	51	60	18%	56	10%	52	2%

Table 4-12 Difference between AADT Values when Correction Factor is Applied to Daily Totals, Hourly Totals, and 15-Minute Totals; High Volume Bicycle Sites (500 < AADT)

Station	*n*	Year	Uncorrected AADT	Day AADT	Percent Difference	Hour AADT	Percent Difference	15-Min AADT	Percent Difference
CRB_LCB	365	2015	539	571	6%	570	6%	565	5%
CRB_LCB	366	2016	539	571	6%	571	6%	566	5%
CRB_LCB	325	2017	559	593	6%	592	6%	588	5%

4.4.4 Weighted Average Percentage Deviation

The Weighted Average Percentage Deviation (WAPD) is a measure of accuracy that accounts for the low-volume bias in the observations and is considered a more reliable statistic than the average percent error (Ryus et al., 2014). The WAPD for each validation study was calculated. Summary statistics for bicycle and pedestrian studies are displayed in Table 4-13. The results show that the mean WAPD for pedestrian sensors was about 9% while the mean WAPD for bicycle sensors was about -5%. However, the standard deviation for bicycle sensors was much higher than pedestrian sensors. These results show that the weighted mean percent error is not consistent between sensors installed in the field and the individual validation studies are still a valuable resource for determining if the accuracy and precision of a counting system is within the programs' tolerance.

Table 4-13 Weighted Average Percentage Deviation

Pedestrian WAPD	
Mean	0.09
Standard Error	0.02
Median	0.10
Standard Deviation	0.16
Sample Variance	0.03
Range	1.36
Minimum	-0.83
Maximum	0.52
Count (Validation Studies)	104

Bicycle WAPD	
Mean	-0.05
Standard Error	0.06
Median	0.04
Standard Deviation	0.71
Sample Variance	0.50
Range	6.81
Minimum	-5.81
Maximum	1.00
Count	121

Correlation - Pedestrian Studies	
	WAPD
Total Users Observed	0.2000

Correlation - Bicycle Studies	
	WAPD
Total Users Observed	0.0009

4.4.5 Sensor Performance Over Time

Sensor performance over time may deteriorate, which would cause the need to re-validate counting systems to determine if performance accuracy and precision are within acceptable ranges. Additional validation studies were performed on systems to determine if sensor performance deteriorated over time. Counting systems tested were required to have components that were all the same age and installed at the same time. Unfortunately, due to sensor maintenance needs, very few counting systems met these criteria. Results of the revalidation studies are summarized in Table 4-14. While these studies show that the correction factor calculated can change depending on when the study was performed, there were too few counting systems that met the criteria needed to make reliable conclusions on the counters' performance due to sensor aging. Other factors that may impact the correction factor calculated could be the number of user observations over the different validation studies and changes in user behavior, such as more groups of people passing the sensor causing higher occlusion rates. Further investigation into counting device performance over time will require analysis of other devices, likely in other regions of the country.

Table 4-14 Correction Factor Changes in Aging Systems

Sensor ID	Original Validation	Revalidation Start	Age of System (Years)	Original Correction Factor	Revalidation Correction Factor	Difference
GSO_SPR_P_S_SW	11/11/2014	8/22/2020	6	1.02	1.02	0
GSO_SPR_B_S_RD	11/11/2014	8/22/2020	6	1.16	0.98	-0.18
GSO_SPR_P_N_SW	11/11/2014	8/22/2020	6	1.05	1.12	0.07
GSO_SPR_B_N_RD	11/11/2014	8/22/2020	6	0.92	0.99	0.07
RAL_HAR_P_N_SW	4/30/2019	8/21/2020	1	1.21	1.13	-0.08
RAL_HAR_B_N_RD	4/30/2019	8/21/2020	1	1.09	0.94	-0.15
RAL_NRG_P_P_SU	6/7/2018	8/21/2020	2	1.17	1.07	-0.1
RAL_NRG_B_P_SU	6/7/2018	8/21/2020	2	0.98	1.07	0.09

4.5 Summary

This analysis shows that bicycle and pedestrian count data is affected by rounding errors due to the application of correction factors at shorter time intervals. The effect of this error is magnified at lower volume bicycle counting sites and when correction factors are applied to hourly data or 15-minute data. The research team recommends mitigating this source of error by reporting raw hourly or 15-minute count data and then separately reporting correction factors calculated from validation studies. Aggregated figures, such as average volumes by the hour or AADT, are more representative of ground truth counts when correction factors are applied to aggregated figures as opposed to raw data.

Correlation coefficients among validation studies showed that Eco-Counter data could be poorly correlated ($r < 0.7$) while the correction factor was in the acceptable tolerance (between 0.6 and 1.4). The research team recommends calculating the correlation of validation data as well as error and determining a tolerance range for correlation.

Validation studies can likely be optimized by reviewing data based on the number of users observed and time intervals where a user is observed, or where the Eco-Counter detects a user as

opposed to collecting two complete days of manual counts. Reviewing 30-time intervals when users were observed resulted in correction factors that were within five percentage points of correction factors calculated from two complete days of data. However, a significant number of validation studies resulted in correction factors with differences greater than five percentage points between the two correction factor calculation methodologies, which shows that more than 30 non-zero intervals are likely needed to create accurate correction factors for more counting systems.

Counter performance over time could not be adequately determined given the availability of systems with original components. The validation studies performed showed that the performance as represented by the correction factor did not change for some counting systems, overcounted more often for other counting systems, and undercounted more often for other counting systems. To determine if the results of this validation effort are due to counting system age, more systems with components that have not been replaced need to be revalidated.

Comparison of WAPD values among available validation studies shows that the average error of systems is consistent with previous research (undercounting by 9% for pedestrian data and overcounting by -5% for bicycle data), but that the variability between studies was high. Higher errors also had no correlation between total users observed, so it is still difficult to predict which counting sites are more prone to higher rates of error. This analysis shows that validation studies of individual counting systems are still valuable for determining the individual system performance and account for that performance in reported counting figures.

5 BICYCLE VOLUME ESTIMATION AND PREDICTION

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 were developed to quantify the relationship between bicycle count data from permanent continuous count stations and Strava bicycle data as well as other relevant variables. Model results were analyzed and bicycle volume on most of the road segments in the City of Charlotte was calculated based on the model estimation results. In addition, a map illustrating the bicycle ridership in the City of Charlotte was 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 predictions for most of the roadway segments in the City of Charlotte and creates a map to give an overall view of the bicycle ridership in the City of Charlotte. Finally, Section 5.5 concludes this chapter with a summary.

5.2 Data Processing

Before bicycle volume modeling, all the data collected including bicycle volume data from both Strava Metro and permanent continuous count stations, and other supporting data were compiled together for the preparation of bicycle volume estimation and prediction. To be specific, three steps were conducted using ArcGIS and SAS.

Step 1:

This step was to combine the bicycle volume data collected from Strava and permanent continuous count stations with the same date and time period. In these two datasets, information including bicycle counts on the specific road segment, and the cycling date and time period is provided. Therefore, to have a temporal relationship between these two datasets, time period variables are created to join the cycling information using SAS. Then bicycle counts were summed up by each location, date, and time period. Finally, the two datasets were combined with the temporal relationship. The detailed data processing procedure can be seen in the following figure.

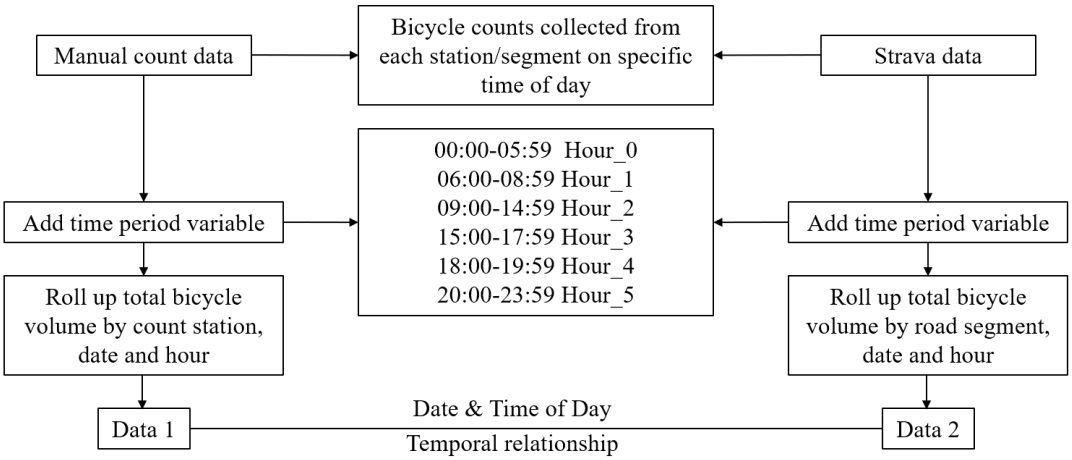

Figure 5-1 First Step of the Data Processing Procedure in SAS

Step 2:

This step was to combine the Strava shapefile data with the other supporting data to build spatial relationship between the datasets. First, a point layer that mapped the locations of the permanent continuous count stations was created. Second, all the datasets were joined together by spatial join in ArcGIS. Finally, all the spatial information was compiled in Data 4 as shown in the following figure.

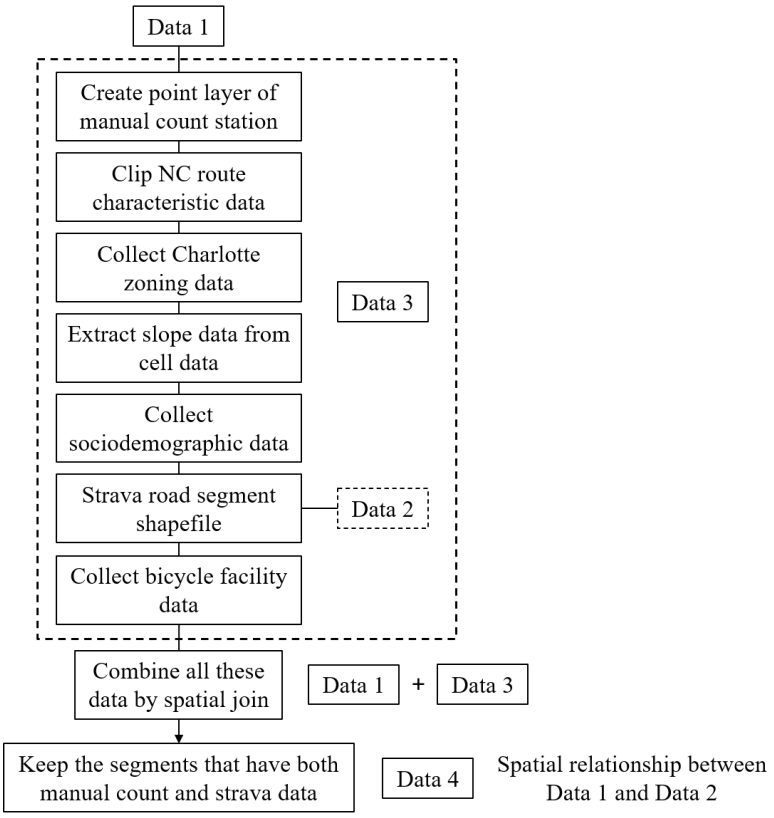

Figure 5-2 Second Step of the Data Processing Procedure in ArcGIS

Step 3:

This step was to generate the final data for model development. Basically, this step combined the final data obtained from the previous two steps to create both temporal and spatial relationships between the datasets. In addition, dummy variables including weekdays and time periods were added to the final dataset. The detailed data processing procedure for this step is shown in the following figure.

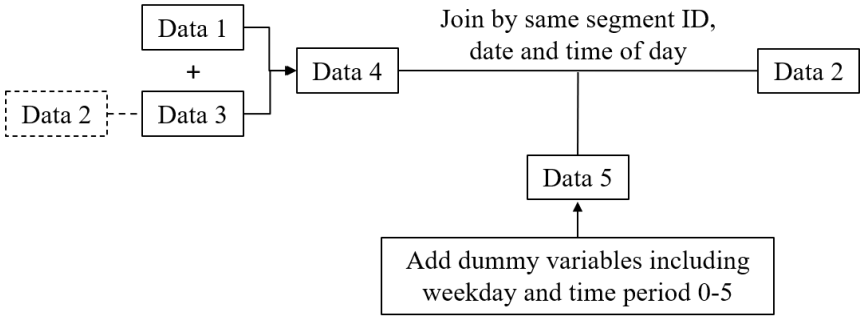

Figure 5-3 Third Step of the Data Processing Procedure in SAS

5.3 Bicycle Volume Regression Models

5.3.1 Simple Linear Regression Model

Based on the compiled data, a simple linear regression model was developed with the bicycle volume data collected from the permanent continuous count stations being the dependent variable, and the Strava bicycle counts being the independent variable. To conduct the model estimation process, SAS 9.4 was used. The model estimation results can be seen in Table 5-1.

Table 5-1 Simple Linear Regression Model Estimation Results

Variable	Label	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	5.72062	0.30014	19.06	<.0001
StravaCounts	Strava	4.45564	0.12772	34.89	<.0001
R-Square		0.3354	**Adj R-Square**		0.3353

Results revealed that total bicycle counts are about 4.46 times as high as the Strava counts on the same road segment. However, this model only shows a basic relationship between the two types of data, which can provide an approximate value of the actual bicycle counts on a specific road segment with the availability of Strava bicycle counts. In addition, the values of R square (0.3354) and adjusted R square (0.3353) are low. That probably means that one cannot simply estimate the bicycle counts from permanent continuous count stations only based on the available Strava data because the actual bicycle volume could be determined by many other factors that are not accounted for in this simple linear regression model. To estimate the bicycle volume on each road segment considering other relevant variables, a multiple linear regression model was developed.

5.3.2 Multiple Linear Regression Model

To examine the association between the bicycle volume collected from permanent continuous count stations and other relevant variables including Strava bicycle counts, road characteristics, geometry, sociodemographic data, zoning data, temporal data, AADT and bicycle facilities, a multiple linear regression model was formulated, and the variables considered in this model based on the literature review are presented in Table 5-2. All of the variables in Table 5-2 are included in the multiple linear regression model to identify the variables that have a significant impact on bicycle count.

$$\text{Bicycle Count} = f(N, G, S, Z, T, B, C) \qquad \text{Eq. 5-1}$$

where:

N = Road characteristics data which include speed limit, road segment length and number of through lanes.

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 greenways.

V = Annual average daily traffic.

C = Strava bicycle count.

Table 5-2 Variable Description

Variable Type	Variable Label	Description
Road Characteristics	Speed Limit	The posted speed limit on a roadway segment.
	Segment length	The length of the segment in miles.
	Through lane	The number of through lanes.
Geometry	Slope	The slope of a road segment at intersection.
Sociodemographic characteristics	TOTPOP_CY	Total population in each census block.
	MEDAGE_CY	The median age in each census block.
	MEDHINC_CY	Median household income in each census block.
Zoning	Residential	Charlotte zoning with residential land use.
	Business	Charlotte zoning with business land use.
	Mixed use	Charlotte zoning with mixed use land use.
Temporal Variables	Hour_0	If cycling time is during 00:00-05:59, then Hour_0 = 1.
	Hour_1	If cycling time is during 06:00-08:59, then Hour_1 = 1.
	Hour_2	If cycling time is during 09:00-14:59, then Hour_2 = 1.
	Hour_3	If cycling time is during 15:00-17:59, then Hour_3 = 1.
	Hour_4	If cycling time is during 18:00-19:59, then Hour_4 = 1.
	Hour_5	If cycling time is during 20:00-23:59, then Hour_5 = 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.
AADT	AADT	Annual average daily traffic.
Strava data	StravaCounts	Strava user count on a road segment.

The multiple linear regression model estimation process conducted in SAS 9.4, and the results are presented in Table 5-3.

Table 5-3 Multiple Linear Regression Model Estimation Results

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.28556	0.87176	0.33	0.7432
Hour_1	4.38905	0.77344	5.67	<.0001
Hour_2	7.62742	0.73108	10.43	<.0001
Hour_3	12.39803	0.74970	16.54	<.0001
Hour_4	12.87091	0.79261	16.24	<.0001
Hour_5	7.01667	0.87302	8.04	<.0001
Weekday	-5.50749	0.33801	-16.29	<.0001
StravaCounts	3.87088	0.10425	37.13	<.0001
Bike_Lanes	-5.07751	0.46008	-11.04	<.0001
Off_Street_Paths	11.77539	0.47767	24.65	<.0001
R-Square	0.6345	**Adj R-Square**		0.6340

Based on the model estimation results shown in Table 5-3, variables including five time periods from 6:00 am to midnight, weekday, Strava bicycle counts, the presence of a bike lane, and off-street path are highly associated with the bicycle volume data collected from permanent continuous count stations. The detailed analysis of the relationship between each explanatory variable and the actual bicycle counts is discussed below.

Time periods except 00:00 to 06:00 am have a positive impact on the actual bicycle volume on a specific road segment. This result indicates that the cycling activities of the bicyclists in the City of Charlotte start early in the morning and end late at night. Based on the negative impact of the weekday variable, it can be interpreted that bicycle volume during weekdays is lower compared to weekends, which is associated with the fact that bicyclists in the City of Charlotte prefer to bike on weekends. This result is probably related to the high proportion of recreational trips generated by Strava users. In addition, bicyclists may need to work during weekdays, which gives them less time for cycling compared with weekends. Therefore, weekdays show a negative impact on the bicycle counts on a specific road segment.

Bicycle facilities are critical impact factors on bicycle counts. From the model estimation results, bike lanes and off-street paths have different impacts on the actual bicycle volume. To be specific, off-street paths have a positive impact, while bike lanes have a negative impact. It can be inferred that bicyclists in the City of Charlotte prefer to bike on off-street paths compared with other bicycle facilities. In addition, bike lanes are always shared with other lanes, which may provide a negative impact on the actual bicycle volume on the specific road segment. The values of R square (0.6345) and adjusted R square (0.6340) of this multiple linear regression model are higher than those of the simple linear regression model, which indicates that this model is better than the previous one.

5.3.3 Bicycle Volume Prediction

Based on the model estimation results in the multiple linear regression model, a bicycle volume prediction on all the road segments in the network of Charlotte with the availability of Strava data and bike facility data can be calculated using the following equation:

$$V = \beta_0 + \beta_k X_k \qquad \text{Eq. 5-2}$$

where:

V = Predicted bicycle volume on each road segment.

β_0 = Constant term of the bicycle volume model.

β_k = Estimated coefficient associated with kth attribute based on the multiple linear regression model.

X_k = Attributes that have significant impacts on bicycle volume based on the multiple linear regression model.

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

Therefore, the AADB prediction can be calculated using the following equation:

$$AADB = V_T / 365 \qquad \text{Eq. 5-3}$$

Based on the AADB prediction, a map illustrating the predicted AADB on most of the road segments in the City of Charlotte is presented in the following figure. The predicted AADB are categorized into five levels, ranging from 0 to156 counts. The red lines represent the top level of AADB, which is 86.03 – 156.04, while the dark green lines represent the bottom level of AADB, which is 0 – 5.5. This map can potentially be helpful for strategic planning of bicycle facility management.

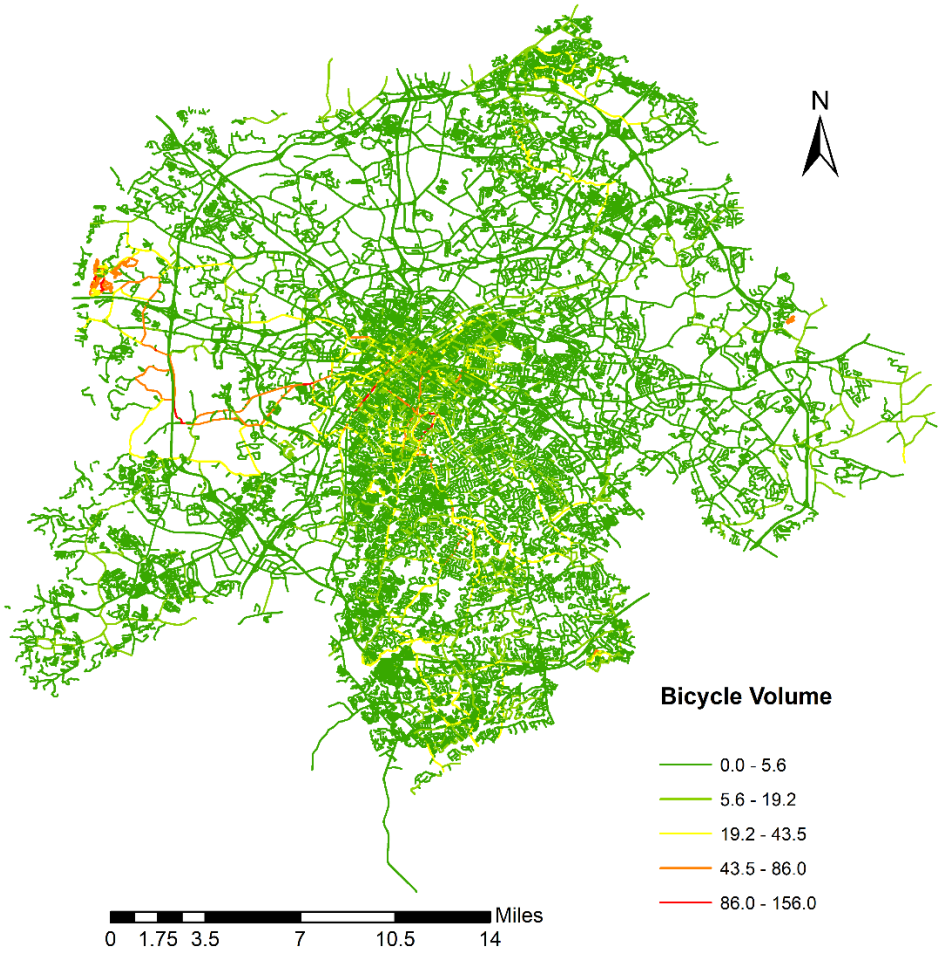

Figure 5-4 AADB Prediction in the City of Charlotte

Table 5-4 AADT Distribution in the City of Charlotte

Categories	AADB				
	0-5.6	5.6-19.2	19.2-43.5	43.5-86.0	86.0-156.0
Percentages	86.14%	9.35%	3.57%	0.81%	0.13%

5.4 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 were developed to quantify the relationship between actual bicycle count data and Strava bicycle data as well as other relevant variables. Model results were analyzed and predicted bicycle volume on most of the road segments in the City of Charlotte were calculated using the developed estimation model. In addition, a map illustrating the bicycle ridership in the City of Charlotte was also created. The cycling activities may vary for different cities/locations, which limits the transferability of the developed model to another place. However, if the efforts to conduct localized modeling is not possible due to budget or resource limitations, the model

can work as a good rule-of-thumb to provide estimates, and in that sense, can be generally applied to other locations as long as the data in Table 5-3 are gathered.

6 BICYCLIST INJURY RISK ANALYSIS

6.1 Introduction

This chapter develops a series of safety performance functions to analyze bicyclist injury risk. Section 6.2 provides the data preparation procedure for the later bicyclist injury risk analysis. Section 6.3 through Section 6.6 present the methodology for analyzing the impact of cycling safety including the Poisson model, the NB model, the ZIP model, and the ZINBmodel. Section 6.7 compares the model estimation results using the goodness of fit and summarizes the model results with the different impacts of various explanatory variables. Section 6.8 concludes this chapter with a summary.

6.2 Data Preparation

The data preparation procedure is similar to the data processing procedure presented in the previous two chapters. This process was conducted mainly using ArcGIS. The primary function used in ArcGIS was spatial join that helps researchers join multiple layers by the same location with different spatial and relevant information. Based on the literature review and the data availability, the following information including bicycle volume, bicycle-vehicle crashes, road characteristics, sidewalk information, bicycle facilities, bus stops, and AADT was collected for the model development of safety performance functions. The detailed data description and sources are presented in Table 6-1.

Table 6-1 Data Description and Sources

Data	Description	Sources
Strava	Bicycle volume data (December 2016 to November 2017) including bicycle counts on each road segment in the City of Charlotte and the Charlotte road network shapefile	Strava Metro
Bike Crashes	Bicycle-vehicle crashes occurred in the City of Charlotte from 2007 to 2017	NCDOT
Road Characteristics	North Carolina road characteristics	NCDOT
Sidewalks	The sidewalk information in the City of Charlotte	Charlotte Open Data Portal
PBIN	Bicycle facilities in North Carolina	NCDOT
AADT	Annual average daily traffic information in North Carolina	NCDOT

To obtain the final combined dataset including the information mentioned above, all the data were imported into ArcGIS and spatial join was used to identify the spatial relationships between each dataset. To be specific, the Strava road segment shapefile created based on the

OpenStreetMap was used as the base of all the spatial joins/table joins. First, layers including road characteristics, AADT, sidewalks, bus stops, and bicycle facilities were joined spatially to the base layer (Strava road segment shapefile). Second, Strava data including the bicycle volume on each segment and all the spatial joined layers were compiled together with the same road segment ID to obtain the combined road shapefile. Finally, each bicycle-vehicle crash was assigned to its closest road segment and the bicycle crash counts on each road segment were rolled up to generate the final complete data for the development of safety performance functions. The data preparation procedure can be seen in Figure 6-1.

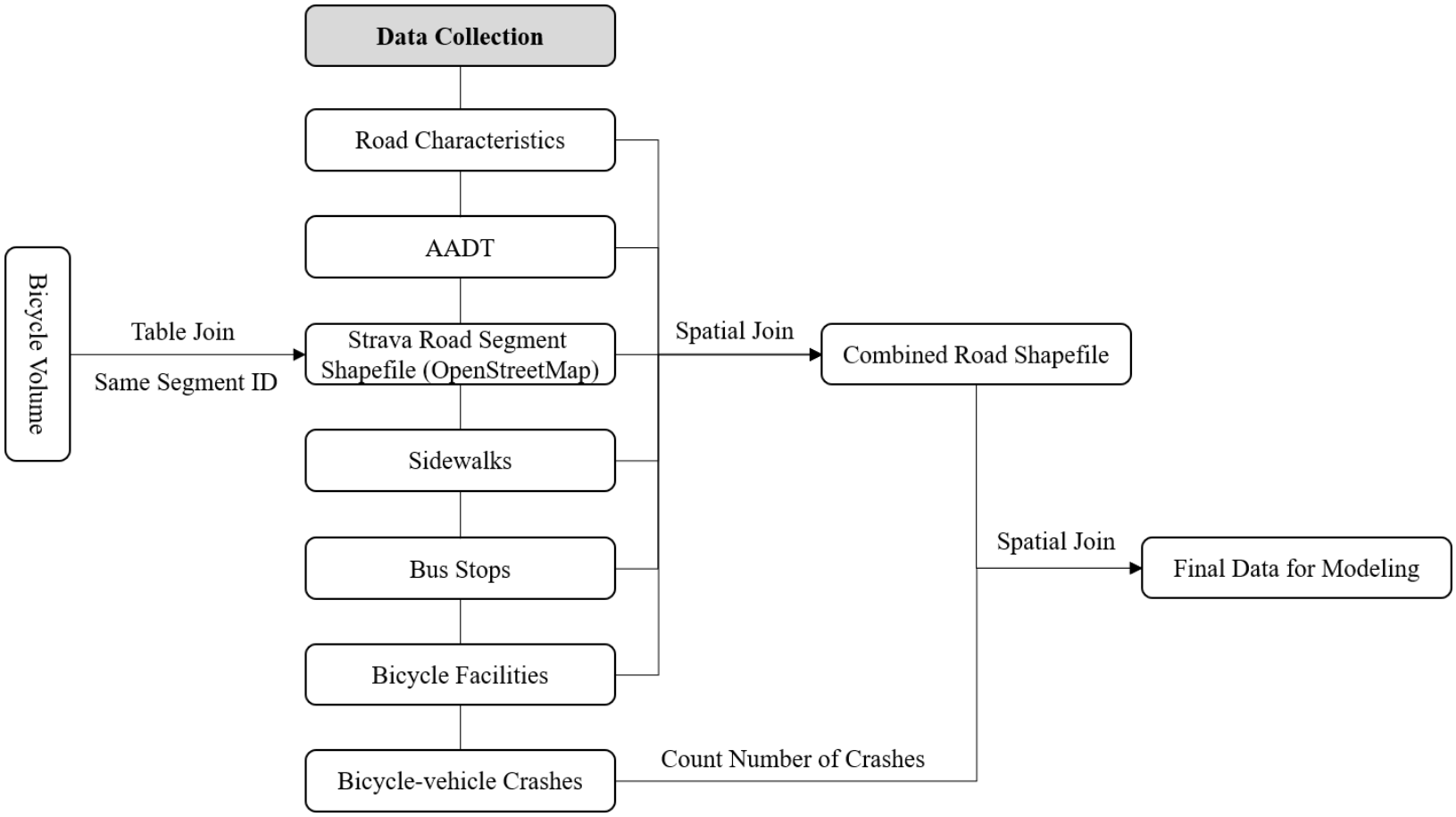

Figure 6-1 Data Preparation Procedure

Before using the combined data, it was imported into SAS to remove the observations with missing values and convert variables into dummy variables. The detailed explanatory variables considered in the following safety performance functions and their descriptions are presented in Table 6-2.

Table 6-2 Explanatory Variables

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

6.3 Poisson Model

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

$$VAR[y_i] = E[y_i] \qquad \text{Eq. 6-1}$$

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

$$\mu_i = EXP(\beta X_i) \qquad \text{Eq. 6-2}$$

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

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

$$P(y_i) = \frac{EXP(-\mu_i)\mu_i{}^{y_i}}{y_i!} \qquad \text{Eq. 6-3}$$

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

6.4 Negative Binomial Model

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

$$\mu_i = EXP(\beta X_i + \varepsilon_i) \qquad \text{Eq. 6-4}$$

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

$$VAR[y_i] = E[y_i]\left[1 + \alpha E[y_i]\right] = E[y_i] + \alpha E[y_i]^2 \qquad \text{Eq. 6-5}$$

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

$$P(y_i|X_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\mu_i}{\frac{1}{\alpha}+\mu_i}\right)^{y_i} \qquad \text{Eq. 6-6}$$

where Γ represents the gamma distribution function.

6.5 Zero-inflated Poisson Model

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

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

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

$$y_i = 0\ with\ probability\ p_i + (1 - p_i)\exp(-u_i) \qquad \text{Eq. 6-7}$$

$$y_i = y\ with\ probability\ \frac{(1 - p_i)\exp(-u_i)\,{u_i}^y}{y!} \qquad \text{Eq. 6-8}$$

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

$$VAR[y_i|X_i, Z_i] = u_i(1 - p_i)(1 + u_i p_i) \qquad \text{Eq. 6-9}$$

6.6 Zero-inflated Negative Binomial Model

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

The ZINB model is presented in the following equations:

$$y_i = 0\ with\ probability\ p_i + (1 - p_i)\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+u_i}\right)^{\frac{1}{\alpha}} \qquad \text{Eq. 6-10}$$

$$y_i = y\ with\ probability\ (1-p_i)\left[\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+u_i}\right)^{\frac{1}{\alpha}}\left(\frac{u_i}{\frac{1}{\alpha}+u_i}\right)^{y_i}\right] \quad \text{Eq. 6-11}$$

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

$$VAR[y_i|X_i, Z_i] = u_i(1-p_i)(1+u_i(p_i+\alpha)) \quad \text{Eq. 6-12}$$

6.7 Model Result Analysis

To analyze the bicyclist injury risk on road segments and explore the impact factors on the bicycle-vehicle crash counts in the City of Charlotte, several safety performance functions including the Poisson model, the Negative Binomial model, the Zero-inflated Poisson model, and the Zero-inflated Negative Binomial model have been developed. Explanatory variables (presented in Table 6-2) are carefully selected for model estimation based on the literature review and data availability.

All the explanatory variables presented in Table 6-2 were first included in the safety performance functions to analyze the probability of certain crash counts. The maximum likelihood estimation method was applied to estimate the model parameters. SAS 9.4 was used to conduct the model estimation procedure. To keep the variables that have significant impacts on the crash counts on the roadway segments, the backward selection method was used in the model estimation procedure. The final model results for the four safety performance functions with significant variables only are presented in the following tables.

Table 6-3 Poisson Model Estimation Results

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

Analysis of Maximum Likelihood Parameter Estimates						
Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Bike_Lane	0.6181	0.1103	0.4020	0.8342	31.43	<.0001

Table 6-4 Negative Binomial Model Estimation Results

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

Table 6-5 Zero-inflated Poisson Model Estimation Results

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

Analysis of Maximum Likelihood Parameter Estimates						
Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1.0395	0.1190	0.8064	1.2726	76.37	<.0001
AADB	-0.0004	0.0001	-0.0005	-0.0002	19.64	<.0001

Table 6-6 Zero-inflated Negative Binomial Model Estimation Results

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

To compare the four safety performance functions, the indicators for model comparison are adopted. Indicators used for model comparison include -2Log-likelihood, the Akaike's information criterion (AIC), and the Bayesian information criterion (BIC).

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

$$AIC = 2p - 2LL \quad \text{Eq. 6-13}$$

$$BIC = pln(Q) - 2LL \quad \text{Eq. 6-14}$$

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

Therefore, the indicators for each model (including the Poisson model, the Negative Binomial model, the Zero-inflated Poisson model, and the Zero-inflated Negative Binomial model) are presented in Table 6-7.

Table 6-7 Indicators for Model Comparison

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

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

Summarizing the model estimation results presented in Table 6-3, Table 6-4, Table 6-5, and Table 6-6, variables that have significant impact on bicyclist injury risk include annual average daily bicycle counts, interstate roads, principal arterials, minor arterials, major collectors, minor collectors, the presence of bus stops, and the presence of bicycle lanes. These are identified and will be interpreted in detail. The explanation of the impact of significant variables on bicyclist injury risk is provided below.

1. Volume variables:

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

2. Road characteristics:

Interstate roads, principal arterials, minor arterials, major collectors, and minor collectors all have significant impact on bicyclist injury risk. The functional classification of a road segment is the major impact on bicycle-vehicle crash counts. Interstate roads have a negative impact on bicyclist injury risk, while principal arterials, minor arterials, major collectors, and minor collectors affect cycling safety positively. This result indicates that the likelihood of crash counts

on principal arterials, minor arterials, major collectors, and minor collectors is higher, while the probability of crashes occurring on interstate roads is lower. State law prohibits bicyclists from traveling on interstate roads, which may explain the lower probability of cyclist crashes on interstate roads.

In addition, the presence of bus stops on a road segment has a positive impact on bicyclist injury risk, which indicates that the presence of bus stops may increase the probability of more bicycle-vehicle crashes. One can imagine that if a bus stop is located on a road segment, the conflict between bicyclists and buses may increase the likelihood of a bicycle-vehicle crash.

3. Bicycle facilities:

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

6.8 Summary

This chapter develops several safety performance functions including the Poisson model, the Negative Binomial model, the Zero-inflated Poisson model, and the Zero-inflated Negative Binomial model to analyze bicyclist injury risk. Model comparison was conducted to select the best model structure for this research study. The results show that the ZINB model has the best performance and therefore it is recommended for use. Factors that have a significant impact on the number of bicycle-vehicle crashes that occurred on road segments in the City of Charlotte were identified and interpreted. The transferability of the developed injury risk model is similar to the bicycle volume prediction models. The same ZINB may be used but the significant variable and their corresponding coefficients may vary for other locations with the availability of localized data. However, if the local data is not available, then the model can work as a good rule-of-thumb to provide estimates, and in that sense, can be generalized and transferred to other locations.

7 SUMMARY AND CONCLUSION

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

To increase the mode share of cycling, research studies are needed to explore the methods for estimating and predicting bicycle volume on a road segment in a city network and bicyclist injury risk. One of the most critical issues that needs to be considered is the data collection method. Traditional data collection methods including travel surveys and data from permanent continuous count stations can be time-consuming and expensive. Novel crowdsourced data can address the issues brought by traditional data collection methods and provide temporal and spatial information on cycling to bridge the data gap.

The primary objectives of this project are to validate the bicycle counts collected from counting machines, to determine the correction factors, to estimate and predict the bicycle volume on each road segment, and to conduct a cycling safety analysis. Based on the crowdsourced data collected from Strava, the descriptive analyses were conducted in terms of the demographic information on Strava users, cycling activities for different trip purposes, and the total cyclist count on each road segment in the City of Charlotte.

Crowdsourced bicycle data from the Strava smartphone application were combined with a series of other relevant data including NC road characteristics data, demographic data, slope data, annual average daily traffic, bicycle count data from permanent continuous count stations in the City of Charlotte, temporal data, the presence of a bus stop, bicycle facility data, etc. Data comparison was conducted to demonstrate the differences between actual bicycle count data and Strava bicycle count data. Data processing and combination procedures were completed using ArcGIS and SAS.

Based on the combined data, two linear regression models were developed. The relationship between actual bicycle count data collected by permanent continuous counters and Strava data as well as other relevant data was analyzed. To estimate the bicycle volume based on the linear regression model results, total bicycle counts are about 4.46 times as high as the Strava counts on the same road segment. However, this only shows a basic relationship between the actual bicycle count data and the crowdsourced bicycle data from Strava. The actual bicycle count data could be determined by many other factors that are not accounted for in the simple linear regression model. To be specific, variables including five time periods from 6:00 am to midnight, weekday, Strava bicycle counts, the presence of a bike lane, and off-street paths were found to be highly associated with bicycle volume on each road segment.

Comparing the values of R square and adjusted R square, the multiple linear regression model has higher values, which indicates better model performance than the simple linear regression model. According to the multiple linear regression model estimation results, it is more likely to have higher bicycle volume on weekends during the daytime. In terms of bicycle facilities, off-

street paths are the preferred ones in the City of Charlotte. Bicycle volume on most of the road segments in the City of Charlotte can be predicted using the developed multiple linear regression model. A bicycle ridership map was created to have a graphical view of the bicycle volume for the whole city network.

This research project also investigated the validation and correction factor calculation methodology used for bicycle and pedestrian count data collected by permanent continuous counters in the North Carolina Non-Motorized Volume Data Program (NC NMVDP). The analysis shows that bicycle and pedestrian count data is affected by rounding errors due to the application of correction factors at shorter time intervals, especially at lower volume bicycle counting sites and when correction factors are applied to hourly data or 15-minute data. Comparison of WAPD values among available validation studies shows that the average error of systems is consistent with previous research.

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

According to the bicycle volume estimation model results, the bicyclist injury risk analysis obtained, and conclusions made in this project, some policy-related recommendations are provided:

- Based on the modeling results which indicate that bicyclists prefer off-street paths, planners can design more off-street paths to offer better bike environments in the City of Charlotte.
- To promote biking to work, the locations of the off-street paths need to be constructed in the uptown area. There is much traffic in Charlotte’s uptown area, and the bicycle volume is higher there compared to other locations, especially during peak hours.
- According to the modeling results, the predicted bicycle volume on road segments in the vicinity of parks and greenways in the City of Charlotte is greater than the predicted bicycle volume on other road segments. In the Strava data, it was found that greenways and parks attract considerable non-commuter bicycle trips. To encourage recreational bicycle trips, the bicycle facilities in parks or greenway areas should be improved.
- It is important to identify the right of way on a roadway segment with bus stops. It is recommended that the city construct separated bike facilities for bicyclists to avoid crashes.
- If the above policy-related recommendations are followed, better bike environments and cycling safety can be provided for the citizens of Charlotte to improve their quality of life and to mitigate traffic congestion to some extent.

REFERENCES

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

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

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

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

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

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

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

Brabham, D. C. (2013). *Crowdsourcing*. Mit Press.

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

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

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

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

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

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

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

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

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

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

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

Estellés-Arolas, E., Navarro-Giner, R., & González-Ladrón-de-Guevara, F. (2015). Crowdsourcing fundamentals: definition and typology. In *Advances in crowdsourcing* (pp. 33-48). Springer, Cham.

Federal Highway Administration Office of Highway Policy Information (2014). "Traffic Monitoring Guide, Chapter 4 Traffic Monitoring for Non-Motorized Traffic," USDOT, Washington, DC.

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

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

Di Leo, G., Pietrosanto, A., & Sommella, P. (2007, July). Estimating Measurement Uncertainty of Traffic Monitoring Systems. In 2007 IEEE International Workshop on Advanced Methods for Uncertainty Estimation in Measurement (pp. 143-148). IEEE. Guan, H. Discrete choice Modeling. 2004.

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

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

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

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

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

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

Hudson, J. G., Duthie, J. C., Rathod, Y. K., Larsen, K. A., and Meyer, J. L. (2012). *Using smartphones to collect bicycle travel data in Texas* (No. UTCM 11-35-69). Texas Transportation Institute. University Transportation Center for Mobility.

Institute for Transportation Research and Education (ITRE). (2016). "North Carolina Non-Motorized Volume Data Program (NC NMVDP): Phase 1 Final Report," NCDOT, Raleigh, NC.

Jackson, S., Miranda-Moreno, L. F., Rothfels, C., & Roy, Y. (2014). *Adaptation and implementation of a system for collecting and analyzing cyclist route data using smartphones* (No. 14-4637).

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of transport geography*, *52*, 90-97.

Jessberger S. (2017) "Traffic Monitoring Statistics, Data Quality, Usage, and Integration: Data Quality and Equipment Calibration," Transportation Research Board, Washington, D.C.

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

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

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

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

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

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

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

LaMondia, J., & Watkins, K. (2017). Using Crowdsourcing to Prioritize Bicycle Route Network Improvements.

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

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

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

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

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

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

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

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

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

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

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

RenoTracks. *RenoTracks*. (2013). Available at http://renotracks.nevadabike.org/.

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

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

Ryus, P., Fergueson, E. Lausten, K., Schneider, R., Proulx, F., Hull, T. Miranda-Moreno, L., "Method and Technologies for Pedestrian and Bicycle Volume Data Collection," Transportation Research Board, Washington, DC, 2014.

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

San Francisco County Transportation Authority. The CycleTracks Smartphone Application. (2013). Available at http://www.sfcta.org/modeling-and-travel-forecasting/cycletracks-iphone-andandroid/cycletracks-smartphone-application.

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

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

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

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

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

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

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

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

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

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

Sun, Y., & Mobasheri, A. (2017). Utilizing Crowdsourced data for studies of cycling and air pollution exposure: A case study using Strava Data. *International journal of environmental research and public health*, *14*(3), 274.

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

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

Transportation Research Board (2007), "Traffic Monitoring Data: Successful Strategies in Collection and Analysis: A Workshop," Transportation Research Board, Washington, DC.

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

Vukovic, M. (2009, July). Crowdsourcing for enterprises. In *2009 congress on services-I* (pp. 686-692). IEEE.

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

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

Zimmermann, M., Mai, T., & Frejinger, E. (2017). Bike route choice modeling using GPS data without choice sets of paths. *Transportation research part C: emerging technologies*, *75*, 183-196.