12-2016

Improved Methodology for Developing Non-Motorized Safety Perfomance Functions

Keneth Morgan Kwayu
Western Michigan University

IMPROVED METHODOLOGY FOR DEVELOPING NON-MOTORIZED SAFETY PERFOMANCE FUNCTIONS

by

Keneth Morgan Kwayu

A thesis submitted to the Graduate College
in partial fulfilment of the requirements
for the degree of Master of Science in Engineering (Civil)
Civil and Construction Engineering
Western Michigan University
December 2016

Thesis Committee:

Valerian Kwigizile, Ph.D., Chair
Jun-Seok Oh, Ph.D.
Ron Van Houten, Ph.D.

IMPROVED METHODOLOGY FOR DEVELOPING NON-MOTORIZED SAFETY PERFOMANCE FUNCTIONS

Keneth Morgan Kwayu, M.S.E.

Western Michigan University, 2016

This study aimed at improving the methodology for developing statewide non-motorized safety performance functions (SPFs). Due to lack of pedestrian and bicyclist counts, the methodology proposed a procedure for developing statewide surrogate non-motorized exposure measures using data that are available at statewide level. Eleven years non-motorized crashes at signalized urban intersections joining Arterial and Collector roads in Michigan were used to test the procedure.

The study also explored the use of Bayesian approach for modeling non-motorized crashes as an alternative to traditional classical count data models. Classical count data models that were considered as potential fit to the data include; Poisson Regression (PRM), Negative Binomial Regression (NBRM), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB). NBRM was selected as the best classical count data model after thorough comparison between competing models using appropriate goodness of fit tests. For Bayesian approach, Poisson likelihood with gamma distribution prior was used for model estimation. The results showed that the Bayesian Poisson-gamma model outperforming classical NBRM model in terms of model estimation and out-of-sample prediction, especially with a relative small sample size.

DEDICATION

I dedicate this thesis to my parents and siblings. I will always be proud of them for their endless love and relentless support in all aspects of my life.

ACKNOWLEDGEMENTS

I would like first and far most to thank God, my Ebenezer, for enabling me to reach this life milestone. He has done it again!

Special appreciation to my thesis supervisors; Dr. Valerian Kwigizile, Dr. Jun-Seok Oh and Dr. Ron Van Houten for their invaluable support and guidance from the beginning of this thesis up to its completion.

I would also like to acknowledge the help, support, friendship and encouragement from my colleagues at Transportation Research Center for Livable Communities (TRCLC) lab.

Special thanks to US Department of Transportation through the Transportation Research Center for Livable Communities (TRCLC), a Tier 1 University Transportation Center for sponsoring this thesis.

Keneth Morgan Kwayu

TABLE OF CONTENTS

Table of Contents-Continued

Table of Contents-Continued

LIST OF TABLES

List of Tables-Continued

LIST OF FIGURES

List of Figures-Continued

CHAPTER 1

INTRODUCTION

Research Problem

Walking and biking are forms of transportation that offers basic mobility for all people. In totality, walking and biking improve quality of life in many ways such as increased physical activities and active lifestyles that consequently reduce obesity and other health related problems. In community were walking and biking is encouraged, it reduces number of motor vehicle trips which are often the cause of air pollution and congestion. It can also boost local economy by inviting retail merchant to invest in places near homes and working places.

In USA, trips that are done by walking and bicycling rose from 9.5% in 2001 to 11.9% in 2009 (National Household Travel Survey, 2009). On the other hand, bicyclist and pedestrian are 2.3 and 1.5 times, respectively, more likely be killed in a crash for each trip as compared to vehicle occupants(Beck et al, 2007).

Therefore, transportation agencies have several prevailing concerns with respect to pedestrian and bicycle safety. Resource constraints make it imperative for such agencies to develop a framework for identifying those locations that are at highest risk for non-motorized crashes. Most importantly, the ability to not only develop, but also to evaluate effectiveness of appropriate countermeasures is crucial for ensuring safety of pedestrians and bicyclists.

Against this backdrop, safety performance functions (SPFs) provide a promising approach for quantifying the risk for non-motorized crashes at specific intersections or road segments. Currently in Michigan, there is no robust safety performance function

developed to cater for statewide non-motorized safety planning. The difficult has been mostly in obtaining necessary data that are available at statewide level for model development such as pedestrian and bicycle volumes counts and Average Annual Daily Traffic (AADT) for arterial, collectors and local roads. These data are essential part of the model as they explain most of variation in non-motorized crashes occurred at different locations.

Therefore, careful sampling plan which captures the randomness of non-motorized crashes and inclusion of reliable proxy exposure measure for pedestrian and bicyclist will help in coming out with the robust statewide safety performance function for bicyclist and pedestrians. These SPFs can be modified over time as more planning agencies within the state are starting to collect pedestrian and bicycle volumes within their jurisdiction for planning purpose.

Objectives of the Project

The main objective was to develop a better methodology for developing statewide safety performance functions for pedestrian and bicyclist at urban intersection. Specifically, the procedure for developing these SPFs addressed the following:

- Proper sampling procedure in coming up with unbiased sample size for model development
- Developing a proxy measure for pedestrian and bicyclist exposure using data that are readily available at statewide level.
- Using appropriate modeling technique to improve SPFs model performance.

Scope of the Study and Thesis Format

Methodology formulated in this research can be used to develop non-motorized SPFs at county level, census tract, census block group and at corridor level for instance at road mid-blocks areas. Transferability of the model to other state is possible if proper local calibration factors are applied. The Safety Performance Function developed are applicable at any signalized urban intersection in Michigan that comprises of collector and arterial roads.

The thesis has seven chapters : Introduction(Chapter 1), Literature review (Chapter 2), Site selection(Chapter 3), Data collection(Chapter 4), Development of surrogate measure of non-motorized exposure(Chapter 5), Comparison of Classical and Bayesian approach in developing non-motorized SPFs (Chapter 6) and Conclusions (Chapter 7)

CHAPTER 2

LITERATURE REVIEW

Overview

This section will cover a review of past studies that have focused on different aspects of non-motorized safety as listed below

- Non-motorized performance measures that use crash data
- Different exposure measures used in past studies to account for level of risk that the pedestrians and bicyclists experience as they interact with other road users.
- The use of factor analysis as part of structure equation modeling in explaining factors that are associated with pedestrian and bicyclist crashes.

Non-Motorized Performance Measures

Performance measures for non-motorized safety refer to the factors that can be used to quantify the level of risk that pedestrians and bicyclists are experiencing for a given roadway environment. Over the past years, different performance measures have been developed from a relative simple to complicated ones. Names have been assigned to those performance measures depending on type of data and methodologies that were used. With regards to data, performance measures have been mainly developed using crash data, behavioral data and safety ratings. This review will mainly focus on studies that have used crash data for developing non-motorized safety performance measures.

Level of risk experienced by pedestrians and bicyclists on given road infrastructure have been often quantified using non-motorized crash data. Crash data are observed incidences and therefore represent the actual facts. However, they are rear and

random events. Therefore, it has been a challenge to develop robust modeling approach to compute the observed variation of non-motorized crashes in given location.

Using non-motorized crash data, safety performance measures have been developed using main two approaches as summarized in Table 1

- Quantifying non-motorized risk by normalizing the crash data with the exposure measure such as pedestrian volume, distance walked and time spent walking.
- Model development that relates number of non-motorized crashes with roadways, demographic, social economic and non-motorized facility characteristics. Performance measures developed using this approach are commonly referred as Safety Performance Functions (SPFs).

Table 1 Non-Motorized Safety Performance Measures Findings from Past Studies

Author	Modeling approach	Model outcome
Schneider et al., 2010	Pedestrian SPFs at signalized intersections Use of Crash rates to quantify risk (crashes per 10 million pedestrians crossing) Negative binomial regression to identify geometric characteristics that has significant relationship with pedestrian-involved crashes.	Factors associated with increase in pedestrian crashes • Vehicle volume • Number of pedestrians crossing • Number of right turn movements • Non-residential driveways within 50 feet • Commercial properties within 0.1miles • Percentage of young residents (age <18 years) within 0.25miles Raised medians was associated with decrease in pedestrian crashes
Nordback, K., Marshall, W. E., & Janson, B. N., 2014	Bicycle SPFs at intersections Negative binomial model using generalized linear model with log link.	Increasing bicycle crashes were significantly associated with: • Bicyclist volume (Annual Average Daily Bicyclist, AADB) • Traffic volume (AADT) Intersections with more than 200 entering cyclists had fewer collisions per cyclist. This demonstrated safety in number concept.
Minikel, 2012	Relative collision rate for Bicycle facility running parallel to the arterial	Collision rates on bicycle boulevards are 2-8 times lower than bike facility that were parallel or adjacent arterial routes. From literature, factor associated with diminished bicyclist safety • High vehicle speed and volume • Presence of heavy vehicles

Table 1-Continued

Author	Modeling approach	Model outcome
Oh et al., 2013	Poisson regression model-Pedestrian intersection SPF Negative binomial Regression-Bicyclist intersection SPF	Increase in pedestrian crashes at intersection were significantly related with: • Decrease in total number of lanes at minor roads • Increase in total number of entering vehicles at the intersection • Increase in number of bars • Decrease in number of people with graduate degree within a quarter mile Increase in bicycle crashes at intersection were associated with: • Number of right turn lanes on the major approach • Bicycle volume • Average daily traffic volume • Presence of bus stop • Business land use
Oh et al, 2013	Negative binomial Method was used for both pedestrian and bicycle midblock SPFs	Increase in pedestrian crashes at the midblock was significantly associated with • Increase in number of access points • Increase in Average Daily Traffic • Increase in pedestrian volume • Decrease in speed limit • Increase in length of the segment Increase in pedestrian crashes at the midblock was significantly associated with: • Increase in bicycle volume • Decrease in speed limit • Increase in number of bus stop

Table 1-Continued

Author	Modeling approach	Model outcome
Turner et al, 2011	Generalized Linear Model-Poisson and Negative Binomial SPFs were developed by crash type	Increase in crashes was significantly related with the following variables: • Increase in bicycle and vehicle volumes • Absence of advanced stop boxes • Increase in intersection depth • Decrease in cycle lane width, curbside lane width • Increase in midblock length
Jonsson, 2013	Non-motorized SPFs for midblock SPF for bike-bike, pedestrian alone using crash and hospital data. Generalized Linear model-Negative binomial distribution was used for modeling	Variables that were significantly associated with the increase of pedestrians crashes include • Segment Length • Traffic volume • Mixed land use
McArthur, A., Savolainen, P., & Gates, T. (2014).	SPF for child pedestrian at school zone(1mile radius) Negative binomial distribution	• Census data; Average Family Size, Children Ages 5 to 14 (increase crashes), Average Parents per Household (decrease crashes), Median Family Income $1000 (decrease crashes), population density (increase crashes) and proportion of non-whites households (decrease crashes). • Number of students enrolled (increase crashes) • Schools located on Local Roadway (increase crashes)

Exposure Measures for Pedestrians and Bicyclists

Federal Highway Administration and National Highway Traffic Safety Administration (NHTSA) has identified bicyclist and pedestrian exposure as one among top four most important research area (Hedlund, 2000). Planners and safety advocates have been using crash data alone in assigning risks that are associated with pedestrians and bicyclists at different facilities. This has led to misallocation of efforts to improve the non-motorized safety. Better comparison of risk across different facilities and modes of transportation could be obtained using non-motorized crashes normalized by either of the following

- Population density
- Number of pedestrians using the facility
- Time spent walking/bicycling
- Distance walked/cycled
- Number of trips
- Other surrogate measures such as number of potential collisions

In essence, there is no single measure that is most suitable to represent pedestrians and bicyclists exposure to traffic unless there was continuous monitoring of pedestrians and bicyclists movements at all time. The choice of exposure is dependent upon the intended purpose of the study. For example, time spent walking will be suitable when evaluating pedestrian risk at different transportation modes. Distance travelled by a pedestrian will be preferred when analyzing the effectiveness of the sidewalks. (Greene-Roesel et al, 2010).

Population Data

Most of the time it is represented as population density in a particular geographic unit. It is used with the underlying assumption that crashes between pedestrians or bicyclists with a motor vehicles will likely to occur as number of residents increase in a given area. It has been widely used, as it is readily available from census data.

It is recommended not to use this exposure measure unless it is impractical to obtain other granular measure of exposure. It is a crude measure of pedestrian and bicyclist exposure and only provide course picture of non-motorized safety. Malino (2000) commented on the insensitivity of population density to location specific factors such as changes in travel behaviors of bicyclists and pedestrians. It also assumes non-motorized exposure is uniform for a given population and does not account for the number of people who actually walk or bike. Distance and time a pedestrian or bicyclist is exposed to traffic are not taken into consideration.

Pedestrian/Bicyclist Volume

Number of pedestrians/bicyclists observed in a roadway at a given location in a specified duration. This exposure measure is usually incorporated in non-motorized safety studies as hourly count, or it can be annualized to account for the time of the day, day of the week, and month of the year.

It can be collected using different ways such manual count, video data and through other sophisticated technologies such as the use of active and passive infrared, inductive loops, pneumatic tubes and computer visioning. The choice of which method to use for counting will depend on the purpose of the count, required level of accuracy and overall cost. Statistical models have been developed which relates pedestrian and bicycle

volume with geometric characteristics of the road, facility information, socio-economic and demographic factors.

Raford et al., (2003) developed a space syntax pedestrian volume-modeling tool for the Oakland city in California. The method utilizes data such as connectivity of street grids, population density, employment density and pedestrian count at some key locations within the pedestrian grid network. The space syntax software correlates and extrapolates the aforementioned data to estimate pedestrian volume at street level.

Nordback et al., (2014) used negative binomial volume model to estimate bicycle hourly count. Independent variables were hourly temperature, parameter to account for working and non-working days in a year, solar radiation and school days. Available continuous count were used to calibrate the model.

Oh et al (2013) developed a model to estimate pedestrian and bicyclist volume as the function of land use, demographic and socio-economic characteristics. Based on the nature of sample data collected, log-linear model and negative binomial model were used for developing pedestrian and bicyclist volume respectively.

Qin and Ivan (2001) used generalized linear regression model to predict pedestrian volume in rural areas as the function of population density, site characteristics, demographic characteristics, land use characteristics and roadway characteristics.

Apart from modeling approach, the product of pedestrian volume and traffic volume at intersection has been used in evaluating the risk associated with variety of pedestrian characteristics and behaviors (Davis et al, 1987 and Tobey et al, 1983). The shortcoming of this exposure metric is that it does not account for the time of separation and how close a motorist is from a pedestrian. A situation might happen when a pedestrian was crossing

the road at different time when motorist passes and pedestrian might be walking far away from the moving traffic thus reducing the chances of the crash to occur (Molino et al, 2000).

Number of Trips

Number of trips made by pedestrians or bicyclists regardless of the time and distance travelled. This exposure metric can be obtained from survey data such as National Household Travel Survey (NHTS), U.S. census journey to work and America Community Survey. It is mostly used to assess the changes in non-motorized behavior across different jurisdictions. When using number of trips as an exposure measure it offers flexibility in analysis since trips can be analyzed at individual, household or location level. However, since most of the information is from survey, the reliability of the data is usually questioned as most of the non-motorized trips are underreported in surveys (Schwartz, 2000).

Distance Travelled

This is the distance that pedestrians or bicyclists travels while exposed to vehicular traffic. Mostly expressed as million person miles travelled when analyzed at individual level (Chu, 2003). It can be obtained in aggregated format by summing their distances travelled in a given defined area to get total miles travelled.

Molino et al (2012) estimated annual pedestrian and bicyclist exposure for Washington DC defined as 100million pedestrian/bicyclist mile. Distances that the pedestrian and bicyclists travelled on the shared facility with motor vehicles and 15 min

raw count data were used for development of this exposure metric. Using distance travelled to account for non-motorized exposure also has its own setbacks. It is not all the time that pedestrians are exposed to traffic when walking. Aggregating distance travelled by pedestrian/bicyclist in a certain geographical unit might overestimate the actual level of exposure to traffic that pedestrians are experiencing. In addition, it does not account for the difference in speed among individuals who are walking which could moderate the individual level of risk to traffic (Chu, 2003).

Time Spent Walking/Bicycling

Time taken walking or bicycling while exposed to vehicular traffic. This exposure metric has been used in comparing pedestrian risk across different transportation modes and in different social groups based on age, and sex (Greene-Roesel et al, 2010). Other useful application can be on quantifying risks that pedestrians are facing while crossing the intersection. Knoblauch et al (1996) suggest time spent crossing at intersection can be a better representation of exposure than a volume count because it takes into account pedestrian age, gender, weather condition, compliance with signal control and signal length. However, it is difficult to obtain this granular data when dealing with large geographical area. Always cost constrain is an impediment for collecting this exposure metric.

The Use of Structural Equation Modeling in Traffic Safety Studies

Structural Equation modeling is the multivariate technique that has been applied for creating and testing the causal models. It is a combination of confirmatory factor analysis, path analysis and regression analysis. In most cases, SEM is used as confirmatory tool that test the theory the researcher has hypothesized during model construction. For this reason, the researcher has to establish the causality between different variables involved in the model. SEM will then test how well the sample used by the researcher support the model specification. Schumacker et al (2004) provide a good introduction to structural equation modeling for beginners. Step by step procedures are elaborated on how to develop the structure equation modeling including model specification and identification, model fit, model estimation, testing and assessing the goodness of fit. It is not the aim of this study to explore in details such steps. Rather the goal is to leverage the benefits of SEM in developing tools for assessing non-motorized safety.

SEM has been widely used due to its ability to model complex phenomena, incorporate latent variables in the model and advance in statistical software with minimal coding efforts. Latent construct can be estimated in the model as a function of measurable variables.

Endogeneity effect among variable is explicitly accounted in the process of explaining complex phenomenon between variables using SEM. Endogeneity exists when there is a loop of causality between variables. In traffic safety studies that involve modeling of crash frequency, often times researchers have been getting results which can be easily judged as counterintuitive. A good example was the one provided by Jonsson

(2005) whereby road having low speed limit were highly associated with high non-motorized crashes as compared to high speed limit roads. This can be due to high pedestrian levels in those low speed roads, which in turn increases the conflicts between vehicular movements and pedestrians.

Example of Structural Equation Modeling Applications in Traffic Safety

Modeling of Crashes

Wang, K., & Qin, X. (2014) used SEM to model severity of single vehicle crashes. Force and speed were introduced as the latent variables that in turn were hypothesized to influence the crash severity. Manifest variables that were used to measure force include the types of object that were hit by the vehicle. Speed as latent construct was estimated using roadway, weather and lighting condition, gender and age. By using this model technique, it was possible to explain some of the relationship that could not be unraveled using normal ordinal models. Inclement weather, poor lighting condition, poor pavement surface condition were found to reduce speed (latent variable) which in turn reduced the injury severity.

Initially, SEM was designed for continuous variables whereby the estimation was done in a sample variance-covariance matrix. Therefore, it was impossible at that time to incorporate other data format such as nominal, ordinal and intervals. Overtime SEM has been modified to handle the aforementioned data format but introducing a link function which defines the type of data used. Application of this type of modification can be found in the study done by Xie et al (2016). They estimated the effect of secondary collision on injury severity levels using SEM. Injury severity is ordinal in nature and therefore had to

be specified in the model. SEM results were compared with ordered probit model. The ordered probit model tends to overestimate the safety effect of confounding variables by lumping their direct and indirect effects. Whereby, using SEM it was possible to separate direct and indirect effects of confounding variables that were related directly to crash severity and occurrences of secondary collisions.

Modeling of Road User Travelling Behavior and Mobility

A study conducted by Kim (2003) used SEM to determine factors that were significantly associated with elderly mobility. Urban form was used as the latent construct estimated by retail employment density, population density, age, gender and household size. Likewise, mobility was measured by non-home activity time, travel time and travel distance of elderly persons. Structure model was used to unveil how urban form affect mobility of elder drivers. With the use of SEM, age and gender showed to have significant effect on older driver mobility. Whereby older women had less mobility than older men and it's more likely for a person to refrain his or her desire for travelling as the age increases.

Ranaiefar et al (2016) estimated bicycle ridership using SEM as a function of different demographic and environmental characteristics surrounding the bike sharing stations. By using SEM, it was possible to forecast origin-destination bike share ridership.

CHAPTER 3

SITE SELECTION

Sampling Strategy and Preliminary Data Collection

There are varieties of sampling strategies that can be used in selecting a sample size from a population. They range from crude sampling procedures such as random sampling which doesn't take into account the sampling error, to more sophisticated sampling techniques such as stratified random sampling.

For this study, whereby the main objective was to improve the methodology for developing statewide safety performance function for bicyclists and pedestrians at signalized urban intersections, it was necessary to come up with a sampling strategy that will represent the aforementioned target group. Details of the criteria that were used to come up with the sample size and selection technique were adopted from the procedure developed by Aggarwal (1988). Figure 1 below summarize the sampling strategy process

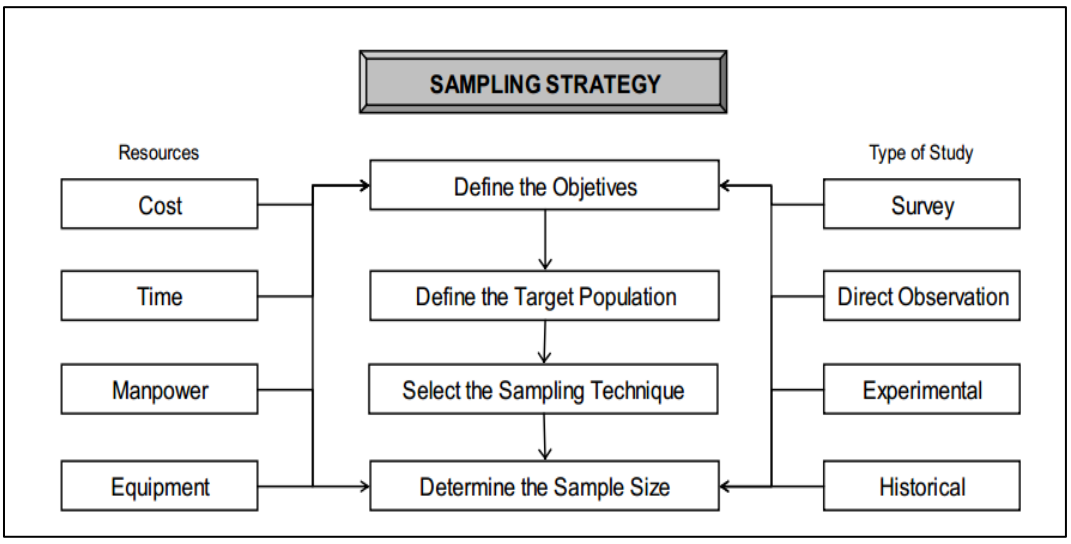


Figure 1 Sampling Strategy Process

In order to come up with sample size and sampling technique, available resources in terms of cost, time, manpower and equipment have to be evaluated. This should go concurrently with the proper understanding of the type of study that will be carried out to achieve the project objective. Upon consideration of all factors, stratified random sampling was selected as sampling technique for the study.

The following section will explain step by step procedure on how stratified random sampling technique was utilized, selection of sample size and finally descriptive statistics of crashes that occurred at signalized urban intersections in Michigan were discussed.

The choice of using signalized urban intersections as the case study was driven by data availability such as average annual daily traffic (AADT) and overrepresentation of non-motorized crashes in urban areas as compared to rural areas. Also there was no statewide non-motorized safety performance function developed for all signalized urban intersections in Michigan. The signalized urban intersections that were included in the analysis involve those joining arterial and collector road segments.

ArcGIS was used as the tool for identifying all urban intersections in Michigan so that the sample could be drawn from it. Sampling procedure, using ArcGIS is summarized below in a concise manner.

Identifying All the Collector and Arterial Road Intersections

Michigan road shapefile, which provide the statewide road network, was used in ArcGIS to identify all road intersection points. As mentioned earlier, only intersections connecting arterial and collector road segments were selected. Therefore based on Road

Functional Classification (NFC), three groups of intersections were identified which were Arterial-Arterial intersections, Arterial-Collector intersections and Collector-Collector intersections. Figure 2 below provide an example of this intersection types as identified in Michigan road shapefile. About four thousands signalized urban intersections were identified.

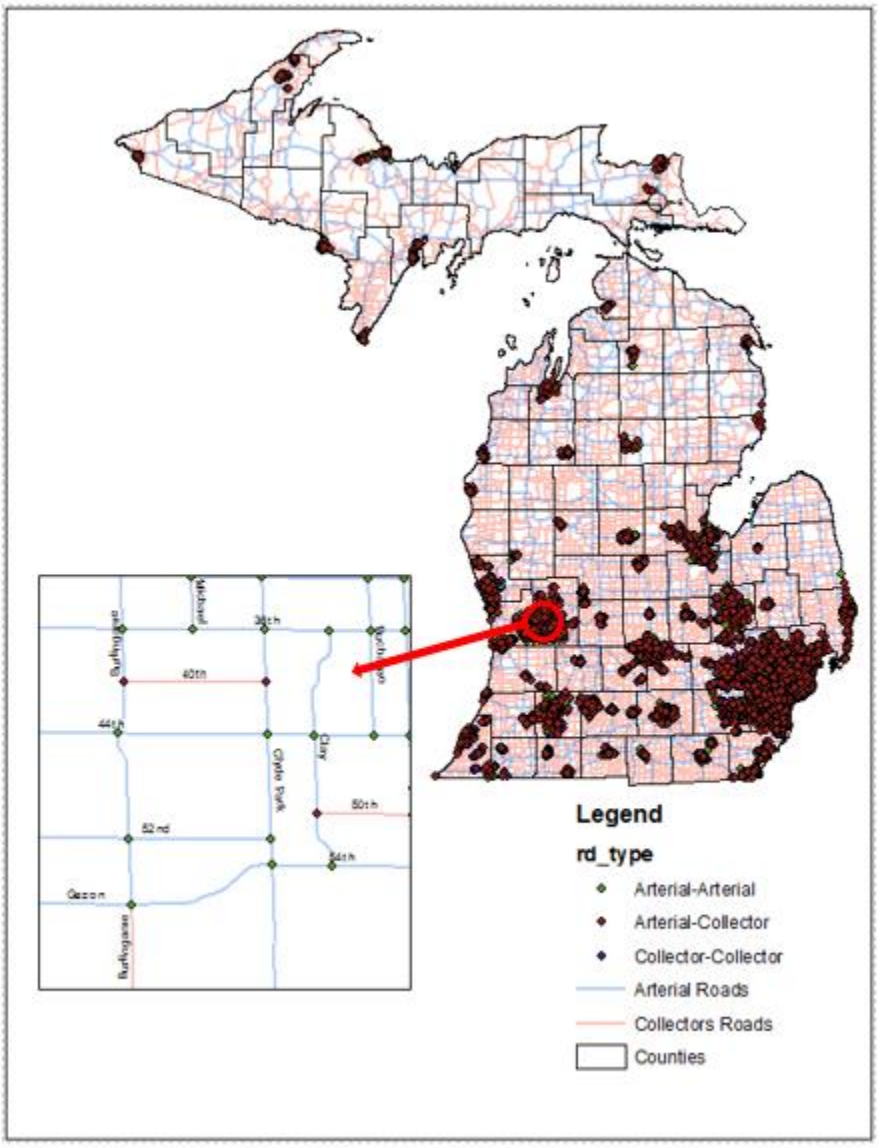


Figure 2 Distribution of Urban Intersections in Michigan

Subdividing the Target Population into Subgroup

Stratified random sampling require the target population to be subdivided into groups each having similar characteristics. To achieve this goal, parameters that were available at statewide level were used as shown in Table 2 below

Table 2 List of Parameters and Subcategories.

Parameters	**Subcategory**
Road function	Intersection connecting arterial roads
	Intersection connecting arterial road and collector road
	Intersection connecting collector roads.
Intersection type	Three leg intersection
	Four leg intersection
Urban population	5000-49,999
	50,000-199,999
	200,000-more
Non-motorized crashes: Pedestrians and Bicyclists crashes(2004-2014)	No crash observed
	1-5 crashes
	6-10 crashes
	>10 crashes

Based on subcategory for each parameter, seventy-two groups were created and each of the signalized urban intersection was placed to its corresponding group.

Sample Size Computation

The decision on the total sample size was based on the available resources such as time frame and manpower for data collection and cost associated with obtaining the data. In order determine the number of intersections each of strata will contribute to the total

sample size, a weighting factor was used. Formula for computing weighting factor and sample size for each of strata is shown below.

$$w_i = \frac{N_i}{N_{tot}}$$

$$S_i = w_i * N$$

Whereby

w_i = Weighted factor for intersections in group i

N_i = Number of intersections in group i

N_{tot} = Total number of intersections for all groups

S_i = Number of intersections drawn from group i

N = Required total sample size from all groups

Table 3 shows the number of intersections from each individual strata. Table 4 provide the output of the sampling process using stratified random sampling. Weighted factors were computed based on all signalized urban intersections in Michigan joining collector and arterial roads. The total number of signalized intersections were 3848 intersections and a sample size of 300 intersections was obtained from it.

Table 3 Placement of Signalized Urban Intersection Into Strata

Number of legs	Road function	Urban population	Non-motorized crashes(2004-2014)				
			No crashes	1-5 crashes	6-10 crashes	>10 crashes	Total
3-legged intersection	Arterial-Arterial	>200,000	272	215	13	1	501
		5,000-49,999	28	25	2	0	55
		50,000-199,999	50	46	3	0	99
	Arterial-Collector	>200,000	76	129	3	1	209
		5,000-49,999	11	15	1	0	27
		50,000-199,999	33	25	1	0	59
	Collector-Collector	>200,000	15	11	0	0	26
		5,000-49,999	2	0	0	0	2
		50,000-199,999	3	1	0	0	4
4-legged intersection	Arterial-Arterial	>200,000	339	793	188	74	1394
		5,000-49,999	57	115	22	2	196
		50,000-199,999	114	222	28	7	371
	Arterial-Collector	>200,000	186	369	49	8	612
		5,000-49,999	38	40	6	0	84
		50,000-199,999	63	76	10	1	150
	Collector-Collector	>200,000	19	26	4	1	50
		5,000-49,999	3	3	0	0	6
		50,000-199,999	3	0	0	0	3
Total			**1312**	**2111**	**330**	**95**	**3848**

Table 4 Sampled Signalized Urban Intersection from each Strata

Number of legs	Road function	Urban population	Non-motorized crashes(2004-2014)				
			No crashes	1-5 crashes	6-10 crashes	>10 crashes	Total
3-legged intersection	Arterial-Arterial	>200,000	22	17	2	1	42
		5,000-49,999	3	2	1	0	6
		50,000-199,999	4	4	1	0	9
	Arterial-Collector	>200,000	6	11	1	1	19
		5,000-49,999	1	2	1	0	4
		50,000-199,999	3	2	1	0	6
	Collector-Collector	>200,000	2	1	0	0	3
		5,000-49,999	1	0	0	0	1
		50,000-199,999	1	1	0	0	2
4-legged intersection	Arterial-Arterial	>200,000	27	62	15	6	110
		5,000-49,999	5	9	2	1	17
		50,000-199,999	9	18	3	1	31
	Arterial-Collector	>200,000	15	29	4	1	49
		5,000-49,999	3	4	1	0	8
		50,000-199,999	5	6	1	1	13
	Collector-Collector	>200,000	2	3	1	1	7
		5,000-49,999	1	1	0	0	2
		50,000-199,999	1	0	0	0	1
Total			111	172	34	13	330

Descriptive Statistics of Non-motorized Crashes at Signalized Urban Intersections

Trend of Pedestrian and Bicycle Crashes at Urban Intersections 2004-2014

Figure 3 below depicts the distribution of pedestrian and bicyclist involved crashes at signalized urban intersections from 2004 through 2014. Overall, there has been a decreasing trend of non-motorized crashes at signalized urban intersections from 2004-2014. Bicyclist-involved crashes occurred more at these intersections compared to pedestrian-involved crashes.

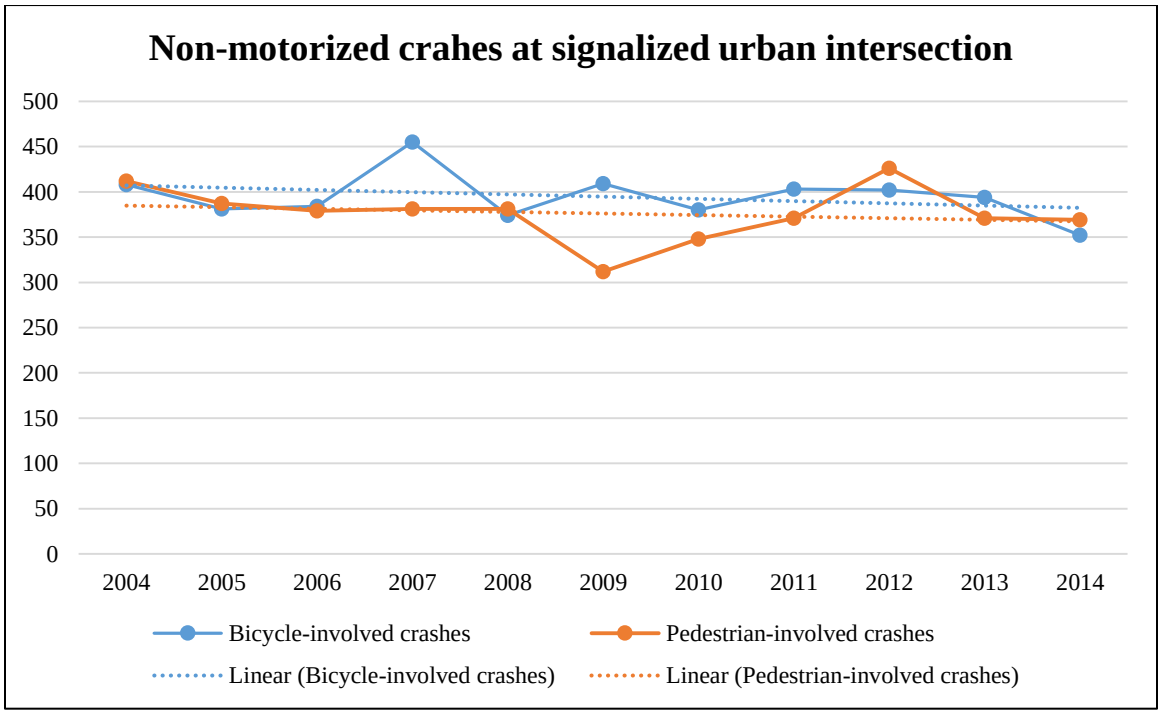
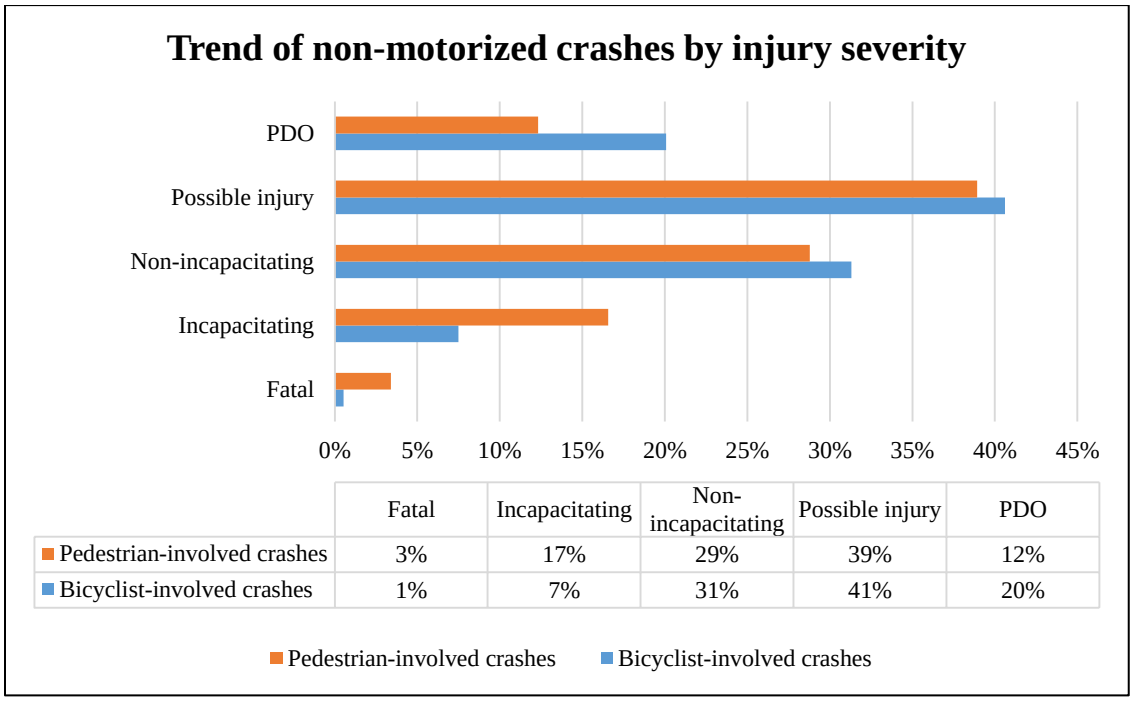


Figure 3 Trend of Non-motorized Crashes Signalized Urban Intersections

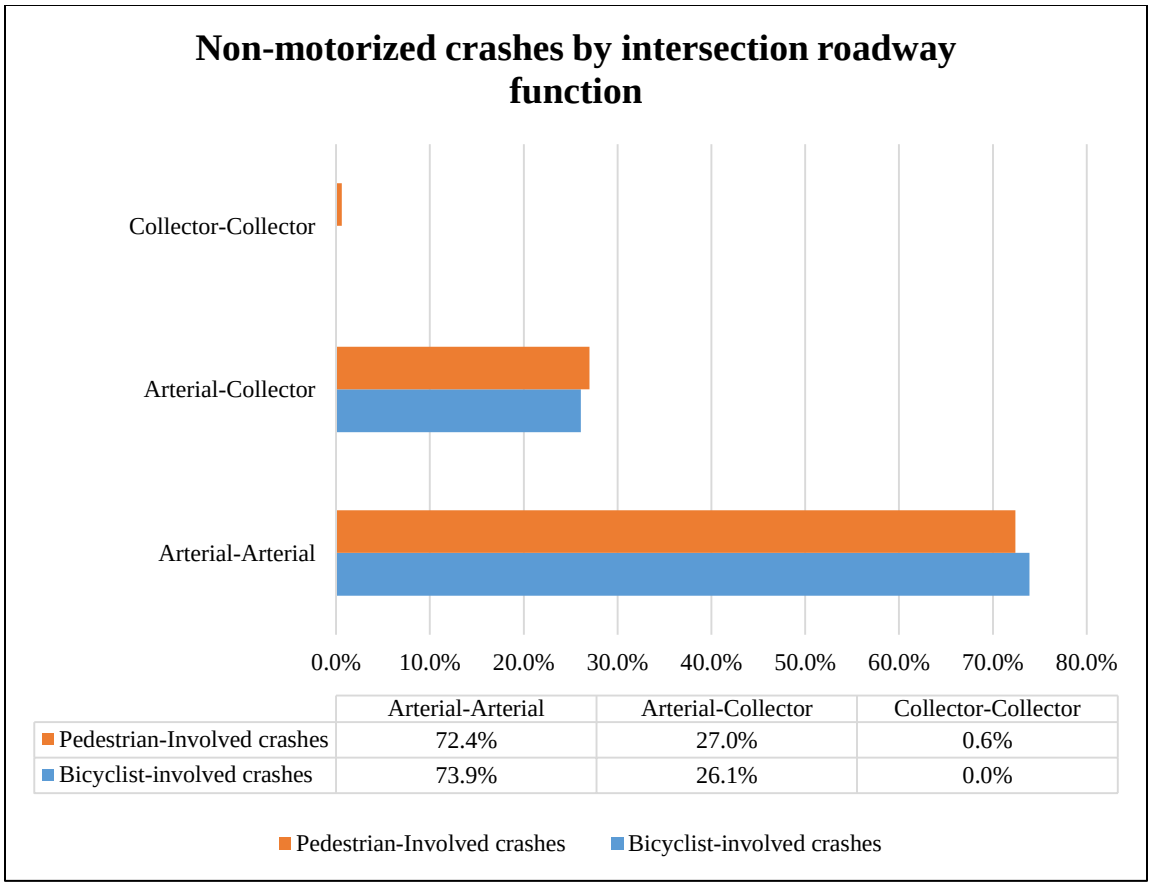
Distribution of Non-Motorized Crashes by Injury Severity Level

In total, there were 141 fatal pedestrian-involved crashes and 23 bicyclist involved crashes from 2004 to 2014 in all signalized urban intersections in Michigan connecting collector and arterial roads as indicated in Figure 4. For pedestrians, it represented 3% of all pedestrian crashes occurred at signalized urban intersections, while for bicyclist it represented 1% of all bicyclist crashes occurred at signalized urban intersections. Based on these statistics, it is evident that pedestrians are more likely to be involved in fatal crashes as compared to bicyclist in such locations. The analysis of fatal crashes distribution by intersection roadway functional type was then conducted. It was found that most of these fatal crashes occurred at intersections joining two arterial roads as shown in Figure 4. High speed associated with such arterial roads was likely to exacerbate the severity of non-motorized crashes.



	Fatal	Incapacitating	Non-incapacitating	Possible injury	PDO
Pedestrian-involved crashes	3%	17%	29%	39%	12%
Bicyclist-involved crashes	1%	7%	31%	41%	20%

Figure 4 Non-motorized Crashes at Signalized Urban Intersections in Michigan



	Arterial-Arterial	Arterial-Collector	Collector-Collector
Pedestrian-Involved crashes	72.4%	27.0%	0.6%
Bicyclist-involved crashes	73.9%	26.1%	0.0%

Figure 5 Non-motorized Intersection Fatal Crashes by Roadway Type

Figure 6 and Figure 7 show the distribution of pedestrian and bicycle involved crashes by roadway type and urban population. For both cases, nearly half of all crashes occurred at intersections joining two arterial roads located in areas with urban population greater than 200,000 people. Densely populated areas are more likely to have high pedestrians and bicyclists movements. The presence of arterial roads in such locations characterized by high volume of traffic,was likely to increase the chances of non-motorized crashes occurrence.

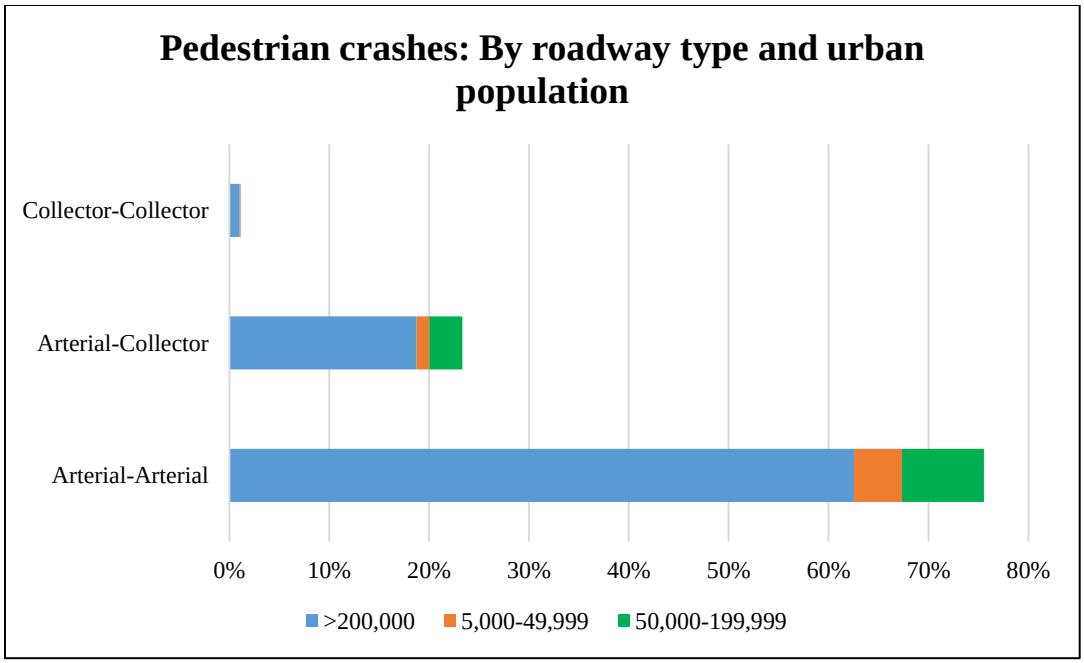


Figure 6 Pedestrian Involved Crashes by Roadway Type and Urban Population

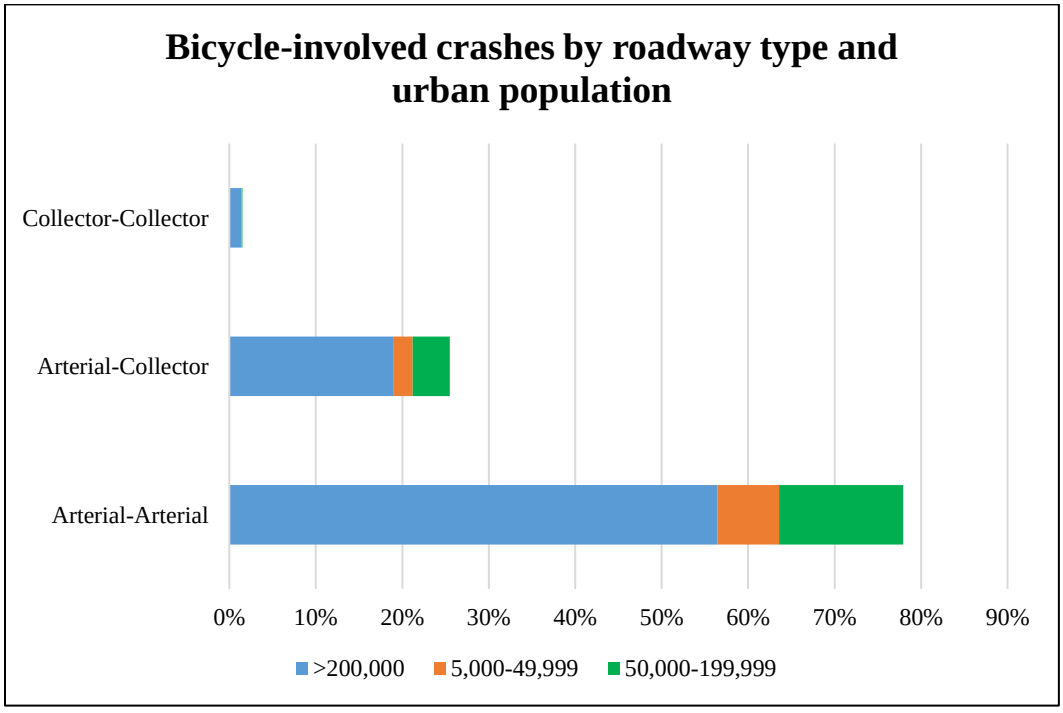


Figure 7 Bicycle Involved Crashes by Roadway Type and Urban Population

CHAPTER 4

DATA COLLECTION

This section covers methods and challenges that were encountered when gathering data. The collected data can be subdivided into six major groups

- Non-motorized crash data
- Demographic data
- Land use data
- Traffic volume data
- Road Geometry data
- Walk score index

Non-motorized Crash Data

Pedestrian and Bicyclist crash data for eleven years (2004-2014) were acquired from Michigan State Police (MSP) in the office of Highways Safety and Planning (OHSP). Only crash data attribute that were considered relevant for this research were kept in order to facilitate efficient handling and processing of the data in tools like ArcGIS. A buffer of 150ft, established from previous study (Dolatsara, 2014) for aggregating non-motorized intersection crashes, was used. ArcGIS provides spatial join option, which is the convenient means of aggregating crashes to each intersection. Figure 8 depicts how the buffer were created in ArcGIS

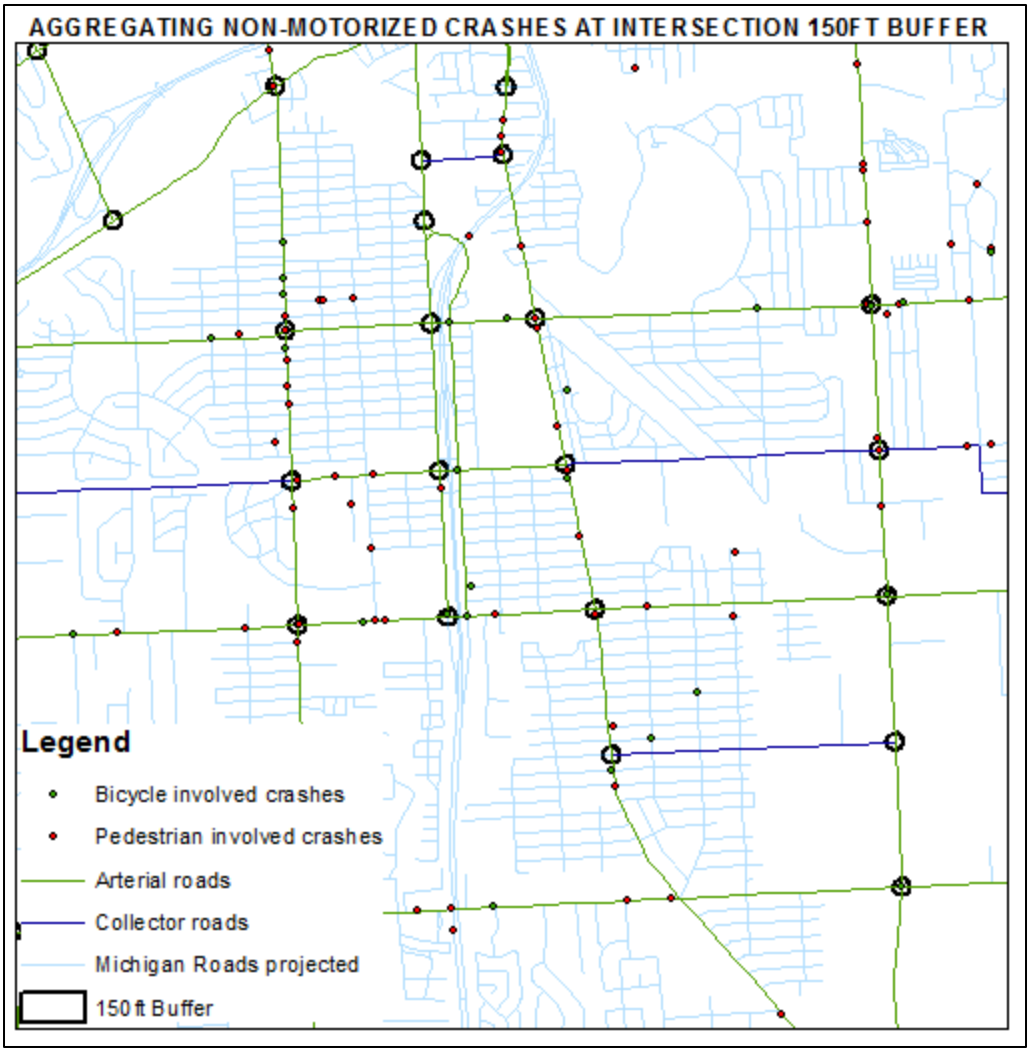


Figure 8 Aggregating Non-motorized Intersection Crashes

Land Use Data

Michigan Land use shapefile was used to obtain the land use data for the selected urban intersections. Four major categories of urban land use data were considered for the analysis. These were commercial, residential, industrial, institutional, and outdoor recreation as shown in Figure 9. Commercial areas include Central Business District (CBD) and neighborhood business.

In order to capture the dominant land use for a given intersection, weighted factors by area were used instead of dummy variables. Each land use area at the intersection was divided by the total area of blocks joining that intersection to obtain the weighted factors. Summation of weighted factors for all land use type in a given intersection will then be equal to one.

In previous studies, intersection with more than one land use type was considered as having mixed land use not considering the fact that the proportions of each land use adjoining to intersection are different. Therefore, area proportion were used to come up with unbiased description of land use characteristics surrounding a given intersection.

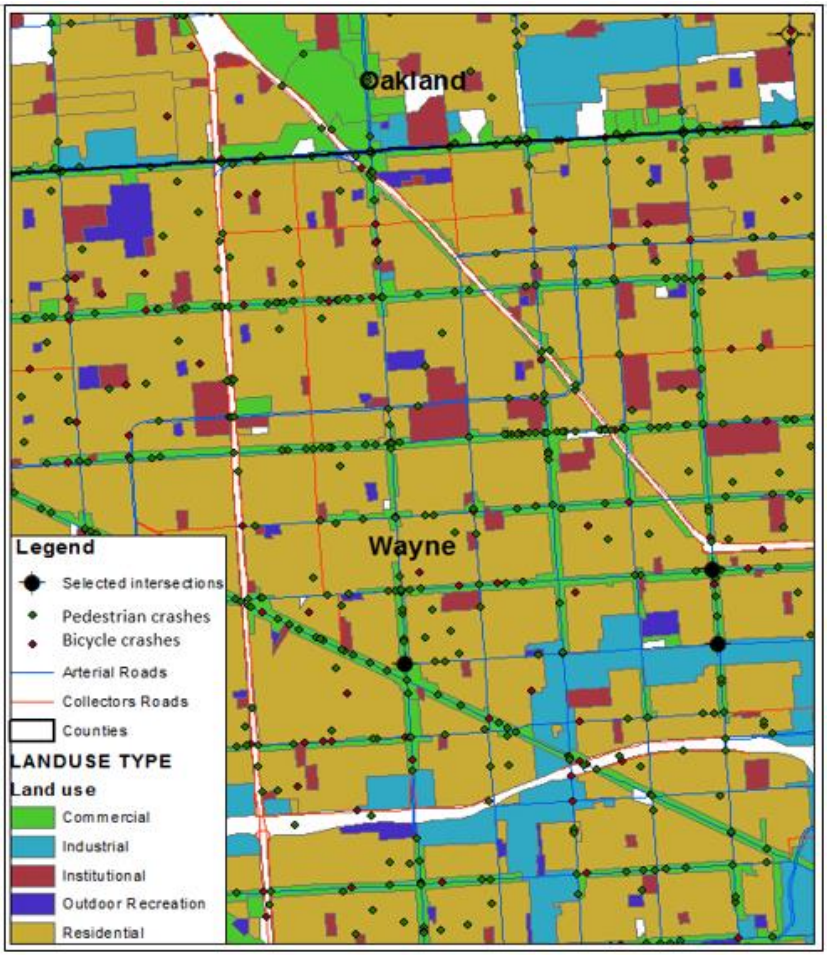


Figure 9 Land Use Distribution

Average Annual Daily Traffic (AADT)

AADT is one of the essential parameter when evaluating risks that a road users are experiencing when using road infrastructure. Most of the AADT data were collected from road commission Transportation Count Database System (TCDS) of each county. The database act like the central hub for storing and disseminating AADT data. Since the data are coming from different agencies within the same county, the data are first cleaned and validated before being available to the public. The level of details such as time of the day, hourly count differs across counties that have adopted this system. Below are some of the counties that have adopted this system in Michigan.

- Counties under SEMCOG (Wayne, Washtenaw, Macomb, Oakland, Monroe, St. Clair and Livingston)
- Counties under Grand Valley Metropolitan Counsel (Kent and Ottawa)
- Genesee County
- Kalamazoo County
- Eaton County
- Ingham County

With good cooperation from South Eastern Michigan Counsel of Governments (SEMCOG), it was possible to obtain AADT shapefiles for counties under SEMCOG. This help to automate the process of assigning AADT data to intersection segments. For other counties the data were recorded manually from their TCDS database.

Geometric Data

A list was created of all road geometric factors that have been established from past studies to have an influence on non-motorized crashes. Google earth was used as the main tool for obtaining road geometric characteristics. Below is the summary of main categories of roadway characteristics that were utilized in subsequent analysis.

*Signal information***:** Consist of the attributes such as signal control type, signal configuration (box or diagonal), left turn protection and no turn on red.

Intersection type: This provide information of whether the intersection was three legged intersection or four legged intersection

Lane uses information: This group consist of attributes that described the designated lane use for each approach. Lane use information such as number of exclusive through lane, number of shared through-right lane, number of shared through-left lane, number of exclusive right lane, number of exclusive through lanes and total number of outgoing lanes were recorded.

Pedestrian facility: For each approach, information about the presence of pedestrian facility was collected. To be more precise, the pedestrian facilities were subdivided into four categories as shown below.

- Pedestrian sidewalk on one side of the road separated from traffic
- Pedestrian sidewalk on one side of the road not separated from traffic
- Pedestrian sidewalk on both sides of the road separated from traffic
- Pedestrian sidewalk on both sides of the road not separated from traffic

The reason for this subdivision of non-motorized facility information was to capture different level of risk that each category of pedestrian facility will have. For example,

presence of sidewalk, which is not separated from the main road, is more dangerous than the separated sidewalk. Also providing sidewalk only on one side of the road might have an implication on non-motorized movements and the way they interact with traffic as compared to providing pedestrian facility at both sides of the road.

Bicycle facility: This include information about presence of bike lane and the position of bike lane for each approach at the intersection. For example, the bike lane can be in-between lanes or the far right side of the approach.

Figure 10 and Figure 11 provides plan view and the street view respectively, in one of the intersection included in the study. The plan view provided geometric characteristics of the intersection, lane designation and facility information while the street view provided the signal information.

Figure 10 Plan View of Intersection as Seen From Google Earth

Figure 11 Google Earth Street View of an Intersection with Signal Information

Walk Score Index

This variable has been used mostly in the field of urban planning, real estate and public health. Walk score Index measures walkability of a given point or area on a scale of one to one hundred. The points are given after analyzing different walking routes to the amenities that are nearby. Distance decay function is used to model score index. Amenity that have 5min walk get the maximum points and the points keep on diminishing up to zero after 30 min walk. In addition, walk score captures pedestrian friendliness of a given location by considering population density, block length and intersection density. Details of whole procedure can be obtained from the Walk score company official website, which developed the procedure (walkscore.com). Table 5 provides a description for different ranges of walk score. It ranges from car dependent areas to what is referred as walker's paradise. Figure 12 and Figure 13 shows two

examples of intersections, one situated in a car dependent community while the other one situated in walker's paradise community.

Table 5 Definition of Walk Score Index

Score	Definition
90-100	Walkers' Paradise Daily trips do not require a car
70-89	Very Walkable Most trip can be accomplished on foot
50-69	Somewhat Walkable Some trips can be accomplished on foot.
25-49	Car Dependent Most trips require a car
0-24	Car Dependent almost all trips require a car

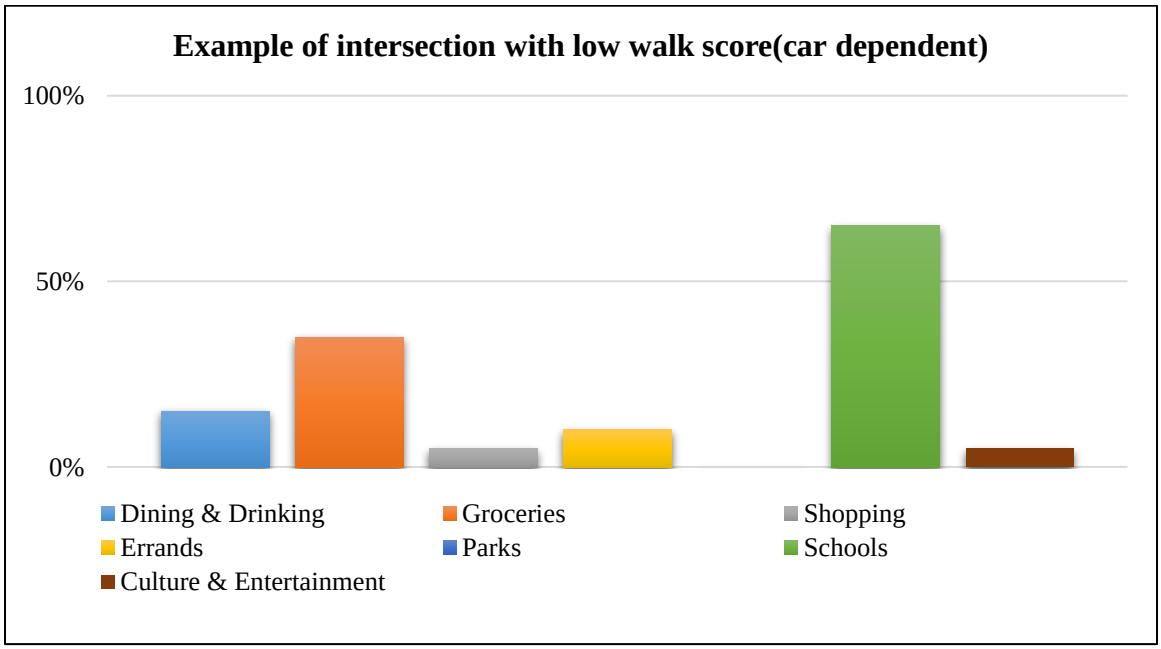


Figure 12 Eastern Ave SE @ 60th St SE Intersection with Walk Score of 17

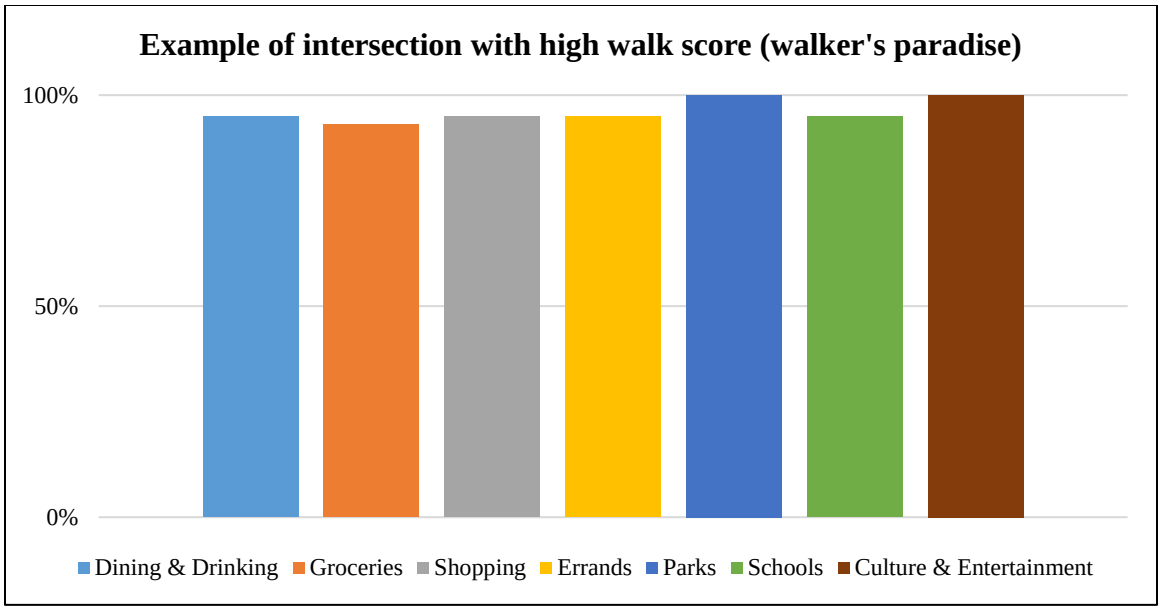


Figure 13: E Fulton St @ Lafayette Ave NE with the Walk Score of 91

Demographic Data

Demographic information at census block level using census shapefile were obtained for all selected urban intersections. Information that was extracted include population by age, educational status, poverty level, means of transportation to work and household income. Figure 14 provides part of Kent County were some of the signalized urban intersections were included in the sample size. It can be observed that non-motorized crashes clustered in areas with high population density, high percentage of people below poverty level area and in areas with relatively high percentage of people who are walking and biking as means of transportation to work.

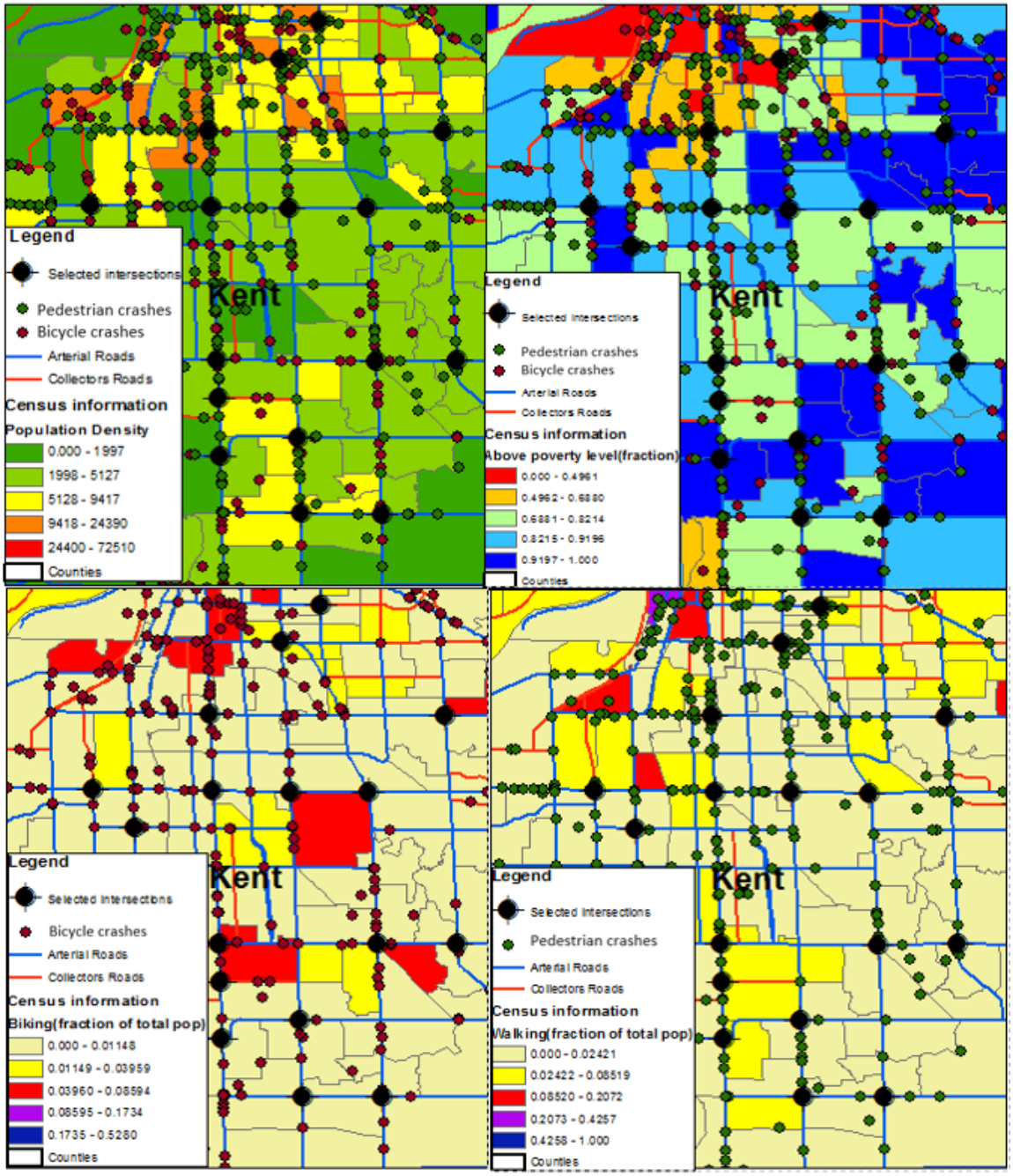


Figure 14 Census Information Extracted from Michigan Census Shapefile

CHAPTER 5

DEVELOPMENT OF SURROGATE MEASURE FOR NON-MOTORIZED EXPOSURE

This section describes how the surrogate measure for pedestrians and bicyclists were developed. The proxy measure can be used to provide pedestrian and bicyclist level as a function of demographic, walkability, bikeability and roadway characteristics of a particular location. The procedure involves the use of factor analysis that incorporate latent variables. This chapter provides a theoretical background of factor analysis and finally how it was used to come with non-motorized proxy measures.

Factor Analysis

This is the multivariate technique that aims at explaining the joint variation and covariation of observed variables using less number of unobserved constructs which are called factors. It is a means of reducing dimensionality of correlated data as it tends to clusters variables into homogeneous sets. These sets of unobserved constructs are unmeasured since we don't have a single perfect measure to represent them. In some instances, they are difficult to measure because of data insufficiency and other practical reasons.

Factor analysis accounts for measurement error when relating observed variables with latent variables as opposed to methods that use Ordinary Least Square (OLS) approach. The error term expresses the percentage of observed variable variance that could not be explained by a factor. The estimation procedure utilizes maximum likelihood approach that estimate model parameters by minimizing the discrepancy

between the observed and predicted variance-covariance matrix. The parameters estimated from the factors analysis include factor loadings, observed variable error variances, factor variances and covariance. Factor loadings inform how each observed variable is related with the factor. It's a slope of regression coefficient between observed variable and a factor when presented in unstandardized form. When factor loadings are standardized, they represent correlation between a factor and an observed variable. Preference on which format of factor loading to used, depends on the type of study and intended outcome of the analysis.

Using matrix notation, factor analysis can be presented as

$$y_{nx1} = \Sigma_{nxm} F_{mx1} + e_{nx1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}_{nx1} = \begin{bmatrix} \lambda_{11} & \cdots & \cdots & \lambda_{1n} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \lambda_{n1} & \cdots & \cdots & \lambda_{nm} \end{bmatrix}_{nxm} \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix}_{mx1} + \begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ e_n \end{bmatrix}_{nx1}$$

Where

y_{nx1} = Observed variables matrix

Σ_{nxm}= Variance-covariance matrix that comprises of factor loadings, λ_{nm}

F_{mx1} = Factor Matrix

e_{nx1} = Error term

Estimation procedure of unknown parameters such as factor loadings and error term utilize Maximum Likelihood (ML) approach, which aims at minimizing the following function.

$$\Gamma_{ml} = ln|\Sigma| - ln|S| + trace|(S)(\Sigma^{-1})| - p$$

Where

Γ_{ml} = Log likelihood function

$|\Sigma|$ = Determinant of predicted covariance-variance matrix

$|S|$ = Determinant of observed covariance-variance matrix

p = Number of input indicators/observed variables

Trace= Sum of the diagonal values in the covariance-variance matrix

In ideal case whereby $|\Sigma| = |S|$, $(S)(\Sigma^{-1})$ will turn out to be an identity matrix in which its trace value will be equal to p. Hence the log likelihood function, Γ_{ml} will be equal to zero (Jaccard et al, 1996)

Model Specification

Due to unavailability of non-motorized volume counts, factor analysis was used in this study to estimate proxy measure of pedestrians and bicyclists volume at urban intersections. Observed variables that were used to form proxy measure of pedestrians and bicyclists exposure were selected based on prior research knowledge. Variables that were significant at 95% confidence level were retained in the final factor analysis model. Table 6 and Table 7 provide the descriptive summary of the variables that were significant for pedestrians and bicyclists factor analysis respectively. Figure 15 and Figure 16 provide a schematic diagram of significant observed variables for pedestrians and bicyclists factor analysis. The error terms ε for each observed variable was estimated in the process.

Table 6 Description of Variables Used in Estimation of Pedestrian Proxy Measure

Variable	Description	Mean	Std. Dev.	Min	Max
Percent using public transport	Percentage of people using public transit in a census block where the intersection is located	0.97	2.39	0	22.15
Population per square mile	Population density for a census block where the intersection is located	420.18	370.61	12.87	2384.90
Percent of poverty below	Percentage of people below poverty level in a census block where the intersection is located	13.47	14.20	0	83.72
Walking per square mile	Walking commuters density in a census block where the intersection is located	36.02	148.45	0	1671.94
Pedestrian facility	Dummy variable for the presence of pedestrian facility separated from roadway within 150ft intersection buffer	0.59	0.49	0	1
Walk score	Walk score index estimated using distance decay function	35.77	24.98	0	94
Proportion of commercial land use	Proportion of commercial land use by area	0.15	0.28	0	1

Table 7 Description of Variables Used in Estimation of Bicycle Proxy Measure

Variable	Description	Mean	Std. Dev.	Min	Max
Bike facility	Presence of bike facility (side path/bike lane) roadway within 150ft intersection buffer	0.60	0.49	0	1
Poverty level below	Percentage of population below poverty level in a given census block group where the intersection is located	13.44	14.19	0	83.72
Population per square mile	Population density for a census block where the intersection is located	419.05	370.75	0	2384.90
Speed limit major	Speed limit in the major intersection approach	42.83	8.98	25	70
Speed limit minor	Speed limit in the minor intersection approach	34.89	8.65	20	55
Proportion of commercial land use	Proportion of commercial land use by area where the intersection is located	0.15	0.28	0	1

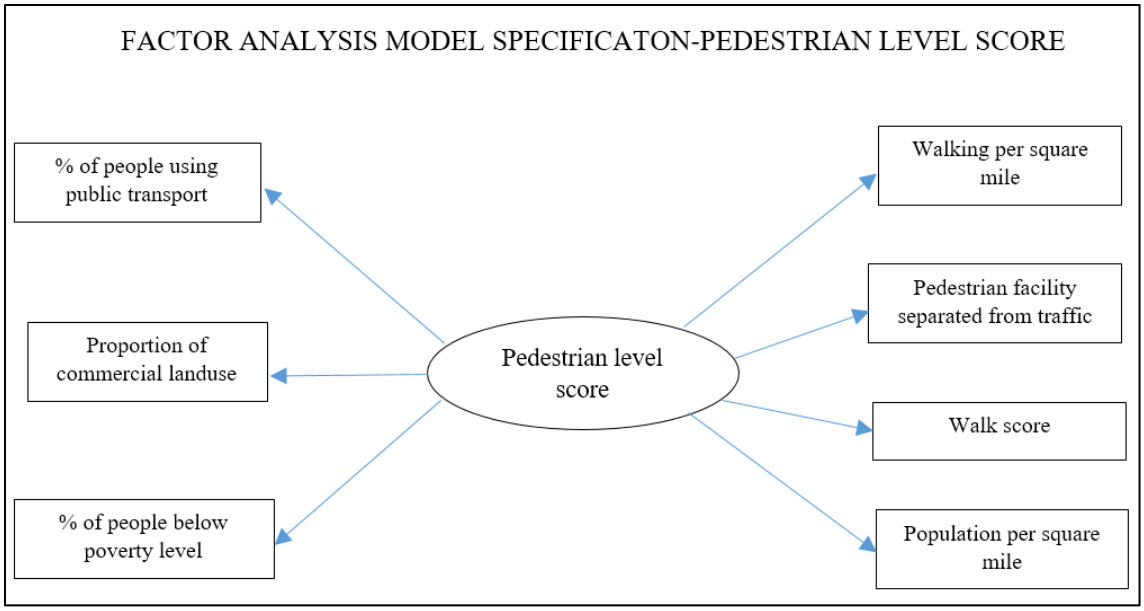


Figure 15 Schematic Diagram of Pedestrians Factor Analysis

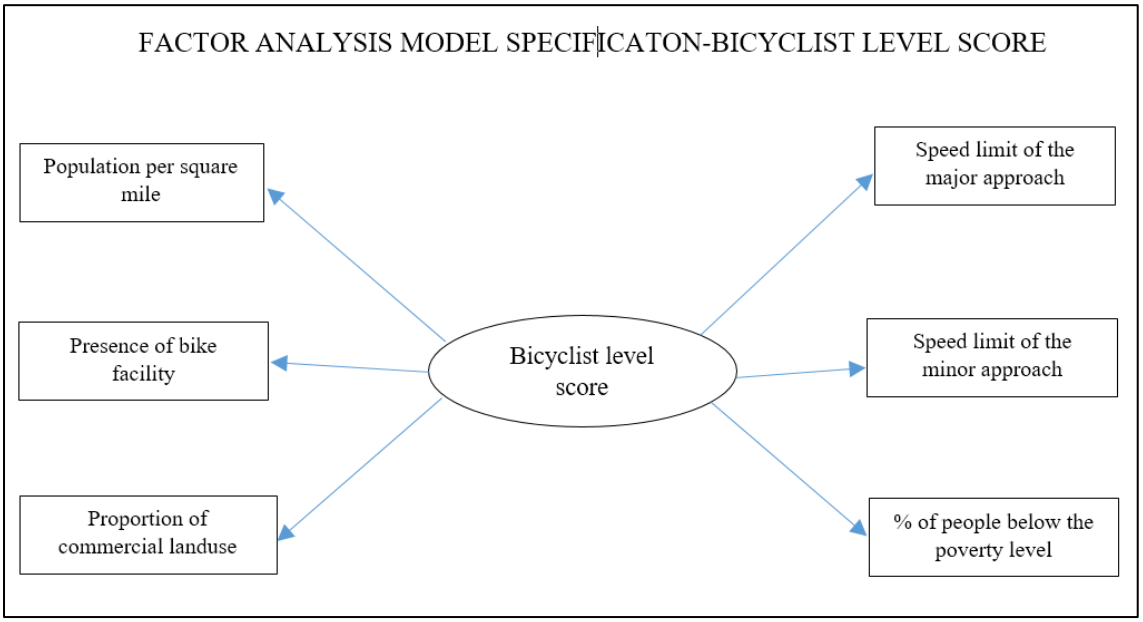


Figure 16 Schematic Diagram of Bicyclists Factor Analysis

Model Estimation

As summarized in Table 8, the increase in pedestrians level score index at a given intersection was manifested by the following factors; increase in percentage of people using the public transit in a given block group where the intersection was situated, population density, percentage of household below poverty level, number of workers commuting to their working places by foot per square mile, walk score index, proportion of commercial land use and presence of pedestrian facility separated from the roadway.

Table 8 Standardized Factor Loadings for Pedestrians Level Score

Variable	Standardized Coef.	Std. Err.	z	P>z
Percent using public transport	0.5397	0.0440	12.26	0
Population per square mile	0.6959	0.0345	20.17	0
Percent of poverty below	0.6131	0.0392	15.65	0
Walking per square mile	0.5299	0.0448	11.82	0
Pedestrian facility	0.2568	0.0545	4.72	0
Walk score	0.8347	0.0288	29.01	0
Proportion of commercial land use	0.3244	0.0518	6.26	0

Table 9 summarized the significant factor loading for bicyclist factor analysis. Bicyclist level score, a proxy measure of bicyclist volume was found to increase with the following factors; presence of bicycle facility which includes bike lanes and sidewalks, increase in percentage of people below poverty level, increase population density, lower speed limit in major and minor intersection approaches and increase in proportion of commercial land use by area in a given census block group were the intersection is situated.

Table 9 Standardized Factor Loadings for Bicyclist Level Score

Variable	Standardized Coef.	Std. Err.	z	P>z
Bike facility	0.3713	0.0547	6.79	0
Poverty level below	0.4860	0.0507	9.59	0
Population per square mile	0.5454	0.0496	11.01	0
Speed limit major	-0.7318	0.0415	-17.61	0
Speed limit minor	-0.6646	0.0423	-15.7	0
Proportion of commercial land use	0.1358	0.0601	2.26	0.024

Estimation of Bicyclists and Pedestrians Level Score

Pedestrians and bicyclists level score were then estimated from their respective significant observed variables. There are different methods in which the factor score can be estimated such as sum score by factor, weighted sum scores, regression scores, Bartlett Scores and Anderson-Rubin Scores. Distefano et al (2009) provides a good description of these factors giving applicability, pros and cons of each. For this study, the estimation procedure adopted was least squares regression approach that is similar to regression score developed by Thomson (1935).

Stata, which was the statistical package that was used in data analysis for this project, utilizes this approach in computing factor scores. In this method, the observed variables are centered to their respective means. The final factor score is the sum of the product between factor score weights and their respective observed variables. The factor score weights are obtained by multiplying the inverse of observed variable covariance matrix by factor-observed variables covariance matrix.

Mathematically estimation of factor score can be expressed as follows

$$f_i = (\Sigma^{-1} * \Lambda)$$

$$Factor\ score = \sum_{i=1}^{i=n} fi * (x_i - \bar{x}_i)$$

Whereby

f_i = the factor score weight for observed variable i

Σ^{-1} = Inverse of observed variable covariance matrix

Λ = Factor-observed variable covariance matrix

x_i = Observed variable i

$\bar{x}_i$ = Mean of observed variable i

Latent pedestrians level score can be computed as

$$\boldsymbol{Pedlevel} = 0.0707\left(perc_{publ} - 0.974\right) + 0.0008\left(pop_{sqmile} - 420.178\right) + 0.0153\left(pov_{tot_{blw}} - 13.473\right) + 0.0011\left(walking_{qmile} - 36.32\right) + 0.1233\left(ped_{faclty} - 0.586\right) + 0.0244(walkscore - 35.772) + 0.2828(pro_{comm} - 0.146)$$

Latent bicyclist level score can be computed as

$$\boldsymbol{Bikelevel} = 0.0415\left(bike_{facility} - 0.598\right) + 0.0021\left(pov_{tot_{blw}} - 13.44\right) + 0.0001(pop\_sqmile - 419.052) - 0.0086(speedlmt\_min - 34.893) - 0.0063(speedlmt\_maj - 42.828) + 0.0231(pro\_comm - 0.146)$$

CHAPTER 6

COMPARISON OF CLASSICAL AND BAYESIAN APPROACH IN DEVELOPING NON-MOTORIZED SPFS

Classical Approach

Using classical approach, parameters of pedestrian and bicycle safety performance function can be estimated using maximum likelihood approach. These parameters are unknown but fixed in a given sample. The probability assigned to a given parameter is considered as the proportion of times that a parameter will occur if an experiment were to be repeated in an infinite number of repetitions (Bolstad, 2013). In defining the precision of our estimates, confidence intervals at 95% are estimated for each parameter obtained in the model. This represent the percent at which the parameter to be estimated will fall into a given range under repeated sampling process.

This section describe how different count models were used in estimation under classical approach. Mathematical formulation are provided for goodness of fit measures that were used in classical approach for comparing model performances.

The counts model that were considered for the analysis are listed below:

- Poisson Regression Model (NRM)
- Negative Binomial Regression Model (NBRM)
- Zero Inflated Poisson Regression Model (ZIP)
- Zero Inflated Negative Binomial Model (ZINB)

Poisson model and negative binomial regression model have been used widely in most of the studies that analyze count data. Equidispersion assumption of Poisson regression model that the mean and variance are identical is often violated. That's why Negative

Binomial regression have been used more often compared to Poisson regression as it accounts for over-dispersion.

Zero inflated models were also considered as the potential fit to the data due to presence of excess zero at the selected intersections. For zero inflated models, there are two types of zero counts. The first type of zero count is predicted by the binary component of the model, whereby it shows locations that will always have zero count. The second type of zero count is predicted by the count model component whereby it shows location that are most likely but not always have zero counts.

Kwigizile et.al (2014) provided a good and simple formulation of four count models that were compared in this analysis as show below.

Poisson and Negative Binomial Regression

The probability of intersection i having pedestrian/bicycle crashes in a given time period can be written as:

$$P(y_i) = \frac{EXP(-\lambda) \cdot \lambda_i^{y_i}}{y_i!}$$

Whereby

λ_i is the Poisson parameter for signalized urban intersection i, which for this study it can be defined as the expected number of pedestrian/bicyclist crashes in eleven years period. This parameter is a function of predictor variables given as

$$\lambda_i = EXP(\boldsymbol{\beta} \boldsymbol{X}_i)$$

Where $\boldsymbol{\beta}$ is the vector of estimable parameter

Estimation of parameters deploy maximum likelihood method given as

$$LL(\beta) = \sum_{i=1}^{N} [-EXP(\boldsymbol{\beta X}_i) + y_i \boldsymbol{\beta X}_i - \ln(y_i!)]$$

Negative binomial regression, which handle cases where mean and variance of the count data are not equal, can be derived from the Poisson model. Generalizing Poisson model by introducing unobserved effect ε_i, the expected Poisson parameter becomes

$$\lambda_i = EXP(\boldsymbol{\beta X}_i + \varepsilon_i)$$

With $\lambda_i = EXP(\varepsilon_i)$ which is known as gamma distributed error term with mean of one and variance of α^2.

Upon modification of mean-variance relationship for expected number of pedestrian/bicycle crashes y_i becomes:

$$Var[y_i] = E(y_i) \cdot [1 + \alpha E(y_i)] = E[y_i] + \alpha E(y_i)^2$$

If α is significantly different from zero, then the bicyclist/pedestrian involved crashes are said to be overdispersed for positive α values and underdispersed for negative α values.

For overdispersion case, the resulting Negative binomial probability distribution becomes

$$P(y_i) = \frac{\Gamma\left(\left(1/\alpha\right) + y_i\right)}{\Gamma\left(1/\alpha\right) y_i!} \left(\frac{1/\alpha}{\left(1/\alpha\right) + \lambda_i}\right)^{1/\alpha} \left(\frac{\lambda_i}{\left(1/\alpha\right) + \lambda_i}\right)^{y_i}$$

Whereby

$\Gamma(x)$ is a value of the gamma function.

α is an overdispersion parameter

y_i is the number of pedestrian/bicyclist involved crashes for intersection *i*

Zero Inflated Models

Zero inflated models are used when there is excess number of zero in the data that tends to violate assumptions used in Poisson or Negative binomial model formulation. For ZIP model the probability for the two component (binary logistic and Poisson regression) can be estimated as follows (Lord et al, 2005)

$$\Pr(y_i = 0) = p_i + (1 - p_i)e^{y}$$

$$\Pr(y_i > 0) = (1 - p_i)\frac{e^{-y}y^{n}}{n!}$$

The probability of zero pedestrian/bicyclist intersection crashes for the binary component of the ZINB model can be computed as:

$$Pr(y_i = 0) = p_i + (1 - p_i)\left[\frac{1/\alpha}{1/\alpha + \lambda_i}\right]^{1/\alpha}$$

The count component of the model with the probability of $y_i > 0$ can be computed as

$$\Pr(y_i = y) = (1 - p_i)\left[\frac{\Gamma\left(\left(1/\alpha\right) + y\right)\psi_i^{1/\alpha}(1 - \psi_i)^{y}}{\Gamma\left(1/\alpha\right)y!}\right]$$

With $\psi_i = \frac{1/\alpha}{1/\alpha + \lambda_i}$

Goodness of Fit Tests

Goodness of fit test that were used to analyze how well the model fits the data are summarized below

- Akaike's Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Vuong test
- Residual probability plot
- Root Mean Square Error (RMSE)

The selection of best model was based on collective assessment of all goodness of fit measures. AIC, BIC and Vuong test were used to test within sample goodness of fit. Residual probability was used for within sample and for cross validation. Hilbe (2011) provide good description and application of BIC, AIC, Vuong test and Residual probability plot. Mathematical formula for the given goodness of fit measures are provided below

Akaike's Information Criterion (AIC)

$$AIC = \frac{-2L + 2k}{n}$$

Bayesian Information Criterion (BIC)

$$BIC = -2L + klog(n)$$

With

k=Number of predictors including the intercept

n= Number of observations

L= Model log-likelihood

Vuong Test

It tests whether the zero inflated models are preferred over non-inflated models. It is the most commonly used test, despite invention of other tests serving a similar purpose. It is conservative and therefore reduces the chances of making incorrect decision (Clarke 2007).

It is given as the log ration of the sum of probability for each observation computed as

$$\phi_i = \ln\left(\frac{\sum_i P_1(y_i/x_i)}{\sum_i P_2(y_i/x_i)}\right)$$

With Vuong test statistics is a calculated as

$$V = \frac{\sqrt{N}(\bar{\phi})}{SD(\phi_i)}$$

Where

$P_1(y_i/x)$= Probability of observing pedestrian/bicyclist involved y crashes on the basis of variable x for model i (inflated model)

$P_1(y_j/x)$ = Probability of observing pedestrian/bicyclist involved y crashes on the basis of variable x for model j (Non-inflated model)

$\bar{\phi}$ = Average of the log ratios

$SD(\phi_i)$ = Standard deviation of the log ratios

If V is greater than 1.96, model inflated model is favored while if V is less than -1.96, model non-inflated model is favored

Residual Probabilities

It was computed as the difference between the average observed probability and average predicted probability for each pedestrian/bicyclist observed crash count at signalized urban intersection. The model with the best performance has residual probabilities close to zero for all observed pedestrian or bicyclist crash counts

Root Mean Square Error (RMSE)

Is the square of the difference between observed values and the values predicted by a model. Individual differences between observed and predicted values are normally called residuals. It can be computed as

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$$

With

$\hat{y}_i$ = predicted pedestrian/bicyclist crashes for intersection i

y_i = observed pedestrian/bicyclist crashes for intersection i

N= total number of intersections

Bayesian Approach

Alternative method for modeling non-motorized crashes is the use of Bayesian approach. This method emanates from Bayes theorem that tends to update the belief about the probability distribution of the parameter based on the prior knowledge. Compared to the frequentist approach, Bayesian modeling hypothesize a parameter as a random variable in a given sample size. Instead of giving out the point estimates as how frequentist approach does, it estimates the distribution of a given parameter.

The main advantages of Bayesian approach as compared to frequentist approach are:

- The ability to use prior knowledge about a parameter and the new evidence from the observed data to update a belief about the parameter. This can be the most useful application in development of SPFs whereby the results from prior studies can be incorporated into the new studies in order to obtain SPFs that are more reliable. The new SPFs developed will represent a collective research effort from different studies over time. With frequentist approach, it is not possible to do this.
- The ability to use relative small sample size as compared to frequentist approach and yet obtain the SPFs with reliable predictive capabilities.

Derivation of Bayesian Inference from Bayes Theorem

Bayesian inference has its root from Bayes theorem, which uses the conditional probability theorem. Suppose we have two events namely A and B. Event A can be thought as the observable event and B as the unobservable event that partition the universe U.

$$p(B/A) = \frac{p(A \cap B)}{p(A)}$$

$p(A)$ is the marginal probability of event A which is the sum of probability of disjoint parts. Suppose we have a set of disjoints parts $B_1, B_2, B_3 \dots \dots \dots B_n$ partition the universe U.

The marginal probability of A

$$p(A) = p(A \cap B_1) + p(A \cap B_2) + p(A \cap B_3) + \cdots \dots \dots \dots p(A \cap B_n)$$

$$p(A) = \sum_{j=1}^{n} p(A \cap B_j) = \sum_{i=1}^{n} p(B_j) x\ p(A/B_j)$$

Therefore, Bayes theorem can be rewritten as

$$p(B/A) = \frac{p(B)\ x\ p(A/B)}{\sum_{i=1}^{n} p(B_j) x\ p(A/B_j)}$$

For Bayesian statistics, $p(B)$ can be referred as the prior belief about our data. $p(A/B)$ is the likelihood of observing event A given our prior belief. The product of prior belief and the likelihood yield a posterior distribution $p(B/A)$ which is our updated belief based on the new observed data, A.

The denominator of Bayes theorem, is the sum of prior probabilities times the likelihood over the entire partition of universe U which is appropriate for the discrete data. For continuous data, it should be taken as the integral of probability distribution of observed data A. This denominator is the normalizing constant that allows the posterior to have a probability distribution with the summation of one. Therefore, by treating the denominator as a constant, Bayes theorem as used in Bayesian statistics can be simply written as

$$posterior \propto likelihood\ x\ prior$$

Selection of Priors

Prior probability distribution of the parameter to be estimated can be categorized as informative or non-informative. For each category there are wide range of prior distributions. The choice of prior to be used can be influenced by availability of prior information and its relation to the likelihood function. A good example, is a prior

distribution of a parameter that is a conjugate to a likelihood function. Prior distribution is considered conjugate of likelihood function if the posterior distribution resulted from the multiplication of likelihood and prior yield the same distribution as the prior. This makes posterior distribution to be analytically tractable and therefore point estimates of the parameters such as posterior mean, median and credible interval can be easily obtained. It should be noted that most of the posterior distributions are analytically intractable and therefore in most cases simulation methods are used for sampling posterior distribution so as make inferences about the parameters.

Non-Informative Priors

These type of priors are used when we don't have any information about the parameter that is to be estimated. They are formal ways of expressing the ignorance about a parameter (Kass & Wasserman, 1996). They are often referred as the vague priors because they do not contribute much to the posterior distribution. When using non-informative priors, it is recommended to use large sample size in order to obtain reliable posterior distribution. The more informative the prior is, the less sample is required to achieve a desired precision of the estimates. Non-informative priors have being criticized for not being objective because they don't real present any past information about the parameter. Instead, they are used as a way of initiating parameters estimation. When using non-informative priors, the point estimates such as expected posterior mean of the parameter will be nearly the same as the expected means of the parameter derived from classical approach.

Non-informative priors that were discussed in this section include

- Normal priors with large variance
- Uniform priors
- Jeffrey's priors

Uniform Priors

The prior distribution of a parameter is assumed to follow uniform distribution. This assign equal probabilities for all possibilities of a parameter values. There are two main concern of using uniform distribution. First, it is variant under transformation of the parameter (Syversveen, 1998). Suppose θ is our parameter of interest. The uniform distribution of θ will not be the same as $1/\theta$. Therefore the posterior distribution of θ cannot be used for making inference about $1/\theta$ via normal transformation of variable formula. The whole process for calculating posterior distribution of $1/\theta$ has to be repeated again from the scratch. Secondly, the parameter space under uniform distribution has infinite bound and that makes uniform prior improper. However, in some cases, even improper priors when multiplied with the likelihood function can results into proper posterior distribution.

Jeffrey's Priors

This another class of uninformative priors. The probability distribution of a parameter is proportional to the square root of determinant of fisher information matrix. Fisher information measure the amount of information that an observable variable has for unknown parameter θ

Normal Distribution

This distribution is discussed here because it is one of the most commonly used distribution for most type of data. This distribution can be used as non-informative priors when we assign large variance to the parameter to be estimated. Large variance specification expresses lack of prior knowledge about the parameter. It can also be used as informative prior by specifying mean and variance of a parameter obtained from prior studies.

Informative Priors

These priors are used when there is a good advance knowledge about the distribution of the parameter to be estimated. Among many informative priors, gamma distribution was considered appropriate prior for model estimation. Gamma distribution is the conjugate prior of Poison distribution (likelihood function) because it yield a posterior which is also gamma distributed.

Suppose data $X_{1,}X_2 \ldots \ldots . X_n$, which are independent and identically distributed follows Poisson distribution with an average rate λ, then Gamma distribution (α, β) on parameter λ can be given as

$$p(\lambda;\ \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

Whereby α is the shape parameter and β is the rate parameter and $\Gamma(\alpha)$ is the gamma function.

The likelihood function of Poisson distribution is given as the product of probabilities of all possible outcomes of parameter to be estimated given the observed data. It can be given as

$$L(\lambda/x) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i\,!}$$

From Bayesian statistics

$$posterior \propto\ likelihood\ x\ prior$$

Combining the likelihood and the prior to obtain the posterior distribution given as

$$Posterior\ \pi(\lambda/x)\ \propto \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i\,!} * \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

Since λ is our parameter of interest, other part of the equation that are not associated with λ are treated as constant.

$$Posterior\ \pi(\lambda/x) \propto\ \lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

Therefore, the posterior distribution is also gamma distribution with shape parameter $(\sum x_i + \alpha - 1)$ and rate parameter $(n + \beta)$

Evaluation of Posterior Distribution

All the inference and information about the parameter of interest are obtained from the posterior distribution. In order for the inference to be made, then the posterior distribution has to be well defined. Unfortunately, most of the posterior distribution are analytically intractable unless the conjugate priors were used. However, only few cases have conjugate prior. Therefore, Bayesian statistician have resorted to the use of simulation methods whereby the random sample with proposed distribution is withdrawn from the posterior distribution for the purpose of making inferences about the parameter. Two popular simulation methods have been used which relies on the principles of

Markov Chain Monte Carlo Simulation Method. These are Metropolis Hasting Algorithm and Gibbs Sampling.

Metropolis Hasting Algorithm

This algorithm utilizes Markov Chain Monte Carlo Simulation. The Monte Carlo part is executed when the parameters are randomly selected from the posterior. The whole process can be viewed as a random walk of selected parameters on the posterior domain. In another context, the parameter can be viewed as wondering around the posterior domain. In order to make sure that the random selection of the parameter adhere to the proposed target distribution, then the Markov Chain part is introduced whereby each current random selection of the parameter depends on its immediate predecessor. In nutshell the Metropolis hasting algorithm is summarized below

i. Selection of proposed distribution, q(*). The Metropolis Hasting Algorithm allows the selection of any distribution.

ii. Selection of the random parameter θ_0 from the posterior domain to begin the simulation process such that $p(\theta_0/y) > 0$.

iii. Computation of acceptance probability of the current selection based on the previous selection. Let θ_t as the accepted current selection and θ_{t-1} as immediate predecessor and θ_* as the proposed value. The acceptance probability can be computed as the ratio of posterior distribution of the proposed value against that of the immediate predecessor. Their posteriors are usually unknown. Therefore, the ratio can be computed using their likelihoods and prior distributions as shown below.

$$r(\theta_*/\theta_{t-1}) = \frac{p(\theta_*/y)q(\theta_{t-1}/\theta_*)}{p(\theta_{t-1}/y)q(\theta_*/\theta_{t-1})}$$

Whereby

$p\,(\theta_*/y)$ and $q(\theta_{t-1}/\theta_*)$ are the likelihood and prior distribution of the proposed value

$p\,(\theta_{t-1}/y)$ and $q(\theta_*/\theta_{t-1})$ are the likelihood and prior distribution of the immediate predecessor.

If $r(\theta_*/\theta_{t-1})$ is greater than 1, then accept the proposed value if not, further analysis is needed to see whether it can be accepted or absolutely rejected.

iv. For the case of $r(\theta_*/\theta_{t-1}) < 1$, draw u ~ uniform (0,1), if u < $r(\theta_*/\theta_{t-1})$ then accept the new proposed value otherwise reject.

The efficiency of the whole process of sampling using Metropolis Hasting algorithm can be assessed mainly by acceptance rate of the proposed values and the degree of autocorrelation. If the acceptance rate is too low, then it implies the chain fail to select values from the regions of posterior domain that will likely to conform to the proposed distribution. On the other hand, if the acceptance rate is too high then the proposed values might have been selected on a small portion of the posterior domain.

Autocorrelation occurs as the result of introducing conditional probabilities of selected parameters based on its immediate predecessor. For univariate distribution, optimal acceptance rate can be taken as 0.45(Roberts et al, 1997).

Some setbacks with this algorithm are also common to almost all algorithms that rely on MCMC simulation. The chain is influenced by the starting value. Suppose the starting value happen to be at the tail of the proposed distribution. It will take a lot of iterations for the chain to move all regions of posterior domain. To counteract this setback, some of the selected samples at the beginning of the chain are discarded. The discarded sample are usually referred as burn-in samples.

Gibbs Sampling

Gibbs sampling is a special case of Metropolis Hasting Algorithm whereby it updates the parameter based on its full conditional distribution. Gibbs sampling has high efficiency as all proposals are accepted. However, few of the posterior distribution are known to have full conditional distribution and therefore limits its application.

Blocking of the Parameters

The Metropolis hasting algorithm runs the MCMC for all the parameters simultaneously. This might result to inefficiency of the chain to produce a sample that covers the whole posterior domain. The problem become more pronounced when parameters are of different scale. To reduce this problem, parameters can be placed into separate blocks whereby the MCMC will run for each block separately. This results into a well-mixed sample having an adequate acceptance rate and low autocorrelation.

Diagnostic Check of the Proposed Distribution

After obtaining the proposed or target posterior distribution via MCMC sampling, different graphical methods can be used to assess the quality of the posterior distribution. This is based on how well the Metropolis algorithm converged to the optimal solution.

Trace Plot

This is one way of visualizing how well the simulated sample traverse through the posterior domain. It is a plot of magnitude of simulated parameters against the number of iterations for the entire simulation. For a well-mixed sample, the mean and variance throughout the iterations should remain constant and the chain should be able to transcend to all the posterior regions after some few number of iterations.

Autocorrelation Plot

Autocorrelation for the case of MCMC can be defined as the correlation between simulated parameters values within the same chain separated by a given number of iterations. Autocorrelation between MCMC sample parameter values is inevitable as the current selection of parameter depends on the immediate predecessor. However, it is desirable to reduce the autocorrelation as much as possible so that the selected parameter values will be somewhat random and independent of each other. It is expected after first few lags while the chain progresses, the autocorrelation will be reduced to the minimal.

Histogram and Kernel Density Plots

These plots are used to show the distribution of the MCMC sample. The density plots can be visually compared with the distribution that was initially proposed to see how well they fit each other.

Inferences from Posterior Distribution

There are different ways of summarizing posterior distribution for the purpose of extracting meaningful information. Among those, are point estimators such as mean, median, mode and credible intervals.

The posterior mean of a parameter θ can be computed as

$$\hat{\theta} = E(\theta/y) = \int \theta p(\theta/y)$$

Whereby

$p(\theta/y)$ is the posterior probability distribution of parameter values, θ

Credible Intervals

They are used to quantify the uncertainty of the estimated parameters. They provide a range that the expected value of the parameter θ lies in a posterior domain at a given probability. The narrower the interval at a given probability the more reliable the expected value of parameter θ will be. The credible intervals can be expressed as Equal-tail interval or Highest Posterior Density (HPD) interval.

With equal-tailed interval, it is selected in such a way the probability of the posterior mean of parameter θ being below the interval is as likely as being above the

interval. It is defined as $(q_{\alpha/2}, q_{1-\alpha/2})$ whereby$\propto \in (0,1)$. The common equal tail credible interval used is $(q_{0.25}, q_{0.975})$ at 0.95 probability.

Highest posterior density interval defines the area of highest posterior density that includes the posterior mode. It is the credible interval with the shortest width. When the posterior distribution is symmetrical, then credible intervals for equal tailed and HPD are the same.

Model Comparison

Based on the specification of prior distribution and likelihood function, more than one model may seem to fit the data well. Using only diagnostic plots will not be sufficient to conclude which model has outperformed the rest. Therefore, other criteria were used, namely Deviance Information Criteria (DIC) and Bayes factors.

Deviance Information Criterion

DIC can be viewed as the measure of within sample predictive accuracy (Gelman et al, 2014). DIC is based on the deviance, which is the similar criterion used in information criteria tests such as AIC and BIC. In Bayesian context, this is the log of predicted density distribution of the data given posterior mean multiplying by -2. It is given as

$$DIC = -2(logp(y/\hat{\theta}) + 2p_{DIC}$$

Where $\hat{\theta}$ is the posterior mean, $p(y/\hat{\theta})$ is the probability distribution of the data given the posterior mean and $2p_{DIC}$ is the effective sample size which accounts for within sample prediction bias. p_{DIC} can be computed as

$$p_{DIC} = 2(logp(y/\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \log p(y/\theta_i)$$

With *n* equals to the number of MCMC iteration.

The preferable model will be the one with smaller DIC value when compared to other competing models.

Bayes Factor

Bayes factor is the ratio of marginal probabilities of the competing models given the observed data.

$$BF_{jk} = \frac{p(y/M_j)}{p(y/M_k)}$$

Where

$p(y/M_j)$ is the marginal likelihood of M_j given observed data *y*

$p(y/M_k)$ is the marginal likelihood of competing model M_k given the observed data y

Kass & Raftery (1995) had a table for making judgement after calculating Bayes factor $BF_{jk}$when comparing competing models.

Table 10 Bayes Factors for Comparing Competing Models

2 ln BF_{jk}	**BF_{jk}**	**Strength of evidence**
0 to 2	1 to 3	not worth more than a bare mention
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
>10	>150	very strong

Discussion of Results

This section covers the results that were obtained using two modeling approaches, which were Classical approach and Bayesian approach. For each approach, different models that could potentially fit the data were analyzed and appropriate statistical tests were applied to assess the model fit and performance. The best model from Bayesian approach was compared with the best model that was developed using frequentist approach. The purpose of this comparison was to test if the use of Bayesian approach with an additional advantage of incorporating prior knowledge would increase the model estimation and predictive performance. In nutshell, this section aimed at demonstrating the following:

- Incorporating prior knowledge about the parameters (coefficients) of factors influencing pedestrian and bicyclist involved crashes at intersection.
- How Bayesian approach can be utilized using small sample size and yet achieve the desired model performance.

Data Description

Two hundred and forty signalized intersections in Michigan were included in the analysis of comparing the performance of the model developed using classical/frequentist approach and Bayesian approach. About 80% of the intersections were used for model estimation and the rest were used for model validation. Figure 17 below shows the distribution of eleven years non-motorized crashes at signalized intersections that were used for model estimations. More than half of the intersections had zero pedestrian-involved crashes. Same distribution was observed for bicycle-involved crashes.

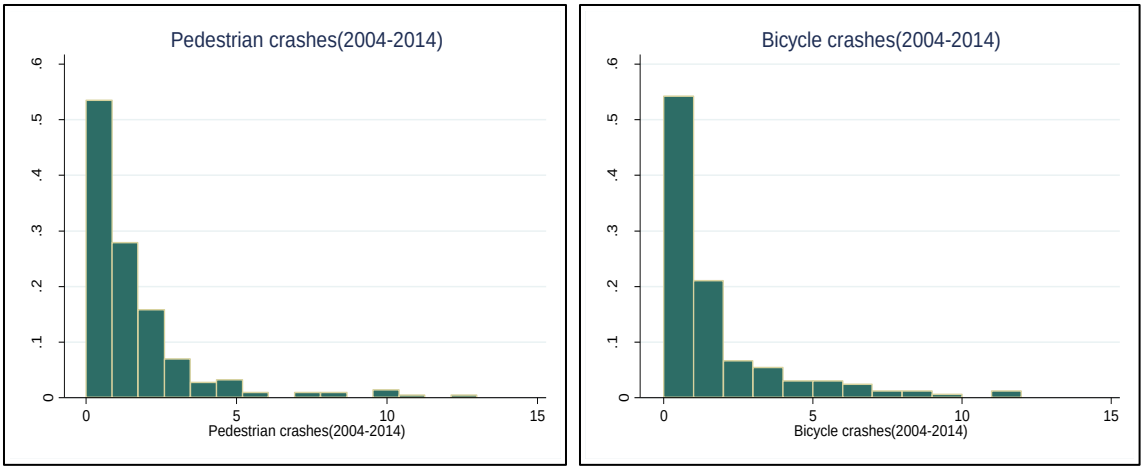


Figure 17 Eleven Years Non-motorized Crashes at 240 Signalized Intersections

Model Estimation Using Classical Approach

The estimation procedure was performed for classical model in order to discern which of the count models fit the data well. Description of the significant variables that were used for both pedestrian and bicycle SPFs are presented in Table 11. For pedestrian SPFs, the variables that were significant at 95% confidence level include AADT at the major approach, AADT in the minor approach, pedestrian level and presence of on-street parking as shown in Table 12. For Bicyclist SPFs, significant variables were total entering traffic, bicycle level and presence of four-legged intersection as depicted in Table 13. All of the predictor variables for Pedestrian and Bicycle SPFs had a positive association with the expected number of pedestrian and bicycle involved crashes.

Table 11 Description of Variables Used for Non-motorized SPFs Modeling

Variable	Definition	Obs	Mean	Std. Dev.	Min	Max
Ln AADT major	Natural logarithm of AADT in the major approach in thousands	165	2.758	0.515	1.120	3.876
Ln AADT minor	Natural logarithm of AADT in the minor approach in thousands	165	1.996	0.686	0.071	3.477
Pedestrian level	Pedestrian level	165	-0.021	0.710	-1.168	2.499
Parking	Dummy variable for the presence of street parking	165	0.030	0.172	0.000	1.000
Total entering traffic	Sum of AADT at major and minor approach in thousands	166	26.020	12.458	4.235	68.118
Bicycle level score	Latent Bicyclist exposure measure	166	-0.005	0.139	-0.326	0.396
Intersectio_4leg	Intersection type:4leg	166	0.77	0.43	0	1

Table 12 Model Estimation for Pedestrian SPF

Variable	PRM	NBRM	ZIP	ZINB
Ln AADT major	0.555 (3.31)	0.634 (2.76)	0.504 (2.75)	0.460 (2.01)
Ln AADT minor	0.513 (4.41)	0.426 (2.72)	0.513 (4.19)	0.471 (3.13)
Pedestrian level	0.666 (8.86)	0.889 (6.46)	0.524 (5.55)	0.695 (5.38)
Parking	1.372 (4.23)	1.413 (3.19)	1.239 (3.64)	1.257 (2.85)
Constant term	-2.723 (-5.81)	-3.172 (-4.87	-2.301 (-4.38)	-2.311 (-3.72)
Over dispersion parameter				
alpha		0.369		0.348
Inflate(For zero-inflated models)				
Pedestrian level score			-1.323 (-2.31)	-8.111 (-1.1)
Constant			-1.415 (-3.39)	-8.776 (-1.06)

Table 13 Model Estimation for Bicyclist SPF

Variable	PRM	NBRM	ZIP	ZINB
Total entering traffic	0.033 (7.01)	0.030 (3.22)	0.020 (3.67)	0.030 (3.26)
Bicycle level score	1.981 (3.89)	2.029 (2.1)	1.361 (0.034)	
Intersection_4leg	1.058 (4.01)	1.078 (3.05)	1.105 (3.68)	1.163 (3.3)
Constant	-1.654 (-5.88)	-1.593 (-3.98)	-0.771 (-0.019)	-1.521 (-3.71)
Over dispersion parameter				
alpha		1.507		1.304
Inflate(For zero-inflated models)				
Bicycle level score			-1.482 (-0.97)	-15.178 (-1.65)
Constant			(-0.302) -1.42	(-3.519) (-1.74)

Figure 18 and Figure 19 show Negative binomial regression model (NBRM) to have the lowest AIC and BIC values which indicated better fit of NBRM compared to other competing models. Residual probability plots obtained after within sample prediction also supported the choice of NBRM as the best model for both bicycle and pedestrian SPF**.** Figure 20 and Figure 21 present residual probability plots for within sample prediction for Pedestrian and Bicyclist-involved crashes respectively. NBRM had the lowest residual probability for a given observed non-motorized crash counts. Figure 22 and Figure 23 show out of sample residual probability for NBRM. Out of sample prediction was fairly similar to within sample prediction.

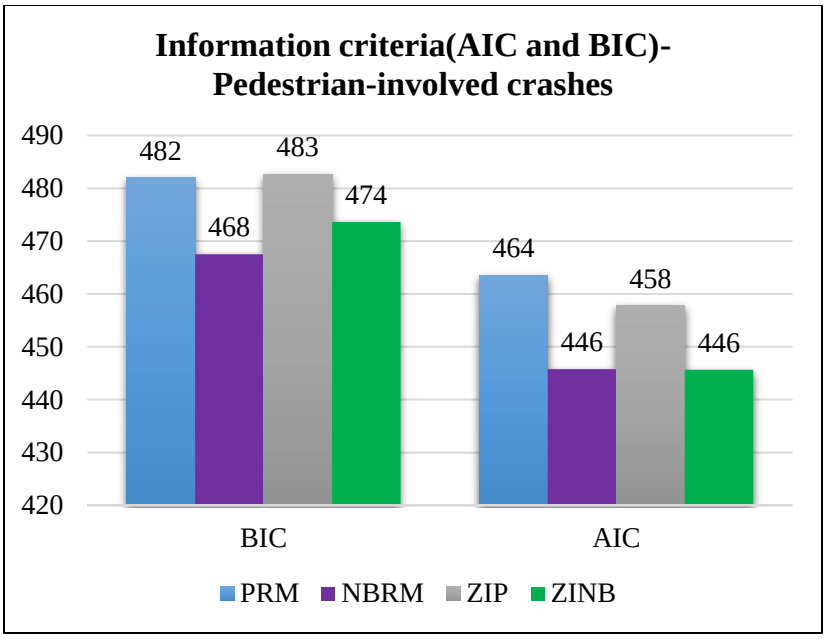


Figure 18 Information Criteria for the Competing Count Models-Pedestrian SPF

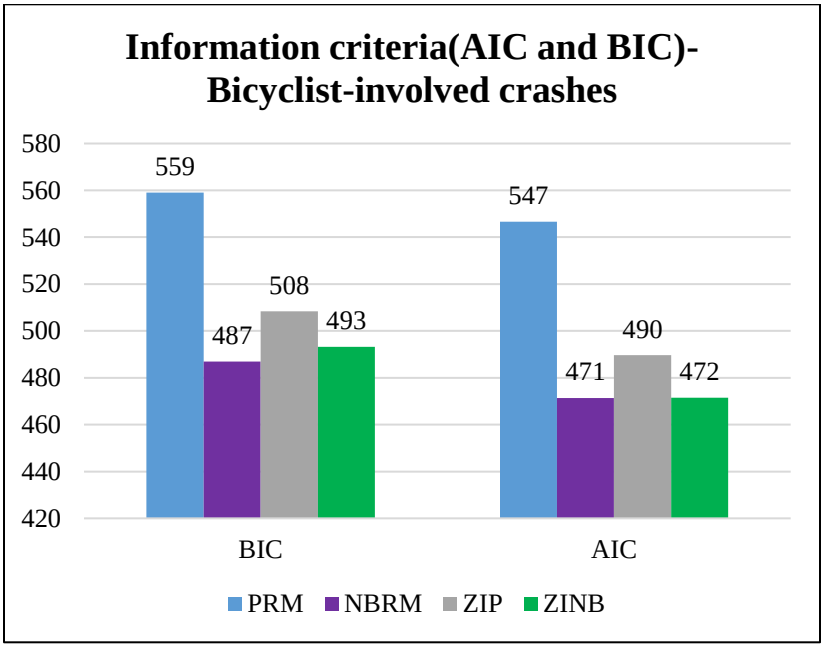


Figure 19 Information Criteria for the Competing Count Models-Bicycle SPF

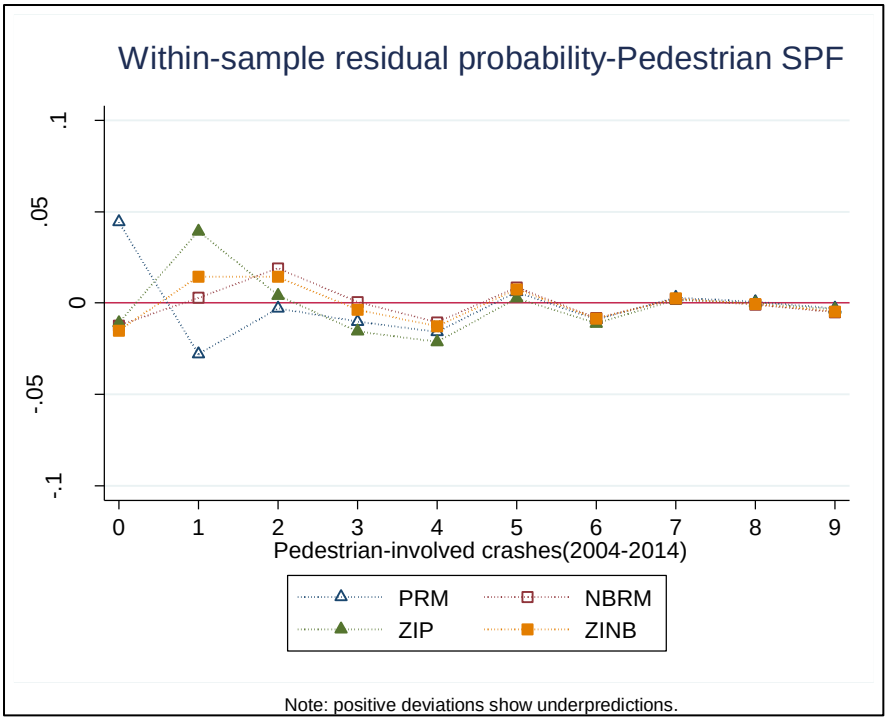


Figure 20 Within-Sample Residual Probability-Pedestrian SPF

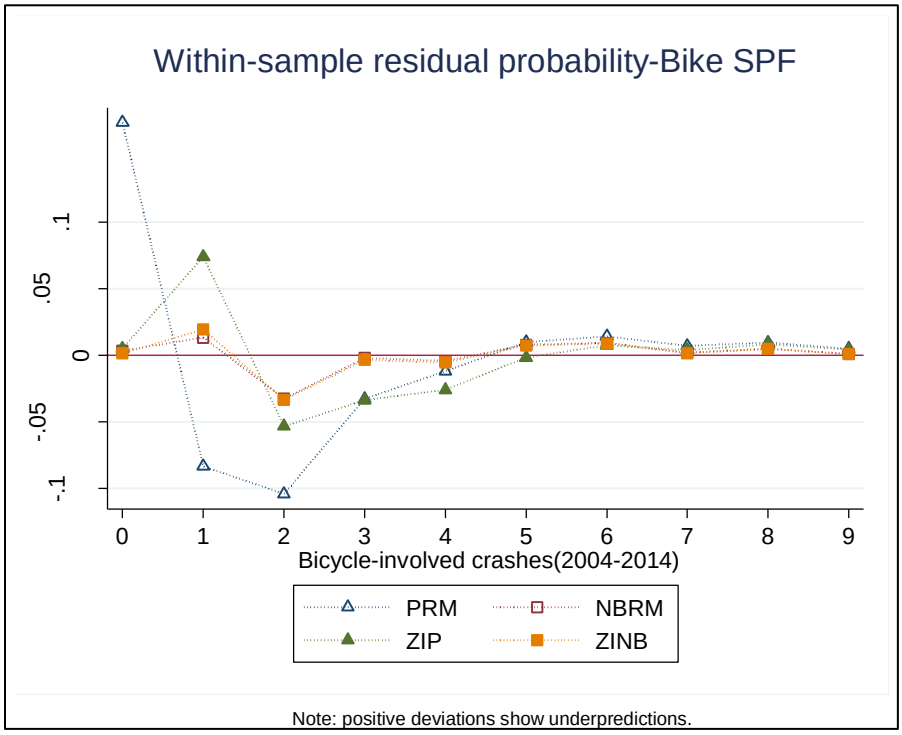


Figure 21 Within-Sample Residual Probability-Bicycle SPF

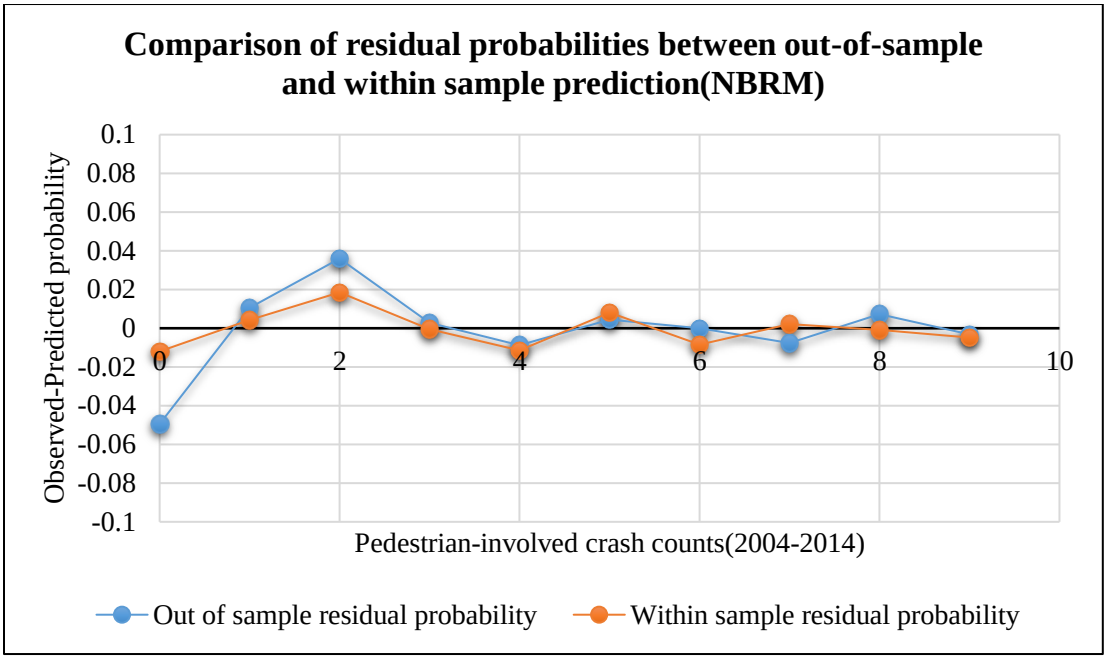


Figure 22 Out-of-Sample Residual Probability plot-Pedestrian SPF

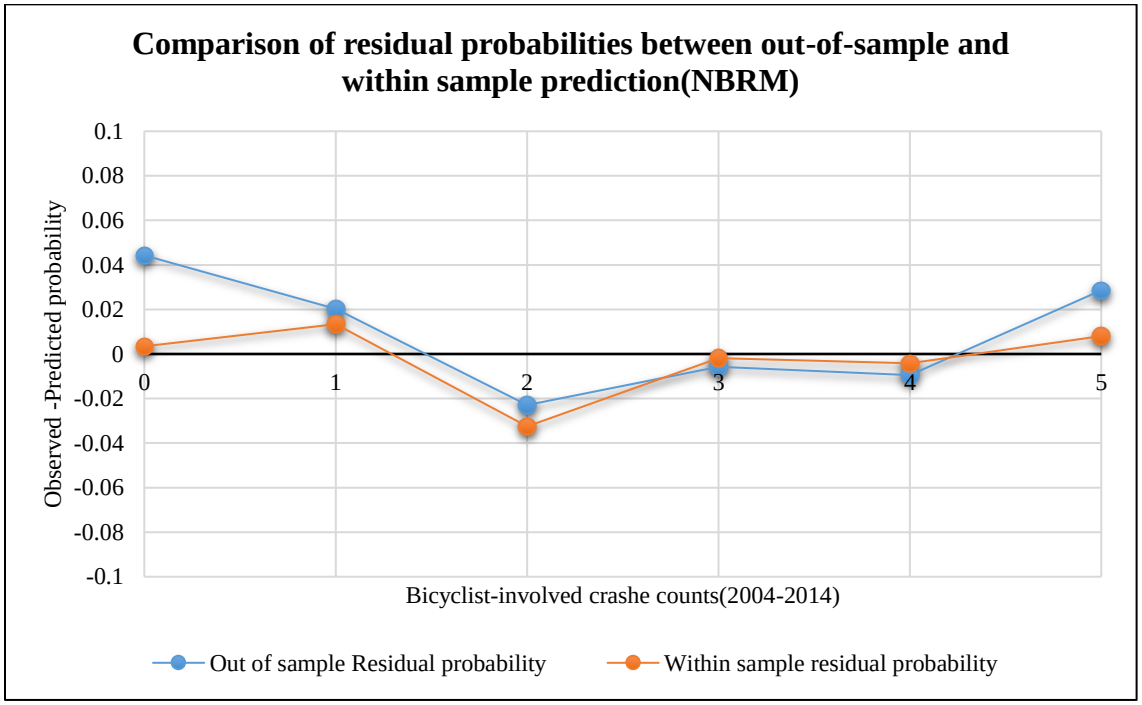


Figure 23 Out-of-Sample Residual Probability plot-Bicyclist SPF

Therefore, based on assessment of the models using goodness of fit measures NBRM had a better fit to the data. The final estimation of SPFs for pedestrian and bicyclist used NBRM.

Mathematically the SPFS can be written as

Pedestrian SPF

$$Pedcrashes = AADT_{maj}{}^{0.634} * AADT\_min^{0.426} * e^{(-3.172+0.889Pedscore+1.413Parkn)}$$

Whereby

AADT_maj – AADT in the major approach (in thousands)

AADT_min- AADT in the minor approach (in thousands)

Pedscore- Pedestrian score computed using factor analysis

Parkn-Presence of on-street parking

Bicycle SPF

$$Bikecrashes = e^{(-1.593+0.0305AADT+2.029bikelevel+1.07int\_typ\_fleg)}$$

Whereby

AADT – Total entering vehicles (in thousands)

Bikelevel- Bicyclist score computed using factor analysis

Int_type_fleg- Dummy variable for four-legged intersection type

Bayesian Model Estimation

Model estimation using Bayesian approach involved the following procedure

- Specification of likelihood function and prior distribution
- Running the model using simulation (MCMC Metropolis hasting algorithm)
- Making inferences from the posterior distribution

Specification of Likelihood Function and Prior Distribution

Poisson distribution was used as the likelihood function, which is appropriate for crash data. Gamma distribution with is the conjugate prior of Poisson distribution was used as the prior.

Graphical Diagnostic Plots

Estimation results involve sampling from posterior distribution by running the Metropolis hasting simulation algorithm. The sample obtained from the simulation was used for point estimation such as the posterior means, standard errors and credible intervals. 45000 iterations were specified for the simulation with the burn-in sample of 15000 iterations. Graphical diagnostics plots were used to investigate if the sample obtained is the good representation of the posterior distribution. Figure 24 and Figure 25 show the trace plots, autocorrelation plots and kernel density plots for the parameters(coefficients) of the variables that were significant in the pedestrian and bicycle SPFs.

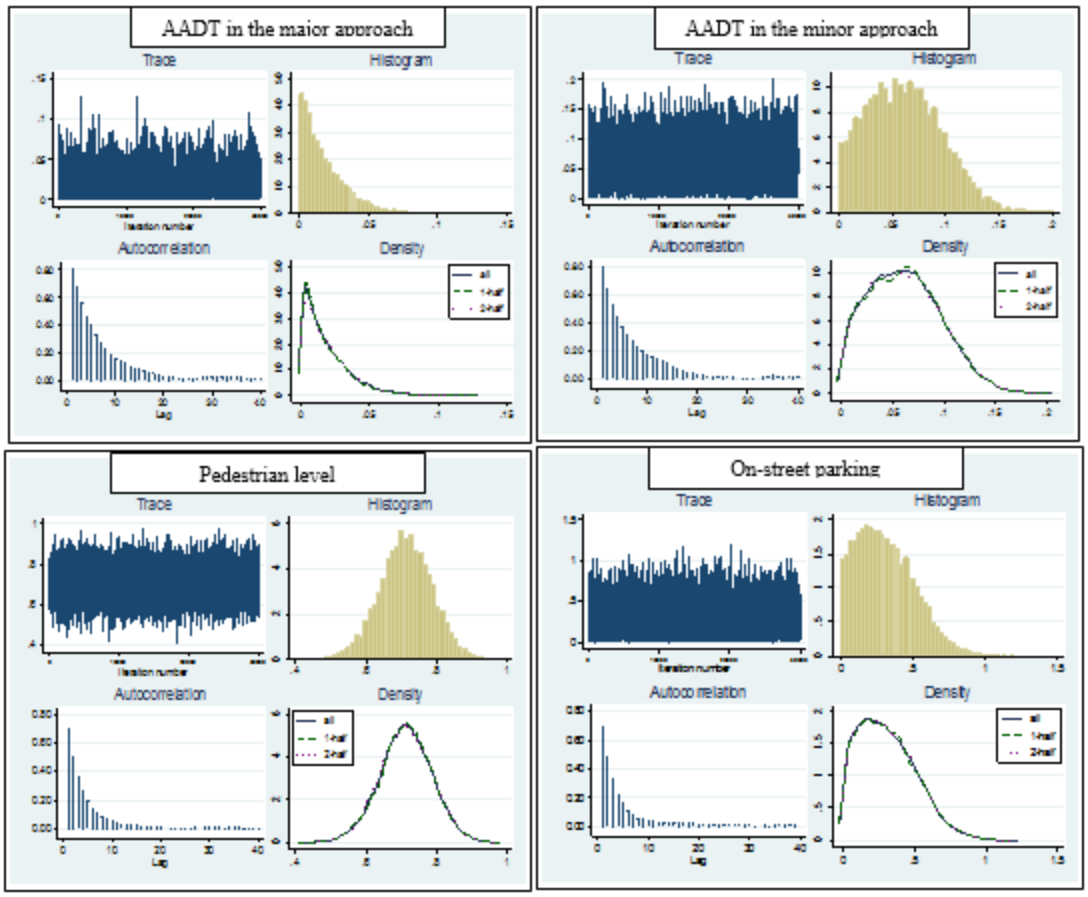


Figure 24 Diagnostics Plots for Coefficient of Significant Factors in Pedestrian SPF

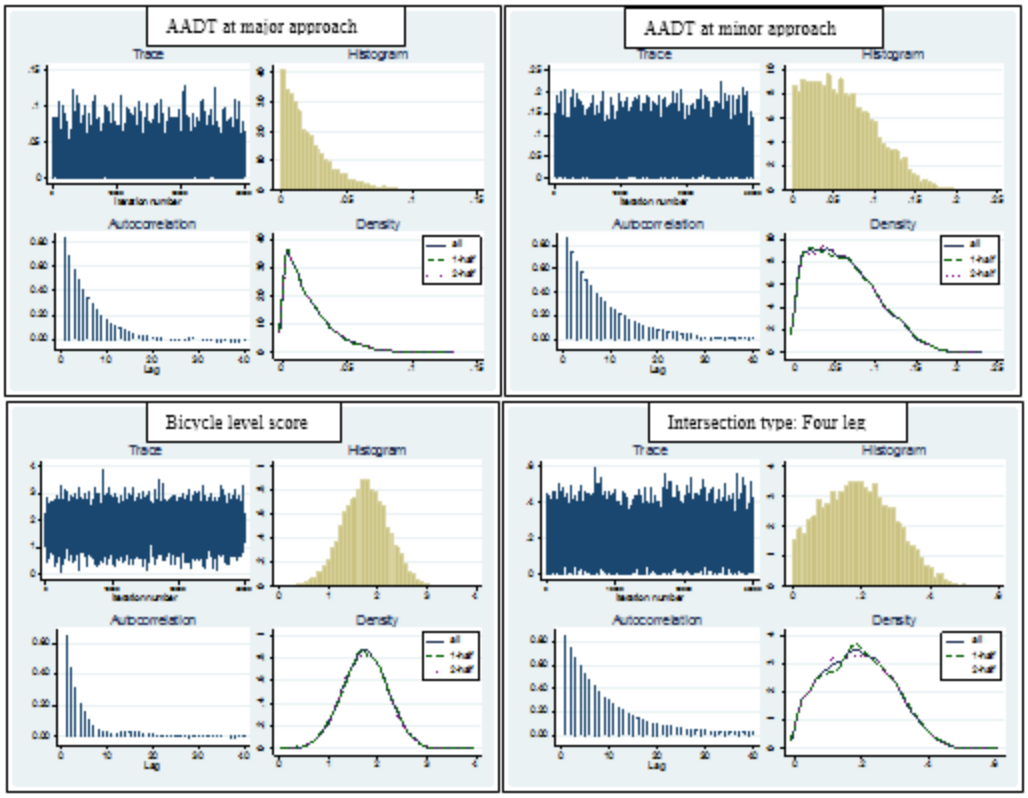


Figure 25 Diagnostics Plots for Coefficient of Significant Factors in Bicycle SPF

Trace Plot

The trace plot for parameter of significant variables showed a good mix of MCMC sample for both pedestrian and bicycle SPF. For the good MCMC mix, the trace plot should show a rapid movement of simulated parameters across the posterior domain. After 15 lags, the parameters had low autocorrelation with pedestrian and bicycle level score having lower autocorrelation compared to other parameters.

Cusum Plots

For the good MCMC sample, the Cumulative sum plot (Cusum) should cross at x-axis at least once as shown in Figure 26. This gives an indication that the MCMC simulation explored all the regions of posterior domain.

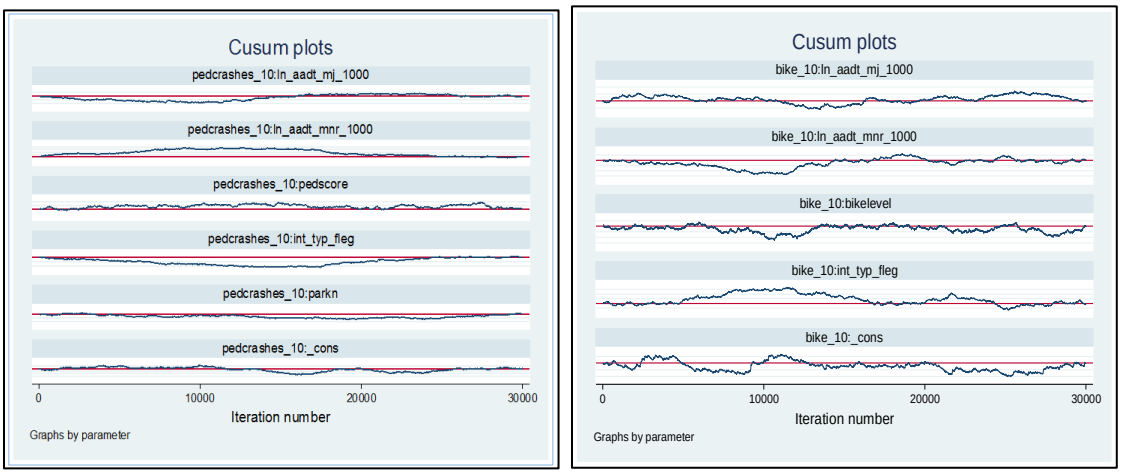


Figure 26 Cusum Plot for the Coefficients of Significant Variables in Pedestrian and Bicycle SPFS

Effective Sample Size

Effective sample size (ESS) shows the percentage of MCMC simulated parameters that were randomly and independently distributed. This sample size was used in making inferences from the posterior distribution such as mean, median and credible interval of significant variables. For Pedestrian SPF, the parameter for the presence of on-street parking had the higher effective sample size compared to other parameters with the sampling efficiency of 18.7%. For Bicycle SPF, bicycle level score had a higher sample size compared to other parameters.

Table 14 Effective Sample Size for Significant Variables-Non-motorized SPF

	Effective Sample size	Corr. time	Efficiency
Pedestrian SPFs			
Ln AADT major	2840	10.56	0.0947
Ln AADT minor	2896	10.36	0.0965
Pedestrian level	4981	6.02	0.166
Intersection_4leg	3053	9.83	0.1018
parkn	5611	5.35	0.187
Constant	2387	12.58	0.0795
Bicyclist SPFs			
Ln AADT major	2803	10.7	0.0934
Ln AADT minor	1971	15.22	0.0657
Bicyclist level	5502	5.45	0.1834
Intersection_4leg	1777	16.88	0.0592
_cons	2455	12.22	0.0818

Making Inferences from the Posterior Distribution

Factors that were found to be significantly associated with pedestrian involved crashes at signalized intersections were Average annual daily traffic at the major and minor approaches, presence of on-street parking at the intersection, pedestrian level score, and intersection type based on number of legs. Both were significant at 95% credible interval. For Bicycle SPF, the same factors were significant with exception of presence of on-street parking. Table 15and Table 16 provides the statistical summary of Pedestrian and Bicycle SPFs respectively. All of the significant factors had positive coefficients indicating a positive association with non-motorized crashes.

Table 15 Bayesian Poisson-Gamma Model for Pedestrian SPF

	Mean	Std. Dev.	MCSE	Median	[95% Cred. Interval]	
Ln AADT major	0.018	0.016	0.000	0.014	0.001	0.050
Ln AADT minor	0.061	0.035	0.001	0.059	0.009	0.122
Pedestrian level	0.713	0.075	0.001	0.713	0.589	0.834
Intersection_4leg	0.056	0.048	0.001	0.044	0.003	0.153
Parking	0.310	0.198	0.003	0.287	0.035	0.666
_cons	0.028	0.027	0.001	0.020	0.001	0.085

Table 16 Bayesian Poisson-Gamma Model for Bicycle SPF

	Mean	Std. Dev.	MCSE	Median	[95% Cred. Interval]	
Ln AADT major	0.021	0.018	0.000	0.015	0.001	0.069
Ln AADT minor	0.061	0.041	0.001	0.057	0.003	0.150
Bicyclist level	1.746	0.461	0.006	1.751	0.835	2.644
Intersection_4leg	0.191	0.103	0.002	0.187	0.016	0.397
_cons	0.032	0.030	0.001	0.023	0.001	0.111

Mathematically, the model can be written as

Pedestrian SPF

$$Pedcrashes = AADT_{maj}{}^{0.018} * AADT\_min^{0.061} * e^{(0.028+0.713Pedscore+0.310Parkn+0.056Int\_typ\_fleg)}$$

Whereby

AADT_maj – AADT in the major approach (in thousands)

AADT_min- AADT in the minor approach (in thousands)

Pedscore- Pedestrian score computed using factor analysis

Parkn-Presence of on-street parking

Int_typ_fleg- Dummy variable for presence 4-legged intersection

Bicycle SPF

$$Bikecrashes = AADT_{maj}{}^{0.021} * AADT\_min^{0.061} * e^{(0.032+1.746Bikelevel+0.191Int\_typ\_fleg)}$$

Whereby

AADT_maj – AADT in the major approach (in thousands)

AADT_min- AADT in the minor approach (in thousands)

Bikelevel- Bicyclist score computed using factor analysis

Int_typ_fleg- Dummy variable for presence 4-legged intersection

Comparison of Model Performance

The comparison between models that were developed using classical approach and Bayesian Poisson-gamma model was performed by looking the effect of sample size on model estimation and out-of-sample model prediction.

Effect of sample size on model estimation.

The models were estimated using sample size of 80, 120 and 160 intersections. For each sample size, variables that were significantly associated with pedestrian crashes at signalized intersections were retained in the final model. Table 17, Table 18 and Table 19 shows the significant variables at different sample sizes for Pedestrian SPFs.

Classical model

At a sample size of 80 intersections only two variables were found to be significant for classical Pedestrian SPF, namely AADT at the minor approach and pedestrian score level. Presence of on-street parking was the additional variable that was found to be significant at a sample size of 120 intersections. AADT at the major approach, which was one of the essential variable, was still not significant at a sample size of 120 intersections. At a sample size of 160 intersections, more variables were found to be significant including AADT at the major approach.

Bayesian model

For Pedestrian SPF, all the variables that were significant at sample size of 160 intersections were also significant at a smaller sample size of 80 intersections. This

demonstrate the ability Poisson-gamma Bayesian model to accommodate more variables at smaller sample size compared to classical Negative Binomial model.

For Bicyclist SPFs, there was no appreciable difference between Bayesian and Classic Bicycle SPFs by comparing the number of variables that were significant at different sample sizes.

Table 17 Model Comparison at a Sample Size of 80 Intersections-Pedestrian SPF

Sample size=80(Negative Binomial-Classical approach)						
	Coef.	**Std. Err.**	**z**	**P>z**	**[95% Conf. Interval]**	
Ln AADT major	0.604	0.230	2.620	0.009	0.153	1.056
Pedscore	0.927	0.206	4.510	0.000	0.524	1.330
_cons	-3.312	0.905	-3.660	0.000	-5.085	-1.539
alpha	0.422	0.199			0.168	1.063
Sample size=80(Poisson-Gamma: Bayesian approach)						
	Mean	**Std. Dev.**	**MCSE**	**Median**	**[95% Cred. Interval]**	
Ln AADT major	0.023	0.020	0.000	0.017	0.001	0.075
Ln AADT minor	0.066	0.043	0.001	0.061	0.003	0.162
Pedscore	0.685	0.103	0.001	0.686	0.483	0.879
Intersecton_4leg	0.084	0.070	0.001	0.066	0.003	0.256
Parking	0.162	0.144	0.003	0.122	0.006	0.534
_cons	0.043	0.040	0.001	0.032	0.001	0.149

Table 18 Model Comparison at a Sample Size of 120 Intersections-Pedestrian SPF

Sample size=120 (Negative Binomial-Classical approach)						
	Coef.	**Std. Err.**	**z**	**P>z**	**[95% Conf.**	**Interval]**
Ln AADT minor	0.535	0.194	2.750	0.006	0.154	0.917
Pedscore	1.009	0.170	5.930	0.000	0.675	1.343
Parking	1.235	0.564	2.190	0.029	0.129	2.342
_cons	-3.027	0.783	-3.870	0.000	-4.562	-1.492
alpha	0.427	0.158			0.207	0.883

Sample size=120(Poisson-Gamma: Bayesian approach)						
	Mean	**Std. Dev.**	**MCSE**	**Median**	**[95% Cred. Interval]**	
Ln AADT major	0.018	0.016	0.000	0.013	0.000	0.059
Ln AADT minor	0.048	0.033	0.001	0.043	0.002	0.124
Pedscore	0.795	0.091	0.001	0.794	0.617	0.969
Intesection_4leg	0.076	0.062	0.001	0.060	0.003	0.228
Parking	0.179	0.146	0.002	0.144	0.006	0.538
_cons	0.034	0.033	0.001	0.024	0.001	0.121

Table 19 Model Comparison at a Sample Size of 160 Intersections-Pedestrian SPF

Sample size=160 (Negative Binomial-Classical approach)						
	Coef.	**Std. Err.**	**z**	**P>z**	**[95% Conf.Interval]**	
Ln AADT major	0.634	0.230	2.760	0.006	0.183	1.084
Ln AADT minor	0.426	0.156	2.720	0.006	0.119	0.732
Pedscore	0.889	0.138	6.460	0.000	0.619	1.159
Parking	1.413	0.442	3.190	0.001	0.546	2.280
_cons	-3.172	0.651	-4.870	0.000	-4.448	-1.896
alpha	0.369	0.132			0.183	0.743

Sample size=160(Poisson-Gamma: Bayesian approach)						
	Mean	**Std. Dev.**	**MCSE**	**Median**	**[95% Cred. Interval]**	
Ln AADT major	0.018	0.016	0.000	0.014	0.001	0.050
Ln AADT minor	0.061	0.035	0.001	0.059	0.009	0.122
Pedscore	0.713	0.075	0.001	0.713	0.589	0.834
Intesection_4leg	0.056	0.048	0.001	0.044	0.003	0.153
Parking	0.310	0.198	0.003	0.287	0.035	0.666
_cons	0.028	0.027	0.001	0.020	0.001	0.085

Table 20 Model Comparison at a Sample Size of 80 Intersections-Bicycle SPF

Sample size=80						
	Coef.	Std. Err.	z	P>z	[95% Conf.Interval]	
Total entering traffic	0.025	0.013	1.96	0.05	0.001	0.05107
Bikelevel	2.575	1.291	1.990	0.046	0.044	5.106
Intersection_4leg	1.199	0.545	2.200	0.028	0.131	2.268
_cons	-1.517	0.542	-2.800	0.005	-2.578	-0.456
alpha	1.678	0.505			0.930	3.028
Sample size=80						
	Mean	**Std. Dev.**	**MCSE**	**Median**	**[95% Cred. Interval]**	
Ln AADT major	0.027	0.024	0.000	0.021	0.001	0.091
Ln AADT minor	0.072	0.050	0.001	0.064	0.003	0.185
Bikescore	1.483	0.548	0.007	1.481	0.397	2.564
Intesection_4leg	0.207	0.124	0.002	0.198	0.012	0.469
_cons	0.049	0.046	0.001	0.036	0.001	0.170

Table 21 Model Comparison at a Sample Size of 120 Intersections-Bicycle SPF

Sample size=120						
	Coef.	Std. Err.	z	P>z	[95% Conf.Interval]	
Total entering traffic	0.033	0.010	3.170	0.002	0.012	0.053
Bikelevel	2.146	1.055	2.030	0.042	0.079	4.213
Intersection_4leg	1.188	0.414	2.870	0.004	0.376	2.000
_cons	-1.747	0.453	-3.850	0.000	-2.635	-0.858
alpha	1.452	0.379			0.870	2.421
Sample size=120						
	Mean	**Std. Dev.**	**MCSE**	**Median**	**[95% Cred. Interval]**	
Ln AADT major	0.022	0.020	0.000	0.017	0.001	0.074
Ln AADT minor	0.081	0.048	0.001	0.077	0.006	0.180
Bikescore	1.612	0.485	0.006	1.616	0.674	2.568
Intesection_4leg	0.194	0.113	0.003	0.184	0.014	0.432
_cons	0.033	0.033	0.001	0.023	0.001	0.124

Table 22 Model Comparison at a Sample Size of 120 Intersections-Bicycle SPF

Sample size=160						
	Coef.	**Std. Err.**	**z**	**P>z**	**[95% Conf.Interval]**	
Total entering traffic	0.030	0.009	3.220	0.001	0.012	0.049
Bikelevel	2.029	0.964	2.100	0.035	0.139	3.919
Intersection_4leg	1.078	0.354	3.050	0.002	0.385	1.771
_cons	-1.593	0.400	-3.980	0.000	-2.378	-0.809
alpha	1.507	0.337			0.973	2.336
Sample size=160						
	Mean	**Std. Dev.**	**MCSE**	**Median**	**[95% Cred. Interval]**	
Ln AADT major	0.021	0.018	0.000	0.015	0.001	0.069
Ln AADT minor	0.061	0.041	0.001	0.057	0.003	0.150
Bikescore	1.746	0.461	0.006	1.751	0.835	2.644
Intesection_4leg	0.191	0.103	0.002	0.187	0.016	0.397
_cons	0.032	0.030	0.001	0.023	0.001	0.111

Cross Validation Results for Different Sample Size

Incorporation of more variables at smaller sample size was not enough to justify the superiority of Bayesian Poisson-gamma model over the Negative Binomial classical model. Therefore, out of sample prediction was performed after estimation, to measure the model performance at different sample sizes. Figure 27 shows the Root mean square error (RMSE) for both Negative binomial and Poisson-gamma model at different sample sizes. For both cases, RMSE values for Bayesian Poisson-gamma model were smaller compared to classical Negative binomial model. The difference kept on diminishing with the increase in sample size. Similar trend was observed for Bicycle SPFS.

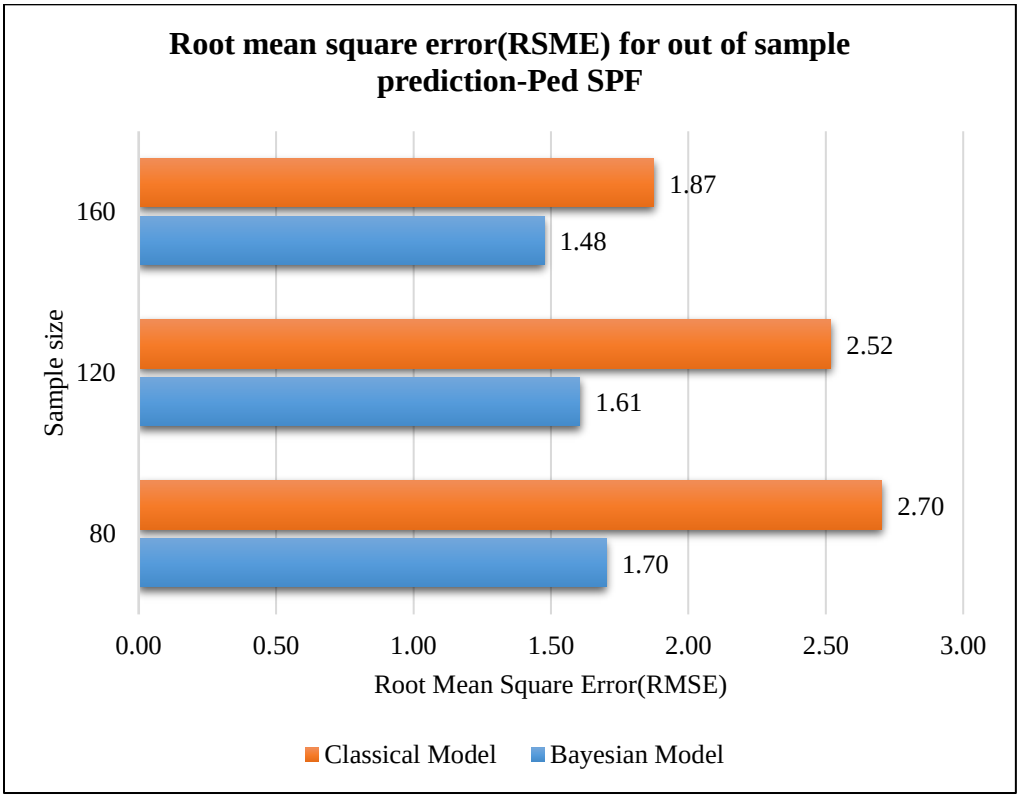


Figure 27 Comparison of NBRM and Bayesian Poisson-Gamma Model-Pedestrian SPF

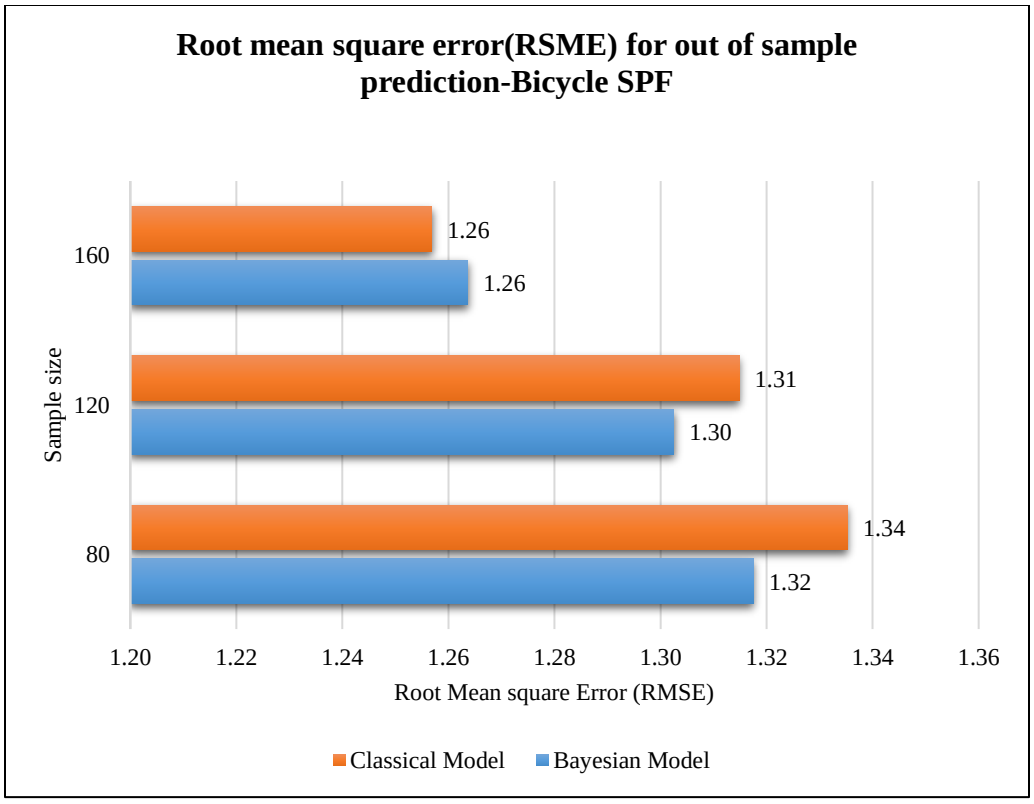


Figure 28 Comparison of NBRM and Bayesian Poisson-Gamma Model-Bicycle SPF

CHAPTER 7

CONCLUSIONS

This project aimed at improving the methodology for developing statewide safety performance function for bicyclist and pedestrian. Specific focus was on signalized urban intersections in Michigan connecting collector and arterial roads.

Proper sampling procedure was needed to come up with the unbiased sample representative of all signalized urban intersections in Michigan. Stratified random sampling was selected as the sampling strategy. All signalized urban intersections in Michigan were placed into strata of similar characteristics. These characteristics were established using parameters that were available at statewide level. These were National Function Classification (NFC) of the roadway forming an intersection, intersection type (three legged or four legged intersection), urban population and number of non-motorized crashes per intersection in eleven years. Seventy-two strata were created from which sample intersections were selected for developing SPFs.

Due to lack of pedestrian and bicycle volume counts at intersections, it was necessary to develop a reliable proxy exposure measure. Factor analysis was used to develop pedestrian and bicycle level score using variables that are readily available at statewide level. Latent bicyclist level score, a proxy measure of bicyclist volume was found to increase with the following parameters; the presence of bicycle facility which includes bike lanes and sidewalks, increase in percentage of people below poverty level, increase population density, lower speed limit in major and minor approach and increase in proportion of commercial land use by area in a given census block group were the

intersection is situated. High pedestrian level score was manifested by the increase in percentage of people using the public transit in a given block group where the intersection was situated, population density, percentage of people below poverty level, number of workers commuting to their working places by foot per square mile, walk score index, proportion of commercial land use and presence of pedestrian facility separated from the roadway.

Analysis was carried out to compare the performance of models developed using Bayesian approach and Classical approach. Eleven-year non-motorized crash data at signalized intersections were used in the analysis. The comparison was specifically based on the model estimation and prediction at different sample sizes. The result showed Bayesian model to have better performance in both model estimation and out-of-sample model prediction at smaller sample size. The ability of Bayesian approach to incorporate prior knowledge provides additional advantage of developing SPFs that reflects the collective efforts of past studies to the present studies.

BIBLIOGRAPHY

Aggarwal, Y. P. (1988). Better sampling: concepts, techniques, and evaluation. Stosius Inc/Advent Books Division.

Beck, L. F., Dellinger, A. M., & O'neil, M. E. (2007). Motor vehicle crash injury rates by mode of travel, United States: using exposure-based methods to quantify differences. American Journal of Epidemiology, 166(2), 212-218.

Chu, X. (2003, February). The fatality risk of walking in America: A time-based comparative approach. In Walk 21 (IV) Conference Proceedings.

Davis, S., King, E., Robertson, D., Mingo, R., & Washington, J. (1987). Measuring pedestrian volumes and conflicts. Volume I. Pedestrian volume sampling. Final report (No. FHWA/RD-88/036).

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. Practical Assessment, Research & Evaluation, 14(20), 1-11.

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

Greene-Roesel, R., Diogenes, M. C., & Ragland, D. R. (2010). Estimating Pedestrian Accident Exposure (No. UCB-ITS-PRR-2010-32).

Hedlund, J., & North, H. S. (2000). NHTSA/FHWA Pedestrian and Bicycle Strategic Planning Research Workshops. Highway Safety North, Ithaca, NY.

Hilbe, J. M. (2014). Modeling count data. Cambridge University Press.

Jaccard, J., & Wan, C. K. (1996). LISREL approaches to interaction effects in multiple regression (No. 114). Sage.

Jonsson, T. (2013). Safety performance models for pedestrians and bicyclists. In 16th International Conference Road Safety on Four Continents. Beijing, China (RS4C 2013). 15-17 May 2013. Statens väg-och transportforskningsinstitut.

Kim, S. (2003). Analysis of elderly mobility by structural equation modeling. Transportation Research Record: Journal of the Transportation Research Board, (1854), 81-89.

Knoblauch, R., Pietrucha, M., & Nitzburg, M. (1996). Field studies of pedestrian walking speed and start-up time. Transportation Research Record: Journal of the Transportation Research Board, (1538), 27-38.

Kwigizile, V., Mulokozi, E., Xu, X., Teng, H. H., & Ma, C. (2014). Investigation of the impact of corner clearance on urban intersection crash occurrence. Journal of transportation and statistics, 10(1), 35-48.

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention, 37(1), 35-46.

McArthur, A., Savolainen, P., & Gates, T. (2014). Spatial Analysis of Child Pedestrian and Bicycle Crashes: Development of Safety Performance Function for Areas Adjacent to Schools. Transportation Research Record: Journal of the Transportation Research Board, (2465), 57-63.

Minikel, E. (2012). Cyclist safety on bicycle boulevards and parallel arterial routes in Berkeley, California. Accident Analysis & Prevention, 45, 241-247.

Molino, J. A., Kennedy, J. F., Inge, P. J., Bertola, M. A., Beuse, P. A., Fowler, N. L., ... & Do, A. (2012). A Distance-Based Method to Estimate Annual Pedestrian and Bicyclist Exposure in an Urban Environment (No. FHWA-HRT-11-043).

Nordback, K., Marshall, W. E., & Janson, B. N. (2014). Bicyclist safety performance functions for a US city. Accident Analysis & Prevention, 65, 114-122.

Oh, J. S., Kwigizile, V., Van Houten, R., McKean, J., Abasahl, F., Dolatsara, H., ... & Clark, M. (2013). Development of Performance Measures for Non-Motorized Dynamics (No. RC-1603).

Qin, X., & Ivan, J. (2001). Estimating pedestrian exposure prediction model in rural areas. Transportation Research Record: Journal of the Transportation Research Board, (1773), 89-96.

Raford, N., & Ragland, D. (2004). Space syntax: Innovative pedestrian volume modeling tool for pedestrian safety. Transportation Research Record: Journal of the Transportation Research Board, (1878), 66-74.

Ranaiefar, F., & Rixey, R. A. (2016). Bike Sharing Ridership Forecast using Structural Equation Modeling. In Transportation Research Board 95th Annual Meeting (No. 16-6573).

Santos, A., McGuckin, N., Nakamoto, H. Y., Gray, D., & Liss, S. (2011). Summary of travel trends: 2009 national household travel survey (No. FHWA-PL-ll-022).

Schneider, R., Diogenes, M., Arnold, L., Attaset, V., Griswold, J., & Ragland, D. (2010). Association between roadway intersection characteristics and pedestrian crash risk in Alameda County, California. Transportation Research Record: Journal of the Transportation Research Board, (2198), 41-51.

Schumacker, R. E., & Lomax, R. G. (2004). A beginner's guide to structural equation modeling. Psychology Press.

Schwartz, W., & Porter, C. (2000). Bicycle and pedestrian data: Sources, needs, and gaps.

StataCorp (2015). Stata Bayesian Analysis Reference Manual Release 14. College Station, TX: StataCorp LP

Thomson, G. H. (1935). The definition and measurement of" g"(general intelligence). Journal of Educational Psychology, 26(4), 241.

Tobey, H. N., Knoblauch, R. L., & Shunamen, E. M. (1983). Pedestrian Trip Making Characteristics and Exposure Measures. Final Report. Federal Highway Administration, Office of Safety and Traffic Operations.

Turner, S., Wood, G., Hughes, T., & Singh, R. (2011). Safety performance functions for bicycle crashes in New Zealand and Australia. Transportation Research Record: Journal of the Transportation Research Board, (2236), 66-73.

Wang, K., & Qin, X. (2014). Use of structural equation modeling to measure severity of single-vehicle crashes. Transportation Research Record: Journal of the Transportation Research Board, (2432), 17-25.

Xie, K., Ozbay, K., & Yang, H. (2016). A Joint Analysis of Secondary Collisions and Injury Severity Levels Using Structural Equation Models. In Transportation Research Board 95th Annual Meeting (No. 16-0206).

APPENDIX

Table 23 Description of Data Used for Modeling

Description	Mean	Std. Dev.	Min	Max
Total number of bicycle crashes(2004-2014)	1.296	1.988	0	11
Total number of pedestrian crashes(2004-2014)	0.474	2.205	0	12
Average pedestrian crashes(crashes/year)	0.299	0.464	0	2
Average Bicycle Crashes(crashes/year)	0.317	0.513	0	2
Intersection type	3.701	0.458	3	4
Intersection type: Three leg	0.299	0.458	0	1
Intersection type: Four leg	0.701	0.458	0	1
AADT of the major approach	14664	8879	1388	57285
AADT of the minor approach	6696.3	5207.7	346	24716
Number of exclusive through lane in the major approach	1.534	1.564	0	8
Number of shared through-right turn lane in the major approach	0.876	0.848	0	2
Number of share through-left turn lane in the major approach	0.126	0.403	0	2
Number of shared through-right-left turn lane in the major approach	0.209	0.584	0	2
Number of shared left-right turn lane in the major approach	0.039	0.193	0	1
Number of exclusive right turn lane in the major approach	0.570	0.739	0	3
Number of exclusive left turn lane in the major approach	1.253	0.849	0	4
Number of lane for leaving traffic in the major approach	2.925	1.367	0	9

Table 23-Continued

Description	Mean	Std. Dev.	Min	Max
Presence of crosswalk in the major approach	0.611	0.488	0	1
Presence of median in the major approach	0.088	0.283	0	1
Presence of pedestrian facility in the major approach	0.773	0.419	0	1
Presence of pedestrian facility separated from traffic in the major approach	0.446	0.498	0	1
Presence of bike lane in the major approach	0.018	0.133	0	1
Presence of on-street parking in the major approach	0.031	0.173	0	1
One way indicator for the major approach	0.008	0.088	0	1
Number of exclusive through lane in the minor approach	0.853	1.314	0	8
Number of shared through-right turn lane in the minor approach	0.784	0.841	0	2
Number of share through-left turn lane in the minor approach	0.147	0.427	0	2
Number of shared through-right-left turn lane in the minor approach	0.312	0.688	0	2
Number of shared left-right turn lane in the minor approach	0.088	0.283	0	1
Number of exclusive right turn lane in the minor approach	0.515	0.713	0	3
Number of exclusive left turn lane in the minor approach	1.080	0.924	0	4

Table 23-Continued

Description	Mean	Std. Dev.	Min	Max
Number of lane for leaving traffic in the minor approach	2.356	1.140	1	8
Presence of crosswalk in the minor approach	0.580	0.494	0	1
Presence of median in the minor approach	0.064	0.246	0	1
Presence of pedestrian facility in the minor approach	0.742	0.438	0	1
Presence of pedestrian facility in the minor approach separated from roadway	0.392	0.489	0	1
Presence of bike lane in the minor approach	0.015	0.124	0	1
Presence of on-street parking in the minor approach	0.039	0.193	0	1
One way indicator for the minor approach	0.018	0.133	0	1
Control type: Traffic signal	0.742	0.438	0	1
Control type: Two way stop sign	0.216	0.412	0	1
Control type: All way stop sign	0.041	0.199	0	1
Control type: Stop sign(Two way and all way combined)	0.258	0.438	0	1
Signal configuration/arrangement: Diagonal	0.466	0.500	0	1
Signal configuration/arrangement: Box	0.276	0.447	0	1
No turn on red on the major approach	0.015	0.124	0	1
Protected left turn on the major approach	0.302	0.460	0	1
No turn on red on the minor approach	0.018	0.133	0	1
Protected left turn on the minor approach	0.291	0.455	0	1
Presence of crosswalk	0.647	0.479	0	1

Table 23-Continued

Description	Mean	Std. Dev.	Min	Max
Presence of median	0.131	0.338	0	1
Presence of pedestrian facility	0.794	0.405	0	1
Presence of pedestrian facility separated from traffic	0.577	0.495	0	1
Presence of bike lane	0.034	0.180	0	1
Presence of on-street parking	0.590	0.492	0	1
One way indicator for the major approach	0.049	0.216	0	1
Presence of pedestrian facility	0.026	0.159	0	1
National functional classification: Arterial: Arterial	0.521	0.500	0	1
National functional classification: Collector-Arterial	0.358	0.480	0	1
National functional classification: Collector-Collector	0.121	0.327	0	1
Speed limit on the major approach	43.144	9.038	25	70
Speed limit on the minor approach	35.180	8.781	20	55
Walk score index	35.188	24.828	0	94
Proportion of land use by census block : Commercial	0.252	0.281	0	1
Proportion of land use by census block: Industrial	0.063	0.171	0	1
Proportion of land use by census block: Institutional	0.078	0.159	0	1
Proportion of land use by census block: Outdoor recreation	0.036	0.136	0	1
Proportion of land use by census block: Residential	0.570	0.316	0	1

Table 23-Continued

Description	Mean	Std. Dev.	Min	Max
Proportion of land use by area: Commercial	0.146	0.283	0	1
Proportion of land use by area: Industrial	0.054	0.184	0	1
Proportion of land use by area: Institutional	0.030	0.117	0	1
Proportion of land use by area: Outdoor recreation	0.025	0.132	0	1
Proportional of land use by area: Residential	0.746	0.357	0	1
Means of transportation: Percentage of worker using cars in a given census block	94.216	8.341	34.961	100
Means of transportation: Percentage of worker using public transport in a given census block	0.940	2.348	0	22.2
Means of transportation: Percentage of worker using bus in a given census block	0.931	2.321	0	22.2
Means of transportation: Percentage of worker using taxi in a given census block	0.055	0.512	0	9.2
Means of transportation: Percentage of worker using motorcycle in a given census block	0.217	0.615	0	5.9
Means of transportation: Percentage of worker biking in a given census block	0.044	0.187	0	2.4
Means of transportation: Percentage of worker walking in a given census block	1.453	4.595	0	40.8
Bicyclist commuter density for a census block	0.596	2.716	0	29.8

Table 23-Continued

Description	**Mean**	**Std. Dev.**	**Min**	**Max**
Walking commuter density for a census block	34.638	145.51	0	1671.9
Percentage of household above poverty level in given census block	79.934	25.587	0.468	100.0
Percentage of household below the poverty level in a given census block	13.128	14.032	0	83.7
Percentage of whites in a census block	79.409	26.327	0	100.0
Percentage of blacks in in a census block	13.446	25.636	0	99.7
Percentage of Indian Alaska in a census block	0.385	1.591	0	28.1
Percentage of Asian in 0.25 mile in a census block	3.167	5.081	0	35.6
Population density in a census block	412.574	366.545	0	2384.9