SUSTAINABLE ROAD SAFETY: DEVELOPING COMMUNITY-BASED MACRO-LEVEL COLLISION PREDICTION MODELS OF INCREASED BICYCLE USE IN THE REGIONAL DISTRICT OF CENTRAL OKANAGAN

by

FENG WEI

B.Eng., Beijing Forestry University, China, 2006

M.Sc., University of Warwick, UK, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

The College of Graduate Studies

(Civil Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA
(Okanagan)

October 2012

ABSTRACT

Since the energy crisis, global warming, transportation congestion, and road safety issues, governments and individuals are more willing to use sustainable transportation modes to support the development of compact neighbourhoods. Bicycling, as one of the most effective modes for short-distance trips, is widely encouraged. However, as vulnerable road users (VRUs), cyclists are more likely to be injured when they are involved in collisions. In order to prevent bicycle collisions from occurring, it is necessary to conduct proactive evaluation of cyclists' road safety. Community-based, macro-level collision prediction models, as prospective empirical tools to evaluate and predict road safety, have been suggested in reactive road safety applications and road safety planning. This research reviews previous studies on road safety and bicycle use and summarizes different regression methods for collision prediction model (CPM) development. On the basis of insights gained in the literature review, community-based, macro-level CPMs related to bicycle use were developed using the generalized linear regression (GLM), zero-inflated count regression (ZIC), geographically weighted regression (GWR), and full Bayesian (FB) methods, based on data from the Regional District of Central Okanagan (RDCO) of British Columbia, Canada. The statistical associations of total/severe/bike-vehicle collisions and their neighbourhood traits, which were derived from these model results, are reasonable. In the reactive road safety application of macro-level black spot study, bike-vehicle collision prone zones (CPZs) were identified and ranked in the RDCO with the developed GLM and FB collision prediction models and preliminary diagnoses and remedies for these CPZs are suggested. Finally, data gaps for macro-level CPM development were identified. In order to improve data quality and linkage, several suggestions about how to build an integrated data warehouse for vulnerable road user collisions are proposed.

PREFACE

Parts of the contents of this thesis have been published in two journal papers, co-authored with Dr. Gordon Lovegrove. The two hypotheses of the relationships between bicycle use and road safety in the first chapter, and the development of macro-level bicycle collision prediction models with negative binomial regression in the third chapter have been presented in "An Empirical Tool to Evaluate the Safety of Cyclists: Community-based, Macro-level Collision Prediction Models Using Negative Binomial Regression" which has been accepted by the Journal of Accident Analysis & Prevention. Also, the literature summary on road safety and bicycle use in second chapter has been presented in another paper "Sustainable Road Safety: A New (?) Neighbourhood Road Pattern That Saves VRU Lives", which has been published in the Journal of Accident Analysis & Prevention (Vol 44, Iss.1, pp. 140–148, 2012).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC = Akaike Information Criterion

AICc = Adjusted Akaike Information Criterion

AREA = Zonal area

ALKP = Percentage of Arterial Lane Kilometres

BC = Province of British Columbia, Canada

BCIRPU = BC Injury and Research Prevention Unit

Black Spot = Hazardous location

BKT = Bicycling Kilometres Travelled

BLKM = Bike Lane Kilometres

BS = Zonal Bus Stops

BSD = Zonal Bus Stop Density

CCR = Collision Risk Ratio

CMF = Collision Modification Factor

CORE = Zonal Core area

CPL = Collision Prone Location

CRP = Core to Zonal Area percentage

CPZ = Collision Prone Zone

CPM = Collision Prediction Model

CSI = Collision Severity Index

DA = Dissemination Area

DIC = Deviance Information Criterion

DoF = Degree of Freedom

DRIVE = Zonal commuters who Drive

DRP = Percentage of Commuters Who Drive

EB = Empirical Bayesian

EMPD = Zonal Employed Density

FB = Full Bayesian

FS = Zonal Average Family Size

GLM = Generalized Linear Regression Modeling

G. SD = Grouped Scaled Deviance

GWR = Geographically Weighted Regression

ICBC = Insurance Corporation of British Columbia

INCA = Zonal Average Income

INT = Zonal Intersections

INTD = Zonal Intersection Density

I3WP = Percentage of 3-way Intersections

IALP = Percentage of Arterial-local Intersections

LLKP = Percentage of Local Lane Kilometres

Macro-level = Area-wide (e.g. neighbourhood)

Macro-reactive = Reactive use of Macro-level CPMs

Macro-proactive = Proactive use of Macro-level CPMs

MAD = Mean Absolute Deviation

MCMC = Markov Chaim Monte Carlo Method

Micro-level = Single location (e.g. intersection)
MSE = Mean Squared Error
MSPE = Mean Squared Prediction Error
NB = Negative Binomial
NHD = Zonal Home Density
NW = Network
OR = Odds Ratio
PCR = Potential Collision Reduction
POPD = Population Density
POP30 = Percentage of Population at 30 years and under
RDCO = Regional District of Central Okanagan
RSA = Road Safety Audits
RSIP = Road Safety Improvement Program
RSPF = Road Safety Planning Framework
RSRI = Road Safety Risk Index
RTM = Regression to the Mean
S-D = Socio-Demographic
SD = Scaled Deviance
SIG = Zonal Signals
SIGD = Zonal Signal Density
SPF = Safety Performance Function
SRS = Sustainable Road Safety
STS = Sustainable Transport Safety
TAZ = Traffic Analysis Zone
TCM = Total zonal Commuters
TDM = Transportation Demand Management
TLKM = Total Lane Kilometres
VKT = Vehicle Kilometres Travelled
VRU = Vulnerable Road Users
ZIC = Zero-inflated Count Regression
ZIP = Zero-inflated Poisson Regression
ZINB = Zero-inflated Negative Binomial Regression

ACKNOWLEDGEMENTS

This thesis has only been possible with the gracious support of many people.

I am deeply grateful to my supervisor Dr. Gordon Lovegrove for his enthusiastic and expert guidance, encouragement in my academic life. I would like to give my thanks to my committee members: Dr. Deborah Roberts, Dr. Rehan Sadiq, and Dr. Tarek Sayed, for their care with what they reviewed the original manuscript; and for conversations that clarified my thinking on this and other matters. Dr. Carolyn Labun also provided technical writing supports for this thesis: my thanks to her too.

I would like to express my appreciations to Insurance Corporation of British Columbia, BC Injury and Research Prevention Unit, City of Kelowna, BC Transit, for their data and for their trusted advice, without which the data for research contained in this thesis would not have been realized. I especially want to thank Dr. Peter Barss at Interior Health, for his cooperation, supports, and valuable suggestions to my research.

My thanks to my landlords Sylvia and Lou, for their care and protection. To my dear friends Molly, Maggie, Jessica, Nathan, Roger, Peter and Xianchang, for their company and academic experience share with me in my studying abroad life. To my dear friends in China: Huan, Xue, Nan, Hong, Yu, and Xia, for their friendship to share everything with me. Special thanks to Yuan, my cousin and best friend, for always be there, to comfort me whenever I need.

Last, but not least, I want to thank all of my family members for their love and supports. Specially, I thank my uncle, Xiaolin Wei, for motivating me to be a better me and the many years of support during my graduate studies that provided the foundation for this work. My deepest gratitude and love to my parents, Xiaoyu Tian and Xiaofu Wei, for their love and dedication since I was born. Their support and encouragement was in the end what made this thesis possible.

CHAPTER 1 INTRODUCTION

Although many articles refer to road safety events in terms of *accidents*, this thesis uses *collisions* to replace *accidents*, because road safety is no accident and can be predicted and prevented if reliable empirical tools are available. Road collisions cause individual fatalities, injuries and property loss. According to World Health Organization reports, road traffic injuries (RTI) were the 11th leading cause of death in 2004, and the 9th leading cause of death in 2008 in the world (WHO, 2004; WHO, 2008). Related research projections indicate that RTI would be the third leading cause of death by 2020, unless there is new commitment to prevention (Peden et al., 2004; Gasper, 2004). The economic cost of RTI was estimated to be 1% of the gross national product in low-income countries, 1.5% in middle-income countries and 2% in high-income countries. The global cost of RTIs was estimated to be US$ 518 billion per year (WHO, 2004). In Canada, the annual economic cost of traffic injury and property damage was estimated at between $11 and $27 billion (Transport Canada, 2005).

Today, sustainable transport modes such as biking, walking, public transit and car-sharing are widely encouraged by governments and individuals, not only to reduce the enormous loss due to road collisions, but also to reduce greenhouse gas emissions, traffic congestion, and non-renewable energy consumption. Developing sustainable land use and transportation environment, including less dependence of single occupancy vehicles and more use of sustainable transport modes, is an effective way to achieve a sustained reduction in road collision risks, frequencies, and severity. The main goal of this research is to develop empirical tools that can be used to test or demonstrate the relationship between road safety and bicycle use. This chapter introduces the research background, research objectives, and the thesis structure.

1.1 Research Background

Previous studies suggested that increased bicycle mode splits were statistically linked to lower road traffic fatalities. Based on the 1990s data, Newman and Kenworthy (1999) reported that Amsterdam and Copenhagen, two European cities with medium levels of bicycle use, had average road traffic fatality rates of 5.8 and 7.5 per 100,000 population

(POP) per year, respectively. These fatality rates were less than half of the average fatality rate for US cities (i.e. 14.6/100,000). Similarly, Osberg and Stiles (1998) compared road safety in Boston, Paris and Amsterdam based on the road traffic fatality data in the 1990s. These three cities are similar in socio-economic term but significantly different in bicycle use: a very low level in Boston, a low-medium level in Paris, and a medium level in Amsterdam. Osberg and Stiles showed that despite holding the highest bicycle fatality rate of the three with 1.8/100,000 POP (0.6/100,000 for Paris and 0.3/100,000 for Boston), Amsterdam had the lowest total road fatality rate, at 5.8/100,000 POP (10.0/100,000 for Paris and 8.8/100,000 for Boston). Additionally, Marshall and Garrick (2011) examined an 11-year road safety dataset (1997-2007) in 24 California cities and found that the cities with high bicycle use had lower total road fatality rates than the cities with low bicycle use.

Although increased bicycle collisions/injuries were associated with increased bicycle use, previous studies suggested that decreased bicycle collision/injury risks were associated with increased bicycle use, which means the increase in bicyclist collisions/injuries was less than the increase in bicycle volume. Two studies focused on the relationship of bicycle collision risk and bicycle use at a micro-level (e.g. intersections). Ekman (1996) found there was an apparently inverse relationship between the collision risk per bicyclist and bicycle volumes by examining bicyclist volume and severe bicycle collision data at 95 intersections in Malmö, Sweden. Leden et al. (2000) conducted a before and after study, then concluded that the collision risk per cyclist decreased by about 20% with a 50% increase in bicycle volume at 45 non-signalized intersections in Gothenburg, Sweden. Other two "safety in number" studies focused on the relationship of bicycle collision risk and bicycle use at a macro level. Based on five datasets (three population level data and two time series data) from 68 Californian cities, 47 Danish towns and 14 European countries, Jacobsen (2003) developed a safety performance model to estimate cyclists' injury risk in a community. He suggested that the injury risk per cyclist would decrease by 34% if cycling doubled in a community. Robinson (2005) reviewed three datasets in Australia and also showed that doubling cycling distance or volume was associated with a 35% decrease in bicyclist fatality or injury risk.

Although these above studies show that more bicycle use is related to more road safety, one study in the Netherlands shows that a 10% shift from cars to bicycles for the trips with distances shorter than 7.5 kilometres would lead to an increase of 4-8 deaths and 500 serious

injuries per year (SWOV, 2010). However, this conclusion is questionable because an incorrect linear relationship between injuries and traffic exposure is assumed in this study.

Based on the above studies, two hypotheses (see Figure 1.1) are proposed to depict the relationship between road safety and increased bicycle use. The upper dashed line in Figure 1.1 depicts the hypothesis regarding the statistical relationship between overall road safety and the bicycle mode split. In this hypothesis, increased bicycle use will lead to a significant reduction in total traffic collisions. However, it is still unknown whether the dashed line would decrease linearly, concavely, or convexly. The lower dashed line in Figure 1.1 depicts the hypothesis of the relationship between cyclists' safety and bicycle mode split. In this hypothesis, initial increases from a low level to some unknown medium level of cycling use will likely cause increased bicycle-related collisions; however, as bicycle use increases beyond that unknown medium level, there would be a decrease in bicycle-related collisions.

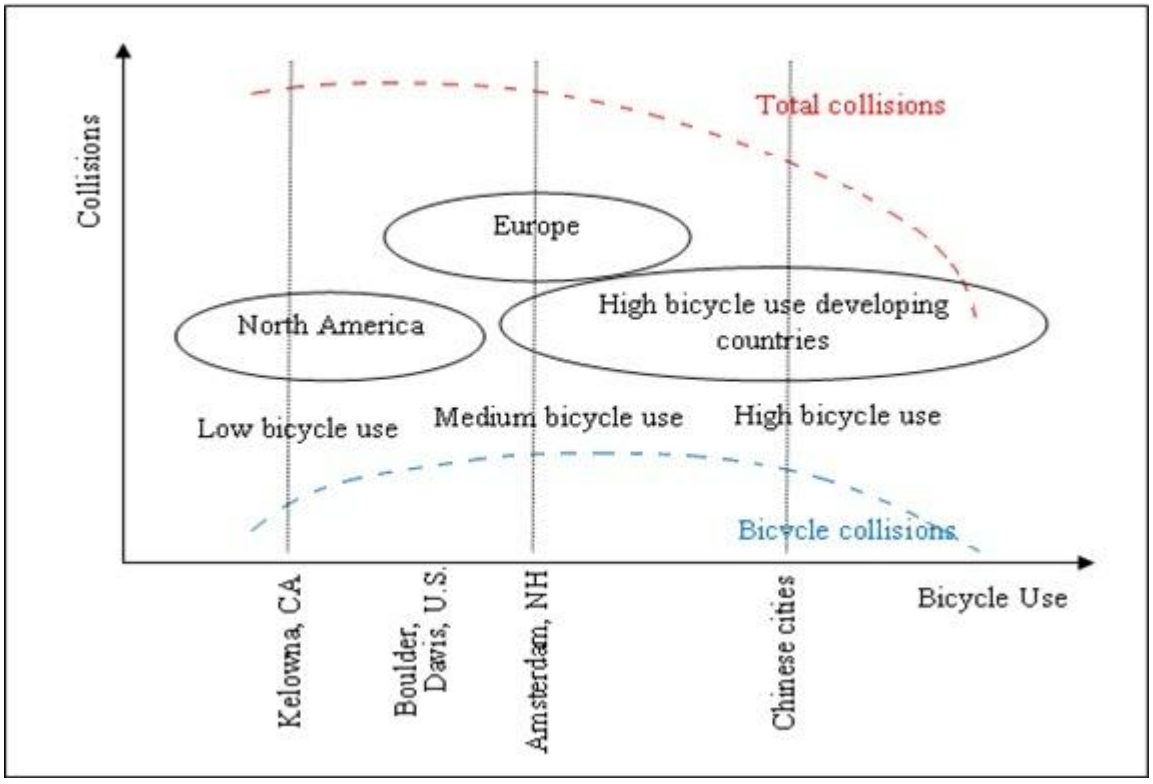

Figure 1.1 Hypotheses on Relationships between Bicycle Use and Road Safety

Extensive research would be needed to demonstrate or test these two hypotheses in different bicycle use levels and regions using reliable empirical tools. This study contributes

to that future work by examining the relationship between bicycle use and road safety in a North American region with a typical low bicycle use level. Macro-level collision prediction models (CPMs) are regarded as reliable empirical tools to be used in proactive and reactive road safety applications (Lovegrove, 2007; Hadayeghi et al., 2003, 2009, 2010). In proactive applications, macro-level CPMs are used for road safety planning to prevent potential road safety problems. In reactive applications, macro-level CPMs are used to identify existing road safety problems in built environment or to measure the effectiveness of any community-based road safety countermeasure.

1.2 Research Objectives

Although a few studies on community-based, macro-level CPMs have been conducted before, they are rarely used to evaluate road safety with increased bicycle use in neighbourhoods. Following the methodologies in previous community-based, macro-level CPM research, this study attempts to investigate the statistical relationship between road safety and bicycle use in a BC regional district. The following three research objectives are included in this study.

1.2.1 Objective 1: Develop Macro-level CPMs Using Different Regression Methods

The first research objective is to use different regression methods to develop community-based, macro-level CPMs related to bicycle use. The literature reveals two main methods for macro-level CPM development: 1) artificial neural network methods, and 2) regression methods. Artificial neural network methods provide a flexible way to model nonlinear relationships. However, they are criticized for acting as 'black boxes' because the knowledge contained in a neural network model is kept in the form of a weight matrix, making it hard to interpret the statistical associations between dependent and independent variables. Therefore, regression methods are favoured for model development as they are able to clearly show the statistical associations between dependent and independent variables in models.

Many regression methods can be used for CPM development. However, only four of them are researched in this study. They are most commonly used regression methods to modelling discrete, non-negative and rare events like traffic collisions. The four regression

methods are generalized linear modelling or regression (GLM), zero-inflated count regression (ZIC), geographically weighted regression (GWR), and full-Bayesian regression (FB). In this study, community-based, macro-level CPMs are developed using these four regression methods, based on data from the regional district of central Okanagan (RDCO). Finally, the model results from different methods are compared to suggest the most reliable method(s).

1.2.2 Objective 2: Conduct a Macro-reactive Safety Application Using Macro-level CPMs.

The second research objective of this thesis is to conduct a macro-level black spot program using the bicycle-vehicle CPMs developed in *Objective 1*. Black spots here are hazardous collision prone zones (CPZs) with significant above-average (95% level) collision frequencies. This macro-reactive black spot study is used in this study to identify and rank bicycle CPZs, diagnosis road safety problems for bicycle collisions, and suggest possible remedies in the RDCO. The bicycle CPZ identification results from different regression models are compared to recommend the 'most practical' (at this time intended to mean the 'most understandable and least complex for use') identification and ranking method(s).

1.2.3 Objective 3: Identify Data Issues and Needs for Model Development

The third research objective is to identify data issues and needs in developing community-based, macro-level CPMs. This research objective will help researchers to pursue an ongoing, long-term, future sustainable transport safety research goal of developing a global model that predicts *total* collisions (i.e. the sum of vehicle-vehicle, bike-vehicle, pedestrian-vehicle, bike-bike, bike-pedestrian collisions) instead of only *total vehicle* collisions.

The development of community-based, macro-level CPMs in *Objective 1* is based on geo-coded collision data from the Insurance Corporation of British Columbia (ICBC), a provincial crown corporation providing mandatory no-fault auto insurance to all BC motorists. Despite being the largest collision database in BC, the ICBC collision database has no non-vehicle collision records. Therefore, for bicycle collision data, the ICBC

database records bike-vehicle collisions but does not record bike-bike, bike-pedestrian, and single bike collisions. According to anecdotal evidence, Stallard (2006) suggested that there were at least as many bicycle collisions, if not more, that did NOT involve vehicles. The bicycle injury data from the British Columbia Injury and Research Prevention Unit (BCIRPU) support this opinion. In this study, the bike collision data from ICBC and the bicycle injury data from BCIRPU are compared to identify the data gaps in current road collision/injury databases. Based on these data gaps, suggestions about data linkage and sharing from different sources are proposed.

1.3 Thesis Structure

This thesis contains six chapters. *Chapter 1* introduces the research background, the research objectives, and the thesis structure. *Chapter 2* reviews the literature on road safety improvement programs, regression methods of CPM development, macro-level CPM development and applications, the factors related to bicycle use, and the data issues in road safety programs. *Chapter 3* describes the data extraction, model development, and model results of community-based, macro-level CPMs using four regression methods, based on the data from the regional district of central Okanagan (RDCO). *Chapter 4* presents a reactive road safety application in the RDCO based on developed macro-level CPMs. *Chapter 5* identifies the data issues and challenges in macro-level CPM development, and clarifies future research needs in data sharing and linkage. *Chapter 6* summarizes the research conclusions, contributions, and limitations, and offers recommendations for future work.

CHAPTER 2 LITERATURE REVIEW

This chapter summarizes the literature related to the collision prediction model development, applications, and collision data issues. The database TRID was mainly used for literature search. It is the largest database in transportation engineering, integrating the scholar article records from Transport Research Board's Transportation Research Information Services Database and Joint Transport Research Centre's International Transport Research Documentation Database. Other databases like ASCE Research Library, Compendex Engineering Village, and MathSciNet were also used for literature search. The search language is English, and the date range is from 1980 to the present. The search key words include: road safety, bicycle safety, bicycle use, cyclists, accidents, collisions, road safety problems, safety in numbers, collision/accident prediction models, macro-level, regional, generalized linear regression, full Bayesian, geographically weighted regression, zero-inflated count regression, road safety planning, sustainable transport, black spot studies, traffic exposures, collision/accident data, and road injury data.

Seven sections are included in this chapter. *Section 2.1* reviews road safety improvement programs. *Section 2.2* summarizes different criteria of road safety performance evaluations and describes the historical process of collision prediction models (CPMs). *Section 2.3* reviews four regression techniques for CPM development and their comparison results in previous studies. *Section 2.4* presents proactive and reactive road safety applications of macro-level CPMs. *Section 2.5* reviews the factors influencing bicycle use. *Section 2.6* summarizes data sharing issues existing in other road safety studies. Finally, a summary of this chapter is provided in *Section 2.7*.

2.1 Sustainable Road Safety Improvement Programs

Road safety improvement programs are achieved by reactive and proactive approaches (de Leur & Sayed, 2003; Lovegrove & Sayed, 2006; Lovegrove, 2007). Traditional road safety improvement programs rely on reactive engineering approaches, which require collision history records to identify hazardous locations (or black spots). Therefore, before any safety improvement countermeasure can be taken in these identified black spots, road collisions have already occurred and caused enormous loss. Proactive approaches are thus

pursued to allow engineers and planners to target road safety at an early planning stage with the potential for significant reductions in collisions below those achieved to date. De Leur & Sayed (2003) suggested that if road safety were addressed before any program was started, it could reduce the number and cost of reactive safety countermeasures needing to be retrofitted into existing communities. Moreover, proactive approaches can be decision-making tools in the selection of possible transport planning programs once road safety is set as an evaluation factor.

Several proactive or sustainable road safety studies have been conducted in Europe and North America. Dutch researchers began to do road safety planning analyses in the early 1990s and launched the Dutch Sustainable Safety Vision (DSSV). In the DSSV, a sustainable road safety (SRS) program on road infrastructure planning and design was proposed (Wegman et al., 2006). Canadian researchers also developed a proactive road safety planning framework (RSPF) that lays further groundwork towards quantifying a planning-level predictive relationship (de Leur & Sayed, 2002; 2003). Following that, American researchers developed a safety conscious planning (SCP) program for a similar purpose (Herbel, 2004). In road safety planning processes, the major challenges are the lack of data and the lack of proactive safety evaluation methods (Chatterjee et al., 2004; Washington et al., 2006; Lovegrove, 2007; Wang, 2011). The above studies are only good for preliminary road safety planning research because they lack reliable empirical tools for proactive road safety evaluation. Without such tools, road safety engineers cannot predict future road safety levels.

Three proactive empirical tools, road safety risk indices, road safety audits, and CPMs, are reviewed as following. Road safety risk indices (RSRI) were proposed by de Leur and Sayed (2002). The motivation of RSRI is to produce a tool to support road safety analyses that do not rely on historical collision data. For each planned or built facility, a RSRI can be quantified using 30 safety planning principles related to exposure (i.e. land use, network shape, model choice), probability (i.e. manoeuvrability, geometric design, functionality, conflicts, road friction, road predictability), and consequence (i.e. vulnerable users, reduce speed, roadside). RSRI combine qualitative and quantitative safety measures in an analytically hierarchical approach; however, these measures are significantly related to observers' subjective assessments on road safety. Road safety audits (RSAs) are proven road

safety engineering tools to provide an explicit, formalized safety evaluation of road programs of any size. They can be applied in existing identified hazardous locations or road planning and design programs. An independent, multi-agency team needs to perform an 'audit' or safety evaluation of a location to identify road safety improvement measures. However, collision potential forecasts and safety improvements involved in RSAs still rely heavily on experience and professional judgment. As the practice of planning level road safety analyses spreads, improved empirical tools are being pursued to provide reliable safety planning evaluation for entire communities. Collision prediction models (CPMs) could be considered as such tools.

There are two kinds of CPMs: micro- and macro-level CPMs. Both of them are empirical tools for road safety evaluations and have similar methodologies in model development and applications. Micro-level CPMs are used for road safety evaluations of road intersections or segments. They have been fully developed and widely used as reactive engineering approaches in traditional road safety improvement programs. Macro-level CPMs are used for road safety evaluations and prediction of communities, districts, or states. Macro-level CPMs are being expected to be researched more. Their development and application methodologies are based on and similar to those of micro-level CPMs. Initially, researchers attempted to use micro-level CPMs in road safety planning (Volk et al., 1999; Lord & Persaud, 2004), but found that they could not fill the gap between what is needed and what is available in terms of reliable safety planning tools for proactive road safety improvement programs. Therefore, macro-level CPMs were suggested for road safety planning (Hadayeghi et al., 2003, 2009, 2010; Ladron de Guevara et al., 2004; and Lovegrove & Sayed, 2006). Lovegrove (2007) developed guidelines for the use of macro-level CPMs in both reactive and proactive road safety programs. More literature about macro-level CPMs is reviewed in the following sections.

2.2 Basic Statistics of Collision Prediction Models

This section reviews basic statistics related to collision prediction model development, including three road safety evaluation measures and statistical techniques for road safety evaluations. All of these techniques have formed to the basis of currently using collision prediction models (CPMs).

2.2.1 Road Safety Evaluation Measures

The safety performance of a location is usually evaluated by three measures, including: collision frequency, rate, and severity, either combined or separately. The collision frequency measure is defined as the number of collisions occurring at a location during a specific period. Hazardous locations may be identified and ranked simply by comparing their collision frequencies to the average collision frequency of a reference group. The period for measuring collision frequencies is important. If the frequency of collision events in a particular location is low, a short sampling period is not recommended because the collision frequency in the short period could be a random fluctuation around its average value (Rodegerdts et al., 2004). Therefore, the use of a longer sampling period is suggested to provide a more stable frequency. Zegeer (1982) suggested that a sampling period of one- to three-years might minimize the effect of random fluctuations. However, even a period of three years may not be enough to account for random fluctuations; therefore, a three- to five year sampling period was recommended by Lovegrove (2007). Collision frequencies are easily computed. In order to accurately identify black spots, it is important to consider both collision frequency and traffic exposure.

The collision rate in a particular location is calculated by dividing collision frequency with some unit of exposure, such as million-vehicle-kilometres, million-entering-vehicles, or million populations. Collision rates can take traffic exposure into account when they are used to identify black spots. However, the ratio of collisions per unit of traffic exposure assumes a linear relationship between collisions and traffic exposure, which has been shown to be incorrect (Hauer, 1995).

The collision severity index (CSI) was proposed to measure collision severity. The CSI is defined as the weighted sum of fatal, injury, and property damage only (PDO) collisions. Different transportation authorities or research organizations assign different weights to collision severity, based on the cost and impact of the collision. For example, the Institute of Transportation Engineers (1999) weights fatal or incapacitating injury collisions at 9.5 times, and non-incapacitating or invisible injury collisions at 3.5 times the severity of PDO collisions. In Alberta and British Columbia, some transportation authorities weight fatal collisions at 100 times, and injury collisions at 10 times the severity of PDO collisions

(Duckworth et al., 2011, Sayed, 1999). If the CSI at one location is greater than the average CSI from a reference group, this location is regarded as a black spot.

2.2.2 Statistical Techniques

Because of the random component in the observed collision frequency at one location, it is better to consider the occurrence of collisions to be a random variable, which cannot be predicted with absolute accuracy. To do that, the mean collision frequency of this location can be estimated as an *expected* value using reasonable assumption and empirical techniques (Sayed, 1998). Statistical techniques are used to screen out randomness in the observed historic mean collision frequency in order to provide an estimate of the expected value of the mean collision frequency with some degree of accuracy (Lovegrove, 2007). The accuracy of an estimate is usually expressed in terms of an estimate's standard deviation or variance.

Classical statistical techniques were used in early engineering practices to identify black spots. Sayed (1999) proposed a confidence interval technique, in which a location is identified as a black spot if

$$X_i > \lambda + k\sigma \tag{2.1}$$

where X_i is the observed collision rate at the site i, λ is the average collision rate from a reference population of sites with similar traits, k is the statistic coefficient obtained from the normal distribution function for a desired confidence interval (e.g. 95% confidence interval), and σ is the standard deviation of the population. The confidence interval technique assumes that the occurrence of collisions per unit period is normally distributed, which does not fit with the historical data (Norden, 1956; Oppe, 1982, 1992). Norden (1956) suggested a rate quality control technique to identify black spots. In this technique, the occurrence of collisions per unit period is assumed to follow a Poisson distribution, which better fits the actual case. A location is identified as a black spot if

$$X_i > \lambda + k\sqrt{\frac{\lambda}{m}} + \frac{1}{m} \tag{2.2}$$

where X_i is the observed collision rate at site i, λ is the average collision rate from a reference group, k is the statistic coefficient related to a desired confidence level, and m is the traffic exposure (i.e. million-entering-vehicle) at the site. These two statistical techniques are able to account for the randomness bias. However, the collision rate used in these two

techniques may be not the best measure to identify black spots, because the incorrect assumption of a linear relationship between collisions and traffic exposure still exists when the collision rate is used. Therefore, collision prediction models (CPMs) were proposed to model the non-linear relationship between collisions and traffic exposure. In CPMs, the collision frequency measure is used to identify black spots.

Considerable research on CPMs has been conducted. Simple CPMs use one traffic exposure variable as the only independent variable to estimate the expected values of road collisions or injuries at locations or communities. For example, based on five datasets from 68 Californian cities, 47 Danish towns and 14 European countries, Jacobsen (2003) used bicycling/walking exposure to estimate bike-auto/pedestrian-auto injuries in communities. The model function is shown as

$$\mathrm{I} = aE^{b} \tag{2.3}$$

where I is the estimated bike-auto/pedestrian-auto injuries in a community, E is the measure of bicycling/walking exposures, and a and b are model parameters. The deficiency of simple CPMs is that they do not involve other potential independent variables in models in addition to the traffic exposure variable. Therefore, more accurate CPMs are preferred to account for other independent variables. As mentioned in *Section 2.2.1*, CPMs are divided into micro- and macro-levels. The development of macro-level CPMs is based on micro-level CPMs. As this study focuses on community-based, macro-level CPMs, previous studies on macro-level CPM development and the basic statistic techniques associated with CPMs are generally reviewed as follows.

Based on the data derived from 463 traffic zones in Toronto, Ontario, Hadayeghi et al. (2003) developed community-based, macro-level CPMs using generalized linear model (GLM) techniques. The model form used in this study is presented as:

$$E(\Lambda) = a_0 VKT^{a_1} e^{\sum_{j=1}^{n} b_j X_j} \tag{2.4}$$

where $E(\Lambda)$ is the expected collision frequency at one zone, a_0, a_1, and b_j are parameters, VKT is the zonal vehicle-kilometres travelled from Emme/2 forecast, and X_j are other explanatory variables (e.g. arterial road lane kilometres, the number of households, area, posted speed, and average zonal congestion). Based on their model results, Hadayeghi et al. (2003) showed that increased collisions were associated with increased vehicle kilometres travelled,

households, major road kilometres, and intersection density. However, decreased collisions were associated with increased average posted speed and average zonal congestion.

Based on data from Tucson, Arizona, Ladron de Guevara et al. (2004) developed community-based, macro-level CPMs using negative binomial regression, which is one of GLM techniques. The model from in this study is presented as:

$$E(\Lambda) = e^{\sum_{j=1}^{n} b_j X_j} \tag{2.5}$$

where X_i are independent variables representing community traits. Ladron de Guevara et al. (2004) found increased collisions/injuries were associated with increased population density, total employment, intersection density, major arterial roads, minor arterial roads, and urban collector roads.

Lovegrove and Sayed (2006) also used negative binomial regression to develop community-based, macro-level CPMs based on data from the Greater Vancouver Regional District (GVRD), British Columbia. The model form in this study is similar to *Eq. 2.4* and written as

$$E(\Lambda) = a_0 Z^{a_1} e^{\sum_{j=1}^{n} b_j X_j} \tag{2.6}$$

where Z is a zonal traffic exposure variable which has a dominant statistical influence on collision numbers (e.g. vehicle kilometers traveled or total lane kilometres). Both of these two model functions (i.e. *Eq. 2.5* and *Eq. 2.6*) guarantee the expected collisions to be non-negative. However, *Eq. 2.6* is better than *Eq. 2.5* because: 1) *Eq. 2.6* ensures that zero exposures generate zero collisions, and 2) *Eq. 2.6* differentiates the effects of the leading variable Z and other independent variables X on the dependent variable.

2.3 Regression Methods for CPM Development

This section reviews four regressions in detail. These regression techniques include generalized linear regression, zero-inflated count regression, geographically weighted regression, and full Bayesian regression. They are most commonly used in CPM development.

2.3.1 Generalized Linear Regression

Generalized linear regression/modelling (GLM) is defined as an extension of the linear regression modelling that allows models to be fitted to data that follow probability distributions other than the Normal distribution, such as binomial, negative binomial, Poisson, Gamma, etc. (Mcullagh & Nelder, 1989). The probability distribution of the dependent variable could be discrete or continuous. Poisson and negative binomial distributions are discrete distributions used to model rare, sporadic, and non-negative collision data. Therefore, Poisson regression and negative binomial regression are agreed to be the two promising GLM methods in developing CPMs (Hauer et al., 1988; Miaou & Lum, 1993; Sawalha & Sayed, 2006; Lovegrove 2007).

In Poisson regression, given an expected mean collision count λ_i, the observed collision count y_i follows a Poisson distribution. The probability mass function of the Poisson distribution is written as:

$$P(Y_i = y_i \mid X_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{2.7}$$

The expected collision λ_i is derived from a model with a series of covariates X_i. The commonly used model forms have been reviewed in *Section 2.2.2*, and repeated here:

$$E(Y_i \mid X_i) = \lambda_i = E(\Lambda_i) = e^{\sum_{j=1}^{n} b_j X_{ij}} \tag{2.5}$$

$$E(Y_i \mid X_i) = \lambda_i = E(\Lambda_i) = a_0 Z^{a_1} e^{\sum_{j=1}^{n} b_j X_{ij}} \tag{2.6}$$

In order to estimate parameters, these model forms need to be transformed into logarithmic functions (see *Eq. 2.8* and *Eq. 2.9*), in which expected collisions have linear relationships with their explanatory covariates.

$$\ln E(Y_i \mid X_i) = \ln \lambda_i = \sum_{j=1}^{n} b_j X_{ij} \tag{2.8}$$

$$\ln E(Y_i \mid X_i) = \ln \lambda_i = \ln a_0 + a_1 \ln(Z) + \sum_{j=1}^{n} b_j X_{ij} \tag{2.9}$$

In a Poisson regression model, the collision mean equals to the collision variance:

$$E(y_i) = Var(y_i) = \lambda_i \tag{2.10}$$

This feature of Poisson regression does not match the over-dispersion nature of collisions. Over-dispersion is the presence of greater variability in a dataset, which is expressed as the sample variance is being greater than the sample mean. Therefore, negative binomial regression is recommended to account for over-dispersion by introducing an unobserved heterogeneity term in a CPM form. The form is presented as:

$$E(Y_i \mid X_i) = \lambda_i \tau_i \tag{2.11}$$

where τ_i follows a gamma distribution with $E(\tau_i)$=1 and $Var(\tau_i)$ =1/κ. Given X_i and τ_i, the dependent variable Y_i still follows a Poisson distribution, whose probability mass function is written as

$$P(Y_i = y_i \mid X_i, \tau_i) = \frac{e^{-\lambda_i \tau_i} \lambda_i \tau_i^{\,y_i}}{y_i!} \tag{2.12}$$

However, if only given X_i, Y_i follows a negative binomial (NB) distribution. The probability mass function of the NB distribution is written as

$$P(Y_i = y_i \mid X_i) = \frac{\kappa^{\kappa} \lambda_i^{\,y_i} \Gamma(\kappa + y_i)}{\Gamma(y_i + 1)\Gamma(\kappa)(\lambda_i + \kappa)^{\kappa + y_i}} \tag{2.13}$$

Therefore, NB regression is termed Poisson-Gamma regression. In a NB regression model, the collision variance is greater than the collision mean, which is formulated as

$$Var(y_i) = E(y_i) + \frac{E(y_i)^2}{\kappa} = \lambda_i + \frac{\lambda_i^2}{\kappa} \tag{2.14}$$

The dispersion parameter σ_d is used to decide which regression to use for model development (McCullagh and Nelder, 1989). This parameter assesses the level of variation in the data. A value of σ_d greater than 1 indicates an apparent dispersion in the data, so a NB regression is used. The dispersion parameter is given as

$$\sigma_d = \frac{Pearson\chi^2}{n - p} \tag{2.15}$$

where p is the number of parameters.

In the Poisson and NB regression procedures, the maximum likelihood estimation method is commonly used for providing model parameter estimates (Lovegrove 2007; Sawalha & Sayed, 2006). Two goodness-of-fit tests, scaled deviance (SD) and Pearson χ^2, are used to check how well a GLM model fits a set of observations (McCullagh & Nelder, 1989). These two goodness-of-fit tests for NB models are written as:

$$SD = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{E(\Lambda_i)}\right) - (y_i + \kappa)\ln\left(\frac{y_i + \kappa}{E(\Lambda_i) + \kappa}\right)\right] \tag{2.16}$$

$$Pearson\chi^2 = \sum_{i=1}^{n}\frac{[y_i - E(\Lambda_i)]^2}{Var(y_i)} \tag{2.17}$$

where y_i is the observed collision frequency at location i, $Var(y_i)$ is the variance for y_i, $E(\Lambda_i)$ is the predicted collision frequency, and κ is the shape parameter in model development, derived from the GLM regression outputs. If the values of SD and Pearson χ^2 of one model are smaller than the standard χ^2 value at a desired level of confidence (e.g. 95%) with *(n-p-1)* degrees of freedom, the model is considered to fit the data well.

Previous studies suggest that the SD measure performs better in a large sample than a small one and does not work well when there are many extreme observations (e.g. zeros) in a population (Maycock and Hall, 1984; Maher & Summersgill, 1996; Wood, 2002; Agrawal & Lord, 2006). Wood (2002) proposed the grouped scaled deviance measure as one goodness-of-fit test for collision samples with too many extreme observations. In order to improve the normality of extreme observations, all data points in a sample are grouped. All the groups are used to create a new sample. The group scaled deviance (GSD) is formulated as:

$$G.SD = 2\sum_{i=1}^{n} r_i\left[\bar{y}_i \ln\left(\frac{\bar{y}_i}{\hat{E}(\Lambda_i)}\right) - (\bar{y}_i + \hat{\kappa})\ln\left(\frac{\bar{y}_i + \hat{\kappa}}{\hat{E}(\Lambda_i) + \hat{\kappa}}\right)\right] \tag{2.18}$$

where $\bar{y}_i$ is the average of observed collisions in each group; $\hat{E}(\Lambda_i)$ is the predicted collision based on the average values of independent variables in each group; and $\hat{\kappa}$ is the over-dispersion parameter for each group based on new observations. When the GSD value in a model is less than the corresponding new critical χ^2 values, the model is considered to be a good fit to the data.

Hadayeghi et al. (2003) used another measure for the goodness-of-fit test of NB models. This measure was proposed by Miaou (1996) and is expressed as

$$R_{\kappa}^{\ 2} = 1 - \frac{\kappa}{\kappa_{\max}} \tag{2.19}$$

where κ is the estimated shape parameter of the model, and κ_{max} is the estimated shape parameter for the most fundamental model. The shape parameter κ is used to determine how

well the data variance is explained in a relative sense. The higher the value of R_κ^2 is, the better the model fits.

If a model does not fit the data according to goodness of fit tests, an outlier analysis is undertaken to identify outliers, which will be removed from the data (McCullagh and Nelder, 1989; Sawalha and Sayed, 2001; Lovegrove, 2007). An outlier is defined as a data point that is numerically distant from other data points in the sample in which they occur (Barnett & Lewis, 1994). The presence of outliers indicates faulty data, erroneous procedures, or some area where a certain theory may not be valid. In a large sample, a small number of outliers is to be expected. Leverage and Cook's distance (CD) are two measures used to identify outliers. The CD measure was suggested by Sayed and Rodriguez (1999) because the leverage of a point can indicate how far this point is from the centroid of other points in the data, but does not reflect the influence of this point on model parameters. In a CD measure, the point with the largest CD value is removed from the data.

In addition to traditional negative binomial (TNB) regression, modified negative binomial (MNB) regression was researched in CPM development. The shape parameter κ is fixed for all individual points in TNB regression, but is different for individual points in MNB regression. Miaou and Lord (2003) suggested that treating κ as a fixed parameter might undermine the goodness-of-fit estimate of individual sites up to 30%, based on intersection data from Toronto, Ontario. They also suggested that similar research should be conducted using data in other regions to determine whether a varying dispersion parameter was a common or isolated situation. Heydecker and Wu (2001) compared the performances of TNB and MNB models based on a set of 3-way intersection data, and found that MNB models fit the data better than TNB models. Miranda-Moreno et al. (2005) compared TNB, MNB and Poisson-lognormal models based on highway-rail grade crossings data, and concluded that MNB and Poisson-lognormal models fit the data better than TNB models. They used these three types of models to identify and rank collision prone locations. The ranking results showed that TNB models agreed more with MNB models and less with Poisson-lognormal models. El-Basyouny and Sayed (2006) used the data from 58 arterials in Vancouver and Richmond, British Columbia, to compare TNB and MNB models. The goodness of fit results in this study showed that both methods fit the data well and MNB models had a better performance than TNB models. However, El-Basyouny and Sayed

(2006) found no big differences between their model application results (i.e. identification and ranking of collision prone locations.)

2.3.2 Zero-inflated Count Regression

Zero-inflated count (ZIC) regression can be seen as an extension of GLM. ZIC models include zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models, which are based on Poisson and NB models, respectively. ZIP and ZINB models are used in collision samples with excess zero counts (Shankar et al., 2003).

ZIP and ZINB models assume a dual-state process. The first process only generates zero counts. The second process generates non-zero counts. Each of them is assumed to follow a Poisson or NB distribution. For the occurrence of collision events at location *i*, if its probability of being a zero is set as φ_i in the first process, the probability of being a non-zero count should be 1-φ_i. The probability of the occurrence of collision events is written as

$$P(Y_i = y_i \mid X_i) = \begin{cases} \varphi + (1-\varphi) g(0 \mid X_i), if, y_i = 0 \\ (1-\varphi) g(y_i \mid X_i), if, y_i > 0 \end{cases} \tag{2.20}$$

where $g(y_i|X_i)$ is the probability mass function of a Poisson or NB distribution. The probability φ_i is usually specified as a logistic function, including a series of variables that are surrogates of the characteristics of observation *i*. This logistic function is written as:

$$\varphi_i = P(y_i \mid Z_i) = \frac{\exp(Z_i'\theta)}{1+\exp(Z_i'\theta)} \tag{2.21}$$

where Z_i is the vector of zero-inflated covariates and θ is the vector of zero-inflated parameters to be estimated. Its link formula is written as

$$\log it(\varphi_i) = Z_i'\theta \tag{2.22}$$

The model form in the second non zero-count modeling process is the same as the form of Poisson or NB models, which are mentioned in *Section 2.3.1*. Finally, the mean and variance of a ZIP model are:

$$E(y_i \mid X_i, Z_i) = \lambda_i (1-\varphi_i) \tag{2.23-a}$$

$$Var(y_i \mid X_i, Z_i) = \lambda_i (1-\varphi_i)(1+\lambda_i \varphi_i) \tag{2.23-b}$$

The mean and variance of a ZINB model are:

$$E(y_i \mid X_i, Z_i) = \lambda_i (1-\varphi_i) \tag{2.24-a}$$

$$Var(y_i \mid X_i, Z_i) = \lambda_i(1-\varphi_i)[1+\lambda_i(\varphi_i+\frac{1}{\kappa})] \quad (2.24\text{-b})$$

where κ is the shape parameter in the ZINB model. Both zero-inflated models could interpret the over-dispersion feature (i.e. $Var(y_i)>E(y_i)$). The maximum likelihood estimation method or the Quasi-Newton method can be used for parameter estimation.

The Vuong test is recommended for comparisons of ZIC models to Poisson or NB models (Shankar et al., 2003). The Vuong test is a likelihood-ratio-based test for model selection. It makes probabilistic statements on two models, and tests which one is closer to the actual model (Voung, 1989). The Vuong test is formulated as

$$V = \frac{\bar{m}\sqrt{N}}{S_m} \quad (2.25)$$

m is the mean of $ln[(f_1/f_2]$, where f_1 is the density function of a ZIP or ZINB distribution and f_2 is the density function of its parent-Poisson or NB distribution, S_m is the standard deviation of m, and N *is* the sample size.

2.3.3 Geographically Weighed Regression

Unlike GLM, geographically weighted regression (GWR) is able to account for spatial variation. GWR was proposed by Fotheringham et al. (2002) and is a local regression method to estimate specific regression parameters for individual sites. To clearly describe GWR, a global classical regression model is presented in a matrix form:

$$Y = X\beta \quad (2.26)$$

where Y is the dependent variable vector including n data points (dimensions of $n\times1$), X is the independent variable matrix with n data points and k explanatory variables (dimensions of $n\times(k+1)$), and β is the vector of parameters (dimensions of $(k+1)\times1$), which is estimated by

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (2.27)$$

However, the corresponding GWR model in a matrix form is written as

$$Y = (X \otimes \beta) \quad (2.28)$$

where $\otimes$ is a logical multiplication operator, making each element of β multiplied by the corresponding element of X. Unlike *Eq. 2.26* in which β is constant for all points, β in

Eq.2.28 is a matrix with different parameters for individual sites. The matrix β consists of n sets of local parameters and is given as:

$$\beta = \begin{bmatrix} \beta_0(u_1,v_1) & \beta_1(u_1,v_1)\cdots\beta_k(u_1,v_1) \\ \beta_0(u_2,v_2) & \beta_1(u_2,v_2)\ldots\beta_k(u_2,v_2) \\ \cdots \quad \cdots & \cdots \quad \cdots \\ \beta_0(u_n,v_n) & \beta_1(u_n,v_n)\ldots\beta_k(u_n,v_n) \end{bmatrix} \tag{2.29}$$

In this matrix, the parameters in each row (i.e. each site) are estimated by

$$\hat{\beta}(i) = \left(X^T W(i) X\right)^{-1} X^T W(i) Y \tag{2.30}$$

As shown in this equation, the regression parameter estimates for a particular site is directly influenced by its spatial weight matrix *W(i)*. *W(i)* is a $n \times n$ matrix, presented as

$$W(i) = \begin{bmatrix} w_{i1} & 0\ldots\ldots & 0 \\ 0 & w_{i2}\ldots\ldots & 0 \\ 0 & 0\ldots\ldots & w_{in} \end{bmatrix} \tag{2.31}$$

where w_{ij} is the weight given to the data point j in the calibration of the model for a particular site i. Fotheringham et al. (2002) suggested two functions to decide the spatial weight for each site. One is Gaussian function, formulated as

$$w_{ij} = e^{-\left(\frac{d_{ij}}{b}\right)^2} \tag{2.32}$$

The other is bi-square function, presented as

$$w_{ij} = \left[1-\left(\frac{d_{ij}}{b}\right)^2\right]^2 \tag{2.33}$$

In these two functions, b is a bandwidth for the regression point i and d_{ij} is the distance between any data point j from the regression point (see Figure 2.1). In GWR, a bandwidth is the distance between a regression point and its furthest data point(s). It can be seen as a parameter to identify how large the spatial context is.

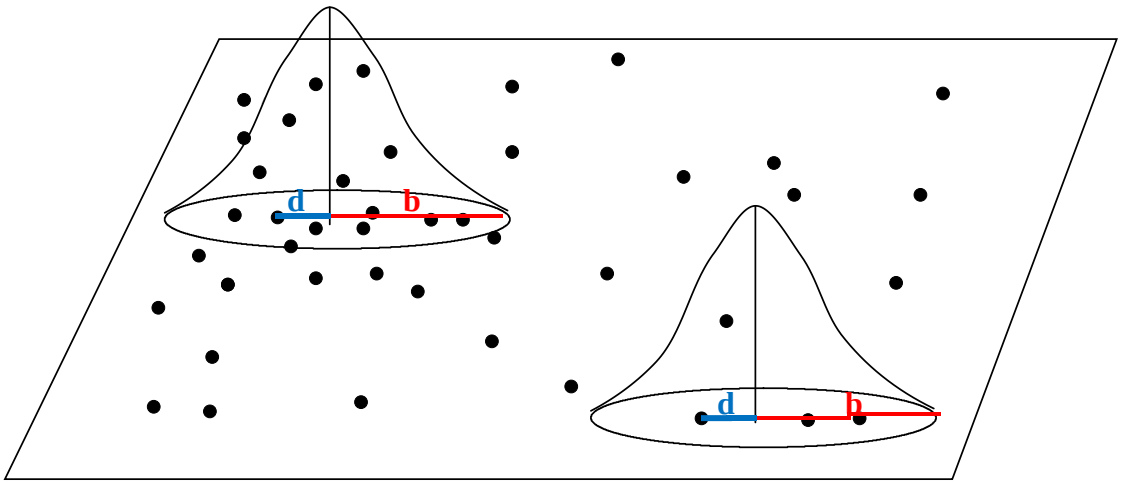

Figure 2.1 GWR Spatial Kernel & Weighting Distribution

The lengths of bandwidths are critical for GWR model development (Fotheringham et al., 2002; Guo et al., 2008). An 'over-smooth' bandwidth generates similar regression parameters for all data points across a study area, while an 'under-smooth' bandwidth generates overly varied parameters such that a spatial pattern in data is hard to identify. Cross-validation (CV) and Akaike information criterion (AIC) were suggested by Fortheringham et al. (2002) to decide the length of a bandwidth. The formula of CV is presented as:

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}_{i \neq i})^2 \tag{2.34}$$

In this formula, y_i is the observed value for location i and $\hat{y}_{i \neq 1}$ is the estimated value without considering the location i. The bandwidth making a minimum CV value is selected. The CV method can only be used when data points are set as regression points. Akaike information criterion (AIC) was proposed by Hirotugu Akaike (1974). It is a measure used for model selection. The formula of AIC is presented as:

$$AIC = \sum_{i=1}^{n} \{2p_i + n[\ln\left(\frac{2\pi \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}\right) + 1]\} \tag{2.35}$$

where y_i is the observed value for location i, $\hat{y}_i$ is the estimated value, and p_i is the number of parameters for each location i. Similarly, the bandwidth making a minimum AIC value is selected.

Finally, the logarithmic function of a GWR model is written as:

$$\ln(E(\Lambda_i)) = \ln(a_0(u_i)) + a_1(u_i)\ln(Z) + \sum b_j(u_i)X_j \quad (2.36)$$

where $a(u_i)$ and $b(u_i)$ are seen as parameter functions, derived from *Eq. 2.29*. In GWR models, the Monte Carlo test is used to examine the significance of spatial variability in local parameter estimates (Fortheringham et al. 2002). The *p*-value is derived from the Monte Carlo test to indicate the probability that the variation occurs by chance. The lower the *p*-value is, the more the spatial variation exists (Charlton et al., 2011). Although GWR models are able to account for the spatial variation issue, a large effective number of parameters in GWR models may lead to a situation where expected values overfit observed values. This over-fitting problem makes GWR models "impractical" in model applications (Fortheringham et al. 2002). In order to avoid the over-fitting problem, the selection of bandwidth lengths is critical. If bandwidths are extremely large, the effective number of parameters for a GWR model would be the same as for their corresponding global models; if bandwidths are asymptotically zero, the effective number of parameters for the GWR model would approach the number of observations in the data so that predicted values would fit observed values perfectly (Fortheringham et al., 2002; Guo et al., 2008). Additionally, Hadayeghi et al. (2009) suggested that GWR models could not be spatially transferred.

2.3.4 Full Bayesian Method

The full Bayesian (FB) method is recognized as a regression method to account for spatial/temporal variation and heterogeneity in collision data (Miaou & Lord, 2003; Aguero-Valverde & Jovanis, 2005; Hadayeghi et al., 2010; El-Basyouny & Sayed, 2009). A FB model consists of a set of global parameter estimates for regression variables and separate coefficients representing error terms. Similar to GLM models, the occurrence of collision events at location *i* in FB models is assumed to follow a Poisson distribution with the parameter λ_i:

$$Y_i|\ \lambda_i \sim Poisson\ (\lambda_i\ ) \quad (2.37)$$

where λ_i is the mean or expected value for collisions in zone *i*. In order to address the extra-variation of collisions, λ_i can be assumed to follow either a gamma distribution or a lognormal distribution:

$$\text{Poisson gamma or NB models: } \lambda_i \sim Gamma\left(\kappa, \frac{\kappa}{\lambda_i}\right) \tag{2.38}$$

$$\text{Poisson lognormal models: } Ln(\lambda_i) \sim N\left(Ln(\lambda_i), \sigma^2\right) \tag{2.39}$$

where κ is the over-dispersion or shape parameter in NB models and σ_α^2 represents the within-site extra variation in Poisson lognormal (PLN) models. If the expected collisions λ_i follow a gamma distribution, the mean and variance of y_i are

$$E\left(y_i\right) = \lambda_i \tag{2.40-a}$$

$$Var\left(y_i\right) = E\left(y_i\right) + \frac{E\left(y_i\right)^2}{\kappa} = \lambda_i + \frac{\lambda_i^2}{\kappa} \tag{2.40-b}$$

If the expected collisions λ_i follow a lognormal distribution, the mean and variance of y_i are

$$E(y_i) = e^{(\ln\lambda_i + 0.5\sigma^2)} = \lambda_i e^{0.5\sigma^2} \tag{2.41-a}$$

$$Var(y_i) = E(y_i) + (e^{\sigma^2} - 1)E(y_i) = \lambda_i e^{0.5\sigma^2} + \lambda_i e^{0.5\sigma^2}(e^{\sigma^2} - 1) \tag{2.41-b}$$

The expected value of λ_i is derived from any FB model.

In one full Bayesian NB or PLN model, the prior distributions of all unknown parameters should be determined before incorporating observed information. Normal distributions, $N\ (0,\ \tau^2)$, are usually set as the prior distributions of regression parameters in the model, where τ^2 can be a large variance (e.g. 1000). Gamma distributions, $Gamma\ (\varepsilon,\ \varepsilon)$ or $Gamma\ (1,\ \varepsilon)$, are commonly set as the prior distributions of σ^2 and κ, where ε is a small number (e.g. 0.001 or 0.01). After being combined with observed data, prior distributions of all parameters are updated to posterior distributions. All statistic inferences such as means, standard deviations, and credible intervals for all parameters and expected collisions of all sites are available according to these posterior distributions.

Unlike GLM models in which a classical likelihood-based statistical inference framework (e.g. maximum likelihood method) is used to estimate parameters, the Markov chain Monte Carlo (MCMC) method is used in FB models to obtain posterior distributions of all parameters. In the MCMC method, random numbers can be drawn from numerically intractable posterior distributions and posterior statistical inferences are obtained from empirical analogue. Metropolis-Hastings (MH) and Gibbs sampling are commonly used MCMC methods to generate samples.

The deviance information criterion (DIC), a generalization of the Akaike information criterion and the Bayesian inference system, is usually recommended for Bayesian model comparison and selection. Superior to other criteria, the DIC is easily calculated from the samples generated by a MCMC simulation (Spiegelhalter et al., 2002). The DIC is formulated as

$$DIC = p_D + \bar{D} = 2\bar{D} - D(\bar{\theta}) \tag{2.42}$$

where θ are the unknown parameters in a model, $D(\theta)$ is the deviance, $\bar{D}$ is the mean of $D(\theta)$, $D(\bar{\theta})$ is the deviance if θ are average values, and p_D is the effective number of parameters. Generally, the lower the DIC of a model, the better the model fits the data. However, it is difficult to determine what would constitute a significant difference in DIC. Spiegelhalter et al. (2005) suggested that a difference of more than 10 between two models definitely rules out the model with a higher DIC, a difference between 5 and 10 is substantial, and a difference of less than 5 makes very different inferences. If the difference is less than 5 between two models, it would be misleading to report the model with a lower DIC.

2.3.5 Comparisons between Regression Methods

The regression methods mentioned above are compared in many studies. Qin et al. (2004), Kumara and Chin (2003), and Lee and Mannering (2002) showed that ZIP regression was more promising than Poisson regression for providing explanatory insights into the causality behind collisions with an excess zero mass. After comparing traditional NB models, GWR models, and full Bayesian NB models, Hadayeghi et al. (2009; 2010) found that the goodness-of-fit test performances of GWR and full Bayesian NB models were much better than traditional NB models. Also, the spatial covariates in GWR and FB models indicate a significant spatial correlation between zones. Li et al., (2008), Miaou and Lord, (2003), Persaud et al. 2010, and El-Basyouny and Sayed (2010) compared the model development and/or application results using the FB and GLM methods and suggested three advantages of the FB method over the GLM method: 1) the FB method has the ability to account for all uncertainties including spatial variation, temporal influences and interaction between covariates; 2) the FB method has more flexibility in selecting collision distributions (i.e. both Poisson gamma and Poisson lognormal distributions can be used in the FB method); and 3)

the FB method offers an integrated procedure to obtain outcomes in CPM applications rather than the GLM method that requires separate steps.

Three goodness-of-fit tests, mean absolute deviation **(**MAD), mean squared prediction error (MPSE), and mean squared error (MSE) were suggested by Hadayeghi et al. (2009; 2010) to compare CPMs using different regression methods. All of these measures indicate the degree of correctness of a model. A measure value of a model close to zero suggests that the model fits the data well. MAD, MPSE, and MSE are formulated as

$$MAD = \frac{\sum_{i=1}^{n} | \hat{Y}_i - Y_i |}{n} \tag{2.43}$$

$$MSPE = \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n} \tag{2.44}$$

$$MSE = \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n - p} \tag{2.45}$$

Additionally, Pearson's product moment correlation coefficient was suggested by Oh et al. (2003) as a way to measure how modeled collision values fit observed collision values. This measure can be used for comparing different regression models. The coefficient equal to 1 in a model indicates that the model fits the data perfectly. Pearson's product moment correlation coefficient is given as

$$r = \frac{\sum (Y_i - Y_{avg})(Y_i - Y_{avg})}{\left[\sum (Y_i - Y_{avg})^2 \sum (Y_i - Y_{avg})^2 \right]^{0.5}} \tag{2.46}$$

2.4 Development and Applications of Macro-level CPMs

This section includes several important parts relating to macro-level CPM development and applications. The first part describes macro-level CPM variable selection and model stratification. The second part summarizes two macro-reactive safety applications: black spot studies and before and after evaluation studies. The third part reviews macro-proactive road safety applications.

2.4.1 Variable Selection and Model Stratification

Dependent variables in traditional macro-level CPMs can be different patterns of zonal collision frequencies over a period, such as total/severe/AM zonal vehicle collision frequencies (Sayed & Lovegrove, 2006; Hadayeghi, 2009). Independent or explanatory variables in macro-level CPMs can be variables representing zonal traffic exposures and other neighbourhood traits. Lovegrove (2007) suggested several criteria to screen independent variables, including: 1) variable data should be sufficient but not expensive to collect, 2) variable data should be collected and extracted accurately, 3) variable definitions should be reasonable and easy to understand, 4) variable data should be predictable for use, which means both current and future values of these variables can be obtainable, and 5) all variables should be relevant and practical for use as road safety descriptors. Following these criteria, Lovegrove (2007) identified over fifty candidate independent variables. These variables are divided into four themes, including exposure, social-demographics (S-D), transportation demand management (TDM), and road network (NW).

Exposure variables provide a surrogate for road traffic exposure, such as vehicle kilometres travelled (VKT), total lane kilometres (TLKM), and zonal area. Most macro-level studies use VKT or TLKM as a leading variable (Sayed & Lovegrove, 2006; Lovegrove, 2007; Hadayeghi, 2009). A leading variable has the strongest statistic associations with the dependent variable in a model. Lovegrove (2007) suggested that VKT was superior to TLKM because the model results with VKT were more stable than those with TLKM. S-D variables are used to represent demographic and economic characteristics in communities. In Lovegrove (2007), candidate S-D variables include population, population density, participation in the labour force, employed residents, employed percentage, employed density, unemployed residents, unemployed rate, average income, home number, and home density, all of which could be from census data.

TDM variables relate to efficient use of road systems, including total commuters, commuter density, core area (CORE), core area percentage, bicycling commuter percentage, driving commuter percentage, car passenger commuter percentage, walking commuter percentage, transit commuter percentage, number of driving commuters, zonal bus stop number, and bus stop density. CORE was defined by van Minnen (1999) as the largest area of one zone not bisected by major roads. NW variables represent road network environment

variables, including arterial lane kilometres, local lane kilometres, connector lane kilometres, proportions of arterial/local/connector lane kilometers out of total lane kilometres, number of intersections, intersection density, number of signals, signal density, percentage of 3-way intersection, percentage of arterial-local road intersections, and the ratio of intersection number to total lane kilometres.

To improve model performances, macro-level CPMs are usually stratified. The stratification can be performed in both independent and dependent variables. Dependent collision variables can be stratified according to collision patterns, such as collision types (e.g. overall, bike, pedestrian, and vehicle collisions), collision severity (e.g. fatality collisions, extremely severe collisions, relatively severe collisions, and lightly severe collisions), or collision locations (e.g. mid-blocks or intersections). Independent variables can be stratified in terms of variable themes (i.e. exposure, S-D, TDM, and NW), land use (i.e. rural or urban), and data derivation (i.e. modeled or measured) (Lovegrove, 2007). Lovegrove (2007) suggested sixteen different groups derived from the three independent data stratification levels, as shown in Table 2.1.

Table 2.1 Model Groups (Lovegrove & Sayed, 2006)

Themes	Land Use	Data Derivation	Group #
Exposure	Urban	Modeled	1
		Measured	2
	Rural	Modeled	3
		Measured	4
Socio-Demographic	Urban	Modeled	5
		Measured	6
	Rural	Modeled	7
		Measured	8
Transportation Demand Management	Urban	Modeled	9
		Measured	10
	Rural	Modeled	11
		Measured	12
Network	Urban	Modeled	13
		Measured	14
	Rural	Modeled	15
		Measured	16

2.4.2 Macro-reactive Applications

Black spot studies, and before and after studies, are two reactive applications of community-based, macro-level CPMs. In macro-level black spot studies, black spots are collision prone zones (CPZs) with significantly above-average (95% level) collision frequency. Black spot studies are used to identify and rank CPZs so that these hazardous zones can be diagnosed and remedied. Before and after studies are used to evaluate the effectiveness of road safety improvement countermeasures. These two studies are usually used together in reactive road safety improvement programs.

2.4.2.1 Black Spot Studies

One problem in ensuring that truly hazardous locations/zones are identified as black spots relates to the regression-to-the-mean bias (Hauer, et al., 1988; Sayed & de Leaur, 2001). Regression-to-the-mean (RTM) is a statistical phenomenon whereby extreme values of a random variable tend to be followed by less extreme values, even though no change has occurred in the underlying causal mechanism (Hauer, et al., 1988; Sayed 1998). In road safety terms, RTM occurs when the observed collision frequency or rate in one location regresses to its long-term mean values of collision frequency or rate as time goes by. Given that black spots are usually identified because of a recorded high occurrence of collisions, the RTM bias may lead to a "false" labelling of a site as hazardous. Therefore, use of the empirical Bayesian (EB) method is suggested in order to reduce the RTM bias (Hauer, et al., 1988, 2002; Higle & Wikowski, 1988; Sayed 1998). The empirical Bayesian (EB) method is expressed as a combination of observed collision potentials (e.g. collision rates or frequencies) and estimated collision potentials derived from the reference group.

Higle and Wikowski (1988) first used collision rates with the EB method to identify black spots. However, there are two problems in Higle and Wikowski's study. First, measuring road safety by using the collision rate of one location to identify black spots can yield misleading results (Hauer, 1995). Second, using the EB method with collision rates makes it difficult to find a suitable reference group for accurate estimates of prior means and variances. Therefore, Hauer (1992) used CPMs with the EB method to identify black spots, as 1) the road safety measure used to identify black spots is collision frequency instead of

collision rate, and 2) CPMs do not require a large reference group to generate prior means and variances.

In the EB method using CPMs to identify black spots, the first step is to get collision priors from CPMs. If the negative binomial (NB) method is used to develop macro-level CPMs, the prior distribution of predicted collisions at each site would be assumed as a gamma distribution with the shape parameter κ and scale parameter $\kappa/E(\Lambda_i)$. The second step is to combine prior collision estimates from NB models with local collision data to get posterior collision estimates. The posterior distribution of collisions at each site is also assumed to follow a gamma distribution with shape and scale parameters, as shown in *Eq. 2.47* (Hauer et al., 1988, 2002).

$$\alpha = \kappa + count \tag{2.47-a}$$

$$\beta = \frac{\kappa}{E(\Lambda_i)} + 1 \tag{2.47-b}$$

The mean and variance of posterior or EB collisions for a site are respectively presented as:

$$EB_i = E(\Lambda \mid Y = count) = \frac{\alpha}{\beta} = \left[\frac{E(\Lambda_i)}{\kappa + E(\Lambda_i)}\right](\kappa + count) \tag{2.48-a}$$

$$Var(EB_i) = Var(\Lambda \mid Y = count) = \frac{\alpha}{\beta^2} = \left[\frac{E(\Lambda_i)}{\kappa + E(\Lambda_i)}\right]^2 (\kappa + count) \tag{2.48-b}$$

The final step is to compare prior collision estimates derived from NB models (i.e. $E(\Lambda_i)$) with their corresponding posterior collision estimates (i.e. EB_i). In macro-level black spot studies, if the EB value at one zone exceeds its $E(\Lambda)$ at a significant confidence level (i.e. δ=0.95), this zone would be considered to be a collision prone zone (CPZ). This process is mathematically presented as

$$1 - \int_0^{E(\Lambda)} f_{EB}(\lambda)d\lambda = \left[1 - \int_0^{E(\Lambda)} \frac{[\kappa / E(\Lambda) + 1]^{(\kappa + count)} \lambda^{(\kappa + count - 1)} e^{-(\kappa/E(\Lambda)+1)\lambda}}{\Gamma(\kappa + count)} d\lambda\right] \geq \delta \tag{2.49}$$

Figure 2.2 illustrates how to identify CPZs.

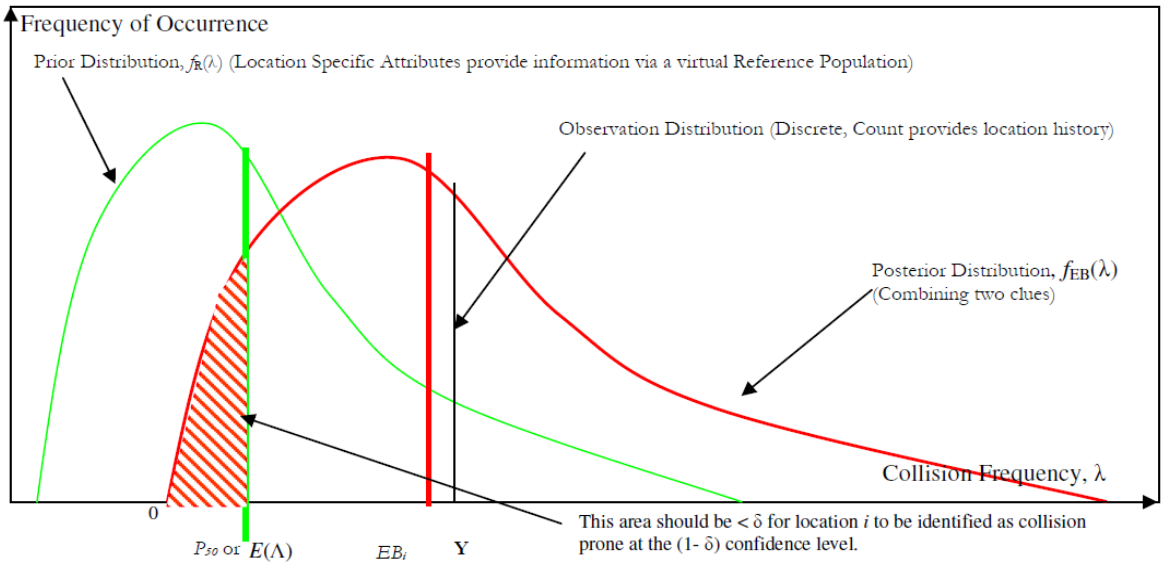

Figure 2.2 Empirical Bayes Identification of CPZs (Lovegrove, 2007)

Potential collision reduction (PCR) and collision risk ratio (CRR) were proposed by Sayed (1998) to rank micro-level black spots (i.e. intersections or road segments). These two ranking measures were also used by Lovegrove (2007) to rank macro-level black spots (i.e. CPZs). They are respectively formulated as

$$PCR = EB - E(\Lambda) \tag{2.50}$$

$$CRR = \frac{EB}{E(\Lambda)} \tag{2.51}$$

Since different groups of macro-level CPMs developed, the identification and rank results from different CPMs may not agree with each other. Therefore, an additional step is required to sum ranks across multiple CPMs to get a total ranking score for each zone. After identifying and ranking CPZs, related diagnoses and remedies for these CPZs should be implemented. The guideline for CPZ diagnoses and remedies is available in Lovegrove (2007).

As of now, no research has been found to use geographically weighted regression (GWR) models to identify black spots and few studies have been conducted to use full Bayesian (FB) models to identify and rank micro-level black spots. Using the FB method (or FB models) to identify and rank micro-level black spots is simpler than using the EB method because posterior collisions can be directly obtained from FB models. Huang et al. (2009) suggested two measures of black spot identification. One measure is to use the mean value of posterior

collision frequency at one site to compare a critical collision frequency. If the mean value of posterior collisions is greater than the critical collision frequency, the site is set as a black spot. However, the determination of critical collision frequencies was not discussed in Huang et al. (2009). The second measure is to set a critical cut-off percentage in a data sample (e.g. 1%, 2.5%, 5%, or 10%). All locations in this data sample are ranked from high to low according to their posterior collision estimates. Then the critical cut-off percentages are used to determine which of the top ranked locations should be identified as black spots. For black spot ranking, Lan et al. (2011) suggested eight criteria. These criteria are listed as follows.

1) Posterior collision mean (λ_i): this is the mean value of posterior collisions for each site.
2) Potential safety improvement (PSI): this criterion is similar to the potential collision reduction in the EB method, which is formulated as

$$PSI = \lambda_i - u_i \tag{2.52}$$

where λ_i represents the posterior collision mean for a specific site and μ_i represents the normal expected value at similar sites.

3) Pseudopotential safety improvement (PPSI): this is similar to PSI but λ_i is replaced by the observed collision y_i. It is formulated as

$$PPSI = y_i - u_i \tag{2.53}$$

4) Posterior expected rank of collision mean: it is based on the rank of λ_i. This rank can be obtained in the software for FB model development directly.
5) Posterior median rank of collision mean.
6) Posterior mode rank of collision mean.
7) Observed collision counts.
8) Probability that one site is the worst of all other considered sites for the posterior collision mean (P_{worst}): it represents the posterior probability that the site *i* has the largest predicted collision estimation of all sites. In other words, it is the probability that the predicted collisions in site *i* is larger than the expected collisions for the remaining sites. It is formulated as:

$$P_{worst} = p(\lambda_i > \lambda, \text{ for all } j \neq i | y) \tag{2.54}$$

In micro-level black spot studies using the FB method, the identification and ranking measures mentioned above are practical because the concerned sites are a batch of

intersections or road segments with similar geometric features and/or traffic volume. However, in macro-level black spot studies, different zones in a study area often have diverse neighbourhood traits (e.g. various traffic exposures, area, and road network traits). Therefore, the identification and ranking measures listed above are not appropriate for collision prone zone (CPZ) identification and ranking. To date, no previous research has been found to use the FB method to identify macro-level black spots. This study fills this gap by using the FB method to identify macro-level black spots, as shown in *Section 4.1.2*.

2.4.2.2 Before and After Evaluation Studies

Road safety improvement countermeasures are usually implemented after CPZs are identified. Before and after studies are conducted to evaluate the countermeasure effectiveness, usually indicated by the percentage of reduced collisions. This percentage ratio is the collision reduction factor (CRF). Its subtraction from the unity (i.e. 1- CRF) is referred to as the collision modification factor (CMF) or odds ratio (OR). The method used to estimate the OR of any given countermeasure is based on the approach described in Hauer (1997), Sayed (1998), and Sayed and de Leur (2001), including the use of CPMs and the EB method. The OR of a certain collision pattern is estimated as:

$$OR = \frac{A/C}{B/D}, \text{ with treatment effect} = CRF = OR\text{-}1 \tag{2.55}$$

where *A/C* is collision frequencies that occurred before/after the countermeasure implementation in the comparison group, *B* is the EB collision frequency that would have occurred in the *after* period in the subject site without treatment, and *D* is collision frequencies that occurred after countermeasure implementation in the subject site. The role of the comparison group is to reduce the bias of the time trend from the before to the after period. The comparison group data should be drawn from a randomly selected sample of sites. All of the quantities in *Eq. 2.55* are observed collision frequencies except for the quantity *B*, which is estimated as:

$$B = EB_a = EB_b \frac{E(\Lambda_a)}{E(\Lambda_b)} \tag{2.57}$$

where EB_a is the EB safety estimate of the treatment group in the *after* period without treatment, EB_b is the EB safety estimate of the subjective site in the *before* period, $E(\Lambda_a)$ is

the collision frequency of the subjective site estimated by the CPM using its traffic exposure in the *after* period, and $E(\Lambda_b)$ is the collision frequency estimated by the CPM using its traffic exposure in the *before* period.

Based on the method reviewed above, Lovegrove (2007) used a similar approach for the macro-level before and after evaluation studies. After evaluating traffic calming strategies in three neighbourhoods in GVRD, Lovegrove (2007) concluded that the average ORs of total and severe collisions were 0.60 and 0.61, respectively, which means the reduction in total and severe collisions in these three neighbourhoods were 40% and 39% due to traffic calming strategies. This result supports the original ICBC research results reported by Geddes et al. (1996). So far, GLM models have been used in micro- and macro-level before and after studies. Although FB models have been used in micro-level and after road safety evaluations (El-Basyouny & Sayed, 2010; Li et al., 2008; Persaud et al. 2010), they have not been used in macro-level studies. Therefore, the methodology about macro-level before and after studies using the FB method is recommended for future research.

2.4.3 Macro-proactive Applications

Community-based, macro-level collision predilection models (CPMs) have been shown to have the capability of proactively evaluating the road safety of any planned transportation program (Lovegrove & Sayed, 2006b; Lovegrove, 2007; Hadayeghi et al., 2003, 2009, 2010). Lovegrove (2007) proposed proactive CPM use guidelines for regional and neighbourhood road safety planning with three steps. The first step is to choose zones influenced by any regional or neighbourhood transportation program. The second step is to assemble and process variable data in selected CPMs for each involved zone. The variable data are different between the scenarios with and without the transportation programs. The last step is to run the CPMs to estimate collisions in each scenario based on variable data.

Lovegrove et al. (2010) used community-based, macro-level CPMs from GVRD to evaluate the road safety effects of TransLink's 2005 to 2007 three-year regional plan. The influence area of this plan covers 400 zones in GVRD. The model results based on two scenarios show that the regional collision estimate from the scenario with this three-year transit plan is much lower than the regional collision estimate from the do-nothing scenario. Lovegrove and Sayed (2006b) used macro-level CPMs in neighbourhood road safety

planning to compare the safety effects of various community road network patterns. After a comparison of CPM estimates between four test scenarios of grid networks, cul-de-sac networks, 3-way offset networks, and modified Dutch SRS networks, Lovegrove and Sayed (2006b) suggested that 3-way offset road networks and modified Dutch SRS road networks were safer than grid networks and cul-de-sac networks. These proactive model application results suggest that macro-level CPMs could provide a solid step in developing new and improved empirical tools for road safety planning.

2.5 Factors Influencing Bicycle Use

Factors influencing bike use are reviewed in this section. These factors are considered to have indirect associations with road safety because bicycle use potentially influences road safety. These factors include demographic, economic, social environmental, engineering, and geographical features.

Demographic factors include gender, age, and experience of cyclists. Generally, males are more likely to bicycle than females because women tend to think bicycle riding has more risks than men (Williams & Larson, 1996; Garrard et al., 2008; Bernhoft & Carstensen, 2009). Young people are more likely to bicycle than older people because young people's age and/or economic status drive them to use bicycles more than adults (Williams & Larson, 1996; Garrard et al., 2008; Bernhoft & Carstensen, 2009). However, these two conclusions were derived from low bicycle use countries. In European countries with medium bicycle use, the differences are not apparent on the number of female and male cyclists, and the number of young and adult cyclists (Baker, 2009).

Usually, higher wealth is associated with lower bicycle use. Winters et al. (2007) suggested that the household income was negatively correlated with bicycle use. Moudon and Lee (2005) found that the household car number was negatively correlated with bicycle use. Rietveld and Daniel (2004) suggested that reducing bicycling costs and increasing private automobile mode costs could promote bicycle use. Pucher and Buehler (2006) found bicycling rates are higher in Canada than in America, partly because Canada has higher overall costs of owning and operating a car than America.

Social environmental factors include bicycle culture, policy, and so on. Traffic culture is affected by historic, economic and political characteristics. In developing countries,

bicycling is one popular transport mode for commuting due to less social wealth. In developed countries with a medium/high level of bicycle use, the generation of bicycle culture is attributed to education, enforcement, and engineering. In the Netherlands and Sweden, bicycling accounts for 35%-40% of all trips in some communities (Pucher & Dijkstra, 2003). Bicycling in these countries is not a fashion sport but a regular transport mode.

Engineering factors are the quality and quantity of bicycling infrastructure and facilities. The engineering factors to reduce bicycle use include too many stops for cyclists, lack of bike lanes and trails, lackof good lighting conditions at night, and inadequate bicycle racks at destinations (CROW, 1998; Dill & Carr, 2003; Stinson & Bhat, 2004; Moudon & Lee., 2005). Miller et al. (2010) suggested that decreasing vehicle lanes or increasing bicycle paths in an urban area might reduce total vehicle collisions and the severity of injuries to cyclists. Reynolds et al. (2009) concluded that purpose-built bicycle facilities could reduce bicycle collisions and injuries.

Geographical factors related to bicycle use include temperature, precipitation, wind, and terrain. In general, warmer and drier places are better for cycling than cooler and rainier places if other factors are the same. Also, flat terrain is better for cycling than hilly terrain, but it does not mean that cycling is impossible in hilly regions (e.g. Copenhagen).

2.6 Data Issues and Data Sharing

In addition to the methodology for CPM development, the data for model development is also important. The quality of data directly influences the accuracy of model results. Accurate and complete road traffic collision and injury data are pursued. "Accurate" means that the collision/injury number and severity should reflect or be close to the actual situation.

Based on a meta-analysis of official road collision reporting in 13 countries, Elvik and Mysen (1999) found that the reported injuries in official collision statistics were incomplete and the number of collisions between single-vehicles and bicycles was more likely to be under-reported. Hauer and Hakkert (1988), James (1991), and Hvoslef (1994) suggested that the reported injury severity was not accurate and the number of vehicles involved in each collision was mostly under-reported.

Different organizations develop individual road safety databases due to different responsibilities and purposes. Major road safety data can be from auto insurance companies, police departments, hospitals, and emergency services. Aptel et al. (1999) compared the hospital and police road injury data of Isle La Réunion, France, from 1993-1994, and found that only 37.3% of traffic-injured patients from hospital records were recorded in the police database. Cryer et al. (2001) checked British police and hospital road collision data, and found that hospital data were more accurate than police data. Reurings and Stipdonk (2010) reviewed the hospital inpatient registry database and the police collision database in the Netherlands from 1993-2008. They discovered that the hospital database was unable to indicate whether a patient was involved in a road collision and the police database was unable to indicate road injury severity. Through linking these two databases, Reurings and Stipdonk (2010) found that 85% of road traffic injuries were recorded in the hospital database, but only 58% of motor vehicle traffic injuries and 4% of non-motor vehicle traffic injuries were recorded in the police database.

Many provincial and national road injury surveillance programs in Canada are used for ongoing injury data collection, analysis, interpretation, and timely dissemination. In BC, road injury data are original saved by the Insurance Corporation of British Columbia (ICBC), hospitals, police, and BC Ambulance Services in different databases. In Alberta, the Office of Traffic Safety and the Alberta Centre for Injury Control and Research have individual databases to record road injury data (Office of Traffic Safety, 2007). In Quebec, the ambulance services and police have separate databases for transport collisions (Miranda-Moreno et al., 2011). At the national level, Transport Canada collects vehicle collision data from provincial and territorial agencies and then stores them in the Traffic Accident Information Database (TRAID). TRAID provides national figures of all reported vehicle collisions in Canada (Transport Canada, 2001), but only covers severe vehicle collisions. Another national injury surveillance program is the Canadian Hospitals Injury Reporting and Prevention Program (CHIRPP), which is funded by the Public Health Agency of Canada (PHAC). CHIRPP is an information system that has collected and analyzed the injury data of people seen at the emergency departments of 14 pilot hospitals in Canada since 1990 (Hayes et al., 2001). Thus, patients injured in road traffic collisions are also included in this program. However, like other databases, CHIRPP cannot include all transport injuries.

2.7 Summary

This chapter reviewed sustainable road safety improvement programs, basic statistics of macro-level CPMs, four regression methods in CPM development, the development and application methodologies of macro-level CPMs, and data issues and challenges in road safety analyses. It provides a theory foundation of developing and applying the community-based, macro-level CPMs related to bicycle use. The methodologies used in *Chapter 3* and *4* are derived from the literature and updated to some extent.

CHAPTER 3 DATA EXTRACTION AND MODEL DEVELOPMENT

This chapter has five sections that describe data extraction and macro-level CPM development. *Section 3.1* describes the data extraction process. *Section 3.2* discusses several model forms, including how to select the most appropriate model form(s). *Section 3.3* discusses the model grouping or stratification. *Section 3.4* individually presents the model development results using four regression methods and then compares them. *Section 3.5* provides a summary of this chapter.

3.1 Data Extraction

A valid data extraction (i.e. data collection and data aggregation) process increases the chance that statistical models correctly reflect underlying causal mechanisms. This section describes the data extraction process of community-based, macro-level CPM development. It includes two parts. The first part describes the geographic scope of the study area and the data aggregation process. The second part introduces variable data sources and provides statistical inferences based on these variable data.

3.1.1 Geographic Scope and Aggregation Units

The study area used for model development is the Regional District of the Central Okanagan (RDCO), in BC, Canada. This region spans approximately 44,000 square kilometres and is comprised of four member municipalities: Kelowna, West Kelowna, Lake Country, and Peachland. Based on 2006 Canada Census data, there were about 160,000 residents, 66,000 households, and 29,000 total lane kilometres (vehicle lanes) in RDCO (Statistic Canada, 2006). *Figure 3.1* shows a map of RDCO, including the four municipalities and their urban and rural land use.

All variable data used for macro-level CPM development were aggregated into community-level areal units. In this study, 500 Traffic Analysis Zones (TAZs) derived from the UBC 2010 VISSUM model (a transportation planning model) are set as aggregation units, as shown in Figure 3.1. TAZs allow planners to obtain long period traffic data for each TAZ to sure a practical neighbourhood focus planning. As this strategic transportation planning objective

coincides with the macro-level road safety planning objective, TAZs are recommended as aggregation units in model development. The boundaries of TAZs usually keep zonal population densities or employment densities at a roughly uniform level and partly overlap with census tract and/or dissemination area boundaries.

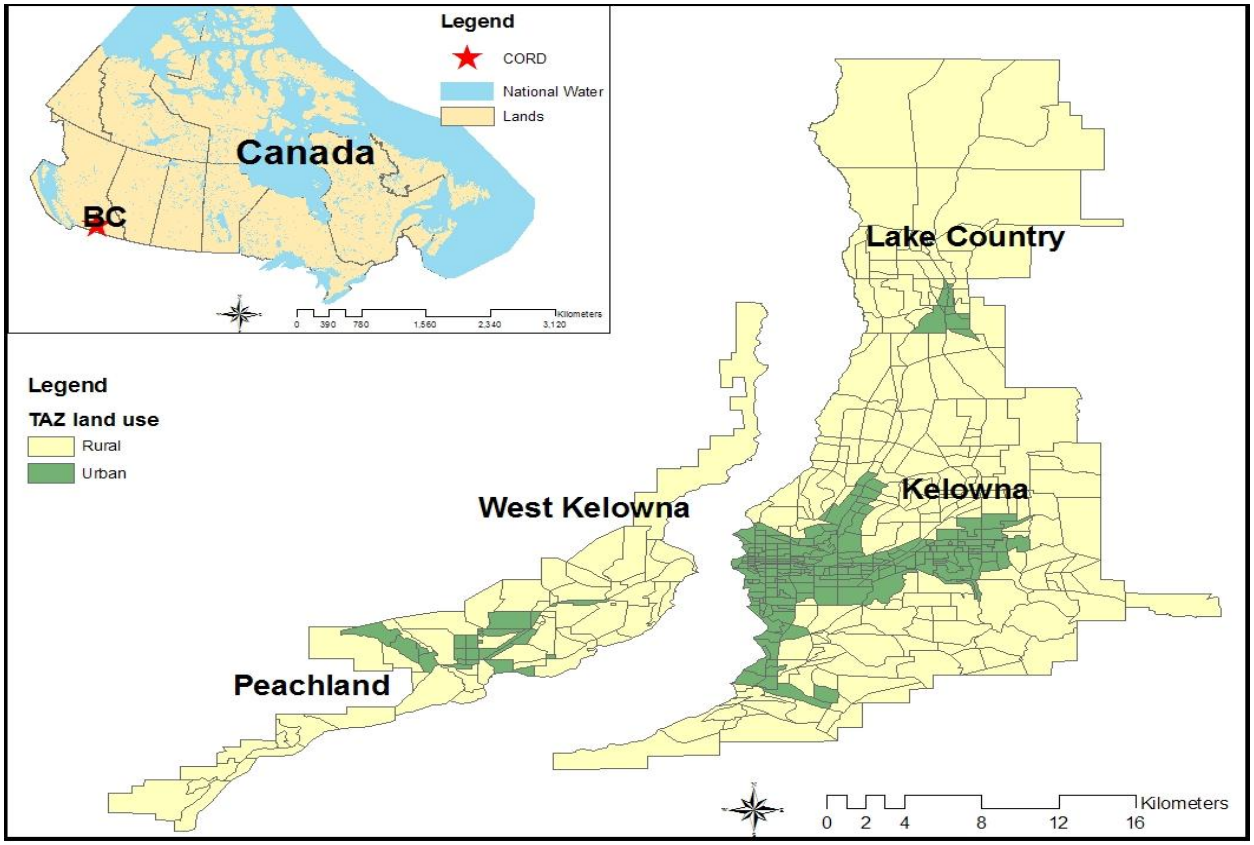

Figure 3.1 Study Area: the Regional District of Central Okanagan

ArcGIS 9.3 was used to aggregate data at the TAZ level. Most data such as traffic exposure and road network data can be directly obtained through geo-processing functions in ArcGIS. However, other data such as demographic data need more complex geo-processing. Demographic data (e.g. population, employment) from Canada Census are recorded in units of dissemination areas (DAs). In order to aggregate these data in TAZ units, two steps are required. First, the residential area in the study region should be identified to determine where the population is distributed. Second, the demographic data in each DA should be assigned into different TAZs that are geographically related to the associated DA, according to the residential area distribution in these TAZs. With the geo-processing aid of ArcGIS, demographic data in

DA units can be transformed to demographic data in TAZ units. All the data in this study were aggregated at least twice and reviewed to ensure an accurate data aggregation.

3.1.2 Variable Data Sources and Statistics

As macro-level CPMs use different patterns of zonal collisions as dependent variables, the information of collision incident locations is necessary to identify how many collisions are located in each zone. Of the two available road safety databases (i.e. the BCIRPU and ICBC databases), the BCIRPU database is rejected for use because it lacks injury incident locations. Finally, the ICBC database is selected for macro-level CPM development. In addition to providing geo-referenced locations for its all collision claims, the ICBC database covers all police- and self-reported collision claims in BC so that they are regarded as 'complete.' In this study, total, severe, and bike collisions are set as three collision patterns. Since only ICBC claims are used, total collisions are limited to fatality, injury, and property-damage-only collisions involving motor vehicles; severe collisions are limited to fatality and injury collisions involving motor vehicles; and bike collisions are limited to bike-vehicle collisions. The records on collision severity (i.e. fatality, injury, and property-damage-only) are from the ICBC database.

Most independent candidate variables used in this study are derived from Lovegrove's study (2007), mentioned in *Section 2.4.1*. These variables are divided into four themes, including exposure, social-demographics (S-D), transportation demand management (TDM), and road network (NW). These variable data are from 2006 Census Canada data, the City of Kelowna, the RDCO community plan, and BC transit. As the signal and intersection GIS data in West Kelowna, Lake Country, and Peachland are not available, signals and intersections in these three municipalities are collected from Google Earth. In addition, the modeled exposure data – vehicle kilometers travelled (VKT) – is not available, so only the measured exposure data – total lane kilometres (TLKM) – is used as one leading exposure variable.

New variables related to bicycle use are considered in this study because they are critical for bicycle-related macro-level CPMs. These *new* variables should be as practical for use and easy to collect as *old* variables. They are derived from the factors influencing bicycle use, which are summarized in *Section 2.5.1*. Bike lane kilometres (BLKM) is a new exposure variable to represent bicycle exposure and is provided by the City of Kelowna (only Kelowna has marked bicycle lanes in the RDCO). Although BLKM is not the best variable to reflect bicycle use, it is

the only available bicycle exposure surrogate. In future, the UBC Sustainable Transport Safety Lab will collect bicycle counts at intersections and use them in transportation planning models to get more accurate bicycle exposure data, such as bicycling kilometres travelled (BKT). The demographic variables such as the percentage of population at and under 30 years (POP30) and the ratio of males and females (M/F) are also suggested as *new* candidate variables. Geographic variables and social environment variables are not considered in this study because these variables should be the same among zones.

Finally, forty old variables and three new variables (i.e. BLKM, POP30, M/F) related to bicycle use are set as independent candidate variables. Generally, any variable that has potential association with collisions may be set as an independent candidate variable; however, statistic techniques are needed to test whether it is valid to be kept in a final model. The data sources and statistical summary including maximum/minimum/average values of all variables are presented in Table 3.1.

Table 3.1(a) Dependent and Exposure Variable Definitions & Data Summary

Variables	Symbol	Source	Years	Zonal min	Zonal max	Zonal Avg.
Dependent Variables						
Bicycle/vehicle collisions	B5	ICBC	02-06[1]	0	6	0.37
Total collisions	T3	ICBC	04-06	0	371	21.74
Severe collisions	S3	ICBC	04-06	0	161	8.26
Exposure						
Total lane km	TLKM	CanMap®	2006	0.00	63.25	5.79
Total bicycle lane km	BLKM	CoK[2]	2006	0.00	7.12	0.70
Zonal Area (Hectares)	AR	RDCO	2005	1.33	163.12	88.72

Notes:
[1] 5-year instead of 3-year data for bike-vehicle collisions were utilized because bike-vehicle collisions are extremely rare events in this study area.
[2] CoK: City of Kelowna

Table 3.1(b) S-D Variable Definitions & Data Summary

Variables	Symbol	Source	Years	Zonal min	Zonal max	Zonal Avg.
Urban zones	URB	RDCO	2005	n/a	n/a	n/a
Rural zones	RUR	RDCO	2005	n/a	n/a	n/a
Population	POP	Census	2006	0	2858	320
Population density (=POP/AR)	POPD	Census	2006	0.00	85.35	11.26
Population aged < 30/POP (%)	POP30	Census	2006	0.00	53.57	31.39
Male/female ratio	M/F	Census	2006	0.50	1.67	1.04
Home	NH	Census	2006	0	1012	132
Home density	NHD	Census	2006	0.00	56.45	5.32
Participation in labour force (=(EMP+UNEMP)/POP15[1]) (%)	PARTP	Census	2006	12.34	84.76	62.13
Employed residents	EMP	Census	2006	1	1539	162
Employed percentage (=EMP/POP15*) (%)	EMPP	Census	2006	12.27	82.72	58.98
Employed density(=EMP/AR)	EMPD	Census	2006	0.01	58.59	5.50
Unemployed residents	UNEMP	Census	2006	0	79	8
Unemployed rate (=UNEMP/(UNEMP+EMP)) (%)	UNEMPP	Census	2006	0.00	19.15	5.04
Average income $	INCA	Census	2006	6100	69600	32000
Notes: [1] POP15: Population aged 15 and over in 2006.						

Table 3.1(c) TDM Variable Definitions & Data Summary

Variables	Symbol	Source	Years	Zonal min	Zonal max	Zonal Avg.
Total commuters	TCM	Census	2006	0	1315	145
Commuter density(=TCM/AR)	TCD	Census	2006	0.00	55.19	5.05
Core area(Hectares)	CORE	CanMap®	2006	0.00	293.67	17.88
Core area percentage	CRP	CanMap®	2006	0.00	100.00	42.34
Car passenger commuter percentage (%)	PASS	Census	2006	0.00	18.11	7.07
Transit commuter percentage (%)	BUS	Census	2006	0.00	19.28	2.45
Biking commuter percentage (%)	BIKE	Census	2006	0.00	14.05	1.88
Pedestrian percentage (%)	WALK	Census	2006	0.00	31.44	5.20
No. of driving commuters	DRIVE	Census	2006	0	1179	118
Driving commuter percentage (%)	DRP	Census	2006	47.17	100.00	78.85
Bus stops	BS	BC Transit	2006	0	14	1.60
Bus stop density	BSD	BC Transit	2006	0.00	2.82	0.07

Table 3.1(d) NW Variable Definitions & Data Summary

Variables	Symbol	Source	Years	Zonal min	Zonal max	Zonal Avg.
No. of Signals	SIG	CoK /GE*	2010**	0	4	0
Signal density	SIGD	CoK /GE	2010	0.00	0.89	0.02
No. of intersections	INT	RDCO/GE	2006	0	50	6.16
Intersection density	INTD	RDCO/GE	2006	0.00	1.50	0.19
No. of intersections/TLKM	INTKD	RDCO/GE	2006	0.00	7.67	1.09
No. of 3 way intersections/INT (%)	I3WP	RDCO/GE	2006	0.00	100.00	66.05
No. of Arterial-local intersections/INT (%)	IALP	RDCO/GE	2006	0.00	100.00	15.02
No. of arterial lane-km	ALKM	CanMap®	2006	0.00	19.43	0.84
No. of collector lane-km	CLKM	CanMap®	2006	0.00	32.57	0.73
No. of local lane-km	LLKM	CanMap®	2006	0.00	37.65	4.16
No. of arterial lane-km/TLKM (%)	ALKP	CanMap®	2006	0.00	100.00	13.75
No. of collector lane-km/TLKM (%)	CLKP	CanMap®	2006	0.00	100.00	11.89
No. of local lane-km/TLKM (%)	LLKP	CanMap®	2006	0.00	10.00	68.13

Notes: [1] GE: Google Earth.

[2] 2010: The data were original from 2010 data, and some adjustment were made to make them more close to 2006 data

3.2 Model Forms

This study focuses on how overall safety and cyclist safety would change with more bicycle use. The bicycle exposure variable (i.e. BLKM in this study) is considered in new model forms to reflect the effects of bicycle splits on road safety. Five possible model forms are proposed initially to predict total vehicle, severe vehicle and bike-vehicle collisions, as shown in Table 3.2.

Table 3.2 Possible Model Forms

$E = a_0 Z^{a_1} e^{a_2 B + \sum b_j X_j}$	[1]
$E = a_0 e^{a_1 Z + a_2 B + \sum b_j X_j}$	[2]
$E = a_0 \left(\frac{Z}{B+1} \right)^{a_1} e^{\sum b_j X_j}$	[3]
$E = a_0 Z^{a_1} (B+1)^{a_2} e^{\sum b_j X_j}$	[4]
$E = a_0 (B+1)^{a_1} e^{a_2 Z + \sum b_j X_j}$	[5]

In these candidate model forms, E represents the collision estimates in different collision patterns; a_0, a_1, a_2, and b_j are model parameters; Z is vehicle exposure variable (i.e. total lane kilometres); B is bicycle exposure variable (i.e. bike lane kilometres); and X_j are other independent variables. All of the model forms support the 'product-of-exposure-to-power' relationship, which has been demonstrated by previous macro-level CPM studies (Jacobson, 2003; Lovegrove, 2007; Hadayeghi et al., 2003; Ladron de Guevara et al., 2004). The traditional model forms [1] and [2] have appeared in previous studies. These two model forms are appropriate for the regions where bicycle use is low (e.g. in North America) because the bicycle exposure in these regions may not be a leading, or even a statistically significant variable to influence road safety. However, the significance of B in these two model forms still needs to be tested using statistical measures. The model forms [3] and [4] are proposed for the regions where bicycle use is medium or high. Given a statistical relationship between collisions and the ratio of vehicle to bicycle exposure, the model form [3] is expected to present a positive effect on road safety with increasing bicycle use and decreasing vehicle use. The model form [4] is modified based on the model form [3], in which the vehicle and bicycle exposure variables have individual exponents. The model form [5] is proposed to predict bike-vehicle collisions, in which the bicycle exposure variable is set as a leading variable. In the model forms [3], [4], and [5], $(B+1)$, rather than B, is used to avoid generating zero collisions caused by zero bicycle exposure.

Based on the data from RDCO, macro-level CPMs using the NB regression method are initially developed to validate these model forms. Given that TLKM is set as Z and BLKM is set as B, three results are revealed. First, the model forms [3] and [4] are inappropriate to predict all three collision patterns (i.e. total, severe, and bicycle collisions) in RDCO, because the variable BLKM in these two forms is not statistically significant. This result confirms that model form [3] and [4] are not valid forms to develop macro-level CPMs in regions with low bicycle use. Second, the model forms [1] and [2] are valid to develop macro-level total, severe, and bike CPMs. Third, the model form [5] is qualified to develop bike-vehicle CPMs.

Although both the model forms [1] and [2] are valid to develop total and severe CPMs, only the model form [1] is used for total and severe CPM development in this study because this form is more reasonable than the other one and its goodness-of-fit performance is superior as well. For macro-level bike CPMs, because they have not been extensively researched in previous

studies, all the valid model forms [1], [2], and [5] for bike CPM development are analyzed in this study. The best model form for bike CPMs is initially selected based on the model results using the NB regression method and then recommended for use in developing bike CPMs using other regression methods.

3.3 Model Grouping

To reduce confounding effects and improve model performances, stratification is conducted in this study. Dependent collision variables are stratified into three collision patterns, including: total collisions, severe collisions, and bike collisions. Independent variables are stratified according to variable themes, land use (i.e. urban or rural), and data derivation (i.e. modeled or measured) (Lovegrove, 2007). Only the CPMs in the urban area of RDCO are developed because bicycle collisions in the rural area are too low (e.g. only 18 of 258 rural TAZs have 1 or 2 bicycle collisions in 2002-2006, while the rest of the TAZs recorded no collisions in this period). Additionally, only measured CPMs including the leading variable of TLKM or BLKM are developed in this study as the modeled data such as VKT or BKT are not available. Finally, a three by four matrix of models are developed, as shown in Table 3.3.

Table 3.3 Model Groups

	Total vehicle collisions	**Severe vehicle collisions**	**Bike-vehicle collisions**
Exposure	Urban, Measured	Urban, Measured	Urban, Measured
S-D	Urban, Measured	Urban, Measured	Urban, Measured
Transport Demand Management	Urban, Measured	Urban, Measured	Urban, Measured
Road Network	Urban, Measured	Urban, Measured	Urban, Measured

3.4 Model Results and Discussions

This section describes the model development process and model results using traditional generalized linear regression (GLM), zero-inflated count (ZIC) regression, geographically weighted regression (GWR), and full Bayesian (FB) regression individually. The principles of these four regression methods are introduced in *Section 2.4*.

3.4.1 Negative Binomial Regression Model Results

The traditional negative binomial (NB) regression method is the most commonly used method in CPM development. The model developed using the NB regression method in this study follows the methodology described in Lovegrove and Sayed (2006). Independent variables in each model are added in a forward step, which involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion (i.e. scaled deviance in this study), adding the variable from the same variable theme group to improve the model the most, and repeating this process until there is no significant improvement. The first variable in the model should be the traffic exposure variable (i.e. total lane kilometres or bike lane kilometres) because of its dominating influence on collisions. According to Sawalha and Sayed (2006), the decision to keep a variable in the model is based on four criteria, including: 1) the logic (i.e. +/-) of the estimated parameter for a variable should be intuitively associated with collisions, 2) the t-statistic of the variable should be significant at the 95% confidence level (i.e. >1.96), 3) independent variables should have little or no correlation (or collinearity) between each other (i.e. any correlation coefficient between two independent variables must be under 0.5 in this study. See Appendix A), and 4) the addition of one variable should make a significant drop in the scaled deviance (SD) measure at the 95% confidence level (i.e. >3.84).

The goodness-of-fit tests used in NB regression models include SD, Pearson χ^2, grouped SD, and the AIC. SD and Pearson χ^2 are two quantitative tests in GLMs, recommended by McCullagh and Nelder (1989). The grouped SD method is specifically used to test bike-vehicle CPMs, following the methodology of Wood (2002). The AIC is used for model form selection for bike-vehicle CPMs. The formulations of the four measures have been shown in *Chapter 2* and are repeated here

$$SD = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{E(\Lambda_i)}\right) - (y_i + \kappa)\ln\left(\frac{y_i + \kappa}{E(\Lambda_i) + \kappa}\right)\right] \tag{2.16}$$

$$Pearson\chi^2 = \sum_{i=1}^{n}\frac{[y_i - E(\Lambda_i)]^2}{Var(y_i)} \tag{2.17}$$

$$G.SD = 2\sum_{i=1}^{n} r_i\left[\bar{y}_i \ln\left(\frac{\bar{y}_i}{\hat{E}(\Lambda_i)}\right) - (\bar{y}_i + \hat{\kappa})\ln\left(\frac{\bar{y}_i + \hat{\kappa}}{\hat{E}(\Lambda_i) + \hat{\kappa}}\right)\right] \tag{2.18}$$

$$AIC = \sum_{i=1}^{n}\{2p_i + n[\ln\left(\frac{2\pi\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}\right) + 1]\} \quad (2.35)$$

Outlier analyses are undertaken to refine models. According to Sawalha and Sayed (2006), the data point with the highest CD value is removed from the data sample until the SD and Pearson χ^2 are smaller than the standard χ^2 value. The number of removed outliers should not exceed 5% of the total sample number, about 10 outliers in this study. The NB models before and after outlier analyses are compared to see if there are apparent changes in parameter estimates.

The regression software GenStat 11 (VSN International, 2008) is used for NB model development. After running models, basic model outputs such as over-dispersion parameter κ, degree of freedom, deviance, mean deviance, F tests, parameter estimates, and t-tests of parameters can be derived from this software. Other model results, such as fitted values, standard residuals, and leverages, can be obtained optionally.

Table 3.4 shows the NB model results before outlier analyses and Table 3.5 shows the NB model results after outlier analyses. The comparison results suggest that parameter estimates are significantly different before and after outlier analyses and the goodness-of-fit performances of refined models are better than original models. Also, the outlier analyses are not applied in bike-vehicle CPMs as their goodness-of-fit tests indicate that they fit data well even without outlier analyses. As shown in Table 3.5(a-b), two CPMs for total vehicle collisions, four CPMs for severe vehicle collisions, and nine CPMs for bike-vehicle collisions are finalized. The total vehicle CPMs in the S-D and TDM groups are invalid because the goodness-of-fit tests (i.e. SD and Pearson χ^2) indicate they do not fit the data well at the 95% confidence level. Bike-vehicle CPMs in three different model forms are shown separately. The bike-vehicle CPMs from the S-D group are not developed since no S-D variables show significant statistical relationships with bike-vehicle collisions according to their t-tests.

As mentioned in *Section 2.3.1*, SD and Pearson χ^2 are not enough to validate the goodness-of-fit results of bike-vehicle CPMs, so the grouped SD (GSD) method is used. It is found that the GSD values in all bike-vehicle CPMs are smaller than their grouped critical χ^2 value (i.e. χ_1^2) except the model in the exposure group from the model form [5]. This result indicates that most bike-vehicle CPMs are valid. Also, the AIC values of bike-vehicle CPMs from the model form

[1] are generally smaller than the AIC values of bike-vehicle CPMs from the model forms [2] and [5], suggesting the model form [1] is the best form for bike-vehicle CPM development.

Table 3.4 Macro-level CPM Results with NB Regression–Total/Severe Vehicle Collisions Before Outlier Analyses

Model Form [1] & t-statistics	**κ**	**DoF**	**SD**	***Pearson*** χ^2	$\chi^2_{0.05,dof}$	**AIC**	**AICc**
EXP	0.5249	236	285.6	354.8	272.8	2412.2	2412.9
$T3 = 23.4336 TLKM^{0.3553}$							
t-statistics: con=23.07, tlkm=3.59							
EXP	0.4254	236	272.4	315.2	272.8	1768.3	1769.1
$S3 = 9.3617 TLKM^{0.3563}$							
t-statistics: con=14.89, tlkm=3.29							
SD	0.5871	233	280.6	405.9	269.6	2365.8	2366.3
$T3 = 118.7809 TLKM^{0.6571} e^{-0.0343 PARTP}$							
t-statistics: con=12.86, tlkm=6.63, partp=-5.30							
SD	0.5290	231	268.4	280.4	268.4	1694.5	1694.9
$S3 = 348.8537 TLKM^{0.7232} e^{-1.1153 FS - 0.0546 EMPD - 0.0000416 INCA}$							
t-statistics: con=9.38, tlkm=6.18, fs=-4.59, empd=-4.45, inca=-2.37							
TDM	0.5807	232	280.8	352.7	268.5	2314.5	2314.8
$T3 = 29.0029 TLKM^{0.5623} e^{-0.0105 CRP + 1.4546 BSD}$							
t-statistics: con=14.93, tlkm=5.82 crp=-3.71, bsd=4.12							
TDM	0.4639	232	269.7	307.5	268.5	1676.1	1677.1
$S3 = 11.9034 TLKM^{0.5474} e^{-0.0103 CRP + 1.3069 BSD}$							
t-statistics: con=9.92, tlkm=5.08, crp=-3.29, bsd=3.33							
NW	0.9016	232	278.8	499.1	268.5	2343.9	2344.5
$T3 = 49.8394 TLKM^{0.4689} e^{-0.0179 LLKP + 0.5663 SIG + 0.0082 IALP - 0.0076 I3WP}$							
t-statistics: con=13.98, tlkm=5.56, llkp=-5.79, sig=4.51, ialp=3.15, i3wp=-3.24							
NW	0.7742	232	272.5	414.7	268.5	1644.5	1644.8
$S3 = 21.1861 TLKM^{0.5008} e^{-0.0188 LLKP + 0.5653 SIG - 0.0094 I3WP + 0.0095 IALP}$							
t-statistics: con=10.01, tlkm=5.25, llkp=-5.50, sig=4.16, i3wp=-3.60, ialp=3.39							

Table 3.5(a) Macro-level CPM Results with NB Regression–Total/Severe Vehicle Collisions

Model Form [1] & t-statistics	**κ**	**DoF**	**SD**	***Pearson*** χ^2	$\chi^2_{0.05,dof}$	**AIC**	**AICc**
EXP	0.6236	228	272.4	278.0	263.1	1927.4	1927.5
$T3 = 14.43 TLKM^{0.5625}$							
t-statistics: con=20.50, tlkm=6.10							
EXP	0.5072	228	261.5	233.5	264.2	1493.6	1493.7
$S3 = 5.578 TLKM^{0.5760}$							
t-statistics: con=11.81, tlkm=5.62							
S-D	0.5267	228	263.4	242.8	264.2	1542.7	1543.1
$S3 = 453.2 TLKM^{0.839} e^{-1.165 FS - 0.051 EMPD - 0.0000529 INCA}$							
t-statistics: con=9.94, tlkm=7.17, fs=-4.86, empd=-4.68, inca=-3.09							
TDM	0.5055	229	265.7	261.4	265.3	1565.3	1565.6
$S3 = 9.585 TLKM^{0.5055} e^{-0.01010 CRP + 1.981 BSD}$							
t-statistics: con=9.27, tlkm=5.34, crp=-3.34, bsd=4.10							
NW	1.1053	230	272.2	222.5	266.4	1885.0	1885.5
$T3 = 31.78 TLKM^{0.6455} e^{-0.01972 LLKP + 0.693 SIG + 0.00906 IALP - 0.046 I3WP}$							
t-statistics: con=13.51, tlkm=8.16, llkp=-6.88, sig=6.13, ialp=3.88, i3wp=-2.12							
NW	0.9553	229	265.1	210.0	265.3	1441.1	1441.6
$S3 = 15.32 TLKM^{0.7074} e^{-0.02087 LLKP + 0.609 SIG - 0.00804 I3WP + 0.00953 IALP}$							
t-statistics: con=9.65, tlkm=7.68, llkp=-6.51, sig=4.89, i3wp=-3.25, ialp=3.72							

Table 3.5(b-1) Macro-level CPM Results with NB Regression–Bike-vehicle Collisions

Model Form [1] & t-statistics	**κ**	**DoF**	**SD**	***Pearson*** χ^2	$\chi^2_{0.05,dof}$	**AIC**	**G. SD**	χ_1^2 **(dof)**
EXP	0.6095	236	193.7	240.6	272.8	535.3	61.7	73.3 (55)
$B5 = 0.4352 TLKM^{0.4140}$								
t-statistics: con=-5.05, tlkm=3.76								
S-D	N/A							
TDM	0.7539	233	195.3	270.6	269.6	522.9	65.1	74.5 (56)
$B5 = 8.5255 TLKM^{0.6061} e^{-0.0434 DRP}$								
t-statistics: con=2.56, tlkm=4.19, drp=-3.58								
NW	1.1892	233	189.5	225.7	269.6	491.8	68.0	77.9 (59)
$B5 = 0.1033 TLKM^{0.4077} e^{0.5229 SIG + 1.9687 INTD + 0.0123 IALP}$								
t-statistics: con=-6.91, tlkm=3.17, sig=3.59, intd=4.13, ialp=3.99								

Table 3.5(b-2) Macro-level CPM Results with NB Regression–Bike-vehicle Collisions

Model Form [2] & t-statistics	**κ**	**DoF**	**SD**	***Pearson*** χ^2	$\chi^2_{0.05,dof}$	**AIC**	**G. SD**	χ_1^2 **(dof)**
EXP	0.5706	240	196.5	234.1	277.1	543.7	69.7	73.3 (55)
$B5 = 0.5402e^{0.2582\,BLKM}$								
t-statistics: con=-4.57, blkm=2.70								
S-D	N/A							
TDM	0.6993	236	197.2	235.5	272.8	532.6	63.8	74.5 (56)
$B5 = 5.2630e^{0.224\,BLKM\,-0.034\,DRP+0.1074\,BS}$								
t-statistics: con=-2.30, blkm=2.18, drp=-3.38, bs=2.79								
NW	1.0693	237	190.5	227.3	273.9	500.3	68.0	74.5 (56)
$B5 = 0.1283e^{0.237\,BLKM\,+0.446\,SIG+2.168\,INTD+0.0126\,IALP}$								
t-statistics: con=-8.77, blkm=2.46, sig=3.27, intd=5.88, ialp=4.85								

Table 3.5(b-3) Macro-level CPM Results with NB Regression–Bike-vehicle Collisions

Model Form [5] & t-statistics	**κ**	**DoF**	**SD**	***Pearson*** χ^2	$\chi^2_{0.05,dof}$	**AIC**	**G. SD**	χ_1^2 **(dof)**
EXP	0.5742	240	196.8	238.4	277.1	543.9	127.0	74.5 (56)
$B5 = 0.5177(BLKM+1)^{0.548}$								
t-statistics: con=-4.51, blkm=2.66								
S-D	N/A							
TDM	0.7047	236	197.5	250.5	272.8	532.8	68.6	74.5 (56)
$B5 = 4.922(BLKM+1)^{0.478}e^{-0.0337\,DRP+0.1098\,BS}$								
t-statistics: con=2.22, blkm=2.20, drp=-3.37, bs=2.89								
NW	1.0688	237	190.4	227.8	273.9	500.6	55.2	74.5 (56)
$B5 = 0.1224(BLKM+1)^{0.511}e^{0.441\,SIG+2.19\,INTD+0.01254\,IALP}$								
t-statistics: con=-8.54, blkm=2.40, sig=3.21, intd=5.87, ialp=4.84								

The statistical relationships of total/severe/bicycle collisions and their independent variables are individually described as follows. In total vehicle CPMs, increased total vehicle collisions are associated with increased total lane kilometers (TLKM), zonal traffic signals (SIG), and the

parentage of arterial-local intersections (IALP). The relationships with SIG and IALP are reasonable as signalized or arterial-local intersections usually have traffic with high speed and high volume, and accordingly represent high collision risks. Conversely, decreased total vehicle collisions are associated with increased three-way intersection percentage (I3WP) and local lane kilometre percentage (LLKP). These relationships are also consistent with the statistical associations described in Lovegrove (2007). The variable BLKM is not included in total vehicle CPMs as it is not statistically significant in model development. This situation is understandable because the vast majority of collisions in RDCO do not involve bicycles.

In severe vehicle CPMs, increased severe vehicle collisions are associated with increased total lane kilometers (TLKM), bus stop density (BSD), zonal traffic signals (SIG), and arterial-local intersection percentage (IALP). Conversely, decreased severe vehicle collisions are associated with increased family size (FS), employment density (EMPD), zonal average income (INCA), core area percentage (CRP), three-way intersection percentage (I3WP), and local lane kilometre percentage (LLKP). The associations with BSD, CRP, IALP, SIG, I3WP, and LLKP confirm earlier findings by van Minnen (1999) and Lovegrove (2007). Collision frequency has been shown to be reduced with FS. It has been suggested that parents of large families are usually more responsible drivers (Ladron de Guevara et al., 2004). However, the associations with EMPD and INCA are difficult to explain because they are counterintuitive. From an intuitive perspective, higher employment densities lead to more commuters, then more traffic exposure and collisions; and higher average income results in more cars, then more traffic exposure and collisions.

In bike-vehicle CPMs, increased bike-vehicle collisions are associated with increased TLKM, BLKM, zonal bus stops (BS), SIG, zonal intersection density (INTD), and IALP. Conversely, decreased bike-vehicle collisions are related to increased drive commuter percentage (DRP). The associations with traffic exposure (i.e. TLKM and BLKM) confirm the intuitive expectation that more traffic exposure contributes is related to more bike-vehicle collisions. The association with BS is consistent with Kim's research (2010). The associations with SIG, INTD, and IALP indicate that intersections with high traffic volume, high speeds, and many turning movements are high risky locations for cyclists. The association with DRP is also reasonable because when more commuters choose to drive, a low bicycle mode share would result, intuitively leading to fewer bike-vehicle collisions.

3.4.2 Geographically Weighted Regression Model Results

Geographically weighted regression (GWR) is one regression method that takes spatial variations into account. In this study, the centroid location of each zone (i.e. longitude and latitude) is provided as the geographical reference of each point in the data sample, as shown in Figure 3.2.

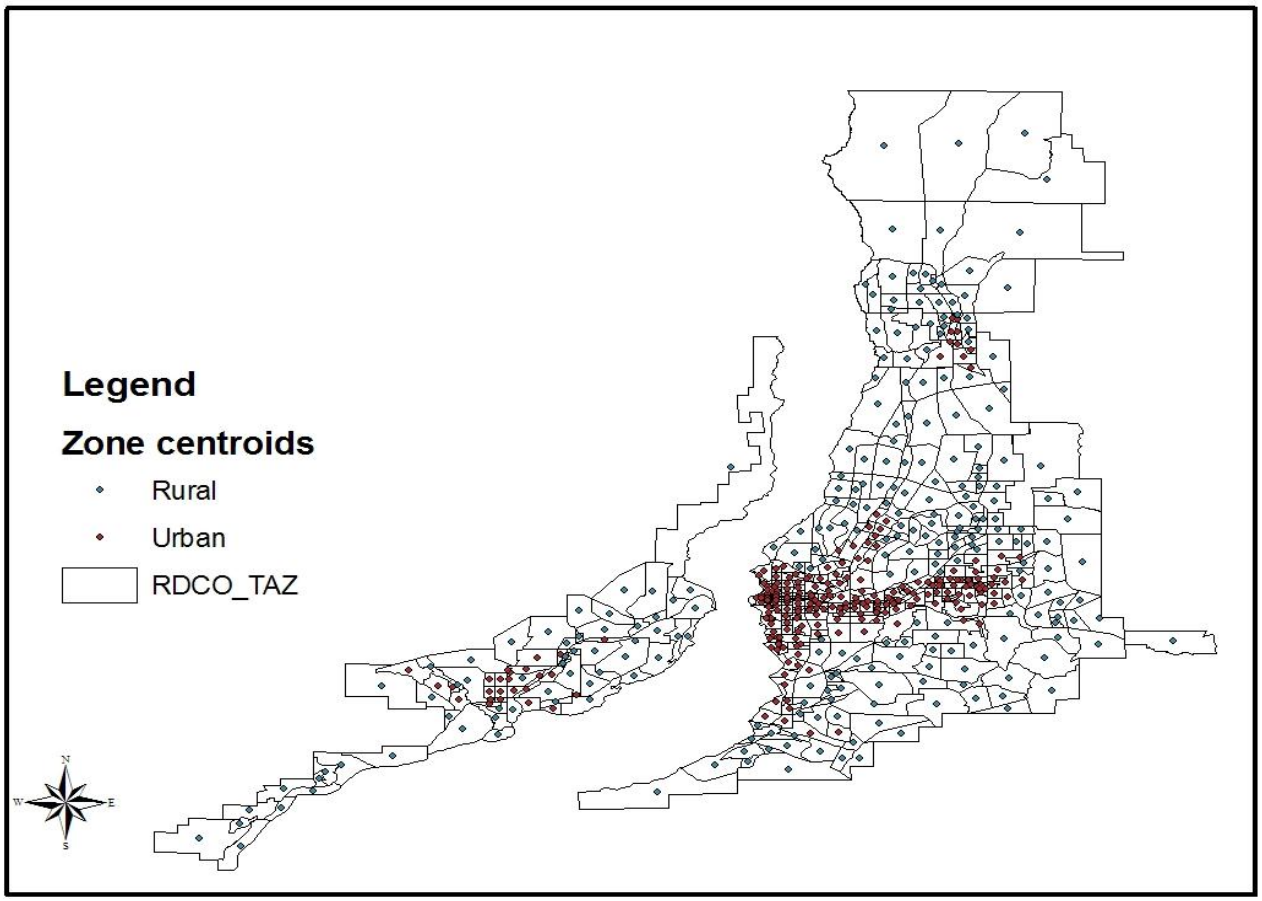

Figure 3.2 TAZ Centroids

The software GWR 3.0 (NCG, 2009) is used for GWR model development. In GWR 3.0, geographically weighted Gaussian regression, geographically weighted Poisson regression, and geographically weighted logistic regression are three options for model development. A geographically weighted Poisson regression is selected as it is more appropriate for CPM development than the other two. The adaptive kernel is set as the kernel type and the AIC is

used to determine the kernel bandwidth for GWR model development in this study. Unfortunately, GWR models are not developed successfully because the modelling simulation stops after 60 iterations of attempts to determine the parameter values. This result may be because of data issues. Meanwhile, it raises doubts as to whether GWR is a reliable regression method to develop multiple variable models and future work should focus on using other data to valid this method.

3.4.3 Zero-inflated Count Regression Model Results

Figure 3.3 shows the distribution of zonal bike-vehicle collisions. The data points (i.e. TAZs) with zero bike-vehicle collisions account for approximately 65% of the sample. Therefore, the zero-inflated count (ZIC) regression method is used to develop macro-level bike-vehicle CPMs.

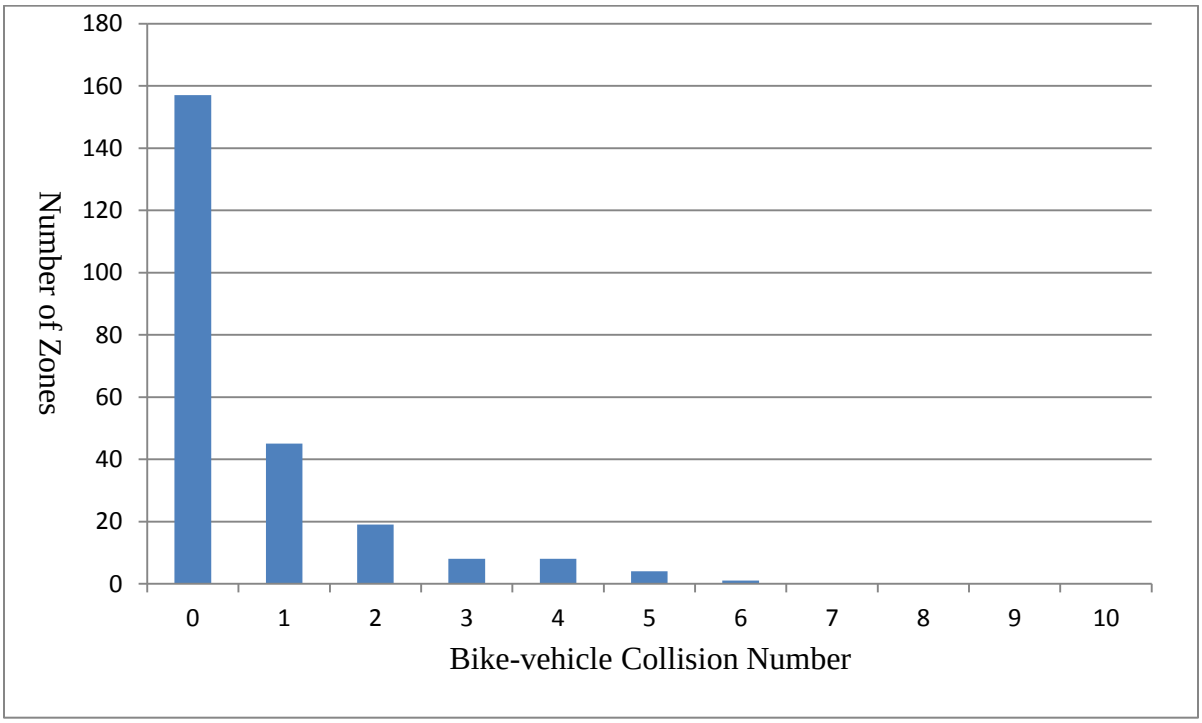

Figure 3.3 Distribution Histogram of Bike-vehicle Collisions

As mentioned before, the model form [1] has better performance than the model forms [2] and [5] to develop NB models for bike-vehicle collisions. In order to make an unbiased comparison (i.e. apples to apples) between different regression methods, the model form [1] is

used to develop ZIC models for bike-vehicle collisions. Following the methodology to develop ZIC models, the predicted bike-vehicle collisions in each TAZ is given as

$$E_B = a_0 TLKM^{a_1} e^{\sum b_j X_j} (1-\varphi) \tag{3.1}$$

where φ is a logistic function to represent the probability of 'zero' collisions in a zone, written as

$$\varphi = \frac{e^{\alpha+\sum \beta X}}{1+e^{\alpha+\sum \beta X}} \tag{3.2}$$

where α and β are parameters, and X are independent variables surrogating traffic exposure and other neighbourhood characteristics. Traffic exposure variables such as TLKM, BLKM, the logarithm form of TLKM, the logarithm form of BLKM, and the binary form of BLKM (i.e. '1' for BLKM if there are bike lanes and '0' for BLKM if there are no bike lanes) are involved in this logistic function individually or in combination. By comparing the Akaike information criteria (AIC) of ZIC models with these different logistics functions, the logistic function in which the logarithm of TLKM is set as an independent variable is finalized because it gives the corresponding ZIC model the smallest AIC.

The software SAS 9.3 (SAS Institute Inc., 2011) is used to develop ZIC models, including zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models. In the regression processes using SAS 9.3, the GENMOD procedure for fitting generalized linear models is used, the Poisson distribution is the underlying distribution for ZIP models, the negative binomial distribution is the underlying distribution for ZINB models, and the default parameter estimation method is the maximum likelihood method. An example of SAS codes for the ZIP model procedure is presented in *Appendix B.* The analysis outputs from SAS provide the basic model information, goodness-of-fit test results, parameter estimation, converge status, and others if specified.

Table 3.6 shows one ZIP model and one ZINB model with the best goodness-of-fit performances. As shown in this table, t statistics of the variable TLKM in the non-zero collision probability state are -0.28 for ZIP and 0.03 in ZINB, which indicate that TLKM in both models are not significant. The Vuong test is used to test whether the ZIP and ZINB regression methods are better than traditional NB method in this study. The Vuong test is discussed in *Section 2.3.2* and its formulation is shown as

$$V = \frac{\bar{m}\sqrt{N}}{S_m} \tag{2.28}$$

Vuong (1989) suggested using correction factors, such as the criteria proposed by Akaike (1974) or Schwarz (1978), to adjust this likelihood ratio statistic. In SAS 9.3, the unadjusted and adjusted Voung tests can be completed. Table 3.7 shows the Voung test results for comparing the two ZIC models in Table 3.6 and their parent NB model.

Table 3.6 ZIP and ZINB Model Results for Bike-vehicle Collisions

	ZIP		**ZINB**	
	Coefficient	**t statistics**	**Coefficient**	**t statistics**
Non zero collision probability state				
Constant (Exposure)	1.5723	2.36	0.9429	-0.13
TLKM	-0.0374	-0.28	0.0068	0.03
Dispersion Parameter κ			1.0362	
Zero collision probability state				
Constant	0.8667	3.13	0.0565	0.07
Ln TLKM	-0.7211	-3.03	-1.4796	-1.84
Goodness of fit tests				
SD	530.6383		521.8494	
Pearson χ^2	264.9743		215.1831	
Log likelihood at convergence	-186.6617		-260.9247	
AIC	538.6383		531.8494	

Table 3.7 Vuong Tests to Compare Zero-inflated Count Models and NB Models

	ZIP vs NB			**ZINB vs NB**		
	V	Pr>|V|	Preferred Model	V	Pr>|V|	Preferred Model
Unadjusted	-0.2005	0.8411	NB	1.5305	0.1259	ZINB
Akaike Adjusted	-0.7842	0.4329	NB	0.7049	0.4809	ZINB
Schwarz Adjusted	-1.7977	0.0722	NB	-0.7275	0.4663	NB

Usually, a confidence level is set in the Voung test for model comparisons. For example, if the confidence level is 95%, a value greater than 1.96 for the Vuong test favours the ZIC model, a value less than -1.96 favours the parent NB model, and a value between 1.96 and -1.96 suggests an inconclusive result. As shown in *Table 3.7*, the unadjusted Vuong value between ZIP and NB is -0.2005 (i.e. 1st row), indicating that the NB model is better than the ZIP model at

a 16% (i.e. 1-0.8411) confidence level; the Akaike adjusted Vuong value between ZIP and NB is -0.7842 (i.e. 2ed row), indicating that the NB model is better than the ZIP model at a 57% (i.e. 1-0.4329) confidence level; and the Schwarz adjusted Vuong value is -1.7977 (i.e. 3rd row), indicating that the NB model is better than the ZIP model at a 93% (i.e. 1-0.0722) confidence level. However, an inconclusive result is suggested to tell which one is better at the 95% confidence level. Similarly, the unadjusted Vuong value between ZINB and NB is 1.5305 (i.e. 1st row), indicating that the ZINB model is better than the NB model at a 88% (i.e. 1-0.1259) confidence level; the Akaike adjusted Vuong value is 0.7049 (i.e. 2ed row), indicating that the ZINB model is better than the NB model at a 52% (i.e. 1-0.4809) confidence level, and the Schwarz adjusted Vuong value (i.e. 3rd row) is -0.7275, indicating that the NB model is better than the ZINB model at a 53% (i.e. 1-0.4663) confidence level.

As the ZIC regression method is considered to be a promising method to develop micro-level CPMs for collision samples with excess 'zeros' (Lee and Mannering, 2003; Shankar et al. 2004; Qin et al., 2004), this study attempts to develop macro-level CPMs. However, the results are not consistent with previous research. These macro-level bike-vehicle CPMs using the ZIC method show that the zonal bike-vehicle collisions characterized by a preponderance of zero values are not indicative of ZIP or ZINB distribution. This situation is possibly because the ZIP and ZINB methods are not appropriate for modeling such collision data in this case study or because potential variables in model development are missed. Future research is needed to test if this situation exists in other macro-level CPM studies.

3.4.4 Full Bayesian Model Results

Kim et al. (2002) suggested that both Poisson-gamma (i.e. NB) and Poisson lognormal (PLN) models could account for extra-Poisson variations, but PLN models performed even better than NB models. In this study, PLN models using the FB method are developed and their methodology is similar to those described in Qin et al. (2005), Li et al. (2008), and El-Basyoung (2009). The model form [1] is used to develop PLN models for three collision patterns. The PLN model form is presented as

$$E(y) = e^{(\ln E(\Lambda)+0.5\sigma^2)} = E(\Lambda)e^{0.5\sigma^2} = a_0 TLKM^{a_1} e^{\sum b_j X_j + 0.5\sigma^2} \quad (3.3)$$

where σ is the extra variation parameter in PLN distribution.

In the FB method, the first step is to determine prior distributions of model parameters, and the second step is to estimate posterior distributions of model parameters via combining prior distributions with observed data. In this study, the normal distribution *N* (0, 1000) is used as the prior distribution for any regression parameter, and the gamma distribution *Gamma* (0, 0.1) is used as the prior distribution of the extra variation parameter σ^2. Markov chain Monte Carlo (MCMC) methods are used to estimate posterior distributions of all parameters in FB models. Deviance information criterion (DIC) is used for Bayesian model comparison and selection. For an unbiased comparison, the same variables and the same datasets for negative binomial (NB) models are used to develop FB models.

The software used to develop FB models is WinBUGS 1.4 (MRC Biostatistics Unit, 2004), in which seven steps are required. The first step is to create a model file, a data file, and a data initial file. Three samples for each them are presented in Appendix C. The second step is to open the three files and check the model file. The third step is to load the data file and then to compile it with the model file. The fourth step is to load the data initial file into the model file. The fifth step is to monitor parameter values. In order to obtain posterior summaries of the model parameters and predicted collisions of each data point, users need to set monitors for them. This step allows WinBUGS to store simulated estimates for parameters and predicted collisions, otherwise, WinBUGS would automatically discard these simulated estimates. The sixth step is to run the simulation to get posterior distributions of all parameters. The last step is to obtain the DIC value for the model. The parameter estimates of PLN models using the FB method are presented in Table 3.8(a-c). As PLN models (i.e. *Eq. 3.3*) need to be transferred into link functions in WinBUGS, the constants generated from the software are link functions' constant estimates. The constant estimates in PLN models should be equal to the value of constant *e* (i.e. 2.7183) raised to the power of constant estimates in link functions. The constants shown in *Table 3.8* have been transformed into the constants in PLN models.

The confidence interval is often used to test the variable significance in FB models. This is quite different from the t-statistics used to test the variable significance in NB models. For example, if the parameter of a variable has the same logic (i.e. -/+) from its 2.5% to 97.5% interval, this variable is significant at a 95% confidence interval; otherwise, it is not. As shown in Table 3.8, some variables that are significant in NB models are not significant in FB models

according to the 95% confidence interval. However, the parameter means of FB models and NB models show the same logic.

Table 3.8(a) Macro-level CPMs with FB method–Total Vehicle Collisions

	Exposure, Urban, Measured	**NW**
Cons.	(4.9382,1.0133)[1] (3.7173,6.2464)[2]	(23.1733,2.2291) (4.7398,145.4744)
TLKM	(0.8003,0.0862) (0.6475,0.9761)	(1.0670,0.4847) (0.1298,2.0250)
LLKP		(-0.0273,0.0408) (-0.1094,-0.00645)
I3WP		(-0.0515,0.0294) (-0.1091, 0.0062)
SIG		(0.5730,0.7240) (-0.8459,1.9825)
IALP		(0.0078,0.00488) (-0.00169,0.01649)
σ^2	(0.5118,0.0580) (0.5095,0.6327)	(0.1814,0.2836) (0.0091,0.9181)
DIC	1380.620	1427.160

Note: 1. (Mean, Standard Deviation)
2. (2.5% confidence interval, 97.5% confidence interval)

Table 3.8(b) Macro-level CPMs with FB method–Severe Vehicle Collisions

	EXP	S-D	TDM	NW
Cons.	(1.9307,1.1665)[1] (0.3531,0.9410)[2]	(242.9851,1.7766) (74.515,626.4068)	(1.6705,1.2415) (1.1468, 2.7871)	(5.6238,1.3050) (3.2349, 9.3278)
TLKM	(0.6979,0.0915) (0.5251,0.8759)	(1.0900,0.1165) (0.8849,1.3130)	(0.6947,0.1149) (0.4676,0.9051)	(0.6949, 0.1015) (0.5069,0.9229)
FS		(-1.5100,0.2332) (-1.8560,-1.0410)		
EMPD		(-0.0350,0.0147) (-0.0642,-0.0054)		
INCA		(-5.24E-5,1.35E-5) (-8.32E-5,-2.99E-5)		
CRP			(-0.0017,0.0029) (-0.0078,0.0030)	
BSD			(2.6470,0.6117) (1.5160,3.8300)	
LLKP				(-0.0175,0.0036) (-0.0244,-0.0111)
I3WP				(-0.0053,0.0027) (-0.0104,3.59E-4)
SIG				(0.6142,0.1153) (0.4008,0.8431)
IALP				(0.0129,0.0028) (0.0076,0.0181)
σ^2	(0.4271,0.0553) (0.3275,0.5438)	(0.4824,0.0610) (0.3706,0.6081)	(0.4263,0.0539) (0.3280,0.5393)	(0.8141,0.1133) (0.6074,1.0530)
DIC	1114.950	1144.230	1152.580	1151.740

Note: 1. (Mean, Standard Deviation)
2. (2.5% confidence interval, 97.5% confidence interval)

Table 3.8(c) Macro-level CPMs with FB method–Bike-vehicle Collisions

	EXP	S-D	TDM	NW
Cons.	(0.2207,1.2928)[1] (0.1290,2.0859)[2]		(4.7115,2.5564) (0.6682, 29.0495)	(0.0707,1.4723) (0.0318,0.1454)
TLKM	(0.4617,0.1394) (0.2011,0.7352)		(0.6759,0.1624) (0.3683, 1.0120)	(0.4464,0.1343) (0.1871,0.7173)
DRP			(-0.0439,0.0138) (-0.0731, -0.01675)	
SIG				(0.5426,0.1448) (0.2669,0.8372)
INTD				(1.7950,0.4718) (0.9221,2.723)
IALP				(0.01223,0.003132) (0.006099,0.0185)
σ^2	(0.8143,0.2185) (0.4685,1.3220)		(0.9352,0.2577) (0.5532, 1.557)	(1.3980,0.4996) (0.7367,2.6060)
DIC	488.039		479.284	462.539

Note: 1. (Mean, Standard Deviation)
2. (2.5% confidence interval, 97.5% confidence interval)

Mean absolute deviation (MAD), mean squared prediction error (MSPE), and mean squared error (MSE), three measures to calculate model mis-prediction, are used to compare FB and NB models. A lesser value in MAD, MSPE, and MSE indicates a better model fit. The formulations of these three measures are given in *Section 2.4.5* and are repeated below

$$MAD = \frac{\sum_{i=1}^{n} | \hat{Y}_i - Y_i |}{n} \tag{2.45}$$

$$MSPE = \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n} \tag{2.46}$$

$$MSE = \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n - p} \tag{2.47}$$

Table 3.9 shows the results of the three measures for NB and FB models.

Table 3.9 MAD, MSPE, and MSE of NB Models and FB Models

Measures	Regression methods	Exposure	S-D	TDM	NW
Total vehicle Collisions					
MAD	NB	25.622	N/A	N/A	32.072
	FB	22.847			31.165
MSPE	NB	1469.918			3797.536
	FB	1549.813			3723.111
MSE	NB	1482.812			3896.602
	FB	1570.295			3836.918
Severe Vehicle Collisions					
MAD	NB	10.721	13.381	12.433	10.880
	FB	9.287	11.057	10.914	10.471
MSPE	NB	257.517	475.039	429.315	655.666
	FB	270.768	479.926	465.183	625.500
MSE	NB	259.775	485.457	436.814	672.845
	FB	274.346	492.611	475.384	644.704
Bike-vehicle Collisions					
MAD	NB	0.851	N/A	0.811	0.771
	FB	0.809		0.795	0.891
MSPE	NB	1.414		1.342	1.446
	FB	1.427		1.350	1.850
MSE	NB	1.426		1.359	1.477
	FB	1.445		1.373	1.898

As shown in Table 3.9, the MAD values of most FB models (except the total vehicle and bike-vehicle CPMs in the NW group) are smaller than those of NB models; however, the MPSE and MSE values of most FB models (except the total and severe vehicle CPMs in the NW group) are greater than those of NB models. These two inverse results make it hard to tell which models, FB or NB, are more competitive. However, the degree of MAD differences between FB and NB models is greater than the degree of MSPE and MSE differences between these two models. From this perspective, FB models perform better than NB models. In the future, hierarchical FB models accounting for spatial and temporal variations should be researched to improve currently developed FB models.

3.5 Summary

This chapter describes data extraction, model form selection, model stratification, and model results using the NB, GWR, ZIC, and FB methods. GWR models fail to be developed probably due to the issues of data quality or software algorithm capability. Although ZIC models for bike-vehicle collisions are developed, the models' statistical results indicate that zonal bike-vehicle collision data in the RDCO do not follow a ZIP or ZINB distribution. Finally, only NB models and FB models are valid. Both NB and FB models provide similar statistical relationships between dependent variables and their independent variables. Given the same model form, datasets, and variables, FB and NB models are compared; however, the inconclusive result cannot show which models are better.

Data extraction and model development methodologies still need to be refined in the future. This future work include: 1) improving collision data completeness, 2) collecting more reliable bicycling exposure variables (e.g. bicycle volumes, bicycle kilometer travelled), 3) refining FB models (e.g. develop hierarchical FB models) to account for spatial and temporal variations and variable interactions, and 4) developing macro-level CPMs for other regions to test and validate these regression methods.

CHAPTER 4 MACRO-REACTIVE SAFETY APLLICATION: BLACK SPOT CASE STUDY

A black spot study is a reactive road safety application used to identify and rank hazardous road sites with abnormally high rates of road collisions, to diagnose road safety problems at these sites, and to suggest possible remedies to improve road safety for these sites. This chapter presents a macro-reactive road safety program that uses macro-level CPMs to identify collision prone zones (CPZs) in the RDCO, unlike traditional black spot programs that use micro-level CPMs to identify hazardous road intersections or road segments. Khondakar (2008) has identified CPZs for total and severe vehicle collisions in this region. Since this study intends to research road safety related to bicycle use, this chapter specifically focuses on how to identify, diagnose, and remedy bike-vehicle CPZs.

Three sections are included in this chapter. *Section 4.1* describes identification and ranking results of bike-vehicle CPZs. *Section 4.2* presents possible diagnoses and remedies for all identified bike-vehicle CPZs. *Section 4.3* gives a summary of this chapter.

4.1 Identification and Ranking

Observed collision rates or frequencies could be used to identify and rank black spots, but these measures are not able to account for regression-to-the-mean (RTM) biases. Therefore, collision prediction models (CPMs) with the empirical Bayesian (EB) method were suggested by Hauer (1997) to identify and rank black spots. Despite the advantage to reduce the RTM bias, the EM method still has inadequacies. For example, the EB method is criticized for using data twice, which means that collision data are first used to develop CPMs to estimate collision prior distributions and then utilized to make further inferences to estimate collision posterior distributions (Hauer, 1997; Carlin & Louis, 2000). Moreover, the EB method may be inadequate to account for all uncertainties of associations between covariates and safety (Miaou, 2003). In this case, the full Bayesian (FB) method is proposed as a promising alternative of the EB method (Miaou, 2003; Huang, 2007). This section describes the identification and ranking methodology and results using the EB and FB methods with CPMs developed in *Chapter 3*, including a comparison of the results generated by these two methods.

4.1.1 CPZ Identification and Ranking Results with the EB Method

The identification and ranking methodology using the EB method follows Lovegrove's study (2007), which has been described in *Section 2.4.2*. In this method, the prior estimations of collision estimates (i.e. $E(\Lambda)$) are derived from the three bike-vehicle NB models in *Table 3.5*(*a*). After being combined with observed collisions, the EB or posterior estimates of collisions are obtained. The EB collision means and variances are given as:

$$EB = E(\Lambda \mid Y = count) = \frac{E(\Lambda)}{\kappa + E(\Lambda)}(\kappa + count) \tag{2.48-a}$$

$$Var(EB) = Var(\Lambda \mid Y = count) = \left[\frac{E(\Lambda)}{\kappa + E(\Lambda)}\right]^2 (\kappa + count) \tag{2.48-b}$$

If the EB collision estimate of one zone exceeds its prior collision estimate at a significant confidence level (i.e. δ=0.95), this zone is regarded as a CPZ. This comparison is mathematically formulated as

$$1 - \int_0^{E(\Lambda)} f_{EB}(\lambda)d\lambda = \left[1 - \int_0^{E(\Lambda)} \frac{[\kappa / E(\Lambda) + 1]^{(\kappa + count)} \lambda^{(\kappa + count - 1)} e^{-(\kappa / E(\Lambda) + 1)\lambda}}{\Gamma(\kappa + count)} d\lambda\right] \geq \delta \tag{2.49}$$

Potential collision reduction (*PCR*=*EB*-*E*(Λ)) and collision risk ratio (*CRR*=*EB*/*E*(Λ)) are used to rank bike-vehicle CPZs in this study. As the identification and rank results may vary with different models, a total ranking score for each zone is required by summing the rankings across models. All CPZs are ranked by sorting the total ranking scores.

Following the approach above, bike-vehicle CPZs and safer zones (SZs) in the RDCO are identified and ranked. SZs represent the lowest PCR and CRR ranked zones. *Table 4.1* shows all CPZs and the top 20 SZs according to the rankings in each NB model. Figure 4.1 and Figure 4.2 show the geographic locations of these CPZs and SZs, respectively.

Table 4.1 Urban Bike-vehicle CPZ and SZ Identification Using EB Method

Bike CPZs	Exposure	TDM	NW	Bike SZs	Exposure	TDM	NW
1	231	231	231	**1**	376	451	489
2	253	171	244	**2**	449	211	273
3	171	253	151	**3**	37	37	216
4	244	244	132	**4**	305	128	99
5	307	307	127	**5**	447	265	465
6	151	179	179	**6**	431	449	346
7	179	151	274	**7**	49	245	259
8	483	132		**8**	182	480	155
9	132	177		**9**	257	457	490
10	177	266		**10**	297	116	480
11	266	141		**11**	216	447	458
12	124			**12**	245	131	301
13	127			**13**	116	257	221
14	459			**14**	301	305	454
15				**15**	128	252	207
16				**16**	252	49	469
17				**17**	87	155	87
18				**18**	50	220	128
19				**19**	118	301	358
20				**20**	135	376	302

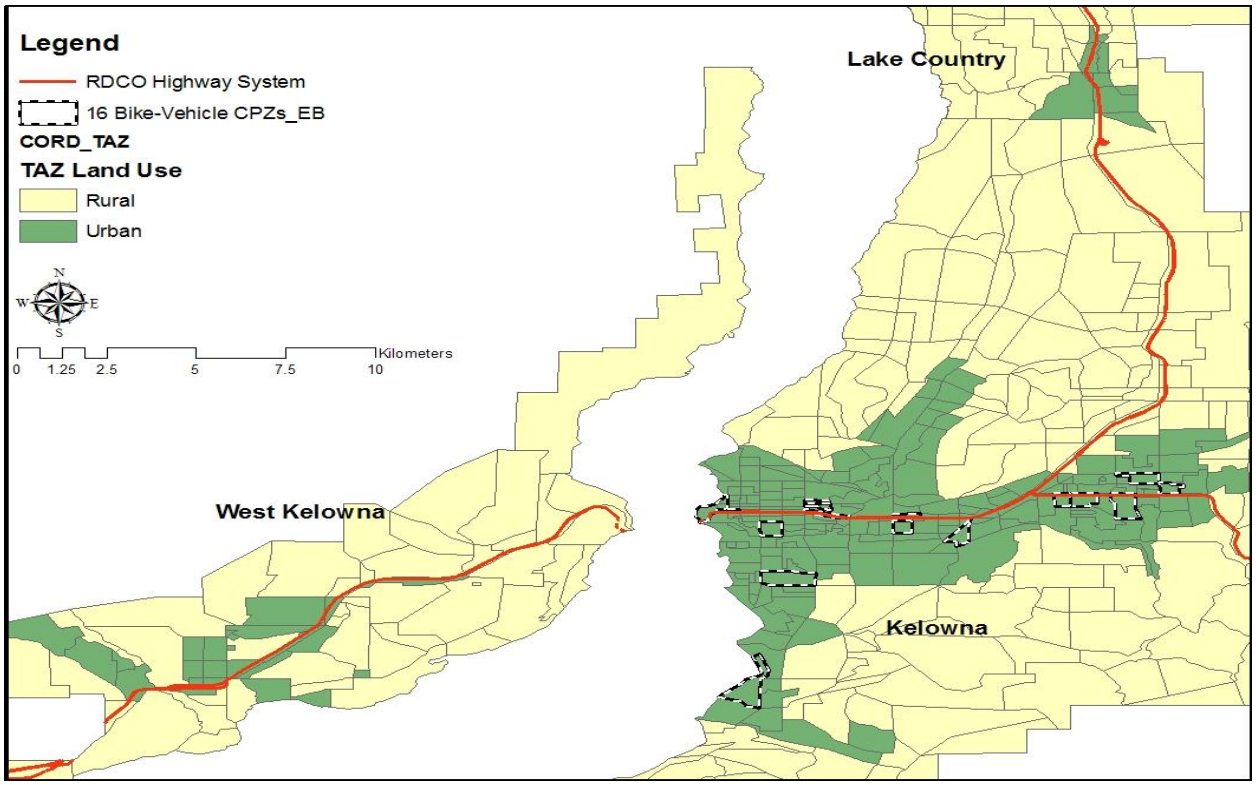

Figure 4.1 Bike-vehicle CPZs Using the EB Method

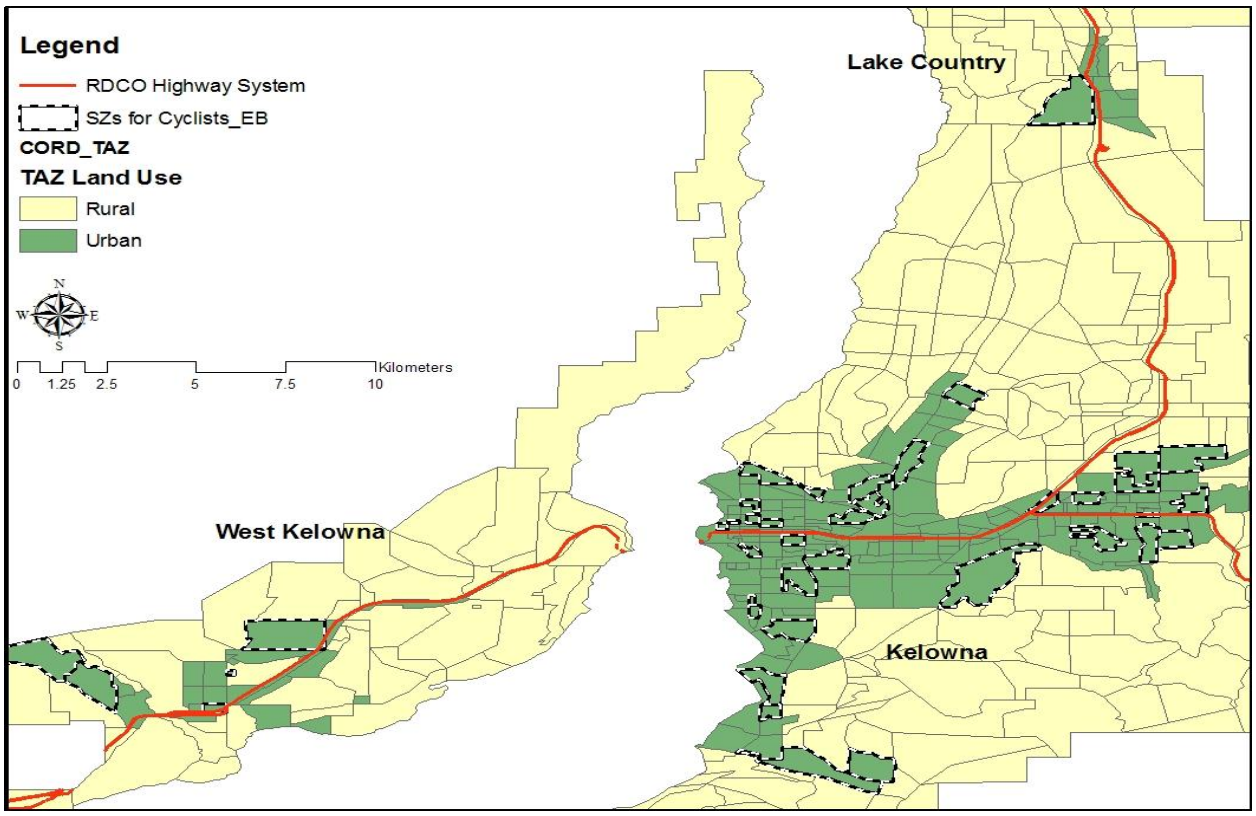

Figure 4.2 Bike-vehicle SZs Using the EB method

Several results related to bike-vehicle CPZ and SZ identification and ranking are observed. First, a total of 16 bike-vehicle CPZs are identified. 14 of 16 CPZs are identified by the model in the exposure group, 11 CPZs are identified by the model in the traffic demand management (TDM) group, and only 7 CPZs are identified by the model in the network (NW) group. Second, although the CPZ identification results from different models cannot fully agree with each other, the average CPZ identification similarity among these models is up to 83%. Moreover, the models from the exposure and TDM groups largely agree with each other in identifying and ranking CPZs. Third, 42 zones are identified as SZs in total if the 20 lowest PCR and CRR ranked zones in each model group are seen as "safer." The identified SZs from the three models are significantly different from each other. For example, only Zone 128 is identified by all three models. Generally, both the CPZ and SZ identification and ranking results generated from the exposure and TDM models are closer to each other than to the NW model. One reason for this might be inaccurate data in the NW model. The data for the NW model development include intersections and signals, which are from 2010 data provided by the City of Kelowna. Although these 2010 data have been adjusted to 2006 data, the data adjustment bias still exists. The other reason might be the exposure model and the TDM model are closer to each other than to the NW model.

4.1.2 CPZ identification and Ranking Results with the FB Method

The FB method used for micro-level black spot identification and ranking (i.e. intersections and road segments) has been described in previous studies (Huang et al., 2009; Lan & Persaud, 2011) and summarized in *Section 2.4.2*. However, no macro-level black spot studies using the FB method are found because few studies on macro-level CPMs using the FB method have been researched so far. As the FB method used in micro-level black spot identification may not be practical for macro-level black spot identification, an approach to identify and rank macro-level black spots using the FB method is proposed in this study. This new approach is derived from the macro-level black spot identification and ranking methodology with the EB method.

In the EB method, NB models are used to generate prior distributions of collisions as the first step, and observed collision counts are combined with the prior distributions to estimate posterior distributions of collisions as the second step. As mentioned before, a zone is identified

as a CPZ in the EB method if its posterior collision mean is greater than its prior collision mean at a confidence level of 95%. In the FB method, posterior distributions of collisions are directly obtained from FB models. Therefore, a similar identification criterion is proposed as follows: if the posterior mean of collisions in one zone exceeds its normal estimate at a confidence level of 95%, this zone is considered to be a CPZ. Figure 4.3 illustrates this process. The posterior collision estimate for each zone is interpreted by a fitted distribution, and the normal collision estimate for each zone is obtained by averaging the possible values relative to all posterior parameter distributions. This identification criteria is mathematically presented as

$$\lambda_{i(5\%)} > E(y_i) \tag{4.1}$$

where $\lambda_{i(5\%)}$ is the 5% posterior value of collisions for zone i, obtained from WinBUGS simulation results; and $E(y_i)$ is the normal collision estimate for zone i, which is written as

$$E(y_i) = e^{(\ln E(\Lambda_i)+0.5\sigma^2)} = E(\Lambda_i)e^{0.5\sigma^2} = a_0 TLKM^{a_1} e^{\sum b_j X_j + 0.5\sigma^2} \tag{3.3}$$

All parameters used (i.e. a_0, a_1, b_j, and σ^2) in this equation are the means of posterior parameter estimates simulated from WinBUGS, as shown in Table 3.8(c). Table 4.2 shows several zones' 5% posterior collision estimates, normal collision estimates, and CPZ identification results.

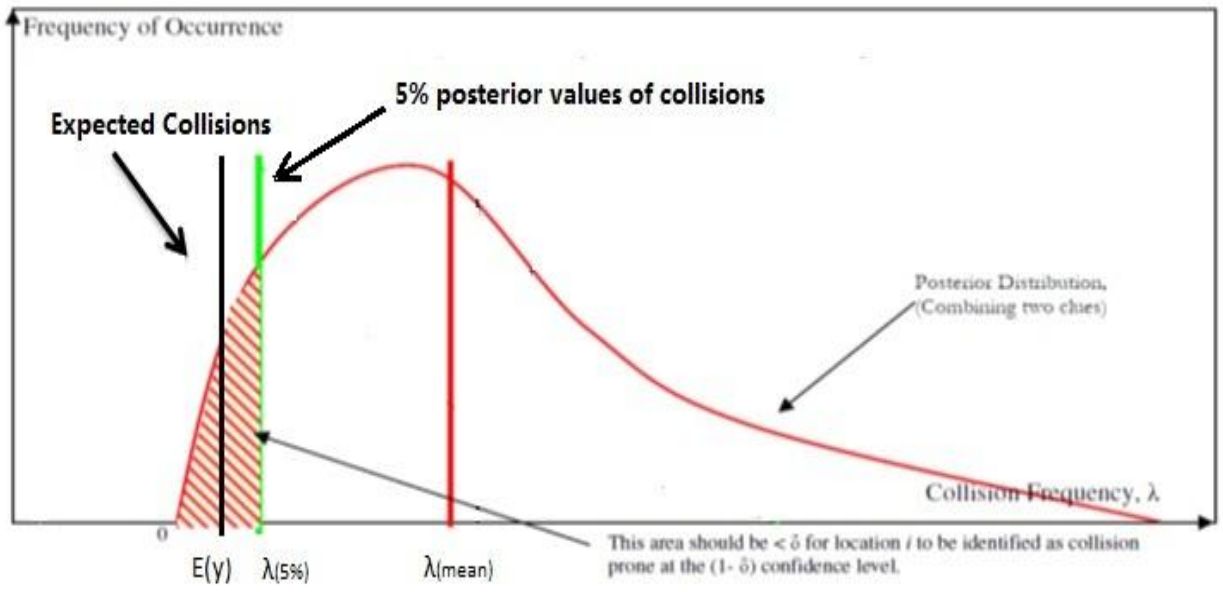

Figure 4.3 Full Bayes Identification of CPZs

Table 4.2 Examples on the Identification Process Using FB Exposure Model

Zone No.	$\lambda_{i(5\%)}$ by WinBUGS	$E(y_i)$ by *Eq. 3.3*	CPZ
231	1.406	0.327	Yes
171	1.135	0.426	Yes
253	1.232	0.596	Yes
244	1.137	0.471	Yes
264	0.493	0.503	No
141	0.500	0.540	No

PCR and CRR are also used to rank CPZs and SZs in the FB method. These two ranking measures are formulated as

$$PCR = \lambda_{i(mean)} - E(y_i) \tag{4.2}$$

$$CRR = \frac{\lambda_{i(mean)}}{E(y_i)} \tag{4.3}$$

where $\lambda_{i\ (mean)}$ represents the posterior mean value of collisions for zone *i*, also obtained from simulation results in WinBUGS; and $E(y_i)$ represents the normal collision estimate for zone *i*. Following the above approach, all urban zones in the RDCO are ranked from high to low based on PCR and CRR. The 20 lowest ranked zones are identified as SZs. Similar to the EB method, three FB models yield different identification and ranking results. Therefore, a total ranking score for each zone is calculated by summing the rankings across the three models. The results of CPZ and SZ identification and ranking by the three models are presented in Table 4.3. Figure 4.4 and Figure 4.5 individually display the identified CPZs and SZs using the FB approach.

Table 4.3 Urban Bike-vehicle CPZ and SZ Identification Using FB Method

Bike CPZs	Exposure	TDM	NW	Bike SZs	Exposure	TDM	NW
1	231	231	244	**1**	449	37	489
2	171	171		**2**	376	128	490
3	253	244		**3**	305	449	480
4	244	253		**4**	37	245	469
5	151	179		**5**	447	265	465
6	179	307		**6**	431	447	458
7	483	151		**7**	49	480	491
8	307	141		**8**	257	305	476
9	266	132		**9**	216	116	457
10	132	177		**10**	182	257	431
11	177			**11**	297	457	378
12	459			**12**	245	376	473
13	124			**13**	128	276	456
14	127			**14**	301	49	358
15				**15**	116	252	447
16				**16**	252	131	301
17				**17**	50	155	368
18				**18**	87	220	449
19				**19**	135	301	486
20				**20**	118	259	460

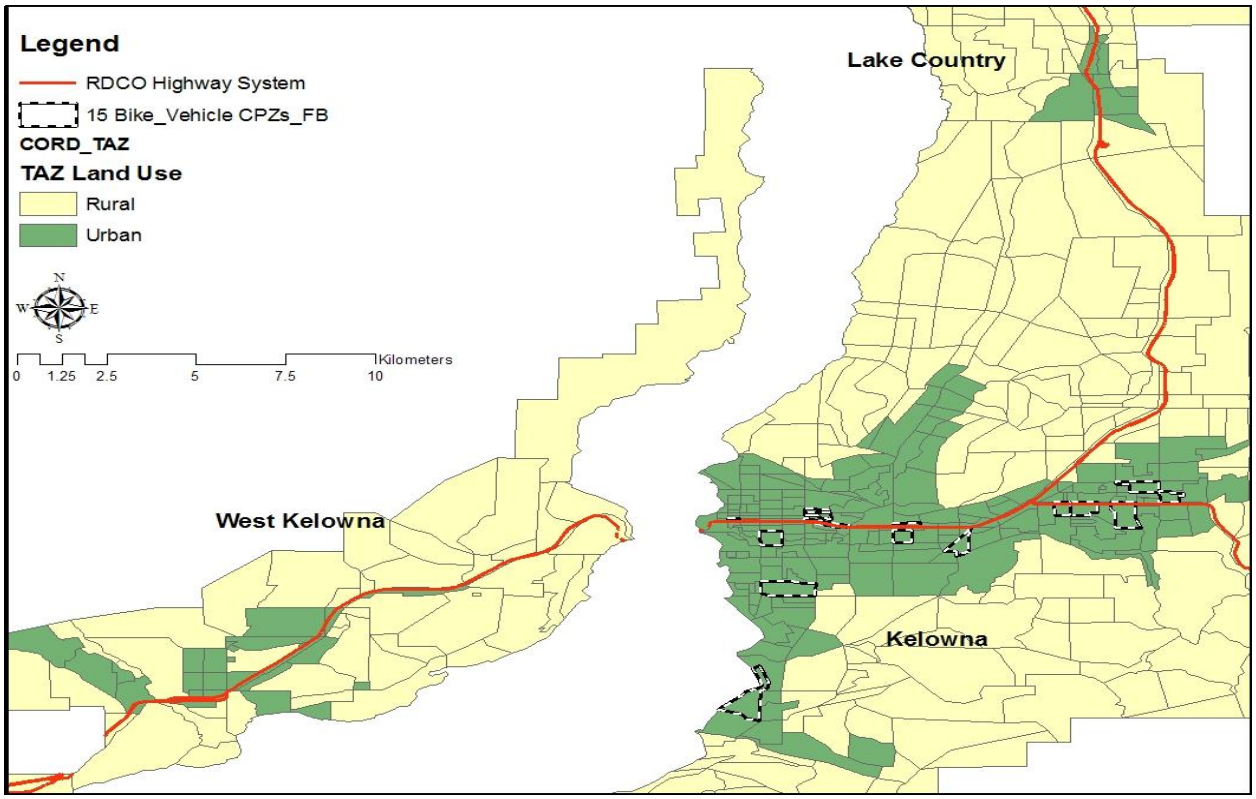

Figure 4.4 Bike-vehicle CPZs Using the FB Method

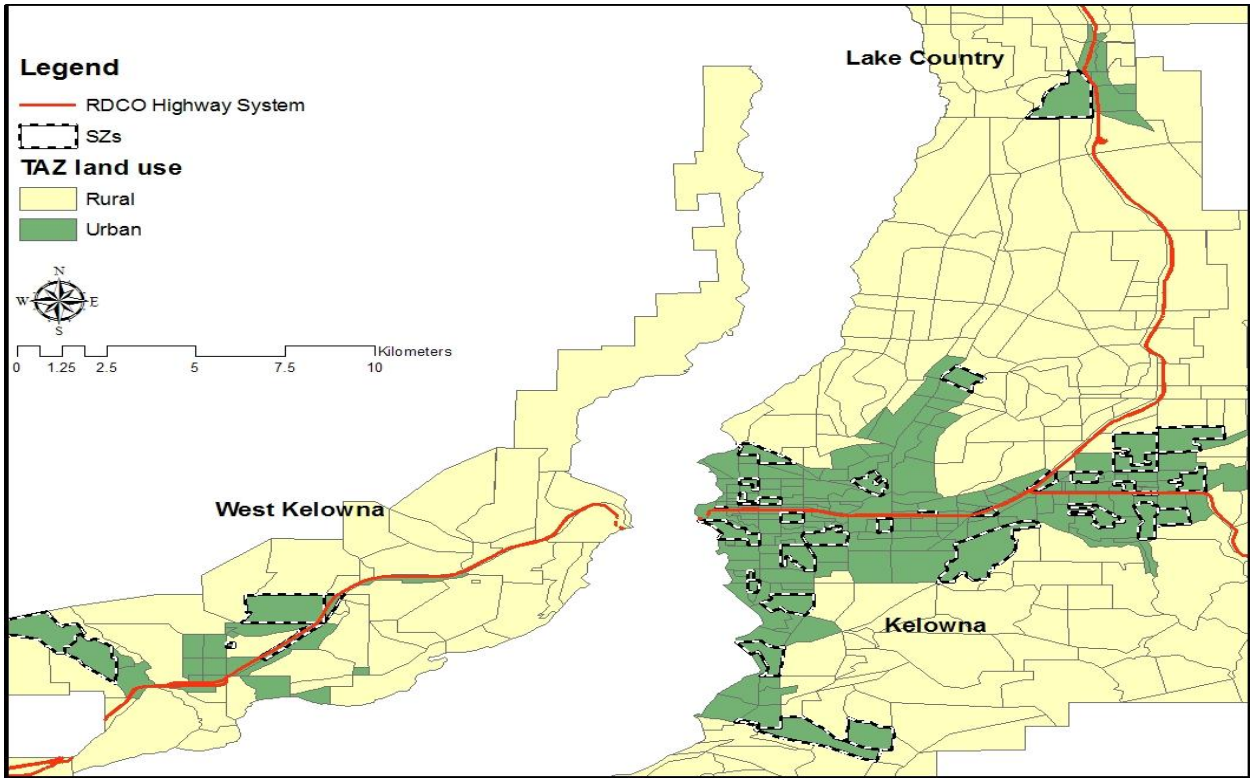

Figure 4.5 Bike-vehicle SZs Using the FB Method

Several CPZ and SZ identification and ranking results are observed. First, a total of 15 CPZs are identified using the FB method. Of the 15, 14 are identified by the model in the exposure group, 10 are identified by the model in the TDM group, and only 1 is identified by the model in the NW group. Second, the average CPZ identification similarity among these three models is up to 85%. The models from the exposure and TDM groups largely agree with each other in identifying and ranking CPZs. Third, a total of 41 SZs are identified. However, the SZ identification results from the three models have big differences. For example, only three same SZs (i.e. Zone 301, 449, 447) are identified by all models, nine same SZs are identified by both exposure and TDM models, and two same SZs are identified by both TDM and NW models. The CPZ and SZ identification and ranking results between the exposure and TDM models are closer to each other than to the NW model. This result is consistent with that from the EB method.

4.1.3 CPZ and SZ Identification Comparison

In this section, the CPZ identification results using the EB and FB methods are compared. In total, 16 CPZs and 15 CPZs were identified using the EB and FB methods respectively. The same 15 CPZs are found in both methods, showing a similarity rate of 97%. This result indicates that the FB and EB methods agree with each other in CPZ identification. Furthermore, the CPZ identification results using these two methods are compared in pairs according to model groups. In the exposure model group, the CPZ identification results between the EB and FB methods are exactly same. In the TDM model group, all of the 10 CPZs derived from the FB method are also identified by the EB method. In the NW model group, 7 zones are identified as CPZs by the EB method; however, only 1 zone is identified as a CPZ by the FB method. Thus, the CPZ identification similarity in the NW model group between these two methods is very low. A possible reason for this may be the data quality issue.

The SZ identification results obtained using the EB and FB methods are also compared. The total identified SZs from these two methods show the SZ similarity rate is 76%. Like CPZ identification, the SZ identifications resulting from the two methods are compared in pairs in each model group. In the exposure model group, the SZ similarity rate between the EB and FB methods is up to 100%. In the TDM model group, the SZ similarity rate between the two

methods is 90%. In the NW model group, the similarity rate is only 40%. Generally, the CPZ and SZ identification results from the EB and FB methods are regarded as roughly consistent.

4.2 Diagnosis and Remedy

After CPZ identification and ranking, diagnoses and remedies for these CPZs are implemented. In-office and on-site analyses are involved in the diagnoses. Over-represented collision patterns and trigger variables are suggested by Lovegrove (2007) as two indicators for in-office diagnoses. The description claims of bike-vehicle collisions in the ICBC database provide information for over-represented collision patterns. The trigger variables of any CPZs are defined as model independent variables whose values are apparently lower or higher than the regional average values. Trigger variables are examined to confirm if they are triggering the collision prone identification. In addition to the in-office analyses, on-site analyses are conducted to verify in-office findings and to supplement other undiscovered causes for safety problems in CPZs (e.g. speeding, vehicle volume, and traffic calming levels). In the identified bike-vehicle CPZs, intersections are usually the hazardous locations for cyclists; therefore, the intersections where bike-vehicle collisions happened were visited. Lovegrove (2007) suggested that trigger variables together with collision patterns and site visits could provide evidence for identifying the overall road safety problems and suitable countermeasures in CPZs.

The diagnosis analyses are followed by strategic remedy analyses at a community-level. Potential countermeasures in this study are categorized according to the themes of traffic exposure, TDM, and road network. In addition to the strategic level of remedy analyses, detailed remedy analyses are also proposed to improve the safety performances at individual intersections. Several countermeasures to improve safety of individual intersections are summarized in Table 4.4. The detailed diagnoses and remedies for 16 bike-vehicle CPZs are described as following.

Table 4.4 Summaries of Remedies for Bike-vehicle CPZs

Remedy	Description	Photo	Applied CPZs
Bike Refuge (TDM)	To provide bike access to cross streets where a median continues through the intersection, usually for use at the intersection of an off-street bike path and a high traffic roadway (TAC, 2010).		231,244
Raised Intersection (TDM)	Often in concert with curb bulges, raised crosswalks provide a vertical visual and sensory cue to slow down and be alert (Lovegrove et al., 2012).		231,244
Elephant Feet (TDM)	Elephants' feet (aka crossbikes) are a novel paint marking that is usually applied on beside and parallel to a pedestrian crosswalk, but meant to indicate an on-street crossing corridor for bicycles (Lovegrove et al., 2012).		307
Warning Signage (TDM)	An activated warning sign set on a narrow high-speed stretch of roads that cautions drivers to watch out for bicyclists on the road (Lovegrove et al., 2012).		179,151, 132, 253, 171, 177, 127,266,
Bike Box (TDM)	An area where cyclists may go ahead of motor vehicles at a red signal to get into position for turning or going straight before other vehicle traffic when the signal turns green (Lovegrove et al., 2012).	 (Bike box in Portland)1	307,141
Bike Signals (NW)	A separate phase to allow bicycles to cross high speed/high volume vehicle traffic corridors (Lovegrove et al., 2012).		483,124, 459
Bike paths (NW)	A path segregated from motorized traffic lanes for the use of bikes, sometimes shared with pedestrians (Lovegrove et al., 2012)		179,151, 132, 307, 127, 266, 483,141
Colored Bike Lanes (NW)	Colored bike lanes are usually installed across high conflict vehicle-bicycle crossing zones to caution drivers that bicycles may be present and to look for them (Lovegrove et al., 2012).		231,244, 253, 171, 177, 266, 124,

Note: [1] Photo Courtesy: City of Kelowna

Zone 231 and 244

Zone 231 and 244 are two adjacent zones (see Figure 4.6). They were identified as CPZs by the NB and FB models in the exposure, TDM, and NW groups. Looking at the land use and road map, together with trigger variables and collision descriptions, provided four clues to possible road safety problems in the two CPZs. First, land use of the two zones consists of residential homes in a larger residential area surrounded by commercial, government, and recreational areas. Zone 231 and zone 244 are divided by Bernard Ave, which is a connector road with medium/high traffic. Each zone consists of two blocks. Second, the roads in these two zones are set in the midst of a discontinuous and offset mixed road network pattern. Also, all of the roads, at the time of this study, had no "marked" bike lanes. Third, the intersection density was a trigger variable with an abnormal "high" value in both zones. Fourth, five micro-level black spots for bike-vehicle collisions were identified in the two zones, including intersections of Gordon Dr. @ Bernard Ave., D'Anjou St. @ Bernard Ave., Richmond St. @ Bernad Ave., Gordon Dr. @ Lawrence Ave., and Centennial Cres St. @ Lawrence Ave. The collision descriptions from the ICBC database show that most bicycle collisions happened when the motor vehicles or bicycles were making turning movements. These four clues suggest that the safety problem might have been due to an unsafe road environment for cyclists and high traffic volume on the perimeter connector roads (i.e. Gordon Rd. and Bernard Ave.). Two possible remedies in TDM and NW themes are suggested to solve these safety problems.

- TDM – Increase traffic calming strategies in the community where the two CPZs are located, to increase drivers' attention to cyclists (e.g. bike refuges on Gordon, raised intersections).
- Network – As there were no "marked" bike lane for cyclists in the two zones, build friendly and conspicuous bicycle facilities (e.g. coloured bike lanes).

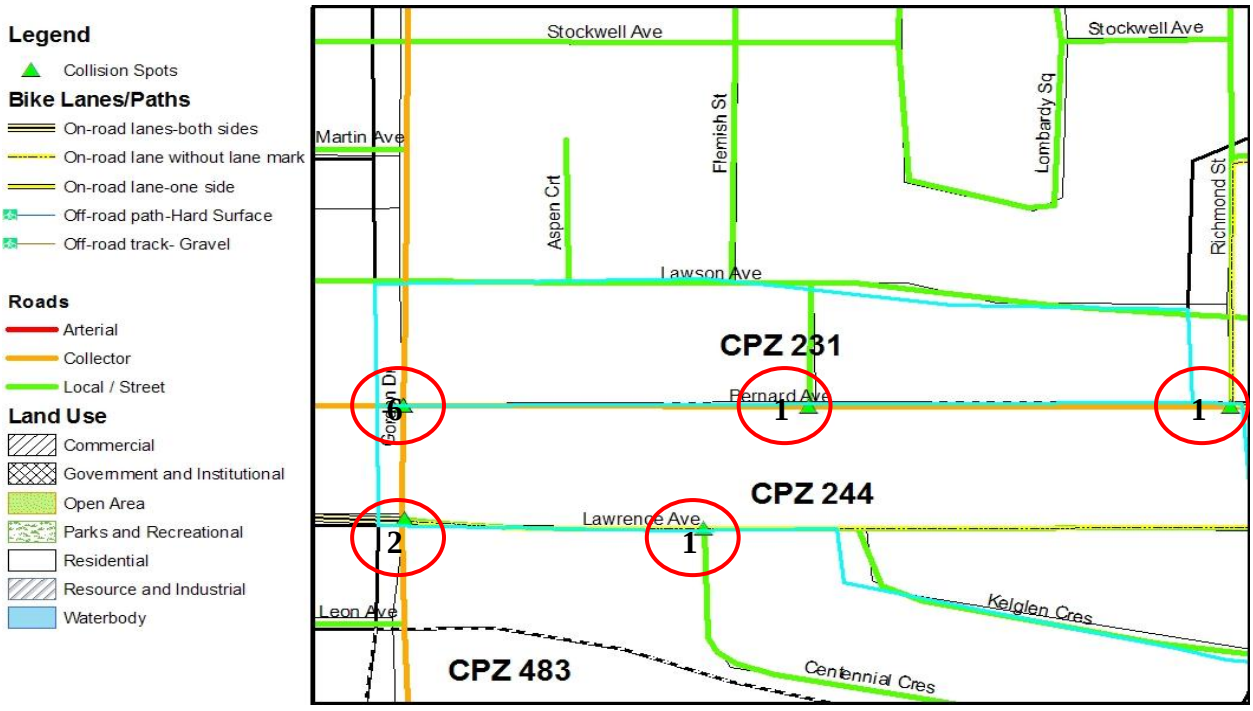

Figure 4.6 Urban Bike CPZs – Zone 231&244

Zone 179

Zone 179 was identified as a CPZ by all NB models and two FB models from the exposure and TDM model groups (see Figure 4.7). Land use, road maps, trigger variables, and collision descriptions provide four clues to possible road safety problems. First, this zone is composed of commercial and industrial area with large parking lots. The Glenmore community, which is one of the largest residential communities in Kelowna, is on the north of this zone and the Parkinson Recreation Park is on the northwest of this zone. Second, in this zone, the arterial road Spall Rd. crosses Highway 97 and the local road Ambrosi Rd. connects the Highway 97 from the south to form a T intersection. Both Spall Rd. and Highway 97 have high traffic volumes, but only the Spall Rd section on the north side of Highway 97 has on-road bike lanes. Third, at the time of this study, zonal signals (high) and the percentage of arterial-local intersections (high) were trigger variables in this zone. Fourth, two black spots, the intersections of Highway 97 @ Spall Rd. (signalized) and Highway 97 @ Ambrosi Rd., were identified. The collision descriptions show that within the four collisions in this zone, two of them happened when motor vehicles were pulling out or getting into the two gas stations located at the two corners of the intersection of Highway 97 @ Spall Rd. These clues suggest that the safety problem might be associated

with the high traffic volume and high speeds along Highway 97. Two possible remedies are suggested to solve this safety problem.

- TDM – Put warning signage (i.e. look left) to increase drivers' attentions at entries and exits of the industrial area in this zone.
- Network – Build off-road bike paths or painted bike lanes along Spall Rd and Highway 97. Design other bike networks to reduce bicycle use along Highway 97.

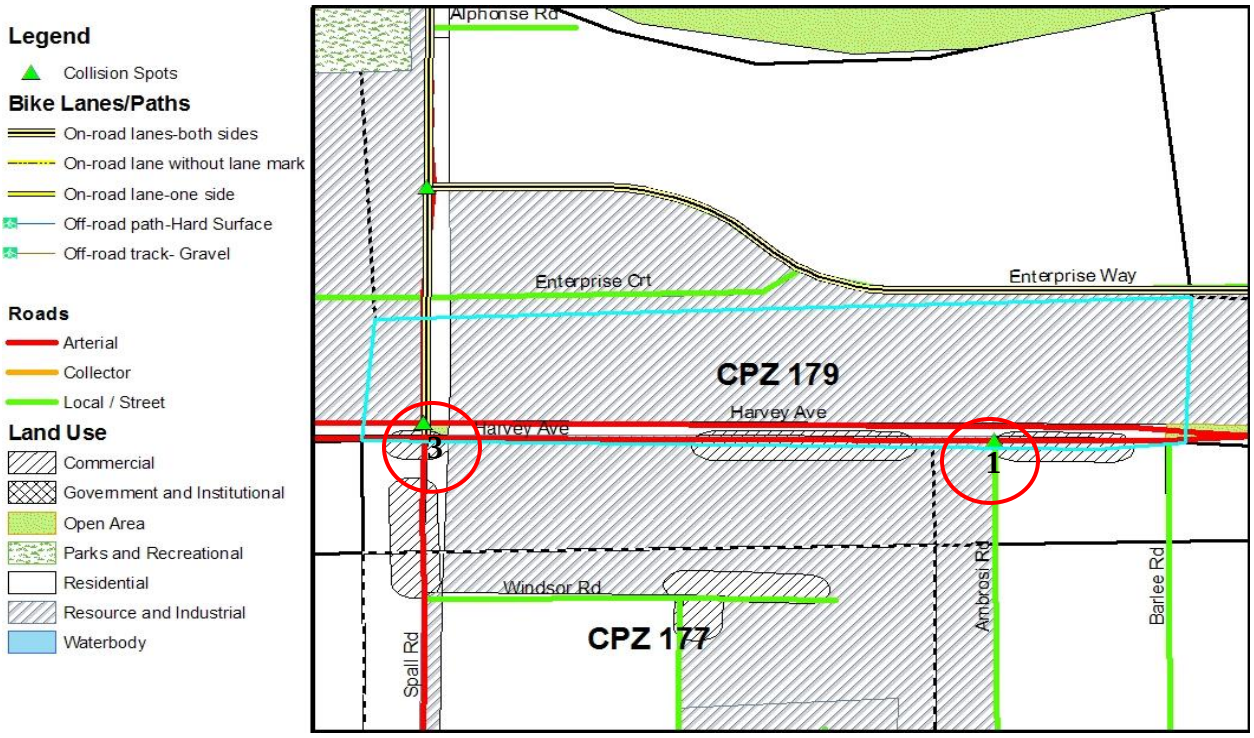

Figure 4.7 Urban Bike CPZ – Zone 179

Zone 151 and 132

Zone 151 and 132 are two adjacent zones (see Figure 4.8). They were identified as bike-vehicle CPZs by the three NB models and the two FB models from the exposure and TDM groups. Four clues are provided as follows to reveal the safety problem. First, these two zones are in a larger residential area surrounded by commercial and park areas. Second, all roads in these two zones are set in a limited access (i.e. discontinuous and offset mixed road pattern) road network pattern. Additionally, there were, at the time of this study, no "marked" bike lanes in these two zones. Third, the percentage of arterial-local intersections (high) was the trigger variable in both zones. Fourth, the hazardous intersections were all along Highway 33, including

Highway 33 @ Dave Rd., Highway 33 @ Gerstmar Rd. (signalized), Highway 33 @ Nickel Rd., and Highway 33 @ Taylor Rd. The collision descriptions show that most bike-vehicle collisions happened when cyclists were going straight in order to cross intersections along Highway 33 and motor vehicle drivers were turning from minor roads to Highway 33. These clues suggest the road safety problem might be an unsafe road environment on Highway 33, where the high traffic volume and high traffic speed have created high risks for cyclists. Therefore, two recommended remedies are proposed:

- TDM – Restrict left-turns and shortcuts at arterial-local intersections along Highway 33, or use signage to educate cyclists on how to share roads with motor vehicle drivers.
- Network – Prohibit cycling along Highway 33 but provide other safer access for cyclists (e.g. build safer bike lanes/paths on other roads parallel to Highway 33 to replace the sidewalks along Highway 33 now used for cycling).

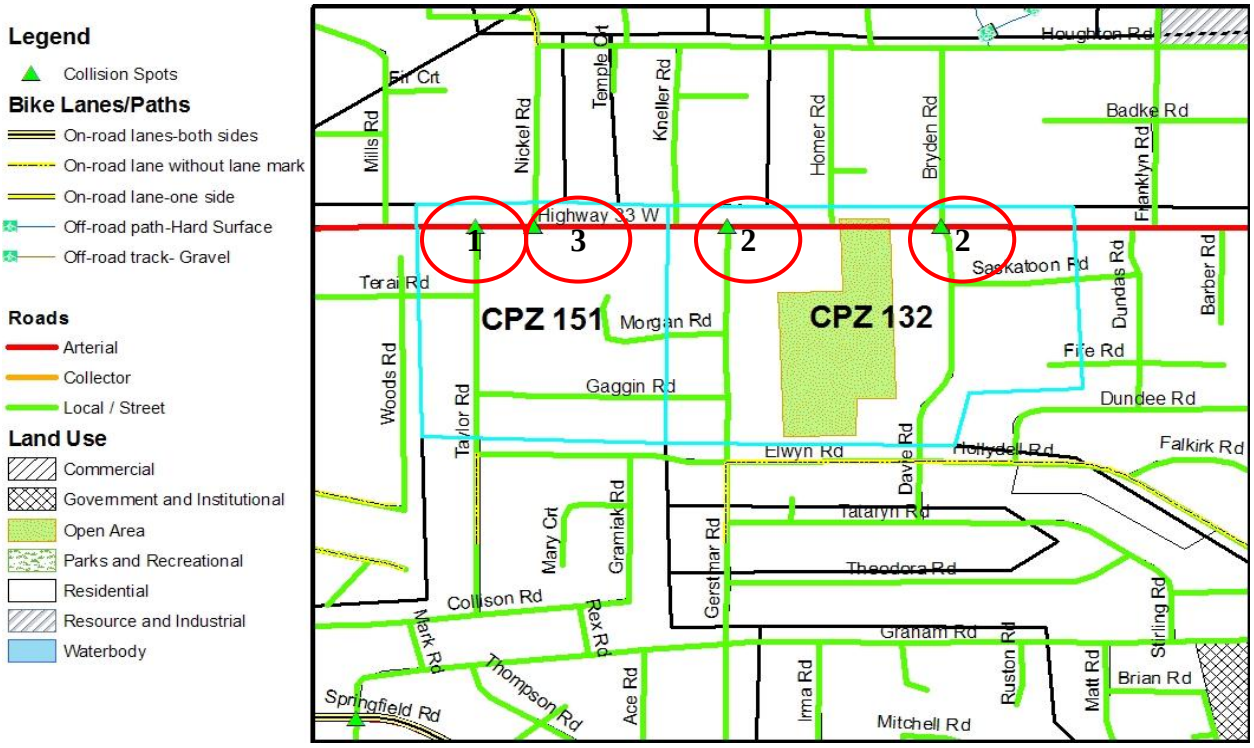

Figure 4.8 Urban Bike CPZs – Zone 151 & 132

Zone 253

Zone 253 (see Figure 4.9) was identified as a CPZ by NB models and FB models from the exposure and TDM model groups. Three clues are provided to reveal the road safety problem.

First, this zone is a school community (e.g. Okanagan College is located here) surrounded by a residential and commercial area, so internal connectivity of motor traffic is precluded in the school community. However, the zone boundaries are composed of arterial roads (e.g. K.L.O. Rd.) and connector roads (e.g. Richter Rd) with high motor traffic. Second, at the time of this study, the internal roads of this zone had no marked bike lanes, while all perimeter roads of this zone had bike lanes for both sides. Third, three black spots in this zone were the intersections of K.L.O. Rd. @ Richter St. (signalized), K.L.O. Rd. @ Casorso Rd., and K.L.O. Rd. @ De Montreuil Ct. Rd.. The collision descriptions reveal that most collisions occurred when motor vehicles were making left turns and bikes were going through or making turning movements. These clues suggest that the safety problem might be due to the high volume of motor vehicles and bicycles along K.L.O. Rd. and Richter Rd. Possible remedies to solve the program are

- TDM – Use signage to educate cyclists and motorists on how to share roads with motor vehicles, and build a bike box at K.L.O @ Richter.
- Network – Paint coloured bike lanes close to/across from the two intersections along K.L.O. Rd.

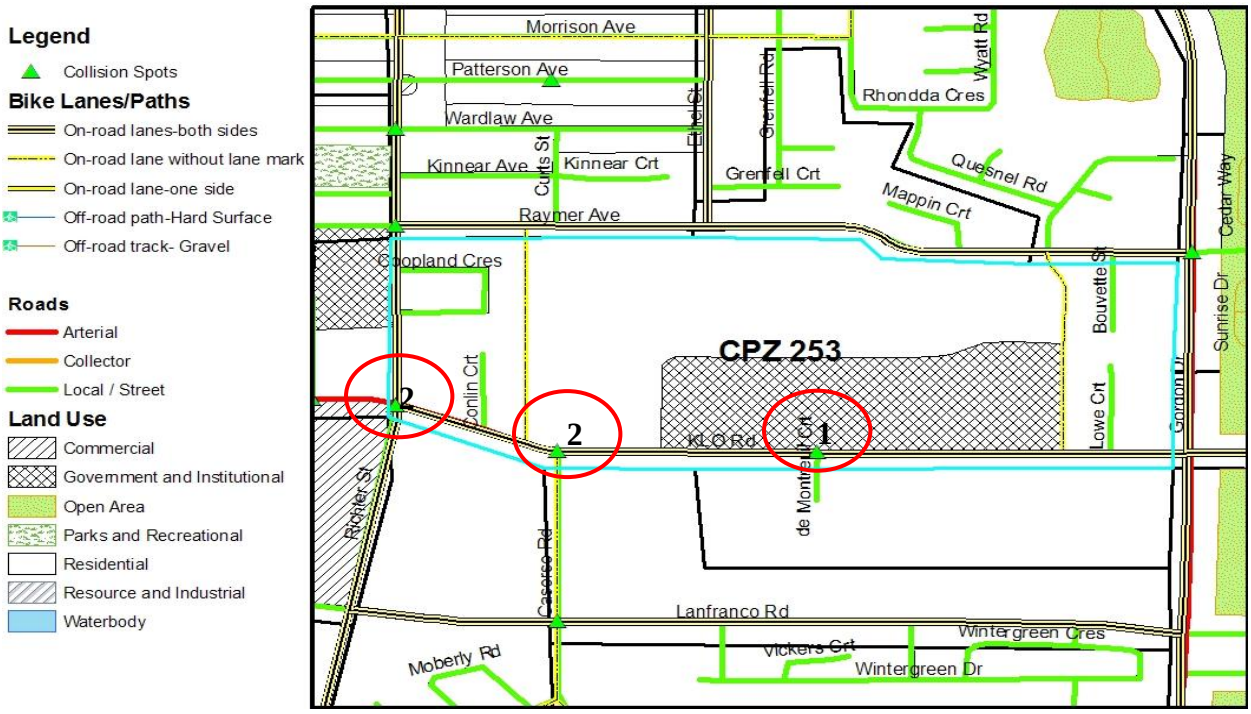

Figure 4.9 Urban Bike CPZ – Zone 253

Zone 171

Zone 171 was identified by the NB and FB models both from the exposure and TDM groups (see Figure 4.10). There are three clues to suggest possible road safety problems. First, a corner of 'Orchard Park', the biggest commercial center in Kelowna, is located in this zone. The remaining area in this zone is open. Second, at the time of this study, the connector road Dilworth Dr. and the arterial road Springfield Rd. individually had marked bike lanes. Benvoulin Rd. obliquely crossed Springfield Rd with a local pedestrian signal. Third, the black spots in this CPZ included the intersections of Springfield Rd. @ Dilworth Rd. (fully signalized), Springfield Rd. @ Benvoulin Rd. (pedestrian actuated signalized), and Dilworth Rd. @ Mayer Rd. All bike-vehicle collisions were caused when motor vehicles were making turning movement (both right and left) at intersections. These clues suggest that the safety problems might be the high traffic volume generated from or passing by Orchard Park, and a lack of safe intersection design for cyclists.

- TDM – Put signage to warn cyclists and motor vehicle drivers how to share roads safely.
- Network – Paint coloured bike lanes close to/across from these three intersections to make them conspicuous.

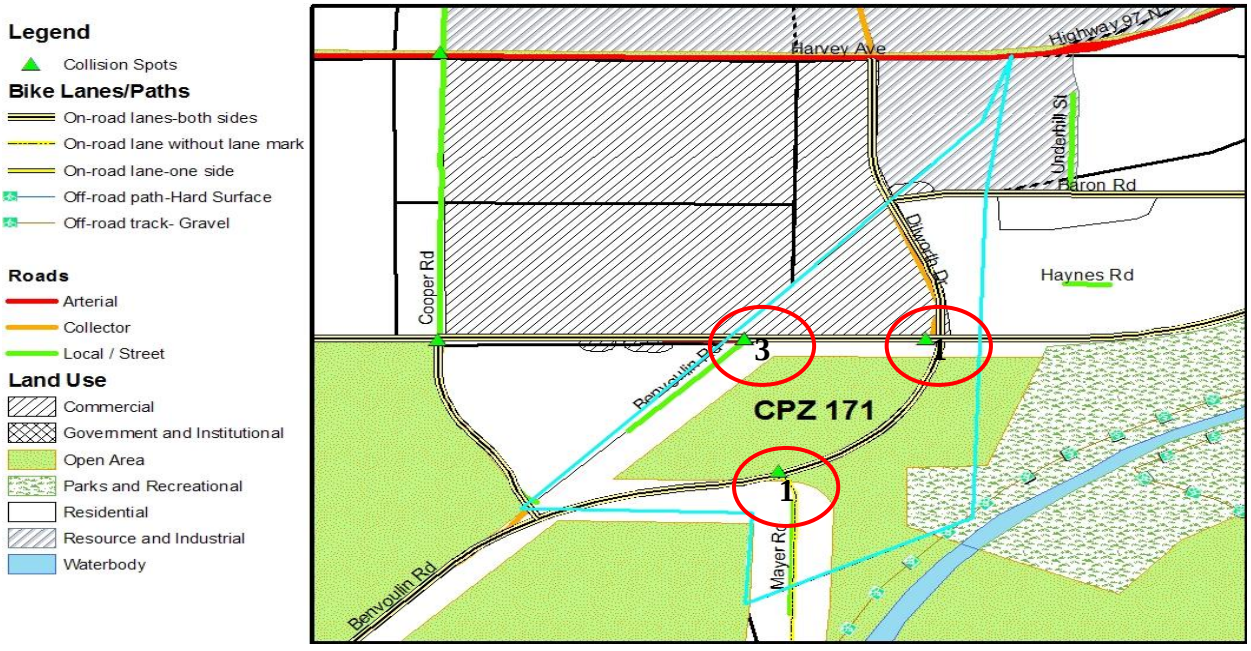

Figure 4.10 Urban Bike CPZ – Zone 171

Zone 307

Zone 307 (see Figure 4.11) was identified as a CPZ by NB and FB models in the exposure and TDM model groups. Four clues are presented as follows to show potential safety problems. First, Zone 307 is a residential zone in Lower Mission, an old residential neighbourhood in Kelowna. This zone is on the east side of Okanagan Lake. It has a high bicycle volume around because this lakeside area attracts people to come for recreation and relaxation. Second, at the time of this study, only the arterial road Lakeshore Rd. had marked bike lanes, while the local roads in this zone did not have marked bike lanes. Third, total lane kilometers (high) was a trigger variable in this zone. Finally, the intersections of Lakeshore Rd. @ Dehart Rd. (signalized), Lakeshore Rd. @ Greene Rd., and Hobson Rd. and Sarsons Rd. were identified as three black spots. Most collisions happened at the intersection of Lakeshore Rd. @ Dehart Rd when motor vehicles were making left turns. These road environmental and geographic clues suggest that the safety problem might have been due to a high number of cyclists crossing intersections. Two remedies in the aspects of TDM and road network are recommended:

- TDM – Increase friendly bicycle facilities to create a safer environment for cyclists when they cross intersections (e.g. raised intersection, elephant's feet), or set up warning signs for motorists and cyclists.
- Network – Build off-road bike paths in order to meet the demand of a relative high level of bicycle use in this zone.

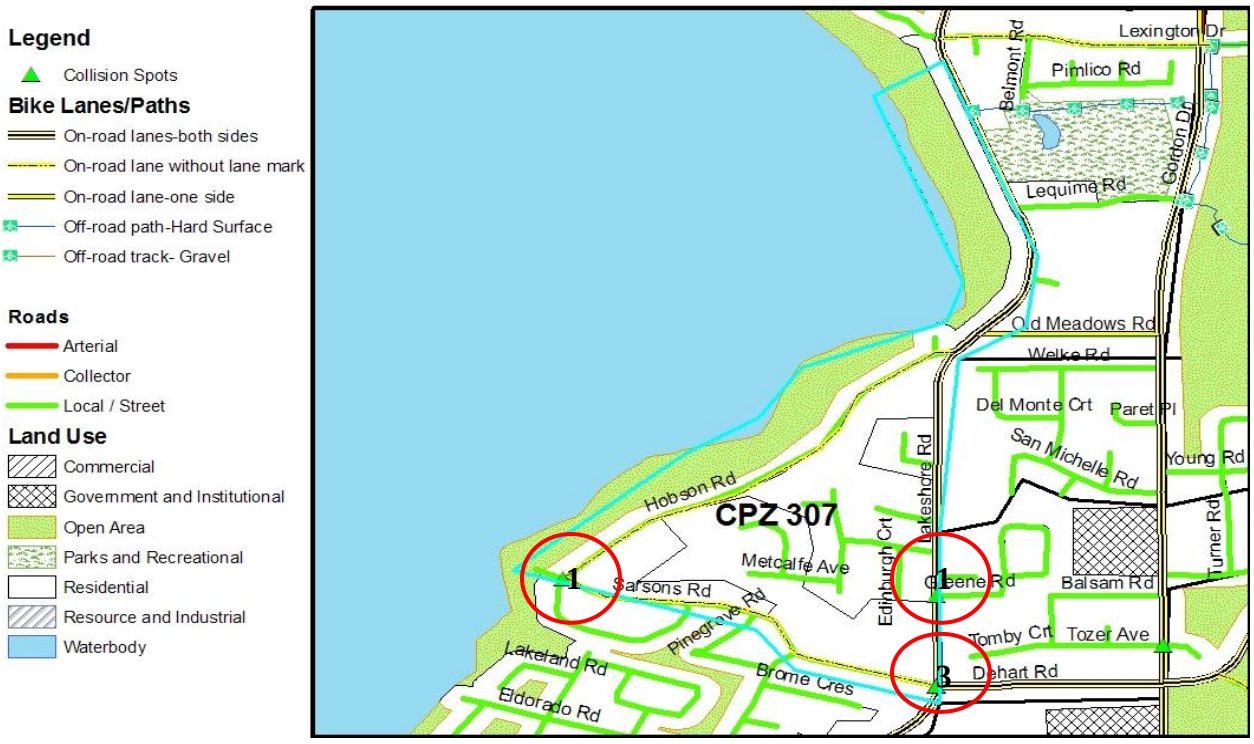

Figure 4.11 Urban Bike CPZ – Zone 307

Zone 177

Zone 177 was identified as a CPZ by NB and FB models in the exposure and TDM model group (see Figure 4.12). Four clues are suggested to reveal possible road safety problems. First, this zone is an area with mixed industrial, commercial, and residential land use, and is surrounded by open, recreation, and industrial areas. Second, this zone has arterial and local roads and all of these roads are set in the midst of a discontinuous road grid pattern. At the time of this study, only the arterial road Springfield Rd. had marked bike lanes. Third, the variable total lane kilometres (high) was a trigger variable in this zone. Fourth, of the four collisions in this zone, three happened at intersections and one happened in a mid-block. All the collisions occurred at intersections were caused when bikes or motorists were making turning movements. These clues suggest that the safety problems might have been due to the high traffic volume and high speed on Springfield Road. Two possible remedies are:

- TDM – Put signage at Springfield Rd. @ Bredin Rd. to remind motor vehicle drivers pay attentions for cyclists.
- Network – Paint coloured bike lanes on Springfield Rd.

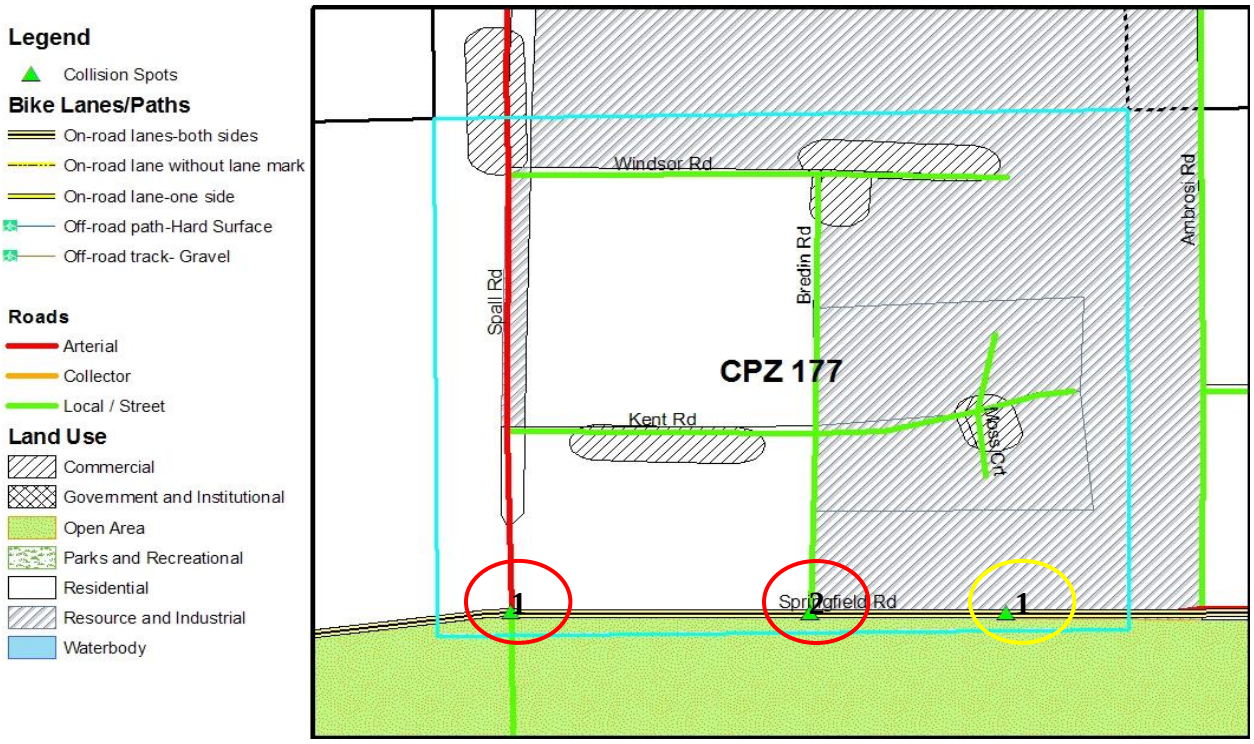

Figure 4.12 Urban Bike CPZ – Zone 177

Zone 127

Zone 127 (see Figure 4.13) was identified as a CPZ by two NB models in the exposure and NW groups, and one FB model in the exposure group. Through the in-office and on-site analyses, this zone had a similar road safety problem to Zone 151 and 132. Therefore, similar remedies to Zone 151 and 132 are suggested as follows:

- TDM – Restrict left-turns and shortcuts at arterial-local intersections, or use signage to educate cyclists on how to share roads with motor vehicle drivers.
- Network – Prohibit cycling along Highway 33 but provide other safer access for cyclists (e.g. build safer bike lanes/paths on the roads parallel to Highway 33 to replace the sidewalks along Highway 33 now used for cycling).

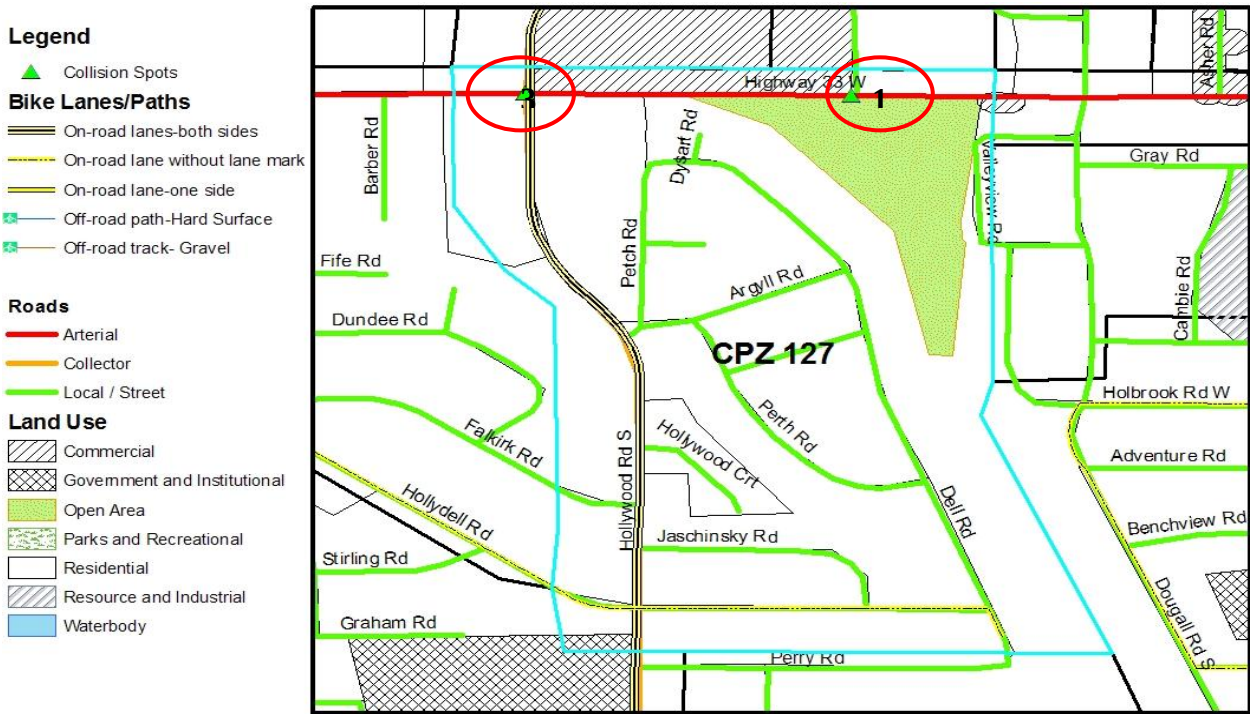

Figure 4.13 Urban Bike CPZ – Zone 127

Zone 266

Zone 266 (see Figure 4.14) was identified as a CPZ by NB models in the exposure and TDM groups, and the FB model in the exposure group. Four clues are provided for road safety problems in this zone. First, this is a small residential zone with mixed commercial, institutional, industrial, and recreation land use. Second, all roads in this zone are set in a discontinuous road grid pattern. At the time of this study, only Richter St. had on-road bike lanes (bike lanes on Cadder Ave. were just painted recently). Third, the percentage of driving commuters (low) was the trigger variable in Zone 266. Fourth, three black spots in this zone included the intersections of Richter St. @ Sutherland Ave. (signalized), Richter St. @ Elliot Ave., and Richter St. @ Cadder Ave. (signalized). These road environmental and geographic clues suggest that the safety problems might have been due to the high traffic volume on Richter St. and the high bicycle volume at intersections. Possible remedies are listed below.

- TDM – Increase traffic calming strategies (e.g. traffic signage at intersections).
- NW – Paint coloured bike lanes on Richter Rd., or build bike lanes/paths to increase internal connectivity and reduce arterial accesses of cyclists.

.

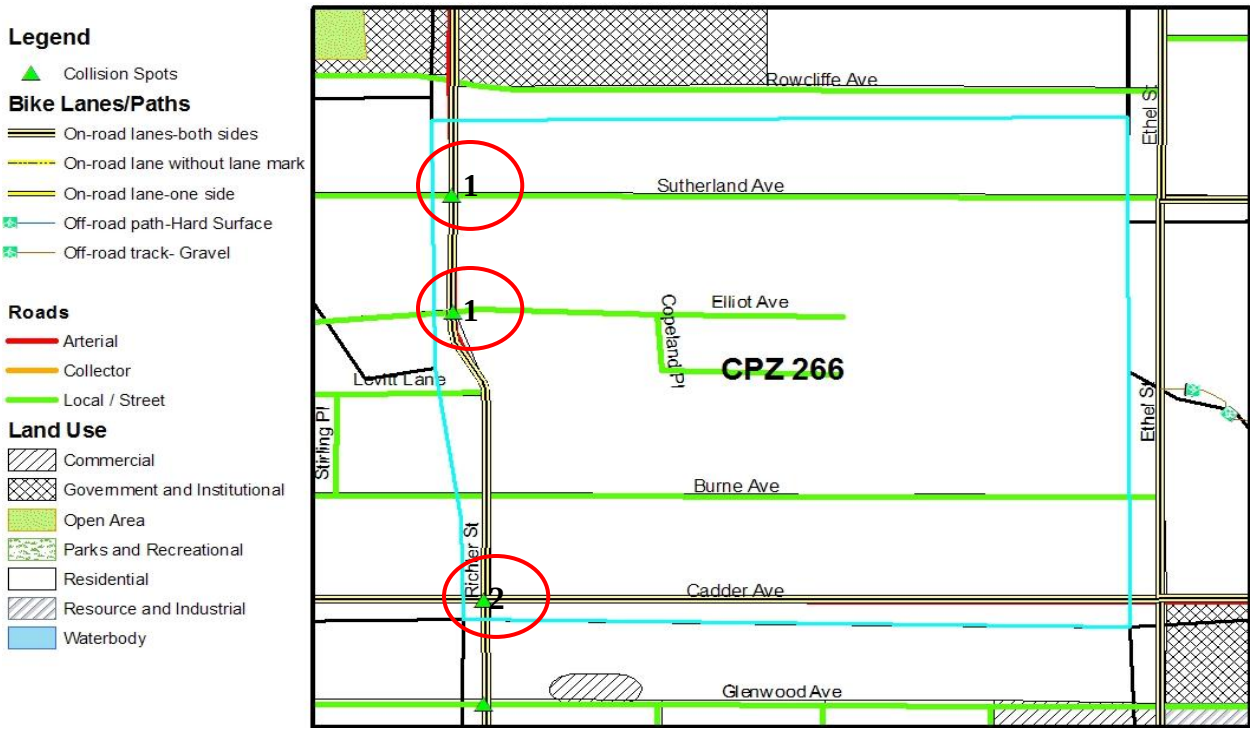

Figure 4.14 Urban Bike CPZ – Zone 266

Zone 483

Zone 483 (see Figure 4.15) was identified as a CPZ by the NB model and the FB model both from the exposure group. Three clues for the road safety problem are provided below. First, this zone is close to Kelowna downtown and consists of commercial and residential areas. Second, this zone is along Harvey Ave. (i.e. Highway 97) and had no marked bike lanes at the time of this study. Third, two black spots were the intersections of Harvey 97 @ Gordon Rd. (signalized) and Harvey 97 @ Capri Rd. All bike-vehicle collisions happened at intersections were caused when motor vehicles were turning into Harvey Ave. from minor roads. These clues suggest that the safety problem might have been due to the high traffic volume on Harvey Ave. and the high bicycle volume in this zone. The following remedies are suggested to solve the safety problem.

- TDM – Build separated signal phases for bicycles, set warning signage, or build a bike box at Harvey Ave. @ Gordon Rd.
- NW – Build safer bike lanes/paths on the roads parallel to Harvey Ave.

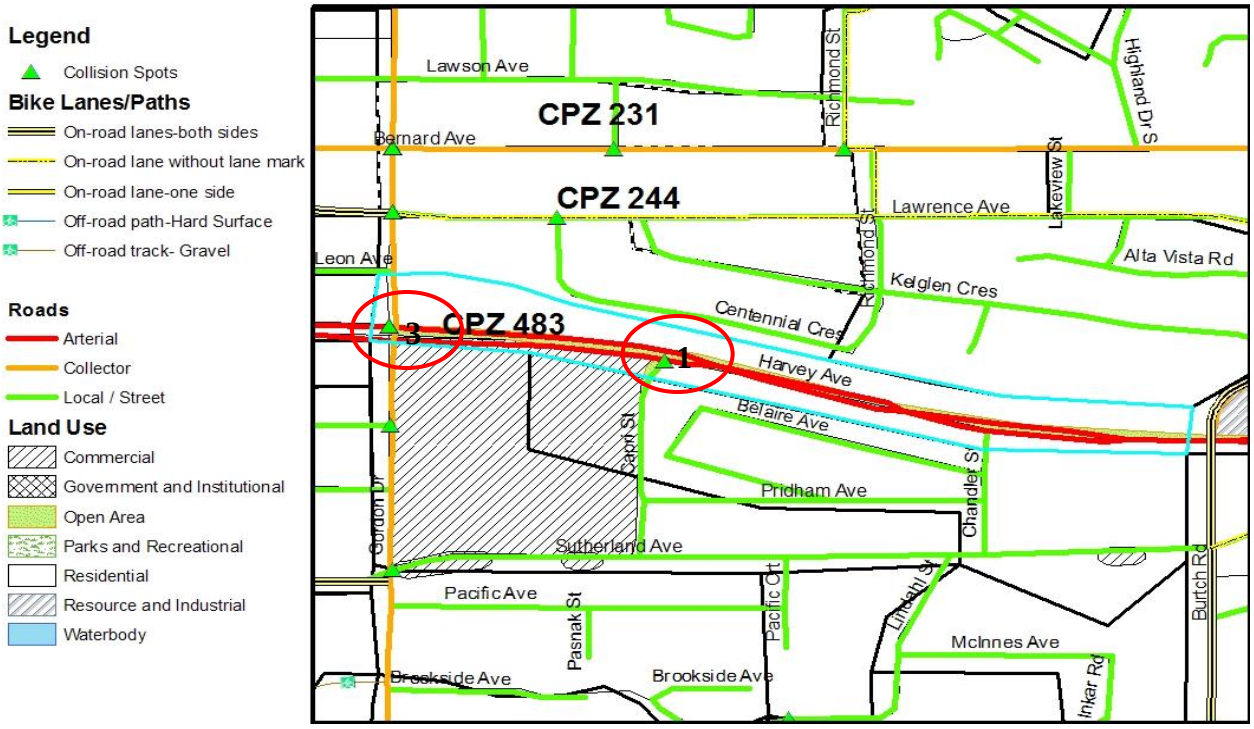

Figure 4.15 Urban Bike CPZ – Zone 483

Zone 124

Zone 124 (see Figure 4.16) was identified as a CPZ by the NB model and the FB model in the exposure group. Four clues are provided to suggest possible road safety problems. First, this zone is a residential area surrounded by recreational (e.g. Ben Lee Park and Rutland Centennial Park) and commercial areas. Second, all roads in this zone are local roads and set in the midst of a discontinuous road grid pattern. At the time of this study, all roads except Leathead Rd did not have marked bike lanes. Third, the total lane kilometers (high) was a trigger variable. Fourth, four black spots in this zone included the intersections along Leathead Rd., Leathead Rd. @ Pinetree Rd., Leathead Rd. @ Dougall N Rd., Leathead Rd. @ Tartan Rd., and Leathead Rd. @ Froelich Rd. According to these clues, the sight limits at these intersections might be a problem to make drivers have difficulties to clearly see cyclists making turning movements from minor roads into Leathead Rd. Possible remedies are:

- TMD – Set warning signage close to these hazardous intersections or build raised intersections to make drivers pay more attentions on potential cyclists. Improve the sight condition along Leathead Rd.
- NW – Paint coloured bike lanes on Leathead Rd.

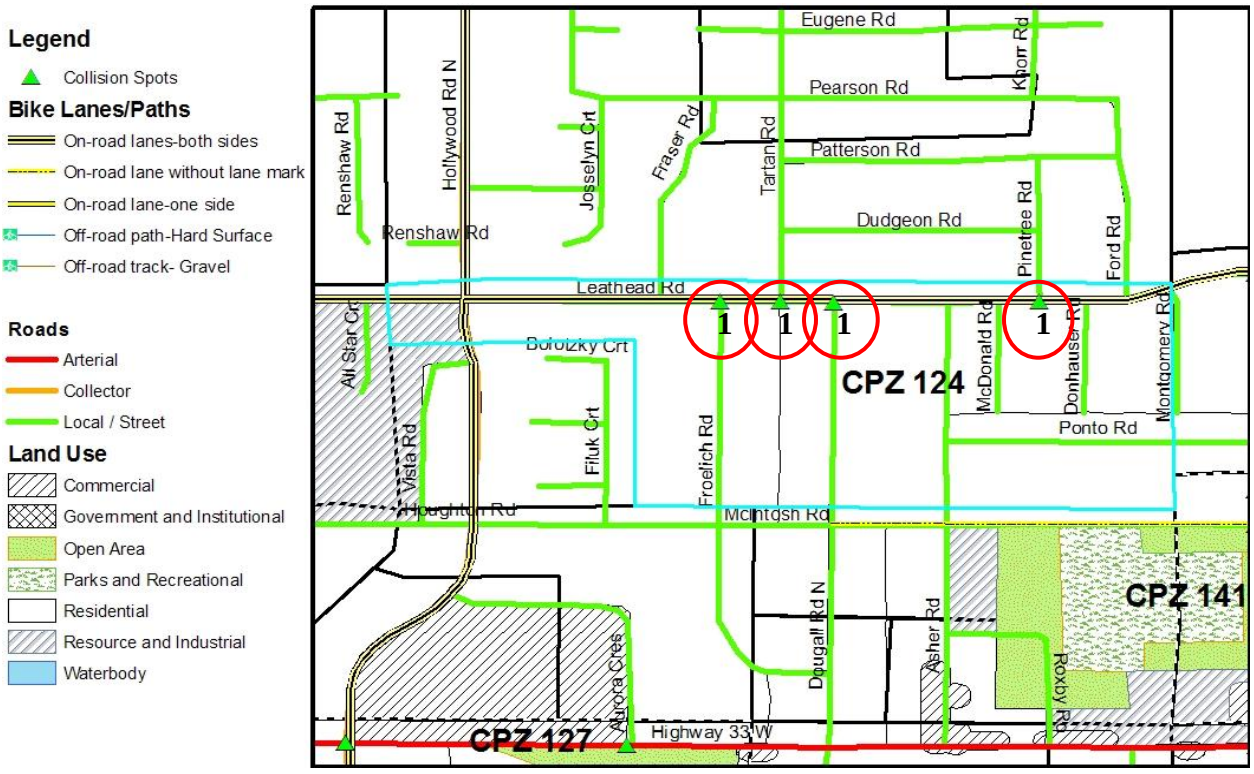

Figure 4.16 Urban Bike CPZ – Zone 124

Zone 459

Zone 459 (see Figure 4.17) was identified as a CPZ by the NB model and the FB model in the exposure group. There are three clues for revealing safety problems. First, this zone is a commercial area at downtown Kelowna. It is located on the east side of the City Park and has Harvey Ave. go through it. Second, all roads in this zone are set in a grid road pattern and had no marked bike lanes at the time of this study. Third, two black spots included Harvey Ave. @ Water St. and Harvey Ave. @ Pandosy St. These clues suggest that the safety problems in this zone might have been due to the high traffic volume on Harvey Ave. and the high bicycle volume in downtown. The following remedies are suggested as follows.

- TDM – Build specific signal phase for bicycles at the two intersections in this zone.
- NW – Restrict cycling access along Harvey Ave. but provide more convenient and safer bike networks.

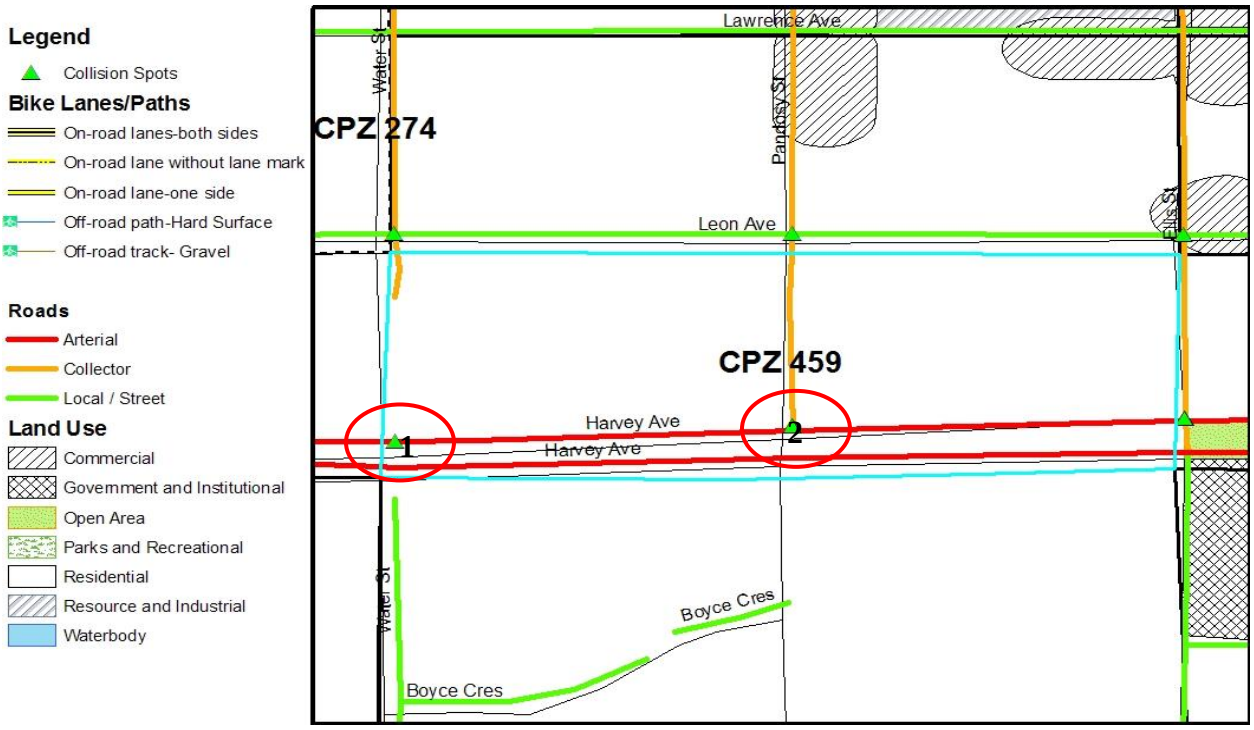

Figure 4.17 Urban Bike CPZ – Zone 459

Zone 141

Zone 141 (see Figure 4.18) was identified as a CPZ by the NB model and the FB model both in the TDM group. Three clues are provided to show possible road safety problems. First, this zone mainly consists of residential and recreational areas. A small piece of commercial area in this zone is along Rutland Rd. Second, all roads in this zone are in the middle of a discontinuous grid road pattern. At the time of this study, only Mugford Rd. had marked bike lanes. Third, two black spots were the intersections of Rutland Rd. @ McIntosh Rd. and Rutland Rd. @ Mugford Rd. The collision descriptions show that collision happened when cyclists were crossing Rutland Rd. As Rutland Middle School is on the north side of this zone, the safety problem might have been due to a relative high bike volume from or to the school. Possible remedies are recommended below.

- TDM – Set pedestrian/bicycle crossing walk on Rutland Rd.
- NW – Set on-road bike lanes or off-road bike paths along Rutland Rd., or set up pedestrian/bicycle actuated signals.

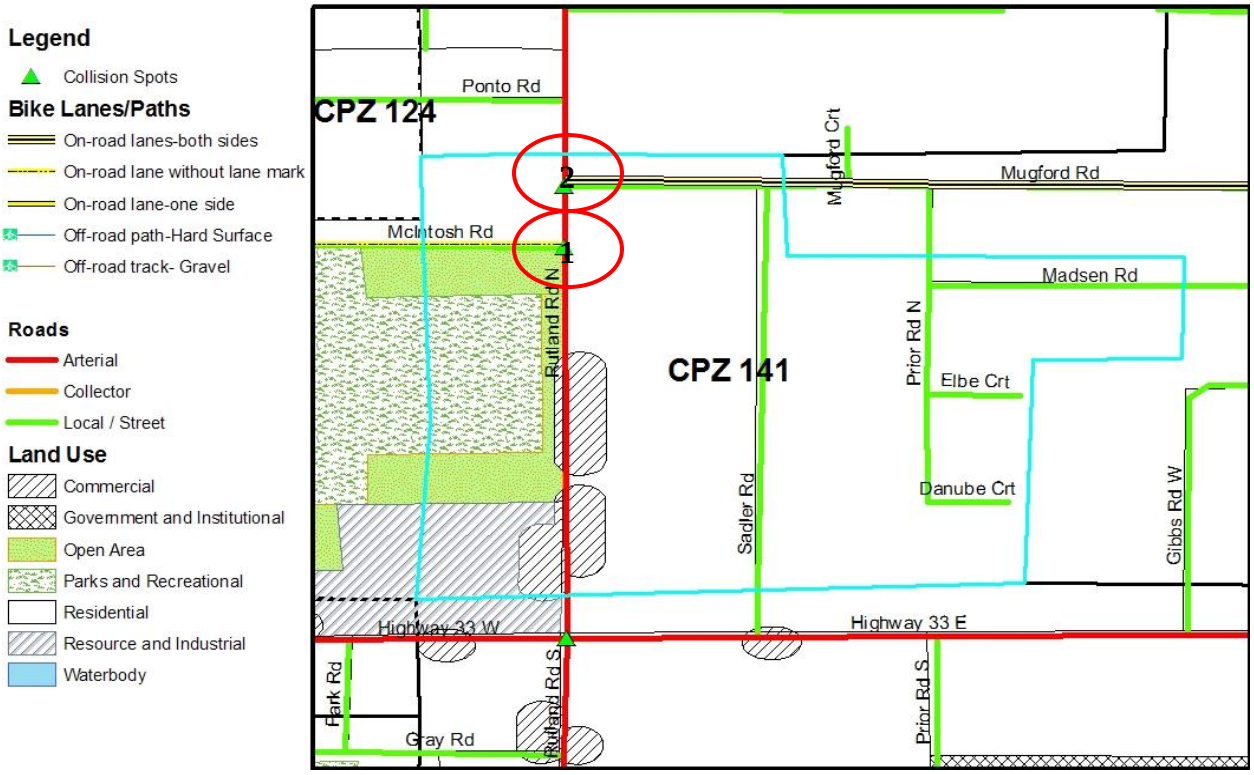

Figure 4.18 Urban Bike CPZ – Zone 141

Zone 274

Zone 274 (see Figure 4.19) was only identified as a CPZ by the NB model from the NW group. Four clues are given as follows. First, this zone is on the east bank of Okanagan Lake and located in downtown Kelowna. A half area of this zone is occupied by City Park. Traffic is busy in this zone because of too many leisure activities. Second, except the arterial road Harvey Ave., other roads in this zone are local or connector roads. An off-road multiple-use path goes though City Park. Third, the pedestrian and bicycle volume in this zone are much higher than the regional average level. Fourth, at the time of this study, two black spots included the intersections of Abbott St. @ Harvey Ave. and Abbott St. @ Leon Ave. Possible remedies for the safety problem are suggested as follows.

- TDM – Make access management (restrict left turns at some intersections), or set signage showing safer cycling routes.
- NW – Design bike lanes/paths along Abbott St. to increase cycling access in this zone.

Figure 4.19 Urban Bike CPZ – Zone 274

4.3 Summary

Based on developed macro-level CPMs for the RDCO, a reactive road safety application was conducted to identify bike-vehicle CPZs and black spots (i.e. intersections) in these CPZs. This macro-reactive black spot study is capable of facilitating early identification of neighbourhood road safety problems. In this road safety application, the EB method with NB models and the FB method with FB models were used to identify and rank CPZs and SZs. The identification results show that the two methods could yield roughly consistent results: the similarity rates of CPZ and SZ identification between the two methods are 97% and 76%, respectively.

After CPZ identification, diagnoses and remedies were preliminarily suggested for each CPZ. Trigger variables, together with land use, road maps, and on-site visits provided clues to possible road safety problems. In most CPZs, the arterial-local intersection percentage (high) is a trigger variable, implying it is necessary to improve the safety of cyclists at arterial-local intersections. Moreover, ICBC data show that cyclists were more likely to suffer collisions along Highway 33 and Highway 97. This result suggests that a safer and more complete bike network should be considered to improve the safety of cyclists who share highways with motor vehicles.

CHAPTER 5 DATA ISSUES AND NEEDS FOR VULNERABLE ROAD USERS' TRANSPORT SAFETY RESEARCH

This chapter refers to data issues in macro-level CPM research and discusses future data needs for transport safety research of vulnerable road users' (VRUs). As mentioned before, the accuracy of VRUs' road safety data is important for VRU safety research (e.g. macro-level CPM development and application). Although this chapter focuses only on bicycle collision data, similar issues may also exist in pedestrian collision data.

Three sections are included in this chapter. *Section 5.1* analyzes bicycle collision data from two sources: BC Injury Research and Prevention Unit Center (BCIRPU) and ICBC. *Section 5.2* describes the differences of these two datasets by comparing the data in the RDCO to identify data gaps. *Section 5.3* discusses promising methods to improve data quality and to promote data linkage from different sources of the VRU road safety data in BC.

5.1 Bike Collision Data Analyses

Two bike-related road collision datasets were provided by ICBC and BCIRPU to UBC-O researchers. The ICBC dataset provides motor vehicle collision claims involving cyclists, while the BCIRPU dataset is an aggregated dataset to keep hospitalized patient and mortality data due to biking-related injuries. The bike-vehicle collision data used in the model development described in *Chapter 3* is from the ICBC dataset. However, it is suspected that in addition to bike-vehicle collisions, there would be other bike-related collisions such as single bike, bike-bike, bike-pedestrian, and bike-fixed object collisions. A global data comparison between ICBC and BCIRPU shows that the number of bike-vehicle collisions was much lower than that of all bike collisions in BC (i.e. 4061 bike-vehicle collisions in ICBC vs. 6498 bike-related collisions in BCIRPU). This information indicates that each data source does not actually cover all bike collision data. The BCIRPU data was considered to be a supplement to the ICBC data in macro-level bicycle CPM development. However, it is virtually impossible to combine them at this time because of different data contents, structures, and information privacy protocols. Table 5.1 shows the sources, structures, and availability of use of the two datasets in our research.

Table 5.1 Data Sources, Structures, and Availability for Use in CPM Research

ICBC	BCIRPU
Claims Customer • Info. Sources • Body region • Conditions • Demographic info. of individuals • Individual roles in crashes • Severity and Injury Results	**BC Ministry of Health** • Hospitalized patient information **BC Vital Statistics Agency** • Mortality information
What we have from ICBC? • Geo-coded collision incident location in ArcGIS format o Date & time o Occurrence place (geo-coded) o General descriptions o Severity o Heavy vehicle/motorcycle/ pedestrian/bike flag • Cyclist data in Excel format o Incident years and cities o Cyclists' residence address (first three digital postal codes) o Cyclists numbers	**What we have from BCIRPU?** • Patient data in Excel format o Fiscal year o Patient gender and age o Local health area ID o Canada health service delivery area ID o Hospital ID o First three digital postal codes of patients' residence o Admitted to hospital & leaving hospital dates o Codes of injury causes.
Can be used in CPM development? Yes	**Can be used in CPM development?** No

The ICBC data warehouse consists of elements extracted from Claims Customer (CCUS) and their Medical Services Plan (MSP) Invoice databases. In addition to information sources, body regions, injury conditions, and injury severity results, the CCUS data also records demographic information of involved individuals such as age, gender, marital status, and the person's 'role' in the collision (i.e. pedestrian, cyclist, driver and passenger). The MSP Invoice data includes the transaction level of fee for medical service for all ICBC customers. The ICBC warehouse is able to link their injured customers to other claims data. The geo-coded shapefile data provided by ICBC for this research shows collision occurrence locations in maps, but does not include the MSP Invoice data. The collision attribute tables in the shapefiles also include collision descriptions, occurrence time/day/date/month/year, collision severity, types of collision impact, and total number of vehicles in each collision. In addition to the geo-coded shapefile

data, ICBC provided a MS Excel format dataset from its data warehouse, which shows the information of cyclists involved in ICBC collision claims. This Excel dataset is separated with the geo-coded collision dataset and includes information such as the first three digits of cyclists' residential postal codes, incident year, and incident cities.

BCIRPU collects biking hospitalization data from the BC Ministry of Health and biking mortality data from the BC Vital Statistics Agency annually. However, biking injuries without hospitalization treatment were not recorded. For example, lightly injured cyclists might be treated at walk-in clinics or hospital emergency departments. Due to the Freedom of Information and Protection of Privacy guidelines (FoI/PoP), BCIRPU could not access all information from the BC Ministry of Health and the BC Vital Statistics Agency. Although the BCIRPU database is a secondary data source, it is the only other available data source we can use for research as of now. In the future, more collision data from other sources will be pursued (e.g. BC Coroner's, hospitals). The dataset provided by BCIRPU includes only patients' gender and age, local health area ID, Canada health service delivery area ID, hospital ID, the first three digits of their residential postal codes, admitted to hospital dates, leaving hospital dates, and injury cause codes in the MS Excel format.

The BCIRPU data and the ICBC data were analyzed individually and the analysis results are described in the following sections. It is noted that the study area for these data analyses is limited to the RDCO.

5.1.1 Geographical Distribution of the Bike Injury Data from BCIRPU

This section describes geographical distribution of bike injury data from BCIRPU. In the BCIRPU dataset, injury occurrence locations were not available, so it is impossible to identify how many bike hospitalization injuries occurred in the RDCO. However, the residential postal codes of patients and hospital ID can provide references for analyses. Given the first three digits of each patient's residential postal codes, the number of injured cyclists who lived in the RDCO is known. According to the first three digits of the postal codes of injured cyclists, the RDCO is divided into nine communities, including V1V, V1X, V1Y, V1W, V1Z, V1P, V4V, V4T, and V0H. However, it should be pointed out that the community with the postal code of V0H (in Peachland) covers three regional districts, including the RDCO, the Regional District of Kootenay Boundary (RDKB), and the Regional District of Okanagan Similkameen (RDOS).

This situation indicates that hospitalized cyclists living in the V0H community not only include some population from RDCO but also include some population from RDKB and RDOS. With current information, it is impossible to separate the injury data of the V0H community from the three regional districts. Therefore, what can be confirmed is that there were less than 341 injured cyclists living in the RDCO from 2002-2006.

Given the hospital ID, the number of injured cyclists who were hospitalized in the RDCO is known. In the RDCO, Kelowna General Hospital (KGH) is the only hospital, located in the Mission district in Kelowna. According to 2002-2006 records, 308 injured cyclists (e.g. visitors and residents) were hospitalized in the RDCO, and 241 local injured cyclists (e.g. only residents living in the RDCO) were hospitalized in the RDCO. Therefore, it is inferred that about 67 (about 22%) visiting cyclists were hospitalized in the RDCO. These 67 cyclists were from the outside the RDCO but probably suffered bike injuries in the RDCO. It is also inferred that less than 100 (less than 29%) local cyclists were not hospitalized in the RDCO. These less than 100 cyclists might suffer bicycle injuries out of the RDCO. These injury numbers are classified and presented in Table 5.2.

Table 5.2 Statistic Comparisons between Counts in Several Classifications (2002-2006)

Injured Cyclists		Counts	Derivation
(1)	Live in the RDCO	<341[1]	In terms of postal codes
(2)	Hospitalized in the RDCO	308	In terms of hospital ID
(3)	Live and Hospitalized in the RDCO	241	In terms of postal codes and hospital ID
(4)	Live but not hospitalized in the RDCO	<100	<341-241
(5)	Hospitalized but not live in the RDCO	67	308-241

Notes: [1] This is the number of patients who live in the 9 communities with the first three digit postal codes of 'V1V', 'V1X', 'V1Y', 'V1W', 'V1Z', 'V1P', 'V4V', 'V4T', and 'V0H'. However, the community of 'V0H' covers three regions including the RDCO, RDKB, and RDOS, so the number of patients just living in 'V0H' community of the RDCO should be smaller than that of patients living in the entire 'V0H' community.

All bike injuries from the BCIRPU database were aggregated into the nine postal code communities. Figure 5.1 uses histograms to show 1) the number of injured cyclists who lived in each community and were hospitalized in the RDCO, and 2) the number of injured cyclists who only lived in each community. The numbers of these two groups in each community are

comparable except the V0H community. This situation is reasonable because the VOH community includes injured cyclists from other regions.

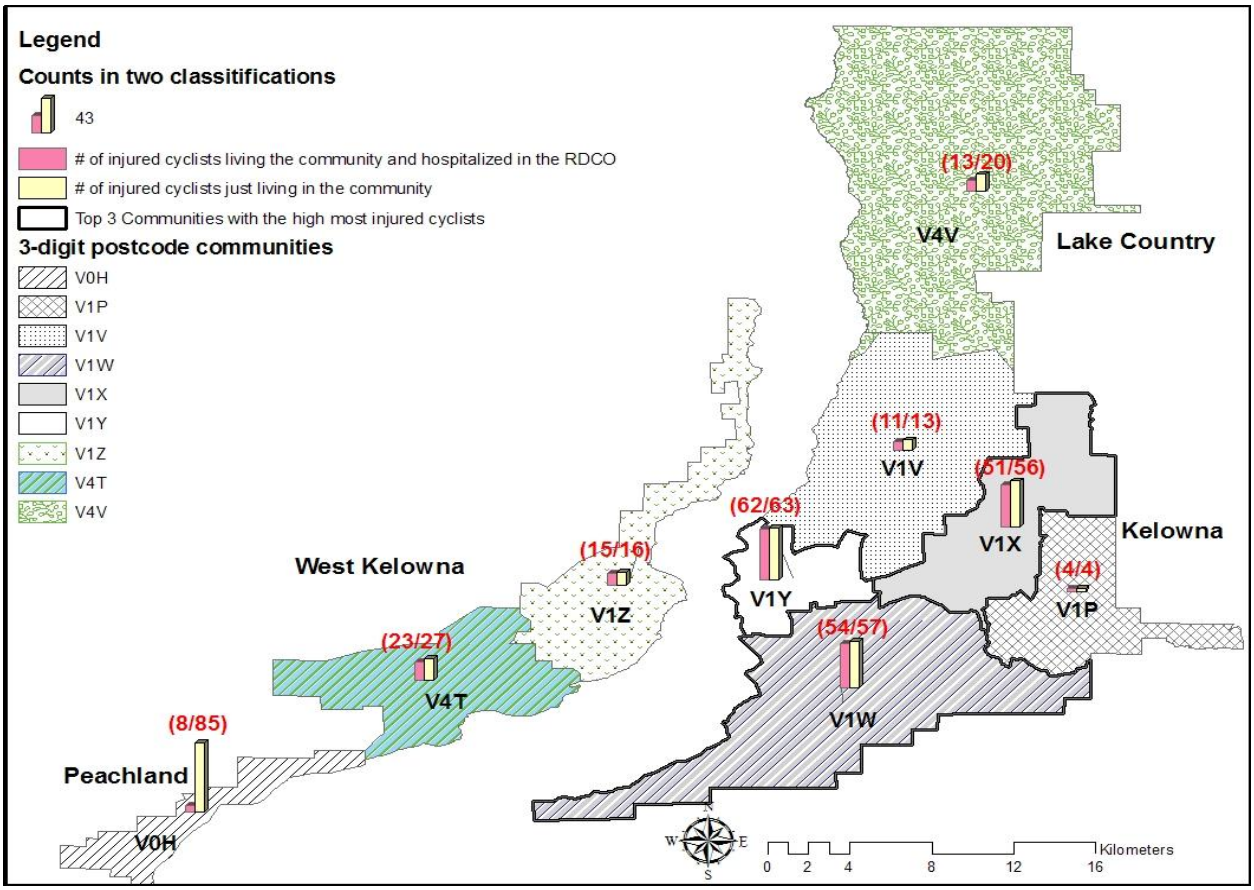

Figure 5.1 BCIRPU Data in 9 Communities in the RDCO

5.1.2 Geographical Distribution of the Bike-vehicle Collision Data from ICBC

This section describes the geographical distribution of bike-vehicle collision data from ICBC. Given the geo-coded data of collisions, the number of ICBC bike-vehicle collisions that occurred in each postal code community is known. Given the first three digits of the residential postal codes of the cyclists who were involved in ICBC claims, the number of cyclists living in each postal code community is known. Based on the ICBC 2002-2006 data, an analysis was conducted to show how many local cyclists suffered collisions outside their residential community. Related results are shown in Table 5.3. In this table, column (1) presents the number of local cyclists who were involved in ICBC claims no matter where these claims happened, column (2) presents the number of local cyclists who were involved in the ICBC

claims that only occurred in the four RDCO municipalities, and column (3) presents the number of local cyclists who were involved in ICBC claims that happened in their own residential municipality. The comparison of column (1) and (2) shows that 17% (i.e. (249-207)/249) of claims involved cyclists who had collisions out of the RDCO. The comparison of column (2) and (3) shows that 12% (i.e. (207-182)/207) of the involved cyclists had collisions in the RDCO but outside of their residential RDCO municipalities. In addition, most cyclists living in the three small municipalities of the RDCO (i.e. Peachland, West Kelowna, and Lake Country) were likely to suffer collisions in the municipality of Kelowna.

Table 5.3 Analysis of ICBC Involved Cyclist Data (2002-2006)

Communities	**(1)[1]**	**(2)[2]**	**(3)[3]**
V0H (Peachland)	25	2	1 (the other 1 in Kelowna)
V1P (Kelowna)	5	5	5
V1V (Kelowna)	14	13	13
V1W(Kelowna)	36	31	31
V1X (Kelowna)	48	45	45
V1Y (Kelowna)	89	84	83 (the other 1 in West Kelowna)
V1Z (West Kelowna)	16	13	1 (other 12 in Kelowna)
V4T (West Kelowna)	12	11	2 (other 9 in Kelowna)
V4V (lake country)	4	3	1 (other 2 in Kelowna)
Total	249	207	182

Notes:

[1] Number of cyclists living in each community, and these cyclists were involved in ICBC bike-vehicle collisions no matter where these collisions happened.

[2] Number of cyclists living in each community, and these cyclists were involved in the bike-vehicle collisions that only occurred in the four RDCO municipalities.

[3] Number of cyclists living in each community, and these cyclists were involved in bike-vehicle collisions that happened in their own residential municipality.

Figure 5.2 shows the number of ICBC bike-vehicle collisions in each postal code community and the number of cyclists who were involved in ICBC claims and lived in each postal code community. Both groups of data indicate that the communities of V1Y, V1X, and V1W have the highest numbers of bike-vehicle collisions and injured cyclists.

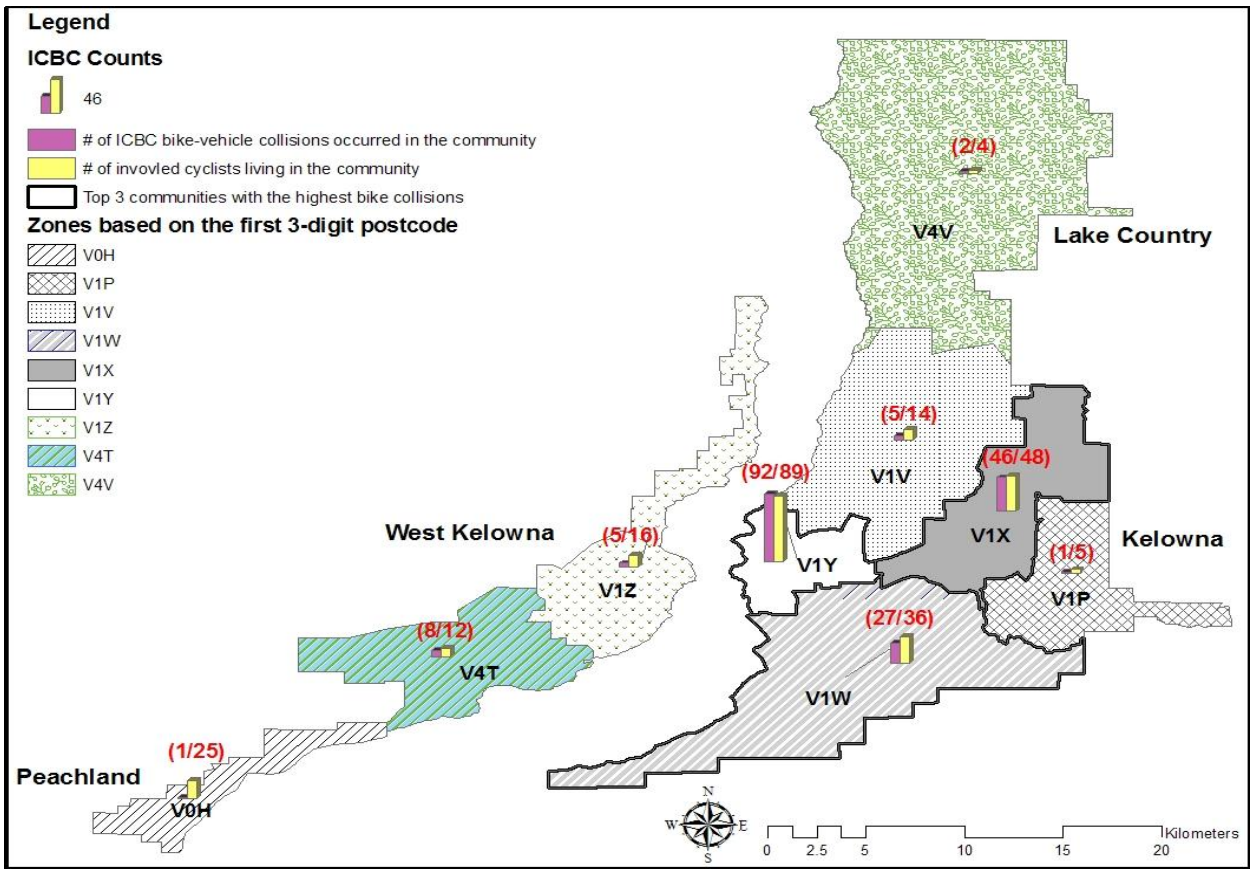

Figure 5.2 ICBC Bike-vehicle Collision Data in 9 Communities in the RDCO

The following describes the geographical relationships between the three collision prone communities (CPCs) discussed above (i.e. the communities of V1Y, V1W and V1X) and collision prone zones (CPZs) described in *Chapter 4*. Figure 5.3 displays the three CPCs and 16 CPZs. Apparently, these three communities cover most urban areas in this region and all CPZs derived from macro-level CPMs are located in the three CPCs. A preliminary traffic exposure analysis in each postal code community was conducted to suggest possible causes of the three CPCs.

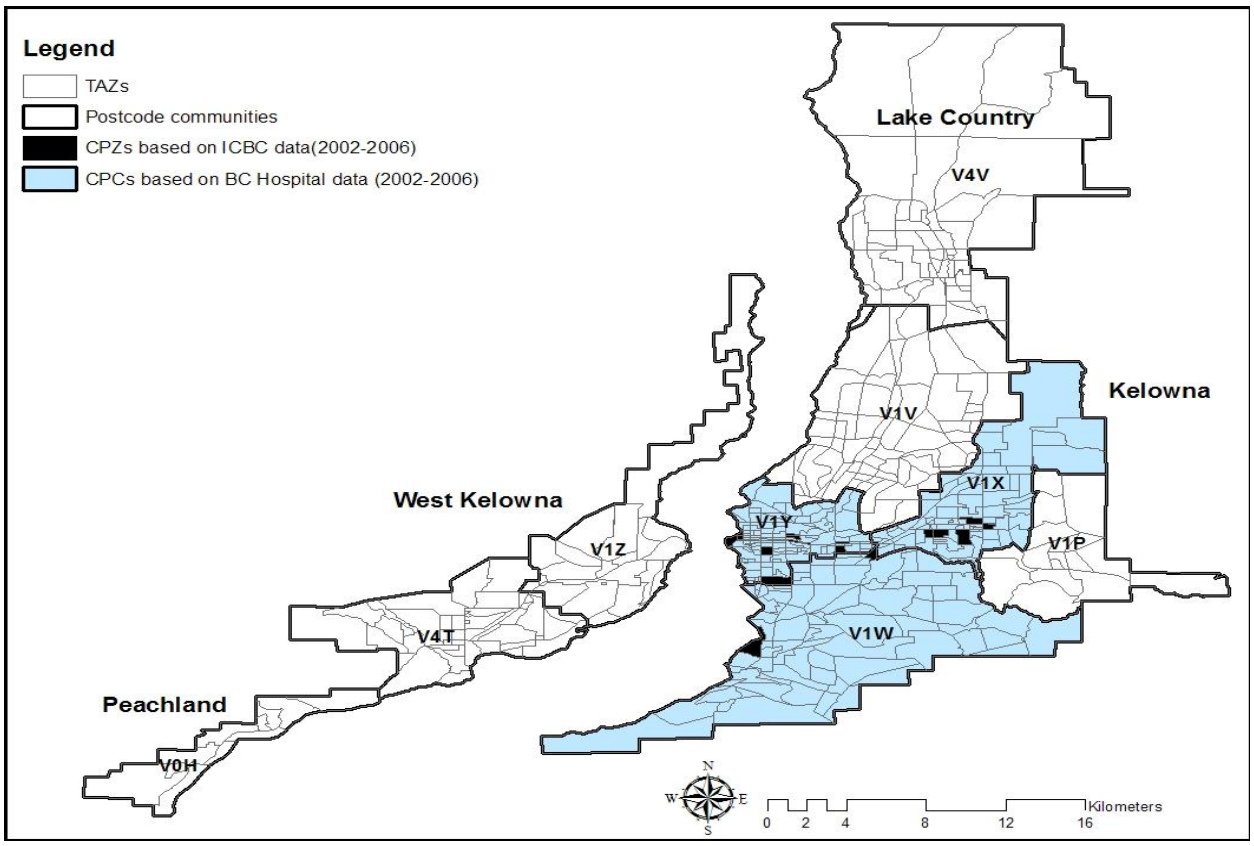

Figure 5.3 Collision Prone Zones/Communities for Bicyclists based on Two Datasets

Table 5.4 shows the traffic exposure in each community, including total lane kilometers (TLKM), bike lane kilometers (BLKM), total lane kilometer densities (TLKD), bike lane kilometer densities (BLKD), availability of bicycle lanes/paths/trails, and subjective evaluations for vehicle kilometres travelled (VKT) and bike kilometres travelled (BKT). Based on a general observation, the evaluations of VKT and BKT were classified into three levels: low, medium, and high. It is found that most trails and parks in the RDCO, such as Mission Creek Trail, Crawford Trail, Gillard Trail, MacDougall Rim Trail, Rose Valley Park, and Knox Mountain Park, cross though the communities of V1Y, V1X, and V1W.

The analysis results suggest a significant association of bike injuries with traffic exposure. Generally, the communities with high or medium vehicle volume and high or medium bike volume have a high number of bike injuries. Also, more trails and parks for mountain biking are related to more non-traffic bike injuries.

Table 5.4 Traffic Exposure Comparison in 9 Postal code Communities

Community	Area[1]	TLKM	BLKM[2]	TLKD	BLKD	Bike lane[3]	Bike paths/ Trails[4]	VKT	BKT
V1Y Downtown	18.3	336.2	59.1	18.4	3.2	Yes	Paths & trails	High	High
V1W Mission	78.6	499.8	131.9	6.4	1.7	Yes	Paths & trails	Medium	High
V1X Rutland	39.0	401.6	60.7	10.3	1.6	Yes	Paths & trails	High	Medium
V1V UBC-O	63.7	310.8	57.3	4.9	0.9	Yes	Only trails	Medium	Medium
V1P Black Mout.	30.9	139.9	10.3	4.5	0.3	Yes	Paths & trails	Low	Low
V4V Lake country	118.0	454.7	0	3.9	0	No	Only trails	Medium	Low
V1Z Westside	39.8	325.7	0	8.2	0	No	Only trails	Medium	Low
V4T Westbank	38.1	355.8	0	9.3	0	No	Only trails	Medium	Low
V0H Peachland	17.2	180.0	0	10.5	0	No	Only trails	Low	Low

Notes:
[1] Unit – Squared Kilometres
[2] The in-park trails and un-marked bike lanes were excluded.
[3] Only on-road marked bike lanes
[4] Off-road bike paths or and trails

5.1.3 Other Analyses from the BCIRPU and ICBC Data

Table 5.5 breaks down the total of 341 cyclists injured in the RDCO according to different collision types. The collision types shown in Table 5.5 are derived from the BCIRPU data. The primary sources of the BCIRPU data use the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) to classify transport collisions; thus, the definitions of 'traffic' and 'non-traffic' collisions shown in Table 5.5 are same as the definition of 'traffic' and 'non-traffic' collisions in ICD-10 (CIHI, 2009). In ICD-10, traffic collisions are defined as any transport collision occurring on the public roads. Non-traffic collisions are defined as any transport collision occurring entirely in any place other than a public road (e.g. transport collisions in mountain bike skills parks or home). For a concise description, injury

causes were classified in this study as bike-bike, bike-pedestrian/animal, bike-fixed object, bike-vehicle, single bike collisions, and other unspecified collisions in individual traffic and non-traffic collision groups.

Table 5.5 Collision Types Causing Bike Injuries[1] (RDCO, 2002-2006)

	Collision Types	Count	Percentage
Traffic	Bike-bike	2	1.3%
	Bike-pedestrian/animals	3	2.0%
	Bike-fixed object	5	3.3%
	Bike-vehicle	52	34.2%
	Single Bike[2]	88	57.9%
	Others	2	1.3%
	Total	**152**	**100%**
Non-traffic	Bike- three-wheeled motor vehicle	1	0.5%
	Bike-fixed object	8	4.3%
	Single Bike[2]	180	95.2%
	Total	**189**	**100%**

Notes:
[1] Secondary data source: BCIRPU database.
[2] 'Single Bike' means collisions such as bike falling collisions, or bike boarding and alighting collisions, which only involve bikes instead of other vehicles.

As shown in Table 5.5, the number of non-traffic bike injuries was 189 in the RDCO from 2002-2006, 10% higher than traffic bike injuries. Of the traffic bike injuries, 57.9% were due to single bike collisions and 34.2% were due to bike-vehicle collisions. Of the non-traffic bike injuries, 95.2% were due to single bike collisions and 4.3% were due to bike-fixed object collisions. Generally, 78.6% of bike injuries (traffic and non-traffic) in the RDCO were caused by single bike collisions. Single bike and bike-vehicle collisions became two main causes for bike injuries in the RDCO. This result suggests that future work should focus on how to improve the safety of non-traffic bike activities (e.g. mountain bike skills).

The BCIRPU data show 281 male cyclists injured in the RDCO from 2002-2006, over four times the 61 female injured cyclists. This finding suggests that males did more cycling activities than females in this region. Meanwhile, the numbers of injured cyclists in different ranges of age are presented in Table 5.6. This table apparently shows that the younger injured cyclists (i.e. <30 years old) account for over 50% of the total injured cyclists in this region. Using population

numbers from 2006, the bike injury rate of the RDCO population under 30 years was 33 injuries per 10,000 RDCO residents, 55% higher than the average bike injury rate of 21 bike injuries per 10, 000 of. Therefore, future research should give more attention to the safety of younger cyclists.

Table 5.6 Numbers of Patients in the RDCO in Terms of Ages (RDCO, 2002-2006)

Age	Patient counts	%
<=15	98	28.7%
16-25	65	19.1%
26-35	42	12.3%
36-45	35	10.3%
46-55	45	13.2%
56-65	21	6.2%
>65	35	10.3%

Based on ICBC data, several observations about RDCO bike-vehicle collisions are revealed. First, 95% of bike-vehicle collisions were severe (i.e. injury and/or fatality). Second, 90% of bike-vehicle collisions occurred at intersections and 10% happened at mid-blocks in the RDCO. Third, 88% of bike-vehicle collisions (165 out of 187) were in the urban area, while 12% were in the rural area. Last, 55% of bike-vehicle collisions occurred on roads with bike lanes, versus 45% on roads without bicycle lanes.

ICBC data also show collision dates and time. Of 187 ICBC bike-vehicle collisions in the RDCO from 2002-2006, 30 collisions happened from Dec. to March, 96 collisions happened from June to Sep., and the others happened in the remaining months. This result is reasonable because the winter has much lower bicycle use than the summer. Of these 187 bike-vehicle collisions, 55 collisions occurred from 3-6 pm, accounting for 29%. A traffic counting study was conducted at 13 intersections in Kelowna by UBC students in the summer of 2011. This study shows that the bicycle volume from 3-6 pm accounted for 40% of total daily bicycle volume, much higher than the bike-vehicle collision proportion at the same period, 29%. This result indicates that bike-vehicle collisions have a non-linear relationship with bicycle volume; in other words, decreased bike-vehicle collision risks are associated with increased bicycle volume. A high number of bike-vehicle collisions at the afternoon rush hours could be explained by the following reasons: 1) the period from 3-6 pm covers the afternoon traffic rush hours in which

motor vehicle volume is high, 2) many cyclists choose 3 to 6 pm as their biking time, 3) drivers suffer from decreased visibility at sunset even if they have good vision, 4) drivers have not adjusted their driving habits to accommodate for the change from daytime to dusk, and 5) drivers experience fatigue from their daily work. A Transport Canada study suggests a similar result to our findings, showing that 17% and 23% of cyclists were killed and seriously injured, respectively, in the same afternoon rush hours in Canada (Transport Canada, 2004).

5.2 Data Comparisons and Issues between BCIRPU and ICBC Datasets

BCIRPU data and ICBC data in the RDCO were compared. This comparison helps researchers understand differences between two datasets, identify data gaps, and will contribute to data cooperation in the future. The BCIRPU data show that 56 cyclists injured in bike-vehicle collisions were hospitalized in the RDCO from 2002-2006. Of these 56 cyclists, 52 lived in the RDCO. The ICBC data show that there were 187 bike-vehicle collisions in the RDCO from 2002-2006. Despite different comparison objects (i.e. injured cyclists vs. collisions), the records in the BCIRPU data are obviously fewer than the records in the ICBC data.

In addition, this comparison raises question about how many of the same items can be identified from the two datasets. To answer this question, it is necessary to find identification codes to link the two datasets. However, based on the current available information, the only "code" that could be used to identify the same items is "event dates" even though it is not a good identification code to link the two datasets. The ICBC data include collision occurrence dates, and the BCIRPU data include admitted hospital dates of patients. If the occurrence dates of bike-vehicle collisions are assumed to be the same as the admitted hospital dates for injured cyclists, the same items from the two datasets would be tentatively identified. Through this speculation, only 17 same items in the two datasets were identified. Figure 5.4 illustrates the differences of the BCIRPU and ICBC data. Four questions related to these differences between the two datasets are discussed as follows.

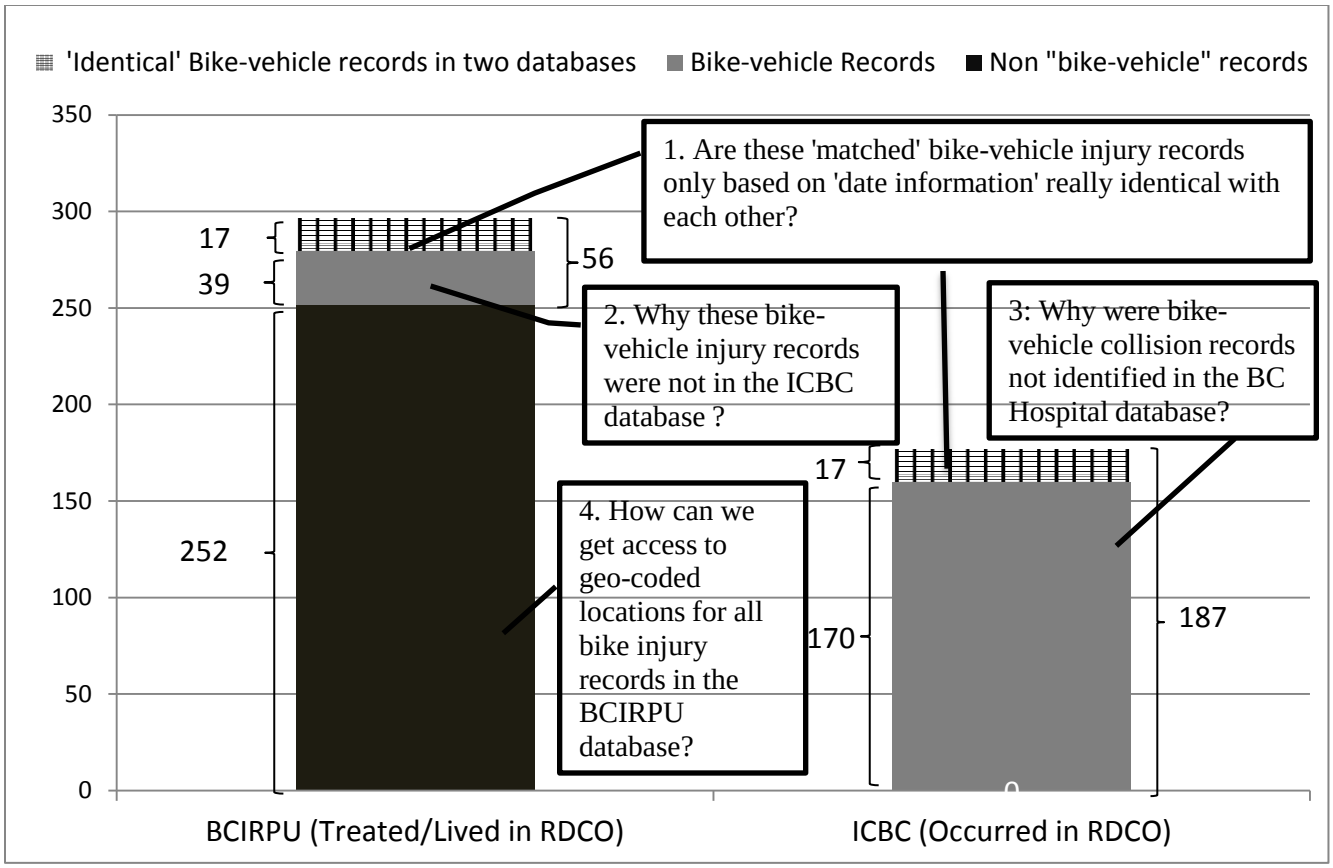

Figure 5.4 Comparisons between BCIRPU and ICBC Data in the RDCO

First, do these 17 pairs of "same" items identified with "event dates" really match each other? As mentioned before, it is not reliable to use 'matching' injury occurrence dates and patients' admitted hospital dates to identify the same items because these two dates for any injury event might be different. In addition, there is confusion with the identification of the 17 pairs of "same" items. Table 5.7 presents three examples to show the confusion. As shown in this table, more than one item in the ICBC data matches one item in the BCIRPU data in Example 1 and 2, and more than one item in the BCIRPU data matches one item in the ICBC data Example 3. Such confusion makes it impossible to identify which one of the several items from one data source is related to the unique item in the other data source. Therefore, unified identification codes are needed to link different sources of road safety data. The patient ID could be a good option for data linkage.

Table 5.7 Examples to Show Why the 'Matching' Dates Does not Work

	ICBC Data		BCIRPU Data	
	Occurrence date	**Occurrence Location**	**Admitted date**	**Residential Postal codes**
Example 1	26/05/2003	Hwy33 @ Ziprick Rd,	26/05/2003	XX
	26/05/2003	Old Okanagan Rd @ Butt Rd		
Example 2	29/05/2003	Springfield Rd @ Benvoulin Rd	29/05/2003	V1Y
	29/05/2003	Harvey Ave @ Sutherland Ave		
	29/05/2003	Glenmore Rd @ High Rd		
Example 3	26/08/2006	Hwy33 @Rutland Rd	26/08/2006	V1Y
			26/08/2006	XX

Second, why are so many bike-vehicle injury records (i.e. 56-17) in the BCIRPU data not identified in the ICBC data? Logically, any bike-vehicle collisions involving ICBC's customers and being claimed should be covered by ICBC's database. Therefore, if the injury records in the BCIRPU data did not involve ICBC customers (e.g. the drivers out of the BC provinces), the collision records of these injuries were not included in the ICBC data warehouse. In addition, if the injured cyclists who were hospitalized in the RDCO suffered collisions out of the RDCO, the collisions of these injuries were not included in ICBC's RDCO records either.

Third, why is the number of bike-vehicle collisions in the ICBC data much higher than the number of bike-vehicle injuries in the BCIRPU data, and why are many ICBC bike-vehicle collisions (i.e. 187-17) not identified in the BCIRPU data? Generally, the BCIRPU data only cover serious bike injuries which were treated in hospitals but do not include slight bike injuries. If cyclists who were involved in ICBC claims were slightly injured, they might be treated at emergency departments or walk-in clinics in which case their records were not kept in BCIRPU's hospitalization data. Also, if cyclists who were involved in ICBC claims were treated in other hospitals out of the RDCO, their records were not kept in the BCIRPU data for the RDCO either.

Finally, how is possible to know the occurrence locations of bike injuries recorded by hospitals (or by BCIRPU)? Currently, the occurrence locations of bike injuries have not been widely collected by hospitals. However, if road safety data from different sources could be linked, the locations of some bike injuries recorded by hospitals might be identified from other data sources, such as ICBC, police, and ambulance services.

The comparison of the BCIROPU and ICBC data reveal three data gaps: first, it was hard to accurately identify the same items between the two databases due to a lack of data linkages; second, each database does not cover all bike collisions in BC; third, the occurrence locations of some bike injury incidents are not available. At present, different organizations just focus on the bike collision data under their responsibilities, and these data from different sources may overlap each other or may not. In order to create a non-overlapped but complete database that includes all bike traffic collisions in BC, the problems in data collection, access, and linkage must be solved.

5.3 Data Collection, Data Access, and Data Linkage in BC

Based on the data gaps described in *Section 5.2*, potential improvements in road safety data collection, access, and linkage are discussed in this section. In BC, different agencies (e.g. ICBC, BCIRPU) collect road safety data in different ways, but they have few linkages to each other. Their databases might record different information of the same collision events. For example, ICBC or the police may record the information of traffic collisions, and BC hospitals may record medical information of patients who were injured in these collisions. Under a legal agreement, all the information from different sources could be linked and combined to create a complete road safety database. In the Canadian Injury Prevention and Safety Promotion Conference (CIPSPC, 2011) in Vancouver, medical researchers from different fields in BC had a sub-meeting on how to launch the BC Road Safety Research and Evaluation Centre (BCRSREC), which will act as an unbiased expert to promote road safety in BC. The goals of BCRSREC are to provide: 1) knowledge generation, 2) communication, 3) networking, 4) database development, and 5) policy update. The members of BCRSREC are expected to include university affiliation, health authority affiliation, and other groups who conduct road safety research in BC. Meanwhile, the structure of BCRSREC is planned to have loose association with some central administrative support. In this meeting, many data harmonization and linkage issues were preliminarily discussed by professionals. Several issues discussed at this meeting, including the primary sources of road safety data, the collection of collision geo-coded information, the possibility of building an integrated data warehouse, the development of an over-arching data dictionary, and privacy issue in data access and linkage, are preliminarily considered below.

5.3.1 Primary Sources of Road Safety Data in BC

In order to build an integrated data warehouse in BC, road safety data should be provided by different sources. In additional to ICBC, other organizations have their own databases to keep road safety and injury data in BC. BC Ambulance Service (BCAS) provides public ambulance service under the authority of the Emergency Health Services Commission of the provincial Ministry of Health, and it records the information of road injury emergency events. The BC Ministry of Health (BCMH) has the BC Health Data Warehouse to provide online access to community-level population health indicators and datasets. BCMH is one of data sources for BCIRPU's data collection. The BC Trauma Registry collects and maintains data on all trauma patients admitted to any of the nine trauma receiving facilities in BC. Population Data BC, which is a multi-university, data and education resource, supports research access to individual-level, de-identified longitudinal data on BC's four million residents.

Two organizations specifically keep fatality data in BC. The BC Vital Statistics Agency registers vital events occurring in BC and provides vital statistics-related products and services. This agency is also one of sources for the BCIRPU data and it supports injury surveillance through its collection of death-related data. The Office of the Chief Coroner reports on all sudden, unexpected, and unexplained deaths in BC.

5.3.2 Possibility to Build an Integrated Data Warehouse

Two options for building an integrated data warehouse for road safety research are suggested. The first option is to develop a central database, which collects all data regarding road transport collisions from different databases. Each data item in the central database should include information regarding the collision, patient(s), emergency service requirements, hospital/medical treatments, insurance status, and financial costs. However, a central database might have potential risks in financial operations (e.g. what happens when funding runs out) and is complicated in practice (e.g. researchers working on the central database have to take on responsibility for keeping it up to date, yet without control over original data sources).

The second option is to develop a "virtual" data warehouse network. From an economic perspective, this option is better than the central database. A "virtual" data warehouse network is comprised of distributed databases, each of which is updated by respective "owners" who have

statutory responsibility and funding to do so. This virtual data warehouse will be a hub providing secure channels of data access to all members in this network. From a long term perspective, more related data providers could also become involved in this network. However, both of these options would be risky due to the issue of FoI/PoP.

5.3.3 Using GPS Technology to Collect Collision Locations

As mentioned in *Section 5.2*, one of the data gaps is a lack of traffic collision/injury locations. In order to fill this gap, advanced technologies such as geographic information systems (GISs), global positioning systems (GPSs), and management information systems (MISs) are more and more used in the collection of collision/injury locations.

Text-based descriptions of collision locations usually are recorded by insurance companies, police, and/or hospitals. With GIS technology, these collision locations are marked on electronic maps based on the text-descriptions. The geo-coded information marks collisions geographically associated with road networks and is widely recommended for use in research now. However, locating geo-references of collisions based on the text-descriptions may cause misallocation problems. Therefore, GPS technology is recommended to locate geo-references of collisions based on latitudes and longitudes. It will be a trend to use GPS to collect more accurate information about traffic collisions. Related findings about GPS applications in emergency services in Alberta, BC, and Quebec are described as follows.

In Alberta, the information of each ambulance event is recorded in the Alberta Ambulance Information Management System (AAIMS) (Government of Alberta, 2011). For each ambulance event, paramedics need to fill a Patient Care Report (PCR) form, in which the latitude and longitude of the event occurrence location can be recorded. However, this information only can be recorded if the corresponding ambulance unit is equipped with a GPS unit. In BC, BCAS has set up GPS units for every ambulance in its fleet (BCAS, 2010). The GPS technology makes paramedics not have to work with dispatch or reference a map book to arrive at collision locations, but helps them quickly and accurately find call locations and determine the best driving routes. Today, cell phone GPS tracking is widely used in emergency services and police. Urgences-Santé, Montreal's ambulance service, uses cell phone GPS tracking to get accurate geo-coded information of responded collisions in Montreal. For every call made to 911 in Montreal, the caller's location is automatically sent to Urgences-Santé, and then the location is

instantly mapped in a GIS to guide the ambulance to get there quickly (Miranda-Moreno et al., 2011).

5.3.4 Data Dictionary and Privacy Issues

A data dictionary and data sharing protocols are both important for an integrated data warehouse. A data dictionary is defined as "*a descriptive list of names, definitions, and attributes of data elements to be collected in an information system or database*" by American Health Information Management Association (AHIMA, 2006). A common data dictionary will make different database users have a common understanding of all data elements and provide a prerequisite to link data from different sources.

Privacy protection protocols guarantee data confidentiality and data use security. Usually, a request for raw data from any organization needs to clarify why and how the data will be used. After an agreement is signed, the data can be provided by data owners and used by demanders. The balance between data access and data confidentiality is always concerned. Lane and Schur (2010) proposed three suggestions to improve data access without affecting data confidentiality. The first suggestion is to use remote access data "enclaves." The "enclaves" should be built to facilitate the productive, high-quality usage of data and should support data collaboration environment. The concept of "enclaves" is similar to the concept of "virtual" data warehouse network. The second suggestion is to reduce the delays of data access. The third suggestion is to build a broad body of knowledge about the availability of existing technologies for data access.

5.4 Summary

This chapter focuses on data issues and needs for the safety of VRUs, especially the traffic safety of cyclists. *Section 5.1* summarizes the geographical distribution of bike injury/collision records from BCIRPU and ICBC individually, and describes other analysis results based on the two datasets. *Section 5.2* compares the two datasets in the RDCO and reveals three data gaps: first, it was hard to identify the same items accurately between the two databases due to a lack of data linkage; second, both databases did not cover all bike collisions in BC; third, the geographical locations of parts of VRU collisions were not available. *Section 5.3* discusses road safety data sources, access, linkage and sharing. An initiative to link various data from different

sources has gained momentum in the CIPSPC in Vancouver in 2011. Developing an integrated road safety data warehouse will be a trend in the future.

CHAPTER 6 CONCLUSIONS, CONTRIBUTIONS & FUTURE RESEARCH

This chapter is divided into three sections. *Section 6.1* presents a summary of research results and conclusions. *Section 6.2* describes the contributions of this research. *Section 6.3* lists limitations and future research work.

6.1 Summary & Conclusions

Macro-level CPMs are suggested as reliable empirical tools for road safety evaluation and planning. They can be used to predict safety performances of planned facilities, identify and rank collision prone zones, and evaluate the effectiveness of road safety countermeasure in communities. In this study, macro-level CPMs were developed to investigate the relationships between bicycle use and road safety in the Regional District of Central Okanagan (RDCO), and were applied to identify bike-vehicle collision prone zones (CPZs) in this region. Additionally, the data issues and needs in model development were discussed. The results and conclusions of this study are summarized below.

6.1.1 Model development with Four Regressions

The first objective of this research was to develop bike-related macro-level CPMs with four regression methods, including the generalized linear model (GLM), geographical weighted regression (GWR), zero-inflated counted regression (ZIC), and the full Bayesian (FB) method. The model development methodologies using these four regression methods were derived from previous studies. The model development data were from ICBC (2002-2006 collision claims), the City of Kelowna (2006 bike lane maps, intersection, and signal data), RDCO Official Community Plan (2005 land use map), Statistics Canada (2006 demographical and mode split data), and BC transit (2006 bus stops and bus routes). To extract and aggregate these data into community-based level, 500 TAZs were selected as areal units. Developed macro-level CPMs were stratified according to land use, data derivation, and variable themes for three collision patterns: total vehicle, severe vehicle, and bike-vehicle collisions. In addition to *old* explanatory variables suggested by Lovegrove (2007), three *new* variables related to bicycle use were also taken into account in model development. Five possible model forms were initially proposed

and shown in Table 3.2. The model form [1] was suggested to be used in model development with different regression methods, due to having best goodness-of-fit statistical results.

The traditional NB regression method is commonly used in macro-level CPM development. Many NB models for different collision patterns were developed but only 14 of them were valid. The GWR method was proposed to account for spatial heterogeneity. However, GWR models were not developed successfully because the GWR modelling simulation failed to yield parameter estimates. The ZIC regression method was used to predict bike-vehicle collisions. In the ZIC regression method, non-zero counts followed a Poisson or an NB distribution, the probability of zero counts was derived from a logistic function. However, these bike-vehicle CPMs using the ZIC method show that the zonal bike-vehicle collisions characterized by a preponderance of zero values are not indicative of zero-inflate negative binomial or zero-inflated Poisson distribution. Finally, Poisson lognormal (PLN) models using the FB method were developed and considered to be valid. To compare the FB method to the NB regression method without biases, the same variables and the same datasets in NB models were used to develop FB models.

Based on these model development results, three main conclusions were drawn. First, the results from NB models and FB models showed that it was possible to quantify a statistically significant association between community collisions and community traits. Second, given the same model form and variables for FB and NB models, it was hard to tell which models are more competitive according to the measures of MAD, MSPE, and MSE. Third, GWR and ZIC regressions were valid to be used in developing macro-level CPMs in this study, but could be researched in the future after data refinement.

6.1.2 Macro-reactive Road Safety Application

The second objective of this research was to apply developed macro-level CPMs to identify hazardous communities for bike-vehicle collisions. In this model application, the EB method with NB models and the FB method with FB models were used to identify and rank collision prone zones (CPZs) and safer zones (SZs). A 95% level of confidence was used in both methods for the CPZ identification. The collision risk ratio and potential collision reduction were used to rank identified CPZs. The identification results show that the two methods yielded roughly consistent results: the CPZ identification similarity was up to 97% and the SZ identification

similarity rate was up to 76%. After bike-vehicle CPZ identification and ranking, road safety diagnoses and remedies were conducted. Following the methodology in Lovegrove (2007), trigger variables, together with land use, road maps, and on-site visits provided clues to possible road safety problems for each CPZ. At most CPZs, the arterial-local intersection percentage (high) was one trigger variable, implying cyclists were more likely to suffer collisions at arterial-local intersections. Additionally, many CPZs were along Highway 33 and Highway 97. According to potential road safety problems of these CPZs, safer and more complete bike routes and facilities should be considered to reduce the collision risk of cyclists.

As the data used for model development is from 2002-2006, the CPZs identified by these macro-level CPMs might no longer be current CPZs. However, if the current data is available, a macro-reactive black spot study to identify current CPZs could be conducted in the same way. This macro-level CPM application enhances traditional black spot programs and facilitates early identification of road safety problems for VRUs.

6.1.3 Data Issues and Needs for VRU Road Safety Research

The third research objective was to identify data gaps and needs in VRUs' road safety research. This objective helps researchers to pursue an ongoing, long-term, future sustainable road safety research goal of developing a global model that predicts *total* collisions (i.e. the sum of vehicle-vehicle, bike-vehicle, pedestrian-vehicle, bike-bike, bike-pedestrian collisions) instead of only vehicle collisions with increased sustainable transport. ICBC is a primary source and BCIRPU is a secondary source to provide bike collision/injury data, which were used for analysis in this study. ICBC data provides motor vehicle collision claims involving cyclists, and BCIRPU data provide hospitalization and mortality data due to bike injuries. The geographical distributions of bike collision/injury data from these two databases were analyzed individually. According to a ranking of collision/injury numbers in nine postal code communities, the communities with the postal codes of V1Y, V1W and V1X had the highest numbers of injured cyclists and bike-vehicle collisions. The main reason for this situation appears to be high motor vehicle/bike volume in these three communities. In addition, the time and location attributes of bike-vehicle collisions, the age and gender of injured cyclists, and bike injury causes were analyzed.

After a comparison between BCIRPU and ICBC data, three data gaps were revealed: first, it was hard to identify the same items between the two databases due to a lack of accurate data linkage codes; second, each of the databases does not cover all bike collisions in BC; third, the occurrence locations of some transport collisions/injuries are not available. Also, several issues in road safety data collection, access, and linkage were discussed. In BC, several primary databases have road safety data, including ICBC, the police, BC Ambulance Services, provincial hospitalisation databases in BC (e.g. the BC Ministry of Health), the BC Trauma Registry, the BC Vital Statistics Agency, and the Office of the Chief Coroner. However, each of these databases does not cover all road safety data in BC and has few linkages to other databases. In order to increase data linkage, an integrated data warehouse for road safety research was suggested. The integrated data warehouse needs financial funding, technical supports, a common data dictionary, and information protection and sharing protocols.

6.2 Research Contributions

This study makes a significant contribution towards addressing the research gap in previous macro-level CPM studies for three reasons. First, the variables (e.g. bike lane kilometres) representing bicycle use were identified in this study. Second, five possible model forms including bicycle exposure variables were proposed, unlike traditional model forms which only include motor vehicle traffic exposure variables as leading variables. Third, four regression methods were used in macro-level CPMs, unlike previous studies in which the traditional NB regression method was dominantly used.

Moreover, although the EB method together with negative binomial (NB) models was previously used in macro-level black spot studies, the FB method with FB models was first attempted to identify and rank CPZs in this study. Using the FB method to identify and rank CPZs does not require an additional step to get the posterior distribution of collisions. This research shows that the FB method yielded similar CPZ identification results to the EB method, indicating that the FB models were as practical as NB models but simpler to use in macro-level black spot studies than NB models.

6.3 Research Limitations and Future Research Recommendations

Sufficient data of sound quality and a reliable methodology are cornerstones of reliable statistical models. In this research, data, statistical regression approaches, and application methodologies of community-based, macro-level CPMs still need to be refined. This section describes research limitations and future research recommendations relating to model development and model applications, respectively.

Limitation 1: In order to develop macro-level CPMs based on traffic analysis zones (TAZs), the social-demographical and model split variable data (e.g. population, the percentage of commuters in one transport mode) were transformed from dissemination areas (DAs) to TAZs. Therefore, the data aggregation bias, as one limitation in data quality, was generated.

Recommendation 1: In order to avoid this data bias, DAs will be set as aggregation areal units in macro-level CPM development in future. It is necessary to compare models based on DAs with models based on TAZs to suggest which areal units, DAs and TAZs, are better.

Limitation 2: The variable omission could be another limitation in model development. In this research, only the data of measured exposure variables (e.g. total lane kilometres, bike lane kilometres) were used for macro-level CPMs; however, the data of modelled exposure variables (e.g. vehicle kilometres travelled, bike kilometres travelled) which are better proxy variables of traffic exposures were not available.

Recommendation 2: Future work can focus on a supplement and refinement of explanatory variable data. For example, the UBC SRS lab could work with the RDCO on collecting more reliable bicycle exposure data.

Limitation 3: The collision data used in model development in this research are limited to vehicle collision data, but not overall collision data. As discussed before, the data from individual databases cannot cover all VRU collision data and some VRU collisions do not have occurrence location records. These data limitations could reduce model reliability.

Recommendation 3: In order to create an integrated collision database, data sharing and linkage are suggested. Collaborative use of a broader database network of road safety data sources would be expected, but database structures and data sharing protocols have not been compatible for BC road safety data. In future, the road safety data linkage should be improved, leading to more reliable macro-level CPMs for reactive and proactive road safety applications. Based on

collision data refinement, further model stratification by bike collision types (e.g. bike-self, bike-pedestrian, and bike-bike collisions) or cyclist ages can be left for future research. In addition, the macro-level CPMs for pedestrian collisions can be addressed too. A similar model stratification to bike collisions can be considered in pedestrian collisions in model development.

Limitation 4: Previous micro-level CPMs with the FB method show that the FB models have the ability to account for all uncertainty in the data (e.g. spatial and temporal influences, all possible interaction effects between covariates). However, the developed FB models have not considered spatial effects and covariate interaction effects. Also, only Poisson lognormal regression was used to develop FB models, but Poisson gamma (i.e. negative binomial) regression has not been used.

Recommendation 4: In future, hierarchical FB models accounting for these variations can be developed. These updated FB models will be compared with basic FB models and traditional NB models. Meanwhile, the Poisson gamma (i.e. NB) regression method should also be used to develop FB models in the future so that the performances of Poisson lognormal models and Poisson gamma models with the same FB method can be tested.

Limitation 5: The macro-level CPMs using geographically weighted regression (GWR) and zero-inflated count (ZIC) regression were not developed successfully. This result was probably caused by the weakness of data quality or the inappropriateness of these two regression methods. However, the detailed reasons have not been scientifically researched.

Recommendation 5: The failure of GWR and ZIC raises a doubt about using these two regression methods, and drives researchers to verify if this situation also exists in other jurisdictions and to look for other reasons for it in the future.

Limitation 6**:** In the development of NB models, the variable selection process followed a traditional process – forward step, but other step processes such as a backward step process have not been attempted.

Recommendation 6**:** Future research can use different variable selection processes to compare their model performances.

Limitation 7: Macro-level CPMs were just developed for the RDCO, which is a typical North American region with very low bicycle use. However, the overarching goal is to demonstrate the relationship of road safety and bicycle use in regions with different levels of bicycle use and we have no data support from other countries or regions to do this currently.

***Recommendation 7*:** In order to verify the statistical relationship of road safety and bicycle use in other levels, empirical tools need to be researched in regions with low-medium and medium-high levels of bicycle use. This future research topic may need international collaboration to support.

Limitation 8: In the macro-reactive road safety application of black spot studies, diagnoses and remedies were conducted in a preliminary level because of limited collision information.

***Recommendation 8*:** The diagnoses and remedies should be researched in detail pending data and model refinement.

***Limitation 9*:** While bicycle-related macro-level CPMs were used in a reactive road safety application – black spot studies - CPMs have not been applied in another macro-reactive road safety application – before and after studies. Also, NB models with empirical methods were often used as tools in traditional macro-level before and after studies.

***Recommendation 9*:** Future research in reactive road safety application can focus on before and after safety evaluation studies of regional or neighbourhood safety improvement countermeasures for cyclists. However, previous macro-level before and after studies were based on the EB method, so how to use the FB method to conduct before and after studies needs to be researched.

***Limitation 10*:** In this research, macro-level CPMs have not been used in proactive road safety applications for road safety planning.

***Recommendation 10*:** Pending data and model refinement, future research should focus on regional or neighbourhood safety planning studies with bicycle intensive programs.

REFERENCES

Aguero-Valverde, J., and P. P. Jovanis, Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. *Accident Analysis & Prevention*, Vol. 38, No. 3, pp. 618-625, 2005.

AHIMA e-HIM Work Group on EHR Data Content. Guidelines for Developing a Data Dictionary. *Journal of AHIMA*, Vol. 77, No. 2, pp. 64A-64D, 2006.

Akaike, H., A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, Vol. 19, Iss. 6, pp. 716–723, 1974.

Aptel I, L. R. Salmi, F. Masson, A. Bourde, G. Henrion, and P. Erny (1999). Road Accident Statistics: Discrepancies between Police and Hospital Data in a French Island. *Accident Analysis & Prevention*, Vol. 31, No. 1/2, pp. 101-108, 1999.

Baker, L. Shifting gears. *Scientific American*, Vol. 301, No. 4, pp. 28-29, 2009.

Barnett, V., Lewis, T. 1994, *Outliers in Statistical Data*, John Wiley & Sons, 1994.

Bernhoft, I. M., and G. Carstensen, Preferences and Behaviour of Eedestrians and Cyclists by Age and Gender. *Transportation Research Part F*, Vol. 11, Iss. 2, pp. 83-95, 2009.

Bronnert, J., J. S. Clark, L. Hyde, et al. Health Data Analysis Toolkit, RHIA American Health Information Management Association, 2011.

Canadian Institute for Health Information, *International Statistical Classification of Diseases and Related Health Problems 10th Revision*, 2009.

Center for Research and Contract Standardization in Civil and Traffic Engineering (CROW), *Recommendations for Traffic Provisions in Built-up Areas*. CROW, Netherlands, 1998.

Charlton, M., A. S. Fotheringham, and C. Brunsdon. *Geographically Weighted Regression Version 3.0, User's Manual and Installation Guide*, National Centre for Geocomputation, 2009.

Chatterjee, A., F. J. Wegmann, N. J. Fortey, and J. D. Everett. Incorporating Safety and Security Issues in Urban Transportation Planning, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1777, pp.75-83, 2001.

Cryer, P. C, S. Westrup, A. C. Cook, V. Ashwell, P. Bridger, and C. Clarke, Investigation of Bias after Data Linkage of Hospital Admissions Data to Police Road Traffic Crash Reports. *Injury Prevention*, Vol. 7, pp. 234-241, 2001.

de Leur, P. and T. Sayed, Development of A Road Safety Risk Index, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1784, pp. 33-42, 2002.

de Leur, P. and T. Sayed, A Framework to Proactively Consider Road Safety Within the Road Planning Process, *Canadian Journal of Civil Engineering*, Vol. 30, iss. 4, pp. 711-719. 2003.

Dill, J., and T. Carr, T. 2003. Bicycle Commuting and Facilities in Major U.S. Cities: If You Build Them, Commuters Will Use Them--Another Look. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1828, pp. 116-123, 2003.

Duckworth, R., M. Imran, and J. Chan, Combined Ranking Method for Screening Collision Monitoring Locations along Alberta Highways, prepared to *2011 Annual Conference Transportation Association of Canada, Edmonton*, Alberta, 2001.

Ekman, L. *On the Treatment of Flow in Traffic Safety Analysis: A Non-parametric Approach Applied on Vulnerable Road Users*. Department of Traffic Planning and Engineering, University of Lund, Lund, Sweden, 1996.

El-Basyouny, K., and T. Sayed, T., Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1950, pp. 9-16, 2006.

El-Basyouny, K., and T. Sayed, Urban Arterial Accident Prediction Models with Spatial Effects. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2102, pp. 27-33, 2009.

El-Basyouny, K., and T. Sayed, Full Bayes Approach to Before-and-After Safety Evaluation with Matched Comparisons: Case Study of Stop-Sign In-Fill Program. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2148, pp. 1–8, 2010.

Elvik, R., A. B. Mysen, Incomplete Accident Reporting: Meta-Analysis of Studies Made in 13 Countries, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1665, pp. 133-140, 1999.

Fotheringham, A.S., C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons, Chichester, 2002.

Garrard, J., G. Rose, and S. K. Lo, Promoting Transportation Cycling for Women: The Role of Bicycle Infrastructure. *Preventive Medicine*, Vol. 46, pp. 55-59, 2008.

Gaspers, K. On the Road to Danger: *WHO Call Traffic Injuries: A Global Public Health Problem*, Safety & Health. National Safety Council, Illinois, USA, 2004.

Geddes, E., S. Hemising, B. Locher, S. Zein, *Safety Benefits of Traffic Calming, Report prepared by Hmilton Associates for the Insurance Corporation of British Columbia*. Hamilton Associates, ICBC, Vancouver, Canada, 1996.

Government of Alberta, *Health Information Standards Committee for Alberta: Emergency Medical Services Patient Care Reporting Minimum Data Set, V1.2*, Health and Wellness Government of Alberta, 2011.

Miranda-Moreno, L. F., J. Strauss, P. Morency, Disaggregate Exposure Measures and Injury Frequency Models for Analysis of Cyclist Safety at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2236, pp. 74-82, 2011.

Hadayeghi, A., A. S. Shalaby, and B. N. Persaud, Macro level Accident Prediction Models for Evaluating the Safety of Urban Transportation Systems, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1840, pp. 87-95, 2003.

Hadayeghi, A., A. S. Shalaby, and B. N. Persaud. Development of Planning Level Transportation Safety Tools using Geographically Weighted Poisson regression. *Accident Analysis & Prevention*, Vol. 42, No. 2, pp. 676-688, 2009.

Hadayeghi, A., A. S. Shalaby, and B. N. Persaud. Development of Planning-level Transportation Safety Models Using Full Bayesian Semiparametric Additive Techniques. *Journal of Transportation Safety & Security*, Vol. 2, No.1, pp. 45-68, 2010.

Hauer, E., J.C.N. Ng, and J. Lovell, Estimation of Safety at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1185 Washington, D.C., pp. 48-61, 1988.

Hauer, E., and A. Hakkert, Extent and Some Implications of Incomplete Accident Reporting. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1185, pp. 1–10, 1988.

Hauer, E. Empirical Bayes Approach to the Estimation of 'Unsafety': The Multivariate Regression Method. *Accident Analysis and Prevention*, Vol. 24, No. 5, pp.457-477, 1992.

Hauer, E. *On Exposure and Accident Rate*, Traffic Engineering and Control, Vol. 36, issue. 3, pp. 134-138. 1995.

Hauer, E. *Observational Before-After Studies in Road Safety-Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Elsevier Science Incorporated, Tarrytown, NY, USA, 1997.

Hauer, E., D.W. Harwood, F.M. Council, and M. S. Griffith, Estimating Safety by the Empirical Bayes Method: A Tutorial. *Transportation Research Record: Journal of the Transportation Research Board,* No. 1784, pp. 126-131, 2002.

Hayes, M., Y. Holder, and W. Pickett, *The Evaluation of the Canadian Hospitals Injury Reporting and Prevention Program (CHIRPP).* Public Health Agency of Canada, 2001.

Herbel, S. B. Planning It Safe to Prevent Traffic Deaths and Injury, *North Jersey Planning Authority*, Spring Issue. pp. 7-27, 2004.

Heydecker, B. G. and J. Wu. Identification of Sites for Accident Remedial Work by Bayesian Statistical Method: An Example of Uncertain Inference. *Advances in Engineering Software*, Vol. 32, pp. 859-869. 2001.

Higle, J.L., and J. M. Witkowski, Bayesian Identification of Hazardous Locations. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1185, pp. 24-36, 1988.

Hirshon, J. M., M. Warner, C. B. Irvin, et al. Research Using Emergency Department-related Data Sets: Current Status and Future Directions. *Academic Emergency Medicine,* Vol. 16, No. 11, pp. 1103-1109, 2009.

Huang, H., H. C. Chin, and M. M. Haque, Empirical Evaluation of Alternative Approaches in Identifying Crash Hot Spots: Naive Ranking, Empirical Bayes, and Full Bayes Methods, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2103, pp. 32-41, 2009.

Hvoslef, H. *Under-reporting of Road Traffic Accidents Recorded by the Police at the International Level*. Norway: Public Roads Administration, 1994.

ITE. *Statistical Evaluation in Traffic Safety Studies*. Washington, DC: ITE, 1999.

Jacobsen, P. L., Safety in Numbers: More Walkers and Bicyclists, Safer Walking and Bicycling. *Injury Prevention*, Vol. 9, pp. 205–209, 2003.

James, H. F. Under-reporting of Road Traffic Accidents. *Traffic Engineering Control*, Vol. 32, pp. 573–83, 1991.

Kim, K., P. Pant, and E. Yamashita, Accidents and Accessibility: Measuring the Influences of Demographic and Land Use Variables in Honolulu, Hawaii. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2147, pp. 9-17, 2010.

Kim, H., D. Sun, and R. K. Tsutakawa, Lognormal vs. Gamma: Extra Variations. *Biometrical Journal*, Vol. 44, Iss. 3, pp. 305-323, 2002.

Khondakar, B. *Transferability of Community-Based Macro-level Collision Prediction Models for Use in Road Safety Planning Applications*, Master Thesis of Applied Science in University of British Columbia, 2008.

Lan, B., and B. Persaud, Fully Bayesian Approach to Investigate and Evaluate Ranking Criteria for Black Spot Identification, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2237, pp. 117–125, 2011.

Lane, J., C. Schur, Balancing Access to Health Data and Privacy: A Review of the Issues and Approaches for the Future, Health Services Research, Vol. 45, Iss. 5p2, pp. 1456-1467, 2010.

Leden, L., P. Garder, and U. Pulkkinen, An Expert Judgment Model Applied to Estimate the Safety Effect of A Bicycle Facility. *Accident Analysis & Prevention*, Vol. 32, No. 2, pp. 589-599, 2000.

Lee, J., and F. L. Mannering, F.L., 2002. Impact of roadside features on the frequency and severity of run off-road accidents: an empirical analysis. *Accident Analysis & Prevention*, Vol. 34, No. 2, pp. 349–361, 2002.

Li, W., A. Carriquiry, M. Pawlovich, and T. Welch, The Choice of Statistical Models in Road Safety Countermeasures Effectiveness Studies in Iowa. *Accident Analysis & Prevention*, Vol. 40, No. 4, pp.1531–1542, 2008.

Lovegrove, G., and T. Sayed, Macro-level Collision Prediction Models for Evaluating Neighbourhood Traffic Safety. *Canadian Journal of Civil Engineering*,Vol. 33, No. 5, pp. 609-621, 2006a.

Lovegrove, G., and T. Sayed, Using Macro-level Collision Prediction Models in Road Safety Planning Applications. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1950, pp. 73–82, 2006b.

Lovegrove, G. *Road Safety Planning: New Tools for Sustainable Road Safety and Community Development*, Verlag Dr. Müller, Germany, 2007.

Lovegrove, G., C. Lim, and T. Sayed, Community-Based, Macro-level Collision Prediction Model Use with a Regional Transportation Plan, *ASCE Journal of Transportation Engineering*, Vol. 136, No. 2, pp. 120-129, 2010.

Lovegrove, G. Review of Canadian Promising Practices in Promoting Safe Use of Roads and Pathways for Vulnerable Road Users, Prepared for Public Health Agency of Canada, 2012.

Ladron de Guevara, F., S. P. Washington, and J. Oh, Forecasting Collisions at the Planning Level: A Simultaneous Negative Binomial Collision Model Applied in Tucson, Arizona. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1897, pp. 191–199, 2004.

Lord, D., and B. N. Persaud, Estimating the Safety Performance of Urban Road Transportation Networks. *Accident Analysis & Prevention*, no. 36, pp. 609-620. 2004.

Marshall, W. E., and N. W. Garrick, Evidence on Why Bike-Friendly Cities Are Safer for All Road Users. *Environmental Practice*, Vol. 13, No.1, pp. 16-26, 2011.

McCullagh, P., and J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, New York, 1989.

Miaou, S. P. and H. Lum, Modelling Vehicle Accident and Highway Geometric Design Relationships, *Accident Analysis & Prevention*, Vol. 25, Iss. 6, pp. 689-709, 1993.

Miaou, S. P., The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions. *Accident Analysis & Prevention*, Vol. 26, No. 4, pp. 471-482, 1994.

Miaou, S. P. *Measuring the Goodness of Fits of Accident Prediction Models*. Report FHWA-RD-96-040. FHWA, U.S. Department of Transportation, McLean, Va., 1996.

Miaou, S. P., and D. Lord, Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes versus Empirical Bayes Methods. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1840, pp. 31-40, 2003.

Miller, J. S., Garber, N. J., Kamatu, J.N. 2010. *Incorporating Safety into The Regional Planning Process in Virginia: Volume I: Development of a Resource*. Publication VTRC-10-R14, Virginia Transportation Research Council.

Miranda-Moreno, L. F., L. Fu, F. Saccomanno, A. Labbe, Alernative Risk Models for Ranking Locations for Safety Improvement. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1908, pp. 1-8, 2005.

Moudon, A.V., and C. Lee, Cycling and the Built Environment, A US Perspective. *Transportation Research Part D*, Vol. 10, Iss. 3, pp. 245-261, 2005.

Newman, P., and J. Kenworthy, *Sustainability and Cities: Overcoming Automobile Dependence*. Island Press, Washington, D.C., 1999.

Norden, M., J. Orlansky, H. Jacobs, Application of Statistical Quality Control Techniques to Analysis of Highway Accident data, *Highway Research Board*, Bulletin, No. 117, 1956.

Office of Traffic Safety, Alberta Traffic Safety Plan, 3-yr 2007-1010 Action Plan, Alberta Infrastructure and Transportation, 2007.

Oh, J., C. Lyon, S. Washington, B. Persaud, and J. Bared, Validation of FHWA Crash Models for Rural Intersection Lessons Learned. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1840, pp. 41-49, 2003.

Oppe, S. *Detection and Analysis of Black Spots with Even Small Accident Figures*, SWOV, Leidschendam, The Netherlands, 1982.

Oppe, S. A Comparison of Some Statistical Techniques for Road Accident Analysis, *Accident Analysis & Prevention*, Vol. 24, No. 4, pp. 397-423, 1992.

Osberg, J. S. and S. C. Stiles, Bicycle Use and Safety in Paris, Boston, and Amsterdam. *Transportation Quarterly Fall*, Vol. 52, No. 4, pp. 61-76, 1998.

Persaud, B., B. Lan, C. Lyon, and R. Bhim, Comparison of Empirical Bayes and Full Bayes Approaches for Before-After Road Safety Evaluations. *Accident Analysis & Prevention*, Vol. 42, No. 1, pp. 38-43, 2010.

Pucher, J., and L. Dijkstra, Promoting Safe Walking and Cycling to Improve Public Health: Lessons from the Netherlands and Germany. *American Journal of Public Health*, Vol. 93, Iss. 9, pp. 1509-1516, 2003.

Pucher, J. and Buehler. R. 2006. Why Canadians Cycle More Than Americans: A Comparative Analysis of Bicycling Trends and Policies. *Transport Policy*, Vol. 13, Iss. 3, pp. 265-279, 2006.

Qin, X., J. N. Ivan, and N. Ravishankar, Selecting Exposure Measures in Crash Rate Prediction for Two-lane Highway Segments. *Accident Analysis & Prevention*, Vol. 36, No. 2, pp. 183–191, 2004.

Qin, X., J. N. Ivan, N. Ravishanker, and J. Liu. Hierarchical Bayesian Estimation of Safety Performance Function for Two-lane Highways using Markova Chain Mote Carlo modeling. *ASCE Journal of Transportation Engineering*, Vol. 131. No. 5, pp. 345-351, 2005.

Reurings, M., and H. L. Stipdonk, Estimating the Number of Serious Road Injuries in the Netherlands. *Annals of Epidemiology*, Vol. 21, Iss. 9, pp. 648-653, 2011.

Reynolds, C., M. A. Harris, K. Teschke, et al. The Impact of Transportation Infrastructure on Bicycling Injuries and Crashes: A Review of the Literature. *Environmental Health*, Vol. 8, pp. 47, 2009.

Rietveld, P., and V. Daniel, Determinants of Bicycle Use: Do Municipal Policies Matter? *Transportation Research Part A*, Vol. 38, Iss. 7, pp. 531-550, 2004.

Robinson, D. L., Safety in Numbers in Australia: More Walkers and Bicyclists, Safer Walking and Bicycling. *Health Promotion Journal of Australia*, No.16, pp. 47–51, 2005.

Rodegerdts, L. A., J. Ringert, P. Koonce et al. *Signalized Intersections: Informational Guide*, Publication No.FHWA-HRT-04-091, Federal Highway Administration, 2004.

Sawalha, Z., and T. Sayed, Evaluating Safety of Urban Arterial Roadways. *ASCE Journal of Transportation Engineering*, Vol. 127, No. 2, pp. 151-158, 2001.

Sawalha, Z. and T. Sayed, Traffic Accident Modelling: Some Statistical Issues. *Canadian Journal of Civil Engineering*, No. 33, pp.1115-1123, 2006.

Sayed, T. *New and Essential Tools: Part 1: Selecting Sites for Treatment*. Seminar on New Tools for Traffic Safety, Section 2, Institute of Transportation Engineers, August 9th, Toronto, Canada, pp.1-13, 1998.

Sayed, T. *ICBC Consultant Training*, prepared for ICBC, 1999.

Sayed, T., and F. Rodriguez, Accident Prediction Models for Urban Unsignalized Intersection in British Columbia. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1665, pp. 93-99, 1999.

Sayed, T., and P. de Leaur, *Program Evaluation Report: Road Improvement Program*. Prepared for the Insurance Corporation of British Columbia, ICBC, North Vancouver, Canada, 2001.

Schwarz, G., Estimating the Dimension of a Model. *Analysis of Statistics*, Vol. 6, No. 2, pp. 461-464, 1978.

Shankar, V.N., G.F. Ulfarsson, R. M. Pendyala, M. B. Nebergall, Modeling Crashes Involving Pedestrians and Motorized Traffic, *Safety Science*, Vol. 41, pp. 627–640, 2003.

Smith, R., D. Harkey, and B. Harris. *Implementation of GIS-Based Highway Safety Analysis: Bridging the Gap.* FHWA-RD-01-039. FHWA, U.S. Department of Transportation, 2001.

Spiegelhalter, D. J.,N. Best, B. P. Carlin, A. van der Linde, 2002. Bayesian Measures of Model Complexity and Fit . Journal of the Royal Statistical Society, Series B (Statistical Methodology) Vol. 63, Iss. 4, pp. 583–639, 2002.

Spiegelhalter, D. J., A. Thomas, N. Best, D. Lunn, *WinBUGS User Manual*. MRC Biostatistics Unit, Cambridge. 2005.

Stallard, S. (Ed.) *Bike Sense: the British Columbia Bicycle Operator's Manual*, 4th Edition he Greater Victoria Cycling Coalition, Victoria, BC, Canada, 2006.

Stinson, M.A. and C. R. Bhat, Frequency of Bicycle Commuting: Internet-based Survey Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1878, pp. 122–130, 2004.

SWOV(Institute for Road Safety Research). From Car to Bicycle: Road Safety Effects, *SWOV Research Activities*, No. 45, pp. 7-8, 2010.

Thacker, S. B., and R. L. Berkelman. Public Health Surveillance in the United States. *Epidemiologic Reviews*, Vol. 10, pp. 164-190, 1988.

Transport Canada, *Trends in Motor Vehicle Traffic Collision Statistics 1988-1997*, Report TP13743E, 2001.

Transport Canada, *Vulnerable Road User Safety: A Global Concern*, Report TP2436E, 2004.

Transport Canada, *Road Safety Vision: Annual Report*, TP 13347 E, 2005.

Transportation Association of Canada. *Bikeway Traffic Control Guidelines for Canada, 2nd Edition*. Ottawa, Canada, 2010.

van Minnen, J., *The Suitable Size of Residential Areas: A Theoretical Study with Testing to Practical Experiences*. SWOV Report No. R-99-25, Leidschendam, Netherlands, 1999.

Volk, K., E. Felipe, G. Ho, and M. Guarnaschelli, *Developing a Road Safety Module for the Regional Transportation Model*, ICBC, Vancouver, Canada, 1999.

Vuong, Q., Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses. *Econometrica*, Vol. *57*, No. 2, pp. 307-333, 1989.

Wang, T. *Incorporating Safety into Transportation Planning for Small and Medium-sized Communities*, Master Thesis, Iowa State University, USA, 2011.

Washington, S., I. V. Schalkwyk, M. Meyer, E. Dumbaugh, and M. Zoll, *Incorporating Safety into Long-Range Transportation Planning*, NCHRP Report 546, Washington D.C., Transportation Research Board, 2006.

Wegman, F., A. Dijkstra, G. Schermers, and P. V. Vliet, Sustainable Safety in the Netherlands: Evaluation of National Road Safety Program, *Transportation Research Record: Journal of the Transportation Research Board*, Vol.1969, pp. 72-78, 2006.

Williams, J. and J. Larson, Promoting Bicycle Commuting: Understanding the Customer. *Transportation Quarterly*, Vol. 350, No. 3, pp. 67-78, 1996.

Winters, M., M.C. Friesen, M. Koehoorn, and K. Teschke. 2007. Utilitarian Bicycling, a Multilevel Analysis of Climate and Personal Influences. *American Journal of Preventive Medicine*, Vol. 32, Iss. 1, pp. 52-58, 2001.

World Health Organization, *World Report on Road Traffic Injury Prevention, Summary*. Geneva, Swiss, 2004.

World Health Organization, *Fact Sheet of Top 10 Causes of Death*. Geneva, Swiss, 2008.

Zegeer, C. V. *Highway accident analysis systems*, NCHRP, 91, Transportation Research Board, Washington, D.C. 1982.

APPENDICES

Appendix A. Collinearity (Correlation) Test Sample of Explanatory Variables

For the NB bike collision model in the urban, measured, exposure group

Correlations between parameter estimates

Parameter	**ref**	**correlations**				
Constant	1	1.000				
TLKM	2	-0.636	1.000			
SIG	3	-0.171	-0.104	1.000		
INTD	4	-0.686	0.193	0.036	1.000	
IALP	5	-0.407	0.169	-0.276	0.125	1.000
		1	2	3	4	5

Appendix B. SAS Code Sample of ZIP Model Development

```
% include "C:\Users\fwei\Documents\My SAS Files\vuong.sas";
proc genmod data = Tmp1.Cord;
  model B5 = LNtlkm /dist=zip;
  zeromodel LNtlkm /link = logit ;
  estimate 'intercept' intercept 1   / exp;
  output out=Vw.B5_expzip pred=predzip pzero=p0;
run;
proc genmod data = Vw.B5_expzip;
  model B5 = LNtlkm /dist=nb;
  output out=Vw.out pred=prednb;
  estimate 'intercept' intercept 1   / exp;
run;
%vuong(data=Vw.out, response=B5,
      model1=zip, p1=predzip, dist1=zip, scale1=1, pzero1=p0,
      model2=nb,   p2=prednb,   dist2=nb,  scale2=1.6408,
      nparm1=4,nparm2=2)
```

Appendix C. WinBUGS Samples of Model File, Data File, and Data Initial File

For the FB total collision model in the urban, measured, exposure group

Model file Sample % in *.txt format

```
model {
    for( i in 1 : N ) {
                        t[i] ~ dpois(lambda2[i])
                        mu[i] ~ dnorm(0.0, tau)
                        log(lambda[i]) <- a0 + a1*tlkm[i] + mu[i]
                            lambda2[i]<-max(lambda[i], 0.0001)
        }
        a0 ~ dnorm(0.0,1000)
        a1 ~ dnorm(0.0,1000)
        tau ~ dgamma(1, 0.1);
        sigma <- 1/sqrt(tau);
    }
```

Data file sample % in *.txt format

```
list(N=230)
t[]     tlkm[]
5       1.496224
0       0.675034
11      0.33961
……..
3       -0.272859
5       -0.010454
0       -2.175952
1       -1.49388
END
```

Data initial file sample % in *.txt format

```
list(
a0 =2.5, a1 = 0.5, tau = 1,
mu = c(0.001, 0.001, 0.001, 0.001, ……, 0.001, 0.001))
```