BICYCLE SAFETY ANALYSIS:
CROWDSOURCING BICYCLE TRAVEL DATA TO ESTIMATE RISK EXPOSURE AND CREATE SAFETY PERFORMANCE FUNCTIONS
(Final Draft)

by

Haizhong Wang, Ph.D., Assistant Professor
Chen Chen, Graduate Research Assistant
School of Civil and Construction Engineering
Oregon State University, Corvallis, OR 97331

Yinhai Wang, Ph.D., Professor and Director
Ziyuan, Pu, Graduate Research Assistant
Civil and Environmental Engineering, University of Washington

Michael B. Lowry, Ph.D., P.E. Associate Professor
Department of Civil Engineering, University of Idaho

Sponsorship
Pacific Northwest Transportation Consortium

for
Pacific Northwest Transportation Consortium (PacTrans)
USDOT University Transportation Center for Federal Region 10
University of Washington
More Hall 112, Box 352700
Seattle, WA 98195-2700

In cooperation with US Department of Transportation-Research and Innovative Technology Administration (RITA)

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Pacific Northwest Transportation Consortium, the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

Technical Report Documentation Page			
1. Report No.	**2. Government Accession No.**	**3. Recipient's Catalog No.**	
4. Title and Subtitle **Bicycle Safety Analysis: Crowdsourcing Bicycle Travel Data to Estimate Risk Exposure and Create Safety Performance Functions**		**5. Report Date** **6. Performing Organization Code**	
7. Author(s) Haizhong Wang, Yinhai Wang, Michael B. Lowry, Ziyuan Pu, Chen Chen		**8. Performing Organization Report No.**	
9. Performing Organization Name and Address PacTrans Pacific Northwest Transportation Consortium University Transportation Center for Region 10 University of Washington More Hall 112 Seattle, WA 98195-2700		**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.** DTRT-13-G-UTC40	
12. Sponsoring Organization Name and Address United States of America Department of Transportation Research and Innovative Technology Administration		**13. Type of Report and Period Covered** Research 1/15/2015-12/15/2016 **14. Sponsoring Agency Code**	
15. Supplementary Notes Report uploaded at www.pacTrans.org			
16. Abstract Around 700 bicycle fatalities happen every year, and the number has steadily increased since 2009. At the same time Seattle, Wash., and Portland, Ore., are promoting cycling as a healthy transportation mode, so bicycle safety has become a more urgent concern for the Pacific Northwest area. The Highway Safety Manual provides an evidence-based approach – safety performance function (SPFs) to evaluate safety for common traffic, but not, however, for bicycles. Therefore, a data-driven and evidence-based bicycle safety evaluation method is needed specifically for the Pacific Northwest region. Based on the first bicycle SPFs created by Krista Nordback in 2013, we used STRAVA bicycle count data (a type of crowdsourced bicycle travel data), other traffic count data, and bicycle crash data to establish Pacific Northwest SPFs in terms of bike count and crash frequency for road intersections. The SPFs demonstrated the relationship between crash frequency and traffic and bike volumes. The intersections with higher traffic volumes had higher bicycle crash frequencies. In addition, this project improved the usability of a GIS tool that had been created during a previous PacTrans project to estimate bicycle exposure. States DOTs and other agencies can use the SPFs to screen and identify previous bicycle black spots in the Pacific Northwest region in order to optimize safety investments.			
17. Key Words Bicycle safety, SPFs, Crowdsourcing data, GIS, STRAVA		**18. Distribution Statement** No restrictions.	
19. Security Classification (of this report) Unclassified.	**20. Security Classification (of this page)** Unclassified.	**21. No. of Pages**	**22. Price** NA

Contents

Figures

Tables

CHAPTER 1 INTRODUCTION

1.1 Problem Statement

In 2012, 726 bicyclists were killed in crashes with motor vehicles in the United States, and there were 25 bicyclist fatalities in the Pacific Northwest (Alaska, Idaho, Oregon, and Washington). The USDOT reported that bicycle fatalities "have steadily increased since 2009," as shown in Figure 1-1 (USDOT 2014). Recent studies have indicated that cyclists are 12 times more likely to be killed per distance traveled (Beck et al., 2007) than automobile occupants. The U.S. rate of bicycle fatalities is double that of Germany and triple that of the Netherlands, both in terms of number of trips and in distance travelled (Pucher and Dijkstra, 2003). Yet despite the dangers, individuals are increasingly choosing to bike throughout the country and especially in the Pacific Northwest (Milne and Melin, 2014).

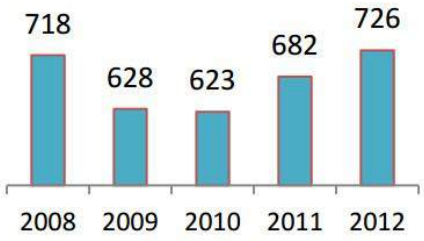

Figure 1-1 National bicycle fatalities in traffic crashes, 2009-2012 (Source: USDOT, 2014)

Engineers and planners face three interrelated challenges when conducting safety analysis for bicyclists. The first is the problem of insufficient data about bicycle crashes, specifically "near miss" crashes or where bicyclists are choosing to ride or not ride because of perceived safety concerns. Schimek (2014) suggested that as many as 89 percent of bicycle accidents go unreported since they often do not incur insurance claims or traffic violations. The second problem is the lack of tools for estimating bicycle volumes. Traditional travel demand models, such as the ubiquitous 4-step model, produce very poor results for bicyclists, and without reliable volume information it is very difficult to prioritize accident locations. For example, an intersection with only a few crashes might deserve highest priority if it exhibited a high *crash rate* (crashes/volume). Third is the lack of tools to analyze proposed improvements. For highway and other road projects, engineers can use safety performance functions (SPF) to predict the expected number of automobile crashes for a given location and compare how different improvement projects might reduce accident rates by using crash modification factors (CMF). The current Highway Safety Manual does not include any SPFs or CMFs for bicycles (AASHTO, 2010).

1.2 Background

Secretary Anthony Foxx has declared bicyclist safety a top priority for the USDOT and two months ago it launched what is being called "the most innovative, forward-leaning, biking-walking safety initiative ever" (Foxx, 2014). The initiative will include increased funding for bicycle infrastructure and research (USDOT, 2014).

In the past, transportation data have largely been collected by expensive fixed mechanical sensors and manual observation, both of which appear increasingly archaic in a world accustomed to mobile communications, instantaneous information sharing, and massive, low cost data collection. In particular, the rise of mobile computing presents an incredible opportunity for extracting vast quantities of useful data for transportation planning and management, without the need to maintain large networks of sensors. Many public and private agencies have identified "crowdsourcing," i.e., outsourcing of information gathering to the public, as a rich, low cost, and highly scalable framework for data acquisition and problem solving (e.g., Krykewycz et al., 2011; Jin et al., 2013). A search of relevant literature provides no universally agreed upon definition of crowdsourcing, but it is generally regarded as an online participatory activity initiated by an organization to engage a group of individuals (i.e., the public or a subset thereof) in the completion of voluntary tasks toward the collaborative resolution of a problem (Estellés-Arolas and González-Ladrón-de-Guevara, 2012). Thus, in the context of transportation, the "crowd" consists of consumers or users of the transportation facility in question, or those who are otherwise invested in the issue at hand (Brabham, 2013). The problem to be addressed, then, usually takes the form of some planning or management research question identified by a public agency or consultant that can benefit from the input of a large number of individual users (Molina, 2014). User input could include self-reported travel data, attitudes and ideas regarding current and planned infrastructure elements, and user reported traffic and infrastructure status information.

A number of recent studies have demonstrated a range of possible applications of crowdsourcing in transportation planning and management. For example, Krykewycz et al. (2011) developed a crowdsourcing framework for evaluating and mapping bicycle level-of-service data in Mercer County, New Jersey. This project demonstrated the benefit of large-scale voluntary user interaction in valuing inherently subjective measures of bikeability and identified several strategies for facilitating active participation by stakeholders. Hood et al. (2011) developed the CycleTracks smart phone application to collect cyclist trajectory and trip purpose data in the San Francisco area using built-in GPS capabilities. The resulting data were used to develop a route choice model, and to estimate marginal rates of substitution for specific features and link characteristics. The CycleTracks application is now in use in the Seattle, Washington, area, providing an excellent data resource for the efforts described in this proposal. This project expanded on previous work to addresses two key challenges in the application of crowdsourcing to non-motorized transportation decision making. First, this project built on the work described in Hood et al. (2011) to enhance a set of generalized tools for the analysis of cyclist exposure

data. Second, a framework for the application of crowdsourced bicycle incident and hazard data to the analysis of user risk exposure was demonstrated.

In a previous 2014 PacTrans project, the University of Idaho (UI) team developed an innovative new method for estimating bicycle volumes and calculating bicyclist exposure to dangerous situations. The team conducted a proof-of-concept-test to compare existing and expected exposure rates for the proposed Bicycle Master Plan for Bellingham, Washington. McDaniel et al. (2014) described the bicycle volume estimation method. Figure 1-2 shows estimated annual average daily bicyclists (AADB) for Bellingham, and Table 1-1 compares exposure rates for bicyclists. This PacTrans project advanced and expanded this new method in various ways, including tailoring the results for the creation of SPFs.

Figure 1-2 Estimated bicycle volumes for Bellingham, Wash.

Table 1-1 Dangerous situation exposure at intersections

Dangerous Situation	Conditions and Thresholds	Scenario 1: Existing Conditions (AADB)	Scenario 2: w/Proposed Improvements (AADB)	Change (AADB)	Percent Change (%)
Hazardous crossing	cross street: > 8,000 AADT, > 50 mph, > 10% heavy vehicle	31,595	33,297	+1,702	+5
Oncoming cross	oncoming left-turning AADT > 2,000	45,577	42,516	-3,061	-7
Right hook	right turning vehicles > 2,000 AADT	51,603	47,737	-3,866	-7
Left sneak	adjacent vehicles > 8,000 AADT oncoming vehicles > 8,000 AADT	9,015	8,798	-217	-2

Safety performance functions (SPFs) are statistical regression models used to estimate the predicted average crash frequency for a specific site type (e.g., an individual roadway segment or an intersection) through the mathematical relationship between frequency of crashes and the most significant causal factors. The Highway Safety Manual (HSM) documents how to predict crashes at similar intersections or road segments by using the SPF as a base and adjusting it with "crash modification factors" based on the specific geometrics or other features of the location (Nordback et al., 2014). The current HSM (2010) includes predictive methods primarily for motor vehicles on rural two-lane, two-way roads, rural multilane highways, intersections, road segments, and urban/suburban arterials, but there is no bicycle-specific safety performance function included. In Oregon, Dixon et al. (2012, 2013) developed specific safety performance functions for driveways and roundabouts to quantify their safety improvement potentials.

In recent years, non-motorized transportation, specifically cycling, has been promoted by authorities in the Pacific Northwest—Seattle, Wash., and Portland, Ore., in particular—as an alternative, healthy mode of travel. The major challenges for developing bicycle-specific SPFs are twofold: insufficient bicycle crash data and bicycle volume data on a wide range of bicycle facility types (Nordback et al., 2014).

Even if bicycle SPFs are developed from other locations, all associated SPFs for these facility types should be calibrated when they are applied to a different location. The bicycle ADT and volume exposure data are typically heterogeneous in nature; as a result, the variance is usually significantly different from the mean, which causes an over-dispersion issue (AASHTO, 2010). Therefore, the negative binomial model was used to develop the bicycle safety performance function in this research. A list of significant variables needs to be identified to develop a bicycle safety performance function, including bike ADT, pedestrian volume (exposure), number of left turn lanes, presence of bike lanes, presence of bus stops, etc.

This project built on a previous 2014 PacTrans project, "Data Collection and Spatial Interpolation of Bicycle and Pedestrian Data," led by the UI team in collaboration with the UW team. In that project, a GIS tool was created to analyze exposure to dangerous situations for bicyclists and tested with case study data from Bellingham, Washington. This PacTrans project improved the tools by incorporating crowdsourced bicycle data and tailoring the analysis and results for SPFs. This project complements a project led by the OSU team on "Risk Factors for Pedestrian and Bicycle Crashes" with Oregon Department of Transportation, and the five-year (2007 – 2012) geo-coded crash data set from ODOT and the risk factors identified from this project will support this PacTrans project in the bicycle safety performance functions development process.

1.3 Project Objectives, Approach/Method, and Report Organization

This project created tools, guidelines, and repeatable processes that engineers and planners can use to

- analyze crowdsourced bicycle data
- calculate bicycle exposure to dangerous situations
- create and analyze safety performance functions for bicyclists.

This project was divided into three tasks, which were completed by three institutes: University of Washington, University of Idaho, and Oregon State University. Each team was responsible for specific milestone deliverables throughout the project. The tasks were interconnected so as to address the three interrelated challenges defined in the problem statement. The intermediate results for each task were crosschecked by the other teams, as the results from one task served as inputs to other tasks. The UW team's analysis of crowdsourced data was used by the UI team to estimate bicycle volumes (exposure) and enhance GIS tools, which were used by the OSU team to create SPFs. The actual subtasks were slightly different from the original plan, and the subtasks for each institute are summarized below. Some tasks required cooperation between institutes, and because of too much detail, the cooperation process is not described here:
University of Washington:

- Create tools for analyzing crowdsourced bicycle data
- Conduct literature review on crowdsourcing non-motorized travel data
- Prepare and preprocess GIS bicycle dataset
- Develop framework for crowdsourcing incident and hazard data
- Write report documenting work and findings

University of Idaho:

- Create tools to calculate bicycle cxposure
- Conduct literature review on bicycle accident exposure
- Prepare GIS data (bicycle counts, street network, etc.)

- Use OD-centrality to estimate bicycle demand
- Analyze dangerous situation metrics
- Write report documenting work and findings.

Oregon State University

- Create bicycle safety performance functions
- Conduct literature review on bicycle SPFs
- Synthesizing the data and other sources
- Develop the state-specific bicycle SPFs
- Calibrate and validate the bicycle SPFs
- Write report documenting work and findings
- Combine and finalize report.

The two case study communities for this project were Seattle, Wash., and Portland, Ore. They both represent urban metropolitan areas, and Seattle has larger size than Portland. Seattle is the largest city in the Pacific Northwest region of North America. The population of Seattle is about 652,405, with around 3.6 million in the greater Seattle Metropolitan area. The bicycle mode share for Seattle rose to 22 percent in 2011, and it is expected to increase. In Chapter 2, literature reviews of bicycle risk exposure, SPF, and crowdsourcing data are described. Chapter 3 describes how engineers improved the usability of a GIS tool that had been created during a previous PacTrans project. Engineers and planners can use the GIS tool to estimate bicycle exposure when they conduct safety analyses. The original tool had some weaknesses. Engineers improved the tool by (1) streamlining the tool's operation and (2) decreasing the tool's computer execution time. Chapters 4 and 5 describe how, based on the first bicycle SPFs created by Krista Nordback in 2013, we used STRAVA bicycle count data (a type of crowdsourced bicycle travel data), other traffic count data, and bicycle crash data to establish Pacific Northwest SPFs, especially for Portland and the Seattle Metropolitan area, in terms of bike counts and crash frequencies for intersections. Other models are were applied, and the best model for SPF was chosen. The last chapter contains conclusion about what was done and provides recommendation for future work.

CHAPTER 2 LITERATURE REVIEW

2.1 Crowdsourcing Literature

Since Jeff Howe coined the term "crowdsourcing" in 2006 (Howe, 2006), it has become a hot topic in the world. Although it is hard to find a universal definition for this relatively recent concept, Enrique Estelles-Arolas provided a consistent and exhaustive definition after qualitatively researching the main elements of crowdsourcing: crowd, initiator, and process (Estellés-Arolas, 2012). Generally, crowdsourcing is regarded as an online participatory activity initiated by an organization to engage a group of individuals who vary in knowledge, heterogeneity, and number in the completion of voluntary tasks toward the collaborative resolution to a problem, providing mutual benefit.

Four problem-based approaches are identified as the most useful for governments as the crowdsourcing initiator: knowledge discovery and management, distributed human intelligence tasking, broadcast search, and peer-vetted creative production (Brabham, 2013). These approaches aim to address different kinds of problems and have distinct requirements for crowd participation. Knowledge discovery and management tasks crowds with gathering and collecting information or data into a common location and format; distributed human intelligence tasking is appropriate for analyzing large amounts of information; broadcast search is meant to handle empirical problems such as scientific problems; and peer-vetted creative production is used to solve problems related to matters of taste or market support, such as design and aesthetic issues. In order to assess whether crowdsourcing is an appropriate tool and which approach is most useful for the problem h to be addressed, scholars developed a framework to evaluate the appropriateness of crowdsourcing for governance. Figure 2-1 provides an illustration of assessing logic. The first question should be whether the problem that needs to be addressed is related to an information management task or ideation task. With an information management task, the researchers need to determine whether the task is more relevant to locating and assembling information or to existing information analysis. The former type of task is best suited to knowledge discovery and management; the latter is suited to distribute human intelligent tasking. With an ideation problem, the question is whether it requires empirical experience or relates to taste and popular support. The former is best suited to broadcast search, and the latter is appropriate for peer-vetted creative production.

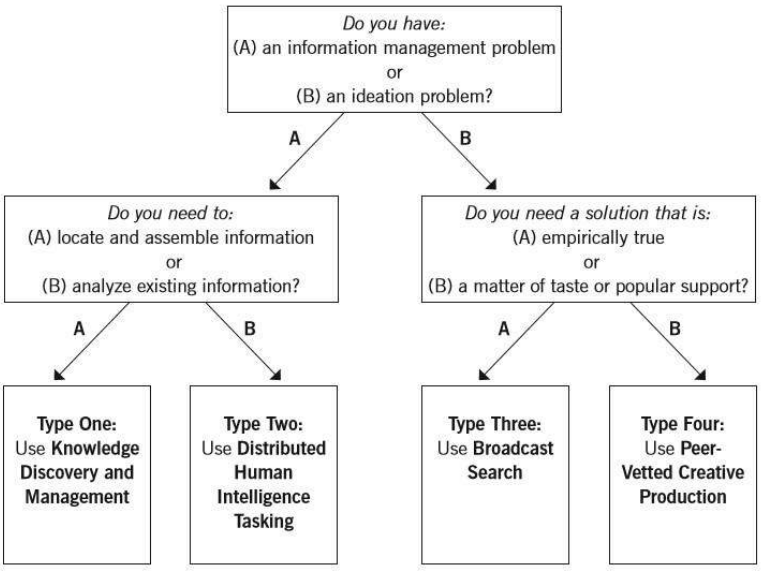

Figure 2-1 Decision tree for crowdsourcing method selection (Source: Brabham, 2013)

Planners and managers are continuously hindered by a lack of bicycle related data, such as bicycle count data, crash data, and trip data. Therefore, crowdsourcing has been being used as a powerful tool to address the data issue because of its capability to gather and collect information through the effort of numerous individuals (Molina, 2014). The Illinois Department of Transportation (IDOT) and Chicago Police Department (CPD) developed a web-based application called Chicago's Bicycle Crash Map for published bicycle crash data and also for collecting bicycle crash and near-miss data from bicyclists' self-reporting (Quartuccio, 2014). The San Francisco County Transportation Authority (SFCTA) developed a phone application called CycleTracks that allowed bicyclists to log their bike trip routes conveniently. It also allowed route choice, trip purpose and demographics data to be collected (Charlton, 2010). Previous research has summarized the application of crowdsourcing to bicycle planning projects (Molina, 2014). Molina (2014) categorized bicycle projects into five main types in which crowdsourcing was implemented, including facility demand, network planning, bike safety, suitability, and route demand modeling. In this research, ten projects were selected by two standards, one being that they were representative of the main five project types and the other that the projects provided available documents online or in print. The main purpose of these ten projects was data collection, and half of them had the goals of gathering the preferences of specific problems. For example, Capital Bike Share project developed a web-based application to collect the preferences for bicycle parking locations by users' votes. Web-based applications and smartphone applications are the most used tools for crowdsourcing in the bicycle planning field.

Several scholars have evaluated the effectiveness and quality of crowdsourcing as a bicycle-related data collection method. Ben Jestico compared data from the crowdsourced fitness app provided by STRAVA.com to those from manual cycling counts in Victoria, British Columbia, to evaluate the representative degree of crowdsourced fitness data for ridership

(Jestico, 2016). The results indicated that the crowdsourced fitness data provided by STRAVA.com were biased in representing the ridership, but they still could be used to predict categories of ridership and map spatial variation. Watkins compared two data sets collected from two smartphone-based apps, Cycle Atlanta and STRAVA, to explore the method of mapping cyclist movements in an urban area by using GPS data (Watkins, 2016). Based on the study's outcomes, the author suggested that the smartphone-based data were likely biased and that the apps could complement but not replace large-scale bicycle volume counting programs.

2.2 Bicycle Crash Literature

A bicycle crash is defined as an event in which the bicyclist hits the ground, a motor vehicle, road infrastructure, or any other solid object for a reason that leads to damage to body or property (Lindman, 2015). Scholars and planners have put a lot of effort into researching bicycle safety problems from several perspectives, such as crash type, severity, and data collection.

2.2.1 Crash Types

A better understanding of bicycle crash types and characteristics would be useful for planning and policy making. The Federal Highway Administration (FHWA) developed a course on bicycle and pedestrian crash types that provides detailed descriptions of crash characteristics, crash rates, exposure, and a grounding in crash typing to engineers, planners, scholars, and law enforcement personnel for clarifying their understanding of how crashes occur and how to avoid them (Hunter, 1996). The National Highway Traffic Safety Administration (NHTSA) summarized the most common bicycle types in research on how to prevent bicycle crashes. Although most previous research has emphasized bicycle-motor vehicle crashes, Paul Schepers conducted an analysis on single-bicycle crash types and characteristics by using a questionnaire study (Schepers, 2012). Single-bicycle crashes were categorized into four types: infrastructure-related crashes, cyclist-related crashes, bicycle malfunctions, or unknown. Table 2.1 lists the most common bicycle crash types by summarizing several studies focused on bicycle crash types (Schepers, 2011).

Table 2-1 Bicycle crash type summary

No.	Crash Type	Description
1	Bicyclist or motorist rides through stop sign or red light	The bicycle or the motorist fails to follow the rules of the road including obeying all signs and signals
2	Wrong way riding	The bicyclist ride on the road or sidewalk against the flow of traffic
3	Bicyclist turns left in front of traffic	The motorist turns right
4	Bicyclist enters road from a driveway, alley, curb or sidewalk	The bicyclist fails to stop, slow and look before entering a roadway from a residential or commercial driveway
5	Motorist passes a bicyclist	A motorist fails to see and avoid the bicyclist until it is too late to avoid a collision
6	Motorist turns right or left into bicyclist	The motorist takes a right or left turn, and the bicyclist rides in either the same or opposing direction
7	Motorist enters road from a driveway or alley	The motorist fails to stop and look before entering a roadway
8	Multiple threads	The bicyclist fails to clear the intersection before the light turns red.

2.2.2 Crash Injury Severity

Generally, bicycle crash injury severity is divided into four levels: fatal injury, incapacitating injury, non-incapacitating injury, and possible/no injury (Reynolds, 2009). Previous research analyzed the factors that affect the level of crash injury severity from several distinct perspectives. Margaret, Attewell and Thompson developed a regression model to analyze the effectiveness of helmets for reducing the injury severity levels of bicycle crashes (Thompson, 1990). The results indicated that bicycle helmet use has a significantly protective effect in reducing crash injury severity, especially for reducing head, brain and neck injuries (Attewel, 2001; Dorsch, 1987).

Several scholars have conducted research to explore the factors that highly affect the injury severity of bicyclists in bicycle-motor vehicle accidents by using regression-based models on police-reported crash data, and the results showed that several factors more than double the probability of fatal injury in a bicycle-motor vehicle accident, including darkness with no streetlights, inclement weather, peak hour in the morning, head-on and angle collision, speeding

involved, vehicle speeds of about 48.3 km/h, a truck involved, bicyclists age 55 or more, roads without a median/division, running over the bicyclist, etc. (Yan, 2011). They also highly recommended the installation of medians, division between the roadway and bikeway, improvements in illumination on road segments, and speed limit reductions.

Additionally, a few previous studies have examined the influence of factors related to bicycle crash injury severity in specific situations. Jeremy R. Klop examined the physical and environmental factors that influence the injury severity level of bicycle crashes on two-lane, divided roadways. An ordered probit model was used to model the crash and inventory data collected by the North Carolina Highway Safety Information System. Analysis results showed that the factors of straight grades, curve grades, darkness, fog, and speed limit would heavily increase injury severity, and the factors of higher AADT, dark conditions with street lighting, and the interaction between speed limit and shoulder-width decrease injury severity level (Klop, 1999). Morteza Asgarzadeh explored the impacts of intersection and street design on the injury severity of bicycle-motor vehicle crashes (Asgarzadeh, 2016). A multivariate log-binomial regression model was used to model 3,266 BMVC data from New York City police records, which included geographical information and latitudes/longitude data. The research found that 1) crashes at non-orthogonal intersections and crashes at non-intersection street segments had a higher risk of severe injury than crashes at orthogonal intersections; 2) crashes with trucks involved and buses involved had more probability of resulting in a higher injury severity level; and 3) there was no relationship between street width and injury severity level.

2.2.3 Data Collection

Traditionally, police reports and hospital reports are the main sources for bicycle crash data. Most previous research on bicycle safety has used these two data source as their research input. Given that a quantitative methodology is widely used in bicycle safety research and that its performance is highly dependent on data accuracy and coverage, several scholars have conducted analyses to examine the data quality of both data sources. Agran (1990) compared police reported data to the data provided by a hospital monitoring system for children under 15 years old injured as pedestrians and bicyclists by bicycle-motor vehicle crashes in Orange County, Calif. The comparison was conducted with respect to demographics and circumstances by using cross-tabulations of corresponding variables. The results indicated that the under-reporting rate for bicyclists was a conservative 10 percent and was composed of non-traffic cases. The author also claimed that police agency reporting requirements led to under-reporting and that the police injury severity scale barely correlated with a scale based on medical diagnoses. Stutts (1990) compared data reported from North Carolina hospital emergency rooms during the summers of 1985 and 1986 with the police-reported bicycle accident data from the North Carolina state government for the same periods. The results were that only 10 percent of the data overlapped; whereas more than 60 percent of bicyclists were 5 to 14 years old, and 70 percent were male in the hospital-reported data, in the police-reported data set less than half the crashes involved bicyclists under 15 and 85 percent were male. In addition, almost all of the police-reported accidents involved motorists, but less than a fifth in did in the hospital-reported

data set. As seen from the result of the literature review, both police-reported and hospital-reported bicycle crash data would cause bias. Scholars and professionals have made efforts to explore the reasons for the biased data and have also developed new reporting methods to improve data quality. It has been suggested that incident form design affects the quality of police-reported crash data. Lusk (2014) redesigned the police reporting template by adding bicycle-crash-scene coded variables that included four bicycle environments, 18 vehicle impact-points, motor vehicle/bicycle crash patterns, in/out of the bicycle environment, and bike/relevant motor vehicle categories and used multiple logistic regression to test the crash data, which were improved by redrawing the crashes and entering the new bicycle-crash-scene details into a corresponding data spreadsheet. The author suggested that bicycle-crash-scene codes should be included in police reporting templates so that crash analysis can be conducted with big data methodology on several novel topics of bicycle safety (Lusk, 2014).

2.3 SPF Literature

SPFs are crash prediction models (Federal Highway Administration, 2013). They are essentially mathematical equations describing the number of crashes of various types and site features, and they always include traffic volume AADT but also may include other site features, for example lane width, horizontal curves, the presence of turn lanes, etc. These models can be used in network safety screening, determining the safety impact of design changes, evaluating the effects of engineering treatments, and so on (Federal Highway Administration, 2013).

Motor vehicle SPFs for normal roadway types are established in the Highway Safety Manual (HSM). These provide an evidence-based tool to estimate motor vehicle crashes by traffic volume and other factors that can influence the results (American Association of State Highway and Transportation, 2010; Nordback, Marshall, & Janson, 2014). However, few studies have addressed SPFs for estimating bicycle crashes.

Nordback et al. 2014 created SPFs for bicycles and applied this method to Boulder city in Colorado. The authors used collision, AADT, and AADB data to build the function describing the relationship between traffic and bicycle volumes with crash frequency at intersections. They found that with the bicycle and motor vehicle volume increases, the frequency of cyclist crashes increases but the crash rate decreases. In other words, at intersections the cyclist crash frequency has a positive relationship, whereas the cyclist crash rate has a negative relationship with traffic and bicycle volumes. This relationship has been previously studied by others has been found to be not linear and is called “safety in numbers” (Ekman, 1996; Jacobsen, 2003; Jonsson, 2005; Nordback et al., 2014; Robinson, 2005).

This relationship, which captures the number of crashes and the exposure to crashes, is known as SPF, which is a more efficient method for prioritizing intersections (Kononov & Allery, 2003; Nordback et al., 2014). This method is also a useful tool for prioritizing segments. Nordback et al. established the process and method of creating SPFs for bicyclists by using a negative binomial generalized linear model with log link, and this model was based on annual average daily traffic (AADT) and annual average daily bicycle (AADB) data. The authors

compared negative binomial regression and poisson regression, and they found that the former can fit the data better because in the collision data sets has feature of variance triples the mean; in other words, the crash data are over-dispersed. In Poisson distribution, the mean equals the variance, but when the variance is larger than mean, the situation is called over-dispersion (Federal Highway Administration, 2013). The three peak hours counted for both bicycles and traffic, provided by the city of Boulder (Boulder & Even, 2012), were adjusted to AADT and AADB by using daily and monthly factors (Ferrara C, 2001). Negative binomial distribution was determined by Long (1997).

Nordback et al. did a sensitivity analysis on changes in AADB. The results showed that with higher AADB, the corresponding parameter was still well under 1, which indicated that the SPF was still sublinear; whereas the parameter for lower AADB was closer to zero, which indicated that the AADB was not an important factor in determining motorist-cyclist crashes. Further analysis should investigate this observation. In addition, the estimations of the parameters for AADT and AADB were at the same magnitude, indicating that the collisions were similarly sensitive to both volumes; however, the AADT exponent was one or two orders of magnitude higher than AADB exponent, so the change of AADB had more critical influence on crashes than same change in AADT. Therefore, getting accurate estimations of bicyclist volumes is more important for analyzing SPF (Nordback et al., 2014). Future work can use larger data sets and more accurate AADB and include facility type in the analysis.

This analysis only captured the connection between volumes and crashes but did not reveal the causation between them. In other words, the reasons connecting traffic and bicycle volumes and crash frequency were not explained. The reasons could be that increasing bicycle volumes may lead to safer motorist and bicyclist behavior; or more bicyclists may be riding on safer facilities. Other studies stated that more bicyclists trigger changes in driver behavior, but the conclusion was based on logical speculation not empirical data analysis (Ekman, 1996).

Other characteristics of the road have been studied, such as bicycle infrastructure, bicycle lane width, street light, and the angle of grade that may influence the crash-volume relationship (Reynolds et al., 2009). In Dolatsara's (2014) study, crash data, volume data, and road geometric data were collected. Traffic and bicycle volumes were provided by the Department of Transportation; geometric data included different lane numbers, bike lane characteristics, posted speeds, bus stops, and so on; the crash data collected came from 164 intersections in four cities in Michigan, and crashes happened within a 500-ft buffer of the center of an intersection (Dolatsara, 2014). Intersection traffic volumes were collected by combining four directions of ADT. The 500 feet were calculated by Stopping Sight Distance (SSD) by Fambro et al, (1997). However, more practically, 250 feet has been used as a diameter for assigning crashes to an intersection (Dolatsara, 2014; Vogt & Bared, 1998). Justifying which threshold should be used to assign the crashes was critical for this project, since that threshold directly influences crash frequency. Portland and Seattle both have relatively small street blocks, so with the crash-determining buffer diameter increasing, more crashes could be incorrectly assigned to intersections. In other words, the error could be large. So the 250-ft threshold as used in our SPF project to identify the intersection where the crash happened.

Dolatsara also mentioned that the Poisson distribution cannot capture the over-dispersion of crash data (American Association of State Highway and Transportation, 2010), so the negative binomial regression was employed (Dolatsara, 2014). The significant variables included in the SPF were ADT, number of left turn lanes, presence of bike lanes, and presence of bus stops (Dolatsara, 2014). This suggests that engineers may include other factors besides traffic and bicycle volumes in an SPF project.

Dolatsara (2014) concluded that a higher exposure of bike volumes, the presence of bike lanes, the presence of bus stops within 0.1 mile from an intersection, and an increased number of left turn lanes are associated with more bicycle crashes. However, that doesn't mean the bikes cause more crashes because there are more bicycles inside bike lanes than outside bike lanes (Dolatsara, 2014). This finding is consistent with the Nordback et al. (2014) paper. Bicycle facility-related studies have also been conducted by others. Reynolds et al. (2009) concluded that there is evidence to support that fact that the purpose-built, bicycle-specific facilities can reduce bicycle collisions, and street lighting, paved surfaces, and low-angled grades are also factors that improve bicycle safety.

The Federal Highway Administration (2013) established the steps for developing SPFs for jurisdictions, and these are as follows:

Step 1: Determine the use of the SPF. Is this SPF for network screening, project level prediction, deriving CMF, or before-after evaluation using the EB method?

Step 2: Identify the facility type. The developers need to choose a specific type of facility: intersection, segment, or ramp?

Step 3: Compile the necessary data. According to the purpose of building the SPF, the sample size and corresponding data set will be different. The guidance on the minimum sample size can be found in the SPF Decision Guide (Srinivasan, Carter, & Bauer, 2013). For our project purposes, the effort requirements for "project level" were recommended. In other words, the samples needed to include 30 to 50 sites, with at least 100 crashes per year for the total group, and three years of data.

Step 4: Prepare and clean up the database. Statistical plotting tools can be used to check for outlier and data entry errors.

Step 5: Develop the SPF. Estimate the regression coefficients, calculate model diagnostics such as goodness-of-fit, and examine residual and Cumulative Residual Plots. The basic SPF form for a segment is:

$$\mathrm{Y} = \mathrm{L} \times e^{a} \times (AADT)$$

The basic SPF form for an intersection is:

$$\mathrm{Y} = e_{a} \times (AAADT_{minor}) \times (AAADT_{major})$$

Step 6: Develop the SPF for the basic condition.

Step 7: Develop CMFs for specific treatment. A CMF is a multiplicative factor used to compute the expected number of crashes after a given countermeasure has been

implemented at a specific site (Crash Modification Factor Clearinghouse, n.d.). This process was not included in this project.

Step 8: Document the SPFs. The content of documentation should include the following:

- Crash type(s)/severity(s) for which the SPF was estimated
- Total number of crashes (by type and severity) used in the estimation
- Purpose of the SPF (e.g., network screening, project level analysis, CMF development, etc.)
- State(s)/county(s)/city(s) that were used
- Facility type (e.g., rural 2 lane, 3 leg stop-controlled intersection, freeway to freeway exit ramp)
- Number of years used in the estimation of SPF
- Number of units (segments, intersections, ramps)
- Minimum, maximum, and average length of segments
- Minimum, maximum, and average AADT
- Minimum, maximum, and average values for key explanatory variables
- Coefficient estimates of the SPF
- Standard errors of the coefficient estimates
- Goodness of fit statistics
- Discussion of potential biases or pitfalls.

Identifying variables in function form is the most critical step in the process of developing the SPF. Forward and backward stepwise regressions can be used to determine significant variables. T-statistic, Chi-square statistic, Akaike's information criterion (AIC), and Bayesian Information Criterion (BIC) can be used to compare models. Cumulative residual Plots was further recommended to obtain insight into whether the functional form has been appropriately selected (Federal Highway Administration, 2013; Hauer, 2004). The potential issues may include over-dispersion which can be overcome with the negative binomial model. Temporal and spatial correlation can be solved by averaging the values of the site characteristics, e.g., averaging lane width for three years for the same location (Federal Highway Administration, 2013).

Srinivasan et al. (2013) determined the data sets needed for developing SPFs for different purposes for traffic SPF, shown in Table 2-2:

Table 2-2 Requirements for data sets for building SPFs

Intended Use	Process	Sample needed	Staff hours needed - data collection and preparation (per SPF)	Staff hrs needed - statistical analyst (per SPF)
Project level	Calibrate SPF	30-50 sites; at least 100 crashes per year for total group[a]. At least 3 years of data are recommended.	150 to 350	n/a[d]
Project level	Develop SPF	100-200 intersections or 100-200 miles; at least 300 crashes per year for total group[c]. At least 3 years of data are recommended.	450 to 1050	16 to 40
Network screening	Calibrate SPF	Must use entire network to be screened. No minimum sample specified. At least 3 years of data are recommended.	24 to 40[b]	n/a[d]
Network screening	Develop SPF	Must use entire network to be screened. Minimum sample would be 100-200 intersections or 100-200 miles; at least 300 crashes per year for total group[c]. At least 3 years of data are recommended.	24 to 40[b]	8 to 24

2.4 STRAVA Data Literature

Traditional bicycle data counting typically refers to manual counts of bicycles during peak hour periods (Jestico, Nelson, & Winters, 2016), and it is used to calculate daily, monthly, or yearly volumes by multiplying daily or seasonal factors. However, traditional count methods lack spatial details and temporal coverage (Jestico et al., 2016; Ryus et al., 2014). Global Positioning Systems (GPS) embedded in mobile devices allow people to track and map their locations, and researchers can use those data can to analyze bicycle behavior and route choice (Broach, Dill, & Gliebe, 2012; Casello & Usyukov, 2014; Hood, Sall, & Charlton, 2011; Jestico et al., 2016; Le Dantec, Asad, Misra, & Watkins, 2015). Crowdsourcing fitness apps in mobile devices provide a new source of data for transportation agencies and increase the temporal and spatial resolution of official counts (Jestico et al., 2016).

STRAVA® crowdsourcing data based on GPS have been used in different bicycle projects and studies all over the world: Queensland, Australia, used them to quantify how a new bicycle pathway changed bicyclists' behaviors; Glasgow, Scotland, analyzed a corridor of bicycle activities to provide evidence for new bicycle infrastructure on a street; Austin, Texas, combined STRAVA® data with bike share data to explore the impacts of its program on streets and on a bike network; the Oregon DOT used them to decide where to build bike counters and to adjust existing bike counter locations to capture bicycle behaviors better; Vermont Transportation used these data as its key layer for statewide planning design; the University of Victoria and University College London used them to model bicycling transportation in their areas (STRAVA, 2016a).

In 2014 STRAVA users accumulated 2,700,000,000 km and 75,700,000 riders all over the world (Scott, 2015). While it seems that STRAVA has taken a large proportion of market share, it is necessary to be careful about using these data. While Oregon DOT paid $20,000 to

purchasing these data, ODOT acknowledged a problem in that STRAVA's target demographic does not represent all bicyclists. It is built for cyclists who treat bicycles as sport but not for bicycle commuters (Hunt, 2015). Hunt (2015) warned that it is important to analyze how STRAVA represents the real story before we fully believe it.

However, some existing papers have verified the representation of STRAVA data for all bicyclists. Jestico et al. (2016) compared STRAVA data with manual counts data in Victoria, British Columbia. The authors compared those two types of data by hourly, AM, and PM peak and peak period totals separated by season. They used a Generalized Linear Model (GLM) to capture the relationship between STRAVA data and traditional manual count data, and the results showed a linear association between them in which one STRAVA count can represent 51 riders from manual counts. They said that the accuracy of categorical cycling volume can be 62 percent, but they also mentioned that STRAVA fitness data are a biased sample of ridership; however, they can represent categories of ridership and map spatial variance in urban areas with high temporal and spatial resolution.

Watkins et al., (2016) compared STRAVA data with data from another transportation agency app called "Cycle Atlanta." They found that Cycle Atlanta only represented 3 percent of manual counts, and there were also differences between STRAVA and Cycle Atlanta. The representation should be carefully analyzed because of the biases related to gender of users, racing or commuter users, age, and income. However, STRAVA data provide opportunity for agencies to obtain data without creating their own app. Watkins et al. concluded that data from STRAVA should be compared to data from local sources and weighted appropriately, and they can be supplemental to other bicycle counts. Selala and Musakwa, (2016) stated in their studies that it is clear that STRAVA data are a useful tool that can provide efficient information for decision making and formulation of policies for non-motorized transportation programs. In their paper, they also mentioned that only 20 percent of the cycling trips were commuting, whereas recreational trips accounted for the other 80 percent in the city of Johannesburg. Therefore, it was obvious that there are some levels of bias in STRAVA data, but conclusive decisions should be made with more information. In relation to trip time, cycling counts from STRAVA had higher numbers in the morning, and the number decreased approaching midday, then started increasing after that, finally declining again after 16:00. They said that the numbers recorded by STRAVA were affected by the availability of gated communities, income levels, crime levels, and the provision of infrastructure (Selala & Musakwa, 2016).

CHAPTER 3 ENHANCEMENT OF GIS TOOL FOR ESTIMATING BICYCLIST EXPOSURE

This chapter documents the process of improving the usability of a GIS tool that had been created during a previous PacTrans project. Engineers and planners can use the GIS tool to estimate bicycle exposure when conducting safety analyses. The original tool had the following weaknesses: (1) used numerous disparate Python scripts that were confusing to the uninitiated, (2) required tedious intermediate data processing activities, and (3) had a very long computer runtime. For example, the bicycle network for Seattle, Washington, was the primary data set used for testing and development, and the computer runtime for that data set alone was over 22 hours. Moreover, the process of preparing intermediate data and calibrating the tool (i.e., repeatedly executing runs to check output and tweak input) would take multiple days and even weeks to complete.

Our goals were to (1) streamline the tool's operation and (2) decrease the tool's computer execution time. The original tool involved eight individual scripts, each requiring various amounts of intermediate data processing (see figure 3-1). As part of this PacTrans project the new tool was streamlined to just one script. The tool-user now only needs to do basic pre-processing of the data.

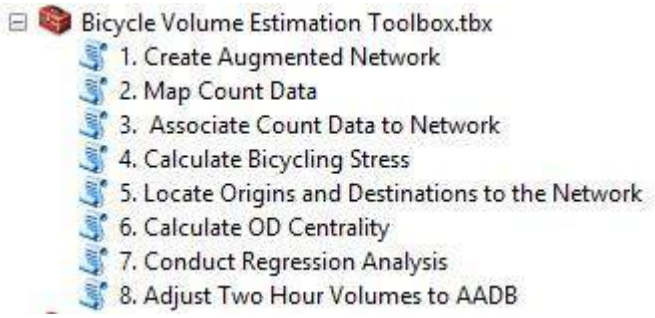

Figure 3-1 Original toolbox (the new tool is just one script)

One of the goals was to decrease the tool's computer execution time for the Seattle data set by 75 percent. Through various innovative changes and optimization of the code we exceeded our goal. Figure 3-2 shows the execution time at various benchmark points throughout the project. The new tool runs in 10 percent of the original time.

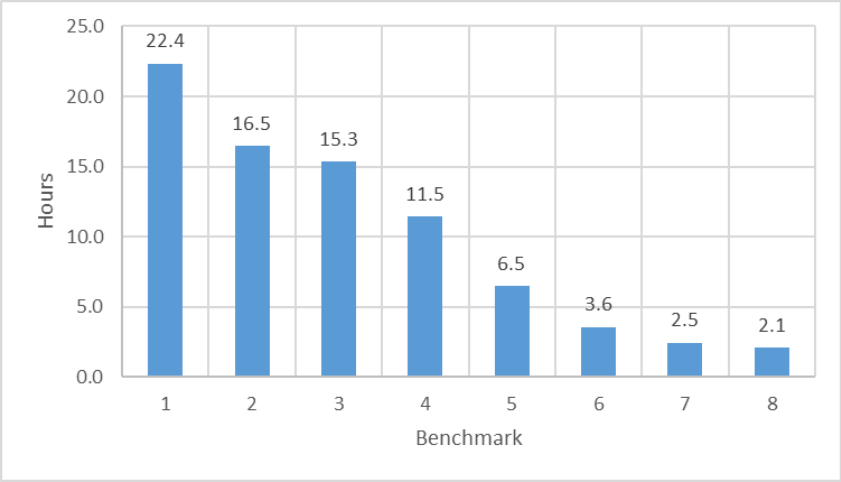

Figure 3-2 Improvements in computer execution time (hours) for Seattle

This report describes the improvements that were made to the tool with a focus on the steps that were taken to improve tool process and computer runtime. The material documented in this report provides useful guidance for practitioners who will use the tool and helpful information for future researchers who might continue to advance the underlying methods.

3.1 Introduction

In a previous PacTrans project, researchers at the University of Idaho created a geographic information system (GIS) tool that can be used to estimate bicycle volumes throughout a city (Lowry et al. 2015). This chapter describes improvements that were made to the GIS tool through this subsequent PacTrans project.

Engineers and planners can use the GIS tool to estimate bicycle exposure on streets and pathways (i.e., demand volumes) in order to conduct safety analyses or for other planning purposes. The goal of this 2014-2015 PacTrans project was to improve the tool so that it could be more easily used by practitioners. The original tool had the following weaknesses: (1) used numerous disparate Python scripts that were confusing to the uninitiated, (2) required tedious intermediate data processing activities, and (3) had a very long computer runtime. For example, to estimate bicycle volumes for the streets of Seattle, Washington, the original tool took 22 hours to run (computer runtime only). The process of calibrating the tool (i.e., repeatedly executing runs to check output and tweak input) would take multiple days and even weeks to accomplish during a regular work schedule.

This chapter describes three specific improvements that dramatically reduced the time required to use the tool. Other improvements were made during the time period of this PacTrans project, but these three improvements are by far the most significant. The process of making these improvements (i.e., identifying problems, proposing solutions, and testing ideas) occurred while examining case study data. Primarily, our case study data were for Seattle, Washington, and Moscow, Idaho, but included numerous other cities in Washington and Idaho, as well as

data from throughout the country as part of a different, but related, project for the Rails-to-Trails Conservancy.

The material documented in this report provides useful guidance for practitioners who will use the tool and helpful information for future researchers who might continue to advance the underlying methods. The next section provides background about demand estimation and the underlying method of the tool. Then, a section for each improvement is presented.

3.2 Background

There are primarily two modeling techniques to estimate bicycle volumes: multi-step travel demand models and direct demand models (Porter et al. 1999). Multi-step models, such as the "four step model," are data intensive, expensive, and complex (Liu et al. 2012). Direct demand models are more simplistic. They often involve linear regression as follows:

$$Facility\ Bicycle\ Volume = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \quad (1)$$

where the explanatory variables, $x_1, \ldots, x_m$, represent characteristics of the bicycle facility, such as adjacent vehicle volume, adjacent land use, and distance to the city center. The regression coefficients $\beta_0, \beta_1, \ldots, \beta_m$ are derived from observed data (Griswold et al. 2011; Jones et al. 2010).

The original GIS tool was made to estimate bicycle volumes using only one type of explanatory variable: origin-destination centrality (McDaniel et al. 2015). OD centrality is calculated for every link in the network (i.e., street or path segment) by finding the "shortest path" between every origin and every destination and counting how many times a link is used. The number of times a link is used is that link´s OD centrality. The tool-user supplies two sets of data representing origins/destinations: residential parcels and non-residential parcels. The tool calculates OD centrality four times for bicycle travel that would go from (1) residential to nonresidential, (2) residential to residential, (3) non-residential to residential, and (4) non-residential to non-residential. Next, a regression model is fit to observed bicycle volumes using the four OD centrality values. Finally, the regression model is applied to every link in the network.

3.3 Improvement One: Seamless Tool Process

The original "tool" was actually a toolbox of eight individual tools. The user had to execute all eight tools in succession, with various amounts of data processing in between each tool. Figure 3-3 shows the original toolbox and ight individual tools. The tools were developed this way because (1) software development was easier through modulation, (2) a few scripts (i.e., sub-processes) could be skipped when conducting repeated analyses to save time, and (3) a few scripts required the user to select certain options. Our goal was to combine all the scripts into one seamless script.

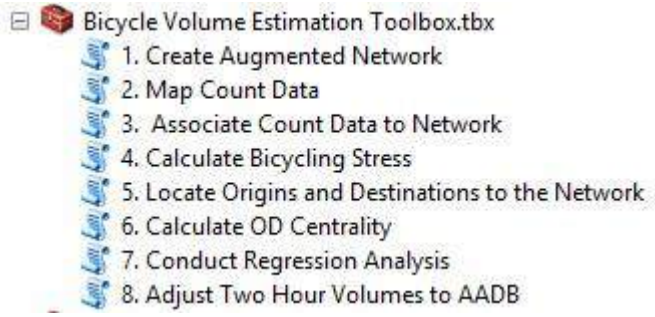

Figure 3-3 Individual tools in original toolbox

The first script, "Create Augmented Network," was extremely time-intensive, but, luckily, could be skipped when conducting repeated analysis. This sub-process creates a copy of the user-supplied street network and adds new "augmented" links that represent two directions along segments and turn-movements at intersections. For example, a four-leg intersection would be punched out and replaced with 12 new links representing turn/crossing movements and labeled with a Movement field NBT, NBR, NBL, WBT, WBR, WBL, etc. (see figure 3-4). The augmented network is necessary to model bicyclist behavior at intersections. However, once an augmented network has been made, it does not need to be made again unless the analyst makes changes to the underlying network. For Seattle, Wash., it took the original tool 3 hours to create an augmented network. Consequently, we had originally modularized the tool so that the analyst would not need to run this sub-process each time. We overcame this challenge by (1) adding an if-clause that checked for the presence of an augmented network and (2) optimizing the code (for the Seattle data set this sub-process now only takes 20 minutes).

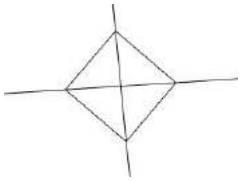

Figure 3-4 Augmented links at an intersection for turn/crossing movements

The next two scripts, "Map Count Data" and "Associate Count Data to the Network" were modularized so that the tool-user can handle different types of count techniques. Common count techniques include screenline, 4-way entering, 4-way exiting, and 12-movement (Lowry et al. 2015). However, after a few years of looking at data for numerous cities across the country, we concluded that using screenline data (or 12-movement converted to screenline) is the most straightforward. (Unfortunately, WSDOT uses a 4-way exiting technique for manual counts. If it continues to use this inferior technique, then additional tools should be made to deal with this data format.)

Similarly, we removed user options from all the remaining tools. In the original tool, the user could choose between different methods to calculate bicycle stress, levels of aggregation when locating origins and destinations, travel distance thresholds when calculating OD

centrality, specific regression techniques, and factors for adjusting short duration counts to Average Annual Daily Bicyclists (AADB). Our decision to remove these options and "hard code" the default choices were based on our experience with data for numerous cities. By doing this, the toolbox lost flexibility but gained in terms of simplicity and execution time.

3.4 Improvement Two: Calculating Bicycling Stress

As mentioned in the Background section, bicycle volumes are estimated on the basis of OD centrality values. OD centrality is determined for every link by finding the "shortest path" between every origin and destination. The "shortest path" between ODs is the path between ODs with the smallest cumulative bicycling stress. The original tool was made to calculate bicycle stress for a street segment in a complicated manner that required the tool-user to pre-code links and supply specific facility data. We significantly changed the calculation of bicycle stress as part of this PacTrans project and in conjunction with a similar project for the Rails-to-Trails Conservancy (Lowry et al. 2016). The method for calculating bicycle stress is based on two roadway characteristics: number of lanes and speed limit. This subsection describes the new calculation method.

For a street and trail network composed of a set of links and a set of nodes, denoted E and V (in graph terminology, links and nodes are called edges and vertices, respectively) we define roadway stress along every link e as:

$$F_{roadway,e} = a \times \left(\frac{S_e}{S_e^*}\right)^b \times \left(\frac{N_e}{N^*}\right)^c$$

where

a = roadway stress intercept parameter,
b = roadway stress speed exponent parameter,
c = roadway stress number of lanes exponent parameter,
$Se*$ = comfortable speed parameter for a street segment,
$Ne*$ = comfortable number of lanes parameter for a street segment,
S_e = vehicle speed associated with link e (e.g. speed limit or prevailing speed), and
N_e = number of lanes on link e.

S_{e^*} and N_{e^*} are basic parameters set by the user on the Stress Excel sheet of Bicycle_Parameters.xlsx.

Stress Parameters	Value
Comfortable Speed (Street)	20
Comfortable Number of Lanes (Street)	2

Figure 3-5 Basic stress parameters for a street segment

The parameters a, b, and c are advanced system parameters with default values of 0.1, 3.0, and 2.0, respectively. The default values are recommended but can be changed on the Advanced Excel sheet of Bicycle_Paramters.xlsx.

Intercept (advanced)	0.10
Speed Exponent (advanced)	3.00
Lane Exponent (advanced)	2.00

Figure 3-6 Advanced bicycle stress parameters

The roadway stress equation produces a number that represents a bicyclist's marginal rate of substitution (MRS) for lanes and speed. In economics, MRS is the rate at which a consumer is willing to give up one good in exchange for another good. Hood et al. (2001) and Broach et al. (2012) placed GPS trackers on cyclists and used logistic regression to empirically identify MRS values for various roadway conditions. For example, Hood et al. (2001) found that bicyclists are willing to travel 51 percent farther in a bike lane than on a similar road without a bike lane. Likewise, Broach et al. (2012) found cyclists are willing to go 140 percent out of the way to avoid a street where Annual Average Daily Traffic (AADT) exceeds 20,000 vehicles per day.

The equation above is a novel way to specify MRS values. It is advantageous because it provides a functional form and includes parameters for user preferences. For example, if S_{e^*} and N_{e^*} are 20 mph and two lanes, respectively, then the following are the MRS values for streets with these speed limits and number of lanes. The baseline is an off-street multi use trail. Thus, for example, a 25 mph and two-lane street has 20 percent MRS, meaning that riding on that street is equivalent to traveling 20 percent farther than if it were a multi-use trail.

	2	3	4	5	>5
<25	10%	23%	40%	63%	90%
25	20%	44%	78%	122%	176%
30	34%	76%	135%	211%	304%
35	54%	121%	214%	335%	482%
>35	80%	180%	320%	500%	720%

Figure 3-7 Roadway bicycle stress

The colors indicate levels of acceptable stress and are based on parameters on the Advanced Excel sheet of Bicycle_Parameters.xlsx. The parameters below show that 10 mph over the comfortable speed is considered unacceptable and that any increase in the number of lanes is unacceptable. Everything in between is tolerable. Thus the colors above are as follows:

white = acceptable low stress, yellow = tolerable moderate stress, and red = unacceptable high stress.

Unacceptable Speed (Street) +10	10
Unacceptable Number of Lanes (Street) +0	0

Figure 3-8 Unacceptable bicycle stress parameters

Roadway stress can be reduced if there is a stress-reducing bicycle accommodation present. Stress-reducing bicycle accommodations include bike lanes, buffered bike lanes, protected bike lanes, etc. The calculation is:

$$F_{stress,} = F_{roadway,} * (1 - F_{bikeaccom,e})$$

where

$F_{stress,}$ = *stress factor for link e,*

$F_{roadway,}$ = *roadway stress factor for link e, and*

$F_{bikeaccom,}$ = *bicycle accommodation stress reduction factor for link e.*

The stress reduction for various bicycle accommodations can be changed on the Advanced Excel sheet of Bicycle_Parameters.xlsx.

Bicycle Facility on Street Segment (advanced)	Reduction %
Local Street	25%
Signed Bicycle Route Only	5%
Wide Curb Lane	5%
Sharrows	5%
Paved Shoulder	5%
Neighborhood Greenway	7%
Bike Lane	40%
Buffered Bike Lane	60%
Protected Bike Lane	90%
Multi use Trail	100%

Figure 3-9 Stress reduction from bicycle accommodations

Continuing the example from above, then a protected bike lane that reduces stress by 90 percent produces the bicycle stress shown in figure 3-10.

90%	2	3	4	5	>5
<25	1%	2%	4%	6%	9%
25	2%	4%	8%	12%	18%
30	3%	8%	14%	21%	30%
35	5%	12%	21%	33%	48%
>35	8%	18%	32%	50%	72%

Figure 3-10 Bicycle stress with a protected bike lane

The same process is repeated to calculate stress for every intersection crossing movement k at intersection $v \in V$ as follows: $S_{v\,b}\, N_{v\,c}$

$$F_{cross,v} = a \times \left(\frac{\quad}{S_v^*}\right) \times \left(\frac{\quad}{N_v^*}\right)$$

where

a = *roadway stress intercept parameter,*

b = *roadway stress speed exponent parameter,*

c = *roadway stress number of lanes exponent parameter,*

S_{v^*} = *comfortable speed parameter for crossing an intersection,*

N_{v^*} = *comfortable number of lanes parameter for crossing an intersection,*

S_v = *vehicle speed associated with the cross street at intersection v (e.g. speed limit or prevailing speed), and*

N_v = *number of lanes on cross street at intersection v.*

S_{v^*} and N_{v^*} are basic parameters set by the user on the Stress Excel sheet of Bicycle_Parameters.xlsx. The parameters a, b, and c are same advanced system parameters used for the street segment.

Comfortable Speed (Intersection)	25
Comfortable Number of Lanes (Intersection)	3

Figure 3-11 Basic stress parameters for an intersection

And likewise, the stress can be reduced because of bicycle accommodations at the intersection.

$$F_{stress,} = F_{cross,} * (1 - F_{crossaccom,v})$$

where

$F_{stress,}$ = *stress factor for crossing intersection v,*

$F_{cross,v}$ = *cross-street stress factor for intersection v, and*

$F_{crossaccom,}$ = *crossing stress reduction factor at intersection v.*

For example, if the comfortable crossing is 25 mph and three lanes, then a median fefuge that reduces crossing stress by 65 percent produces the intersection stress shown in figure 3-12.

65%	2	3	4	5	6
20	1%	2%	3%	5%	7%
25	2%	4%	6%	10%	14%
30	3%	6%	11%	17%	24%
35	4%	10%	17%	27%	38%
40	6%	14%	25%	40%	57%

Figure 3-12 Intersection section with a median refuge

In addition to colored tables like the one above, the Advanced Excel sheet in Bicycle_Parameters.xlsx provides two charts to help visualize acceptable stress levels. The user can change the stress reduction in the upper left corner orange cell of the table associated with the chart. Below the orange line is acceptable low stress, between orange and red is tolerable moderate stress, and above the red line is unacceptable high stress. For example, with a bike lane, less than 25 mph is only acceptable at two lanes and becomes unacceptable at five lanes. Above 35 mph does show on the chart, so is unacceptable even at two lanes. Speeds of 35 mph and 25 mph are tolerable at two lanes but unacceptable at three lanes; 25 mph is unacceptable at four lanes.

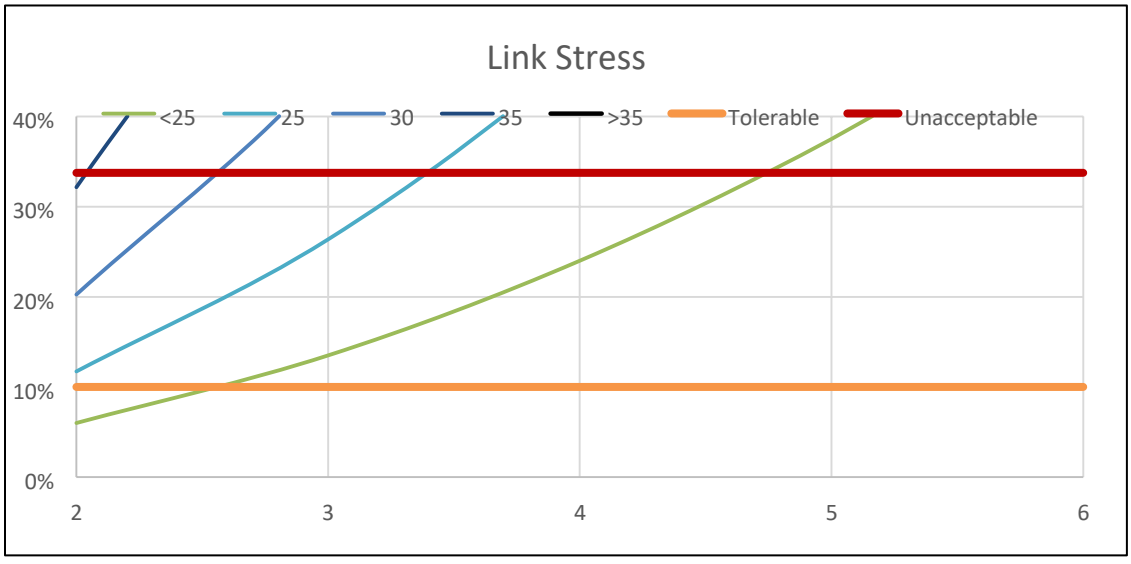

Figure 3-13 Acceptable stress levels for street segment with a Bike Lane

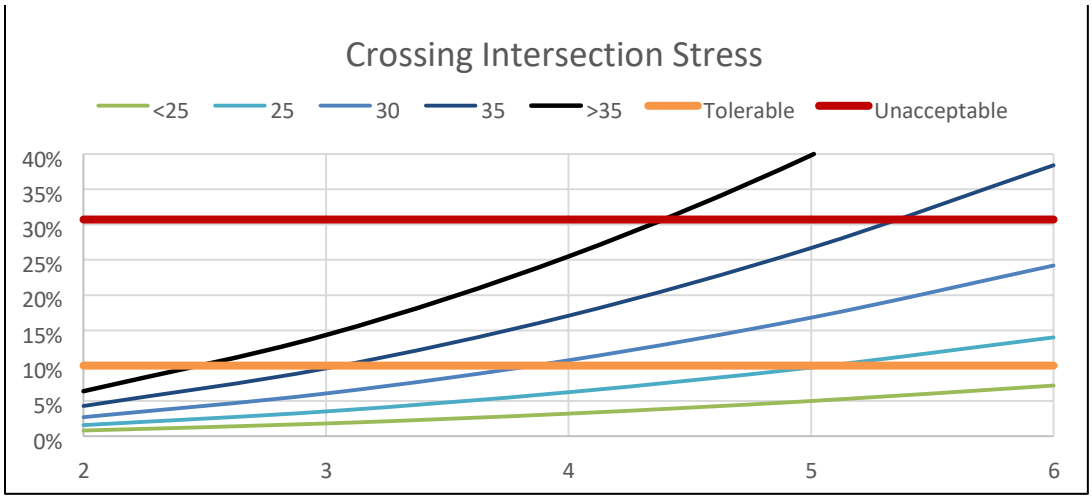

Figure 3-14 Acceptable stress levels for an intersection with a median refuge

3.5 Improvement Three: Program Optimization

We made various improvements to the computer code that dramatically decreased the computer execution time. Many of the improvements were minor and were due to our increasingly better understanding of Python and how ArcGIS integrates with Python. Nevertheless, collectively the many minor improvements had significant impact. We also accomplished a handful of major changes in the code that each individually produced significant enhancement. Over the course of the project we documented the computer execution time required to estimate bicycle volumes for Seattle, Washington. Figure 3-15 shows that with each benchmark the execution time decreased. The original tool took over 22 hours, while the current tool takes 1/10th the amount of time or about 2 hours. Likewise, figure 3-16 shows how the improvements decreased the computer time required to estimate bicycle volumes throughout Moscow, Idaho. The improvements to the scripts are summarized in table 3-1 and described below.

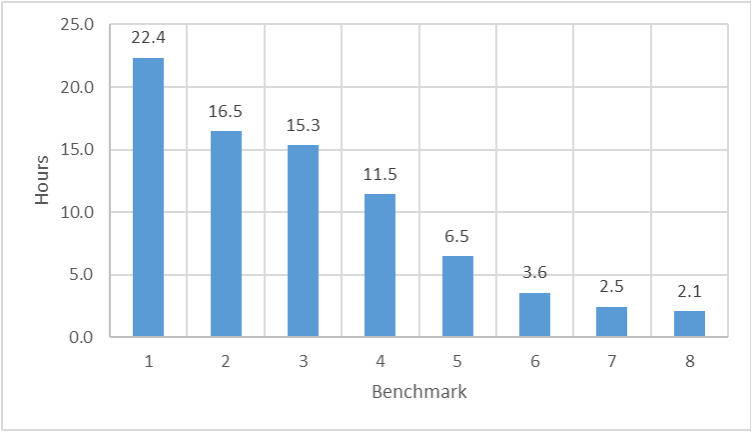

Figure 3-15 Improvements in computer execution time (hours) for Seattle

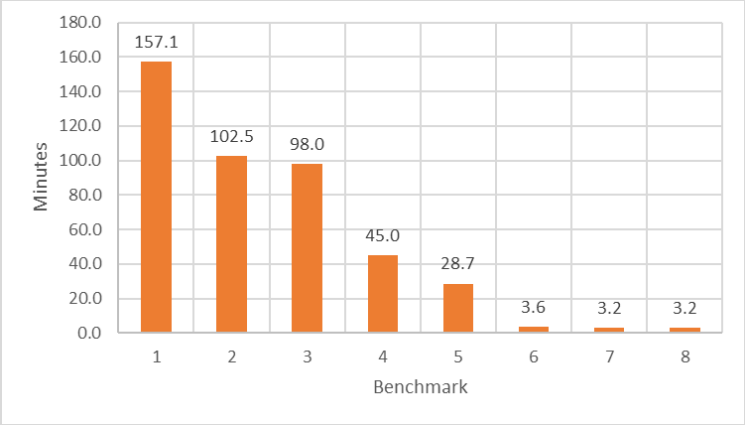

Figure 3-16 Improvements in computer execution time (minutes) for Mosco

Table 3-1 Code Optimization Improvements

Script	Improvement
Create Augmented Network	New ArcGIS function
Calculate Bicycling Stress	IDs generated with augmented network
Locate Origins and Destinations to the Network	k-d tree algorithm, changed looping
Calculate OD Centrality	Python default dictionary, changed looping

We made a significant change in “Create Augmented Network” that reduced that module´s execution time from 3 hours down to 20 minutes for Seattle. For Moscow the change made the execution time negligible (15 minutes to 1 minute). The change concerned how the algorithm draws the lines that that represent turn movements at intersections (See figure 2.2). In the original script the lines were drawn using ArcGIS´s draw editing. The new tool uses ArcGIS´s built in function called “XY to Line.” Thus instead of looping through each individual line, the script now draws all lines in one single process (it is not clear how the ArcGIS tool accomplishes this).

The original script for “Calculate Bicycling Stress” had a very time consuming loop in the code to assign to every intersection link the bicycling stress associated with crossing a street. The original tool looped through every link to obtain the coordinates of the endpoints. Then it looped through every endpoint to determine which links were connected, and then it looped through every link again to determine whether links were crossing each other. This series of looping took a very long time. In the new tool, every link is given an ID during the creation of

the Augmented network and the links they cross are determined as well. Then, when calculating bicycling stress, the IDs are used.

The script for "Locate Origins and Destinations to the Network" was very time intensive because it found which network node to be the nearest node to the centroid of an OD through an algorithm that looped over Euclidean distance between an OD and every node. For Seattle, this process would take about 20 minutes. The new technique takes less than 5 seconds. We now incorporate the k-d tree algorithm first developed by Maneewongvatana and Mount (1999). Furthermore, when using the original toolbox, the analyst needed to execute this script eight times, for origins and destinations for every combination of OD pair (residential to non-residential, residential to residential, non-residential to residential, and non-residential to non-residential).

The new tool only does this process once and retains the information.

Another significant change concerned the code to calculate OD centrality. The original tool was based on the algorithm by Brandes (2001). The authors of the Python code that we obtained from NetworkX assumed that the entire network would be searched, so before each loop of a source-to-all, they initialized an empty Python dictionary to represent the nodes of the entire network. However, since we constrain bicyclists to only travel 5 miles, we only need the nodes within that distance. Therefore, we were able to use an innovative Python technique called "default dictionaries" to populate only the nodes as they are searched. This minor change in the code (changing one line of code) saved about 2 hours of computation time for the Seattle network.

3.6 Conclusion

This chapter describes improvements to a GIS tool that engineers and planners can use to estimate bicycle exposure when conducting safety analyses. Indeed, the improvements achieved the goals of this PacTrans project to make the tool more user-friendly and decrease the computation time. The original tool consisted of eight individual scripts that required extensive intermediate data processing. The improved tool is now just one script. This change sacrificed tool flexibility by removing various options that the tool-user can no longer select; however, the tool is now significantly more user-friendly.

The original tool took more than 22 hours to run (computer time only) for the bicycle network of Seattle, Washington. Consequently, the process of calibrating the tool (i.e., repeatedly running it and adjusting the output) could have taken a few days or even weeks to accomplish. The new tool runs in a fraction of the time, just over 2 hours. Thus, calibration can be done within a single work day. The improvements documented in this report can provide guidance for practitioners who will use the tool and provide helpful information for future researchers who might continue to advance the underlying methods.

CHAPTER 4 DATA COLLECTION AND ANALYSES

4.1 Data Collection for Portland

This section documents the data collection and analysis process for building SPFs in the Portland metropolitan area. The data used in building SPFs included traffic volume data, STRAVA bicycle volumes, and six years of crash data (2009 to 2014). It should be noted that after the random sampling process and data collection process, engineers found that not enough crashes had happened in segments in Portland within the six-year period to build SPFs for segments, so engineers were only able to create an SPF for intersections. However, the procedure of building SPFs for segments is similar to that for intersections, so jurisdictions can follow the same process to establish SPFs for any site.

4.1.1 Collect Annual Average Daily Traffic (AADT)

AADT is one of two critical components of bicycle SPFs. The difficulty of collecting AADT is that not all roads have AADT available. Since crowdsourced STRAVA Metro data provide high resolution bicycle volume data for Portland, Oregon, the locations for the sample sites depended on where AADT was available. Permanent Automatic Traffic Recorder Stations (ATR) are the first choice for collecting AADT data. In this project, engineers used ODOT TransGIS to collect AADT and used those locations where the ATRs were located as the sample locations (more explanation will be presented later). Figure 4-1 shows the non-state (not state highway) locations in Portland.

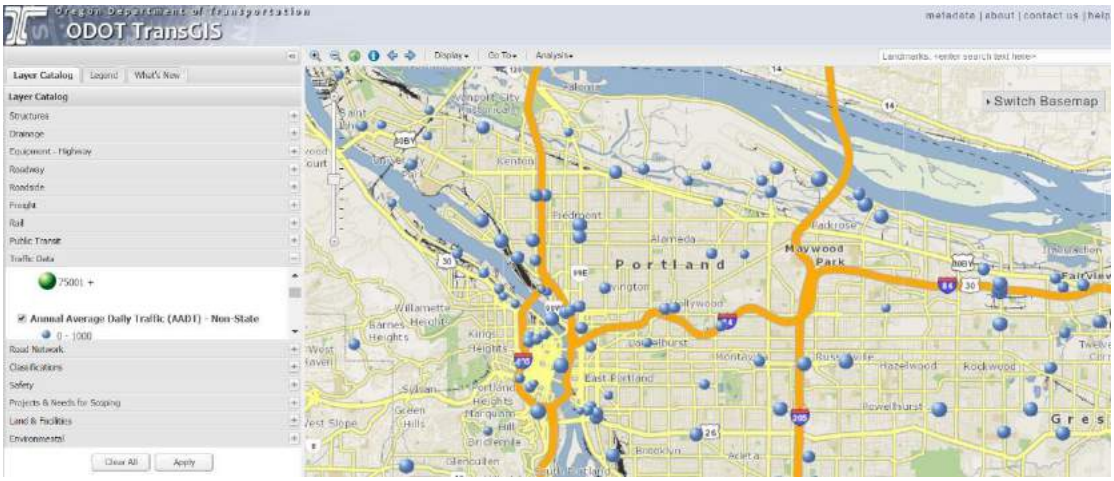

Figure 4-1 Non-state ATR locations in Portland, Oregon.

One limitation of choosing ATRs for sample locations is that the locations may not perfectly represent all road situations in Portland. The ideal method is to choose those locations randomly from all roads and intersections; however, it is not practical to choose sites totally randomly because data are not available for every site. Holding to the“best available science” principle, engineers decided to choose intersections where ATRs were available. As shown in

Figure 4-1, non-state ATR locations in Portland, Oregon, were spread out evenly, so those locations could generally represent the population. Those ATRs are spread on the basis of road density and usage rates, and there are more ATRs at locations of high population density or higher functional roads. Thus the ATR locations could basically represent the population of Portland.

However, because ATR locations are installed according to road usage and population density, another limitation was that there was no sample site for low AADT roads, e.g., local functional roads. Therefore, engineers chose some sample sites on those low usage roads to balance the bias. The way they were chosen will be discussed later.

Ideally, all ATR locations would be used to collect data; however, some ATRs were located on the state highway ramps, where bicycles are not allowed to ride. Therefore, ATRs located on state highway ramps or that were highly influenced by (very close to) state highways were excluded from the samples.

Some ATRs were close to each other, as shown in Figure 4-2. If that was the case, engineers chose only one or two of them to avoid cluster issues. Cluster issues are a data selection problem in spatial correlation in which the samples are spatially close, causing the characteristics and feature of the surrounding environment of samples to be similar. If many samples are selected in one cluster, the representation will be biased because of the over-weighting of those samples. Again, the sample selected cannot perfectly represent all populations, but the "best available science" idea was used to address the data availability issue.

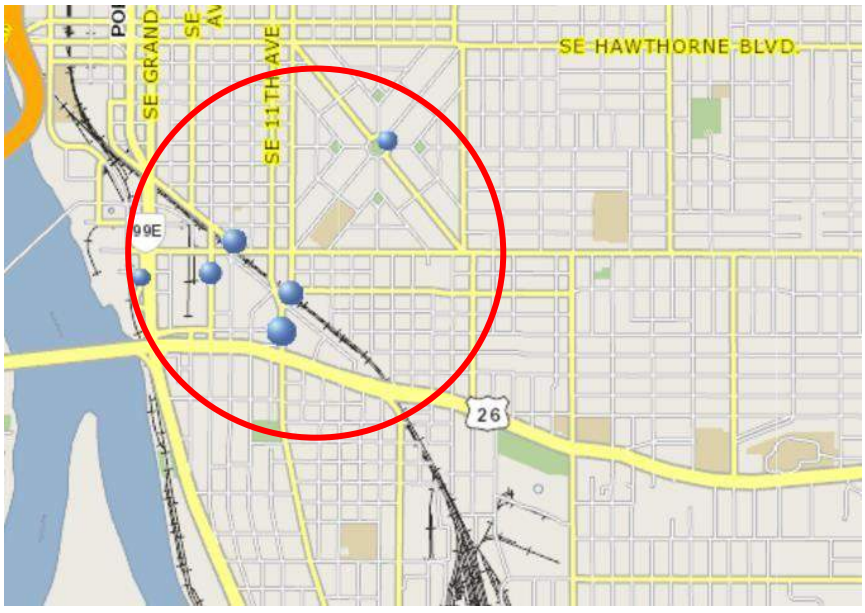

Figure 4-2 A cluster of ATRs close to each other.

4.1.2 Identify Segments

Once an ATR has been chosen, the street on which the ATR is located is chosen as a segment sample, but the start and end points of the segment should be chosen carefully. The segment needs be homogeneous (American Association of State Highway and Transportation, 2010); in other words, anything that can influence the consistency of data may be the breaking point of a segment. For instance, any change in the presence of a bike lane, the number of traffic lanes, the presence of a high population building or organization (e.g., schools), the presence of another arterial road crossing, etc., can change the AADT or AADB data significantly.

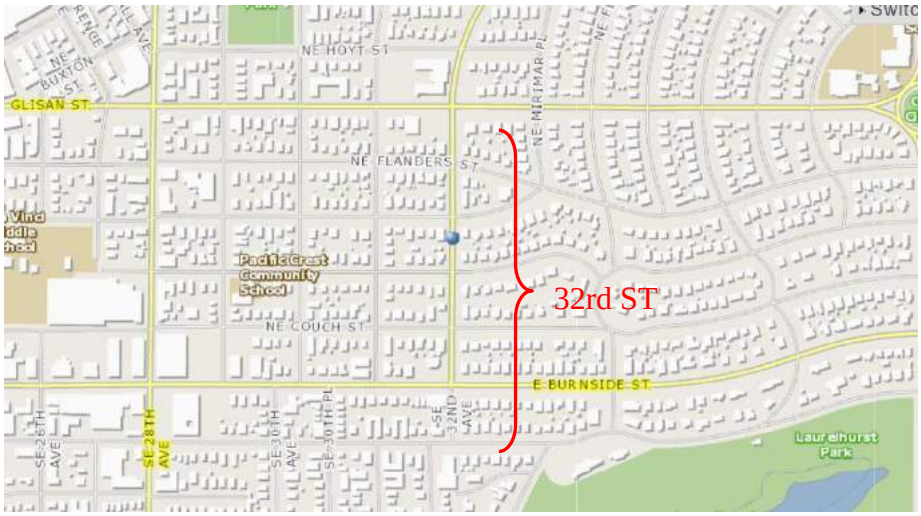

Figure 4-3 The segment chosen as a sample site based on ATR location.

Figure 4-3 shows a segment chosen on the basis of an ATR location, which was broken at Glian St and E Burnside St because those streets are city arterials that could change traffic volumes significantly. Sometimes not only the presence of an arterial but also the presence of collector roads can influence traffic volumes or bike volumes greatly, so engineering judgment is needed to determine the relationship between the sample segment and the surrounding land-use environment to decide whether to break a segment.

4.1.3 Identify Intersections

Similar to identifying segments, identifying intersections is based on the locations of ATRs. After an ATR has been chosen, the closest intersection whose minor road has ADT data available from the Portland Bureau of Transportation (Portland Bureau of Transportation, 2016) is chosen as an intersection sample. Figure 4-4 shows the same segment identified by an ATR site in figure 4-3, and the red circle highlights the ADT available for Church and Burnside roads. Either of the two roads can be the minor road of the intersection that an engineer will choose. The ADT can be converted to AADT by multiplying by a seasonal factor, which will be

addressed later. Therefore, the intersection will have AADT for both the major road and the minor road.

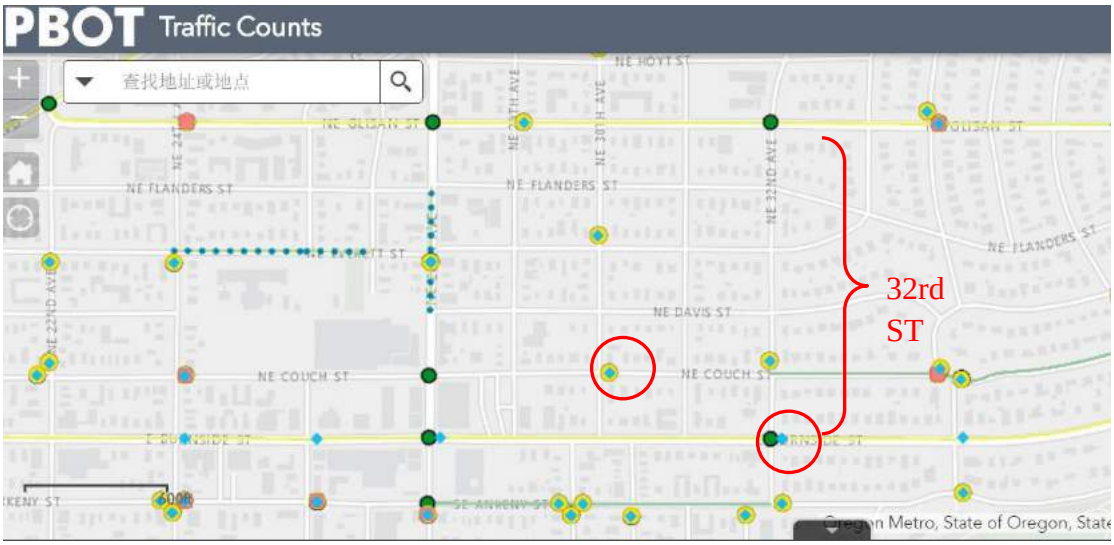

Figure 4-4 Traffic count data available for a site from PBOT, the city of Portland.

4.1.4 Convert ADT to AADT

Average Daily Traffic (ADT) is obtained from short-term traffic counts, and it is typically 72 hours of traffic collected on Tuesday to Thursday. Sometimes ADT can be obtained by 48-hour counts or at least 24-hour counts. In order to convert ADT to AADT the Seasonal Correction Factor (SF) and Axle Correction Factor are employed (Department of Transportation, 2012). Weekly Seasonal Factor should be found from the local permanent count station.

$$\text{AADT} = \text{ADT} \times \text{Weekly Seasonal Factor} \times \text{Axle Correction Factor}$$

4.1.5 Obtain STRAVA Bicycle Counts

The Oregon Department of Transportation (ODOT), collaborating with STRAVA, created an Oregon bicycle network map that aggregates all bicycle records on roads from the app to a GIS shapefile. ODOT purchased these data for research. This product provides information about bicycle counts for days, weeks or weekends, trip times, and some basic demographic information. Figure 4-5 shows the STRAVA count map.

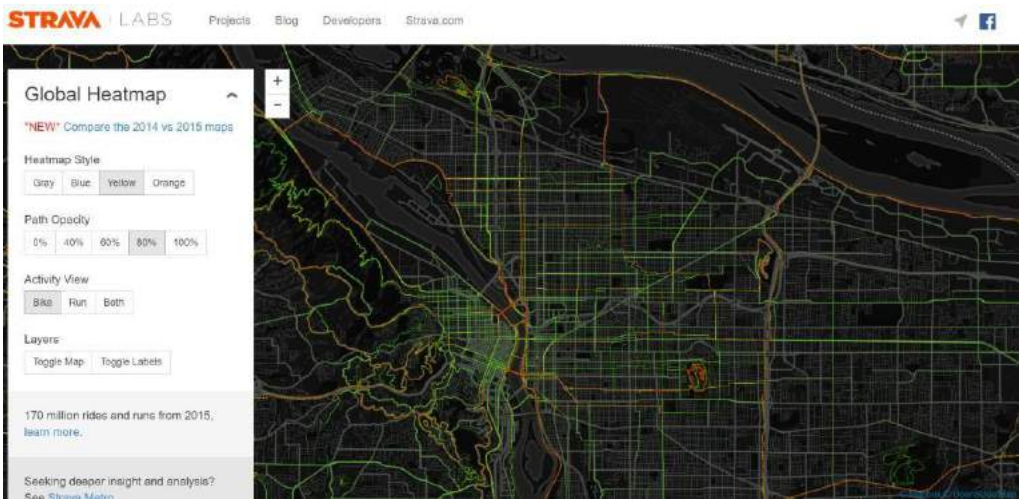

Figure 4-5 STRAVA count map (STRAVA, 2016b).

STRAVA data are obtained from ArcGIS® 10.2.2, which is widely used in research. Similar data collection was done in ODOT bicycle and pedestrian project SPR779, and since one of the authors of this project also participated in SPR779, the process is directly documented below:

"The bike volume can be roughly represented via the STRAVA bike count, but the accuracy of representation is one limitation of the data, even though STRAVA Company has differentiated the commuter count and cyclist count. ODOT has been doing tests on STRAVA and the results show the STRAVA count can represent 1 percent of total bike volume without considering the difference between commuter and cyclist. However, the tests are based on a few locations, so the finding is not conclusive. Future work can focus on validating the representation of STRAVA, but we assume it can represent basic information of real bike volume in this project.

One issue of STRAVA data in GIS is that there are more than one lines representing the same link at some segments. For example, figure X shows that there are three count links (in red) on a bridge in Portland Downtown area, and each of them has bike count 3473 bike trip/year, 5264 bike trip/year, and 2983 bike trip/year from top to bottom, respectively. This issue may come from the bike count assignment process, since STRAVA built buffers around GPS signal to assign bike count to segments. Thus we manually checked all of our sample and only used the link with highest bike count to address the problem." (Monsere, Wang, Wang, & Chen, 2016)

Figure 4-6 Multiple bike links on the same segments in the Portland downtown area (Monsere et al., 2016).

4.1.6 Obtain Crash Data

Crash data for all Oregon were provided by ODOT from 2009 to 2014 (ODOT, 2016). The shapefile was separated for each year, then engineers used ArcGIS® to aggregate them together. In this project, crash severity was not considered; instead the crash frequency, i.e., the number of crashes, was collected. Ideally, the crash for and intersection was assigned to the interaction by using a 250-ft buffer, but if the distance between two intersections was shorter than 250 feet, then engineers checked the crash manually to assign the data. In other words, a 250-ft radius buffer was built to define each intersection midpoint and crashes that happened in the buffer were counted as happening at the intersection. However, the segments between intersections in Portland are typically shorter than segments in other cities, and many of them are shorter than 250 feet. Therefore, it was necessary for engineers to assign crash data to intersections by hand. Crashes that did not happen at intersections were assigned to segments.

ATR sites on bridges were not counted in the sample because of the complexity of bridges. For example, some bridges have ramps for traffic only, which would cause inaccuracy in the AADT data. In addition, most bridges have dedicated bicycle paths where motorist-bicycle crashes would be unlikely to happen.

4.2 Data Collection for Seattle.

Twelve intersections were selected as the representative group for Seattle due to data availability. Bicycle crash data in 2009 to 2014, 2014 average annual daily traffic (AADT) volume data, and 2014 average annual daily bicycle (AADB) volume data for each intersection were collected. The following sections provide the data collection methods and the detailed data description for each kind of data set.

4.2.1 Intersection Bicycle Crashes

Generally, bicycle crash data were collected on the basis of reported data, such as from policy reports, hospitals, and insurance companies. Unfortunately, the quality of reported crash data is weakened by underreporting; for example, policy reports often underreport single-bicycle crashes and crashes that do not involve in insurance indemnity (Juhra, 2012). Crowdsourcing is a novel method for collecting bicycle crash data, deploying Internet applications to collect self-reported data through the collaboration of participants. Since self-reported data have much more selection bias than traditionally reported data (Roberts, 1995), scholars have suggested that crowdsourced data are not able to substitute for traditionally reported data (Jestico, 2016).

The bicycle crash data set used in this project was collected from police reports of bicycle collisions. Since the bicycle safety performance function was developed only for intersections, an intersection bicycle crash data set was selected from the original data set. Figure 4-7 shows the spatial distribution of 2014 bicycle crash data in Seattle. The detailed data set description is introduced in the next section.

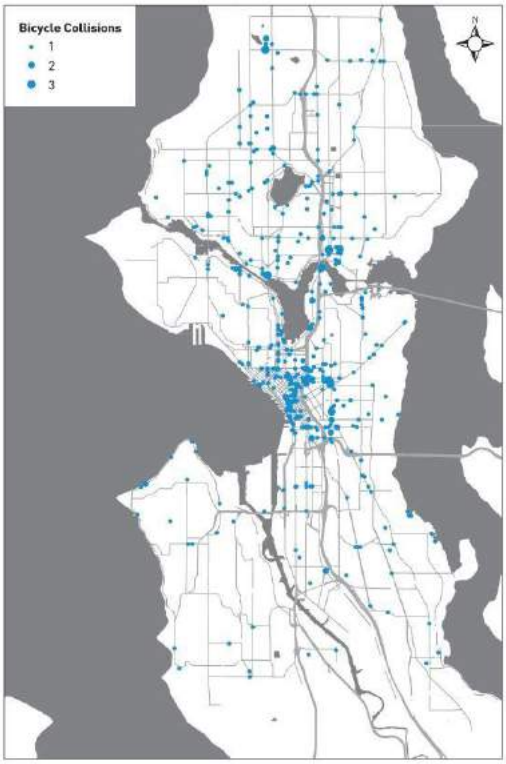

Figure 4-7 2014 bicycle crash spatial distribution in Seattle (Source: SDOT, 2015)

4.2.2 AADT

AADT is the total volume of vehicle traffic of a road for a year divided by 365 days, which is one of the exposure variables of the safety performance function. The AADT data from the Seattle area were calculated based on traffic count data, which were measured for 20 controlled count locations, 164 screen line count locations, and 111 additional count locations

(SDOT, 2015). The intersection AADT data were calculated by summing up traffic flow data form major roads and minor roads. Figure 4-8 and figure 4-9 show the traffic count locations and 2014 average annual daily traffic in Seattle.

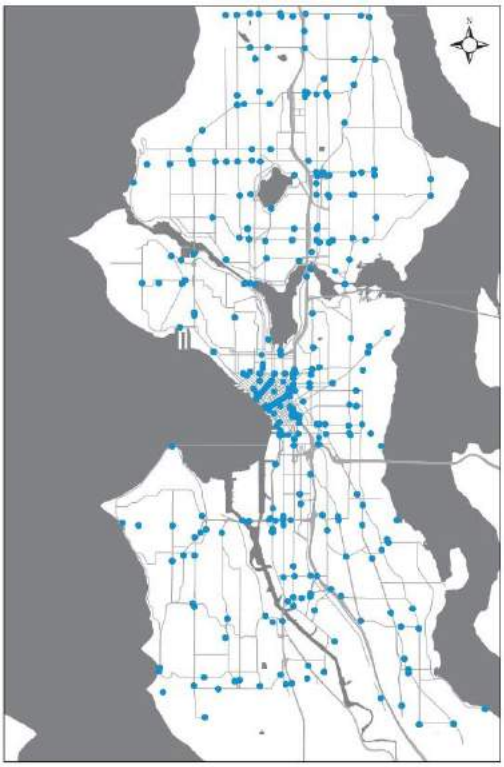

Figure 4-8 Traffic count locations (Source: SDOT, 2014)

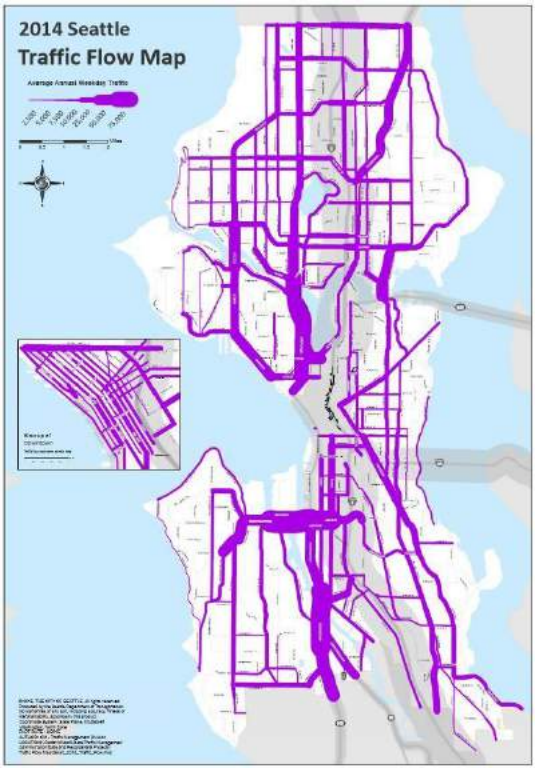

Figure 4-9 2014 average annual daily traffic in Seattle (SDOT, 2015)

4.2.3 AADB

AADB data were collected from a Seattle Department of Transportation (SDOT) open data source. SDOT deployed three bicycle counting methods to calculate citywide bicycle volumes, automated permanent bicycle counts, multiday short counts, and spot bicycle counts (SDOT, 2015). Automated permanent bicycle counting has been conducted at 12 locations in Seattle since the end of 2013. The type, variety, and spatial diversity of the locations enable them to be good representatives for characterizing the features of bicycle volumes in Seattle. Figure 4-10 shows the locations that were equipped with automated permanent bicycle counters. The hourly, daily, weather, seasonal factors, and other influencing factors of bicycle volume could be demonstrated by permanent counting results, which enabled us to create daily volume factors. Then, the AADB was calculated on the basis of multiday short counts and spot bicycle counts (SDOT, 2014). Figure 4-11 shows 2014 calculated average annual daily bicycle volumes in Seattle.

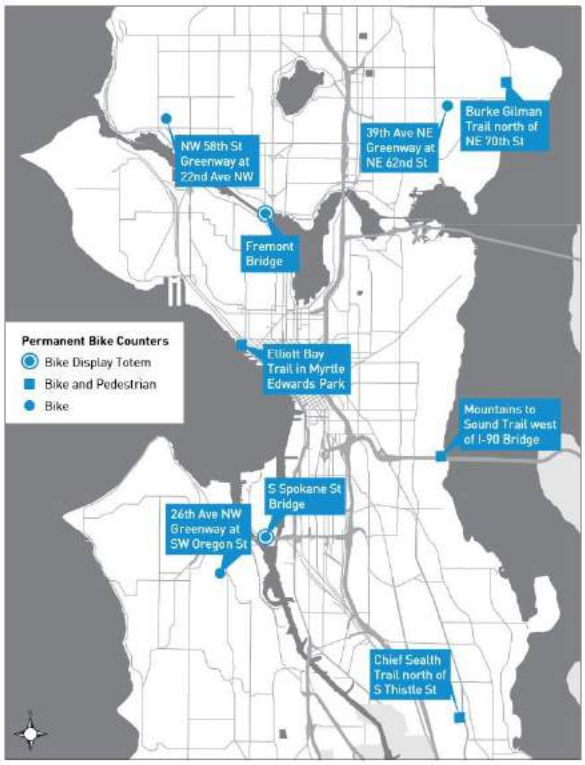

Figure 4-10 Automated permanent bicycle counting locations (SDOT, 2014)

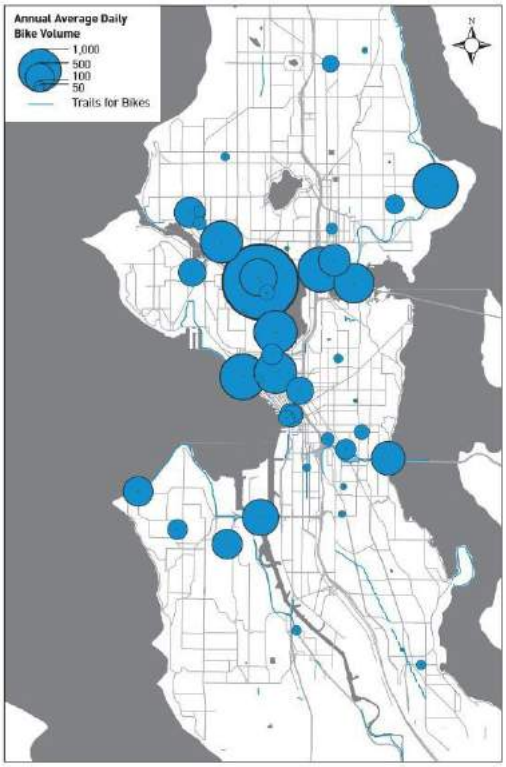

Figure 4-11 2014 calculated average daily bicycle volume in Seattle (SDOT, 2015)

4.3 Data Description and Analysis for Portland Data

In this section, details of data that were collected for intersections in Portland will be discussed. Data included intersection locations, crash data, AADT, and bicycle count data.

4.3.1 Crash Data

Crashes that happened at intersections or related to intersections were collected and assigned to intersections. Previous studies collected intersection crash data by a buffer and used the buffer to decide whether crashes belonged to an intersection; however, in this project, engineers found that this method was not accurate because of the short distances between intersections in the Portland metropolitan area. In other words, because intersections are too close to each other, engineers could not assign crash data spatially. So each crash possibly related to an intersection was reviewed by engineers to decide its location. Figure 4-12 shows the intersection sample locations with crashes that happened from 2009 to 2014.

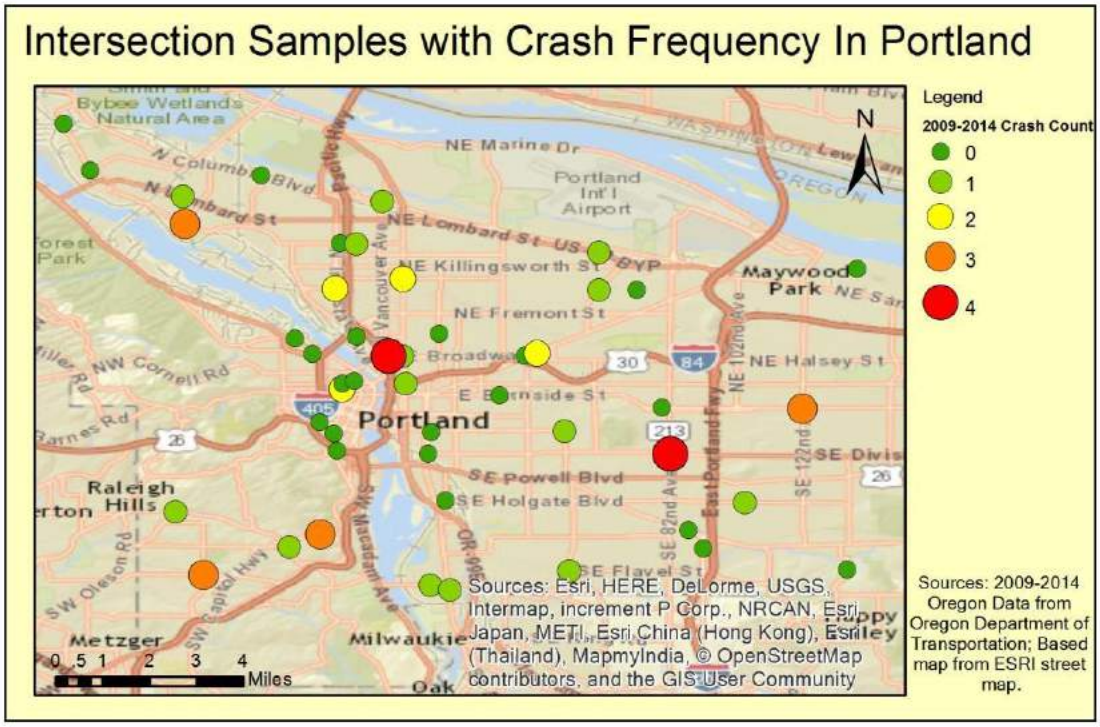

Figure 4-12 Intersection sample with crash counts from 2009 to 2014.

Even though the intersection samples in figure 4-12 were not totally random, the spread of intersection sample locations can generally represent the whole population in the Portland metropolitan area. More justification can be found in the Data Collection section.

As shown in figure 4-13, the crash counts for all intersection samples were fewer than four, indicating that fewer bicycle crashes than traffic crashes happened every year. From 2009

to 2014, the entire Oregon bicycle crash number was around 800, and there were certainly fewer in the Portland area. Another reason for the fewer crashes shown out is under-reporting. Most the bicycles are not covered by insurance, so a lot of bicycle crashes are not reported to either an insurance company or a police department if the crash is not severe. In other words, many PropertyDamage-Only or single bicyclist crashes have not been reported and so remain unknown.

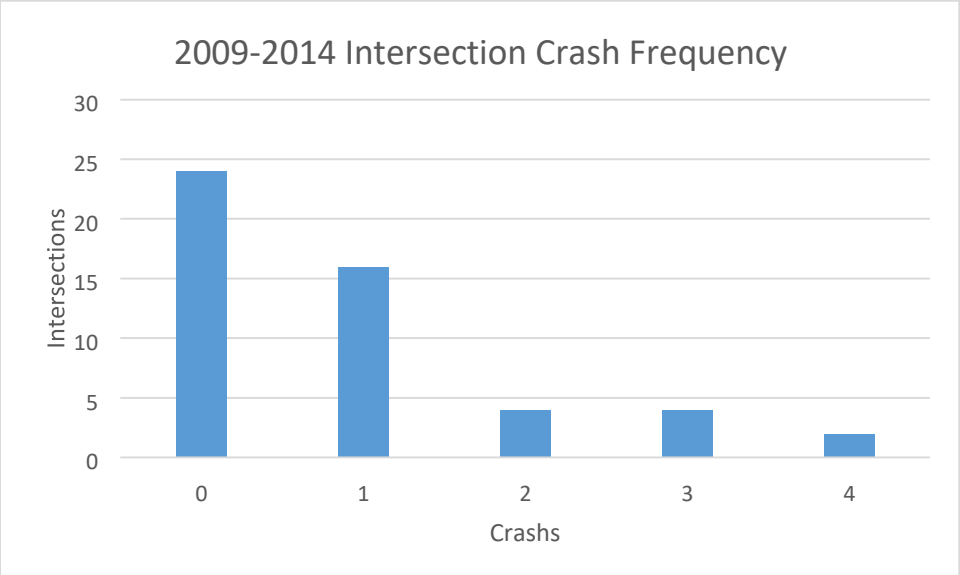

Figure 4-13 Crash frequency for each intersection sample.

In many intersections, no crash happened between 2009 and 2014, which agrees with the over-dispersion feature of bicycle crashes observed by others. Other studies' observations can be found in the Literature Review. Figure 4-13 shows the distribution of crash frequencies for all samples. The figure shows that about half of the intersections sampled had 0 crashes during the six years, and as the crash frequency increased, the intersection number decreased significantly. In other words, the majority of intersections had fewer crashes from 2009 to 2014, which verifies the over-dispersion characteristic of crash data. The crash frequency has a mean of 0.88 crashes/intersections and a variance of 1.25 crashes. The variance is larger than the mean, indicating over-dispersion.

Figure 4-14 shows the total number of crashes for each year for all intersections. The average crash rate is about 7 or 8 crashes/year for all 50 sites, with a peak of 12 crashes in 2012.

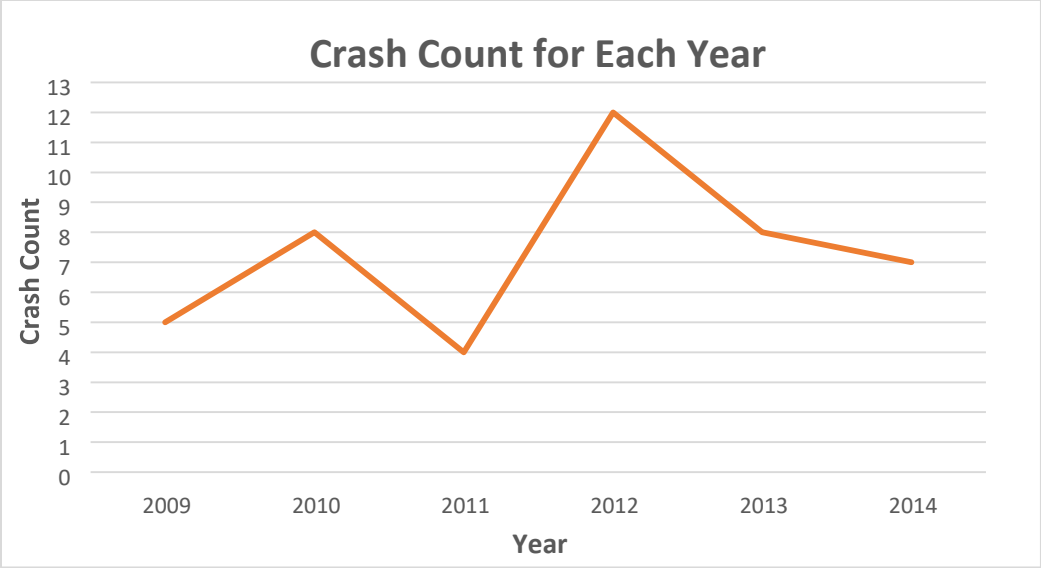

Figure 4-14 Crash count by year

Figure 4-15 summarizes the functional classifications of the roads where the crashes happened. The functional classifications were collected from crash reports in the ODOT crash data set. The functional classifications were based on which road the crashing bicycle was riding on. For example, if a bicycle was on an urban minor arterial at an intersection and it crashed with a car from the perpendicular urban collector leg, then the crash was defined as happening on the urban minor arterial at the intersection. As shown in the figure, the majority of crashes happened on arterials. This phenomenon may result from two reasons: 1) there are many more bicycles on higher functional classification roads; 2) data selection bias; in other words, because the selected data sites were near ATR locations that normally were selected to build on higher functional classification roads, most intersections had at least two legs with an arterial functional classification.

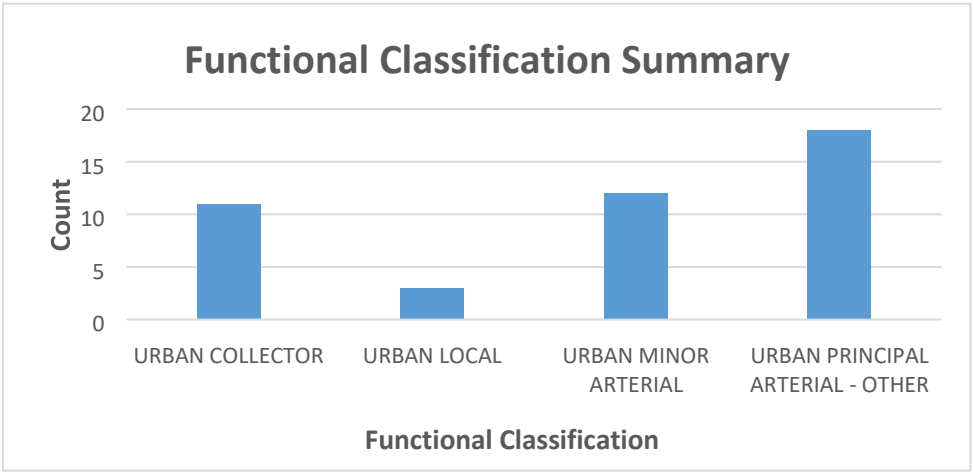

Figure 4-15 The functional classifications of intersection leg where the crashes happened

Figure 4-16 shows the collision types of all crashes that happened in those intersections. Most of crashes happened during turning movements, a situation similar to traffic collisions at intersections. Surprisingly, the least frequent collision type is rear-end collision. One reasonable hypothesis is that bicycles typically move more slowly than traffic and take less distance to stop, so fewer rear-end crashes happen.

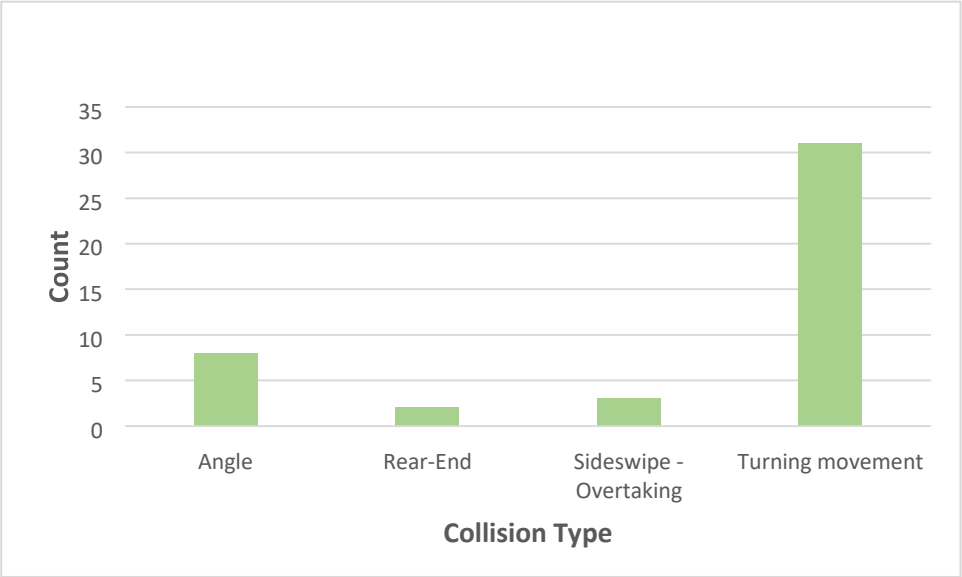

Figure 4-16 Collision types of crashes

Crash severity was not addressed by this project, but it is worth summarizing. Figure 4-17 shows the crash severity types of all crashes that happened at the intersections from 2009 to 2014. Most of the crashes were non-fatal injury, with very few property damage only (PDO). However, in Oregon, typically about half of traffic crashes are PDO. The demographic difference may be the result of less protection for bicyclists. In other words, drivers are typically protected by the vehicle frame but not bicyclists, which causes bicyclists to be more vulnerable than vehicle drivers. In addition, under-reporting may be an issue. Typically, bicycles have no insurance, so bicyclists normally will not report an accidents if they have very little loss.

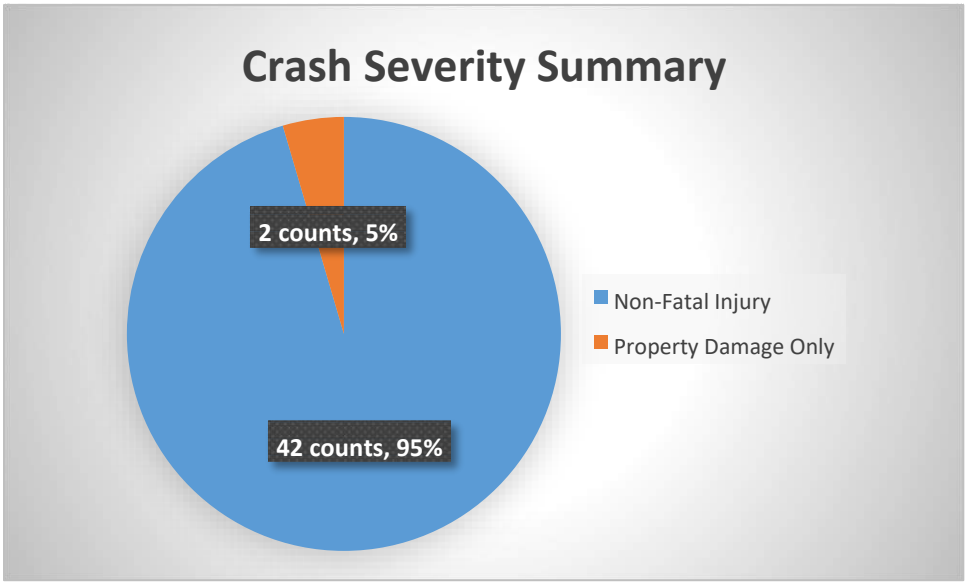

Figure 4-17 Crash severity type for six years of crashes at intersections.

Figure 4-18 and figure 4-19 summarize the weather and road surface conditions related to the crashes, respectively. Most crashes happened in better weather and surface conditions. There may be three reasons for that: 1) fewer bad weather or surface days than good weather days; in other words, a larger number of days in the year have good weather, such as clear or cloudy; 2) drivers and bicyclists pay more attention and are more careful when driving and riding in bad weather; 3) there are more traffic and bicycle volumes in good weather and on good surface conditions. Bicyclists, both commuters and recreational cyclists, typically choose clear or cloudy weather instead of riding in the rain. The weather can be a significant reason for changes in bicycle volumes, so certainly there are fewer bicycles crashes when there are fewer bicycles on the road.

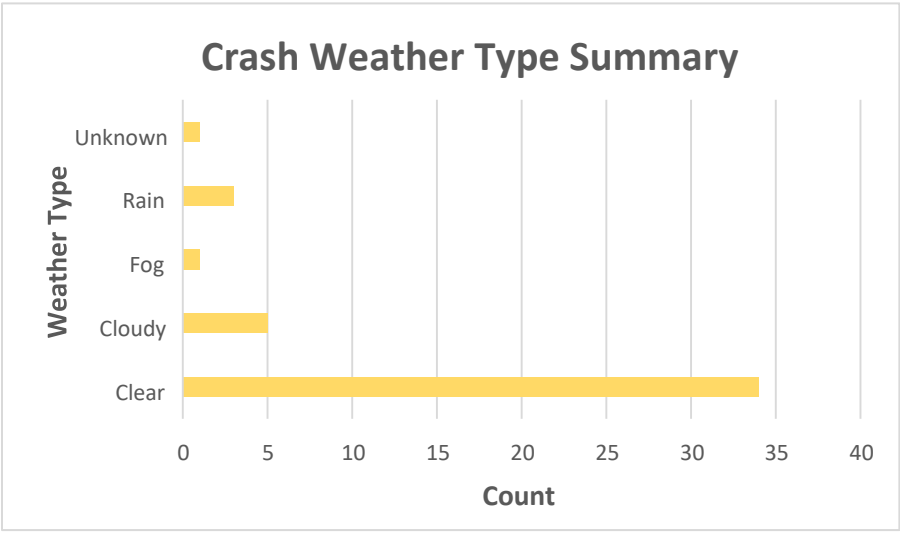

Figure 4-18 Weather conditions of the crashes at intersections.

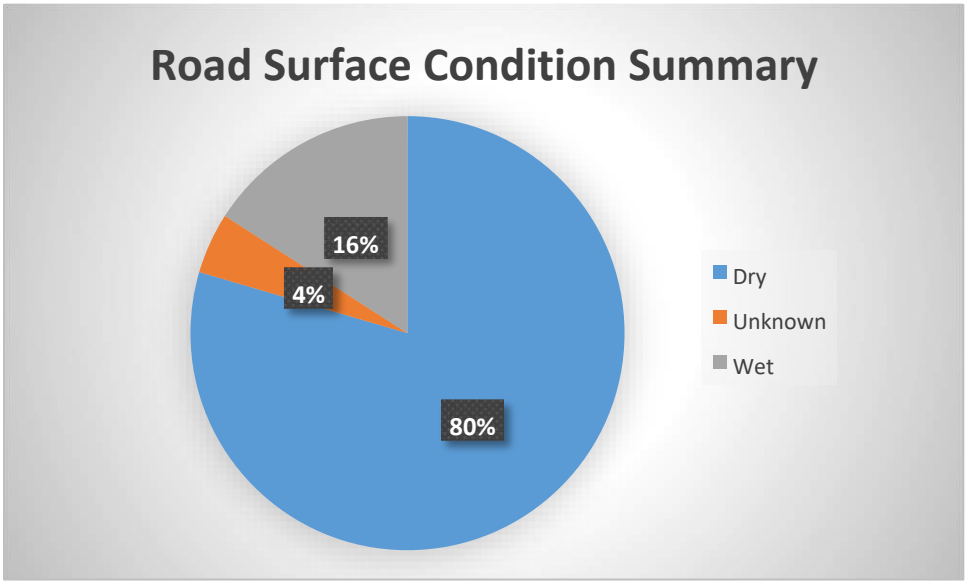

Figure 4-19 Road surface conditions of the crashes at intersections.

Figure 4-20 shows the lighting conditions for crashes that happened at intersections. It shows that most crashes happened in daylight. This phenomenon may have reasons similar to those for weather and surface conditions: 1) most bicycles and traffic move in the day instead of at night; lighting is a very important factor that can influence bicyclists choosing riding time, since bicycles are much less visible than vehicles; 2) bicyclsts and riders may be more careful at night or under bad lighting conditions.

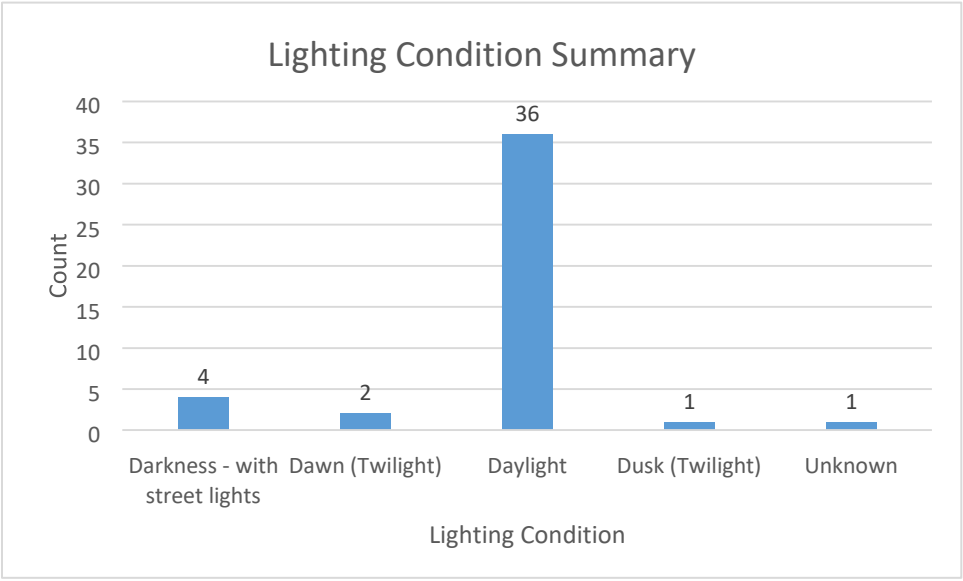

Figure 4-20 Lighting conditions of crashes at intersections.

Figure 4-21 shows the control type of the intersections where each crash happened. Most of the crashes happened at intersections with a traffic signal. This may result from two reasons: 1) drivers and bicycles move less carefully at intersections with traffic signals; for example, drivers may speed up while the traffic light changes from green to yellow; 2) data selection bias: many of the intersection sites that were selected were on arterial roads where traffic signals are typically used.

Figure 4-21 Intersection control types associated with crashes.

Table 4-1 shows the causes of crashes happening at intersections. Most crashes resulted from no yield to right of way, but the report doesn't show whether it was because of the driver or cyclist. However, it can logically be inferred from the fact that most collision types were "Turning Movement" that right-turning traffic didn't yield to the straight-moving bicycles. It is very normal the drivers may ignore straight-moving bicycle for two reasons: 1) bicycles are less visible due to less body volume and less reflective surface; 2) drivers do not check their mirrors for bicycles. It is normal for drivers to think there is no bicycle coming on the right if there is no visible bike lane. Many drivers are not used to checking their mirror for bicycles when turning right in places with fewer bicycles.

Table 4-1 The causes of crashes at intersections

Crash Cause Summary	**Count**
Careless Driving (per PAR)	1
Did not yield right-of-way	29
Disregarded R-A-G traffic signal.	6
Followed too closely	2
Made improper turn	2
Not Visible	1
Other improper driving	1
Passed stop sign or red flasher	1
Unknown	1

4.3.2 AADT Data

All intersection AADT data collected in this project are summarized In Figure 4-22. . In this data set, some of the minor roads did not have AADT from permanent count stations, so ADT was collected. Ideally, the ADT should be converted to AADT, but the seasonal table could not be found, so ADT was used in the modeling process. Furthermore, the assumption that ADT almost equals AADT was reasonable at this project's scale.

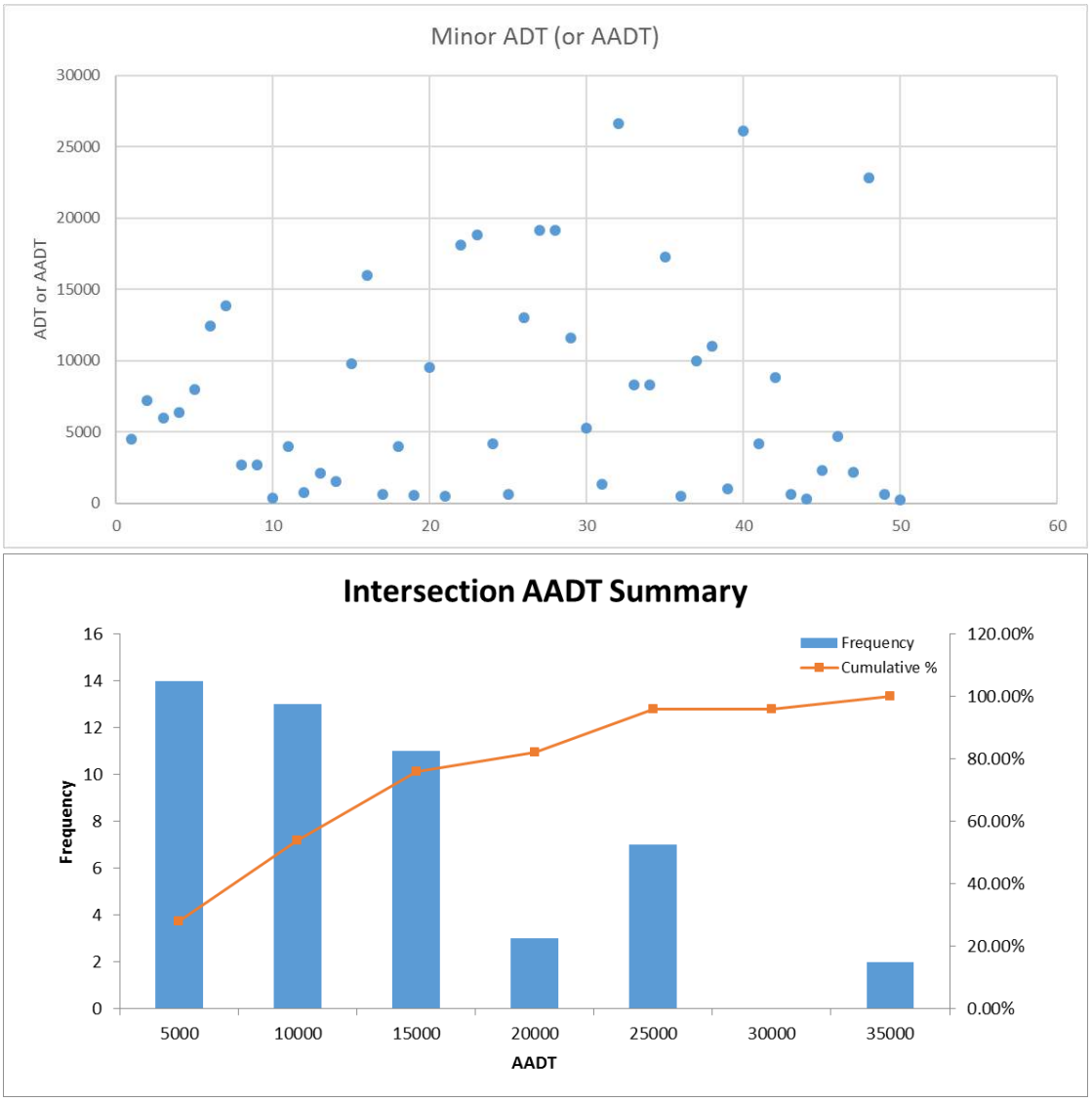

Figure 4-22 (a) Intersection major road AADT scatter and histogram graphs

The scatter graph shows that the major roads' AADTs are within the range of 0 and 30,000 traffic count, but they are not evenly spread out within this range. The histogram shows fewer samples when the AADT increases, and the majority of the major roads at interections have AADT of less than 20,000.

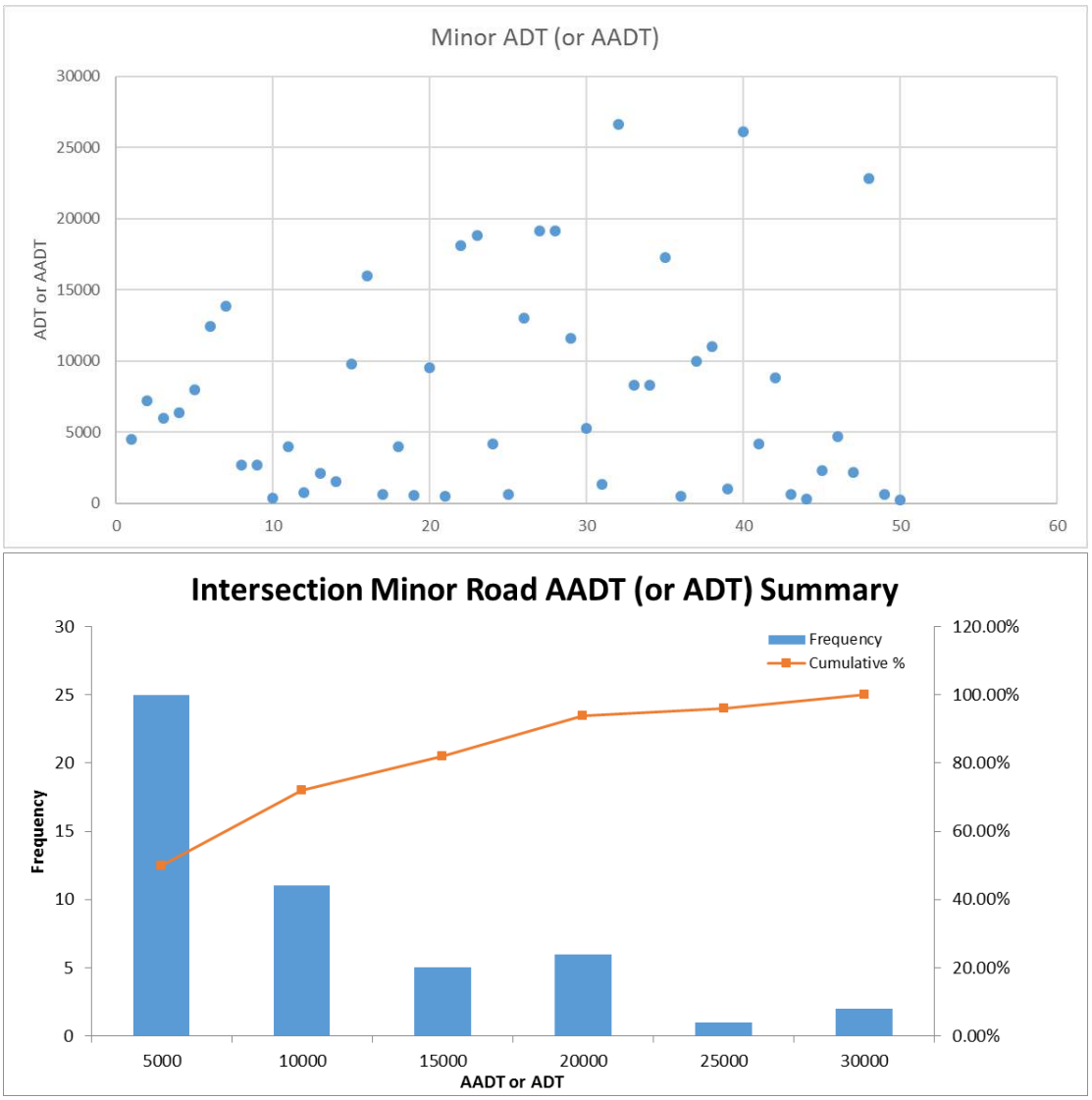

Figure 4-22 (b), Intersection minor road AADT scatter and histogram graphs.

Figure 4-22 (b) demonstrates the scatter graph and histogram graph AADTs from intersections with minor roads. In the histogram, the AADT decreases sharply from 5000 to 10,000. In other words, the range is similar to that of major roads, but the majority of them have an AADT of less than 10,000. This may come from the lower functional classification of minor roads.

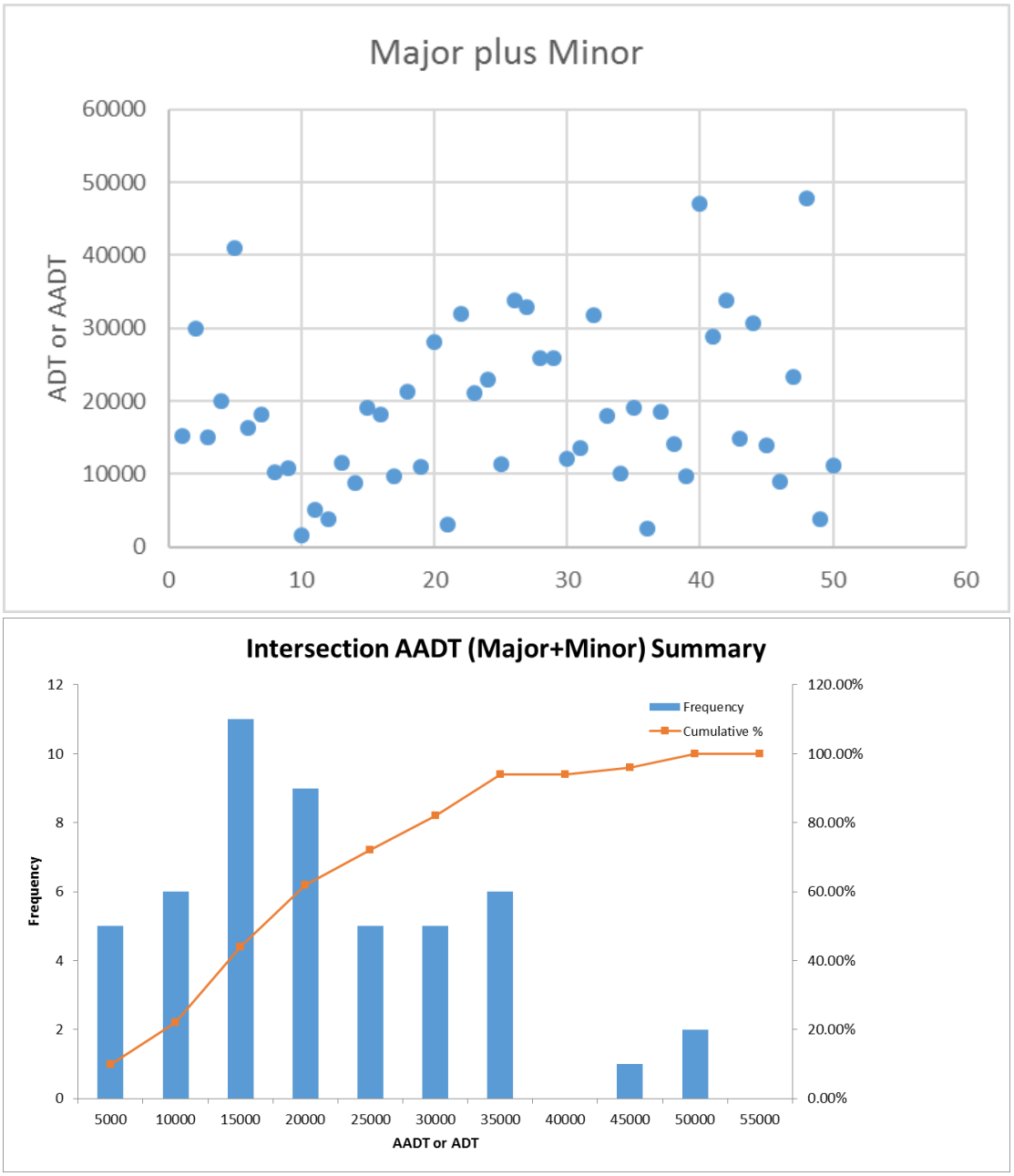

Figure 4-23 Intersection total AADT scatter and histogram graphs

In this project, engineers used both minor road and major road AADT or ADT to represent intersection traffic counts. Figure 4-23 shows the total traffic counts of adding major AADT and minor AADT to ADT. The histogram shows that AADTs peaked around 15,000 and have a right-skewed distribution. In other words, the major road AADTs have less than 25,000.

4.3.3 Bicycle Count Data

This section summarizes STRAVA bicycle count data for all intersections, both major and minor roads. The STRAVA counts were organized by year originally, so these data describe all-year counts. Average daily STRAVA counts can be obtained by dividing the yearly count by 365. However, this project used original all-year STRAVA counts as the model variable. The only difference between using yearly counts and daily counts is the number of coefficients for STRAVA data, but that did not significantly the variable. Transportation agencies can use either average daily STRAVA data or yearly STRAVA data.

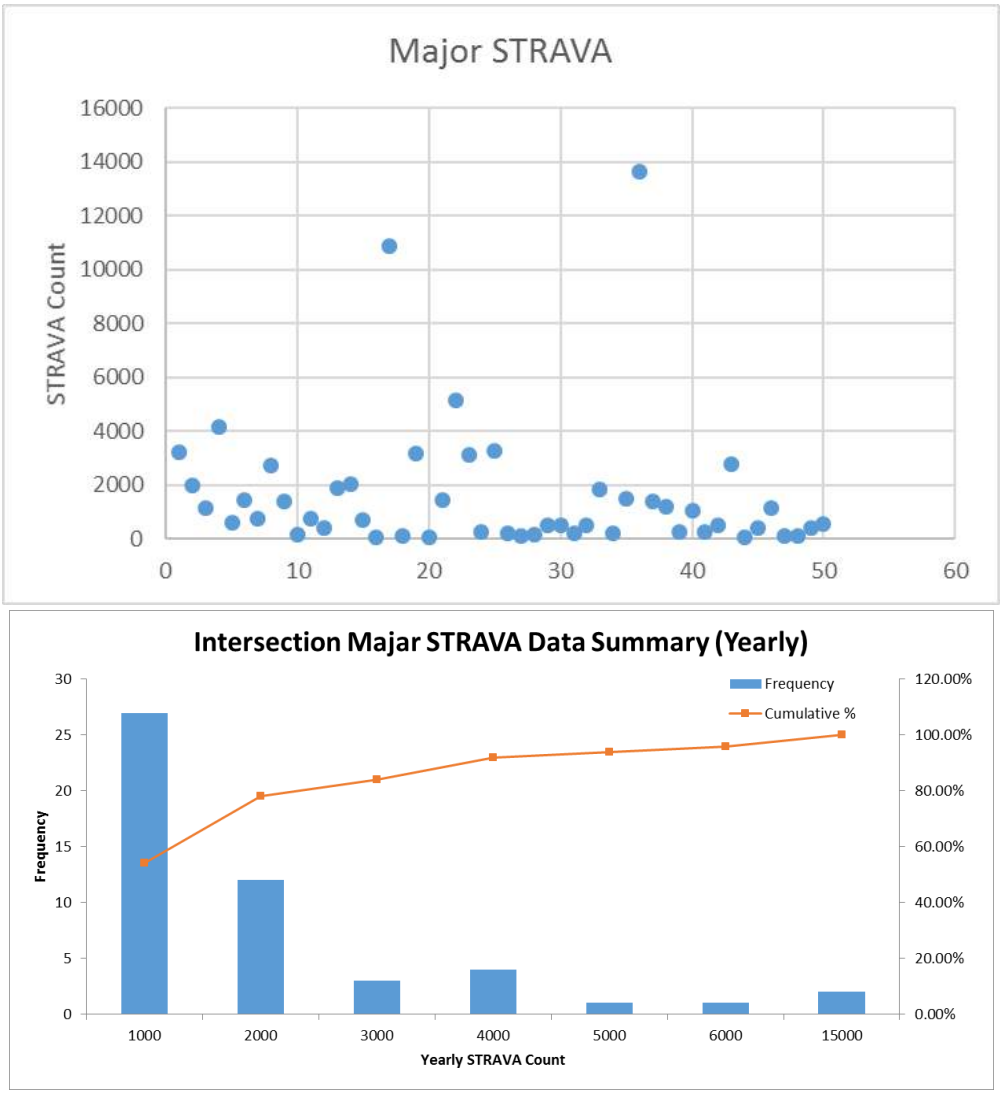

Figure 4-24 Intersection major road STRAVA scatter and histogram graphs.

The scatter graph and histogram graph in figure 4-24 both illustrate the range and distribution of major road STRAVA data. Most of those data were within the range of 0 to 4,000 bicycles, with high concentrations on less than 2,000 bicycles. In other words, the majority of

the sampled intersections had STRAVA data on a small scale. This is because 1) there are fewer bicycle users than vehicle users; 2) STRAVA data only represent a small proportion of all bicycles.

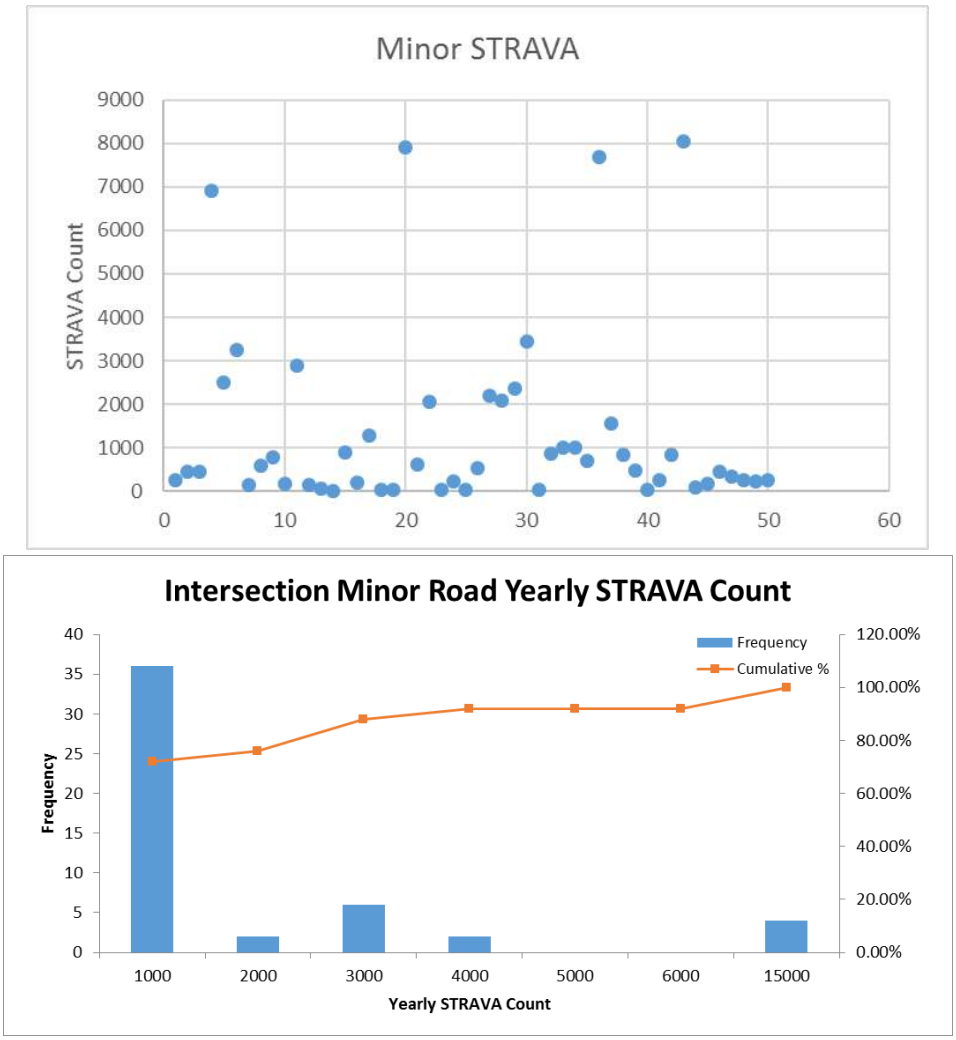

Figure 4-25 Intersection minor road STRAVA scatter and histogram graphs.

The majority of samples from smaller road STRAVA counts were similar to those from major road STRAVA counts. However, unlike STRAVA counts on major roads, almost all minor road STRAVA counts were within 1,000. This may be a result of the lower functional classification of minor roads.

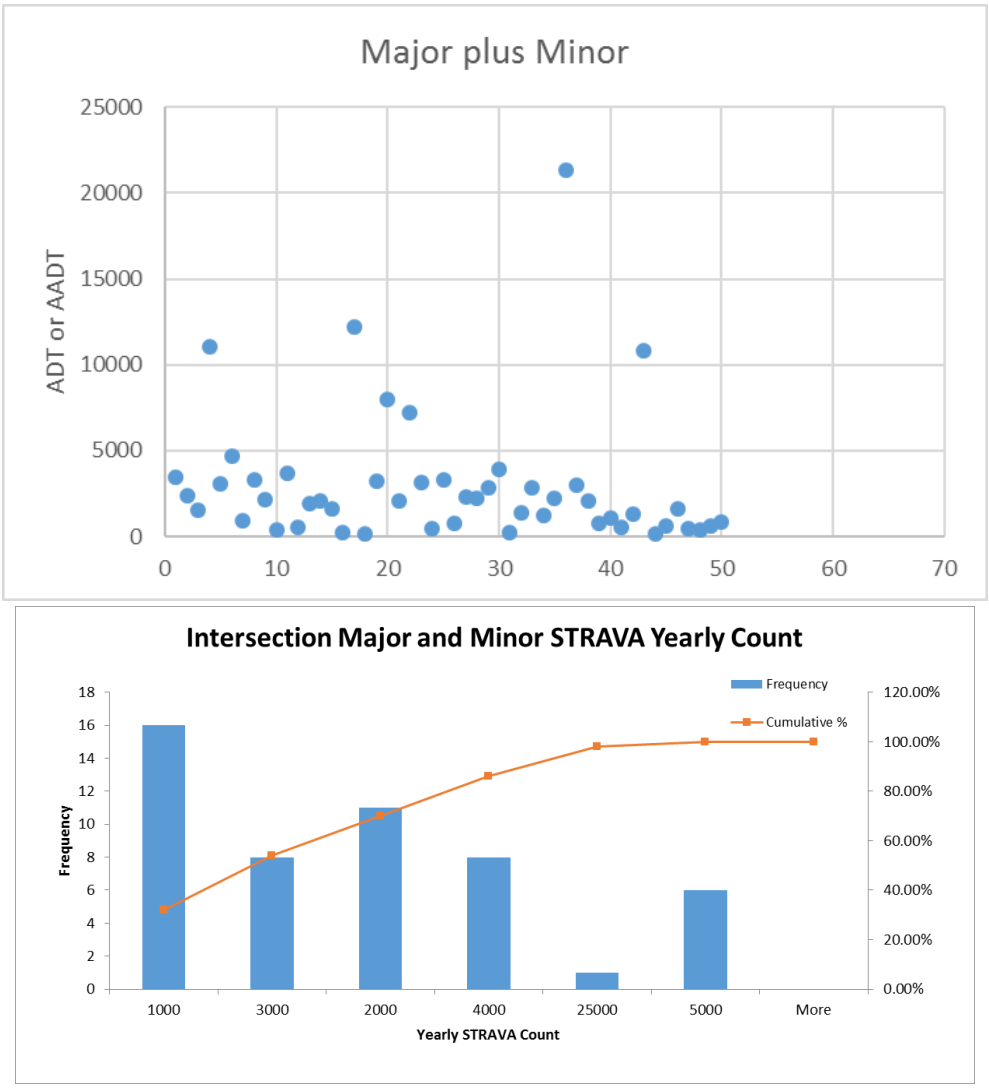

Figure 4-26 Intersection total STRAVA scatter and histogram graphs.

In this project, engineers used both minor road and major road STRAVA counts to represent intersection bicycle counts. Figure 4-26 shows the total bicycle counts, including both major and minor road STRAVA counts. The histogram shows the shape decreasing with an increase in the number on the X-axis, indicating that fewer intersections have larger STRAVA counts. Most intersections had STRAVA counts within the range of 0 to 4,000 bicycles.

4.4 Data Description and Analysis for Seattle

Table 4-2 summarizes data from the Seattle area that were used in this project. We collected six years of bicycle crash data from 2009 to 2014, and AADT and AADB data from 2014. The following sections describe the detailed features, respectively, of bicycle crash, AADT, and AADB.

Table 4-2 Data summary of Seattle data

Location	AADB (2014)	AADT (2014)	Bicycle Crash (2009 ~ 2014)
Montlake Bridge	900	57400	0
Gilman Ave W NB n/o W Bertona	470	16200	0
Mercer St and Aurora Ave N	290	118700	0
3rd Ave s/o Madison NB	210	22600	0
Pike St w/o Terry Ave	460	14600	1
2nd Ave PBL s/o Madison St	370	30000	1
NE 125th St e/o 12th Ave NE	200	24800	1
12th Ave NE n/o NE 50th St	100	24500	2
12th Ave S s/o S Weller St NB	130	26200	2
S Jackson Btwn 23rd and 25th	160	11300	3
Fremont Bridge	2760	33900	4
S Spokane St at 11th Ave S	780	26600	7
Total	6830	406800	21

4.4.1 Bicycle Crash Data

In all, approximately 1950 bicycle crashes occurred in Seattle from 2009 to 2014, and 21 bicycle crashes occurred at the 12 selected intersections. In comparison to motor vehicle crashes, the number of bicycle crashes was much lower. Figure 4-27 and figure 4-28 show the bicycle crash counts for each year in the study areas and in Seattle. Obviously, they have the same pattern of variation trend from 2009 to 2013. Namely, crashes increased from 2009 to 2013, reached a peak in 2013, and then dramatically dropped in 2014. The increasing trend of bicycle crashes from 2009 to 2013 was caused by growing ridership and population, and the rapid drop in 2014 was caused by the implementation of a bicycle safety management project that started at the end of 2013 (SDOT, 2015).

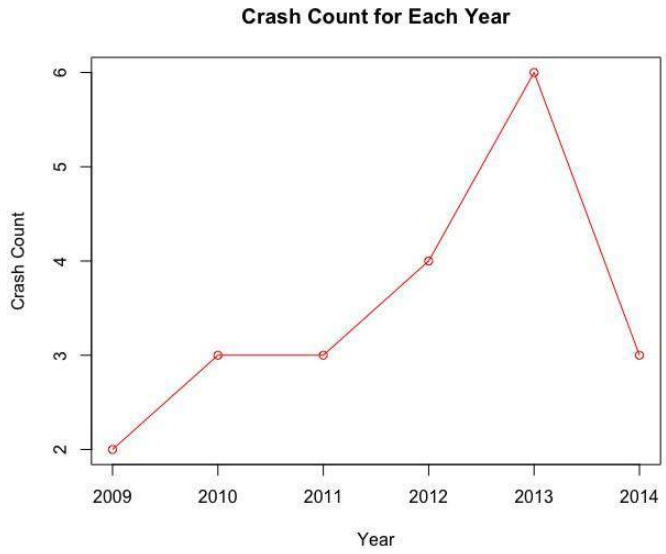

Figure 4-27 Bicycle crash countd for each year in study area

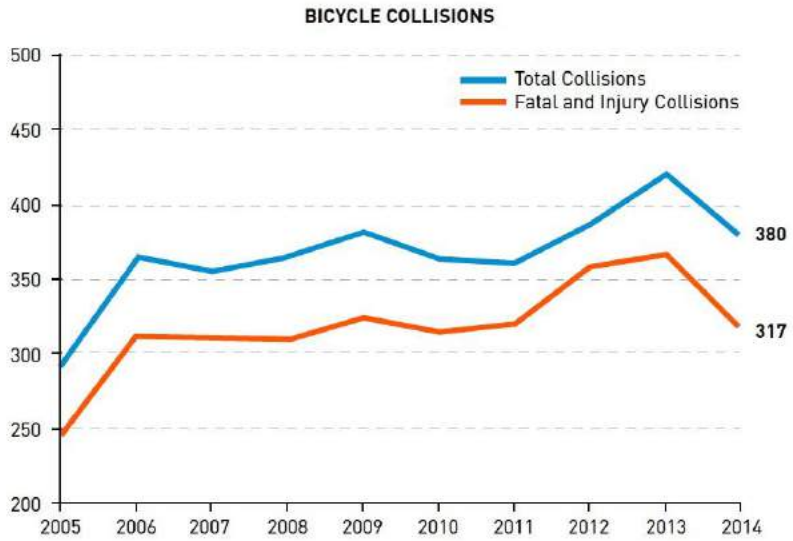

Figure 4-28 Bicycle crash countd for each year in Deattle (Source: SDOT, 2015)

Figure 4-29 shows the crash frequency in the study areas. More than half of the selected intersections had no crash or only had one crash within the intersection area. The main reason is the over-dispersion feature of bicycle crash data (Kim, 2006). Over-dispersion is the characteristic by which variance is larger than the mean value of the data set (Gardner, 1995). Seen from the statistics in this data set, the mean was 1.75 and the variance was 4.39, which verifies the over-dispersion feature of the bicycle crash data set.

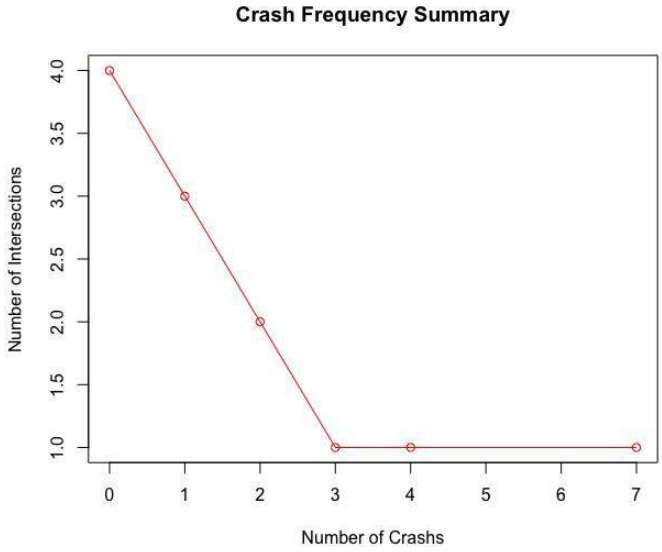

Figure 4-29 Crash frequency summary

Figure 4-30 shows a summary of the collision types from the bicycle crash data related to the study intersections. According to the left figure, most crashes were front end at angle, and rear-end crashes and left-side sideswipe crashes were, respectively, responsible for 10 percent of all crashes. For motorized vehicle crashes, rear-end crashes were the most common (Wang, 2006); however, that is different from the data sample. The main reason is that the speed of bicycles is much lower than motorized vehicle speeds, so bicycles can stop more easily when accidents occur. The right figure shows that 67 percent of all crashes were caused by cyclists striking motorized vehicles, and 33 percent of were caused by motorized vehicles striking cyclists.

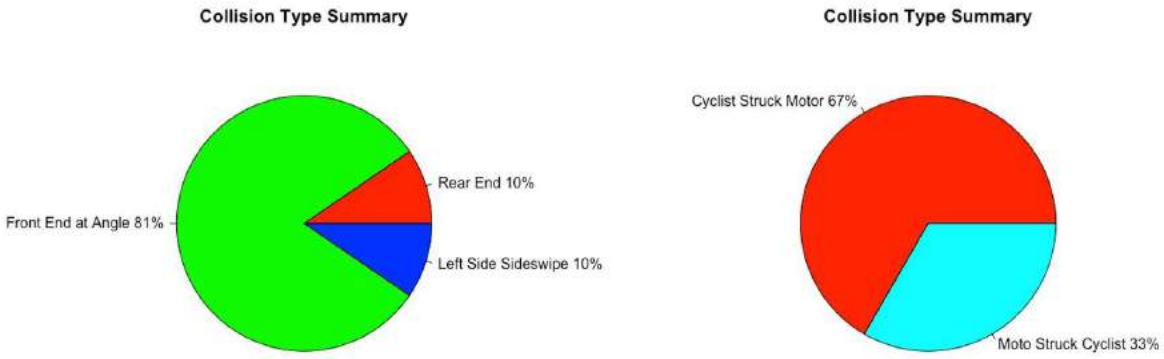

Figure 4-30 Collision type summary

Figure 4-31 summarizes the crash injury severity from the data sample. About 86 percent of all crashes were involved in bodily injuries, and only 14 percent of them were property damage only collisions. The composition is caused by the underreporting of bicycle crashes. Since the crash data set was police-reported, crashes that did not involve bodily injury and insurance indemnity would not be reported to police department (Juhra, 2012). Therefore, reports of property damage only collisions were a small portion of total crashes, and the rest of the crashes were almost all bodily injury involved.

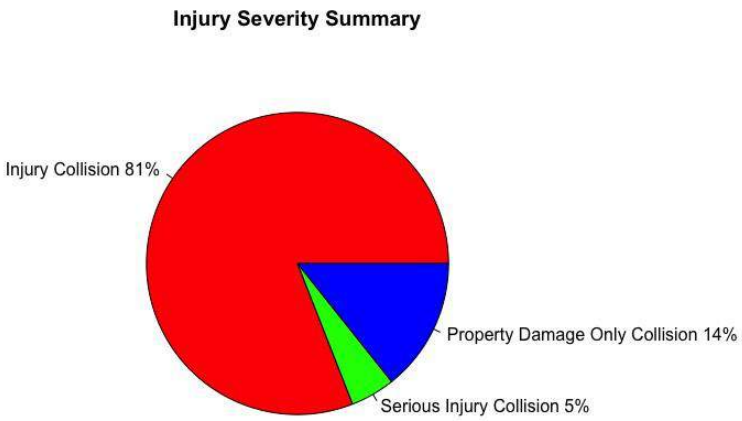

Figure 4-31 Collision injury severity summary

Figure 4-32 depicts the composition of road conditions in the crash data sample and light conditions. As seen in the right figure, 90 percent of crashes occurred under dry road conditions, and only 10 percent happened under wet road conditions. Intuitively, wet road conditions would more frequently lead to crashes; however, the data sample revealed the opposite circumstance. The potential reason for this phenomenon is that people are more willing to choose bikes as a commute travel mode when road conditions are dry, and cycling demand increases when road is in good condition (Thomas, 2009). On the other thand, fewer people travel by bicycle when the road is wet. For the light condition summary, most crashes happened in daylight conditions, and only one-fourth happened on dark streets. The reason is that the demand for bicycling in daylight conditions is higher than demand in street light conditions (Spencer, 2013).

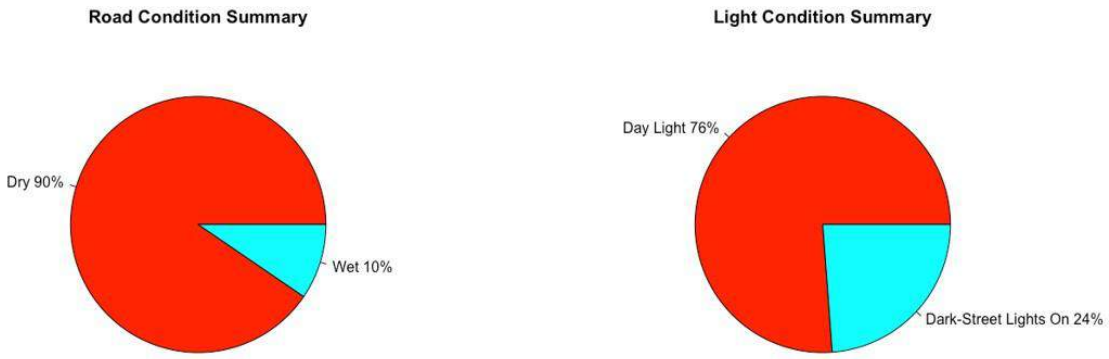

Figure 4-32 Road condition and light condition summary

Figure 4-33 summarizes the weather conditions for the data sample. About 81 percent of crashes happened in clear or partly cloudy weather, and 10 percent of crashes happened, respectively, on overcast days and rainy days. As mentioned above, road conditions and light conditions impact bicycle travelling demand, and weather conditions also significantly affect bicycling demand (Rose, 2011). Therefore, even though the total number of bicycle crashes for clear days is higher than the number for overcast and rainy days, the crash rate is less for clear or partly cloudy days.

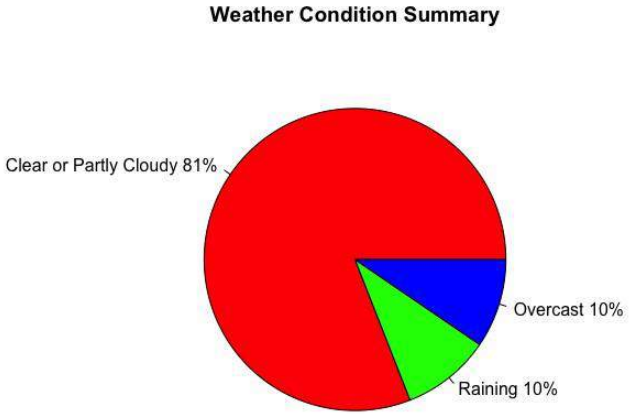

Figure 4-33 Weather condition summary

4.4.2 AADT and AADB

This section summarizes the features of the AADT and AADB data for the Seattle area that were used in this project. Figure 4-34 shows the distribution of AADT. The AADT for the study area ranged from 11,300 per day to 118,700 per day, and the AADT for most of the intersections was between 20,000 per day and 30,000 per day. The mean value was 33,900, with a standard deviation of 29,180.

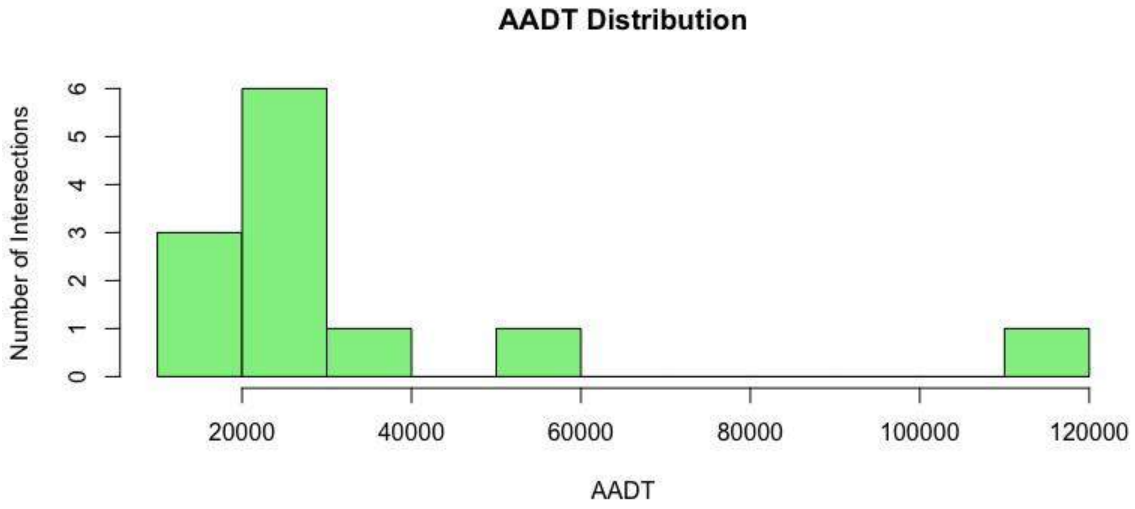

Figure 4-34 Average annual daily traffic volume distribution

Figure 4-35 shows the distribution of AADB data in 2014. Almost all the intersections in the study area had AADB of below 1000 per day, except for the Fremont Bridge, which had 2760 per day in 2014. The most frequent level was between 0 and 200. The mean value was 569, with a standard deviation of 734.

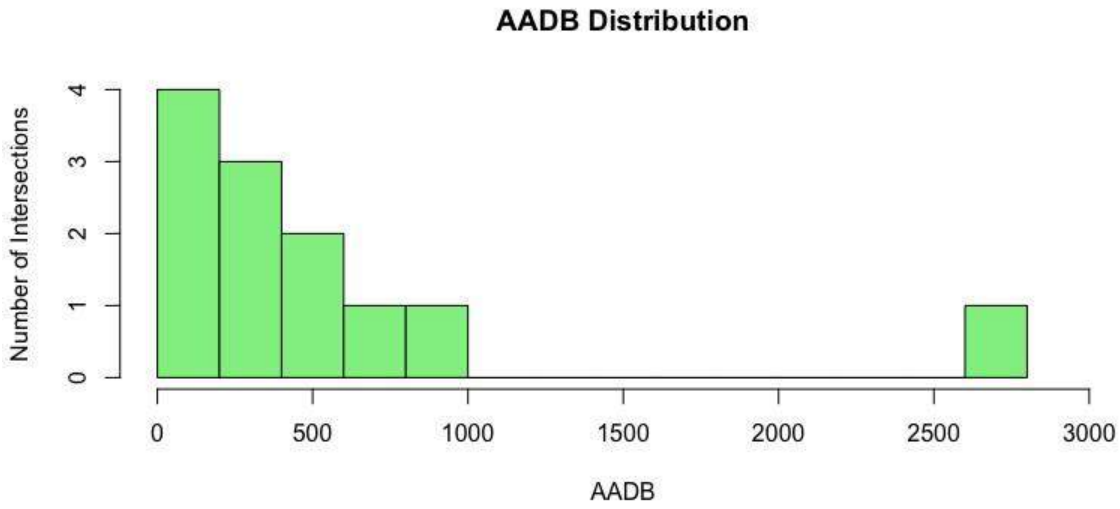

Figure 4-35 Average annual daily bicycle volume distribution

Figure 4-36 shows a preliminary analysis of the relationship among AADT, AADB, and bicycle crashes. Even though the small sample size and the simple linear relationship might not support a finding of the true relationships among them, the basic trend still might be revealed. Figure 4-36 shows the relationship between AADT and AADB. Intuitively, AADT and AADB at the same intersection should have some kind of relationship, for example, some intersections that are responsible for large traffic volumes would not have considerable bicycle volumes (Landis, 1996). However, according to the flat line in figure 4-36, AADT was barely related to AADB in the study area, with a correlation of 0.049. The probably reason for that is that the sample size was too small to reveal the true relationship between AADT and AADB.

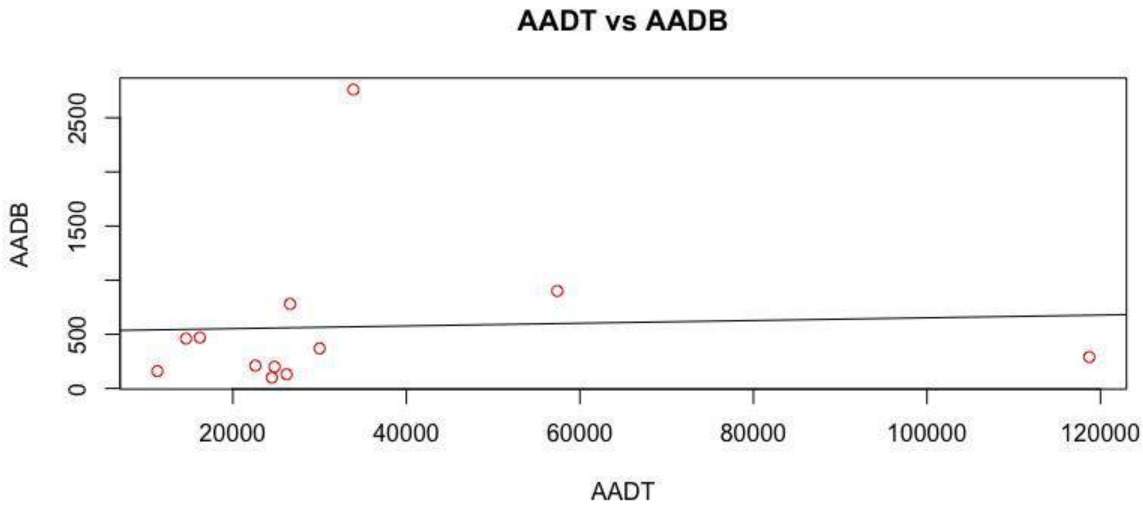

Figure 4-36 Scatter plot of AADT vs AADB

Figure 4-37 shows the relationship between AADT and bicycle crashes. As seen from the scatter plot, there was a weak negative relationship between AADT and bicycle crashes in study area, with a correlation of -0.261. The negative sign indicates that bicycle crashes decrease when AADT increases. The potential reason for that is that the geometric design of the intersections that are responsible for considerable traffic volumes is probably not suitable for bicycle travelling, such as high traffic speed limits (Pucher, 2010).

Figure 4-38 depicts the relationship of AADB and bicycle crashes. As seen from the figure, the line shows a positive relationship between AADB and bicycle crashes, with a correlation of 0.411. The positive relationship indicates that bicycle crashes would increase when AADB increases.

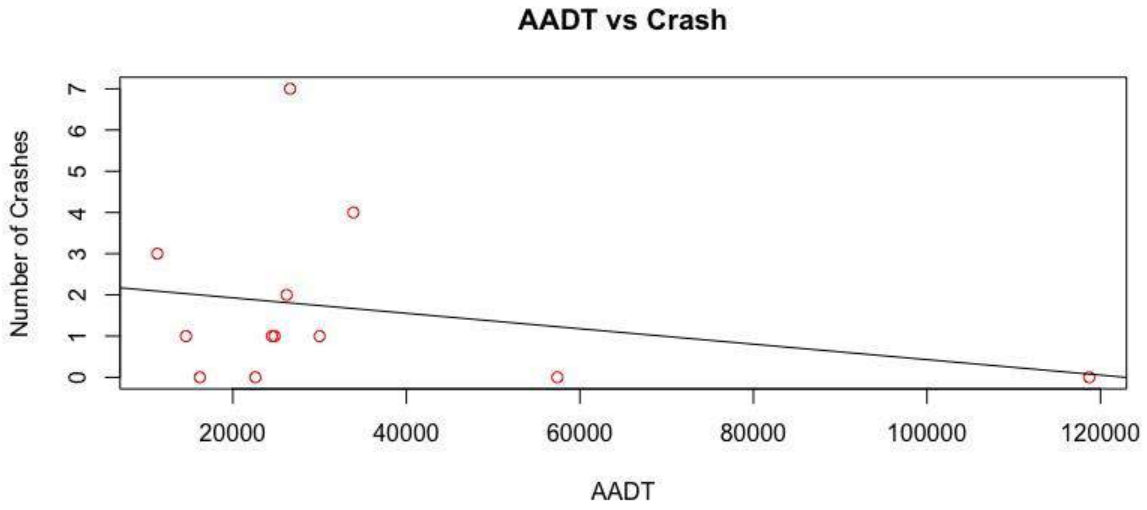

Figure 4-37 Scatter plot of AADT vs crashes

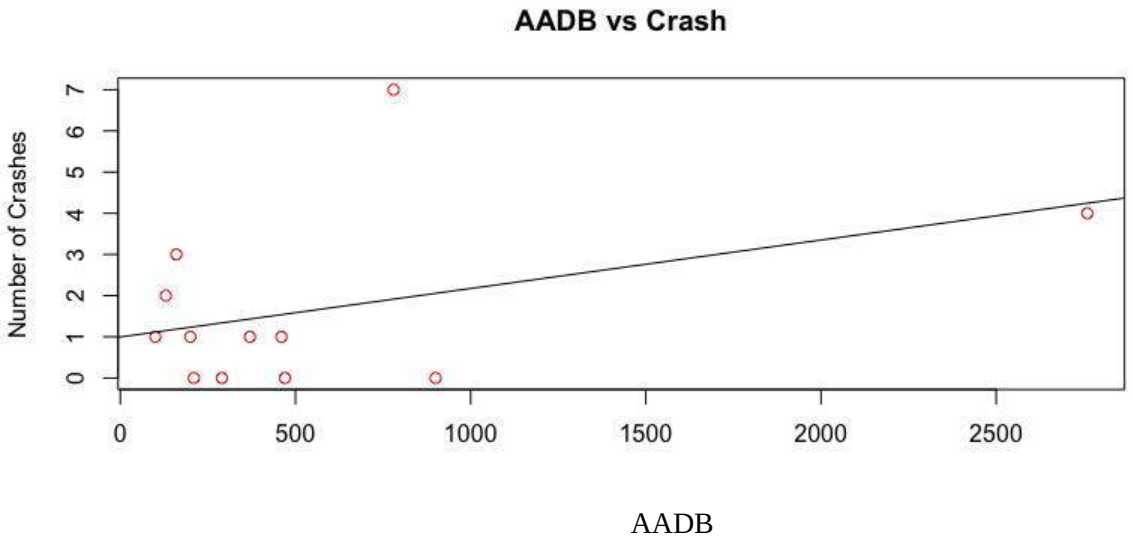

Figure 4-38 Scatter plot of AADB vs crashes

CHAPTER 5 METHODOLOGY

This chapter documents the process of building a jurisdiction bicycle SPF using crowdsourced data. Then different methods that were conducted in the modeling process are documented, including the negative binomial regression model (NBRM), Poisson regression model (PRM), zero-inflated negative binomial regression model, and other hypothesis tests.

5.1 Procedure for Building the SPF Using Crowdsourced Data

The steps of using crowdsourced data to build the SPF was partly different from the steps for establishing an SPF from typical data described by the Federal Highway Administration (2013). The steps were adopted and conducted as described below:

Step 1: Determine the use of the SPF and facility type

Engineers need to identify the use and facility type of the SPF before other steps. Depending on the purpose for the SPF the project scale will be different. The scale difference will influence the following steps, labor requirements, and time requirements. For example, if engineers decide to focus on building a project-level intersection SPF, then the data collection and other following steps will be concentrated in the project area for the intersection. If a statewide network screen is the goal, then the data should be randomly collected at the statewide level. The uses of the SPF may include but not be limited to network screening, project-level prediction, drive CMF, before-after evaluation using the EB method, etc. The facility types may include but not limited to intersections, segments of non-highways, highway segments, ramps, etc. Specifically, most bicycle projects will not include freeways as target facilities because of fewer bicycles on highways and the illegality of riding on freeways.

Step 2: Identify the necessary data

Depending on the uses of the SPF and facility types, the required data will be different. The differences may include sample size and the corresponding data set. Guidance on the minimum sample size can be found in the SPF Decision Guide (Srinivasan et al., 2013). For our project purposes, the effort requirement was larger than “project level” but smaller than the statewide level. In other words, the sample needed to include more than 50 sites, with at least 100 crashes per year for a total group and three years of data. For example, a statewide SPF will require a much larger data set than a local, project-level SPF. However, in some projects, because of labor and time limitations, engineers can slightly change the original requirements to be practical. For example, because of too fewbicycle crashes, in this project, engineers could not meet the 100 crashes per year requirement in either Portland or in Seattle if the samples were chosen randomly, since the total crash number per year for all of Oregon was about 800.

Step 3: Identify the corresponding crowdsourced data

After determining necessary data, engineers need to identify corresponding crowdsourced data. This step may be the critical and most difficult step, since crowdsourced

data are the foundation of this SPF. Engineers may need to use proper judgment with a short-term prediction of what the project would be if a certain type of crowdsourced data were chosen. Depending on project scale, sample size, and data set requirements, specific crowdsourced data need to be found. Crowdsourced data can be retrieved from an online open source, provided by DOTs, purchased from an agency, or adopted from another project. The data should meet the requirements of sample size and facility type. For example, if the project is to build a statewide intersection SPF, then the crowdsourced data should be able to represent all intersections for the state. In this project, the scale focused on two cities, so STRAVA data were chosen for Portland bicycle count since they could represent all city situations. The representation will be discussed later. Another main reason for using STRAVA data was that ODOT had purchased STRAVA data for 2014. However, because no STRAVA data were available for Seattle for this project, engineers decided to use typical count data. In this way, non-crowdsourced data were compared to crowdsourced data at the end.

Step 4: Verify the crowdsourced data

The most important step in using a crowdsourced data is to justifying it. In other words, the crowdsourced data chosen by engineers should be verified for use, including their representation and reliability. Crowdsourced data should be able to represent the population in a project, and the source should be reliable enough to meet engineering requirements. For example, an agency properly collecting crowdsourced data is more reliable than data retrieved from online new media without authors. In this project, STRAVA data were chosen because of their high reliability and fairly good representation. Justifying the representation is necessary since most crowdsourced data can only represent a small proportion of users. User types should be clarified while verifying the representation. Normally, bicyclists can be divided into two large types: recreational cyclists and commuters. For example, in this project, only subpopulation of recreational cyclists and commuters were using the STRAVA application, so the data from them could only represent those users instead of the entire population. Thus the way that the STRAVA data represented the entire population was verified; more detail can be found in the data collection description. Generally, the STRAVA data can represent 1 percent of total users in Oregon.

Step 5: Prepare and clean up the crowdsourced data

Engineers must decide which details from the data should be used to build the SPF. Since crowdsourced data will include a lot of information, unnecessary data detail should be deleted. For instance, the STRAVA bicycle data contained a lot of detail information about each user that would not contribute toward this project's purpose. For basic SPF, traffic counts, bicycle counts, and crash data should be included. Some statistical plotting, such as scatter plots, can be used to determine outliers and human error.

Step 6: Data analysis

Engineers develop a sense of how the data look in order to make decisions. Data analysis is accomplished mainly though statistical plotting and analyzing. Mean, variance, scatter plots, histogram graphs, etc. are ways to analyze collected data. An example can be found in the data description for in this project.

Step 7: Develop the SPF for basic conditions

The first step of establishing the SPF regression is the basic SPF. The basic SPF for bicycles includes traffic counts and bicycle counts as independent variables and crash frequency as the dependent variable. The basic SPF for traffic at intersections is:

$$Y = e^{a} \times (AADT_{minor})^{b} \times (AADT_{major})^{c}$$

where
Y: crash frequency per year
a: estimation of coefficient of exponent
b: estimation of coefficient of AADT on minor road
c: estimation of coefficient of AADT on major road.

According to Nordback et al., (2014) the bicycle SPF for intersections looks similar to that for traffic but includes AADB:

$$C = e^{a}(AADT)^{b}(AADB)^{c}$$

where
C: number of intersection motorist-cyclist collisions during the study period;
AADT: the annual average daily motorized traffic passing through the intersection;
AADB is the annual average daily bicycle traffic passing through the intersection;
a b c: the exponents estimated by the model.

Model coefficients can be estimated in statistical software, such as R® programming or Microsoft Excel®.

Step 8: Develop CMFs for specific treatments

A crash modification factor (CMF) is a multiplicative factor used to compute the expected number of crashes after a given countermeasure has been implemented at a specific site (Crash Modification Factor Clearinghouse, n.d.). As an important component in SPFs, CMFs address specific and detailed conditions of SPFs. For instance, if a basic SPF was built on the basis of a two-leg intersection, then if the target site had a five-leg intersection, the prediction of SPF would change because the five legs would influence crash occurrence. This difference can be addressed with a CMF. This process was not included in the case for this project.

Step 9: SPF interpretation and discussion

Interpretation of coefficients, understanding the dependent and independent variables, and how to interpret log transformation and exponents etc. are the main components for building basic SPFs. The model may be interpreted within local jurisdiction and the constrains of the scale of the data unless the representation and extrapolation at a larger scale can be justified. For example, the SPF built on STRAVA data should be only interpreted for STRAVA users; however, because the representation for the whole population was verified in Portland, the discussion can refer to Portland bicyclists.

Step 10: Document establishment of the SPF

According to the Federal Highway Administration, (2013) documentation should include but not limited to the following:

- Crash type(s)/severity(s) for which the SPF was estimated
- Total number of crashes (by type and severity) used in the estimation
- Purpose of the SPF
- State(s)/county(s)/city(s) that were used
- Facility type
- Number of years used in the estimation of SPF
- Number of units (segments, intersections, ramps)
- Minimum, maximum, and average length of segments
- Minimum, maximum, and average AADT
- Minimum, maximum, and average values for key explanatory variables
- Coefficient estimates of the SPF
- Standard errors of the coefficient estimates
- Goodness of fit statistics
- Discussion of potential biases or pitfalls.

Depending on the project scale and purpose, engineers can adopt the documentation process for their uses. The main purpose of documentation is to help others and the engineers themselves to recall and repeat the process as accurately as possible.

5.2 Statistic Regression

Many statistical regressions can be used to describe the count distribution, but success in analyzing traffic counts and bicycle counts seems rare. Since Nordback et al., (2014) used the negative binomial regression model (NBRM) successfully in Boulder, Colorado, this project mainly employed this model for the Portland and Seattle cases. However, other models should

be tested if NBRM is not significant, including zero-inflated negative binomial regression, Poisson regression, etc.

5.2.1 Negative Binomial Regression Model

The negative binomial regression model is a popular generalization of the Poisson regression model since it does not follow the features of equality of mean and variance (Hilbe, 2011). Previously, bicycle crash data proved to be over-dispersed (Ladron, 2004), which is characterized by sample variance that is larger than the sample mean. Therefore, the negative binomial regression model is more suitable for bicycle safety performance function development. According to Long, (1997), the NBRM can be used to address data with over-dispersion characteristics, which the Poisson regression model (PRM) cannot fit. The details of how the NBRM can address over-dispersion can be found in Long's book but are not included here. Poisson distribution has one important property of equal-dispersion (Long, 1997):

$$\mathrm{Var(y)} = \mathrm{E(y)} = \mathrm{u}$$

where:
var: variance u: the rate that number of times an event has occurred per site during a period.

Long (1997) said that the PRM can barely fit in practice because within major applications the condition variance is larger then the condition mean. Gourieroux et al., (1984) mentioned that in the situation that the mean is correct but with over-dispersion, then estimates from PRM are consistent but not efficient. In addition, the standard errors from the PRM will be biased downward, which can result in spuriously large z test values (Cameron & Trivedi, 1986), and that will mislead interpretation and judgment of fitness of the model. In other words, over-dispersion may cause bias and inefficiency in the modeling process when PRM is used for bicycle and traffic count data that have the property of over-dispersion. The over-dispersion of the data in this project was described in the data collection and description section.

The negative binominal model handles over-dispersion in the data sample by introducing a stochastic component to the log-linear Poisson mean function relationship (Hilbe, 2011). The basic model specification is given by the following equation:

$$\ln u_i = \varepsilon + \sum X_i \beta_n$$

where $u_i$is the expected value of the response variable, which is the number of bicycle crashes at intersection i;

ε is the random error term that aims to deal with over-dispersion in the data sample;

$X_i$is the independent variable that is the AADT and AADB at intersection i;

β_n is the estimated coefficient of each independent variable.

The model assumes the response of the variable follows the negative binominal probability distribution that is given by the equation below:

$$P(y_i|X_i) = \frac{\Gamma(y_i + \upsilon_i)}{y_i!\,\Gamma(\upsilon_i)} \frac{\upsilon_i^{\upsilon_i}}{(\upsilon_i + \mu_i)} \frac{\mu_i^{y_i}}{(\upsilon_i + \mu_i)}$$

where Γ is the gamma distribution function;

$\upsilon_i$is the gamma distribution parameter divided by the dispersion parameter;

y_i is the expected value of the response variable.

The variance and the standard deviation of the the response variable are given by:

$$Var(y_i|X_i) = \mu_i + \frac{\mu_{i2}}{\upsilon_i}$$

$$\sigma_i = \sqrt{\mu_i + \frac{\mu_i^2}{\upsilon_i}}$$

5.2.2 Poisson Model

Poisson regression is a member of the class of models known as generalized linear models (GLM), which is the standard method used to analyze count data (Cameron, 2013). The model assumes the response variable has a Poisson distribution that is characterized by the equality of mean and variance (Zou, 2004). The basic model specification is given below:

$$\log(\mu) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

where μ is the expected value of the response variable, in this project, the expected number of bicycle crashes for a given average annual daily traffic and bicycle volume at intersection i;
X_i are the independent variables of group i, in this project, traffic volume and bicycle volume;
β_i is the estimated coefficient of independent variable X_i.

The Poisson probability distribution is given by:

$$P(y_i|x_i) = \frac{\lambda_{y_i}}{y_i!} e^{-\lambda}$$

where λ is the sample mean and variance;
x_i is the given independent variable;

y_i is the number of bicycle crash at intersection i.

5.2.3 *Zero-Inflated Negative Binomial Model*

By the observation that some intersections in the study area had zero bicycle crashes during the study periods, the bicycle crash generation process might differ for the intersections that had zero bicycle crashes and the intersections that had plenty of bicycle crashes. The zero-inflated negative binomial model is not only able to handle over-dispersion in the data sample but can also provide an effective way to model the excess zeros (Lambert, 1992). The model assumes that there are two possible data generation processes for each data point.

$$\varphi^i \; y_i \sim \begin{cases} 0 & with\ probability\ \ \varphi_i \\ g(y_i \mid X_i) & with\ probability\ \ 1-\varphi_i \end{cases}$$

where y_i is the expected value of the response variable, which is the number of bicycle crashes at intersection i.

In this case, for each intersection i, there are two bicycle crash generation processes. One is responsible for generating zero bicycle crashes with probability φ_i and another one is in charge of generating bicycle crashes from the negative binomial distribution with probability $1-\varphi_i$.

The probability of $\{Y_i = y_i | X_i\}$ is given by following equations.

$$(Y_i = y_i|X_i, Z_i) = \begin{cases} (\gamma' Z_i) + \{\{11 - - \varphi\varphi((\gamma\gamma'' ZZ_{ii}))\}\}gg((0y|_i X|X_i)_i) & ifif \quad yy_{ii} => 00 \end{cases}$$

When the probability φ_i depends on the characteristics of observation i, φ_i is written as a function of γ' , where Z_i is the vector of zero-inflated covariates and γ' is the vector of zero-inflated coefficients to be estimated. The mean and variance of the zero-inflated negative binomial model are calculated by the equations below.

$$(y_i|X_i, Z_i) = \mu(1-\varphi_i)$$
$$V(y_i|X_i, Z_i) = \mu_i(1-\varphi_i)(1+\mu_i(\varphi_i+\alpha))$$

5.3 Measures of Goodness of Fit

In order to verify and compare between regressions, basic measures of goodness of fit are introduced in this section, but the results analyses may include more ways to judgt than described in this section.

5.3.1 Likelihood Ratio Test

The likelihood ratio test is a statistical model that aims to compare the goodness of fit of two nested models, and one of two models is the special case of the other (Anisimova, 2006). In this project, the likelihood ratio test was used to compare the performance of the Poisson regression model and negative binomial model. The test makes two hypotheses, a null hypothesis and an alternative hypothesis, and the null hypothesis would be rejected or be accepted by the test results based on the likelihood ratio function, which is how many times more likely the data fit one model better than the other model. The test statistic is given by the equation:

$$\Lambda(x) = \frac{\sup\{L(\theta|x) : \theta \in \Theta_0\}}{\sup\{L(\theta|x) : \theta \in \Theta\}}$$

where $L(\theta|x)$ is the likelihood function that is calculated by the probability density function of the model. The likelihood ratio statistic follows chi-square distribution, then, p-value can be computed according to each significance levels.

5.3.2 Vuong Non-Nested Hypothesis Test

The Vuong non-nested test is based on a comparison of the predicted probabilities of two models that do not nest, for examples the comparison of zero-inflated count models with their non-zero-inflated analogs (e.g., zero-inflated negative binomial model versus ordinary negative binomial model); in this project, it would be used for comparing the performance of negative binomial regression model and the zero-inflated negative binomial regression model. It is a likelihood ratio-based test for model selection using the Kullback-Leibler information criterion (Vuong, 1989), and the test statistic is given by the equations below.

$L(\beta_{ML,1}, \beta_{ML,2})$

$$Z = \frac{}{\sqrt{N}\overline{\omega_N}}$$

where the numerator of the statistic is the difference between the maximum likelihoods of the two models, which is calculated as

$$L(\beta_{ML,1}, \beta_{ML,2}) = L_N - L_{2N} - \frac{}{logN\ K}_1 - K_2$$

1

2

where K_1 and K_2 are the numbers of parameters in model 1 and model 2, respectively. The term in the denominator of the expression for Z is defined by setting ω_N^2 equal to either the mean of the squares of the pointwise log-likelihood ratios l_i, or to the sample variance of these values, where

$$l_i = log \frac{f_1(y_i|x_i, \beta_{ML,1})}{f_2(y_i|x_i, \beta_{ML,2})}$$

The test statistic asymptotically follows standard normal distribution, then, the p-value can be computed according to each significance levels.

CHAPTER 6 RESULTS AND DISCUSSION

The model regressions were applied to Portland and Seattle separately because no crowdsourcing was valid in Seattle. Fortunately, ODOT purchased 2014 STRAVA data for Oregon, and engineers were able to apply the method to this crowdsourced data set. In this chapter, regression results are analyzed for both cities.

6.1 Results and Analyses for Portland

In this section, the Poisson regression model and negative binomial model are applied to the Portland data set. The modeling processes were completed in the R® program. Results are included for both models and are interpreted separately. Also included is a comparison of the two models and suggestions for the better model for SPF.

6.1.1 Poisson Regression Model

PRM was conducted and the results are shown in Table 6-1**,** Poisson regression results (log link). The second column shows the estimated coefficient for each variable, and it indicates the level of a variable's influence on the dependent variable (crash number). For example, the estimated coefficient of AADT means that for each one-unit increase in AADT, the expected log count of crashes increases by 5.245e-05 (e-05 is equal to 0.00005), holding other variables constant. Note that it is "log count" instead of a "count" because there is a log link function between the dependent and independent variables in regression. (More details are shown in the Methodology section.) Another important component in this table is the p value and the significance level. The number 5.245e-05 seems small, and it is because 1) an increase in AADT by one unit does not have a lot of influence on crashes, but AADT normally changes on a larger scale; 2) the bicycle crash number has a smaller order of magnitude than motorized vehicle crashes. The significance level is shown based on p-value, which indicates whether the variable is significant. Three stars, two stars, one star, and no star represent "convincing evidence," "convincing evidence," "moderate evidence," and "week evidence," respectively. For instance, the p-value for AADT is 4.82e-05 and the corresponding significance level is ***, which means there is convincing evidence that the AADT has influence on the number of crashes.

Table 6-1 Poisson regression results (log link)

Variable	Estimated coefficient	SE	z	Pr(>|z|)	Significance level
Intercept	-1.571	4.237e-01	-3.708	0.000209	***
AADT	5.239e-05	1.289e-05	4.064	4.82e-05	***
STRAVA	8.877e-05	3.549e-05	2.501	0.012371	*

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 6-2 and table 6-3 show the 95 percent interval of each variable with log function and without, respectively. In table 6-2, using AADT as an example, we can conclude with 95 percent confidence that the expected log count of crashes increases between 2.702283e-05 to 7.778322e-05 with one-unit increases in AADT. Again, the numbers are small because the unit is changing by one AADT. With a change of 1,000 or 10,000 AADT, the number would be correspondingly larger. When we transform the log count back to a normal count, the interpretation is that for every unit increase in AADT, the crash occurrence increases 1.0000525 times than before, holding other variables constant, and with a 95 percent confidence, expected crash counts will increase between 2.702283e-05 to 7.778322e-05 times with a one-unit increase in AADT.

Table 6-2, 95 percent interval results (log link)

Variable	Estimation	2.5%	97.5%
Intercept	-1.571	2.454689	-7.873487e-01 - 7.778322e-05
AADT	5.239e-05	2.702283e-05	1.544616e-04
STRAVA	8.877e-05	1.324681	

Table 6-3, 95 percent interval results

Variable	Estimation	2.5%	97.5%
Intercept	0.2077537	0.08588987	0.4550497
AADT	1.0000524	1.00002702	1.0000778
STRAVA	1.0000888	1.00001325	1.0001545

6.1.2 Negative Binomial Regression Model

NBRM was conducted, and the results are shown in table 6-4. The second column shows the estimated coefficient for each variable, and it indicates the levels of a variable's influence on the dependent variable (crash number). For example, the estimated coefficient of STRAVA means that for each one-unit increase in STRAVA, the expected log count of crashes will increase by 8.942e-05 (e-05 is equal to 0.00005), holding other variables constant. The number 8.942e-05 seems small, and it is because 1) increasing STRAVA by one unit does not have a lot of influence on crashes, but STRAVA typically is on a large scale; 2) there are many fewer

bicycle crashes than motorized vehicle crashes; 3) STRAVA only represents a proportion of all bicycle counts. The p-value for STRAVA is 0.012836 and the corresponding significance level is *, which means we have only moderate evidence that the STRAVA count has influence on the number of crashes.

Table 6-4, Negative binomial regression results (log link)

Variable	Estimation	SE	z	Pr(>|z|)	Significance level
Intercept	-1.575	4.281e-01	-3.679	0.000234	***
AADT	5.245e-05	1.311e-05	4.000	6.34e-05	***
STRAVA	8.942e-05	3.594e-05	2.488	0.012836	*

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 6-5 and table 6-6 show the 95 percent interval for each variable for NBRM. In table 6-5, using STRAVA as an example, we can conclude with 95 percent confidence that the expected log count of crashes will increase between 1.278175e-05 to 1.566332e-04 with a one-unit increase in STRAVA count. Again, the numbers are small because there is only a one-unit change in STRAVA. If the change were 1,000 or 10,000 STRAVA, the number would be correspondingly larger. When we transform the log count back to a normal count, the interpretation is that for every unit increase in STRAVA, the crash occurrence will increase 1.0000525 times than before, holding other variables constant, and with 95 percent confidence, expected crash counts will increase between 1.00001278 to 1.0001566 times with a one-unit increase in STRAVA count.

Table 6-5, Negative binomial regression results 95 percent interval (log link)

Variable	Estimation	2.5%	97.5%
Intercept	-1.575	-2.467539	-7.822395e-01
AADT	5.245e-05	2.670806e-05	7.834061e-05
STRAVA	8.942e-05	1.278175e-05	1.566332e-04

Table 6-6, Negative binomial regression results 95 percent interval.

Variable	Estimation	2.5%	97.5%

			0.4573805
Intercept	0.2070074	0.08479324	
			1.0000783
AADT	1.0000525	1.00002671	
			1.0001566
STRAVA	1.0000894	1.00001278	

6.1.3 Comparison of NBRM and PRM

The information of residual deviances is used to test the goodness of fit for both models to test the overall model. Table 6-7 shows the deviance, degree of freedom, and significant value (P) for both. The residual deviance represents the difference between the deviance of the current model and the maximum deviance of the ideal model in which the predicted value is totally the same as the observed value. So when the residual is small, the goodness of fit test will not be significant, and the p value will be larger than 0.05. The p values for both models are large, which indicates that both models fit the data well. Comparing between the two models, NBRM has smaller residual deviance and better goodness of fit, even if the difference is not very obvious, so it performs better than PRM.

Table 6-7, Goodness of fit test by deviance

model	Residual Deviance	Degree of freedom	P
NBRM	52.15687	47	0.280344
PRM	53.19661	47	0.2478945

As mentioned in the methodology and literature review chapter, the NBRM has an assumption that the variance is not equal to the mean. Table 6-8 shows the dispersion type of the two models. Residual deviance larger than degree of freedom indicates over-dispersion. In the two model, the results show so, but only slightly. In addition, the two time difference of likelihoods show the dispersion as well. The two time difference of likelihoods is a little larger than 0, indicating slight over-dispersion. The slight over-dispersion is consistent with previous results that the NBRM fits the data set only slightly better than PRM.

Table 6-8, Check dispersion

model	Residual deviance	Degree of Freedom	likelihoods	Two times difference

NBRM	52.157	47	52.15687	0.01133067
PRM	53.197	47	53.19661	

6.2 Results and Analyses for Seattle

This section discusses the analyses of the modeling results for tje Seattle data. Similar to the previous analysis of the Portland data, this section also compares different models and uses the model with the best fitness as the final model.

6.2.1 Poisson Model

Table 6-9 shows the estimation results of the Poisson regression model. According to the modeling results, both AADB and AADT have impacts on the number of bicycle crashes at intersections. The sign of estimated coefficient of variable for AADB is positive, which indicates the relationship between AADB and bicycle crashes is positive. The value of coefficient means that the number of bicycle crashes at intersections in the study area will increase 0.00061 when AADB increases by one unit. The hypothesis test indicates that the impact of AADB on the number of bicycle crashes is significant at the significance level of 0.05. The variable AADT has a negative effect on the response variable, and the hypothesis test reveals the impact is significant at the 0.1 significance level. The modeling results are consistent with the preliminary analysis of the relationship between AADB, AADT, and the number of bicycle crashes, which was discussed in data description.

Table 6-9 Poisson regression results

Variable	**Estimation**	**Robust SE**	**z**	**Pr(>|z|)**
Intercept	1.040285	0.5005479	1.822	0.03768248
AADB	0.000612304	0.000131466	2.437	3.2003E-06
AADT	-3.54547E-05	1.80875E-05	-1.479	0.04997486

Table 6-10 provides the value of the 95 percent confidence interval of estimated coefficients. The results means that 95 percent of the time, the true amount of the impact of the independent variables on the response variable would be between 2.5 percent and 97.5 percent.

Table 6-10 Poisson regression model estimation 95 percent confidence interval

Variable	Estimation	2.50%	97.50%
Intercept	1.040285	0.05921079	2.021359
AADB	0.0006123	0.00035463	0.00086998
AADT	-3.545E-05	7.091E-05	-3.24E-09 -

6.2.2 Negative Binomial Model

As mentioned previously, bicycle crash count data have the feature of over-dispersion. Thus, since the Poisson regression model assumes that the response variable is distributed Poisson, and the mean is the same as the variance, the over-dispersion could not be handled perfectly by the Poisson regression model. However, the negative binomial regression model is able to handle the over-dispersed data sample by introducing a stochastic component to the loglinear Poisson mean function relationship. In order to check the assumption of the over-dispersed bicycle crash data sample, the scatter plot of estimated mean value versus the estimated variance is provided in figure 6-1. According to the figure, most data points are located under the line, which means the data sample is over-dispersed. If the mean and the variance of the distribution that the sample data follow are the same, then the data points should be located along the line. We also calculated the sum of square of residuals of Poisson model divided by the degree of freedom of residuals, which should be no more than 1 if the data are not over-dispersed. The value of 2.39 indicates that the data are over-dispersed. Therefore, the negative binomial model should be more suitable for this kind of data.

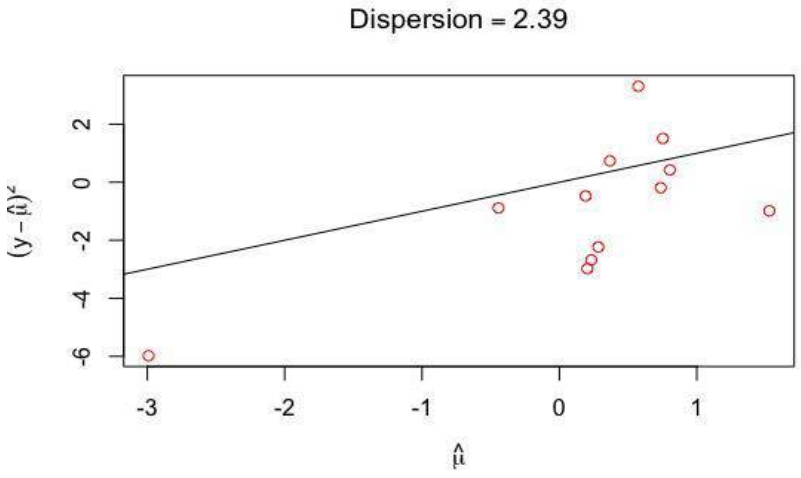

Figure 6-1 Dispersion of data

Table 6-11 provides the modeling results for the negative binomial regression model. The modeling results who that both AADT and AADB have an impact on the response variable. The signs of estimated coefficients are consistent with the Poisson regression model results, positive impact for AADB and negative impact for AADT. The significant level for AADT and AADB, respectively, are 0.05 and 0.1. Table 6-12 provides the 95 percent confidence interval for the estimated coefficients. The results mean that 95 percent of the time, the true amount of impacts for the independent variables on the response variable would be between 2.5 percent and 97.5 percent.

Table 6-11 Negative binominal regression results

Variable	Estimation	Standard Deviation	z	Pr(>|z|)
Intercept	1.18E+00	7.70E-01	1.532	0.1256
AADB	6.90E-04	3.88E-04	1.777	0.0756
AADT	-4.30E-05	3.18E-05	-1.354	0.1759

Table 6-12 Negative binomial regression model estimation 95 percent confidence interval

Variable	Estimation	2.50%	97.50%
Intercept	1.18E+00	-2.08E-01	3.06E+00
AADB	6.90E-04	-6.71E-05	1.61E-03
AADT	-4.30E-05	-1.27E-04	2.03E-06

6.2.3 Zero-Inflated Negative Binomial Model

Since excess zeros occurred in the bicycle crash count data set, the zero-inflated negative binomial model was introduced to deal with the excess zeros. The results of the performance comparison of the zero-inflated negative binomial model and normal negative binomial model are provided in the section of measures of goodness of fit by using the Vuong test.

Table 6-13 shows the results of the negative binomial regression coefficients for each of the variables, along with standard errors, z-scores, and p-values for the coefficients. The estimates have the same signs, with the signs of both the Poisson model and normal negative binomial model. The impacts on the response variable for AADT and AADB are at significance level of, respectively, 0.1 and 0.05.

Table 6-13 Zero-inflated negative binominal - count model coefficients

Variable	Estimation	Std. Error	z	Pr(>|z|)
Intercept	1.18E+00	4.32E-01	2.734	0.00626
AADT	-4.30E-05	1.05E+00	0.786	0.43202
AADB	6.90E-04	4.61E-04	1.497	0.13429

Table 6-14 Zero-inflated negative binominal - count model doefficients 95 percent CL

Variable	Estimation	2.50%	97.50%
Intercept	1.18E+00	3.34E-01	2.03E+00
AADT	-4.30E-05	-1.27E-04	2.03E-06
AADB	6.90E-04	-2.13E-04	1.59E-03

Table 6-15 provides the results that correspond to the inflation model, which includes logit coefficients for predicting excess zeros along with their standard errors, z-scores, and p-values. Table 6-16 shows the 95 percent confidence interval of estimated coefficients.

Table 6-15 Zero-inflated negative binominal - zero-inflation model coefficients

Variable	Estimation	Std. Error	z	Pr(>|z|)
Intercept	-1.77E+01	1.41E+03	-0.013	0.99
AADB	-4.19E-04	3.00E+00	0	1

Table 6-16 Zero-inflated negative binominal - zero-inflation model coefficients 95 percent CL

Variable	Estimation	LL	UL
Intercept	-1.77E+01	-2.77E+03	2.74E+03
AADB	-4.19E-04	-5.87E+00	5.87E+00

6.2.4 Measures of Goodness of Fit

As mentioned previously, the likelihood ratio test aims to compare the performance of nested models, such as the Poisson model and negative binomial model. Table 6-17 shows the results of the comparison of two models. According to the modeling results, even the significance level of the test result is not very high, but the negative binomial model handled the over-dispersion feature of the sample data better than the Poisson regression model at a 0.1 significance level,

Table 6-17 Likelihood ratio test results

Model	Degree of Freedom	Log-likelihood	Degree of Freedom	Chisq	Pr(>Chisq)
1	3	-19.933			
2	4	-18.801	1	2.2624	0.1325

As described previously, the Vuong non-nested hypothesis test is meant to compare the performance of non-nested models, such as the zero-inflated negative binomial model and the normal negative binomial model. Table 6-18 shows the test results, where model 2 is the normal negative binomial model and model 1 is the zero-inflated negative binomial model. According to the results, model 2 performed better than model 1 at a significance level for both AIC and BIC criteria. The potential reason is that the zero-inflated negative binomial model assumes that a response variable has two different generation processes, but the assumption might be violated in the bicycle crash count data set, and another potential reason is that the sample size was too small to get the true relationship between the independent variables and response variables.

Table 6-18 Vuong non-nested hypothesis test-statistic

Parameter	Vuong z-statistic	H_A	p-value
Raw	-3.93E-01	model2 > model1	0.34707
AIC-corrected	-3.59E+07	model2 > model1	< 2e-16
BIC-corrected	-4.46E+07	model2 > model1	< 2e-16

CHAPTER 7 CONCLUSION AND RECOMMENDATION

The chapter provides recommendation for engineers and city planners for building an SPF with crowdsourced data. The conclusions and recommendations include suggestions for enhancement of GIS tools, procedures for building an SPF using crowdsourced data, data collection to build an SPF, and SPF modeling.

7.1 Enhancement of GIS Tools

This project improved the usability of a GIS tool that had been created during a previous PacTrans project. Engineers and planners can use this tool to estimate bicycle exposure when conducting safety analyses. This achievement makes the tool more user-friendly and decreases computation time. The new tool only needs 2 hours for the bicycle network of Seattle, which previously required 22 hours in the original tool. Thus, calibration can be done within a single work day. The improvements documented in this report can provide guidance for practitioners who will use the tool and provide helpful information for future researchers who might continue to advance the underlying methods.

7.2 Establishing an SPF by Using Crowdsourced Data

One of the most important achievements was to build a repeatable process for building an SPF using crowdsourced data. This project used STRAVA counts as an example of crowdsourced data to create a procedure for establishing a bicycle SPF. Other DOTs and transportation agencies can follow the process of building an SPF for different jurisdictions. The requirements for building an SPF for different scale jurisdictions and different purposes for the SPF were summarized and proved by the Portland study case.

7.3 Data Collection for Building an SPF

This project used crowdsourced data – STRAVA bicycle counts— to build an SPF for Portland, Oregon. Even though the STRAVA data were not available in Seattle and engineers had to use count data to build an SPF for Seattle, this limitation provided engineers with a chance to compare the processes of building an SPF based on crowdsourced data and non-crowdsourced data.

The most important part of using crowdsourced data is to verify the representation of them. Representation is a very basic foundation to making decisions. In addition, even though it may be relatively easier for engineers to get crowdsourced data that include massive amounts of information and that cover a much larger spatial area, we recommend that engineers clean up the data set as early as possible, since a lot of information from the crowdsourced data may not be useful for achieving project goals and could distract attention and even hide more importation information Engineers should also focus on the errors in crowdsourced data. For instance, the

STRAVA data had some routes that were double- or triple-counted because of GPS assignment errors.

7.4 Modeling for SPF

Several different regressions were applied in this project to build an SPF, including negative binomial regression model, Poisson regression model, and zero-inflated negative binomial model. Different tests were conducted to verify and compare those models. The results showed that the negative binomial regression model and Poisson regression model fit better to the Portland data set than the Seattle data set. Engineers suggest that using STRAVA data to build an SPF for Pacific Norwest cities is justified and is more efficient than using traditional count data. The model results regarding significance were similar to those of previous studies that used traditional count data, which indicates that jurisdictions can use crowdsourced data rather than labor-consuming traditional count data, but with attention to representation of the population.

7.5 Limitations and Future Work

Because of data availability and time limitations, this project faced several limitations described below, and work to address those limitations also is listed:

1. Due to AADT data availability, even though the overall data set could represent the population, part of the data set was not collected randomly.
2. STRAVA data can represent a proportion of total bicycle counts, but the percentage of representation and demographic representation have not been proven.
3. Engineers can use a larger data set to build an SPF and may find other significant results.
4. STRAVA data are not an open source data, and steps to cooperation and investment are necessary. Other researchers can try other open source crowdsourced data to build an SPF.

ACKNOWLEDGMENTS

PacTrans funded this project, and the authors are grateful to ODOT/TPAU for purchasing and providing the STRAVA data. Alex Bettinardi in ODOT/TPAU provided the initial representation of STRAVA data.

REFERENCES

Agran, Phyllis F., Dawn N. Castillo, and Dianne G. Winn. "Limitations of data compiled from police reports on pediatric pedestrian and bicycle motor vehicle events." Accident Analysis & Prevention 22.4 (1990): 361-370.

American Association of State Highway and Transportation. (2010). Highway Safety Manual (1st ed.).

Anisimova, Maria, and Olivier Gascuel. "Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative." Systematic biology 55.4 (2006): 539-552.

Asgarzadeh, Morteza, et al. "The role of intersection and street design on severity of bicyclemotor vehicle crashes." Injury prevention (2016): injuryprev-2016.

Attewell, Robyn G., Kathryn Glase, and Michael McFadden. "Bicycle helmet efficacy: a metaanalysis." Accident Analysis & Prevention 33.3 (2001): 345-352.

Beck, L. F., A. M. Dellinger, and M. E. O'Neil. Motor Vehicle Crash Injury Rates by Mode of Travel, United States: Using Exposure-Based Methods to Quantify Differences. American Journal of Epidemiology, Vol. 166, No. 2, Jul. 2007, pp. 212–218.

Boulder, M., & Even, S. (2012). Safe Streets Boulder, (February).

Brabham, D.C. (2013). Using crowdsourcing in government. Washington, DC: IBM Center for Business Government: Collaborating Across Boundaries Series. Retrieved from http://www.businessofgovernment.org/sites/default/files/Using%20Crowdsourcing%20In %20Government.pdf

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

Brandes, U. A Faster Algorithm for Betweenness Centrality. Journal of Mathematical Sociology, Vol. 25, 2001, pp. 163–177.

Broach, J., Dill, J., & Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. Transportation Research Part A: Policy and Practice, 46(10), 1730–1740. https://doi.org/10.1016/j.tra.2012.07.005\

Cameron, A. C., & Trivedi, P. K. (1986). Econometric Models Based on Count Data: Comparisons and Application of some Estimators and Tests. Journal of Applied Econometrics, 1(1), 29–53.

Cameron, A. Colin, and Pravin K. Trivedi. Regression analysis of count data. Vol. 53. Cambridge university press, 2013.

Casello, J. M., & Usyukov, V. (2014). Modeling Cyclists' Route Choice Based on GPS Data. Transportation Research Record: Journal of the Transportation Research Board, (2430), pp 155–161. https://doi.org/10.3141/2430-16

Charlton, Billy, et al. "CycleTracks: a bicycle route choice data collection application for GPSenabled smartphones." 3rd Conference on Innovations in Travel Modeling, a Transportation Research Board Conference, Tempe, AZ. 2010.

Crash Modification Factor Clearinghouse. (n.d.). CMF Clearinghouse >> About CMFs. Retrieved December 15, 2016, from http://www.cmfclearinghouse.org/about.cfm Department of Transportation, D. of F. (2012). Project Traffic Forecasting Handbook.

Department of Transportation, 2012

Dixon, K., Zheng, J., 2013. Development Safety Performance Measures for Roundabout Applications in the State of Oregon. SPR733 Final Report. Oregon Department of Transportation

Dolatsara, H. A. (2014). Development of Safety Performance Functions for Non-Motorized Traffic Safety, 1–91.

Dorsch, Margaret M., Alistair J. Woodward, and Ronald L. Somers. "Do bicycle safety helmets reduce severity of head injury in real crashes?." Accident Analysis & Prevention 19.3 (1987): 183-190.

Ekman, L. (1996). On the treatment of flow in traffic safety analysis: a non-parametric approach applied on vulnerable road users. Bulletin, (136), 99 p. Retrieved from http://trid.trb.org/view/687009

Estellés-Arolas, Enrique, and Fernando González-Ladrón-De-Guevara. "Towards an integrated crowdsourcing definition." Journal of Information science 38.2 (2012): 189-200.

Fambro, D., Fitzpatrick, K., & Koppa, R. (1997). Determination of Stopping Sight Distances. NCHRP Report, (400). Retrieved from http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_400.pdf

Federal Highway Administration. (2013). Safety Performance Function Development Guide : Developing Jurisdiction-Specific SPFs, (September), 47.

Ferrara C, T. (2001). Statewide Safety Study of Bicycles and Pedestrians on Freeways Expressways, Toll Bridges, and Tunnels. MTI Report ; 01-01, ii, 156 p. Retrieved from http://transweb.sjsu.edu/mtiportal/research/publications/documents/BikesAndPeds.htm%5Cnhttp://ntl.bts.gov/lib/11000/11800/11851/BikesAndPeds2.pdf%5Cnhttp://ntl.bts.gov/l ib/18000/18600/18660/PB2002101184.pdf%5Cnhttp://trid.trb.org/view/714012

Foxx, A. (2014) “Bicycling, walking should be as safe as any other transportation” Fast Lane: The Official Blog of the U.S. Department of Transportation. Accessed 10/24/2014 http://www.dot.gov/fastlane/bicycling-and-walking-should-be-safe

Gardner, William, Edward P. Mulvey, and Esther C. Shaw. "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models." Psychological bulletin 118.3 (1995): 392.

Gourieroux, A. C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: applications to poisson models. Econometrica, 52(3), 701–720. https://doi.org/10.2307/1913472

Griswold, J, A. Medury, and R. Schneider. Pilot Models for Estimating Bicycle Intersection Volumes. In Transportation Research Record: Journal of the Transportation Research Board No. 2247, TRB of the National Academies, Washington, D.C., 2011, pp. 1–7.

Hauer, E. (2004). Statistical Road Safety Modeling. Transportation Research Record, 1897(1), 81–87. https://doi.org/10.3141/1897-11

Hauer, E., 2014. The art of regression modeling in road safety. SPF workshop lecture notes.

Hilbe, Joseph M. Negative binomial regression. Cambridge University Press, 2011.

Hood, J., Sall, E., & Charlton, B. (2011). A GPS-based bicycle route choice model for San Francisco, California. Transportation Letters: The International Journal of Transportation Research, 3(1), 63–75. https://doi.org/10.3328/TL.2011.03.01.63-75

Howe, Jeff. "The rise of crowdsourcing." Wired magazine 14.6 (2006): 1-4.

Hunt, K. (2015). STRAVA training app not the best way to gather cycling data. Retrieved from http://www.metronews.ca/views/ottawa/your-ride/2015/11/23/STRAVA-training-appnot-the-best-way-to-gather-cycling-data.html

Hunter, William W., et al. Pedestrian and bicycle crash types of the early 1990's. No. FHWARD-95-163. 1996.

Jacobsen, P. (2003). Safety in numbers: more walkers and bicyclists, safer walking and bicycling. Injury Prevention, 9(3), 205–209.

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. Journal of Transport Geography, 52, 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006

Jin, Peter J., Dan Fagnant, Andrea Hall, C. M. Walton, Jon Hockenyos, and Mike Krusee. Developing Emerging Transportation Technologies in Texas. No. FHWA/TX-13/0-68031. 2013.

Jones, M., S. Ryan, J. Donlon, L. Ledbetter, D. Ragland, and L. Arnold. Seamless Travel: Measuring Bicycle and Pedestrian Activity in San Diego County and Its Relationship to Land Use, Transportation, Safety, and Facility Type, PATH Report UCB-ITS-PRR-201012, 2010.

Jonsson, T. (2005). Predictive models for accidents on urban links - A focus on vulnerable road users. Lund Institute of Technology, Department of Technology and Society ., 226. Retrieved from https://lup.lub.lu.se/search/publication/24269

Juhra, C., et al. "Bicycle accidents–Do we only see the tip of the iceberg?: A prospective multicentre study in a large German city combining medical and police data." Injury 43.12 (2012): 2026-2034.

Kim, Karl, I. Brunner, and Eric Yamashita. "Influence of land use, population, employment, and economic activity on accidents." Transportation Research Record: Journal of the Transportation Research Board 1953 (2006): 56-64.

Klop, Jeremy, and Asad Khattak. "Factors influencing bicycle crash severity on two-lane, undivided roadways in North Carolina." Transportation Research Record: Journal of the Transportation Research Board 1674 (1999): 78-85.

Kononov, J., & Allery, B. (2003). Level of Service of Safety: Conceptual Blueprint and Analytical Framework. Transportation Research Record, 1840(1), 57–66.

Krykewycz, Gregory R., Christopher Pollard, Nicholas Canzoneri, and Elizabeth He. "Web-Based" Crowdsourcing" Approach to Improve Areawide" Bikeability" Scoring." Transportation Research Record: Journal of the Transportation Research Board 2245, no. 1 (2011): 1-7.

Ladron de Guevara, Felipe, Simon Washington, and Jutaek Oh. "Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson,

Arizona." Transportation Research Record: Journal of the Transportation Research Board 1897 (2004): 191-199.
Lambert, Diane. "Zero-inflated Poisson regression, with an application to defects in manufacturing." Technometrics 34.1 (1992): 1-14.
Landis, Bruce W. "Bicycle system performance measures." ITE journal 66.2 (1996): 18-26.
Le Dantec, C. A., Asad, M., Misra, A., & Watkins, K. E. (2015). Planning with Crowdsourced Data. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15, 1717–1727. https://doi.org/10.1145/2675133.2675212
Lindman, M., et al. "Cyclists interacting with passenger cars; a study of real world crashes." IRCOBI Conference Proceedings. No. IRC-15-10. 2015.
Liu, F., J. Evans, and T. Rossi. Recent Practices in Regional Modeling of Nonmotorized Travel. In Transportation Research Record: Journal of the Transportation Research Board No. 2303, TRB of the National Academies, Washington, D.C., 2012, pp. 1–8.
Long, J. S. (1997). Regression models for categorical and limited dependent variables. American Journal of Sociology. https://doi.org/10.1086/231290
Lowry, M., Furth, P., and Hadden-Loh, T. (2016). "Prioritizing new bicycle facilities to improve low-stress network connectivity" Transportation Research Part A: Policy and Practice, Vol. 86, pp. 124-140.
Lowry, M., McGrath, R., Scruggs, P. and Paul, D. (2016). "Practitioner survey and measurement error in manual bicycle and pedestrian count programs" International Journal of Sustainable Transportation.
Lusk, Anne C., Morteza Asgarzadeh, and Maryam S. Farvid. "Database improvements for motor vehicle/bicycle crash analysis." Injury prevention (2015): injuryprev-2014.
Maneewongvatana and Mount. (1999) "On the Efficiency of Nearest Neighbor Searching with Data Clustered in Lower Dimensions" Proceeding ICCS '01 Proceedings of the International Conference on Computational Sciences-Part I. Pages 842-851
McDaniel, S., Lowry, M., and Dixon, M. (2014). "Using Origin-Destination Centrality to Estimate Directional Bicycle Volumes." Transportation Research Record: Journal of the Transportation Research Board.
Milne, A., and M. Melin. (2014) Bicycling and Walking in the United States: 2014 Benchmarking Report. Aliance for Biking and Walking, 2014.
Molina, Jennifer. The Case for Crowdsourcing in Bicycle Planning: An Exploratory Study. Diss. Tufts University, 2014.
Monsere, C., Wang, H., Wang, Y., & Chen, C. (2016). RISK FACTORS FOR PEDESTRIAN AND BICYCLE CRASHES.
Nordback, K., Marshall, W. E., & Janson, B. N. (2014). Bicyclist safety performance functions for a U.S. city. Accident Analysis and Prevention, 65, 114–122. https://doi.org/10.1016/j.aap.2013.12.016
ODOT, O. D. of T. (2016). ODOT GIS File TransDATA. Retrieved December 18, 2016, from ftp://ftp.odot.state.or.us/tdb/trandata/GIS_data/

Porter, C, J. Suhrbier, and W.L. Schwartz. Forecasting Bicycle and Pedestrian Travel: State of the Practice and Research Needs. In Transportation Research Record: Journal of the Transportation Research Board, 1674, TRB of the National Academies, Washington, D.C., 1999, pp. 94–101.

Portland Bureau of Transportation. (2016). Traffic Counts | Services | The City of Portland, Oregon. Retrieved December 18, 2016, from https://www.portlandoregon.gov/transportation/article/180473

PSRC (2014) CycleTracks: Understanding Bicyclist Needs, Puget Sound Regional Council, Available online at http://www.psrc.org/transportation/bikeped/cycletrack/

Pucher, J., and Dijkstra, L. (2003). Promoting safe walking and cycling to improve public health: lessons from The Netherlands and Germany. American Journal of Public Health, 93(9), 1509–16.

Pucher, John, Jennifer Dill, and Susan Handy. "Infrastructure, programs, and policies to increase bicycling: an international review." Preventive medicine 50 (2010): S106-S125.

Quartuccio, Jacob, et al. "Seeing is Believing The Use of Data Visualization to Identify Trends for Cycling Safety." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 58. No. 1. SAGE Publications, 2014.

Reynolds, C. C. O., Harris, M. A., Teschke, K., Cripton, P. A., & Winters, M. (2009). The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature. Environmental Health Perspectives, 8(1), 47. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2776010{&}tool=pmcentrez{&}rendertype=abstract$\$nhttp://www.ehjournal.net/content/8/1/47

Roberts, R. J. "Can self-reported data accurately describe the prevalence of overweight?." Public health 109.4 (1995): 275-284.

Robinson, D. L. (2005). Safety in numbers in Australia: more walkers and bicyclists, safer walking and bicycling. Health Promotion Journal of Australia : Official Journal of Australian Association of Health Promotion Professionals, 16(1), 47–51.

Rose, Geoffrey, et al. "Quantifying and comparing effects of weather on bicycle demand in Melbourne, Australia, and Portland, Oregon." Transportation Research Board 90th Annual Meeting. No. 11-3205. 2011.

Ryus, P., Ferguson, E., Laustsen M, K., Schneider J, R., Proulx R, F., Hull, T., & MirandaMoreno, L. (2014). Guidebook on Pedestrian and Bicycle Volume Data Collection. NCHRP Report. Retrieved from http://www.trb.org/Publications/Blurbs/171973.aspx%5Cnhttps://trid.trb.org/view/1342012

Schepers, J. P., et al. "Road factors and bicycle–motor vehicle crashes at unsignalized priority intersections." Accident Analysis & Prevention 43.3 (2011): 853-861.

Schepers, Paul, and K. Klein Wolt. "Single-bicycle crash types and characteristics." Cycling research international 2.1 (2012): 119-135.

Scott, G. (2015). STRAVA 2014: The year in numbers. Retrieved from http://roadcyclinguk.com/sportive/STRAVA-2014-year-

numbers.html#OLMfeEHYSHlruhLO.97%5Cnhttp://roadcyclinguk.com/sportive/STRAVA-2014-year-numbers.html
SDOT (2015). 2015 Seattle Traffic Report. Version Zero. Seattle Department of Transportation, Seattle, WA.
Selala, M. K., & Musakwa, W. (2016). The potential of STRAVA data to contribute in nonmotorised transport (NMT) planning in Johannesburg. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, 41(July), 587–594. https://doi.org/10.5194/isprsarchives-XLI-B2-587-2016
Spencer, Phoebe, et al. "The effect of environmental factors on bicycle commuters in Vermont: influences of a northern climate." Journal of Transport Geography 31 (2013): 11-17.
Srinivasan, R., Carter, D., & Bauer, K. (2013). Safety Performance Function Decision Guide: SPF Calibration versus SPF Development, (September), 1–31.
STRAVA. (2016a). Data-Driven Bicycle and Pedestrian Planning. Retrieved from http://metro.STRAVA.com/
STRAVA. (2016b). STRAVA Global Heatmap. Retrieved December 18, 2016, from http://labs.STRAVA.com/heatmap/#12/-122.67626/45.54243/yellow/bike
Stutts, Jane C., et al. "Bicycle accidents and injuries: a pilot study comparing hospital-and police-reported data." Accident Analysis & Prevention 22.1 (1990): 67-78.
Thomas, Tom, Rinus Jaarsma, and Bas Tutert. "Temporal variations of bicycle demand in the Netherlands: The influence of weather on cycling." Transportation Research Board 88th Annual Meeting. No. 09-1545. 2009.
Thompson, Diane C., et al. "A case-control study of the effectiveness of bicycle safety helmets in preventing facial injury." American Journal of Public Health 80.12 (1990): 1471-1474.
Turner, S., Roozenburg, A. P., Francis, T., 2006. Predicting Accidents Rates for Cyclists and Pedestrians. Land Transport New Zealand, Wellington, New Zealand.
Turner, S., Wood, G., Hughes, T., Singh, R., 2011b. Safety performance functions for bicycle crashes in New Zealand and Australia. Transportation Research Record: Journal of the Transportation Research Board No. 2236, 66-73.
USDOT (2014) Safe People, Safer Streets: Summary of U.S. Department of Transportation Action Plan to Increase Walking and Biking and Reduce Pedestrian and Bicyclist Fatalities, USDOT September 2014. Available online: http://www.dot.gov/sites/dot.gov/files/docs/safer_people_safer_streets_summary_doc_acc_v1-11-9.pdf
Vogt, A., & Bared, J. (1998). Accident Models for Two-Lane Rural Segments and Intersections. Transportation Research Record: Journal of the Transportation Research Board, 1635(-1), 18–29. https://doi.org/10.3141/1635-03
Vuong, Quang H. "Likelihood ratio tests for model selection and non-nested hypotheses." Econometrica: Journal of the Econometric Society (1989): 307-333.
Wang, Xuesong, and Mohamed Abdel-Aty. "Temporal and spatial analyses of rear-end crashes at signalized intersections." Accident Analysis & Prevention 38.6 (2006): 1137-1150.

Watkins, K., Ammanamanchi, R., LaMondia, J., & LeDantec, C. A. (2016). Comparison of Smartphone-based Cyclist GPS Data Sources. Transportation Research Record.
Watkins, Kari, et al. "Comparison of Smartphone-based Cyclist GPS Data Sources." Transportation Research Board 95th Annual Meeting. No. 16-5309. 2016.
Yan, Xinping, et al. "Motor vehicle–bicycle crashes in Beijing: Irregular maneuvers, crash patterns, and injury severity." Accident Analysis & Prevention 43.5 (2011): 1751-1758.
Zou, Guangyong. "A modified poisson regression approach to prospective studies with binary data." American journal of epidemiology 159.7 (2004): 702-706.