AN ABSTRACT OF THE THESIS OF

Chen Chen for the degree of Master of Science in Civil Engineering presented on May 31, 2017.

Title: Crowdsourcing Data-driven Development of Bicycle Safety Performance Functions (SPFs): Microscopic and Macroscopic Scales.

Abstract approved:

_____

Haizhong Wang

While riding bicycles has been promoted for its health, economic, and environmental benefits, it also complements other modes to complete a safe, efficient, and reliable transportation system. However, the dramatic increase of bicycle usage in the U.S. is accompanied by a growth of bicycle crashes. The U.S. Department of Transportation, therefore, is focusing on providing safer riding environments. Providing a more bicycle friendly environment means more investment in (but not limited to) bicycle infrastructure. A correct prediction of bicycle crashes can increase the return on this investment. One useful tool to understand the causality and predict crashes is Safety Performance Functions (SPFs), but no sophisticated SPFs have been established for bicycles. Therefore, the objective of this thesis is to establish SPFs for microscopic (intersection) and macroscopic (corridor) scales in medium and large size cities using crowdsourced bicycle data, with a case study in the Portland and Eugene-Springfield metropolitan, which overcomes the challenge of insufficient bicycle volume data and crash data. Specifically, in this research 1) bicycle SPFs are created for intersections and corridors that have not been sufficiently studied; 2) bicycle crash severity distributions are used the first time to predict the number of bicycle crashes with

different crash severity levels; 3) affordable crowdsourced bicycle volume data – STRAVA® is chosen to solve the problem of limited data; 4) STRAVA® data was verified to be able to represent general bicyclists by comparison with automatic bike count station data; 5) a general framework for building SPFs was developed for jurisdictions.

Crowdsourcing Data-driven Development of Bicycle Safety Performance Functions (SPFs): Microscopic and Macroscopic Scales

by
Chen Chen

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented May 31, 2017
Commencement June 2017

Master of Science thesis of Chen Chen presented on May 31, 2017

APPROVED:

Major Professor, representing Civil Engineering

Head of the School of Civil and Construction Engineering

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Chen Chen, Author

ACKNOWLEDGEMENTS

I express sincere appreciation to people and organizations contributing help in this study. The instruction from Dr. Haizhong Wang, my academic advisor, made the work possible. Dr. Haizhong Wang provides me the opportunity to pursue my master degree and has been advising me for the last two years. Without him, I could not go this far in my academic journey.

I am also grateful to thesis committee members: Dr. Katharine Hunter-Zaworski, Dr. David Hurwitz, Dr.Yue Zhang. PacTrans and ODOT provided funding for my master academic years. Part of the data sets and ideas are from Pacific Northwest Transportation Consortium (PacTrans) project and Oregon Department of Transportation (ODOT) project SPR 779. ODOT/TPAU also provided STRAVA® data and other road characteristic data. Many of people have also contributed to this study. Josh Roll and Alexander Bettinardi from ODOT gave advice on data choice and paper objectives. Jason Anderson and Dr. Salvador Hernandez provided important suggestions on modeling process. Jenessa Duncombe, Shangjia Dong, Jason Anderson offered help on my writing.

I also appreciate all the help and supports from my friends and family. The greatest appreciation I want to give to my fiancée, Zi Li, who is seeking a master degree in the U.S. too. During the three years I am in OSU, no matter what challenges I suffered from, whenever I was disappointed, she was always the first person who give me a hand.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continues)

LIST OF TABLES

1. Introduction

As riding bicycles has been promoted for its health, economic and environmental benefits (Simmons et al., 2015), more people are selecting it for commuting or recreation. Data show that the number of bicycle trips increased from 1.7 billion in 2001 to 4 billion in 2009 in the U.S., according to National Household Travel Survey (League of Amercian Bicyclists, 2015). Figure 1-1 shows the increase of bicycle

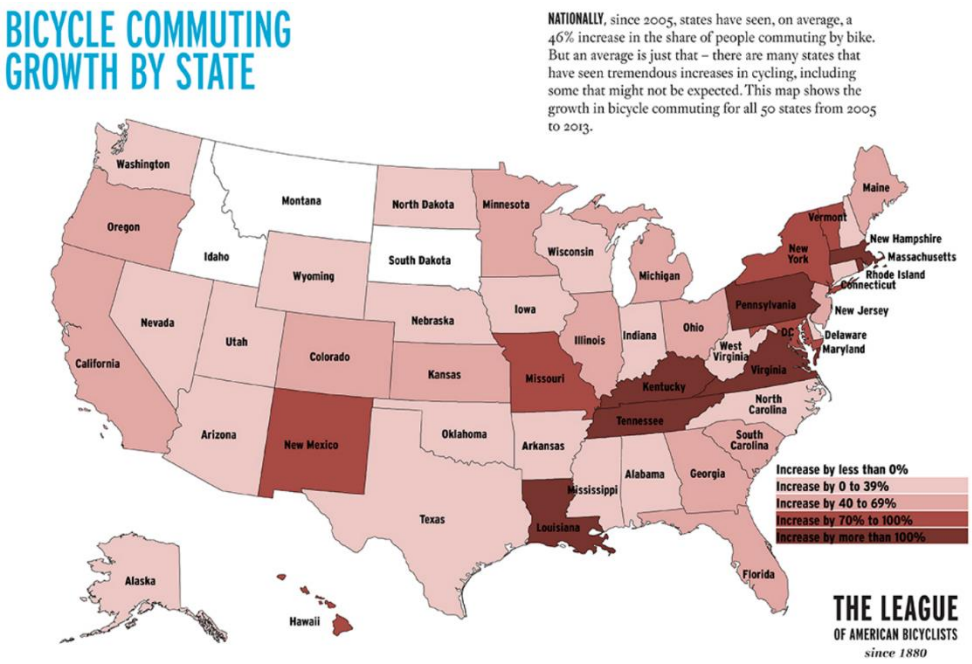

Figure 1-1 The growth of bicycle commuting by state from 2005 to 2013 (League of Amercian Bicyclists, 2015)

commuting in all states in the U.S. from 2005 to 2013. Portland, Oregon, with 408% percentage growth rate, has the fastest rate of increase of bicycle commuter share in the U.S. (League of Amercian Bicyclists, 2015). While the growth rate of bicycle usage is different from state to state, the general trend shows bicycles have become an important mode choice.

Unfortunately, the dramatic increase of bicycle usage is accompanied with a growth of bicycle crashes (National Highway Traffic Safety Adminstriation, 2014; Wang et al., 2016). There were 726 people who lost their lives in bicycle crashes in 2014 and the percentage of total fatalities has increased from 2005 to 2014 in the U.S. (National

Highway Traffic Safety Adminstriation, 2016). In the past decades, engineers and researchers have focused more on designing and deploying countermeasures to mitigate motor-vehicle crashes, but less attention has been paid to bicycle safety until recently. Figure 1-2 demonstrates that the fatality rate for bicyclist per year in the U.S. has been increasing steadily since 2005. The increase in the percentage of bicycle fatalities results from two reasons: the increase bicycle usage and slight decrease of traffic fatalities of other modes (National Center for Statistics and Analysis, 2017). Even though bicycle trips only account for one percent of total trips, more than two percent of total road fatalities are from bicycle trips (Nordback et al., 2014). This trend drives an urgent need for engineers and planners to increase bicycle safety.

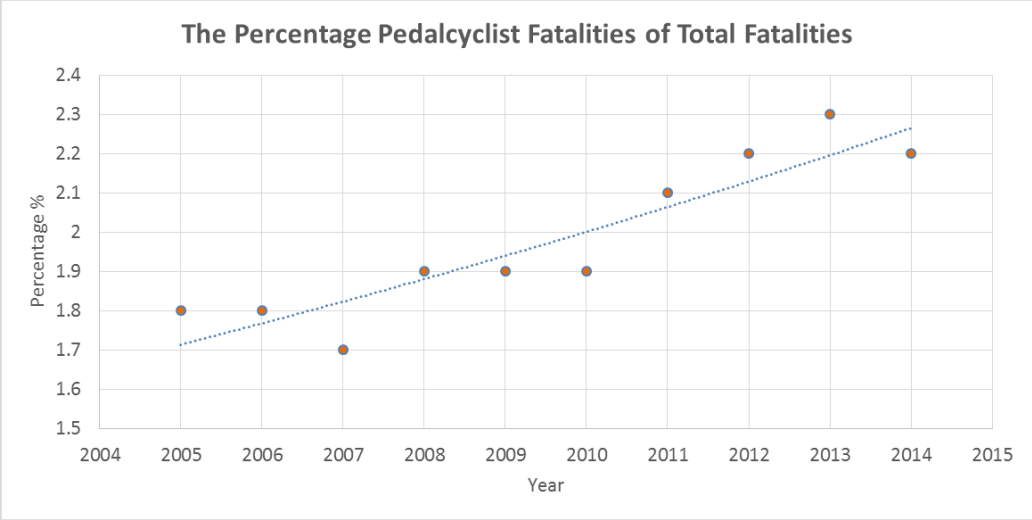

Figure 1-2 The percentage bicycle fatalities out of the total traffic fatalities (data source: (National Highway Traffic Safety Adminstriation, 2016))

The goal of this thesis is to create crowdsourcing data-driven bicycle safety solutions for engineers and planners to improve bicycle safety specifically building a predictive bicycle crash tool – Safety Performance Function (SPF). This chapter provides the basic information of this work including problem definition, objectives, and thesis organization.

1.1 Problem Definition

Bicycling is not only a healthy and environmental friendly transportation mode, it also complements other modes to complete a safe, efficient, and reliable transportation system (U.S. Department of Transportation, 2017). As previously mentioned, more bicycle crashes appear with increasing bicycle usage. The U.S. Department of Transportation therefore is focusing on providing safer riding environment (U.S. Department of Transportation, 2017). Providing a more bicycle friendly environment means more investment in bicycle infrastructure. Thus, improved understanding of the relationship between the number of crashes, crash severity, exposure to collision and other factors contributing to accidents can lead to more effective mitigation strategies and prioritize efficient investment to improve bicycle safety.

This relationship is also known as Safety Performance Functions (SPFs). SPFs are statistical regression models that can predict the crash frequency for one or more specific sites, such as intersections or two-way urban streets. SPFs describe the number of crash of various types of sites with different features. SPFs always include traffic volume AADT, but also may include other site features, such as lane width, horizontal curve, presence of turn lane, etc. Those models can be used in network safety screening, determining the safety impact of design changes, evaluating the effect of engineering treatments and so on (Federal Highway Administration, 2013; Wang et al., 2017). The misunderstanding of the relationship between crash number and bicycle volume can cause engineers to simply calculate the accidents per vehicle by using number of crash divided by the number of bicycles or vehicles (Hauer, 1995; Nordback et al., 2014).

Only motor-vehicle SPFs were established in the first edition of Highway Safety Manual (HSM). This method provides an evidence-based tool to estimate motor vehicle crashes by traffic volume and other factors that could influence the results (American Association of State Highway and Transportation Officials, 2010; Nordback et al., 2014). However, there are few studies that address the SPFs for estimating bicycle crashes (Wang et al., 2017).

There are three interrelated challenges of establishing SPFs for bicycle transportation: 1) there is a general lack of bicycle crash data (due to not enough crashes, missing data and the issue of under-reporting) and the sporadic nature of bicycle crashes; 2) there are no accurate bicycle volume data; 3) it is difficult to decide what statistical models should be used to establish SPFs due to insufficient data availability.

1.2 Objectives

To address the challenges of building SPFs for bicycle, this research uses crowdsourced data (i.e., STRAVA®) to establish bicycle SPFs at urban intersections and segments. Compared to previous studies, the unique contribution of this research includes:

1. Establishing SPFs for intersections (micro scale) and corridors (macro scale) that have not been sufficiently studied;
2. Bicycle crash severity distributions are used the first time to predict bicycle crash severity level in combination with crash frequency;
3. Using affordable data resources – crowdsourced bicycle volume data – to solve the issue of insufficient data;
4. Embedding factors that could influence bicycle crash prediction model rather than building separate Crash Modification factors (CMFs);
5. Establishing the model selection process when building SPFs for jurisdictions;
6. Investigating the representativeness of crowdsourced data.

Detailed explanations for the six points are presented as follow. Previous studies (Turner et al., 2011; Dolatsara, 2014; Nordback et al., 2014) built bicycle SPFs for intersections but not for segments. The main reason is that there is insufficient crash data for segments. To solve this issue, this present research establishes SPFs on a long corridor which contains multiple segments and intersections, termed “Macroscopic SPFs” here because the SPFs are built to predict crashes for the whole corridor from a larger scale rather than only focusing on one intersection. Transportation agencies can combine the micro and macro PSFs to predict bicycle crash frequency.

This paper uses STRAVA® bicycle count data as the bicycle exposure to build SPFs. STRAVA® count data, as a type of crowdsourced data, has advantages over traditional count data. For example, it is easier to access and affordable for small jurisdictions. The idea of using crowdsourced data to build SPFs comes from the PacTrans project (Wang et al., 2017) that the author involved. Further, the data is available for each city's transportation network which means researchers can build SPFs on all roads with different functional classifications. The paper also investigates how STRAVA® represents the whole population of the studied area (Metropolitan Portland and Eugene-Springfield).

Another highlight of this paper is that the author documents the procedure of establishing SPFs on crowdsourced data with a detail explanation of how to choose a correct model. Other jurisdictions could follow the process to build their own SPFs to obtain more accurate prediction results.

The present study builds bicycle SPFs on both intersection (microscopic scale) and corridor (macroscopic scale) based on data collected from Portland, OR and Eugene-Springfield Metropolitan area. Figure 1-3 demonstrates the difference between the ideas of macroscopic and microscopic SPFs. The idea of macro level SPFs is inspired by Wei and Lovegrove (2013) who built predictive models of collisions in Canada. This idea also matches the safety evaluation scope of DOTs in the U.S. For example, Kittelson & Associates, Inc. cooperated with Oregon Department of Transportation to create a pedestrian and bicycle safety implementation plan on macro (corridor) level in 2014 (Kittelson&Associates Inc and ODOT, 2014). Thus, the author decided to adopt this idea and create macro level SPFs to meet both research and practical needs.

The data include: STRAVA® bicycle count data, traffic volume, road characteristics, geometric data, land-use data, crash data, etc. The results of this study can essentially provide a bicycle crash prediction tool to evaluate safety, determine the impact of changing design, and screen transportation networks to identify the most efficient investment regarding locations and means. Additionally, the repeatable procedure of building of bicycle SPFs on crowdsourced data is established.

Figure 1-3 The difference between microscopic SPFs and macroscopic SPFs.

It should be noted that parts of this study (including a small proportion of data and documentation part) come from two projects and the author of the present paper is one of the main author of both projects: 1) PacTrans project: Bicycle Safety Analysis – Crowdsourcing Bicycle Travel Data to Estimate Risk Exposure and Create Safety Performance Functions in 2017 (unpublished), and 2) ODOT SPR 779 project: Risk Factors for Pedestrian and Bicycle Crashes in 2017 (unpublished).

1.3 Paper Organization

This paper presents the literature review, methodology, results, conclusion and recommendation. Chapter 2 reviews literature and provides a comprehensive recording and comparison of existing studies relating to present research; Chapter 3 illustrates the methods that the author used to build the SPFs; Chapter 4 documents the data preparing process and analyses of the data; Chapter 5 discusses the details of results from bicycle SPFs in this research; Chapter 6 provides recommendations and conclusions for engineers and decision makers.

2. Literature Review

This chapter provides a comprehensive record of existing SPFs studies, discusses a comparison and contrast of predictive models, a documentation of bicycle crash analyses and also a record of how others studies chose and collected data. Thus, the main purpose of this section is to review the suggestions and methods of building SPFs from other studies.

2.1 Existing SPFs Studies and SPFs History

SPFs are mathematical equations predicting the number of the crash at various sites. Traffic volume is always included in SPFs, yet they may also include other features, for example, the width of lanes, the number of lanes, intersection control, etc. Those equations can be applied to assess the effect of treatments, screen network safety, determine the safety impact of changing designs and so on (Federal Highway Administration, 2013). This predictive approach can provide crash estimates on sites that have not been constructed or have been constructed too recent to have crash data (American Association of State Highway and Transportation Officials, 2010).

Other methods besides SPFs can be used to predict crashes. Tegge, Jo and Ouyang (2010) mentioned the method of Controlled Studies or Experiments can support accurate analysis to understand causal relation between variables and the number of crashes; however, due to a significant amount of variation involved, those methods are very difficult to apply. Specifically, this causality involves a combination of various factors, such as weather, road condition, drivers' behaviors, etc. Therefore, the alternative method of Observational Studies has been the most practical method to explore the casual relationship rather than Controlled Studies. Statistical tools are used to analyze crash data in Observational Studies to discover the correlation between variables. With data from crash report, traffic volume, and geometric information, SPFs can provide a statistical relationship between expected yearly crash count and roadway features (Tegge et al., 2010).

When Highway Safety Manual (HSM) in 2010 established sophisticated SPFs but only for vehicle, Nordback et al. 2014 applied this idea and created a basic SPFs for bicycle to Boulder city in Colorado. The authors used collision, AADT, AADB data

to build the function describing the relationship between traffic and bicycle volume with crash frequency in intersections. They found that with the bicycle and motor vehicle volume increase, the frequency of cyclist crash 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 volume. This relationship previously studied by others has been found not linear and is called “safety in number” (Ekman, 1996; Jacobsen, 2003; Jonsson, 2005; Robinson, 2005; Nordback et al., 2014).

Nordback et al. in 2014 established the process and method of creating SPFs for bicyclist by a negative binomial generalized linear model with a log link, and this model was based on data of Annual Average Daily Traffic (AADT) and Annual Average Daily bicycle (AADB). The authors compare negative binomial regression and Poisson regression, and they found that the former can fit the data better because of the collision datasets has the feature of variance triple the mean, in other words, the crash data is over-dispersed. In Poisson distribution, the mean equal to the variance, but when variance is larger than mean, the situation is called over-dispersion (Federal Highway Administration, 2013). Speaking about dataset, three peak hours count for both bicycle and traffic, provided by the city of Boulder (Boulder and 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).

The authors did a sensitivity analysis on change of AADB. The results show: with higher AADB, the corresponding parameter is still well under one which indicates the SPF is still sub-linear; whereas the parameter for lower AADB is closed to zero, which indicates the AADB is not a major factor in determining motorist-cyclist crash. Further analysis can investigate this observation. In addition, the estimations of parameters of AADT and AADB are at the same magnitude, indicating the collision is similar sensitive to both volumes; however, the AADT exponent is one or two orders of magnitude higher than AADB exponent, so the change of AADB has more critical influence on crash than same change of AADT. Thus getting accurate estimation of bicyclist volume is more important to analyze the SPF (Nordback et al.,

2014). Future work can use larger dataset and more accurate AADB and include facility type into the analysis.

This analysis only captures the connection between volume and crash but cannot reveal the causation between them. In other words, the reason between traffic & bicycle volume and crash frequency is not explained. The reasons can be: the increasing bicycle volume may lead safer behavior of motorists and bicyclists, or more bicyclist riding on safer facilities. Other studies state more bicyclists triggers changing the behavior of drivers, but it is based on logical speculation not empirical data analysis (Ekman, 1996). Another potential improvement is the accuracy of AADB. While there is not actual AADB available, the estimated from two-hour data can be not accurate.

Furthermore, Nordback et al. 2014 did not consider other factors such as geometric data, road characteristics, land-use, etc. Therefore, Dolatsara's (2014) study built SPFs for the intersections on crash data, the volume data and the road geometric data, as an improvement from previous work. Department of Transportation provided traffic and bicycle volume; the geometric data include different lane number, bike lane characters, post speed, bus stop and so on; the crash data collected in this study came from 164 intersections of four cities in Michigan and crashes happened within 500 feet buffer of the center if intersection were collected (Dolatsara, 2014). Intersection traffic volume was collected by combining four directions of ADT in this study. The 500 feet was calculated by Stopping Sight Distance (SSD) by (Fambro et al., 1997). However, more practically, 250 feet has been used as a diameter to assign crashes into an intersection (Vogt and Bared, 1998; Dolatsara, 2014).

The author also mentioned that the Poisson distribution could not capture the over-dispersion of crash data (American Association of State Highway and Transportation Officials, 2010), so the negative binomial regression was employed (Dolatsara, 2014). The significant variables are included in the SPF are ADT, the number of the left turn lane, presence of bike lane, presence of bus stop (Dolatsara, 2014). This would suggest that this present study may include other factors besides traffic and bicycle volume into our SPF. Dolatsara (2014) conclude that 1) the higher exposure of bike

volume; 2) presence of bike lane; 3) presence of bus stop within 0.1 miles within an intersection; 4) increase number of left turn lane are associated with more bicycle crash. However, that does not mean the bike cause more crash because there are more bicycles attaching bike lane than without bike lane (Dolatsara, 2014). This finding is consistent with Nordback et al. (2014) paper. Others have also studied bicycle facility related studies. Reynolds et al. (2009) concluded that there was evidence support that the purpose-built bicycle-specific facilities can reduce bicycle collision. Street lighting, paved surfaces and low-angled grade are also the factors that can improve bicycle safety. This study also used manual bicycle count exposure data which may have same estimated error as Nordback et al.'s (2014) work.

Turner et al., (2011) used generalized linear model and before-after-control impact method to study the safety performance of intersections in New Zealand and Australia. Similar to the study from Dolatsara, (2014), the Turner et al. also considered geometric information in the model. However, the study considered fewer variables and used a simple model that may not completely capture the impact of factors. Again, they also used manual count data for bicycle volume which can raise the issue of misrepresenting the real bicycle volume. The manual count included turning movement which can be used to evaluate bicycle safety on each movement; however, this manual count requires much more effort input than automatic count loop or crowdsourced data.

In 2002, Midwest Research Institute (MRI) developed a software tool to Federal Highway Administration (FHWA), called *Safety Analyst*, to analyze road safety. This tool includes four module: Network Screening, Diagnosis and countermeasure selection, Economic appraisal and priority-raking, and Evaluation. The first module used SPFs to estimate expected accident frequencies and developed the procedure of building SPFs for segment ramps and intersection. In addition, it combined SPFs and Empirical Bayes (EB) method together to predict crash rate (Midwest Research Institute et al., 2002). Later on, the combination of SPFs and EB method was included in HSM (American Association of State Highway and Transportation Officials, 2010).

Table 2-1 SPFs developing history (Jo et al., 2009; Tegge et al., 2010)

Study	Year	Site
Zegeer et al.	1987	segment
Persaud	1992	segment
Forkenbrock et al.	1994	segment
Tarko et al.	1999	segment
Harwood et al.	2000	segment
Minnesota [Vogt and Bared]	1998	intersection
Washington [Vogt and Bared]	1998	intersection
IHSDM Model [Bauer and Harwood]	1999	intersection
California, Michigan [Vogt]	1999	intersection
Iowa [Harwood et al.]	2002	intersection
Illinois [Harwood et al.]	2002	intersection
Louisana, Nebrask, Virginia [Harwood et al.]	2002	intersection
North Carolina [Harwood et al.]	2002	intersection
Oregon [Harwood et al.]	2002	intersection
Minnesota [Harwood et al.]	2002	intersection
HSM	2010	Segment and intersection

2.2 Crash Modification Factor

Basic SPFs capture the relationship between traffic volume and frequency of different types of crash, but cannot predict crash frequency accurately for a specific condition. A CMF is a multiplicative factor that used to adjust the predicted mean crash frequency estimated by SPFs for a specific condition. Typically, each road feature has one CMF, such as CMF of the presence of bicycle buffer line. In addition, it can also be used to compute the expected number of the crashes after installing a treatment at a specific site. In other words, it also can be used to multiply the estimated crash frequency without treatment by basic SPFs (American Association of State Highway and Transportation Officials, 2010; Gross et al., 2010).

CMFs are multiplicative based on one critical assumption that the feature which a CMF represent is independent of other features. However, there is little research existing justify the independence between these effects. The limited understanding the interrelationship between road characteristics or countermeasures installed on the street need to be addressed carefully when applying CMFs. It is possible to

overestimate the effect of a combination of countermeasures existing the same place on the street. For example, shown in Figure 2-1, increasing shoulder width, add rumble strips, and increasing reflection of marking can all reduce single vehicle crashes (hitting on roadside objects). The countermeasures' effects are over-estimated when the three CMFs are multiplied by and estimate crash frequency at the same time due to the overlaps. Engineers should use engineering judgment to assess the interrelation or interdependence between elements or treatments (American Association of State Highway and Transportation Officials, 2010). The lack of knowledge of understanding the interrelationship between factors paly the main reason that author of this present paper decided not using CMF to build bicycle SPFs.

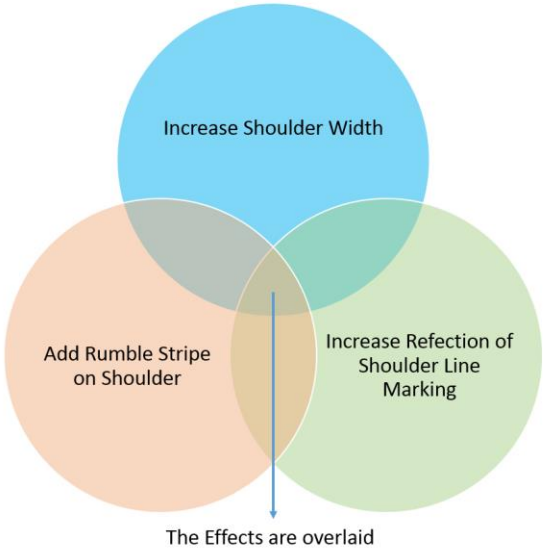

Figure 2-1 The overlaps of CMFs

There are various approaches to develop CMFs, for example, Before-After with Comparison Group method, Full Bayes, Cross-sectional study, Cohort method, etc. Using Empirical Bayes (EB) method can compensate the bias caused by interdependence between elements; however, this approach mainly estimates average crash frequency of past and feature either on a specific site or project level after calculating prediction from SPFs (American Association of State Highway and Transportation Officials, 2010). EB method provides a mean of a combination of prediction from SPFs and observed crashes data. Midwest Research Institute et al.

(2002) apply EB method after SPFs in a network screening tool called "*SafetyAnalyst*" to predict crash frequency.

2.3 Bicycle Crash Type, Frequency, and Severity

Bicycle crash is defined as an accident that bicyclist has a collision with the ground, objectives such as a vehicle, road facilities and others and causes damage property or human body (Lindman et al., 2015). Researchers have been studying crashes type, frequency and severity for a long time. The better understanding the bicycle crash types, the better engineers and decision makers can make enhancement of bicycle safety. A course on this topic was developed by Federal Highway Administration (FHWA) and described the crash rates, exposure, characteristics, and so forth. for engineers, planners, researchers to better clarify the reasons of crashed happening and how to avoid them (Hunter, 1996). Most bicycle crashes are involving single bicycle and are categorized into four types: infrastructure-related crashes, cyclist-related crashes, bicycle malfunction, other or unknown (Schepers et al., 2011). Table 2-2 summarizes the crash types as below.

Table 2-2 The Bicycle Crash Types (Wang et al., 2017)

Crash Type	Description
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
Wrong way riding	The bicyclist ride on the road or sidewalk against the flow of traffic
Bicyclist left turn in front of traffic	The motorist right turn
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
Motorist passes a bicyclist	A motorist fails to see and avoid it the bicyclist until it is too late to avoid a collision
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
Motorist enters road from a driveway or alley	The motorist fails to stop and look before entering a roadway
Multiple threads	The bicyclist fails to clear the intersection before the light turns red.

Besides those studies (Turner et al., 2011; Dolatsara, 2014; Nordback et al., 2014) mentioned in the previous section, there are also other scholars has been focused on studying crash frequency. Count models are usually used to investigate the relationship. For example, Oh et al. (2008) use the Poisson model to analyze the

bicycle crash at intersections in the urban area. More other studies can be found in Section 2.1.

According to Wang et al. (2017) and Reynolds et al. (2009), bicycle crash injury severity divided into four different levels: fatal injury, incapacitating injury, non-incapacitating injury and possible/no injury. Previous researchers analyze the factors affecting the level of crash injury severity from various perspectives. Thompson et al. developed a regression model to investigate the impact of helmets on reducing the injury severity levels of bicycle crashes. The results indicated that bicycle helmet has a protective effect for reducing crash injury severity (Thomas et al., 2009).

Yan et al., (2011) used a regression model to explore the factors that affect the injury severity of bicyclist in bicycle-motor vehicle accidents on police-reported crash data. The results showed that several factors doubled the probability of fatal injury in a bicycle-motor vehicle accident, including darkness with no street lights, inclement weather, peak hour in the morning, head-on and angle collision, speeding-involved, vehicle speeds about 48.3 km/h, truck involved, bicyclist age 55 or more, roads without median/division, running over bicyclist and etc. Klop and Khattak, (1999) studied the physical and environmental factors that influenced the injury severity level of bicycle crash on the two-lane and divided roadway. The author concluded the factors such as straight grades, curve grades, darkness, fog and speed limit might increase injury severity and the factors including higher AADT, lighting condition and speed limit at intersections and shoulder-width could lower injury severity level (Wang et al., 2017).

2.4 Statistic model

This section will demonstrate several useful count model that are widely used in safety research. Poisson, Negative Binomial (NB) and Zero-inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB) are introduced along with existing studies. In addition, some other potential methods are also reviewed regarding building SPFs.

2.4.1 Count model

If the story documented in *The role of intersection and street design on severity of bicycle-motor vehicle crashes* by Todhunter (1865) is true, Pierre De Montmort firstly mentioned NB distribution in the contest in 1713. He mentioned it captured the number of failures before a number of success in a series of binary trials. Based on Negative Binomial distribution, the Poisson distribution first introduced by Simeon Poisson (1781-1840) in the study about cranial and civil matters in 1838 (Hilbe, 2011). The Poisson distribution was also widely used to investigate variables that can influence the crash count. Oh et al. (2008) used the Poisson distribution to analyze bicycle collisions at signalized intersections in the urban area. In this study, bicycle variables were considered and the authors mentioned there could be more risk factors found if driver characteristics had been considered. The Poisson distribution also was used for analyzing the factors influence the risk level of bicyclists in the major cities in New Zealand (Tin Tin et al., 2013).

Later on, NB was derived from Poisson distribution because of its capturing the over-dispersed shape of data, which Poisson cannot. Poisson distribution has to have an assumption that the mean is equal to the variance. Therefore, NB distribution became a standard method of describing count data in most empirical data (Hilbe, 2011). Nordback et al. (2014) focused on finding (SPFs) for bicycles in cities in the United Stated and states that Poisson distribution can create a logical fit for the accident data but cannot capture the over-dispersed shape. Other studies analyzing crash count can be found in Figure 2-2 which shows part of the development of using models on crash data in transportation.

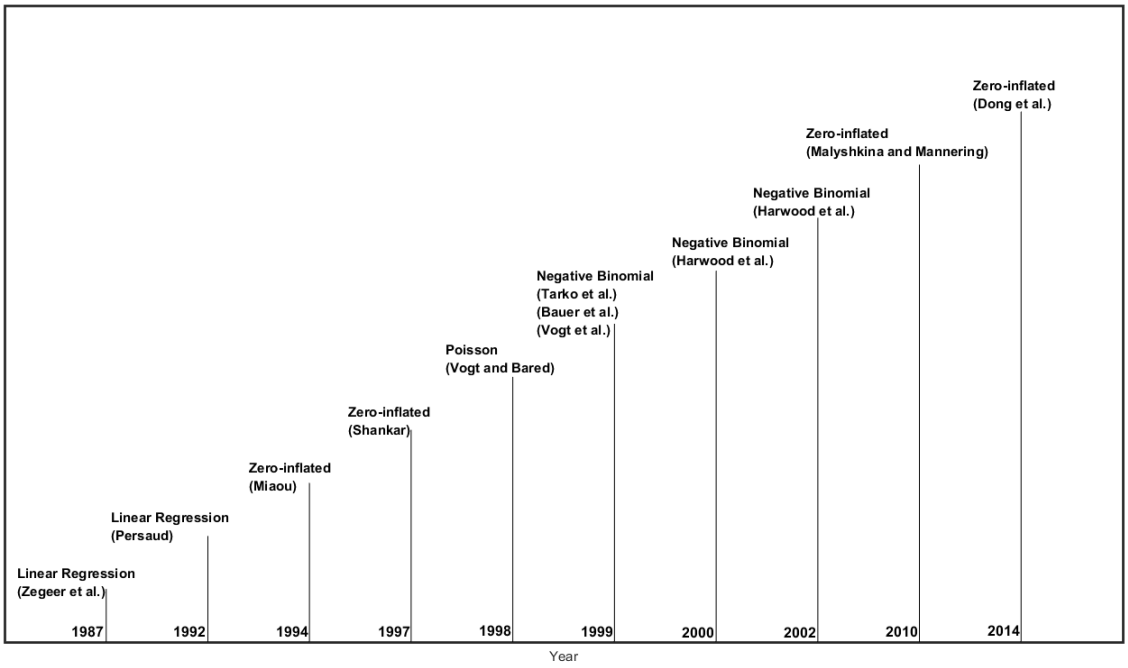

Figure 2-2 The development history of statistic model on crash count.

Late on, more scholars have been found a phenomenon where an observation of zero events happened in a period can arise a new condition. For example, zero crash can be noticed at many intersections. This interesting phenomenon may result from two reasons: 1) failure to be observed; or 2) unable to ever experience an event. Mullahy in 1986 introduced the idea of modeling with zero and Lambert (1992) Greene, (1994) later on extend the zero model to ZIP and ZINB model. This model allows additional zero states where the model can estimate the zero count separately from other non-zero observation. Recently, Dong et al. (2014) used multivariate random-parameters ZINB model to investigate the relationship between crash frequency with pavement condition, traffic factors, and geometric design. They found that this model has the ability to accommodate excess zero counts in crash data. Other researchers such as Miaou (1994), Shankar et al. (1997), Malyshkina and Mannering (2010) applied the zero-inflated model in the transportation field.

2.4.2 Alternative Approaches

Instead of the Poisson, NB, ZIP and ZINB models mentioned in the previous section, there are other statistical and mathematical methods available for predicting count data. Full Bayes method (FB) is one of them. FB is a modeling approach that can be

used the similar way as generalized modeling approach. Instead of using a point to predict the expected frequency and variance for a site, FB can provide a distribution of possible value, which can be combined with observed crash frequency to expected longer term crash frequency. FB has several advantages over other methods: 1) having the ability to specify complex models; 2) smaller sample size required; 3) having ability to consider spatial correlation; 4) having ability to incorporate prior knowledge, such as introducing previous reliable estimation in a new prediction model. However, it has disadvantages including very high statistical knowledge is needed and it is difficult to build a practical software based on FB (Gross et al., 2010). Some researchers have proposed FB method in various topics regarding safety including crash rate, ranking sites and identifying high-risk sites in road segments (Carriquiry and Pawlovich, 2004).

2.5 Data Collection Review

This section reviews the data collection process, significant variables in other studies, the way to assign crash to sample sites. The purpose of this review is essentially giving engineers a reference of collection data.

2.5.1 Sampling Process and Aggregating Crash Data

The first step of collecting data is a sampling process. Random sample process is normal way in research. Some issues arise along with the random sampling process. For example, spatial correlation indicates that one location is proximity to other places resulting an impact on predicting crash frequency (Gross et al., 2010). In 2006, a county level crash study in Pennsylvania found the significance of spatial correlation (Aguero-Valverde and Jovanis, 2006; Gross et al., 2010). Those findings indicate the spatial correlation may influence this study.

How to assign crashes to intersection sites is an critical step for building SPFs. One common and convenient way is to create a 2-dimensional buffer at an intersection within which crash will be assigned to the intersection. Fambro et al. (1997) first calculated 500 feet as Stopping Sight Distance (SSD) for intersection which can be the diameter for this buffer, but Dolatsara's (2014) demonstrated the 500 feet buffer is so large that it may cover crashes from adjacent intersections, shown in Figure 2-3.

Crash data, volume data and road geometric data are collected in Dolatsara's (2014) study. The crash data in this study were from 164 intersections in four cities in Michigan. The author used 250 feet buffer as a threshold to assign crash to intersection site. In other words, crashes happened within 250 feet buffer of the center point of an intersection were collected and assigned to the intersection (Dolatsara, 2014). This 250 feet has been widely used as a diameter to assign crashes (Vogt and Bared, 1998; Bauer and Harwood, 2000; Dolatsara, 2014).

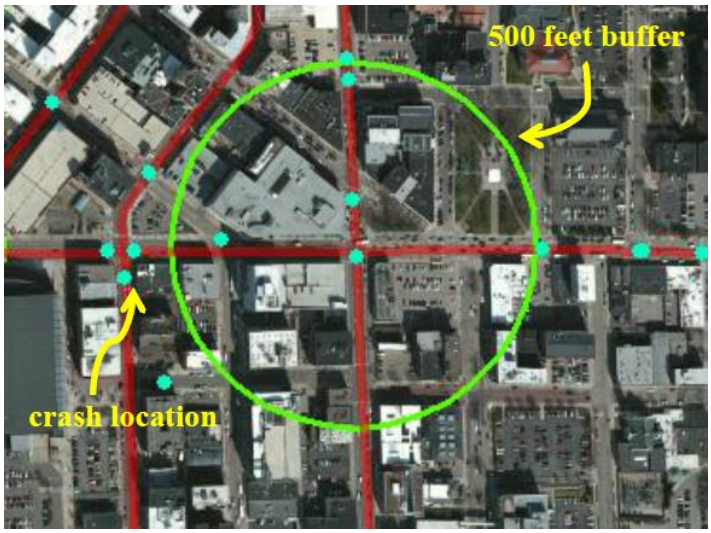

Figure 2-3 500 feet buffer used to assign crashes into an intersection (Dolatsara, 2014).

Justifying which threshold of buffer can be used to assign the crashes is a critical step. Street blocks within both downtown Portland and Eugene are smaller than areas outside downtown, and the sample sites in this research locate in all area of Eugene and Portland metropolitans, so the threshold of the buffer should be various depending on different land use. Therefore another way to avoid this error should be utilized in this present study.

2.5.2 Micro Scale Variables

Determining what data need to be collected for developing intersection SPF is critical since this step influences the data collection period, the final results, even the general research idea. Related researches and projects are reviewed in the section to provide engineer experience to make decisions on what variables are necessary.

Even though some of the bicycle researches or projects are not directly relating to building SPF, they still capture factors that may influence the bicycle risk or probability of bicycle crash. Monsere et al. (2016) analyzed the risk factors for pedestrian and bicycle crashes for Oregon. They collected the geometric data, STRAVA data, land use data and found several variables are significant, shown in Table 2-3.

Table 2-3 Variables and their significance in Oregon bicycle and pedestrian risk factors project (Monsere et al., 2017)

Data Element	Collection Method or Source	Significance
Traffic volume (AADT, factored 2014)	ODOT Databases, Local files, other sources	
Estimated bicycle volume per day	(STRAVA) STRAVA Database	×
Functional class of roadway	ODOT Databases	
Number/presence of left-turn lanes	Google Earth/ODOT Digital Video Log	
Number/presence of right turn lanes	Google Earth/ODOT Digital Video Log	×
Number of total traffic lanes		×
Presence of bicycle lanes	Google Earth/ODOT Digital Video Log	
Number of total traffic lanes (including left and right turn lanes) on all approaches	Google Earth/ODOT Digital Video Log	
Posted speed limit	Google Earth/ODOT Digital Video Log	
Presence of lighting by approach	Google Earth/ODOT Digital Video Log	
Number of total traffic lanes (including left and right turn lanes) on all approaches	Google Earth/ODOT Digital Video Log	
Presence of school area within 1000 feet	Google Earth	
Presence of green bicycle markings	Google Earth/ODOT Digital Video Log	
Number of bus stops within 1000 feet	Google Transit	×
Presence of median	Google Earth/ODOT Digital Video Log	
Functional class		×
Neighborhood concepts	GIS geodatabase Currans et al. 2014	
3-Leg Intersection Density	Environmental Protection Agency (EPA)'s Smart Location Database	
4-Leg Intersection Density	Environmental Protection Agency (EPA)'s Smart Location Database	
Retail Density	Environmental Protection Agency (EPA)'s Smart Location Database	
Total Population Density	Environmental Protection Agency (EPA)'s Smart Location Database	
Household Density	Environmental Protection Agency (EPA)'s Smart Location Database	
Household Size	Environmental Protection Agency (EPA)'s Smart Location Database	

FHWA developed Pedestrian and Bicyclist Intersection Safety Index (Ped ISI and Bike ISI) to help engineers, planner and other transportation agencies to prioritize the

intersection based on safety. The higher the score is, the higher priority the potential problem of corresponding intersection crosswalk and approaches need to be addressed. This bicycle analysis covered 67 intersection approaches from Gainesville, FL; Philadelphia, PA; Portland and Eugene, OR. Instead of taking intersection as a whole, this project evaluated the safety on crosswalks and each approach (through, left turn, right turn) in the intersection. Table 2-4 summarizes the variables required in this method.

Table 2-4 The data required in bicycle ISI method (Carter et al., 2006)

Variables	**Data type**
Bike lane presence	Yes/no
Cross street traffic volume	ADT numerical
Number of through lanes on cross street	Numerical
Number of traffic lanes for cyclists to cross to make a left turn	Numerical
Number, type, and configuration of traffic lanes on main street approach	Numerical
Street speed limit	Numerical
On-street parking on main street approach	Yes/no
Type of traffic control on the approach of interest (signal or no signal).	Signal/stop control

In 2012, Mekuria et al. developed Bicycle Level of Traffic Stress (BLTS) indirectly measured the stress of bicyclist based on cycling comfort level and surrounding environment. It has been an efficient tool to analyze bicycle behaviors, route choices and safety (Mekuria et al., 2012; Dill and McNeil, 2013). Based on the concept and content from Mekuria et al., a flow chart is created to demonstrate criteria and required data to evaluate BLTS in Figure 2-4 below.

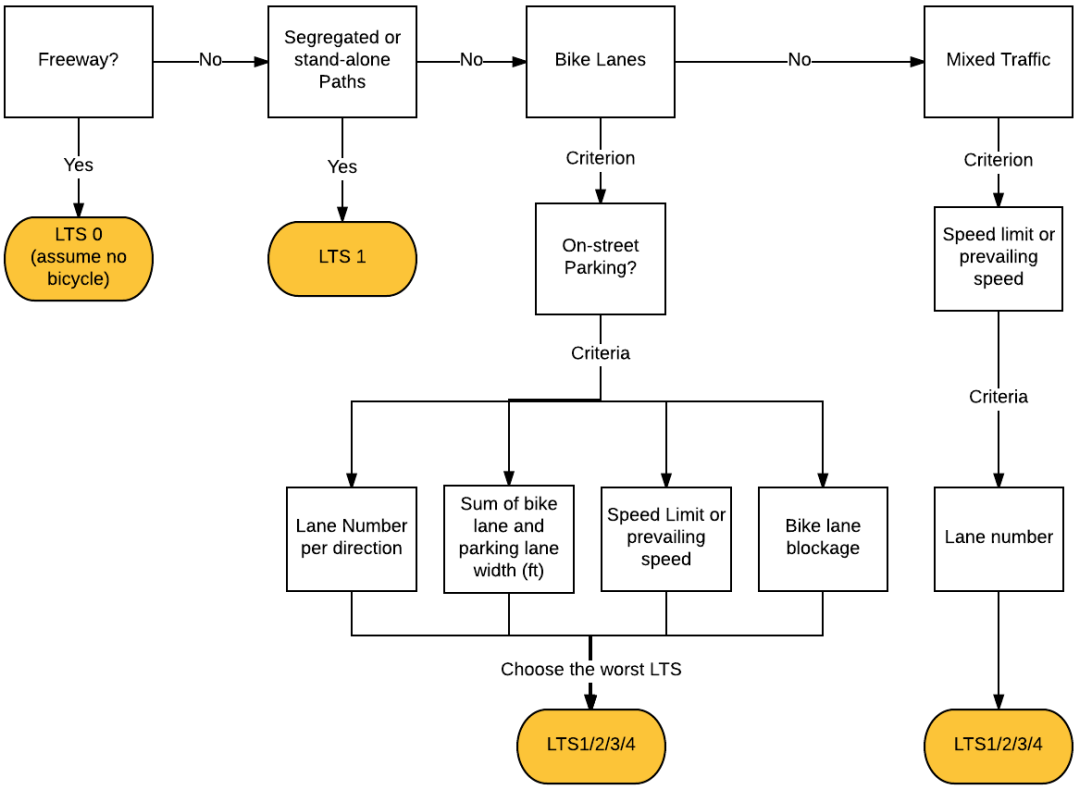

Figure 2-4 The BLTS criteria and data requirements.

Dolatsara (2014) built SPFs for non-motorized crashes in four Michigan cities. Besides traffic volume and bicycle volume, various geometric characteristics were collected in this study, and the significant variables in the final model are summarized in Table 2-5.

Table 2-5 Variables collected and corresponding significances for developing bicycle intersection SPFs.

Variables	**Significance**
ADT	×
Bicycle volume	×
Number of Left Lanes	×
Number of Through Lanes	
Number of Right Lanes	
Total Number of Lanes	
Presence of Bike Lane	×
Presence of Median	
Width of Corridors	
Length of Unpainted Crossing	
Number of Access	
Presence of On-Street Parking	
Presence of Speed Sign	
Posted Speed	
Presence of Bus Stop within 0.1 mile	×

In New Zealand, Turner et al. (2006) analyzed the prediction model for pedestrians and bicyclists. The variables include: intersection control types, intersection layout (four or three legs), and the number of traffic directions. Various geometric data turn out to play an important role in building intersection SPFs but not all. As one component of geometrics, horizontal curve tends to have an impact on segment safety (Findley et al., 2012). However, the horizontal curve is defined in segment rather than in intersection.

Validating existing SPFs in different areas or different time is important, since there is possible that the prediction model may not be able to estimate crash as accurate under temporal and spatial changes. Dixon and Avelar (2015) used new data sets to validate the old segment arterial SPFs model in Oregon that was developed in 2012 by ODOT. This function was designed to assess the safety performance of driveways located on arterial highways. Dixon and Avelar (2015) evaluated spatial transferability, spatial–temporal transferability, and individual coefficient stability and significance to verify the performance of old SPFs. The variables included in the 2012 Oregon urban arterial SPFs model are shown in Table 2-6. The validation method can be used in the research to verify the model prediction ability after development.

Table 2-6 The variables in 2012 ODOT arterial highway driveway SPFs (Dixon and Avelar, 2015)

Variables	**Significance**
AADT (Ln transformed)	×
Segment length (Ln transformed)	×
Post speed over 35 mph	×
Two-way left turn lanes	×
Four travel lanes	×
Two-way left turn lanes for four lanes	×
number of commercial plus industrial driveways	×
driveways that are not commercial or industrial	×

El-basyouny and Sayed (2011) used traffic conflicts instead of traffic volume as a predictor to develop SPF. They compared the conflict-based and volume-based SPFs, and the results showed those two approaches have similar crash prediction capability. It comes from the high correlation between conflict count and crash count, shown in

Figure 2-5. This paper disclosed the traffic conflicts could be alternative measures of traffic crash while developing bicycle SPFs.

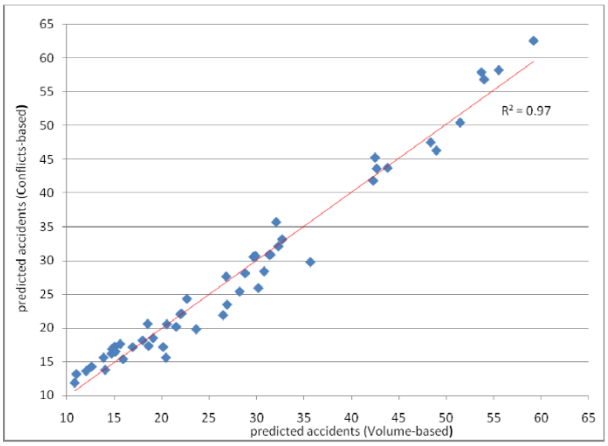

Figure 2-5 A scatter plot of predicted accidents obtained from conflict-based and volume-based SPFs model (El-basyouny and Sayed, 2011)

SPFs can also be developed for a specific area, such as land surrounding a school. McArthur et al. (2014) used NB model to establish pedestrian and bicycle SPFs for school, and the data collected are summary in Table 2-7:

Table 2-7 Variables and their significances for building school pedestrian and bicycle SPFs (McArthur et al., 2014)

Variables	**Significance**
Average Family Size	×
Children Ages 5 to 14	×
Average Parents per Household	×
Median Family Income ($)	×
Population Density (1,000 per sq. mi.)	×
Proportion of Non-White Households	×
Local Roads (proportion)	×
Collectors (proportion)	
Arterials (proportion)	
Students Enrolled	×

Turner et al. (2011) developed bicycle crash SPFs for two major cities in New Zealand and Australia. Those SPFs were developed for each movement for every city rather than looking the intersection as a whole. Crash, traffic and bicycle volume, geometric layout data were collected for road segments and intersections, shown as below:

- Volume for each vehicle movement
- Volume for each cycle movement
- Cycle lane width in meters plus 1m
- Depth of advanced cycle box in meters plus 1m
- Width in meters of lane closest to curb
- Presence of approach cycle facility
- Presence of transition cycle facility on approach
- Storage treatments present on approach
- Colored treatments
- Shared right-turn lane on motor vehicle movement approach
- Shared LT lane on approach
- Presence of free left turn on approach
- Fully / partially protected phasing arrangement at intersection

2.5.3 Macro Scale Variables

The variables that are significant impacting on intersection safety may differ with that on segments or corridors. In 2014, Kittelson&Associates Inc. prepared a plan For ODOT to prioritize the high-risk corridor for pedestrian and bicycle. Instead of analyzing each short segment, this study summarized the risk factor and score for each corridor (includes multiple segments). They stated that the main challenges faced by them include few crash number and lack of exposure data. Therefore, in order to overcome them, engineers combined network screening method and risk-based systemic safety planning process (identifying risk according to the roadway characteristics) to identify high-risk corridors. The variables included in the network screening method include:

- Driveway density
- Undivided 4-lane roadways in urban areas
- Lack of bicycle facility on at least one side of the roadway
- Presence of a traffic signal
- Average daily traffic (ADT or AADT)
- Posted speed limit
- Crash frequency and severity

In 2012, Teschke et al. recruited 690 city residents injured from cycling in Toronto and Vancouver in Canada to compare bicycle risk factors of 14 route type infrastructure features. They found: no presence of on-street parking and presence of bicycle lane on the major roads have negative impacts on risk; local streets also have lower risk; public transit, trucks, downhill grades, and construction are associated with increased risks.

Eluru et al. (2008) applied a mixed generalized ordered response logit (MGORL) on non-motorist crash severity using the 2004 crashes in the USA. The authors suggested that the general pattern of analyzing pedestrian and bicyclist crash were similar. In other words, there are some similarities between the impacts on bicycle crashes and pedestrian crashes. The significant variables include: gender, age, under the influence of alcohol, vehicle type, speed limit, time of day, traffic direction.

Kim et al. (2007) investigated the factors impacting bicycle crash severities by using a multinomial logit model on police-reported accident data from 1997 to 2002 from North Carolina, USA. Significant results include: age, the influence of intoxication, using helmet, speed, truck, bicycle direction, turning movement, curved road, two-way divided, land use type – institutional area, the weekend of week, time of day, weather, and lighting situation.

Yan et al., (2011) used a multinomial logit model to analyze bicycle crash severities and binary logit model to analyze bicycle crash patterns in Beijing, China. The significant variables of crash severity analysis are: crash pattern, age, vehicle type, speed limit, lighting situation, peak traffic hour.

Moore et al. (2011) used standard multinomial logit model and mixed logit model to estimate the influence of bicyclist, driver, motor vehicle, geometric, environmental, and crash type characteristics on bicycle injury severity for both intersection and non-intersection location. Non-intersection mixed logit model suggests significant variables include: gender, age, the influence of drug, the influence of alcohol, speed, dry/wet, driveways, seasonal of a year.

While researcher study bicycle safety on intersections (points) or segments (line), there are other engineers focus on macro level – spatial area unit (area). In 2015, Peng

Chen in the University of Washington used Poisson lognormal random effects model to analyze the correlation between building environment factor and motor vehicles involved bicycle crash frequency. He assembled datasets a rich source of datasets include road network, street elements, traffic controls, travel demand, land use, and socio-demographics. The significant variables with clear impacts (have either negative or positive influence on crash frequency, and can be determined simply by 95% credible interval sign) in the model include:

- 3-way intersection density
- 4-way intersection density
- On and off arterial bike lane
- Bus stop density
- Speed limits
- Traffic signal
- Number of automobile trips
- Mixing land use
- Household density

Wei and Lovegrove, (2013) aggregated data into TAZ zone to build a macro level crash predictive tool using negative binomial regression in Okanagan Regional District of British Columbia, Canada. They concluded that total lane length in area, bus stop number, traffic signals, intersection density, arterial-local street intersection percentage were associated with an increase in bicycle-motor vehicle crashes; while it has a negative relationship with the driving commuter number in the area.

2.6 Crowdsourced STRAVA data

Traditional bicycle data count is typically calculated from manual bicycle count during peak hours (Jestico et al., 2016), and it was calculated by multiplying daily or seasonal factors. However, traditional manual count data could be problematic. The manual count has a significant level of error. Nordback et al. (2013) found that when AADB is obtained by manual count more than one week of the hourly manual count, the average error is 30%, but the error can be 54% when AADB is only estimated by one hour of bicyclist count. Roll in 2013 also found that two-hour manual count

produced significant error in Oregon unless there is 24-hour count to determine the factors that can produce the least amount of error. Furthermore, traditional count method lacks spatial detail and temporal coverage (Ryus et al., 2014; Jestico et al., 2016). Global Positioning System (GPS) embedded in mobile devices allows people to track and map their locations and those data can be used for researchers to do analyses on bicycle behavior and route choice (Hood et al., 2011; Broach et al., 2012; Casello and Usyukov, 2014; Le Dantec et al., 2015; Jestico et al., 2016). Crowdsourced fitness app in mobile device provide a new source of data for transportation agencies and increase the temporal and spatial resolution of official counts (Jestico et al., 2016).

STRAVA® as one crowdsourced data using GPS has been used in different bicycle projects and researches all over the world: Queensland, Australia used it to quantify how a new bicycle pathway changed bicyclist behaviors; Glasgow, Scotland analyze 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 impact of its program on streets and on bike network; Oregon DOT used it to decide where to build bike counters and to adjust existing bike counter location to capture bicycle behavior better; Vermont Transportation used this data as their key layer for statewide planning designs; University of Victoria and University College London use it to model bicycling transportation in their area (Strava, 2016a).

In 2014 STRAVA® user accumulate 2,700,000,000 km and 75,700,000 riders all over the world (Scott, 2015). When it seems like STRAVA® has been taken a large proportion of market share, it is very necessary to be careful when using this data. When Oregon DOT paid $20,000 dollar a year data, ODOT acknowledged a problem that STRAVA's targeting demographic doesn't represent all cyclists. It is built for cyclists who treat bicycle as a recreation tool but not for bicycle commuters (Hunt, 2015). This paper warned us it is important to analyze how STRAVA® represent the real story before we fully believe it.

There are some existing papers verifying the representation of STRAVA® data for all bicyclists. Jestico et al. (2016) compared STRAVA® data with manual counts data in

Victoria, British Columbia in their study. The authors compared those two type of data by hourly, AM and PM peak and peak period totals separated by season. They use Generalized Linear Model (GLM) to capture the relationship between STRAVA® data and traditional manual count data and the results showed that there is 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%, but they also mentioned STRAVA® fitness data are a biased sample of ridership; however, it can represent categories of ridership and map spatial variance in an urban area with high temporal and spatial resolution.

Watkins et al., (2016) compared STRAVA® data with another transportation agency installing an app called "Cycle Atlanta." They found that Cycle Atlanta only represented 3% of manual counts, and there are also differences between STRAVA® and Cycle Atlanta. The representation should be carefully analyzed because of it has a bias on gender users, racing or commuting users, age, and income. However, STRAVA® data provides an opportunity for agencies to obtain data without creating their own app. They concluded that data from STRAVA® should be compared to local data sources and weighted appropriately and it can be a supplement resources of bicycle count. Selala and Musakwa, (2016) stated in their studies that it is clear that STRAVA® data is a useful tool that can provide efficient information when it comes to decision making and formulation of policies for Non-Motorized transportation program. In their paper, they also mentioned that only 20% of the cycling trips are commuting whereas recreational trips account for the left 80% in the city of Johannesburg. It is obvious that there are some levels of bias in STRAVA® data, but conclusive decision should be made with more information. Speaking about trip time, cycling count from STRAVA® has a higher number in the morning, and the number decrease as it approaches midday, then start increasing after that, finally decline again after 16:00. They said that the number of recorded by STRAVA® are affected by the availability of gated communities, income levels, crime level and the provision of infrastructure (Selala and Musakwa, 2016).

3. Methodology

This chapter documents the methodology used to build SPFs in this study. Poisson, NB, ZIP, ZINB models are applied and compared to find the best-fitted model. Crash distribution is utilized to estimate crash severity based on those model estimation. Different model assessment and goodness-of-fit measures are used to compare those models.

Different from other existing studies that normally use one model, this study identify the best-fitted models for various regressions. Then the model assessment and goodness-of-fit tools are used to compare different models in order to choose the best model among the various regressions – Poisson, NB, ZIP, and ZINB. Jurisdictions can use the established process to identify the suitable model for local bicycle crash data to build SFPs.

3.1 Poisson Model

Poisson regression is one of the most popular methods for count data. It has been applied to a wide range of transportation count data. For instance, Poisson was used estimate rate-event count data – accident occurrence, failure in manufacturing or processing, and the number of vehicles waiting in the queue at an intersection. This model has an assumption that the mean equals to its variance, expressed in equation 3.1.1.

$$VAR[y_i] = E[y_i] \tag{3.1.1}$$

where VAR represents variance; y_i represents site i has y time of event happens in a period of time; E indicates the expected mean. The number of event y has a Poisson distribution with a condition mean and depends on an individual's characteristics.

The expected value of y and its relationship with independent variables are written as (Long, 1997):

$$\mu_i = EXP(\beta X_i) \tag{3.1.2}$$

where μ_i represents the expected value of the response variable indicating the number of crash happens at a given site (an intersection or corridor) with characteristics; EXP is the exponential; β is the estimated coefficient of independent variable X_i.

The probability of a site i having y_i accidents in the study period is given by (Washington et al., 2011):

$$P(y_i) = \frac{EXP(-\mu_i)\mu_i{}^{y_i}}{y_i!} \tag{3.1.3}$$

where $P(y_i)$ is the probability of a site i having y_i accidents per year; μ_i indicates the Poisson parameter for this site, and it equals to the expected number of crash in the study period, $E[y_i]$. The Poisson model is estimated by the specifying the Poisson parameter μ_i as a function of explanatory variables such as road geometry, land-use, or traffic and bicycle volume.

Poisson model can be estimated by standard maximum likelihood function or log-likelihood function. Log-likelihood function is easier to be solved (Long, 1997) and is given by:

$$LL(\beta) = \sum_{i=1}^{n}[-\exp(\beta X_i) + y_i \beta X_i - Ln(y_i!)] \tag{3.1.4}$$

Where LL is log-likelihood; β is the estimated coefficient of independent variable X_i; Ln indicates taking log value.

3.2 Negative Binomial Model

Crash count especially bicycle crash count data is often found to be over-dispersed which means the sample variance is larger than sample mean. Poisson regression has an assumption that the mean equals to the variance, mentioned in the previous section; whereas NB regression can address data with the over-dispersed feature. Therefore, NB model has been a popular generation of Poisson regression, especially for count data (Long, 1997; Hilbe, 2011; Washington et al., 2011). The relationship between the independent variables and the dependent variable is given by:

$$\mu_i = EXP(\beta X_i + \varepsilon_i) \tag{3.2.1}$$

where ε is a random error that is assumed to be uncorrelated with X. It can be interpreted as unobserved variables omitted in model (Long, 1997); $EXP(\varepsilon_i)$ is a Gamma-distributed disturbance term with mean 1 and variance α. This addition term, different from the Poisson regression, allows the variance not equal to the mean and can be expressed as:

$$VAR[y_i] = E[y_i]\left[1 + \alpha E[y_i]\right] = E[y_i] + \alpha E[y_i]^2 \qquad (3.2.2)$$

The Poisson model can be interpreted as when α equals to 0 then the variance equals to mean plus zero. The parameter α often infers as the overdispersoin parameter so the NB distribution has probability function:

$$P(y_i|X_i) = \frac{\Gamma(y_i+1/\alpha)}{y_i!\Gamma(1/\alpha)}\left(\frac{1/\alpha}{1/\alpha+\mu_i}\right)^{1/\alpha}\left(\frac{\mu_i}{1/\alpha+\mu_i}\right)^{y_i} \qquad 3.2.3$$

where Γ is the gamma distribution function. This formulation has the likelihood function:

$$L(\beta) = \prod_i P(y_i|X_i) = \prod_i \frac{\Gamma(y_i+1/\alpha)}{y_i!\Gamma(1/\alpha)}\left(\frac{1/\alpha}{1/\alpha+\mu_i}\right)^{1/\alpha}\left(\frac{\mu_i}{1/\alpha+\mu_i}\right)^{y_i} \qquad (3.2.4)$$

where $\prod i$ represent the production of i individual likelihoods.

3.3 Zero-inflated Poisson and Negative Binomial

As mentioned in Literature Review Chapter, scholars have found a phenomenon that observations of zero events happened in a period can arise a new condition. In this case, zero crash is noticeable at many intersections or road segments and this interesting phenomenon is attributable to two primary reasons: 1) failure to observe; or 2) unable to ever experience an event. For instance, each state in the U.S. has a monetary threshold (i.e., $1000 in Oregon?) that requires users to report accidents to state, but the cost of accidents under the threshold will not be reported. This underreporting issue can cause "failure to be observed" because of not being reported. To solve the zero state issue, Mullahy in 1986 introduced the idea of modeling with zero and Lambert (1992) Greene (1994) later extended the zero model to ZIP and

ZINB model. This model allows additional zero state where the model can estimate the zero count separately from other non-zero observation.

Zero-inflated model implies that the underlying data-generating process has two splitting regimes for two types of zero (Washington et al., 2011). In other words, counts are generated by two processes in the zero state model. The splitting process is assumed to follow either Probit (normal) or Logit (logistic) cumulative density function (Long, 1997).

ZIP model assumes the events happening $Y = (y_1, y_2 \dots y_n)$ are independent and the model is given by (Washington et al., 2011):

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

$$y_i = y \; with \; probability \;\; \frac{(1-p_i)\exp(-u_i)u_i^y}{y!} \qquad (3.3.1)$$

where p_i is the probability of being in the zero state; y_i is the number of event in study period, and it is bicycle crash number at intersections or segments here; $u_i = \exp(\beta X_i)$. The variance is given by:

$$Var(y_i|X_i, Z_i) = \mu_i(1 - p_i)(1 + \mu_i p_i) \qquad (3.3.2)$$

The ZINB model is given by equation:

$$y_i = 0 \; with \; probability \;\; p_i + (1 - p_i)\left(\frac{1/\alpha}{1/\alpha + \mu_i}\right)^{1/\alpha}$$

$$y_i = y \; with \; probability \; (1 - p_i)\left[\frac{\Gamma(y_i + 1/\alpha)}{y_i!\Gamma(1/\alpha)}\left(\frac{1/\alpha}{1/\alpha+\mu_i}\right)^{1/\alpha}\left(\frac{\mu_i}{1/\alpha+\mu_i}\right)^{y_i}\right] \qquad (3.3.3)$$

where Gamma-distributed disturbance term with the mean 1 and the variance α. The variance of ZINB can be written as:

$$Var(y_i|X_i, Z_i) = \mu_i(1 - p_i)(1 + \mu_i(p_i + \alpha)) \qquad (3.3.4)$$

3.4 Crash Severity Distribution

Crash severity is another importance component of SPFs. It provides the proportion of different severities for sites. For instance, if the proportion of fatal crash at an

intersection from last year is 5%, then the estimated fatal crash number by using SPFs is 5% multiplying the predicted total crash number at this intersection. HSM included the crash severity distribution of motor-vehicle crash but not the bicycle crash severity. This study uses crashes that occurred on sampled sites (either intersections or corridors) to create the crash severity distribution.

3.5 Model Assessment

This section provides several model evaluation tools to assess models within one type of regression or models between different types of regressions. Assessment tools can be used to compare the difference between regressions to find the best regression, such as comparing between Poisson and Negative Binomial models. Goodness-of-fit statistics, Vuong statistic, and over-dispersion test are documented in this section.

3.5.1 Goodness-of-fit Statistics

McFadden Pseudo R-squared and Likelihood Ratio Test (LRT) are used to compare models in this study. Normal R-square is defined as the proportion of the variance in dependent variable y that can be explained by the $x's$ in a model. The R-square is given by:

$$R^2 = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2} \quad (3.5.1.1)$$

Where SSE is sum of squared errors; SSR is the sum of square; SST is the total sum of squares. When SSE=0 and $R^2 = 1$, then all variances are explained by the model; if SSR = 0 and $R^2 = 0$, then no association is found between independent variables and dependent variable (Washington et al., 2011). Therefore, the model is better when R^2 is closer to 1.

McFadden Pseudo R-squared, also called ρ^2, is given by:

$$\rho^2 = 1 - \frac{LL(\beta)}{LL(0)} \quad (3.5.1.2)$$

where $LL(\beta)$ is the log-likelihood for fitted model with coefficient vector β; $LL(0)$ represents the log-likelihood of reduced mode with only constant as independent variable. The interpretation of McFadden Pseudo R-squared is slightly different from R-squared and the relationship can be found in Figure 3-1. It should be mentioned

that the McFadden Pseudo R-squared between 0.2 and 0.4 could be interpreted as perfect fit, so McFadden Pseudo R-squared is normally found to be low (Domencich and McFadden, 1975).

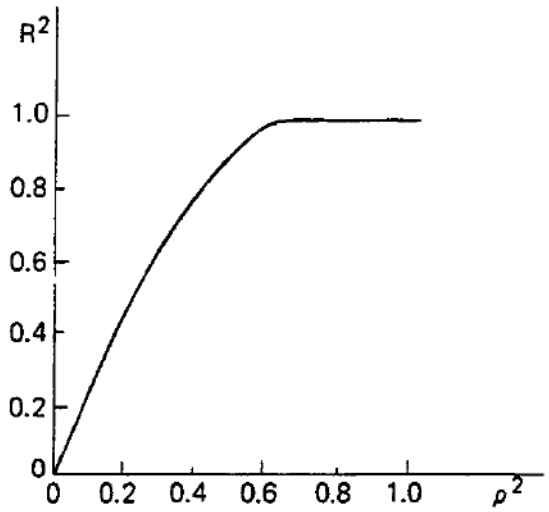

Figure 3-1 the relationship between McFadden Pseudo R-squared and R-squared (Domencich and McFadden, 1975)

LRT is a statistic hypothesis test used to compare two competing models based on the Chi-square distribution. It is used to compare the best-fitted model and the restricted model (only with constant as the independent variable). The test is given as:

$$LRT = -2[LL(0) - LL(\beta)]$$

where $LL(\beta)$ is the log-likelihood for fitted models with coefficient vector β; $LL(0)$ represents the log-likelihood of reduced models with only constant as the independent variable.

3.5.2 Vuong Non-Nested Hypothesis Statistic

Vuong (1989) proposed a statistic test for non-nested models to verify the appropriateness of using a zero-inflated model rather than a traditional model. This tool is mainly used to compare the Poisson, ZIP, NB, and ZINB models. The statistic calculation is given as:

$$m_i = LN(\frac{f_1(y_i|x_i)}{f_2(y_i|x_i)}) \quad (3.5.2.1)$$

Where $f_1(y_i|x_i)$ represents the estimated probability density function of model 1; whereas $f_2(y_i|x_i)$ represents the estimated probability density function of model 2; LN indicates taking log value. Using the approach, Vuong statistic of model 1 versus model 2 is given by (Long, 1997; Washington et al., 2011):

$$V = \frac{\sqrt{n}[(1/n)\sum_{i=1}^{n} m_i]}{\sqrt{(1/n)\sum_{i=1}^{n}(m_i-\bar{m})^2}} = \frac{\sqrt{n}(\bar{m})}{S_m} \qquad (3.5.2.2)$$

Where m_i represents the mean and S_m is the standard deviation; n is the number of observation. If V number is larger than critical value 1.96 (95% confidence level) then the first model is preferred; if the V is less than – 1.96 then the second model is favored; if the V is between -1.96 and 1.96, then the results is inconclusive and neither of two models is preferred (Washington et al., 2011).

3.5.3 Over-dispersion Test

Using Poisson regression or NB is based on whether the data is over-dispersed. A test for over-dispersion was introduced by Cameron and Trivedi (1990). This test relies on the assumption that under the Poisson model, the difference between the variance and the mean $(y_i - E[y_i])^2 - E[y_i]$, is zero; where $E[y_i]$ is predicted count. The null and alternative hypotheses are used to test the significance of over-dispersion:

$$H_0 = VAR[y_i] = E[y_i]$$

$$H_A = VAR[y_i] = E[y_i] + \alpha g(E[y_i]) \qquad (3.5.3.1)$$

where $g(E[y_i])$ is a function of predicted count and it is given by $E[y_i]$ or $(E[y_i])^2$. A t-statistic is used to justify the significance of null and alternative hypotheses. The absolute result of t-statistic larger than 1.96 indicate the appropriation of NB model and the rejection of Poisson; otherwise the Poisson is preferred (Washington et al., 2011).

The useful summary of over-dispersion parameter and Vuong-statistic is provided by Washington, Karlaftis and Mannering (2011) and is shown in Figure 3-2.

Decision Guidelines for Model Selection (Using the 95% Confidence Level) among Negative Binomial (NB), POISSON, Zero-Inflated POISSON (ZIP) and Zero-Inflated Negative Binomial (ZINB) Models Using the Vuong Statistic and the Overdispersion Parameter α

		t-Statistic of the NB Overdispersion Parameter α	
		$<\|1.96\|$	$>\|1.96\|$
Vuong statistic for ZINB($f_1(\cdot)$) and NB($f_2(\cdot)$) comparison	<-1.96	ZIP or Poisson as alternative to NB	NB
	>1.96	ZIP	ZINB

Figure 3-2 Vuong-statistic and Over-dispersion test

3.6 Other Methods

In the data collection process, the author found that not all intersections and corridors have AADT available, but the Average Daily Traffic (ADT) is available much more often than AADT. ADT is not an accurate representation of the traffic count as AADT. However, ADT can be converted to AADT by:

$$AADT = ADT \times F_d \times F_m \times F_y \quad (3.6.1)$$

where F_d is the daily factor in transportation and is the ratio of the daily traffic volume to average daily traffic for the whole week; F_m is the monthly factor and is calculated by the ratio of the average daily traffic in a certain month to the AADT for the year; F_y is the yearly factor is calculated by the ratio of the average daily traffic in a study year to the average daily traffic for the data year.

Another critical component of this study is to compare the crowdsourced STRAVA® data with actual AADB recorded by the automatic count stations or loops. The STRAVA® is compared with AADB from an existing count station (i.e., Eugene or Portland?) to get the representativeness proportion. More details can be found in section Data Preparation and Analyses.

4. Data Preparation and Analyses

This chapter discusses the data collection process and data analyses results. This chapter presents the selection of sample sites, discussion of variables, representativeness and bias in data collected, data visualization and analyses. Several assumptions are made for the data collection process:

- Road characteristics, geometric data, and land-use data keep constant in study period – 6 years from 2009 to 2014;
- The change of AADT and AADB in study period is relatively small;
- Data from sources are reasonably accurate.

Each data has its features and bias and will be discussed in this section.

4.1 Sample Sites Selection

Samples were selected from Portland and Eugene-Springfield metropolitans in Oregon, because those two cities 1) have enough bicycle volume to justify the statistical model, and 2) represent different types of cities (Portland represents an economic center of Oregon and Eugene represents a medium size city with University).

The statistically proved sample sites selection method is a random sampling process; however, issues arise when the author applied the random sampling process: 1) there is no intersection GIS ((Geographic Information System) files or other types of files available to conduct random sampling approach in Oregon; 2) the spatial correlation issue arises in downtowns with higher density of intersections and road segments in Portland and Eugene.

In order to solve those two problems, a "random and systematic" sample selection approach is proposed. No intersection GIS files available motivate engineers to use segments as an alternative to select intersections: the random sampling process is conducted on segments (available in STRAVA® GIS file), and the closest intersection is selected as the sample. Furthermore, areas with high density of intersections and segments have higher probability to be selected, which can cause the spatial correlation issue (mentioned in Literature Review). Therefore, a systematic sampling

method is used to mitigate the spatial correlation issue: before an intersection is selected, Portland and Eugene metropolitans were divided into hexagons with a radius of 0.5 miles using ArcGIS®, shown in Figure 4-1. Intersections located near to the selected intersection were excluded from the sampling population. In other words, only one intersection will be chosen in one hexagon to mitigate the spatial correlation issue.

Figure 4-1 Systematic and random sampling process in Portland using ArcGIS®.

Different from selecting an intersection, a pure random sampling process is used to select corridor samples. Once a segment was selected, the nearest ADT or AADT point is selected, then the corridor on this ADT or AADT point was selected as a sample. The samples are shown in Figure 4-2 and Figure 4-3.

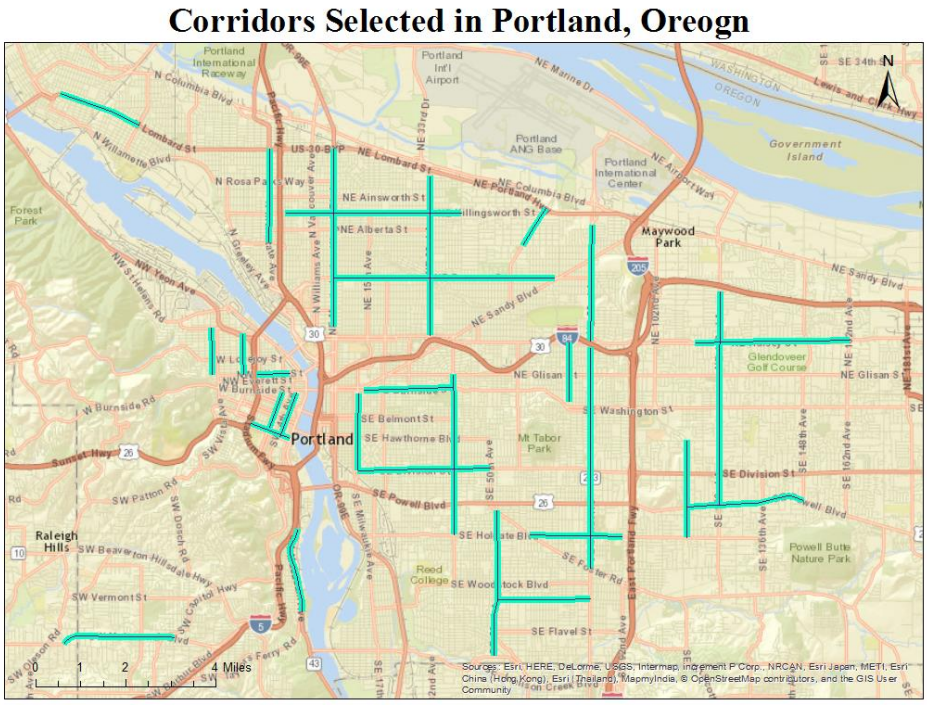

Figure 4-2 Corridor samples selected in Portland, Oregon.

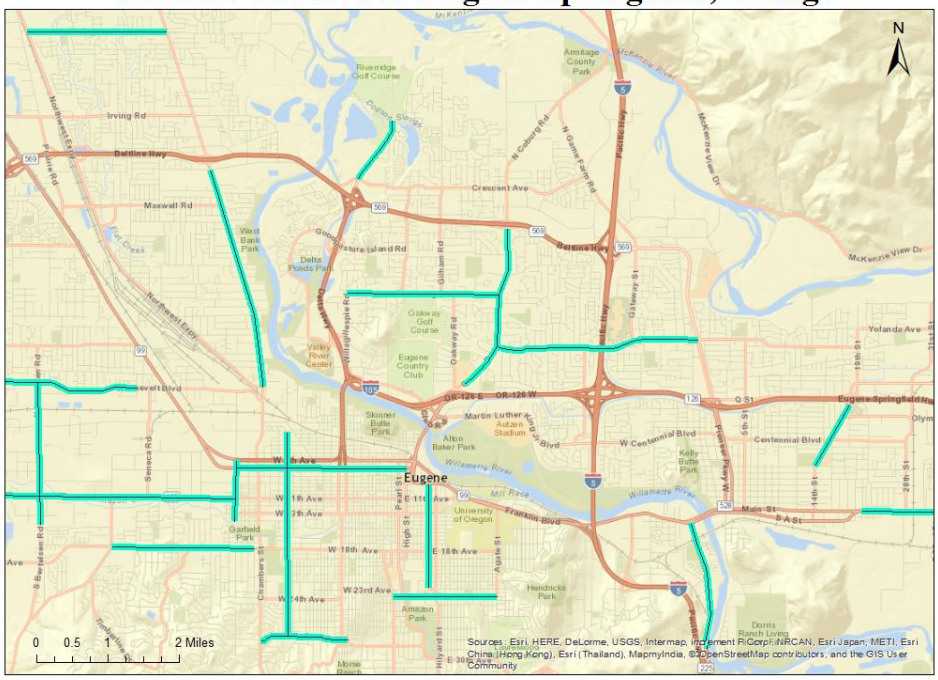

Figure 4-3 Corridor samples selected in Eugene-Springfield, Oregon.

It should be noted that because of the lack of ADT or AADT on lower functional classification streets, the random sample was not chosen when traffic volume is not available. Thus the corridor samples selected have a potential of bias of over-representing higher functional classification streets, such as major arterials or major collectors. Intersection samples have a similar issue. Additionally, intersection closed to or under the influence from freeway ramp are excluded.

4.2 Independent and Dependent Variables

The section documents the process of collecting and converting each variable in this study including traffic volume, bicycle volume, geometric data, land-use data, and other road characteristics.

4.2.1 AADT

AADT (or ADT) for intersection and corridor, as one of the most critical components of SPFs, is collected from ODOT (Oregon Department of Transportation, 2017), City of Portland (Portland Bureau of Transportation, 2016), and Eugene-Springfield MPO (Central Lane Metropolitan Planning Organization, 2017). Most AADT data are from ODOT as GIS file; whereas other data sources only with ADT available are then converted into AADT using Equation 3.6.1. All ADT or AADT were converted into 2014 using weekly and monthly or yearly factor from ODOT Traffic Volume Table 2015 (Oregon Department of Transportation, 2016). 2014 is used because STRAVA® data is only available for this year.

AADT for intersections contains two parts: major and minor road. A major road is defined having a higher functional classification or having higher traffic volume over the minor road. For example, in Figure 4-4, the intersection of Martin Luther King Rd. and NE Killingsworth St. is selected as a sample, and they both have functional classifications of major arterial. Martin Luther King Rd has AADT of 21800 versus NE Killingsworth St. has 14000. Therefore Martin Luther King Rd are collected as the major road and the other is the minor road at this intersection. If the AADT site is too far away from the selected intersection or a major arterial existing between an intersection and an AADT site, then engineers found another closer AADT site unless the arterial in between is believed to have negligible influence on the AADT.

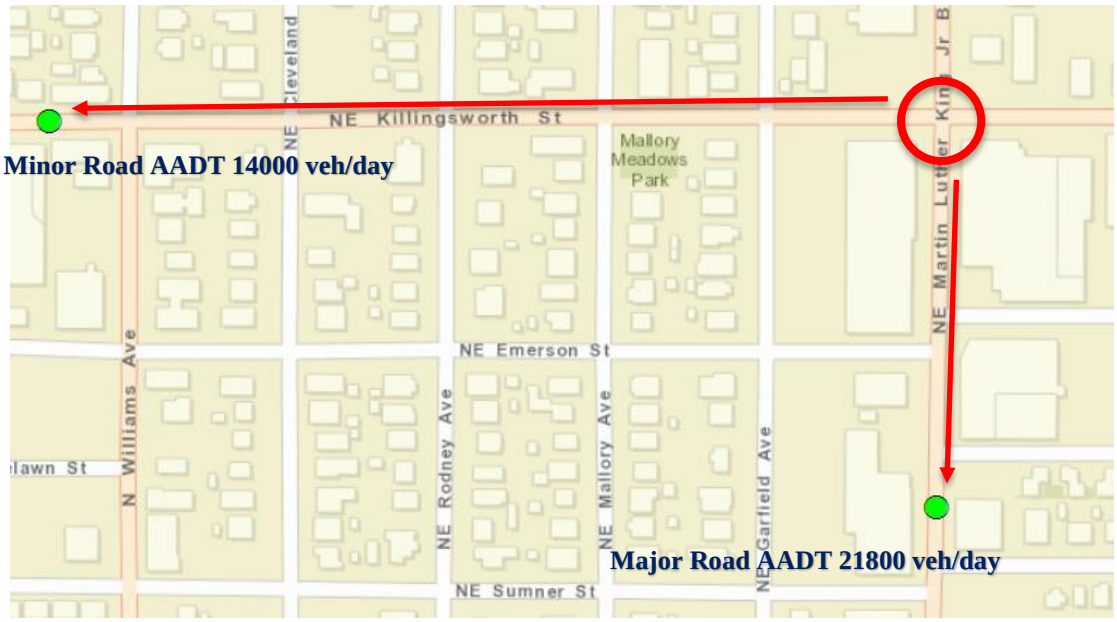

Figure 4-4 Collecting AADTs for an intersection on both major and minor roads.

The way to Collect AADT or ADT for a corridor is different. All available AADT sites on selected corridors are collected, and the average AADT is calculated as the AADT for the whole corridor. Shown in figure…, for instance, the Martin Luther King Rd is selected as a corridor sample, then AADT site 1 and AADT site 2 are both collected and average value of them is the AADT for the whole Martin Luther King Rd corridor.

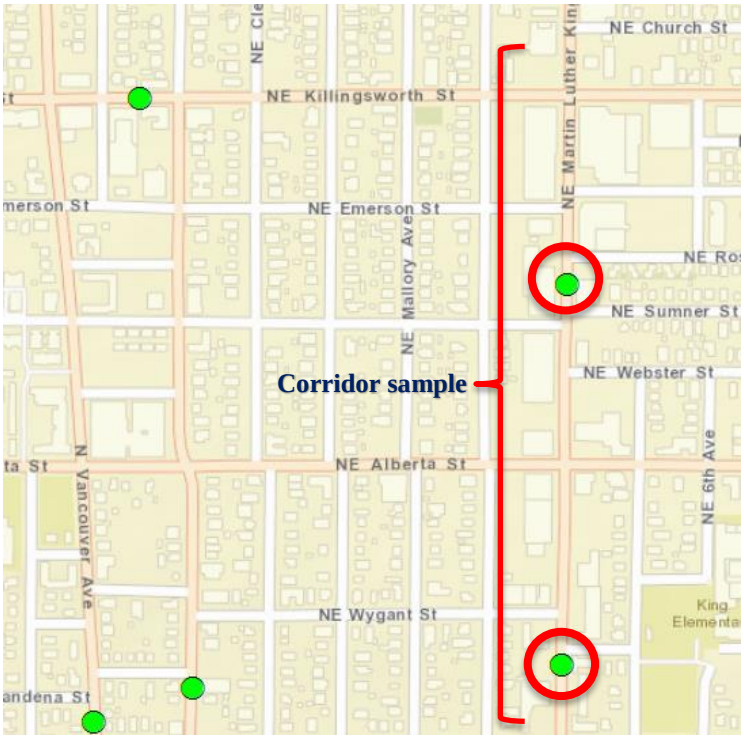

Figure 4-5 All available AADT on a corridor are used to calculated average AADT.

4.2.2 AADB: STRAVA® Data

STRAVA® is a mobile application that can track athletic activities including cycling and running through Global Positing System (STRAVA, 2017). STRAVA® app can record the detailed information, such as the speed, route, location and time, while athletes are doing exercises with the app active on their mobile devices. Bike count data is not widely available through Oregon, also the majority of cities in the U.S., so an efficient and affordable way to collect bike data is necessary for this study. Since STRAVA® has social network features by which users can communicate and involve with other users and groups, it attracts lots of people to use this mobile application. Thus it creates an opportunity for researchers to use STRAVA® bicycle count to represent the bike volume in building SPFs.

STRAVA® Metro is created through the cooperation with various Department of Transportation and STRAVA®. This tool aggregates all the cycling records from STRAVA® members into GIS files. ODOT purchased STRAVA® Metro product of Oregon in 2014 for research and project purposes. In this data, locations and time frames are aggregated into the street network and compiled to shapefiles that can be

used in Geographic Information System. The GIS map provides the information from cycling records for each segment, including location, time, month, year, week or weekend, gender, commuter or cyclist. Figure 4-6 demonstrates the bike count on each link in Oregon.

Figure 4-6 STRAVA® cyclist count in Oregon (left) and in Portland Metropolitan area (right) (Strava, 2016b)

Similar to collect AADT, STRAVA® data on both major road and minor road are collected for intersections. One advantage of this type of crowdsourced data is that each segment has bicycle volume available. Therefore, the bicycle volume on each leg of intersection is known. For instance, shown in Figure 4-7, when the intersection of 5th St. and Alder St. in Portland Downtown is selected, then the bicycle volume on all four legs of the intersection can be obtained through STRAVA® data in ArcGIS. When bicycle volume is collected for all legs at an intersection, each bicyclist is actually counted twice, so the correct bicycle volume is using the total volume of all legs divided by two.

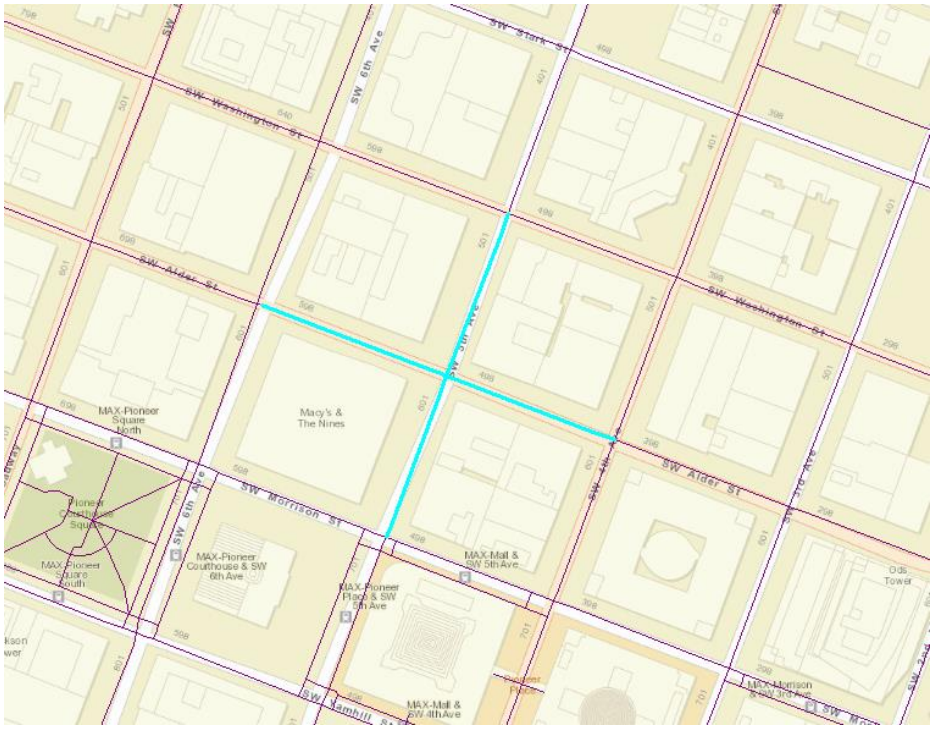

Figure 4-7 Example of collecting bicycle Volume for an intersection in Portland.

One issue of STRAVA® data: more than one link are representing the same line at some segments, especially when the road is wide. For instance, Figure 4-8 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 results from the bike count assignment process. STRAVA® built buffers around GPS signal to assign bike counts to segments which causes the double counting problem when two parallel bicycle paths are close to each other (Monsere et al., 2017). Thus, to address the problem, the author manually collect and check each bicycle volume through STRAVA® GIS product, instead of building buffers to collect in GIS which may cause double counting issue.

Figure 4-8 Multiple bike links on the same road in Portland Downtown area in GIS.

4.2.3 Geometric, Land-use, and Road Characteristic Data

Geometric, land-use, and road characteristics data, such as lane number and posted speed, are collected from Google Earth street view, ODOT TransGIS®, City of Portland and Eugene-Springfield MPO. Those data collection method and sources for intersection and corridor are summarized in Table 4-1 and Table 4-2.

Table 4-1 Geometric and Road Characteristic Data for intersection

Data Element for intersection		Collection Method or Source
Volume	Traffic volume (AADT, or ADT, factored 2014)	ODOT TransGIS Databases; City of Portland ADT; Eugene MPO;
	STRAVA data 2014	(STRAVA) STRAVA Database
Functional Class	Major road Functional class	ODOT TransGIS
	Minor road Functional class	
Intersection configuration for Major and Minor roads	One/two way on Major	Google Earth
	One/two way on Minor	
	Presence of bicycle lanes on Major	
	Presence of bicycle lanes on Minor	
	Presence of left turn lanes on minor road	
	Presence of right turn lanes on major road	
	Number of total traffic lanes on Major	
	Number of total traffic lanes on Minor	
	Number of total traffic lanes (including left and right turn lanes) on all approaches	
	Presence of on-street parking	
	Presence of median	
Posted speed limit	Major	ODOT TransGIS/Google Earth
	Minor	
Type of traffic Control	Signal, two-way/four-way stop	Google Earth
Number of transit stops within 5000 feet	Number of transit stops within 500 feet	Google Transit
Land use	3-Leg Intersection Density (per square mile)	Environmental Protection Agency (EPA)'s Smart Location Database (D3bmm3)
	4-Leg Intersection Density (per square mile)	Environmental Protection Agency (EPA)'s Smart Location Database (D3bmm4)
	Total Road network density (per square mile)	Environmental Protection Agency (EPA)'s Smart Location Database (D3a)
	Population Density (per square mile)	Environmental Protection Agency (EPA)'s Smart Location Database (D1b)

Table 4-2 Geometric and Road Characteristic Data for corridor.

Data Element for intersection	**Collection method and sources**

Geometric	Length (mile)	Google Earth
	Functional classification	
	One/two Way	
	Presence of on-street Bicycle Lane	
	Number of Total Traffic Lanes	
	Presence of Median	
	Presence of Two -way left turn lane	
	Presence of On-street Parking	
	Presence of Curve	
	Percentage of Curve of segment	
	speed limit	ODOT TransGIS Databases; City of Portland ADT; Eugene MPO;
	intersection number	Google Earth
Operation	bus stop number	ODOT TransGIS
	number of bus route on segment	
	AADT on segment	ODOT TransGIS Databases; City of Portland ADT; Eugene MPO;
	Average STRAVA on segment	(STRAVA) STRAVA Database
	Number of intersection with Signal Control	Google Earth
Land use	3-Leg Intersection Density (per square mile)	Environmental Protection Agency (EPA)'s Smart Location Database
	4-Leg Intersection Density (per square mile)	
	Total Road Network Density (per square mile)	
	Population Density (per square mile)	
	household density	

4.3 Representativeness and Bias

Every research based on samples must answer this question: how the samples represent the total population? This study provides the answer in this section. Representativeness is analyzed based on sample selection process and variables collected. Crowdsourced data -- STRAVA® is specifically analyzed regarding users and comparison with automatic count stations and loops. Some unobserved issues from crash data are also provided.

4.3.1 Representativeness of Samples

Even though authors combined random sampling and systematic sampling approaches to avoid spatial correlation issue and to be as unbiased as possible, there are still some issues arising from the sampling process based on AADT availability.

As aforementioned in the Sample Sites Selection section, the AADT availability influences the sampling process. Sample sites are generated randomly, but then would not be chosen if there is no traffic volume available at the intersection or on the corridor. In other words, the sample site location is somewhat correlating to AADT or ADT locations from data resources. Since most traffic volume sites are located on arterials and collectors, the samples represent more on roads with higher functional classifications. Figure 4-9 shows the percentage of functional classification for intersection and corridor samples. On corridors and major roads at intersections, the majority of samples are on arterials, and only 2-4% are local roads. Thus, it is reasonable that fewer minor roads than major roads are collected in data set.

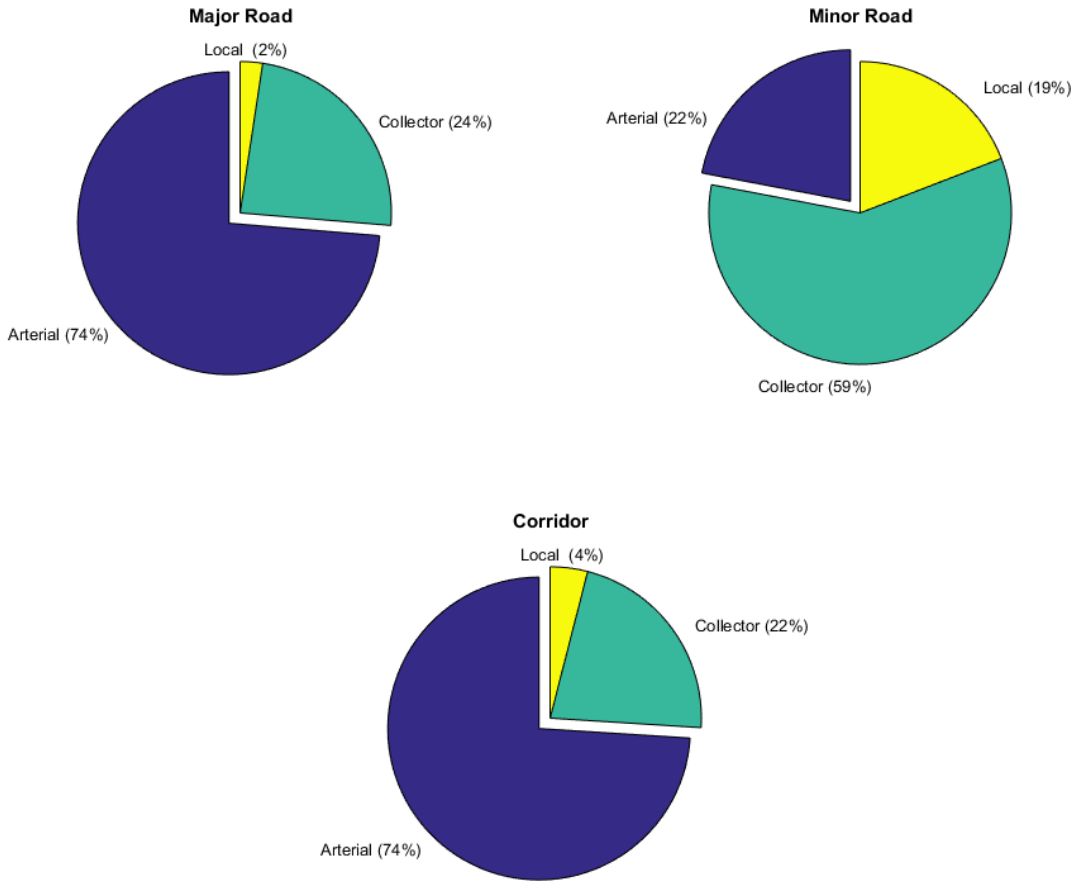

Figure 4-9 Functional classification of intersection legs and corridors

This functional classification distribution may indicate the results from SPFs of this study highly represent arterials and collectors. Since the data samples of local road may be limited, the predictive ability of SPFs on the local road may not be as accurate as that on the major roads. However, the overall predictive ability of SPFs may not be influenced because the equations are also built on traffic volume and other factors. Furthermore, there are more accidents and bicycle volume on the higher class road, so SPFs are still targeting on the majority of users on roads.

4.3.2 Bias in STRAVA®

As mentioned previously, STRAVA® is a mobile application that can track athletic activities including cycling and running through Global Positing System (STRAVA, 2017). Since STRAVA® has social network features by which users can communicate and involve with other users and groups, it attracts lots of people to use this mobile application, and ODOT purchased 2014 STRAVA® Metro product of Oregon.

It is because only 2014 STRAVA® is available, the first bias appears because the study period is six years from 2009 to 2014. Bicycle crashes from 2009 to 2014 were collected for this study, and that assumes the STRAVA® would not change during the six years, but zero change is impossible in reality. Since the bicycle trip has been increasing recently, the STRAVA® from 2009 to 2014, if existing, have a high possibility to have an increasing trend. Therefore, only use 2014 data can result in over-represent bicycle volume of average bicycle volume in those six years.

The second bias arises as using STRAVA® to represent all bicyclist population because the majority of STRAVA® users are biking for recreation than commuting. In other words, this type of data under-represent bicycle commuter population. However, this bias exists when bicyclists are divided into two groups: recreational bicyclists and bike commuters. Recreational bicyclists have higher biking skill and choose different route comparing to commuters.

The STRAVA® bike count can roughly represent the bike volume for an percentage. STRAVA® has differentiated the commuting count and recreational count in the data set. To check what is the percentage of STRAVA® data is commuting count, a describe statistic test was done and showed in Figure 4-10 and Figure 4-11. The figures show that on average 30-40% of STRAVA® strips were done for commuting purpose in Portland and Eugene in 2014, but with a range of variance.

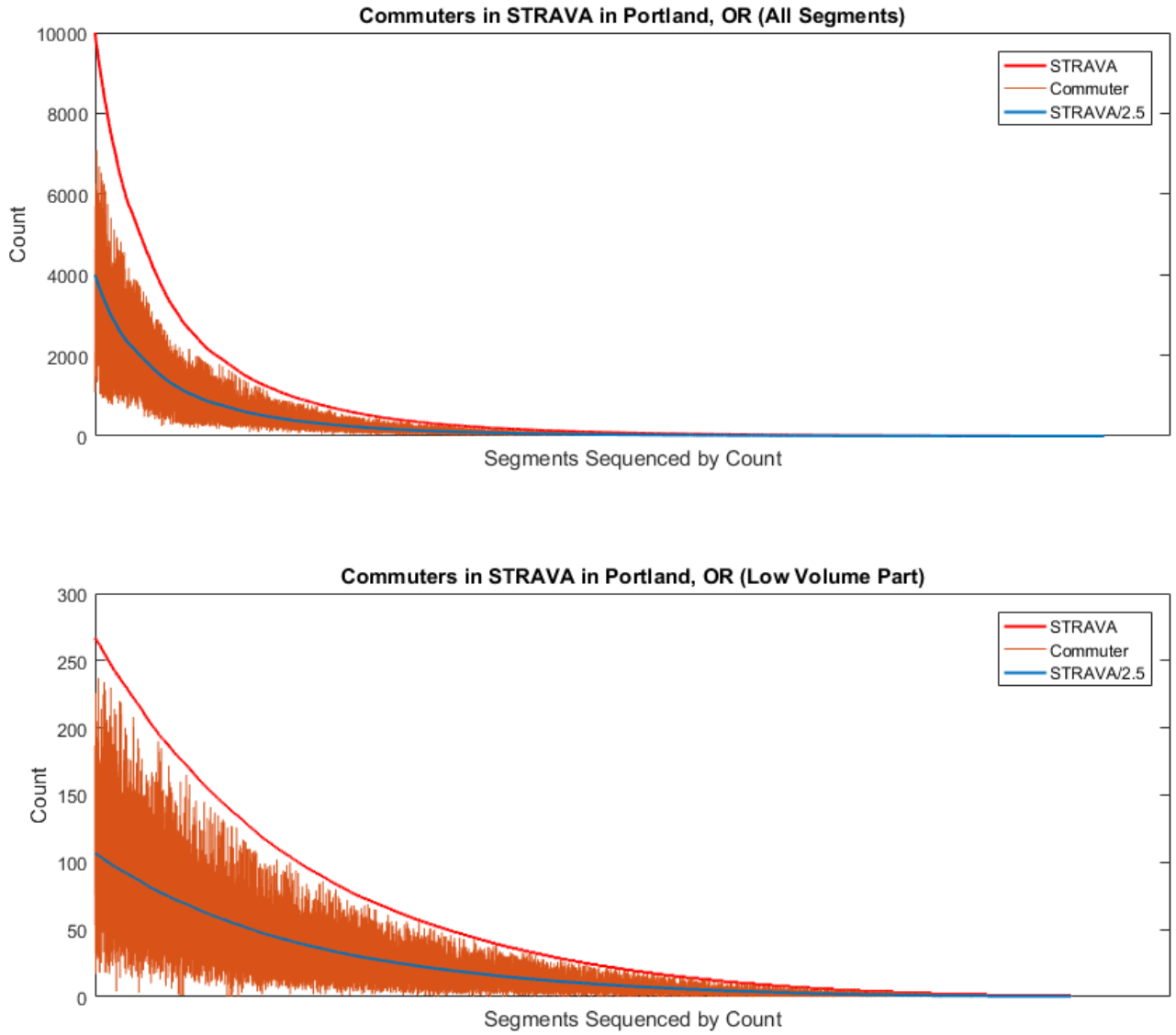

Figure 4-10 Commuters in total STRAVA count in Portland in 2014.

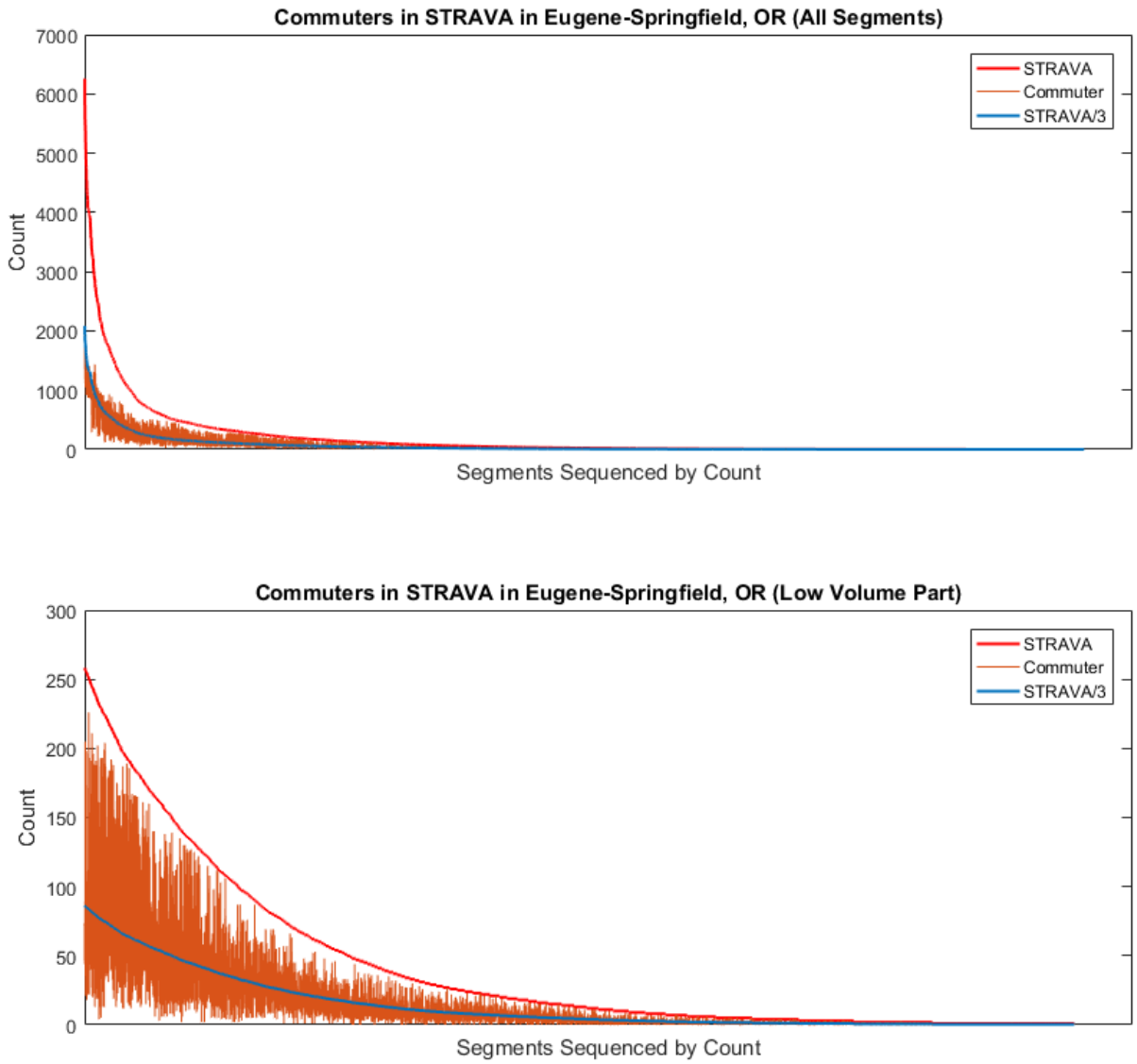

Figure 4-11 Commuters in total STRAVA count in Eugene-Springfield in 2014

Tow figures also show that the average commuter percentage in Portland (40%) is slightly different from that in Eugene-Springfield (33%). This indicates the difference in popularity of using STRAVA® and trip purpose in different cities. Even though STRAVA® provided commuting count in row data, the author investigated that this percentage should be utilized with caution, since STRAVA® does not mention how they differentiate commuters from other bicyclists in data user manual. It may form two approach: 1) embedded button in STRAVA® App.; 2) STRAVA® has an algorithm to calculate user types by different speed. However, there are concerns regarding both of the potentials. Shown in Figure 4-12, a button is existing for users to switch the trip is for the recreational or the commuting purpose. When it seems reasonable to differentiate user type from this button, there is no evidence support that a user knows or remember switching to "commute" option when they ride bikes for commuting.

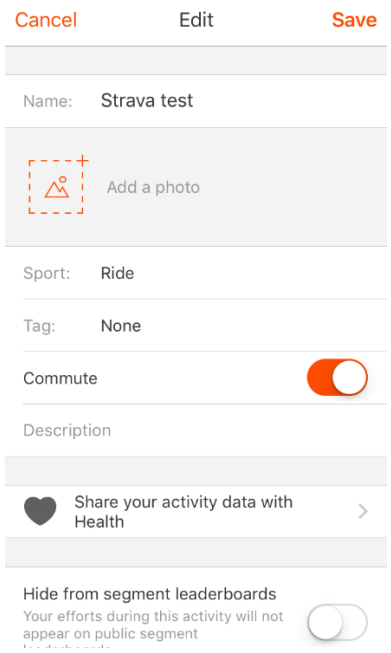

Figure 4-12 STRAVA App. Smartphone interface.

The other possible method to separate user types – by using speed can raise an issue: it is difficult to use a speed threshold to decide whether a trip is for recreation or commuting. Cyclists with higher biking skill can ride with high speed but other ordinary users can also bike for recreation with lower speed.

Thus, when STRAVA® is used represent all types of users, it represents more recreational trips than commuting trips. The percentage of representativeness is on average from 30-40% but with variance and unobserved uncertainties.

4.3.3 Comparing STRAVA® with Automatic Count Station

Another critical assumption of this study is the STRAVA® can partially represent the total bicyclists. In order to justify this assumption, STRAVA® data is comparing with automatic bicycle counter on Hawthorne Bridge in Portland. The Hawthorne Bridge bicycle counter is a 24-hour automatic counter record bicycle volume since 2012, shown in Figure 4-13.

Figure 4-13 Hawthorne Bridge bicycle counter (Haberman, 2017)

To comprehensively understand the representativeness of STRAVA® of the different time in a year, bicycle volume of each month retrieved from STRAVA® are compared with month volume from Hawthorne Bridge. The result, demonstrated in Figure 4-14, shows that STRAVA® can generally represent 1.4% of the total bicyclists in on this bridge and also with similar percentage of each month. Even though the bridge may not represent all locations' situation in Oregon, the stable percentage ratio between STRAVA® and auto-counter in each month justifies that STRAVA® can represent a certain proportion of the total bicyclists in Portland.

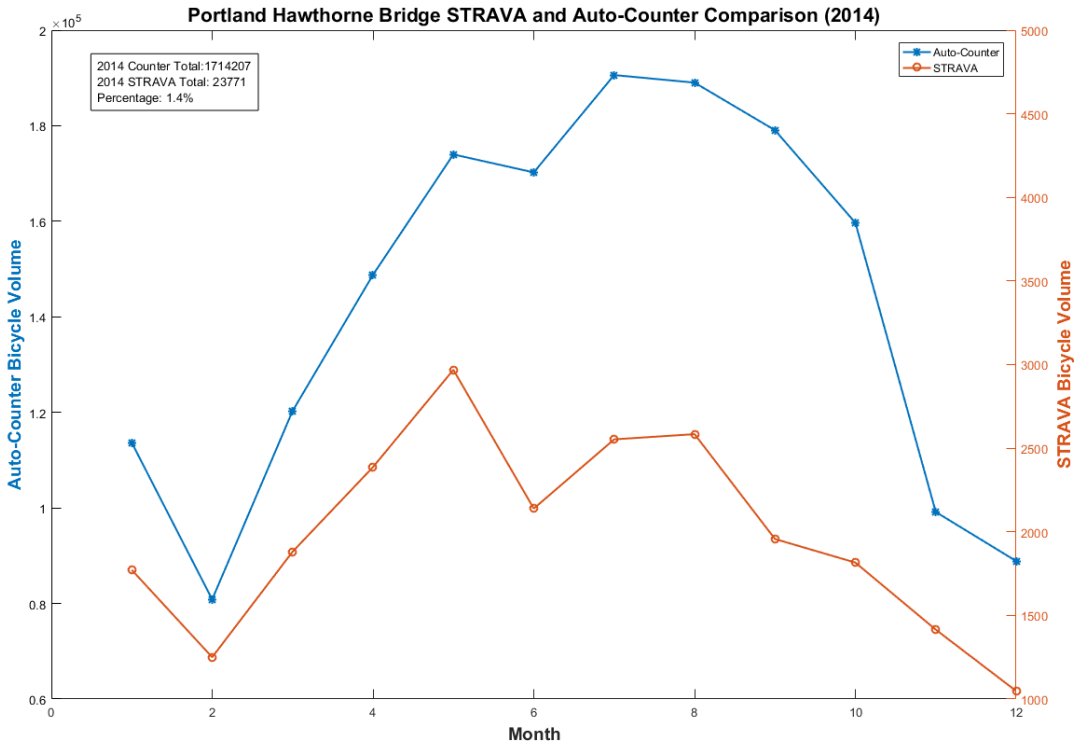

Figure 4-14 STRAVA count and Auto-counter volume

4.3.4 Under-Reporting Issue in Bicycle Crash Data

The reported bicycle crash data from Oregon is used in this study. We understood that this dataset is limited in certain aspects (i.e., underreporting issue). Under-reporting issue reduces the sample sizes which could affect the predictive power of the established SPFs. To understand the under-reporting issue in crash data in Oregon, the author has interviewed the staff members in Crash Analysis and Reporting Unit within ODOT. The important factors learned from the interview that may influence the under-report issue are summarized below (more details can be found in Appendix A):

- Oregon is a self-reporting state which means drivers must report; however, police are not required to investigate. This indicates it is possible that not all users would report the accident if the damage is low.

- Users must report only when Property Damage Only (PDO) accident with more than $1,500 damage to any vehicle, or there is an injury or death from the accident. Bicycle PDO crash often has damage lower than $1,500, which means a large portion of PDO accident have not been reported to DMV.
- Bicyclists, pedestrians, and owners of parked vehicles are not required to report, but information can be obtained from drivers involved in those accidents. This can also increase the under-reported PDO crashes for bicyclists.
- Bicycle/vehicle crashes require no medical transport or emergency response may cause under-reporting issue as well.

Thus, there could be a large part of PDO bicycle accidents which are not reported to DMV. This could cause the crash severity distribution skewed, especially lacking PDO crashes. However, the injury and fatal crash data would be still accurate.

4.4 Data Analyses

This section visualizes the independent and dependent variables at intersections and on corridors. Over-dispersion issue is discussed which could influence the decision to choose the proper statistical models for SPFs. Then correlation between variables are also investigated.

4.4.1 Data Visualization and distribution

Visualization of independent and dependent variables can help engineers: 1) understanding the features of data by descriptive statistics (shape, mean, range, variance, etc.); 2) choosing proper models to build SPFs. Instead of analyzing every single variable, the section focus on variables that are found significant in the modeling process including crash frequency, bicycle volume, traffic volume, site characteristics of selected intersections and corridors.

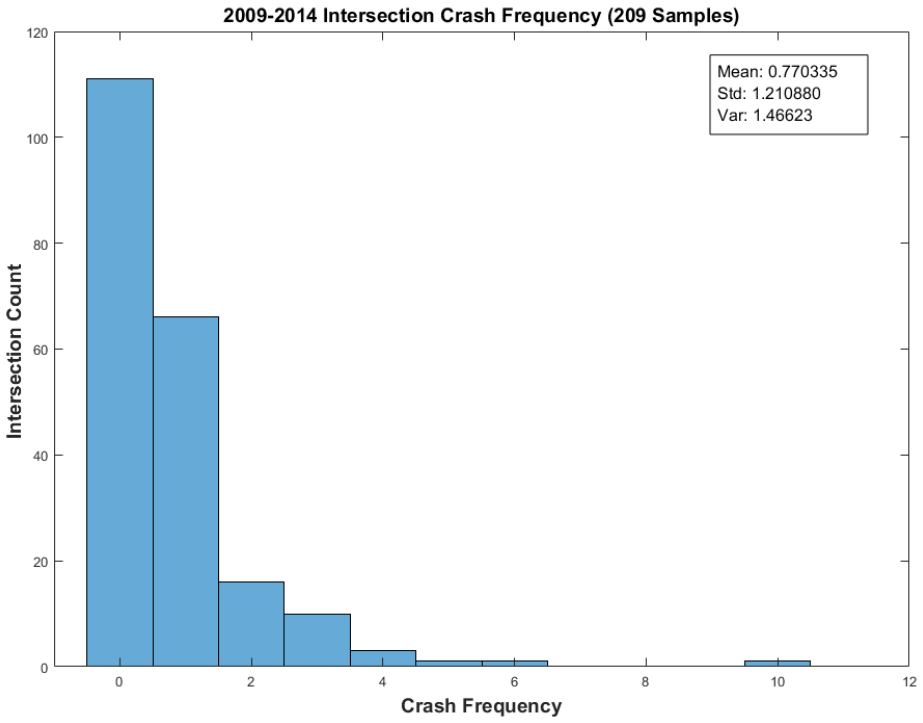

Figure 4-15 Intersection crash frequency histogram.

Figure 4-15 shows the crash frequency of intersection samples as a histogram skewed to the right, an average of 0.77, the standard deviation of 1.21, and the variance of 1.46. Figure 4-16, as follow, demonstrates the bicycle volume and traffic volume of both major and minor road of intersections. Total intersection bicycle volume and traffic volume are also demonstrated. The author found that the more crashes are likely to happen when there is a large bicycle volume on minor road. Larger bicycle volume on the minor road indicates more conflicts at an intersection when bicycles on the minor road are crossing or turning. Therefore, to capture this factor, the author created two new variables: 1) the ratio of minor bicycle volume to major road AADT; 2) the ratio of the minor road bicycle volume to major road bicycle volume. The histograms of those two factors are also shown in Figure 4-16.

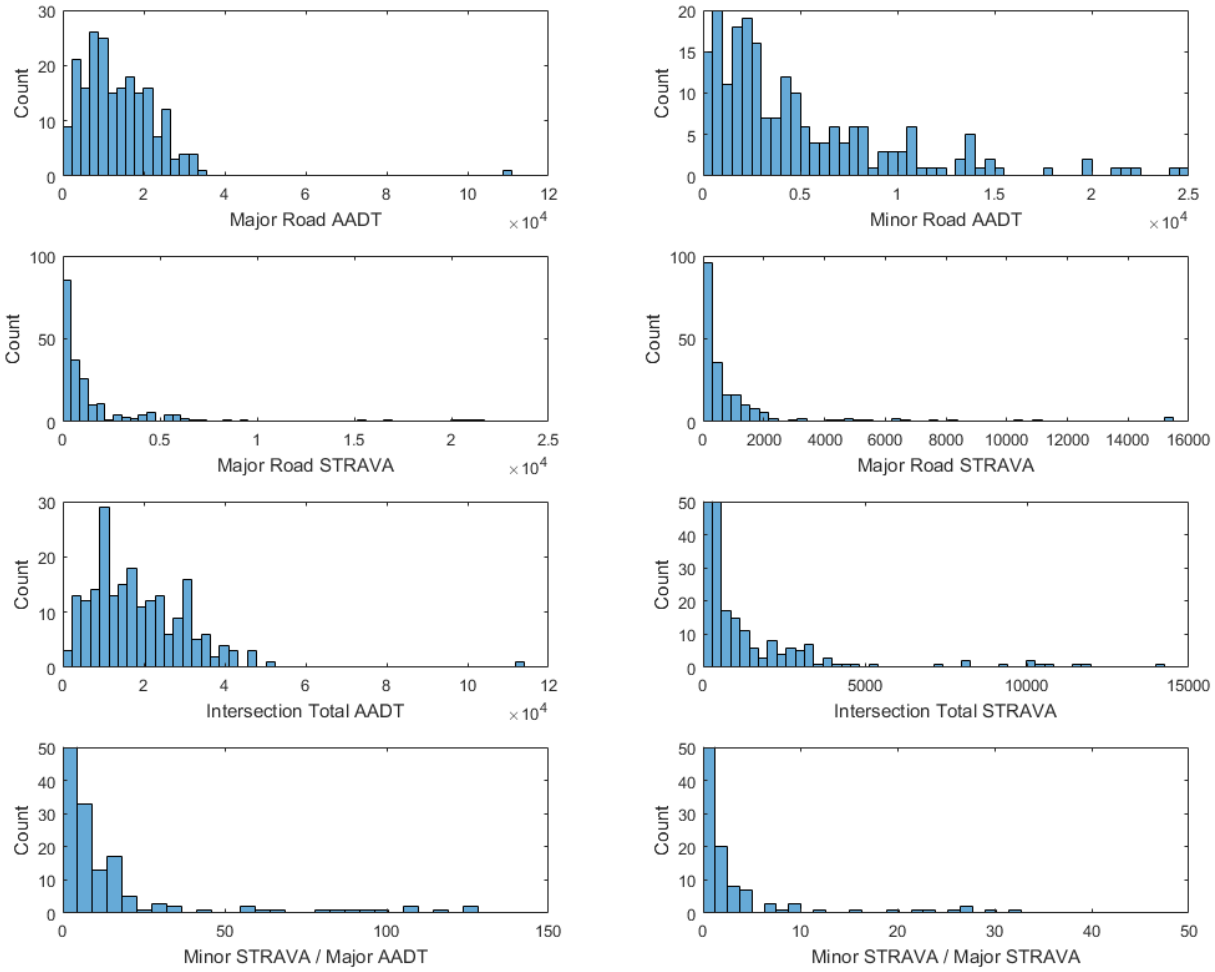

Figure 4-16 Bicycle and traffic volume at intersections as histograms.

Figure 4-17 summarizes other dependent variables that are found to be significant. Not all geometric, land-use, and road characteristic variables that collected are converted to histograms. For binary variables, 0 represents that there is no presence of the feature, whereas 1 represents the presence of the feature at intersections.

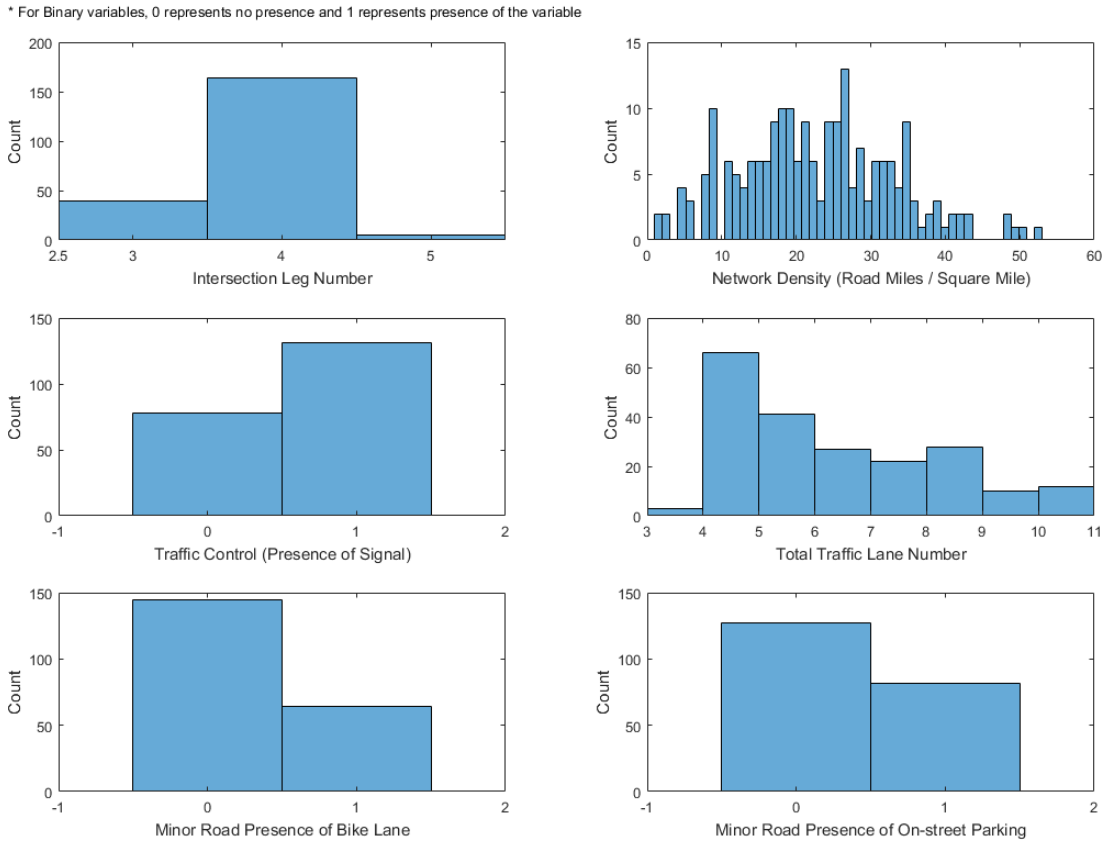

Figure 4-17 Histograms of partial intersection characteristic variables.

Figure 4-18 shows the histogram of crash frequency and crash number per mile on corridors. Even though the dependent variable is crash frequency, crash frequency per mile histogram is shown here to understand the crash rate. Crash frequency has the mean of 8.66, the standard deviation of 7.89 and variance of 62.31.

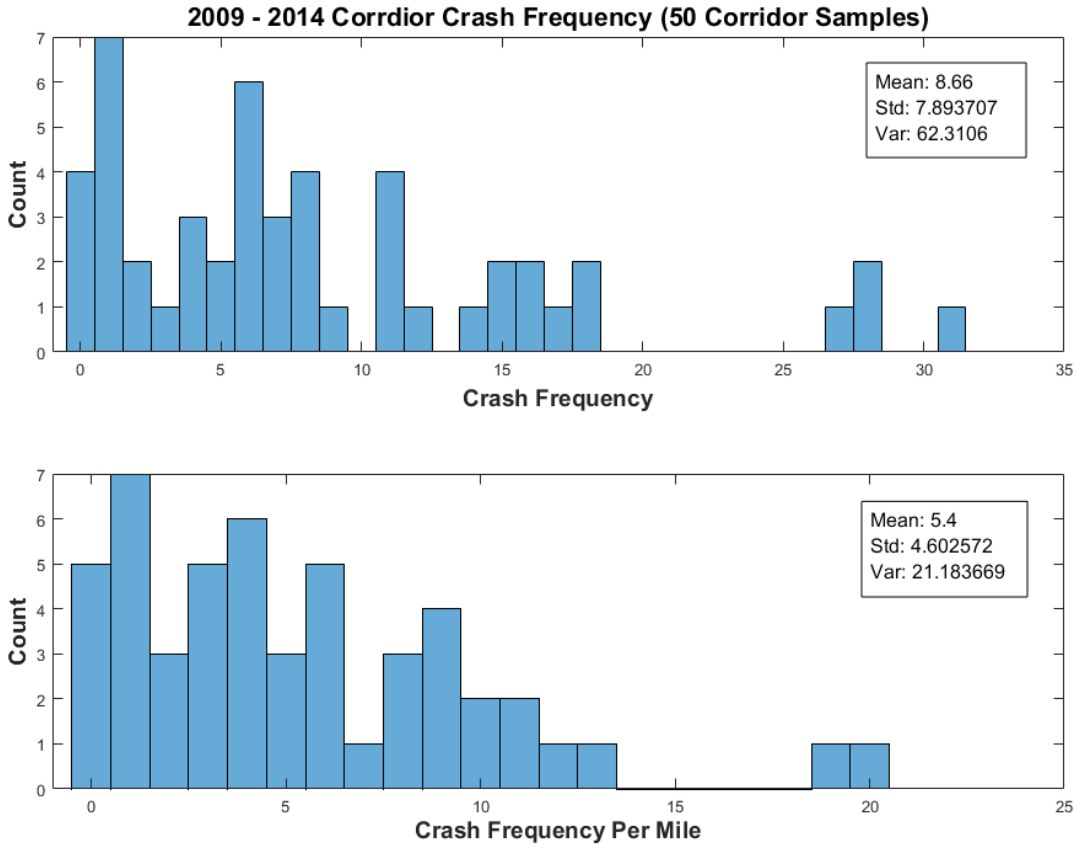

Figure 4-18 Histogram of corridor crash frequency.

Figure 4-19 illustrates the histogram of average bicycle and traffic volume on corridor, and Figure 4-20 shows the histogram of road characteristic variables that are significant in models.

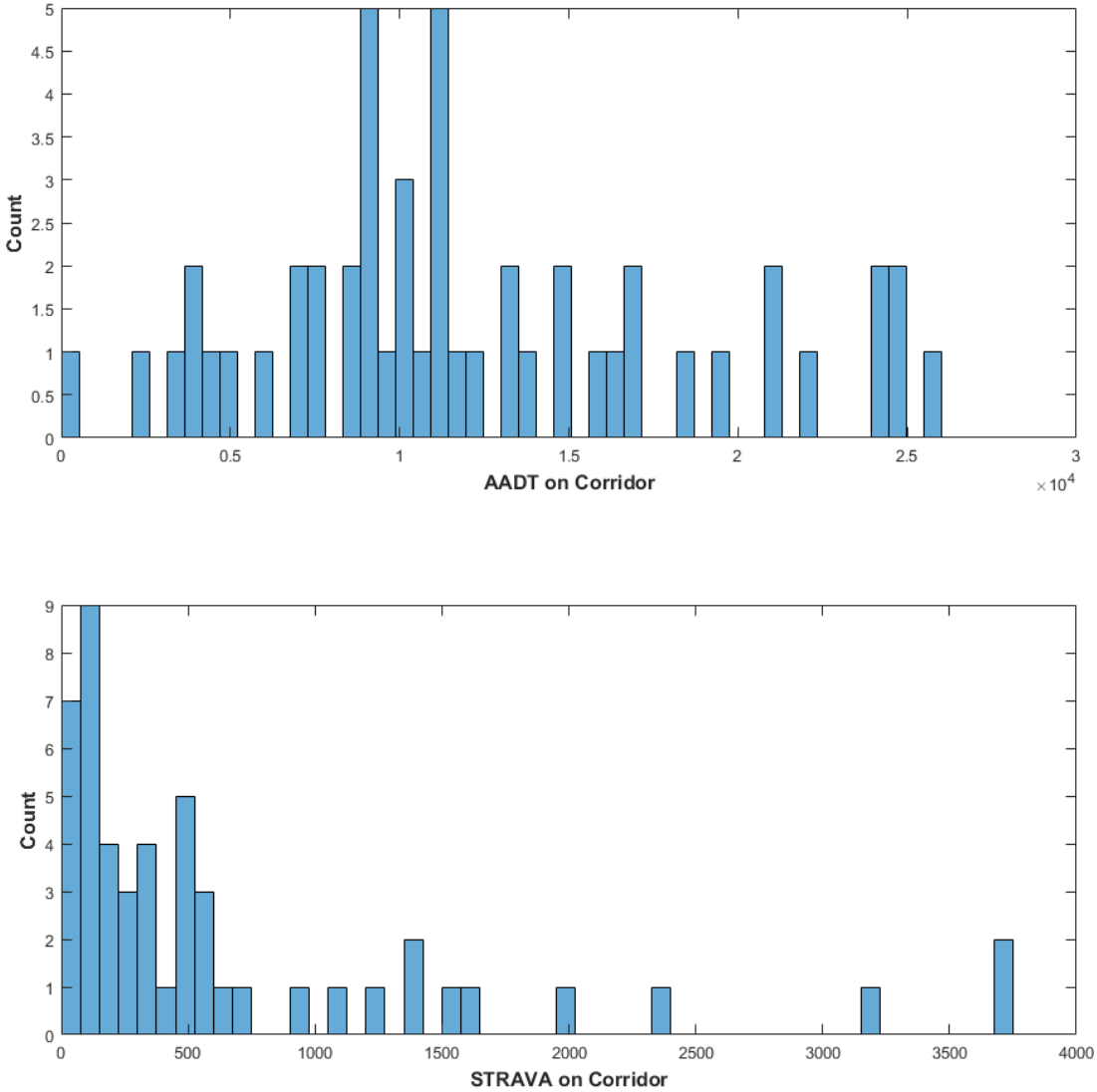

Figure 4-19 Histogram of bicycle and traffic volume on corridor samples.

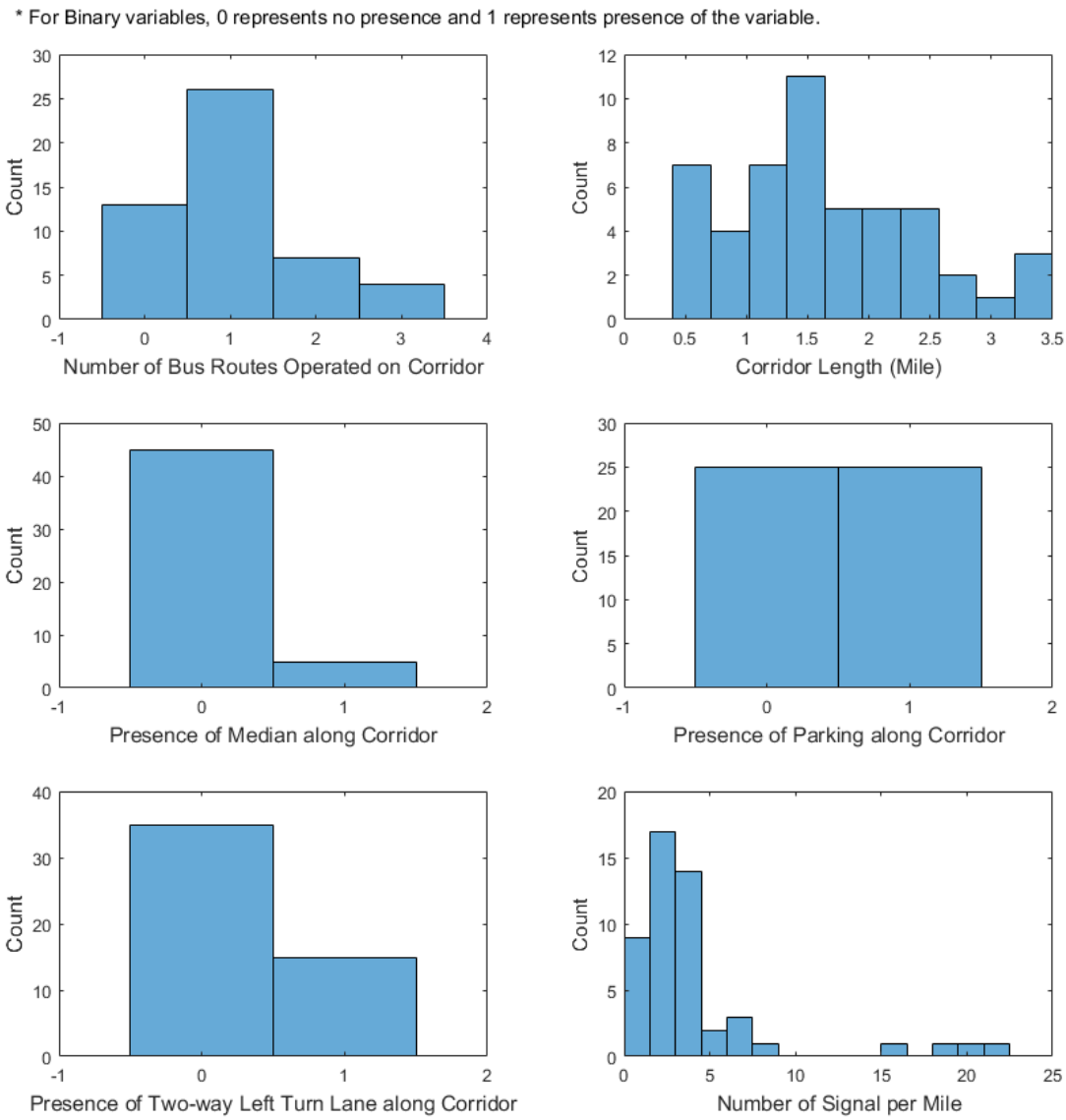

Figure 4-20 Histograms of partial corridor characteristic variables.

4.4.2 Dispersion of Dependent Variables

Researchers care about the dispersion type of dependent variable because of its influence on model choices. There are three different types of dispersion: equal-dispersion, under-dispersion and over-dispersion. Over-dispersion, meaning the mean equals to the variance, is the most common feature in crash count data. Over-dispersion can be investigated from three steps:

- Observing the data distribution;

- Calculating the mean and variance;
- Running an over-dispersion test.

Observing the data distribution, using histogram to plot crash count data, is the first step to identify the existence of over-dispersion. When a histogram is skewed to the right, or when there are a larger amount of zero counts, it is likely to be over-dispersed. Figure 4-15 and Figure 4-18 capture the shape of the crash count for intersections and corridors, respectively. A right-skewed shape is shown in the intersection histogram and there are many zero counts, which indicates that data may be over-dispersed. However, the range of crash count (0 to 10) is not large, which indicates the variance may not be significantly greater than the mean. For a histogram of corridor crash count, it shows the right-skewed shape and with a broad range of count (0 to 35), so the corridor data has high possibility to have over-dispersion feature.

Comparing the mean and variance is the second step to confirm the over-dispersion. In intersection crash data, shown in Figure 4-15, has a mean of 0.77 and variance of 1.46. The variance is larger than mean but not significantly. Therefore, the data may not be very over-dispersed. An over-dispersion test is needed to confirm existing of over-dispersion. Corridor data, shown in Figure 4-18, has the mean of 8.66 and variance of 62.31. Variance is eight times larger than mean which indicates the over-dispersion is evident. Over-dispersion test results are shown in chapter 5 to confirm the speculation.

4.4.3 Correlation between Variables

Correlation in statistic indicates whether and how strong one variable relates to another variable (Creative Research Systems, 2016). If one variable has a high correlation with another, this often means the two variables have either positive or negative relationships and only one of them should be included in the model. Therefore, it is critical to test the variables' correlations. Correlation matrix for intersection and corridor variables are created in MATLAB and are shown as follow.

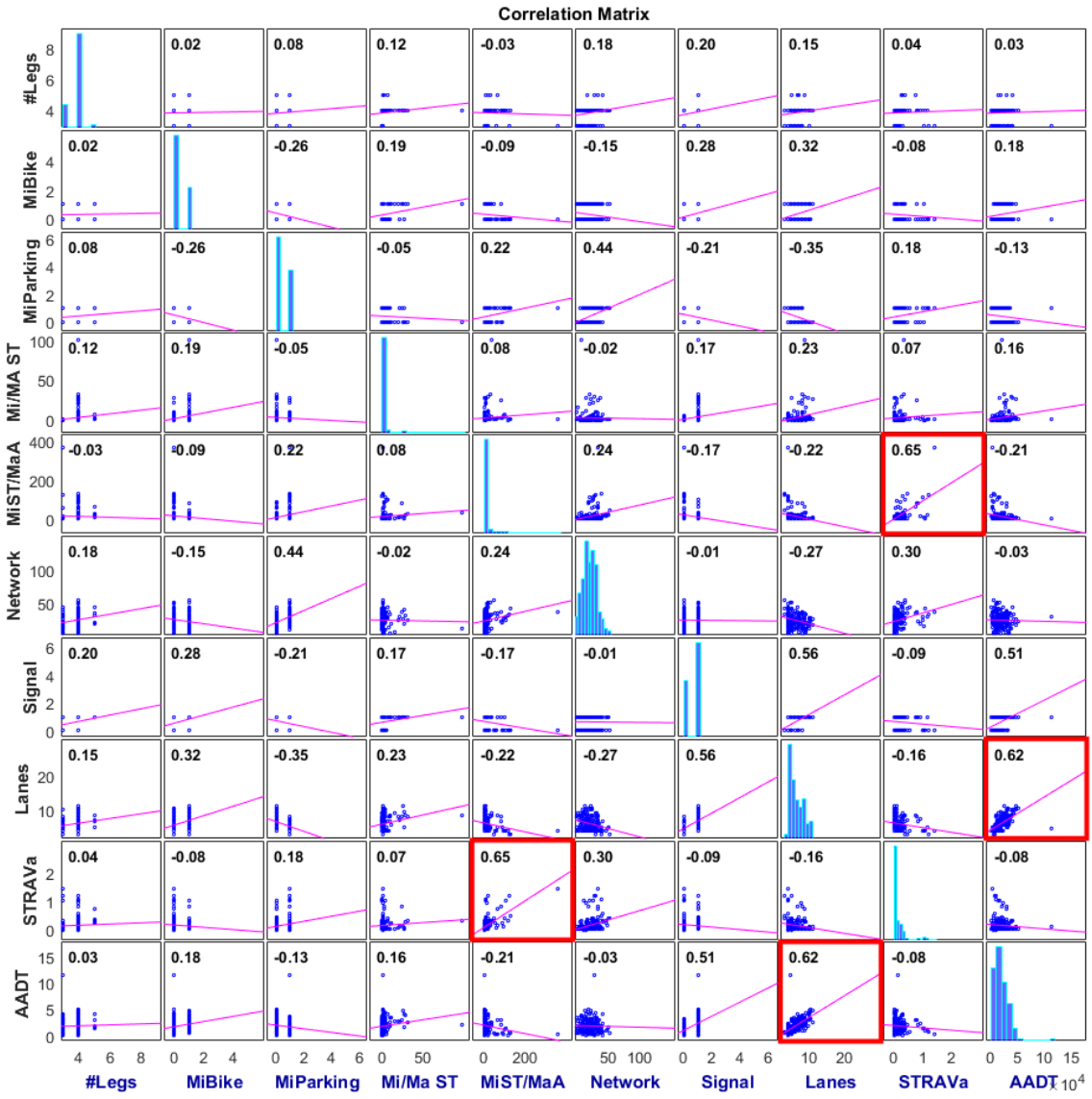

Figure 4-21 Intersection variable correlation matrix

Correlation between significant variables for intersection data are demonstrated in Figure 4-21 and the high correlation is highlighted (using 0.6 as the threshold). STRAVA® data has 0.65 correlation with the variable "minor road STRAVA/major road AADT." The author expected the correlation between those two variables because the later variable is created from the former. Total traffic lanes number has a positive correlation with AADT, which makes sense because the more lanes a road have, the more traffics will appear. However, since the purpose of this study is

building SPFs to predict crash frequency, the two correlations are acceptable. Generally speaking, the correlation matrix shows the variables included in the model are reasonable.

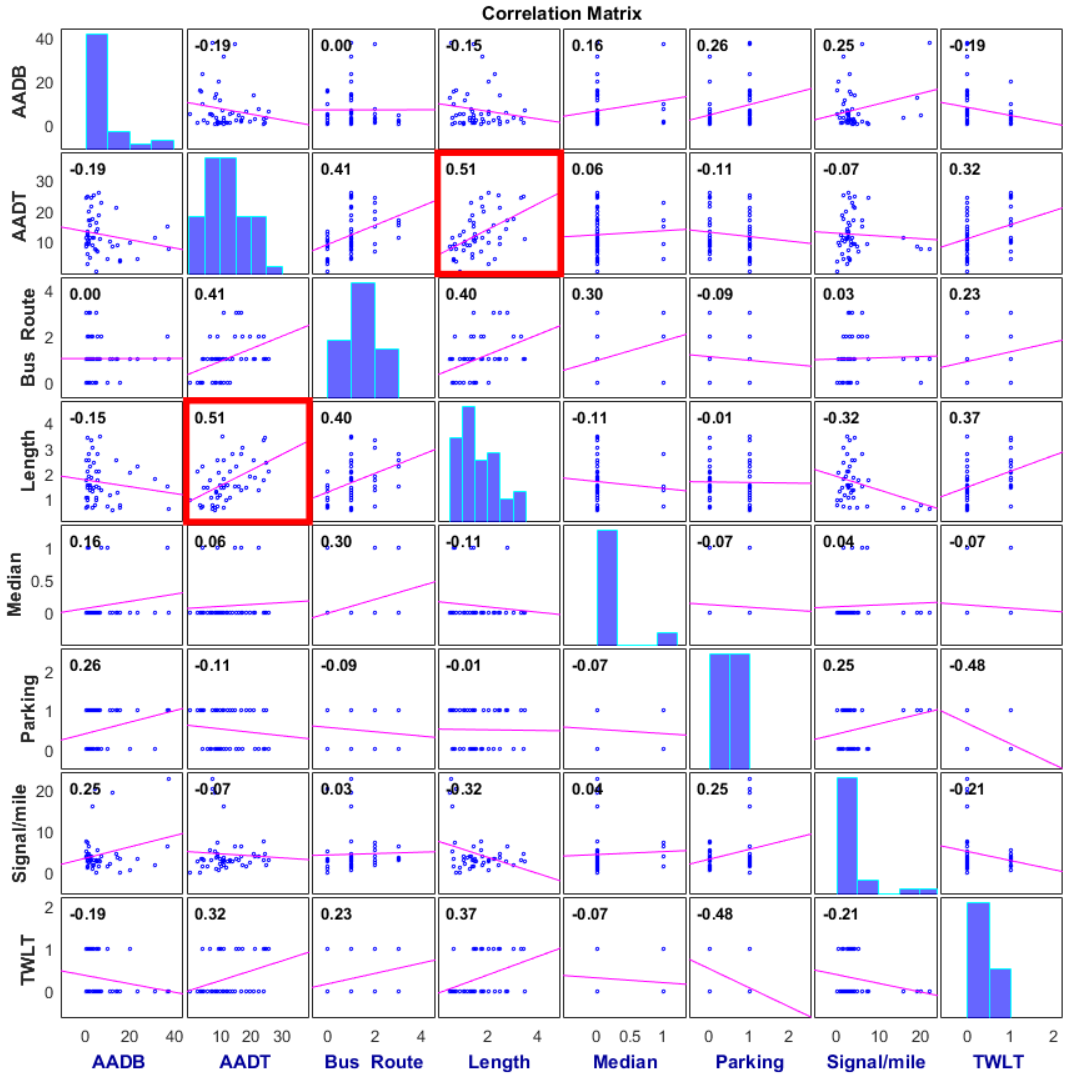

Figure 4-22 Correlation matrix of corridor variables.

Similar with correlations in intersection variables, the correlation of corridor variables justifies those significant variables are reasonable. AADT has 0.51 correlation with corridor length. The possible reason is that roads with higher function classification usually have larger AADT, and engineers tend to keep those major road characteristics consistent. Again, the purpose of this study is to create SPFs to predict crash frequency, so the correlations are acceptable.

5. Results and Discussion

The goal of this chapter is to present the results of this study and establish a repeatable process of building bicycle SPFs for transportation agencies. Specifically, this chapter discusses the results from modeling process including crash frequency predictive model for microscopic scale (intersections) and macroscopic scale (corridors). Crash severity distribution is created based on crashes happened at sample sites and then used to combine with predictive crash frequency results to estimate crash severity. The chapter then will summarize the established SPFs and the process of building bicycle SPFs.

5.1 Microscopic Model: Intersection Crash Frequency

Poisson, NB, ZIP and ZINB models for intersections are all created and run in NLOGIT®. Over-dispersion parameter and Vuong statistic are used to determine the best model for the data set. The justification of best-fitted significant variables is through the Likelihood Ratio Test (LRT) and the model fitness is measured by the McFadden Pseudo R-squared. The sign of coefficients and partial effects are used to discuss the impacts of explanatory variables on response variable – crash frequency.

Previous results (Nordback et al., 2014) show that the increase in bicycle volume (estimated AADB) is associated with the increase in bicycle crashes but decrease bicycle crash rate. To justify this phenomenon, the study did similar analyses shown in Figure 5-1 and Figure 5-2. As shown in Figure 5-1, when the bicycle volume increases, so does the intersection bicycle crash number; however, the crash rate tells a different story. As shown in Figure 5-2, Poisson regressions were fitted for different levels of AADT and different STRAVA® counts. Indeed, the results show that the increase of STRAVA® bicycle count can decrease the crash rate. This phenomenon is called “Safety in Number.” The reasonable explanations of this phenomenon are that the increasing bicycle volume may lead to safer behavior of motorists and bicyclists, or more bicyclist riding on safer facilities (Nordback et al., 2014).

However, our results show the impact of AADT change on crash rate is different from Nordback’s result. In the “High Risk” zone (STRAVA® <2000), the increase of

AADT can increase bicycle crash rate; in the “Moderate Risk” and “Low Risk” zones (STRAVA® >2000), the growth of AADT shows an opposite impact. A possible explanation is that when AADT is very high, the average traffic speed is low (due to congestion), and drivers can stop vehicles easier to avoid accidents.

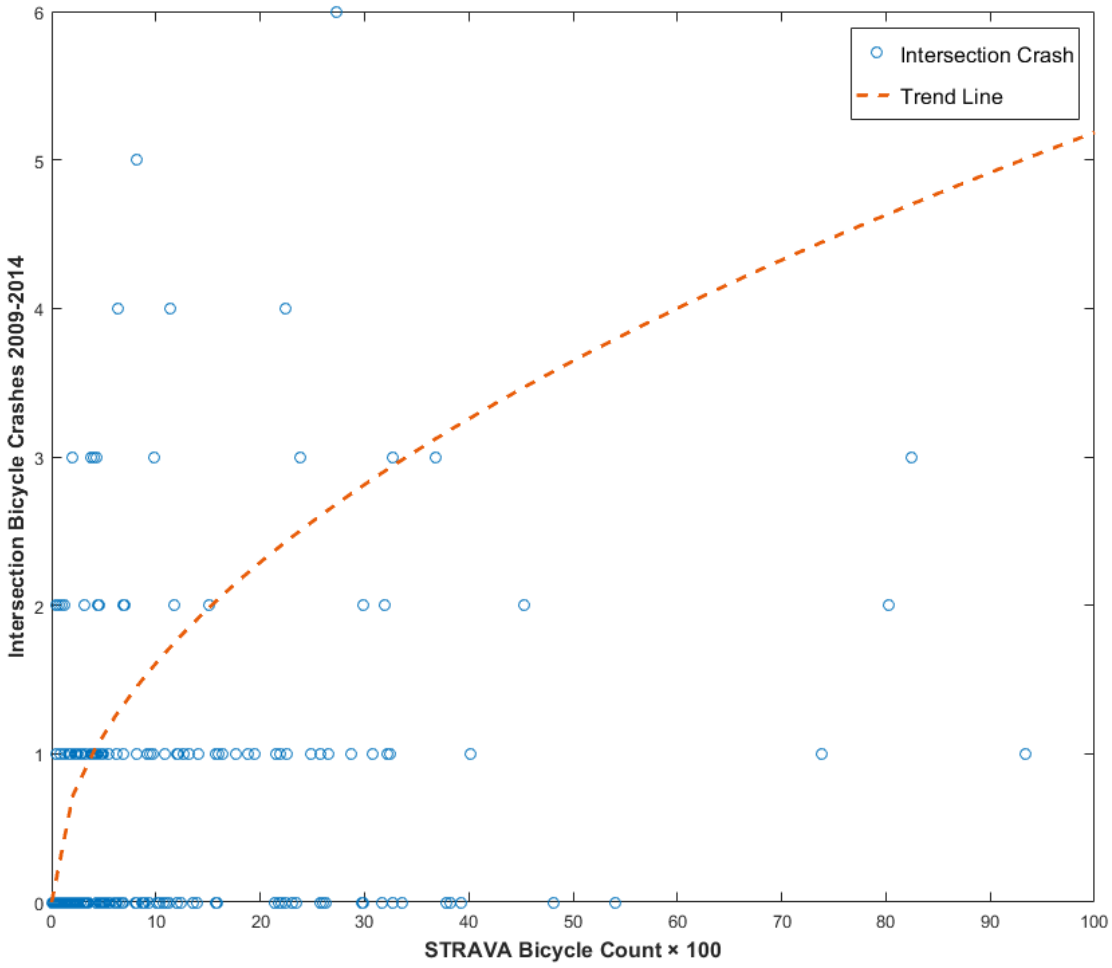

Figure 5-1 The relation between intersection bicycle crash count and STRAVA bicycle count

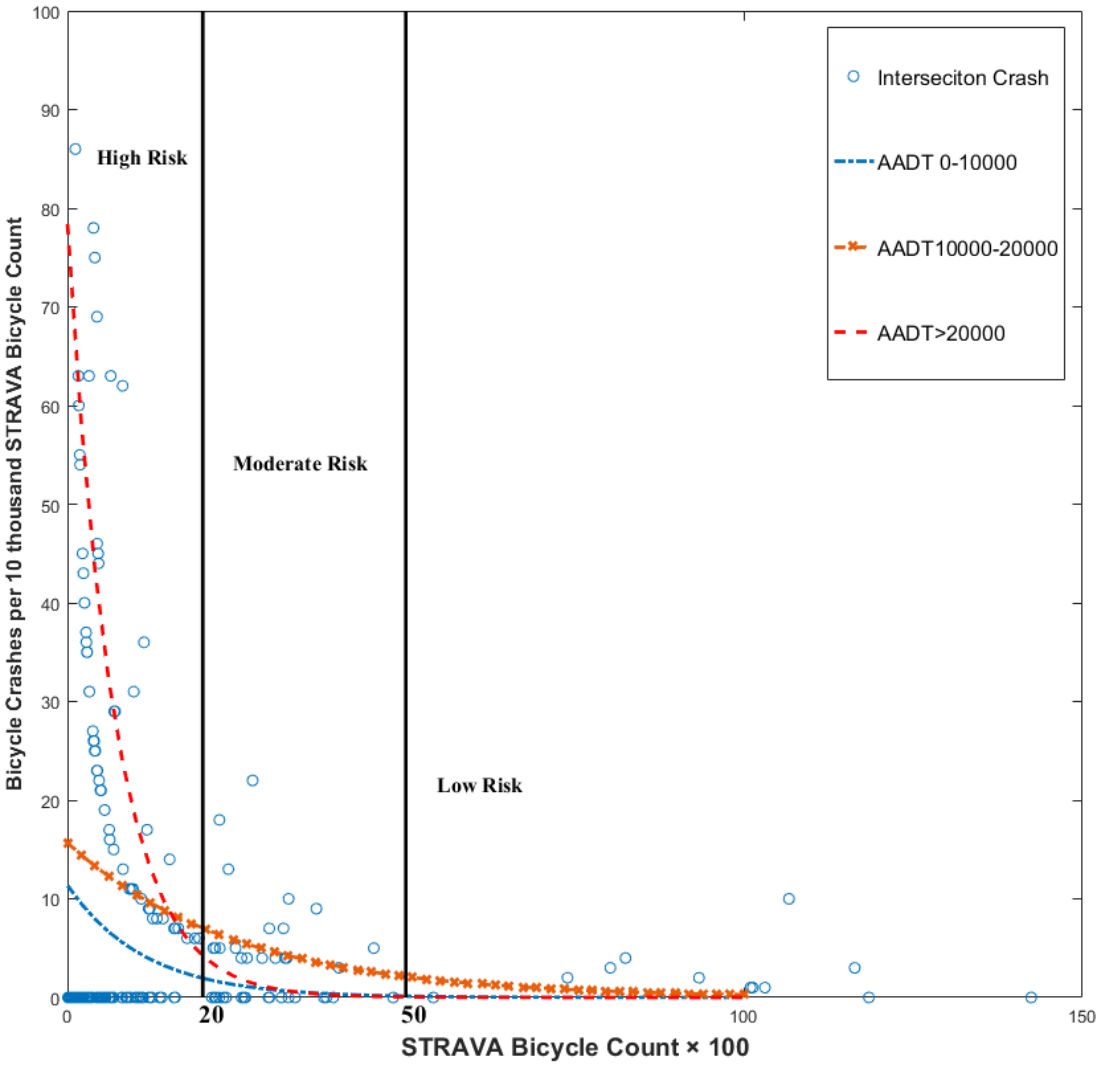

Figure 5-2 Poisson regression: more bicycle volume can increase bicycle crashes but decrease crash rate.

Table 5-1 Intersection Crash Frequency Poisson Model Results.

Poisson Regression

Dependent variable Crash Frequency 2009-2014 N=209
Log likelihood function -206.51824
Restricted log likelihood -268.02298
Chi squared 123.00948 Significant Level 0.00000
McFadden Pseudo R-squared .2294756

Variables	Coefficient	z	Prob. |z|>Z*	95% Confidence Interval		Partial Effect
Constant	-7.17456***	-5.54	0.0000	-9.71095	-4.63818	-
Total STRAVA (100)	.02351***	7.87	0.0000	0.01765	0.02937	0.01811
Total AADT(K)	.01269*	1.65	0.0999	-0.00243	0.02781	0.00978
Network Density	.02792***	3.01	0.0026	0.00972	0.04611	0.0215
Minor Directions(1:two-way; 0:one-way)	1.59389***	2.88	0.0039	0.51048	2.67729	0.67368
Minor Bike Lane(1: bike lane; 0:no Bike Lane)	.56335***	3.15	0.0016	0.21283	0.91388	0.4721
Minor STRAVA/Major AADT	-.01559***	-3.2	0.0014	-0.02513	-0.00605	-0.01201
Total Lanes	.13755**	2.42	0.0156	0.02604	0.24906	0.10596
Signal(1:signal; 0:non-signal)	.50249**	1.96	0.0496	0.00092	1.00406	0.33634
Leg Number	.68367***	2.72	0.0065	0.19118	1.17616	0.52665

Partial effect for Binary variable is E[y| x, d=1] - E[y| x, d=0]
Note: ***, **, * represent Significance at 1%, 5%, 10% level.

When other variables are added to the model, the scenario becomes complicated. A 2-D figure cannot capture the relationships anymore. Thus, Poisson model results for intersection crash frequency are summarized in Table 5-1. LRT with a significant level of <0.0000 indicates there is significant difference between this fitted model and the model only with constant. McFadden Pseudo R-squared of 0.229 indicates the model is greatly fitted, according to Domencich and McFadden (1975). Those two tests, in other words, suggest the model performs very well regarding the explanation of the crash data set.

Variable "Total STRAVA (100)" with a partial effect of 0.01811 and a positive coefficient indicates an increase in every 100 STRAVA count will increase the intersection crash count by 0.01811 on average. Similarly, variable "Total AADT (K)" with a partial effect of 0.00978 means an increase in every 1000 AADT will result in the intersection crash count increasing by 0.00978 on average. Nordback, Marshall and Janson (2014) found similar results that the increase of AADT and AADB can cause crash count to increase, but crash rate decreases. It should be noted that even though the impact seems small, such as 0.01811 crashes for every 100

STRAVA® count change, but the difference of STRAVA® count can be several thousand in some intersections, and same as AADT.

The author found an interesting phenomenon while collecting data: there is a trend that fewer bicycle crashes occurred when bicycle STRAVA® volume on the minor road increase. In order to justify this finding, the author defined a new variable for the model by using STRAVA® on the minor street divided by traffic volume on the major road and it is called "Minor STRAVA/Major AADT." Indeed, the model results justified the phenomenon that "Minor STRAVA/Major AADT" has a negative impact on crash frequency at intersections. With every unit increase in the ratio, the crash frequency decreases by 0.01201. This result makes logical sense, because the more bicycle volume on the minor roads, drivers on the major roads could be increasingly aware of the crossing bicyclists from the minor roads. In other words, drivers expect more bicycles from the minor roads while driving and they would operate carefully to avoid collision with bicyclists. To the author's best knowledge, this finding is the first time to be noticed and can be a new guide for engineering design and policy making, which will be discussed in chapter 6.

"Network Density" with a partial effect of 0.0215 suggests that for every 1 mile/square mile increasing in Network Density, the crash frequency in intersections would increase by 0.0215 on average. That may result from higher traffic and bicycle volume when the road density increases.

Results from variable "Minor Directions" represent that the two-way direction can cause 0.67368 more crashes than the one-way direction on a minor road. One possible explanation is that there are more conflicts when there are more directions of traffic, especially bicycles conflicting with left turning vehicles, shown in Figure 5-3. According to FHWA, this type of accident results in 5.9% of the total crash and 24% of severe injuries and fatalities. Other studies also discovered the turning movement could increase the conflicts between bicycle and motor vehicle. According to Hurwitz, et al. (2015), the right-hook crashes represented over 500 of reported crashes related to bicycles from 2007 to 2011 in Oregon. Warner et al. (2017) states that one of the more prevalent bicycle-motor vehicle crash type at intersections is the

right-hook crash which is caused by the conflicts between right turning vehicle and through bicyclists. Warner et al. (2017) analyzed the effectiveness of engineering treatments, such as signage and pavement marking, and found that those engineering treatments have positive effects on improving safety. The decrease of conflicts between bicycles and vehicles can improve bicycle safety. Therefore, this thesis recommended low-cost engineering treatments to mitigate the conflicts in chapter 6.

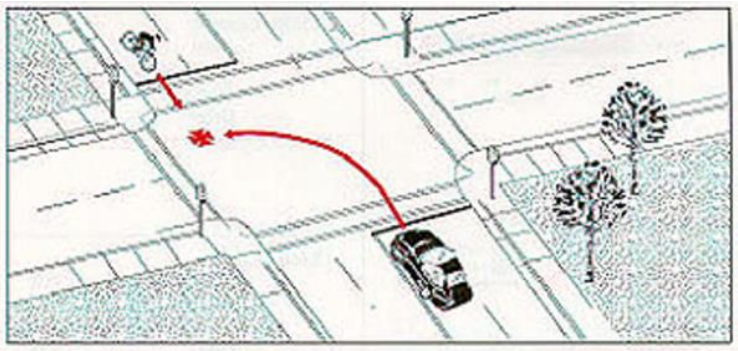

Figure 5-3 Bicyclists conflicts with upcoming left turning traffic (Federal Highway Administration, 2003).

Another interesting finding is from variable "Minor Road Presence of Bike Lane". The result indicates that the presence of bike lane on the minor road can increase the crash number at intersections by 0.4721. This finding violates the common understanding that bicycle lane is safer on bike lanes. However, the bike lane is designed for protecting the bicyclists from the high volume roads. In other words, bike lane normally appears on the major arterials or collectors with a higher traffic volume. Furthermore, the presence of bike lanes can attract more bicyclists to that road which will increase the bicycle volume and crashes. Therefore, it is reasonable to observe more crashes on roads with bicycle lanes. Indeed, other researchers also found the association between bicycle crash and bike lane (Wei and Lovegrove, 2013; Dolatsara, 2014). This finding can be used to guide engineering design and policy making, and more details can be found in chapter 6.

Total number of lanes, the presence of traffic signal, and more leg number at intersections all have positive impacts on bicycle crash frequency. The possible reason is that they are all associated with more conflicts between vehicles and

bicycles. Total number of lanes and the presence of signal can be interpreted together for engineers because the road with more lanes typically has a higher functional classification and more signal traffic controls. For leg number, this result is reasonable as well. Intersection with three legs has fewer conflicts between bicycles and traffic because bicycles on minor roads would not do crossing movements, shown in Figure 5-4.

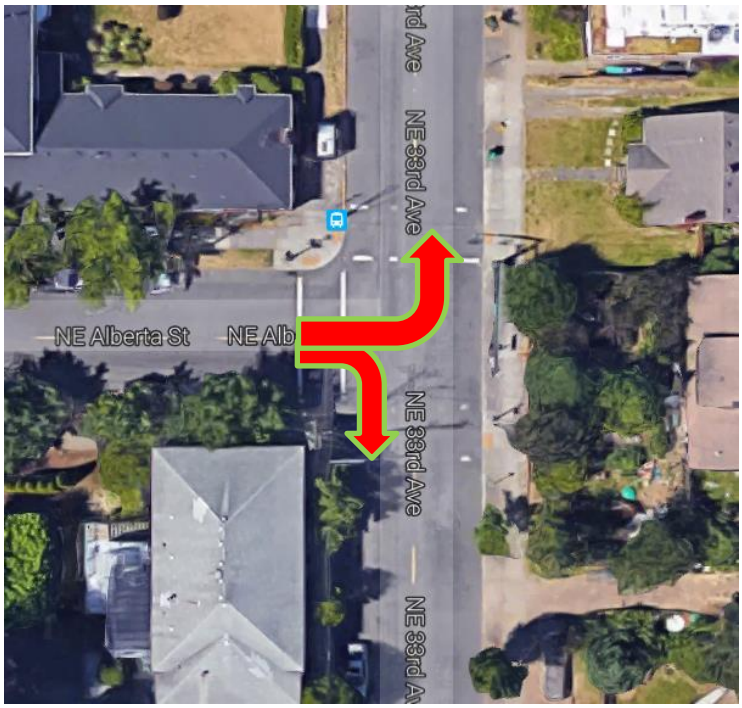

Figure 5-4 Bicycle movement are three-leg intersection.

The next step is to compare the Poisson model with NB mode by using the over-dispersion parameter to identify which one is preferred. Table 5-2 shows the results of NB model including over-dispersion test, LRT, and McFadden Pseudo R-squared. Dispersion parameter is not significant which indicates the over-dispersion is not significant in the data, and NB is rejected in favor of Poisson model. This result justifies the finding in data analyses that crash data from intersections may not be over-dispersed though the shape seems over-dispersed. Comparing LRT and McFadden Pseudo R-squared of NB and that of Poisson model in Table 5-1, it suggests that the Poisson model fits and predicts better than NB model. Thus, the Poisson is chosen between NB model and Poisson model.

Table 5-2, NB results of intersection crash frequency

Negative Binomial Regression

Dependent variable Crash Frequency 2009-2014 N = 209
Log likelihood function -213.78906
Restricted log likelihood -216.79074
Chi squared 6.00337 Significance level .01428
McFadden Pseudo R-squared .0138460

Variable	Coefficient	z	Prob.|z|>Z*	95% Confidence Interval	
Constant	-2.78070***	-5.93	0.000	-3.69944	-1.86196
Total STRAVA (100)	.02624***	9.3	0.000	0.02071	0.03177
Minor Bike Lane	.40183*	1.78	0.0751	-0.04063	0.84429
Minor STRAVA/Major AADT	-.01556**	-2.42	0.0154	-0.02813	-0.00298
Total Lanes	.19068***	2.81	0.005	0.05749	0.32387
Signal(1:signal; 0:non-signal)	.86770***	2.78	0.0055	0.25497	1.48042
Minor Parking(1: parking; 0:no parking)	.39993*	1.71	0.0868	-0.05781	0.85768
Dispersion parameter for count data model					
Alpha α	0.21551	1.45	0.1468	-0.07564	0.50666

Partial effect for Binary variable is E[y| x, d=1] - E[y| x, d=0]
Note: ***, **, * represent Significance at 1%, 5%, 10% level.

Since Poisson model is better than NB model in this case, then Vuong-statistic is used to compare between Poisson and ZIP models. Vuong-statistic result shows the Vuong-statistic of -0.5149 which indicates the ZIP model is rejected in favor of Poisson model. The result table is not reported here because the model is rejected, but reported in **Appendix B**. Therefore, Poisson model is chosen reasonably to establish the SPFs for intersections.

5.2 Macroscopic Model: Corridor Crash Frequency

Similar to the development of the intersection crash frequency model, The Poisson, ZIP, NB and ZINB models are compared to identify which model is suitable for corridor crash frequency model. Poisson model and NB model results for corridor crash were created and summarized in Table 5-3 and Table 5-4, respectively. Dispersion parameter is significant which indicates that NB model is preferred. Another clue to reject Poisson is the McFadden Pseudo R-squared value of 0.44 in Poisson model. This value between 0.2 and 0.4 means the model is excellent fitted, and anything above 0.4 may involve unobserved issue (Domencich and McFadden, 1975). The underserved issue could be over-dispersion. Vuong-statistic is then

applied to compare ZINB and NB models. Vuong-statistic shows a result of 1.0696 which is under the 90% significant level. Therefore, NB is preferred than ZINB model.

Table 5-3 Poisson model for corridor crash frequency.

Poisson Regression

Dependent variable Crash Frequency 2009-2014 N=50
Log likelihood function -142.25440
Restricted log likelihood -257.89010
Chi squared 231.27139 Significance level .00000
McFadden Pseudo R-squared .4483914

Variables	Coefficient	z	Prob.|z|>Z*	95% Confidence Interval	
Constant|	0.14516	0.81	0.4182	-0.20626	0.49657
Length	.81447***	11.46	0	0.67512	0.95382
Signal/Mile	.09235***	7.94	0	0.06956	0.11515
Median	-.49667**	-2.46	0.014	-0.89292	-0.10043
two-way Left Turn Lane(1:TWLT; 0:no TWLT)	-.45303***	-3.1	0.002	-0.73986	-0.16619
Bus Route Number	.37679***	5.26	0	0.23627	0.5173
ON-street Parking (1:parking; 0:no parking)	-.42164***	-3.47	0.0005	-0.65983	-0.18346

Partial effect for Binary variable is E[y| x, d=1] - E[y| x, d=0]
Note: ***, **, * represent Significance at 1%, 5%, 10% level.

Table 5-4 NB model for corridor crash frequency.

Negative Binomial Regression

Dependent variable Crash Frequency 2009-2014 N=50
Log likelihood function -133.58044
Restricted log likelihood -142.25440
Chi squared 17.34793 Significance level .00003
McFadden Pseudo R-squared .0609750

Variables	Coefficient	z	Prob.|z|>Z*	95% Confidence Interval		Partial Effect
Constant|	0.09144	0.37	0.7128	-0.39551	0.57839	-
Length	.84148***	5.28	0	0.52919	1.15378	7.38959
Signal/Mile	.09847***	4.48	0	0.05535	0.14158	0.86472
Median	-.76336*	-1.86	0.0632	-1.5688	0.04207	-5.16692
two-way Left Turn Lane(1:TWLT; 0:no TWLT)	-.66950***	-2.62	0.0088	-1.1701	-0.1689	-5.76652
Bus Route Number	.45500***	2.89	0.0039	0.14611	0.76388	3.99561
ON-street Parking (1:parking; 0:no parking)	-.47638**	-2.01	0.0441	-0.94028	-0.01248	-4.31573
Dispersion parameter for count data model						
Alpha	.15458*	1.86	0.0633	-0.00858	0.31773	

Partial effect for Binary variable is E[y| x, d=1] - E[y| x, d=0]
Note: ***, **, * represent Significance at 1%, 5%, 10% level.

Variable “Length” with a partial effect of 7.38959 indicates with one mile increasein corridor length, the number of crashes for six years increases by 7.40, holding other variables constant. This result is easy to understand because the longer the corridor is, the more crashes it is likely to have.

Signal per mile is an interesting variable that has a positive impact on corridor bicycle crash frequency. With one signalized intersection increase per mile on a corridor, the crash frequency increase by 0.86472 on average. This finding is consistent with the finding that more crashes are likely to happen at intersections (in Table 5-1) and also justified by other studies (Carter et al., 2006; Wei and Lovegrove, 2013; Chen, 2015).

The presence of median and the presence of two-way left turn lane on corridor both can decrease the crash frequency. They can be interpreted using the same logic: the presence of median provides a barrier to isolate the traffic from two different directions, which not only decreases the conflicts between bicycles and vehicles, but also prevents bicycles from illegal crossing the street; presence of two-way left turn lane provide more space for bicycles and vehicles to avoid collision.

With one more bus route operating on a corridor, there are on average 3.99561 more crashes occurring on the corridor for 6 years. The possible reason is that the bus stop along the road by occupying the bike lane, which could force bicycles to use the traffic lanes instead of the bike lane to pass the stopped bus. Wei and Lovegrove (2013) and Teschke et al. (2012) also discovered that more bus stops are associated with more bicycle crashes.

Interestingly, the presence of on-street parking is found to have a negative impact on crash number. The possible explanation is that on-street parking is typically present on a low functional class road, such as a local road, where there are less vehicular traffic and bicycles. However, this phenomenon can also result from the insufficient sample size (50 corridors). Generally, the more sites are collected the more accurate the model result will be. Future work needs to justify this finding.

5.3 Crash Severity Distribution

Crash severity distribution (percentage) is used to predict the number of different crash severities (PDO, Injury, and Fatal). The crash severity percentages were calculated based on the crashes happened at intersections and on corridors that collected in samples. The results are summarized in Figure 5-5 that shows the corridor severity distribution keeps consistent with intersection crash severity. Therefore, the combination of both is applied to simply the prediction process.

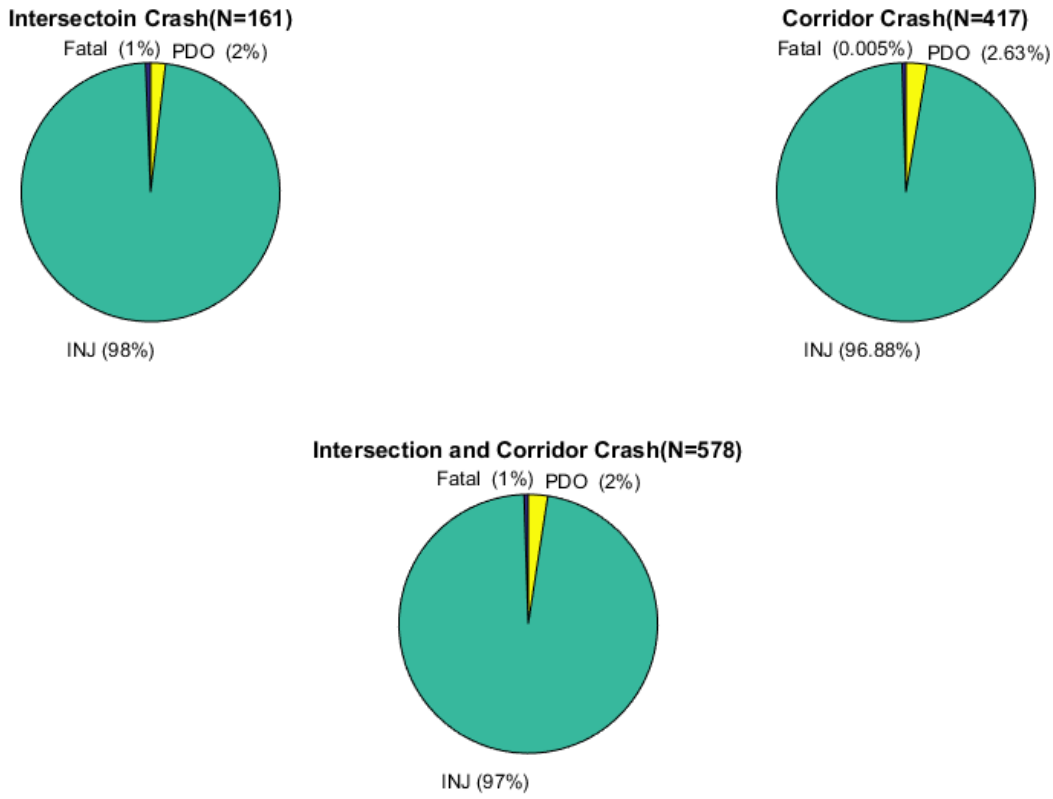

Figure 5-5 Crash severity distribution for intersections and corridors.

The percentage of PDO crash – 2% of all crashes – indicates under-report issue may exist in the data set, which has been explained in section 4.3.4. Therefore, the distribution could underestimate the number of actual PDO crash in reality. However, the results can be interpreted to infer “reported” crashes instead of actual bicycle crashes. We could also assume the damage is low and could be ignored, if an accident is not reported. Thus, the distribution is still useful in this case.

5.4 SPFs Summary

Since the objective is predicting the crash frequency for intersections and corridors, it is more important to the use the predictive functions. Therefore, SPFs for intersections from this study is created as in Equation 5.4.1, and SPFs for the corridors is established in Equation 5.4.2.

Micro Scale SPFs for Intersections

$$\boldsymbol{Intersection\ Crash_{total}} = EXP(-7.17456 + 0.02351X_1 + 0.01269X_2 + 0.02792X_3 + 1.59389X_4 + 0.56335X_5 + (-0.01559)X_6 + 0.13755X_7 + 0.50249X_8 + 0.68367X_9)$$

$$\boldsymbol{Crash_{PDO}} = 0.02 \times Intersection\ Crash_{total}$$

$$\boldsymbol{Crash_{INJ}} = 0.97 \times Intersection\ Crash_{total}$$

$$\boldsymbol{Crash_{Fatal}} = 0.01 \times Intersection\ Crash_{total} \quad (5.4.1)$$

Where:

X_1	Total STRAVA (100)
X_2	Total AADT(K)
X_3	Network Density
X_4	Minor Directions(1:two-way; 0:one-way)
X_5	Minor Bike Lane(1: bike lane; 0:no Bike Lane)
X_6	Minor STRAVA/Major AADT
X_7	Total Lanes
X_8	Signal(1:signal; 0:non-signal)
X_9	Leg Number

Macro Scale SPFs for Corridors

$$\boldsymbol{Corridor\ Crash_{total}} = EXP(0.09144 + 0.84148\ X_1 + 0.09847X_2 + (-0.76336)X_3 + (-0.66950)X_4 + 0.455X_5 + (-0.47638)X_6)$$

$$\boldsymbol{Crash_{PDO}} = 0.02 \times Corridor\ Crash_{total}$$

$$\boldsymbol{Crash_{INJ}} = 0.97 \times Corridor\ Crash_{total}$$

$$\boldsymbol{Crash_{Fatal}} = 0.01 \times Corridor\ Crash_{total} \quad (5.4.2)$$

Where:

X_1	Length
X_1	Signal/Mile
X_1	Median
X_1	Two-way Left Turn Lane(1:TWLT; 0:no TWLT)
X_1	Bus Route Number
X_1	On-street Parking (1:parking; 0:no parking)
ε_i	Can be calculated by a Gamma-distribution with mean 1 and variance α.The over-dispersion parameterα, in this model, is 0.15458.

To simplify the difference of microscopic and macroscopic SPFs, the process of using those two SPFs to diagnose high-risk sites are summarized in Figure 5-6.

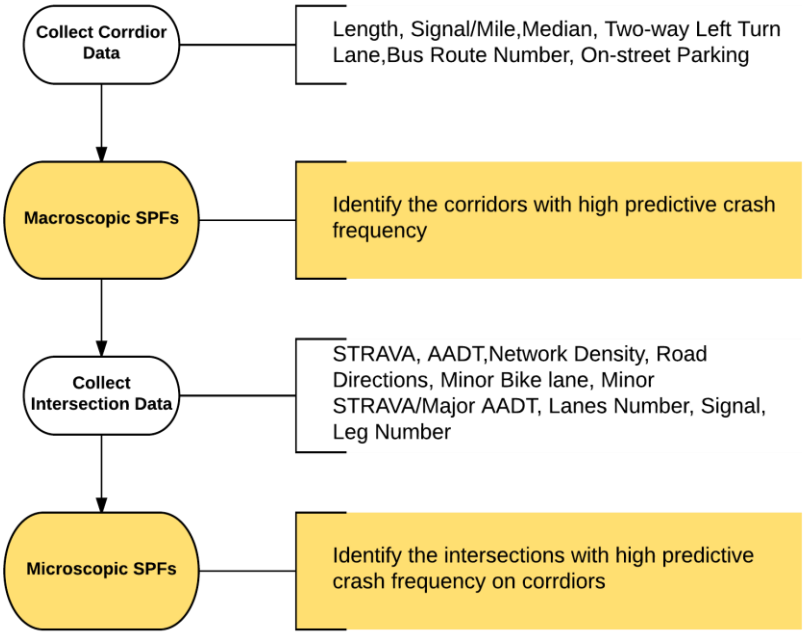

Figure 5-6 The process of using macroscopic and microscopic SPFs

5.5 Establishing SPFs Procedure

A repeatable procedure of building bicycle SPFs is another goal of this research. Bicycle SPFs procedure is created based on the procedure from Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs (Federal Highway Administration, 2013), but adopted to bicycle and using crowdsourced data. The process are summarized below:

Step 1: Determine facility type to build SPFs

Firstly, engineers need to identify the use and facility type of SPFs. The scale difference will influence other steps, labor requirement, and time requirement. The facility types for bicycle include but not limited to intersections, segments of non-highway, corridors, ramps, etc. For instance, in this study, urban non-freeway corridors and intersections are facility types.

Step 2: Identify necessary data

Depending on the uses of SPF and facility types, the required data will be different. The differences include sample size and corresponding datasets. The guidance on the

minimum sample size can be found in SPF Decision Guide (Srinivasan et al., 2013). Generally, more data sets are required when the scale is larger. Statewide SPFs require much larger data set than local project level SPF. In this present study, 209 samples with 161 crashes from 6-year study period were collected for intersection; 50 sites with 417 crash from six years were collected for corridor SPFs. However, in some studies, due to the labor and time limitation, engineers can slightly change the original requirements to be practical.

Step 3: Identify crowdsource data

After determining necessary data, another critical step is to identify the appropriate crowdsourced data. Depending on project scale, sample size, data set requirements, specific crowdsourced data need to be found to meet the purposes. Crowdsourced data can be retrieved from online open sources, provided by DOTs, purchased from agencies, or adopted from other projects. Some of the crowdsourced data, only include a small proportion of actual data and cannot represent overall population. For example, BIKEMAP accident self-report data (BikeMaps.org, 2017) only have few data point for some small cities, which cannot be a proper data choice. In this project, STRAVA® data is chosen as an alternative of annual average daily bicycle count.

Step 4: Verify and clean-up crowdsource data

It is necessary to justify the representativeness and accuracy of crowdsourced data. Crowdsourced data provides emerging opportunities to represent the general population in a project and the sources are becoming more reliable to meet engineering project needs. In this study, STRAVA® is chosen because it is highly reliable and representative. User types of bicyclists should be clarified while verifying the representativeness. For example, bicyclists can be divided into two groups: cyclists for recreation and commuters. In this study, only a portion of all cyclists and commuters in Portland and Eugene are using the STRAVA® application, so the author verified how well the STRAVA® can represent the cycling population. Additionally, since crowdsourced data includes plenty of information, unnecessary data details and confidential personal information should be removed. STRAVA®

bicycle data contains detailed information about each user which was not useful to this project scope and was cleaned by the author.

Step 5: Data analysis

This step provides engineers a sense of how the data look like in order to make decisions on modeling. The mean, variance, scatter plot, histogram graph, etc. are general ways to analyze collected data. An example can be found section 4.4.

Step 6: Model Comparison

Different models should be considered as potential options for bicycle crash dataset since bicycle crashes are sporadic. The occurrence of a bicycle crash on one location is generally not a great indicator of future crashes. . The best models are chosen from potential model options to fit the dataset best. For example, in this study, around 110 zero crash intersections are in the data set, which can influence the choice of models. Poisson, NB, ZIP and ZINB models are potential options for count models. Over-dispersion test and Vuong-statistic are used to identify which model should be chosen, and detailed examples can be found in Section 3.5.

Step 7: Develop bicycle crash severity distribution

The crash severity distribution is used to estimate the number of crashes with different crash types. It provides the percentage of types of crashes of total crashes. In this study, it was calculated based on the crashes happened at intersections and corridors that are collected in the samples.

Step 8: Develop the SPFs

SPFs are established based on the model regression from step 7 and severity distribution from step 8. Frequency and severity predictive equations (SPFs) should be created in this step. In this way, transportation agencies only need to calculate the predictive crashes from equations rather than run the model in statistical tools. An example could be found in section 5.4.

Step 9: Interpret the SPFs

Besides building SPFs for jurisdictions, understanding the impacts of variables on the crash frequency and severity can help engineers and decision makers to change designs and policies accordingly. Engineering and policy recommendations are provided based on the results and discussion. The example can be found in chapter 6.

6. Recommendation and Conclusion

Engineering and policy recommendations for building SPFs in the U.S. and what has been learned from this study are provided in this chapter. In addition, the author concludes the overall study with major findings and discusses the potential research directions in the future work.

6.1 Engineering Recommendation of Bicycle Safety

The results from SPFs for intersections and corridors provide evidence of how factors are impacting bicycle safety. The identified major influencing factors that affect bicycle crash frequency can be further used to facilitate the future road design, especially for the cities that have the similar scale as Portland and Eugene, OR. The general rule is that engineers could consider mitigating the factors that have positive effects on bicycle crash frequency and could also promote the designs that have negative impacts on bicycle crash frequency. Positive coefficient of a variable for intersection Poisson model (see Table 5-1) and corridor NB model (see Table 5-4) implies a positive impact, and vice versa. The recommendations of bicycle safety design at intersections include:

1. Engineers may build more bicycle buffers on one-way roads rather two-way roads since the one-way roads decrease the crash frequency at intersections, according to the model results in section 5.1.
2. Countermeasures are necessary for intersections with more legs or signal traffic control to help bicyclists mitigating the conflicts with motor vehicles, such as bicycle traffic lights and median at 5-leg intersections.
3. The ratio of minor street AADB to major street AADT is negative suggesting that when drivers are aware of more bicycles crossing a street, fewer bicycle

crashes happen. A vertical sign is an option to influence on the driver expectancy and warn drivers that bicycles are crossing.

4. Bicycle traffic lights could be installed at intersections to mitigate the conflicts between bicycles and vehicles. Figure 6-1 shows a bicycle priority traffic light that decreases the bicycle risk in Portland, Oregon.

Figure 6-1 Bicycle traffic light in Portland, Oregon (Maus, 2011).

5. Dedicated bicycle buffer lanes and bicycle buffer lane with vertical structure could be implemented near intersections with more traffic lanes to mitigate the conflicts between two modes of traffic.

When the intersection SPFs inspect safety performance from a microscopic scale, the corridor SPFs results provide different perspectives from a macroscopic level. The difference between the two levels of scale could lead to different interpretations. The engineering recommendation of bicycle safety design on the corridors are summarized as below:

1. Engineers may consider isolating bicycles from buses while designing bus stops since the number of bus routes has a positive impact on corridor bicycle crash frequency. A bus stop island could be the potential option to mitigate the conflicts between bus and bicycle (National Association of City Transportation Officials, 2013), shown in Figure 6-2.

Figure 6-2 Bus island to mitigate the conflicts with bicycles (National Transport Authority, 2011).

2. Median could be a countermeasure to increase bicycle safety, especially when a road has multiple lanes, it was found to have negative impact on bicycle crash frequency in Portland and Eugene in this thesis.

3. The increase of intersections with signal control on corridors result in more bike crashes. Therefore, strategies need to implement to attract bicyclists riding on the roads with fewer signal controlled intersections. Bike lanes could be eliminated on the roads with a large number of intersections with signal control, because of its attraction of bicyclists.

6.2 Policy Recommendation of Bicycle Safety

Similar to the recommendation for engineers, some decision makers could also learn strategies to improve bicycle safety based on the SPFs but at a higher level. The recommendations of bicycle safety policy for intersections and corridors include:

1. The ratio between minor street STRAVA and major street AADT having a negative impact on bicycle crashes suggests that policy strategies could focus on aggregating bicyclists crossing the same intersection when they have to.

2. Higher functional classed roads with more lanes and signal control are found to be associated with more bicycle crashes. Strategies could target attracting bicyclists riding on streets with lower functional classifications.

3. When public transit and bicycle transportation have been promoted as more sustainable modes, city planners need to consider the relationship between

them as they are operating on roads. It seems plausible that building bike lanes on roads where buses operate can connect the modes; however, it was found to increase the bicycle crash frequency.

4. Investing in building median barrier (e.g. trees and concrete barriers) and two-way left turn lanes, especially on streets with multiple lanes, could help improving overall bicycle safety. The overall improvement is larger than other variables in this study.

6.3 Recommendation of Building SPFs by Using Crowdsourced Data

Building bicycle SPFs by using crowdsourced data is an affordable and efficient way for of jurisdictions of any size. This study discussed what was learned by building bicycle SPFs and these findings are summarized below:

- Different affordable crowdsourced data need to be compared to choose the best one that can represent most types of bicycle users;
- STRAVA®, representing a small proportion of users, is an affable and efficient crowdsourced data;
- Raw crowdsourced data is normally including plenty of additional information that should be cleaned before use;
- Errors in crowdsourced data, especially double count issue, should be taken care before use;
- Reported bicycle crash data from DOTs has under-report issue, especially on PDO crash;
- Building bicycle SPFs on segments suffers insufficient crash data, so SPFs on corridors (multiple homogenous segments with same features) are recommended as an alternative;
- Multiple statistical models could be compared to find the best-fitted regression;

- When data sample site increase, model shows better fitted. A recommendation sample size is more than 200 sites for building SPFs in the urban area. More sample may be needed for the rural area due to fewer bicycle crashes.

6.4 Conclusion

This study established microscopic scale (intersection) and macroscopic scale (corridor) bicycle Safety Performance Functions for medium and large size cities by using crowdsourced data, with a case study in Portland and Eugene-Springfield metropolitans, which overcame the challenges of no sufficient bicycle volume data and crash data. Specifically, in this research 1) bicycle SPFs is built for intersections and corridors that have not been sophistically studied; 2) bicycle crash severity distributions are used the first time to predict the number of bicycle crashes with different crash severity levels; 3) an affordable crowdsourced bicycle volume data – STRAVA® is chosen to solve the problem of limited data; 4) STRAVA® data was verified to be able to represent general bicyclists by comparison with automatic bike count station data; 5) a general framework of building SPFs for was created for jurisdictions.

Transportation agencies, city planners, and engineers can use SPFs to evaluate bicycle safety for both microscopic and macroscopic levels, determine the impact of changing design, screen transportation network, and identify the most efficient investment regarding locations and means. Since the procedure of building bicycle SPFs by using crowdsourced data is established, other jurisdictions can also repeat the same process if they found new SPFs are necessary.

6.5 Limitation and Future Work

Even though several critical challenges have been overcome in this study, there are still some parts that future researcher can focus on. The author admits bicycle SPFs cannot draw the complete causality of bicycle crashes, and the interactions between variables are not clear studied in this research. However, bicycle SPFs established in this study are essential crash prediction tools, so improving the predictive ability is the most important future task.

- More corridor sample sites could be collected to improve the model predictive accuracy;
- The availability of AADT impacts the sites selection process and only the sites having ADT or AADT were selected. Therefore, the prediction results may be biased. Since the majority of ADT or AADT are only existing on high-class roads (e.g. urban major arterials), the SPFs can predict more accurately on those roads rather than local roads. Future work can improve the random sampling process by using a more comprehensive traffic volume data.
- Using other automatic bicycle count station data on different roads to verity how well the STRAVA® data can represent all bicycle population;
- Other types of regressions type can be applied to improve model predictive ability;
- Building bicycle SPFs for various types of cities with different biking cultures, such as a university town, a capital city, or a super large metropolitan.

Reference

Aguero-Valverde, J. and Jovanis, P. P. (2006) 'Spatial analysis of fatal and injury crashes in Pennsylvania', *Accident Analysis and Prevention*, 38(3), pp. 618–625. doi: 10.1016/j.aap.2005.12.006.

American Association of State Highway and Transportation Officials (2010) *Highway Safety Manual*. 1st edn.

Bauer, K. M. and Harwood, D. W. (2000) 'Statistical Models of At-Grade Intersection Accidents. Addendum', (March).

BikeMaps.org (2017) *BikeMaps*. Available at: https://bikemaps.org/ (Accessed: 19 May 2017).

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

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

Cameron, A. C. and Trivedi, P. K. (1990) 'Regression-based tests for overdispersion in the Poisson model', *Journal of Econometrics*, 46(3), pp. 347–364.

Carriquiry, A. L. and Pawlovich, M. (2004) *From Empirical Bayes to Full Bayes: Methods for Analyzing Traffic Safety Data*. doi: http://www.iowadot.gov/crashanalysis/eb_fb_comparison.htm.

Carter, D. L., Hunter, W. W., Zegeer, C. V., Stewart, J. R. and Huang, A. H. F. (2006) 'Pedestrian and Bicyclist Intersection Safety Indices', *Security*, (April), p. 99. Available at: https://www.fhwa.dot.gov/publications/research/safety/pedbike/06130/06130.pdf.

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

Central Lane Metropolitan Planning Organization (2017) *Traffic Volumes*. Available at:

http://www.sdslane.org/645/Traffic-Volumes (Accessed: 15 May 2017).

Chen, P. (2015) 'Built environment factors in explaining the automobile-involved bicycle crash frequencies: A spatial statistic approach', *Safety Science*. Elsevier Ltd, 79, pp. 336–343. doi: 10.1016/j.ssci.2015.06.016.

Creative Research Systems (2016) *Correlation*. Available at: https://www.surveysystem.com/correlation.htm.

Le Dantec, C. A., Asad, M., Misra, A. and Watkins, K. E. (2015) 'Planning with Crowdsourced Data', *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, pp. 1717–1727. doi: 10.1145/2675133.2675212.

Dill, J. and McNeil, N. (2013) 'Four Types of Cyclists? Examination of Typology for Better Understanding of Bicycling Behavior and Potential', *Transportation Research Record: Journal of the Transportation Research Board*, (2387), p. pp 129–138. doi: 10.3141/2387-15.

Dixon, K. K. and Avelar, R. E. (2015) 'Validation Technique Applied to Oregon Safety Performance Function Arterial Segment Models', 5(November), pp. 1–18. doi: http://dx.doi.org/10.3141/2515-15.

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

Domencich, T. and McFadden, D. (1975) 'Statistical Estimation of Choice Probability Functions', *Urban Travel Demand: A Behavioral Analysis*, pp. 101–25.

Dong, C., Clarke, D. B., Yan, X., Khattak, A. and Huang, B. (2014) 'Multivariate Random-Parameters Zero-Inflated Negative Binomial Regression Model: An Application to Estimate Crash Frequencies at Intersections', *Accident Analysis and Prevention*, 70, pp. 320–329. doi: 10.1016/j.aap.2014.04.018.

Ekman, L. (1996) 'On The Treatment of Flow in Traffic Safety Analysis: A Non-Parametric Approach Applied on Vulnerable Road Users', *Bulletin*. Lund Institute of Technology, Department of Technology and Society, Traffic Engineering, (136), p. 99 p. Available at: http://trid.trb.org/view/687009.

El-basyouny, K. and Sayed, T. (2011) 'Conflicts-Based Safety Performance Functions', *Transportation Research Record*, (2583), pp. 1–12.

Eluru, N., Bhat, C. R. and Hensher, D. A. (2008) 'A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes', *Accident Analysis and Prevention*, 40(3), pp. 1033–1054. doi: 10.1016/j.aap.2007.11.010.

Fambro, D., Fitzpatrick, K. and Koppa, R. (1997) 'Determination of Stopping Sight Distances', *NCHRP Report*, (400). Available at: http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_400.pdf.

Federal Highway Administration (2003) 'Pedestrian and Bicycle Facility Guidelines', pp. 198–201.

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

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

Findley, D. J., Hummer, J. E., Rasdorf, W., Zegeer, C. V. and Fowler, T. J. (2012) 'Modeling the impact of spatial relationships on horizontal curve safety', *Accident Analysis and Prevention*. Elsevier Ltd, 45, pp. 296–304. doi: 10.1016/j.aap.2011.07.018.

Greene, W. H. (1994) 'Accounting for excess zeros and sample selection in Poisson and Negative Binomial regression models', *NYU Working Paper No. EC-94-10*, 9, pp. 265–265. doi: 10.1007/BF00857937.

Gross, F., Persaud, B. and Lyon, C. (2010) *A Guide to Developing Quality Crash Modification Factor*. doi: FHWA-SA-10-032.

Haberman, M. (2017) *Portland bike counter: Nudging 1 million trips over the Hawthorne Bridge | OregonLive.com*, *2017 Oregon Live LLC*. Available at:

http://www.oregonlive.com/cycling/index.ssf/2013/03/portland_bike_counter_nudging.html (Accessed: 16 May 2017).

Hauer, E. (1995) 'On exposure and accident rate', *Traffic engineering & control*, 36(January 1995), pp. 134–138.

Hilbe, J. M. (2011) 'Negative Binomial Regression', *Public Administration Review*, 70, pp. 1–6. doi: 10.1111/j.1540-6210.2010.02207.x.

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

Hunt, K. (2015) *Strava training app not the best way to gather cycling data*, *Ottawa Metro*. Available at: http://www.metronews.ca/views/ottawa/your-ride/2015/11/23/strava-training-app-not-the-best-way-to-gather-cycling-data.html.

Hunter, W. W. (1996) *Pedestrian and bicycle crash types of the early 1990's*. doi: 99176134250001451.

Hurwitz, D., Jannat, M., Warner, J., Monsere, C. and A., R. (2015) 'Towards Effective Design Treatment For Right Turns At Intersections With Bicycle Traffic', *Oregon Department of Transportation*, SPR 767, p. 283.

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

Jestico, B., Nelson, T. and Winters, M. (2016) 'Mapping ridership using crowdsourced cycling data', *Journal of Transport Geography*, 52, pp. 90–97. doi: 10.1016/j.jtrangeo.2016.03.006.

Jo, J.-H., Lee, J.-S., Ouyang, Y. and Li, Z. (2009) *Crash Data Analysis and Engineering Solutions for Local Agencies*, *September*. Available at: https://apps.ict.illinois.edu/projects/getfile.asp?id=3016.

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. Available at: https://lup.lub.lu.se/search/publication/24269.

Kim, J. K., Kim, S., Ulfarsson, G. F. and Porrello, L. A. (2007) 'Bicyclist injury severities in bicycle-motor vehicle accidents', *Accident Analysis and Prevention*, 39(2), pp. 238–251. doi: 10.1016/j.aap.2006.07.002.

Kittelson&Associates Inc and ODOT (2014) *Pedestrian and Bicycle Safety Implementation Plan*. Available at: https://www.oregon.gov/ODOT/HWY/TRAFFIC-ROADWAY/docs/pdf/13452_report_final_partsA+B.pdf.

Klop, J. R. and Khattak, A. J. (1999) 'Factors Influencing Bicycle Crash Severity on Two-Lane, Undivided Roadways in North Carolina', *Transportation Research Record: Journal of the Transportation Research Board*, 1674, pp. 78–85. doi: 10.3141/1674-11.

Lambert, D. (1992) 'Zero-Inflated Poisson With an Regression , in Manufacturing to Defects Application', *Technometrics*, 34(1), pp. 1–14.

League of Amercian Bicyclists (2015) *Bicycle Commuting Data*. Available at: http://bikeleague.org/commutingdata.

Lindman, M., Jonsson, S., Jakobsson, L., Karlsson, T., Gustafson, D. and Fredriksson, A. (2015) 'Cyclists interacting with passenger cars; a study of real world crashes', *International Research Council on the Biomechanics of Injury*, pp. 1–12. Available at: http://www.ircobi.org/downloads/irc15/pdf_files/10.pdf.

Long, J. S. (1997) 'Regression models for categorical and limited dependent variables', *American Journal of Sociology*, p. 328. doi: 10.1086/231290.

Malyshkina, N. V. and Mannering, F. L. (2010) 'Zero-state Markov switching count-data models: An empirical assessment', *Accident Analysis and Prevention*, 42(1), pp. 122–130. doi: 10.1016/j.aap.2009.07.012.

Maus, J. (2011) *On January 1, bike traffic signals get the green light in Oregon, BikePortland.org*. Available at: https://bikeportland.org/2011/12/27/on-january-1-bike-traffic-signals-get-the-green-light-in-oregon-64283 (Accessed: 6 February 2017).

McArthur, A., Savolainen, P. T. and Gates, T. J. (2014) 'Spatial Analysis of Child Pedestrian and Bicycle Crashes Development of Safety Performance Function for Areas

Adjacent to Schools', *Transportation Research Record*, (2465), pp. 57–63. doi: 10.3141/2465-08.

Mekuria, M. C., Furth, P. G. and Nixon, H. (2012) 'Loss-Stress Bicycling and Network Connectivity', *Mineta Transportation Institute Report 11-19*, p. 68. Available at: http://transweb.sjsu.edu/PDFs/research/1005-low-stress-bicycling-network-connectivity.pdf.

Miaou, S. P. (1994) 'The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions', *Accident Analysis and Prevention*, 26(4), pp. 471–482. doi: 10.1016/0001-4575(94)90038-8.

Midwest Research Institute, iTRANS Consulting, I., Human Factors North, I. and Hauer, E. (2002) 'SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites (Task K)'.

Monsere, C., Wang, H., Wang, Y. and Chen, C. (2017) *Risk Factors for Pedestrian and Bicycle Crashes (unpublished)*.

Moore, D. N., Schneider IV, W. H., Savolainen, P. T. and Farzaneh, M. (2011) 'Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations', *Accident Analysis and Prevention*. Elsevier Ltd, 43(3), pp. 621–630. doi: 10.1016/j.aap.2010.09.015.

Mullahy, J. (1986) 'Specification and testing of some modified count data models', *Journal of Econometrics*, 33(3), pp. 341–365.

National Association of City Transportation Officials (2013) *Urban Street Design Guide*.

National Center for Statistics and Analysis (2017) *Bicyclists and other cyclists: 2015 data. (Traffic Safety Facts. Report No. DOT HS 812 382), Washington, DC: National Highway Traffic Safety Administration.* Washington, DC. Available at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812382.

National Highway Traffic Safety Adminstriation (2014) *TRAFFIC SAFETY FACTS 2012 Data, National Highway Transportation Safety Administration*. doi: Report No. DOT HS 812 032.

National Highway Traffic Safety Adminstriation (2016) *Traffic Safety Facts, NHTSA's National Center for Statistics and Analysis*. doi: http://dx.doi.org/10.1016/j.annemergmed.2013.12.004.

National Transport Authority (2011) *National Cycle Manual*. Available at: https://www.cyclemanual.ie/manual/thebasics/.

Nordback, K., Marshall, W. E. and Janson, B. N. (2014) 'Bicyclist safety performance functions for a U.S. city', *Accident Analysis and Prevention*. Elsevier Ltd, 65, pp. 114–122. doi: 10.1016/j.aap.2013.12.016.

Nordback, K., Marshall, W. E., Janson, B. N. and Stolz, E. (2013) 'Estimating Annual Average Daily Bicyclists', *Transportation Research Record: Journal of the Transportation Research Board*, 2339(1), pp. 90–97. doi: 10.3141/2339-10.

Oh, J., Jun, J., Kim, E. and Kim, M. (2008) 'Assessing Critical Factors Associated with Bicycle Collisions at Urban Signalized Intersections', *Transportation Reserach Board*, pp. 1–17.

Oregon Department of Transportation (2016) *Summary of Traffic Trends at Automatic Traffic Recorder Stations 2015*. Available at: http://www.oregon.gov/ODOT/td/tdata/Pages/tsm/tvt.aspx.

Oregon Department of Transportation (2017) *ODOT GIS TransData*. Available at: ftp://ftp.odot.state.or.us/tdb/trandata/GIS_data/ (Accessed: 4 March 2017).

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

Reynolds, C. C. O., Harris, M. A., Teschke, K., Cripton, P. A. and Winters, M. (2009) 'The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature.', *Environmental Health Perspectives*, 8(1), p. 47. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2776010%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract$%5C$nhttp://www.ehjournal.net/content/8/1/47.

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), pp. 47–51.

Roll, J. F. (2013) 'Bicycle Traffic Count Factoring: An Examination of National, State and Locally Derived Daily Extrapolation Factors', p. 175. Available at: http://pdxscholar.library.pdx.edu/open_access_etds.

Ryus, P., Ferguson, E., Laustsen M, K., Schneider J, R., Proulx R, F., Hull, T. and Miranda-Moreno, L. (2014) 'Guidebook on Pedestrian and Bicycle Volume Data Collection', *NCHRP Report*, p. 159p. Available at: http://www.trb.org/Publications/Blurbs/171973.aspx%5Cnhttps://trid.trb.org/view/1342012.

Schepers, J. P., Kroeze, P. A., Sweers, W. and Wüst, J. C. (2011) 'Road factors and bicycle-motor vehicle crashes at unsignalized priority intersections', *Accident Analysis and Prevention*, 43(3), pp. 853–861. doi: 10.1016/j.aap.2010.11.005.

Scott, G. (2015) *Strava 2014: The year in numbers*, *Roadcycling UK*. Available at: http://roadcyclinguk.com/sportive/strava-2014-year-numbers.html#OLMfeEHYSHlruhLO.97%5Cnhttp://roadcyclinguk.com/sportive/strava-2014-year-numbers.html.

Selala, M. K. and Musakwa, W. (2016) 'The potential of strava data to contribute in non-motorised transport (NMT) planning in johannesburg', *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41(July), pp. 587–594. doi: 10.5194/isprsarchives-XLI-B2-587-2016.

Shankar, V., Milton, J. and Mannering, F. (1997) 'Modeling accident frequencies as zero-altered probability processes: An empirical inquiry', *Accident Analysis and Prevention*, 29(6), pp. 829–837. doi: 10.1016/S0001-4575(97)00052-3.

Simmons, E., Kay, M., Ingles, A., Khurana, M., Sulmont, M. and Lyons, W. (2015) 'White Paper : Evaluating the Economic Benefits of Nonmotorized Transportation', (March).

Srinivasan, R., Carter, D. and Bauer, K. (2013) 'Safety Performance Function Decision Guide: SPF Calibration versus SPF Development', (September), pp. 1–31.

Strava (2016a) 'Data-Driven Bicycle and Pedestrian Planning'. Available at:

http://metro.strava.com/.

Strava (2016b) *Strava Global Heatmap*. Available at: http://labs.strava.com/heatmap/#12/-122.67626/45.54243/yellow/bike (Accessed: 18 December 2016).

STRAVA (2017) *Strava | About Us*. Available at: https://www.strava.com/about (Accessed: 23 April 2017).

Tegge, R. A., Jo, J.-H. and Ouyang, Y. (2010) 'Development and Application of Safety Performance Functions for Illinois', (10).

Teschke, K., Harris, M. A., Reynolds, C. C. O., Winters, M., Babul, S., Chipman, M., Cusimano, M. D., Brubacher, J. R., Hunte, G., Friedman, S. M., Monro, M., Shen, H., Vernich, L. and Cripton, P. A. (2012) 'Route infrastructure and the risk of injuries to bicyclists: A case-crossover study', *American Journal of Public Health*, 102(12), pp. 2336–2343. doi: 10.2105/AJPH.2012.300762.

Thomas, T., Jaarsma, R. and Tutert, B. (2009) 'Temporal Variations of Bicycle Demand in the Netherlands: Influence of Weather on Cycling', *Transportation Research Board 88th Annual Meeting*.

Tin Tin, S., Woodward, A. and Ameratunga, S. (2013) 'Incidence, risk, and protective factors of bicycle crashes: findings from a prospective cohort study in New Zealand.', *Preventive medicine*. The Authors, 57(3), pp. 152–61. doi: 10.1016/j.ypmed.2013.05.001.

Todhunter, I. (1865) *History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Cambridge, London, Macmillan and Co.

Turner, S. a, Roozenburg, a P. and Francis, T. (2006) *Predicting Accident Rates for Cyclists and Pedestrians, Land Transport New Zealand Research Report 289*.

Turner, S., Wood, G., Hughes, T. and Singh, R. (2011) 'Safety Performance Functions for Bicycle Crashes in New Zealand and Australia', *Transportation Research Record: Journal of the Transportation Research Board*, 2236, pp. 66–73. doi: 10.3141/2236-08.

U.S. Department of Transportation (2017) *Safer People, Safer Streets: Pedestrian and*

Bicycle Safety Initiative | Department of Transportation. Available at: https://www.transportation.gov/safer-people-safer-streets (Accessed: 10 May 2017).

Vogt, A. and Bared, J. (1998) 'Accident Models for Two-Lane Rural Segments and Intersections', *Transportation Research Record: Journal of the Transportation Research Board*, 1635(1), pp. 18–29. doi: 10.3141/1635-03.

Vuong, Q. H. (1989) 'Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses', *Econometrica*, 57(2), p. 307. doi: 10.2307/1912557.

Wang, H., Chen, C., Wang, Y., Ziyuan, P. and Lowry, M. B. (2017) *Bicycle Safety Analysis: Crowdsourcing Bicycle Travel Data to Estimate Risk Exposure and Create Safety Performance Functions (Unpublished)*.

Wang, H., Palm, M., Chen, C., Vogt, R. and Wang, Y. (2016) 'Does bicycle network level of traffic stress (LTS) explain bicycle travel behavior? Mixed results from an Oregon case study', *Journal of Transport Geography*. Elsevier B.V., 57, pp. 8–18. doi: 10.1016/j.jtrangeo.2016.08.016.

Warner, J., Hurwitz, D. S., Monsere, C. M. and Fleskes, K. (2017) 'A simulator-based analysis of engineering treatments for right-hook bicycle crashes at signalized intersections', *Accident Analysis & Prevention*. Elsevier, 104(November 2016), pp. 46–57. doi: 10.1016/j.aap.2017.04.021.

Washington, S., Karlaftis, M. G. and Mannering, F. L. (2011) *Statistical and Econometric Methods for Transportation Data Analysis [2rd Edition]*.

Watkins, K., Ammanamanchi, R., LaMondia, J. and LeDantec, C. A. (2016) 'Comparison of Smartphone-based Cyclist GPS Data Sources', *Transportation Research Record*.

Wei, F. and Lovegrove, G. (2013) 'An empirical tool to evaluate the safety of cyclists: Community based, macro-level collision prediction models using negative binomial regression', *Accident Analysis and Prevention*. Elsevier Ltd, 61, pp. 129–137. doi: 10.1016/j.aap.2012.05.018.

Yan, X., Ma, M., Huang, H., Abdel-Aty, M. and Wu, C. (2011a) 'Motor vehicle-bicycle crashes in Beijing: Irregular maneuvers, crash patterns, and injury severity', *Accident Analysis and Prevention*. Elsevier Ltd, 43(5), pp. 1751–1758. doi:

10.1016/j.aap.2011.04.006.

Yan, X., Ma, M., Huang, H., Abdel-Aty, M. and Wu, C. (2011b) 'Motor vehicle-bicycle crashes in Beijing: Irregular maneuvers, crash patterns, and injury severity', *Accident Analysis and Prevention*. Elsevier Ltd, 43(5), pp. 1751–1758. doi: 10.1016/j.aap.2011.04.006.

Appendix A: Oregon DOT Crash Data Interview Supplement

Crash Data Strengths, Limitations, and Assumptions

- The Statewide crash data file only includes "reportable" crashes that meet the current State's legislated reporting threshold.
    - As of 01/01/2004 - Over $1,500 damage to any one vehicle; any vehicle is towed from the scene as a result of damage from the accident; injury or death resulted from the accident; or damage to any one person's property other then a vehicle involved in the accident, is over $1,500.
    - Oregon is a self-reporting state by statute which means drivers must report. However, police are not required to attend or to investigate or to report all accidents they attend.
- DMV is the regulatory authority over driver records, which include liability insurance law compliance. They are the owners of the crash reports and have the responsibility of assembling/matching crash reports; to create cases for all drivers included in one crash, verifying insurances, licenses, and tracking citation information, and administrative court proceedings license suspensions. Drivers have 72 hours to report, and police have two weeks to send in reports. This process requires time and is part of the reason crash data is delayed. We have partnered with DMV to improve this process and our objective is to get it coding back to a three-month lag from the month of the crash.
- Current year data is available as soon as it is added to the reportable database. It has been subjected to 300+ system generated validations. However, it has not gone through the more rigorous final year-end QA / QC process. This makes it preliminary and subject to change. Although we do not encourage using preliminary data for final safety project decisions or litigation, it can be suitable for preliminary review, research, and evaluation. Generally, five previous years of crash data is quite sufficient for most project purposes.
- There are approximately 50,000 crashes added to the state file annually. This includes all jurisdictions of public roads; some special jurisdictions i.e. tribal land, NFS, BLM; 36 counties; and 247 incorporated cities. There are 10 crash data technicians who evaluate and code the reports.
- Bicyclist, pedestrians, and owners of parked vehicles are not required or expected to report crashes. *If they can get the information from the vehicle drivers and include it in a report, that will encourage DMV to pursue the driver of the vehicle.*
- Not all "reportable" crashes are reported through the appropriate channels; hence, all anecdotal stories can not be supported with data.
- "Close calls" are not reported.
- Reported crash locations are viewed with aerial imagery to confirm reliability of the information. Locations are adjusted as supporting visual information indicates. This procedure applies to reports from police and drivers.
- Each crash site is assigned a latitude/longitude on the roadway to allow for geo-locating and mapping purposes.
- Crash locations that are not precise enough to assign a true value are assigned a predetermined latitude/longitude (point) placing the crash in a non-roadway area, such as a field. But these crashes can still be included in GIS summary queries. *An example of such a location would be on 12th Street, in Salem but not cross-referenced from another fixed site or intersecting road.* This crash would be assigned to a "default unlocatable point" for the city of Salem.
- Crashes on state highways which are not precise enough to assign a milepoint, are coded to a 999.99 milepoint value and coded to the county and/or city they reportedly occurred in.

Created on 10/16/2013 12:47:00 PM
Created by Robin Ness, Manager Crash Analysis and Reporting Unit

- Rural jurisdictions have less bicycle and pedestrian collisions, and more animal collisions for obvious reasons.

Police Reporting Considerations

- Not all crashes are attended by law enforcement – not enough resources to cover all crashes and not required by law. About 50% +- include a police report.
- Crashes that involve fatalities or injury transport and citations are most likely to be police reported.
- Police reports do not provide precise crash location information unless they are fatalities or very serious and involve crash reconstruction investigation. Their investigation priorities at the crash scene are not with precise location information but rather general location, unless it results in further serious investigation or death
- Coordinate values provided on police reports are frequently not precise enough for engineering purposes. They may be recording where they stopped which can be off road and/or 100+- feet away.
- 99.9% of fatals are reported by police. The exceptions are those fatalities on tribal lands that are investigated only by tribal police.

Citizen Driver Self-Reporting Considerations

- Because of self-reporting some contributing causes, locations and driver issues will be under – reported i.e.,
  - Cell phone use / texting / hand held devices
  - Distraction
  - Alcohol and drugs
  - Bicycle / vehicle collisions that require no medical transport or emergency response
  - Hit-n-run with parked vehicles or fixed objects – no driver to assign crash in driver records.
  - Driver license status
  - Rural crashes
- Driver accident reporting forms are created by DMV with the input of a stakeholder group; police, the Crash Analysis and Reporting (CAR) Unit, Fatal Analysis Reporting System (FARS), Traffic Safety Division and insurance groups. They are developed to guide the reporting driver to use check boxes and specific codes for key information and describe crash locations in a simple, consistent manner. This makes the interpretation of crash locations and the circumstances of driver reported crashes more consistent.
- Frequently you have more then one driver report and the information on the reports are compared to one another to determine the location consistencies or inconsistencies. Then viewed with aerial imagery to further determine applicable traffic control, number of lanes, etc. If it's a fixed object, one car collision, the technician would use the imagery to locate the 'fixed object' if possible.
- Drivers may more accurately identify possible and non-incapacitating injuries as police do not follow-up once the crash site is cleared and do not attend more minor crashes.
- My observation over several years of experience with citizen reports is that a fair portion of drivers report as accurately as they can because they have an interest in explaining the circumstances for insurance and their driver records.
- Determination of fault or error is not decided by the drivers; it's decided by the overwhelming evidence at the location and considering all circumstances reported on the crash. If a police report is available and complete, the fault /errors can be derived from it.

Created on 10/16/2013 12:47:00 PM
Created by Robin Ness, Manager Crash Analysis and Reporting Unit

Interpreting Crash Data and Using Safety Tools

- Data can be reported out in hundreds of different report formats. These different formats may reflect overlapping data elements or overlapping location. If a query overlaps a previous location you will get crashes that were identified in the previous query. If you ask for drivers aged 16-20 and then ask for drivers 20-30 you will get overlapping data. It's critical that data users know what they are querying.

- Access databases and the DECODE database are readily available. Unless you have a knowledgable and skilled Access expert you may get incorrect query results. Please consult with Kelly Hawley or Theresa Heyn, CAR Unit Data Analysts to get tips in using the tools before concluding you have bad data.

- Geodatabases are available upon request. Please consult with Theresa Heyn a CAR Unit Data Analyst to ensure the best results form your efforts.

- There are various safety data tools available for your use. It's recommended you consult with the subject experts to ensure you are correctly applying the tools to the data.

Created on 10/16/2013 12:47:00 PM
Created by Robin Ness, Manager Crash Analysis and Reporting Unit

Appendix B: Modeling Results

Intersection Poisson Model

```
|-> poisson
    ;lhs= Crash
    ;rhs= one, TT_STR2, TTAADT2, Net_D, MI_Dir2, MI_Bike, MiSMaA, TTLane, Signal, Legs
    ;marginal effect$

------------------------------------------------------------------------------
Poisson Regression
Dependent variable                CRASH
Log likelihood function      -206.51824
Restricted log likelihood    -268.02298
Chi squared [   9 d.f.]       123.00948
Significance level               .00000
McFadden Pseudo R-squared      .2294756
Estimation based on N =    209, K =  10
Inf.Cr.AIC  =    433.0 AIC/N =    2.072
Model estimated: May 20, 2017, 11:48:03
Chi- squared =   199.17660  RsqP= .4969
G  - squared =   188.20636  RsqD= .3953
Overdispersion tests: g=mu(i)  :  -.559
Overdispersion tests: g=mu(i)^2:  1.480
--------+---------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   CRASH|  Coefficient       Error       z    |z|>Z*         Interval
--------+---------------------------------------------------------------------
Constant|   -7.17456***     1.29410    -5.54  .0000    -9.71095  -4.63818
 TT_STR2|     .02351***      .00299     7.87  .0000      .01765    .02937
 TTAADT2|     .01269*        .00771     1.65  .0999     -.00243    .02781
   NET_D|     .02792***      .00928     3.01  .0026      .00972    .04611
 MI_DIR2|    1.59389***      .55277     2.88  .0039      .51048   2.67729
 MI_BIKE|     .56335***      .17884     3.15  .0016      .21283    .91388
  MISMAA|    -.01559***      .00487    -3.20  .0014     -.02513   -.00605
  TTLANE|     .13755**       .05689     2.42  .0156      .02604    .24906
  SIGNAL|     .50249**       .25591     1.96  .0496      .00092   1.00406
    LEGS|     .68367***      .25127     2.72  .0065      .19118   1.17616
--------+---------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
------------------------------------------------------------------------------


------------------------------------------------------------------------------
Partial derivatives of expected val. with
respect to the vector of characteristics.
Effects are averaged over individuals.
Observations used for means are All Obs.
Sample average conditional mean     .7703
Scale Factor for Marginal Effects   .7703
--------+---------------------------------------------------------------------
        |     Partial      Standard            Prob.      95% Confidence
   CRASH|      Effect        Error       z    |z|>Z*         Interval
--------+---------------------------------------------------------------------
 TT_STR2|     .01811***      .00271     6.69  .0000      .01280    .02342
 TTAADT2|     .00978         .00599     1.63  .1027     -.00197    .02152
   NET_D|     .02150***      .00735     2.93  .0034      .00710    .03591
 MI_DIR2|     .67368***      .12215     5.52  .0000      .43426    .91310   #
 MI_BIKE|     .47210***      .16545     2.85  .0043      .14783    .79638   #
  MISMAA|    -.01201***      .00387    -3.11  .0019     -.01959   -.00443
  TTLANE|     .10596**       .04462     2.37  .0176      .01851    .19340
  SIGNAL|     .33634**       .15020     2.24  .0251      .04196    .63073   #
    LEGS|     .52665***      .19796     2.66  .0078      .13865    .91466
--------+---------------------------------------------------------------------
#  Partial effect for dummy variable is E[y|x,d=1] - E[y|x,d=0]
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
```

Intersection ZIP model

```
------------------------------------------------------------------------------
Zero Inflated Poisson      Regression Model
Logistic distribution used for splitting model.
ZIP term in probability is F[tau x ln LAMBDA]
Comparison of estimated models
```

```
           Pr[0|means]       Number of zeros        Log-likelihood
Poisson          .57580   Act.=   111 Prd.=   120.3      -215.95559
Z.I.Poisson      .53434   Act.=   111 Prd.=   111.7      -218.32992
Note, the ZIP log-likelihood is not directly comparable.
ZIP model with nonzero Q does not encompass the others.
Vuong statistic for testing ZIP vs. unaltered model is      -.5149
Distributed as standard normal. A value greater than
+1.96 favors the zero altered Z.I.Poisson model.
A value less than -1.96 rejects the ZIP model.
--------+-------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   CRASH|  Coefficient       Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
        |Poisson/NB/Gamma regression model
Constant|   -1.00080***      .34506    -2.90  .0037    -1.67710   -.32450
 TT_STR2|     .01174***      .00148     7.93  .0000      .00884    .01464
MA_RIGHT|    -.18092*        .09601    -1.88  .0595     -.36909    .00726
 MI_BIKE|     .14957*        .08106     1.85  .0650     -.00930    .30843
  MISMAA|    -.00622***      .00217    -2.87  .0041     -.01048   -.00197
  TTLANE|     .05612**       .02429     2.31  .0209      .00851    .10374
  SIGNAL|     .25395**       .10314     2.46  .0138      .05180    .45610
    LEGS|     .13389*        .08045     1.66  .0961     -.02380    .29157
        |Zero inflation model
     Tau|   -7.26667***     2.45567    -2.96  .0031   -12.07969  -2.45364
--------+-------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------------


-----------------------------------------------------------------------------
Partial derivatives of expected val. with
respect to the vector of characteristics.
Effects are averaged over individuals.
Observations used for means are All Obs.
Sample average conditional mean     .8326
Scale Factor for Marginal Effects  1.8647
--------+-------------------------------------------------------------------
        |     Partial      Standard            Prob.      95% Confidence
   CRASH|      Effect        Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
 TT_STR2|     .02189***      .00459     4.77  .0000      .01290    .03087
MA_RIGHT|    -.33735*        .17939    -1.88  .0600     -.68894    .01425   #
 MI_BIKE|     .27890*        .15091     1.85  .0646     -.01688    .57467   #
  MISMAA|    -.01161*        .00623    -1.86  .0623     -.02381    .00060
  TTLANE|     .10465**       .04527     2.31  .0208      .01592    .19339
  SIGNAL|     .47353**       .19204     2.47  .0137      .09715    .84991   #
    LEGS|     .24965*        .14971     1.67  .0954     -.04378    .54308
--------+-------------------------------------------------------------------
#  Partial effect for dummy variable is E[y|x,d=1] - E[y|x,d=0]
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------------
```

Intersection NB model

```
-----------------------------------------------------------------------------
Negative Binomial Regression
Dependent variable                CRASH
Log likelihood function      -213.78906
Restricted log likelihood    -216.79074
Chi squared [   1 d.f.]         6.00337
Significance level               .01428
McFadden Pseudo R-squared      .0138460
Estimation based on N =    209, K =   8
Inf.Cr.AIC  =    443.6 AIC/N =    2.122
Model estimated: May 20, 2017, 11:50:46
NegBin form 2; Psi(i) = theta
Tests of Model Restrictions on Neg.Bin.
Model               Logl ChiSquared[df]
Poisson(b=0)     -268.02  ******** [**]
Poisson          -216.79     102.5 [ 6]
Negative Bin.    -213.79       6.0 [ 1]
--------+-------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   CRASH|  Coefficient       Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|   -2.78070***      .46875    -5.93  .0000    -3.69944  -1.86196
 TT_STR2|     .02624***      .00282     9.30  .0000      .02071    .03177
 MI_BIKE|     .40183*        .22575     1.78  .0751     -.04063    .84429
```

```
  MISMAA|    -.01556**       .00642    -2.42  .0154     -.02813   -.00298
  TTLANE|     .19068***      .06796     2.81  .0050      .05749    .32387
  SIGNAL|     .86770***      .31262     2.78  .0055      .25497   1.48042
 MI_PARK|     .39993*        .23355     1.71  .0868     -.05781    .85768
        |Dispersion parameter for count data model
   Alpha|     .21551         .14855     1.45  .1468     -.07564    .50666
--------+------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
-------------------------------------------------------------------------------
```

Intersection ZINB model

```
-------------------------------------------------------------------------------
Zero Inflated Neg.Binomial Regression Model
Logistic distribution used for splitting model.
ZIP term in probability is F[tau x ln LAMBDA]
Comparison of estimated models
            Pr[0|means]       Number of zeros        Log-likelihood
Poisson          .56849   Act.=   111 Prd.=   118.8      -219.04454
Neg. Bin.        .75261   Act.=   111 Prd.=   157.3      -215.50139
Z.I.Neg_Bin      .51741   Act.=   111 Prd.=   108.1      -220.63289
Note, the ZIP log-likelihood is not directly comparable.
ZIP model with nonzero Q does not encompass the others.
Vuong statistic for testing ZIP vs. unaltered model is     -1.2708
Distributed as standard normal. A value greater than
+1.96 favors the zero altered Z.I.Neg_Bin model.
A value less than -1.96 rejects the ZIP model.
--------+----------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   CRASH|  Coefficient       Error       z    |z|>Z*         Interval
--------+----------------------------------------------------------------------
        |Poisson/NB/Gamma regression model
Constant|    -.41566**       .16248    -2.56  .0105     -.73412   -.09720
 TT_STR2|     .01086***      .00211     5.15  .0000      .00673    .01498
 MI_BIKE|     .13497*        .07986     1.69  .0910     -.02156    .29150
  MISMAA|    -.00500**       .00235    -2.13  .0332     -.00961   -.00040
  TTLANE|     .04251*        .02372     1.79  .0731     -.00398    .08900
  SIGNAL|     .23502**       .10858     2.16  .0304      .02221    .44783
        |Dispersion parameter
   Alpha|     .11523         .21124      .55  .5854     -.29880    .52925
        |Zero inflation model
     Tau|   -8.25982**      3.93435    -2.10  .0358   -15.97100   -.54864
--------+----------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
-------------------------------------------------------------------------------
```

Corridor Poisson model

```
-------------------------------------------------------------------------------
Poisson Regression
Dependent variable                CRASH
Log likelihood function      -142.25440
Restricted log likelihood    -257.89010
Chi squared [   6 d.f.]       231.27139
Significance level              .00000
McFadden Pseudo R-squared      .4483914
Estimation based on N =     50, K =   7
Inf.Cr.AIC  =    298.5 AIC/N =    5.970
Model estimated: May 20, 2017, 11:53:17
Chi- squared =   103.25974  RsqP= .7071
G  - squared =   113.24771  RsqD= .6713
Overdispersion tests: g=mu(i)  :  3.248
Overdispersion tests: g=mu(i)^2:  2.462
--------+----------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   CRASH|  Coefficient       Error       z    |z|>Z*         Interval
--------+----------------------------------------------------------------------
Constant|     .14516         .17930      .81  .4182     -.20626    .49657
  LENGTH|     .81447***      .07110    11.46  .0000      .67512    .95382
 SIGMILE|     .09235***      .01163     7.94  .0000      .06956    .11515
  MEDIAN|    -.49667**       .20217    -2.46  .0140     -.89292   -.10043
    TWLT|    -.45303***      .14635    -3.10  .0020     -.73986   -.16619
    BUSR|     .37679***      .07169     5.26  .0000      .23627    .51730
 PARKING|    -.42164***      .12152    -3.47  .0005     -.65983   -.18346
--------+----------------------------------------------------------------------
```

```
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
--------------------------------------------------------------------------------
```

Corridor NB model

```
--------------------------------------------------------------------------------
Negative Binomial Regression
Dependent variable                CRASH
Log likelihood function      -133.58044
Restricted log likelihood    -142.25440
Chi squared [   1 d.f.]        17.34793
Significance level               .00003
McFadden Pseudo R-squared      .0609750
Estimation based on N =     50, K =   8
Inf.Cr.AIC  =    283.2 AIC/N =    5.663
Model estimated: May 20, 2017, 11:53:17
NegBin form 2; Psi(i) = theta
Tests of Model Restrictions on Neg.Bin.
Model               Logl ChiSquared[df]
Poisson(b=0)     -257.89  ******** [**]
Poisson          -142.25     231.3 [ 6]
Negative Bin.    -133.58      17.3 [ 1]
--------+-----------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   CRASH|  Coefficient       Error       z    |z|>Z*         Interval
--------+-----------------------------------------------------------------------
Constant|     .09144         .24845      .37  .7128     -.39551    .57839
  LENGTH|     .84148***      .15934     5.28  .0000      .52919   1.15378
 SIGMILE|     .09847***      .02200     4.48  .0000      .05535    .14158
  MEDIAN|    -.76336*        .41094    -1.86  .0632    -1.56880    .04207
    TWLT|    -.66950***      .25541    -2.62  .0088    -1.17010   -.16890
    BUSR|     .45500***      .15760     2.89  .0039      .14611    .76388
 PARKING|    -.47638**       .23669    -2.01  .0441     -.94028   -.01248
        |Dispersion parameter for count data model
   Alpha|     .15458*        .08324     1.86  .0633     -.00858    .31773
--------+-----------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
--------------------------------------------------------------------------------


--------------------------------------------------------------------------------
Partial derivatives of expected val. with
respect to the vector of characteristics.
Effects are averaged over individuals.
Observations used for means are All Obs.
Sample average conditional mean    8.7816
Scale Factor for Marginal Effects  8.7816
--------+-----------------------------------------------------------------------
        |     Partial      Standard            Prob.      95% Confidence
   CRASH|      Effect        Error       z    |z|>Z*         Interval
--------+-----------------------------------------------------------------------
  LENGTH|    7.38959***     1.83070     4.04  .0001     3.80148  10.97769
 SIGMILE|     .86472        1.75733      .49  .6227    -2.57959   4.30902
  MEDIAN|   -5.16692        8.47571     -.61  .5421   -21.77901  11.44517   #
    TWLT|   -5.76652        7.67868     -.75  .4527   -20.81645   9.28341   #
    BUSR|    3.99561        3.79230     1.05  .2921    -3.43716  11.42837
 PARKING|   -4.31573        3.60458    -1.20  .2312   -11.38058   2.74911   #
--------+-----------------------------------------------------------------------
#  Partial effect for dummy variable is E[y|x,d=1] - E[y|x,d=0]
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
--------------------------------------------------------------------------------
```

Corridor ZINB model

```
--------------------------------------------------------------------------------
Zero Inflated Neg.Binomial Regression Model
Logistic distribution used for splitting model.
ZIP term in probability is F[tau x ln LAMBDA]
Comparison of estimated models
           Pr[0|means]       Number of zeros        Log-likelihood
Poisson          .00153   Act.=     4 Prd.=      .1      -142.25440
Neg. Bin.        .56105   Act.=     4 Prd.=    28.1      -133.58044
Z.I.Neg_Bin      .01817   Act.=     4 Prd.=      .9      -132.40685
Note, the ZIP log-likelihood is not directly comparable.
ZIP model with nonzero Q does not encompass the others.
Vuong statistic for testing ZIP vs. unaltered model is      1.0696
Distributed as standard normal. A value greater than
+1.96 favors the zero altered Z.I.Neg_Bin model.
A value less than -1.96 rejects the ZIP model.
```

CRASH	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Poisson/NB/Gamma regression model					
Constant	.18927	.23810	.79	.4267	-.27739	.65594
LENGTH	.81493***	.14273	5.71	.0000	.53517	1.09468
SIGMILE	.09429***	.02052	4.59	.0000	.05407	.13452
MEDIAN	-.69641*	.39482	-1.76	.0778	-1.47024	.07742
TWLT	-.62027**	.24470	-2.53	.0113	-1.09987	-.14067
BUSR	.42443***	.14890	2.85	.0044	.13260	.71627
PARKING	-.46520**	.21919	-2.12	.0338	-.89480	-.03560
	Dispersion parameter					
Alpha	.12912	.09005	1.43	.1516	-.04737	.30561
	Zero inflation model					
Tau	-2.46846**	1.05568	-2.34	.0194	-4.53755	-.39937

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.