AN EMPIRICAL TOOL TO EVALUATE THE SAFETY OF CYCLISTS: COMMUNITYY BASED, MACRO-LEVEL COLLISION PREDICTION MODELS USING NEGATIVE BINOMIAL REGRESSION

Feng Wei, MSc, PhD student
Graduate Research Assistant, School of Engineering, Faculty of Applied Science, UBC

Dr. Gordon Lovegrove, P.Eng., MBA, PhD (Corresponding Author)
Assistant Professor, School of Engineering, Faculty of Applied Science
University of British Columbia, Kelowna, BC Canada
Tel: +1.250.807.8717, Email: gord.lovegrove@ubc.ca

Submitted to the 3rd International Conference on Road Safety and Simulation, September 14-16, 2011, Indianapolis, USA

ABSTRACT

Today, North American governments are more willing to consider compact neighbourhoods with sustainable transportation. Bicycling, one of the most effective modes for short trips with distances less than 5 kilometres, is encouraged widely in our neighbourhoods. However, as vulnerable road users (VRUs), cyclists are more likely to be injured in collisions. In order to create a safe road environment for them, evaluating cyclists' road safety at a *macro* level in a proactive way is necessary. In this paper, different generalized linear regression methods for collision prediction model (CPM) development are reviewed and previous studies on *micro*-level and *macro*-level bicycle-related CPMs are summarized. On the basis of insights gained in the exploration stage, this paper also reports on efforts to develop negative binomial models for bicycle-auto collisions at a community-based, macro-level. Data came from the Central Okanagan Regional District (CORD), of British Columbia, Canada. The model results revealed several statistical associations between collisions and explanatory variables: 1) an increase in bicycle-auto collisions is associated with an increase in each of total lane kilometres (TLKM), bicycle lane kilometres (BLKM), bus stops (BS), traffic signals (SIG), intersection density (INTD), and arterial-local intersection percentage (IALP), an intuitive result; 2) Surprisingly, an increase in each of drive commuters (DRIVE) and drive commuter percentage (DRP) were found to be associated with a decrease in bicycle collisions, somewhat counterintuitive. One possible reason is that these models were developed in a North American community with low bicycle use (< 4%). To test this hypothesis and to further explore the statistical relationships between bicycle mode split and overall safety, in future, macro-level CPMs for communities with medium and high bicycle use will also be pursued.

Key words: bicycle safety, cyclists, macro-level collision prediction models, generalized linear regression, negative binomial regression, and bicycle use levels.

1. INTRODUCTION

Emerging global problems of climate change, peak oil, traffic congestion, and road safety are forcing governments to consider ways to encourage sustainable transportation systems. Sustainable transportation systems, including walking, bicycling, public transit, green vehicles, and car sharing, make more positive contributions to the society, economy and environment than automobile-dominated transportation systems. Bicycling, featuring low costs, zero-emissions, non-carbon-based fuel use, health benefits, and convenient parking, is one of the most effective modes for short trips with distances less than 5 kilometres. However, as vulnerable road users (VRUs), cyclists are more likely to be injured in traffic collisions. Research programs about VRUs' safety assessment at road intersections have been undertaken for many years (e.g. Ekman, 1996; Leden et al., 2002; Grey et al., 2010). Most research programs addressed the impacts of intersection traffic volume and geometric design on VRUs' the safety. Although these reactive road safety measures are efficient in improving the safety level of existing road facilities, they lack of the early-planning ability and are unable to measure the safety impacts of non-motorized mode split on the whole transportation system. In order to fill these gaps and encourage more bicycles in our communities, it is necessary to develop reliable methodology to evaluate the road safety of VRUs in a proactive way.

This paper presents a modeling technique to predict bicycle collisions at the community-based, macro-level, which can be used as reliable science-based decision aid tools by community planners and engineers. Three objectives of this paper are summarized here: (1) to conduct a literature review on micro-level and macro-level bicycle collision prediction models (CPMs); (2) to use negative binomial (NB) regression to develop community-based, macro-level bicycle CPMs based on data of Central Okanagan Region District (CORD), BC, Canada; and (3) to discuss some statistical and data issues in the model development, which will be covered in future. This study is a part of the research program about quantifying road safety benefits of increasing bicycle use. According to empirical observations, two hypotheses describing the relationships between bicycle mode split and safety are derived. One hypothesis is for the overall safety, assuming that increasing bicycle use may lead to a significant reduction in total traffic collisions (see the red line in Figure 1); the other one is for the VRUs' safety, supposing that a beginning increase in bicycle use from the low level to the medium level may cause an increase in bicycle collisions, but with the continuous increase in bicycle use from the medium to the high level, the bicycle collisions would drop down (see the blue line in Figure 1). Although these hypotheses are derived from empirical observations, they still need to be tested by reliable empirical tools. Therefore, macro-level collision prediction models (CPMs), as an empirical tool to evaluate road safety and test the hypotheses, are researched this program.

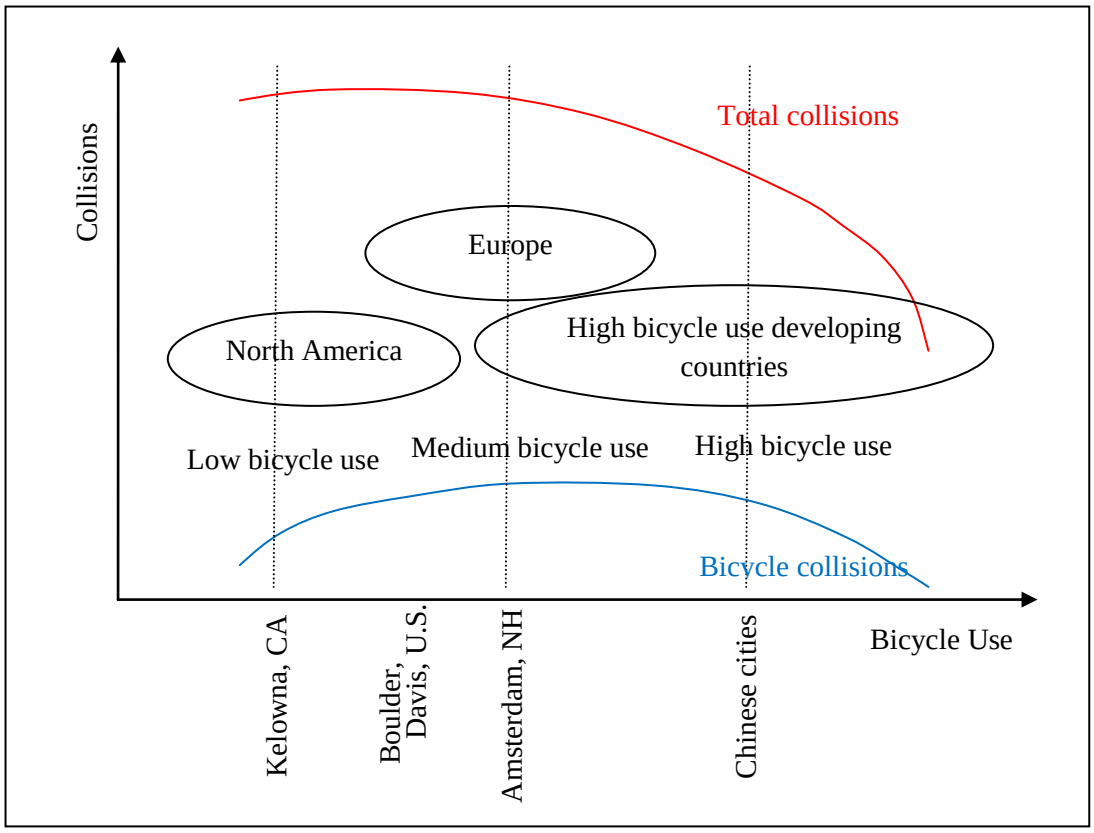

Figure1 Hypothesis on bicycle use and Overall level of road safety

2. LITERATURE REVIEW

2.1 Generalized Linear Regression Approaches for CPMs

In previous CPM studies, generalized linear models (GLMs) are commonly used and proved successfully as they could effectively model the rare, random, sporadic, and non-negative collision data. The generalized linear regression methods for CPM development mainly include Poisson regression and its various extensions such as zero-inflated Poisson regression, Poisson gamma regression; and Poisson lognormal regression.

2.1.1 Poisson and ZIP Regression

Miaou et al. (1992) found that the Poisson regression approach was more effective to predict truck collisions when compared to the regular linear regression techniques. However, Poisson regression assumes that the mean value equals to the variance value, which is not consistent with the over-dispersion of collision data. Therefore, several other regression techniques based on Poisson regression were proposed. The zero-inflated Poisson model (ZIP) is one extension of the Poisson model. It is used to solve the issue of "excess zeros" that can characterize collision data (Shankar et al., 2003; Qin et al., 2004; Kumara and Chin, 2003; Lee and Mannering, 2002). ZIP models assume a dual-state process which is responsible for generating collision data. The first process generates only zero counts and the second process generates non-zero counts from a Poisson model. The empirical results from related studies show that ZIP regression was more

promising for providing explanatory insights into the causality behind collisions than Poisson regression (Shankar et al., 2003; Qin et al., 2004; Kumara and Chin, 2003; Lee and Mannering, 2002).

2.1.2 NB Regression

The second extensional approach for developing CPMs uses Poisson-gamma hierarchy, also called negative binomial (NB) regression. This regression specifically accounts for extra Poisson variation of collisions, and is widely used in many studies for both micro and macro-level CPMs. Model results from these studies demonstrated that an NB model was superior to a Poisson model (Miaou & Lord, 2003; Lord, 2006; Sawalha & Sayed, 2006; Hadayeghi et al., 2006; Lovegrove &Sayed, 2006; Ladron de Guevara et al., 2004). The formulations for NB regression are presented as:

$$Y_i|E_i \sim Poisson(E_i) \quad (1)$$

$$E_i = \mu_i \quad (2)$$

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} \quad (3)$$

$$E_i \sim Gamma(\lambda_i, \kappa_i) \quad (4)$$

where, Y_i = the number of collisions at location i; E_i , λ_i, κ_i= the distribution parameters, X_{ij} = explanatory covariates. So in Equation 1, the observed number of collisions at location i, Y_i, follows a Poisson probability distribution with the parameter of expected number of collisions, E_i. And the parameter E_i, seen as another random variable, is presented in Eq. 3 and assumed to follow a Gamma distribution with parameters λ_i and κ_i(see Eq.4). Under the NB model, the mean and the variance are presented in Eq. 5 as:

$$E(Y_i) = \mu_i \quad and \quad Var(Y_i) = \mu_i + \frac{{\mu_i}^2}{\kappa_i} \quad (5)$$

2.1.3 PLN Regression

Poisson lognormal regression (PLN) model also can be reflective of the extra-Poisson variations (Kim et al., 2002; El-Basyouny & Sayed 2009). In PLN models, the collision data still follow a Poisson distribution (presented in Eq. 6); however, the parameter of Poisson distribution ----the mean value of collisions, E_i, should be derived from an exponential function (Eq. 7), in which the variable u_i follows a lognormal distribution (shown in Eq. 9). The formulations for PLN regression are presented in Eq. 6-9:

$$Y_i|E_i \sim Poisson(E_i) \quad (6)$$

$$E_i = \mu_i \exp(u_i) \quad (7)$$

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} \quad (8)$$

$$\exp(u_i) \sim Lognormal(0, {\sigma_u}^2) \quad or \quad \mu_i \sim N(0, {\sigma_u}^2) \quad (9)$$

where, the term $exp(u_i)$ represents a multiplicative random effect. Kim et al. (2002) found that both Poisson-gamma and Poisson lognormal can provide more reasonable collision predictions and account for extra-Poisson variations; also, their comparison results indicate that the difference of analysis results between these two methods are negligible. In PLN models, the mean and the variance of collision counts are depicted as:

$$E(Y_i) = \mu_i \exp(0.5{\sigma_u}^2) \quad and \quad Var(Y_i) = E(Y_i) + [E(Y_i)]^2(\exp({\sigma_u}^2 - 1)) \quad (10)$$

2.2 Micro- and Macro-level Models for Bicycle Collisions

Regression models for collision prediction can be divided into two types according to the study area. Study area may be micro-level locations (e.g. intersections or road segments) or macro-level regions (e.g. neighbourhoods, cities, or districts). Although micro-level CPMs have been widely researched for a long period and gradually formed a mature technical system, they are only used in reactive road safety improvement programs (RSIPs) (e.g. improve the safety performance of existing road facilities). However, macro-level CPMs are not only used in reactive RSIPs, but also can be a road safety planning tool in proactive RSIPs. (de Leur & Sayed, 2003; Lovegrove & Sayed, 2006). Therefore, macro-level CPMs can allow engineers and planners to target the road safety level at an early planning stage, with potential for significant reductions in collision frequencies below that achieved to date using reactive techniques. De Leur & Sayed (2003) indicated that if road safety was addressed as one of the evaluation factors before a project is built, it could reduce the number and cost of reactive safety countermeasures that have to be retrofitted into existing communities. Lovegrove (2007) suggested that lower-cost road safety planning using macro-level CPMs in the long term might be a more effective and sustainable road safety engineering approach than reactive safety improvement measures using micro-level CPMs. Next, previous studies on bicycle CPM development are summarized.

2.2.1 Micro-level Bicycle CPMs

Brude and Larsson (1993) collected data at 377 intersections with more than 100 pedestrian or cycle movements per annual average day, from 30 towns in Sweden. Their model forms were power functions with two leading variables as bases: the incoming motor vehicles (AADT) and the number of passing cyclists per average annual day. The least squares method was used to estimate model parameters. Results showed that the risk to cyclists (i.e. the number of collisions involving cyclists per million passing cyclists) increased with increased motor traffic flow, but decreased with increased cyclist flow. Turner & Francis (2003) developed micro-level CPMs with Poisson or NB regression methods for pedestrian and cyclists, based on data from three cities in New Zealand. Their research objective was to estimate the likely changes in collision frequencies and collision rates due to a mode shift from motor vehicle trips to pedestrian or cycle trips. Study results indicated that the overall pedestrian and bicycle collision *frequencies* increased with increased VRU flows, but that collision *rates* per pedestrian and per cyclist decreased with increased VRU flows.

2.2.2 Macro-level Bicycle CPMs

From the macro-level aspect, Jacobson (2003) examined the relationship between the numbers of pedestrians or cyclists (which are called VRUs here) and their collisions with motor vehicles based on five data sets from all over the world. For each data set, the measure of VRU injuries was determined by a power function with the measure of walking or bicycling as an explanatory variable. The model parameters were estimated by the least squares analysis. Results showed that the number of VRU collisions would increase at roughly 0.4 power of the measure of people walking or bicycling. For example, a community doubling its bicycle use could expect a 32% increase in injuries ($2^{0.4} = 1.32$). Although the VRU injury frequency increased with increases in walking and bicycling measures, the probability that a motorist might collide with an individual VRU (i.e. injury rate) would decline with the roughly 0.6 power of the measure of people walking or cycling. Robinson (2005) reviewed three datasets in Australia, and verified that Australian data also produced similar results to Jacobsen's model.

Lovegrove (2007) developed a series of community-based, macro-level CPMs using NB regression for the Greater Vancouver Regional District (GVRD) in BC, Canada. A unique bicycle-auto collision model for rural area was found, revealing that increased bicycle collisions were associated with increased bicycle mode share in rural areas. While this result was intuitive, Lovegrove (2007) also indicated that an association between bicycle use and total collisions (i.e. the sum of bicycle, pedestrian, and vehicle collisions) was not revealed and needed further research. Kim et al. (2010) examined the relationships between different types of collisions (i.e. total/injury/fatal/pedestrian/bicycle collisions) and independent variables in demographic, land use, and roadway accessibility fields in Honolulu. A binary logistic regression was chosen to model such relationships after failures using Poisson and NB regression. The results from bicycle collision models suggested that demographic variables such as job count and the number of people living below the poverty level were significant and positively associated with bicycle collisions; accessibility variables such as the number of bus stops, the bus route length, and the number of intersections were also positively associated with bicycle collisions.

3. METHODOLOGY

Traditional macro-level CPMs mainly focus on motor vehicle collisions, but few on bicycle collisions. The methodology used for predicting bicycle collisions in this study was derived from previous community-based, macro-level CPM studies using NB regression (Lovegrove, 2007; Hadayeghi et al., 2003; Ladron de Guevara et al., 2004), with updates to fit bicycling characteristics. In the process of model development, several inherent problems related to model regression method, model forms, variable selection, and model tests have been addressed.

3.1 Negative Binomial Regression &Model Form

The NB regression method has been mentioned previously. It is a discrete distribution with the mean and variance values given in Eq. 5, where κ is a positive constant known as the dispersion parameter. In this study, bicycle CPMs with NB regression were developed using GenStat, a general statistics software package. Previous studies (Lovegrove, 2007; Hadayeghi et al., 2003; Ladron de Guevara et al., 2004) proposed a generalized linear model form for motor collisions, which is presented as:

$$E = a_0 Z^{a_1} e^{\sum b_j X_j} \tag{11}$$

Where E = the predicted collision frequency (over 3 years for motor collisions); a_0, a_1, b_j = model parameters; Z = leading exposure variables (i.e. VKT-vehicle kilometres for modeled data or TLKM-total lane kilometres for measured data; and, X_j = other explanatory variables. This form not only takes account of the influence weights of different independent variables but also meets the non-negative, nonlinear, and non-normal collision features.

Based on the general form, four possible model forms predicting bicycle collisions were proposed at the early time. Then, related datasets would be used to examine the statistical performance of these model forms to find which form(s) appear promising. These possible forms were given as

$$E_B = a_0 e^{a_1 B + a_2 Z + \sum b_i X_i} \qquad \text{(Model form 1)}$$

$$E_B = a_0 Z^{a_2} e^{a_1 B + \sum b_i X_i} \qquad \text{(Model form 2)}$$

$$E_B = a_0(B+1)^{a_1} e^{a_2 Z + \sum b_i X_i} \quad \text{(Model form 3)}$$
$$E_B = a_0(B+1)^{a_1} Z^{a_2} e^{\sum b_i X_i} \quad \text{(Model form 4)}$$

where, E_B= the predicted bicycle collision frequency over 5-year period; B = leading exposure variables of bicycle use (i.e. BLKM-bicycle lane kilometres meters). 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). In the Model form 3 and 4, the leading variable is set as ($B+1$) instead of B to avoid the zero logic error (i.e. zero 'B' leads to zero 'E_B') as bicycle collisions could happen in the locations without bicycle lanes. In the Model form 1, 2, and 3, set the variable Z or B as a less-influenced variable instead of a leading variable because bicycle collisions may not be strongly influenced by traffic exposures in communities with a low bicycle use (e.g. North America).

3.2 Variable Selection

To develop each CPM, selecting significant variables from numerous candidate variables is a critical first step. The variable selection is a forward stepwise procedure by which all candidate variables were added to a model one by one. Sawalha and Sayed (2006) recommended that the first variable to be added should be the leading exposure variable(s) due to its dominating prediction influence. In this study, the leading variable included: TLKM and BLKM. The decision to keep a variable in the model was based on it meet all four of the following criteria: (1) the logic (i.e. +/-) of the estimated parameter was intuitively associated with collisions; (2) the t-statistic for each parameter was significant at the 95% confidence level (i.e. >1.96); (3) the added variable had little or no correlation (i.e. < 0.3) with any other independent variables in the same model; and (4) the added variable should make a significant drop in scaled deviance (SD) at 95% confidence level (i.e. >3.84). In this way model over-fitting was also avoided (McCullagh & Nelder, 1989; Sawalha & Sayed, 2006; Lovegrove, 2007).

3.3 Model fit tests & Outlier Analysis

As each candidate variable was added, the model fit was re-assessed. Scaled Deviance (SD) and Pearson χ^2 have been two common goodness of fit measures for Poisson or NB regression (Sawalha & Sayed, 2006; Lovegrove, 2007), defined as follows:

$$SD = 2\sum_{i=1}^{n}\left[y_i ln\left(\frac{y_i}{E(\Lambda_i)}\right) - (y_i + \kappa) ln\left(\frac{y_i+\kappa}{E(\Lambda_i)+\kappa}\right)\right] \quad (12)$$

$$Pearson\chi^2 = \sum_{i=1}^{n}\frac{[y_i - E(\Lambda_i)]^2}{Var(y_i)} \quad (13)$$

where, y_i and $E(\Lambda_i)$ are observed and predicted collision frequency in location i, respectively; $Var(y_i)$ is the variance for location i; and κ is the over dispersion parameter for the model (an output of the GLM regression analysis). Both SD and Pearson χ^2 approximate a χ^2distribution with (n-p) degrees of freedom, where p is the number of parameters. In this case study, a 95% confidence level was always used, which meant as long as the SD and Pearson χ^2values of any model were smaller than $\chi^2(0.05, n\text{-}p)$, this model was seen as qualified.

If a test was not successful, outlier analysis was undertaken. In GenStat, three indicators were available for defining outliers: residual, leverage and Cook's distance (CD) value. Following from previous studies (Sawalha & Sayed, 2006; Lovegrove, 2007), Cook's distance was used in

this case. Cook's distance uses leverage values and Pearson residuals. After a point with the largest CD value was removed, regression was re-run but with the dispersion parameter fixed to the value derived via the last model regression, and the change in SD was checked to see if it dropped greater than $\chi^2_{0.05,1} = 3.84$ (i.e. 95% level of confidence test). If this 95% test was passed, regression was re-run to provide: 1) new estimates for all parameters, 2) a new dispersion parameter, and, 3) a new series of CD measures for remaining data points. These steps were repeated until the reduction in SD was less than 3.84. Throughout, each variable's t-statistic were monitored to ensure that individual variables remained significant.

4. DATA DESCRIPTION & EXTRACTION

The study area, the Regional District of the Central Okanagan (RDCO) in the Province of British Columbia, Canada is shown in Figure 2. This area is roughly 44,000 hectares and comprised of 4 member municipalities (i.e. Kelowna, West Kelowna, Lake Country and Peachland). Based on Canada census data 2006, there were about 160,000 residents, 66,000 households, and 29,000 total lane kilometres in the RDCO (Statistic Canada, 2006). The major corridor through this region is Highway 97, which crosses the commercial areas of all four municipalities.

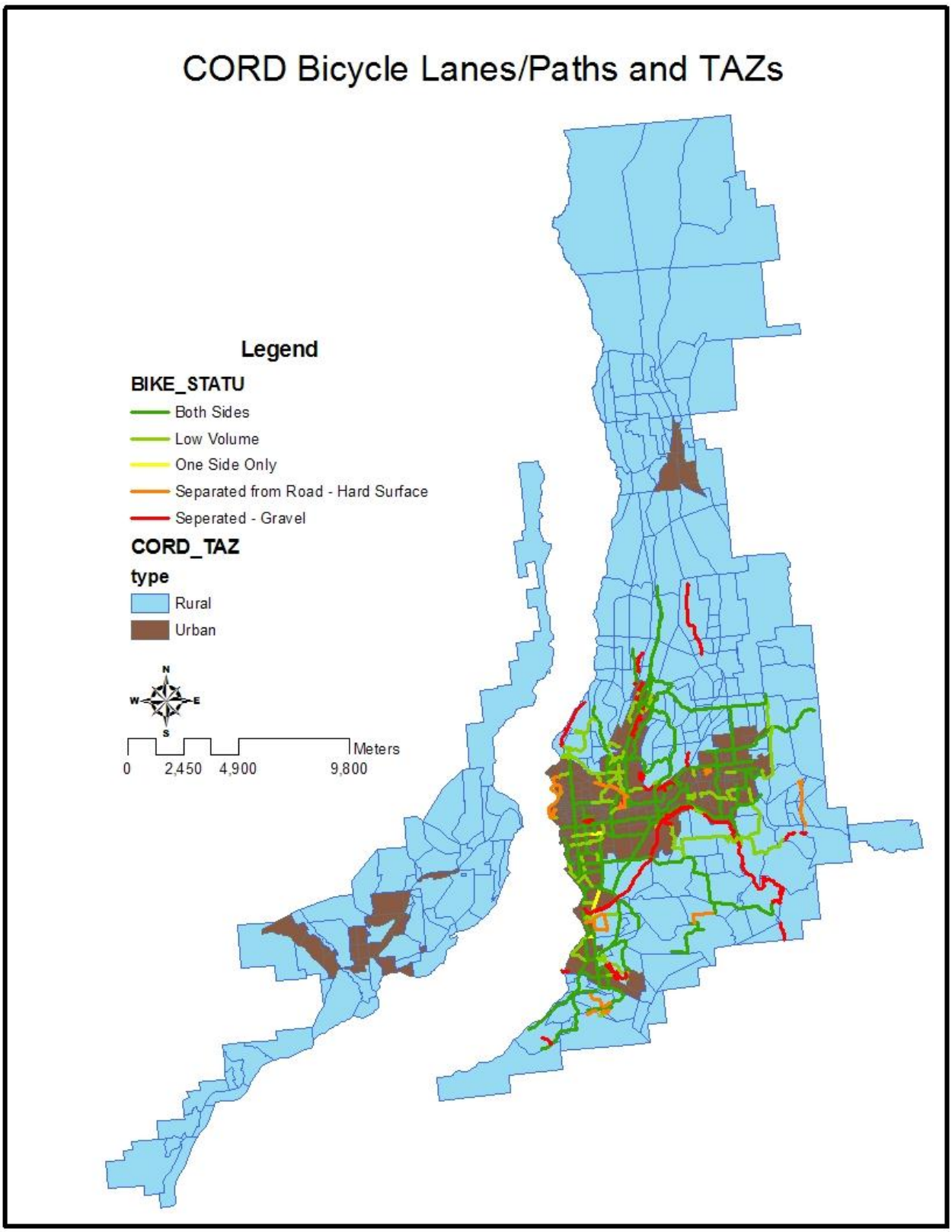

Figure 2: CORD bicycle lanes/paths and TAZs

For development of macro-level CPMs, collision data and all independent variable data needed to be aggregated into areal units. The aggregation units in the RDCO were 500 Traffic Analysis Zones (TAZs) derived from the 2005 RDCO transportation planning model. TAZs were chosen as aggregation units because their layouts would keep population and employment densities for each zone at a roughly uniform level. Also, most TAZ boundaries overlap the boundaries of census tracts or dissemination areas. In this way, data quality and relevance were maximized, and integration of disparate data sources was facilitated.

The collision data in this study was bicycle-auto collisions from 2002 to 2006, provided by the Insurance Corporation of British Columbia (ICBC). ICBC is a provincial crown corporation, and provides mandatory no-fault auto insurance to virtually all B.C. motorists. Five years of collision data was used, versus the usual three when studying vehicle collisions, because bicycle collisions are extremely rare events compared to vehicle collisions. As is well known, if the time period is too long (>> 3 to 5 years), a time trend bias is more likely to happen. Alternatively, a shorter time period (<<3 years) may introduce random extreme values away from the true long term mean and impair data quality.

Three criteria: variable themes, land use types and data derivations, were used for model stratification based on Lovegrove's research (2007). The objective of stratification was to make the models more specific and accurate so that the chances of causality-based, empirical relationships could be maximized. In Lovegrove's original research, models were stratified into 16 groups, as shown in Table 1. As modeled data (i.e. VKT, VC) for the RDCO was not available for this study, only odd group # models (i.e. using TLKM derived from measured data) were developed. Model variable definitions and their statistical summary are presented in Table 2.

Table 1 Model Groups (Lovegrove, 2007)

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 (TDM)	Urban	Modeled	9
		Measured	10
	Rural	Modeled	11
		Measured	12
Network	Urban	Modeled	13
		Measured	14
	Rural	Modeled	15
		Measured	16

Table 2 Variable Definitions & Data Summary (n=500 TAZs, urban=242, rural=258)

Variables	**Symbol**	**Source**	**Years**	**Zonal min**	**Zonal max**	**Zonal Avg.**
Bicycle/vehicle collisions	B5	ICBC	02-06	0	6	0.37
Exposure						
Total lane km		Census	2006	0.00	63.25	5.79
Total bicycle lane km	TLKM	RDCO	2006	0.00	7.12	0.70
Total bicycle lane km–off road	BLKMO	RDCO	2006	0.00	3.61	0.11
Total bicycle lane km–on road	BLKMF	RDCO	2006	0.00	6.39	0.58
Zonal Area (Hectares)		RDCO	2005	1.33	163.12	88.72
Socio-Demographics						
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 labor force (=(EMP+UNEMP)/POP15) (%)	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)) (%)	UNEMP P	Census	2006	0.00	19.15	5.04
Average income $	INCA	Census	2006	6100	69600	32000
Transportation Demand Management						
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
Road network						
No. of Signals	SIG	RDCO/GE	2006	0	4	0.3

Signal density	SIGD	RDCO/GE	2006	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: GE: Google Earth. POP15: Population aged 15 and over in 2006.						

5. MODEL RESULTS & DISCUSSION

In the process of model development, when reviewing the available data, several observations were made that should be noted. First, all of the models developed in this study predict only total collisions, and more specifically, only total bicycle-vehicle collisions. Further model predictions by collision types (i.e. bicycle-auto, bicycle-pedestrian, and bicycle-bicycle collisions), and/or by collision severities (i.e. fatality, injury, and property damage only collisions) were not pursued due to data limitations. Moreover, our preliminary data checks suggested that 95% of all bicycle-vehicle collisions are severe (i.e. involving injury and/or fatalities). Therefore this further stratification was left as a topic for future research, pending improved bicycle collision data. Second, only models in urban areas were developed, due to extremely low observed bicycle collisions are in rural areas. These low observed counts precluded establishment of significant statistical correlations. For example, in the RDCO's 258 rural TAZs, there were only 18 TAZs with 1 or 2 bicycle collisions recorded in five years. Third, about 90% of bicycle-vehicle collisions occurred at intersections, suggesting that how to improve VRU safety at intersection should also be a focus of future research. Fourth, 55% of bicycle-vehicle collisions in the RDCO occurred on roads with on-road bicycle lanes, versus only 45% on roads without bicycle lanes, suggesting that on-road bicycle lanes may not always have the intended safety benefit. Further research to differentiate this statistic by road class (i.e. arterial, collector, local) would be needed to verify this claim. Last, as only 16% of the RDCO total bicycle path length consisted of off-road bicycle paths (versus on-road), separated statistical associations were not possible between collisions and on-road bicycle lane length and off-road bicycle path length.

As mentioned above, four model forms were tested. Models of the first three forms were successfully developed in GenStat. Due to negligible differences among their sum of residuals, it was hard to tell which model form was better. Therefore, all of the first three model forms were considered to be acceptable, and a sample of several developed bicycle CPMs along with goodness of fit statistics has been given in Table 3.

Table 3 Model fit Comparisons in Different Model Forms

Model Form 1	**df**	κ	**SD**	**Pearson** χ^2	χ^2
Group 2	240	0.5773	198.0	235.2	277.1
Group 6	N/A				
Group 10	236	0.6993	197.2	235.5	272.8
Group 14	237	1.0693	190.5	227.3	273.9
Integrated Group	235	1.0751	190.3	225.3	271.8
ModelForm 2	**df**	κ	**SD**	**Pearson** χ^2	χ^2
Group 2	236	0.6095	193.7	240.6	272.8
Group 2	234	0.7563	192.4	244.2	270.7
Group 6	N/A				
Group 10	231	0.7809	194.5	235.6	267.5
Group 14	233	1.1892	189.5	196.1	269.6
Integrated Group	230	1.3547	187.7	212.0	266.4
ModelForm 3	**df**	κ	**SD**	**Pearson** χ^2	χ^2
Group 2	240	0.5742	196.8	238.4	277.1
Group 6	N/A				
Group 10	236	0.7047	197.5	250.5	272.8
Group 14	237	1.0688	190.4	227.0	273.9
Integrated Group	235	1.0743	190.1	225.8	271.8
Bicycle-auto CPM Examples					
Model form 1-Group2	$E_B = 0.5402e^{0.2582BLKM}$				
t- statistics	Con: -4.57, BLKM: 2.70				
Model form 2-Group10	$E_B = 6.598TLKM^{0.737}e^{-0.0408DRP+0.099BS-0.002375DRIVE}$				
t- statistics	Con: 2.62, TLKM: 4.39, DRP: -3.89, BS: 2.58, DRIVE:-2.46				
Model form 3-Group14	$E_B = 0.1224(BLKM+1)^{0.511}e^{0.01254ALP+2.19INTD+0.441SIG}$				
t- statistics	Con: -8.54, BLKM: 2.40, IALP: 4.84, INTD: 5.87, SIG: 3.21				
Model form 1-Integrated	$E_B = 0.1283e^{0.2331BLKM+0.01249IALP+2.141INTD+0.446SIG}$				
t- statistics	Con: -8.63, BLKM: 2.41, IALP: 4.80, INTD: 5.77, SIG: 3.26				

Model results revealed logical relationships between bicycle-vehicle collisions and explanatory variables. The direct associations between bicycle-vehicle collisions and traffic exposure — total lane kilometres (TLKM), and bicycle lane kilometres (BLKM), confirmed intuitive expectations. The direct relationship between collisions and bus stops (BS) was consistent with Kim's research (2010). Increases in signals (SIG) and intersection density (INTD) were also each associated directly with increases in collisions. Arterial-local intersection percentage (IALP) had a positive relationship with bicycle-vehicle collisions probably because arterial-local intersections have high traffic volumes, high speeds, and many conflicting un-signalized turning movements, potentially making high risks to cyclists. Inverse associations were observed between bicycle-vehicle collisions and drive commuters (DRIVE) and drive commuter percentage (DRP). As more commuters choose to drive, a low bicycle mode share would result, intuitively leading to fewer bicycle-vehicle collisions. This result could also demonstrate support for the bicycle safety

hypothesis, lower left end of the blue line in Figure 1. Models from the social demographic group could not be successfully developed, suggesting that perhaps social demographics do not play a large role in bicycle-vehicle collisions. On the other hand, the most statistically significant variable associations were found in the road network group, which suggested that bicycle-vehicle collisions are highly influenced by road network patterns, a topic left for further research to verify.

As discussed previously, scaled deviance and Pearson χ^2 were the two common statistics used for model fit tests, as recommended in the literature. However, previous studies suggest that for samples from a Poisson or NB distribution with low mean values, the scaled deviance and s may not perform very well. For example, SD performs better in large samples than small samples, and does not work well when there are many extreme observations (such as zeroes) (Maycock and Hall, 1984; Maher & Summersgill, 1996; Agrawal & Lord, 2006). In our case study, many observations were zeroes, as discussed earlier, suggesting that the use of an SD test may not be wise. A subsequent comparison of the developed CPM collision predictions versus observed bicycle-vehicle collisions was conducted, and verified that in this case, the models developed using these two methods were acceptable. However, this SD and Pearson χ^2 model test statistic issue should be addressed in future research to identify more appropriate tests if available.

On the issue of data quality, this paper presents initial results from a comprehensive research program just underway to test the hypothesis on the relationships between bicycle mode split and safety shown in Figure 1. The hypothesis stems from observations the researchers have made regarding bicycle use and road safety worldwide. It is necessary to consider reliable empirical tools to test and demonstrate this point of view. Using negative binomial models, this paper only presents bicycle collision prediction in communities with a low bicycle use, so collecting additional data in European and Chinese communities with medium/high bicycle use would be a critical next step to investigate and validate whether the relationship between safety and bicycle use in Figure 1 actually exists.

6. CONCLUSIONS

This paper proposed an empirical tool to estimate community-based, macro-level bicycle-vehicle collisions. First, it reviewed GLM regression methods for CPMs and summarized previous studies in bicycle collision prediction at micro and macro-level. Based on NB regression methods, several bicycle CPMs in urban areas were developed. The models revealed that bicycle-auto collisions had directly proportional relationships with total lane kilometres (TLKM), bicycle lane kilometres (BLKM), bus stops (BS), signals (SIG), intersection density (INTD), and arterial-local intersections (IALP); but had inverse relationships with drive commuters (DRIVE) and drive commuter percentage (DRP). Furthermore, this paper discussed model fit statistics and data issues in the research process that need to be addressed in future research.

To build (or rebuild!) a road environment that is safer for bicyclists in our auto-dominated North American culture, major bicycle infrastructure and facility investment is needed. Bicycle-vehicle, community-based, macro-level CPMs may be an effective empirical way to provide economic justifications of that bicycle-related investment, with predicted outcomes that allow for ongoing monitoring and validation of program success.

REFERENCES

Agawal, R., and Lord, D. (2006) "Effects of Sample Size on the Goodness of Fits Statistic and Confidence Intervals of Crash Prediction Models Subjected to Low Sample Mean Values". *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1950, 35-43.

BC Transit.(2006) *Geo-referenced Bus Stop & Bus Line Data of CORD,* (May, 2010).

Brude, U., and Larsson, J. (1993) "Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit?"*.Accident Analysis &Prevention,* Vol. 25,499-509.

City of Kelowna, (2009) *Geo-referenced Intersections of Kelowna, BC,* (March 4th, 2011).

City of Kelowna, (2009) *Geo-referenced Bicycle Lanes of Kelowna,* BC, (March6th, 2011).

De Leur, P., and Sayed, T. (2003) "A Framework to Proactively Consider Road Safety within the Road Planning Process". *Canadian Journal of Civil Engineering*, Vol. 30(4), 711-719.

DMTI Spatial Inc. (2006) CanMap Street files v2006.3.ABACUS Data Set, http://hdl.handle.net/10573/41234 (Dec. 7th, 2010).

Ekman, L. (1996) *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.

El-Basyouny, K., and Sayed, T. (2009) "Accident Prediction Models with Random Corridor Parameters". *Accident Analysis and Prevention*, Vol. 41, 1118-1123.

Grey, J., Raford, N., Ragland, D., and Pham, T. (2010) "Safety in Numbers: Data from Oakland, California". *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1982. Washington, D.C., 150-154.

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

Insurance Corporation of British Columbia (2006*) 1996-2006 Collision Geo-referenced Data for BC for UBC Research Only*. (2006)

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

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

Kim, K., Pant, P., and Yamashita, E. (2010) "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, 9-17.

Kumara S., and Chin, H.(2003) "Modeling Accident Occurrence at Signalized Tee Intersections with Special Emphasis on Excess Zeros". *Traffic Injury Prevention*, Vol.4 (1),53-57.

Ladron de Geuvara, F. L., Washington, S. P., and Oh, J. (2004) "Forecasting Crashes at the Planning Level: A Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona". *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1897, 191-199.

Leden, L., Garder, P., and Pulkkinen, U. (2000) "An Expert Judgment Model Applied to Estimate the Safety Effect of a Bicycle Facility". *Accident Analysis & Prevention,* Vol. 32, 589-599.

Lee, J. and Mannering, F. (2002) "Impact of Roadside Features on the Frequency and Severity of Run-off-roadway Accidents: an Empirical Analysis". *Accident Analysis and Prevention*, Vol. 34, 149-161.

Lord, D., Washington, S., and Ivan, J. (2005) "Poisson, Poisson-gamma, and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory". *Accident Analysis and Prevention*, Vol. 37, 35-46.

Lord, D. (2006) "Modeling Motor Vehicle Crashes using Poisson-Gamma Models: Examining the Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the fixed dispersion parameter". *Accident Analysis and Prevention*, Vol. 38, 751-766.

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

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

Maher, M.J. and Summersgill, I.A. (1996) "Comprehensive Methodology for the Fitting of Predictive Accident Models". *Accident Analysis & Prevention*, 28(3), 281-296.

Maycock, G. and Hall, R.D. (1984). *Accidents at 4-Arm Roundabouts*. Laboratory Report LR 1120. Crowthorne, Berks, U.K. Transport Research Laboratory.

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

Miaou, S.-P., Hu, P. S., Wright, T., Rathi, A. K., and Davis, S. (1992) "Relationships between Truck Accidents and Highway Geometric design: a Poisson Regression Approach". *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1376, 10-18.

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

Osberg, J.S. and Stiles, S.C. (1998) “Bicycle use and safety in Paris, Boston, and Amsterdam”. *Transportation Quarterly Fall* 52 (4), 61–76.

Qin, X., Ivan, J., and Ravishanker, N. (2004) “Selecting Exposure Measures in Crash rate Prediction for Two-Lane Highway Segments”. *Accident Analysis & Prevention*, 36, 183-191.

Robinson, D.L. (2005) “Safety in Numbers in Australia: More Walkers and Cyclists, Safer Walking and Bicycling”. *Health Promotion Journal of Australia*, Vol.16, 47–51.

Sawalha, Z., and Sayed, T. (2006) “Traffic Accident Modeling: Some Statistical Issues”. *Canadian Journal of Civil Engineering*, Vol. 33, 1115-1123.

Shankar, V., Ulfarsson, G., Pendyala, R., and Nebergall, M. (2003) “Modeling Crashes Involving Pedestrians and Motorized Traffic” *Safety Science*, Vol. 41(7), pp. 627-640.

Statistic Canada.(2006) *Census of Canada, Cumulative Profile and Release Components*, ABACUS Data Set, http://hdl.handle.net/10573/41852 (Dec. 7th, 2010).

Statistic Canada.(2006) *Census of Canada. Road Network and Geographic Attribute File*, ABACUS Data Set, http://hdl.handle.net/10573/41629 (Dec. 7th, 2010).

Turner, S.A., and Francis, T. (2003) *Predicting Accident Rates for Cyclists and Pedestrians*, Land Transport New Zealand Research Report 289, Land Transport New Zealand.

July 14, 2011

To members of the review committee,

RSS Paper No: 0068-000271

Paper Title: Collision Prediction Models Used for Evaluating the Safety of Cyclists at Community-based Level

Thank you again for your time and effort providing valuable comments to improve our paper. Below we give you our responses, and attached you will find our paper. We trust all is now in order. All the best.

Wei, F. & Lovegrove, Gordon, authors

Authors' response to reviewer comments:

Collision Prediction Models Used for Evaluating the Safety of Cyclists at Community-based LevelComments Field:

Reviewer 1 - In my comments I refer to page X paragraph Y line Z as px #y lZ. -----The authors present the problem of traffic safety for VRU and describe the micro and macro level of possible analysis (for example using the GLM). The subject is relevant for developing of a sustainable transport systems and I like the initiative of the authors. However, some aspects can be improved or further discussed.

Comment 1:I would like to start with the title. The paper deals with NB distribution but nothing is said about it in the title. It deals with the statistical aspects of creating a model but, again, that is missing. An important issue is to highlight that this study uses empirical data, which are not estimations or simulated. This should be underscored in the title. "Community based" is correctly used.

Answer for response: Thanks for your comments. We have revised the title of "Collision prediction models for evaluating the safety of cyclists at community-level" into "An empirical tool to evaluate the safety of cyclists: community-based, macro levelcollision prediction models using negative binomial regression". Hope this new title can capture the key points.

Comment 2: The authors highlights that earlier methods cannot capture the impacts of increments bicycling, however in p5 #2 l5 an early study that relates to number of acc per million passing cyclist is mentioned.

Answer for response: Thanks for your comments. There are some earlier methods capturing the impact of increments bicycling, as you mentioned, Jacobson's study (2003). But his model is too simple to capture the other explanatory variables besides bicycling exposure measures. We do not notice that we did mention the earlier methods cannot capture the impacts of increments, if we did, please let us know where.

Comment 3: Section 3.2, In line 10, what does the authors mean with relative independent? In the same section, how the including criteria can avoid over-fitting of the existing conditions (that include noise)

Answer for response: Thanks for your comments. For the first question, "relative independent" means any two variables in the same model should be has no correlation or little correlation in order to minimize the correlation problem in the multivariate regression. For example, transit mode split would be the complement of, and therefore highly correlated with, auto mode split; hence, only one mode split variable (i.e.auto or transit) could be included in any one CPM. Correlation between variables was checked by viewing correlation results in the GenStat software in our case. So the correlation between any two variable in the model should not exceed 20% generally. To avoid such expression confusion, we changed the sentence into "the added variables should have little or no correlation with any other independent variable in the same model in order to minimize the correlation problem". Hope it will be better in understanding.
For the second question, to avoid overfitting the model is always a concern. We have added a sentence to clarify how we avoided this following recommendations in the literature (see McCullagh and Nelder (1989),Sayed and Sawalha (2006), and Lovegrove (2007), wherein they recommend that only variables that cause a significant drop in Scaled Deviance at the 95% level (i.e. SD drop > 3.84) be added into the CPM, and to stop adding more variables if and when no further significant drops in SD are observed.

Comment 4: Section 3.3 presents 2 goodness of fit indicators. There is no reference recommending them. Later in the model results (p11 #2) there are some references that actually do not recommend these indicators. Why the authors considered them?

Answer for response: Thanks for your comments. The 2 goodness of fit indicators have been mentioned in the micro-level CPM (Sawalha&Sayed, 2006) and macro –level CPM development (Lovegrove, 2007). Sorry formissing the references here and now I have clarified this in page 5, line 35-36. The second question is very good.These 2 indicators have limitations in small sample and the sample with many "zeros". Now we are actually exploring much better tests for these small samples. However, as mentioned in the paper, this issue definitely need lots of work to address it in future. Currently, we considered these two indicators because (1) we have not found reliable model fit methods yet; (2) after comparing the prediction results and observations, we found the model results which are tested by these 2 goodness-of-fit tests are not bad. Therefore, we usedthese two "traditional" goodness-of-fit indicators temporarily. And we have clarified this point in the page 14, line 14-17.

Comment 5: If despite the disadvantages of the indicators of goodness of fit, what does the authors means with successfully in section 6 in p12 # 2 l-5.

Answer for response: Thanks for your comments. Still, please see the answer for Comment 4. Develop models "successfully" might be not used correctly as we recognized there are some limitation for the goodness of fit tests and need to update our research in future.We have revised this sentence in page 14, line 41-42.

Comment 6: P10, it is good that the authors avoid regression to the mean problems, however, P10 # 2 l-8 à10, how is it the empirical relationships maximized and the biased minimized???

Answer for response: Thanks for your comments. Regression to the mean is a statistical phenomenon where extreme values of a random variable tend to be followed by less extreme values. RTM problem may lead to a bias or false on the road safety improvement programs such as before and after study or the identification of collision prone locations as the collected data of collision frequency or rate in one location are possibly in the period where higher values or lower values appear. So basically, we use the empirical Bayesian method to minimize this bias. However, the Bayesian method is used only do the CPM applications to reduce RTM and selection bias. This paper is not about CPM applications but about CPM development. So we would like to reduce the regression to the mean bias at the data collection stage for model development. If the period for collision collection is too long (e.g. >> 3 to 6 years), there may be a time trend problem bias as background factors (e.g. socio-demographics, growth) change; if the period is too short (<< 3 years), regression to the mean bias may influence collision data (e.g. extreme values randomly occurring). That's why we recommended 5 years: not too long and too short, reasonable to balance two biases from time trend and regression to the mean. We have clarified it in page 10, line 14-17. Hope this will clarify your confusion.

Comment 7: Some short aspects that can improve the document can be: - Why to use "sustainable" #1 lane 3- # 2 presents the problem earlier studies (i.e. they are reactive) however nothing about the un-reliability of this effort is mentioned before.

Answer for response: Thanks for your comments. Yes, we have extended the aspect about why "sustainable". For the second point, we agree that the micro-level CPMs are reactive and lack proactive ability, but we still think they are reliable when properly applied. We also clarify this point in page 5, paragraph 1.

Comment 8:P4 #1, is there any reference for the affirmation that closes the paragraph. P4 #2 Any reference for "This regression specifically accounts for extra Poisson variation of collisions and has become one of the most popular and reliable modeling techniques for both micro and macro-level CPMs"

Answer for response: Thanks for your comments. Actually, we concluded with this affirmation because many studies used the NB regression method. But his may be not accurate so we rewrote it as "This regression specifically accounts for extra Poisson variation of collisions and is widely used in many studies for both micro and macro-level CPMs." in page 4, line 8-9.The references were provided after the next sentence.

Comment 9: Just above Eq 1, PLN is not previously defined.

Answer for response: Thanks for your comments. It was a type mistake, it should be "NB" instead of "PLN". We have corrected it in Page 4,Line 12.

Comment 10:P5 #1. I am skeptical to the affirmation "macro-level CPMs minimize the road safety risk at an early planning stage and to preclude black spots from occurring at all". Would not be black spots something more a micro level?

Answer for response: Thanks for your comments. We reconsider this sentence and think the previous description is not so accurate and overstates the effort of macro-level CPMs. Here,we would like to emphasise that the application of macro-level CPMs can allow engineers and planners to evaluate the road safety level at an early planning stage, before construction occurs, allowing for design improvements to increase safety levels by precluding the black spot from happening at all. We agree that this is a theoretical hypothesis that our research seeks to validate. Of course, it is impossible to know what would have happened in the absence of these design improvements, but our model predictions can at least be validated. In any case, our overall intent is a proactive engineering approach that will lead to significant reductions in collision frequencies below that achieved to date using reactive techniques. We have rewritten this sentence in page 5, line 8-10.

Comment 12:P11 #0 l1.what does the "effective" mean?

Answer for response: Thanks for your comments. "effective" is not correct here, we changed it to "acceptable" in page 12, line 22, which means these model form are accepted after testing.

Comment 13: Some short editorial comment. P6 #2 l-4, correct Guevara instead of Geuvara

Answer for response: Thanks for your comments. We revised such spell errors.

Reviewer 2 -The paper does not respectthe guidelines for submitted texts.

Comment 1: The title does not have the correct font style and is misplaced, and the legend "3rd internationalconference on road safety..." is heading the document instead. *Page numbering should be centered and in the bottom of the pages, as established. * Lines are not number, as asked for revision.

Answer for response: Thanks for your comments. We have followed the guidelines for submitted texts strictly. And add the line number for revision.

Comment 2: Abstract exceeds 250words.

Answer for response: Thanks for your comments. We reduced the abstract words below 250 words.

Comment 3: Topics on micro and macromodels, and theoretical approaches in probability tools should be betterorganized in Section 2 “Literature Review”, and the creation of two subsectionsfor the development of them should be considered.

Answer for response: Thanks for your comments. We consider a better organization for micro- and macro- level models. According to your suggestion, we two subsections are created for a clear organization.

Comment 4: This text possesses manywriting errors of different kind as missed or misplaced prepositions, articles,or conjunctions, as in “With (the) development of automobile”, “Threeobjectives of this paper are (to):” in Introduction section, or “... there aremany observations (that) are '0'...” in Model Results & Discussion, among others.

Answer for response: We thank you for your patience in reviewing the previous version. As you could tell, English is not the first language for the first author. Hence, we have reviewed this manuscript many (more) times. We hope that we have now caught and corrected all grammar/writing errors. Please let us know what you think!

Comment 5: Syntax errors must bewatched, as in “... geometric design related to VRUs' the safety issue atmicro-level”, in the Introduction

Answer for response: Thanks for your comments. We have revised such errors.

Comment 6: Equations have a lack of presentation, as expressions (1) and (5). Equation numbering is not properlyordered.

Answer for response: Thanks for your comments. We have presented these equations in page 4 line and 18-23 and 29-33. Also we have re-numbered the equations in the right order.

Comment 7: Figures and Tables arealso non congruent with the numbering in the respective text.

Answer for response: Thanks for your comments. We have followed the guideline to give the correct figure and table numbering.

Comment 8:“And” or “&” are usedindistinctly in the references, and sometimes neither of them appeared whereshould be convenient. * Some cited references donot appear in the respective list, as the Census Canada, 2006 and CanMap, 2006

Answer for response: Thanks for your comments. We have re-edited the references and also added the omitted references in the paper.

Comment 9: Discussion on results shouldbe more precise. It is not clear whether the relation among different types ofaccidents by cyclists are sufficiently described by the models presented.

Answer for response: Thanks for your comments.We are unclear about whether you are referring to total versus severe and PDO collision types, or whether you are referring to auto-bike, bike-bike, and bike-pedestrian collision types. If the former, we have added text in page 12, line 5-10 to clarify that our models present Total collisions only, we have not yet pursued other model types the predict Severe or PDO collision types. This would be a step for future research, pending additional bicycle collision data. If on the latter, we have added text in page 12, line 11-12 to clarify that in this paper, we have only used bicycle collision data related to bicycle-auto collisions. Specifically, we have not yetobtained and therefore not yet included data on bicycle-pedestrian, and bicycle- bicycle. Hence, the presented models predict only the total of bicycle-auto collision frequencies. We hope that these combined responses address your concern.

Comment 10: Reference values and parameters should be pointed out. I think this document mustbe revised by authors in order not only to correct all these observations, but in order to raise the level of their research and results.

Answer for response: Thanks for your comments, your constructive comments are truly appreciated and will no doubt help us to raise our level of research and results.By reference values and parameters, we are unclear whether you are referring to model regression results, and/or data summary statistics, or ??. If the former, parameter estimates and goodness of fit statistics for each variable and model are given in Table 3 in detail. If the latter, all summary statistics for data are given in Table 2. As we mentioned, this paper is an early result of our research. Our response is not meant to be arrogant, please advise if we missed the point of your comment 10.

Reviewer 3 - Good paper. Some comments:

Comment 1: Literature review about GLMs might be shortened.

Answer for response: Thanks for your comments, length is always an issue so we have tried to be as concise as possible, yet maintain clarity. Please see if it works for you now.

Comment 2: Options of the software GenStat should be removed, since are not of scientific and general interest.

Answer for response: Thanks for your comments. We have removed the options of GenStat in page 6: the first paragraph in section 3.1 we used.

Comment 3: Outlier exclusion procedure is not very clear.

Answer for response: Thanks for your comments. We have added more details in the part of outlier exclusion procedure in page 8, paragraph 1.

Comment 4: In section 5, first paragraph, it is stated that "it is hard find potential causality". I do not agree with the principle that CPMs identify causality. Really, they identify statistical dependencies.

Answer for response: Thanks for your revisions. We agree with you. Actually, these covariates are not really causality for the collision frequency but the correlated dependent relationship with the collisions. We have revised this point in page 12, line 11-13.